Efficient Implementation of MCMC When Using An Unbiased Likelihood Estimator

Michael Pitt
University of Oxford
Joint work with A. Doucet, G. Deligiannidis & R. Kohn

Cambridge, 22/04/14
MCMC with intractable likelihood functions.
1. MCMC with intractable likelihood functions. 
   In physics, first appeared in Lin, Liu & Sloan (2000). In statistics, 
   Beaumont (2003), Andrieu, Berthelesen, D., Roberts (2006), Andrieu 
   & Roberts (2009).

2. Inefficiency of the MCMC scheme with likelihood estimator against 
   known likelihood.

Inefficiency of the MCMC scheme with likelihood estimator against known likelihood.

Guidelines on an optimal precision for the estimator of the log-likelihood.
Bayesian Inference

- Likelihood function $p(y; \theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$. 
Bayesian Inference

- Likelihood function $p(y; \theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.
Bayesian Inference

- Likelihood function $p(y; \theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.
- Prior distribution of density $p(\theta)$.
- Bayesian inference relies on the posterior

$$\pi(\theta) = p(\theta | y) = \frac{p(y; \theta) p(\theta)}{\int_{\Theta} p(y; \theta') p(\theta') d\theta'}.$$
Likelihood function $p(y; \theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^d$.

Prior distribution of density $p(\theta)$.

Bayesian inference relies on the posterior

$$p(\theta|y) = \frac{p(y; \theta) p(\theta)}{\int_{\Theta} p(y; \theta') p(\theta') d\theta'}.$$ 

For non-trivial models, inference relies typically on MCMC.
Set $\theta^{(0)}$ and iterate for $i = 1, 2, ...$
Metropolis-Hastings algorithm

1. Set $\theta^{(0)}$ and iterate for $i = 1, 2, ...$

2. Sample $\theta \sim q(\cdot | \theta^{(i-1)})$.
Set $\vartheta^{(0)}$ and iterate for $i = 1, 2, \ldots$

1. Sample $\vartheta \sim q \left( \cdot | \vartheta^{(i-1)} \right)$.

2. Compute

$$\alpha = 1 \wedge \frac{p(y; \vartheta) p(\vartheta)}{p(y; \vartheta^{(i-1)}) p(\vartheta^{(i-1)})} \frac{q\left(\vartheta^{(i-1)} | \vartheta\right)}{q\left(\vartheta | \vartheta^{(i-1)}\right)}.$$
Set $\theta^{(0)}$ and iterate for $i = 1, 2, ...$

1. Sample $\theta \sim q(\cdot | \theta^{(i-1)})$.

2. Compute

$$\alpha = 1 \land \frac{p(y; \theta) p(\theta)}{p(y; \theta^{(i-1)}) p(\theta^{(i-1)})} \frac{q(\theta^{(i-1)} | \theta)}{q(\theta | \theta^{(i-1)})}.$$ 

3. With probability $\alpha$, set $\theta^{(i)} := \theta$ and $\theta^{(i)} := \theta^{(i-1)}$ otherwise.
In numerous scenarios, $p(y; \theta)$ cannot be evaluated pointwise; e.g.

$$p(y; \theta) = \int p(x, y; \theta) \, dx$$

where the integral cannot be evaluated.
In numerous scenarios, \( p(y; \theta) \) cannot be evaluated pointwise; e.g.

\[
p(y; \theta) = \int p(x, \theta) \, dx
\]

where the integral cannot be evaluated.

A standard “solution” consists of using MCMC to sample from

\[
p(\theta, x \mid y) = \frac{p(x, \theta \mid \theta) \, p(\theta)}{p(y)}
\]

by updating iteratively \( x \) and \( \theta \).
In numerous scenarios, \( p(y; \theta) \) cannot be evaluated pointwise; e.g.

\[
p(y; \theta) = \int p(x, y; \theta) \, dx
\]

where the integral cannot be evaluated.

A standard “solution” consists of using MCMC to sample from

\[
p(\theta, x | y) = \frac{p(x, y; \theta) \, p(\theta)}{p(y)}
\]

by updating iterately \( x \) and \( \theta \).

Gibbs sampling strategies can be slow mixing and difficult to put in practice.
Let $\hat{p}(y; \theta, U)$ be an unbiased non-negative estimator of the likelihood where $U \sim m(\cdot)$; i.e.

$$p(y; \theta) = \int_U \hat{p}(y; \theta, u) m(\theta, u) \, du.$$
Let $\hat{p}(y; \theta, U)$ be an unbiased non-negative estimator of the likelihood where $U \sim m_\theta(\cdot)$; i.e.

$$p(y; \theta) = \int_U \hat{p}(y; \theta, u) m_\theta(u) \, du.$$ 

Introduce a target distribution on $\Theta \times U$ of density

$$\bar{\pi}(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_\theta(u) = \frac{p(\theta) \hat{p}(y; \theta, u) m_\theta(u)}{p(y)}$$

then unbiasedness yields

$$\int_U \bar{\pi}(\theta, u) \, du = \pi(\theta)$$
Let \( \hat{p}(y; \theta, U) \) be an unbiased non-negative estimator of the likelihood where \( U \sim m_\theta(\cdot) \); i.e.

