1 Monte Carlo Estimation

Monte Carlo sampling is based on the *Law of Large Numbers* which states that if $\mathbf{X}$ is a random variable then we can approximate the expectation $\mathbb{E}(\mathbf{X})$ arbitrarily well by averages of independent samples from $\mathbf{X}$ i.e. if we take sample sets $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ then

$$\frac{1}{N}\sum_i \mathbf{x}_i \to \mathbb{E}(\mathbf{X}) \text{ as } N \to \infty$$

Also for all N the sample mean $\frac{1}{N}\sum_i \mathbf{x}_i$ is an unbiased estimator for the expectation i.e.

$$\mathbb{E}(\frac{1}{N}\sum_i \mathbf{x}_i) = \mathbb{E}(\mathbf{X})$$

If the samples are independent, the variance

$$\mathbb{V}(\frac{1}{N}\sum_i \mathbf{x}_i) = \frac{1}{N^2}\sum_i \mathbb{V}(\mathbf{x}_i) = \frac{1}{N^2} N\mathbb{V}(\mathbf{X}) = \frac{1}{N}\mathbb{V}(\mathbf{X})$$

Remark that this computation requires the samples to be independent otherwise the variance of a sum will involve some covariance terms

So Monte Carlo Estimation basically comes down to be able to sample from the appropriate distribution

Generating random samples from a distribution uses a random number generator which generates numbers that are (approximately) uniformly distributed.

Say we want to generate samples from a 1-dimensional random variable $\mathbf{X}$ with CDF

$$G(x) = \int_{-\infty}^{x} p(t)dt$$

This is an increasing function $G : \mathbb{R} \to [0, 1]$ and so it has an inverse G^{-1}.

Let $\mathbf{U} \sim Uniform([0, 1])$ then

$$Prob(G^{-1}(\mathbf{U}) < x) = Prob(\mathbf{U} < G(x)) = G(x)$$

so $G^{-1}(\mathbf{U})$ has the distribution of $\mathbf{X}$ and so if $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ are uniformly distributed in $[0, 1]$, $G^{-1}(\mathbf{u}_1), G^{-1}(\mathbf{u}_2), \ldots, G^{-1}(\mathbf{u}_N)$ are samples from $\mathbf{X}$

If $\mathbf{X}$ has density function p then

$$\mathbb{E}(f(\mathbf{X})) = \int f(x)p(x)dx$$

Suppose now that q is another distribution we can easily sample from, like a Gaussian for instance.

We can write

$$\begin{aligned}\mathbb{E}_p(f(\mathbf{X})) &= \int f(x)p(x)dx \\ &= \int f(x)\frac{p(x)}{q(x)}q(x)dx \\ &= \mathbb{E}_q(f(\mathbf{X})\frac{p(\mathbf{X})}{q(\mathbf{X})}) \\ &\approx \frac{1}{N}\sum f(\mathbf{y}_i)\frac{p(\mathbf{y}_i)}{q(\mathbf{y}_i)}\end{aligned}$$

where $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N \sim q$ are independent samples from the distribution q.

This is the principle of *importance sampling*.

The distribution q is sometimes called the *proposal distribution* and the ratios $w(\mathbf{y}) = \dfrac{p(\mathbf{y})}{q(\mathbf{y})}$ the *importance weights*

Suppose we want to sample from a distribution with density of the form $p = \dfrac{g}{Z}$ where $g : \mathbb{R}^d \to \mathbb{R}$ is a positive function and $Z = \int_{\mathbb{R}^d} g$ is the intractable normalization factor.

Let

$$A = \{(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R} | 0 < y < g(\mathbf{x})\}$$

(This is just the open set bounded by the graph of g and the subspace $\mathbb{R}^d \times \{0\}$)

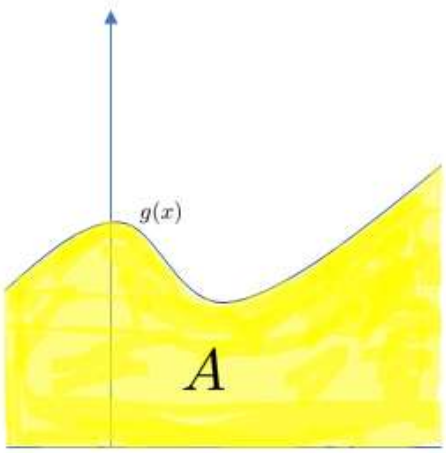

Figure 1:

If $\mathbf{X}$ and $\mathbf{Y}$ are random variables, $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}$, such $(\mathbf{X}, \mathbf{Y}) \in \mathbb{R}^{d+1}$ is uniformly distributed over A then $\mathbf{X} \sim p$.

