BUSI 573: Stochastic Models in Operations Management **Spring 2026**

Lecture 3: Concentration Inequalities

Lecturer: Süleyman Kerimov *Date: January 27, 2026*

Disclaimer: *These notes are primarily adapted from expositional texts, including work by Michel Goemans and Anupam Gupta. These notes are not meant to be complete or fully rigorous; some proofs are not given, incomplete, or only outlined, as they are discussed in class.*

3.1 Classical Bounds

We are interested in concentration inequalities, which help us understand how close random variables are to their expected values (or to other values). So far, we have studied Markov's and Chebyshev's inequalities, which apply to a single random variable, and the law of large numbers, which characterizes the behavior of sums of many random variables.

3.1.1 Chernoff Bound

The generic Chernoff bound is inspired by applying Markov's inequality to the exponential of a random variable. Note that for any random variable $X \geq 0$ and $a, t > 0$, we have $\mathbb{P}(X \geq a) = \mathbb{P}(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}}$, since e^{tX} is monotonically increasing. If $S_n = \sum_{k=1}^n X_k$, then we also have $\mathbb{P}(S_n \geq a) \leq e^{-ta}\mathbb{E}[\prod_{i=1}^n e^{tX_i}]$, which looks useful when the random variables are independent. One gets useful (sometimes tight) bounds when optimizing the right-hand side over t.

Example 3.1. *Let $(X_i)_{i=1}^n$ be a sequence of i.i.d. Bernoulli random variables with $\mathbb{P}(X_k = 1) = p$, and let $X = \sum_{i=1}^n X_i$. First, note that*

$$\begin{aligned} \mathbb{E}\left[e^{tX_i}\right] &= pe^t + (1-p)e^0 \\ &= 1 + p(e^t - 1) \\ &\leq e^{p(e^t-1)}, \end{aligned} \tag{3.1}$$

where we used $1 + x \leq e^x$ with $x = p(e^t - 1)$.

$$\begin{aligned}\mathbb{P}(X \geq a) &\leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{at}} \\ &\leq e^{-at}\,\mathbb{E}\left[e^{t\sum_i X_i}\right] \\ &\leq e^{-at}\,\mathbb{E}\left[e^{tX_1}\right]\mathbb{E}\left[e^{tX_2}\right]\cdots\mathbb{E}\left[e^{tX_n}\right] \\ &\leq e^{-at}\,e^{\sum_{i=1}^n p(e^t-1)}, \qquad (3.2)\end{aligned}$$

where (3.2) *follows from* (3.1). *Now taking* $a = (1+\delta)\mathbb{E}[X] = (1+\delta)np$ *and* $t = \log(1+\delta)$, *we get*

$$\begin{aligned}\mathbb{P}(X \geq (1+\delta)np) &\leq \frac{e^{np(1+\delta-1)}}{(1+\delta)^{(1+\delta)np}} \\ &\leq \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{np}. \qquad (7)\end{aligned}$$

Using arguments analogous to the ones in Example 3.1, we can get the following two general results.

Theorem 3.2 (Chernoff bound for Bernoulli trials)**.** *Let* $X = \sum_{i=1}^n X_i$, *where* $X_i = 1$ *with probability* p_i *and* $X_i = 0$ *with probability* $1 - p_i$, *and suppose that* $X_1, \ldots, X_n$ *are independent. Let* $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. *Then we have*

(i) ***Upper Tail:***

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2}{2+\delta}\,\mu\right), \qquad \textit{for all } \delta > 0.$$

(ii) ***Lower Tail:***

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2}\right), \qquad \textit{for all } 0 < \delta < 1.$$

Combining both bounds for $\delta \in (0,1)$ *yields*

$$\mathbb{P}(|X-\mu| \geq \delta\mu) \leq 2\exp\Big(-\frac{\mu\delta^2}{3}\Big).$$

Theorem 3.3 (Chernoff bound for bounded random variables)**.** *Let* $X_1, X_2, \ldots, X_n$ *be random variables such that*

$$a \leq X_i \leq b \quad \textit{for all } i.$$

Let

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

Then, for all $\delta > 0$*:*

(i) ***Upper Tail:***

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{2\delta^2\mu^2}{n(b-a)^2}\right).$$

(ii) ***Lower Tail:***

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{2\delta^2\mu^2}{n(b-a)^2}\right).$$

