Particle-based exploration of potentials

Anton Lebedev

July 26, 2021

0.1 NUTS - 1 particle

Given an ensemble of m non-interacting particles in an external potential the probability of *one* particular configuration state (q_i, p_i) is provided via the microcanonical ensemble:

$$\mathbb{P}[\{(q_i, p_i)\}] = \frac{e^{-\beta H(q_i,p_i)}}{\sum_{j=1}^{m} e^{-\beta H(q_j,p_j)}} \ . \tag{1}$$

Steps to determine one particle'S path:

1. Hamiltonian Monte Carlo.
2. ODE symplectic integrator (LF,SV, others.)
3. Step-size adaptation (after a trajectory in phase space has been determined).
4. Estimation of the initial step size.

0.1.1 Hamiltonian Monte Carlo

Given M particles at positions $\{\vec{x}_i\}_{i=1}^{M}$:

1. Draw momenta/velocities from a thermal distribution

$$\vec{p}_i \sim \mathcal{N}(0, \Sigma) \quad \forall i \in \{1, \ldots, M\} \ . \tag{2}$$

2. Propagate the ensemble in time for L steps

$$\forall i \quad (\vec{x}_i^{new}, \vec{p}_i^{new}) \leftarrow \text{Sympl.Int}(\{(\vec{x}_j, \vec{p}_j)\}_{j=1}^{M}, \Delta t) \ . \tag{3}$$

3. The probability (in thermal equilibrium) of a particle being in a state $(\vec{x}, \vec{p})$ is

$$\mathbb{P}[(\vec{x}, \vec{p})] = \frac{e^{-\beta H(\vec{x},\vec{p})}}{\mathcal{Z}} \ , \tag{4}$$

where $\mathcal{Z}$ is the partition function. Hence we accept the move if:

(a) $E_i^{new} = H(\vec{x}_i^{new}, \vec{p}_i^{new}) < H(\vec{x}_i, \vec{p}_i)$ or

(b) $\alpha \sim U[0,1] \qquad \wedge \qquad \alpha < \frac{\mathbb{P}[(\vec{x}_i^{new}, \vec{p}_i^{new})]}{\mathbb{P}[(\vec{x}_i, \vec{p}_i)]}$

If the move is accepted set

$$\vec{x}_i \leftarrow \vec{x}_i^{new}$$
$$\vec{p}_i \leftarrow -1\vec{p}_i^{new} ,$$

where the question remains as to *why* the direction of the momentum is introduced in the last line. According to Hoffman et. al. the reversal of the momentum is necessary to maintain time-reversibility, but *can* be omitted if one is only interested in sampling $p(\theta)$.

4. Set acceptance to $e^{-\beta(H(\vec{x}_i^{new},\vec{p}_i^{new})-H(\vec{x}_i,\vec{p}_i))}$.

0.1.2 NUTS

The goal of NUTS is to eliminate the need to specify the number of time steps manually and hence the criterion for terminating the propagation in step is: Stop when the particle's path (in real space) turns on itself:

$$(\vec{x}_i^{new} - \vec{x}_i) \cdot \vec{p}_i < 0 . \tag{5}$$

Extra (not yet fully understood): Slice sampling $u \sim U[0, \frac{e^{-\beta H(\vec{x}_i,\vec{p}_i)}}{\mathcal{Z}}]$ augments $\mathbb{P}[(\theta, r); u] \propto \chi_{0,E(\theta,r)}(u)$ with $(?)E(\theta, r) = e^{-H(\theta,r)}$. Integration over u yields $p(\theta, r)$. The parameter u appears to yield, via $\ln(u)$, an energy bound such that $-\ln(u) \in [H(\vec{x}, \vec{p}), \infty)$. Specifically sampling $-\ln u$ [in which way?] will provide an 'upper bound' on the valid range of energies the particle could attain (in the thermal environment), giving us a valid energy shell $[H(\vec{x}, \vec{p}), -\ln(u)]$, which can be related to the energy shells of the microcanonical description.

