Introduction to Markov Chain Monte Carlo

Introduction

- What is Markov chain Monte Carlo (MCMC)?
  - Run an ergodic Markov chain with invariant distribution π,
  - Use sample averages from this Markov chain to compute expectations
- Contributed to the success of Bayesian Statistics
  - and is still very popular with practioners in many applications

Introduction

- Long and classic literature
  - N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller 53
  - Hastings 71, Geman and Geman 84, Gelfand and Smith 90, Tierney 94, ...
- Very popular topic with theorists from Applied Probability
  - very elegant and interesting theory: Markov chains on general state spaces

Outline

- Introduction to Markov chain Monte Carlo (MCMC)
  - Basic principle, and Metropolis Hastings algorithm
- Different implementations of MCMC
  - MH with RW, MALA, pCN
  - Gibbs Sampling
- Assessing the performance of MCMC

Setting

- Consider an arbitrary distribution on $\mathcal{X}$ with a density π w.r.t to dx, such that

$$\pi(dx) = \frac{\gamma(x)}{Z} dx$$

and is Z **unknown.**

- We want to compute expectations:

$$\pi(\varphi) = \int_{\mathcal{X}} \varphi(x)\pi(dx)$$

here $\varphi : \mathcal{X} \longrightarrow \mathbb{R}^{n_x}$, examples: $\varphi = x^n$, $\varphi = 1_A$,...

Bayesian Inference

- Bayesian inference
  - Parameter X is a random variable and Y is some dataset
  - Bayes rule: **posterior**$\propto$ likelihood $\times$ prior

$$p(x|y) \propto \underbrace{p(y|x)p(x)}_{\gamma(x)}$$

- MCMC simulates an appropriate ergodic Markov chain $(X_k)_{k\geq 0}$ with stationary distribution $p(x|y)dx$

Introduction to MCMC

- lets say we have access to
  - a Markov Probability kernel K such that $\pi K = \pi$, i.e.

$$\int \pi(dx)K(x, dy) = \pi(dy)$$

  - an initial distribution ν (possibly δ_x)
- MCMC sampling procedure:

$$X_0 \sim \nu, X_1 \sim K(X_0, \cdot), X_2 \sim K(X_1, \cdot), \ldots, X_N \sim K(X_{N-1}, \cdot), \ldots$$

- Approximation:

$$\widehat{\pi}(\varphi) = \frac{1}{N} \sum_{i=1}^{N} \varphi(X_i)$$

  mixing of chain is important as we want algorithm to converge fast

Introduction to MCMC

- What principles does it make sense to invoke for $\widehat{\pi}(\varphi)$?

1. convergence of K^n to π in some sense (e.g. L^2, total variation norm, Wasserstein distance,...)
2. SLLN $\widehat{\pi}(\varphi) \rightarrow_{N\rightarrow\infty} \pi(\varphi)$ for $\varphi \in L^1(\pi)$
3. CLT for $\sqrt{N}(\widehat{\pi}(\varphi) - \pi(\varphi)) \rightarrow \mathcal{N}(0, \sigma^2)$, $\varphi \in L^2(\pi)$,
   - 3.1 CLT variance useful to characterise asymptotic sampling error in $\widehat{\pi}(\varphi)$
   - 3.2 can be used to derive measure of Effective Sample Size

Introduction to MCMC theory

- Above clearly relate with theory of **Markov Chains in general spaces**
  - (...,Revuz 75, Nummelin 84, Kipnis & Varhadhan 86, Meyn & Tweedie 92, ...)
- Given K and x_0, one typically checks
  - π is unique invariant distribution
  - irreducibility, aperiodicity, reversibility
- Rates of convergence of $\widehat{\pi}(\varphi)$
  - uniform or geometric ergodicity
- Significant MCMC theory relate with tuning of K in various contexts
  - e.g. diffusive limits of Roberts et. al.

Introduction to MCMC

- On faces the following issues:

1. need to design K
   1.1 this might involve certain tuning parameters
2. we start from initial distribution ν (and not π)
   2.1 so it might take a while for chain to reach stationarity
3. we would like to get quick convergence to π
   3.1 the speed of which can depend on ν or
   3.2 at stationarity we can safely treat samples from K as approx. samples of π

An incomplete list of MCMC designs

- Many MCMC approaches:
  - Metropolis-Hastings,
  - Gibbs sampling,
  - Langevin diffusions,
  - independence sampler,
  - Metropolis within Gibbs,
  - reversible jump MCMC
  - hybrid (or Hamiltonian) Monte Carlo,
  - Multiple try Metropolis sampling
  - random scan Gibbs sampler
  - slice sampler
  - simulated tempering
  - simulated annealing
  - pseudo-marginal MCMC (particle MCMC)
  - ABC-MCMC
  - delayed acceptance MCMC
  - pre-conditioned Crank-Nicholson MCMC
  - piecewise deterministic MCMC
  - ....

