Lecture #21: Tail Bounds
Proofs of Claims

Proof of the Basic Inequality

Theorem 3 (Basic Inequality)**.** *Let Ω be a* **finite** *sample space with probability distribution* $\mathsf{P} : \Omega \to \mathbb{R}$*, let $X : \Omega \to \mathbb{R}$ be a random variable, and let $h : \mathbb{R} \to \mathbb{R}$ be a total function such that*

$$h(x) \geq 0 \quad \textit{for all } x \in \mathbb{R}.$$

Then, for every real number a such that $a > 0$,

$$\mathsf{P}\left(h(X) \geq a\right) \leq \frac{\mathsf{E}\left[h(X)\right]}{a}.$$

Proof. Suppose, to obtain a contradiction, that

$$\mathsf{P}\left(h(X) \geq a\right) > \frac{\mathsf{E}\left[h(X)\right]}{a}.$$

Then, since $a > 0$ it follows (by multiplying both sides of the inequality by a) that

$$a \times \mathsf{P}\left(h(X) \geq a\right) > \mathsf{E}\left[h(X)\right]. \tag{1}$$

Now (since Ω is a finite, and one can reorder the terms in a finite sum without changing its value)

$$\begin{aligned}
\mathsf{E}\left[h(X)\right] &= \sum_{\mu \in \Omega} h(X(\mu)) \times \mathsf{P}\left(\mu\right) \\
&= \sum_{\substack{\mu \in \Omega \\ h(X(\mu)) < a}} h(X(\mu)) \times \mathsf{P}\left(\mu\right) + \sum_{\substack{\mu \in \Omega \\ h(X(\mu)) \geq a}} h(X(\mu)) \times \mathsf{P}\left(\mu\right) && \text{(splitting the sum)} \\
&\geq \sum_{\substack{\mu \in \Omega \\ h(X(\mu)) < a}} 0 \times \mathsf{P}\left(\mu\right) + \sum_{\substack{\mu \in \Omega \\ h(\mu) \geq a}} h(X(\mu)) \times \mathsf{P}\left(\mu\right) \\
& && \text{(since } h(X(\mu)) \geq 0 \text{ and } \mathsf{P}\left(\mu\right) \geq 0 \text{ for every outcome } \mu \in \Omega\text{)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{\substack{\mu\in\Omega \\ h(X(\mu))\geq a}} h(X(\mu)) \times \mathsf{P}\left(\mu\right) \\
&\geq \sum_{\substack{\mu\in\Omega \\ h(X(\mu))\geq a}} a \times \mathsf{P}\left(\mu\right) && \text{(again, since } \mathsf{P}\left(\mu\right) \geq 0 \text{ for every outcome } \mu \in \Omega\text{)} \\
&= a \times \sum_{\substack{\mu\in\Omega \\ h(X(\mu))\geq a}} \mathsf{P}\left(\mu\right) \\
&= a \times \mathsf{P}\left(h(X) \geq a\right) \\
&> \mathsf{E}\left[X\right] && \text{(by the inequality at line (1), above).}
\end{aligned}
$$

Thus $\mathsf{E}\left[h(X)\right] > \mathsf{E}\left[h(X)\right]$ — which is impossible, since a real number cannot be strictly greater than itself. Since a ***assumption*** has been obtained, the assumption must be false — and

$$\mathsf{P}\left(h(X) \geq a\right) \leq \frac{\mathsf{E}\left[h(X)\right]}{a},$$

as claimed. □

Proof of Markov's Inequality

Corollary 4 (Markov's Inequality)**.** *Let Ω be a* **finite** *sample space with probability distribution* $\mathsf{P} : \Omega \to \mathbb{R}$*, and let $X : \Omega \to \mathbb{R}$ be a random variable. Then, for every* **positive** *real number a,*

$$\mathsf{P}\left(|X| \geq a\right) \leq \frac{\mathsf{E}\left[|X|\right]}{a}.$$

Proof. This follows immediately from Theorem 3: In particular, Markov's Inequality follows by an application of the Basic Inequality, using the function $h : \mathbb{R} \to \mathbb{R}$ such that $h(x) = |x|$ for every real number x. □

