2

Numerical integration and importance sampling

2.1 Quadrature

Consider the numerical evaluation of the integral

$$I(a,b) \;=\; \int_a^b dx\, f(x)$$

- Rectangle rule: on small interval, construct interpolating function and integrate over interval.
  - Polynomial of degree 0 using mid-point of interval:

$$\int_{ah}^{(a+1)h} dx\, f(x) \approx h\, f\left((ah + (a+1)h)/2\right).$$

  - Polynomial of degree 1 passing through points $(a_1, f(a_1))$ and $(a_2, f(a_2))$: Trapezoidal rule

$$f(x) = f(a_1) + \frac{x - a_1}{a_2 - a_1}\left(f(a_2) - f(a_1)\right) \;\longrightarrow\; \int_{a_1}^{a_2} dx\, f(x) = \left(\frac{a_2 - a_1}{2}\right)\left(f(a_1) + f(a_2)\right).$$

  - Composing trapezoidal rule n times on interval (a,b) with even sub-intervals $[kh, (k+1)h]$ where $k = 0, \ldots, n-1$ and $h = (b-a)/n$ gives estimate

$$\int_a^b dx\, f(x) \approx \frac{b-a}{n}\left(\frac{f(a) + f(b)}{2} + \sum_{k=1}^{n-1} f(a + kh)\right).$$

- Simpsons rule: interpolating function of degree 2 composed n times on interval (a, b):

$$\int_a^b dx\, f(x) \approx \frac{b-a}{3n}\left[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + 2f(a+4h) + \cdots + f(b)\right].$$

  - Error bounded by

$$\frac{h^4}{180}(b-a)\left|f^{(4)}(\xi)\right|.$$

    where $\xi \in (a, b)$ is the point in the domain where the magnitude of the 4th derivative is largest.

- Rules based on un-even divisions of domain of integration
  - $w(x) =$ weight function.
  - $I(a, b) \approx \sum_{i=1}^{n} w_i f_i$, where $w_i = w(x_i)$ and $f_i = f(x_i)$ with choice of x_i and w_i based on definite integral of polynomials of higher order.
  - Example is Gaussian quadrature with n points based on polynomial of degree $2n-1$: well-defined procedure to find $\{x_i\}$ and $\{w_i\}$ (see *Numerical Recipes*).
  - Error bounds for n-point Gaussian quadrature are

$$\frac{(b-a)^{2n+1}}{(2n+1)!}\frac{(n!)^4}{[(2n)!]^3}\left|f^{(2n)}(\xi)\right| \text{ for } \xi \in (a, b)..$$

- For multi-dimensional integrals, must place n_i grid points along each i dimension.
  - Number of points in hyper-grid grows exponentially with dimension.
  - Unsuitable for high-dimensional integrals.

2.2 Importance Sampling and Monte Carlo

Suppose integrand $f(\mathbf{x})$ depends on multi-dimensional point $\mathbf{x}$ and that integral over hyper-volume

$$I = \int_V d\mathbf{x}\, f(\mathbf{x})$$

is non-zero only in specific regions of the domain.

- We should place higher density of points in region where integrand is large.
- Define a weight function $w(\mathbf{x})$ that tells us which regions are significant.

  - Require property $w(\mathbf{x}) > 0$ for any point $\mathbf{x}$ in volume.
  - Sometimes use normalized weight function so $\int_V d\mathbf{x}\, w(\mathbf{x}) = 1$, though not strictly necessary.
  - Re-express integral as:

$$I = \int_V d\mathbf{x}\, \frac{f(\mathbf{x})}{w(\mathbf{x})}\, w(\mathbf{x}).$$

- Idea: Draw a set of N points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ from the weight function $w(\mathbf{x})$ then

$$\overline{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(\mathbf{x}_i)}{w(\mathbf{x}_i)}.$$

- As $N \to \infty$, $\overline{I} \to I$.

- How does this improve the rate of convergence of the calculation of I? We will see that the statistical uncertainty is related to the variance σ_I^2 of the estimate of I, namely

$$\sigma_{\overline{I}}^2 \;=\; \frac{1}{N} \sum_{i}^{N} \langle \Delta I_i \Delta I_i \rangle \qquad \text{where} \qquad \Delta I_i = \frac{f(\mathbf{x}_i)}{w(\mathbf{x}_i)} - \overline{I}.$$

  and we have assumed that the random variables ΔI_i are statistically independent. Here $\langle \cdots \rangle$ represents the average over the true distribution of f/w that is obtained in the limit $N \to \infty$.

  - Vastly different values of ratio $f(\mathbf{x}_i)/w(\mathbf{x}_i)$ lead to large uncertainty.

- The error is minimized by minimizing $\sigma_{\overline{I}}^2$.
  - If $\alpha\, w(\mathbf{x}_i) = f(\mathbf{x}_i)$, then $f(\mathbf{x}_i)/w(\mathbf{x}_i) = \alpha$ and

$$\left\langle \frac{f(\mathbf{x}_i)}{w(\mathbf{x}_i)} \right\rangle = I = \alpha \qquad \left\langle \left(\frac{f(\mathbf{x}_i)}{w(\mathbf{x}_i)} \right)^2 \right\rangle = \alpha^2,$$

    and $\sigma_{\overline{I}}^2 = 0$.
  - Note that writing $w(\mathbf{x}_i) = f(\mathbf{x}_i)/\alpha$ requires knowing $\alpha = I$, the problem we are trying to solve.
  - Generally desire all $f(\mathbf{x}_i)/w(\mathbf{x}_i)$ to be roughly the same for all sampled points $\mathbf{x}_i$ to mimimize $\sigma_{\overline{I}}^2$.

