CSCI-6971 Lecture Notes: Probability theory*

Kristopher R. Beevers
Department of Computer Science
Rensselaer Polytechnic Institute
beevek@cs.rpi.edu

January 31, 2006

1 Properties of probabilities

Let, A, B, C be events. Then the following properties hold:

- $A \subseteq B \Rightarrow P(A) \leq P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, so $P(A \cup B) \leq P(A) + P(B)$

Definition 1.1. Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{1}$$

Definition 1.2. The Law of Total Probability: if $A_1, \ldots, A_n$ are *disjoint* events that partition the sample space, then

$$P(B) = P(A_1 \cap B) + \ldots + P(A_n \cap B) \tag{2}$$

Definition 1.3. Bayes' Rule: By the def of conditional probability,

$$P(A \cap B) = P(A|B)\,P(B) = P(B|A)\,P(A) \tag{3}$$

so

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{4}$$

and by the Law of Total Probability

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(A)\,P(B|A) + P(A)\,P(B|\neg A)} \tag{5}$$

Definition 1.4. Independence: A and B are *independent* iff $P(A \cap B) = P(A)\,P(B)$ or equivalently $P(A|B) = P(A)$.

Definition 1.5. Conditional independence: A and B are independent when *conditioned on* C iff $P(A \cap B|C) = P(A|C)\,P(B|C)$. Note that independence and conditional independence do not imply each other.

*The primary sources for most of this material are: "Introduction to Probability," D.P. Bertsekas and J.N. Tsitsiklis, Athena Scientific, Belmont, MA, 2002; and "Randomized Algorithms," R. Motwani and P. Raghavan, Cambridge University Press, Cambridge, UK, 1995; and the author's own notes.

2 Random variables

Let X and Y be *random variables*.

Definition 2.1. A *probability density function* (PDF) is a function $f_X(x)$ such that:

- For every $B \subseteq \mathbb{R}$, $P(X \in B) = \int_B f_X(x)\ dx$
- For all x, $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)\ dx = 1$
- Note that $f_X(x) \neq$ the probability of an event; in particular, $f_X(x)$ may be greater than one.

Definition 2.2. A *cumulative density function* (CDF) is defined as:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\ dt \tag{6}$$

So a CDF is defined in terms of a PDF, and given a CDF, the PDF can be obtained by differentiating, i.e.: $f_X(x) = dF_X(x)/dx$.

Definition 2.3. The *expectation* (expected value or mean) of X is defined as:

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\ dx \tag{7}$$

Some properties of the expectation:

- $\mathbf{E}[\sum_i X_i] = \sum_i \mathbf{E}[X_i]$ regardless of independence
- For $\alpha \in \mathbb{R}$, $\mathbf{E}[\alpha X] = \alpha \mathbf{E}[X]$
- $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$ iff X and Y are independent
- Linearity of expectation: given $Y = aX + b$, a linear function of the random variable X, $\mathbf{E}[Y] = a\mathbf{E}[X] + b$, which we show for the discrete case:

$$\begin{aligned} \mathbf{E}[Y] &= \sum_x (ax+b) f_X(x) && (8)\\ &= a\sum_x x f_X(x) + b\sum_x f_X(x) && (9)\\ &= a\mathbf{E}[X] + b && (10) \end{aligned}$$

- Law of iterated expectations or law of total expectation: if X and Y are random variables in the same space, then $\mathbf{E}[\mathbf{E}[X|Y]] = \mathbf{E}[X]$, shown as follows:

$$\begin{aligned} \mathbf{E}[\mathbf{E}[X|Y]] &= \mathbf{E}\left[\sum_x xP(X = y|Y = y)\right] && (11)\\ &= \sum_y \left(\sum_x xP(X = x|Y = y)\right) P(Y = y) && (12)\\ &= \sum_y \sum_x xP(Y = y|X = x)\,P(X = x) && (13)\\ &= \sum_x xP(X = x) \cdot \sum_y P(Y = y|X = x) && (14)\\ &= \sum_x xP(X = x) && (15)\\ &= \mathbf{E}[X] && (16) \end{aligned}$$

Note that $\mathbf{E}[X|Y]$ is itself a random variable whose value depends on Y, i.e. $\mathbf{E}[X|Y]$ is a function of y.

