Chapter 7

Markov Chain Monte-Carlo (MCMC)

7.1 The Problem

Quite often integrals like

$$I = \int d^N x \, f(\mathbf{x}) g(\mathbf{x}), \qquad \mathbf{x} \in \mathbb{R}^N, \tag{7.1}$$

with $g(x) \geq 0$ a PDF, are to be evaluated. Typically, $N \gg 1$. The standard example for this kind of an integral stems from statistical physics and is given by

$$\langle F \rangle = \int d^N x \, f(\mathbf{x}) \underbrace{\frac{e^{-\beta E(\mathbf{x})}}{Z}}_{=g(\mathbf{x})},$$

with the partition sum Z:

$$Z = \int d^N x \, e^{-\beta E(\mathbf{x})}.$$

Here, N is the number of particles in the one dimensional problem and equal to three times the number of particles in a 3D problem. As it is virtually impossible to calculate such integrals analytically, the value I is to be estimated using Monte-Carlo techniques.

Another important problem is defined by

$$I = \sum_{x_1,\ldots,x_n \in X} f(x) P(X). \tag{7.2}$$

Here, $X = \{x_1, \ldots, x_n\}$ is a set discrete states (for instance, the ISING model has $X = \{+1, -1\}$) and $P(X)$ is the probability which is given by

$$P(X) = Z^{-1} e^{-\beta E(X)}$$

as follows from statistical physics. Furthermore,

$$Z = \sum_X e^{-\beta E(X)}.$$

and the sum I is to be calculated using Monte-Carlo methods.

The POTTS Model, an Example

Many spin models can be found in the literature; some were born from theoretical models, some are based on experiments. The POTTS model is a classical model and is related to the physisorption of Krypton atoms on a graphite surface.

The POTTS model is defined by a q-state variable, $\sigma = 1, 2, 3, \ldots, q$, which exists on each lattice site. The interaction between spins is described by the Hamiltonian

$$\mathcal{H} = -\sum_{ij} J_{ij} \delta_{\sigma_i, \sigma_j}. \tag{7.3}$$

The sum goes over all lattice sites and J_{ij} describes the exchange interaction between lattice site i and j. Furthermore,

$$-J_{ij}\delta_{\sigma_i \sigma_j} = \begin{cases} -J_{ij} & \sigma_i = \sigma_j \quad \text{ferromagnetism} \\ 0 & \sigma_i \neq \sigma_j. \end{cases}$$

In the special case of nearest neighbor interaction, the Hamiltonian (7.3) simplifies into:

$$\mathcal{H} = -J \sum_{\langle ij \rangle} \delta_{\sigma_i, \sigma_j}. \tag{7.4}$$

Obviously, the POTTS model has q equivalent ground states in which all spins are identical but can acquire any of the q possible values. The model is identical to the spin-1/2 ISING model for $q = 2$ but it is no longer equivalent to the spin-1 ISING model for $q = 3$ because the three states of the spin-1 ISING model are not equivalent to each other. Increasing the temperature results in a transition to a paramagnetic phase. This transition is continuous for $q \leq 4$ but is of first order for $q > 4$. This makes the POTTS model a model for collective magnetism, and it is also a valuable test model for phase transitions.

If an external B-field is applied in the 1-direction we get a modified Hamiltonian (7.4):

$$\mathcal{H}_B = -B \sum_i \delta_{\sigma_i, 1}.$$

There are the following observables:

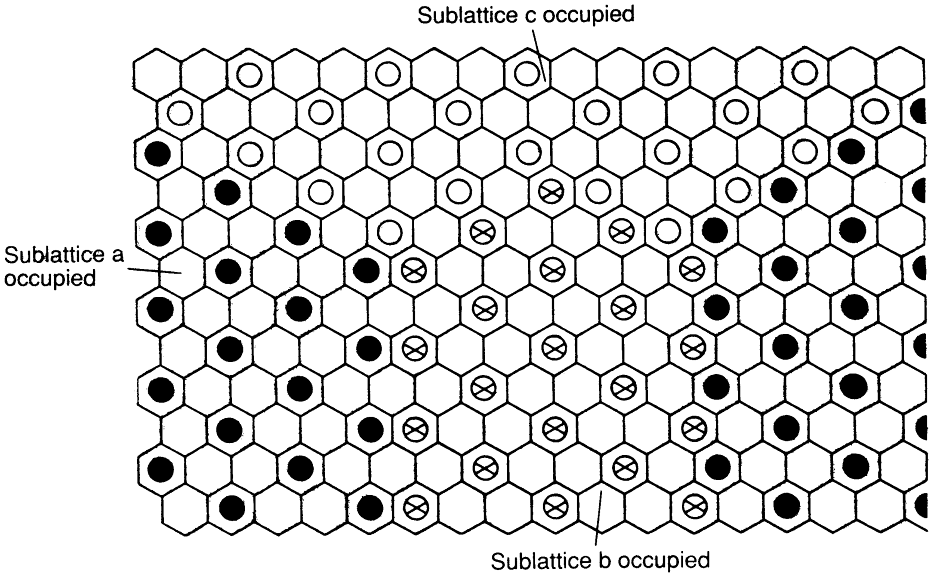

Figure 7.1: Krypton atoms adsorbed on the basal plane of graphite show coexisting regions of three ground states. According to M. KARDAR and A.N. BERKER, Phys. Rev. Lett. **48**, 1552 (1982).

- The internal energy

$$\langle E\rangle = \sum_{\mathbf{S}} \left(-J \sum_{\langle ij\rangle} \delta_{\sigma_i,\sigma_j} \right) P(\mathbf{S}),$$

  with S the complete set of states.

- Magnetization:

$$\langle M\rangle = \sum_{\mathbf{S}} \left(B \sum_{i} \delta_{\sigma_i,1} \right) P(\mathbf{S}).$$

- Magnetic susceptibility:

$$\chi = \frac{\partial}{\partial B}\langle M\rangle = \beta \left\langle (\Delta M)^2 \right\rangle .$$

Thus, POTTS model is a good example for expressions of type (7.2).

The POTTS model can be realized by Krypton atoms which have been adsorbed on the basal plane of graphite. The surface of graphite exists of hexagonal rings of carbon atoms and it is advantageous for an adsorbed krypton atom to come to rest inside of such a ring. On the other hand, krypton atoms are rather large and is not favorable for another krypton atom to populate a carbon ring in the immediate neighborhood of an already occupied carbon ring. Thus, the krypton atoms occupy, at best, only one third of the triangular lattice. Nevertheless, there are three completely equivalent positions for these triangular krypton lattices, the sub-lattices a, b, and c shown in Fig. 7.1. Thus, the system shows the symmetry of a $q = 3$ POTTS model. The lattice position i corresponds to a triplet of adsorption rings and $\sigma_i = 1, 2, 3$ corresponds to the three possibilities that the adsorbed krypton atoms belong to the sub-lattices a, b, or c.

7.2 Solution

An estimate of

$$\langle f(X) \rangle = \frac{\sum\!\!\!\!\!\!\int_x f(x)g(x)}{\int dx\, g(x)} \tag{7.5}$$

is to be determined using Monte-Carlo integration. Here, $g(x)$ is a PDF and $f(x)$ is the function one is interested in. In Monte-Carlo integration samples $\{X_t, t = 1, \ldots, n\}$ from $g(x)$ are generated and Eq. (7.5) is approximated by:

$$\langle f(x) \rangle \approx \frac{1}{n} \sum_{i=1}^{n} f(X_t). \tag{7.6}$$

This replaces the expectation value $\langle f(X) \rangle$ by the sample expectation value $\langle f(x) \rangle$. If the set of samples $\{X_t\}$ is independent (uncorrelated) it is guaranteed by the central limit theorem that it is possible to make the approximation as exact as it is required by simply increasing n.

Nevertheless, there is no need for the set $\{X_t\}$ to be independent. This set may be generated by any process which allows to draw samples correctly weighted by the PDF $g(x)$. Obviously, one possibility exists in the use of a *Markov chain* which uses $g(x)$ as the stationary PDF. This possibility resulted in the development of the *Markov Chain Monte-Carlo (MCMC) Method* or *Dynamic Monte Carlo Simulation*.

It is, in principle, rather easy to set up a dynamic Monte Carlo method for generating samples from an equilibrium distribution π. It suffices to invent a stochastic matrix $\boldsymbol{P}$ satisfying the following two conditions:

(a) *Ergodicity (Irreducibility).* For each pair x_α, $x_\beta \in \mathcal{S}$, there exists an $n \geq 0$ for which $p_{\alpha\beta}^{(n)} > 0$.

(b) *Stationarity of* π. For each $x_\beta \in \mathcal{S}$,

$$\sum_\alpha \pi_\alpha p_{\alpha\beta}^{(n)} = \pi_\beta.$$

Then Theorem 1.4 shows that simulation of the MARKOV chain $\boldsymbol{P}$ constitutes a legitimate Monte Carlo method for estimating averages with respect to π. We can start the system in any state x_α, and the system is guaranteed to converge to equilibrium as $t \to \infty$ [at least in the averaged sense of Eq. (1.25)]. Long-time averages of any possible f will converge with probability 1 to π-averages (strong law of large numbers), and will do so with fluctuations of size $\sim n^{-1/2}$ (central limit theorem).