Alternative Form of Variance

Theorem 5. *Let Ω be a* **finite** *sample space, let* $\mathsf{P} : \Omega \to \mathbb{R}$ *be a probability distribution for Ω, and let X be a random variable. Then X^2 is also a random variable, and*

$$\mathrm{var}(X) = \mathsf{E}\left[X^2\right] - \mathsf{E}\left[X\right]^2.$$

Proof. Since Ω is a finite sample space,

$$\begin{aligned}
\text{var}(X) &= \sum_{\mu\in\Omega}(X(\mu) - \mathsf{E}[X])^2 \times \mathsf{P}(\mu) \\
&= \sum_{\mu\in\Omega}(X(\mu)^2 - 2\mathsf{E}[X] \times X(\mu) + \mathsf{E}[X]^2) \times \mathsf{P}(\mu) \\
&= \left(\sum_{\mu\in\Omega} X(\mu)^2 \times \mathsf{P}(\mu)\right) - \left(\sum_{\mu\in\Omega} 2\mathsf{E}[X] \times X(\mu) \times \mathsf{P}(\mu)\right) + \left(\sum_{\mu\in\Omega} \mathsf{E}[X]^2 \times \mathsf{P}(\mu)\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(reordering terms)} \\
&= \mathsf{E}\left[X^2\right] - 2\mathsf{E}[X] \times \left(\sum_{\mu\in\Omega} X(\mu) \times \mathsf{P}(\mu)\right) + \mathsf{E}[X]^2 \times \sum_{\mu\in\Omega} \mathsf{P}(\mu) \\
&= \mathsf{E}\left[X^2\right] - 2\mathsf{E}[X] \times \mathsf{E}[X] + \mathsf{E}[X]^2 \times 1 \\
&= \mathsf{E}\left[X^2\right] - 2\mathsf{E}[X]^2 + \mathsf{E}[X]^2 \\
&= \mathsf{E}\left[X^2\right] - \mathsf{E}[X]^2,
\end{aligned}$$

as claimed. ☐

Using Pairwise Independence

Theorem 6. *Let Ω be a* **finite** *sample space with probability distribution* $\mathsf{P} : \Omega \to \mathbb{R}$ *and let* $X_1, X_2, \ldots, X_n : \Omega \to \mathbb{R}$ *be random variables (for some positive integer n). If $X_1, X_2, \ldots, X_n$ are* **pairwise independent** *then*

$$\text{var}(X_1 + X_2 + \cdots + X_n) = \text{var}(X_1) + \text{var}(X_2) + \cdots + \text{var}(X_n).$$

Recall, from Lecture #20, that the expected values of random variables are not generally "multiplicative". As the following claim states, they *are* multiplicative when the random variables are independent.

Claim. *Let Ω be a* **finite** *sample space with probability distribution* $\mathsf{P} : \Omega \to \mathbb{R}$*, and let $X, Y : \Omega \to \mathbb{R}$ be* **independent** *random variables (with respect to* P*). Then*

$$\mathsf{E}[X \times Y] = \mathsf{E}[X] \times \mathsf{E}[Y].$$

Proof. Since the sample space Ω is finite there is a finite set of values

$$V_X = \{\alpha_1, \alpha_2, \ldots \alpha_k\} \subseteq \mathbb{R}$$

such that, if

$$S_i = \{\mu \in \Omega \mid X(\mu) = \alpha_i\}$$

for $1 \le i \le k$, then $S_i \neq \emptyset$ for $1 \le i \le k$ and

$$S_1 \cup S_2 \cup \cdots \cup S_k = \Omega.$$

Assuming (as the above notation may suggest) that $\alpha_1, \alpha_2, \ldots, \alpha_k$ are distinct (so that $|V_X| = k$), $S_i \cap S_j = \emptyset$, as well, for $1 \le i, j \le k$ such that $i \neq j$.

Now

$$\begin{aligned}
\mathsf{E}\left[X\right] &= \sum_{\mu \in \Omega} X(\mu) \times \mathsf{P}\left(\mu\right) \\
&= \sum_{i=1}^{k} \left(\sum_{\mu \in S_i} X(\mu) \times \mathsf{P}\left(\mu\right) \right) && \text{(reordering terms)} \\
&= \sum_{i=1}^{k} \left(\sum_{\mu \in S_i} \alpha_i \times \mathsf{P}\left(\mu\right) \right) && \text{(since } X(\mu) = \alpha_i \text{ for every outcome } \mu \in S_i\text{)} \\
&= \sum_{i=1}^{k} \left(\alpha_i \times \sum_{\mu \in S_i} \mathsf{P}\left(\mu\right) \right) \\
&= \sum_{i=1}^{k} \alpha_i \times \mathsf{P}\left(S_i\right) \\
&= \sum_{i=1}^{k} \alpha_i \times \mathsf{P}\left(X = \alpha_i\right).
\end{aligned}$$

