Monte Carlo Methods

for Econometric Inference I

Institute on Computational Economics

July 19, 2006

John Geweke, University of Iowa

The problem

$p(\boldsymbol{\theta}, \boldsymbol{\omega} \mid I)$ Distribution of interest

$\boldsymbol{\theta} \in \Theta$, $\boldsymbol{\omega} \in \Omega$; $I =$ "Information"

$$p(\boldsymbol{\theta}, \boldsymbol{\omega} \mid I) = p(\boldsymbol{\theta} \mid I) \cdot p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)$$

$p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)$ Tractable for simulation

$p(\boldsymbol{\theta} \mid I)$ Not so easy

Leading example:

$$I = \{\text{Model specification}\} \cup \{\text{Data}\}$$

Origins of the problem in econometric inference

Complete model

$$\left.\begin{array}{r} p\left(\mathbf{y} \mid \boldsymbol{\theta}, A\right) \\ p\left(\boldsymbol{\theta} \mid A\right) \end{array}\right\} \Longrightarrow p\left(\boldsymbol{\theta} \mid \mathbf{y}^o, A\right)$$

Vector of interest $\boldsymbol{\omega}$

$$p\left(\boldsymbol{\omega} \mid \mathbf{y}^o, \boldsymbol{\theta}, A\right)$$

Simulation problem

$$\begin{array}{rcl} \boldsymbol{\theta}^{(m)} & \sim & p\left(\boldsymbol{\theta} \mid \mathbf{y}^o, A\right) \\ \boldsymbol{\omega}^{(m)} & \sim & p\left(\boldsymbol{\omega} \mid \mathbf{y}^o, \boldsymbol{\theta}, A\right) \end{array}$$

Direct Sampling

Some familiar tools

Strong law of large numbers:

If $\boldsymbol{\omega}^{(m)} \overset{iid}{\sim} p(\omega)$ and $\mathsf{E}[h(\boldsymbol{\omega})] = \int h(\boldsymbol{\omega}) p(\omega) d\nu(\boldsymbol{\omega})$ exists, then

$$\overline{h}^{(M)} = M^{-1} \sum_{m=1}^{M} h\left(\boldsymbol{\omega}^{(m)}\right) \overset{a.s.}{\to} \mathsf{E}[h(\boldsymbol{\omega})] = \overline{h};$$

that is, $P\left\{\lim_{M\to\infty} \overline{h}^{(M)} = \overline{h}\right\} = 1$, a stronger condition than $\overline{h}^{(M)} \overset{p}{\to} \overline{h}$.

Lindeberg-Levy central limit theorem:

If in addition $\sigma_h^2 = \mathsf{var}\left[h\left(\boldsymbol{\omega}\right)\right]$ exists, then

$$M^{1/2}\left[\overline{h}^{(M)} - \overline{h}\right] \xrightarrow{d} N\left(0, \sigma_h^2\right).$$

Generic notation for posterior simulation

$$\begin{aligned} \boldsymbol{\theta} &\sim p(\boldsymbol{\theta} \mid I);\ \boldsymbol{\theta} \in \Theta \\ \boldsymbol{\omega} &\sim p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I);\ \boldsymbol{\omega} \in \Omega \end{aligned}$$

"I" denotes the distribution of interest.

Direct sampling is possible if there are practical algorithms for

$$\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p\left(\boldsymbol{\theta} \mid I\right) \quad \text{and} \quad \boldsymbol{\omega}^{(m)} \overset{iid}{\sim} p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}, I\right).$$

Example:

(1) Mother of all random number generators (Geweke 1996):

$$u^{(m)} \overset{iid}{\sim} uniform\left(0,1\right)$$

(2) Inverse c.d.f. transformation to simulate θ : $P\left(\theta \leq c\right) = F\left(c\right)$:

$$\begin{aligned} \theta^{(m)} &= F^{-1}\left(u^{(m)}\right) \\ \theta^{(m)} &\leq c \iff u^{(m)} = F\left(\theta^{(m)}\right) \leq F\left(c\right) \\ &\Longrightarrow P\left(\theta^{(m)} \leq c\right) = P\left[u^{(m)} \leq F\left(c\right)\right] = F\left(c\right) \end{aligned}$$

General case:

$$\boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid I\right) \quad \text{and} \quad \boldsymbol{\omega}^{(m)} \mid \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}, I\right)$$

where $\boldsymbol{\theta}^{(m)} \in \Theta$, $\boldsymbol{\omega}^{(m)} \in \Omega$, and all draws are independent.

For given h and M, let $\overline{h}^{(M)} = M^{-1} \sum_{m=1}^{M} h\left(\boldsymbol{\omega}^{(m)}\right)$.

Theorem 4.1(a): $\boldsymbol{\omega}^{(m)} \overset{i.i.d.}{\sim} p\left(\boldsymbol{\omega} \mid I\right)$

Theorem 4.1(b): If $\overline{h} = \mathsf{E}\left[h\left(\boldsymbol{\omega}\right) \mid I\right]$ exists then $\overline{h}^{(M)} \overset{a.s.}{\to} \overline{h}$ (strong law of large numbers).

Theorem 4.1(c): If $\overline{h} = \mathsf{E}\left[h\left(\boldsymbol{\omega}\right) \mid I\right]$ and $\mathsf{var}\left[h\left(\boldsymbol{\omega}\right) \mid I\right] = \sigma^2$ exist,

then $M^{1/2}\left(\overline{h}^{(M)} - \overline{h}\right) \xrightarrow{d} N\left(0, \sigma^2\right)$ and

$$\widehat{\sigma}^{2(M)} = M^{-1} \sum_{m=1}^{M} \left[h\left(\boldsymbol{\omega}^{(m)}\right) - \overline{h}^{(M)}\right]^2 \xrightarrow{a.s.} \sigma^2.$$

(Lindeberg-Levy central limit theorem and strong law of large numbers)

Theorem 4.1(d): If for given $p \in (0,1)$, there is a unique h_p such that the statements

$$P\left[h\left(\boldsymbol{\omega}\right) \leq h_p \mid I\right] \geq p \text{ and } P\left[h\left(\boldsymbol{\omega}\right) \geq h_p \mid I\right] \geq 1-p$$

are both true, then

$$\widehat{h}_p^{(M)} \overset{a.s.}{\rightarrow} h_p$$

where $\widehat{h}_p^{(M)}$ is any real number such that

$$M^{-1}\sum_{m=1}^{M} I_{\left(-\infty,\widehat{h}_p^{(M)}\right]}\left[h\left(\boldsymbol{\omega}^{(m)}\right)\right] \geq p;\ \ M^{-1}\sum_{m=1}^{M} I_{\left[\widehat{h}_p^{(M)},\infty\right)}\left[h\left(\boldsymbol{\omega}^{(m)}\right)\right] \geq 1-p$$

(Follows from in Rao (1965), result 6f.2(i).)