Definition 2.4. The *variance* of X is defined as:

$$\text{var}(X) = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] \tag{17}$$

This can be rewritten into the often useful form $\text{var}(X) = \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2$, which we will illustrate for the discrete case:

$$\begin{aligned}
\text{var}(X) &= \mathbf{E}\left[(X - \mathbf{E}[X])^2\right] && (18)\\
&= \sum_x (x - \mathbf{E}[X])^2 f_X(x) && (19)\\
&= \sum_x \left(x^2 - 2x\mathbf{E}[X] + (\mathbf{E}[X])^2\right) f_X(x) && (20)\\
&= \sum_x x^2 f_X(x) - 2\mathbf{E}[X] \sum_x x f_X(x) + (\mathbf{E}[X])^2 \sum_x f_X(x) && (21)\\
&= \mathbf{E}\left[X^2\right] - 2(\mathbf{E}[X])^2 + (\mathbf{E}[X])^2 && (22)\\
&= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 && (23)
\end{aligned}$$

The law of total variance asserts that $\text{var}(X) = \mathbf{E}[\text{var}(X|Y)] + \text{var}(\mathbf{E}[X|Y])$, which we can show using the law of iterated expectation:

$$\begin{aligned}
\text{var}(X) &= \mathbf{E}\left[X^2\right] - (\mathbf{E}[X])^2 && (24)\\
&= \mathbf{E}\left[\mathbf{E}\left[X^2|Y\right]\right] - \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] && (25)\\
&= \mathbf{E}[\text{var}(X|Y)] + \mathbf{E}\left[(\mathbf{E}[X|Y])^2\right] - \mathbf{E}[\mathbf{E}[X|Y]]^2 && (26)\\
&= \mathbf{E}[\text{var}(X|Y)] + \text{var}(\mathbf{E}[X|Y]) && (27)
\end{aligned}$$

Definition 2.5. The *covariance* of X and Y is defined as:

$$\text{cov}(X, Y) = \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] \tag{28}$$

which can be rewritten:

$$\begin{aligned}
\text{cov}(X, Y) &= \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])] && (29)\\
&= \mathbf{E}[XY - \mathbf{E}[X]\,Y - \mathbf{E}[Y]\,X + \mathbf{E}[X]\,\mathbf{E}[Y]] && (30)\\
&= \mathbf{E}[XY] - \mathbf{E}[\mathbf{E}[X]\,Y] - \mathbf{E}[\mathbf{E}[Y]\,X] + \mathbf{E}[X]\,\mathbf{E}[Y] && (31)\\
&= \mathbf{E}[XY] - \mathbf{E}[X]\,\mathbf{E}[Y] && (32)
\end{aligned}$$

Note that if X and Y are independent, $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$ so $\text{cov}(X, Y) = 0$.

Definition 2.6. The *correlation coefficent* of X and Y is obtained from the covariance:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} \tag{33}$$

The correlation coefficient can be thought of as a "normalized" measure of the covariance of X and Y. If $\rho(X, Y) = 1$ X and Y are fully positively correlated; if $\rho(X, Y) = -1$ they are fully negatively correlated.

2.1 The variance of sums of random variables

Let $\tilde{X}_i = X_i - \mathbf{E}[X_i]$. Then

$$\begin{aligned}
\text{var}\left(\sum_{i=1}^{n} \tilde{X}_i\right) &= \mathbf{E}\left[\left(\sum_{i=1}^{n} \tilde{X}_i\right)^2\right] &(34)\\
&= \mathbf{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{X}_i \tilde{X}_j\right] &(35)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbf{E}\left[\tilde{X}_i \tilde{X}_j\right] &(36)\\
&= \sum_{i=1}^{n} \mathbf{E}\left[\tilde{X}_i^2\right] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbf{E}\left[\tilde{X}_i \tilde{X}_j\right] &(37)\\
&= \sum_{i=1}^{n} \text{var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \text{cov}\left(X_i, X_j\right) &(38)
\end{aligned}$$