\[
p(y; \theta) = \int_U \hat{p}(y; \theta, u) m_\theta(u) \, du.
\]

Introduce a target distribution on \( \Theta \times U \) of density

\[
\overline{\pi}(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_\theta(u) = \frac{p(\theta) \hat{p}(y; \theta, u) m_\theta(u)}{p(y)}
\]

then unbiasedness yields

\[
\int_U \overline{\pi}(\theta, u) \, du = \pi(\theta)
\]

Any MCMC algorithm sampling from \( \overline{\pi}(\theta, u) \) yields samples from \( \pi(\theta) \).
Pseudo-Marginal Metropolis-Hastings algorithm

Set \( (\theta^{(0)}, U^{(0)}) \) and iterate for \( i = 1, 2, ... \)
Set \( (\varnothing^{(0)}, U^{(0)}) \) and iterate for \( i = 1, 2, \ldots \)

1. Sample \( \varnothing \sim q (\cdot | \varnothing^{(i-1)}) \), \( U \sim m_\varnothing (\cdot) \) to obtain \( \hat{p} (y; \varnothing, U) \).
Pseudo-Marginal Metropolis-Hastings algorithm

- Set \( \left( \vartheta^{(0)}, U^{(0)} \right) \) and iterate for \( i = 1, 2, \ldots \)

1. Sample \( \vartheta \sim q \left( \cdot \mid \vartheta^{(i-1)} \right) \), \( U \sim m_{\vartheta} (\cdot) \) to obtain \( \hat{p}(y; \vartheta, U) \).

2. Compute

\[
\alpha = 1 \wedge \frac{\hat{p}(y; \vartheta, U)}{\hat{p}(y; \vartheta^{(i-1)}, U^{(i-1)})} \frac{p(\vartheta)}{p(\vartheta^{(i-1)})} \frac{q(\vartheta^{(i-1)} \mid \vartheta)}{q(\vartheta \mid \vartheta^{(i-1)})}
\]
Pseudo-Marginal Metropolis-Hastings algorithm

- Set \( (\theta^{(0)}, U^{(0)}) \) and iterate for \( i = 1, 2, \ldots \)

1. Sample \( \theta \sim q \left( \cdot \mid \theta^{(i-1)} \right) \), \( U \sim m_{\theta} (\cdot) \) to obtain \( \hat{p} (y; \theta, U) \).

2. Compute

\[
\alpha = 1 \land \frac{\hat{p} (y; \theta, U)}{\hat{p} (y; \theta^{(i-1)}, U^{(i-1)})} \frac{p (\theta)}{p (\theta^{(i-1)})} \frac{q \left( \theta^{(i-1)} \mid \theta \right)}{q \left( \theta \mid \theta^{(i-1)} \right)}
\]

3. With proba \( \alpha \), set \( \left( \theta^{(i)}, \hat{p} (y; \theta^{(i)}, U^{(i)}) \right) \) := \( (\theta, \hat{p} (y; \theta, U)) \) and stay where you are otherwise.
For latent variable models, one has

\[ p(y; \theta) = \int p(x, y; \theta) \, dx = \int \frac{p(x, y; \theta)}{q_\theta(x)} q_\theta(x) \, dx \]

where \( q_\theta(x) \) is an importance sampling density.
Importance Sampling Estimator

- For latent variable models, one has

\[
p(y; \theta) = \int p(x, y; \theta) \, dx = \int \frac{p(x, y; \theta)}{q_\theta(x)} q_\theta(x) \, dx
\]

where \(q_\theta(x)\) is an importance sampling density.

- An unbiased estimator is given by

\[
\hat{p}(y; \theta, U) = \frac{1}{N} \sum_{k=1}^{N} \frac{p(X^k, y; \theta)}{q_\theta(X^k)}, \quad X^k \overset{i.i.d.}{\sim} q_\theta(\cdot)
\]
For latent variable models, one has

\[ p(y; \theta) = \int p(x, y; \theta) \, dx = \int \frac{p(x, y; \theta)}{q_\theta(x)} q_\theta(x) \, dx \]

where \( q_\theta(x) \) is an importance sampling density.

An unbiased estimator is given by

\[ \hat{p}(y; \theta, U) = \frac{1}{N} \sum_{k=1}^{N} \frac{p(X^k, y; \theta)}{q_\theta(X^k)}, \quad X^k \text{ i.i.d. } q_\theta(\cdot) \]

Here \( U := (X^1, ..., X^N) \) and \( m_\theta(u) = \prod_{k=1}^{N} q_\theta(x^k) \).
Importance Sampling Estimator

- For latent variable models, one has

\[ p(y; \theta) = \int p(x, y; \theta) \, dx = \int \frac{p(x, y; \theta)}{q_\theta(x)} q_\theta(x) \, dx \]

where \( q_\theta(x) \) is an importance sampling density.