Theorem 4.1(e): If for given $p \in (0,1)$, there is a unique h_p such that the statements

$$P\left[h\left(\boldsymbol{\omega}\right) \leq h_p \mid I\right] \geq p \text{ and } P\left[h\left(\boldsymbol{\omega}\right) \geq h_p \mid I\right] \geq 1-p$$

are both true, and moreover p, $p\left[h\left(\boldsymbol{\omega}\right) = h_p \mid I\right] > 0$, then

$$M^{1/2}\left[\widehat{h}_p^{(M)} - h_p\right] \xrightarrow{d} N\left\{0, p\left(1-p\right)/p\left[h\left(\boldsymbol{\omega}\right) = h_p \mid I\right]^2\right\}.$$

(Follows from in Rao (1965), result 6f.2(i).)

Approximation of Bayes actions by direct sampling – General case

The setting:

- $\left\{\boldsymbol{\theta}^{(m)}, \boldsymbol{\omega}^{(m)}\right\}$ i.i.d. with $\boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid I\right)$, $\boldsymbol{\omega}^{(m)} \mid \left(\boldsymbol{\theta}^{(m)}, I\right) \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}, I\right)$.
- Loss function $L\left(\mathbf{a}, \boldsymbol{\omega}\right) \geq 0$ defined on $\Omega \times A$, A an open subset of $\mathbb{R}^q$.
- The risk function

$$R\left(\mathbf{a}\right) = \int_{\Omega} \int_{\Theta} L\left(\mathbf{a}, \boldsymbol{\omega}\right) p\left(\boldsymbol{\theta} \mid I\right) p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I\right) d\boldsymbol{\theta} d\boldsymbol{\omega}$$

  has a strict global minimum at $\widehat{\mathbf{a}} \in A \subseteq \mathbb{R}^m$.

Let A_M be the set of all roots of $M^{-1}\sum_{m=1}^{M}\partial L\left(\mathbf{a},\boldsymbol{\omega}^{(m)}\right)/\partial\mathbf{a}=\mathbf{0}$.

Theorem 4.2(a): If

(1) $M^{-1}\sum_{m=1}^{M}L\left(\mathbf{a},\boldsymbol{\omega}^{(m)}\right)$ converges uniformly to $R(\mathbf{a})$ for all $\mathbf{a}\in N(\widehat{\mathbf{a}})$, almost surely;

(2) $\partial L(\mathbf{a},\boldsymbol{\omega})/\partial\mathbf{a}$ exists and is a continuous function of $\mathbf{a}$, for all $\boldsymbol{\omega}\in\boldsymbol{\Omega}$ and all $\mathbf{a}\in N(\widehat{\mathbf{a}})$;

then for any $\varepsilon>0$

$$\lim_{M\to\infty}P\left[\inf_{\mathbf{a}\in A_M}(\mathbf{a}-\widehat{\mathbf{a}})'(\mathbf{a}-\widehat{\mathbf{a}})>\varepsilon\mid I\right]=0.$$

Theorem 4.2 (b): If, in addition,

(3) $\partial^2 L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a} \partial \mathbf{a}'$ exists and is a continuous function of $\mathbf{a}$, for all $\boldsymbol{\omega} \in \Omega$ and all $\mathbf{a} \in \mathbf{N}(\widehat{\mathbf{a}})$;

(4) $\mathbf{B} = var\left[\partial L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a}|_{\mathbf{a}=\widehat{\mathbf{a}}} \mid I\right]$ exists and is finite;

(5) $\mathbf{H} = E\left[\partial^2 L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a} \partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}} \mid I\right]$ exists and is finite and nonsingular;

(6) For any $\varepsilon > 0$, there exists M_ε such that

$$P\left[\sup_{\mathbf{a} \in N(\widehat{\mathbf{a}})} \left|\partial^3 L(\mathbf{a}, \boldsymbol{\omega}) / \partial a_i \partial a_j \partial a_k\right| < M_\varepsilon \mid I\right] \geq 1 - \varepsilon$$

for all $i, j, k = 1, \ldots, m$;

then:

1. $M^{1/2}\left(\widehat{\mathbf{a}}_M - \widehat{\mathbf{a}}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}\right)$,

2. $M^{-1}\sum_{m=1}^{M} \partial L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \cdot \partial L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \xrightarrow{p} \mathbf{B}$,

3. $M^{-1}\sum_{m=1}^{M} \partial^2 L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}\partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \xrightarrow{p} \mathbf{H}$.

... This all follows from Amemiya (1985), Chapter 4, in straightforward fashion.

Acceptance and Importance Sampling

Motivation and approach

We have no practical method to simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid I)$.

We can simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid S)$, where $p(\boldsymbol{\theta} \mid S)$ is "similar" to $p(\boldsymbol{\theta} \mid I)$.

Can we use $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid S)$ to simulate from $p(\boldsymbol{\theta} \mid I)$?

Yes, **but** the sense in which $p(\boldsymbol{\theta} \mid S)$ is similar to $p(\boldsymbol{\theta} \mid I)$ is critical.

Acceptance Sampling

Density of interest: $p(\boldsymbol{\theta} \mid I)$

We cannot simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid I)$,

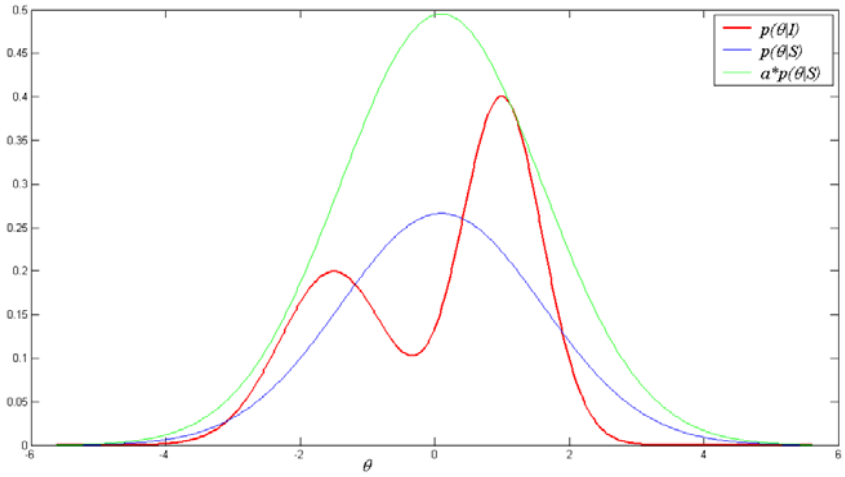
We can evaluate $k(\boldsymbol{\theta} \mid I) \propto p(\boldsymbol{\theta} \mid I)$