- Example in 1-dimensional integral $I = \int_a^b dx\, f(x) = \int_a^b dx\, \frac{f(x)}{w(x)}\, w(x)$.

- **Monte-Carlo** sampling: use random sampling of points x with weight $w(x)$ to estimate ratio $f(x)/w(x)$.
- How do we draw sample points with a given weight $w(x)$? Consider simplest case of a uniform distribution on interval (a, b):

$$w(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in (a, b). \\ 0 & \text{otherwise.} \end{cases}$$

- Use a *random number generator* that gives a pseudo-random number *rand* in interval $(0, 1)$.

$$x_i = a + (b - a) \times rand,$$

then $\{x_i\}$ are distributed uniformly in the interval (a, b).

- We find that the estimator of f is then:

$$\overline{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{f(x_i)}{w(x_i)} = \frac{b-a}{N} \sum_{i=1}^{N} f(x_i) = \frac{1}{N} \sum_{k=1}^{N} I_k,$$

where each $I_k = (b - a)\, f(x_k)$ is an independent estimate of the integral.

- Like trapezoidal integration but with randomly sampled points $\{x_i\}$ from (a, b) each with uniform weight.

- How about other choices of weight function $w(x)$?
  - It is easy to draw uniform $y_i \in (0, 1)$. Now suppose we map the y_i to x_i via

$$y(x) = \int_a^x dz\, w(z).$$

  - Note that $y(a) = 0$ and $y(b) = 1$ if $w(x)$ is normalized over (a, b).
  - How are the x distributed? Since the y are distributed uniformly over $(0, 1)$, the probability of finding a value of y in the interval is

$$dy(x) = w(x)\, dx,$$

so the x are distributed with weight $w(x)$.

  - It then follows that the one-dimensional integral can be written in the transformed variables y as:

$$I = \int_a^b dx\, w(x)\, \frac{f(x)}{w(x)} = \int_{y(a)}^{y(b)} dy\, \frac{f(x(y))}{w(x(y))} = \int_0^1 dy\, \frac{f(x(y))}{w(x(y))}$$

  - Integral easily evaluated by selecting uniform points y and then solving $x(y)$. Must be able to solve for $x(y)$ to be useful.

- Procedure:

  1. Select N points y_i uniformly on $(0,1)$ using $y_i = rand$.
  2. Compute $x_i = x(y_i)$ by inverting $y(x) = \int_a^x dz\, w(z)$.
  3. Compute estimator for integral $\overline{I} = \frac{1}{N}\sum_{i=1}^N f(x_i)/w(x_i)$.

- This procedure is easy to do using simple forms of $w(x)$. Suppose the integrand is strongly peaked around $x = x_0$. One good choice of $w(x)$ might be a Gaussian weight

$$w(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-x_0)^2}{2\sigma^2}},$$

  where the variance (width) of the Gaussian is treated as a parameter.

  – If $I = \int_{-\infty}^{\infty} dx\, f(x) = \int_{-\infty}^{\infty} dx\, w(x) f(x)/w(x)$, can draw randomly from the Gaussian weight by:

$$\begin{aligned} y(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} d\tilde{x}\, e^{-\frac{(\tilde{x}-x_0)^2}{2\sigma^2}} \\ &= \frac{1}{2} + \frac{1}{\sqrt{2\pi\sigma^2}} \int_{0}^{x} d\tilde{x}\, e^{-\frac{(\tilde{x}-x_0)^2}{2\sigma^2}} \\ &= \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x-x_0}{\sqrt{2}\sigma}} dw\, e^{-w^2} = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-x_0}{\sqrt{2}\sigma}\right)\right). \end{aligned}$$

  – Inverting gives $x_i = x_0 + \sqrt{2}\sigma\, \mathrm{ierf}(2y_i - 1)$, where ierf is the inverse error function with series representation

$$\mathrm{ierf}(z) = \frac{\sqrt{\pi}}{2}\left(z + \frac{\pi}{12}z^3 + \frac{7\pi^2}{480}z^5 + \ldots\right)$$

  – The estimator of the integral is therefore

$$\overline{I} = \frac{1}{N}\sqrt{2\pi\sigma^2}\sum_{i=1}^{N} f(x_i) e^{\frac{(x_i-x_0)^2}{2\sigma^2}}.$$

  – This estimator reduces the variance $\sigma_{\overline{I}}^2$ if $w(x_i)$ and $f(x_i)$ resemble one another.

- Another way to draw from a Gaussian:

  – Draw 2 numbers, y_1 and y_2 uniformly on $(0,1)$. Define $R = \sqrt{-2\ln y_1}$ and $\theta = 2\pi y_2$. Then $x_1 = R\cos\theta$ and $x_2 = R\sin\theta$ are distributed with density

$$w(x_1, x_2) = \frac{1}{2\pi} e^{-x_1^2} e^{-x_2^2}$$

since

$$dy_1 dy_2 = \begin{vmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{vmatrix} dx_1 dx_2 = w(x_1, x_2) dx_1 dx_2$$

