SEEM5020 Algorithms for Big Data
A Review of Probability Concepts and Concentration Bounds

Sibo WANG

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Table of Contents

1 Review of Probability Concepts

2 Concentration Bounds
Table of Contents

1. Review of Probability Concepts

2. Concentration Bounds
Definition 1 (Sample Space)
The **sample space** $\Omega$ is the set of all possible outcomes of an experiment. The outcomes are also called **elementary events**.

Definition 2 (Events)
- An event is a subset of the sample space $\Omega$.
- The event $\Omega$ is the **certain event**, and the event $\emptyset$ is called the **null event**.
- Two events $A$ and $B$ are **disjoint**, a.k.a, mutually exclusive, if $A \cap B = \emptyset$.
- We can treat an elementary event $\omega \in \Omega$ as a set $\{\omega\}$. We will abuse the notation $\omega$ to indicate an elementary event (an event should be a set). Elementary events are disjoint by definition, as we can only see one possible outcome at a time.
Example 1

Experiment: Flip two coins, with each individual flip resulting in a head (H) or a tail (T). We can view the sample space as consisting of the set of all possible outcomes:

\[ \Omega = \{HH, HT, TH, TT\} . \]

Here, \( HH \), \( HT \), \( TH \), and \( TT \) are the elementary events of the experiments. Event \( A = \{HH, HT\} \) and \( B = \{TH, TT\} \) are disjoint.
A **probability distribution** $\mathbb{P}[]$ on a sample space $\Omega$ is a mapping from events in $\Omega$ to $[0, 1]$. Given an event $A$, it indicates the likelihood that at least one elementary element in $A$ occurs. It satisfies the following **probability axioms**.

- **Non-negative.** $\mathbb{P}[A] \geq 0$ for any $A \subseteq \Omega$.
- **Unitarity.** $\mathbb{P}[\Omega] = 1$, i.e., the probability that at least one elementary event in the sample space $\Omega$ occurs is 1.
- **Additivity.** If $A$ and $B$ are two disjoint events, then the probability of their union satisfies that:

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B].$$

The above property can be extended to an arbitrary number of pairwise disjoint events $A_1, A_2, \ldots, A_i, \ldots$. 
Properties of Discrete Probability

**Fact:** Let $A$ and $B$ be two events, then

$$P[A \cup B] = P[A] + P[B] - P[A \cap B].$$

**Theorem 1 (Union Bound)**

**Main result:** Let $A_1, A_2, \ldots, A_k$ be events. Then, we have that:

$$P[A_1 \cup A_2 \cup \cdots \cup A_k] \leq P[A_1] + P[A_2] + \cdots + P[A_k].$$

**Definition 3 (Independence)**

Two events $A$ and $B$ are independent if and only if $P[A \cap B] = P[A] \cdot P[B]$. More generally, events $A_1, A_2, \cdots, A_k$ are mutually independent if and only if, for any subset $I \in \{1, 2, \cdots, k\}$,

$$P[\cap_{i \in I} A_i] = \prod_{i \in I} P[A_i].$$
Exercise

- Assume that we roll two fair six-sided dice. What is the probability that either the sum is even or one of the dice is odd?

- Assume that in a data center we maintain 10 machines and each machine fails with probability 0.00001. Then, what is the probability that all these 10 machines do not fail in ten days?

- Assume that in a data center we maintain 10 machines and each machine fails with probability 0.00001. The failure of each machine is independent of each other, and each day the failure of a machine is independent of the previous days. Then, what is the probability (lower bound) that all these 10 machines do not fail in ten days?
Definition 4 (Conditional Probability)

The **conditional probability** that event $A$ occurs given that the event $B$ occurs is:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}.$$ 

Example 2

Given a six-sided dice, the possible outcomes are $\{1, 2, 3, 4, 5, 6\}$. Assume that it is fair dice so that each number can be drawn with equal probability. Define event $A$ as the outcome that the numbers are odd, i.e., $A = \{1, 3, 5\}$. Further define $B$ as the event that the outcome is no larger than 4, i.e., $B = \{1, 2, 3, 4\}$. What is the probability of event $A$ conditioned on $B$?
Exercise

- You roll two fair six-sided dice. What is the probability that the sum of the two dice is 7, given that at least one of the dice shows an even number?