Similarly, there is a finite set of values

$$V_Y = \{\beta_1, \beta_2, \ldots \beta_\ell\} \subseteq \mathbb{R}$$

such that, if

$$T_i = \{\mu \in \Omega \mid Y(\mu) = \beta_i\}$$

for $1 \le i \le \ell$, then $T_i \neq \emptyset$ for $1 \le i \le \ell$ and

$$T_1 \cup T_2 \cup \cdots \cup T_\ell = \Omega.$$

Assuming (as the above notation may suggest) that $\beta_1, \beta_2, \ldots, \beta_\ell$ are distinct (so that $|V_Y| = \ell$), $T_i \cap T_j = \emptyset$, as well, for $1 \le i, j \le \ell$ such that $i \neq j$. Repeating the above argument (replacing X with Y and the set V_X with V_Y) that

$$\mathsf{E}\left[Y\right] = \sum_{j=1}^{\ell} \beta_j \times \mathsf{P}\left(Y = \beta_j\right).$$

Note that, for every outcome $\mu \in \Omega$, there exists *exactly one* pair of integers i and j such that $\mu \in S_i \cap T_j$. Thus

$$\begin{aligned}
\mathsf{E}[X \times Y] &= \sum_{\mu \in \Omega} (X \times Y)(\mu) \times \mathsf{P}(\mu) \\
&= \sum_{\mu \in \Omega} X(\mu) \times Y(\mu) \times \mathsf{P}(\mu) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} \left(\sum_{\mu \in S_i \cap T_j} X(\mu) \times Y(\mu) \times \mathsf{P}(\mu) \right) && \text{(reordering terms)} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} \left(\sum_{\mu \in S_i \cap T_j} \alpha_i \times \beta_j \times \mathsf{P}(\mu) \right) && \text{(by the definitions of } S_i \text{ and } T_j\text{)} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} \left(\alpha_i \times \beta_j \times \sum_{\mu \in S_i \cap T_j} \mathsf{P}(\mu) \right) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} (\alpha_i \times \beta_j \times \mathsf{P}(\mu \in S_i \cap T_j)) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} (\alpha_i \times \beta_j \times \mathsf{P}(\mu \in S_i \text{ and } \mu \in S_j)) \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} (\alpha_i \times \beta_j \times \mathsf{P}(X = \alpha_i \text{ and } Y = \beta_j)) && \text{(by the definitions of } S_i \text{ and } T_j\text{)} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} (\alpha_i \times \beta_j \times \mathsf{P}(X = \alpha_i) \times \mathsf{P}(Y = \beta_j)) \\
& && \text{(since } X \text{ and } Y \text{ are } \textbf{\textit{independent}} \text{ random variables)} \\
&= \sum_{i=1}^{k} \sum_{j=1}^{\ell} (\alpha_i \times \mathsf{P}(X = \alpha_i)) \times (\beta_j \times \mathsf{P}(Y = \beta_j)) \\
&= \left(\sum_{i=1}^{k} \alpha_i \times \mathsf{P}(X = \alpha_i) \right) \times \left(\sum_{j=1}^{\ell} \beta_j \times \mathsf{P}(Y = \beta_j) \right) \\
& && \text{(reordering terms, once again)} \\
&= \mathsf{E}[X] \times \mathsf{E}[Y],
\end{aligned}$$

as claimed. $\square$

Proof of Theorem 6. Let

$$X = X_1 + X_2 + \cdots + X_n.$$

Then, by Theorem 5, above,

$$\begin{aligned}
\mathrm{var}(X) &= \mathsf{E}\left[X^2\right] - \mathsf{E}\left[X\right]^2 \\
&= \mathsf{E}\left[\left(\textstyle\sum_{i=1}^n X_i\right)^2\right] - \left(\mathsf{E}\left[\textstyle\sum_{i=1}^n X_i\right]\right)^2 \\
&= \mathsf{E}\left[\textstyle\sum_{i=1}^n \sum_{j=1}^n X_i \times X_j\right] - \left(\mathsf{E}\left[\textstyle\sum_{i=1}^n X_i\right]\right)^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n \mathsf{E}\left[X_i \times X_j\right] - \left(\sum_{i=1}^n \mathsf{E}\left[X_i\right]\right)^2 && \text{(by Linearity of Expectation)} \\
&= \sum_{i=1}^n \sum_{j=a}^n \mathsf{E}\left[X_i \times X_j\right] - \sum_{i=1}^n \sum_{j=1}^n \mathsf{E}\left[X_i\right] \times \mathsf{E}\left[X_j\right] && \text{(reordering terms)} \\
&= \sum_{i=1}^n \left(\mathsf{E}\left[X_i^2\right] - \mathsf{E}\left[X_i\right]^2\right) + \sum_{\substack{1 \le i,j \le n \\ i \neq j}} \left(\mathsf{E}\left[X_i \times X_j\right] - \mathsf{E}\left[X_i\right] \times \mathsf{E}\left[X_j\right]\right) \\
& && \text{(reordering terms, again)} \\
&= \sum_{i=1}^n \left(\mathsf{E}\left[X_i^2\right] - \mathsf{E}\left[X_i\right]^2\right) + \sum_{\substack{1 \le i,j \le n \\ i \neq j}} \left(\mathsf{E}\left[X_i\right] \times \mathsf{E}\left[X_j\right] - \mathsf{E}\left[X_i\right] \times \mathsf{E}\left[X_j\right]\right) \\
& && \text{(by the above claim, since } X_i \text{ and } X_j \text{ are } \textbf{\textit{independent}} \text{ if } i \neq j\text{)} \\
&= \sum_{i=1}^n \left(\mathsf{E}\left[X_i^2\right] - \mathsf{E}\left[X_i\right]^2\right) + \sum_{\substack{1 \le i,j \le n \\ i \neq j}} 0 \\
&= \sum_{i=1}^n \left(\mathsf{E}\left[X_i^2\right] - \mathsf{E}\left[X_i\right]^2\right) \\
&= \sum_{i=1}^n \mathrm{var}(X_i).
\end{aligned}$$

