Adaptive Increasingly Rarely Markov Chain Monte Carlo (AirMCMC)

Krys Latuszynski
(University of Warwick, UK)

Cyril Chimisov Gareth O. Roberts
(both Warwick)

Durham - June 28th, 2017

Adaptive MCMC

Air MCMC

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

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 impractical
- 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
- it is **difficult** to design a transition kernel P with **good convergence properties**, especially if $\mathcal{X}$ is high dimesional
- Trying to find an optimal P would be a disaster problem

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider

$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$

- $Z_t^{(d)}$ converges to the solution Z of the SDE

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$

- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider

$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$

- $Z_t^{(d)}$ converges to the solution Z of the SDE

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$

- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider

$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$

- $Z_t^{(d)}$ converges to the solution Z of the SDE

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$

- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider

$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$

- $Z_t^{(d)}$ converges to the solution Z of the SDE

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$

- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider
$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$
- $Z_t^{(d)}$ converges to the solution Z of the SDE
$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$
- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Optimal Scaling of Metropolis-Hastings P

- for Metropolis chains there is a "prescription" of how to scale proposals as dimension $d \to \infty$.
- If $\sigma_d^2 = l^2 d^{-1}$ then based on an elegant mathematical result (Roberts 1997)
- Consider

$$Z_t^{(d)} := X_{\lfloor td \rfloor}^{(d,1)}, \qquad \text{then as } d \to \infty,$$

- $Z_t^{(d)}$ converges to the solution Z of the SDE

$$dZ_t = h(l)^{1/2} dB_t + \frac{1}{2} h(l) \nabla \log f(Z_t) dt$$

- so maximise $h(l)$ to optimise Metropolis-Hastings.
- one-to-one correspondence between l_{opt} and mean acceptance rate of 0.234.

Adaptive MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{.7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{\cdot 7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{\cdot 7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{.7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{.7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly.**
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{.7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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 MCMC

- Use scale γ of the proposal such that mean acceptance rate of M-H is 0.234
- one needs to learn π to apply this
- Trial run? High dimensions? Metropolis within Gibbs?
- Adaptive MCMC: update the scale **on the fly**.
- For adaptive scaling Metropolis-Hastings one may use

$$\log(\gamma_n) = \log(\gamma_{n-1}) + n^{\cdot 7}(\alpha(X_{n-1}, Y_n) - 0.234)$$

- so 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

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

Success story of Adaptive Metropolis-Hastings

- The adaptation rule is mathematically appealing (diffusion limit)
- The adaptation rule is computationally simple (acceptance rate)
- It works in applications (seems to improve convergence significantly)
- Improves convergence even in settings that are neither high dimensional, nor satisfy other assumptions needed for the diffusion limit
-
- Adaptive scaling beyond Metropolis-Hastings?
- **YES.** Similar optimal scaling results are available for MALA, HMC, etc. Each yields an adaptive version of the algorithm!
-
- What can you optimise beyond scale?
- E.g. **covariance matrix of the proposal.**

The fly in the ointment

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- Adaptive MCMC is not Markovian
- The standard MCMC theory does not apply
- Theoretical properties of adaptive MCMC have been studied using a range of techniques, such as: coupling, martingale approximations, stability of stochastic approximation (Roberts, Rosenthal, Moulines, Andrieu, Vihola, Saksman, Fort, Atchade, ...)
- Still, the theoretical underpinning of Adaptive MCMC is (even) weaker and (even) less operational than that of standard MCMC

The fly in the ointment

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- Adaptive MCMC is not Markovian
- The standard MCMC theory does not apply
- Theoretical properties of adaptive MCMC have been studied using a range of techniques, such as: coupling, martingale approximations, stability of stochastic approximation (Roberts, Rosenthal, Moulines, Andrieu, Vihola, Saksman, Fort, Atchade, ...)
- Still, the theoretical underpinning of Adaptive MCMC is (even) weaker and (even) less operational than that of standard MCMC

The fly in the ointment

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- Adaptive MCMC is not Markovian
- The standard MCMC theory does not apply
- Theoretical properties of adaptive MCMC have been studied using a range of techniques, such as: coupling, martingale approximations, stability of stochastic approximation (Roberts, Rosenthal, Moulines, Andrieu, Vihola, Saksman, Fort, Atchade, ...)
- Still, the theoretical underpinning of Adaptive MCMC is (even) weaker and (even) less operational than that of standard MCMC