To see this, remark that

$$vol(A) = \int_{\mathbb{R}^d} g(\mathbf{x}) d\mathbf{x} = Z$$

Let q be the density of $(\mathbf{X}, \mathbf{Y}) \sim Uniform(A)$. Then

$$q(\mathbf{x}, y) = \begin{cases} \dfrac{1}{vol(A)} \text{ if } (\mathbf{x}, y) \in A \\ 0 \text{ otherwise} \end{cases}$$

Thus

$$q(\mathbf{x}, y) = \frac{\mathbb{1}(0 < y < g(\mathbf{x})))}{vol(A)}$$

Marginalizing

$$p_{\mathbf{X}}(x) = \int_{-\infty}^{\infty} q(\mathbf{x}, y) dy = \frac{1}{vol(A)} \int_0^{g(\mathbf{x})} dy = \frac{g(\mathbf{x})}{Z}$$

We can use this to make a *proposal-acceptance/rejection* procedure to generate samples from a distribution p proportional to a positive function g.

- Choose a proposal distribution q which is close to proportional to g
- Choose a constant c such that $cq(\mathbf{x}) \geq g(\mathbf{x})$ for all $\mathbf{x}$
- Sample $\mathbf{x}$ from the distribution q
- Sample y from $Uniform(0, cq(\mathbf{x}))$
- Only samples which satisfy $y < g(\mathbf{x})$ are accepted

The accepted samples are then samples from the distribution p

Indeed let

$$A = \{(\mathbf{x}, y) \in \mathbb{R}^{d+1} | 0 < y < g(\mathbf{x})\}$$

and

$$B = \{(\mathbf{x}, y) \in \mathbb{R}^{d+1} | 0 < y < cq(\mathbf{x})\}$$

so $A \subset B$.

Now

$$q(\mathbf{x}) Uniform(y, [0, cq(\mathbf{x})]) = q(\mathbf{x}) \frac{\mathbb{1}\{0 < y < cq(\mathbf{x})\}}{cq(\mathbf{x})} = \frac{\mathbb{1}\{(\mathbf{x}, y) \in A\}}{c}$$

and $c = c \int_{\mathbb{R}^d} q(\mathbf{x}) d\mathbf{x} = \int_{\mathbb{R}^d} cq(\mathbf{x}) d\mathbf{x} = vol(B)$. So $(\mathbf{X}, \mathbf{Y}) \sim Uniform(B)$.

The rejection part only keeps samples $(\mathbf{x}, y)$ that lies in A.

The uniform distribution on B is characterized by

$$Prob((\mathbf{x}, y) \in C) = \frac{vol(C)}{vol(B)}$$

for any measurable subset $C \subset B$. If $C \subset A$ and it is given that $(\mathbf{x}, y) \in A$ then

$$Prob((\mathbf{x}, y) \in C | A) = \frac{Prob((\mathbf{x}, y) \in C)}{Prob((\mathbf{x}, y) \in A)} = \frac{vol(C)}{vol(B)} / \frac{vol(A)}{vol(B)} = \frac{vol(C)}{vol(A)}$$

This shows that the accepted samples are uniform on A and so the $\mathbf{x}$'s are samples from p.

1.1 Markov Chain Monte Carlo (MCMC)

Recall that a Markov Process consists of a state space $\mathcal{S}$ and for each $\mathbf{s} \in \mathcal{S}$ a transition probability

$$T(\mathbf{s}, \widetilde{\mathbf{s}})$$

which is the probability of going from state $\mathbf{s}$ at time t to state $\widetilde{\mathbf{s}}$ at time $t+1$. So for each $\mathbf{s}$, the function $\widetilde{\mathbf{s}} \to T(\mathbf{s}, \widetilde{\mathbf{s}})$ is a distribution over the state space $\mathcal{S}$.

A Markov Chain is generated by starting at a state $\mathbf{s}_0$ and then sample $\mathbf{s}_1$ from the distribution $T(\mathbf{s}_0, \widetilde{\mathbf{s}})$. Then sampling a $\mathbf{s}_2$ from $T(\mathbf{s}_1, \widetilde{\mathbf{s}})$ and in general, sample $\mathbf{s}_{t+1}$ from $T(\mathbf{s}_t, \widetilde{\mathbf{s}})$.

Assume we are given a distribution p_0 on $\mathcal{S}$. Then we get another distribution p_1 by

$$p_1(\widetilde{\mathbf{s}}) = \int T(\mathbf{s}, \widetilde{\mathbf{s}}) p_0(\mathbf{s}) d\mathbf{s}$$

and in general

$$p_{t+1}(\widetilde{\mathbf{s}}) = \int T(\mathbf{s}, \widetilde{\mathbf{s}}) p_t(\mathbf{s}) d\mathbf{s}$$

In case $\mathcal{S}$ is a finite set the integral is replaced by a sum.

A distribution q on $\mathcal{S}$ is said to be a limiting or stable distribution if

$$q(\widetilde{\mathbf{s}}) = \int T(\mathbf{s}, \widetilde{\mathbf{s}}) q(\mathbf{s}) d\mathbf{s}$$

Theorem Assume that for any $\mathbf{s}, \widetilde{\mathbf{s}} \in \mathcal{S}$ there exists a chain

$$\mathbf{s} = \mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_u = \widetilde{\mathbf{s}}$$

for some u depending on $\mathbf{s}, \widetilde{\mathbf{s}}$.
Assume further that the process is non-periodic.