- You meet a family with two children. Assume that the chance to have a boy or a girl is equal. Given that at least one of them is a boy, what is the probability that they have two boys?
Lemma 1 (Multiplication Rule)

Let $A_1, A_2, \cdots, A_k$ be $k$ events. We have that:

$$\mathbb{P}[\cap_{i=1}^k A_i] = \mathbb{P}[A_1] \cdot \mathbb{P}[A_2|A_1] \cdot \mathbb{P}[A_3|A_1 \cap A_2] \cdots \mathbb{P}[A_k|A_1 \cap A_2 \cap A_3 \cdots \cap A_{k-1}]$$

Theorem 2 (Law of Total Probability)

Let $A_1, A_2, \cdots, A_k$ be multiple disjoint events in sample space $\Omega$ and $\bigcup_{i=1}^k A_i = \Omega$. Given an event $B$, then we have that:

$$\mathbb{P}[B] = \sum_{i=1}^k \mathbb{P}[B \cap A_i] = \sum_{i=1}^k \mathbb{P}[B|A_i] \mathbb{P}[A_i].$$
Example 3
Three cards are drawn from an ordinary 52-card deck without replacement. What is the probability that none of them is heart?

Example 4
Assume that we have three bags, where each bag includes 8 balls. The three bags include 3, 5, and 2 green balls (the remaining balls are red), respectively. If we first randomly sample a bag and then randomly sample a ball, what is the probability that we sample a green ball?
Theorem 3 (Bayes’ Theorem)

Let \( A_1, A_2, \ldots, A_k \) be multiple disjoint events in the sample space \( \Omega \) and \( \bigcup_{i=1}^{k} A_i = \Omega \). Given an event \( B \), then we have that:

\[
P[A_j | B] = \frac{P[A_j \cap B]}{P[B]} = \frac{P[B | A_j] \cdot P[A_j]}{\sum_{i=1}^{k} P[B | A_i] \cdot P[A_i]}
\]

Example 5

A blood test is used to detect a banned drug, where the test produces 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose 0.5% of people are drug users. What is the chance that a randomly selected person who tests positive is a drug user?
Theorem 3 (Bayes’ Theorem)

Let $A_1, A_2, \cdots, A_k$ be multiple disjoint events in the sample space $\Omega$ and $\bigcup_{i=1}^{k} A_i = \Omega$. Given an event $B$, then we have that:

$$P[A_j | B] = \frac{P[A_j \cap B]}{P[B]} = \frac{P[B | A_j] \cdot P[A_j]}{\sum_{i=1}^{k} P[B | A_i] \cdot P[A_i]}.$$ 

Example 5

A blood test is used to detect a banned drug, where the test produces 99% true positive results for drug users and 99% true negative results for non-drug users. Suppose 0.5% of people are drug users. What is the chance that a randomly selected person who tests positive is a drug user?

$$P[A|+] = \frac{P[+|A] \cdot P[A]}{P[+|A] \cdot Pr[A] + P[+|A^c] \cdot Pr[A^c]} = \frac{0.99 \times 0.005}{0.99 \times 0.005 + 0.01 \times 0.995}.$$
Let $X$ be a discrete random variable. Let $a$ be a real value. The event ”$X = a$” includes all basic events in the sample space $\Omega$ so that $X$ has the value of $a$. That is, $X = a$ represents the set $\{s \in \Omega | X(s) = a\}$ and

$$p_X(a) = \mathbb{P}[X = a] = \sum_{s \in \Omega, X(s) = a} \mathbb{P}[s].$$

### Example 6

Let $X$ be the sum of two six-sided dice rolls. Then the event $X = 4$ corresponds to the set of basic events $\{(1, 3), (2, 2), (3, 1)\}$. Hence

$$\mathbb{P}[X = 4] = \frac{3}{36} = \frac{1}{12}.$$
**Definition 5 (Expectation)**

The expectation of a discrete random variable $X$, denoted by $\mathbb{E}[X]$ is

$$\mathbb{E}[X] = \sum_i i \cdot P[X = i],$$

where the summation is over all possible values of $X$. Notice that the expectation might be unbounded if $\sum_i |i| \cdot P[X = i]$ does not converge.