2.3 Markov Chain Monte Carlo

2.3.1 Ensemble averages

- Generally, we cannot find a simple way of generating the $\{\mathbf{x}_i\}$ according to a known $w(\mathbf{x}_i)$ for high-dimensional systems by transforming from a uniform distribution.
  - Analytical integrals for coordinate transformation may be invertable only for separable coordinates.
  - Many degrees of freedom are coupled so that joint probability is complicated and not the product of single probabilities.
- Typical integrals are ensemble averages of the form:

$$\langle A \rangle = \frac{1}{Z} \int_V d\mathbf{r}^{(N)} e^{-\beta u(\mathbf{r}^{(N)})} A(\mathbf{r}^{(N)}) = \frac{\int_V d\mathbf{r}^{(N)} e^{-\beta u(\mathbf{r}^{(N)})} A(\mathbf{r}^{(N)})}{\int_V d\mathbf{r}^{(N)} e^{-\beta u(\mathbf{r}^{(N)})}}$$

  - Typical potentials are complicated functions of configuration $\mathbf{r}^{(N)}$.
  - A good importance sampling weight would set $w(\mathbf{r}^{(N)}) = e^{-\beta u(\mathbf{r}^{(N)})}$.
- How do we sample configurations from a general, multi-dimensional weight?
- **Goal:** Devise a method to generate a sequence of configurations $\{\mathbf{r}_1^{(N)}, \ldots, \mathbf{r}_n^{(N)}\}$ in which the probability of finding a configuration $\mathbf{r}_i^{(N)}$ in the sequence is given by $w(\mathbf{r}_i^{(N)}) d\mathbf{r}_i^{(N)}$.
- We will do so using a stochastic procedure based on a random walk.

2.3.2 Markov Chains

We will represent a general, multi-dimensional configurational coordinate that identifies the state by the vector $\mathbf{X}$. We wish to generate a sequence of configurations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ where each $\mathbf{X}_t$ is chosen with probability density $P_t(\mathbf{X}_t)$. To generate the sequence, we use a *discrete-time, time-homogeneous, Markovian random walk*.

- Consider the configuration $\mathbf{X}_1$ at an initial time labelled as 1 that is treated as a random variable drawn from some known initial density $P_1(\mathbf{X})$.

- We assume the stochastic dynamics of the random variable is determined by a *time-independent* transition function $\mathsf{K}(\mathbf{Y} \to \mathbf{X})$, where the transition function defines the probability density of going from state $\mathbf{Y}$ to state $\mathbf{X}$ in a time step.
- Since K is a probability density, it must satisfy

$$\int d\mathbf{X}\, \mathsf{K}(\mathbf{Y} \to \mathbf{X}) = 1 \qquad\qquad \mathsf{K}(\mathbf{Y} \to \mathbf{X}) \geq 0.$$

- The state $\mathbf{X}_t$ at time t is assumed to be obtained from the state at time $\mathbf{X}_{t-1}$ by a realization of the dynamics, where the probability of all transitions is determined by K.
- At the second step of the random walk, the new state $\mathbf{X}_2$ is chosen from $P_2(\mathbf{X}|\mathbf{X}_1) = \mathsf{K}(\mathbf{X}_1 \to \mathbf{X})$. At the third step, the state $\mathbf{X}_3$ is chosen from $P_3(\mathbf{X}|\mathbf{X}_2) = \mathsf{K}(\mathbf{X}_2 \to \mathbf{X})$, and so on, generating the random walk sequence $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$.
- If infinitely many realizations of the dynamics is carried out, we find that the distribution of $\mathbf{X}_i$ for each of the i steps are

$$\begin{array}{ll} P_1(\mathbf{X}) & \text{for } \mathbf{X}_1 \\ P_2(\mathbf{X}) = \int d\mathbf{Y}\, \mathsf{K}(\mathbf{Y} \to \mathbf{X}) P_1(\mathbf{Y}) & \text{for } \mathbf{X}_2 \\ P_3(\mathbf{X}) = \int d\mathbf{Y}\, \mathsf{K}(\mathbf{Y} \to \mathbf{X}) P_2(\mathbf{Y}) & \text{for } \mathbf{X}_3 \\ \vdots & \vdots \\ P_t(\mathbf{X}) = \int d\mathbf{Y}\, \mathsf{K}(\mathbf{Y} \to \mathbf{X}) P_{t-1}(\mathbf{Y}) & \text{for } \mathbf{X}_t. \end{array}$$

- The transition function K is *ergodic* if any state $\mathbf{X}$ can be reached from any state $\mathbf{Y}$ in a finite number of steps in a non-periodic fashion.
- If K is ergodic, then the *Perron-Frobenius Theorem* guarantees the existence of a unique *stationary distribution* $P(\mathbf{X})$ that satisfies

$$P(\mathbf{X}) = \int d\mathbf{Y}\, \mathsf{K}(\mathbf{Y} \to \mathbf{X}) P(\mathbf{Y}) \qquad \text{and} \qquad \lim_{t \to \infty} P_t(\mathbf{X}) = P(\mathbf{X}).$$

- Implications
  1. From any initial distribution of states $P_1(\mathbf{X})$, an ergodic K guarantees that states $\mathbf{X}_t$ will be distributed according to the unique stationary distribution $P(\mathbf{X})$ for large t (many steps of random walk).
  2. $P(\mathbf{X})$ is like an eigenvector of "matrix" K with eigenvalue $\lambda = 1$.
  3. Goal is then to design an ergodic transition function K so that the stationary or *limit distribution* is the Bolztmann weight $w(\mathbf{X})$.

Finite State Space

To make the Markov chain more concrete, consider a finite state space in which only m different configurations of the system exist. The phase space then can be enumerated as $\{\mathbf{X}_1, \ldots \mathbf{X}_m\}$.

