Bayesian Parameter Inference in State-Space Models using Particle MCMC

Arnaud Doucet
Department of Statistics, Oxford University

University College London

5th October 2012
In most scenarios of interest, the state-space model contains an unknown static parameter $\theta \in \Theta$ so that

$$X_1 \sim \mu_\theta (\cdot) \text{ and } X_t \mid (X_{t-1} = x) \sim f_\theta (\cdot \mid x_{t-1}).$$
In most scenarios of interest, the state-space model contains an unknown static parameter $\theta \in \Theta$ so that

$$X_1 \sim \mu_{\theta} (\cdot) \quad \text{and} \quad X_t | (X_{t-1} = x) \sim f_{\theta} (\cdot | x_{t-1}).$$

The observations $\{Y_t\}_{t \geq 1}$ are conditionally independent given $\{X_t\}_{t \geq 1}$ and $\theta$

$$Y_t | (X_t = x_t) \sim g_{\theta} (\cdot | x).$$
In most scenarios of interest, the state-space model contains an unknown static parameter $\theta \in \Theta$ so that

$$X_1 \sim \mu_{\theta}(\cdot) \quad \text{and} \quad X_t \mid (X_{t-1} = x) \sim f_{\theta}(\cdot \mid x_{t-1}).$$

The observations $\{Y_t\}_{t \geq 1}$ are conditionally independent given $\{X_t\}_{t \geq 1}$ and $\theta$

$$Y_t \mid (X_t = x_t) \sim g_{\theta}(\cdot \mid x).$$

**Aim:** Given observations $y_1:T$, we want to infer $\theta$ in a Bayesian framework.
Bayesian Parameter Inference in State-Space Models

Set a prior $p(\theta)$ on $\theta$ so inference relies now on

$$p(\theta, x_{1:T} | y_{1:T}) = \frac{p(\theta, x_{1:T}, y_{1:T})}{p(y_{1:t})}$$

where

$$p(\theta, x_{1:T}, y_{1:T}) = p(\theta) p_\theta(x_{1:T}, y_{1:T})$$

with

$$p_\theta(x_{1:T}, y_{1:T}) = \mu_\theta(x_1) \prod_{k=2}^{T} f_\theta(x_k | x_{k-1}) \prod_{k=1}^{T} g_\theta(y_k | x_k)$$
Set a prior $p(\theta)$ on $\theta$ so inference relies now on

$$p(\theta, x_1:T \mid y_1:T) = \frac{p(\theta, x_1:T, y_1:T)}{p(y_1:T)}$$

where

$$p(\theta, x_1:T, y_1:T) = p(\theta) p_{\theta}(x_1:T, y_1:T)$$

with

$$p_{\theta}(x_1:T, y_1:T) = \mu_{\theta}(x_1) \prod_{k=2}^{T} f_{\theta}(x_k \mid x_{k-1}) \prod_{k=1}^{T} g_{\theta}(y_k \mid x_k)$$

Standard approaches rely on MCMC.
**MCMC Idea:** Simulate an ergodic Markov chain \( \{ \theta (i), X_{1:T}(i) \}_{i \geq 0} \) of invariant distribution \( p ( \theta, x_{1:T} | y_{1:T} ) \)... infinite number of possibilities.
**MCMC Idea:** Simulate an ergodic Markov chain $\{\theta (i), X_{1:T} (i)\}_{i \geq 0}$ of invariant distribution $p (\theta, x_{1:T} | y_{1:T})$... infinite number of possibilities.

Typical strategies consists of updating iteratively $X_{1:T}$ conditional upon $\theta$ then $\theta$ conditional upon $X_{1:T}$.
Common MCMC Approaches and Limitations

- **MCMC Idea**: Simulate an ergodic Markov chain \( \{\theta(i), X_{1:T}(i)\}_{i \geq 0} \) of invariant distribution \( p(\theta, X_{1:T} | y_{1:T}) \)...
  infinite number of possibilities.

- Typical strategies consists of updating iteratively \( X_{1:T} \) conditional upon \( \theta \) then \( \theta \) conditional upon \( X_{1:T} \).

- To update \( X_{1:T} \) conditional upon \( \theta \), use MCMC kernels updating subblocks according to \( p_\theta \left( x_{t:t+K-1} | y_{t:t+K-1}, x_{t-1}, x_{t+K} \right) \).
Common MCMC Approaches and Limitations

- **MCMC Idea**: Simulate an ergodic Markov chain \( \{ \theta (i), X_{1:T} (i) \}_{i \geq 0} \) of invariant distribution \( p(\theta, X_{1:T} \mid y_{1:T}) \) ... infinite number of possibilities.

- Typical strategies consists of updating iteratively \( X_{1:T} \) conditional upon \( \theta \) then \( \theta \) conditional upon \( X_{1:T} \).

- To update \( X_{1:T} \) conditional upon \( \theta \), use MCMC kernels updating subblocks according to \( p_{\theta} ( x_{t:t+K-1} \mid y_{t:t+K-1}, x_{t-1}, x_{t+K} ) \).

- Standard MCMC algorithms are inefficient if \( \theta \) and \( X_{1:T} \) are strongly correlated.
**MCMC Idea**: Simulate an ergodic Markov chain \( \{ \theta (i) , X_{1:T} (i) \}_{i \geq 0} \) of invariant distribution \( p ( \theta , X_{1:T} \mid y_{1:T} ) \)... infinite number of possibilities.

Typical strategies consists of updating iteratively \( X_{1:T} \) conditional upon \( \theta \) then \( \theta \) conditional upon \( X_{1:T} \).

To update \( X_{1:T} \) conditional upon \( \theta \), use MCMC kernels updating subblocks according to \( p_\theta ( x_{t:t+K-1} \mid y_{t:t+K-1} , x_{t-1} , x_{t+K} ) \).

Standard MCMC algorithms are inefficient if \( \theta \) and \( X_{1:T} \) are strongly correlated.

Strategy impossible to implement when it is only possible to sample from the prior but impossible to evaluate it pointwise.
To bypass these problems, we want to update jointly $\theta$ and $X_{1:T}$. 
To bypass these problems, we want to update jointly $\theta$ and $X_{1:T}$.

Assume that the current state of our Markov chain is $(\theta, x_{1:T})$, we propose to update simultaneously the parameter and the states using a proposal

$$q((\theta^*, x_{1:T}^*) | (\theta, x_{1:T})) = q(\theta^* | \theta) \cdot q_{\theta^*}(x_{1:T}^* | y_{1:T}).$$
Metropolis-Hastings (MH) Sampling

- To bypass these problems, we want to update jointly $\theta$ and $X_{1:T}$.
- Assume that the current state of our Markov chain is $(\theta, x_{1:T})$, we propose to update simultaneously the parameter and the states using a proposal

$$q((\theta^*, x_{1:T}^*) | (\theta, x_{1:T})) = q(\theta^* | \theta) \ q_{\theta^*}(x_{1:T}^* | y_{1:T}).$$

- The proposal $(\theta^*, x_{1:T}^*)$ is accepted with MH acceptance probability

$$1 \wedge \frac{p(\theta^*, x_{1:T}^* | y_{1:T}) \ q((x_{1:T}, \theta) | (x_{1:T}^*, \theta^*))}{p(\theta, x_{1:T} | y_{1:T}) \ q((x_{1:T}^*, \theta^*) | (x_{1:T}, \theta))}.$$
To bypass these problems, we want to update jointly $\theta$ and $X_{1:T}$.

