Noisy Monte Carlo algorithms

Richard Everitt

University of Reading

January 7th, 2016

Types of intractable likelihood

- A likelihood is intractable when it is difficult to evaluate pointwise at θ.

1. Big data
$$f(y|\theta) = \prod_{i=1}^{N} f_i(y_i|\theta).$$
2. When there are a large number of latent variables x, with
$$f(y|\theta) = \int_x f(y,x|\theta)dx.$$
3. When, for an intractable $Z(\theta)$ (e.g for a *Markov random field*),
$$f(y|\theta) = \frac{1}{Z(\theta)}\gamma(y|\theta).$$
4. Where $f(\cdot|\theta)$ can be sampled, but not evaluated.

Exact-approximate methods

- Suppose that, for any θ, it is possible to compute an unbiased estimate $\widehat{f}(y|\theta)$ of $f(y|\theta)$. Then...

1. Using the acceptance probability

$$\alpha\left(\theta^{(p)},\theta^{*}\right)=\min\left\{1,\frac{\widehat{f}(y|\theta^{*})p(\theta^{*})q(\theta^{(p)}|\theta^{*})}{\widehat{f}(y|\theta^{(p)})p(\theta^{(p)})q(\theta^{*}|\theta^{(p)})}\right\}$$

   yields an MCMC algorithm with target distibution $\pi(\theta|y)$.

2. Using the weight

$$w^{(p)}=\frac{\widehat{f}(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}$$

   yields an importance sampling algorithm with target distribution $\pi(\theta|y)$.

Beaumont (2003), Andrieu and Roberts (2009), Fearnhead et al. (2010).

Type 3: "doubly intractable" distributions

- Coined by Murray et al. (2006).
- Intractable in that we need to resort to simulation.
- Doubly intractable since the acceptance probability in MH

$$\min\left\{1, \frac{\gamma(y|\theta^*)}{\gamma(y|\theta^{(p)})}\frac{p(\theta^*)}{p(\theta^{(p)})}\frac{q(\theta^{(p)}|\theta^*)}{q(\theta^*|\theta^{(p)})}\frac{1}{Z(\theta^*)}\frac{Z(\theta^{(p)})}{1}\right\}$$

requires evaluating the intractable term Z.

The single auxiliary variable (SAV) method

- Møller et al. (2006) use

$$\frac{q_u(u^*|\theta^*,y)}{\gamma(u^*|\theta^*)}$$

with some distribution q_u and $u^* \sim f(.|\theta^*)$, as an unbiased importance sampling estimator of $\frac{1}{Z(\theta^*)}$.

- This gives an acceptance probability of

$$\min\left\{1, \frac{\gamma(y|\theta^*)}{\gamma(y|\theta^{(p)})}\frac{p(\theta^*)}{p(\theta^{(p)})}\frac{q(\theta^{(p)}|\theta^*)}{q(\theta^*|\theta^{(p)})}\frac{q_u(u^*|\theta^*,y)}{\gamma(u^*|\theta^*)}\frac{\gamma(u|\theta^{(p)})}{q_u(u|\theta^{(p)},y)}\right\}.$$

SAV importance sampling

- Everitt et al. (2016) use
$$\frac{q_u(u^*|\theta^*,y)}{\gamma(u^*|\theta^*)}$$
with some distribution q_u and $u^* \sim f(.|\theta^*)$, as an unbiased importance sampling estimator of $\frac{1}{Z(\theta^*)}$.
- Using $\frac{q_u(u|\theta^*,y)}{\gamma(u|\theta^*)}$ as an IS estimator of $\frac{1}{Z(\theta^*)}$ we obtain
$$w^{(p)} = \frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}\frac{q_u(u|\theta^{(p)},y)}{\gamma(u|\theta^{(p)})}.$$
- Note: we may use multiple importance points, i.e. use
$$\frac{1}{M}\sum_{m=1}^{M}\frac{q_u(u^{(m)}|\theta^*,y)}{\gamma(u^{(m)}|\theta^*)}.$$

Noisy methods

- The use of “inexact approximate” or “noisy” methods in which an exact method is approximated **without** resulting in an exact target distribution.
- Focus on doubly intractable problems
  - strong link to work on other types of intractable likelihood.
- In particular, that an exact sampler does not exist for $u^* \sim f(.|\theta^*)$.
- Alternatives:
  - Russian roulette (Lyne et al., 2015);
  - use a long run of an MCMC in place of an exact sampler (Caimo and Friel, 2011; Everitt, 2012).