- An unbiased estimator is given by

\[ \hat{p}(y; \theta, U) = \frac{1}{N} \sum_{k=1}^{N} \frac{p(X^k, y; \theta)}{q_\theta(X^k)}, \quad X^k \text{i.i.d.} q_\theta(\cdot) \]

- Here \( U := (X^1, \ldots, X^N) \) and \( m_\theta(u) = \prod_{k=1}^{N} q_\theta(x^k) \).

- Whatever being \( N \geq 1 \), the pseudo-marginal MH admits \( \pi(\theta) \) as invariant distribution.
Sequential Monte Carlo Estimator

- \{X_t\}_{t \geq 0} is a \mathbb{X}\text{-valued} latent Markov process with \(X_0 \sim \mu(\cdot; \theta)\) and \(X_{t+1}|X_t \sim f(\cdot|X_t; \theta)\).
\{X_t\}_{t \geq 0} \text{ is a } X\text{-valued latent Markov process with } X_0 \sim \mu(\cdot; \theta) \text{ and } X_{t+1} | X_t \sim f(\cdot | X_t; \theta).

Observations \{Y_t\}_{t \geq 1} \text{ are conditionally independent given } \{X_t\}_{t \geq 0} \text{ with } Y_t | \{X_k\}_{k \geq 0} \sim g(\cdot | X_t, \theta).
Sequential Monte Carlo Estimator

- \{X_t\}_{t \geq 0} is a \mathbb{X}-valued latent Markov process with \(X_0 \sim \mu(\cdot; \theta)\) and \(X_{t+1} | X_t \sim f(\cdot | X_t; \theta)\).
- Observations \{Y_t\}_{t \geq 1} are conditionally independent given \{X_t\}_{t \geq 0} with \(Y_t | \{X_k\}_{k \geq 0} \sim g(\cdot | X_t, \theta)\).
- Likelihood of \(y_{1:T} = (y_1, \ldots, y_T)\) is

\[
p(y_{1:T}; \theta) = \int_{\mathbb{X}^{T+1}} p(x_{0:T}, y_{1:T}; \theta) dx_{0:T}.
\]
Sequential Monte Carlo Estimator

- \( \{X_t\}_{t \geq 0} \) is a \( \mathcal{X} \)-valued latent Markov process with \( X_0 \sim \mu(\cdot; \theta) \) and \( X_{t+1} | \tilde{X}_t \sim f(\cdot | X_t; \theta) \).

- Observations \( \{Y_t\}_{t \geq 1} \) are conditionally independent given \( \{X_t\}_{t \geq 0} \) with \( Y_t | \{X_k\}_{k \geq 0} \sim g(\cdot | X_t, \theta) \).

- Likelihood of \( y_{1:T} = (y_1, \ldots, y_T) \) is

\[
p(y_{1:T}; \theta) = \int_{\mathcal{X}^{T+1}} p(x_{0:T}, y_{1:T}; \theta) dx_{0:T}.
\]

- SMC provides an unbiased estimator of relative variance \( \mathcal{O}(T/N) \) where \( N \) is the number of particles.
Sequential Monte Carlo Estimator

- $\{X_t\}_{t \geq 0}$ is a $\mathcal{X}$-valued latent Markov process with $X_0 \sim \mu(\cdot; \theta)$ and $X_{t+1}|X_t \sim f(\cdot|X_t; \theta)$.
- Observations $\{Y_t\}_{t \geq 1}$ are conditionally independent given $\{X_t\}_{t \geq 0}$ with $Y_t|\{X_k\}_{k \geq 0} \sim g(\cdot|X_t, \theta)$.
- Likelihood of $y_{1:T} = (y_1, ..., y_T)$ is
  \[ p(y_{1:T}; \theta) = \int_{\mathcal{X}^{T+1}} p(x_{0:T}, y_{1:T}; \theta) dx_{0:T}. \]
- SMC provides an unbiased estimator of relative variance $O\left(\frac{T}{N}\right)$ where $N$ is the number of particles.
- Whatever being $N \geq 1$, the pseudo-marginal MH admits $\pi(\theta)$ as invariant distribution.
A Nonlinear State-Space Model

- Standard non-linear model

\[ X_t = \frac{1}{2} X_{t-1} + 25 \frac{X_{t-1}}{1 + X_{t-1}^2} + 8 \cos(1.2t) + V_t, \quad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_V^2), \]

\[ Y_t = \frac{1}{20} X_t^2 + W_t, \quad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_W^2). \]
A Nonlinear State-Space Model

- Standard non-linear model

\[
X_t = \frac{1}{2} X_{t-1} + 25 \frac{X_{t-1}}{1+X^2_{t-1}} + 8 \cos(1.2t) + V_t, \quad V_t \sim \mathcal{N}(0, \sigma_V^2),
\]

\[
Y_t = \frac{1}{20} X_t^2 + W_t, \quad W_t \sim \mathcal{N}(0, \sigma_W^2).
\]

- \(T = 200\) data points with \(\theta = (\sigma_V^2, \sigma_W^2) = (10, 10)\).
A Nonlinear State-Space Model