- The transition function is then an $m \times m$ matrix K, where the element $\mathsf{K}_{ij} \geq 0$ is the transition probability of going from state $\mathbf{X}_j$ to state $\mathbf{X}_i$.
- Since the matrix elements represent transition probabilities, for any fixed value of j,

$$\sum_{i=1}^{m} \mathsf{K}_{ij} = 1.$$

- The distribution $P_t(\mathbf{X})$ corresponds to a column vector $\mathbf{P}_t = \mathrm{col}[a_t^{(1)}, \ldots a_t^{(m)}]$, where $a_t^{(i)}$ is the probability of state $\mathbf{X}_i$ at time t.
- The distribution evolves under the random walk as $\mathbf{P}_t = \mathsf{K} \cdot \mathbf{P}_{t-1}$.
- The matrix K is *regular* if there is an integer t_0 such that K^{t_0} has all positive (non-zero) entries. Then K^t for $t \geq t_0$ has all positive entries.
- Suppose the matrix K is a regular transition matrix. The following properties hold:

Statements following from the Frobenius-Perron theorem

1. The multiplicity of the eigenvalue $\lambda = 1$ is one (i.e. the eigenvalue is *simple*.)

   Proof. Since K is regular, there exists a transition matrix $\mathcal{K} = \mathsf{K}^{t_0}$ with all positive elements. Note that if $\mathsf{K}\mathbf{e} = \lambda\mathbf{e}$, then $\mathcal{K}\mathbf{e} = \lambda^{t_0}\mathbf{e}$, so that all eigenvectors of K with eigenvalue λ are also eigenvectors of $\mathcal{K}$ with eigenvalue λ^{t_0}. Let $\mathbf{e}_1 = \mathrm{col}[1, 1, \ldots, 1]/m$. Since $\sum_i \mathsf{K}_{ij} = 1$ for any j, we have $\sum_i (\mathsf{K} \cdot \mathsf{K})_{ij} = 1$, and hence $\sum_i \mathcal{K}_{ij} = 1$ for any j. The transpose of $\mathcal{K}$ therefore satisfies $\sum_i \mathcal{K}^T_{ji} = 1$ for all j. The vector $\mathbf{e}_1$ is the right eigenvector with eigenvalue $\lambda = 1$ since $(\mathcal{K}^T\mathbf{e}_1)_j = 1/m \sum_i \mathcal{K}^T_{ji}$, and hence $\mathcal{K}^T\mathbf{e}_1 = e_1$. Now both $\mathcal{K}$ and $\mathcal{K}^T$ are $m \times m$ square matrices and have the same eigenvalues (since the characteristic equation is invariant to the transpose operation), so $\lambda = 1$ is an eigenvalue of both $\mathcal{K}$ and $\mathcal{K}^T$. Now suppose $\mathcal{K}^T\mathbf{v} = \mathbf{v}$ and all components v_j of $\mathbf{v}$ are not equal. Let k be the index of the largest component, v_k, which we can take to be positive without loss of generality. Thus, $v_k \geq |v_j|$ for all j and $v_k > |v_l|$ for some l. It then follows that $v_k = \sum_j \mathcal{K}^T_{kj} v_j < \sum_j \mathcal{K}^T_{kj} v_k$, since all components of $\mathcal{K}^T$ are non-zero and positive. Thus we conclude that $v_k < v_k$, a contradiction. Hence all v_k must be

the same, corresponding to eigenvector $\mathbf{e}_1$. Hence $\mathbf{e}_1$ is a unique eigenvector of $\mathcal{K}^T$ and $\lambda = 1$ is a simple root of the characterstic equation. Hence $\mathcal{K}$ has a single eigenvector with eigenvalue $\lambda = 1$. □

2. If the eigenvalue $\lambda \neq 1$ is real, then $|\lambda| < 1$.

Proof. Suppose $\mathcal{K}^T\mathbf{v} = \lambda^{t_0}\mathbf{v}$, with $\lambda^{t_0} \neq 1$. It then follows that there is an index k of vector $\mathbf{v}$ such that $v_k \geq |v_j|$ for all j and $v_k > |v_l|$ for some l. Now $\lambda^{t_0} v_k = \sum_j \mathcal{K}^T_{kj} v_j < \sum_j \mathcal{K}^T_{kj} v_k = v_k$, or $\lambda^{t_0} v_k < v_k$. Thus $\lambda^{t_0} < 1$, and hence $\lambda < 1$, since λ is real. Similarly, following the same lines of argument and using the fact that $-v_k < v_l$ for some l, we can establish that $\lambda > -1$. Hence $|\lambda| < 1$. A similar sort of argument can be used if λ is complex. □

3. Under the dynamics of the random walk $\mathbf{P}_t = \mathsf{K}\mathbf{P}_{t-1}$, $\lim_{t\to\infty} \mathbf{P}_t = \mathbf{P}_1$, where $\mathbf{P}_1$ is the right eigenvector of K with simple eigenvalue $\lambda = 1$.

