Langevin MCMC and its M-H acceptance probability

Adrià Garriga-Alonso

August 4, 2020

Abstract

MCMC via Langevin dynamics, in its usual form, has an acceptance probability of 0. In this document is an alternative discretisation of its stochastic differential equation (SDE) that is time-reversible, and thus has a non-zero acceptance probability

1 Introduction

The acceptance probability of an MCMC sampler that follows the Langevin stochastic differential equations (SDE) is always 1. However, to be able to simulate the SDE, we discretise the time steps, and in so introduce error. This error can be exactly corrected Särkkä and Solin, 2019 by adding a Metropolis-Hastings (M-H) step, that applies rejection sampling for the new state of the SDE. Additional error is introduced by the approximate floating-point arithmetic, but we will disregard it.

In Langevin MCMC, we have some parameters θ to be inferred (also named *position variables*), and their corresponding *momentum variables* m. The generalised Metropolis-Hastings acceptance probability Bussi and Parrinello, 2007, which negates the final momentum, is as follows. If a transition goes from $(\theta_0, m_0) \to (\theta_*, m_*)$, its acceptance probability is

$$\Pr_{\text{accept}}(\theta_*, m_*) = \min\left(1, \frac{\pi(\theta_*)T(\theta_0, -m_0|\theta_*, -m_*)}{\pi(\theta_0)T(\theta_*, m_*|\theta_0, m_0)}\right) \tag{1}$$

for transition probability density T and target distribution density π. If a sample (θ_*, m_*) is rejected, we have to set the next state of the chain to $(\theta_0, -m_0)$ (and not (θ_0, m_0) like ordinary Metropolis-Hastings).

It is important to note that the transition probability distribution T, and thus the acceptance probability, very much depend on the scheme we use to discretise time in the SDE. That is, the integrator.

2 Symplectic Euler integrator

The Symplectic Euler integrator used in Wenzel et al., 2020 is not time reversible in the limit of zero friction. This implies that the probability of transitioning $(\theta_*, -m_*) \to (\theta_0, m_0)$ is always zero, strictly speaking. This is because the value of θ_* is determined entirely by the previous parameters, θ_0, and its **contemporaneous** momentum m_*. The forward transition for the parameters is

$$\theta_* = \theta_0 + hM^{-1}m_*. \tag{2}$$

Now, consider what the backward transition would be. The momentum $-m_*$ evolves to $-m_0$, and the parameters θ_* evolve to θ_0. The transition, now applied backward, is

$$\theta_0 = \theta_* + hM^{-1}(-m_0), \tag{3}$$

which cannot be true if $m_0 \neq m_*$.

3 Related work

Mannella (2004) published a stochastic integrator for Langevin dynamics that reduces to the normal Störmer-Verlet integrator on the limit of $\gamma \to 0$. However, it is in position form, that is, it first updates the position, then the velocity, then the position again. Thus, it is not easily adapted to an MCMC algorithm. Grønbech-Jensen (2020) describes a form of Mannella's integrator which corresponds to velocity Verlet. However, neither of these schemes is suitable for MCMC in Langevin dynamics: they only use one Gaussian random variable per time step, and Langevin dynamics is not measure-preserving. This implies that the backwards transition also has probability 0 (Bussi and Parrinello, 2007).

Bussi and Parrinello (2007) also wrote an integration scheme that reduces to the velocity Störmer-Verlet integrator. Their explicit intention is to correct the error introduced by discretising the time-step, using the Metropolis-Hastings correction. They note that, since the dynamics of Langevin dynamics are not measure preserving, unlike those of Hamiltonian flow, the acceptance probability with discrete time steps cannot be simply the reduction in log-joint probability. Instead, the forward and backward probability of transition have to be explicitly calculated. Unfortunately, it can be shown (Leimkuhler and Matthews, 2013) that their integration scheme overestimates the temperature of the velocity in a simple 1D harmonic oscillator.

Sivak, Chodera, and Crooks (2014) solve this problem by scaling the gradient time step down by a factor of $0.95 \leq b \leq 1$, which we define in the next section. Their proposed OVRVO integrator (so named by the order in which various operators are applied) is the approach we employ in this document.

4 OVRVO integration in one dimension

We use the OVRVO integrator by Sivak, Chodera, and Crooks, 2014. It is symmetric under time reversal and employs two independent Gaussian RVs per time step, so its probability of acceptance is never zero and can be evaluated.