Integration of the path by recursive doubling and propagation forward and backward in time[1] and terminate if outermost states of the tree start to turn(5):

$$(\vec{x}_i^{max} - \vec{x}_i^{min}) \cdot \vec{p}_i^{max} \leq 0$$
$$\text{or } (\vec{x}_i^{max} - \vec{x}_i^{min}) \cdot \vec{p}_i^{min} \leq 0 .$$

Additionally check if $(E^{new} - E_{max}) > \Delta E$ i.e., out of the energy slice. [???]. Then select a set of new positions from the trajectory if they fulfill the energy condition. Selection with uniform probability.

The way NUTS is presented in the Hoffman paper it consists of two separate parts:

[1]Actually fwd/bwd on a trajectory, time reversal not necessary, momentum reversal would suffice.

- NUTS - avoids choosing the number of steps for HMC.
- 'Dual averaging scheme' to determine the step-size for HMC and hence for NUTS.

In accordance to the paper the algorithm propagates a particle symmetrically forward and backward in time until the the trajectory from the *most distant past* to *most distant future* states starts to curve back on itself. Once this (or a maximum number of states) has been achieved the next state is selected from *all* points of the trajectory in a way as to preserve detailed balance.

The (naive) algorithm will select the valid states from the set of all visited states by applying the constraints of the energy slice and those of detailed balance. After this subset has been selected a new proposed state will be selected from this restricted set with uniform probability. In this version the *all* candidate states must be stored, which is unsustainable w.r.t. memory consumption.

An improved version (Alg. 3) will select the states on the fly when building the sub-trees with weights given by the number of states in that subtree.

0.2 Models used for testing

0.2.1 Linear Model with Normal Noise

Consider the position of a particle moving along the vertical direction z without air resistance. Its position is given by

$$z(t) = z_0 + v_0(t - t_0) + \frac{a}{2}(t - t_0)^2 \,. \tag{6}$$

In general we assume that the observed positions $Z_i \sim \mathcal{N}(z_i, \sigma_0^2)$ are normally distributed around their true values, i.e., our observations are noisy.

"Experiment" set-up

Consider an arrow being fired directly upward on a building of height h. A GPS/altitude tracker is affixed to it. The tracker has a standard deviation/uncertainty of $3[m]$. And we assume that the internal clock has a vanishing spread. Altitudes are logged every $10[ms]$.

Query: Given the observed positions determine h, v_0, a, i.e., the altitude at which the arrow has been fired, its initial velocity and the gravitational acceleration.

Specific set-up For testing the following parameters were used:

$$\begin{aligned} g &= 9.80665[m/s^2] \\ v_0 &= 85[m/s] \\ h &= 100[m] \\ dt &= \frac{1}{100}[s] \\ \sigma_{obs} &= 0.01[m] \end{aligned}$$

Classical Approach

This problem is fairly easy to solve using linear least-squares fitting, e.g., by utilising normal equations $A^t A\vec{x} = A^t\vec{b}$, where $\vec{x}$ is the parameter vector and $\vec{b}$ the vector containing observed positions.

Stochastic Approach

In this case we use the following model

$$Z_i = Z_0 + V_0(t - t_0) + A(t - t_0)^2 + W_i \tag{7}$$

where

$$\begin{aligned} W_i &\sim \mathcal{N}(0, \sigma_{obs}^2) \\ Z_0 &\sim \mathcal{N}(h, \sigma_{obs}^2) \\ V_0 &\sim \mathcal{N}(v_0, \sigma_{obs}^2) \\ A &\sim \mathcal{N}(a, \sigma_{obs}^2) \ , \end{aligned}$$

with $\sigma_{obs} = 3[m]$.

Rationale All measurements are performed with an observation uncertainty of $\sigma_{obs} = 3[m]$. Since times are assumed to be exact the variation in v, a will depend on σ_{obs}.