Assume that the current state of our Markov chain is $(\theta, x_{1:T})$, we propose to update simultaneously the parameter and the states using a proposal

$$q((\theta^*, x_{1:T}^*) | (\theta, x_{1:T})) = q(\theta^* | \theta) \cdot q_{\theta^*}(x_{1:T}^* | y_{1:T}).$$

The proposal $(\theta^*, x_{1:T}^*)$ is accepted with MH acceptance probability

$$1 \wedge \frac{p(\theta^*, x_{1:T}^* | y_{1:T}) \cdot q((x_{1:T}, \theta) | (x_{1:T}^*, \theta^*))}{p(\theta, x_{1:T} | y_{1:T}) \cdot q((x_{1:T}^*, \theta^*) | (x_{1:T}, \theta))}.$$ 

**Problem**: Designing a proposal $q_{\theta^*}(x_{1:T}^* | y_{1:T})$ such that the acceptance probability is not extremely small is very difficult.
Consider the following so-called marginal Metropolis-Hastings (MH) algorithm which uses as a proposal

$$q ((x^*_1:T, \theta^*) | (x_1:T, \theta)) = q (\theta^* | \theta) p_{\theta^*} (x^*_1:T | y_1:T).$$
Consider the following so-called marginal Metropolis-Hastings (MH) algorithm which uses as a proposal

\[
q \left( (x_{1:T}^*, \theta^*) \mid (x_{1:T}, \theta) \right) = q(\theta^* \mid \theta) \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}.
\]

The MH acceptance probability is

\[
1 \wedge \frac{p(\theta^*, x_{1:T}^* \mid y_{1:T})}{p(\theta, x_{1:T} \mid y_{1:T})} \frac{q(\theta^* \mid (x_{1:T}^*, \theta^*))}{q(\theta \mid (x_{1:T}^*, \theta^*))} = 1 \wedge \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}.
\]
Consider the following so-called marginal Metropolis-Hastings (MH) algorithm which uses as a proposal

\[ q((x^*_1:T, \theta^*)| (x_1:T, \theta)) = q(\theta^*| \theta) p_{\theta^*} (x^*_1:T | y_{1:T}). \]

The MH acceptance probability is

\[
1 \wedge \frac{p(\theta^*, x^*_1:T | y_{1:T})}{p(\theta, x_1:T | y_{1:T})} \frac{q((x_1:T, \theta)| (x^*_1:T, \theta^*))}{q((x^*_1:T, \theta^*)| (x_1:T, \theta))} \\
= 1 \wedge \frac{p_{\theta^*} (y_{1:T})}{p_{\theta} (y_{1:T})} \frac{p(\theta^*)}{p(\theta)} \frac{q(\theta| \theta^*)}{q(\theta^*| \theta)}
\]

In this MH algorithm, \( X_{1:T} \) has been essentially integrated out.
**Problem 1:** We do not know $p_\theta (y_{1:T}) = \int p_\theta (x_{1:T}, y_{1:T}) \, dx_{1:T}$ analytically.
Implementation Issues

- **Problem 1**: We do not know $p_\theta (y_{1:T}) = \int p_\theta (x_{1:T}, y_{1:T}) \, dx_{1:T}$ analytically.

- **Problem 2**: We do not know how to sample from $p_\theta (x_{1:T} \mid y_{1:T})$. 

"Idea": Use particle approximations of $p_\theta (x_{1:T}, y_{1:T})$ and $p_\theta (y_{1:T})$. 

A. Doucet (UCL Masterclass Oct. 2012)
**Problem 1**: We do not know \( p_{\theta} (y_{1:T}) = \int p_{\theta} (x_{1:T}, y_{1:T}) \, dx_{1:T} \) analytically.

**Problem 2**: We do not know how to sample from \( p_{\theta} (x_{1:T} \mid y_{1:T}) \).

**“Idea”**: Use particle approximations of \( p_{\theta} (x_{1:T} \mid y_{1:T}) \) and \( p_{\theta} (y_{1:T}) \).
Given \( \theta \), particle methods provide approximations of \( p_\theta (x_{1:T} \mid y_{1:T}) \) and \( p_\theta (y_{1:T}) \).
Given $\theta$, particle methods provide approximations of $p_\theta (x_{1:T} \mid y_{1:T})$ and $p_\theta (y_{1:T})$.

At time $T$, we obtain the following approximation of the posterior of interest

$$
\hat{p}_\theta (x_{1:T} \mid y_{1:T}) = \frac{1}{N} \sum_{k=1}^{N} \delta_{x_{1:T}^{(k)}} (x_{1:T})
$$

and an approximation of $p_\theta (y_{1:T})$ is given by

$$
\hat{p}_\theta (y_{1:T}) = \hat{p}_\theta (y_1) \prod_{t=2}^{T} \hat{p}_\theta (y_t \mid y_{1:t-1}) = \prod_{t=1}^{T} \left( \frac{1}{N} \sum_{k=1}^{N} g_\theta (y_t \mid X_t^{(k)}) \right)
$$

if we use $f_\theta (x_t \mid x_{t-1})$ as a proposal.
A Few Theoretical Results

We have

$$\mathbb{E} \left[ \hat{p}_\theta(y_{1:T}) \right] = p_\theta(y_{1:T}).$$

Under mixing assumptions, we have

$$V \left[ \hat{p}_\theta(y_{1:T}) \right] p_2 \theta(y_{1:T}) D_{\theta} T_{\mathbb{N}}.$$
A Few Theoretical Results

- We have
  \[ \mathbb{E} [\hat{p}_\theta (y_{1:T})] = p_\theta (y_{1:T}). \]

- Under *mixing assumptions*, we have
  \[ \frac{\nabla [\hat{p}_\theta (y_{1:T})]}{p^2_\theta (y_{1:T})} \leq D_\theta \frac{T}{N}. \]
A Few Theoretical Results

- We have
  \[ \mathbb{E} [ \hat{p}_\theta (y_{1:T}) ] = p_\theta (y_{1:T}) . \]

- Under \textit{mixing assumptions}, we have
  \[ \frac{\nabla [\hat{p}_\theta (y_{1:T})]}{p^2_\theta (y_{1:T})} \leq D_\theta \frac{T}{N} . \]

- Under \textit{mixing assumptions}, we also have
  \[ \int | \mathbb{E} [ \hat{p}_\theta (x_{1:T} \mid y_{1:T}) ] - p_\theta (x_{1:T} \mid y_{1:T}) | \, dx_{1:T} \leq C_\theta \frac{T}{N} . \]

so if I run a particle method to obtain \( \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \) then \( X_{1:T} \sim \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \), unconditionally \( X_{1:T} \sim \mathbb{E} [ \hat{p}_\theta (\cdot \mid y_{1:T}) ] \).
A Few Theoretical Results

- We have
  \[ E \left[ \hat{p}_\theta (y_{1:T}) \right] = p_\theta (y_{1:T}). \]

- Under \textit{mixing assumptions}, we have
  \[ \frac{\nabla \left[ \hat{p}_\theta (y_{1:T}) \right]}{p_\theta^2 (y_{1:T})} \leq D_\theta \frac{T}{N}. \]

- Under \textit{mixing assumptions}, we also have
  \[ \int \left| E \left[ \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \right] - p_\theta (x_{1:T} \mid y_{1:T}) \right| dx_{1:T} \leq C_\theta \frac{T}{N} \]
  so if I run a particle method to obtain \( \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \) then \( X_{1:T} \sim \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \), unconditionally \( X_{1:T} \sim E [\hat{p}_\theta (\cdot \mid y_{1:T})] \).