**Theorem 4 (Linearity of Expectations)**

For any finite collection of discrete random variables $X_1, X_2, \ldots, X_k$ with bounded expectations,

$$\mathbb{E}\left[\sum_{i=1}^k X_i\right] = \sum_{i=1}^k \mathbb{E}[X_i].$$
Example 7

Given $n$ balls and $n$ bins, we throw each ball independently to the $n$ bins. Assume that each ball has an equal chance of falling into each of these $n$ bins. What is the expected number of empty bins?

Let $X_i$ be the random variable such that it is 1 if the $i$-th bin is empty and otherwise is 0. For $E[X_i]$, we have that:

$$E[X_i] = 1 \times P[X_i = 1] + 0 \times P[X_i = 0] = \frac{n-1}{n}$$

After throwing $n$ balls, the total number of empty bins is $\sum_{i=1}^{n} X_i$. The expected number of empty bins is hence:

$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

(By linearity of expectation)

Then, the expected number of bins is $n \cdot \left( \frac{n-1}{n} \right)$. 

Sibo WANG (CUHK)
Example 7

Given $n$ balls and $n$ bins, we throw each ball independently to the $n$ bins. Assume that each ball has an equal chance of falling into each of these $n$ bins. What is the expected number of empty bins?

Let $X_i$ be the random variable such that it is 1 if the $i$-th bin is empty and otherwise is 0. For $\mathbb{E}[X_i]$, we have that:

$$\mathbb{E}[X_i] = 1 \times P[X_i = 1] + 0 \times P[X_i = 0]) = 1 \times \left(\frac{n-1}{n}\right)^n + 0$$

After throwing $n$ bins, the total number of empty bins is $\sum_{i=1}^{n} X_i$. The expected number of empty bins is hence:

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]. \text{ (By linearity of expectation)}$$

Then, the expected number of bins is $n \cdot \left(\frac{n-1}{n}\right)^n$. 

Exercise

- **Coupon Collector Problem.** Suppose that each cereal box contains one of \( n \) different coupons. Once you obtain one of every type of coupon, you can get a prize. Assume that the coupon in each box is chosen independently and uniformly at random from the \( n \) possibilities and that you do not collaborate with others to collect coupons. How many boxes of cereal must you buy before you obtain at least one of every type of coupon?

- **Analysis of QuickSort.** Assume that the elements in the array are distinct and Quicksort randomly selects the pivot. Given the pivot, we compare all elements with the pivot and then divide the array into two subsets: the elements smaller than the pivot and the elements larger than the pivot. The Quicksort then solves these two sub-problems recursively. Prove that the expected number of comparisons is \( O(n \log n) \).
Assume that initially, we have $n$ fair coins. We flip the coin in rounds. In the first round, we flip all of these $n$ coins (once for each coin); if a coin shows the tail, it is removed. In the next round (Round 2), we have the coins that are left (which shows the head in the last round) and flip each of these coins. Again, if a coin shows the tail, it is removed. Then, we turn to Round 3, Round 4, etc. until there is no coin left. Answer the following questions.

- **Question 1:** What is the expected number of coins left at the end of round 3?
- **Question 2:** What is the expected number of coin tosses we have flipped in total?
- **(Optional) Question 3:** Prove that the expected number of rounds we have played is bounded by $\log n + 2$. Useful inequality: $(1 + x)^r \geq 1 + x \cdot r$ for any $x \geq -1$ and $r > 1$. 

