

AM261 – Recent Applications of Probability and Statistics

Project 1: The maximum-entropy principle, the large-deviation principle, and the exponential family

(Don't forget to show your work and explain your reasoning.)

1. Imagine rolling a fair die a large number (N) times. Let $X_1, \dots, X_N \in \{1, 2, 3, 4, 5, 6\}$ be the outcomes. Typically,

$$\frac{1}{N} \sum_{i=1}^N X_i \approx \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

Suppose you are told that, instead, the event

$$3.0 < \frac{1}{N} \sum_{i=1}^N X_i < 3.2 \tag{1}$$

occured.

- (a) Use the large-deviation principle (equivalent, here, to the maximum-entropy principle) to make a guess (call it q^*) at the value of the six relative frequencies

$$f_k(X_1^N) \doteq \frac{1}{N} \#\{i : X_i = k\} \quad 1 \leq k \leq 6$$

(i.e. guess the empirical distribution). You will need to find an appropriate lambda in the Gibbs representation of q^* . Do this by gradient descent on the log of the partition function (as opposed to using a numerical package). Recall that $\log(Z_\lambda)$, with a suitably defined “energy” $\mathcal{E}(X)$, is convex, has a minimum at the desired λ , and

$$\frac{d}{d\lambda} \log(Z_\lambda) = E_\lambda[\mathcal{E}(X)]$$

Hence, an iteration of the form

$$\lambda(n+1) = \lambda(n) - \epsilon E_{\lambda(n)}[\mathcal{E}(X)]$$

will approach an approximate solution, provided ϵ is small enough and enough iterations are used. Ten thousand iterations with $\epsilon = .001$ should suffice. Check your λ by checking that the appropriate constraint is (nearly) satisfied.

- (b) Perform a series of Monte Carlo experiments: For each of $N = 50, 100, 150, \dots, 400$ repeatedly sample N rolls of a fair die.¹ For each sample of size N check for an instance of the rare event defined in Equation 1. Accumulate $M = 50$ examples of this event for each value of N .

For each set of M examples, compute three quantities:

- (i) The ratio of M to the number of samples needed to get M rare events (and hence an estimate of the probability of the rare event);
- (ii) The average of the M distances between the empirical distribution, $f(X_1^N)$, and the distribution q^* predicted in (a) above, where

$$distance = d(q^*, f(X_1^N)) \doteq \max_{1 \leq k \leq 6} |q_k^* - f_k(X_1^N)|$$

- (iii) The standard deviation of the M distances.

Plot each of the three quantities against N . Interpret the plots. What is the asymptotic ($N \rightarrow \infty$) value of each quantity? Why?

Remark. Observe that

$$\frac{1}{N} \sum_{i=1}^N X_i - 3.5 = \frac{1}{N} \sum_{i=1}^N (X_i - 3.5) = \frac{1}{\sqrt{N}} \left(\frac{1}{\sqrt{N}} \sum_{i=1}^N (X_i - 3.5) \right),$$

and this, by virtue of the central limit theorem, has distribution approximately $\frac{1}{\sqrt{N}}W$, where W is normal with mean zero and variance $\sigma^2 = \frac{1}{6} \sum_{k=1}^6 (k - 3.5)^2 \approx 2.9167$. Hence, with $N = 400$ (the largest value) $\frac{1}{400} \sum_{i=1}^{400} X_i - 3.5$ is approximately normal with mean zero and standard error $\sqrt{2.9167}/\sqrt{400} \approx .085$, and

$$\begin{aligned} Pr\{3.0 < \frac{1}{N} \sum_{i=1}^N X_i < 3.2\} &= Pr\left\{\frac{1}{N} \sum_{i=1}^N X_i - 3.5 \in (-.5, -.3)\right\} \\ &= Pr\{Z \in (-.5/.085, -.3/.085)\} \\ &\quad (Z \text{ standard normal}) \\ &\approx .0002, \end{aligned}$$

meaning we can expect about one large deviation result in every 5,000 samples.

¹One way to sample N rolls of a fair die is with the Matlab command `ceil(6.*rand(1,N))`.

2. This time, imagine rolling an *un*-fair die with $p = (p_1, \dots, p_6) = (.1, .1, .2, .1, .2, .3)$. Typically,

$$\frac{1}{N} \sum_{i=1}^N X_i \approx E_p[X] = 4.1$$

and

$$\frac{1}{N} \sum_{i=1}^N X_i^2 \approx E_p[X^2] = 19.7$$

Consider the event that, instead of the typical values,

$$3.6 < \frac{1}{N} \sum_{i=1}^N X_i < 3.8 \tag{2}$$

and

$$\frac{1}{N} \sum_{i=1}^N X_i^2 < 17 \tag{3}$$

- (a) As in 1(a), apply the LDP to predict the relative frequencies $f_k(X_1^N)$, $1 \leq k \leq 6$. (There are two constraints. One, the other, or both might be needed. Try each one individually and then check if the other is satisfied. If neither one works by itself, you'll need to apply both simultaneously.)
- (b) Repeat 1(b), using $N = 50, 100, 150, \dots, 300$, and $M = 50$.²

3. Approximately what would be the probability of observing a 4 followed immediately by a 6, at any given position in a string of 300 rolls of the die in the previous problem that happens to satisfy both of the conditions in Equations 2 and 3? Compare your prediction to the empirical probability of the sequence (4, 6), as derived collectively from the $M = 50$ samples obtained in 2(b) when $N = 300$.

²One way to sample N rolls of this particular die is to first set up the table `lookup=[1 2 3 3 4 5 5 6 6 6]` and then use `lookup(ceil(10.*rand(1,N)))`.