Adaptation Subsume the model variables into a vector

$$\vec{\Theta}^t = (Z_0, V_0, A) \qquad \vec{\theta}^t = (z_0, v_0, a) \ . \tag{8}$$

Inference

Given a set of observations $\{(t_i, z_i)\}_{i=1}^N$ we want to estimate Θ_i using

$$f_{\vec{\Theta}|\vec{Z}}(\vec{\theta}|\vec{z}) = \frac{f_{\vec{\Theta}}(\vec{\theta}) f_{\vec{Z}|\vec{\Theta}}(\vec{z}|\vec{\theta})}{f_{\vec{Z}}(\vec{z})} \ . \tag{9}$$

Assuming we were given z_0, v_0, a, then

$$Z_i = z_0 + v_0(t-t_0) + a(t-t_0)^2 + W_i \Longrightarrow Z_i \sim \mathcal{N}(z_0 + v_0(t-t_0) + a(t-t_0)^2, \sigma_{obs}^2) \ . \tag{10}$$

Hence

$$f_{Z_i|\vec{\Theta}}(z_i|\vec{\theta}) = c e^{-\frac{(z_i - (z_0 + v_0(t-t_0) + a(t-t_0)^2))^2}{2\sigma_{obs}^2}} \ , \tag{11}$$

wherefrom follows

$$\begin{aligned}
f_{\vec{\Theta}|\vec{Z}}(\vec{\theta}|\vec{z}) &= \frac{1}{f_{\vec{Z}}(\vec{z})} \prod_{j=0}^{2} f_{\Theta_j}(\theta_j) \prod_{i=1}^{N} f_{Z_i|\vec{\Theta}}(z_i|\vec{\theta}) \\
&= \frac{1}{f_{\vec{Z}}(\vec{z})} \prod_{j=0}^{2} c_j e^{-\frac{(\theta_j - \mu_j)^2}{2\sigma_j^2}} \prod_{i=1}^{N} \tilde{c}_i e^{-\frac{(z_i - \dots)^2}{2\sigma_{obs}^2}} \\
c_j &= \frac{1}{\sigma_j \sqrt{2\pi}} \\
\tilde{c}_i &= \frac{1}{\sigma_{obs}\sqrt{2\pi}} \\
\Rightarrow f_{\vec{\Theta}|\vec{Z}}(\vec{\theta}|\vec{z}) &= \frac{1}{f_{\vec{Z}}(\vec{z})} \prod_{j=0}^{2} \frac{1}{\sigma_j\sqrt{2\pi}} e^{-\frac{(\theta_j - \mu_j)^2}{2\sigma_j^2}} \prod_{i=1}^{N} \frac{1}{\sigma_{obs}\sqrt{2\pi}} e^{-\frac{(z_i - \dots)^2}{2\sigma_{obs}^2}} \ .
\end{aligned}$$

Hence for the logarithm we obtain

$$\begin{aligned}
\ln(f_{\vec{\Theta}|\vec{Z}}(\vec{\theta}|\vec{z})) &= \sum_{j=0}^{2} \Big(\ln(e^{-\dots}) - \ln(\sqrt{2\pi}\sigma_j) \Big) + \sum_{i=1}^{N} \Big(\ln(e^{\dots}) - \ln(\sqrt{2\pi}\sigma_{obs}) \Big) + \ln(f_Z) \\
&= \sum_{j=0}^{2} -\frac{(\theta_j - \mu_j)^2}{2\sigma_j^2} + \sum_{i=1}^{N} -\frac{(z_i - \theta_0 - \theta_1(t-t_0) - \theta_2(t-t_0)^2)^2}{2\sigma_{obs}^2} \\
&\quad - \sum_{j=0}^{2} \ln(\sqrt{2\pi}\sigma_j) - \sum_{i=1}^{N} \ln(\sqrt{2\pi}\sigma_{obs}) - \ln(f_z) \\
&=: g_\Theta(\vec{\theta}) \ .
\end{aligned}$$