2.2 Joint probability density functions

Given two random variables X and Y, their *joint PDF* is defined as:

$$f_{X,Y}(x,y) = P(X = x, Y = y) \quad (39)$$

We also define the *marginal PDFs* $f_X(x)$ and $f_Y(y)$ and the *conditional PDFs* $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$. We can obtain $f_X(()\,x)$ by *marginalizing* the joint PDF:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy \quad (40)$$

The definition of conditional probability can be applied to obtain:

$$f_{X|Y}(x,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad (41)$$

Combining these, a different expression for the marginal PDF is:

$$f_X(x) = \int_{-\infty}^{\infty} f_Y(y) f_{X|Y}(x|y)\, dy \quad (42)$$

2.3 Convolutions

Definition 2.7. Suppose X and Y are independent random variables with PDFs f_X, f_Y, respectively. The PDF f_W representing the distribution of $W = X + Y$ is known as the *convolution* of f_X and f_Y. To derive the distribution f_W we start with the CDF:

$$\begin{aligned}
P(W \leq w|X = x) &= P(X + Y \leq w|X = x) &(43)\\
&= P(x + Y \leq w|X = x) &(44)\\
&\overset{\text{independence}}{=} P(x + Y \leq w) &(45)\\
&= P(Y \leq w - x) &(46)
\end{aligned}$$

This is a CDF of Y. Next we differentiate both sides with respect to w to obtain the PDF:

$$f_{W|X}(w|x) = f_Y(w-x) \tag{47}$$
$$f_X(x)f_{W|X}(w|x) = f_X(x)f_Y(w-x) \tag{48}$$
$$f_{X,W}(x,w) \overset{\text{conditional prob.}}{=} f_X(x)f_Y(w-x) \tag{49}$$
$$f_W(w) \overset{\text{marginalization}}{=} \int_{-\infty}^{\infty} f_X(x)f_Y(w-x)\,dx \tag{50}$$

3 Least squares estimation

Suppose we are given the value of a random variable Y that is somehow related to the value of an unknown variable X. In other words, Y is some form of "measurement" of X. How can we compute an estimate c of the value of X given Y that minimizes the squared error $(X-c)^2$?

First, consider an arbitrary c. Then the *mean squared error* is:

$$\mathbf{E}\left[(X-c)^2\right] = \text{var}\,(X-c) + (\mathbf{E}\,[X-c])^2 = \text{var}\,(X) + (\mathbf{E}\,[X]-c)^2 \tag{51}$$

by Equation 23. If we are given no measurements, we should pick the value of c that minimizes this equation. Since $\text{var}\,(X)$ is independent of c, we choose $c = \mathbf{E}\,[X]$ which eliminates the second term.

Now suppose we are given a measurement $Y = y$. Then to minimize the *conditional* mean squared error, we should choose $c = \mathbf{E}\,[X|Y=y]$. This value is the *least squares estimate of X given Y*. (The proof is omitted.) Note that we have said nothing yet about the relationship between X and Y. In general, the estimate $\mathbf{E}\,[X|Y=y]$ is a function of y, which we refer to as an *estimator*.

3.1 Estimation error

Let $\hat{X} = \mathbf{E}\,[X|Y]$ be the least squares estimate of X, and $\tilde{X} = X - \hat{X}$ be the *estimation error*. The estimation error exhibits the following properties:

- $\tilde{X}$ is *zero mean*:

$$\mathbf{E}\left[\tilde{X}|Y\right] = \mathbf{E}\left[X-\hat{X}|Y\right] = \mathbf{E}\,[X|Y] - \mathbf{E}\left[\hat{X}|Y\right] = \hat{X}-\hat{X} = 0 \tag{52}$$

  (Note that $\mathbf{E}\left[\hat{X}|Y\right] = \hat{X}$ since $\hat{X}$ is completely determined by Y.)