Often the MARKOV chain is started in some chosen configuration x_α. For instance, in an ISING model, x_α might be the configuration with "all spins up"; this is sometimes called an *ordered* or *cold* start. Alternatively, the MARKOV chain might be started in a random configuration chosen according to some simple probability distribution p_α. For instance, in an ISING model, we might initialize the spins randomly and independently, with equal probabilities of up and down; this is sometimes called a *random* or *hot* start. In all these cases, the initial distribution p_α is clearly not equal to the equilibrium distribution π. Therefore, the system is initially "out of equilibrium". Theorem 1.4 guarantees that the system approaches equilibrium as $t \to \infty$, but we need to know something about the *rate* of convergence to equilibrium.

Lacking rigorous knowledge of the rate of convergence one assumes that after a long *equilibration phase*, say, after m iterations the set of samples $\{X_t, t = m+1, \ldots, n\}$ to be approximately distributed according to the equilibrium distribution π. The results of this equilibration phase will be discarded and we find the estimate

$$\langle f(x) \rangle = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t);$$

the *ergodic mean*.

There is a last problem to be solved: how to construct a MARKOV chain which has a equilibrium distribution π which mimics the PDF $g(x)$ of Eq. (7.5)? A possibility can be found in Algorithm 4. The probability of acceptance of a test state X_t is given by:

$$P(AX_t|X) = \min\left(1, \frac{g(X_t)}{g(X)}\right).$$

If we have to deal with a BOLTZMANN distribution, we get

$$\frac{g(X_t)}{g(X)} = e^{-\beta E(X_t) + \beta E(X)} = e^{-\beta \Delta E},$$

with ΔE the energy difference between the two states $\{X_t\}$ and $\{X\}$.

7.3 Statistical Analysis

A simulation can nowadays be divided into two parts, the data generation part and the data analysis part. The interface between these two parts consists of time series of measurements of relevant physical observables taken during the actual simulation. Once the system is in equilibrium (which, in general, is non-trivial to assure), we save $X_n = X[\{x_i\}]_n$, where n labels the measurements.

7.3.1 Estimates

If the time series data result from an importance sampling MC simulation, the expectation value $\langle X \rangle$ can be estimated as as simple arithmetic mean over the MARKOV chain:

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j, \tag{7.7}$$

where we assumed that the time series contains a total of N measurements. Conceptually, it is important to distinguish between the expectation value $\langle X \rangle$ which is an ordinary number, and the *estimator* $\bar{x}$ which is a random number fluctuating around the theoretically expected value. Of course, in practice one does not probe the fluctuations of the mean value directly (which would require repeating the whole MC simulation many times) but rather estimate its variance

$$\mathrm{var}(\bar{x}) = \left\langle [\bar{x} - \langle \bar{x} \rangle]^2 \right\rangle = \left\langle \bar{x}^2 \right\rangle - \langle \bar{x} \rangle^2 \tag{7.8}$$

from the distribution of the individual measurements X_j. If the N subsequent measurements were all uncorrelated, then the relation would simply be

$$\mathrm{var}(\bar{x}) = \frac{\mathrm{var}(X_j)}{N}, \quad \mathrm{var}(X_j) = \left\langle X_j^2 \right\rangle - \langle X_j \rangle^2 . \tag{7.9}$$

$\mathrm{var}(X_j)$ is the variance of the individual measurement. Here, one assumes, of course, that the simulation is in equilibrium and uses time-translational invariance of the MARKOV chain. (See Sec. 1.3.2.) Equation (7.9) is true for any distribution $P(X_j)$ of X_j.

Whatever form the distribution $P(X)$ assumes, by the central limit theorem, the distribution of the mean value is Gaussian, at least for uncorrelated data in the asymptotic limit of large N. The variance of the mean, $\mathrm{var}(\bar{x})$, is the squared width of this (N dependent) distribution which is usually taken as the "one-sigma" squared error $\varepsilon_{\bar{x}}^2 = \mathrm{var}(\bar{x})$, and quoted together with the mean value $\bar{x}$. Under the assumption of a Gaussian distribution, the interpretation is that about 68% of all simulations under the same conditions would yield a mean value in the range $[\bar{x} - \mathrm{std}(\bar{x}), \bar{x} + \mathrm{std}(\bar{x})]$.