Then starting with any distribution p_0 there exists a t such that $p_t = p_{t+1}$ so p_t is a stable distribution. Furthermore the stable distribution is unique and is independent of the starting distribution p_0

We shall only sketch the proof in the case $\mathcal{S}$ is finite.

To make the exposition simpler assume that $\mathcal{S} = \{1, 2, \ldots, n\}$ and let $a_{ij} = T(i, j)$ i.e. the probability of transitioning from state i to state j.

The matrix $\mathbf{A} = \{a_{ij}\}_{ij}$ is a *stochastic matrix* in the sense that all rows sum to 1:

$$\sum_j a_{ij} = \sum_j T(i, j) = 1 \text{ for all } i$$

A distribution on $\mathcal{S}$ is simply a vector $\mathbf{w}$ with non-negative entries such that $\sum_j w_j = 1$

If $\mathbf{w}_t$ is a distribution then $\mathbf{w}_{t+1} = \mathbf{w}_t \mathbf{A}$ and so $\mathbf{w}$ is a stable distribution when

$$\mathbf{w}\mathbf{A} = \mathbf{w}$$

The conditions mean that $\mathbf{A}$ is an *irreducible* non-negative *aperiodic* matrix and so we can use a theorem, the *Perron-Frobenius Theorem* to conclude that 1 is the unique maximal left and right eigenvalue of $\mathbf{A}$ and both the left and right eigenvectors $\mathbf{w}$ and $\mathbf{v}$ have non=negative entries and the eigenspaces are both 1-dimensional (since the rows sum to 1 the right eigenspace is spanned by $(1, 1, \ldots, 1)$) .

The left eigenvector $\pi = \mathbf{w}/\sum w_j$ is the stable distribution.

Furthermore, for any distribution $\mathbf{w_0}$

$$\mathbf{w_0}\mathbf{A}^k \to \pi$$

The idea of $MCMC$ is, for a given distribution, p, we want to sample from, construct a Markov Process with p as its stable distribution.

Generating a chain $\{s_t\}_t$, s_t will be samples from the stable distribution for $t >> 1$

The samples $s_t, s_{t+1}, s_{t+2}, \ldots$ are obviously not independent so one might want to take samples $s_t, s_{t+n}, s_{t+2n}, \ldots$ for some, sufficiently large n, to make the samples have less correlation.

There are several ways to make Markov Processes with a specific stable distribution but they are all more or less based on the *Metropolis-Hastings* method.

We say that the MC is *mixing* at t if we have reached the stable distribution i.e. if $p_t = p_{t+1}$.
Unfortunately the existence theorem does not specify the index t where the process is mixing.

If π is a distribution such that

$$T(s, s')\pi(s') = \pi(s)T(s', s)$$

then π is the stable distribution.

Indeed

$$\int T(s', s)\pi(s)ds = \int T(s', s)\pi(s')ds = \pi(s') \int T(s', s)ds = \pi(s')$$

We want to generate samples from a distribution p of the form $\frac{g}{Z}$ where g is again a positive function $g : \mathbb{R}^d \to \mathbb{R}$ and $Z = \int_{\mathbb{R}^d} g(\mathbf{x})d\mathbf{x}$

Start with a random point $\widetilde{\mathbf{s}}$ and choose a *proposal distribution*

$$Q(\mathbf{s}, \widetilde{\mathbf{s}})$$

which is symmetric in the sense $Q(\mathbf{s}, \widetilde{\mathbf{s}}) = Q(\widetilde{\mathbf{s}}, \mathbf{s})$, for instance a Gaussian $\mathcal{N}(\mathbf{s}|\widetilde{\mathbf{s}}, \sigma^2 I)$

Next let

$$A(\widetilde{\mathbf{s}}, \mathbf{s})$$

be an acceptance probability i.e.$A(\widetilde{\mathbf{s}}, \mathbf{s})$ is the probability that a proposed sample from the proposal distribution is accepted

We define a MC with transition probabilities

$$T(\widetilde{\mathbf{s}}, \mathbf{s}) = \begin{cases} Q(\widetilde{\mathbf{s}}, \mathbf{s}) A(\widetilde{\mathbf{s}}, \mathbf{s}) \text{ if } \mathbf{s} \neq \widetilde{\mathbf{s}} \\ Q(\widetilde{\mathbf{s}}, \widetilde{\mathbf{s}}) + \int Q(\widetilde{\mathbf{s}}, \mathbf{s})(1 - A(\widetilde{\mathbf{s}}, \mathbf{s})) d\mathbf{s} \end{cases}$$

