

## Probability inequalities<sup>11</sup>

There is an adage in probability that says that behind every limit theorem lies a probability inequality (i.e., a bound on the probability of some undesired event happening). Since a large part of probability theory is about proving limit theorems, people have developed a bewildering number of inequalities.

Luckily, we'll only need a few key inequalities. Even better, three of them are really just versions of one another. **Exercise 29:** Can you think of example distributions for which each of the following inequalities are tight (that is, the inequalities may be replaced by equalities)? Are these "extremal" distributions unique?

### Markov's inequality



Figure 5: Markov.

For a nonnegative r.v.  $X$ ,

$$P(X > u) \leq \frac{E(X)}{u}.$$

So if  $E(X)$  is small and we know  $X \geq 0$ , then  $X$  must be near zero with high probability. (Note that the inequality is *not* true if  $X$  can be negative.)

---

<sup>11</sup>HMC 1.10

The proof is really simple:

$$uP(X > u) \leq \int_u^\infty tp_X(t)dt \leq \int_0^\infty tp_X(t)dt = E(X). \quad \square$$

### Chebyshev's inequality



Figure 6: Chebyshev.

$$P(|X - E(X)| > u) \leq \frac{V(X)}{u^2},$$

aka

$$P\left(\frac{|X - E(X)|}{\sigma(X)} > u\right) \leq \frac{1}{u^2}.$$

Proof: just look at the (nonnegative) r.v.  $(X - E(X))^2$ , and apply Markov.

So if the variance of  $X$  is really small,  $X$  is close to its mean with high probability.

### Chernoff's inequality

$$P(X > u) = P(e^{sX} > e^{su}) \leq e^{-su} M(s).$$

So the mgf controls the size of the tail of the distribution — yet another surprising application of the mgf idea.



Figure 7: Chernoff.

The really nice thing about this bound is that it is easy to deal with sums of independent r.v.'s (recall our discussion above of mgf's for sums of independent r.v.'s). **Exercise 30:** Derive Chernoff's bound for sums of independent r.v.'s (i.e., derive an upper bound for the probability that  $\sum_i X_i$  is greater than  $u$ , in terms of  $M(X_i)$ ).

The other nice thing is that the bound is exponentially decreasing in  $u$ , which is much stronger than Chebyshev. (On the other hand, since not all r.v.'s have mgf's, Chernoff's bound can be applied less generally than Chebyshev.)

The other other nice thing is that the bound holds for all  $s$  simultaneously, so if we need as tight a bound as possible, we can use

$$P(X > u) \leq \inf_s e^{-su} M(s),$$

i.e., we can minimize over  $s$ .

### Jensen's inequality

This one is more geometric. Think about a function  $g(u)$  which is curved upward, that is,  $g''(u) \geq 0$ , for all  $u$ . Such a  $g(u)$  is called "convex." (Downward-curving functions are called "concave." More generally, a convex



Figure 8: Jensen.

function is bounded above by its chords:

$$g(tx + (1 - t)y) \leq tg(x) + (1 - t)g(y),$$

while a concave function is bounded below.)

Then if you draw yourself a picture, it's easy to see that

$$E(g(X)) \geq g(E(X)).$$

That is, the average of  $g(X)$  is always greater than or equal to  $g$  evaluated at the average of  $X$ . **Exercise 31:** Prove this. (Hint: try subtracting off  $f(X)$ , where  $f$  is a linear function of  $X$  such that  $g(X) - f(X)$  reaches a minimum at  $E(X)$ .)

**Exercise 32:** What does this inequality tell you about the means of  $1/X$ ? of  $-X \log X$ ? About  $E_i(X)$  vs.  $E_j(X)$ , where  $i > j$ ?

### Cauchy-Schwarz inequality

$$|C(X, Y)| \leq \sigma(X)\sigma(Y),$$

that is, the correlation coefficient is bounded between  $-1$  ( $X$  and  $Y$  are anti-correlated) and  $1$  (correlated).

The proof of this one is based on our rules for adding variance:

$$C(X, Y) = \frac{1}{2}[V(X + Y) - V(X) - V(Y)]$$

(assuming  $E(X) = E(Y) = 0$ ). **Exercise 33:** Complete the proof. (Hint: try looking at  $X$  and  $-X$ , using the fact that  $C(-X, Y) = -C(X, Y)$ .)

Exercise for the people who have taken linear algebra: interpret the Cauchy-Schwarz inequality in terms of the angle between the vectors  $X$  and  $Y$  (where we think of functions — that is, r.v.'s — as vectors, and define the dot product as  $E(XY)$  and the length of a vector as  $\sqrt{E(X^2)}$ ). Thus this inequality is really geometric in nature.

## Limit theorems<sup>12</sup>

*An approximate answer to the right question is worth a great deal more than a precise answer to the wrong question.*

The first golden rule of applied mathematics, sometimes attributed to John Tukey

### (Weak) law of large numbers

Chebyshev's simple inequality is enough to prove perhaps the fundamental result in probability theory: the law of averages. This says that if we take the sample average of a bunch of i.i.d. r.v.'s, the sample average will be close to the true average. More precisely, under the assumption that  $V(X) < \infty$ , then

$$P\left(\left|E(X) - \frac{1}{N} \sum_{i=1}^N X_i\right| > \epsilon\right) \rightarrow 0$$

as  $N \rightarrow \infty$ , no matter how small  $\epsilon$  is.

The proof:

$$E\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = E(X),$$

by the linearity of the expectation.

$$V\left(\frac{1}{N} \sum_{i=1}^N X_i\right) = \frac{V(X)}{N},$$

by the rules for adding variance and the fact that  $X_i$  are independent.

Now just look at Chebyshev.  $\square$

Remember, the LLN does not hold for all r.v.'s: remember what happened when you took averages of i.i.d. Cauchy r.v.'s? **Exercise 34:** What goes wrong in the Cauchy case?

### Stochastic convergence concepts

In the above, we say that the sample mean  $\frac{1}{N} \sum_{i=1}^N X_i$  "converges in probability" to the true mean. More generally, we say r.v.'s  $Z_N$  converge to  $Z$  in

---

<sup>12</sup>HMC chapter 4