Justification?

1. Is the distribution targeted by the noisy algorithm close to the exact target?
2. What is the error in estimates produced by the noisy algorithm?
   - given a fixed computational budget, how should it be allocated to minimise the error of estimates?

- Everitt R. G. (2012). Bayesian Parameter Estimation for Latent Markov Random Fields and Social Networks, Journal of Computational and Graphical Statistics, 21(4), 940-960, or arXiv(1203.3725)
- Alquier, P., Friel, N., Everitt, R. G., Boland, A. (2015). Noisy Monte Carlo: Convergence of Markov chains with approximate transition kernels, Statistics and Computing, or arXiv(1403.5496).
- Everitt, R. G., Johansen, A. M., Rowing, E., Evdemon-Hogan, M. (2016). Bayesian model comparison with un-normalised likelihoods, arXiv(1504.00298).

Noisy MCMC

- MCMC involves simulating a Markov chain $(\theta_n)_{n\in\mathbb{N}}$ with transition kernel P such that π is invariant under P.
- In some situations there is a natural kernel P with this property, but which we cannot draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n.
- A natural idea is to replace P by an approximation $\hat{P}$.
- This leads to the obvious question:

  Can we say something on how close the resultant Markov chain with transition kernel $\hat{P}$ is that resulting from P? Eg, is it possible to upper bound?

$$\left\| \delta_{\theta_0} \hat{P}^n - \pi \right\|.$$

- It turns out that a useful answer is given by the study of the stability of Markov chains.

Noisy MCMC

- MCMC involves simulating a Markov chain $(\theta_n)_{n\in\mathbb{N}}$ with transition kernel P such that π is invariant under P.
- In some situations there is a natural kernel P with this property, but which we cannot draw $\theta_{n+1} \sim P(\theta_n, \cdot)$ for a fixed θ_n.
- A natural idea is to replace P by an approximation $\hat{P}$.
- This leads to the obvious question:

  Can we say something on how close the resultant Markov chain with transition kernel $\hat{P}$ is that resulting from P? Eg, is it possible to upper bound?

$$\left\| \delta_{\theta_0} \hat{P}^n - \pi \right\|.$$

- It turns out that a useful answer is given by the study of the stability of Markov chains.

Stability result

Theorem (Mitrophanov (2005), Corollary 3.1)

- *If* **(H1)** *the MC with transition kernel P is uniformly ergodic:*

$$\sup_{\theta_0} \|\delta_{\theta_0} P^n - \pi\| \leq C\rho^n$$

for some $C < \infty$ and $\rho < 1$.

Then we have, for any $n \in \mathbb{N}$, for any starting point θ_0,

$$\|\delta_{\theta_0} P^n - \delta_{\theta_0} \hat{P}^n\| \leq \left(\lambda + \frac{C\rho^\lambda}{1-\rho}\right) \|P - \hat{P}\|$$

where $\lambda = \left\lceil \frac{\log(1/C)}{\log(\rho)} \right\rceil$.

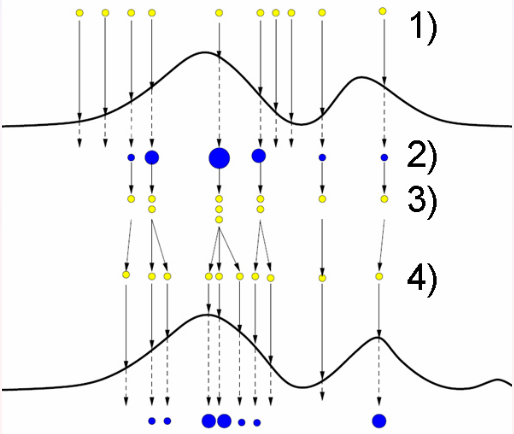
Noisy MCMC: uniform ergodicity

- So, if something can be said about $\|P - \hat{P}\|$ we know something about:
  - the distance between the iterated noisy and exact kernels
  - when the invariant distribution of $\hat{P}$ exists, the distance between the noisy and exact targets.
- In particular:
  - Everitt (2012) shows that when the burn in increases, the distance goes to zero;
  - same argument is used in Andrieu and Roberts (2009) for Monte Carlo within Metropolis;
  - Alquier et al. (2015) give cases where the bound on $\|P - \hat{P}\|$ can be given in terms of M (where the quality of the approximation goes to zero as $M \to \infty$).