- Standard non-linear model

\[ X_t = \frac{1}{2} X_{t-1} + 25 \frac{X_{t-1}}{1 + X_{t-1}^2} + 8 \cos(1.2t) + V_t, \quad V_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_V^2), \]

\[ Y_t = \frac{1}{20} X_t^2 + W_t, \quad W_t \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_W^2). \]

- \( T = 200 \) data points with \( \theta = (\sigma_V^2, \sigma_W^2) = (10, 10) \).
- Difficult to perform standard MCMC as \( p(x_{1:T} | y_{1:T}, \theta) \) is highly multimodal.
A Nonlinear State-Space Model

- Standard non-linear model

\[ X_t = \frac{1}{2} X_{t-1} + 25 \frac{X_{t-1}}{1+X_{t-1}^2} + 8 \cos(1.2t) + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_V), \]

\[ Y_t = \frac{1}{20} X_t^2 + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2_W). \]

- \( T = 200 \) data points with \( \theta = (\sigma^2_V, \sigma^2_W) = (10, 10) \).

- Difficult to perform standard MCMC as \( p(x_{1:T} | y_{1:T}, \theta) \) is highly multimodal.

- We sample from \( p(\theta | y_{1:T}) \) using a random walk pseudo-marginal MH where \( p(y_{1:T}; \theta) \) is estimated using SMC with \( N \) particles.
A Nonlinear State-Space Model

Figure: Autocorrelation of $\{\sigma_V^{(i)}\}$ and $\{\sigma_W^{(i)}\}$ of the MH sampler for various $N$. 

(Cambridge, 22/04/14)
A key issue from a practical point of view is how to select $N$?
A key issue from a practical point of view is how to select $N$?

If $N$ is too small, then the algorithm mixes poorly and will require many MCMC iterations.
How to Select the Number of Samples

- A key issue from a practical point of view is how to select \( N \)?
- 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.
Consider the error in the log-likelihood estimator:

\[ Z = \log \hat{p}(y; \theta, U) - \log p(y; \theta) \sim g_\theta(\cdot) \]
Consider the error in the log-likelihood estimator

\[ Z = \log \hat{p}(y; \theta, U) - \log p(y; \theta) \sim g_\theta(\cdot) \]

In the \((\theta, Z)\) parameterization, the target is

\[ \bar{\pi}(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_\theta(u) \Rightarrow \bar{\pi}(\theta, z) = \pi(\theta) \exp(z) g_\theta(z). \]
Consider the error in the log-likelihood estimator

\[ Z = \log \hat{p}(y; \theta, U) - \log p(y; \theta) \sim g_\theta(\cdot) \]

In the \((\theta, Z)\) parameterization, the target is

\[ \pi(\theta, u) = \pi(\theta) \frac{\hat{p}(y; \theta, u)}{p(y; \theta)} m_\theta(u) \Rightarrow \pi(\theta, z) = \pi(\theta) \exp(z) g_\theta(z). \]

The pseudo-marginal MH algorithm is

\[
Q \{(\theta, z), (d\theta, dw)\} = q(\theta|\theta) g_\theta(w) \min \{1, r(\theta, \vartheta) \exp(w - z)\} \, d\theta \, dw \\
+ \{1 - q_Q(\theta, z)\} \delta_{(\theta, z)}(d\theta, dw)
\]

where \(r(\theta, \vartheta) = \pi(\theta) q(\theta|\vartheta) / \{\pi(\theta) q(\vartheta|\theta)\} \).
Consider the Markov chain \( \{\theta_i, Z_i\} \) of transition \( Q \) with invariant distribution \( \pi(\theta, z) \) and \( h : \Theta \rightarrow \mathbb{R} \). Define

\[
\pi(h) := \mathbb{E}_\pi[h(\theta)] \quad \text{and} \quad \hat{\tau}_n(h) = n^{-1} \sum_{i=1}^{n} h(\theta_i).
\]
Consider the Markov chain \( \{\theta_i, Z_i\} \) of transition \( Q \) with invariant distribution \( \pi(\theta, z) \) and \( h : \Theta \to \mathbb{R} \). Define

\[
\pi(h) := \mathbb{E}_\pi[h(\theta)] \quad \text{and} \quad \hat{\tau}_n(h) = n^{-1} \sum_{i=1}^{n} h(\theta_i).
\]

**Proposition** (KV 1986). If \( \{\theta_i, Z_i\} \) is stationary and ergodic, \( \mathbb{V}_\pi[h(\theta)] < \infty \) and

\[
IF_h^Q := 1 + 2 \sum_{\tau=1}^{\infty} corr_{\pi, Q} \{ h(\theta_0), h(\theta_\tau) \}
\]

then

\[
\sqrt{n} \{ \hat{\tau}_n(h) - \pi(h) \} \to \mathcal{N} \left( 0, \mathbb{V}_\pi[h(\theta)] \cdot IF_h^Q \right).
\]
Inefficiency Measure

- Consider the Markov chain \( \{\theta_i, Z_i\} \) of transition \( Q \) with invariant distribution \( \pi(\theta, z) \) and \( h: \Theta \rightarrow \mathbb{R} \). Define

\[
\pi(h) := \mathbb{E}_\pi[h(\theta)] \quad \text{and} \quad \hat{\tau}_n(h) = n^{-1} \sum_{i=1}^{n} h(\theta_i).
\]