Metropolis Hastings (MH)

Sample $X_0 \sim \nu$.
For $n \geq 1$

1. Sample a candidate proposal: $Y_n \sim Q(X_{n-1}, \cdot)$
2. Compute acceptance ratio

$$\alpha(X_{n-1}, Y_n) = 1 \wedge \frac{\gamma(Y_n)Q(Y_n, X_{n-1})}{\gamma(X_{n-1})Q(X_{n-1}, Y_n)}$$

3. With probability $\alpha(X_{n-1}, Y_n)$
   - Set $X_n = Y_n$ **(accept proposal)**
4. otherwise (with probability $1 - \alpha(X_{n-1}, Y_n)$)
   - Set $X_n = X_{n-1}$ **(reject sample)**

On the construction of MH

- Design of Q is crucial
  - determines at each step how far each exploration goes
  - shapes acceptance ratio
  - hence determines convergence speed and efficiency of MCMC
- We will assume that Q has a density w.r.t dx
  - but one can write more general formulations as in (Tierney 98)

Understanding MCMC

- Resulting Markov transition kernel:

$$K(x, dy) = \alpha(x, y)Q(x, dy) + \delta_x(dy) \int (1 - \alpha(x, y))Q(x, dy)$$

- The aim here is to deduce **reversibility** of K with π, i.e.

$$\pi(dx)K(x, dy) = \pi(dy)K(y, dx)$$

- Reversibility implies

$$\pi K = \pi$$

- This is done by ensuring

$$\pi(dx)Q(x, dy)\alpha(x, y) = \pi(dy)Q(y, dx)\alpha(y, x)$$

On the acceptance probability

- Taking densities w.r.t dx: let $dQ = q dx$

$$\alpha(x,y) = \begin{cases} 1 \wedge \frac{\gamma(y)q(y,x)}{\gamma(x)q(x,y)} & \gamma(x)q(x,y) > 0 \\ 0 & \gamma(x)q(x,y) = 0 \end{cases}$$

- To see reversibility:

$$\begin{aligned} \pi(dx)Q(x,dy)\alpha(x,y) &= \frac{1}{Z}\gamma(x)q(x,y)\alpha(x,y)dxdy \\ &= \frac{1}{Z}\gamma(x)q(x,y)\left[1 \wedge \frac{\gamma(y)q(y,x)}{\gamma(x)q(x,y)}\right] dxdy \\ &= \frac{1}{Z}\left[\gamma(x)q(x,y) \wedge \gamma(y)q(y,x)\right] dxdy \end{aligned}$$

- Similarly $\pi(dy)Q(y,dx)\alpha(y,x)$ gives the same expression.

Metropolis Hastings (MH): implementation

- Steps 3. & 4. can be implemented as:
  - $U_n \sim U[0,1)$
  - if $U_n < \alpha(X_{n-1}, Y_n)$ accept: $X_n = Y_n$
  - else reject: $X_n = X_{n-1}$
- Often a small (but significant) number of samples (of order 10^2 to 10^3 or more some cases) is not taken into account when computing estimates, so

$$\widehat{\pi}(\varphi) = \frac{1}{N - M + 1} \sum_{i=M}^{N} \varphi(X_i)$$

  - This is called burn-in: wait to allow for the chain to reach stationarity.
- How big usually should N be?
  - from 10^4 to 10^6 depending on the problem, dimensionality, etc.

Designing proposals: random walk proposal

- Most common choice for Q is random walk (RW)

$$Y_n = X_{n-1} + \varrho Z_n, \quad Z_n \sim h(\cdot)$$

- if h is symmetric $q(x, y) = q(y, x)$ then

$$\alpha(x, y) = 1 \wedge \frac{\gamma(y)}{\gamma(x)}$$

- Very often a Gaussian random walk is used: $h = \mathcal{N}(0, I)$
- ϱ is a tuning parameter for step size
  - what values should one aim for $\alpha(X_{n-1}, Y_n)$? for MH around 0.2-0.3.

RW-MCMC in practice

- Compute posterior $p(x|y)$ for the following model
  - Prior $x \sim IG(a, b)$, $a, b = 1$,
  - Observations $Y_i \sim \mathcal{N}(0, X)$, $i = 1, \ldots, 5$.
- Consider a Gaussian RW proposal
  - with $\varrho = 0.01, 0.07, 10, 30, 50$
- Resulting average acceptance ratio:
  $0.9881, 0.9352, 0.1229, 0.0299, 0.0134$