That is,

$$\mathrm{var}(X_1 + X_2 + \cdots + X_n) = \mathrm{var}(X_1) + \mathrm{var}(X_2) + \cdots + \mathrm{var}(X_n),$$

as claimed. □

Chebyshev's Inequality and Cantelli's Inequality

The lecture notes also include results that can be applied, to used the expected values of variances of random variables to establish tail bounds, namely, ***Chebyshev's Inequality*** and ***Cantelli's Inequality*** (stated as Theorem 7 and Theorem 8, respectively). These will be considered in the lecture presentation and the tutorial exercise for this topic.

Proof of the Chernoff Bound

Let e be ***Euler's number*** — so that e is approximately (but not exactly) 2.71828.

Several of the following results rely on the following result from calculus:

Theorem (Mean Value Theorem)**.** *Let $a, b \in \mathbb{R}$ such that $a < b$, and let f be a real-valued function that is defined on the closed interval*

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}$$

and differentiable on the open interval

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}.$$

Then there exists a real number c such that $a < c < b$ and

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Theorem 9 (The Chernoff Bound)**.** *Let Ω be a finite sample space with probability distribution $\mathsf{P} : \Omega \to \mathbb{R}$. Suppose that $X_1, X_2, \ldots, X_n$ are mutually independent random variables such that $X_i : \Omega \to \{0, 1\}$ for $1 \leq i \leq n$, and suppose that $\mathsf{P}\left(X_i = 1\right) = p_i$ for every integer i such that $1 \leq i \leq n$ — so that $p_1, p_2, \ldots, p_n$ are real numbers such that $0 \leq p_i \leq 1$ for $1 \leq i \leq n$. Let $X = X_1 + X_2 + \cdots + X_n$, and let $\mu = \mathsf{E}\left[X\right] = p_1 + p_2 + \cdots + p_n > 0$.*

(a) For every real number δ such that $\delta > 0$,

$$\mathsf{P}\left(X \geq (1 + \delta)\mu\right) \leq \left(\frac{e^{\delta}}{(1 + \delta)^{1+\delta}}\right)^{\mu}.$$

(b) For every real number δ such that $0 < \delta < 1$,

$$\mathsf{P}\left(X \leq (1 - \delta)\mu\right) \leq \left(\frac{e^{-\delta}}{((1 - \delta)^{1-\delta}}\right)^{\mu},$$

Sketch of Proof. Let $X_1, X_2, \dots, X_n$ and X be as in the statement of the claim.

(a) Let t be any positive real number. Then, since X is a random variable, e^{tX} is a non-negative random variable — and, for any real number δ such that $\delta > 0$,

$$\mathsf{P}\left(X \geq (1+\delta)\mu\right) = \mathsf{P}\left(e^{tX} \geq e^{t(1+\delta)\mu}\right).$$

Now, since $X = X_1 + X_2 + \cdots + X_n$,

$$\mathsf{E}\left[e^{tX}\right] = \mathsf{E}\left[e^{tX_1} \times e^{tX_2} \times \cdots \times e^{tX_n}\right],$$

and, since the random variables $X_1, X_2, \dots, X_n$ are mutually independent, so are the random variable $e^{tX_1}, e^{tX_2}, \dots, e^{tX_n}$. This can be used to show, by induction on n, that

$$\mathsf{E}\left[e^{tX}\right] = \mathsf{E}\left[e^{tX_1} \times e^{tX_2} \times \cdots \times e^{tX_n}\right] = \prod_{i=1}^{n} \mathsf{E}\left[e^{tX_i}\right]. \tag{2}$$

