Pseudo-marginal Metropolis-Hastings: a simple explanation and (partial) review of theory

Chris Sherlock

Motivation

Imagine a stochastic process V which arises from some distribution with density $p(v|\theta_1)$.

Imagine noisy observations y of this stochastic process with conditional density $p(y|\theta_2, v)$. Set $\theta = (\theta_1, \theta_2)$ and suppose that $p(y|\theta) = \int p(v|\theta_1)p(y|v, \theta_2)\ \mathrm{d}v$ is intractable. From now on we do not distinguish θ_1 and θ_2 and simply condition on θ: $p(y|\theta) = \int p(v|\theta)p(y|v, \theta)\ \mathrm{d}v$.

Let the parameters have a prior, $\pi_0(\theta)$. We wish to obtain a sample from the posterior $\pi(\theta)$. Ideally, we would run a Metropolis-Hastings algorithm targeting $\pi(\theta)$, but the intractability of the likelihood prevents this.

Unbiased estimators

Whilst $p(y|\theta)$ is intractable, we can create an estimate, $\hat{p}(y|\theta; u) := p(y|\theta, v)$, where v has a density of $p(v|\theta)$, and u represents all of the auxiliary variables (*e.g.* Unif(0,1)) needed to create a realisation of v. The corresponding estimator is unbiased since

$$\mathbb{E}\left[\hat{p}(y|\theta; U)\right] = \mathbb{E}\left[p(y|\theta, V)\right] = \int p(v|\theta)p(y|\theta, v)\ \mathrm{d}v = p(y|\theta).$$

Clearly, an average of such estimators is also unbiased. Unbiased estimators may also be obtained, for example, from importance sampling (i.e. not sampling from $p(v|\theta)$, but then reweighting) or, for hidden Markov models, by a particle filter.

From now on we simply assume that we have an unbiased estimator of the likelihood $\hat{p}(y|\theta; U)$ where auxiliary variable U is sampled from some density $\tilde{q}(u|\theta)$.

This leads to the following unbiased (up to a fixed constant) estimator of the posterior, $\pi(\theta)$:

$$\hat{\pi}(\theta; U) = \pi_0(\theta)\hat{p}(y|\theta; U).$$

Algorithm

Start with $\theta, \hat{\pi}(\theta|u)$ and at each iteration:

1. Propose θ' from some $q(\theta'|\theta)$.
2. Propose u' from some $\tilde{q}(u'|\theta')$ and hence create $\hat{\pi}(\theta'|u')$.
3. Accept $(\theta', \hat{\pi}(\theta'; u'))$ with probability
$$\alpha(\theta, u; \theta', u') = 1 \wedge \frac{\hat{\pi}(\theta'; u')q(\theta|\theta')}{\hat{\pi}(\theta; u)q(\theta'|\theta)}.$$

Amazingly (Beaumont, 2003; Andrieu and Roberts, 2009), the stationary distribution of the resulting Markov chain has a marginal density of $\pi(\theta)$.

Extended target

In the final section we show that the chain actually targets the joint density

$$\tilde{\pi}(\theta, u) := \hat{\pi}(\theta; u)\tilde{q}(u|\theta) = \pi_0(\theta)\tilde{q}(u|\theta)\hat{p}(y|\theta; u).$$

Since $\hat{p}(y|\theta; u)$ is unbiased, the marginal for this is then

$$\pi_0(\theta) \int \tilde{q}(u|\theta)\hat{p}(y|\theta; u) = \pi_0(\theta)p(y|\theta) \propto \pi(\theta),$$

as required,

Detailed balance

The chain targets $\tilde{\pi}(\theta, u)$ because detailed balance holds with respect to $\tilde{\pi}(\theta, u)$ since

$$\tilde{\pi}(\theta, u)\ q(\theta'|\theta)\tilde{q}(u'|\theta')\ \alpha(\theta, u; \theta', u') = \tilde{q}(u|\theta)\tilde{q}(u'|\theta') \times \left[\hat{\pi}(\theta; u)q(\theta'|\theta) \wedge \hat{\pi}(\theta'; u')q(\theta|\theta')\right],$$

which is invariant to $(\theta, u) \leftrightarrow (\theta', u')$.

One-dimensional representation

The estimator of the likelihood can be rewritten as $\hat{p}(y|\theta;U) = Wp(y|\theta)$, implictly defining

$$W := \frac{\hat{p}(y|\theta;U)}{p(y|\theta)} \quad \text{with} \quad \mathbb{E}\left[W\right] = 1$$

because the estimator is unbiased. The acceptance probability is therefore

$$\alpha(\theta, w; \theta', w') = 1 \wedge \frac{\pi(\theta')q(\theta|\theta')w'}{\pi(\theta)q(\theta'|\theta)w},$$

where w and w' are the multiplicative noises in the estimates of the likelihood at the current and proposed θ values.

