Design Strategies for Adaptive MCMC

Radu Craiu

Department of Statistics
University of Toronto
www.utstat.toronto.edu/craiu/

Statistical Society of Canada Meeting, May-June 2016

My Collaborators

Elif Acar
Shelley Bull
Zhijian Chen
Andryi Derkach
Reihaneh Entezari
Lawrence Gray
Thomas Lee
Evgeny Levi
Neal Madras
Benjamin Reiser
Jeffrey Rosenthal
Mian Wei
Fang Yao
Yan Bai
Roberto Casarin
Mariana Craiu
Antonio Di Narzo
Osvaldo Espin-Garcia
Caren Hasler
Fabrizio Leisen
Li Li
Xiao-Li Meng
Louis-Paul Rivest
Avideh Sabeti
Chao Yang
Jialin Zou
Sophie Baillargeon
Bo Chen
Wei Deng
Thierry Duchesne
Daniel Fortin
Krzysztof Latuszynski
Christiane Lemieux
Lizhen Li
Andrew Paterson
Gareth Roberts
Lei Sun
Jinyoung Yang

Outline

Markov Chain Monte Carlo

- A search for Markov chain Monte Carlo (or MCMC) articles on Google Scholar yields over 100,000 hits.
- A general web search on Google yields 1.7 million hits. Why so popular?
- MCMC algorithms are used to solve problems in many scientific fields, including physics (where many MCMC algorithms originated) and chemistry and computer science.
- The widespread popularity of MCMC samplers is largely due to their impact on solving statistical computation problems related to Bayesian inference.

Markov Chain Monte Carlo

- Given a sample $\vec{x} = \{x_1, \ldots, x_n\}$ from a *parametric sampling density* $f(x|\theta)$, where $x \in \mathcal{X} \subset \mathbf{R}^k$ and $\theta \in \Theta \subset \mathbf{R}^d$ AND a *prior density* $p(\theta)$ we are interested in the *posterior density*

$$\pi(\theta|\vec{x}) = \frac{p(\theta)f(\vec{x}|\theta)}{\int_\Theta p(\theta)f(\vec{x}|\theta)d\theta} \tag{1}$$

- Very often, denominator in (1) cannot be computed exactly so π cannot be studied.
- The Monte Carlo solution is to sample from π.

Markov Chain Monte Carlo

- We construct and run an aperiodic and irreducible Markov chain with transition $T(x_{old}, x_{new})$ that leaves π invariant

$$\int_{\mathcal{S}} \pi(x) T(x, y) dx = \pi(y).$$

- Unlike traditional Monte Carlo where the samples are independent, MCMC samplers produce dependent draws.
- A number of initial realizations from the chain are discarded (burn-in) and the remaining are used to estimate expectations or quantiles of functions of X.

A good friend: The Metropolis-Hastings algorithm

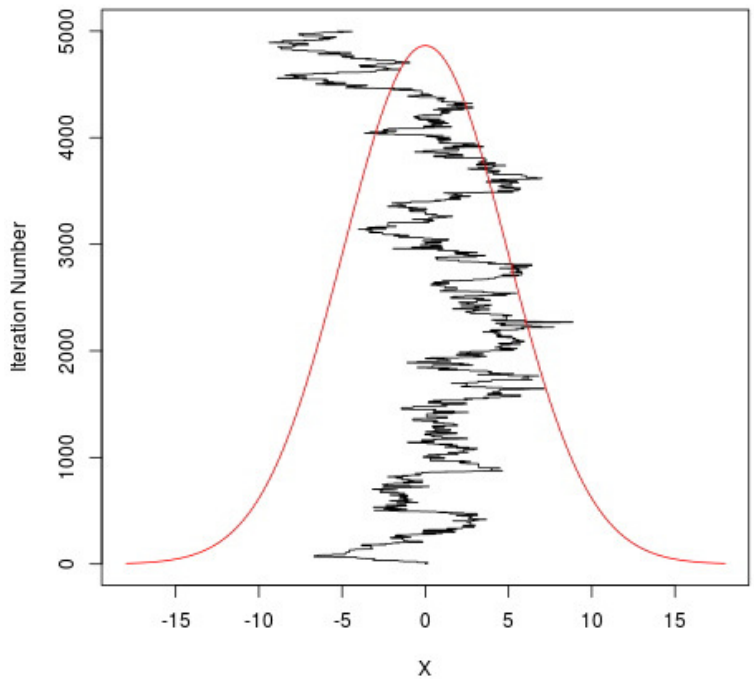
- The Metropolis-Hastings sampler is one of the most used algorithms in MCMC. It operates as follows:
  - Given the current state of the MC, x, a "proposed sample" y is drawn from a proposal density $q(y|x)$.
  - The proposal y is accepted with probability