Source density: $p(\boldsymbol{\theta} \mid S)$

We can simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid S)$

We can evaluate $k(\boldsymbol{\theta} \mid S) \propto p(\boldsymbol{\theta} \mid S)$

Acceptance sampling: Key idea



$$\sup_\theta p(\theta \mid I)/p(\boldsymbol{\theta} \mid S) = p(1.16 \mid I)/p(1.16 \mid S) = a = 1.86$$

Draw θ^* from $p(\theta|S)$, and accept the draw with probability $p(\theta^*|I) / [a \cdot p(\theta^*|S)]$

Theorem 4.2.1 Acceptance sampling. Let

$$k(\boldsymbol{\theta}|I) = c_I \cdot p(\boldsymbol{\theta}|I)$$

be a kernel of the density of interest $p(\boldsymbol{\theta}|I)$ and let

$$k(\boldsymbol{\theta}|S) = c_S \cdot p(\boldsymbol{\theta}|S)$$

be a kernel of the source density $p(\boldsymbol{\theta}|S)$. Let $r = \sup_{\boldsymbol{\theta}\in\Theta} k(\boldsymbol{\theta}|I)/k(\boldsymbol{\theta}|S) < \infty$. Suppose that $\boldsymbol{\theta}^{(m)}$ is drawn as follows.

(1) Draw u uniform on $[0, 1]$.
(2) Draw $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|S)$.
(3) If $u > k(\boldsymbol{\theta}^*|I) / [r \cdot k(\boldsymbol{\theta}^*|S)]$ then return to step 1.
(4) Set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*$.

If the draws in steps 1 and 2 are independent, then $\boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}|I)$.

$$k(\boldsymbol{\theta}\,|I) = c_I \cdot p(\boldsymbol{\theta}\,|I),\;\; k(\boldsymbol{\theta}\,|S) = c_S \cdot p(\boldsymbol{\theta}\,|S),\;\; \Theta^* = \{\boldsymbol{\theta} : p(\boldsymbol{\theta}\,|S) > 0\}$$

(1) Draw u uniform on $[0, 1]$.
(2) Draw $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}\,|S)$.
(3) If $u > k(\boldsymbol{\theta}^*|I) \,/\, [r \cdot k(\boldsymbol{\theta}^*|S)]$ then return to step 1.
(4) Set $\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*$.

$$P[(3) \rightarrow (4)] = \int_{\boldsymbol{\Theta}^*} \{k(\boldsymbol{\theta}\,|I) \,/\, [rk(\boldsymbol{\theta}\,|S)]\}\, p(\boldsymbol{\theta}\,|S)\, d\nu(\boldsymbol{\theta}) = c_I / r c_S$$

$$\begin{aligned} P[(3) \rightarrow (4),\, \boldsymbol{\theta} \in A] &= \int_A \{k(\boldsymbol{\theta}\,|I) \,/\, [rk(\boldsymbol{\theta}\,|S)]\}\, p(\boldsymbol{\theta}\,|S)\, d\nu(\boldsymbol{\theta}) \\ &= \int_A k(\boldsymbol{\theta}\,|I)\, d\nu(\boldsymbol{\theta}) / r c_S \\ P[\,\boldsymbol{\theta} \in A \mid (3) \rightarrow (4)] &= \int_A k(\boldsymbol{\theta}\,|I)\, d\nu(\boldsymbol{\theta}) / c_I = \int_A p(\boldsymbol{\theta}\,|I)\, d\nu(\boldsymbol{\theta}) \\ &= P(\boldsymbol{\theta} \in A \mid I) \end{aligned}$$

Efficiency of acceptance sampling

$$k(\boldsymbol{\theta}\,|I) = c_I p(\boldsymbol{\theta}\,|I), k(\boldsymbol{\theta}\,|S) = c_S p(\boldsymbol{\theta}\,|S), r = \sup_{\boldsymbol{\theta}\in\Theta} \frac{k(\boldsymbol{\theta}\,|I)}{k(\boldsymbol{\theta}\,|S)}, \quad a = \sup_{\theta\in\Theta} \frac{p(\theta \mid I)}{p(\boldsymbol{\theta}\,|S)}$$

$$\begin{aligned} P[(3) \to (4)] &= \int_{\boldsymbol{\Theta}^*} \{k(\boldsymbol{\theta}\,|I) / [rk(\boldsymbol{\theta}\,|S)]\} p(\boldsymbol{\theta}\,|S)\, d\nu(\boldsymbol{\theta}) = \frac{c_I}{rc_S} \\ &= \frac{c_I}{c_S \cdot \sup_{\boldsymbol{\theta}\in\Theta} k(\boldsymbol{\theta}\,|I)/k(\boldsymbol{\theta}\,|S)} = \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{c_I}{c_S \cdot k(\boldsymbol{\theta}\,|I)/k(\boldsymbol{\theta}\,|S)} \\ &= \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{p(\boldsymbol{\theta}\,|S)}{p(\boldsymbol{\theta}\,|I)} = a^{-1} \end{aligned}$$

$$\frac{\text{Draws from } p(\boldsymbol{\theta}\,|S)}{\text{Draws from } p(\boldsymbol{\theta}\,|I)} \to a$$

Example:

Sampling from a standard normal distribution truncated to the interval (a, b)

Characteristics of the truncation			Source distribution
$a < 0 < b < \infty$	$a \geq -t_1$ and $b \leq t_1$		Uniform(a, b)
	$a < -t_1$ or $b > t_1$		$N(0, 1)$
$0 \leq a < b < \infty$	$\phi(a)/\phi(b) \leq t_2$		Uniform(a, b)
	$\phi(a)/\phi(b) > t_2$	$a \leq t_3$	$\lvert N(0, 1)\rvert$
		$a > t_3$	$a + exp\left(a^{-1}\right)$
$b = \infty$	$a \leq t_4$		$N(0, 1)$
	$a > t_4$		$a + exp\left(a^{-1}\right)$
$t_1 = 0.375$, $t_2 = 2.18$, $t_3 = 0.725$, and $t_4 = 0.45$			
2 to 6 times as fast as inverse c.d.f. method			

Importance Sampling

Density of interest: $p(\boldsymbol{\theta} \mid I)$

We cannot simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid I)$,

We can evaluate $k(\boldsymbol{\theta} \mid I) \propto p(\boldsymbol{\theta} \mid I)$

Source density: $p(\boldsymbol{\theta} \mid S)$

We can simulate $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid S)$

We can evaluate $k(\boldsymbol{\theta} \mid S) \propto p(\boldsymbol{\theta} \mid S)$

Weighting function: $w(\boldsymbol{\theta}) = k(\boldsymbol{\theta} \mid I) / k(\boldsymbol{\theta} \mid S)$

Theorem 4.2 (a,b) Approximation of moments by importance sampling.