At the same time, if the exact momenta are not needed, the Gaussian RVs can be fused for an amortised computational cost of 1 random draw per time step. This changes the long-term temperature statistics, but in an easily correctable manner.

The Langevin dynamics differential equations are

$$dr(t) = v(t) \tag{4}$$

$$dv(t) = -\gamma v(t) + m^{-1} f(r(t)) + m^{-\frac{1}{2}} \sqrt{2\gamma T} dW \tag{5}$$

where dW is a Wiener process.

For discrete time step h, the corresponding OVRVO updates are

$$a \triangleq e^{-\gamma h} \tag{6}$$

$$b \triangleq \sqrt{\frac{2}{\gamma h} \tanh\left(\frac{\gamma h}{2}\right)} \tag{7}$$

$$v^{\left(n+\frac{1}{2}\right)} \triangleq \sqrt{a} v^{(n)} + \frac{bh}{2} m^{-1} f^{(n)} + m^{-\frac{1}{2}} \sqrt{(1-a)T} \epsilon^{n+\frac{1}{2}} \tag{8}$$

$$r^{\left(n+\frac{1}{2}\right)} \triangleq r^{(n)} + \frac{bh}{2} v^{\left(n+\frac{1}{2}\right)} \tag{9}$$

$$r^{(n+1)} \triangleq r^{\left(n+\frac{1}{2}\right)} + \frac{bh}{2} v^{\left(n+\frac{1}{2}\right)} \tag{10}$$

$$v^{(n+1)} \triangleq \sqrt{a} v^{\left(n+\frac{1}{2}\right)} + \frac{bh\sqrt{a}}{2} m^{-1} f^{(n+1)} + m^{-\frac{1}{2}} \sqrt{(1-a)T} \epsilon^{n+1} \tag{11}$$

We may also combine the two updates to v to give:

$$v^{\left(n+\frac{1}{2}\right)} = a v^{\left(n-\frac{1}{2}\right)} + \frac{bh(1+a)}{2} m^{-1} f^{(n)} + m^{-\frac{1}{2}} \sqrt{(1-a^2)T} \epsilon^{n+\frac{1}{2}} \tag{12}$$

And define an auxiliary quantity u:

$$u^{(n)} \triangleq v^{\left(n+\frac{1}{2}\right)} - \frac{bh}{2} m^{-1} f^{(n)} \tag{13}$$

$$= \sqrt{a} v^{(n)} + m^{-\frac{1}{2}} \sqrt{(1-a)T} \epsilon^{n+\frac{1}{2}} \tag{14}$$

4.1 Empirical temperature estimation

Langevin dynamics allows for estimation of the temperature using the position and velocity variables.

$$T = \mathbb{E}\left[r(t)f(t)\right] = \mathbb{E}\left[mv(t)^2\right] \tag{15}$$

for any T. We know that the velocity in particular is Gaussian-distributed (Bussi and Parrinello, 2007; Wenzel et al., 2020)

$$v(t) \sim \mathcal{N}\left(0, \frac{T}{m}\right) \tag{16}$$

In the OVRVO integrator, the value of r is more accurate at half time steps, and the value of v is more accurate at full time steps. For this reason, we defined the quantity $u^{(n)}$. Its distribution is

$$u^{(n)} \sim \mathcal{N}\left(0, a\frac{T}{m} + \frac{(1-a)T}{m}\right) = \mathcal{N}\left(0, \frac{T}{m}\right). \tag{17}$$

which means we can also check the temperature by empirically averaging $u^{(n)}$.

5 Multiple dimensions

We introduce an unnormalised posterior density $\pi(\boldsymbol{\theta})$ that we want to sample from. This induces a potential $U(\boldsymbol{\theta}) = -\log \pi(\boldsymbol{\theta})$. The new, vector-valued, variables are

$$\begin{aligned} \boldsymbol{\theta}(t) &\triangleq r(t) & \qquad \boldsymbol{q}(t) &\triangleq \boldsymbol{M}^{\frac{1}{2}} v(t) \\ \boldsymbol{M} &\triangleq m & \qquad \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}(t)) &\triangleq -f(t) \end{aligned}$$

At state $(\boldsymbol{\theta}, \boldsymbol{q})$, the energy of the system is

$$H(\boldsymbol{\theta}, \boldsymbol{q}) \triangleq U(\boldsymbol{\theta}) + \frac{1}{2}\boldsymbol{q}^{\mathsf{T}}\boldsymbol{q}, \tag{18}$$

and the canonical distribution of the dynamical system, for temperature T, is $p(\boldsymbol{\theta}, \boldsymbol{q}) \propto \exp(-\frac{1}{T}H(\boldsymbol{\theta}, \boldsymbol{q}))$.