Example 3.4. *Suppose that we are repeatedly tossing a fair coin. Let S_n be the number of heads we observe from the first n tosses. Per Chebyshev's inequality, we get $\mathbb{P}(|S_n/n - 1/2| \geq \epsilon) \leq 1/4n\epsilon^2$. For example, when $\epsilon = 1/2$, we get $\mathbb{P}(|S_n/n - 1/2| \geq 1/2) \leq 4/n$. If we use Chernoff bound instead, per Theorem 3.2, we have*

$$\mathbb{P}\Big(\Big|S_n - \frac{n}{2}\Big| \geq \delta\frac{n}{2}\Big) \leq 2\exp\Big(-\frac{n\delta^2}{6}\Big).$$

When $\delta = 1/2$, we obtain $\mathbb{P}(|S_n - 1/2| \geq 1/4) \leq 2\exp(-n/24)$, which is a much better bound.

3.1.2 Hoeffding's Inequality

Theorem 3.5 (Hoeffding's Inequality)**.** *Let $(X_i)_{i=1}^n$ be a sequence of independent bounded random variables with $a_i \leq X_i \leq b_i$ for all $i = 1, ..., n$ with probability 1. Let $S_n = \sum_{i=1}^n X_i$. Then for any $t > 0$, we have*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof of Theorem 3.5. We start with a useful result (Hoeffding's Lemma), which you will prove in your homework: if Y is a random variable with $\mathbb{E}[Y] = 0$ and $a \leq Y \leq b$, then we have $\mathbb{E}[\exp(tY)] \leq \exp(\frac{t^2(b-a)^2}{8})$. The proof simply uses the convexity of the exponential function, i.e., $e^{ty} \leq \frac{y-a}{b-a}e^{tb} + \frac{b-y}{b-a}e^{ta}$ for all $a \leq y \leq b$.

Now, note that by Markov's inequality, for all $s > 0$ we have

$$\Pr[S_n - \mathbb{E}[S_n] \geq t] = \Pr\Big[e^{s(S_n - \mathbb{E}[S_n])} \geq e^{st}\Big] \leq \frac{\mathbb{E}\big[e^{s(S_n - \mathbb{E}[S_n])}\big]}{e^{st}} \tag{3.3}$$

Letting $Y_i = X_i - \mathbb{E}[X_i]$, we get

$$\mathbb{E}\Big[e^{s(S_n - \mathbb{E}[S_n])}\Big] = \mathbb{E}\Big[e^{s\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}\Big] = \mathbb{E}\left[\prod_{i=1}^n e^{s(X_i - \mathbb{E}[X_i])}\right] = \mathbb{E}\left[\prod_{i=1}^n e^{sY_i}\right]$$

Per Hoeffding's Lemma, we get

$$\mathbb{E}\left[\prod_{i=1}^n e^{sY_i}\right] \leq \prod_{i=1}^n e^{\frac{s^2(b_i - a_i)^2}{8}} \tag{3.4}$$

Using the bound (3.4) in (3.3), we get

$$\Pr[S_n - \mathbb{E}[S_n] \geq t] \leq \frac{\exp(\frac{s^2}{8}\sum_{i=1}^n (b_i - a_i)^2)}{\exp(st)}. \tag{3.5}$$

Now it is time to optimize the right-hand side. It is easy to see that $s = \frac{4t}{\sum_{i=1}^n (b_i - a_i)^2}$ is the minimizer of the right-hand side. Hence, we have

$$\Pr[S_n - \mathbb{E}[S_n] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{3.6}$$

In a similar fashion, one can prove that for all $t > 0$

$$\Pr[S_n - \mathbb{E}[S_n] \leq -t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{17}$$

Combining both inequalities, we get

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

■

3.2 Martingale Inequalities

The family of random variables $\{X(t) : t \in T\}$ is called a stochastic process, where the parameter t is interpreted as time, and $X(t)$ is interpreted as the state of the process at time t. Throughout the semester, we will study numerous stochastic processes to model queueing systems, dynamic matching markets, online resource allocation settings, etc. We finish this lecture with an important family of stochastic processes: martingales.

Definition 3.6. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Then a filtration on the probability space is an increasing family of sub-σ-fields $(\mathcal{F}_n)_{n\geq 0}$ of $\mathcal{F}$ such that $\mathcal{F}_n \subset \mathcal{F}_{n+1} \subset \mathcal{F}$ for all $n \geq 1$. We further say that a stochastic process $X = (X_n)_{n\geq 1}$ is adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$ if X_n is $\mathcal{F}_n-$measurable.*

What are we trying to achieve via filtration here? If you consider the parameter n as time, then intuitively you can interpret $\mathcal{F}_n$ as all historical information that is available to us up to (and including) time n. In other words, the value of X_n only depends on what has already happened up to time n. The sigma fields are increasing over time, because we do not forget the history.

