Pseudo-Marginal Slice Sampling

IAIN MURRAY (i.murray@ed.ac.uk) MATTHEW M. GRAHAM (m.m.graham@ed.ac.uk)

THE UNIVERSITY of EDINBURGH informatics

PROBLEM

We want to sample from some target distribution

$$\pi(\boldsymbol{\theta}) = \frac{f(\boldsymbol{\theta})}{C}$$

but only have a noisy unbiased estimator for $f(\boldsymbol{\theta})$

$$\hat{f}(\boldsymbol{\theta};\, \boldsymbol{u}) \quad : \quad \mathbb{E}_{q(\boldsymbol{u})}\left[\hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})\right] = f(\boldsymbol{\theta})$$

$\boldsymbol{u}$ random variates used in estimator $\sim q(\cdot)$

SPECIAL CASES

LATENT VARIABLE MODELS

Marginal likelihood $p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \int p(\boldsymbol{y}\,|\,\boldsymbol{z}, \boldsymbol{\theta})\, p(\boldsymbol{z}\,|\,\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{z}$

$$f(\boldsymbol{\theta}) = p(\boldsymbol{y}\,|\,\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \quad \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u}) = \frac{p(\boldsymbol{y}\,|\,\mathrm{z}(\boldsymbol{u}), \boldsymbol{\theta})\, p(\mathrm{z}(\boldsymbol{u})\,|\,\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{\mathcal{N}(\mathrm{z}(\boldsymbol{u});\, \boldsymbol{\mu},\, \boldsymbol{\Sigma})} \quad q(\boldsymbol{u}) = \mathcal{N}(\boldsymbol{u};\, \mathbf{0},\, \mathbf{I})$$

z deterministic function $\mathrm{z}(\boldsymbol{u}) = \boldsymbol{\mu} + \mathrm{chol}(\boldsymbol{\Sigma})\,\mathbf{u}$

$\boldsymbol{\mu}, \boldsymbol{\Sigma}$ Gaussian approximation $\mathcal{N}(\boldsymbol{z};\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \approx p(\boldsymbol{z}\,|\,\boldsymbol{y}, \boldsymbol{\theta})$

DOUBLY-INTRACTABLE DISTRIBUTIONS

Likelihood $p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \frac{g(\boldsymbol{y};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$ with $Z(\boldsymbol{\theta})$ intractable

$$f(\boldsymbol{\theta}) = g(\boldsymbol{y};\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \frac{Z(\hat{\boldsymbol{\theta}})}{Z(\boldsymbol{\theta})} \quad \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u}) = g(\boldsymbol{y};\boldsymbol{\theta})\, p(\boldsymbol{\theta}) \frac{g(\mathrm{y}(\boldsymbol{u});\hat{\boldsymbol{\theta}})}{g(\mathrm{y}(\boldsymbol{u});\boldsymbol{\theta})} \quad q(\boldsymbol{u}) = \mathcal{U}(\boldsymbol{u})$$

y deterministic function $\boldsymbol{u} \sim \mathcal{U}(\cdot) \Rightarrow \mathrm{y}(\boldsymbol{u}) \sim p(\cdot\,|\,\boldsymbol{\theta})$

$\hat{\boldsymbol{\theta}}$ fixed set of reference parameters

STANDARD APPROACH

Pseudo-Marginal (PM) MCMC, as analysed in Andrieu and Roberts (2009):

Inputs: current parameters $\boldsymbol{\theta}$, previous estimate of unnormalized target probability $\hat{f}$, proposal dist. $r(\boldsymbol{\theta}'; \boldsymbol{\theta})$, unbiased estimator : $\mathbb{E}_{\epsilon(\hat{f};\,\boldsymbol{\theta})}\left[\hat{f}\right] = f(\boldsymbol{\theta}) \;\forall\, \boldsymbol{\theta}$,

Output: new state-estimate pair $(\boldsymbol{\theta}, \hat{f})$.

1. Propose new state and estimate its probability:

$$\boldsymbol{\theta}' \sim r(\,\cdot\,; \boldsymbol{\theta})$$
$$\hat{f}' \sim \epsilon(\,\cdot\,; \boldsymbol{\theta}')$$

2. Metropolis–Hastings style acceptance rule,

with probability $\min\left(1,\, \frac{\hat{f}'}{\hat{f}} \frac{r(\boldsymbol{\theta};\, \boldsymbol{\theta}')}{r(\boldsymbol{\theta}';\, \boldsymbol{\theta})}\right)$:

Accept: **return** $(\boldsymbol{\theta}', \hat{f}')$

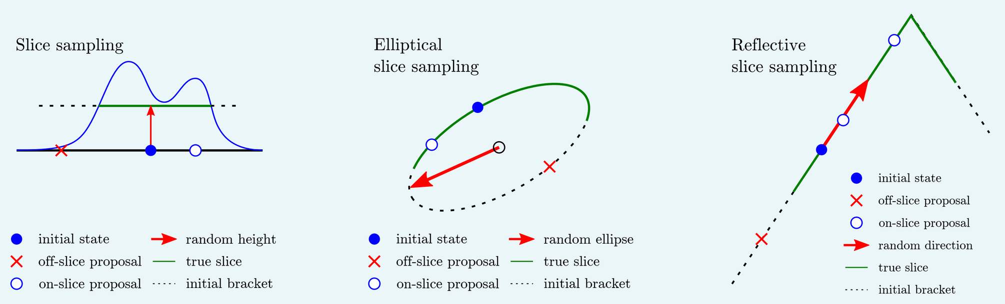
else: Reject: **return** $(\boldsymbol{\theta}, \hat{f})$

If estimator distribution is heavy-tailed, moves to large $\hat{f}$ are occasionally accepted which it is hard to move away from.

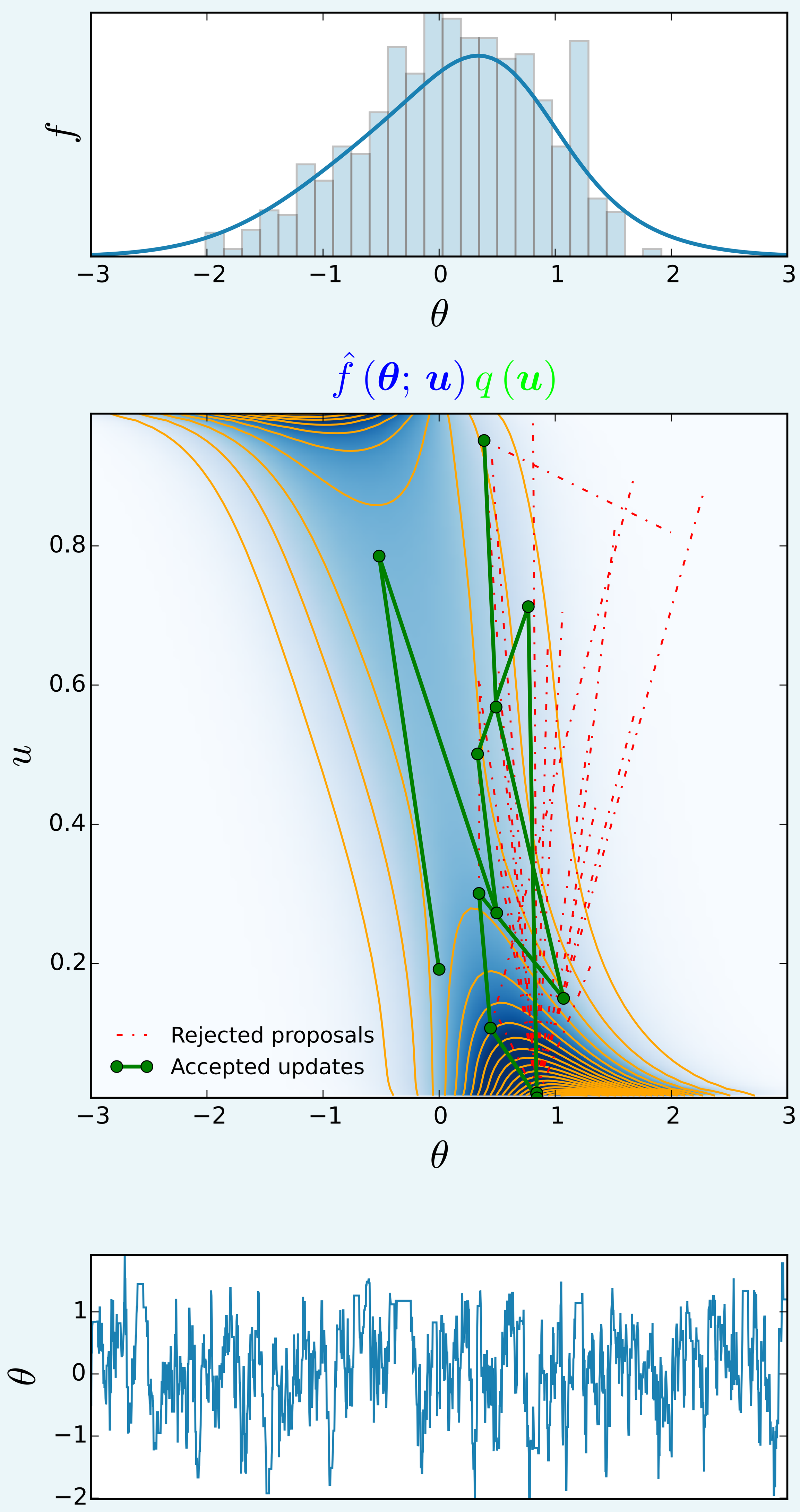
PM methods notorious for this 'sticking' behaviour.