Suppose $\mathsf{E}[h(\boldsymbol{\omega}) \mid I] = \overline{h}$ exists, the support of $p(\boldsymbol{\theta} \mid S)$ includes Θ, and $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}, I\right)$. Then

$$\overline{h}^{(M)} = \frac{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) h\left(\boldsymbol{\omega}^{(m)}\right)}{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)} \overset{a.s.}{\to} \overline{h}.$$

Proof.

$$w\left(\boldsymbol{\theta}^{(m)}\right) \text{ i.i.d., } \mathsf{E}[w(\boldsymbol{\theta}) \mid S] = \int_{\Theta} \frac{k(\boldsymbol{\theta} \mid I)}{k(\boldsymbol{\theta} \mid S)} p(\boldsymbol{\theta} \mid S)\, d\nu(\boldsymbol{\theta}) = \frac{c_I}{c_S} = \overline{w}$$

$$\Longrightarrow \quad \overline{w}^{(M)} = M^{-1} \sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) \overset{a.s.}{\to} \overline{w}.$$

Proof (concluded) So far we have

$$\overline{h}^{(M)} = \frac{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) h\left(\boldsymbol{\omega}^{(m)}\right)}{\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)}, \quad \overline{w}^{(M)} = M^{-1} \sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) \overset{a.s.}{\to} \overline{w}$$

Then $w(\boldsymbol{\theta}) = k(\boldsymbol{\theta} \mid I)/k(\boldsymbol{\theta} \mid S)$, $\left\{w\left(\boldsymbol{\theta}^{(m)}\right),\ h\left(\boldsymbol{\omega}^{(m)}\right)\right\}$ i.i.d. $\Longrightarrow$

$$\begin{aligned}
\mathsf{E}\left[w(\boldsymbol{\theta})\, h(\boldsymbol{\omega}) \mid S\right] &= \int_{\Theta} w(\boldsymbol{\theta}) \left[\int_{\Omega} h(\boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)\, d\nu(\boldsymbol{\omega})\right] p(\boldsymbol{\theta} \mid S)\, d\nu(\boldsymbol{\theta}) \\
&= (c_I/c_S) \int_{\Theta} \int_{\Omega} h(\boldsymbol{\omega})\, p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)\, p(\boldsymbol{\theta} \mid I)\, d\nu(\boldsymbol{\omega})\, d\nu(\boldsymbol{\theta}) \\
&= (c_I/c_S)\, \mathsf{E}\left[h(\boldsymbol{\omega}) \mid I\right] = \overline{w} \cdot \overline{h} \\
\Longrightarrow \quad \overline{wh}^{(M)} &= M^{-1} \sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) h\left(\boldsymbol{\omega}^{(m)}\right) \overset{a.s.}{\to} \overline{w} \cdot \overline{h}
\end{aligned}$$

Theorem 4.2 (c,d) Suppose

$$\mathsf{E}\left[h\left(\boldsymbol{\omega}\right) \mid I\right] = \overline{h} \quad \text{and} \quad \mathsf{var}\left[h\left(\boldsymbol{\omega}\right) \mid I\right] = \sigma^2$$

exist, and

$$w\left(\boldsymbol{\theta}\right) = k\left(\boldsymbol{\theta} \mid I\right) / k\left(\boldsymbol{\theta} \mid S\right)$$

is bounded above on Θ. Then

$$M^{1/2}\left(\overline{h}^{(M)} - \overline{h}\right) \xrightarrow{d} N\left(0, \tau^2\right)$$

and

$$\widehat{\tau}^{2(M)} = \frac{M \sum_{m=1}^{M} \left[h\left(\boldsymbol{\omega}^{(m)}\right) - \overline{h}^{(M)}\right]^2 w\left(\boldsymbol{\theta}^{(m)}\right)^2}{\left[\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)\right]^2} \xrightarrow{a.s.} \tau^2.$$

Proof.

$$\begin{aligned}
\mathsf{E}\left[w(\boldsymbol{\theta})^2 h(\boldsymbol{\omega})^2 \mid S\right] &= \int_\Theta w(\boldsymbol{\theta})^2 \left[\int_\Omega h(\boldsymbol{\omega})^2 p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)\, d\nu(\boldsymbol{\omega})\right] p(\boldsymbol{\theta} \mid S)\, d\nu(\boldsymbol{\theta}) \\
&= \int_\Theta w(\boldsymbol{\theta}) \frac{k(\boldsymbol{\theta} \mid I)}{k(\boldsymbol{\theta} \mid S)} \left[\int_\Omega h(\boldsymbol{\omega})^2 p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)\, d\nu(\boldsymbol{\omega})\right] p(\boldsymbol{\theta} \mid S)\, d\nu(\boldsymbol{\theta}) \\
&= (c_I/c_S) \int_\Theta w(\boldsymbol{\theta}) \left[\int_\Omega h(\boldsymbol{\omega})^2 p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I)\, d\nu(\boldsymbol{\omega})\right] p(\boldsymbol{\theta} \mid I)\, d\nu(\boldsymbol{\theta}) \\
&\leq (c_I/c_S)\, \mathsf{E}\left[h(\boldsymbol{\omega})^2 \mid I\right] \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} w(\boldsymbol{\theta}) < \infty
\end{aligned}$$

Substitute $h(\boldsymbol{\omega}) = 1$ and conclude $\mathsf{E}\left[w(\boldsymbol{\theta})^2 \mid S\right] < \infty$ as well.

Proof (continued). Now we know

$$\mathbf{V} = \text{var}\left[\begin{array}{c} w(\boldsymbol{\theta}) \\ w(\boldsymbol{\theta}) h(\boldsymbol{\omega}) \end{array} \mid S \right]$$

is finite. Apply Lindeberg-Levy Central Limit Theorem:

$$M^{1/2}\left[\begin{pmatrix} \overline{w}^{(M)} \\ \overline{wh}^{(M)} \end{pmatrix} - \begin{pmatrix} \overline{w} \\ \overline{w}\overline{h} \end{pmatrix} \right] \xrightarrow{d} N(\mathbf{0}, \mathbf{V})$$