RW-MCMC in practice: trace plots

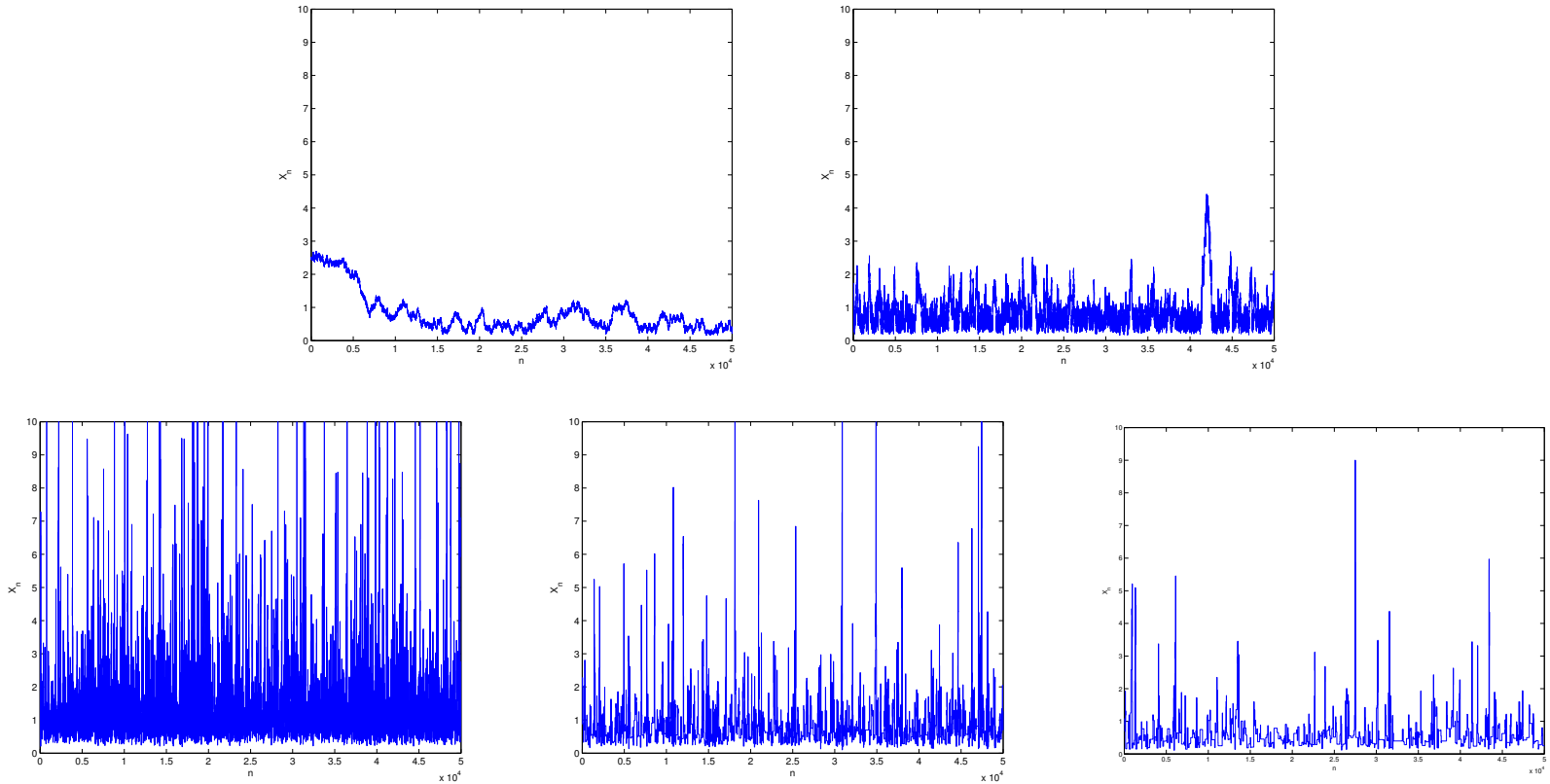

Figure: Trace plots $\varrho = 0.01, 0.07, 10, 30, 50$

RW-MCMC in practice: trace plots

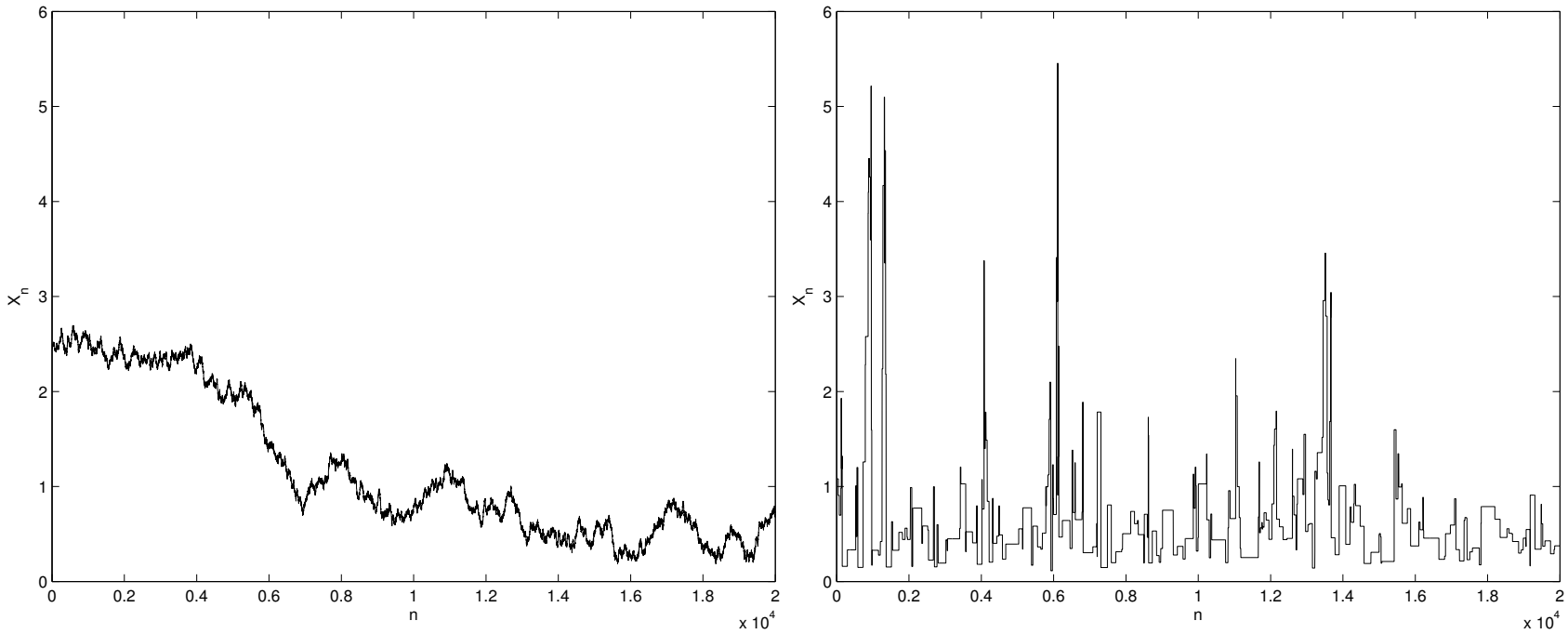