SLICE SAMPLING

Slice sampling (Neal, 2003) - family of MCMC algorithms with updates that locally adapt to the target density.



Slice sampling algorithms always move the state being updated: can we apply them in the pseudo-marginal setting?



OUR APPROACH

Consider $\boldsymbol{u}$ as auxiliary variables as in Chopin and Singh (2015). This gives a new state $(\boldsymbol{\theta}, \boldsymbol{u})$ with joint target density

$$\pi(\boldsymbol{\theta},\, \boldsymbol{u}) = \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})\, q(\boldsymbol{u})/C$$

$$\int \pi(\boldsymbol{\theta},\, \boldsymbol{u})\, \mathrm{d}\boldsymbol{u} = \int \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})\, q(\boldsymbol{u})/C\, \mathrm{d}\boldsymbol{u} = f(\boldsymbol{\theta})/C$$

Standard PM MCMC: Metropolis–Hastings update on the joint target with proposal

$$\tilde{r}\left[(\boldsymbol{\theta}',\, \boldsymbol{u}')\,|\,(\boldsymbol{\theta},\, \boldsymbol{u})\right] = r(\boldsymbol{\theta}';\, \boldsymbol{\theta})\, q(\boldsymbol{u}')$$

Our *Auxiliary Pseudo-Marginal* (APM) framework:

Inputs: current state: parameters $\boldsymbol{\theta}$, randomness $\boldsymbol{u}$; unbiased estimator : $\mathbb{E}_{q(\boldsymbol{u})}\left[\hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})\right] = f(\boldsymbol{\theta}) \;\forall\, \boldsymbol{\theta}$,

Output: new state $(\boldsymbol{\theta}, \boldsymbol{u})$.

1. Update $\boldsymbol{u}$ leaving invariant its target conditional:

$$\pi(\boldsymbol{u}\,|\,\boldsymbol{\theta}) \;\propto\; \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})\, q(\boldsymbol{u})$$

2. Update $\boldsymbol{\theta}$ leaving invariant its target conditional:

$$\pi(\boldsymbol{\theta}\,|\,\boldsymbol{u}) \;\propto\; \hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})$$

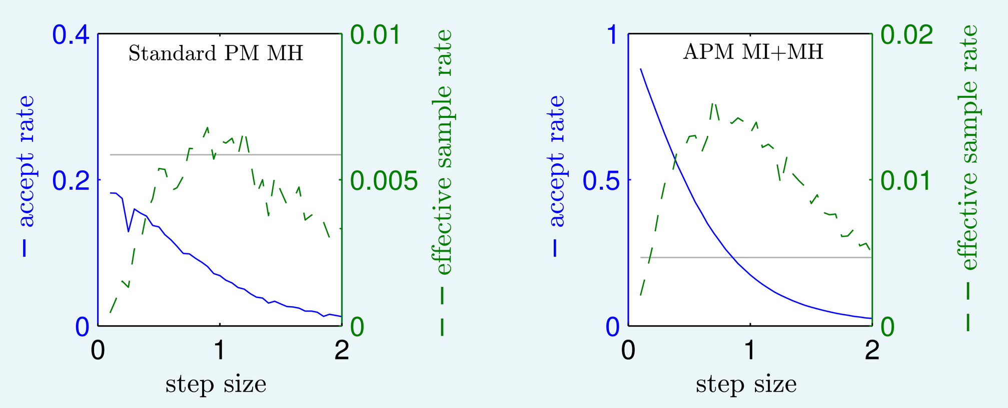
Free to mix and match any standard MCMC operators for the two updates.

With $\boldsymbol{u}$ now clamped during the $\boldsymbol{\theta}$ updates, usual case of a conditional distribution proportional to a deterministic function.

Slice sampling will move $\boldsymbol{\theta}$ if $\hat{f}(\boldsymbol{\theta};\, \boldsymbol{u})$ is continuous almost everywhere.