- $\tilde{X}$ and the estimate $\hat{X}$ are uncorrelated; using $\mathbf{E}\left[\tilde{X}|Y\right] = 0$:

$$\text{cov}\left(\hat{X},\tilde{X}\right) = \mathbf{E}\left[(\hat{X}-\mathbf{E}\left[\hat{X}\right])(\tilde{X}-\mathbf{E}\left[\tilde{X}\right])\right] \tag{53}$$
$$\overset{\text{iter. exp.}}{=} \mathbf{E}\left[(\hat{X}-\mathbf{E}\,[X|Y])\tilde{X}\right] \tag{54}$$
$$= \mathbf{E}\left[(\hat{X}-\mathbf{E}\,[X])\tilde{X}|Y\right] \tag{55}$$
$$= (\hat{X}-\mathbf{E}\,[X])\mathbf{E}\left[\tilde{X}|Y\right] \tag{56}$$
$$= 0 \tag{57}$$

- Because $X = \tilde{X} + \hat{X}$, the $\text{var}\,(X)$ can be decomposed based on Equation 38:

$$\text{var}\,(X) = \text{var}\left(\hat{X}\right) + \text{var}\left(\tilde{X}\right) + 2\text{cov}\left(\hat{X},\tilde{X}\right) = \text{var}\left(\hat{X}\right) + \text{var}\left(\tilde{X}\right) \tag{58}$$

3.2 Linear least squares

Suppose we have the *linear estimator* $X = aY + b$. In other words, the random variable X is a linear function of the random variable Y. Our goal is to find values for the coefficients a and b that minimize the mean squared estimation error $\mathbf{E}\left[(X - aY - b)^2\right]$.

First, suppose a is fixed. Then by Equation 51 we choose:

$$b = \mathbf{E}[X - aY] = \mathbf{E}[X] - a\mathbf{E}[Y] \tag{59}$$

Substituting this into our objective and manipulating, we obtain:

$$\begin{aligned}
\mathbf{E}\left[(X - aY - \mathbf{E}[X] + a\mathbf{E}[Y])^2\right] &= \mathrm{var}(X - aY) & (60)\\
&= \mathrm{var}(X) + a^2\mathrm{var}(Y) + 2\mathrm{cov}(X, -aY) & (61)\\
&= \mathrm{var}(X) + a^2\mathrm{var}(Y) - 2a\mathrm{cov}(X, Y) & (62)
\end{aligned}$$

Our goal is to minimize this quantity with respect to a. Since it is quadratic in a, it is minimized when its derivative with respect to a is zero, i.e.:

$$\begin{aligned}
0 &= 2a\mathrm{var}(Y) - 2\mathrm{cov}(X, Y) & (63)\\
\frac{\mathrm{cov}(X, Y)}{\mathrm{var}(Y)} &= a & (64)\\
\rho\frac{\mathrm{var}(X)}{\mathrm{var}(Y)} &= a & (65)
\end{aligned}$$

The mean squared error of our estimate is then:

$$\begin{aligned}
& \mathrm{var}(X) + a^2\mathrm{var}(Y) - 2a\mathrm{cov}(X, Y) & (66)\\
=\ & \mathrm{var}(X) + \rho^2\frac{\mathrm{var}(X)}{\mathrm{var}(Y)}\mathrm{var}(Y) - 2\rho\frac{\sqrt{\mathrm{var}(X)}}{\sqrt{\mathrm{var}(Y)}}\rho\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)} & (67)\\
=\ & \left(1 - \rho^2\right)\mathrm{var}(X) & (68)
\end{aligned}$$

The basic idea behind the linear least squares estimator is to start with the baseline estimate $\mathbf{E}[X]$ for X, and then adjust the estimate by taking into account the value of $Y - \mathbf{E}[Y]$ and the correlation between X and Y.