7.3.2 Autocorrelation Times

Things become more involved for correlated measurements. Starting from Eq. (7.8) and by inserting Eq. (7.7) we obtain

$$\text{var}(\bar{x}) = \left\langle \bar{x}^2 \right\rangle - \left\langle \bar{x} \right\rangle^2 = \frac{1}{N^2} \sum_{i,j=1}^{N} \left\langle X_i X_j \right\rangle - \frac{1}{N^2} \sum_{i,j=1}^{N} \left\langle X_i \right\rangle \left\langle X_j \right\rangle .$$

By collecting diagonal and off-diagonal terms, we arrive at

$$\text{var}(\bar{x}) = \frac{1}{N^2} \sum_{i=1}^{N} \left(\left\langle X_i^2 \right\rangle - \left\langle X_i \right\rangle^2 \right) + \frac{1}{N^2} \sum_{i \neq j}^{N} \left(\left\langle X_i X_j \right\rangle - \left\langle X_i \right\rangle \left\langle X_j \right\rangle \right) .$$

The first term is identified as $\text{var}(X_i)/N$. In the second term we first use the symmetry $i \longleftrightarrow j$ to reduce the summation

$$\sum_{i \neq j}^{N} \cdots = 2 \sum_{i=1}^{N} \sum_{j=i+1}^{N} \cdots$$

Then we reorder the summation and use time translational invariance to derive

$$\text{var}(\bar{x}) = \frac{1}{N} \left[\text{var}(X_i) + 2 \sum_{k=1}^{N} \left(\left\langle X_1 X_{k+1} \right\rangle - \left\langle X_1 \right\rangle \left\langle X_{k+1} \right\rangle \right) \left(1 - \frac{k}{N} \right) \right] ,$$

where, due to the last factor the $k = N$ term may trivially be kept in the summation. Factoring out $\text{var}(X_i)$, we get

$$\text{var}(\bar{x}) = \frac{\text{var}(X_i)}{N} 2 \tau'_{X,\text{int}} . \tag{7.10}$$

Here we have introduced the so-called (proper) *integrated autocorrelation time*

$$\tau'_{X,\text{int}} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \left(1 - \frac{k}{N} \right) \tag{7.11}$$

with

$$A(k) = \frac{\left\langle X_i X_{i+k} \right\rangle - \left\langle X_i \right\rangle \left\langle X_{i+k} \right\rangle}{\left\langle X_i^2 \right\rangle - \left\langle X_i \right\rangle \left\langle X_i \right\rangle} , \tag{7.12}$$

denoting the *renormalized autocorrelation function*, $A(0) = 1$. For large time separations the autocorrelation function decays exponentially

$$A(k) \overset{k \to \infty}{\longrightarrow} a e^{-k/\tau_{X,\text{exp}}} ,$$

where $\tau_{X,\text{exp}}$ is the *exponential autocorrelation time* and a a constant.

Due to the exponential decay of $A(k)$ as $k \to \infty$ in any meaningful simulation with $N \gg \tau_{X,\mathrm{exp}}$ the correction term in parenthesis in Eq. (7.11) can safely be neglected. The usually employed definition of the integrated autocorrelation time is thus

$$\tau_{X,\mathrm{int}} = \frac{1}{2} + \sum_{k=1}^{N} A(k).$$

Notice that, in general, $\tau_{X,\mathrm{int}}$ (and also $\tau'_{X,\mathrm{int}}$) is different from $\tau_{X,\mathrm{exp}}$. In fact, one can show that $\tau_{X,\mathrm{int}} \leq \tau_{X,\mathrm{exp}}$ in realistic models.

The important point of Eq. (7.10) is that due to temporal correlations of the measurements the statistical error $\varepsilon_{\bar{x}}$ on the MC estimator $\bar{x}$ is enhanced by a factor of $\sqrt{2\tau_{X,\mathrm{int}}}$. This can be rephrased by writing the statistical error similar to the uncorrelated case as $\varepsilon_{\bar{x}} = \sqrt{\mathrm{var}(X_j)/N_{\mathrm{eff}}}$ but now with a parameter

$$N_{\mathrm{eff}} = \frac{N}{2\tau_{X,\mathrm{int}}} \leq N,$$

describing the effective statistics. This shows more clearly that only every $2\tau_{X,\mathrm{int}}$ iterations the measurements are approximately uncorrelated and gives a better idea of the relevant effective size of the statistical sample.