Since the random variable X_i only assumes values 0 and 1, with probabilities $1 - p_i$ and p_i respectively, the random variable e^{tX_i} only assumes values $e^0 = 1$ and e^t, with probabilities $1 - p_i$ and p_i respectively, so that

$$\mathsf{E}\left[e^{tX_i}\right] = (1 - p_i) \cdot 1 + p_i \cdot e^t = 1 + p_i(e^t - 1). \tag{3}$$

It now follows by the equations at lines (2) and (3) that

$$\mathsf{E}\left[e^{tX}\right] = \prod_{i=1}^{n} \mathsf{E}\left[e^{tX_i}\right] = \prod_{i=1}^{n}(1 + p_i(e^t - 1)).$$

Now recall, by Markov's Inequality, that

$$\mathsf{P}\left(e^{tX} \geq k \cdot \mathsf{E}\left[e^{tX}\right]\right) \leq \tfrac{1}{k}$$

for any positive real number k. In particular, this is true when $k = e^{t(1+\delta)\mu} \cdot \mathsf{E}\left[e^{tX}\right]^{-1}$, so that

$$\begin{aligned} \mathsf{P}\left(e^{tX} \geq e^{t(1+\delta)\mu}\right) &\leq \frac{\mathsf{E}\left[e^{tX}\right]}{e^{t(1+\delta)\mu}} \\ &= \frac{\prod_{i=1}^{n}(1 + p_i(e^t - 1))}{e^{t(1+\delta)\mu}}. \end{aligned}$$

Consider the function $h : \mathbb{R} \to \mathbb{R}$ such that $h(x) = e^x - x - 1$ for every non-negative real number x. This function is defined and continuous and differentiable over the set of non-negative real numbers (so that it is defined, and continuous over the closed interval

$[0, a]$ and differentiable over the open interval $(0, a)$, for every positive real number a). Furthermore, $h'(x) = e^x - 1$, so that $h'(a) \geq 0$ for every non-negative real number a — and it follows by the Mean Value Theorem that $h(a) \geq h(0) = 0$ for every non-negative real number a, as well. Thus $1+x \leq e^x$ for every positive real number x, so that $1+p_i(e^t-1) \leq e^{p_i(e^t-1)}$ for every integer i such that $1 \leq i \leq n$. It now follows, from the above inequality, that

$$\begin{aligned}\mathsf{P}\left(e^{tX} \geq e^{t(1+\delta)\mu}\right) &\leq \frac{\prod_{i=1}^n e^{p_i(e^t-1)}}{e^{t(1+\delta)\mu}} \\ &= \frac{e^{\sum_{i=1}^n p_i(e^t-1)}}{e^{t(1+\delta)\mu}} \\ &= \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}}.\end{aligned}$$

Now let $t = \ln(1+\delta)$ — which is a positive real number, since $\delta > 0$. Then it follows from the above that

$$\begin{aligned}\mathsf{P}\left(X \geq (1+\delta)\mu\right) &= \mathsf{P}\left(e^{tX} \geq e^{t(1+\delta)\mu}\right) \\ &= \frac{e^{(e^t-1)\mu}}{e^{t(1+\delta)\mu}} \\ &= \frac{e^{(e^t-1)\mu}}{(e^t)^{(1+\delta)\mu}} \\ &= \frac{e^{(1+\delta-1)\mu}}{(1+\delta)^{(\delta+1)\mu}} && \text{(for } t = \ln(1+\delta)\text{, as chosen above)} \\ &= \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}} \\ &= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu\end{aligned}$$

as claimed.

(b) Now let δ be a real number such that $0 < \delta < 1$. Then $X \leq (1-\delta)\mu$ if and only if $(\mu - X) \geq \mu - (1-\delta)\mu = \delta\mu$. Once again, let t be any positive real number. Then, since X is a random variable, $e^{t(\mu-X)}$ is a non-negative random variable — and, for any real number δ such that $0 < \delta < 1$, $X \leq (1-\delta)\mu$ if and only if $\mu - X \geq \delta\mu$, so that

$$\mathsf{P}\left(X \leq (1-\delta)\mu\right) = \mathsf{P}\left(e^{t(\mu-X)} \geq e^{t\delta\mu}\right).$$

