Adaptive Gibbs Samplers

Krzysztof Latuszynski
(University of Warwick, UK)

joint work with

Gareth O. Roberts Jeffrey S. Rosenthal
(Warwick, Toronto)

MCQMC 2010, Warsaw

Adaptive MCMC

Adaptive Gibbs samplers

Ergodicity results

A specific Metropolis-within-Gibbs adaptation

Markov chain Monte Carlo

- let π be a target probability distribution on $\mathcal{X}$, e.g. to evaluate

$$\theta := \int_{\mathcal{X}} f(x)\pi(dx).$$

- direct sampling from π is not possible or inefficient
- MCMC approach is to simulate $(X_n)_{n\geq 0}$, an ergodic Markov chain with **transition kernel** P and limiting distribution π, and take ergodic averages as an estimate of θ.
- it is **easy** to design an **ergodic** transition kernel P, e.g. using generic Metropolis or Gibbs recipes
- is is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional

Markov chain Monte Carlo

- let π be a target probability distribution on $\mathcal{X}$, e.g. to evaluate

$$\theta := \int_{\mathcal{X}} f(x)\pi(dx).$$

- direct sampling from π is not possible or inefficient
- MCMC approach is to simulate $(X_n)_{n\geq 0}$, an ergodic Markov chain with **transition kernel** P and limiting distribution π, and take ergodic averages as an estimate of θ.
- it is **easy** to design an **ergodic** transition kernel P, e.g. using generic Metropolis or Gibbs recipes
- is is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional

Markov chain Monte Carlo

- let π be a target probability distribution on $\mathcal{X}$, e.g. to evaluate

$$\theta := \int_{\mathcal{X}} f(x)\pi(dx).$$

- direct sampling from π is not possible or inefficient
- MCMC approach is to simulate $(X_n)_{n\geq 0}$, an ergodic Markov chain with **transition kernel** P and limiting distribution π, and take ergodic averages as an estimate of θ.
- it is **easy** to design an **ergodic** transition kernel P, e.g. using generic Metropolis or Gibbs recipes
- is is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional

Markov chain Monte Carlo

- let π be a target probability distribution on $\mathcal{X}$, e.g. to evaluate

$$\theta := \int_{\mathcal{X}} f(x)\pi(dx).$$

- direct sampling from π is not possible or inefficient
- MCMC approach is to simulate $(X_n)_{n\geq 0}$, an ergodic Markov chain with **transition kernel** P and limiting distribution π, and take ergodic averages as an estimate of θ.
- it is **easy** to design an **ergodic** transition kernel P, e.g. using generic Metropolis or Gibbs recipes
- is is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional

Markov chain Monte Carlo

- let π be a target probability distribution on $\mathcal{X}$, e.g. to evaluate

$$\theta := \int_{\mathcal{X}} f(x)\pi(dx).$$

- direct sampling from π is not possible or inefficient
- MCMC approach is to simulate $(X_n)_{n\geq 0}$, an ergodic Markov chain with **transition kernel** P and limiting distribution π, and take ergodic averages as an estimate of θ.
- it is **easy** to design an **ergodic** transition kernel P, e.g. using generic Metropolis or Gibbs recipes
- is is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

optimizing the transition kernel P

- for Metropolis chains there are "prescriptions" of how to scale proposals as dimension $\to \infty$. (optimal scaling results)
- "prescriptions" depend on unknown characteristics of π
- one has to learn π to apply them
- for **random scan Gibbs samplers** a further design decision is choosing selection probabilities

- a promising approach is: **adaptive** MCMC algorithms
- the transition kernel P_n used for obtaining $X_n|X_{n-1}$ may depend on $\{X_0, \ldots, X_{n-1}\}$
- however now the process is **not** Markovian, so the possible benefit comes at the price of more involving theoretical analysis

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Metropolis and adaptive Gibbs