- \textbf{Problem}: We cannot compute analytically the particle filter proposal
  \[ q_\theta (x_{1:T} \mid y_{1:T}) = E \left[ \hat{p}_\theta (x_{1:T} \mid y_{1:T}) \right] \] as it involves an expectation w.r.t all the variables appearing in the particle algorithm...
At iteration $i$

- Sample $\theta^* \sim q(\theta | \theta(i - 1))$. 
At iteration $i$

- Sample $\theta^* \sim q(\theta|\theta(i-1))$.
- Sample $X_{1:T}^* \sim p_{\theta^*}(X_{1:T}|y_{1:T})$. 
At iteration $i$

- Sample $\theta^* \sim q(\theta | \theta(i-1))$.
- Sample $X_{1:T}^* \sim p_{\theta^*}(X_{1:T} | y_{1:T})$.
- With probability

$$1 \wedge \frac{p_{\theta^*}(y_{1:T}) p(\theta^*) q(\theta(i-1)|\theta^*)}{p_{\theta(i-1)}(y_{1:T}) p(\theta(i-1)) q(\theta^*|\theta(i-1))}$$

set $\theta(i) = \theta^*$, $X_{1:T}(i) = X_{1:T}^*$ otherwise set $\theta(i) = \theta(i-1)$, $X_{1:T}(i) = X_{1:T}(i-1)$. 
At iteration $i$

- Sample $\theta^* \sim q(\theta|\theta(i-1))$ and run an particle filter to obtain $\hat{p}_{\theta^*}(x_{1:T}|y_{1:T})$ and $\hat{p}_{\theta^*}(y_{1:T})$. 
At iteration $i$

- Sample $\theta^* \sim q(\theta|\theta(i-1))$ and run an particle filter to obtain $\hat{p}_{\theta^*}(x_{1:T}|y_{1:T})$ and $\hat{p}_{\theta^*}(y_{1:T})$.
- Sample $X^*_{1:T} \sim \hat{p}_{\theta^*}(x_{1:T}|y_{1:T})$. 
At iteration $i$

- Sample $\theta^* \sim q(\theta|\theta(i-1))$ and run an particle filter to obtain $\hat{p}_{\theta^*}(x_{1:T}|y_{1:T})$ and $\hat{p}_{\theta^*}(y_{1:T})$.
- Sample $X_{1:T}^* \sim \hat{p}_{\theta^*}(x_{1:T}|y_{1:T})$.
- With probability

$$1 \wedge \frac{\hat{p}_{\theta^*}(y_{1:T}) p(\theta^*)}{\hat{p}_{\theta(i-1)}(y_{1:T}) p(\theta(i-1))} \frac{q(\theta(i-1)|\theta^*)}{q(\theta^*|\theta(i-1))}$$

set $\theta(i) = \theta^*$, $X_{1:T}(i) = X_{1:T}^*$ otherwise set $\theta(i) = \theta(i-1)$, $X_{1:T}(i) = X_{1:T}(i-1)$. 

A. Doucet (UCL Masterclass Oct. 2012)
Proposition. Assume that the ‘idealized’ marginal MH sampler chain is ergodic then, under very weak assumptions, the PMMH sampler chain is ergodic and admits
\[ p(\theta, x_{1:T} | y_{1:T}) \]
whatever being \( N \geq 1 \).
**Proposition.** Assume that the ‘idealized’ marginal MH sampler chain is ergodic then, under very weak assumptions, the PMMH sampler chain is ergodic and admits $p(\theta, x_{1:T} \mid y_{1:T})$ whatever being $N \geq 1$.

It is easy to show the simpler result that the PMMH admits $p(\theta \mid y_{1:T})$ as invariant distribution whatever being $N \geq 1$. 
Proposition. Assume that the ‘idealized’ marginal MH sampler chain is ergodic then, under very weak assumptions, the PMMH sampler chain is ergodic and admits $p(\theta, x_{1:T} | y_{1:T})$ whatever being $N \geq 1$.

It is easy to show the simpler result that the PMMH admits $p(\theta | y_{1:T})$ as invariant distribution whatever being $N \geq 1$.

Let $U$ denote all the r.v. introduce to build the SMC estimate then write $\hat{p}_{\theta} (y_{1:T}) = \hat{p}_{\theta} (y_{1:T}; U)$ and from unbiasedness

$$\int \hat{p}_{\theta} (y_{1:T}; u) q_{\theta} (u) \, du = p_{\theta} (y_{1:T}).$$
The PMMH targets the distribution

$$\hat{\pi}(\theta, u) \propto p(\theta) \hat{p}_\theta(y_{1:T}; u) q_\theta(u)$$

which satisfies

$$\hat{\pi}(\theta) = p(\theta | y_{1:T}).$$
The PMMH targets the distribution
\[ \hat{\pi}(\theta, u) \propto p(\theta) \hat{p}_\theta(y_{1:T}; u) \ q_\theta(u) \]
which satisfies
\[ \hat{\pi}(\theta) = p(\theta|y_{1:T}). \]

The PMMH sampler uses as a proposal
\[ q((\theta^*, u^*)| (\theta, u)) = q(\theta^*| \theta) \ q_{\theta^*}(u^*) \]
and
\[ \frac{\hat{\pi}(\theta^*, u^*) \ q((\theta, u)| (\theta^*, u^*))}{\hat{\pi}(\theta, u) \ q((\theta^*, u^*)| (\theta, u))} = \frac{p(\theta^*) \hat{p}_{\theta^*}(y_{1:T}; u^*) \ q(\theta|\theta^*)}{p(\theta) \hat{p}_\theta(y_{1:T}; u) \ q(\theta^*| \theta)}. \]
The PMMH targets the distribution

$$\pi(\theta, u) \propto p(\theta) \hat{p}_\theta(y_{1:T}; u) q_\theta(u)$$

which satisfies

$$\pi(\theta) = p(\theta | y_{1:T}).$$

The PMMH sampler uses as a proposal

$$q((\theta^*, u^*) | (\theta, u)) = q(\theta^* | \theta) q_{\theta^*}(u^*)$$

and

$$\frac{\hat{\pi}(\theta^*, u^*) q((\theta, u) | (\theta^*, u^*))}{\hat{\pi}(\theta, u) q((\theta^*, u^*) | (\theta, u))} = \frac{p(\theta^*) \hat{p}_{\theta^*}(y_{1:T}; u^*) q(\theta | \theta^*)}{p(\theta) \hat{p}_\theta(y_{1:T}; u) q(\theta^* | \theta)}.$$

Trivial but deep result: if you plug any unbiased likelihood estimate within a MCMC scheme, you do not perturb the invariant distribution.
Two species $X_t^1$ (prey) and $X_t^2$ (predator)

\[
\begin{align*}
\Pr \left( X_{t+dt}^1 = x_t^1 + 1, X_{t+dt}^2 = x_t^2 \bigg| x_t^1, x_t^2 \right) &= \alpha x_t^1 dt + o(dt), \\
\Pr \left( X_{t+dt}^1 = x_t^1 - 1, X_{t+dt}^2 = x_t^2 + 1 \bigg| x_t^1, x_t^2 \right) &= \beta x_t^1 x_t^2 dt + o(dt), \\
\Pr \left( X_{t+dt}^1 = x_t^1, X_{t+dt}^2 = x_t^2 - 1 \bigg| x_t^1, x_t^2 \right) &= \gamma x_t^2 dt + o(dt),
\end{align*}
\]

with

\[
Y_k = X_{k\Delta T}^1 + W_k \text{ with } W_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).
\]
Two species \( X_t^1 \) (prey) and \( X_t^2 \) (predator)

\[
\Pr \left( X_{t+dt}^1 = x_{t}^1 + 1, \ X_{t+dt}^2 = x_{t}^2 \ \middle| \ x_t^1, x_t^2 \right) = \alpha \ x_t^1 \ dt + o \ (dt),
\]

\[
\Pr \left( X_{t+dt}^1 = x_{t}^1 - 1, \ X_{t+dt}^2 = x_{t}^2 + 1 \ \middle| \ x_t^1, x_t^2 \right) = \beta \ x_t^1 \ x_t^2 \ dt + o \ (dt),
\]

\[
\Pr \left( X_{t+dt}^1 = x_{t}^1, \ X_{t+dt}^2 = x_{t}^2 - 1 \ \middle| \ x_t^1, x_t^2 \right) = \gamma \ x_t^2 \ dt + o \ (dt),
\]

with

\[
Y_k = X_{k\Delta t}^1 + W_k \text{ with } W_k \overset{\text{i.i.d.}}{\sim} \mathcal{N} (0, \sigma^2).
\]