The fly in the ointment

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- Adaptive MCMC is not Markovian
- The standard MCMC theory does not apply
- Theoretical properties of adaptive MCMC have been studied using a range of techniques, such as: coupling, martingale approximations, stability of stochastic approximation (Roberts, Rosenthal, Moulines, Andrieu, Vihola, Saksman, Fort, Atchade, ...)
- Still, the theoretical underpinning of Adaptive MCMC is (even) weaker and (even) less operational than that of standard MCMC

The fly in the ointment

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- Adaptive MCMC is not Markovian
- The standard MCMC theory does not apply
- Theoretical properties of adaptive MCMC have been studied using a range of techniques, such as: coupling, martingale approximations, stability of stochastic approximation (Roberts, Rosenthal, Moulines, Andrieu, Vihola, Saksman, Fort, Atchade, ...)
- Still, the theoretical underpinning of Adaptive MCMC is (even) weaker and (even) less operational than that of standard MCMC

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0$, in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}$.
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}$.
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**! (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0$, in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}$.
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}$.
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**! (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0,$ in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}.$
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}.$
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**! (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0$, in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}$.
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}$.
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**!
  (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0$, in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}$.
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}$.
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**!
  (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- Standard assumptions to validate Adaptive MCMC are e.g. as follows:
- **(DA) Diminishing Adaptation**:
  $\lim_{n\to\infty} D_n = 0$, in probability, where
  $D_n = \sup_{x\in\mathcal{X}} \|P_{\gamma_{n+1}}(x,\cdot) - P_{\gamma_n}(x,\cdot)\|_{TV}$.
- **(C) Containment**:
  The sequence $\{M_\varepsilon(X_n,\gamma_n)\}_{n=0}^{\infty}$ is bounded in probability, where
  $M_\varepsilon : \mathcal{X}\times\Gamma \to \mathbb{N}$ is defined as
  $M_\varepsilon(x,\gamma) := \inf\{n \geq 1 : \|P_\gamma^n(x,\cdot) - \pi(\cdot)\|_{TV} \leq \varepsilon\}$.
- DA + C guarantee ergodicity, i.e. convergence in distribution (Roberts + Rosenthal 2007)
- and also nondeterioration (KL + Rosenthal 2014)
- but for SLLN, you **need additional conditions**! (Roberts + Rosenthal 2007; Fort + Moulines + Priouret 2011; Atchade + Fort 2010)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x)$,
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C$,
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C$,
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C$,
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x),$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C,$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x),$
  $P_\gamma(x,\cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C,$
  $P_\gamma(x,\cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \le \lambda V(x) + bI_C(x)$,
  $P_\gamma(x,\cdot) \ge \delta\nu(\cdot)$ for all $x \in C$,
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \le -cV^\alpha(x) + bI_C$,
  $P_\gamma(x,\cdot) \ge \delta\nu(\cdot)$ for all $x \in C$,
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x)$,
  $P_\gamma(x,\cdot) \geq \delta\nu(\cdot)$ for all $x \in C$,
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C$,
  $P_\gamma(x,\cdot) \geq \delta\nu(\cdot)$ for all $x \in C$,
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x),$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C,$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Post-mortem of the fly

- **Two setting to verify containment**:
- **(SGE) Simultaneous Geometric Ergodicity**:
  $P_\gamma V(x) \leq \lambda V(x) + bI_C(x),$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $\lambda,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- **(SPE) Simultaneous Polynomial Ergodicity**:
  $P_\gamma V(x) - V(x) \leq -cV^\alpha(x) + bI_C,$
  $P_\gamma(x, \cdot) \geq \delta\nu(\cdot)$ for all $x \in C,$
  same $c,\ \alpha,\ b,\ C,\ \delta,\ \nu$ for all $\gamma \in \Gamma$
- How do you verify SGE or SPE?
- You ask Jim Hobert!
- It has been done for fairly general classes of Adaptive Metropolis under tail decay conditions of π (Bai + Roberts + Rosenthal 2011)
- For similarly general Adaptive Metropolis within Adaptive Gibbs (KL + Roberts + Rosenthal 2013)

Adaptive Gibbs Sampler - a generic algorithm

- AdapRSG
  1. Set $p_n := R_n(p_{n-1}, X_{n-1}, \dots, X_0) \in \mathcal{Y} \subset [0,1]^d$
  2. Choose coordinate $i \in \{1, \dots, d\}$ according to selection probabilities p_n
  3. Draw $Y \sim \pi(\cdot|X_{n-1,-i})$
  4. Set $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$
- Given the target distribution π, what are the optimal selection probabilities p?
- Pretend π is a Gaussian - optimal p is known for Gaussians - and it works outside the Gaussian class.
  (Chimisov + KL + Roberts 2017)
- How to verify containment?
- Simultaneous Geometric Drift will not work!!!