Proof (continued)

$$M^{1/2}\left[\begin{pmatrix} \overline{w}^{(M)} \\ \overline{wh}^{(M)} \end{pmatrix} - \begin{pmatrix} \overline{w} \\ \overline{w}\overline{h} \end{pmatrix}\right] \xrightarrow{d} N(\mathbf{0}, \mathbf{V}),$$

$$\frac{\overline{wh}^{(M)}}{\overline{w}^{(M)}} = \frac{\overline{w}\overline{h}}{\overline{w}} + \frac{\overline{wh}^{(M)} - \overline{w}\overline{h}}{\overline{w}} - \frac{\overline{w}\overline{h} \cdot \left(\overline{w}^{(M)} - \overline{w}\right)}{\overline{w}^2} + o_p\left(M^{-1/2}\right)$$

$$\Longrightarrow \quad M^{1/2}\left(\frac{\overline{wh}^{(M)}}{\overline{w}^{(M)}} - \overline{h}\right) \xrightarrow{d} N\left(0, \tau^2\right)$$

with

$$\tau^2 = \overline{w}^{-2}\left\{\mathsf{var}\left[w(\boldsymbol{\theta})\, h(\boldsymbol{\omega}) \mid S\right] - 2\overline{h}\mathsf{cov}_S\left[w(\boldsymbol{\theta})\, h(\boldsymbol{\omega}), w(\boldsymbol{\theta}) \mid S\right] + \overline{h}^2\mathsf{var}_S\left[w(\boldsymbol{\theta}) \mid S\right]\right\} = \overline{w}^{-2}\mathsf{var}\left\{\left[w(\boldsymbol{\theta})\, h(\boldsymbol{\omega}) - \overline{h}w(\boldsymbol{\theta})\right] \mid S\right\}.$$

Proof (concluded)

$$\tau^2 = \overline{w}^{-2}\left\{\mathsf{var}\left[w\left(\boldsymbol{\theta}\right)h\left(\boldsymbol{\omega}\right)\mid S\right] - 2\overline{h}\mathsf{cov}_S\left[w\left(\boldsymbol{\theta}\right)h\left(\boldsymbol{\omega}\right), w\left(\boldsymbol{\theta}\right)\mid S\right]\right.$$
$$\left.+\overline{h}^2\mathsf{var}_S\left[w\left(\boldsymbol{\theta}\right)\mid S\right]\right\} = \overline{w}^{-2}\mathsf{var}\left\{\left[w\left(\boldsymbol{\theta}\right)h\left(\boldsymbol{\omega}\right) - \overline{h}w\left(\boldsymbol{\theta}\right)\right]\mid S\right\}$$

$$\begin{aligned}\widehat{\tau}^{2(M)} &= \frac{M\sum_{m=1}^{M}\left[h\left(\boldsymbol{\omega}^{(m)}\right) - \overline{h}^{(M)}\right]^2 w\left(\boldsymbol{\theta}^{(m)}\right)^2}{\left[\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)\right]^2} \\ &= \frac{M^{-1}\sum_{m=1}^{M}\left[w\left(\boldsymbol{\theta}^{(m)}\right)h\left(\boldsymbol{\omega}^{(m)}\right) - \overline{h}^{(M)}w\left(\boldsymbol{\theta}^{(m)}\right)\right]^2}{\left[M^{-1}\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)\right]^2} \xrightarrow{a.s.} \tau^2\end{aligned}$$

Example: Reweighting to a different prior distribution

Model	Prior	Likelihood
A_1	$p(\boldsymbol{\theta}_A \mid A_1)$	$p(\mathbf{y} \mid \boldsymbol{\theta}_A, A)$
A_2	$p(\boldsymbol{\theta}_A \mid A_2)$	$p(\mathbf{y} \mid \boldsymbol{\theta}_A, A)$

$$\begin{aligned} p(\boldsymbol{\theta}_A \mid \mathbf{y}^o, A_2) &= p(\boldsymbol{\theta}_A \mid I) \\ \boldsymbol{\theta}_A^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta}_A \mid \mathbf{y}^o, A_1) &= p(\boldsymbol{\theta}_A \mid S) \end{aligned}$$

$$w(\boldsymbol{\theta}_A) = \frac{p(\boldsymbol{\theta}_A \mid \mathbf{y}^o, A_2)}{p(\boldsymbol{\theta}_A \mid \mathbf{y}^o, A_1)} = \frac{p(\boldsymbol{\theta}_A \mid A_1)\, p(\mathbf{y} \mid \boldsymbol{\theta}_A, A)}{p(\boldsymbol{\theta}_A \mid A_2)\, p(\mathbf{y} \mid \boldsymbol{\theta}_A, A)} = \frac{p(\boldsymbol{\theta}_A \mid A_1)}{p(\boldsymbol{\theta}_A \mid A_2)}$$

A useful summary measure of computational efficiency

$$\text{var}\left[h\left(\boldsymbol{\omega}\right) \mid I\right] = \sigma^2 \quad \text{var}\left[\overline{h}^{(M)} \mid S\right] = \frac{\tau^2}{M}$$

$$w\left(\boldsymbol{\theta}\right) \propto 1 \; \forall \; \boldsymbol{\theta} \iff p\left(\boldsymbol{\theta} \mid S\right) = p\left(\boldsymbol{\theta} \mid I\right) \implies \tau^2 = \sigma^2$$

Relative numerical efficiency (*RNE*) of the simulator is

$$RNE = \frac{\sigma^2}{\tau^2}.$$

$$\frac{\sigma^2}{M^*} = \frac{\tau^2}{M} \implies \frac{M^*}{M} = \frac{\sigma^2}{\tau^2} = RNE$$

Theorem 4.3 (a) Approximation of Bayes actions by importance sampling.

Conditions:

(1) $\boldsymbol{\theta}^{(m)} \overset{iid}{\sim} p(\boldsymbol{\theta} \mid S)$, $\boldsymbol{\omega}^{(m)} \mid \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}, I\right)$

(2) $w(\boldsymbol{\theta}) = k(\boldsymbol{\theta} \mid I) / k(\boldsymbol{\theta} \mid S) < c < \infty \; \forall \; \boldsymbol{\theta}$

(3) $L(\mathbf{a}, \boldsymbol{\omega}) \geq 0$ defined on $\Omega \times A$, $A \subseteq \mathbb{R}^q$

(4) $R(\mathbf{a}) = \int_{\Omega} \int_{\Theta} L(\mathbf{a}, \boldsymbol{\omega}) \, p(\boldsymbol{\theta} \mid I) \, p(\boldsymbol{\omega} \mid \boldsymbol{\theta}, I) \, d\boldsymbol{\theta} d\boldsymbol{\omega}$

(5) $R(\hat{\mathbf{a}}) < R(\mathbf{a}) \; \forall \; \mathbf{a} \in A$

Theorem 4.3 (a) (concluded)

More conditions:

(A1) $M^{-1}\sum_{m=1}^{M} L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right) \rightarrow R(\mathbf{a})$ uniformly $\forall\, \mathbf{a} \in N(\widehat{\mathbf{a}})$ (almost surely)

(A2) $\partial L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a}$ is a continuous function of $\mathbf{a}$ $\forall\, \mathbf{a} \in N(\widehat{\mathbf{a}})$

Denote

$$A_M = \left\{ \mathbf{a} : M^{-1} \sum_{m=1}^{M} \left[\partial L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right) / \partial \mathbf{a}\right] w\left(\boldsymbol{\theta}^{(m)}\right) = \mathbf{0} \right\}$$

Then for any $\varepsilon > 0$,

$$\lim_{M \rightarrow \infty} P\left[\inf_{\mathbf{a} \in A_M} (\mathbf{a} - \widehat{\mathbf{a}})'(\mathbf{a} - \widehat{\mathbf{a}}) > \varepsilon \mid S \right] = 0.$$

Theorem 4.3 (b) Approximation of Bayes actions by importance sampling.