We are interested in the kinetic rate constants \( \theta = (\alpha, \beta, \gamma) \) a priori distributed as (Boys et al., 2008; Kunsch, 2011)

\[
\alpha \sim \mathcal{G} (1, 10), \quad \beta \sim \mathcal{G} (1, 0.25), \quad \gamma \sim \mathcal{G} (1, 7.5).
\]
Two species $X_t^1$ (prey) and $X_t^2$ (predator)

\[
\begin{align*}
\Pr (X_{t+dt}^1 = x_t^1 + 1, X_{t+dt}^2 = x_t^2 | x_t^1, x_t^2) &= \alpha x_t^1 dt + o(dt), \\
\Pr (X_{t+dt}^1 = x_t^1 - 1, X_{t+dt}^2 = x_t^2 + 1 | x_t^1, x_t^2) &= \beta x_t^1 x_t^2 dt + o(dt), \\
\Pr (X_{t+dt}^1 = x_t^1, X_{t+dt}^2 = x_t^2 - 1 | x_t^1, x_t^2) &= \gamma x_t^2 dt + o(dt),
\end{align*}
\]

with

\[Y_k = X_{k\Delta T}^1 + W_k \text{ with } W_k \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2).\]

We are interested in the kinetic rate constants $\theta = (\alpha, \beta, \gamma)$ a priori distributed as (Boys et al., 2008; Kunsch, 2011)

\[\alpha \sim \mathcal{G}(1, 10), \quad \beta \sim \mathcal{G}(1, 0.25), \quad \gamma \sim \mathcal{G}(1, 7.5).\]

MCMC methods require reversible jumps, Particle MCMC requires only forward simulation.
Experimental Results

Simulated data

Posterior distributions

A. Doucet (UCL Masterclass Oct. 2012)
Autocorrelation of $\alpha$ (left) and $\beta$ (right) for the PMMH sampler for various $N$.  

A. Doucet (UCL Masterclass Oct. 2012)
Offline Bayesian parameter inference is feasible by using particle filter proposals within MCMC.
Offline Bayesian parameter inference is feasible by using particle filter proposals within MCMC.

Particle MCMC allow us to perform Bayesian inference for dynamic models for which only forward simulation is possible.
Offline Bayesian parameter inference is feasible by using particle filter proposals within MCMC.

Particle MCMC allow us to perform Bayesian inference for dynamic models for which only forward simulation is possible.

Computationally intensive but several implementations on GPU already available and applications in control, ecology, econometrics, biochemical systems, epidemiology, water resources research etc.
Offline Bayesian parameter inference is feasible by using particle filter proposals within MCMC.

Particle MCMC allow us to perform Bayesian inference for dynamic models for which only forward simulation is possible.

Computationally intensive but several implementations on GPU already available and applications in control, ecology, econometrics, biochemical systems, epidemiology, water resources research etc.

This approach does not suffer from degeneracy problem and \( N \) scales roughly linearly with \( T \).
Selection of $N$ is a key issue.
Selection of $N$ is a key issue.

If $N$ is too small, then the algorithm mixes poorly and will require many MCMC iterations.
Selection of \( N \) is a key issue.

If \( N \) is too small, then the algorithm mixes poorly and will require many MCMC iterations.

If \( N \) is too large, then each MCMC iteration is expensive.
Selection of $N$ is a key issue.

If $N$ is too small, then the algorithm mixes poorly and will require many MCMC iterations.

If $N$ is too large, then each MCMC iteration is expensive.

**Aim**: We would like to provide guidelines on how to select $N$. 
Selection of $N$ is a key issue.

If $N$ is too small, then the algorithm mixes poorly and will require many MCMC iterations.

If $N$ is too large, then each MCMC iteration is expensive.

**Aim:** We would like to provide guidelines on how to select $N$.

Joint work with Mike Pitt (Warwick) and Robert Kohn (UNSW).
Let $Z = \log \hat{p}_\theta(y; U) - \log p_\theta(y)$ be the error in log-likelihood.
Let $Z = \log \hat{p}_\theta(y; U) - \log p_\theta(y)$ be the error in log-likelihood.

The proposal from which $Z$ arises is denoted $g(z|\theta)$. 
Let $Z = \log \hat{p}_\theta(y; U) - \log p_\theta(y)$ be the error in log-likelihood.

The proposal from which $Z$ arises is denoted $g(z|\theta)$.

We can rewrite the extended target

$$\hat{\pi}(\theta, z) = p(\theta|y)\exp(z)g(z|\theta)$$

which is directly related to $\hat{\pi}(\theta, u)$ through the many-to-one transformation from $u$ to $z$. 
Let $Z = \log \hat{p}_\theta (y; U) - \log p_\theta (y)$ be the error in log-likelihood.

The proposal from which $Z$ arises is denoted $g(z|\theta)$.

We can rewrite the extended target

$$\widehat{\pi}(\theta, z) = p(\theta|y) \exp(z) g(z|\theta)$$

which is directly related to $\widehat{\pi}(\theta, u)$ through the many-to-one transformation from $u$ to $z$.

The previous algorithm proposes $\theta' \sim q(\cdot|\theta)$ and $Z' \sim g(\cdot|\theta')$, accepting $(\theta', z')$ w.p.

$$\alpha_Q(\theta, Z; \theta', Z') = \min \left\{ 1, \exp(Z' - Z) \omega(\theta'; \theta) / \omega(\theta; \theta') \right\},$$

where $\omega(\theta'; \theta) = \pi(\theta') / q(\theta'|\theta)$. 
Inefficiency Measure

Consider a stationary Markov chain \( \{ \theta_j \} \) with invariant density \( \pi(\theta) \) and \( h : \Theta \to \mathbb{R} \) with \( \mathbb{V}_\pi[h(\theta)] < \infty \). Define

\[
\mu_h = \mathbb{E}_\pi[h(\theta)] \quad \text{and} \quad \hat{\mu}_{h,n} = n^{-1} \sum_{j=1}^{n} h(\theta_j).
\]

Then, under regularity conditions, we have

\[
\lim_{n \to \infty} n\mathbb{V} (\hat{\mu}_{h,n}) = \mathbb{V}_\pi[h(\theta)] \cdot \text{IF}_h \quad \text{with} \quad \text{IF}_h = 1 + 2 \sum_{\tau=1}^{\infty} \rho_h(\tau),
\]

where \( \rho_h(\tau) \) is the autocorrelation at lag \( \tau \) of the stationary sequence \( \{ h(\theta_j) \} \).
Consider a stationary Markov chain \( \{ \theta_j \} \) with invariant density \( \pi(\theta) \) and \( h : \Theta \to \mathbb{R} \) with \( \mathbb{V}_\pi [h(\theta)] < \infty \). Define

\[
\mu_h = \mathbb{E}_\pi [h(\theta)] \quad \text{and} \quad \hat{\mu}_{h,n} = n^{-1} \sum_{j=1}^{n} h(\theta_j).
\]

Then, under regularity conditions, we have

\[
\lim_{n \to \infty} n \mathbb{V} (\hat{\mu}_{h,n}) = \mathbb{V}_\pi [h(\theta)] \quad \text{IF}_h \quad \text{with} \quad \text{IF}_h = 1 + 2 \sum_{\tau=1}^{\infty} \rho_h(\tau),
\]

where \( \rho_h(\tau) \) is the autocorrelation at lag \( \tau \) of the stationary sequence \( \{ h(\theta_j) \} \).