An estimator $\hat{A}(k)$ for the autocorrelation function is obtained by replacing in Eq. (7.12) the expectation values by mean values, e.g.: $\langle X_i X_{i+k}\rangle$ by $\overline{X_i X_{i+k}}$. With increasing k the relative variance of $\hat{A}(k)$ diverges rapidly. To get, at least, an idea of the order of magnitude of $\tau_{X,\mathrm{int}}$ and, thus, the correct error estimate (7.10) it is useful to record the "running" autocorrelation time estimator

$$\tau_{X,\mathrm{int}}(k_{\mathrm{max}}) = \frac{1}{2} + \sum_{k=1}^{k_{\mathrm{max}}} \hat{A}(k)$$

which approaches $\tau_{X,\mathrm{int}}$ in the limit of large k_{max} where, however, the statistical error increases rapidly.

7.3.3 Conclusion

There are two fundamental - and quite distinct - issues in dynamic MC simulation:

- *Initialization bias.* If the MARKOV chain is started in a distribution $p(x_\alpha)$ which is not equal to the stationary distribution π, then there is an "initial transient" in which the data do not reflect the desired equilibrium distribution π. This results in a systematic error (bias).

- *Autocorrelation in equilibrium.* The MARKOV chain, once it reaches equilibrium, provides *correlated* samples from π. This correlation causes the *statistical error* (variance) to be a factor $2\tau_{X,\mathrm{int}}$ larger than in independent sampling.

Initialization bias

The system is initially out of equilibrium. Theorem 1.4 guarantees that the system approaches equilibrium as $t \to \infty$ and we need to know the rate of convergence to equilibrium. Using the exponential autocorrelation time τ_{exp} one can set an upper bound on he amount of time we have to wait before equilibrium is "for all practical purposes" attained. There are two difficulties with this bound. Firstly, it is usually impossible to apply it in practice, since one almost never knows τ_{exp}. Secondly, even if we can apply it, it may be overly conservative; indeed, there exist perfectly reasonable algorithms with $\tau_{\text{exp}} = \infty$.

Lacking rigorous knowledge of the autocorrelation time τ_{exp} one should try to estimate it both theoretically and empirically. To make a heuristic theoretical estimate of τ_{exp}, we attempt to understand the physical mechanism(s) causing slow convergence to equilibrium. To make a rough empirical estimate, we measure the (unnormalized) autocorrelation function for a suitably large set of observables X. In both cases there is always the chance to overlook something and to grossly underestimate τ_{exp}. Nevertheless, it is usual to determine *empirically* when "equilibrium" has been achieved by plotting selected observables as a function of time and noting when the initial transient disappears.

Once we know (or guess) the time needed to attain "equilibrium", we simply discard the data from the initial transient up to some time and include only the subsequent data in the averages.

Autocorrelation in Equilibrium

As already explained in the preceeding section, the variance of the sample mean $\bar{x}$ in a dynamic MC method is a factor $2\tau_{X,\text{int}}$ higher than it would be in independent sampling. Otherwise put, a run of length N contains only $N/(2\tau_{X,\text{int}})$ "effectively independent data points". This means that the *computational efficiency* of the algorithm is determined principally by its autocorrelation time. The knowledge of $\tau_{X,\text{int}}$ is also essential for determining run lengths and for setting error bars on the estimates of the expectation values. Roughly speaking, the error bars will be of the order $\sqrt{\tau/N}$. Above all, there is a basic self-consistency requirement: the run length N must be much greater than the estimates of τ produced by that same run, otherwise **none** of the results from that run should be believed. While self-consistency is a *necessary* condition for the trustworthiness of MC data, it is not a *sufficient* condition.

7.4 The METROPOLIS-HASTINGS Algorithm

This method samples the test state $\{X^T\}$ from a particular proposition probability.

1. We start with the state $\{X_\nu\}$ of the MARKOV chain after ν steps.

2. The test state $\{X^T\}$ is sampled according to a proposition probability $q(X^T|X_\nu)$:

$$q(X^T|X) \geq 0, \qquad \int dX^T \, q(X^T|X) = 1.$$

   (a) We get, for instance for the POTTS model:

      i. Choose a lattice site using:

$$i = \text{int}(rN) + 1,$$

      ii. choose the state at lattice site i

$$\sigma_i = \text{int}(rq) + 1,$$

   using an equally distributed random number $r \in [0, 1)$. This generates a new test state $\mathbf{S}_\nu^T$ from $\mathbf{S}_\nu$.

