Discussion of the paper 'Riemann manifold Langevin and Hamiltonian Monte Carlo methods' by M. Girolami and B. Calderhead

M. K. Titsias

University of Manchester, Manchester, UK.

The stochastic volatility and log-Gaussian Cox process models are examples of latent Gaussian models with joint density

$$p(\mathbf{y},\mathbf{x}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}),$$

where $p(\mathbf{y}|\mathbf{x})$ is a non-Gaussian likelihood and $p(\mathbf{x})$ a Gaussian prior distribution over the latent vector $\mathbf{x}$. Here, we only wish to discuss sampling the latent vector. A potential limitation of the proposed algorithms is that they require the first and second derivatives of the full joint density. This is restrictive because in certain applications we may need to deal with limited information regarding the geometry of the likelihood. For instance, the use of second derivatives of $\log p(\mathbf{y}|\mathbf{x})$ can often be undesirable because of high computational cost. In contrast, the full information geometry of the Gaussian prior can always be taken into account.

Consider the proposal distribution

$$Q(\mathbf{x}'|\mathbf{x}) \propto H(\mathbf{x}',\mathbf{x})p(\mathbf{x}'),$$

which proposes a new $\mathbf{x}'$ given the current $\mathbf{x}$. $H(\mathbf{x}',\mathbf{x})$ is such that its logarithm is quadratic in $\mathbf{x}'$, thus $Q(\mathbf{x}'|\mathbf{x})$ is Gaussian. By construction the proposal distribution is invariant to the Gaussian prior. $H(\mathbf{x}',\mathbf{x})$ should be set to incorporate some properties of the non-Gaussian likelihood $p(\mathbf{y}|\mathbf{x})$. Auxiliary variables can be employed for such construction. The idea is to approximate the non-Gaussian likelihood by an auxiliary Gaussian likelihood $p(\mathbf{z}|\mathbf{x})$ where $\mathbf{z}$ are auxiliary variables that can be regarded as pseudo data. The simplest choice is to use $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{x}, \frac{\sigma^2}{2}I)$ which says that $\mathbf{z}$ is a noisy version of $\mathbf{x}$. The sampling scheme iterates between updating $\mathbf{z}$ and $\mathbf{x}$ according to

(a) $\mathbf{z} = \mathbf{x} + \frac{\sigma}{\sqrt{2}}\eta$, $\eta \sim \mathcal{N}(\mathbf{0}, I)$
(b) $\mathbf{x}' \sim \frac{1}{\mathcal{Z}(\mathbf{z})}\mathcal{N}(\mathbf{z}|\mathbf{x}', \frac{\sigma^2}{2}I)p(\mathbf{x}')$ and accept/reject using MH

This iteration leaves $p(\mathbf{x}|\mathbf{y})$ invariant and implies a symmetric form for $H(\mathbf{x}',\mathbf{x})$, i.e. $H(\mathbf{x}',\mathbf{x}) = H(\mathbf{x},\mathbf{x}')$. When the variance of the Gaussian $p(\mathbf{x})$ tends to infinity, step (b) reduces to $\mathbf{x}' = \mathbf{z} + \frac{\sigma}{\sqrt{2}}\eta$ and both steps combined yield $\mathbf{x}' = \mathbf{x} + \sigma\eta$. The above algorithm can be thought of as a *Gaussian scaled random walk Metropolis* (GS-RWM). Similarly, we can obtain a *Gaussian scaled Metropolis adjusted Langevin algorithm* (GS-MALA) by sampling $\mathbf{z}$ in step (a) (while keeping (b) unchanged) according to

$$\mathbf{z} = \mathbf{x} + \frac{\sigma^2}{2}\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}) + \frac{\sigma}{\sqrt{2}}\eta.$$

This implies that the auxiliary likelihood is now $p(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mathbf{x} + \frac{\sigma^2}{2}\nabla_{\mathbf{x}} \log p(\mathbf{y}|\mathbf{x}), \frac{\sigma^2}{2}I)$ and the scheme reduces to standard MALA when the variance of $p(\mathbf{x})$ approaches infinity. When second

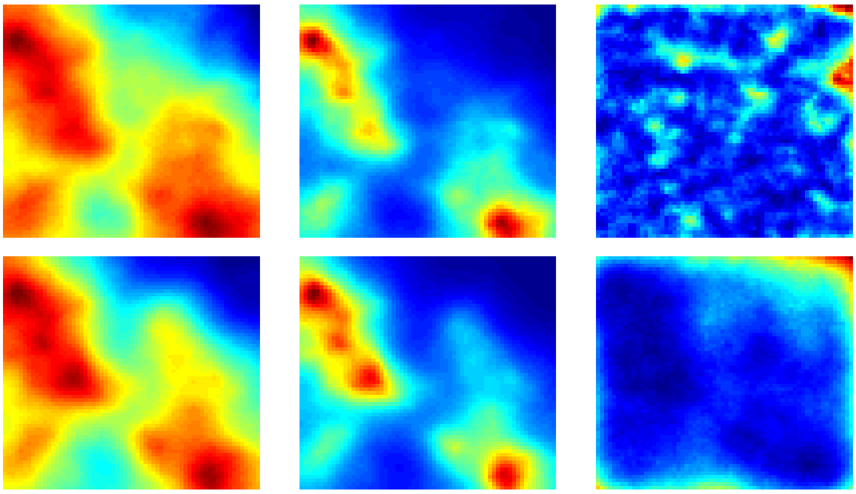

Fig. 1. Posterior Monte Carlo estimates obtained by GS-RWM (plots in the first row) and GS-MALA (second row). The format of the figure and the experiment follows exactly Fig. 8. The results are obtained from 5000 posterior samples after a burn-in period in which the scale σ^2 of the proposal distribution was adapted to achieve a certain acceptance rate. For GS-RWM, σ^2 was tuned to achieve an acceptance rate of between 20% to 30%, while for GS-MALA the range was 50% to 60%. Notice that GS-MALA provides quite satisfactory results.

Table 1. Effective sample sizes of GS-RWM and GS-MALA on the log-Gaussian Cox point process example. Notice that the GS-MALA that uses gradient information about the log likelihood has significantly better performance than GS-RWM and also slightly outperforms MMALA (see Table 10). Running times include also the adaptive phase that tunes σ^2. GS-RWM had larger running time since it required longer adaptive phase.

Method	*Time(s)*	*ESS(minimum, median, maximum)*	*s/minimum ESS*
GS-RWM	1311	(4, 29, 109)	327.7
GS-MALA	942	(36, 205, 524)	26.1

derivatives of $\log p(\mathbf{y}|\mathbf{x})$ are easy to compute, further algorithms can be obtained by following the above framework. Preliminary results using GS-RWM and GS-MALA on the log-Gaussian Cox point process example are shown in Fig. 1 and Table 1.