We show that this is indeed a MC in the finite case. So we need to show

$$\sum_{\mathbf{s} \in \mathcal{S}} T(\mathbf{s}, \widetilde{\mathbf{s}}) = 1$$

Write the sum as

$$\begin{aligned} &\sum_{\mathbf{s} \in \mathcal{S}/\{\widetilde{\mathbf{s}}\}} Q(\mathbf{s}, \widetilde{\mathbf{s}}) A(\mathbf{s}, \widetilde{\mathbf{s}}) + T(\widetilde{\mathbf{s}}, \widetilde{\mathbf{s}}) \\ &= \sum_{\mathbf{s} \in \mathcal{S}/\{\widetilde{\mathbf{s}}\}} Q(\mathbf{s}, \widetilde{\mathbf{s}}) A(s, \widetilde{\mathbf{s}}) + Q(\widetilde{\mathbf{s}}, \widetilde{\mathbf{s}}) + \sum_{\mathbf{s} \in \mathcal{S}/\{\widetilde{\mathbf{s}}\}} Q(\widetilde{\mathbf{s}}, \mathbf{s})(1 - A(\widetilde{\mathbf{s}}, \mathbf{s})) \\ &= \sum_{\mathbf{s} \in \mathcal{S}/\{\widetilde{\mathbf{s}}\}} Q(\mathbf{s}, \widetilde{\mathbf{s}}) + Q(\widetilde{\mathbf{s}}, \widetilde{\mathbf{s}}) = 1 \end{aligned}$$

We define

$$A(\widetilde{\mathbf{x}}, \mathbf{x}) = \min\left(1, \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}\right) = \min\left(1, \frac{g(\mathbf{x})}{g(\widetilde{\mathbf{x}})}\right)$$

Remark how the troublesome normalization factor drops out in the fraction.

Starting with any distribution p_0,let

$$p_1(\mathbf{x}) = \int T(\widetilde{\mathbf{x}}, \mathbf{x}) p_0(\widetilde{\mathbf{x}}) d\widetilde{x}$$

and continuing

$$p_{t+1}(\mathbf{x}) = \int T(\widetilde{\mathbf{x}}, \mathbf{x}) p_t(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}}$$

we get distributions p_t for very t.

The stable distribution is reached when $p_{t+1} = p_t$

We shall show that p is the stable distribution of the Markov process we just constructed (remark that this does not depend on the choice of the proposal distribution Q as long as it is symmetric)

Assume first $\widetilde{\mathbf{x}} \neq \mathbf{x}$ so the proposed sample $\mathbf{x}$ is accepted.

Then

$$\frac{T(\widetilde{\mathbf{x}}, \mathbf{x})}{T(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{Q(\widetilde{\mathbf{x}}, \mathbf{x})A(\widetilde{\mathbf{x}}, \mathbf{x})}{Q(\mathbf{x}, \widetilde{\mathbf{x}})A(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{A(\widetilde{\mathbf{x}}, \mathbf{x})}{A(\mathbf{x}, \widetilde{\mathbf{x}})}$$

Assume $p(\widetilde{\mathbf{x}}) > p(\mathbf{x})$ so $\dfrac{p(\widetilde{\mathbf{x}})}{p(\mathbf{x})} > 1$ and $\dfrac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})} < 1$.

Hence

$$A(\widetilde{\mathbf{x}}, \mathbf{x}) = \min\left(1, \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}\right) = \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}$$

and

$$A(\mathbf{x}, \widetilde{\mathbf{x}}) = \min\left(1, \frac{p(\widetilde{\mathbf{x}})}{p(\mathbf{x})}\right) = 1$$

so

$$\frac{T(\widetilde{\mathbf{x}}, \mathbf{x})}{T(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{A(\widetilde{\mathbf{x}}, \mathbf{x})}{A(\mathbf{x}, \widetilde{\mathbf{x}})} = A(\widetilde{\mathbf{x}}, \mathbf{x}) = \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}$$

If $p(\widetilde{\mathbf{x}}) < p(\mathbf{x})$ then

$$A(\widehat{\mathbf{x}}, \mathbf{x}) = \min(1, \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}) = 1 \text{ and } A(\mathbf{x}, \widehat{\mathbf{x}}) = \frac{p(\widehat{\mathbf{x}})}{p(\mathbf{x})}$$

so

$$\frac{T(\widetilde{\mathbf{x}}, \mathbf{x})}{T(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{A(\widetilde{\mathbf{x}}, \mathbf{x})}{A(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{1}{A(\mathbf{x}, \widetilde{\mathbf{x}})} = \frac{p(\mathbf{x})}{p(\widetilde{\mathbf{x}})}$$

In either case

$$T(\widetilde{\mathbf{x}}, \mathbf{x})p(\widetilde{\mathbf{x}}) = p(\mathbf{x})T(\mathbf{x}, \widetilde{\mathbf{x}})$$

If the proposed $\mathbf{x}$ is rejected, the next sample is just $\widehat{\mathbf{x}}$ itself and there is nothing to show.

The equality

$$T(\widetilde{\mathbf{x}}, \mathbf{x})p(\widetilde{\mathbf{x}}) = p(\mathbf{x})T(\mathbf{x}, \widetilde{\mathbf{x}})$$

shows that p is the stable distribution of the Markov process.

The Metropolis-Hastings algorithm is generally very slow and a faster method to generate samples is the *Hamiltonian Monte Carlo*.

This is based on Hamiltonian Physics:

If a particle with mass= m is moving in a potential field $V(q)$, where q is the position of the particle, with momentum = p, the total energy of the particle is the potential energy $V(q)$, plus the kinetic energy $T = ||p||^2/2m$. For simplicity we take $m = 1$.