The IACT, \( \text{IF}_h \), quantifies how many times more samples are required from the Markov chain relative to using iid samples from \( \pi(\theta) \) to achieve a given precision.
For the sake of analysis, we introduce an alternative $Q^*$ chain.
For the sake of analysis, we introduce an alternative $Q^*$ chain.

Let $(\theta, Z)$ be the current state of the Markov chain.
For the sake of analysis, we introduce an alternative $Q^*$ chain.

Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\theta' \sim q(\cdot | \theta)$ and $Z' \sim g(\cdot | \theta')$. 
For the sake of analysis, we introduce an alternative $Q^*$ chain.

Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\theta' \sim q(\cdot | \theta)$ and $Z' \sim g(\cdot | \theta')$.

2. Accept $\theta'$ w.p. $\alpha_{Q_{EX}}(\theta; \theta') = \min \{1, \omega(\theta'; \theta) / \omega(\theta; \theta')\}$. 

$\theta_0, Z_0$ is accepted if and only if there is acceptance in both criteria.
For the sake of analysis, we introduce an alternative $Q^*$ chain. Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\theta' \sim q(\cdot | \theta)$ and $Z' \sim g(\cdot | \theta')$.
2. Accept $\theta'$ w.p. $\alpha_{Q_{\text{EX}}}(\theta; \theta') = \min \{1, \omega(\theta'; \theta) / \omega(\theta; \theta')\}$.
3. Accept $Z'$ w.p. $\alpha_{Q_{\text{z}}}(Z; Z') = \min \{1, \exp(Z' - Z)\}$. 
A Bounding Chain

- For the sake of analysis, we introduce an alternative $Q^*$ chain.
- Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\theta' \sim q(\cdot|\theta)$ and $Z' \sim g(\cdot|\theta')$.
2. Accept $\theta'$ w.p. $\alpha_{Q^{EX}}(\theta; \theta') = \min \{1, \omega(\theta'; \theta) / \omega(\theta; \theta')\}$.
3. Accept $Z'$ w.p. $\alpha_{Q^{Z}}(Z; Z') = \min \{1, \exp(Z' - Z)\}$.
4. $(\theta', Z')$ is accepted if and only if there is acceptance in both criteria.
Lemma: The Markov chain $Q^*$ has the following properties:

1. We have $\alpha_{Q}(\theta, z; \theta_0, z_0) = \alpha_{Q}(\theta, z; \theta_0, z_0) = \alpha_{Q}(\theta; \theta_0) \alpha_{Q}(z; z_0)$.
2. $Q^*$ is a reversible Markov chain with invariant density $\pi(\theta, z)$.
3. For any function $h(\theta)$ the IACT is higher for the $Q^*$ chain than for the $Q$ chain; i.e. $IF_{Q^*}h > IF_Qh$ (Peskun, 1973; Tierney 1998).

Remark: We have $\alpha_{Q}(\theta, z; \theta_0, z_0) = \alpha_{Q}(\theta, z; \theta_0, z_0) = \alpha_{Q}(\theta; \theta_0) \alpha_{Q}(z; z_0)$ when the likelihood is known exactly and $\alpha_{Q}(\theta, z; \theta_0, z_0) = \alpha_{Q}(\theta, z; \theta, z) = \alpha_{Q}(z; z)$ when the proposal is perfect, i.e. $q(\theta_0|\theta) = \pi(\theta_0)$. 
**Lemma**: The Markov chain $Q^*$ has the following properties:

1. We have

\[ \alpha_Q(\theta, z; \theta', z') \geq \alpha_{Q^*}(\theta, z; \theta', z') = \alpha_{Q^\text{EX}}(\theta; \theta') \times \alpha_{Q^z}(z; z'). \]
Lemma: The Markov chain $Q^*$ has the following properties:

1. We have

$$\alpha_Q(\theta, z; \theta', z') \geq \alpha_{Q^*}(\theta, z; \theta', z') = \alpha_{Q^E}(\theta; \theta') \times \alpha_{Q^Z}(z; z').$$

2. $Q^*$ is a reversible Markov chain with invariant density $\hat{\pi}(\theta, z).$
Lemma: The Markov chain $Q^*$ has the following properties:

1. We have

$$
\alpha_Q(\theta, z; \theta', z') \geq \alpha_{Q^*}(\theta, z; \theta', z') = \alpha_{Q^E}(\theta; \theta') \times \alpha_{Q^z}(z; z').
$$

2. $Q^*$ is a reversible Markov chain with invariant density $\hat{\pi}(\theta, z)$.

3. For any function $h(\theta)$ the IACT is higher for the $Q^*$ chain than for the $Q$ chain; i.e. $IF_{h}^{Q^*} \geq IF_{h}^{Q}$ (Peskun, 1973; Tierney 1998).
Lemma: The Markov chain $Q^*$ has the following properties:

1. We have
   \[
   \alpha_Q(\theta, z; \theta', z') \geq \alpha_{Q^*}(\theta, z; \theta', z') = \alpha_{Q^{\text{EX}}}(\theta; \theta') \times \alpha_{Q^z}(z; z').
   \]

2. $Q^*$ is a reversible Markov chain with invariant density $\hat{\pi}(\theta, z)$.

3. For any function $h(\theta)$ the IACT is higher for the $Q^*$ chain than for the $Q$ chain; i.e. $IF_h^{Q^*} \geq IF_h^Q$ (Peskun, 1973; Tierney 1998).

Remark: We have $\alpha_{Q^*}(\theta, z; \theta', z') = \alpha_Q(\theta, z; \theta', z') = \alpha_{Q^{\text{EX}}}(\theta; \theta')$ when the likelihood is known exactly and $\alpha_{Q^*}(\theta, z; \theta', z') = \alpha_Q(\theta, z; \theta^*, z^*) = \alpha_{Q^z}(z; z^*)$ when the proposal is perfect, i.e. $q(\theta'|\theta) = \pi(\theta')$. 

A. Doucet (UCL Masterclass Oct. 2012)
**Assumption.** Let $Z = \log \hat{p}_\theta (y; U) - \log p_\theta (y)$ be the error in the estimator of the log likelihood.
Assumption. Let $Z = \log \hat{p}_\theta (y; U) - \log p_\theta (y)$ be the error in the estimator of the log likelihood.

We have for $N$ particles

$$g(z|\theta) = \phi \left( z; -\gamma^2(\theta)/2N, \gamma^2(\theta)/N \right),$$
$$\hat{\pi}(z|\theta) = \exp(z)g(z|\theta) = \phi \left( z; \gamma^2(\theta)/2N, \gamma^2(\theta)/N \right)$$

where $\phi(z; a, b^2)$ is a univariate normal of mean $a$, variance $b^2$. 
Assumption. Let $Z = \log \hat{p}_\theta (y; U) - \log p_\theta (y)$ be the error in the estimator of the log likelihood.

1. We have for $N$ particles

$$g(z|\theta) = \phi \left(z; -\gamma^2(\theta)/2N, \gamma^2(\theta)/N\right),$$

$$\hat{\pi}(z|\theta) = \exp(z)g(z|\theta) = \phi \left(z; \gamma^2(\theta)/2N, \gamma^2(\theta)/N\right)$$

where $\phi(z; a, b^2)$ is a univariate normal of mean $a$, variance $b^2$.