Sibo WANG (CUHK)  
SEEM5020 Algorithms for Big Data  
Fall 2023
Theorem 5 (Jensen’s Inequality)

If $f$ is a convex function, i.e., $f$ is twice differentiable and $f''(x) \geq 0$, then

$$\mathbb{E}[f(x)] \geq f(\mathbb{E}[X]).$$

Example 8

Let $X$ be a positive random variable. Then, it is easy to verify that

$$\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2.$$

Think:

More generally, please compare $\mathbb{E}[X^a]$ and $(\mathbb{E}[X])^a$ for all $a \in \mathbb{R}$ given that $X$ is a positive random variable.
Conditional Expectation

Definition 6 (Conditional Expectation)
Given two random variables $Y$ and $Z$, we have that

$$E[Y \mid Z = z] = \sum_y y \cdot P[Y = y \mid Z = z]$$

Lemma 2 (Expectation and Conditional Expectation)
For any random variables $X$ and $Y$, $E[X] = \sum_y P[Y = y] \cdot E[X \mid Y = y]$.  

Lemma 3 (Linearity of Conditional Expectation)
For finite discrete random variables $X_1, X_2, \ldots, X_k$ with finite expectations and for any random variable $Y$, we have that

$$E[\sum_{i=1}^k X_i \mid Y = y] = \sum_{i=1}^k E[X_i \mid Y = y].$$
Examples: Conditional Expectation

**Example 10**
Roll two fair six-sided dice. Let $X$ be the sum of the two dice and $Y$ be the number of the first dice. What is the expectation of $X$ given that $Y$ is 2?

**Example 11**
Let $X$ be the sum of 2 six-sided dice rolls, $Y$ be the result of the 1st dice.

\[
E[X|Y = 1] = 2 \cdot P[X = 2|Y = 1] + 3 \cdot P[X = 3|Y = 1] \cdots = 4.5
\]

\[
E[X|Y = 2] = 3 \cdot P[X = 3|Y = 1] + 4 \cdot P[X = 3|Y = 2] \cdots = 5.5
\]

\[
E[X] = \sum_{y \in \{1,2\cdots 6\}} P[Y = y] \cdot E[X|Y = y]
\]

\[
= \frac{1}{6} \times 4.5 + \cdots \frac{1}{6} \times 9.5 = 7
\]
Definiton 7

Given two random variables $Y$ and $Z$, the expression $E[Y|Z]$ is a random variable $f(Z)$ whose value is $E[Y|Z = z]$ when $Z = z$.

Theorem 6 (Law of Total Expectation)

Given random variables $Y$ and $Z$, we have that:

$$E[Y] = E[E[Y|Z]]$$
Exercise

- You are playing a dice game with the following rules: You roll a fair six-sided die, and if it shows an even number, you win twice the number rolled; if it shows an odd number, you lose three times the number rolled. What is the conditional expectation of your winnings given that you rolled an even number?

- What is the expected gain/loss by playing the above game?

- Assume that $N$ students attend a driving class and we want to compute the expected number $X$ of students that can pass the test. Assume that we now only have the expected number of these $N$ students that can pass the test under different ages. How can you use the total law of expectation to compute $\mathbb{E}[X]$?
### Definition 8 (Variance)
The variance of a random variable $X$ is defined as

$$\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2.$$ 

The standard deviation $\sigma[X] = \sqrt{\text{Var}[X]}$.

### Definition 9 (Moment)
The $k$-th moment of a random variable $X$ is $\mathbb{E}[X^k]$.

### Definition 10 (Covariance)
The covariance of two random variables $X$ and $Y$ is:

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])].$$
Lemma 4 (Expectation of Independent Variables)

If $X$ and $Y$ are independent random variables, then we have:

$$\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

Theorem 7

If $X$ and $Y$ are independent random variables, then $\text{Cov}(X, Y) = 0$ and $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$. More generally, for $k$ mutually independent random variables, we have that:

$$\text{Var}\left[\sum_{i=1}^{k} X_i\right] = \sum_{i=1}^{k} \text{Var}[X_i].$$
Exercise

- Given a constant $a$, show that $\text{Var}[a \cdot X] = a^2 \text{Var}[X]$.