The Hamiltonian is the total energy $H(p, q) = V(q)+T(p)$ and the equations of motion in the coordinates of the phase space (p, q), are Hamilton's equations

$$\frac{dq}{dt} = \frac{\partial H}{\partial p} = \frac{\partial T}{\partial p} \text{ and } \frac{dp}{dt} = -\frac{\partial H}{\partial q} = -\frac{\partial V}{\partial q}$$

If the particle has coordinates (p, q) at time t and Δt is a small time increment then at time $t + \Delta t$ the coordinates in phase space are approximately

$$(p + \Delta t \frac{dp}{dt}, q + \Delta t \frac{dq}{dt}) = (p - \Delta t \frac{\partial V}{\partial q}, q + \Delta t \frac{\partial T}{\partial p})$$

In general we cannot solve the Hamiltonian equations to find the paths of motion so we must resort to numerical integration methods

Now we shall apply this formalism to MCMC.

Assume we have a dataset $\mathcal{D}$ and we hypthesize that the data are samples from a distribution $p(\mathbf{x}|\theta)$ with an unknown parameter vector θ.

We want to create samples from a posterior distribution

$$p_{post}(\theta|\mathcal{D})$$

to allow us to approximately compute things like Maximum Apostiori Probability (MAP) which is $\theta_0 = \underset{\theta}{\text{argmax}}\, p(\theta|\mathcal{D})$, mean and variance. Choosing an

appropriate prior $p_{prior}(\theta)$, we can use Bayes' formula

$$p_{post}(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)\, p_{prior}(\theta)}{p(\mathcal{D})}$$

Where the likelihood

$$p(\mathcal{D}|\theta) = \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x}|\theta)$$

As usual the denominator

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta) p_{prior}(\theta)\, d\theta$$

is intractable so we can only compute the posterior up to the normalization factor

$$p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)\, p_{prior}(\theta)$$

Now θ plays the role of the position q and the potential energy is defined to be

$$V(\theta) = -\log p(\mathcal{D}|\theta)\, p_{prior}(\theta) = -\sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x}|\theta) - \log p_{prior}(\theta)$$

Momentum is a vector r of the same dimension as θ and the Hamiltonian is

$$H(r,\theta) = -\log p(\mathcal{D}|\theta)\, p_{prior}(\theta) + ||r||^2/2$$

Now we lift the posterior distribution $p(\theta|\mathcal{D})$ to a distribution on the phase space by

$$p(r,\theta) = p(r|\theta)p(\theta|\mathcal{D})$$

and

$$p(r|\theta) = \mathcal{N}(0, I)$$

If we have samples from this distribution $\{(r,\theta)\}$ then the θ coordinates are samples from the marginal distribution

$$p(\theta) = \int p(r,\theta)dr = \int \mathcal{N}(r;0,I)p_{post}(\theta|\mathcal{D})dr = p_{post}(\theta|\mathcal{D}) \int \mathcal{N}(r;0,I)dr = p_{post}(\theta|\mathcal{D})$$

We shall construct a Markov sequence on the phase space with limit distribution $p(r,\theta)$.

So we need a proposal distribution and an acceptance distribution.

We set an interval length L and a step-size Δ.

Assume we have constructed (r_{m-1}, θ_{m-1}), and put $\theta^0 = \theta_{m-1}$. We want to construct the next step. The idea is to view this as a particle at position θ^0 at time t. At time t we give the system a push by generating a momentum r^0 as a random sample from $\mathcal{N}(0, I)$ and then compute the path in the phase space using the Hamiltonian equations of motion with starting state (r^0, θ^0) , and compute the (deterministic) orbit of the particle to time $t + L$.

In general we cannot integrate the Hamiltonian equations analytically so we have to use numerical approximations.

We shall use the following *LeapFrog* algorithm.

We divide L into n small subintervals of length Δ, and let $n = \dfrac{L}{\Delta}$ be the number of these subintervals

1. Let $r^0 \sim \mathcal{N}(0, I)$ and $\theta^0 = \theta_{m-1}$
2. For $i = 0, 1, \ldots, n-1$. If (r^i, θ^i) has been computed, set

$$\widetilde{r} = r^i - \frac{\Delta}{2} \nabla_\theta V(\theta^i)$$

3. Set $\theta^{i+1} = \theta^i + \Delta\, \widetilde{r}$
4. Set $r^{i+1} = \widetilde{r} + \frac{\Delta}{2} \nabla_\theta V(\theta^{i+1})$
5. $(r_m, \theta_m) = (-r^n, \theta^n)$

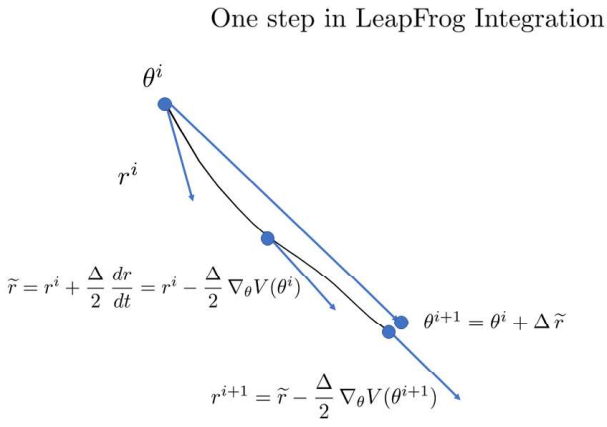

Figure 2:

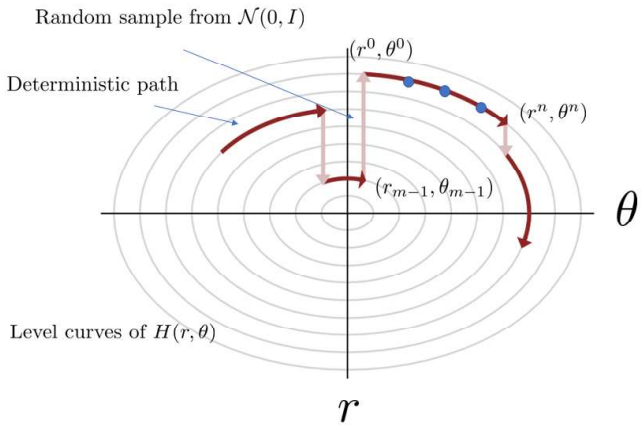

Figure 3:

The reason for negative sign is to make the proposal distribution symmetric. Indeed the construction of $(r^0, \theta^0) \to (r_m, \theta_m))$ is deterministic so

$$p(r', \theta' | r^0, \theta^0) = \delta(r' - r_m)\delta(\theta' - \theta_m) = \begin{cases} 1 \text{ if } (r', \theta') = (-r^n, \theta^n) \\ 0 \text{ otherwise} \end{cases}$$

We have

$$p(r_m, q_m | r^0, q^0) = 1$$

But it is clear that the LeapFrog is reversible, so if we start with $(-r^n, \theta^n)$ we end up with $(-r^0, \theta^0)$ and so going backwards in the process we get (r^0, θ^0) (remark that we have to change sign on r^0). Hence

$$p(r^0, \theta^0 | -r^n, \theta^n) = \delta(r^0 - r^0)\delta(\theta^0 - \theta^0) = 1$$

.

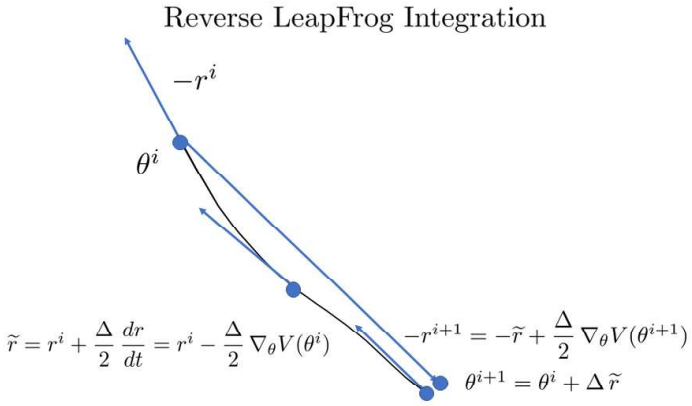

Figure 4:

The transition distribution is

$$p(r', \theta'|r^0, \theta^0)A(r', \theta'|r^0, \theta^0)$$

where the acceptance probability $A(r', \theta'|r^0, \theta^0)$ is

$$\begin{aligned}
A(r_m, \theta_m|r^0, \theta^0) &= \min\left(1, \frac{p((r_m, \theta_m)|(r^0, \theta^0))p(r_m, \theta_m)}{p(r^0, \theta^0|(r_m, \theta_m))p(r^0, \theta^0)}\right) \\
&= \min\left(1, \frac{p(r_m, \theta_m)}{p(r^0, \theta^0)}\right) \\
&= \min\left(1, \frac{\mathcal{N}(r_m; 0, I)p(\theta_m|\mathcal{D})}{\mathcal{N}(r^0; 0, I)p(\theta^0|\mathcal{D})}\right) \\
&= \min\left(1, \frac{\mathcal{N}(r_m; 0, I)\frac{p(\mathcal{D}|\theta_m)p_{prior})(\theta_m)}{p(\mathcal{D})}}{\mathcal{N}(r^0; 0, I)\frac{p(\mathcal{D}|\theta^0)p_{prior}(\theta^0)}{p(\mathcal{D})}}\right) \\
&= \min\left(1, \frac{\mathcal{N}(r_m; 0, I)p(\mathcal{D}|\theta_m)p_{prior})(\theta_m)}{\mathcal{N}(r^0; 0, I)p(\mathcal{D}|\theta^0)p_{prior}(\theta^0)}\right)
\end{aligned}$$

Now

$$V(\theta) = -\log p(\mathcal{D}|\theta)p_{prior}(\theta))$$

so

$$p(\mathcal{D}|\theta)p_{prior}(\theta)) = \exp(-V(\theta))$$

$$\mathcal{N}(r;0,I) = const.\exp(-\frac{r\cdot r^T}{2}) = const.\exp(-\frac{||r||^2}{2})$$

It follows that we can write

$$\begin{aligned}\mathcal{N}(r;0,I)p(\mathcal{D}|\theta)p_{prior}(\theta) &= const.\exp(-V(\theta) - \frac{||r||^2}{2}) \\ &= const.\exp(-H(r,\theta))\end{aligned}$$

and finally

$$A((r_m,\theta_m)|(r_0,\theta_0)) = \min\left(1, \frac{\exp(-H(r_m,\theta_m))}{\exp(-H(r^0,\theta^0))}\right)$$

We sample an α from $Uniform([0,1])$.