W' arises from some (hypothetical) proposal distribution

$$\tilde{q}(w'|\theta') := \int_{u':\hat{p}(y|\theta;u')=wp(y|\theta)} \tilde{q}(u'|\theta')\mathrm{d}u'.$$

Of course w, $\tilde{q}(w|\theta)$ or $\pi(\theta)$ are unknown. However, this representation provides intuition into the behaviour of pseudo-marginal MH and is used in theoretical analyses of the algorithm.

Firstly we realise that the pseudo-marginal algorithm can be viewed as a Markov chain on (θ, w). The extended target is in fact

$$\tilde{\pi}(\theta, w) := \pi(\theta)w\tilde{q}(w|\theta), \tag{1}$$

and, at stationarity, the conditional density of $W|\theta$ is $w\tilde{q}(w|\theta)$; this is a density as $\mathbb{E}_{\tilde{q}}\left[W\right] = 1$.

Ordering pseudo-marginal algorithms

Since $1 \wedge kW'$ is a concave function of W' and $W \wedge k$ is a concave function of W, we may apply Jensen's inequality twice to find (Andrieu and Vihola, 2015):

$$\begin{aligned}
\mathbb{E}_{w\tilde{q}(w|\theta),\tilde{q}(w'|\theta')}\left[\alpha(\theta, W; \theta', W')\right] &= \int \mathrm{d}w\mathrm{d}w' \; w\tilde{q}(w|\theta)\tilde{q}(w'|\theta') \; \alpha(\theta, W; \theta', W') \\
&= \mathbb{E}_{\tilde{q}(w|\theta)}\left[\mathbb{E}_{\tilde{q}(w'|\theta')}\left[W \wedge \left(\frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}W'\right)\right]\right] \leq 1 \wedge \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}.
\end{aligned}$$

Therefore the acceptance rate of a pseudo-marginal MH algorithm is never greater than that of the ideal MH algorithm. In fact, this ordering extends to the spectral gap and to the variance of the estimator of $\mathbb{E}_{\pi}\left[f(\theta)\right]$ for any $f \in L_0^2(\pi)$.

Andrieu and Vihola (2015) generalise these results to pairs of pseudo-marginal algorithms: whenever one algorithm can be viewed as a noisy version of another then the noisier one is always less efficient. In particular, a PMMH algorithm that uses an average of two or more unbiased estimators is always more efficient than an algorithm which uses just one of the estimators.

Tuning m when $\hat{p}$ is obtained using a particle filter

The multiplicative noise in the log-posterior, W, can, in general, have any distribution provided it is non-negative and $\mathbb{E}[W] = 1$. However, when $\hat{p}(y|\theta, U)$ is obtained via a particle filter (or SMC) then in the limit as the number of data points, $T \to \infty$ and with the number of particles $m = t/\beta$, for some $\beta > 0$ then, subject to mixing conditions (Bérard et al., 2014) the noise in a new proposal satisfies:

$$\log W' \Rightarrow \mathsf{N}\left(-\frac{1}{2}\sigma^2, \sigma^2\right),$$

for some $\sigma^2 > 0$ which, typically, depends on the parameters, θ, well as the data generating process. We will provide a heuristic for this result, but first let us note some consequences.

Suppose that σ does not depend on θ. [1] For convenience, set $V := \log W$ and $V' := \log W'$. Thus $V' \sim \mathsf{N}(-\sigma^2/2, \sigma^2)$ and immediately from (1) and the line beneath, the conditional (and marginal) density of V is

$$\exp[v] \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(v + \sigma^2/2)^2\right] = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(v - \sigma^2/2)^2\right],$$

so that $V' \sim \mathsf{N}(\sigma^2/2, \sigma^2)$, or

$$\log W \sim \mathsf{N}\left(\frac{1}{2}\sigma^2, \sigma^2\right) \quad \text{and} \quad \log W' - \log W \sim \mathsf{N}\left(-\sigma^2, 2\sigma^2\right).$$

Thus, the ratio W'/W in the pseudo-marginal acceptance probability has a lognormal distribution. This is the starting point for several papers (Pitt et al., 2012; Sherlock et al., 2015; Doucet et al., 2015; Nemeth et al., 2016) that provide advice on tuning PMMH algorithms when using a particle filter. All recommend choosing m to give some approximately optimal $\hat{\sigma}^2$ value, with the recommended $\hat{\sigma}^2$ somewhere between 0.8 and 3.3.