- **for the adaptive Metropolis algorithm** a lot is known:
- optimal scaling results are available (Roberts et al 97, Roberts Rosenthal 98, Bédard 07, Bédard 08, ...)
- ergodicity of the adaptive Metropolis has been extensively studied (Haario et al 2001, Atchade Rosenthal 2005, Saksman Vihola 08, Bai et al 08, Vihola 09, Vihola 10, ...)
- the **adaptive random scan Gibbs sampler** has not been studied much...
- (Levine Casella 06) build on [Liu at al 95] to give advice of how to choose coordinate probabilities
- however the general ergodicity result for adaptive Gibbs samplers that they establish and use, is wrong
- validity of their suggested adaptive update of selection probabilities remains open

Adaptive Gibbs sampler - a generic algorithm

`AdapRSG`

1. Set $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0) \in \mathcal{Y} \subset [0,1]^d$
2. Choose coordinate $i \in \{1, \ldots, d\}$ according to selection probabilities α_n
3. Draw $Y \sim \pi(\cdot | X_{n-1,-i})$
4. Set $X_n := (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})$

a cautionary example that disproves L-C 06

- Let $\mathcal{X} = \{(i,j) \in \mathbb{N} \times \mathbb{N} : i = j \text{ or } i = j+1\}$,
- with target distribution given by $\pi(i,j) \propto j^{-2}$
- consider a class of adaptive random scan Gibbs samplers with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i,j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j+1, \end{cases}$$

  for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$
- if $a_n \to \infty$ slowly enough, then X_n is **transient** with positive probability, i.e. $\mathbb{P}(X_{1,n} \to \infty) > 0$.

a cautionary example that disproves L-C 06

- Let $\mathcal{X} = \{(i,j) \in \mathbb{N} \times \mathbb{N} : i = j \text{ or } i = j+1\}$,
- with target distribution given by $\pi(i,j) \propto j^{-2}$
- consider a class of adaptive random scan Gibbs samplers with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i,j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j+1, \end{cases}$$

  for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$
- if $a_n \to \infty$ slowly enough, then X_n is **transient** with positive probability, i.e. $\mathbb{P}(X_{1,n} \to \infty) > 0$.

a cautionary example that disproves L-C 06

- Let $\mathcal{X} = \{(i,j) \in \mathbb{N} \times \mathbb{N} : i = j \text{ or } i = j+1\}$,
- with target distribution given by $\pi(i,j) \propto j^{-2}$
- consider a class of adaptive random scan Gibbs samplers with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i,j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j+1, \end{cases}$$

  for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$
- if $a_n \to \infty$ slowly enough, then X_n is **transient** with positive probability, i.e. $\mathbb{P}(X_{1,n} \to \infty) > 0$.

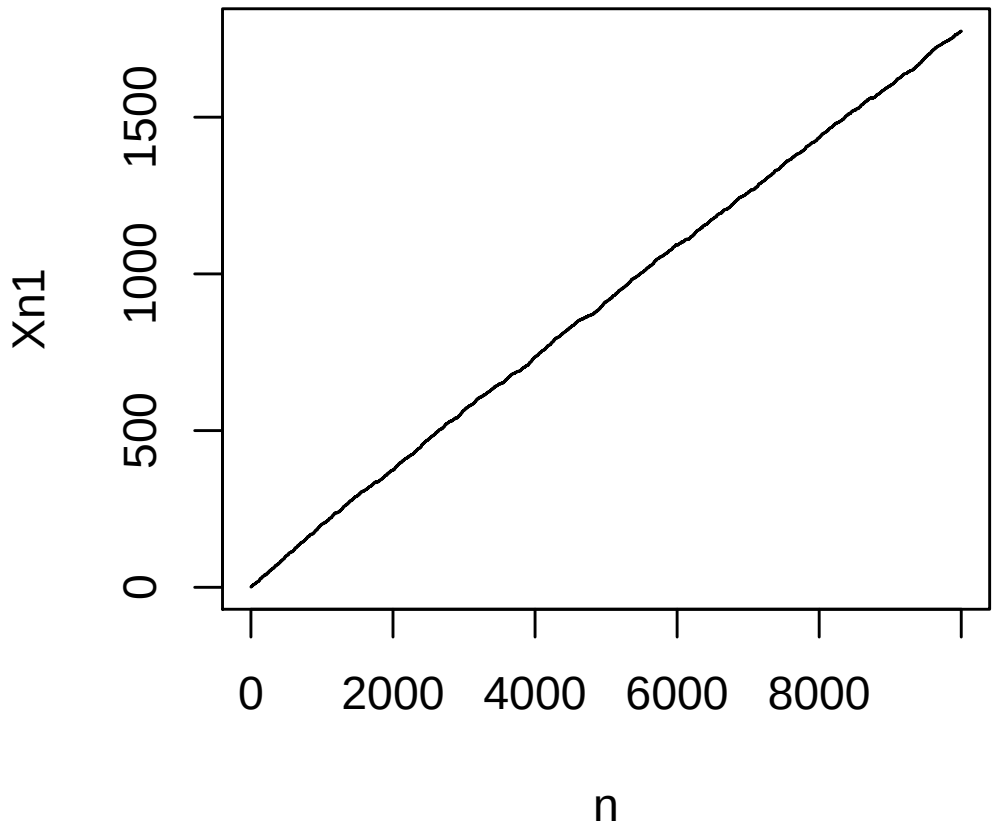
a cautionary example that disproves L-C 06

