

## Large deviations

This is a companion to Durrett's approach to Section 1.9, just done in a different order. You should read both this and Durrett's version, since I will not duplicate all the proofs.

The laws of large numbers show us that  $\mathbb{P}(S_n > an) \rightarrow 0$  for any  $a > \mu := \mathbb{E}X_1$ . Suppose we want to know how fast this goes to zero. If  $\mathbb{E}X_1^2 < \infty$ , we can use Chebyshev's inequality to see that  $\mathbb{P}(S_n > an) \leq \text{Var}(X_1)/a^2$ . The Central limit theorem gives much better bounds, but only in the range  $\mathbb{P}(S_n > \mu n + a\sqrt{n})$ . The range  $S_n > an$  with  $a > \mu$  fixed and  $n \rightarrow \infty$  is called a **large deviation**. When  $\mathbb{E}\exp(\lambda X_1)$  exists for some  $\lambda > 0$ , then it is possible to get bounds that are pretty sharp, as well as being exponentially small. The steps are as follows.

1. Compute an upper bound, depending on  $\lambda$ , using Markov's inequality.
2. Optimize in  $\lambda$ . This gives  $\mathbb{P}(S_n > an) \leq \exp(-h(a)n)$  for some positive function  $h$ . This turns out to be sharp in the sense that  $n^{-1} \log \mathbb{P}(S_n > an) \rightarrow h(a)$ .
3. To see this is sharp, we show "how it happens that  $S_n > na$ ". That is, we find an event whose probability we can compute that is contained in the event  $\{S_n > an\}$  and this is the desired lower bound on  $\mathbb{P}(S_n > an)$ .

Finding this event involves understanding the **tilted** distributions, whose Radon-Nikodym derivative with respect to the original (see below for explanation) is  $ce^{\lambda x}$ . One final step is to make sure all the computations work the way we expect; it is these lemmas that Durrett begins with; I don't like doing the computations before knowing why we do them.

### The optimal upper bound

Markov's inequality gives, for any fixed  $\lambda > 0$ ,

$$\mathbb{P}(S_n > an) \leq e^{-\lambda an} \mathbb{E}e^{\lambda S_n}.$$

Taking logs and dividing by  $n$ , and recalling that  $\mathbb{E}e^{\lambda S_n} = (\mathbb{E}e^{\lambda X_1})^n$ , gives

$$\frac{1}{n} \log \mathbb{P}(S_n > an) \leq -\lambda a + \psi(\lambda)$$

where

$$\psi(\lambda) := \log \phi(\lambda) := \log \mathbb{E} e^{\lambda X_1}.$$

Suppose, for the moment, that we have found the optimal such  $\lambda$ ; call it  $\lambda_0(a)$ . Here's what we will prove:

**Theorem 1** *The function  $\psi$  is convex. If  $\inf_{\lambda} \psi'(\lambda) < a < \sup_{\lambda} \psi'(\lambda)$  then there is a unique  $\lambda_0$  with  $\psi'(\lambda_0) = a$ , and this is also the unique  $\lambda_0$  minimizing  $\psi(\lambda) - \lambda a$ . For this value of  $\lambda_0$ ,*

$$\frac{1}{n} \log \mathbb{P}(S_n > an) \rightarrow -h(a) := \psi(\lambda_0) - \lambda_0 a.$$

Note: we have already proved one direction of this inequality, and the facts about the dually unique  $\lambda_0$  turn out to involve elementary calculus, so the only hard part is in seeing the other direction of the inequality:  $\liminf_n n^{-1} \log \mathbb{P} \geq -h(a)$ .

One further remark: for any convex function  $\psi$ , the **convex conjugate** is defined to be the function  $a \mapsto -\inf_{\lambda} \psi(\lambda - a\lambda)$ . This is always convex; thus in particular, the large deviation rate function  $h$  is convex. A wealth of information about this may be found in [Roc66].

## Radon-Nikodym derivatives

Let  $\pi$  denote the law of  $X_1$  and let us define a new measure  $\nu$ , whose Radon-Nikodym derivative with respect to  $\pi$  is  $c e^{\lambda x}$ , for the value of  $c$  that makes  $\nu$  a probability measure. For those of you who haven't seen this before, a measure  $\nu$  on  $(\Omega, \mathcal{F})$  has Radon-Nikodym derivative  $g$  with respect to  $\pi$  on  $(\Omega, \mathcal{F})$  (written  $d\nu/d\pi = g$ ) if and only if

$$\nu(A) = \int_A g d\pi$$

for all measurable sets  $A \in \mathcal{F}$ .