Let $k = e^{t\delta\mu}/\mathsf{E}\left[e^{t(\mu-X)}\right]$. Since X is a random variable, the function $e^{t(\mu-X)}$ is a non-negative random variable and it follows by an application of Markov's inequality, with this choice of k, that

$$\begin{aligned}
\mathsf{P}\left(X \leq (1-\delta)\mu\right) &= \mathsf{P}\left(e^{t(\mu-X)} \geq e^{t\delta\mu}\right) \\
&= \mathsf{P}\left(e^{t(\mu-X)} \geq k \cdot \mathsf{E}\left[e^{t(\mu-X)}\right]\right) && \text{(for } k \text{ as above)} \\
&\leq k^{-1} && \text{(by Markov's Inequality)} \\
&= e^{-t\delta\mu} \times \mathsf{E}\left[e^{t(\mu-X)}\right] \\
&= e^{-t\delta\mu} \times e^{t\mu} \times \mathsf{E}\left[e^{-tX}\right] && \text{(by Linearity of Expectation)} \\
&= e^{t\mu(1-\delta)} \times \mathsf{E}\left[e^{-tX}\right].
\end{aligned}$$

Once again, since $X = X_1 + X_2 + \cdots + X_n$ and the random variables $X_1, X_2, \dots, X_n$ are mutually independent, the random variables $e^{-tX_1}, e^{-tX_2}, \dots, e^{-tX_n}$ are mutually independent as well. It can be shown, by induction on n, that

$$\begin{aligned}
\mathsf{E}\left[e^{-tX}\right] &= \mathsf{E}\left[\prod_{i=1}^{n} e^{-tX_i}\right] \\
&= \prod_{i=1}^{n} \mathsf{E}\left[e^{-tX_i}\right].
\end{aligned}$$

Now, since $X_i = 0$ with probability $1 - p_i$ and $X_i = 1$ with probability p_i, $e^{-tX_i} = e^0 = 1$ with probability $1 - p_i$ and $e^{-tX_i} = e^{-t}$ with probability p_i. Thus

$$\mathsf{E}\left[e^{-tX_i}\right] = 1 \times (1 - p_i) + e^{-t} \times p_i = 1 - p_i(1 - e^{-t}).$$

Consider the real-valued function $\widehat{h}$ such that, for every real number x, $\widehat{h}(x) = e^{-x} + x - 1$: This function is both continuous and differentiable over the interval $[-1, 1]$. Furthermore, $\widehat{h}'(x) = -e^{-x} + 1$, so that $\widehat{h}'(x) < 0$ if $-1 \leq x < 0$, $\widehat{h}'(0) = 0$, and $\widehat{h}'(x) > 0$ if $0 < x \leq -1$. It follows by the Mean Value Theorem that $\widehat{h}(x) \geq \widehat{h}(0) = 0$, so that $1 - x \leq e^{-x}$, for every real number x such that $|x| \leq 1$. In particular, when $x = p_i(1 - e^{-t})$,

$$1 - p_i(1 - e^{-t}) \leq e^{-p_i(1 - e^{-t})}.$$

It now follows, from the above, that

$$\begin{aligned}
\mathsf{E}\left[e^{-tX}\right] &= \prod_{i=1}^{n} \mathsf{E}\left[e^{-tX_i}\right] \\
&= \prod_{i=1}^{n} \left(1 - p_i(1 - e^{-t})\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \prod_{i=1}^{n} e^{-p_i(1-e^{-t})} \\
&= e^{-(1-e^{-t})\sum_{i=1}^{n} p_i} \\
&= e^{-\mu(1-e^{-t})}.
\end{aligned}
$$

Let $t = -\ln(1-\delta)$. Since $0 < \delta < 1$, $0 < 1-\delta < 1$ as well, $\ln(1-\delta) < 0$, and $t > 0$. It now follows by the above that

$$
\begin{aligned}
\mathsf{P}\left(X \leq (1-\delta)\mu\right) &= e^{t\mu(1-\delta)} \times \mathsf{E}\left[e^{-tX}\right] \\
&\leq e^{t\mu(1-\delta)} \times e^{-\mu(1-e^{-t})} \\
&= (e^{-t})^{-\mu(1-\delta)} \times e^{-\mu(1-e^{-t})} \\
&= (1-\delta)^{-\mu(1-\delta)} \times e^{-\mu(1-(1-\delta)} && \text{(since } t = -\ln(1-\delta)\text{)} \\
&= \left(\left(\frac{1}{1-\delta}\right)^{1-\delta} \times e^{-\delta}\right)^{\mu} \\
&= \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu}
\end{aligned}
$$

as claimed. □

The inequalities given in Theorem 9 might not seem to be very useful — or easy to apply — when δ is small. The following inequalities are slightly weaker but might be easier to use.

Let $\mathbb{R}^{\geq 0}$ be the set of non-negative real numbers.