   (b) If there is a stationary number of degrees of freedom in $X \in \mathbb{R}^N$ one proceeds along the following steps:

      i. Sample states using the proposition probability

$$q(X^T|X_\nu) = (2\pi\sigma^2)^{-N/2} \exp\left\{ -\frac{\left(X^T - X_\nu\right)^2}{2\sigma^2} \right\},$$

      ii. and define a random direction, $\hat{\mathbf{n}}$ in $\mathbb{R}^N$:

$$\mathbf{X}^T = \mathbf{X}_\nu + \lambda \hat{\mathbf{n}},$$

      with λ sampled from a CAUCHY-distribution:

$$p(\lambda) = \frac{\beta}{\pi(\beta^2 + \lambda^2)}, \quad \beta > 0, \quad -\infty < \lambda < \infty.$$

3. The probability of acceptance is given by:

$$P(AX^T|X_\nu) = \min\left(1, \frac{g(X^T)}{g(X_\nu)} \frac{q(X_\nu|X^T)}{q(X^T|X_\nu)}\right). \tag{7.13}$$

   Obviously, for $q(X_\nu|X^T) = q(X^T|X_\nu)$, the METROPOLIS-HASTINGS algorithm is equivalent to the standard METROPOLIS-algorithm 4.

Proof of Eq. (1.24)**, "Detailed Balance"**

We apply the marginalization rule and rewrite the transition probability as:

$$P(X_{t+1}|X_t) = \int dX^T \, P(X_{t+1} \sqcap X^T|X_t).$$

The product rule, furthermore, results in

$$P(X_{t+1}|X_t) = \int dX^T \, P(X_{t+1}|X^T, X_t) \underbrace{P(X^T|X_t)}_{=q(X^T|X_t)},$$

with $q(X^T|X_t)$ the proposition probability. We mark the acceptance of a trial state X^T by the symbol A and its rejection by the symbol $\bar{A}$ and find applying, again, the marginalization rule:

$$\begin{aligned}
P(X_{t+1}|X_t) &= \int dX^T \, P(X_{t+1} \sqcap (A \sqcup \bar{A})|X^T, X_t) q(X^T|X_t) \\
&= \int dX^T \, \big[P(X_{t+1} \sqcap A|X^T, x_t) \\
&\quad + P(X_{t+1} \sqcap \bar{A}|X^T, x_t)\big] \, q(X^T|X_t) \\
&= \int dX^T \, \Bigg[\underbrace{P(X_{t+1}|AX^T, X_t)}_{=\delta(X^T - X_{t+1})} \underbrace{P(A|X^T, X_t)}_{=\alpha(X^T, X_t)} \\
&\quad + \underbrace{P(X_{t+1}|\bar{A}X^T, X_t)}_{=\delta(X_{t+1} - X_t)} \underbrace{P(\bar{A}|X^T, X_t)}_{=1-\alpha(X^T, X_t)}\Bigg] \, q(X^T|X_t) \\
&= \alpha(X_{t+1}, X_t) q(X_{t+1}|X_t) \\
&\quad + \delta(X_{t+1} - X_t) \int dX^T \, \left[1 - \alpha(X^T, X_t)\right] q(X^T|X_t).
\end{aligned} \tag{7.14}$$

Exchanging X_{t+1} and X_t in Eq. (7.14) results in:

$$\begin{aligned}
P(X_t|X_{t+1}) &= \alpha(X_t, X_{t+1}) q(X_t|X_{t+1}) \\
&\quad + \delta(X_{t+1} - X_t) \int dX^T \, \left[1 - \alpha(X^T, X_t)\right] q(X^T|X_t).
\end{aligned} \tag{7.15}$$

In order to arrive at Eq. (1.24) we multiply Eq. (7.14) with $P(X_t)$, the probability for the existence of state X_t and Eq. (7.15) with $P(X_{t+1})$. This results in:

$$\begin{aligned} P(X_{t+1}|X_t)P(X_t) &= \alpha(X_{t+1},X_t)q(X_{t+1}|X_t)P(X_t) \\ &\quad +\delta(X_{t+1}-X_t)P(X_t) \\ &\quad \times \int dX^T \left[1-\alpha(X^T,X_t)\right] q(X^T|X_t), \end{aligned} \tag{7.16}$$

$$\begin{aligned} P(X_t|X_{t+1})P(X_{t+1}) &= \alpha(X_t,X_{t+1})q(X_t|X_{t+1})P(X_{t+1}) \\ &\quad +\delta(X_{t+1}-X_t)P(X_{t+1}) \\ &\quad \times \int dX^T \left[1-\alpha(X^T,X_t)\right] q(X^T|X_t). \end{aligned} \tag{7.17}$$