4 Normal random variables

The univariate Normal distribution with mean μ and variance σ^2, denoted $N(\mu, \sigma)$, is defined as:

$$N(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2} \tag{69}$$

The Standard Normal distribution is the particular case where $\mu = 0$ and $\sigma = 1$, i.e.:

$$N(0, 1) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{70}$$

The cumulative density function of the Standard Normal (The Standard Normal CDF), denoted Φ, is thus:

$$\Phi(y) = P(Y \le y) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} e^{-t^2/2}\,dt \tag{71}$$

Note that since $N(0,1)$ is symmetric, $\Phi(-y) = 1 - \Phi(y)$:

$$\Phi(-y) = P(Y \leq -y) = P(Y \geq y) = 1 - P(Y < y) = 1 - \Phi(y) \tag{72}$$

Finally, the CDF of any random variable $X \sim N(\mu, \sigma)$ can be expressed in terms of the Standard Normal CDF. First, by simple manipulation:

$$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) \tag{73}$$

We see that

$$\mathbf{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{\mathbf{E}[X]-\mu}{\sigma} = 0 \tag{74}$$

$$\text{var}\left(\frac{X-\mu}{\sigma}\right) = \frac{\text{var}(X)}{\sigma^2} = 1 \tag{75}$$

So $Y = (X-\mu)/\sigma \sim N(0,1)$ and the CDF is:

$$P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right) \tag{76}$$

5 Limit theorems

We first examine the asymptotic behavior of sequences of random variables. Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed, each with mean μ and variance σ^2, and let $S_n = \sum_i X_i$. Then

$$\text{var}(S_n) = \sum_i \text{var}(X_i) = n\sigma^2 \tag{77}$$

So as n increases, the variance of S_n does not converge. Instead, consider the *sample mean* $M_n = S_n/n$. M_n converges as follows:

$$\mathbf{E}[M_n] = \frac{1}{n}\sum_i \mathbf{E}[X_i] = \mu \tag{78}$$

$$\text{var}(M_n) = \sum_i \text{var}(X_i)\, n = \frac{1}{n^2}\sum_i \text{var}(X_i) = \frac{\sigma^2}{n} \tag{79}$$

So $\lim_{n\to\infty} \text{var}(M_n) = 0$, i.e. as the number of samples n increases, the sample mean tends to the true mean.

5.1 Central limit theorem

Suppose X_i are defined as above. Let

$$Z_n = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \tag{80}$$

The *Central limit theorem*, which we will not prove, states that as n increases, the CDF of Z_n tends to $\Phi(z)$ (the Standard Normal CDF). In other words, *the sum of a large number of random variables is approximately normally distributed*.

5.2 Markov inequality

For a random variable $X > 0$, define random variable Y as follows:

$$Y = \begin{cases} 0 & \text{if } X < a \\ 1 & \text{otherwise} \end{cases} \tag{81}$$

Clearly $Y \leq X$ so $\mathbf{E}[Y] \leq \mathbf{E}[X]$. Furthermore, by the definition of expectation, $\mathbf{E}[Y] = 0 \cdot P(X < a) + aP(X \geq a)$ so

$$aP(X \geq a) \leq \mathbf{E}[X] \tag{82}$$

$$P(X \geq a) \leq \frac{\mathbf{E}[X]}{a} \tag{83}$$

Equation 83 is known as the *Markov inequality*, which essentially asserts that if a nonnegative random variable has a small mean, the probability that variable takes a large value is also small.

5.3 Chebyshev inequality

Let X be a random variable with mean μ and variance σ^2. By the Markov inequality,

$$P\left((X-\mu)^2 \geq c^2\right) \leq \frac{\mathbf{E}\left[(X-\mu)^2\right]}{c^2} = \frac{\sigma^2}{c^2} \tag{84}$$

Since $P\left((X-\mu)^2 \geq c^2\right) = P(|X-\mu| \geq c)$,

$$P(|X-\mu| \geq c) \leq \frac{\sigma^2}{c^2} \tag{85}$$

Equation 85 is known as the *Chebyshev inequality*. The Chebyshev inequality is often rewritten as:

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{86}$$

In other words, the probability that a random variable takes a value more than k standard deviations from its mean is at most $1/k^2$.

5.4 Weak law of large numbers

Applying the Chebyshev inequality to the sample mean M_n, and using Equations 78 and 79, we obtain:

$$P(|M_n-\mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2} \tag{87}$$

In other words, for large n, the bulk of the distribution of M_n is concentrated near μ. A common application is to fix ϵ and compute the number of samples needed to guarantee that the sample mean is an accurate estimate.