The Langevin differential equations are now

$$d\boldsymbol{\theta} = \boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{q}dt \tag{19}$$

$$d\boldsymbol{q} = -\boldsymbol{M}^{-\frac{1}{2}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta})dt - \gamma\boldsymbol{q}dt + \sqrt{2\gamma T}d\boldsymbol{W}. \tag{20}$$

We also reparameterize the time step h and friction γ by introducing the momentum decay β and the learning rate ℓ. As listed by Wenzel et al. (2020), $h = \sqrt{\ell/n}$ and $\gamma = (1-\beta)\sqrt{n/\ell}$.

The reparameterised OVRVO scheme is

$$a \triangleq e^{\beta - 1} \tag{21}$$

$$b \triangleq \sqrt{\frac{2}{1-\beta}\tanh\left(\frac{1-\beta}{2}\right)} \tag{22}$$

$$\boldsymbol{q}^{\left(n+\frac{1}{2}\right)} \triangleq \sqrt{a}\boldsymbol{q}^{(n)} - \frac{bh}{2}\boldsymbol{M}^{-\frac{1}{2}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) + \sqrt{(1-a)T}\boldsymbol{\epsilon}^{\left(n+\frac{1}{2}\right)} \tag{23}$$

$$\boldsymbol{\theta}^{(n+1)} \triangleq \boldsymbol{\theta}^{(n)} + bh\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{q}^{\left(n+\frac{1}{2}\right)} \tag{24}$$

$$\boldsymbol{q}^{(n+1)} \triangleq \sqrt{a}\boldsymbol{q}^{\left(n+\frac{1}{2}\right)} - \frac{bh\sqrt{a}}{2}\boldsymbol{M}^{-\frac{1}{2}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n+1)}) + \sqrt{(1-a)T}\boldsymbol{\epsilon}^{(n+1)} \tag{25}$$

For computational convenience, we will use the parameters $\boldsymbol{\theta}$ only at whole time steps and the momenta $\boldsymbol{q}$ at half time steps, even though the integrator is slightly more accurate at half time steps for the parameters and whole time steps for the momenta in simple cases (Sivak, Chodera, and Crooks, 2014). A momentum update that uses only one Gaussian is

$$\boldsymbol{q}^{\left(n+\frac{1}{2}\right)} \triangleq a\boldsymbol{q}^{\left(n-\frac{1}{2}\right)} - \frac{bh(1+a)}{2}\boldsymbol{M}^{-\frac{1}{2}}\nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) + \sqrt{(1-a^2)T}\boldsymbol{\epsilon}^{\left(n+\frac{1}{2}\right)} \tag{26}$$

For a d-dimensional parameter and momentum, the configurational and kinetic temperatures are

$$\hat{T}_c^{(n)} = \frac{1}{d}\left\langle \boldsymbol{\theta}^{(n)}, \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) \right\rangle \qquad \hat{T}_k^{\left(n+\frac{1}{2}\right)} = \frac{1}{d}\left\langle \boldsymbol{q}^{\left(n+\frac{1}{2}\right)}, \boldsymbol{q}^{\left(n+\frac{1}{2}\right)} \right\rangle. \tag{27}$$

5.1 Metropolis-Hastings acceptance probability

Writing $x^{(n)} = \left(\boldsymbol{\theta}^{(n)}, \boldsymbol{q}^{(n)}\right)$, we associate with each point $x^{(n)}$ an importance weight $w^{(n)}$ (Bussi and Parrinello, 2007). The ratio of weights can be used in a Metropolis-Hastings acceptance probability $x^{(n)} \to x^{(n+1)}$, as

$$\Pr\left(\text{accept } x^{(n+1)} \,\middle|\, x^{(n)}\right) = \min\left(1, \frac{w^{(n+1)}}{w^{(n)}}\right). \tag{28}$$

We may express $w^{(n)}$ in terms of an effective energy $\tilde{H}^{(n)} = -T \log w^{(n)}$, which starts as $\tilde{H}^{(0)} = H(\boldsymbol{\theta}^{(0)}, \boldsymbol{q}^{(0)})$ and evolves as

$$\Delta\tilde{H} = \tilde{H}^{(n+1)} - \tilde{H}^{(n)} = -T \log \frac{M(x^{(n)} \,|\, x^{(n+1)})}{M(x^{(n+1)} \,|\, x^{(n)})} + H\left(x^{(n+1)}\right) - H\left(x^{(n)}\right). \tag{29}$$

This effective energy should fluctuate during the simulation, but not exhibit a systematic drift. If the momentum is resampled, it should be reset.