- Given a random variable $X$ and another random variable
  $X_{\text{avg}} = \frac{X_1 + X_2 + \cdots + X_k}{k}$ where $X_1, X_2, \cdots, X_k$ are independent random variables sampled from the same distribution. What is the relationship between $\text{Var}[X]$ and $\text{Var}[X_{\text{avg}}]$?

- Binomial distribution: The binomial distribution with parameters $n$ and $p$ is the discrete probability distribution of the number of successes in a sequence of $n$ independent experiments.
  Example: Given a coin that shows the head with probability $p$ and the tail with probability $1 - p$, what is the expected number of times that the coin shows the heads (A single trial is the Bernoulli trial)? Show that the expectation of binomial distribution is $n \cdot p$ and the variance is $n \cdot p \cdot (1 - p)$. 
Continuous Random Variables

**Definition 11 (Probability density function)**

A random variable $X$ is continuous if its probability law can be described with a non-negative function $f_X$, called the probability density function (PDF) of $X$ so that for every subset $B$ of the real line, the following holds:

$$P[X \in B] = \int_B f_X(x) \, dx$$

The probability of $X$ falling into an interval $[a, b]$ is:

$$P[a \leq X \leq b] = \int_a^b f_X(x) \, dx$$

**Definition 12 (Expectation)**

The expectation $E[X]$ of a continuous random variable $X$ is:

$$E[X] = \int_{-\infty}^{+\infty} xf_X(x) \, dx$$
Example 12

Let $r$ be a positive constant. Let $Y$ be a random variable sampled from $[0, r]$ uniformly at random. Define $f(x) = \lceil (x + Y)/r \rceil$. Given two variables $x_1$ and $x_2$ ($0 < x_1 \leq x_2$), what is the probability that $f(x_1) = f(x_2)$ under the random choice of $Y$?

Example 13

Let $X$ have a range of $[0, 2]$ with a density function $f_X(x) = \frac{3x^2}{8}$ in the range. Compute $\mathbb{E}[X]$. 
Table of Contents

1. Review of Probability Concepts

2. Concentration Bounds
Markov’s and Chebyshev’s Inequality

Theorem 8 (Markov’s Inequality)
Let \( X \) be a random variable that only has non-negative values. Then, for all \( a > 0 \),

\[
P[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.
\]

Theorem 9 (Chebyshev’s Inequality)
For any \( a > 0 \), we have that:

\[
P[|X - \mathbb{E}[X]| \geq a] \leq \frac{\text{Var}[X]}{a^2}.
\]
Exercise

1. Assume that we flip $n$ fair coins. What is the probability that more than $\frac{3n}{4}$ of the coins show the head by using Markov’s inequality and Chebyshev’s inequality respectively?

2. Assume that we know QuickSort runs with $n \log n$ comparisons in expectation. What is the probability that the QuickSort makes $3n \log n$ comparisons? How can we use this probability to avoid the worst-case $O(n^2)$ time complexity with high probability by making $O(n \log^2 n)$ comparisons?
Definition 13 (Moment Generating Function (MGF))

The moment generating function $M_X(t)$ of a random variable $X$ is

$$M_X(t) = \mathbb{E}[e^{tX}].$$

Theorem 10 (Properties of MGF)

- Let $X$ be a random variable with moment generating function $M_X(t)$. Under the assumption that exchanging the expectation and differentiation operands is legitimate, for all $n > 1$, we have that

$$\mathbb{E}[X^n] = M_X^{(n)}(0),$$

where $M_X^{(n)}(0)$ is the $n$-th derivative of $M_X(t)$ evaluated at $t = 0$.