2. For a given value of $\sigma^2$ we set $N = N_{\sigma^2}(\theta) = \gamma(\theta)^2 / \sigma^2$. 
Making Assumptions to Move Forward

**Assumption.** Let $Z = \log \hat{p}_\theta(y; U) - \log p_\theta(y)$ be the error in the estimator of the log likelihood.

1. We have for $N$ particles

$$
g(z|\theta) = \phi\left(z; -\gamma^2(\theta)/2N, \gamma^2(\theta)/N\right),$$
$$
\hat{\pi}(z|\theta) = \exp(z)g(z|\theta) = \phi\left(z; \gamma^2(\theta)/2N, \gamma^2(\theta)/N\right)
$$

where $\phi(z; a, b^2)$ is a univariate normal of mean $a$, variance $b^2$.

2. For a given value of $\sigma^2$ we set $N = N_{\sigma^2}(\theta) = \gamma(\theta)^2/\sigma^2$.

Under this assumption, both $g(z|\theta)$ and $\hat{\pi}(z|\theta)$ are functions of $\sigma^2$ only and we write $g(z|\theta)$ and $\pi(z|\theta)$ as

$$
g(z|\sigma^2) = \phi\left(z; -\sigma^2/2, \sigma^2\right), \quad \pi(z|\sigma^2) = \phi\left(z; \sigma^2/2, \sigma^2\right).
$$

and $\theta$ and $Z$ are independent under $\hat{\pi}(\theta, z)$. 
Empirical vs Asymptotic Distribution of Log-Likelihood Estimator

Figure: Histograms of proposed (red) and accepted (pink) values of $z$ in PMCMC scheme. Overlayed are Gaussian pdfs from our simplifying Assumption for a target of $\sigma = 0.92$. 
Lemma (Pitt et al., 2012). Under the previous assumption, the following results hold for the chain \( \{\theta_j, Z_j\} \) arising from \( Q^* \) that uses the perfect proposal \( q(\theta' | \theta) = \pi(\theta') \) denoted \( Q^Z \).
Lemma (Pitt et al., 2012). Under the previous assumption, the following results hold for the chain \( \{\theta_j, Z_j\} \) arising from \( Q^* \) that uses the perfect proposal \( q(\theta'|\theta) = \pi(\theta') \) denoted \( Q^Z \).

Let \( p(z; \sigma^2) \) be the probability of rejection given the current value \( z \). Then

\[
p(z; \sigma^2) = 1 - \int \alpha_{Q^Z} (z; z') g(z'|\sigma^2) dz' \\
= \Phi(z/\sigma + \sigma/2) - \exp(-z)\Phi(z/\sigma - \sigma/2).
\]
Lemma (Pitt et al., 2012). Under the previous assumption, the following results hold for the chain \{\theta_j, Z_j\} arising from Q^* that uses the perfect proposal \( q(\theta' | \theta) = \pi(\theta') \) denoted \( Q^Z \).

1. Let \( p(z; \sigma^2) \) be the probability of rejection given the current value \( z \). Then

\[
p(z; \sigma^2) = 1 - \int a_{Q^Z} (z; z') g(z' | \sigma^2) dz' \\
= \Phi(z / \sigma + \sigma / 2) - \exp(-z)\Phi(z / \sigma - \sigma / 2).
\]

2. \( IF^Z(\sigma^2) = \mathbb{E}_{\pi(\cdot | \sigma^2)} \left( \frac{1 + p(z; \sigma^2)}{1 - p(z; \sigma^2)} \right) = \int \frac{1 + p^*(w, \sigma)}{1 - p^*(w, \sigma)} \phi(w) dw, \)

where \( p^*(w, \sigma) = \Phi(w + \sigma) - \exp(-w\sigma - \sigma^2 / 2)\Phi(w) \), \( \Phi(\cdot) \) standard normal cdf. In particular, \( IF^Z(\sigma^2) \) is independent of \( h \).
Main Result

- Under the previous assumption and additional regularity conditions,

\[ IF_h^Q(\sigma^2) \leq IF_h^{Q^*}(\sigma^2) \leq IF_h^U(\sigma^2), \]

\[ IF_h^U(\sigma^2) = \frac{1}{2} (1 + IF_h^{EX})(1 + IF_h^Z(\sigma^2)) - 1. \]
Main Result

- Under the previous assumption and additional regularity conditions,

\[ IF_Q^h(\sigma^2) \leq IF_Q^*(\sigma^2) \leq IF_U^h(\sigma^2), \]

\[ IF_U^h(\sigma^2) = \frac{1}{2}(1 + IF_{h}^{EX})(1 + IF^{Z}(\sigma^2)) - 1. \]

- The inequality becomes exact when \( \sigma \rightarrow 0 \) (as \( IF^{Z}(\sigma^2) \rightarrow 1 \) for \( \sigma \rightarrow 0 \)) and when the proposal is perfect (as \( IF_{h}^{EX} = 1 \) when \( q(\theta'|\theta) = \pi(\theta') \)).
Main Result

- Under the previous assumption and additional regularity conditions,

\[ IF^Q_h(\sigma^2) \leq IF^{Q^*}_h(\sigma^2) \leq IF^U_h(\sigma^2), \]

\[ IF^U_h(\sigma^2) = \frac{1}{2} (1 + IF^{EX}_h)(1 + IF^{Z}_h(\sigma^2)) - 1. \]

- The inequality becomes exact when \( \sigma \rightarrow 0 \) (as \( IF^{Z}_h(\sigma^2) \rightarrow 1 \) for \( \sigma \rightarrow 0 \)) and when the proposal is perfect (as \( IF^{EX}_h = 1 \) when \( q(\theta'|\theta) = \pi(\theta') \)).

- For the unconditional acceptance probability, we have

\[ P^Q(A|\sigma^2) \geq P^{Q^*}(A|\sigma^2) = 2\Phi(-\sigma/\sqrt{2}) P^{EX}(A). \]

with the bound getting tighter if either \( P^{EX}(A) \rightarrow 1 \) or \( P^{Z}(A|\sigma^2) \rightarrow 1. \)
Relative Inefficiency

We have

\[
\frac{IF^Q_h(\sigma^2)}{IF^E_h} \leq \frac{IF^U_h(\sigma^2)}{IF^E_h} = RIF^U_h(\sigma^2),
\]

\[
RIF^U_h(\sigma^2) = \frac{1}{2} \frac{IF^Z_h(\sigma^2) - 1}{IF^E_h} + \frac{1}{2} \left(1 + IF^Z_h(\sigma^2)\right).
\]
Relative Inefficiency

We have

\[
\frac{\text{IF}_h^Q(\sigma^2)}{\text{IF}_h^\text{EX}} \leq \frac{\text{IF}_h^U(\sigma^2)}{\text{IF}_h^\text{EX}} = \text{RIF}_h^U(\sigma^2),
\]

\[
\text{RIF}_h^U(\sigma^2) = \frac{1}{2} \frac{(\text{IF}_h^Z(\sigma^2) - 1)}{\text{IF}_h^\text{EX}} + \frac{1}{2} (1 + \text{IF}_h^Z(\sigma^2)).
\]

For fixed \( \sigma \), \( \text{RIF}_h^U(\sigma^2) \) decreases as \( \text{IF}_h^\text{EX} \) increases from a value of \( \text{IF}_h^Z(\sigma^2) \) for \( \text{IF}_h^\text{EX} = 1 \) to

\[
\text{RIF}_h^U(\sigma^2) \longrightarrow \frac{1}{2} (1 + \text{IF}_h^Z(\sigma^2)) \leq \text{IF}_h^Z(\sigma^2) \text{ as } \text{IF}_h^\text{EX} \longrightarrow \infty.
\]
Relative Inefficiency