If $A((r_m,\theta_m)|(r^0,\theta^0)) \geq \alpha$ we accept the new point (r_m,θ_m) and if $A((r_m,\theta_m)|(r^0,\theta^0)) < \alpha$ we reject it and set $(r_m,\theta_m) = (r_{m-1},\theta_{m-1})$

From our discussion of the Metropolis-Hastings theory it follows that this defines a Markov Chain with limiting distribution = posterior, $p_{post}(\theta|\mathcal{D})$

Example Assume we have a *Gaussian Mixture Model (GMM)* with three components with unknown parameters consisting of means μ_1,μ_2,μ_3, standard deviations $\sigma_1,\sigma_2,\sigma_3$ and weights $\pi(1),\pi(2),\pi(3)$. We let θ denote the set of parameters $\theta = (\mu_1,\mu_2,\mu_3,\sigma_1,\sigma_2,\sigma_3,\pi)$

We are given a set of samples $\mathcal{D} = \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N\}$ from the GMM. The density function is of the form

$$p(\mathcal{D}|\theta) = \prod_{\mathbf{x}\in\mathcal{D}} \pi(1)\,\mathcal{N}(\mathbf{x};\mu_1,\sigma_1^2) + \pi(2)\,\mathcal{N}(\mathbf{x};\mu_1,\sigma_2^2) + \pi(3)\,\mathcal{N}(\mathbf{x};\mu_3,\sigma_3^2)$$

We sample from this distribution by first sample an $i = 1,2,3$ from the distribution π and then sampling from the Gaussian $\mathcal{N}(\mu_i,\sigma_i^2)$

We want to construct samples from the posterior distribution

$$p_{post}(\theta|\mathcal{D})$$

Choosing a prior $p_{prior}(\theta)$ the posterior is

$$p_{post}(\theta|\mathcal{D}) = \frac{\prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x}|\theta)}{p(\mathcal{D})} p_{prior}(\theta)$$

As usual the denominator

$$p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$$

is intractable so we cannot compute the posterior analytically but we can use HMC to generate samples from $p_{post}(\theta|\mathcal{D})$

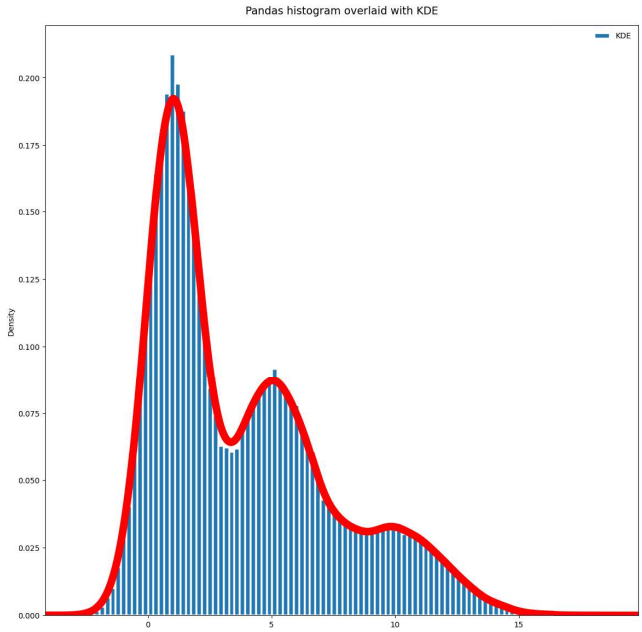

Figure 5:

We use Pyro, which can generate HMC samples from a model without having to write a lot of code.

Example

We shall do another example where we can use either MCMC or Variational Inference

We consider a Neural Network. Basically it is just a parametrized function that takes an input vector $\mathbf{x}$ and produces an output $\mathbf{y}$

$$\mathbf{y} = f(\mathbf{x}, \theta)$$

where θ denotes all the weights and biases of the NN.

A *Bayesian Neural Network* is a Neural Network where the parameters θ are random variables with some joint *distribution*. Thus for each input vector, $\mathbf{x}$, $f(\mathbf{x}, \theta)$ is a distribution over the possible outputs.

Given a training set $\mathcal{D}$ and an input $\mathbf{x}$ we have the conditional distribution of the output

$$p(\mathbf{y}|\mathbf{x},\mathcal{D}) = \int p(\mathbf{y}|\mathbf{x},\theta)p(\theta|\mathcal{D})d\theta = \mathbb{E}_{p(\theta|\mathcal{D})}(p(\mathbf{y}|\mathbf{x},\theta))$$

Training the network then becomes the problem of determining the posterior distribution

$$p(\theta|\mathcal{D})$$

We shall code a very simple *de-noising* model.