- Let $\mathcal{X} = \{(i,j) \in \mathbb{N} \times \mathbb{N} : i = j \text{ or } i = j+1\}$,
- with target distribution given by $\pi(i,j) \propto j^{-2}$
- consider a class of adaptive random scan Gibbs samplers with update rule given by:

$$R_n\Big(\alpha_{n-1}, X_{n-1} = (i,j)\Big) = \begin{cases} \left\{\frac{1}{2} + \frac{4}{a_n}, \frac{1}{2} - \frac{4}{a_n}\right\} & \text{if} \quad i = j, \\ \\ \left\{\frac{1}{2} - \frac{4}{a_n}, \frac{1}{2} + \frac{4}{a_n}\right\} & \text{if} \quad i = j+1, \end{cases}$$

  for some choice of the sequence $(a_n)_{n=0}^{\infty}$ satisfying $8 < a_n \nearrow \infty$
- if $a_n \to \infty$ slowly enough, then X_n is **transient** with positive probability, i.e. $\mathbb{P}(X_{1,n} \to \infty) > 0$.

a cautionary example...



Adaptive random scan Metropolis within Gibbs

`AdapRSMwG`

1. Set $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0) \in \mathcal{Y}$
2. Choose coordinate $i \in \{1, \ldots, d\}$ according to selection probabilities α_n
3. Draw $Y \sim Q_{X_{n-1,-i}}(X_{n-1,i}, \cdot)$
4. With probability

$$\min\left(1, \quad \frac{\pi(Y|X_{n-1,-i})\; q_{X_{n-1,-i}}(Y, X_{n-1,i})}{\pi(X_{n-1}|X_{n-1,-i})\; q_{X_{n-1,-i}}(X_{n-1,i}, Y)}\right), \tag{1}$$

accept the proposal and set

$$X_n = (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d});$$

otherwise, reject the proposal and set $X_n = X_{n-1}$.

Adaptive random scan adaptive Metropolis within Gibbs

`AdapRSadapMwG`

1. Set $\alpha_n := R_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0, \gamma_{n-1}, \ldots, \gamma_0) \in \mathcal{Y}$
2. Set $\gamma_n := R'_n(\alpha_{n-1}, X_{n-1}, \ldots, X_0, \gamma_{n-1}, \ldots, \gamma_0) \in \Gamma_1 \times \ldots \times \Gamma_n$
3. Choose coordinate $i \in \{1, \ldots, d\}$ according to selection probabilities α, i.e. with $\Pr(i = j) = \alpha_j$
4. Draw $Y \sim Q_{X_{n-1,-i},\gamma_{n-1}}(X_{n-1,i}, \cdot)$
5. With probability (1),

$$\min\left(1, \quad \frac{\pi(Y|X_{n-1,-i})\; q_{X_{n-1,-i},\gamma_{n-1}}(Y, X_{n-1,i})}{\pi(X_{n-1}|X_{n-1,-i})\; q_{X_{n-1,-i},\gamma_{n-1}}(X_{n-1,i}, Y)}\right),$$

accept the proposal and set

$$X_n = (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d});$$

otherwise, reject the proposal and set $X_n = X_{n-1}$.