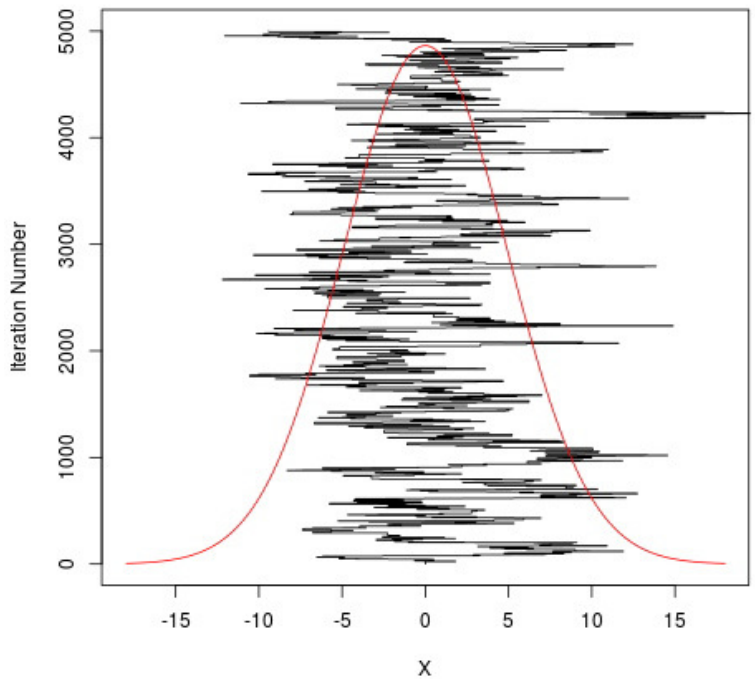
$$\min\left\{1, \frac{\pi(y)q(x|y)}{\pi(x)q(y|x)}\right\}.$$

  - If y is accepted, the next state is y, otherwise it is (still) x.
- The *random walk Metropolis (RWM)* is obtained when $y = x + \epsilon$ with $\epsilon \sim f$, f symmetric, usually $N(0, V)$.
- If $q(y|x) = q(y)$ we get the *Independent Metropolis (IM)* sampler.

Which proposal?

- How to find a good proposal distribution? We know that acceptance rates should be between 20-40%.
- Both RWM chains below accept 24% of the proposals.





Adaptive Metropolis

- The optimality results for the variance of a RWM in the case of Gaussian targets recommends that $V \propto Var(\pi)$ and the acceptance rate is 23.4% (Roberts and Rosenthal, Stat. Sci., '01).
- Non-Markovian Adaptation (Haario, Saksman and Tamminen (HST); Bernoulli, 2001). Involves re-using the past realizations of the Markov chain to modify the proposal distribution of a (RWM) algorithm.
- For instance, choose $q_t(y|x_t) = N(x_t, \Sigma_t)$ where $\Sigma_t \propto \text{SamVar}(\tilde{X}_t)$ and $\tilde{X}_t = (X_1, \ldots, X_t)$.

Adaptive MCMC

- Adaptive MCMC algorithms tune "on the go" the parameters of the proposal distribution (e.g. the variance V for RWM with Gaussian proposal) based on the available MC draws.
- Many MCMC methods perform local adaptation but "dance" around the Markovian property and manage to preserve it. Not AMCMC!
- Makes validation of an adaptive scheme more involved, but...
- ... Frees us to seek other practically useful designs!