Proof. If the matrix K is diagonalizable, the eigenvectors of K form a complete basis (since it can be put in *Jordan canonical form*). Thus, any initial distribution $\mathbf{P}_s$ can be expanded in terms of the complete, linearly independent basis $\{\mathbf{P}_1, \mathbf{P}_2, \ldots \mathbf{P}_m\}$ of eigenvectors as $\mathbf{P}_s = b_1\mathbf{P}_1 + \cdots + \mathbf{b}_m\mathbf{P}_m$, where $\mathsf{K}\mathbf{P}_i = \lambda_i\mathbf{P}_i$ with $\lambda_1 = 1$ and $|\lambda_i| < 1$. Now $\mathbf{P}_t = \mathbf{b}_1\mathbf{P}_1 + b_2\lambda_2^t\mathbf{P}_2 + \cdots + \mathbf{b}_m\lambda_m^t\mathbf{P}_m$, but $\lim_{t\to\infty} \lambda_i^t = 0$ exponentially fast for $i \geq 2$. Thus, $\lim_{t\to\infty} \mathsf{K}^t\mathbf{P}_s = b_1\mathbf{P}_1$. Note that if $\sum_i (\mathbf{P}_s)_i = 1$, then $\sum_i (\mathbf{P}_t)_i = 1$ as K maintains the norm. This implies that $b_1 = (\sum_i (\mathbf{P}_1)_i)^{-1}$. It turns out that the condition of *detailed balance*, which we will define to mean $\mathsf{K}_{ij}(\mathbf{P}_1)_j = \mathsf{K}_{ji}(\mathbf{P}_1)_i$ allows one to define a symmetric transition matrix K', which is necessarily diagonalizable.

More generally, any square matrix is similar to a matrix of Jordan form, with isolated blocks of dimension of the multiplicity of the eigenvalue. Thus the matrix K can be written as $\mathsf{K} = \mathsf{P}\tilde{\mathsf{K}}\mathsf{P}^{-1}$ where the columns of P are the (generalized) eigenvectors of K and $\tilde{\mathsf{K}}$ is of form

$$\tilde{\mathsf{K}} = \begin{pmatrix} 1 & & & 0 \\ & J_1 & & \\ & & \ddots & \\ 0 & & & J_I \end{pmatrix},$$

where $I+1$ is the number of independent eigenvectors. For each eigenvector with

corresponding eigenvalue λ_i, J_i is of the form

$$J_i = \begin{pmatrix} \lambda_i & 1 & & 0 \\ 0 & \lambda_i & 1 & \\ & 0 & \lambda_i & \ddots \\ 0 & & \ddots & \ddots \end{pmatrix}.$$

The dimension of the square matrix J_i depends on the original matrix K. For any given eigenvalue λ, the sum of the dimensions of the J_i for which $\lambda = \lambda_i$ is equal to the algebraic multiplicity of the eigenvalue. It is easily shown that the nth power of a block J_i has elements bounded by $\binom{n}{k}\lambda_i^{n-k}$ for a block of size k, which goes to zero exponentially has n goes to infinity. Hence, the matrix $\tilde{\mathsf{K}}$ goes to

$$\tilde{\mathsf{K}} = \begin{pmatrix} 1 & & & 0 \\ & 0 & & \\ & & \ddots & \\ 0 & & & 0 \end{pmatrix}$$

Thus it follows that as n goes to infinity

$$(\mathsf{K}^n)_{\alpha\beta} = \left(\mathsf{P}\tilde{\mathsf{K}}^n\mathsf{P}^{-1}\right)_{\alpha\beta} \to \mathsf{P}_{\alpha 1}\mathsf{P}^{-1}_{1\beta} = (\mathbf{P}_1)_\alpha,$$

where $\mathbf{P}_1$ is the stationary distribution, since the matrix $\mathsf{P}^{-1}_{1\beta}$ is the β component of the left eigenvector of K with eigenvalue of 1, which is the constant vector. □

- Consider the explicit case $m = 3$, with K_{ij} given by

$$\mathsf{K} = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{3} \end{pmatrix}.$$

  - $(\mathsf{K}\cdot\mathsf{K})_{ij} > 0$ for all i and j, so K is regular. Note also that $\sum_i \mathsf{K}_{ij} = 1$ for all j.
  - The eigenvalues and eigenvectors of K are

$$\begin{aligned} \lambda &= 1 \quad \Longrightarrow \quad \mathbf{P}_1 = (2/7, 2/7, 3/7) \\ \lambda &= -\frac{1}{6} \quad \Longrightarrow \quad \mathbf{P}_2 = (-3/14, -3/14, 6/14) \\ \lambda &= \frac{1}{2} \quad \Longrightarrow \quad \mathbf{P}_3 = (-3/7, 3/7, 0). \end{aligned}$$

– If initially $\mathbf{P}_s = (1, 0, 0)$, then $\mathbf{P}_2 = (1/2, 0, 1/2)$ and

$$\mathbf{P}_{10} = (0.28669\cdots, 0.28474\cdots, 0.42857\cdots),$$

which differs from $\mathbf{P}_1$ by 0.1%.

2.3.3 Construction of the transition matrix $\mathsf{K}(y \to x)$

We wish to devise a procedure so that the limiting (stationary) distribution of the random walk is the Boltzmann distribution $P_{eq}(x)$.

- Break up the transition matrix into two parts, generation of trial states $\mathsf{T}(y \to x)$ and acceptance probability of trial state $\mathsf{A}(y \to x)$

$$\mathsf{K}(y \to x) = \mathsf{T}(y \to x)\mathsf{A}(y \to x).$$

  – Will consider definition of acceptance probability given a specified procedure for generating trial states with condition probability $\mathsf{T}(y \to x)$.