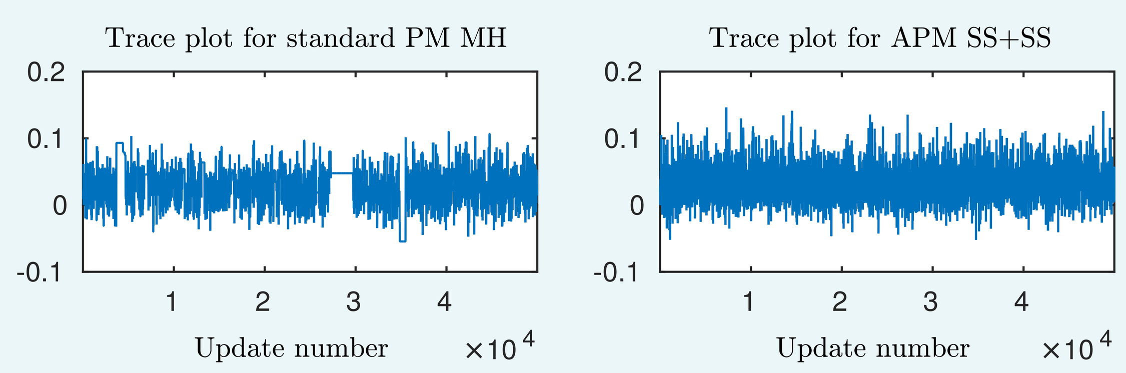
Equally can make perturbative moves to the random variates $\boldsymbol{u}$ with slice sampling or other adaptive methods.

EXPERIMENTS



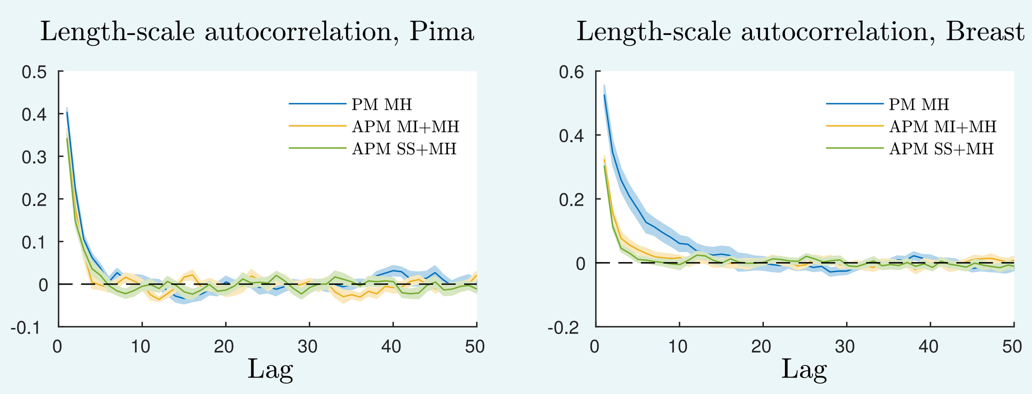
Clearer step-size selection

- Optimal efficiency at step size giving acceptance rate 0.234 for standard Metropolis proposals in high dimensions (Roberts et al. 1997).
- Not applicable to PM approach - in Gaussian test case 0.234 acceptance not even achievable.
- Using APM with independent proposals for $\boldsymbol{u}$ update and Gaussian proposals for $\boldsymbol{\theta}$ update, optimal acceptance rate now close to 0.234.



Parameter inference in Ising model

- Doubly-intractable distribution - 30 × 10 lattice.
- Data generated using exact sampler.
- Using standard PM approach chains stick for long periods.
- Using APM framework with slice sampling updates for $\boldsymbol{u}$ and $\boldsymbol{\theta}$ eliminates sticking and gives much improved cost-scaled autocorrelations.
- With lower-variance estimator using annealed importance sampling, standard PM performs better than APM.



Parameter inference in Gaussian process classifier

- Two UCI datasets, Pima and Breast, modelled following Filippone and Girolami, (2014).
- Estimator for marginal likelihood - importance sampler with using Gaussian latent posterior approximation as importance distribution.
- Adaptive tuning of step size significantly more consistent with APM approaches.
- Cost-normalised effective sample size when using APM updates significantly better than standard PM approach.

CONCLUSION

- General method for clamping and updating random numbers used in a pseudo-marginal method's unbiased estimator.
- Within this framework it is possible to use slice sampling and other adaptive methods.
- Resulting Markov chains are more robust and often mix more quickly.

REFERENCES

Andrieu & Roberts. The pseudo-marginal approach for efficient Monte Carlo computations. *Annals of Statistics*, 2009.
Chopin & Singh. On particle Gibbs sampling. *Bernoulli*, 2015
Filippone & Girolami. Pseudo-marginal Bayesian inference for Gaussian Processes. *IEEE TPAMI*, 2014.
Neal. Slice Sampling. *Annals of Statistics*, 2003.
Roberts, Gelman & Wilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Annals of Applied Probability*, 1997.

This work was supported in part by grants EP/F500385/1 and BB/F529254/1 for the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational Neuroscience (www.anc.ac.uk/dtc) from the UK Engineering and Physical Sciences Research Council (EPSRC), UK Biotechnology and Biological Sciences Research Council (BBSRC), and the UK Medical Research Council (MRC).