- **Proposition** (KV 1986). If \( \{\theta_i, Z_i\} \) is stationary and ergodic, \( \mathbb{V}_\pi[h(\theta)] < \infty \) and

\[
IF^Q_h := 1 + 2 \sum_{\tau=1}^{\infty} \text{corr}_{\pi,Q} \{h(\theta_0), h(\theta_\tau)\}
\]

then

\[
\sqrt{n} \left\{ \hat{\tau}_n(h) - \pi(h) \right\} \rightarrow \mathcal{N}\left(0, \mathbb{V}_\pi[h(\theta)] \text{ IF}^Q_h \right).
\]

- The **Integrated Autocorrelation Time** \( IF^Q_h \) is a measure of inefficiency of \( Q \) which we want to minimize for a fixed computational budget.
Simplifying Assumption: The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $Z \sim \mathcal{N} \left( -\sigma^2/2; \sigma^2 \right)$:

$$
\bar{\pi} (\theta, z) = \pi (\theta) \exp (z) g (z) = \pi (\theta) \mathcal{N} \left( z; \sigma^2/2; \sigma^2 \right).
$$
Aim of the Analysis

- **Simplifying Assumption**: The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $Z \sim \mathcal{N}(-\sigma^2/2; \sigma^2)$:

\[
\overline{\pi}(\theta, z) = \pi(\theta) \exp(z) g(z) = \pi(\theta) \mathcal{N}(z; \sigma^2/2; \sigma^2) \cdot \pi_Z(z)
\]

- **Aim**: Minimize the computational cost

\[
CT^Q_h(\sigma) = IF^Q_h(\sigma) / \sigma^2
\]

as $\sigma^2 \propto 1/N$ and computational efforts proportional to $N$. 
Aim of the Analysis

- **Simplifying Assumption:** The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $Z \sim \mathcal{N} \left( -\sigma^2 / 2; \sigma^2 \right)$:

  $$\tilde{\pi} (\theta, z) = \pi (\theta) \exp (z) g (z) = \pi (\theta) \mathcal{N} (z; \sigma^2 / 2; \sigma^2) .$$

- **Aim:** Minimize the computational cost

  $$CT_h^Q (\sigma) = IF_h^Q (\sigma) / \sigma^2$$

  as $\sigma^2 \propto 1 / N$ and computational efforts proportional to $N$.

- **Special cases:**
Aim of the Analysis

- **Simplifying Assumption:** The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $Z \sim \mathcal{N}(-\sigma^2/2; \sigma^2)$:

  $$\bar{\pi}(\theta, z) = \pi(\theta) \exp(z) g(z) = \pi(\theta) \mathcal{N}(z; \sigma^2/2; \sigma^2).$$

- **Aim:** Minimize the computational cost

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

  as $\sigma^2 \propto 1/N$ and computational efforts proportional to $N$.

- **Special cases:**

  1. When $q(\theta|\theta) = p(\theta|y)$, $\sigma_{opt} = 0.92$ (Pitt et al., 2012).
Aim of the Analysis

- **Simplifying Assumption**: The noise $Z$ is independent of $\theta$ and Gaussian; i.e. $Z \sim \mathcal{N} \left(-\sigma^2/2; \sigma^2\right)$:

  $$\bar{\pi}(\theta, z) = \pi(\theta) \exp(z) g(z) = \pi(\theta) \mathcal{N}(z; \sigma^2/2; \sigma^2).$$

- **Aim**: Minimize the computational cost

  $$CT_h^Q(\sigma) = IF_h^Q(\sigma) / \sigma^2$$

  as $\sigma^2 \propto 1/N$ and computational efforts proportional to $N$.

- **Special cases**:

  1. When $q(\vartheta|\theta) = p(\vartheta|y)$, $\sigma_{opt} = 0.92$ (Pitt et al., 2012).

  2. When $\pi(\theta) = \prod_{i=1}^{d} f(\theta_i)$ and $q(\vartheta|\theta)$ is an isotropic Gaussian random walk then, as $d \to \infty$, $\sigma_{opt} = 1.81$ (Sherlock et al., 2013).
For general proposals and targets, direct minimization of
\[ CT_h^Q (\sigma) = IF_h^Q (\sigma) / \sigma^2 \] impossible so minimize an upper bound over it.
For general proposals and targets, direct minimization of $CT_h^Q (\sigma) = IF_h^Q (\sigma) / \sigma^2$ impossible so minimize an upper bound over it.

We introduce an auxiliary $\pi(\theta, z)$—reversible kernel

$$Q^* \{(\theta, z), (d\theta, dw)\} = q(\theta|\theta)g(w)\alpha_{EX} (\theta, \theta) \alpha_{Z} (z, w) \, d\theta \, dw$$

$$+ \{1 - q_{EX} (\theta) q_{Z} (z)\} \delta_{(\theta, z)} (d\theta, dw)$$

where

$$\alpha_{EX} (\theta, \theta) = \min \{1, r (\theta, \theta)\}, \quad \alpha_{Z} (z, w) = \min \{1, \exp (w - z)\}$$
Sketch of the Analysis

- For general proposals and targets, direct minimization of $CT_h^Q(\sigma) = IF_h^Q(\sigma) / \sigma^2$ impossible so minimize an upper bound over it.