Next we consider each component of a gradient of $g_\Theta(\vec{\theta})$ separately:

$$\partial_{\theta_0} g_\Theta(\vec{\theta}) = -\frac{1}{\sigma_0^2}(\theta_0 - \mu_0) + \sum_{i=1}^{N} \frac{z_i - \theta_0 - \theta_1(t-t_0) - \theta_2(t-t_0)^2}{\sigma_{obs}^2}$$

$$\partial_{\theta_1} g_\Theta(\vec{\theta}) = -\frac{1}{\sigma_1^2}(\theta_1 - \mu_1) + \sum_{i=1}^{N} \frac{1}{\sigma_{obs}^2}\left(z_i - \theta_0 - \theta_1(t-t_0) - \theta_2(t-t_0)^2\right)(t-t_0)$$

$$\partial_{\theta_2} g_\Theta(\vec{\theta}) = -\frac{1}{\sigma_2^2}(\theta_2 - \mu_2) + \sum_{i=1}^{N} \frac{1}{\sigma_{obs}^2}\left(z_i - \theta_0 - \theta_1(t-t_0) - \theta_2(t-t_0)^2\right)(t-t_0)^2$$

0.2.2 Inferring the bias of a coin

Consider the following experiment:

A coin is tossed N times and the outcome of each random experiment is recorded.

Given that 'heads' is observed K times during the N tosses, what is the bias (probability of obtaining 'heads') of the coin?

Forward derivation

Experiment: Toss a coin N times. Assumptions:

1. $\mathbb{P}[\{H\}] = p$
2. All tosses are independent.

Events of interest: $\{H_1, \ldots, H_k, \ldots\}$, i.e., any sequence containing *exactly* K 'heads'.

Since the coin tosses are independent the following holds

$$\mathbb{P}[\{HT\}] = \mathbb{P}[\{H\}]\mathbb{P}[\{T\}] = p(1-p)\ . \tag{12}$$

Hence the probability of obtaining any *one* particular sequence of K 'heads' out of N tosses is

$$\mathbb{P}[\{H_1, \ldots, H_k, T_{K+1}, \ldots, T_N\}] = p^K(1-p)^{N-K}\ . \tag{13}$$

And the probability of obtaining a K-head sequence is equal to (13) weighted by the number of such sequences:

$$\mathbb{P}\left[\{K heads\}\right] = \binom{N}{K} p^K(1-p)^{N-K}\ , \tag{14}$$

where $\binom{N}{K}$ specifies in how many ways we can select K heads from a sequence of N possible outcomes. In this experiment N, p are set as experimental resp. model parameters and K is the outcome of a 'measurement', resp. execution of the experiment. Hence K, since it is random, is the random variable of the experiment and counts the number of 'successes' (='heads') in N independent trials, each of which has a binary outcome. Thus

1. Mapping $T \to 0, H \to 1$ results in each coin being represented by a random variable distributed according to the Bernoulli distribution.

2. K is a binomial random variable, hence

$$p_K(k) = \binom{N}{k} p^k (1-p)^{N-k} \; . \tag{15}$$

Inference

We know that our observable K is discrete. By model design, given that N is fixed, K will depend on the bias $y \in [0, 1]$. Given the model we assume absolute ignorance of the possible bias, hence

$$Y \sim U(0, 1) \; . \tag{16}$$

Once we know y the probability of observing K heads is given by the PMF (14), here *conditioned* on knowing y:

$$p_{K|Y}(k|y) = \binom{N}{k} y^k (1-y)^{N-k} \; . \tag{17}$$

According to Bayes' rule:

$$p_{Y|K}(y|k) = \frac{p_{Y,K}(y,k)}{p_K(k)} = \frac{p_{K|Y}(k|Y) p_Y(y)}{p_K(k)} \tag{18}$$

where

$$p_K(k) = \int_y p_{K|Y}(k|y') p_Y(y') dy' \; . \tag{19}$$

Since $Y \sim U(0, 1)$ we get $p_Y(y) = \chi_{[0,1]}(y)$and we therefore obtain

$$\begin{aligned} p_K(k) &= \int_0^1 \binom{N}{k} y^k (1-y)^{N-k} dy = \binom{N}{k} \int_0^1 y^k (1-y)^{N-k} dy \\ &= \binom{N}{k} \frac{k!(N-k)!}{(k+N-k+1)!} = \frac{N!}{k!(N-k)!} \frac{k!(N-k)!}{(N+1)!} = \frac{1}{N+1} \; . \end{aligned}$$