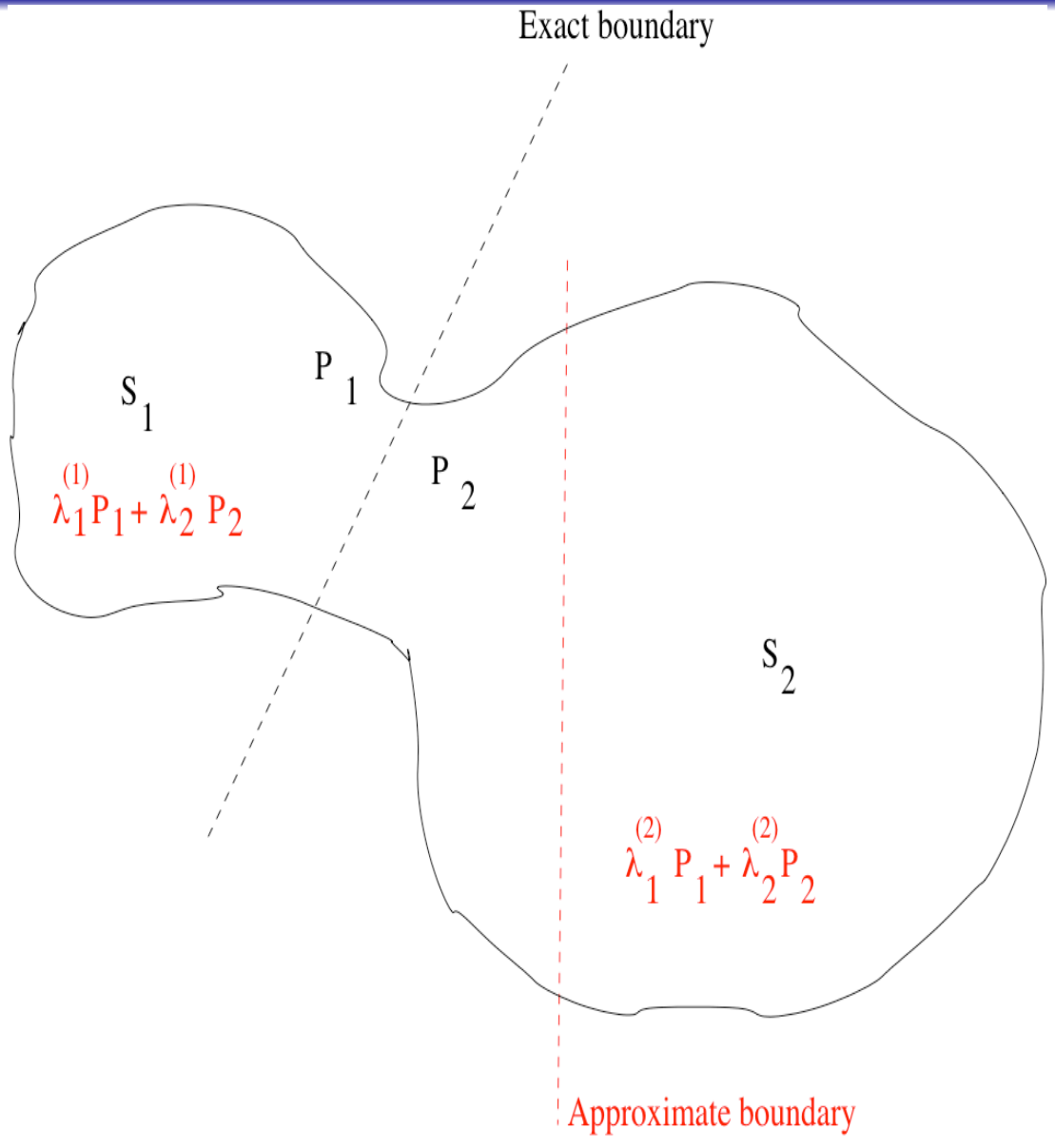
Example 1: Multimodal targets and regional AMCMC

- Multimodality is a never-ending source of headaches in MCMC.
- Chains constructed via generic algorithms often have trouble switching between modal regions.
- For multimodal distributions the "optimal" proposal may vary across regions of the sample space.

Regional AdaPTation (RAPT)

- Consider sampling using RWM from a distribution π with support in $\mathcal{S} \subset R^d$.
- Suppose $\mathcal{S} = \mathcal{S}_1 \uplus \mathcal{S}_2$ is such that depending on whether the current value of the chain is in $\mathcal{S}_1$ or in $\mathcal{S}_2$ the optimal variance of the RWM proposal is different.
- What type of adaptive algorithms can we design to address this problem?
- Adaptive MCMC must:
  - EXPLOIT: use efficiently the information about the target it collects;
  - EXPLORE: always look for new regions of the sample space that may not have been found yet.

RAPT (cont'd)


Exact boundary
S 1
P 1
P 2
S 2
λ(1) 1 P 1 + λ(1) 2 P 2
λ(2) 1 P 1 + λ(2) 2 P 2
Approximate boundary

RAPT (cont'd)

- If we approximate $\mathcal{S}_i$ then we must allow for some uncertainty regarding the distribution to be used in each $\mathcal{S}_i$ by sampling from a mixture of proposals.
- The mixture proportions are allowed to vary between regions and are adaptively adjusted based on the past realizations.
- In addition, the distributions entering the mixture are also adapted based on past realizations.

RAPT (cont'd)

- In region $\mathcal{S}_j$ we sample using the proposal

$$\tilde{P}_{jt}(X_t, \cdot) = \sum_{i=1}^{2} \lambda_i^{(j)} P_{it}(X_t, \cdot), \; j = 1, 2.$$

- Each P_{it} is adapted using samples from $\mathcal{S}_i$.
- The mixture weights $\lambda_i^{(j)}(t)$ are also adapted.