Figure: Trace plots $\varrho = 0.01, 50$

RW-MCMC in practice: trace plots

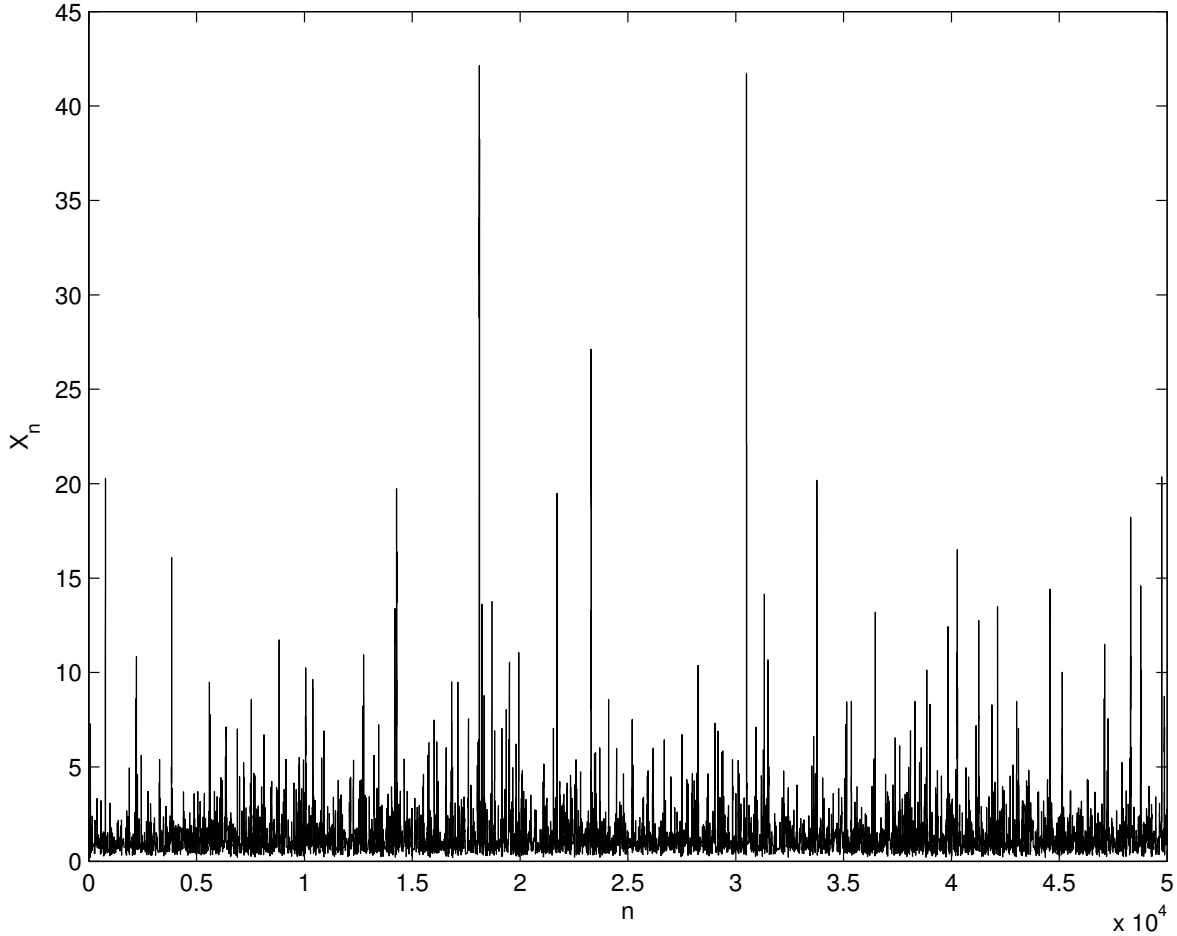

Figure: Trace plots $\varrho = 10$

RW-MCMC in practice: trace plots

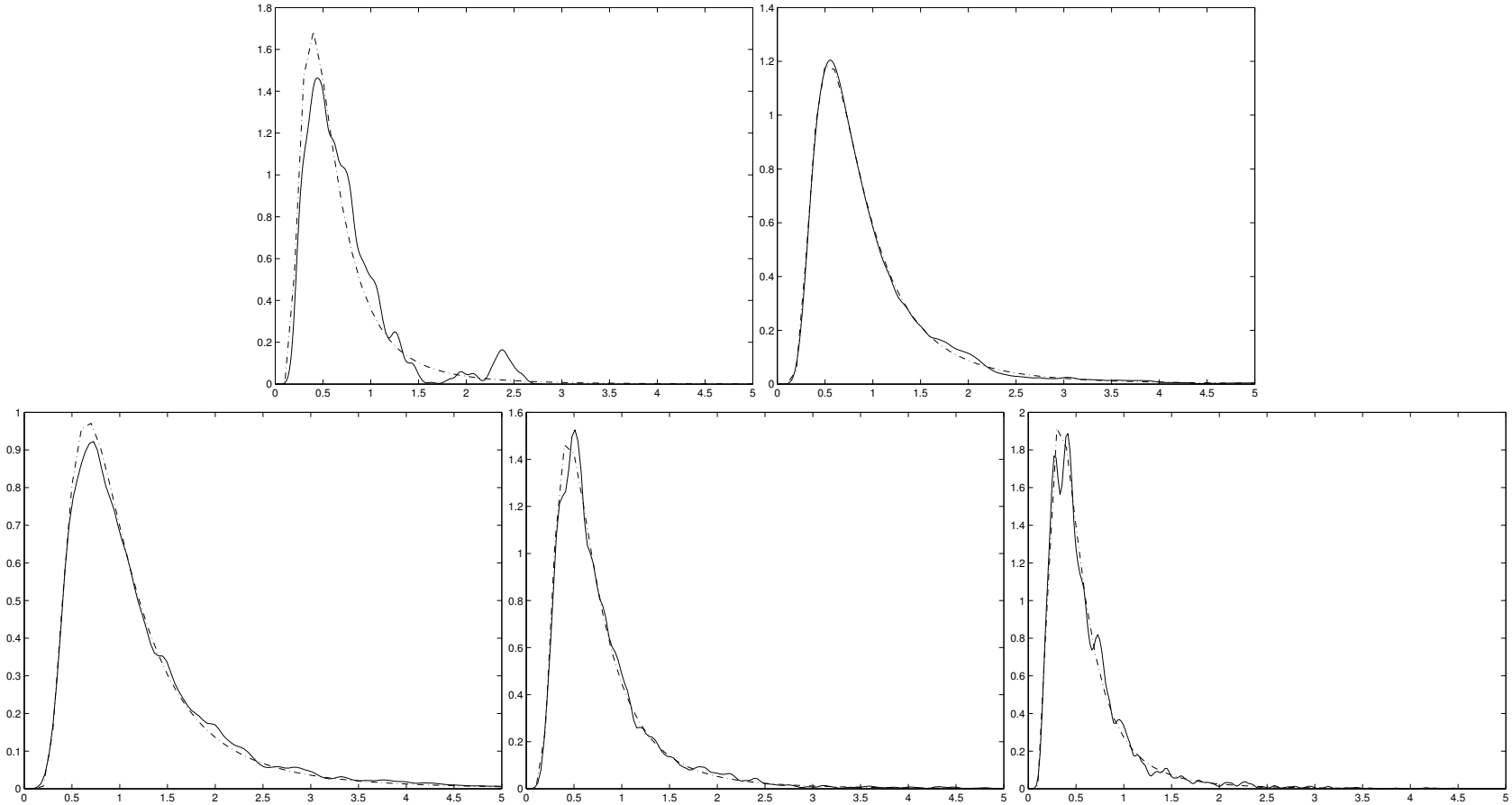