- We have

\[
\frac{IF_h^Q(\sigma^2)}{IF_h^{EX}} \leq \frac{IF_h^U(\sigma^2)}{IF_h^{EX}} = RIF_h^U(\sigma^2),
\]

\[
RIF_h^U(\sigma^2) = \frac{1}{2} \frac{(IF^Z(\sigma^2) - 1)}{IF_h^{EX}} + \frac{1}{2} (1 + IF^Z(\sigma^2)).
\]

- For fixed \( \sigma \), \( RIF_h^U(\sigma^2) \) decreases as \( IF_h^{EX} \) increases from a value of \( IF^Z(\sigma^2) \) for \( IF_h^{EX} = 1 \) to

\[
RIF_h^U(\sigma^2) \rightarrow \frac{1}{2} (1 + IF^Z(\sigma^2)) \leq IF^Z(\sigma^2) \text{ as } IF_h^{EX} \rightarrow \infty.
\]

- The loss in efficiency from using the estimated likelihood goes down as the proposal deteriorates.
The Computing Time (CT) for $Q$ is defined as

$$CT^Q_h(\sigma^2) = \frac{IF^Q_h(\sigma^2)}{\sigma^2};$$

i.e. we take into account the computational efforts associated to $\sigma^2 \propto 1/N$. 

---

A. Doucet (UCL Masterclass Oct. 2012)
Computing Time

- The Computing Time (CT) for $Q$ is defined as

$$CT^Q_h(\sigma^2) = IF^Q_h(\sigma^2) / \sigma^2;$$

i.e. we take into account the computational efforts associated to $\sigma^2 \propto 1 / N$.

- $CT^Q_h(\sigma^2) \leq CT^U_h(\sigma^2)$ where $CT^U_h(\sigma^2) := IF^U_h(\sigma^2) / \sigma^2$. 

A. Doucet (UCL Masterclass Oct. 2012)
The Computing Time (CT) for $Q$ is defined as

$$CT^Q_h(\sigma^2) = IF^Q_h(\sigma^2) / \sigma^2;$$

i.e. we take into account the computational efforts associated to $\sigma^2 \propto 1/N$. 

1. $CT^Q_h(\sigma^2) \leq CT^U_h(\sigma^2)$ where $CT^U_h(\sigma^2) := IF^U_h(\sigma^2) / \sigma^2$. 

2. If $IF^E_h = 1$, then $CT^U_h(\sigma^2)$ is minimized at $\sigma^U_{opt} = 0.92$ and $IF^Z(\sigma^U_{opt}) = 4.54$, $P^Z(A|\sigma^U_{opt}) = 0.5153$. 

A. Doucet (UCL Masterclass Oct. 2012)
Computing Time

The Computing Time (CT) for $Q$ is defined as

$$CT^Q_h(\sigma^2) = IF^Q_h(\sigma^2) / \sigma^2;$$

i.e. we take into account the computational efforts associated to $\sigma^2 \propto 1/N$.

1. $CT^Q_h(\sigma^2) \leq CT^U_h(\sigma^2)$ where $CT^U_h(\sigma^2) := IF^U_h(\sigma^2) / \sigma^2$.
2. If $IF^\text{EX}_h = 1$, then $CT^U_h(\sigma^2)$ is minimized at $\sigma^U_{opt} = 0.92$ and $IF^Z(\sigma^U_{opt}) = 4.54, \ P^Z(A|\sigma^U_{opt}) = 0.5153$.
3. The minimizing $\sigma^U_{opt}$ rises with $IF^\text{EX}_h$ to $\sigma^U_{opt} = 1.0206$ as $IF^\text{EX}_h \rightarrow \infty$. 
Computing Time

- The Computing Time (CT) for $Q$ is defined as

$$CT^Q_h(\sigma^2) = IF^Q_h(\sigma^2) / \sigma^2;$$

i.e. we take into account the computational efforts associated to $\sigma^2 \propto 1/N$.

1. $CT^Q_h(\sigma^2) \leq CT^U_h(\sigma^2)$ where $CT^U_h(\sigma^2) := IF^U_h(\sigma^2) / \sigma^2$.

2. If $IF^E_h = 1$, then $CT^U_h(\sigma^2)$ is minimized at $\sigma^U_{opt} = 0.92$ and $IF^Z(\sigma^U_{opt}^2) = 4.54$, $P^Z(A|\sigma^U_{opt}) = 0.5153$.

3. The minimizing $\sigma^U_{opt}$ rises with $IF^E_h$ to $\sigma^U_{opt} = 1.0206$ as $IF^E_h \rightarrow \infty$.

- Define the relative computing time for the inefficiency bound $IF^U_h(\sigma^2)$ as

$$RCT^U_h(\sigma^2) = \frac{RIF^U_h(\sigma^2)}{\sigma^2}.$$

Both $RIF^U_h(\sigma^2)$ and $RCT^U_h(\sigma^2)$ are decreasing functions of $IF^E_h$. 
Relative Upper Bounds on Inefficiency and Computing Time

**Figure:** $RCT_h^U$ (top) and $RIF_h^U$ (bottom) against $1/\sigma^2$ (left) and $\sigma$ (right). Different values of $IF_h^{EX}$ are shown on each plot.
Example 1: Probit Model

- We use a simple Bernoulli Probit model, where for \( t = 1, \ldots, T \)
  \[
  Y_t = \mathbb{I}(X_t > 0), \quad X_t \overset{iid}{\sim} N(\theta; 1).
  \]
Example 1: Probit Model

- We use a simple Bernoulli Probit model, where for \( t = 1, \ldots, T \)
  \[ Y_t = \mathbb{I}(X_t > 0), \quad X_t \sim \text{iid } \mathcal{N}(\theta; 1). \]

- The likelihood is known explicitly as \( \Pr(Y_t = 1) = \Phi(\theta) \) but is estimated through
  \[ \hat{\Pr}_\theta(Y_t = 1) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(X_t^{(k)} > 0), \quad X_t^{(k)} \sim \text{iid } \mathcal{N}(\theta; 1) \]
  and set \( \theta \sim \mathcal{N}(0, \sigma_\theta^2) \) with \( \sigma_\theta^2 \gg 1. \)
Example 1: Probit Model

- We use a simple Bernoulli Probit model, where for \( t = 1, \ldots, T \)
  \[ Y_t = \text{I}(X_t > 0), \quad X_t \overset{iid}{\sim} N(\theta; 1). \]
- The likelihood is known explicitly as \( \Pr(Y_t = 1) = \Phi(\theta) \) but is estimated through
  \[ \hat{\Pr}_\theta(Y_t = 1) = \frac{1}{N} \sum_{k=1}^{N} \text{I}(X_t^{(k)} > 0), \quad X_t^{(k)} \overset{iid}{\sim} N(\theta; 1) \]
  and set \( \theta \sim \mathcal{N}(0, \sigma^2_\theta) \) with \( \sigma^2_\theta \gg 1. \)
- Autoregressive Metropolis proposal
  \[ \theta' = \hat{\theta} + \rho(\theta - \hat{\theta}) + \sqrt{\frac{\sigma^2(1 - \rho^2)}{\nu - 2}} t_5, \]
  where \( \hat{\theta} \) is the posterior mode and \( \hat{\sigma}^2 \) is chosen as the negative inverse of the second derivative of the log posterior.
Example 1: Probit Model

- We use a simple Bernoulli Probit model, where for $t = 1, \ldots, T$
  \[ Y_t = \mathbb{I}(X_t > 0), \quad X_t \overset{iid}{\sim} N(\theta; 1). \]

- The likelihood is known explicitly as $\Pr(Y_t = 1) = \Phi(\theta)$ but is estimated through
  \[ \hat{\Pr}_\theta(Y_t = 1) = \frac{1}{N} \sum_{k=1}^{N} \mathbb{I}(X_t^{(k)} > 0), \quad X_t^{(k)} \overset{iid}{\sim} N(\theta; 1) \]
  and set $\theta \sim N\left(0, \sigma_\theta^2\right)$ with $\sigma_\theta^2 \gg 1$.