Additional conditions:

(B1) $\partial^2 L(\mathbf{a}, \boldsymbol{\omega}) \partial \mathbf{a} \partial \mathbf{a}'$ is a continuous function of $\mathbf{a}$ $\forall$ $\mathbf{a} \in N(\widehat{\mathbf{a}})$, $\forall$ $\boldsymbol{\omega} \in \Omega$

(B2) $\mathbf{B} = \text{var}\left[\partial L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a}|_{\mathbf{a}=\widehat{\mathbf{a}}} w(\boldsymbol{\theta})^{1/2} \mid S\right]$ exists and is finite

(B3) $\mathbf{H} = \text{E}\left[\partial^2 L(\mathbf{a}, \boldsymbol{\omega}) / \partial \mathbf{a} \partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}} \mid S\right]$ exists and is finite and nonsingular

(B4) For any $\varepsilon > 0$, there exists M_ε such that

$$P\left[\sup_{\mathbf{a} \in N(\widehat{\mathbf{a}})} \left|\partial^3 L(\mathbf{a}, \boldsymbol{\omega}) / \partial a_i \partial a_j \partial a_k\right| < M_\varepsilon \mid S\right] \geq 1 - \varepsilon$$

for all $i, j, k = 1, \ldots, m$.

Theorem 4.3 (b) (concluded)

Then if $\widehat{\mathbf{a}}_M$ is any element of A_M such that $\widehat{\mathbf{a}}_M \xrightarrow{p} \widehat{\mathbf{a}}$,

(1) $M^{1/2}\left(\widehat{\mathbf{a}}_M - \widehat{\mathbf{a}}\right) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{H}^{-1}\mathbf{B}\mathbf{H}^{-1}\right)$

(2) $M^{-1}\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right)^2 \partial L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \cdot \partial L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \xrightarrow{p} \mathbf{B}$

(3) $M^{-1}\sum_{m=1}^{M} w\left(\boldsymbol{\theta}^{(m)}\right) \partial^2 L\left(\mathbf{a}, \boldsymbol{\omega}^{(m)}\right)/\partial \mathbf{a}\partial \mathbf{a}'|_{\mathbf{a}=\widehat{\mathbf{a}}_M} \xrightarrow{p} \mathbf{H}$.

Markov Chain Monte Carlo Methods

The central idea

$$\boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m-1)}, C\right) \quad (m = 1, 2, 3, \ldots)$$

If C is specified correctly, then

$$\boldsymbol{\theta}^{(m-1)} \sim p\left(\boldsymbol{\theta} \mid I\right), \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m-1)}, C\right) \implies \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid I\right).$$

Better yet, if

$$\boldsymbol{\theta}^{(m-1)} \sim p\left(\boldsymbol{\theta} \mid J\right), \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(m-1)}, C\right) \implies \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid J\right)$$

then $J = I$. And even better,

$$p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(0)}, C\right) \xrightarrow{d} p\left(\boldsymbol{\theta} \mid I\right) \quad \forall \quad \boldsymbol{\theta}^{(0)} \in \Theta.$$

The central idea (continued)

If $p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(0)}, C\right) \xrightarrow{d} p\left(\boldsymbol{\theta} \mid I\right) \ \forall \ \boldsymbol{\theta}^{(0)} \in \Theta$,

then we can approximate $\mathsf{E}\left[h\left(\boldsymbol{\omega}\right) \mid I\right]$ by

(1) iterating the chain B ("burn-in") times;

(2) drawing $\boldsymbol{\omega}^{(m)} \sim p\left(\boldsymbol{\omega} \mid \boldsymbol{\theta}^{(m)}\right)$ $(m = 1, \ldots, M)$;

(3) computing

$$\overline{h}^{(M)} = M^{-1} \sum_{m=1}^{M} h\left(\boldsymbol{\omega}^{(m)}\right).$$

The Gibbs sampler

Blocking: $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}'_{(1)}, ..., \boldsymbol{\theta}'_{(B)}\right)$.

Some notation: Corresponding to any subvector $\boldsymbol{\theta}_{(b)}$,

$$\begin{aligned}
\boldsymbol{\theta}'_{<(b)} &= \left(\boldsymbol{\theta}'_{(1)}, ..., \boldsymbol{\theta}'_{(b-1)}\right)(b = 2, \ldots, B)\,, \;\; \boldsymbol{\theta}_{<(1)} = \varnothing \\
\boldsymbol{\theta}'_{>(b)} &= \left(\boldsymbol{\theta}'_{(b+1)}, ..., \boldsymbol{\theta}'_{(B)}\right)(b = 1, \ldots, B-1)\,, \;\; \boldsymbol{\theta}_{>(B)} = \varnothing \\
\boldsymbol{\theta}'_{-(b)} &= \left(\boldsymbol{\theta}'_{<(b)}, \boldsymbol{\theta}'_{>(b)}\right)
\end{aligned}$$

Very important: Choose the blocking so that

$$\boldsymbol{\theta}_{(b)} \sim p\left(\boldsymbol{\theta}_{(b)} \mid \boldsymbol{\theta}_{-(b)}, I\right)$$

is possible.

Intuitive argument for the Gibbs sampler

Imagine $\boldsymbol{\theta}^{(0)} \sim p\left(\boldsymbol{\theta} \mid I\right)$, and then in succession

$$
\begin{aligned}
\boldsymbol{\theta}_{(1)}^{(1)} &\sim p\left(\boldsymbol{\theta}_{(1)} \mid \boldsymbol{\theta}_{-(1)}^{(0)}, I\right), \\
\boldsymbol{\theta}_{(2)}^{(1)} &\sim p\left(\boldsymbol{\theta}_{(2)} \mid \boldsymbol{\theta}_{<(2)}^{(1)}, \boldsymbol{\theta}_{>(2)}^{(0)}, I\right), \\
\boldsymbol{\theta}_{(3)}^{(1)} &\sim p\left(\boldsymbol{\theta}_{(3)} \mid \boldsymbol{\theta}_{<(3)}^{(1)}, \boldsymbol{\theta}_{>(3)}^{(0)}, I\right), \\
&\vdots \\
\boldsymbol{\theta}_{(b)}^{(1)} &\sim p\left(\boldsymbol{\theta}_{(b)} \mid \boldsymbol{\theta}_{<(b)}^{(1)}, \boldsymbol{\theta}_{>(b)}^{(0)}, I\right) \\
&\vdots \\
\boldsymbol{\theta}_{(B)}^{(1)} &\sim p\left(\boldsymbol{\theta}_{(B)} \mid \boldsymbol{\theta}_{-(B)}^{(1)}, I\right)
\end{aligned}
$$

We have $\boldsymbol{\theta}^{(1)} \sim p\left(\boldsymbol{\theta} \mid I\right)$.