Figure: Trace plots $\varrho = 0.01, 0.07, 10, 30, 50$

RW-MCMC in practice: trace plots

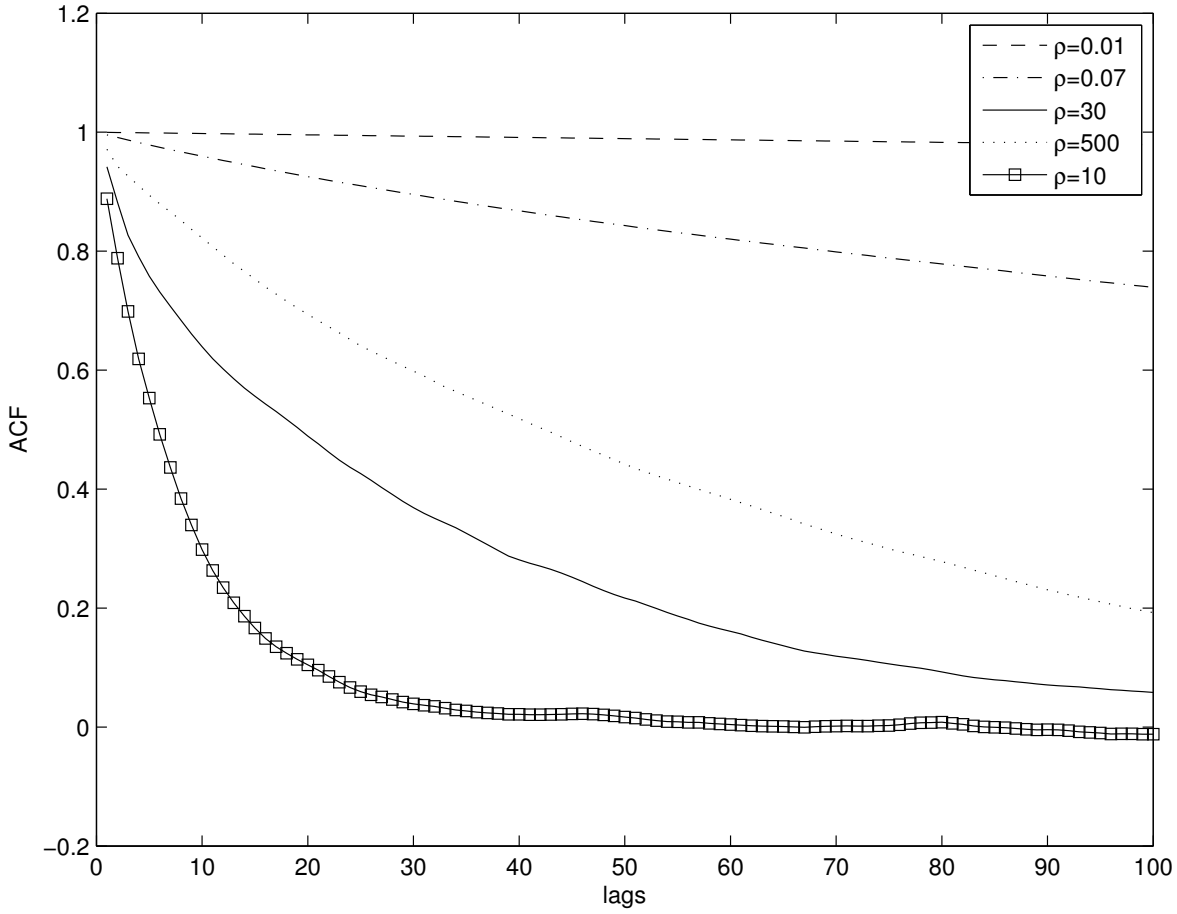

Figure: Autocorrelation function (ACF) for different lags

Metropolis adjusted Langevin algorithm (MALA)