Tools for establishing ergodicity

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (simultaneous uniform ergodicity) $\Rightarrow$ ergodicity.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (containment) $\Rightarrow$ ergodicity.

Tools for establishing ergodicity

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P^N_\gamma(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P^n_\gamma(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (simultaneous uniform ergodicity) $\Rightarrow$ ergodicity.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (containment) $\Rightarrow$ ergodicity.

Tools for establishing ergodicity

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (simultaneous uniform ergodicity) $\Rightarrow$ ergodicity.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (containment) $\Rightarrow$ ergodicity.

Tools for establishing ergodicity

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (simultaneous uniform ergodicity) $\Rightarrow$ ergodicity.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (containment) $\Rightarrow$ ergodicity.

Tools for establishing ergodicity

- **(Diminishing Adaptation)** Let $D_n = \sup_{x \in \mathcal{X}} \|P_{\Gamma_{n+1}}(x, \cdot) - P_{\Gamma_n}(x, \cdot)\|$ and assume $\lim_{n \to \infty} D_n = 0$ in probability
- **(Simultaneous uniform ergodicity)** For all $\varepsilon > 0$, there exists $N = N(\varepsilon)$ s.t. $\|P_\gamma^N(x, \cdot) - \pi(\cdot)\| \leq \varepsilon$ for all $x \in \mathcal{X}$ and $\gamma \in \mathcal{Y}$
- **(Containment condition)** Let $M_\varepsilon(x, \gamma) = \inf\{n \geq 1 : \|P_\gamma^n(x, \cdot) - \pi(\cdot)\| \leq \varepsilon\}$ and assume $\{M_\varepsilon(X_n, \gamma_n)\}_{n=0}^\infty$ is bounded in probability, i.e. given $X_0 = x_*$ and $\Gamma_0 = \gamma_*$, for all $\delta > 0$, there exists N s.t. $\mathbb{P}[M_\varepsilon(X_n, \Gamma_n) \leq N | X_0 = x_*, \Gamma_0 = \gamma_*] \geq 1 - \delta$.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (simultaneous uniform ergodicity) $\Rightarrow$ ergodicity.

Theorem (Roberts Rosenthal 2007)

(diminishing adaptation) + (containment) $\Rightarrow$ ergodicity.

Ergodicity

- Assuming that `RSG(`β`)` is uniformly ergodic and $|\alpha_n - \alpha_{n-1}| \to 0$, we prove ergodicity of
  - `AdapRSG`
  - `AdapRSMwG`
  - `AdapRSadapMwG`

  by establishing diminishing adaptation and simultaneous uniform ergodicity.
- Assuming that $|\alpha_n - \alpha_{n-1}| \to 0$ and regularity conditions for the target and proposal distributions (in the spirit of Roberts Rosenthal 98, Fort et al 03) we prove ergodicity of
  - `AdapRSMwG`
  - `AdapRSadapMwG`

  by establishing diminishing adaptation and containment (by simultaneous geometrical ergodicity, using results of Bai et al 2008)

Ergodicity

- Assuming that `RSG(`β`)` is uniformly ergodic and $|\alpha_n - \alpha_{n-1}| \to 0$, we prove ergodicity of
  - `AdapRSG`
  - `AdapRSMwG`
  - `AdapRSadapMwG`

  by establishing diminishing adaptation and simultaneous uniform ergodicity.
- Assuming that $|\alpha_n - \alpha_{n-1}| \to 0$ and regularity conditions for the target and proposal distributions (in the spirit of Roberts Rosenthal 98, Fort et al 03) we prove ergodicity of
  - `AdapRSMwG`
  - `AdapRSadapMwG`

  by establishing diminishing adaptation and containment (by simultaneous geometrical ergodicity, using results of Bai et al 2008)