Corollary 10. *Let Ω be a a finite sample space with probability distribution $P : \Omega \to \mathbb{R}$. Suppose that $X_1, X_2, \dots, X_n$ are mutually independent random variables such that $X_i : \Omega \to \{0,1\}$ for $1 \leq i \leq n$, and suppose that $\mathsf{P}\left(X_i = 1\right) = p_i$, for $1 \leq i \leq n$ — so that $p_1, p_2, \dots, p_n$ are real numbers such that $0 \leq p_i \leq 1$ for $1 \leq i \leq n$. Let $X = X_1 + X_2 + \cdots + X_n$ and let $\mu = \mathsf{E}\left[X\right] = p_1 + p_2 + \cdots + p_n > 0$.*

(a) If $0 < \delta \leq 1$ then $\mathsf{P}\left(X \geq (1+\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{3}}$.

(b) For all positive real numbers γ and δ, if $\delta \geq \gamma$ then $\mathsf{P}\left(X \geq (1+\delta)\mu\right) \leq \left(e^{\gamma-(1+\gamma)\ln(1+\gamma)}\right)^{\mu}$.

(c) If $0 < \delta < 1$ then

$$
\mathsf{P}\left(X \leq (1-\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{2}}.
$$

Sketch of Proof. Let $X_1, X_2, \dots, X_n$ and X be as in the statement of the claim. It follows by part (a) of Theorem 9 that, if δ is a positive real number, then

$$
\mathsf{P}\left(X \geq (1+\delta)\mu\right) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu} = e^{f(\delta)\mu}
$$

for $\mu = \mathsf{E}[X]$ and for the function $f : \mathbb{R}^{\geq 0} \to \mathbb{R}$ such that $f(x) = x - (1+x)\ln(1+x)$. The function f is continuous, and (twice) differentiable, over $\mathbb{R}^{\geq 0}$: For every non-negative real number x, $f'(x) = -\ln(1+x)$ and $f''(x) = -\frac{1}{1+x}$.

(a) Suppose that $0 < \delta < 1$, and consider the function $g : \mathbb{R}^{\geq 0} \to \mathbb{R}$ such that

$$g(x) = f(x) + \frac{x^2}{3} = x - (1+x)\ln(1+x) + \frac{x^2}{3}$$

for $x \in \mathbb{R}$ such that $x \geq 0$. Differentiating g, one can see that

$$g'(x) = -\ln(1+x) + \frac{2x}{3}$$

and

$$g''(x) = -\frac{1}{1+x} + \frac{2}{3}$$

— so that $g''(x) < 0$ when $0 < x < \frac{1}{2}$, $g''(1/2) = 0$, and $g''(x) > 0$ when $\frac{1}{2} < x \leq 1$. This can be used, along with the Mean Value Theorem, to argue that the function g' must be decreasing over the interview $\left[0, \frac{1}{2}\right]$ and increasing over the interval $\left[\frac{1}{2}, 1\right]$.

Now, $g'(0) = 0$, $g'\left(\frac{1}{2}\right) = -\ln\left(\frac{3}{2}\right) + \frac{1}{3} < 0$, and $g'(1) = -\ln 2 + \frac{2}{3} > 0$. It follows that there exists some real number α such that $\frac{1}{2} < \alpha < 1$, $g'(x) < 0$ for every real number x such that $0 < x < \alpha$, $g'(\alpha) = 0$, and $g'(x) > 0$ for every real number x such that $\alpha < x < 1$.

Applying the Mean Value Theorem once again, it can be argued that the function g is decreasing over the interval $[0, \alpha]$ and increasing over the interval $[\alpha, 1]$ — so that

$$g(x) \leq \max(g(0), g(1))$$

for every real number x such that $0 < x < 1$. Now, $g(0) = 0$ and $g(1) = \frac{4}{3} - 2\ln 2 < 0$, so it follows by the above that $g(x) \leq 0$ for every real number x such that $0 < x < 1$. Thus

$$x - (1+x)\ln(1+x) \leq -\frac{x^2}{3}$$

for every real number x such that $0 \leq x \leq 1$. Since the function $h(x) = e^x$ is an increasing function of x, it also follows that

$$e^{x-(1+x)\ln(1+x)} \leq e^{-\frac{x^2}{3}}$$

— that is,

$$\frac{e^x}{(1+x)^{1+x}} \leq e^{-\frac{x^2}{3}},$$

when $0 < x < 1$. In this is true when $x = \delta$ (if $0 < \delta < 1$, as above), so that

$$\mathsf{P}(X \geq (1+\delta)\mu) \leq \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^\mu$$

$$
\begin{aligned}
&\leq \left(e^{-\frac{\delta^2}{3}}\right)^{\mu} \\
&= e^{\frac{\delta^2 \mu}{3}},
\end{aligned}
$$

as required to establish the inequality in part (a) of the claim.