In the special case where  $\pi$  is supported on a finite set  $\{x_1, \dots, x_n\}$ , then so is  $\nu$  and the Radon-Nikodym derivative is just the likelihood ratio:

$$\frac{d\nu}{d\pi}(x_j) = \frac{\nu(x_j)}{\pi(x_j)}.$$

This case is all you really need to understand: in the original large deviation paper [Che52], the entire proof is done approximating  $\pi$ , and hence  $\nu$  by discrete distributions. Another familiar example is when  $\pi$  is Lebesgue measure: then  $\nu$  is the measure with density  $g$ . The measure  $\nu$  is a probability measure if and only if  $\int g d\pi = 1$ . In general, when  $\Omega$  is the real numbers, then the CDF's,  $F$  and  $G$ , of  $\nu$  and  $\pi$  are related by (note that neither  $F$  nor  $G$  need have a density):

$$G(x) := \int_{-\infty}^x g(t) dF(t). \quad (1)$$

The appendix section A.8 will tell you all you need to know about Radon-Nikodym derivatives, and in particular, how to start with  $\nu$  and  $\pi$  and determine whether there is a  $g$  such that  $d\nu/d\pi = g$ . Going the other way (getting  $\nu$  from  $\pi$  and  $g$ ) is all we need this semester, and is much easier; those of you staying for next semester will get the full explanation. The one fact I will quote without proof is that if  $g > 0$  and we construct the measure  $\nu$  to have  $d\nu/d\pi = g$  as above, then it will also be true that  $d\pi/d\nu = 1/g$ .

If we try to let  $g(x) = e^{\lambda x}$  and let  $\pi$  the law of  $X_1$ , the CDF of  $\nu$  comes out to be:

$$G_0(x) := \int_{-\infty}^x e^{\lambda t} dF(t).$$

This is nondecreasing, right continuous, and goes to zero at  $-\infty$ , but is not necessarily a CDF because its value at  $+\infty$  is  $\int_{-\infty}^{\infty} e^{\lambda t} dF(t)$ . Another name for this is  $\phi(\lambda)$ . Thus, if we divide by the constant  $\phi(\lambda)$ , we see that

$$G(x) := \frac{1}{\phi(\lambda)} \int_{-\infty}^x e^{\lambda t} dF(t)$$

defines a probability CDF. This is the so-called **tilted** distribution, which we call  $\nu$ . The density  $(\phi(\lambda))^{-1} e^{\lambda x}$  will be denoted  $g_\lambda(x)$ .

Let  $\{Y_n\}$  be IID with distribution  $\nu$  and let  $T_n := \sum_{k=1}^n Y_k$  be the partial sums of the  $\{Y_n\}$ . The key is to see how the laws of  $T_n$  and  $S_n$  are related.

**Lemma 2**

$$\frac{dT_n}{dS_n} := \frac{d\mathcal{L}_{T_n}}{d\mathcal{L}_{S_n}} = (\phi(\lambda))^{-n} e^{\lambda x}.$$

PROOF: Using  $\mathbb{P}$  for probability will be confusing, so let us construct  $S_n$  and  $T_n$  as follows. Both are constructed on  $\mathbb{R}^n$ , the first on  $(\mathbb{R}, \mathcal{B}, \pi)^n$  and the second on  $(\mathbb{R}, \mathcal{B}, \nu)^n$ . The

coordinate functions are  $\{X_k\}$  on the first space and  $\{Y_k\}$  on the second, and both  $S_n$  and  $T_n$  are the same map,  $t$ , defined by  $t(x_1, \dots, x_n) := \sum_{k=1}^n x_k$ .

Let  $A' = t^{-1}[A]$  and let  $h_n(x)$  denote  $(\phi(\lambda))^{-n} e^{\lambda x}$ . We need to show that

$$\nu^n(T_n \in A) = \int_A h_n(x) d\mathcal{L}_{S_n}(x).$$

By the change of variables formula, since  $\mathcal{L}_{S_n}$  is the pushforward of  $\pi^n$  by  $t$ , the RHS is equal to

$$\int_{A'} h_n(t(x_1, \dots, x_n)) d\pi^n(x_1, \dots, x_n).$$

The LHS is equal to  $\nu^n(A')$ . Thus the statement to be shown is equivalent to  $d\nu^n/d\pi^n = h_n \circ t$ . This is more or less immediate from the definitions, as follows. For any rectangle  $A_1 \times \dots \times A_n$ ,

$$\nu^n(A) = \prod_{k=1}^n \nu(A_k) = \prod_{k=1}^n \int_{A_k} g_\lambda(x) d\pi(x) = \int_A \prod_{k=1}^n g_\lambda(x_k) d\pi^n(x_1, \dots, x_n).$$

Hence

$$\frac{d\nu^n}{d\pi^n} = \prod_{k=1}^n g_\lambda(x_k) = (\phi(\lambda))^{-n} e^{\lambda t(x_1, \dots, x_n)}.$$

□

One more fact whose proof is deferred until later is:

**Lemma 3** *The mean  $\int Y_1 d\nu$  of the tilted variables is equal to  $\psi(\lambda)$ .*

□

The proof of Theorem 1 is almost very easy. Suppose we could show that

$$\nu^n(T_n \approx an) \geq \epsilon$$

for some  $\epsilon > 0$  (which is plausible because the mean of  $Y_1$  is  $a$ ). It would follow that

$$\begin{aligned} \pi^n(S_n \geq an) &\geq \epsilon \frac{dS_n}{dT_n}(an) \\ &= \epsilon \phi(\lambda_0)^n e^{-an\lambda_0} \\ &= \epsilon \exp(-h(a)n). \end{aligned}$$

This is a little too good: while  $h(a)$  is the correct exponential decay rate, the above inequality is actually wrong by a factor of order  $n^{1/2}$ . To fix this, note that the theorem follows if we know that

$$\liminf n^{-1} \log \mathbb{P}(S_n \geq an) \geq -h(a) - \epsilon$$

for every  $\epsilon > 0$ . With this fudge factor, the easiest way to fix the proof is to tilt by an amount  $\lambda$  that is just a shade more than  $\lambda_0$ . In particular for any  $\delta > 0$ , we may choose  $\lambda \in (\lambda_0, \lambda_0 + \delta)$  such that  $\phi(\lambda)$  is still finite and moreover,  $\int Y_1 d\nu < a + \delta$ . The weak law of large numbers then gives

$$\mathbb{P}(an < T_n < (a + \delta)n) \rightarrow 1.$$

Thus

$$\begin{aligned} \pi^n(S_n \geq an) &\geq (1 + o(1)) \inf_{an < x < (a+\delta)n} \frac{dS_n}{dT_n}(x) \\ &\geq \phi(\lambda_0)^n e^{-(a+\delta)n\lambda}. \end{aligned}$$

For sufficiently small  $\delta > 0$ , the right-hand side is at least  $\exp[(-h(a) - \epsilon)n]$ , which proves the theorem.  $\square$

To go back and fill in the holes, we follow Durrett. Durrett gives a soft argument (9.1) for the existence of the limit  $\lim n^{-1} \log \mathbb{P}(S_n \geq an)$ . We don't need this, but the subadditivity argument is a standard tool, so you should read it. Durrett then constructs the tilted distributions from the CDF formula (1). Using dominated convergence to differentiate under the integral sign, he shows that  $\psi$  (which he calls  $\kappa$ ) is continuous at zero, differentiable between zero and  $\theta_+$  (the supremum of arguments for which it is finite), strictly convex, and has derivative (from the right) equal to  $\mu$  at zero (Lemma 9.4). From these results follow Lemma 3 above, the two characterizations of  $\lambda_0$  in Theorem 1 above, and the fact, used in the proof, that  $\lambda > \lambda_0$  may be chosen to make  $\int Y_1 d\nu$  less than  $a + \delta$ .

## Better estimates

Suppose we can show that  $\mathbb{P}(T_n \in [an, an + t]) \leq Cn^{-1/2}t$ . Then we would have

$$\begin{aligned} \pi^n(S_n \geq an) &= \int_{an}^{\infty} \frac{dS_n}{dT_n}(x) d\mathcal{L}_{T_n}(x) \\ &\leq \phi(\lambda_0)^n \exp(-\lambda_0 x) Cn^{-1/2} dx \\ &\leq C\lambda_0^{-1} n^{-1/2} dx \phi(\lambda_0)^n \exp(-\lambda_0 an). \end{aligned}$$

If we find that in fact  $\mathbb{P}(T_n \in [an, an + 1]) \geq Cn^{-1/2}$ , then we would obtain

$$\begin{aligned} \pi^n(S_n \geq an) &\geq \phi(\lambda_0)^n \exp(-\lambda_0(an + 1)) Cn^{-1/2} dx \\ &\geq C \exp^{-\lambda_0} \exp(-nh(a)). \end{aligned}$$

Various central limit theorems show these hypotheses to be correct, in which case  $\pi^n(S_n \geq an) = \Theta(n^{-1/2} \exp(-nh(a)))$ .