Noisy MCMC: geometric ergodicity

- Alquier et al. (2015) also note that similar results can hold in the geometrically ergodic case
  - from result in Ferré, Hervé and Ledoux (2013)
  - taken much further by Medina-Aguayo et al. (2015).
- Further developments in the geometrically ergodic case, and using Wasserstein distance rather than total variation
  - Pillai and Smith (2014);
  - Rudolf and Schweizer (2015).

Sequential Monte Carlo



SMC samplers

An iteration of an SMC algorithm at target $t+1$.

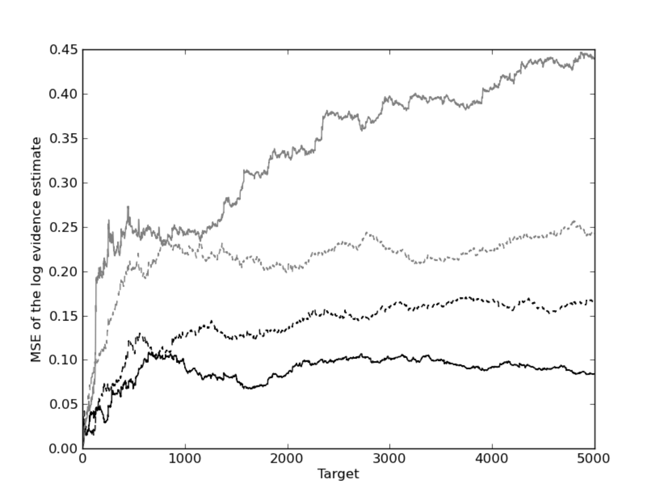
- For $p = 1 : P$
  - Update $\theta_t^{(p)}$ to $\theta_{t+1}^{(p)}$ using some kernel K.
- For $p = 1 : P$
  - Reweight: find $\widetilde{w}_{t+1}^{(p)}$, so that the $\left(\theta_{t+1}^{(p)}, \widetilde{w}_{t+1}^{(p)}\right)$ are (unnormalised) weighted points from $p_{t+1}(.|y)$.
- Normalise $\left\{\widetilde{w}_{t+1}^{(p)}\right\}_{p=1}^{P}$ to give $\left\{w_{t+1}^{(p)}\right\}_{p=1}^{P}$.
- Resample the weighted points if some threshold is met.

- An estimate of the marginal likelihood is given by $\prod_{t=1}^{T} \sum_{p=1}^{P} \widetilde{w}_t^{(p)}$.

Noisy SMC: strong mixing assumptions

- In Everitt et al (2016), we
  - use biased weights at every step of the SMC;
  - are interested in how the error accumulates as the SMC algorithm iterates.
- Under strong mixing assumptions (stronger than a global Doeblin condition) we obtain a uniform bound on total-variation discrepancy between the iterated target distributions of the exact and noisy methods
  - strong mixing can prevent the accumulation of error even in systems with biased weights.

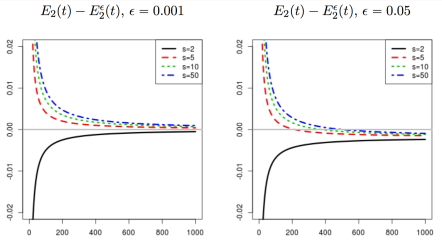
Noisy SMC: empirical study



Error of estimates: noisy MCMC

- The noisy method is more efficient (in terms of mean squared error) if

$$\frac{3}{s_\varepsilon P}\left(1+\frac{1}{s_\varepsilon P}\right)+\frac{3}{4}\varepsilon^2<\frac{1}{P}\left(1+\frac{1}{P}\right).$$



Johndrow, J. E., Mattingly, J. C. Mukherjee, S. Dunson, D. (2015) Approximations of Markov Chains and High-Dimensional Bayesian Inference, arXiv.

