

ECE 235: Problem Set 4 Addendum
(and brief notes on the Chernoff bound)

Assigned: Thursday, October 29

Due: Thursday November 5 (by noon, in course homework box)

Reading: Hajek, Chapter 2; Lecture notes

Topics: LLN and Chernoff bound

Reminder: Midterm is in class, Tue November 10.

Summary of Chernoff bound: Let X_1, X_2, \dots be i.i.d. with mean $\mathbb{E}[X_i] \equiv m$, and let $\Lambda(\theta) = \log \mathbb{E}[e^{\theta X_1}]$ denote the log moment generating function.

The Chernoff bound for a single random variable is

$$P[X_1 > a] \leq e^{-(a\theta - \Lambda(\theta))}, \quad \theta > 0 \quad (1)$$

We can now optimize this over $\theta > 0$ to get the best bound. This bound is only useful for tail probabilities of being larger than the mean. For $a < m$, the optimal value of the bound is one, a trivial answer.

Similarly,

$$P[X_1 < a] \leq e^{-(a\theta - \Lambda(\theta))}, \quad \theta < 0 \quad (2)$$

We can now optimize this over $\theta < 0$ to get the best bound. This bound is only useful for tail probabilities of being smaller than the mean. For $a > m$, the optimal value of the bound is one, a trivial answer.

When we are in the non-trivial regime (i.e., for tail probabilities of being away from the mean), the optimized Chernoff bound for a single random variable is given in both cases by the same expression:

$$\begin{aligned} P[X_1 > a] &\leq e^{-\ell(a)}, & a > m \\ P[X_1 < a] &\leq e^{-\ell(a)}, & a < m \end{aligned} \quad (3)$$

where

$$\ell(a) = \max_{\theta} a\theta - \Lambda(\theta)$$

As we showed in class, the maximum above is attained for $\theta > 0$ (and is positive) if $a > m$, and it is attained for $\theta < 0$ (and is positive) if $a < m$, so that (3) is consistent with (1) and (2).

Large deviations: Applying this to a sum of i.i.d. random variables, we get

$$\begin{aligned} P[S_n = X_1 + \dots + X_n > na] &\leq e^{-n\ell(a)}, & a > m \\ P[S_n = X_1 + \dots + X_n < na] &\leq e^{-n\ell(a)}, & a < m \end{aligned} \quad (4)$$

That is, the probability of large (i.e., $O(n)$) deviations away from the mean nm decays exponentially fast with n . It can actually be shown that the exponent of decay given by the Chernoff bound is asymptotically tight. That is,

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \log P[S_n = X_1 + \dots + X_n > na] &= -\ell(a), & a > m \\ \lim_{n \rightarrow \infty} \frac{1}{n} \log P[S_n = X_1 + \dots + X_n < na] &= -\ell(a), & a < m \end{aligned} \quad (5)$$

This result is called Cramer's theorem.

The purpose of the addendum to this problem set is to help you fill in some details, and to give you a glimpse of some applications.

Problem 7 (do not turn in): For any random variable X , show that $\Lambda(\theta) = \log \mathbb{E}[e^{\theta X}]$ is convex, using the following steps:

(a) Find the second derivative of $\Lambda(\theta)$, assuming that derivative with respect to θ can be exchanged with expectation with respect to the distribution of X .

(b) Show that the second derivative is nonnegative if and only if

$$\mathbb{E}[e^{\theta X}] \mathbb{E}[X^2 e^{\theta X}] \geq \left(\mathbb{E}[X e^{\theta X}] \right)^2$$

(c) Use the Cauchy-Schwarz inequality to show that the above inequality holds.

Remark: Recall that the Cauchy-Schwarz inequality states that $(\mathbb{E}[UV])^2 \leq \mathbb{E}[U^2]\mathbb{E}[V^2]$.

(d) Conclude that $\Lambda(\theta)$ is convex.