- Let q come directly from a discretised Langevin diffusion

$$Y_n = X_{n-1} + \delta \frac{1}{2} \nabla \log \pi(X_{n-1}) + \sqrt{\delta} Z_n, \quad Z_n \sim \mathcal{N}(0, I)$$

- Note here q is **not** symmetric
- Metropolis accept reject step
  - corrects the discretisation bias
  - restores reversibility with π

Independence Sampler

- Q does need to be a Markov kernel.
- **Independence sampler:**
  - Replace step 1. in MH with $Y_n \sim Q(\cdot)$ and step 2. with

$$\alpha(X_{n-1}, Y_n) = 1 \wedge \frac{\gamma(Y_n)Q(X_{n-1})}{\gamma(X_{n-1})Q(Y_n)}$$

- Similarly to Importance and Rejection sampling, in this case Q needs to be similar to π for this algorithm to accept often
  - if accepted one ends with a fresh new sample
- Uniformly ergodic when $\gamma(x)/q(x) < \infty$

Auto-regressive proposals

- Let $\frac{d\pi}{d\lambda}(x) = \vartheta(x)$ and assume $\lambda Q = \lambda$
  - Keep step 1. as in MH with $Y_n \sim Q(X_{n-1}, \cdot)$ and
  - Replace step 2. with

$$\alpha(X_{n-1}, Y_n) = 1 \wedge \frac{\vartheta(Y_n)}{\vartheta(X_{n-1})}$$

- Q is invariant to prior λ so α corrects for likelihood only
  - useful for high dimensional x

Auto-regressive proposals

- When $\lambda = N(\mu, S)$ this is implemented as

$$Y_n = \mu + \rho\,(X_{n-1} - \mu) + (1 - \rho^2)^{\frac{1}{2}} Z, \quad Z \sim \mathcal{N}(0, S)$$

  with $\rho \in (0, 1)$
- recent name: pre conditioned Crank Nicolson sampler
  - very popular in high dimensional inverse problems for PDEs/ODEs
  - λ can be infinitely dimensional Gaussian
  - (Stuart, Law, Cotter et. al.)

The Gibbs Sampler

- Let $x = (x^1, \ldots, x^d)$ and $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_d$
- Define also $x^{-i} = (x^1, \ldots, x^{i-1}, x^{i+1}, \ldots, x^d)$ and assume that the **Gibbs full conditional distributions**

$$\pi_i(x^i | x^{-i}) = \frac{\gamma(x)}{\int \gamma(x) dx^i} = \frac{\pi(x)}{\pi_i(x^{-i})}$$

  are available.
- A Gibbs sampler samples iteratively from each conditional distribution

The Gibbs Sampler

Sample $X_0 \sim \nu$.

For $n \geq 1$. Sample:

- $X_n^1 \sim \pi_1(\cdot | X_{n-1}^{-1})$
- $X_n^2 \sim \pi_2(\cdot | X_n^1, X_{n-1}^3, \ldots, X_{n-1}^d)$
- $X_n^3 \sim \pi_3(\cdot | X_n^1, X_n^2, X_{n-1}^4, \ldots, X_{n-1}^d)$

  $\vdots$

- $X_n^d \sim \pi_d(\cdot | X_n^{-d})$

Understanding the Gibbs Sampler

- Sampler is only π-invariant and **not** reversible w.r.t. π
- **Only** each **conditional move** is reversible to π

$$\begin{aligned}
\pi(x)q_i(x,y) &= \pi(x)\pi_i(y^i|y^{-i})1_{x^{-i}=y^{-i}} \\
&= \frac{\pi(x)\pi(y)}{\pi_i(y^{-i})}1_{x^{-i}=y^{-i}} \\
&= \frac{\pi(x)\pi(y)}{\pi_i(x^{-i})}1_{x^{-i}=y^{-i}} \\
&= \pi(y)q_i(y,x)
\end{aligned}$$

Understanding the Gibbs Sampler

- Each move q_i acts on $\mathcal{X}$, is hence π invariant
  - so composition of all moves means $K = q_1 q_2 \cdots q_d$ is also π invariant.
- Every coordinate moves so can have irreducibility
- No tuning parameters, so method can be quite efficient
  - **if** a good parameterisation of x in terms of π is chosen so that sampler explores nicely areas where π is high

Gibbs sampler in practice

- Compute $\pi = \mathcal{N}(\mu, \Sigma)$ with $\mu = [\mu_1, \mu_2]$, $\quad \Sigma = \begin{bmatrix} 1 & \rho_{12} \\ \rho_{21} & 1 \end{bmatrix}$
- Gibbs conditionals:

$$\pi_1(x_1|x_2) = \mathcal{N}(\mu_1 + \rho_{21}(x_2 - \mu_2), \sqrt{1 - \rho_{21}^2})$$

and

$$\pi_2(x_2|x_1) = \mathcal{N}(\mu_2 + \rho_{12}(x_1 - \mu_1), \sqrt{1 - \rho_{12}^2}),$$

- Lets look at example

$$\mu = [0, 0], \quad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

- Lets take 5000 samples (a bit low so in principle sample more)