Adaptive Gibbs Sampler - a generic algorithm

- AdapRSG
  1. Set $p_n := R_n(p_{n-1}, X_{n-1}, \dots, X_0) \in \mathcal{Y} \subset [0,1]^d$
  2. Choose coordinate $i \in \{1, \dots, d\}$ according to selection probabilities p_n
  3. Draw $Y \sim \pi(\cdot | X_{n-1,-i})$
  4. Set $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$
- Given the target distribution π, what are the optimal selection probabilities p?
- Pretend π is a Gaussian - optimal p is known for Gaussians - and it works outside the Gaussian class.
  (Chimisov + KL + Roberts 2017)
- How to verify containment?
- Simultaneous Geometric Drift will not work!!!

Adaptive Gibbs Sampler - a generic algorithm

- `AdapRSG`
  1. Set $p_n := R_n(p_{n-1}, X_{n-1}, \dots, X_0) \in \mathcal{Y} \subset [0,1]^d$
  2. Choose coordinate $i \in \{1, \dots, d\}$ according to selection probabilities p_n
  3. Draw $Y \sim \pi(\cdot | X_{n-1,-i})$
  4. Set $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$
- Given the target distribution π, what are the optimal selection probabilities p?
- Pretend π is a Gaussian - optimal p is known for Gaussians - and it works outside the Gaussian class.
  (Chimisov + KL + Roberts 2017)
- How to verify containment?
- Simultaneous Geometric Drift will not work!!!

Adaptive Gibbs Sampler - a generic algorithm

- `AdapRSG`
  1. Set $p_n := R_n(p_{n-1}, X_{n-1}, \dots, X_0) \in \mathcal{Y} \subset [0,1]^d$
  2. Choose coordinate $i \in \{1, \dots, d\}$ according to selection probabilities p_n
  3. Draw $Y \sim \pi(\cdot | X_{n-1,-i})$
  4. Set $X_n := (X_{n-1,1}, \dots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \dots, X_{n-1,d})$
- Given the target distribution π, what are the optimal selection probabilities p?
- Pretend π is a Gaussian - optimal p is known for Gaussians - and it works outside the Gaussian class.
  (Chimisov + KL + Roberts 2017)
- How to verify containment?
- Simultaneous Geometric Drift will not work!!!

Adaptive Gibbs Sampler - a generic algorithm

- `AdapRSG`
  1. Set $p_n := R_n(p_{n-1}, X_{n-1}, \ldots, X_0) \in \mathcal{Y} \subset [0, 1]^d$
  2. Choose coordinate $i \in \{1, \ldots, d\}$ according to selection probabilities p_n
  3. Draw $Y \sim \pi(\cdot|X_{n-1,-i})$
  4. Set $X_n := (X_{n-1,1}, \ldots, X_{n-1,i-1}, Y, X_{n-1,i+1}, \ldots, X_{n-1,d})$
- Given the target distribution π, what are the optimal selection probabilities p?
- Pretend π is a Gaussian - optimal p is known for Gaussians - and it works outside the Gaussian class.
  (Chimisov + KL + Roberts 2017)
- How to verify containment?
- Simultaneous Geometric Drift will not work!!!

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - Adapting increasingly rarely

- $P_\gamma, \gamma \in \Gamma$ - a parametric family of π-invariant kernels;
  Adaptive MCMC steps:
  (1) Sample X_{n+1} from $P_{\gamma^n}(X_n, \cdot)$.
  (2) Given $\{X_0, .., X_{n+1}, \gamma^0, .., \gamma^n\}$ update γ^{n+1} according to some adaptation rule.
- How tweak the strategy to make theory easier?
- Do we need to adapt in every step?
- How about adapting increasingly rarely?
- **AirMCMC Sampler**
  Initiate $X_0 \in \mathcal{X}$, $\gamma^0 \in \Gamma$. $\overline{\gamma} := \gamma^0$ $k := 1$, $n := 0$.
  (1) For $i = 1, .., n_k$
    1.1. sample $X_{n+i} \sim P_{\overline{\gamma}}(X_{n+i-1}, \cdot)$;
    1.2. given $\{X_0, .., X_{n+i}, \gamma_0, .., \gamma_{n+i-1}\}$ update γ_{n+i} according to some adaptation rule.
  (2) Set $n := n + n_k$, $k := k + 1$. $\overline{\gamma} := \gamma_n$.
- Will such a strategy be efficient? With say $n_k = ck^\beta$
- Will it be mathematically more tractable?