We now make use of Eq. (7.13) which defines the probability of acceptance and find:

$$\begin{aligned} q(X|Y)P(Y)\alpha(X,Y) &= q(X|Y)P(Y)\text{min}\left(1,\frac{P(X)q(Y|X)}{P(Y)q(X|Y)}\right) \\ &= \text{min}\left(q(X|Y)P(Y)\frac{P(X)q(Y|X)}{P(Y)q(X|Y)},q(X|Y)P(Y)\right) \\ &= \text{min}\left(q(X|Y)P(Y),q(Y|X)P(X)\right). \end{aligned} \tag{7.18}$$

This expression is symmetric in X and Y and, after exchanging X and Y in Eq. (7.18), we arrive at:

$$\begin{aligned} q(Y|X)P(X)\alpha(Y,X) &= \text{min}\left(q(Y|X)P(X),q(X|Y)P(Y)\right) \\ &= q(X|Y)P(Y)\alpha(X,Y). \end{aligned}$$

This is, finally, used in Eq. (7.16) and we get the result:

$$P(X_{t+1}|X_t)P(X_t) = P(X_t|X_{t+1})P(X_{t+1}). \tag{7.19}$$

Thus, the METROPOLIS-HASTINGS-algorithm obeys Eq. (1.24) and it is, therefore, guaranteed that elements $P(X_{t+1}|X_t)$ of a stochastic matrix with distribution $P(X)$ are generated.

7.5 Importance Sampling

This problem is easily motivated: we define

$$\begin{aligned} \Theta &= \int_0^1 dx\, f(x) = \int_0^1 dx\, \frac{f(x)}{g(x)}g(x) \\ &= \int_0^1 dG(x)\frac{f(x)}{g(x)}, \end{aligned}$$

for any function $g(x)$, and, furthermore,

$$G(x) = \int_0^x dy\, g(y).$$

Let $g(x) > 0, \forall x \in [0,1]$ and

$$G(1) = \int_0^1 dy\, g(y) = 1. \tag{7.20}$$

Thus, $G(x)$ is a CDF in the interval $0 \leq x \leq 1$ and if r is a random number sampled on basis of $G(r)$ we conclude that $f(r)/g(r)$ has the expectation value Θ and the variance

$$\text{var}(f/g) = \int_0^1 dG(x) \left(\frac{f(x)}{g(x)} - \Theta \right)^2 .$$

Importance sampling is based on the idea to concentrate trial states in those parts of the interval which are "most important" for the problem at hands. (Usually, one would distribute the trial states equally over the interval.) If $f(x) \geq 0, \forall\, x \in [0,1]$ were valid one could choose $g(x)$ proportional to $f(x)$, for instance $g(x) = c\, f(x)$. As a consequence $c = 1/\Theta$ and $\text{var}(f/g) = 0$! This appears to be the perfect Monte Carlo method which always generates the exact result. But, if we want to make use of this method, we will have to sample $f(x)/g(x)$ and, thus, have to know $g(x)$. This requires the knowledge of Θ which we want to calculate.

Nevertheless, the consequence of this example is that one gets an estimate for Θ using *any* positive function $g(x)$. Thus, one has to find a function $g(x)$ which results in a variance reduction of our estimate. This estimate is in this case the mean value of observed values of $f(x)/g(x)$ and the sampling variance will be small if $f(x)/g(x)$ is as 'constant' as possible. Consequently, $g(x)$ should follow $f(x)$ as closely as possible. On the other hand, functions used for $g(x)$ are supposed to be easily integrable in order to comply with Eq. (7.20) without particular efforts.

Interesting enough, the MCMC method which was discussed in the previous section corresponds to importance sampling as we will demonstrate in the following paragraphs.

A set of states

$$C = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\}, \quad \mathbf{x}_i \in \mathbb{R}^N,$$

is generated according to the probability $P(\mathbf{x})$ [or according to the PDF

$g(\mathbf{x})$] using a Markov chain. As a result, Eq. (7.6) results in:

$$I = \sum\!\!\!\!\!\!\!\int_{\mathbf{x}} f(\mathbf{x})g(\mathbf{x}) = \lim_{M\to\infty}\left(\underbrace{\frac{1}{M}\sum_{i=1}^{M} f(\mathbf{x}_i)}_{=\mathcal{F}_M(\{\mathbf{x}\})}\right).$$

Here, $\mathcal{F}_M(\{\mathbf{x}\})$ depends on the samples $\{\mathbf{x}\}$ of the Markov chain.