Proposal variance

- use a Metropolis-within-Gibbs sampler and update coordinate i by proposing a normal increment to $X_{n-1,i}$, i.e. the proposal

$$Y_{n,i} \sim N(X_{n-1,i}, \sigma^2_{n,i}).$$

- The proposal variance $\sigma^2_{n,i}$ is subject to adaptation.
- Haario et al. 05 use

$$\sigma^{2,\mathrm{HST}}_{n,i} = (2.4)^2(s^2_{n,i} + 0.05),$$

where $s^2_{n,i}$ is the sample variance of $X_{0,i}, \ldots, X_{n-1,i}$,
- whereas Roberts and Rosenthal 06 take

$$\sigma^{2,\mathrm{RR}}_{n,i} = e^{ls_i},$$

and ls_i is updated every batch of 50 iterations by adding or subtracting $\delta(n) = O(n^{-1/2})$. Specifically, ls_i is increased by $\delta(n)$ if the fraction of acceptances of variable i was more then 0.44 on the last batch and decreased if it was less.

Proposal variance

- use a Metropolis-within-Gibbs sampler and update coordinate i by proposing a normal increment to $X_{n-1,i}$, i.e. the proposal

$$Y_{n,i} \sim N(X_{n-1,i}, \sigma^2_{n,i}).$$

- The proposal variance $\sigma^2_{n,i}$ is subject to adaptation.
- Haario et al. 05 use

$$\sigma^{2,\mathrm{HST}}_{n,i} = (2.4)^2(s^2_{n,i} + 0.05),$$

where $s^2_{n,i}$ is the sample variance of $X_{0,i}, \ldots, X_{n-1,i}$,
- whereas Roberts and Rosenthal 06 take

$$\sigma^{2,\mathrm{RR}}_{n,i} = e^{ls_i},$$

and ls_i is updated every batch of 50 iterations by adding or subtracting $\delta(n) = O(n^{-1/2})$. Specifically, ls_i is increased by $\delta(n)$ if the fraction of acceptances of variable i was more then 0.44 on the last batch and decreased if it was less.

Proposal variance

- use a Metropolis-within-Gibbs sampler and update coordinate i by proposing a normal increment to $X_{n-1,i}$, i.e. the proposal

$$Y_{n,i} \sim N(X_{n-1,i}, \sigma^2_{n,i}).$$

- The proposal variance $\sigma^2_{n,i}$ is subject to adaptation.
- Haario et al. 05 use

$$\sigma^{2,\mathrm{HST}}_{n,i} = (2.4)^2(s^2_{n,i} + 0.05),$$

where $s^2_{n,i}$ is the sample variance of $X_{0,i}, \ldots, X_{n-1,i}$,

- whereas Roberts and Rosenthal 06 take

$$\sigma^{2,\mathrm{RR}}_{n,i} = e^{ls_i},$$

and ls_i is updated every batch of 50 iterations by adding or subtracting $\delta(n) = O(n^{-1/2})$. Specifically, ls_i is increased by $\delta(n)$ if the fraction of acceptances of variable i was more then 0.44 on the last batch and decreased if it was less.

Proposal variance

- use a Metropolis-within-Gibbs sampler and update coordinate i by proposing a normal increment to $X_{n-1,i}$, i.e. the proposal

$$Y_{n,i} \sim N(X_{n-1,i}, \sigma^2_{n,i}).$$

- The proposal variance $\sigma^2_{n,i}$ is subject to adaptation.
- Haario et al. 05 use

$$\sigma^{2,\mathrm{HST}}_{n,i} = (2.4)^2(s^2_{n,i} + 0.05),$$

where $s^2_{n,i}$ is the sample variance of $X_{0,i}, \ldots, X_{n-1,i}$,

- whereas Roberts and Rosenthal 06 take

$$\sigma^{2,\mathrm{RR}}_{n,i} = e^{ls_i},$$

and ls_i is updated every batch of 50 iterations by adding or subtracting $\delta(n) = O(n^{-1/2})$. Specifically, ls_i is increased by $\delta(n)$ if the fraction of acceptances of variable i was more then 0.44 on the last batch and decreased if it was less.