- Write dynamics in term of probabilities at time t

$$\begin{aligned}
\text{Probability of starting in } y \text{ and ending in } x &= \int dy\, P_t(y)\, \mathsf{K}(y \to x) \\
\text{Probability of starting in } x \text{ and remaining in } x &= P_t(x) \left(1 - \int dy\, \mathsf{K}(x \to y)\right)
\end{aligned}$$

so

$$\begin{aligned}
P_{t+1}(x) &= \int dy\, P_t(y)\, \mathsf{K}(y \to x) + P_t(x) \left(1 - \int dy\, \mathsf{K}(x \to y)\right) \\
&= P_t(x) + \int dy \left(P_t(y)\mathsf{K}(y \to x) - P_t(x)\mathsf{K}(x \to y)\right).
\end{aligned}$$

- Thus, to get $P_{t+1}(x) = P_t(x)$, we want

$$\int dy \left(P_t(y)\mathsf{K}(y \to x) - P_t(x)\mathsf{K}(x \to y)\right) = 0.$$

- This can be accomplished by requiring *microscopic reversibility* or *detailed balance*:

$$\begin{aligned}
P_t(y)\mathsf{K}(y \to x) &= P_t(x)\mathsf{K}(x \to y) \\
P_t(y)\mathsf{T}(y \to x)\mathsf{A}(y \to x) &= P_t(x)\mathsf{T}(x \to y)\mathsf{A}(x \to y).
\end{aligned}$$

– We desire $P_t(x) = P_{t+1}(x) = P_{eq}(x)$. This places restriction on the acceptance probabilities

$$\frac{\mathsf{A}(y \to x)}{\mathsf{A}(x \to y)} = \frac{P_{eq}(x)\mathsf{T}(x \to y)}{P_{eq}(y)\mathsf{T}(y \to x)}.$$

– Note that $P_{eq}(x)\mathsf{K}(x \to y)$ is the probability of observing the sequence $\{x, y\}$ in the random walk. This must equal the probability of the sequence $\{y, x\}$ in the walk.

- Metropolis Solution: Note that if $P_{eq}(x)\mathsf{T}(x \to y) > 0$ and $P_{eq}(y)\mathsf{T}(y \to x) > 0$, then we can define

$$\begin{aligned} \mathsf{A}(y \to x) &= \min\left(1, \frac{P_{eq}(x)\mathsf{T}(x \to y)}{P_{eq}(y)\mathsf{T}(y \to x)}\right) \\ \mathsf{A}(x \to y) &= \min\left(1, \frac{P_{eq}(y)\mathsf{T}(y \to x)}{P_{eq}(x)\mathsf{T}(x \to y)}\right). \end{aligned}$$

– This definition satisfies details balance.

Proof. Verifying explicitly, we see that

$$\begin{aligned} P_{eq}(y)\mathsf{T}(y \to x)\mathsf{A}(y \to x) &= \min\left(P_{eq}(y)\mathsf{T}(y \to x), P_{eq}(x)\mathsf{T}(x \to y)\right) \\ P_{eq}(x)\mathsf{T}(x \to y)\mathsf{A}(x \to y) &= \min\left(P_{eq}(x)\mathsf{T}(x \to y), P_{eq}(y)\mathsf{T}(y \to x)\right) \\ &= P_{eq}(y)\mathsf{T}(y \to x)\mathsf{A}(y \to x) \end{aligned}$$

as required. □

– Procedure guaranteed to generate states with probability proportional to Boltzmann distribution if proposal probability $\mathsf{T}(y \to x)$ generates an ergodic transition probability K.

Simple example

- Suppose we want to evaluate the equilibrium average $\langle A \rangle$ at inverse temperature β of some dynamical variable $A(x)$ for a 1-dimensional harmonic oscillator system with potential energy $U(x) = kx^2/2$.

$$\langle A(x) \rangle = \int_{-\infty}^{\infty} dx\, P_{eq}(x) A(x) \qquad P_{eq}(x) = \frac{1}{Z} e^{-\beta k x^2/2}$$

- Monte-Carlo procedure

1. Suppose the current state of the system is x_0. We define the proposal probability of a configuration y to be

$$\mathsf{T}(x_0 \to y) \;=\; \begin{cases} \frac{1}{2\Delta x} & \text{if } y \in [x_0 - \Delta x, x_0 + \Delta x] \\ 0 & \text{otherwise} \end{cases}$$

   - Δx is fixed to some value representing the maximum displacement of the trial coordinate.
   - y chosen uniformly around current value of x_0.
   - Note that if y is selected, then $\mathsf{T}(x_0 \to y) = 1/(2\Delta x) = \mathsf{T}(y \to x_0)$ since x_0 and y are within a distance Δx of each other.

2. Accept trial y with probability

$$\mathsf{A}(x_0 \to y) = \min\left(1, \frac{\mathsf{T}(y \to x_0) P_{eq}(y)}{\mathsf{T}(x_0 \to y) P_{eq}(x_0)}\right) = \min\left(1, e^{-\beta \Delta U}\right),$$

   where $\Delta U = U(y) - U(x_0)$.

   - Note that if $\Delta U \leq 0$ so the proposal has lower potential energy than the current state, $\mathsf{A}(x_0 \to y) = 1$ and the trial y is always accepted as the next state $x_1 = y$.
   - If $\Delta U > 0$, then $\mathsf{A}(x_0 \to y) = e^{-\beta \Delta U} = q$. We must accept the trial y with probability $0 < q < 1$. This can be accomplished by picking a random number r uniformly on $(0, 1)$ and then:
     (a) If $r \leq q$, then accept configuration $x_1 = y$
     (b) If $r > q$, then reject configuration and set $x_1 = x_0$ (keep state as is).

3. Repeat steps 1 and 2, and record states $\{x_i\}$.

- Markov chain of states (the sequence) $\{x_i\}$ generated, with each x_i appearing in the sequence with probability $P_{eq}(x_i)$ (after some number of equilibration steps).