[1] More realistically, $\sigma(\theta)$ varies slowly with θ so if $q(\theta'|\theta)$ is a local move, $\sigma^2(\theta') \approx \sigma^2(\theta)$ and the following result holds approximately.

Sketch proof of the Gaussian limit

For simplicity, suppose that the data, $Y_{1:T} := (Y_1, \ldots, Y_t)$ are iid. Conditional on the tth data point, y_t, we generate m independent auxiliary variables, $U_{t,i}$, $(i = 1, \ldots, m)$. Our estimator of the likelihood is

$$\hat{p}(y_{1:T}|\theta, U) = \prod_{t=1}^{T} \frac{1}{m} \sum_{i=1}^{m} \hat{p}_1(y_t|\theta, U_{t,i}) = \prod_{t=1}^{T} \frac{1}{m} \sum_{i=1}^{m} \hat{p}_1(y_t|\theta) W_{t,i} = p(y_{1:T}|\theta) \prod_{t=1}^{T} \frac{1}{m} \sum_{i=1}^{m} W_{t,i},$$

where $\hat{p}_1(y|\theta, u)$ is the unbiased estimator of the likelihood of a single observation, $p_1(y|\theta)$, given the auxiliary variable u, and $W_{t,i} := \hat{p}_1(y_t|\theta, U_{t,i})/p(y_t|\theta)$.

Applying a second-order Taylor expansion, $\log \hat{p}(y_{1:T}|\theta, U) - \log p(y_{1:T}|\theta)$ is

$$\sum_{t=1}^{T} \log \left\{ 1 + \left[\frac{1}{m} \sum_{i=1}^{m} W_{t,i} - 1 \right] \right\} \approx \sum_{t=1}^{T} \left[\frac{1}{m} \sum_{i=1}^{m} W_{t,i} - 1 \right] - \frac{1}{2} \left[\frac{1}{m} \sum_{i=1}^{m} W_{t,i} - 1 \right]^2 .$$

The $W_{t,i}$ are independent; set $\tau_t^2 := \mathrm{Var}(W_{t,i}) < \infty$, and denote $\tau^2 = \mathbb{E}\left[\tau_t^2\right]$, where expectation is over the distribution of Y_t. For simplicity, we ignore the detail that $T = [m\beta]$ rather than $T = m\beta$. The first term in the expansion is

$$\sum_{t=1}^{T} \left[\frac{1}{m} \sum_{i=1}^{m} W_{t,i} - 1 \right] = \sqrt{\beta} \times \frac{1}{\sqrt{\beta m}} \sum_{t=1}^{\beta m} A_t \Rightarrow \mathsf{N}\left(0, \beta\tau^2\right),$$

by the SLLN, where $A_t := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (W_{t,i} - 1) \Rightarrow \mathsf{N}(0, \tau_t^2)$ by the CLT. Similarly, by the SLLN we obtain

$$\sum_{t=1}^{T} \left[\frac{1}{m} \sum_{i=1}^{m} W_{t,i} - 1 \right]^2 = \beta \frac{1}{\beta m} \sum_{t=1}^{\beta m} B_t \xrightarrow{\text{a.s.}} \beta\tau^2,$$

where $B_t := \left[\frac{1}{\sqrt{m}} \sum_{i=1}^{m} (W_{t,i} - 1) \right]^2$ are independent with finite means of τ_t^2. Combining these two limits leads to the required result.

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *Ann. Statist.*, 37(2):697–725.

Andrieu, C. and Vihola, M. (2015). Convergence properties of pseudo-marginal markov chain monte carlo algorithms. *Ann. Appl. Probab.*, 25(2):1030–1077.

Andrieu, C. and Vihola, M. (2015). Establishing some order amongst exact approximations of MCMCs. *ArXiv e-prints.*

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Bérard, J., Moral, P. D., and Doucet, A. (2014). A lognormal central limit theorem for particle approximations of normalizing constants. *Electron. J. Probab.*, 19:no. 93, 1–28.

Doucet, A., Pitt, M., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. *Biometrika.* To appear.

Nemeth, C., Sherlock, C., and Fearnhead, P. (2016). Particle Metropolis-adjusted Langevin algorithms. *Biometrika.* Accepted for publication.

Pitt, M. K., dos Santos Silva, R., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134 – 151.

Sherlock, C., Thiery, A., Roberts, G. O., and Rosenthal, J. S. (2015). On the efficiency of pseudo-marginal random walk Metropolis algorithms. *Ann. Stat.*, 43(1):238–275.