AirMCMC - a simulation study

- $\pi(x) = \frac{l(|x|)}{|x|^{1+r}}, \; x \in \mathbb{R},$
- Air version of RWM adaptive scaling
- The example is polynomially ergodic (not easy for the sampler)
- **AirRWM**
  Initiate $X_0 \in \mathbb{R}$, $\overline{\gamma} \in [q_1, q_2]$. $k := 1$, $n := 0$, a sequence $\{c_k\}_{k \geq 1}$.
  (1) For $i = 1, .., n_k$
    (1.1.) sample $Y \sim N(X_{n+i-1}, \overline{\gamma})$, $a_{\overline{\gamma}} := \frac{\phi(Y)}{\phi(X_{n+i-1})}$;
    (1.2.) $X_{n+i} := \begin{cases} Y & \text{with probability} \quad a_{\overline{\gamma}}, \\ X_{n+i-1} & \text{with probability} \quad 1 - a_{\overline{\gamma}}; \end{cases}$
    (1.3.) $a := a + a_{\overline{\gamma}}$.
    If $i = n_k$ then
    $\overline{\gamma} := \exp\left(\log(\overline{\gamma}) + c_n\left(\frac{a}{n_k} - 0.44\right)\right)$.
  (2) Set $n := n + n_k$, $k := k + 1$, $a := 0$.

AirMCMC - a simulation study

- $\pi(x) = \frac{l(|x|)}{|x|^{1+r}}, \ x \in \mathbb{R},$
- Air version of RWM adaptive scaling
- The example is polynomially ergodic (not easy for the sampler)
- **AirRWM**
  Initiate $X_0 \in \mathbb{R}$, $\overline{\gamma} \in [q_1, q_2]$. $k := 1$, $n := 0$, a sequence $\{c_k\}_{k \geq 1}$.
  (1) For $i = 1, .., n_k$
    (1.1.) sample $Y \sim N(X_{n+i-1}, \overline{\gamma})$, $a_{\overline{\gamma}} := \frac{\phi(Y)}{\phi(X_{n+i-1})}$;
    (1.2.) $X_{n+i} := \begin{cases} Y & \text{with probability} \quad a_{\overline{\gamma}}, \\ X_{n+i-1} & \text{with probability} \quad 1 - a_{\overline{\gamma}}; \end{cases}$
    (1.3.) $a := a + a_{\overline{\gamma}}$.
    If $i = n_k$ then
    $\overline{\gamma} := \exp\left(\log(\overline{\gamma}) + c_n\left(\frac{a}{n_k} - 0.44\right)\right)$.
  (2) Set $n := n + n_k$, $k := k + 1$, $a := 0$.

AirMCMC - a simulation study

- $\pi(x) = \frac{l(|x|)}{|x|^{1+r}},\ x \in \mathbb{R},$
- Air version of RWM adaptive scaling
- The example is polynomially ergodic (not easy for the sampler)
- **AirRWM**
  Initiate $X_0 \in \mathbb{R}$, $\overline{\gamma} \in [q_1, q_2]$. $k := 1$, $n := 0$, a sequence $\{c_k\}_{k \geq 1}$.
  (1) For $i = 1, .., n_k$
    (1.1.) sample $Y \sim N(X_{n+i-1}, \overline{\gamma})$, $a_{\overline{\gamma}} := \frac{\phi(Y)}{\phi(X_{n+i-1})}$;
    (1.2.) $X_{n+i} := \begin{cases} Y & \text{with probability} \quad a_{\overline{\gamma}}, \\ X_{n+i-1} & \text{with probability} \quad 1 - a_{\overline{\gamma}}; \end{cases}$
    (1.3.) $a := a + a_{\overline{\gamma}}$.
    If $i = n_k$ then
    $\overline{\gamma} := \exp\left(\log(\overline{\gamma}) + c_n\left(\frac{a}{n_k} - 0.44\right)\right)$.
  (2) Set $n := n + n_k$, $k := k + 1$, $a := 0$.