- We introduce an auxiliary $\pi(\theta, z)$—reversible kernel

$$Q^* \{ (\theta, z), (d\theta, dw) \} = q(\theta|\theta)g(w)\alpha_{EX}(\theta, \theta)\alpha_Z(z, w)\, d\theta\, dw$$

$$+ \{1 - q_{EX}(\theta)\, q_Z(z)\} \delta_{(\theta, z)}(d\theta, dw)$$

where

$$\alpha_{EX}(\theta, \theta) = \min \{1, r(\theta, \theta)\} , \quad \alpha_Z(z, w) = \min \{1, \exp(w - z)\}$$

- Peskun’s theorem (1973) guarantees that $IF_h^Q(\sigma) \leq IF_{h}^{Q^*}(\sigma)$ so that $CT_h^Q(\sigma) \leq CT_{h}^{Q^*}(\sigma)$. 

(Cambridge, 22/04/14)
The $Q^*$ chain can be interpreted in the following two step procedure:

1. Propose $\vartheta^q(\vartheta_j | \theta)$ and $w^g(z; w)$.
2. Accept $\vartheta$ w.p. $\alpha_{EX}(\theta; \vartheta) = \frac{\pi(\vartheta)}{\pi(\theta) q(\theta | \vartheta) q(\vartheta | \theta)}$.
3. Accept $w$ w.p. $\alpha_Z(z; w) = \exp(wz)$.
4. $(\vartheta, w)$ is accepted if and only if there is acceptance in both criteria.
The $Q^*$ chain can be interpreted in the following two step procedure:

1. Propose $\vartheta \sim q(\vartheta_j|\theta)$ and $w \sim g(z; \vartheta)$.
2. Accept $\vartheta$ w.p. $\alpha_{\theta} = \pi(\vartheta)/\pi(\theta) q(\theta_j|\vartheta) q(\vartheta_j|\theta)$.
3. Accept $w$ w.p. $\alpha_Z(z; w) = \exp(w z)$.
4. $(\vartheta, w)$ is accepted if and only if there is acceptance in both criteria.

Let $(\theta, Z)$ be the current state of the Markov chain.
The \( Q^* \) chain can be interpreted in the following two step procedure.

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

2. Propose \( \vartheta \sim q(\vartheta | \theta) \) and \( w \sim g(\cdot; \vartheta) \).
• The $Q^*$ chain can be interpreted in the following two step procedure.
• Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\vartheta \sim q(\vartheta | \theta)$ and $w \sim g(\cdot; \vartheta)$.
2. Accept $\vartheta$ w.p. $\alpha_{EX}(\theta; \vartheta) = 1 \wedge \frac{\pi(\vartheta)q(\theta|\vartheta)}{\pi(\theta)q(\vartheta|\theta)}$. 

(Cambridge, 22/04/14)
The $Q^*$ chain can be interpreted in the following two step procedure
Let $(\theta, Z)$ be the current state of the Markov chain.

1. Propose $\vartheta \sim q(\vartheta | \theta)$ and $w \sim g(\cdot; \theta)$.
2. Accept $\vartheta$ w.p. $\alpha_{EX}(\theta; \vartheta) = 1 \wedge \frac{\pi(\vartheta) \cdot q(\theta | \vartheta)}{\pi(\theta) \cdot q(\theta | \theta)}$.
3. Accept $w$ w.p. $\alpha_Z(z; w) = 1 \wedge \exp(w - z)$.
The $Q^*$ chain can be interpreted in the following two step procedure:

1. Propose $\vartheta \sim q(\vartheta|\theta)$ and $w \sim g(\cdot; \vartheta)$.
2. Accept $\vartheta$ w.p. $\alpha_{EX}(\theta; \vartheta) = 1 \wedge \frac{\pi(\vartheta) q(\theta|\vartheta)}{\pi(\theta) q(\vartheta|\theta)}$.
3. Accept $w$ w.p. $\alpha_{Z}(z; w) = 1 \wedge \exp(w - z)$.
4. $(\vartheta, w)$ is accepted if and only if there is acceptance in both criteria.
Sketch of the Analysis

- Let \((\theta_i, Z_i)_{i \geq 1}\) be generated by \(Q^*\).
Let \((\theta_i, Z_i)_{i \geq 1}\) be generated by \(Q^*\).