Definition 3.7. *Let $(\mathcal{F}_n)_{n\geq 1}$ be a filtration, and X is adapted to the filtration. Assume that $\mathbb{E}[|X_n|] < \infty$ for all $n \geq 1$. Then,*

1. *X is called a martingale if $\mathbb{E}[X_n|\mathcal{F}_{n-1}] = X_{n-1}$ a.s. for all $n \geq 2$.*

2. *X is called a supermartingale if $\mathbb{E}[X_n|\mathcal{F}_{n-1}] \leq X_{n-1}$ a.s. for all $n \geq 2$.*

3. *X is called a submartingale if $\mathbb{E}[X_n|\mathcal{F}_{n-1}] \geq X_{n-1}$ a.s. for all $n \geq 2$.*

If you are gambling in Las Vegas, then you are playing supermartingale games, while the casino is playing submartingale games. This course is a fair game so we have a martingale. Note that in Definition 3.7[1] by the tower property, $\mathbb{E}[\mathbb{E}[X_n|\mathcal{F}_{n-1}]] = \mathbb{E}[X_n] = \mathbb{E}[X_{n-1}]$, and by recursively we get $\mathbb{E}[X_n] = \mathbb{E}[X_1]$. Similarly, one can also show that $\mathbb{E}[X_{n+m}|\mathcal{F}_n] = X_n$ for any $m \geq 1$.

Example 3.8. *Let $X_1, X_2, ...$ be a sequence of independent random variables with $\mathbb{E}[|X_n|] < \infty$ for all $n \geq 1$, and $\mathbb{E}[X_n] = 0$ for all $n \geq 1$. Let $S_n = \sum_{i=1}^n X_i$ and $\mathcal{F}_n = \sigma(X_1, X_2, ..., X_n)$. Then almost surely we have $\mathbb{E}[S_n|\mathcal{F}_{n-1}] = \mathbb{E}[X_n|\mathcal{F}_{n-1}] + \mathbb{E}[S_{n-1}|\mathcal{F}_{n-1}] = \mathbb{E}[X_n] + S_{n-1} = S_{n-1}$. Thus, S_n is a martingale.*

Example 3.9. *Let $X_1, X_2, ...$ be a sequence of independent random variables with $\mathbb{E}[X_n] = 1$ for all $n \geq 1$. Let $M_0 = 1$, $\mathcal{F}_0 = \{\emptyset, \Omega\}$, $M_n = \prod_{k=1}^n X_k$, and $\mathcal{F}_n = \sigma(X_1, X_2, ..., X_n)$. Then $M = (M_n)_{n\geq 0}$ is a martingale, since $\mathbb{E}[M_n|\mathcal{F}_{n-1}] = \mathbb{E}[M_{n-1}X_n|\mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}[X_n|\mathcal{F}_{n-1}] = M_{n-1}\mathbb{E}[X_n] = M_{n-1}$.*

Example 3.10. *Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables with $\mathbb{E}[X_1] = 0$ and $Var(X_1) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$ and define $Z_n := S_n^2 - n\sigma^2$ for $n \geq 1$. Let $\mathcal{F}_n = \sigma(X_1, X_2, ..., X_n)$. Then Z_n is a martingale, since*

$$\begin{aligned}
\mathbb{E}\big[S_{n+1}^2 \mid \mathcal{F}_n\big] &= \mathbb{E}\big[(S_n + X_{n+1})^2 \mid \mathcal{F}_n\big] \\
&= \mathbb{E}\big[S_n^2 \mid \mathcal{F}_n\big] + \mathbb{E}[2S_nX_{n+1} \mid \mathcal{F}_n] + \mathbb{E}\big[X_{n+1}^2 \mid \mathcal{F}_n\big] \\
&= S_n^2 + 2S_n\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] + \mathbb{E}\big[X_{n+1}^2\big] \\
&= S_n^2 + 2S_n\mathbb{E}[X_{n+1}] + \sigma^2 \\
&= S_n^2 + \sigma^2.
\end{aligned}$$

Definition 3.11. *A map $T : \Omega \to \mathbb{Z}_{\geq 0} \cup \{\infty\}$ is called a stopping time if $\{T \leq n\} = \{\omega : T(w) \leq n\} \in \mathcal{F}_n$ for all $n \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$.*

It is a simple exercise to show that in Definition 3.11, $\{T \leq n\} \in \mathcal{F}_n$ is equivalent to $\{T = n\} \in \mathcal{F}_n$ or $\{T \geq n\} \in \mathcal{F}_{n-1}$. Intuitively, whether the process will stop at time n according to your stopping time T depends only on the history up to (and including) time n.