- Autoregressive Metropolis proposal
  \[ \theta' = \hat{\theta} + \rho(\theta - \hat{\theta}) + \sqrt{\frac{\hat{\sigma}^2(1 - \rho^2)}{\nu - 2}} t_5, \]
  where $\hat{\theta}$ is the posterior mode and $\hat{\sigma}^2$ is chosen as the negative inverse of the second derivative of the log posterior.

- We will consider $\rho \in \{0, 0.4, 0.6, 0.9, 0.97\}$. 
We want to assess experimentally whether our upper/lower bounds are sharp.
Setting the Number of Samples

- We want to assess experimentally whether our upper/lower bounds are sharp.
- We select $N$ so as $\sigma$ to be roughly equal to a pre-specified value.
We want to assess experimentally whether our upper/lower bounds are sharp.

We select $N$ so as $\sigma$ to be roughly equal to a pre-specified value.

1. Choose a large initial value for the number of samples, $N_S$. 

Next steps:

- Run the MCMC scheme for a fixed number of iterates recording $\theta$.
- Record the estimated variance of the log of the likelihood estimator, $V(\theta, N_S) = bV[\log b p_{\theta}(y; U)]$.
- Set $N_{\theta} = V(\theta, N_S) / \sigma^2$. 

A. Doucet (UCL Masterclass Oct. 2012)
Setting the Number of Samples

- We want to assess experimentally whether our upper/lower bounds are sharp.
- We select $N$ so as $\sigma$ to be roughly equal to a pre-specified value.

1. Choose a large initial value for the number of samples, $N_S$.
2. Run the MCMC scheme for a fixed number of iterates recording $\bar{\theta}$.
Setting the Number of Samples

- We want to assess experimentally whether our upper/lower bounds are sharp.
- We select $N$ so as $\sigma$ to be roughly equal to a pre-specified value.

1. Choose a large initial value for the number of samples, $N_S$.
2. Run the MCMC scheme for a fixed number of iterates recording $\theta$.
3. Record the estimated variance of the log of the likelihood estimator, $V(\theta, N_S) = \hat{V} \left[ \log \hat{p}_\theta(y; U) \right]$. 

A. Doucet (UCL Masterclass Oct. 2012)
We want to assess experimentally whether our upper/lower bounds are sharp.

We select $N$ so as $\sigma$ to be roughly equal to a pre-specified value.

1. Choose a large initial value for the number of samples, $N_S$.
2. Run the MCMC scheme for a fixed number of iterates recording $\bar{\theta}$.
3. Record the estimated variance of the log of the likelihood estimator, $V(\bar{\theta}, N_S) = \hat{V}[\log \hat{p}_{\bar{\theta}}(y; U)]$.
4. Set $N_{\bar{\theta}} = V(\bar{\theta}, N_S) / \sigma^2$. 
Acceptance Probabilities for Probit Example

Figure: Probit example $T = 100$, $\theta = 0.5$. Accept. Proba vs $\sigma(\bar{\theta})$. Estim. proba for the exact MCMC scheme is shown (constant), estim. proba from the simulated likelihood scheme (red) and lower bound given as proba exact scheme times $2\Phi(-\sigma/\sqrt{2})$ (blue).
Relative Inefficiency and Computing Time

Figure: $RCT^Q_h$ (top) and $RIF^Q_h$ (bottom) against $N$ (left) and $\sigma(\bar{\theta})$ (right) for various values of $\rho$. 
Example 2: Noisy Autoregressive Example

We have

\[ X_{t+1} = \mu(1 - \phi) + \phi X_t + \sigma_\eta \eta_t \quad \text{and} \quad Y_t = X_t + \sigma_\varepsilon W_t \]

where \( \eta_t \) and \( W_t \) are independent standard normal and
\( \theta = \left( \phi, \mu, \sigma^2_\eta \right) \).
Example 2: Noisy Autoregressive Example

- We have
  \[ X_{t+1} = \mu (1 - \phi) + \phi X_t + \sigma_\eta \eta_t \quad \text{and} \quad Y_t = X_t + \sigma_\varepsilon W_t \]

  where \( \eta_t \) and \( W_t \) are independent standard normal and

  \[ \theta = (\phi, \mu, \sigma_\eta^2). \]

- The likelihood can be computed exactly using the Kalman filter.
Example 2: Noisy Autoregressive Example

- We have

\[ X_{t+1} = \mu (1 - \phi) + \phi X_t + \sigma_\eta \eta_t \quad \text{and} \quad Y_t = X_t + \sigma_\varepsilon W_t \]

where \( \eta_t \) and \( W_t \) are independent standard normal and

\[ \theta = (\phi, \mu, \sigma_\eta^2) \].

- The likelihood can be computed exactly using the Kalman filter.

- Autoregressive Metropolis proposal for \( \theta \) based on multivariate t-distribution.
We have

\[ X_{t+1} = \mu(1 - \phi) + \phi X_t + \sigma_\eta \eta_t \quad \text{and} \quad Y_t = X_t + \sigma_\varepsilon W_t \]

where \( \eta_t \) and \( W_t \) are independent standard normal and \( \theta = \left( \phi, \mu, \sigma_\eta^2 \right) \).

- The likelihood can be computed exactly using the Kalman filter.
- Autoregressive Metropolis proposal for \( \theta \) based on multivariate t-distribution.
- \( N \) is selected in the same manner so as to obtain an approximately constant \( \sigma \left( \bar{\theta} \right) \).
Figure: AR1 plus noise example with $T = 300, \phi = 0.8, \mu = 0.5, \sigma^2_{\eta} = 1, \sigma^2_{\varepsilon} = 0.5$. Probabilities of acceptance displayed against $\sigma(\bar{\theta})$. 
Relative Inefficiency and Computing Time

Figure: From left to right: $RCT^Q_h$ vs $N$, $RCT^Q_h$ vs $\sigma(\bar{\theta})$, $RIF^Q_h$ against $N$ and $RIF^Q_h$ against $\sigma(\bar{\theta})$ for various values of $\rho$ and different parameters.
We have provided an approximate analysis of the PMMH sampler.
We have provided an approximate analysis of the PMMH sampler.

For a general proposal and under simplifying assumptions on the likelihood estimator, we can get guidelines on how to select $\sigma$: *as long as $\sigma$ is around 1 then you are fine.*
We have provided an approximate analysis of the PMMH sampler.

For a general proposal and under simplifying assumptions on the likelihood estimator, we can get guidelines on how to select $\sigma$: \textit{as long as $\sigma$ is around 1 then you are fine.}

Coming up with a nicer way to adapt $N$ would be useful; e.g. Lee, 2011.
We have provided an approximate analysis of the PMMH sampler.

For a general proposal and under simplifying assumptions on the likelihood estimator, we can get guidelines on how to select $\sigma$: *as long as $\sigma$ is around 1 then you are fine.*

Coming up with a nicer way to adapt $N$ would be useful; e.g. Lee, 2011.

Particle Gibbs samplers are a powerful alternative not yet well understood (Andrieu, D. & Holenstein, 2010; Whiteley, Andrieu, D., 2010; Lindsten, Jordan, Schon, 2012).