- After collecting N total configurations, the equilibrium average $\langle A \rangle$ can be estimated from

$$\langle A \rangle = \frac{1}{N} \sum_{i=1}^{N} A(x_i)$$

  since the importance function $w(x) = P_{eq}(x)$.

- Rejected states **are** important to generate states with correct weight.

- Typically, should neglect the first N_s points of Markov chain since it takes some iterations of procedure to generate states with stationary distribution of $\mathsf{K}(y \to x)$.

  - Called equilibration or "burn in" time.
  - Often have correlation among adjacent states. Can record states every N_{corr} Monte-Carlo steps (iterations).

- Why must rejected states be counted? Recall that $\mathsf{K}(x \to y)$ must be normalized to be a probability (or probability density).

$$\int dy\, \mathsf{K}(x \to y) \;\; = \;\; 1 = \int dy\, \left[\delta(x-y) + (1-\delta(x-y))\right] \mathsf{K}(x \to y)$$

  hence, the probability to remain in the current state is

$$\mathsf{K}(x \to x) = 1 - \int dy\, (1-\delta(x-y))\, \mathsf{K}(x \to y)$$

  - $\mathsf{K}(x \to x)$ is non-zero to insure normalization, so we **must** see the sequence $\{\ldots, x, x, \ldots\}$ in the Markov chain.

2.4 Statistical Uncertainties

We would like some measure of the reliability of averages computed from a finite set of sampled data. Consider the average $\overline{x}$ of a quantity x constructed from a finite set of N measurements $\{x_1, \ldots, x_N\}$, where the x_i are random variables drawn from a density $\rho(x)$.

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Suppose the random variables x_i are drawn independently, so that the joint probability satisfies

$$\rho(x_1, \ldots x_N) = \rho(x_1)\rho(x_2) \cdots \rho(x_N).$$

- Suppose that the intrinsic mean $\langle x \rangle$ of the random variable is $\langle x \rangle = \int dx\, x\rho(x)$ and that the intrinsic variance is σ^2, where

$$\sigma^2 = \int dx\, \big(x - \langle x \rangle\big)^2 \rho(x).$$

- A mesaure of reliability is the standard deviation of $\overline{x}$, also known as the *standard error* $\sigma_E = \sqrt{\sigma_E^2}$, where σ_E^2 is the variance of the finite average $\overline{x}$ around $\langle x \rangle$ in the finite set $\{x_1, \ldots, x_N\}$

$$\sigma_E^2 = \langle \big(\overline{x} - \langle x \rangle\big)^2 \rangle.$$

- We expect
  1. Variance (error) decreases with increasing N
  2. Error depends on the intrinsic variance σ^2 of the density ρ. Larger variance should mean slower convergence of $\overline{x}$ to $\langle x\rangle$.

Suppose all the intrinsic moments $\langle x^n\rangle$ of $\rho(x)$ exist. If we define the dimensionless variable

$$z = \frac{\overline{x} - \langle x\rangle}{\sigma_E},$$

and note that the probability density of z can be written as

$$\begin{aligned} P(\tilde{z}) &= \langle \delta(z-\tilde{z})\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, \langle e^{-it(z-\tilde{z})}\rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, e^{it\tilde{z}}\langle e^{-itz}\rangle \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} dt\, e^{it\tilde{z}}\chi_N(t), \end{aligned} \tag{2.1}$$

where $\chi_N(t)$ is the *characteristic function* of $P(z)$. Using the fact that $\sigma_E = \sigma/\sqrt{N}$, as will be shown shortly, z can be expressed as

$$z = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\frac{x_i - \langle x\rangle}{\sigma} = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{x_i - \langle x\rangle}{\sigma}\right),$$

and hence

$$\chi_N(t) = \left\langle e^{\frac{-it}{\sqrt{N}}\sum_{i=1}^{N}\left(\frac{x_i-\langle x\rangle}{\sigma}\right)}\right\rangle = \left(\left\langle e^{\frac{-it}{\sqrt{N}}\left(\frac{x-\langle x\rangle}{\sigma}\right)}\right\rangle\right)^N = \chi(t/\sqrt{N})^N,$$

where

$$\chi(u) = \left\langle e^{-iu(x-\langle x\rangle)/\sigma}\right\rangle,$$

which, for small arguments u can be expanded as $\chi(u) = 1 - u^2/2 + iu^3\kappa_3/(6\sigma^3) + \ldots$, where κ_3 is the *third cumulant* of the density $\rho(x)$. Using this expansion, we find $\ln\chi_N(t) = -t^2/2 + O(N^{-1/2})$, and hence $\chi_N(t) = e^{-t^2/2}$ for large N. Inserting this form in Eq. (2.1) gives

$$P(\tilde{z}) = \frac{1}{\sqrt{2\pi}}e^{-\tilde{z}^2/2}$$

and hence we find the *central limit theorem* result that the averages $\overline{x}$ are distributed around the intrinsic mean according to the normal distribution

$$N(\overline{x};\langle x\rangle,\sigma_E^2) = \frac{1}{\sqrt{2\pi}\sigma_E}e^{-\frac{(\overline{x}-\langle x\rangle)^2}{2\sigma_E^2}},$$

provided the number of samples N is large and all intrinsic moments (and hence cumulants) of $\rho(x)$ are finite.