Now repeat

$$
\begin{aligned}
\boldsymbol{\theta}_{(1)}^{(2)} &\sim p\left(\boldsymbol{\theta}_{(1)} \mid \boldsymbol{\theta}_{-(1)}^{(1)}, I\right), \\
\boldsymbol{\theta}_{(2)}^{(2)} &\sim p\left(\boldsymbol{\theta}_{(2)} \mid \boldsymbol{\theta}_{<(2)}^{(2)}, \boldsymbol{\theta}_{>(2)}^{(1)}, I\right), \\
\boldsymbol{\theta}_{(3)}^{(2)} &\sim p\left(\boldsymbol{\theta}_{(3)} \mid \boldsymbol{\theta}_{<(3)}^{(2)}, \boldsymbol{\theta}_{>(3)}^{(1)}, I\right), \\
&\vdots \\
\boldsymbol{\theta}_{(b)}^{(2)} &\sim p\left(\boldsymbol{\theta}_{(b)} \mid \boldsymbol{\theta}_{<(b)}^{(2)}, \boldsymbol{\theta}_{>(b)}^{(1)}, I\right) \\
&\vdots \\
\boldsymbol{\theta}_{(B)}^{(2)} &\sim p\left(\boldsymbol{\theta}_{(B)} \mid \boldsymbol{\theta}_{-(B)}^{(2)}, I\right).
\end{aligned}
$$

We have $\boldsymbol{\theta}^{(2)} \sim p\left(\boldsymbol{\theta} \mid I\right)$.

The general step in the Gibbs sampler is

$$\boldsymbol{\theta}_{(b)}^{(m)} \sim p\left(\boldsymbol{\theta}_{(b)} \mid \boldsymbol{\theta}_{<(b)}^{(m)}, \boldsymbol{\theta}_{>(b)}^{(m-1)}, I\right)$$

for $b = 1, \ldots, B$ and $m = 1, 2, \ldots$.

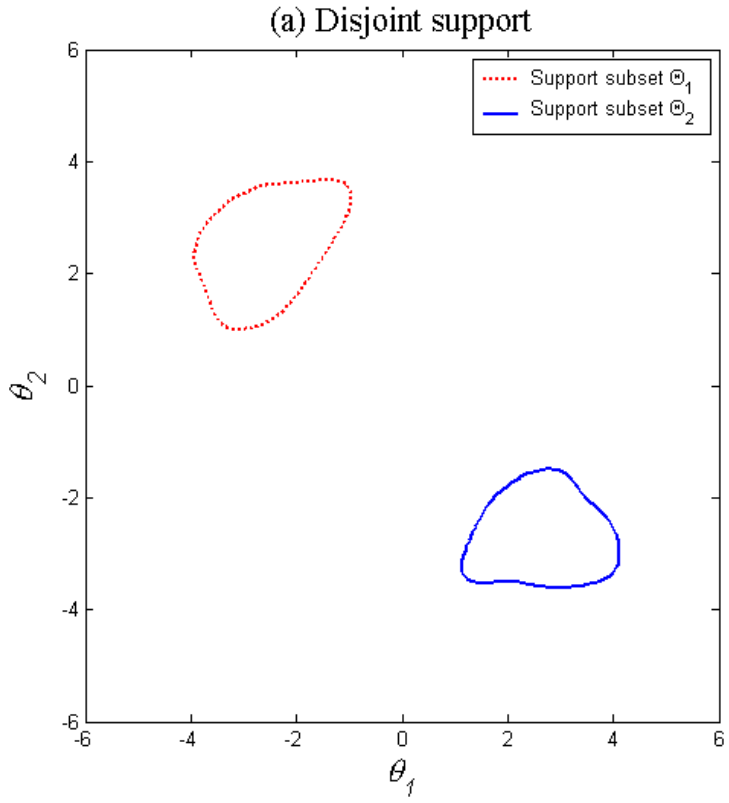
This defines the Markov chain

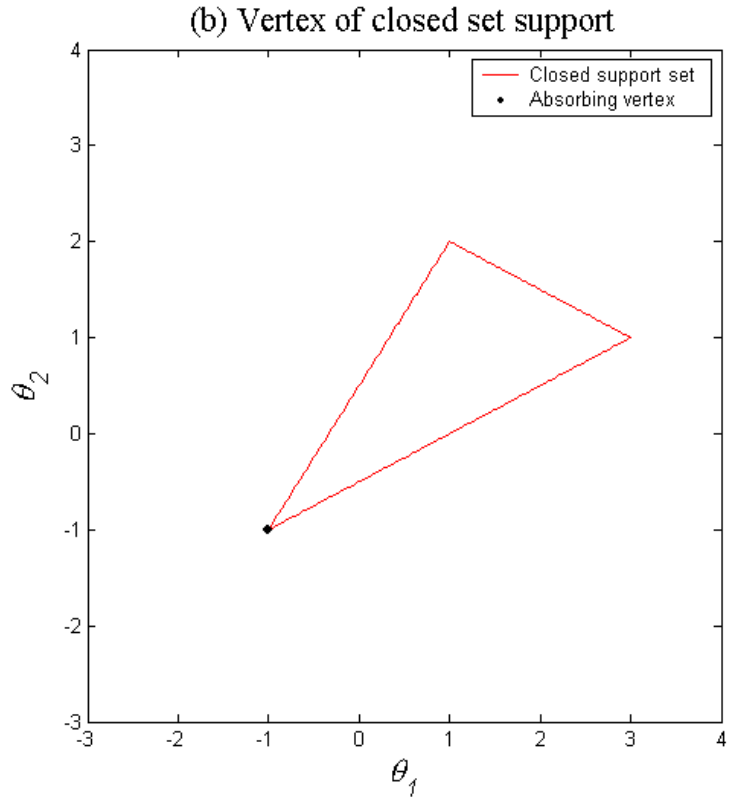
$$p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, G\right) = \prod_{b=1}^{B} p\left[\boldsymbol{\theta}_{(b)}^{(m)} \mid \boldsymbol{\theta}_{<(b)}^{(m)}, \boldsymbol{\theta}_{>(b)}^{(m-1)}, I\right].$$

Key property:

$$\boldsymbol{\theta}^{(0)} \sim p\left(\boldsymbol{\theta} \mid I\right) \implies \boldsymbol{\theta}^{(m)} \sim p\left(\boldsymbol{\theta} \mid I\right)$$

Potential problems:





The Metropolis-Hastings Algorithm

What it does: $\boldsymbol{\theta}^* \sim q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right)$

Then

$$\begin{aligned} P\left(\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^*\right) &= \alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right), \\ P\left(\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)}\right) &= 1 - \alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) \end{aligned}$$

where

$$\alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) = \min\left\{\frac{p\left(\boldsymbol{\theta}^* \mid I\right) / q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right)}{p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) / q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^*, H\right)}, 1\right\}.$$

Some aspects of the Metropolis-Hastings algorithm

If we define

$$u\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) = q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) \alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right)$$

then

$$P\left(\boldsymbol{\theta}^{(m)} = \boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{(m-1)} = \boldsymbol{\theta}, H\right) = r\left(\boldsymbol{\theta} \mid H\right) = 1 - \int_{\Theta} u\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) d\nu\left(\boldsymbol{\theta}^{*}\right).$$

Notice that

$$P\left(\boldsymbol{\theta}^{(m)} \in A \mid \boldsymbol{\theta}^{(m-1)} = \boldsymbol{\theta}, H\right) = \int_{A} \left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) d\nu\left(\boldsymbol{\theta}^{*}\right) + r\left(\boldsymbol{\theta} \mid H\right) I_{A}\left(\boldsymbol{\theta}\right).$$

$$u\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) = q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right) \alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right)$$

We can write the transition density in one line making use of the Dirac delta function, an operator with the property

$$\int_{A} \delta_{\boldsymbol{\theta}}\left(\boldsymbol{\theta}^{*}\right) f\left(\boldsymbol{\theta}^{*}\right) d\nu\left(\boldsymbol{\theta}^{*}\right) = f\left(\boldsymbol{\theta}\right) I_{A}\left(\boldsymbol{\theta}\right).$$

Then

$$\begin{aligned} p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, H\right) &= u\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, H\right) \\ &\quad + r\left(\boldsymbol{\theta}^{(m-1)} \mid H\right) \delta_{\boldsymbol{\theta}^{(m-1)}}\left(\boldsymbol{\theta}^{(m)}\right). \end{aligned}$$

Special cases of the Metropolis-Hastings algorithm

$$\alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) = \min\left\{\frac{p\left(\boldsymbol{\theta}^* \mid I\right)/q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right)}{p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right)/q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^*, H\right)}, 1\right\}$$

Special case 1, original Metropolis (1953):

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}, H\right) = q\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*, H\right)$$
$$\Longrightarrow \alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) = \min\left[p\left(\boldsymbol{\theta}^* \mid I\right)/p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right), 1\right]$$

Important example: *random walk Metropolis chain*

$$q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}, H\right) = q\left(\boldsymbol{\theta}^* - \boldsymbol{\theta} \mid H\right),$$

where $q\left(\cdot \mid H\right)$ is symmetric about zero.

Special cases of the Metropolis-Hastings algorithm

$$\alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right)=\min\left\{\frac{p\left(\boldsymbol{\theta}^{*} \mid I\right) / q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right)}{p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) / q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right)}, 1\right\}$$

Special case 2, *Metropolis independence chain*:

$$q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}, H\right)=q\left(\boldsymbol{\theta}^{*} \mid H\right)$$

$$\begin{aligned}\Longrightarrow \alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right) &= \min\left\{\frac{p\left(\boldsymbol{\theta}^{*} \mid I\right) / q\left(\boldsymbol{\theta}^{*} \mid H\right)}{p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) / q\left(\boldsymbol{\theta}^{(m-1)} \mid H\right)}, 1\right\} \\ &= \min\left\{\frac{w\left(\boldsymbol{\theta}^{*}\right)}{w\left(\boldsymbol{\theta}^{(m-1)}\right)}, 1\right\}\end{aligned}$$

where $w(\boldsymbol{\theta})=p(\boldsymbol{\theta} \mid I) / q(\boldsymbol{\theta} \mid H)$.

Why does the Metropolis-Hastings algorithm work?

A two part argument – Part 1:

Suppose any transition probability density function $p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, T\right)$ satisfies the *reversibility condition*

$$p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, T\right) = p\left(\boldsymbol{\theta}^{(m)} \mid I\right) p\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{(m)}, T\right)$$

with respect to $p(\boldsymbol{\theta} \mid I)$. Then

$$\begin{aligned}
&\int_{\Theta} p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, T\right) d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) \\
&= \int_{\Theta} p\left(\boldsymbol{\theta}^{(m)} \mid I\right) p\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{(m)}, T\right) d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) \\
&= p\left(\boldsymbol{\theta}^{(m)} \mid I\right) \int_{\Theta} p\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{(m)}, T\right) d\nu\left(\boldsymbol{\theta}^{(m-1)}\right) = p\left(\boldsymbol{\theta}^{(m)} \mid I\right).
\end{aligned}$$

and so $p(\boldsymbol{\theta} \mid I)$ is an *invariant density* of the Markov chain.

Part 2 of the argument (How Hastings did it):

Suppose we don't know the probability $\alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right)$, but we want

$p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, H\right)$ to be reversible with respect to $p(\boldsymbol{\theta} \mid I)$:

$$p\left(\boldsymbol{\theta}^{(m-1)}|I\right) p\left(\boldsymbol{\theta}^{(m)} \mid \boldsymbol{\theta}^{(m-1)}, H\right) = p\left(\boldsymbol{\theta}^{(m)}|I\right) p\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{(m)}, H\right).$$

Trivial if $\boldsymbol{\theta}^{(m-1)} = \boldsymbol{\theta}^{(m)}$. For $\boldsymbol{\theta}^{(m-1)} \neq \boldsymbol{\theta}^{(m)}$ we need

$$\begin{aligned} & p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) q\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) \alpha\left(\boldsymbol{\theta}^* \mid \boldsymbol{\theta}^{(m-1)}, H\right) \\ = \; & p\left(\boldsymbol{\theta}^* \mid I\right) q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^*, H\right) \alpha\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^*, H\right). \end{aligned}$$

$$\begin{aligned} & p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right) \alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right) \\ = \; & p\left(\boldsymbol{\theta}^{*} \mid I\right) q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right) \alpha\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right) \end{aligned}$$

Suppose without loss of generality that

$$p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right) > p\left(\boldsymbol{\theta}^{*} \mid I\right) q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right).$$

Set $\alpha\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right) = 1$ and

$$\alpha\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right) = \frac{p\left(\boldsymbol{\theta}^{*} \mid I\right) q\left(\boldsymbol{\theta}^{(m-1)} \mid \boldsymbol{\theta}^{*}, H\right)}{p\left(\boldsymbol{\theta}^{(m-1)} \mid I\right) q\left(\boldsymbol{\theta}^{*} \mid \boldsymbol{\theta}^{(m-1)}, H\right)}.$$