In Section 2.5, we will see some local central limit theorems. In the case of non-lattice distributions (see Durrett for definitions), Theorem 5.4 of Chapter 2 (with  $x_n = 0$ ) states that for fixed  $0 < u < v$ ,

$$\sqrt{n} \mathbb{P}(T_n \in [an + u, an + v]) \rightarrow (2\pi\sigma^2)^{-1/2}(v - u)$$

where  $\sigma^2$  is the variance of  $Y_1$  under the distribution  $\nu$ . It is immediate that this holds simultaneously for finitely many intervals  $[u_j, v_j]$ . Applying this with  $\lambda = \lambda_0$ ,  $u_j = \epsilon j$ ,  $v_j = \epsilon(j + 1)$  and  $0 \leq j \leq N$  with  $N := \epsilon^{-2}$  shows that

$$\begin{aligned} \pi^n(S_n \geq an) &\geq \sum_{j=0}^N \nu^n(T_n \in [\epsilon j, \epsilon(j + 1)]) \frac{dS_n}{dT_n}(\epsilon(j + 1)) \\ &\rightarrow (2\pi\sigma^2)^{-1/2} \frac{1}{\lambda_0} \exp(-nh(a)) \end{aligned}$$

because

$$\sum_{j=0}^N \epsilon \frac{dS_n}{dT_n}(\epsilon(j + 1)) \rightarrow \phi(\lambda_0)^n \int_0^\infty \exp(-\lambda_0 x) dx = \frac{\exp(-nh(a))}{\lambda_0}.$$

It is possible to show that  $n^{1/2} \exp(nh(a)) \mathbb{P}(S_n \geq an + L) \rightarrow 0$  as  $L \rightarrow \infty$  uniformly in  $n$ , which shows that the constant is correct:

$$\mathbb{P}(S_n \geq an) \sim \frac{(2\pi\sigma^2)^{-1/2}}{\lambda_0} n^{-1/2} \exp(-nh(a)).$$

When  $\pi$  is a lattice distribution, the constant term may oscillate.

## Example

A baseball player's **slugging percentage** is the average number of bases per at bat, where a home run counts as 4, a triple as 3, a double as 2, a single as 1, and an out as 0. Derek

Jeter has 6790 career at bats, of which 4640 are outs, 1570 are singles, 347 are doubles, 50 are triples, and 183 are home runs. His slugging percentage is 0.4633. How unlikely is it that he will have a slugging percentage of 1.000 or higher in a given stretch of at bats, if his performance is modeled by IID samples from his career totals? We compute his probabilities for each of the five categories, then compute  $\phi$  and  $\psi$  and  $\psi'$ ,  $\lambda_0$ , and the rate function (I used  $t$  instead of  $\lambda$ ).

```
> y0 := evalf(x0/AB); y1 := evalf(x1/AB); y2 := evalf(x2/AB); y3 := evalf(x3/AB); y4 := evalf(x4/AB);
      y0 := 0.6833578792
      y1 := 0.2312223859
      y2 := 0.05110456554
      y3 := 0.007363770250
      y4 := 0.02695139912

> phi := y0 + y1 * exp(t) + y2 * exp(2*t) + y3 * exp(3*t) + y4 * exp(4*t);

phi := 0.6833578792 + 0.2312223859 exp(t) + 0.05110456554 exp(2 t)
      + 0.007363770250 exp(3 t) + 0.02695139912 exp(4 t)

> psi := log(phi); m := diff(psi,t); evalf(subs(t=0,m)):
> t0 := fsolve(m=1,t);
      t0 := 0.4548794719

> rate := evalf(subs(t=t0,psi-t));
      rate := -0.1402335230
```

Thus over a stretch of 16 at bats, the probability of his slugging at least 1.000 is roughly  $(2\pi 16)^{-1/2}(\sigma\lambda_0^{-1})\exp(-16 * 0.14) \approx 0.023$  (I didn't bother computing  $\sigma$ , but used the approximation  $\sigma = 1$ ; the actual value of  $\sigma$  is greater). In the playoffs this year, Jeter slugged nearly 1.000; there is, therefore, weak evidence that he responded to playoff conditions.

In a prolonged stretch in which Jeter maintains a 1.000 slugging percentage, how should his hits be distributed? I computed the tilted distribution, below, to see that he should hit a home run roughly 1/8 of the time, a double about 1/10 of the time, and a single about 1/4 of the time.

```
> z4 := y4 * exp(4*t0) / evalf(subs(t=t0,phi)); z3 := y3 * exp(3*t0)/evalf(subs(t=t0,phi));  
z2 := y2 * exp(2*t0)/evalf(subs(t=t0,phi)); z1 := y1 * exp(1*t0)/evalf(subs(t=t0,phi));  
z0 := evalf(y0)/evalf(subs(t=t0,phi));
```

```
z4 := 0.1213776975  
z3 := 0.02104292777  
z2 := 0.09266462882  
z1 := 0.2660311693  
z0 := 0.4988835764
```

```
> sigma := z4*16 + z3*9 + z2*4 + z1 - 1;
```

```
sigma := 1.768119194
```

```
> evalf(exp(-16*0.14)/sqrt(2*Pi*16)) / (t0*sigma);
```

```
0.01320147805
```

## References

- [Che52] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23:493–507, 1952.
- [Roc66] R.T. Rockefellar. *Convex analysis*. Princeton University Press, Princeton, 1966.