5.5 Jensen's inequality

Let $f(x)$ be a convex function, i.e. $d^2f/dx^2 > 0$ for all x. First, note that if $f(x)$ is convex, then the first order Taylor approximation of $f(x)$ is an underestimate:

$$f(x) \overset{\text{Fund. Thm. of Calculus}}{=} f(a) + \int_a^x f'(t)\,dt \tag{88}$$

$$\overset{\text{Taylor approx.}}{\geq} \quad f(a) + \int_a^x f'(a)\,dt \tag{89}$$

$$= \quad f(a) + (x-a)f'(a) \tag{90}$$

Thus if X is a random variable,

$$f(a) + (X-a)f'(a) \leq f(X) \tag{91}$$

Now, let $a = \mathbf{E}[X]$. Then we have

$$f(\mathbf{E}[X]) + (\mathbf{E}[X] - \mathbf{E}[X])f'(\mathbf{E}[X]) \leq \mathbf{E}[f(X)] \tag{92}$$

$$f(\mathbf{E}[X]) \leq \mathbf{E}[f(X)] \tag{93}$$

Equation 93 is known as *Jensen's inequality*.

5.6 Chernoff bound

Finally we turn to the Chernoff bound, a powerful technique for bounding the probability that a random variable deviates far from its expectation. First, observe that the Chebyshev inequality provides a *polynomial* bound on the probability that X takes a value in the "tails" of its density function.

The "Chernoff-type" bounds, on the other hand, are *exponential*. We define such a bound as follows. Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables. Assume that

$$\mathbf{E}[X_1] = \mathbf{E}[X_2] = \ldots = \mathbf{E}[X_n] = \mu < \infty$$

and that

$$\text{var}(X_1) = \text{var}(X_2) = \ldots \text{var}(X_n) = \sigma^2 < \infty$$

Further, let $X = \sum_{i=1}^{n} X_i$, so that $\mathbf{E}[X] = n\mu$ and $\text{var}(X) = n\sigma^2$. The Chernoff bound states that, for $t > 0$ and $0 \leq X_i \leq 1$, $\forall i$ such that $1 \leq i \leq n$,

$$P(|X - n\mu_X| \geq nt) \leq 2e^{-2nt^2} \tag{94}$$

Note that this bound is significantly better than that of the Chebyshev inequality. Chebyshev decreases in a manner inversely proportional to n, whereas the Chernoff bound decreases exponentially with n.

We now proove the bound stated in equation 94. In particular, we will proove the bound for the case

$$P(X - n\mu \geq nt) \leq e^{-2nt^2}$$

The proof for the second case,

$$P(X - n\mu \leq -nt) \leq e^{-2nt^2}$$

is very similar. The complete bound is merely the sum of these two probabilities.

Proof: We first define the function

$$f(x) = \begin{cases} 1 & \text{if } X - n\mu \geq nt \\ 0 & \text{if } X - n\mu < nt \end{cases}$$

Note that

$$\mathbf{E}[f(x)] = P(X - n\mu \geq nt) \tag{95}$$

which is exactly the probability we are interested in computing.

Lemma 5.1. *For all positive reals h,*

$$f(x) \leq e^{h(X-n\mu-nt)}$$

Proof: *If $X - n\mu - nt \geq 0$, then $f(x) = 1$ and $e^{h(X-n\mu-nt)} \geq 1$. Note that this condition holds only for* all positive reals. □

So, we now have that

$$\mathbf{E}\left[f(x)\right] \leq \mathbf{E}\left[e^{h(X-n\mu-nt)}\right] \tag{96}$$

We will concentrate on bounding the above expectation, and then minimizing it with respect to h. Let us first manipulate the expectation as follows:

$$\begin{aligned}
\mathbf{E}\left[e^{h(X-n\mu-nt)}\right] &= \mathbf{E}\left[e^{h[(X_1+X_2+\ldots+X_n)-n\mu-nt]}\right] \\
&= \mathbf{E}\left[e^{-hnt}\cdot e^{h(X_1-\mu)+h(X_2-\mu)+\ldots+(X_n-\mu)}\right] \\
&= e^{-hnt}\mathbf{E}\left[\prod_{i=1}^{n} e^{h(X_i-\mu)}\right]
\end{aligned}$$