Solving for $\boldsymbol{\epsilon}^{\left(n+\frac{1}{2}\right)}, \boldsymbol{\epsilon}^{(n)}$ in equation 26, changing the variables in the probability density and calculating the ratio, we get (Bussi and Parrinello, 2007):

$$\begin{aligned} -T \log \frac{M(x^{(n)} \,|\, x^{(n+1)})}{M(x^{(n+1)} \,|\, x^{(n)})} = &-\frac{1}{2}\left\langle \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) + \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n+1)}), \boldsymbol{\theta}^{(n+1)} - \boldsymbol{\theta}^{(n)} \right\rangle \\ &+ \frac{h^2 b^2}{8}\left(\left\langle \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n+1)}), \boldsymbol{M}^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n+1)}) \right\rangle - \left\langle \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}), \boldsymbol{M}^{-1} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}^{(n)}) \right\rangle\right) \\ &- \frac{1}{2}\left\langle \boldsymbol{q}^{(n+1)}, \boldsymbol{q}^{(n+1)} \right\rangle + \frac{1}{2}\left\langle \boldsymbol{q}^{(n)}, \boldsymbol{q}^{(n)} \right\rangle. \end{aligned}$$

Together with equation 18, the result is

$$\Delta\tilde{H} = \Delta U - \frac{1}{2}\Delta\boldsymbol{\theta}^{\mathsf{T}}\left(\nabla U^{(n)} + \nabla U^{(n+1)}\right) + \frac{h^2 b^2}{8}\Delta\left(\nabla U^{\mathsf{T}} \boldsymbol{M}^{-1} \nabla U\right).$$

The change in effective energy does not depend on the momentum at all. Thus, we can avoid calculating the momentum at whole time steps in equation 26, and draw only one Gaussian RV per time step.

In the implementation, we substitute $\boldsymbol{\theta}^{(n+1)} - \boldsymbol{\theta}^{(n)} = bh\boldsymbol{M}^{-\frac{1}{2}}\boldsymbol{q}^{\left(n+\frac{1}{2}\right)}$ to save computation.

References

Bussi, Giovanni and Michele Parrinello (2007). "Accurate sampling using Langevin dynamics". In: *Phys. Rev. E* 75. DOI: `10.1103/PhysRevE.75.056707`.

Grønbech-Jensen, Niels (2020). "Complete set of stochastic Verlet-type thermostats for correct Langevin simulations". In: *Molecular Physics* 118.8, e1662506. DOI: `10.1080/00268976.2019.1662506`. eprint: `https://doi.org/10.1080/00268976.2019.1662506`. URL: `https://arxiv.org/abs/1909.04380`.

Leimkuhler, Benedict and Charles Matthews (2013). "Robust and efficient configurational molecular sampling via Langevin dynamics". In: *The Journal of Chemical Physics* 138.

Mannella, R. (2004). "Quasisymplectic integrators for stochastic differential equations". In: *Physical Review E* 69. DOI: `10.1103/PhysRevE.69.041107`.

Särkkä, Simo and Arno Solin (2019). *Applied Stochastic Differential Equations.* Institute of Mathematical Statistics Textbooks. Cambridge University Press.

Sivak, David A., John D. Chodera, and Gavin E. Crooks (2014). "Time Step Rescaling Recovers Continuous-Time Dynamical Properties for Discrete-Time Langevin Integration of Nonequilibrium Systems". In: *The Journal of Physical Chemistry B* 118. DOI: `10.1021/jp411770f`.

Wenzel, Florian, Kevin Roth, Bastiaan S. Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin (2020). "How Good Is the Bayes Posterior in Deep Neural Networks Really?" In: *CoRR*. arXiv: `2002.02405 [stat.ML]`. URL: `http://arxiv.org/abs/2002.02405v1`.