(b) Part (b) of the claim is a straightforward consequence of part (a) of Theorem 9 (considering γ in place of δ) and the fact that if δ and γ are real numbers such that $0 < \delta \leq \gamma$, then the event "$X \geq (1-\delta)\mu$" implies the event "$X \geq (1-\gamma)\mu$".

(c) Consider, now, the real-valued function $\widetilde{h}$ such that $\widetilde{h}(x) = -(1-x)\ln(1-x) - x + \frac{x^2}{2}$ for every real number x such that $0 \leq x < 1$. This function is continuous and differentiable over the interval $(0,1)$, and $\widetilde{h}'(x) = \ln(1-x) + x$ and $\widetilde{h}''(x) = -\frac{1}{1-x} + 1$ for every real number x such that $0 < x < 1$. Thus $h''(x) \leq 0$ for every real number x such that $0 < x < 1$, so it follows by the Mean Value Theorem that $h'(x) \leq h'(0) = 0$ for every real number x such that $0 < x < 1$ as well. Applying the Mean Value Theorem again, it follows that $h(x) \leq h(0) = 0$, when $0 < x < 1$, as well. Thus if $0 < \delta < 1$ then $h(\delta) \leq 0$, so that

$$
-(1-\delta)\ln(1-\delta) - \delta \leq -\frac{\delta^2}{2}
$$

as well. Since the function $f(x) = e^x$ is an increasing function of x it follows that

$$
e^{-(1-\delta)\ln(1-\delta)-\delta} \leq e^{-\frac{\delta^2}{2}}
$$

when $0 < \delta < 1$ as well, that is,

$$
\left(\frac{1}{1-\delta}\right)^{1-\delta} \times e^{-\delta} \leq e^{-\frac{\delta^2}{2}}.
$$

It follows by this inequality, and part (b) of Theorem 9, that

$$
\mathsf{P}\left(X \leq (1-\delta)^{\mu}\right) \leq e^{-\frac{\delta^2 \mu}{2}},
$$

as claimed. ☐

Suppose, now, that $p_i = p$, for $1 \leq i \leq n$, for a real number $p/$ such that $0 < p < 1$. Then $\mu = \mathsf{E}[X] = pn$. Thus the following is a straightforward consequence of the above results. In particular, parts (a) and (b) of the following claim are consequences of parts (a) and (b) of Theorem 9, respectively, and parts (c), (d), and (r) of the claim are consequences of parts (a), (b), and (c) of Corollary 10, respectively.

Corollary 11. *Let Ω be a a finite sample space with probability distribution P : $\Omega \to \mathbb{R}$. Suppose that $X_1, X_2, \dots, X_n$ are mutually independent random variables such that $X_i : \Omega \to \{0,1\}$ for $1 \leq i \leq n$, and suppose that $\mathsf{P}(X_i = 1) = p$, for $1 \leq i \leq n$ — for a real number p such that $0 < p < 1$. Let $X = X_1 + X_2 + \cdots + X_n$.*

(a) If $\delta > 0$ then

$$\mathsf{P}(X \geq (1+\delta)pn) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \right)^{pn}.$$

(b) If $0 < \delta < 1$ then

$$\mathsf{P}(X \leq (1-\delta)pn) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \right)^{pn}.$$

(c) If $0 \leq \delta \leq 1$ then $\mathsf{P}(X \geq (1+\delta)pn) \leq e^{-\frac{\delta^2 pn}{3}}$.

(d) For all positive real numbers γ and δ, if $\delta \geq \gamma$ then $\mathsf{P}(X \geq (1+\delta)pn) \leq \left(e^{\gamma-(1+\gamma)\ln(1+\gamma)}\right)^{pn}$.

(e) If $0 < \delta < 1$ then $\mathsf{P}(X \leq (1-\delta)pn) \leq e^{-\frac{\delta^2 pn}{2}}$.

More about Chernoff Bounds

The versions of a “Chernoff bound”, given here are special cases of a much more general bound (applying to a much wider variety of probability distributions, that was given by Chernoff [1].

The proof of Theorem 9 is a modified version of the proof given by Hagerup and Rüb [3]. The proof of Corollary 10 is based on the work of Hagerup and Rüb [3] as well.

For a variety of other (possibly easier to apply) “Chernoff bounds”, see the work of Dillencourt, Goodrich, and Mitzenmacher [2].

References

[1] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis bound on the sum of observations. *The Annals of Mathematical Statistics*, 23:403–507, 1952.

[2] Michael Dillencourt, Michael T. Goodrich, and Michael Mitzenmacher. Leveraging parameterized Chernoff bounds for simplfied algorithm analyses. *Information Processing Letters*, January 2025.

[3] Torben Hagerup and Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33:305–308, 1989/90.