Chapter 4

Probability

4.1 Discrete random variable

1. *Random deposition:* A mirror is plated by evaporating a gold electrode in vaccum by passing an electric current. The gold atoms fly off in all directions, and a portion of them sticks to the glass (or to other gold atoms already on the glass plate). Assume that each column of deposited atoms is independent of neighboring columns, and that the average deposition rate is d layers per second.

   (a) What is the probability of m atoms deposited at a site after a time t? What fraction of the glass is not covered by any gold atoms?

   (b) What is the variance in the thickness?

*****

4.2 Continuous random variable

1. *Characteristic functions:* Calculate the characteristic function, the mean, and the variance of the following probability density functions:

   (a) *Uniform* $\quad p(x) = \frac{1}{2a} \quad$ for $\quad -a < x < a \quad$, and $\quad p(x) = 0 \quad$ otherwise;

   (b) *Laplace* $\quad p(x) = \frac{1}{2a} \exp\left(-\frac{|x|}{a}\right) \quad$;

   (c) *Cauchy* $\quad p(x) = \frac{a}{\pi(x^2+a^2)} \quad$.

   The following two probability density functions are defined for $x \geq 0$. Compute only the mean and variance for each.

   (d) *Rayleigh* $\quad p(x) = \frac{x}{a^2} \exp(-\frac{x^2}{2a^2}) \quad$,

   (e) *Maxwell* $\quad p(x) = \sqrt{\frac{2}{\pi}} \frac{x^2}{a^3} \exp(-\frac{x^2}{2a^2}) \quad$.

*****

2. *Diode:* The current I across a diode is related to the applied voltage V via $I = I_0 \left[\exp(eV/kT) - 1\right]$. The diode is subject to a random potential V of zero mean and variance σ^2 which is Gaussian distributed. Find the probability density $p(I)$ for the current I flowing through the diode. Find the most probable value for I, the mean value of I, and indicate them on a sketch of $p(I)$.

*****

3. *Tchebycheff inequality:* Consider any probability density $p(x)$ for $(-\infty < x < \infty)$, with mean λ, and variance σ^2. Show that the total probability of outcomes that are more than $n\sigma$ away from λ is less than $1/n^2$, i.e.

$$\int_{|x-\lambda|\geq n\sigma} dx p(x) \leq \frac{1}{n^2}.$$

Hint: Start with the integral defining σ^2, and break it up into parts corresponding to $|x-\lambda| > n\sigma$, and $|x-\lambda| < n\sigma$.

*****

4.3 Many random variables

1. *Optimal selection:* In many specialized populations, there is little variability among the members. Is this a natural consequence of optimal selection?

   (a) Let $\{r_\alpha\}$ be n random numbers, each independently chosen from a probability density $p(r)$, with $r \in [0,1]$. Calculate the probability density $p_n(x)$ for the largest value of this set, i.e. for $x = \max\{r_1, \cdots, r_n\}$.

   (b) If each r_α is uniformly distributed between 0 and 1, calculate the mean and variance of x as a function of n, and comment on their behavior at large n.

*****

2. *Benford's law* describes the observed probabilities of the *first digit* in a great variety of data sets, such as stock prices. Rather counter-intuitively, the digits 1 through 9 occur with probabilities 0.301, .176, .125, .097, .079, .067, .058, .051, .046 respectively. The key observation is that this distribution is invariant under a change of scale, e.g. if the stock prices were converted from dollars to persian rials! Find a formula that fits the above probabilities on the basis of this observation.

*****

3. *Directed random walk:* The motion of a particle in three dimensions is a series of independent steps of length ℓ. Each step makes an angle θ with the z axis, with a probability density $p(\theta) = 2\cos^2(\theta/2)/\pi$; while the angle ϕ is uniformly distributed between 0 and 2π. (Note that the solid angle factor of $\sin\theta$ is already included in the

definition of $p(\theta)$, which is correctly normalized to unity.) The particle (walker) starts at the origin and makes a large number of steps N.

(a) Calculate the expectation values $\langle z \rangle$, $\langle x \rangle$, $\langle y \rangle$, $\langle z^2 \rangle$, $\langle x^2 \rangle$, and $\langle y^2 \rangle$, and the covariances $\langle xy \rangle$, $\langle xz \rangle$, and $\langle yz \rangle$.

(b) Use the central limit theorem to estimate the probability density $p(x, y, z)$ for the particle to end up at the point (x, y, z).

*****

4.4 From probability to certainty

1. *Information:* Consider the velocity of a gas particle in one dimension ($-\infty < v < \infty$).

   (a) Find the unbiased probability density $p_1(v)$, subject only to the constraint that the average *speed* is c, i.e. $\langle |v| \rangle = c$.

   (b) Now find the probability density $p_2(v)$, given only the constraint of average kinetic energy, $\langle mv^2/2 \rangle = mc^2/2$.

   (c) Which of the above statements provides more information on the velocity? Quantify the difference in information in terms of $I_2 - I_1 \equiv \left(\langle \ln p_2 \rangle - \langle \ln p_1 \rangle \right) / \ln 2$.

*****

2. *Dice:* A dice is loaded such that 6 occurs twice as often as 1.

   (a) Calculate the unbiased probabilities for the 6 faces of the dice.

   (b) What is the information content (in bits) of the above statement regarding the dice?

*****

3. *Random matrices:* As a model for energy levels of complex nuclei, Wigner considered $N \times N$ symmetric matrices whose elements are random. Let us assume that each element M_{ij} (for $i \geq j$) is an independent random variable taken from the probability density function

$$p(M_{ij}) = \frac{1}{2a} \quad \text{for} \quad -a < M_{ij} < a \,, \quad \text{and} \quad p(M_{ij}) = 0 \quad \text{otherwise.}$$

   (a) Calculate the characteristic function for each element M_{ij}.