AirMCMC - a simulation study

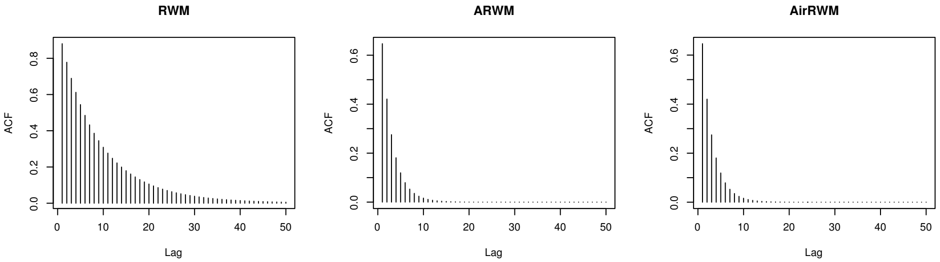
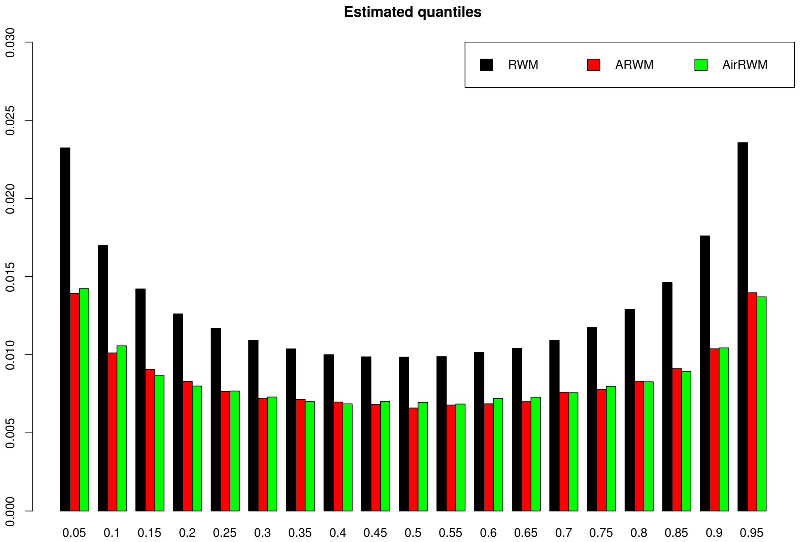
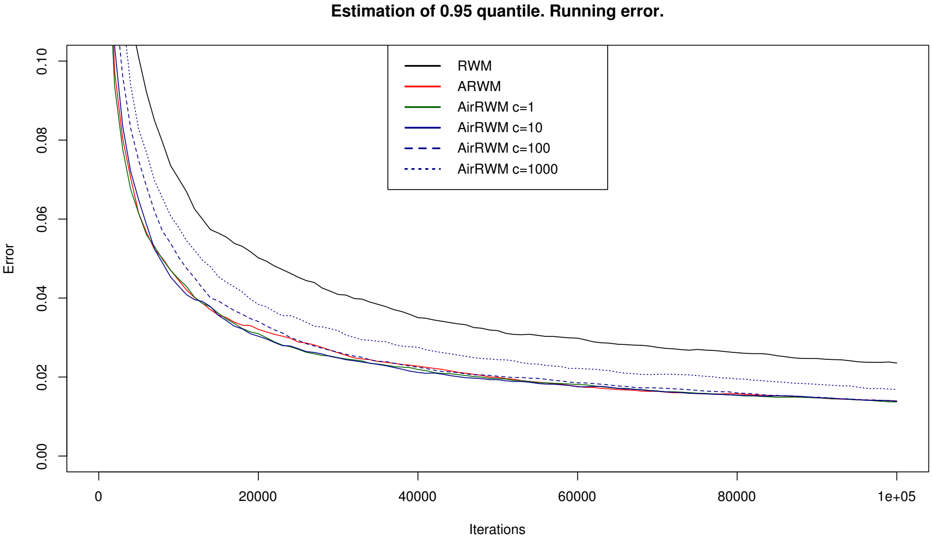


Figure: Autocorrelations (ACF)

AirMCMC - a simulation study



AirMCMC - a simulation study



AirMCMC - theory - the SGE case

- **Theorem 1**
  - Kernels SGE
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{1/2}(x)} < \infty$

  Then WLLN holds, and also for any $\delta \in (0, 2)$

  $\lim_{N\to\infty} \mathbf{E}\left|\frac{1}{N}\sum_{i=0}^{N-1} f(X_i) - \phi(f)\right|^{2-\delta} = 0,$
- **Theorem 2**
  - Kernels SGE and reversible
  - $\frac{d\nu}{d\pi} \in L_2(\mathcal{X}, \pi)$
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{\frac{\beta}{2(\beta+1)}-\delta}(x)} < \infty$, for some $\delta > 0$,

  Then SLLN holds.
- Note that diminishing adaptation is not needed!