Gibbs sampler in practice

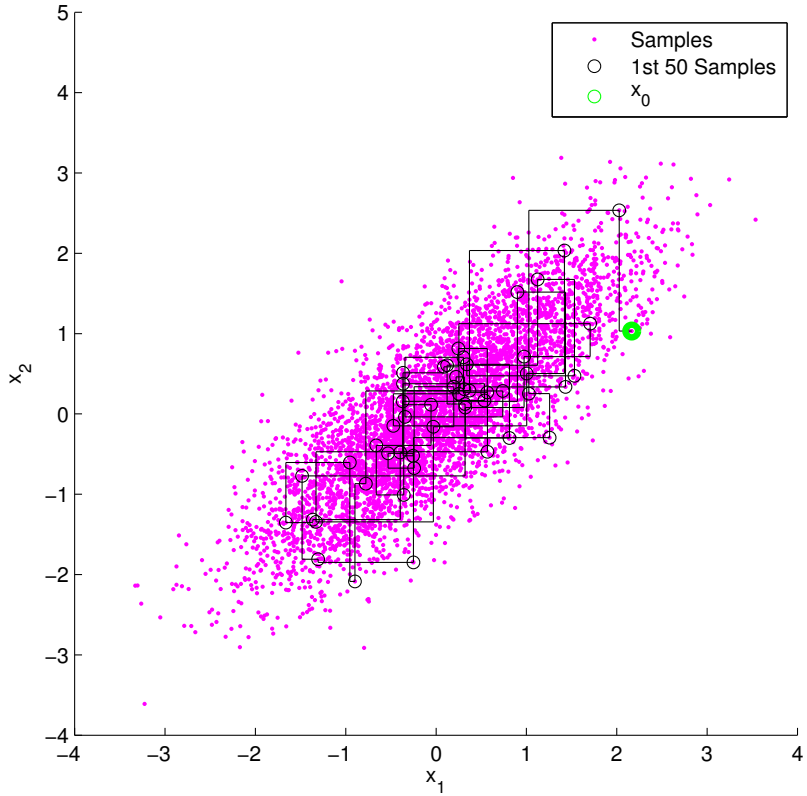

Figure: Scatter plot and illustration of Gibbs trajectory

Gibbs sampler in practice

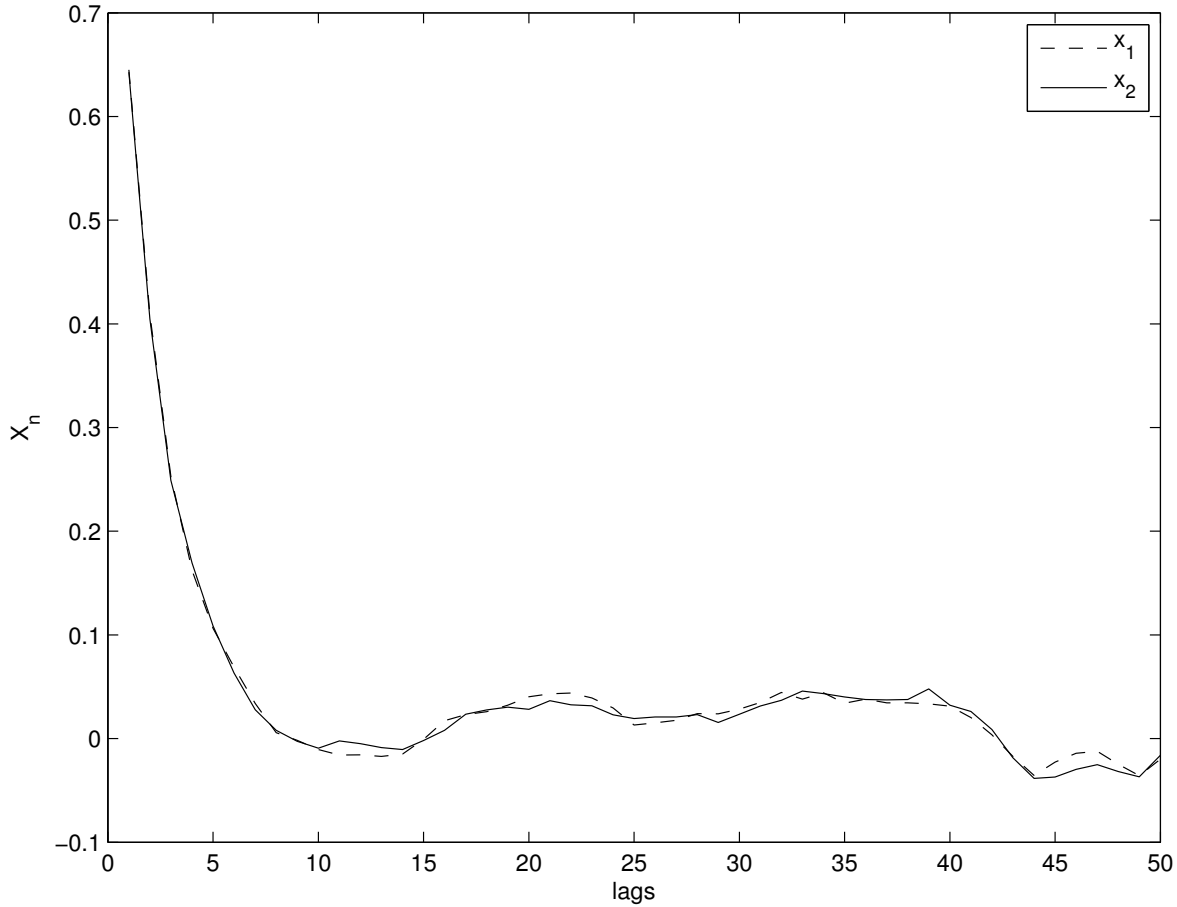

Figure: ACFs

Gibbs sampler in practice

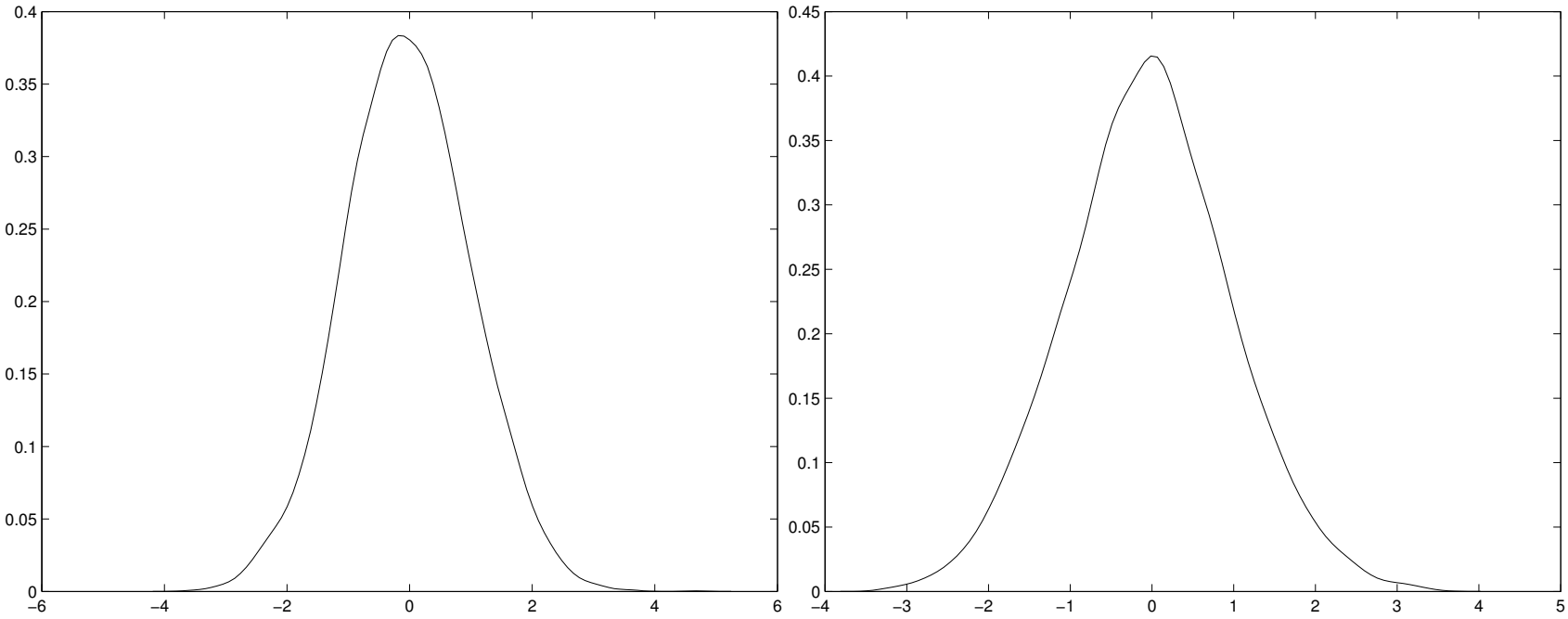

Figure: Kernel density estimates of pdfs

The Gibbs Sampler: extensions

- Metropolis Within Gibbs
  - What happens when cannot sample from π_i directly?
  - use MH accept reject step for each π_i instead
- Random Scan Gibbs
  - can randomly permute order of steps
  - can pick randomly i and then sample from π_i
- Collapsed Gibbs sampler
  - often one does not need full conditional
  - can marginalises over one or more variables when sampling for some other variable x^i
  - care needs to be taken so that all moves are invariant to π

Measuring efficiency using autocorrelations

- Define the **integrated auto-correlation time** for φ as

$$\begin{aligned}\tau_\varphi &= \frac{v(\varphi, K)}{\mathbb{V}ar_\pi[\varphi]} \\ &= 1 + 2\sum_{i \geq 1} Cor[\varphi(X_0), \varphi(X_i)]\end{aligned}$$

  with $v(\varphi, K)$ being an asymptotic CLT variance
- Lower τ_φ means higher efficiency
  - faster mixing and decorrelation reduces Monte Carlo variance
  - loosely speaking samples τ_φ apart are "closer to independent".
- It makes sense to look at ACF to assess mixing

Simple diagnostics of MCMC

- So far diagnostics we have seen
  - trace plot
  - ACF
  - $ESS = \frac{N}{\tau_\varphi}$
  - Acceptance ratio
  - Expected square jumping distance. Use samples to approximate
    $$ESJD = E\left[(X_n - X_{n-1})^2\right]$$
    i.e. just look at first order correlation and linear test functions
- More elaborate diagnostics exist (e.g. Robert & Casella Ch. 8) but often these test whether MCMC samples can be viewed as samples of π
  - often these require multiple chains

Discussion

- MCMC is a very powerful and simple algorithm
- Some weaknesses
  - it is quite hard to parallelise
  - it can often struggle in multimodal settings, and it might take long time to switch between modes.
    - could use simulated tempering and multiple chains
  - iterative/batch method: If new data arrives need to re-run everything again
  - tricky to estimate normalising constant Z
- This is often motivation for using Sequential Monte Carlo
  - note MCMC is still extremely valuable
  - can use MCMC within SMC and SMC within MCMC