   (b) Calculate the characteristic function for the trace of the matrix, $T \equiv \mathrm{tr} M = \sum_i M_{ii}$.

   (c) What does the central limit theorem imply about the probability density function of the trace at large N?

(d) For large N, each eigenvalue λ_α ($\alpha = 1, 2, \cdots, N$) of the matrix M is distributed according to a probability density function

$$p(\lambda) = \frac{2}{\pi\lambda_0}\sqrt{1 - \frac{\lambda^2}{\lambda_0^2}} \quad \text{for} \quad -\lambda_0 < \lambda < \lambda_0, \quad \text{and} \quad p(\lambda) = 0 \quad \text{otherwise},$$

(known as the Wigner semi-circle rule). Find the variance of λ.

(**Hint:** Changing variables to $\lambda = \lambda_0 \sin\theta$ simplifies the integrals.)

(e) If in the previous result, we have $\lambda_0^2 = 4Na^2/3$, can the eigenvalues be independent of each other?

*****

4. *Mutual information:* Consider random variables x and y, distributed according to a joint probability $p(x, y)$. The mutual information between the two variables is defined by

$$M(x, y) \equiv \sum_{x,y} p(x, y) \ln\left(\frac{p(x, y)}{p_x(x)p_y(y)}\right),$$

where p_x and p_y denote the *unconditional* probabilities for x and y.

(a) Relate $M(x, y)$ to the entropies $S(x, y)$, $S(x)$, and $S(y)$ obtained from the corresponding probabilities.

(b) Calculate the mutual information for the joint Gaussian form

$$p(x, y) \propto \exp\left(-\frac{ax^2}{2} - \frac{by^2}{2} - cxy\right).$$

*****

5. *Semi-flexible polymer in two dimensions* Configurations of a model polymer can be described by either a set of vectors $\{\mathbf{t}_i\}$ of length a in two dimensions (for $i = 1, \cdots, N$), or alternatively by the angles $\{\phi_i\}$ between successive vectors.The polymer is at a temperature T, and subject to an energy

$$\mathcal{H} = -\kappa \sum_{i=1}^{N-1} \mathbf{t}_i \cdot \mathbf{t}_{i+1} = -\kappa a^2 \sum_{i=1}^{N-1} \cos\phi_i \quad ,$$

where κ is related to the bending rigidity, such the probability of any configuration is proportional to $\exp(-\mathcal{H}/k_BT)$.

(a) Show that $\langle \mathbf{t}_m \cdot \mathbf{t}_n \rangle \propto \exp(-|n - m|/\xi)$, and obtain an expression for the *persistence length* $\ell_p = a\xi$. (You can leave the answer as the ratio of simple integrals.)

(b) Consider the end–to–end distance $\mathbf{R}$ as illustrated in the figure. Obtain an expression for $\langle R^2 \rangle$ in the limit of $N \gg 1$.

(c) Find the probability $p(\mathbf{R})$ in the limit of $N \gg 1$.

(d) If the ends of the polymer are pulled apart by a force $\mathbf{F}$, the probabilities for polymer configurations are modified by the Boltzmann weight $\exp\left(\frac{\mathbf{F}\cdot\mathbf{R}}{k_B T}\right)$. By expanding this weight, or otherwise, show that

$$\langle \mathbf{R} \rangle = K^{-1}\mathbf{F} + \mathcal{O}(F^3) \quad ,$$

and give an expression for the Hookian constant K in terms of quantities calculated before.

*****

6. Jensen's inequality and Kullback–Liebler divergence: A convex function $f(x)$ always lies above the tangent at any point, i.e. $f(x) \geq f(y) + f'(y)(x-y)$ for all y.

(a) Show that for a convex function $\langle f(x) \rangle \geq f(\langle x \rangle)$.

(b) The *Kullback–Liebler divergence* of two probability distributions $p(x)$ and $q(x)$ is defined as $D(p|q) \equiv \int dx\ p(x) \ln\left[p(x)/q(x)\right]$. Use Jensen's inequality to prove that $D(p|q) \geq 0$.

*****

7. *The book of records:* Consider a sequence of random numbers $\{x_1, x_2, \cdots, x_n, \cdots\}$; the entry x_n is a *record* if it is larger than all numbers before it, i.e. if $x_n > \{x_1, x_2, \cdots, x_{n-1}\}$. We can then define an associated sequence of indicators $\{R_1, R_2, \cdots, R_n, \cdots\}$ in which $R_n = 1$ if x_n is a record, and $R_n = 0$ if it is not (clearly $R_1 = 1$).

(a) Assume that each entry x_n is taken independently from the same probability distribution $p(x)$. [In other words, $\{x_n\}$ are *IIDs* (independent identically distributed).] Show that, irrespective of the form of $p(x)$, there is a very simple expression for the probability P_n that the entry x_n is a record.

(b) The records are entered in the *Guinness Book of Records*. What is the average number $\langle S_N \rangle$ of records after N attempts, and how does it grow for, $N \gg 1$? If the number of trials, e.g. the number of participants in a sporting event, doubles every year, how does the number of entries asymptotically grow with time.

(c) Prove that the record indicators $\{R_n\}$ are *independent* random variables (though not identical), in that $\langle R_n R_m \rangle_c = 0$ for $m \neq n$.

(d) Compute all moments, and the first three cumulants of the total number of records S_N after N entries. Does the central limit theorem apply to S_N?

(e) The first record, of course occurs for $n_1 = 1$. If the third record occurs at trial number $n_3 = 9$, what is the mean value $\langle n_2 \rangle$ for the position of the second record? What is the mean value $\langle n_4 \rangle$ for the position of the fourth record?

*****