Thus we have

$$p_{Y|K}(y|k) = (N+1)\binom{N}{k} y^k (1-y)^{N-k} \ . \tag{20}$$

Using this distribution we may consider the quantities necessary for Hamiltonian Monte Carlo:

$$f_Y(y) := -\ln(p_{Y|K}(y|k)) \ . \tag{21}$$

Note: Since $p : \mathbb{N} \to [0,1]$ we will have $\ln(p(y)) \in (-\infty, 0]$ and hence $f_Y(y) \in [0,\infty)$ with $f_Y(y) = 0$representing the *certain* event.

The explicit expression for f is:

$$\begin{aligned} f_Y(y) &= -1\left[\ln((N+1)\binom{N}{k}) + \ln(y^k) + \ln((1-y)^{N-k})\right] \\ &= -\ln\left((N+1)\binom{N}{k}\right) - k\ln(y) - (N-k)\ln(1-y) \ . \end{aligned}$$

The gradient of the potential function f is

$$\partial_y f_Y(y) = -\frac{k}{y} + \frac{N-k}{1-y} \ . \tag{22}$$

The maximum of $f_Y(y)$ is determined as

$$\begin{aligned} 0 = \partial_y f_Y(y) &= -\frac{k}{y} + \frac{N-k}{1-y} \\ \frac{N-k}{1-y} &= \frac{k}{y} \\ \Rightarrow 0 = (N-k)y - k(1-y) &= Ny - k \Rightarrow y = \frac{k}{N} \ . \end{aligned}$$

Hence the MAP estimator is

$$\hat{Y}_{MAP} := \frac{K}{N} \ . \tag{23}$$

Inference with Beta Distribution

Next we assume that

$$\begin{aligned} Y &\sim Beta(a,b) \qquad Y \in [0,1] \\ p_Y(y) &= \frac{1}{c} y^a (1-y)^b \ . \end{aligned}$$

First we need to determine the normalization constant $1/c$:

$$1 = \int_0^1 p_Y(y)dy = \frac{1}{c}\int_0^1 y^a(1-y)^b dy = \frac{1}{c}\frac{a!b!}{(a+b+1)!} \tag{24}$$

where we used Feynman's trick. Hence $\frac{1}{c} = \frac{(a+b+1)!}{a!b!}$ and therefore

$$p_Y(y) = \frac{(a+b+1)!}{a!b!}y^a(1-y)^b \ . \tag{25}$$

Then for the given case:

$$p_{Y|K}(y|k) = \frac{(a+b+1)!}{a!b!}y^a(1-y)^b\binom{N}{k}y^k(1-y)yN-k\frac{1}{p_K(k)} \tag{26}$$

$$= \frac{(a+b+1)!}{a!b!}\binom{N}{k}y^{k+a}(1-y)yN-k+b\frac{1}{p_K(k)} \tag{27}$$

$$p_K(k) = \frac{(a+b+1)!}{a!b!}\binom{N}{k}\int_0^1 y^{k+a}(1-y)^{N-k+b}dy \tag{28}$$

$$= \frac{(a+b+1)!}{a!b!}\binom{N}{k}\frac{(k+a)!(N-k+b)!}{(N+b+a+1)!} \tag{29}$$

$$\Rightarrow p_{Y|K}(y|k) = \frac{(N+b+a+1)!}{(N-k+b)!(k+a)!}y^{k+a}(1-y)^{N-k+b} \ . \tag{30}$$

Hence the potential function f will be

$$f_Y(y) = -\left[\ln\left(\frac{(N+b+a+1)!}{(N-k+b)!(N+a)!}\right) + (k+a)\ln(y) + (N-k+b)\ln(1-y)\right] \tag{31}$$

which yields the gradient

$$\partial_y f_Y(y) = -\frac{k+a}{y} + \frac{N-k+b}{1-y} \tag{32}$$

where we must note that the differences to (22) are minimal.