Denote \((\tilde{\theta}_i, \tilde{Z}_i)_{i \geq 1}\) the accepted proposals and \((\tau_i)_{i \geq 1}\) the associated sojourn times; i.e. \((\tilde{\theta}_1, \tilde{Z}_1) = (\theta_1, Z_1) = \cdots = (\theta_{\tau_1}, Z_{\tau_1})\), \((\tilde{\theta}_2, \tilde{Z}_2) = (\theta_{\tau_1+1}, Z_{\tau_1+1}) = \cdots = (\theta_{\tau_2}, Z_{\tau_2})\) and so on where \((\tilde{\theta}_{i+1}, \tilde{Z}_{i+1}) \neq (\tilde{\theta}_i, \tilde{Z}_i)\) a.s.
Sketch of the Analysis

- Let \((\theta_i, Z_i)_{i \geq 1}\) be generated by \(Q^*\).
- Denote \((\tilde{\theta}_i, \tilde{Z}_i)_{i \geq 1}\) the accepted proposals and \((\tau_i)_{i \geq 1}\) the associated sojourn times; i.e. \((\tilde{\theta}_1, \tilde{Z}_1) = (\theta_1, Z_1) = \cdots = (\theta_{\tau_1}, Z_{\tau_1}),\) 
  \((\tilde{\theta}_2, \tilde{Z}_2) = (\theta_{\tau_1+1}, Z_{\tau_1+1}) = \cdots = (\theta_{\tau_2}, Z_{\tau_2})\) and so on where \((\tilde{\theta}_{i+1}, \tilde{Z}_{i+1}) \neq (\tilde{\theta}_i, \tilde{Z}_i)\) a.s.
- \(IF_{h}^{Q^*}(\sigma)\) can be re-expressed in terms of \(IF_{h/(Q_{EX}Q_{Z})}^{Q^*_j}(\sigma)\) where

\[
Q_j^* \left\{ (\theta, z), (d\vartheta, dw) \right\} = \frac{q(d\vartheta|\theta)\alpha_{EX}(\theta, \vartheta)}{\rho_{EX}(\theta)} \frac{g(dw)\alpha_{Z}(z, w)}{\rho_{Z}(z)}
\]
Sketch of the Analysis

Let \((\theta_i, Z_i)_{i \geq 1}\) be generated by \(Q^*\).

Denote \((\tilde{\theta}_i, \tilde{Z}_i)_{i \geq 1}\) the accepted proposals and \((\tau_i)_{i \geq 1}\) the associated sojourn times; i.e. \((\tilde{\theta}_1, \tilde{Z}_1) = (\theta_1, Z_1) = \cdots = (\theta_{\tau_1}, Z_{\tau_1})\),
\[ (\tilde{\theta}_2, \tilde{Z}_2) = (\theta_{\tau_1+1}, Z_{\tau_1+1}) = \cdots = (\theta_{\tau_2}, Z_{\tau_2}) \]
and so on where
\[ (\tilde{\theta}_{i+1}, \tilde{Z}_{i+1}) \neq (\tilde{\theta}_i, \tilde{Z}_i) \text{ a.s.} \]

\(IF^Q_h(\sigma)\) can be re-expressed in terms of \(IF^{Q_j}_{h/(\varrho_{EX} \varrho_Z)}(\sigma)\) where

\[
Q^*_j \{(\theta, z), (d\vartheta, dw)\} = \frac{q(d\vartheta|\theta) \alpha_{EX}(\theta, \vartheta)}{\varrho_{EX}(\theta)} \frac{g(dw) \alpha_Z(z, w)}{\varrho_Z(z)}
\]

Proof exploits spectral representations of these kernels + Positivity of \(Q^Z_j(z, dw)\).
Main Results

- **Proposition:**

\[
\frac{IF_h^Q(\sigma)}{IF_h^{EX}} \leq \frac{IF_h^{Q^*}(\sigma)}{IF_h^{EX}} \leq \frac{1}{2} \left(1 + \frac{1}{IF_h^{EX}}\right) \left(1 + IF_h^{Z}(\sigma)\right) - \frac{1}{IF_h^{EX}}
\]

and the bound is tight as \(IF_h^{EX} \to 1\) or \(\sigma \to 0\).

As \(IF_{J,h/\rho_{EX}}^{EX} \to \infty\),

\[
\frac{IF_h^{Q^*}(\sigma)}{IF_h^{EX}} \to \frac{1}{\pi^\sigma_Z(q_Z(\sigma))}.
\]
Main Results

- **Proposition:**

\[
\frac{IF^Q_h(\sigma)}{IF^EX_h} \leq \frac{IF^{Q*}_h(\sigma)}{IF^EX_h} \leq \frac{1}{2} \left(1 + \frac{1}{IF^EX_h}\right) \left(1 + IF^Z(\sigma)\right) - \frac{1}{IF^EX_h}
\]

and the bound is tight as \(IF^EX_h \to 1\) or \(\sigma \to 0\).

As \(IF^EX_{J,h/\varrho_{EX}} \to \infty\),

\[
\frac{IF^{Q*}_h(\sigma)}{IF^EX_h} \to \frac{1}{\pi^\sigma_Z(\varrho_Z(\sigma))}.
\]

- Results used to minimize w.r.t \(\sigma\) upper bounds on \(CT^Q_h(\sigma) = IF^Q_h(\sigma) / \sigma^2\).
Bounds on Relative Computational Costs

**Figure:** Bounds on $\frac{IF_h^Q(\sigma)}{\left(\sigma^2 IF_h^{EX}\right)}$
For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$. 