Example 3.12. *Given a sequence of independent random variables $(X_n)_{n\geq 1}$, the first time X_n hits an arbitrary set A, i.e., $T_A = \min\{n : X_n \in A\}$, is clearly a stopping time. But the last time that X_n visits an arbitrary set A, i.e., $T_A = \max\{n : X_n \in A\}$, or, the time when X_n reaches its maximum, i.e., $T = \min\{n : X_n = \max_{k\geq 1} X_k\}$, are not stopping times.*

Theorem 3.13. *Let X be a martingale, and T be a stopping time. Then $X_{T\wedge n}$, where $T \wedge n = \min\{T, n\}$, is a martingale.*

Proof of Theorem 3.13. It is easy to verify that $X_{T\wedge n} = X_{T\wedge(n-1)} + \mathbf{1}_{\{T\geq n\}}(X_n - X_{n-1})$. Thus,

$$\begin{aligned}
\mathbb{E}[X_{T\wedge n} \mid \mathcal{F}_{n-1}] &= \mathbb{E}\big[X_{T\wedge(n-1)} \mid \mathcal{F}_{n-1}\big] + \mathbb{E}\big[\mathbf{1}_{\{T\geq n\}}(X_n - X_{n-1}) \mid \mathcal{F}_{n-1}\big] \\
&= X_{T\wedge(n-1)} + \mathbf{1}_{\{T\geq n\}}\,\mathbb{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] \\
&= X_{T\wedge(n-1)}.
\end{aligned}$$

■

Theorem 3.14 (Martingale Stopping Theorem)**.** *Let $(X_n)_{n\geq 1}$ be a martingale adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$, and suppose that T is a stopping time for this filtration. Then*

$$\mathbb{E}[X_T] = \mathbb{E}[X_1],$$

if any of the following three sufficient conditions hold:

1. *T is bounded, i.e., there exists a constant C such that $T(w) \leq C$ for all $w \in \Omega$;*
2. *X_n is bounded for all n and $\mathbb{P}(T < \infty) = 1$;*
3. *$\mathbb{E}[T] < \infty$ and X has bounded increments, i.e., there exists $M < \infty$ such that for all $n \geq 1$,*

$$\mathbb{E}[\,|X_{n+1} - X_n||\,\mathcal{F}_n] < M.$$

In the homework, you will prove the following result, which is a corollary of Theorem 3.14.

Theorem 3.15 (Wald's Identity)**.** *Let $X_1, X_2, ...$ be a sequence of independent and identically distributed random variables, and let T be a stopping time for the filtration $\mathcal{F}_n = \sigma(X_1, X_2, ..., X_n)$ for all $n \geq 1$. If $\mathbb{E}[T] < \infty$, then*

$$\mathbb{E}[\sum_{i=1}^{T} X_i] = \mathbb{E}[X_1] \cdot \mathbb{E}[T].$$

Example 3.16 (Ballot counting problem)**.** *Assume that we have two candidates A and B, and let N_A and N_B be the number of votes for each of the candidates A and B, respectively, with $N_A + N_B = N$. Assume that $N_A > N_B$. We start counting votes one by one in a uniformly random ordering. We want to find the probability that candidate A is always ahead when counting votes. Let Y_k be the difference between the number of votes for candidates A and B after counting k votes. Let $X_k = \frac{Y_{N-k}}{N-k}$. In the homework, you will first show that $X = (X_k)_{k\geq 0}$ is a martingale. Then you will define a stopping time $T = \min\{k \in [0, N] : X_k = 0\}$, or $T = N - 1$ if there is no such k and apply Theorem 3.14.*

The following result provides a concentration bound for martingales, where note that we are not imposing independence (martingales can have dependent increments!).