We are given a noisy set of points from some curve in $\mathbb{R}^2$

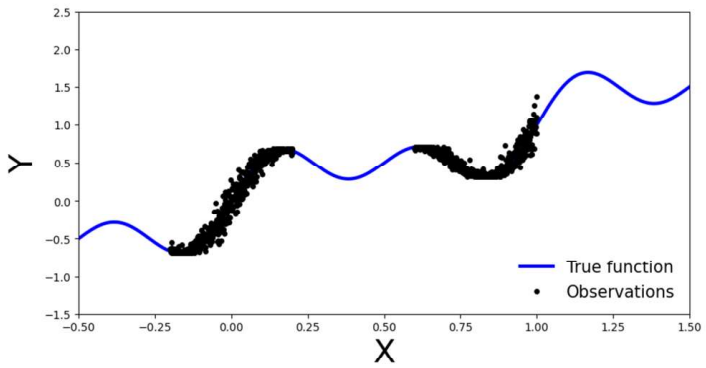

Figure 6:

To estimate $p(\mathbf{y}|\mathbf{x},\mathcal{D})$ which is now expressed as an expectation we can do a Monte Carlo estimation by generating samples from $p(\theta|\mathcal{D})$ and for a given $\mathbf{x}$ average $p(\mathbf{y}|\mathbf{x},\theta)$

We can now use pyro's `MCMC` to generate samples from the posterior distribution. Remark that the parameters of the trained Bayesian NN are samples of distributions so constructiong samples form the posterior $p(\theta|\mathcal{D})$ is in fact training the BNN, without using gradient descent.

```
from pyro.infer import MCMC, NUTS

model = BNN()

pyro.set_rng_seed(42)

nuts_kernel = NUTS(model)

mcmc = MCMC(nuts_kernel,num_samples=50)

x_train = torch.from_numpy(x_obs).float()
y_train = torch.from_numpy(y_obs).float()

mcmc.run(x_train,y_train)
```

Figure 7:

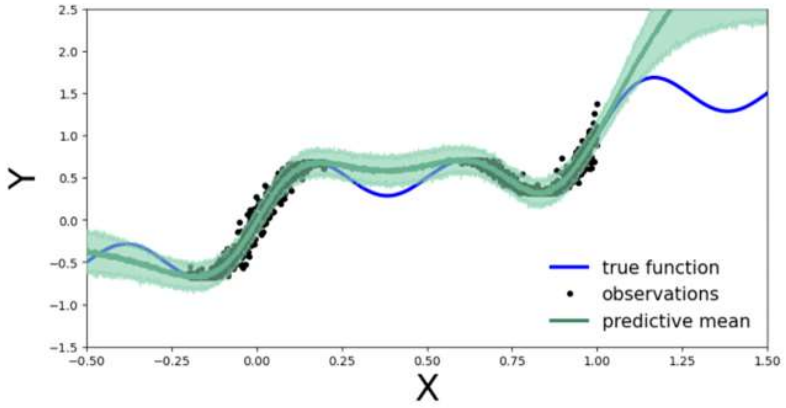

Figure 8:

The graph shows the mean predictions of the y-coordinates with a two *std* interval.

Instead of using $MCMC$ we can use Variational Inference i.e. we fit a distribution from a (simpler) parametrized family, to the posterior distribution.

We will use a *Mean Field Variational Distribution*, this is just a product of independent Gaussian distributions.

Instead of coding the Variational distribution ourselves we can use pyro's `pyro.infer.autoguide`

```
from pyro.infer import SVI, Trace_ELBO
from pyro.infer.autoguide import AutoDiagonalNormal
from tqdm import trange
pyro.clear_param_store()
```

Figure 9:

We can then use gradient descent to optimize the $ELBO$ to minimize the KL-divergence between the variational distribution and the intractable posterior.

The rest of the code is exactly the same as for the $MCMC$ method, use the `Prediction` object to generate samples from the variational distribution and plot the samples

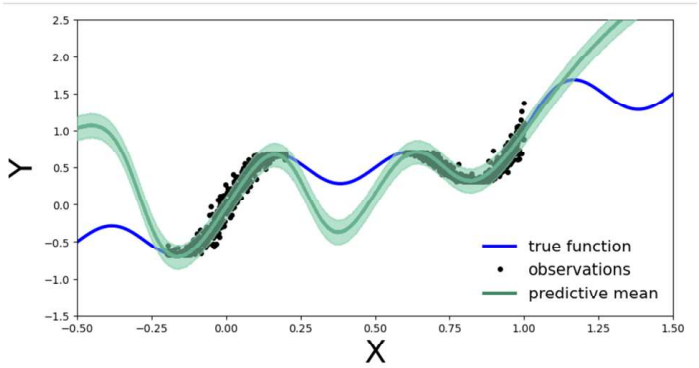

Figure 10:

Notice that the standard deviations band is much narrower.