When you have no clue about the proposal efficiency, 

1. If $\sigma_{opt} = 1$ and you pick $\sigma = 1.7$, computing time increases by 150%.

2. If $\sigma_{opt} = 1.7$ and you pick $\sigma = 1$, computing time increases by 50%.

3. If $\sigma_{opt} = 1$ or $\sigma_{opt} = 1.7$ and you pick $\sigma = 1.2$, computing time increases by 15%.
For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.

When you have no clue about the proposal efficiency,
For good proposals, select \( \sigma \approx 1 \) whereas for poor proposals, select \( \sigma \approx 1.7 \).

When you have no clue about the proposal efficiency,

1. If \( \sigma_{\text{opt}} = 1 \) and you pick \( \sigma = 1.7 \), computing time increases by \( \approx 150\% \).
For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.

When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1$ and you pick $\sigma = 1.7$, computing time increases by $\approx 150\%$.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1$, computing time increases by $\approx 50\%$. 
For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.

When you have no clue about the proposal efficiency,

1. If $\sigma_{\text{opt}} = 1$ and you pick $\sigma = 1.7$, computing time increases by $\approx 150\%$.
2. If $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1$, computing time increases by $\approx 50\%$.
3. If $\sigma_{\text{opt}} = 1$ or $\sigma_{\text{opt}} = 1.7$ and you pick $\sigma = 1.2$, computing time increases by $\approx 15\%$. 
Consider

\[ X_t = \mu(1 - \phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \eta), \]

\[ Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2 \epsilon), \]

where \( \theta = (\phi, \mu, \sigma^2 \eta) \).
Example: Noisy Autoregressive Example

- Consider

\[ X_t = \mu(1 - \phi) + \phi X_t + V_t, \quad V_t \sim \mathcal{N}(0, \sigma^2_\eta), \]

\[ Y_t = X_t + W_t, \quad W_t \sim \mathcal{N}(0, \sigma^2_\varepsilon), \]

where \( \theta = (\phi, \mu, \sigma^2_\eta) \).

- Likelihood can be computed exactly using Kalman.
Example: Noisy Autoregressive Example

Consider

\[ X_t = \mu(1 - \phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2_{\eta}\right), \]

\[ Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2_\varepsilon\right), \]

where \( \theta = \left(\phi, \mu, \sigma^2_{\eta}\right) \).

- Likelihood can be computed exactly using Kalman.
- Autoregressive Metropolis proposal of coefficient \( \rho \) for \( \theta \) based on multivariate t-distribution.
Consider

\[ X_t = \mu (1 - \phi) + \phi X_t + V_t, \quad V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N} \left( 0, \sigma^2_\eta \right), \]

\[ Y_t = X_t + W_t, \quad W_t \overset{\text{i.i.d.}}{\sim} \mathcal{N} \left( 0, \sigma^2_\varepsilon \right), \]

where \( \theta = \left( \phi, \mu, \sigma^2_\eta \right). \)

- Likelihood can be computed exactly using Kalman.
- Autoregressive Metropolis proposal of coefficient \( \rho \) for \( \theta \) based on multivariate t-distribution.
- \( N \) is selected so as to obtain \( \sigma(\bar{\theta}) \approx \text{constant} \) where \( \bar{\theta} \) posterior mean.
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 $\sigma = 0.92$. 
Relative Inefficiency and Computing Time

Figure: From left to right: $RCT_h^Q$ vs $N$, $RCT_h^Q$ vs $\sigma(\bar{\theta})$, $RIF_h^Q$ against $N$ and $RIF_h^Q$ against $\sigma(\bar{\theta})$ for various values of $\rho$ and different parameters.
Acceptance Probability

![Graph showing acceptance probability against different values of ρ (0, 0.4, 0.6, 0.9)]

- ρ = 0
- ρ = 0.4
- ρ = 0.6
- ρ = 0.9

Figure: Acceptance probability against σ. Theoretical (upper bound) vs Actual.

(Cambridge, 22/04/14)
Simplified quantitative analysis of pseudo-marginal MCMC.
Discussion

- Simplified quantitative analysis of pseudo-marginal MCMC.

1 Much weaker assumptions than previous work on proposal and target.
Discussion

- Simplified quantitative analysis of pseudo-marginal MCMC.
  1. Much weaker assumptions than previous work on proposal and target.
  2. Assumption on orthogonality of the noise.
Discussion

- Simplified quantitative analysis of pseudo-marginal MCMC.
  
  1 Much weaker assumptions than previous work on proposal and target.
  
  2 Assumption on orthogonality of the noise.
  
  For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$. 

Discussion

- Simplified quantitative analysis of pseudo-marginal MCMC.
  
  1. Much weaker assumptions than previous work on proposal and target.
  
  2. Assumption on orthogonality of the noise.

- For good proposals, select $\sigma \approx 1$ whereas for poor proposals, select $\sigma \approx 1.7$.

- In scenarios where the proposal efficiency is unknown, select $\sigma \approx 1.2$. 