AirMCMC - theory - the SGE case

- **Theorem 1**
  - Kernels SGE
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{1/2}(x)} < \infty$

  Then WLLN holds, and also for any $\delta \in (0, 2)$

  $\lim_{N\to\infty} \mathbf{E}\left|\frac{1}{N}\sum_{i=0}^{N-1} f(X_i) - \phi(f)\right|^{2-\delta} = 0,$
- **Theorem 2**
  - Kernels SGE and reversible
  - $\frac{d\nu}{d\pi} \in L_2(\mathcal{X}, \pi)$
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{\frac{\beta}{2(\beta+1}-\delta}(x)} < \infty$, for some $\delta > 0$,

  Then SLLN holds.
- Note that diminishing adaptation is not needed!

AirMCMC - theory - the SGE case

- **Theorem 1**
  - Kernels SGE
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{1/2}(x)} < \infty$

  Then WLLN holds, and also for any $\delta \in (0,2)$
  $\lim_{N\to\infty} \mathbf{E}\left|\frac{1}{N}\sum_{i=0}^{N-1} f(X_i) - \phi(f)\right|^{2-\delta} = 0,$
- **Theorem 2**
  - Kernels SGE and reversible
  - $\frac{d\nu}{d\pi} \in L_2(\mathcal{X}, \pi)$
  - $n_k \geq ck^{\beta}, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{\frac{\beta}{2(\beta+1}-\delta}(x)} < \infty$, for some $\delta > 0$,

  Then SLLN holds.
- Note that diminishing adaptation is not needed!

AirMCMC - theory - the local SGE case

- **Theorem 3**
  - Γ is compact
  - Kernels are locally SGE
  - $n_k \geq ck^\beta, \quad \beta > 0$, and adaptation takes place if in a compact set B
  - $\sup \frac{|f(x)|}{V_i^{1/2}(x)} < \infty$

  Then WLLN holds, and also for any $\delta \in (0, 2)$

  $\lim_{N\to\infty} \mathbf{E}\left|\frac{1}{N}\sum_{i=0}^{N-1} f(X_i) - \phi(f)\right|^{2-\delta} = 0,$

- **Theorem 4**
  - Γ is compact
  - Kernels are locally SGE and reversible
  - $\frac{d\nu}{d\pi} \in L_2(\mathcal{X}, \pi)$
  - $n_k \geq ck^\beta, \quad \beta > 0$, and adaptation takes place if in a compact set B
  - $\sup \frac{|f(x)|}{V_i^{\frac{\beta}{2(\beta+1)}-\delta}(x)} < \infty$, for some $\delta > 0$,

  Then SLLN holds.

AirMCMC - theory - the local SGE case

- **Theorem 3**
  - Γ is compact
  - Kernels are locally SGE
  - $n_k \geq ck^{\beta}, \quad \beta > 0$, and adaptation takes place if in a compact set B
  - $\sup \frac{|f(x)|}{V_i^{1/2}(x)} < \infty$

  Then WLLN holds, and also for any $\delta \in (0,2)$

  $$\lim_{N\to\infty} \mathbf{E}\left|\frac{1}{N}\sum_{i=0}^{N-1} f(X_i) - \phi(f)\right|^{2-\delta} = 0,$$

- **Theorem 4**
  - Γ is compact
  - Kernels are locally SGE and reversible
  - $\frac{d\nu}{d\pi} \in L_2(\mathcal{X}, \pi)$
  - $n_k \geq ck^{\beta}, \quad \beta > 0$, and adaptation takes place if in a compact set B
  - $\sup \frac{|f(x)|}{V_i^{\frac{\beta}{2(\beta+1}-\delta}(x)} < \infty$, for some $\delta > 0$,

  Then SLLN holds.

AirMCMC - theory - the SPE case

- **Theorem 5**
  - Kernels SPE with $\alpha > 2/3$
  - $\beta > \frac{2\alpha(1-\alpha)}{2\alpha-1}$ *if* $\alpha < \frac{3}{4}$
    $\beta > \frac{\alpha}{4\alpha-2}$ *if* $\alpha \geq \frac{3}{4}$.
  - $n_k \geq ck^\beta, \quad \beta > 0$
  - $\sup \frac{|f(x)|}{V^{3/2\alpha-1}(x)} < \infty$

  Then WLLN holds.