- If $X$ and $Y$ are independent random variables, then

$$M_{X+Y}(t) = M_X(t) \cdot M_Y(t).$$
Chernoff Bound

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent $\{0, 1\}$ random variables. Let $p_i$ be the probability that $X_i$ is 1. Let $X = \sum_{i=1}^{n} X_i$. Let

$$
\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p_i.
$$

For $X_i$, the MGF of $X_i$ satisfies that:

$$
M_{X_i}(t) = \mathbb{E}[e^{tX_i}] = p_i \cdot e^t + (1 - p_i) \cdot e^0
$$

$$
= 1 + p_i(e^t - 1) \leq e^{p_i(e^t-1)},
$$

where the last inequality is due to the fact that for any $y$, $1 + y \leq e^y$. 
Chernoff Bound (Upper Tail)

Consider random variable $X$ and its MGF

$$M_X(t) = \prod_{i=1}^{n} M_{X_i}(t) \text{ (Applying Theorem 10)}$$

$$\leq \prod_{i=1}^{n} e^{p_i(e^t-1)} = e^{\mu(e^t-1)}.$$

Theorem 11 (Chernoff bound (Upper Tail))

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent \{0, 1\} random variables. Let $p_i$ be the probability that $X_i$ is 1. Let $X = \sum_{i=1}^{n} X_i$, $\mu = \sum_{i=1}^{n} p_i$.

- For any $\delta > 0$,

$$\mathbb{P}[X \geq (1 + \delta)\mu] \leq \left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^\mu;$$

- For $0 < \delta \leq 1$,

$$\mathbb{P}[X \geq (1 + \delta)\mu] \leq e^{-\mu\delta^2/3}.$$
Proving the Upper Tail of the Chernoff Bound

\[ \mathbb{P}[X \geq (1 + \delta) \cdot \mu] = \mathbb{P}[e^{tX} \geq e^{t \cdot (1 + \delta) \cdot \mu}] \leq \frac{\mathbb{E}[e^{tX}]}{e^{t \cdot (1 + \delta) \cdot \mu}} \leq \frac{e^{\mu \cdot (e^t - 1)}}{e^{t \cdot (1 + \delta) \cdot \mu}}. \]

The above inequality holds for an arbitrary \( t \). Hence, we can choose the \( t \) that minimizes \( \frac{e^{\mu \cdot (e^t - 1)}}{e^{t \cdot (1 + \delta) \cdot \mu}} \). Please prove by yourself that when \( t = \ln(1 + \delta) \), we have the smallest value. Putting \( t = \ln(1 + \delta) \) to above inequality:

\[ \mathbb{P}[X \geq (1 + \delta) \cdot \mu] \leq \frac{e^{\mu \cdot \delta}}{e^{\ln(1 + \delta) \cdot (1 + \delta) \cdot \mu}} = \left( \frac{e^\delta}{(1 + \delta)^{1 + \delta}} \right)^\mu. \]

To prove the second one. We only need to prove that:

\[ \left( \frac{e^\delta}{(1 + \delta)^{1 + \delta}} \right)^\mu \leq e^{-\delta^2 \mu / 3}. \]

Think: How to prove this inequality?
Theorem 12 (Chernoff Bound (Lower Tail))

Let $X_1, X_2, \cdots, X_n$ be a sequence of independent $\{0, 1\}$ random variables. Let $p_i$ be the probability that $X_i$ is 1. Let $X = \sum_{i=1}^{n} X_i$, $\mu = \sum_{i=1}^{n} p_i$.

For any $1 > \delta > 0$, we have that:

$$
P[X \leq (1 - \delta)\mu] \leq \left( \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right)^\mu;
$$

$$
P[X \leq (1 - \delta)\mu] \leq e^{-\mu \delta^2 / 2}.
$$

The proof is similar to the proof of the upper bound and is left as self-exercise.

**Remark:** More concentration bounds will be discussed in later lectures if used in the corresponding topics.
Exercise

- Assume that we flip $n$ fair coins. What is the probability that more than $\frac{3n}{4}$ of the coins show the head by using the Chernoff bound?

- In an election with two candidates using paper ballots, each vote is independently misrecorded with probability $p = 0.02$. Use the Chernoff bound to give an upper bound on the probability that more than 4% of the votes are misrecorded in an election of 1,000,000 ballots.