Assumptions

The following conditions hold.

(i) The stationary distribution on $\mathcal{X} = \mathbb{R}^d$ is of the product form

$$\pi(x) \quad = \quad \prod_{i=1}^{d} C_i \, g(C_i x_i), \tag{2}$$

where g is a one dimensional density and C_i, $i = 1, \ldots, d$, are unknown, strictly positive constants.

(ii) The second moment of g exists, i.e. $\sigma^2 := Var_g Z < \infty$.

We consider an adaptive random scan adaptive random walk Metropolis-within-Gibbs algorithm `AdapRSadapMwG`, with Gaussian proposals, for estimating expectation of a linear target function

$$f(x) \quad = \quad a_0 + \sum_{i=1}^{d} a_i x_i. \tag{3}$$

Adaptation

- In this simplistic setting we can compute the asymptotic variance explicitly and conclude that

$$\sigma^2_{\text{as}} \quad \propto \quad \sum_{i=1}^{d} \frac{\sigma^{2,\text{HST}}_{n,i} a_i^2}{\alpha_i} \quad \propto \quad \sum_{i=1}^{d} \frac{\sigma^{2,\text{RR}}_{n,i} a_i^2}{\alpha_i}.$$

- Which is minimised for

$$\alpha_i \propto \left(\sigma^{2,\text{HST}}_{n,i} a_i^2\right)^{1/2} \quad \propto \quad \left(\sigma^{2,\text{RR}}_{n,i} a_i^2\right)^{1/2},$$

and yields a very intuitive prescription for adapting selection probabilities

$$\alpha^{\text{HST}}_{n,i} := \frac{\left(\sigma^{2,\text{HST}}_{n,i} a_i^2\right)^{1/2}}{\sum_{k=1}^{d} \left(\sigma^{2,\text{HST}}_{n,k} a_k^2\right)^{1/2}} \qquad \alpha^{\text{RR}}_{n,i} := \frac{\left(\sigma^{2,\text{RR}}_{n,i} a_i^2\right)^{1/2}}{\sum_{k=1}^{d} \left(\sigma^{2,\text{RR}}_{n,k} a_k^2\right)^{1/2}}.$$

Adaptation

- In this simplistic setting we can compute the asymptotic variance explicitly and conclude that

$$\sigma^2_{\text{as}} \propto \sum_{i=1}^{d} \frac{\sigma^{2,\text{HST}}_{n,i} a_i^2}{\alpha_i} \propto \sum_{i=1}^{d} \frac{\sigma^{2,\text{RR}}_{n,i} a_i^2}{\alpha_i}.$$

- Which is minimised for

$$\alpha_i \propto \left(\sigma^{2,\text{HST}}_{n,i} a_i^2\right)^{1/2} \propto \left(\sigma^{2,\text{RR}}_{n,i} a_i^2\right)^{1/2},$$

and yields a very intuitive prescription for adapting selection probabilities

$$\alpha^{\text{HST}}_{n,i} := \frac{\left(\sigma^{2,\text{HST}}_{n,i} a_i^2\right)^{1/2}}{\sum_{k=1}^{d}\left(\sigma^{2,\text{HST}}_{n,k} a_k^2\right)^{1/2}} \qquad \alpha^{\text{RR}}_{n,i} := \frac{\left(\sigma^{2,\text{RR}}_{n,i} a_i^2\right)^{1/2}}{\sum_{k=1}^{d}\left(\sigma^{2,\text{RR}}_{n,k} a_k^2\right)^{1/2}}.$$