Theorem 3.17 (Azuma–Hoeffding Inequality)**.** *Let $(X_n)_{n\geq 1}$ be a martingale adapted to the filtration $(\mathcal{F}_n)_{n\geq 1}$ and assume that almost surely $|X_n - X_{n-1}| \leq c_i$ for all n, where define $X_0 = \mathbb{E}[X_1]$. Then, for all $t > 0$,*

$$\Pr(X_n - X_0 \geq t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right),$$

and

$$\Pr(X_n - X_0 \leq -t) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Combining both bounds yields

$$\Pr(|X_n - X_0| \leq t) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\right).$$

Note that one can view Theorem 3.17 as providing a concentration bound on the sum function of random variables: $X_n - X_0 = \sum_{i=1}^{n} Y_i$, where $Y_i = X_1 - X_0$. The following result lets us getting other bounds when the functions are relatively more general.

Theorem 3.18 (McDiarmid's Inequality)**.** *Consider n independent random variables $X_1, \ldots, X_n$ taking values in A_i for each i, and a function $f : \prod_{i=1}^{n} A_i \to \mathbb{R}$ satisfying*

$$|f(x) - f(x')| \leq c_i \quad \text{whenever } x \text{ and } x' \text{ differ only in the } i\text{th coordinate.}$$

Let

$$\mu = \mathbb{E}[f(X_1, \ldots, X_n)]$$

be the expected value of the random variable $f(X)$. Then for any $\beta > 0$,

$$\Pr(f(X) \geq \mu + \beta) \leq \exp\left(-\frac{2\beta^2}{\sum_{i=1}^{n} c_i^2}\right),$$

and

$$\Pr(f(X) \leq \mu - \beta) \leq \exp\left(-\frac{2\beta^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Example 3.19. *Consider the following balls and bins problem, where n balls uniformly at random and independently into n bins. Let L_i denote the number of balls in bin i. Note that L_i is distributed as a $\mathrm{Bin}(n, 1/n)$ random variable. Per Markov's inequality, we have*

$$\Pr[L_i \geq 1 + \lambda] \leq \frac{1}{1+\lambda} \approx \frac{1}{\lambda}.$$

However, Chebyshev's inequality gives a much better bound, since

$$\Pr[|L_i - 1| \geq \lambda] \leq \frac{(1 - 1/n)}{\lambda^2} \approx \frac{1}{\lambda^2}.$$

Setting $\lambda = 2\sqrt{n}$, the probability of any fixed bin having more than $2\sqrt{n} + 1$ balls is at most $\frac{1}{4n}$. Taking a union bound over all bins, we have that with probability at least $1 - n \cdot \frac{1}{4n} \leq \frac{1}{4}$, the load on every bin is at most $1 + 2\sqrt{n}$.

Now, if we apply Chernoff bound for Bernoulli trials, we get

$$\Pr[L_i \geq 1 + \lambda] \leq \exp\left(-\frac{\lambda^2}{2+\lambda}\right).$$

If we set $\lambda = \Theta(\log n)$, the probability that bin i has more than $1 + \lambda$ balls is at most $1/n^2$. Taking a union bound over all bins, the probability that any bin has at least $1 + \lambda$ balls is at most $1/n$, i.e., the maximum load is $O(\log n)$ balls with high probability.

Now define $f(X) = f(X_1, X_2, ..., X_n)$ to be the number of empty bins. Note that $\mathbb{E}[f(X)] = n(1 - 1/n)^n$. Noting that changing the assignment of one ball can only change the number of empty bins by at most 1, we get the following bound by McDiarmid's inequality:

$$\mathbb{P}(|f(X) - n(1 - 1/n)^n| \geq \beta) \leq 2\exp\left(\frac{-2\beta^2}{n}\right).$$

Discussion 3.20. *Suppose that you are drawing samples from a distribution X with $\mathbb{E}[X] = \mu$ and $Var(X) = \sigma^2$. Let's say you have a bound on the variance, i.e., $\sigma \leq C$. How large should the sample size be to ensure that with probability $1 - p$, the sample average is 2 away from the mean μ? What if you have a bound on the distribution instead, i.e., $|X| \leq C$?*

Discussion 3.21. *Consider the symmetric random walk. Denote the position of the random walk after n steps by $S_n = \sum_{i=1}^{n} X_i$, where $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = 1/2$. What are the tail bounds under Chebyshev's and Hoeffding's inequalities?*

Discussion 3.22. *Recall the coupon collector problem: given n different types of coupons, how many coupons in expectation do we need to draw with replacement before having drawn each coupon at least once? Using linearity of expectation, one can easily show that $\mathbb{E}[X] = \sum_{i=1}^{n} \frac{n}{n-i+1} = nH(n) \approx n\log(n)$, where $H(n)$ is the harmonic number. What does Markov's inequality, Chebyshev's inequality imply on tail bounds?*