We calculate the variance of $\mathcal{F}_M(\{\mathbf{x}\})$:

- The central limit theorem results in

  $$P(\mathcal{F}_M) = \mathcal{N}\left(\mu = I, \sigma = \frac{\sigma_f}{\sqrt{M}}\right), \quad M \gg 1,$$

  with $\mathcal{N}$ the normal distribution. We find immediately:

  $$\begin{aligned} \langle \mathcal{F}_M \rangle &= I \\ \text{var}\,(\mathcal{F}_M) &= \frac{\sigma_f^2}{M} = \frac{1}{M}\sum\!\!\!\!\!\!\!\int_{\mathbf{x}} (f(\mathbf{x}) - I)^2\, g(\mathbf{x}). \end{aligned}$$

  Nevertheless, the central limit theorem is only valid for $M \to \infty$ and $\text{cov}(\mathbf{x}_i, \mathbf{x}_j) = 0, \forall i \neq j$, and, thus, only for uncorrelated states $\{\mathbf{x}\}$. Therefore, it cannot be applied to make an estimate of the variance achieved with MCMC-methods which make use of correlated states on purpose.

- We start an alternative analysis with

  $$\langle \mathcal{F}_M \rangle = \frac{1}{M}\sum_{i=1}^{M} \langle f(\mathbf{x}_i) \rangle,$$

  and all $\mathbf{x}_i$ obey the PDF $g(\mathbf{x})$. We find

  $$\begin{aligned} \left\langle \mathcal{F}_M^2 \right\rangle &= \frac{1}{M^2}\sum_{i,j=1}^{M} \langle f(\mathbf{x}_i) f(\mathbf{x}_j) \rangle \\ &= \frac{1}{M^2}\sum_{i,j=1}^{M}\left(\underbrace{\langle \Delta f(\mathbf{x}_i)\Delta f(\mathbf{x}_j)\rangle}_{=C_{ij}} + \langle f(\mathbf{x}_i)\rangle \langle f(\mathbf{x}_j)\rangle\right) \\ &= \frac{1}{M^2}\sum_{i,j=1}^{M} C_{ij} + \langle f(\mathbf{x})\rangle^2, \end{aligned}$$

with the elements of the covariance matrix C_{ij}. We find for the variance

$$\begin{aligned} \text{var}(\mathcal{F}_M) &= \frac{1}{M^2}\sum_{i,j=1}^{M} C_{ij} \\ &= \frac{\text{var}(f(\mathbf{x}))}{M}\frac{1}{M}\sum_{i,j=1}^{M}\underbrace{\frac{\langle \Delta f(\mathbf{x}_i)\Delta f(\mathbf{x}_j)\rangle}{\text{var}(f(\mathbf{x})}}_{=R_{ij}}, \end{aligned}$$

with the autocorrelation coefficient R_{ij}. We get the important result:

$$\text{var}(\mathcal{F}_M) = \frac{\text{var}(f(\mathbf{x}))}{M}\left(1 + \frac{1}{M}\sum_{i\neq j=1}^{M} R_{ij}\right).$$

The autocorrelation coefficient is symmetric

$$R_{ij} = R_{ji} = a(|i-j|),$$

and of the general form

$$a(d) = \sum_{\nu=1}^{M} c_\nu e^{-d/\tau_\nu},$$

with the eigenvalues τ_ν of the covariance matrix. This expression is, for $d \gg 1$, dominated by the slow decay and we approximate

$$a(d) \sim e^{-d/\tau_{exp}},$$

where τ_{exp} is the exponential correlation time. This, furthermore, results in:

$$\begin{aligned} \frac{1}{M}\sum_{i\neq j=1}^{M} R_{ij} &= \frac{2M}{M}\sum_{d=1}^{\infty}\left(\underbrace{e^{-1/\tau_{exp}}}_{=q}\right)^{d} \\ &= 2\frac{q}{1-q}. \end{aligned}$$

Thus, we get in the most disadvantageous case $\tau_{exp} \gg 1$

$$\begin{aligned} q &\approx 1 - \frac{1}{\tau_{exp}} \\ \frac{1}{M}\sum_{i,j} R_{ij} &\sim 2\tau_{exp}, \end{aligned}$$

and consequently:

$$\text{var}(\mathcal{F}_M) \sim \frac{\text{var}(f(\mathbf{x}))}{M}(1 + 2\tau_{exp}) = \frac{\text{var}(f(\mathbf{x}))}{M_{eff}}.$$

Here, M_{eff} is the effective number of uncorrelated MCMC states.

This proves that the MCMC method can indeed result in a variance reduction.