So,

$$\mathbf{E}\left[e^{h(X-n\mu-nt)}\right] \overset{\text{independence}}{=} e^{-hnt}\prod_{i=1}^{n}\mathbf{E}\left[e^{h(X_i-\mu)}\right] \tag{97}$$

Lemma 5.2. *Let Y be a random variable such that $0 \leq Y \leq 1$. Then, for any real number $h \geq 0$,*

$$\mathbf{E}\left[e^{hY}\right] \leq (1-\mathbf{E}\left[Y\right]) + \mathbf{E}\left[Y\right]e^{h}$$

Proof: *This follows directly from the definition of convexity.* □

So, using equation 97 and lemma 5.2, we have that

$$e^{-hnt}\prod_{i=1}^{n}\mathbf{E}\left[e^{h(X_i-\mu)}\right] \leq e^{-hnt}\prod_{i=1}^{n}\mathbf{E}\left[e^{-h\mu}\left((1-\mu)+\mu e^{h}\right)\right]$$

Lemma 5.3.

$$e^{-h\mu}\left((1-\mu)+\mu e^{h}\right) \leq e^{h^2/8} \tag{98}$$

Proof: *First,*

$$e^{-h\mu}\left((1-\mu)+\mu e^{h}\right) = e^{-h\mu+\ln\left((1-\mu)+\mu e^{h}\right)}$$

Let

$$L(h) = -h\mu + \ln\left((1-\mu)+\mu e^{h}\right)$$

Taking the Taylor series expansion,

$$\begin{aligned}
L'(h) &= -\mu + \frac{\mu e^{h}}{(1-\mu)+\mu e^{h}} = -\mu + \frac{\mu}{(1-\mu)e^{-h}+\mu} \\
L''(h) &= \frac{u(1-\mu)e^{-h}}{\left((1-\mu)e^{-h}+\mu\right)^2} \leq \frac{1}{4}
\end{aligned}$$

So, we see that the Taylor series is

$$\begin{aligned} L(h) &= L(0) + L'(0)h + L''(0)\frac{h^2}{2!} + \dots \\ &\le \frac{h^2}{8} \end{aligned}$$

□

Combining equations 95,96,97 and 98, we have that

$$\begin{aligned} \mathbf{E}\left[f(x)\right] &= P\left(X - n\mu \ge nt\right) \\ &\le e^{-hnt}\prod_{i=1}^{n} e^{h^2/8} \\ &= e^{-hnt}e^{nh^2/8} \\ &= e^{-hnt+nh^2/8} \end{aligned}$$

So,

$$\mathbf{E}\left[f(x)\right] \le e^{-hnt+nh^2/8} \tag{99}$$

Now we minimize this equation over all positive reals h. Taking the derivative of $(-hnt + nh^2/8)$, we find that $(e^{-hnt+nh^2/8})$ is minimized when $h = 4t$. Subsituting this into 99, we see that

$$P\left(X - n\mu \ge nt\right) \le e^{-2nt^2} \tag{100}$$

which is our objective. □

5.6.1 Extension of the Chernoff Bound

One of the conditions for the Chernoff bound we have just proven to hold is that $0 \le X_i \le 1$. We can generalize the bound to address this constraint. If $X_1, X_2, \dots, X_n$ are independent, identically distributed random variables such that $\mathbf{E}\left[X_i\right] = \mu < \infty, \forall i$ and $\text{var}\left(X_i\right) = \sigma^2 < \infty, \forall i$, and $a_i \le X_i \le b_i$ for some constants a_i and b_i for all i, then for all $t > 0$

$$P\left(|X - n\mu| \ge nt\right) \le 2e^{\frac{-2n^2t^2}{\sum_{i=1}^{n}(a_i-b_i)^2}} \tag{101}$$

We will not prove this bound here.