Problem 8 (do not turn in): Consider $g(\theta) = a\theta - \Lambda(\theta)$, the function to be optimized for the Chernoff bound. From Problem 7, $\Lambda(\theta)$ is convex, and hence $g(\theta)$ is concave. Thus, the global maximum of g can be found by setting its derivative to zero.

(a) Show that $g(0) = 0$ and $g'(0) = a - m$, where $m = \mathbb{E}[X]$.

(b) Conclude that the global maximum of $g(\theta)$ occurs for $\theta > 0$ if $a > m$, and for $\theta < 0$ if $a < m$. In each case, the maximum $\ell(a)$ is positive. These are the scenarios in which the Chernoff bound is non-trivial.

(c) Conclude also that $\max_{\theta > 0} g(\theta) = g(0) = 0$ when $a < m$, and that $\max_{\theta < 0} g(\theta) = g(0) = 0$ when $a > m$, which are the scenarios in which the Chernoff bound is trivial.

Problem 9: Suppose that $\{X_i\}$ are i.i.d. Bernoulli(p).

(a) Show that the optimized Chernoff bound on $P[X_1 + \dots + X_n < na]$ ($a < m$) or $P[X_1 + \dots + X_n > na]$ ($a > m$) is given by $e^{-n\ell(a)}$, where

$$\ell(a) = a \log \frac{a}{p} + (1-a) \log \frac{1-a}{1-p}$$

(This was a group exercise in class.)

(b) We now want to prove that $\ell(a) > 0$ for $a \neq p$. Let us take a more general approach. Let $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$, where p, q are two probability mass functions with argument x taking values from the same set. This is a fundamental quantity in hypothesis testing and information theory, and is called the divergence between the two distributions. Using the fact that $\log x \leq x - 1$ (equality if and only if $x = 1$), show that $-D(p||q) = \sum_x p(x) \log \frac{q(x)}{p(x)} \leq 0$, or $D(p||q) \geq 0$, with equality if and only if p and q are the same.

(c) Apply (b) to (a) to conclude that $\ell(a) > 0$ if $a \neq p$.

Problem 10: Suppose that there are two coins, Coin 1 is fair and Coin 2 is biased, with probability of heads equal to $p > 1/2$. We are given one of these coins and asked to figure out which one it is by doing repeated tosses (assumed to be conditionally independent, once the coin is chosen). Let X_1, X_2, \dots denote the outcomes of the tosses, with $X_i = 1$ denoting a head on the i th toss, and $X_i = 0$ denoting a tail. We use the following decision rule: if $\frac{X_1 + \dots + X_n}{n} > t$, then we guess that Coin 2 was chosen. Otherwise we guess that Coin 1 was chosen.

(a) For $p = 0.7$, for what choices of t are we guaranteed that the probability of error (i.e., of guessing the wrong coin) tends to zero as the number of tosses gets large.

(b) For $p = 0.7$ and $t = 0.55$, use the Chernoff bound to estimate the smallest number of tosses needed to get the probability of error below 10^{-4} .

Problem 11: Suppose that Coin 2 is chosen in the setting of Problem 10. Let $p(x_1, \dots, x_n)$ denote the probability of the first n outcomes being x_1, \dots, x_n .

(a) Show that we can write the log of the joint pmf as follows:

$$\log p(x_1, \dots, x_n) = \sum_{i=1}^n (x_i \log p + (1-x_i) \log(1-p))$$

(b) Replacing the specific values x_1, \dots, x_n by the random variables X_1, \dots, X_n , find the limit

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log p(X_1, \dots, X_n)$$

as $n \rightarrow \infty$. This is called the *entropy* of the sequence. Provide a numerical value for the entropy when $p = 0.7$.

Remark: You can find out more about entropy and similar quantities in classes on information theory, data compression, and physics. The LLN result in (b) can be used to conclude that, even though there are 2^n possible results of n coin tosses, as n gets large, with overwhelming probability, only a fraction of these sequences, called the *typical sequences*, occur, when the coin is biased. This implies, for example, that we can *compress* these sequences, or represent them with fewer than n bits, as long as we can tolerate some inaccuracy.