- If the central limit theorem holds so that the finite averages are *normally distributed* as above, then the standard error can be used to define confidence intervals since the probability p that the deviations in the average, $\overline{x} - \langle x \rangle$, lie within a factor of c times the standard error is given by

$$\int_{-c\sigma_E}^{c\sigma_E} d(\overline{x} - \langle x \rangle) N(\overline{x}; \langle x \rangle, \sigma_{\overline{x}}^2) = p.$$

- If $p = 0.95$, defining the 95% confidence intervals, then we find that $c = 1.96$, and hence the probability that the intrinsic mean $\langle x \rangle$ lies in the interval $(\overline{x} - 1.96\sigma_E, \overline{x} + 1.96\sigma_E)$ is $P(\overline{x} - 1.96\sigma_E \leq \langle x \rangle \leq \overline{x} + 1.96\sigma_E) = 0.95$.

- If the data are not normally distributed, then higher cumulants of the probability density may be needed (skewness, kurtosis, ...). The confidence intervals are defined in terms of integrals of the probability density.

- How do we calculate σ_E or σ_E^2? We defined the variance as

$$\sigma_E^2 = \langle \left(\overline{x} - \langle x \rangle\right)^2 \rangle \qquad \text{where} \qquad \overline{x} = \frac{1}{N} \sum_i x_i$$

- Expanding the square, we get $\sigma_E^2 = \langle \overline{x}^2 \rangle - \langle x \rangle^2$ but

$$\begin{aligned} \langle \overline{x}^2 \rangle &= \frac{1}{N^2} \sum_{i,j} \langle x_i x_j \rangle = \frac{1}{N^2} \left(\sum_i \langle x_i^2 \rangle + \sum_i \sum_{j \neq i} \langle x_i \rangle \langle x_j \rangle \right) \\ &= \frac{1}{N^2} \left(N(\sigma^2 + \langle x \rangle^2) + N(N-1) \langle x \rangle^2 \right) = \frac{\sigma^2}{N} + \langle x \rangle^2, \end{aligned}$$

  hence $\sigma_E^2 = \sigma^2 / N$.

- However we don't really know what σ^2 is, so we must estimate this quantity. One way to do this is to define the estimator of the standard deviation $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{N} \sum_i \left(x_i - \overline{x}\right)^2 .$$

- The average of this estimator is

$$\langle\hat{\sigma}^2\rangle = \frac{1}{N}\sum_i \langle(x_i-\overline{x})^2\rangle = \frac{1}{N}\sum_i \langle x_i^2 - 2x_i\overline{x} + \overline{x}^2\rangle.$$

  but using the facts that

$$\langle x_i\overline{x}\rangle = \frac{\langle x^2\rangle}{N} + \frac{N-1}{N}\langle x\rangle^2 \qquad \langle\overline{x}^2\rangle = \frac{\sigma^2}{N} + \langle x\rangle^2,$$

  we get $\langle\hat{\sigma}_E^2\rangle = (N-1)\sigma^2/N$ and hence $\sigma^2 = N\langle\hat{\sigma}^2\rangle/(N-1)$.

- The standard error can therefore be estimated by

$$\sigma_E = \frac{\sqrt{\hat{\sigma}^2}}{\sqrt{N-1}}$$

- Note that σ_E decreases as $1/\sqrt{N-1}$
  - Narrowing confidence intervals by a factor of 2 requires roughly 4 times more data.
  - If the x_i are **not** indepenent, then $N \sim N_{eff}$, where N_{eff} is the effective number of independent configurations. If τ_c is the correlation length (number of Monte-Carlo steps over which correlation $\langle x_i x_j\rangle$ differs from $\langle x\rangle^2$), then $N_{eff} = N/\tau_c$.

- Another way to estimate the intrinsic variance σ^2 is to use the Jackknife Method: see B. Efron, *The annals of statistics* **7**, 1 (1979).
  - Jackknife procedure is to form N averages from the set $\{x_1,\ldots,x_N\}$ by omitting one data point. For example, the jth average is defined to be

$$\overline{x}^{(j)} = \frac{1}{N-1}\sum_{i\neq j}^{N} x_i$$

    The result is to form a set of averages $\{\overline{x}^{(1)},\overline{x}^{(2)},\ldots\overline{x}^{(N)}\}$.
  - The average $\overline{x}$ over this Jackknifed set is the same as before since

$$\overline{x} = \frac{1}{N}\sum_{j=1}^{N}\overline{x}^{(j)} = \frac{1}{N(N-1)}\sum_{j=1}^{N}\sum_{i\neq j} x_i = \frac{1}{N}\sum_{i=1}^{N} x_i.$$

– We define an estimator of the variance Σ^2 of the Jackknifed set to be

$$\Sigma^2 = \frac{1}{N}\sum_{j=1}^{N}\left(\overline{x}^{(j)} - \overline{x}\right)^2.$$

which converges to $\langle\Sigma^2\rangle$ as N gets large.

– Inserting the definitions of $\overline{x}^{(j)}$ and $\overline{x}$ and using the independence of the x_i, we see that the estimator is related to the intrinsic variance by

$$\langle\Sigma^2\rangle = \left(\frac{1}{N-1} - \frac{1}{N}\right)\sigma^2 = \frac{\sigma^2}{N(N-1)} \qquad \sigma^2 = N(N-1)\langle\Sigma^2\rangle.$$

– The standard error can therefore be estimated using the Jackknifed set as

$$\Sigma_E = \sqrt{\frac{N-1}{N}\sum_{j=1}^{N}\left(\overline{x}^{(j)} - \overline{x}\right)^2}.$$