Error of estimates: noisy IS

- Noisy importance sampling and sequential Monte Carlo: Everitt et al (2016).
- Under some simplifying assumptions, noisy importance sampling is more efficient (in terms of mean squared error) compared to an exact-approximate algorithm if

$$\frac{1}{P}\left(\mathrm{Var}_q\left[w(\theta)+b(\theta)\right]+\mathbb{E}_q[\grave{\sigma}_\theta^2]\right)+\mathbb{E}_q[b(\theta)]^2 < \frac{1}{P}\left(\mathrm{Var}_q\left[w(\theta)\right]+\mathbb{E}_q[\acute{\sigma}_\theta^2]\right),$$

where $b(\theta)>0$ is the bias of the noisy weights, $\grave{\sigma}_\theta^2$ is the variance of the noisy weights, $\acute{\sigma}_\theta^2$ is the variance of the exact-approximate weights and

$$w(\theta) := \frac{p(\theta)\gamma(y|\theta)}{Z(\theta)q(\theta)}.$$

SAV importance sampling

- Recall SAVIS.
- Use it to estimate the marginal likelihood $p(y)$.
- We obtain

$$\widehat{p(y)} = \frac{1}{P}\sum_{p=1}^{P}\frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}\sum_{m=1}^{M}\frac{q_u(u^{(m,p)}|\theta^{(p)},y)}{\gamma(u^{(m,p)}|\theta^{(p)})},$$

where the $u^{(m,p)}$ are generated by taking the final point of a long MCMC run (length B) targeting $f(\cdot|\theta^{(p)})$.

SAV importance sampling

- Recall SAVIS.
- Use it to estimate the marginal likelihood $p(y)$.
- We obtain

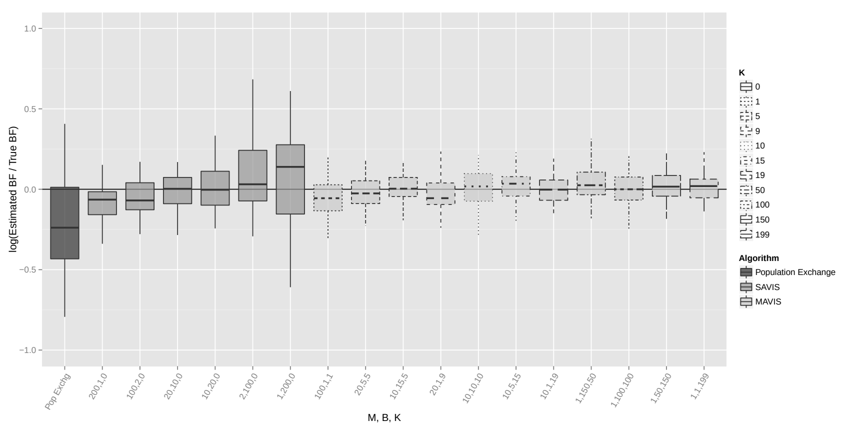
$$\widehat{p(y)} = \frac{1}{P}\sum_{p=1}^{P}\frac{\gamma(y|\theta^{(p)})p(\theta^{(p)})}{q(\theta^{(p)})}\sum_{m=1}^{M}\frac{q_u(u^{(m,p)}|\theta^{(p)},y)}{\gamma(u^{(m,p)}|\theta^{(p)})},$$

where the $u^{(m,p)}$ are generated by taking the final point of a long MCMC run (length B) targeting $f(\cdot|\theta^{(p)})$.

Application to Ising models

- An Ising model is a pairwise Markov random field with binary variables.
- Reanalyse the data from Friel (2013), which consists of 20 realisations from a first-order 10×10 Ising model and 20 realisations from a second-order 10×10 Ising model.
- Compare
  - population exchange;
  - SAVIS and variations on this idea.

Ising models: results



Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Promising results, but many open questions:
  - what one should do in practice is not obvious;
  - potential accumulation of bias in SMC (mitigated by mixing well);
  - in both cases the theory requires very strong assumptions.

Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Promising results, but many open questions:
  - what one should do in practice is not obvious;
  - potential accumulation of bias in SMC (mitigated by mixing well);
  - in both cases the theory requires very strong assumptions.

Conclusions

- Use exact methods where possible...
- ... however the bias from a noisy method may be small compared to errors resulting from commonly accepted approximate techniques such as ABC (and also the Monte Carlo variance).
- What is the best we can do fo some finite computational budget?
- Promising results, but many open questions:
  - what one should do in practice is not obvious;
  - potential accumulation of bias in SMC (mitigated by mixing well);
  - in both cases the theory requires very strong assumptions.

Acknowledgements

- Noisy MCMC: Pierre Alquier, Nial Friel and Aiden Boland (UCD).
- Noisy IS and SMC: Adam Johansen (Warwick), Melina Evdemon-Hogan and Ellen Rowing (Reading).