RAPT (cont'd)

- For instance, $\lambda_i^{(j)} = \frac{n_i^{(j)}(t)}{\sum_{h=1}^{K} n_h^{(j)}(t)}$ and

$$n_i^{(j)}(t) = \#\{ \text{accepted moves up to time } t \text{ from } \mathcal{S}_j \text{ when the proposal dist'n is } P_i \ .$$

- Will tend to favour proposals with high acceptance rates; these are usually the ones creating "small jumps" and thus not necessarily the best for our purpose.
- A better alternative is $\lambda_i^{(j)} = \frac{d_i^{(j)}(t) n_i^{(j)}(t)}{\sum_{h=1}^{K} d_h^{(j)}(t) n_h^{(j)}(t)}$ where,

$$d_i^{(j)}(t) = \text{average square root jump distance up to time } t \text{ from } \mathcal{S}_j \text{ when the proposal dist'n is } P_i$$

RAPT

- In addition to local optimality (within each region) we seek good global traffic (between regions) so we add a global component to the proposal distribution:
- The proposal distribution for RAPT is

$$\tilde{P}_t(X_t, \cdot) = (1-\alpha) \sum_{j=1}^{2} 1_{\mathcal{S}_j}(X_t) \left[\sum_{i=1}^{2} \lambda_{it}^{(j)} P_{it}(X_t, \cdot) \right] + \alpha Q_t(X_t, \cdot).$$

- The parameter $\alpha \in (0.1, 0.3)$ is fixed throughout.
- Caveat: Regions do not evolve. A bad guess will result in loss of efficiency.

RAPTOR - RAPT with Online Recursion

- Suppose

$$\tilde{q}_\eta(x) = \sum_{k=1}^{K} \beta^{(k)} N(x; \mu^{(k)}, \Sigma^{(k)}),$$

  where $\beta^{(k)} > 0$ for all $1 \leq k \leq K$ and $\sum_{k=1}^{K} \beta^{(k)} = 1$, is a good approximation for the target π.

- At each time n during the simulation process one has available n dependent Monte Carlo samples which are used to fit the mixture $\tilde{q}_\eta$ via an Online EM algorithm (Bai, Craiu and Di Narzo, JCGS 2011)

Definition of Regions (K=2)

- We define the partition $\mathcal{S} = \cup_{k=1}^{2} \mathcal{S}^{(k)}$ so that, on each set $\mathcal{S}^{(k)}$, π is more similar to $N(x; \mu^{(k)}, \Sigma^{(k)})$ than to any other mixture component.
- Define

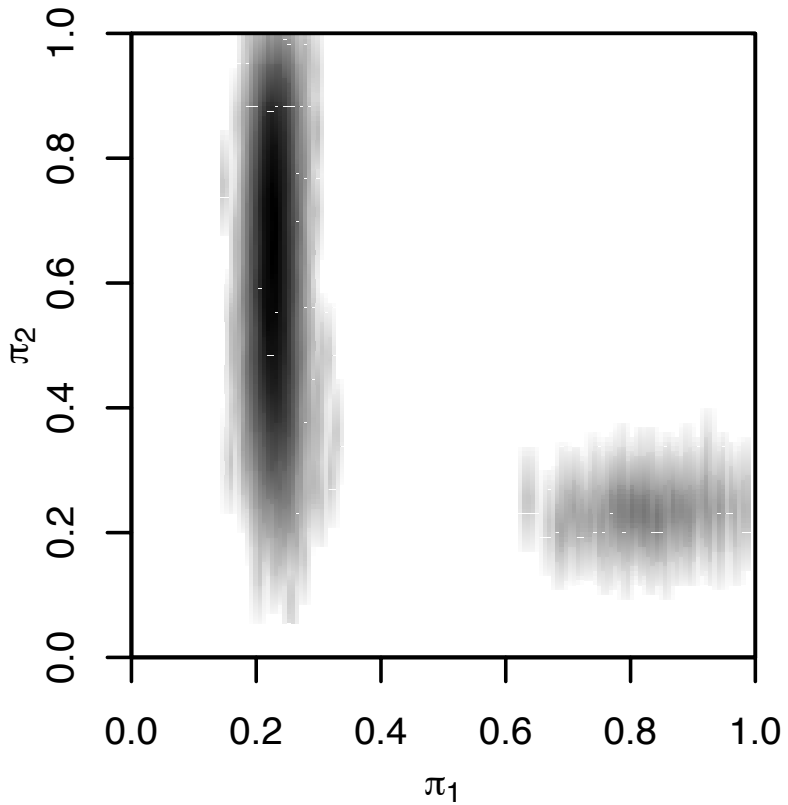
$$\mathcal{S}_n^{(1)} = \{x : \ N(x; \mu_n^{(1)}, \Sigma_n^{(1)}) > N(x; \mu_n^{(2)}, \Sigma_n^{(2)})\},$$
$$\mathcal{S}_n^{(2)} = \mathcal{S} \backslash \mathcal{S}_n^{(1)}.$$

Example: Genetic Instability of Esophageal Cancers

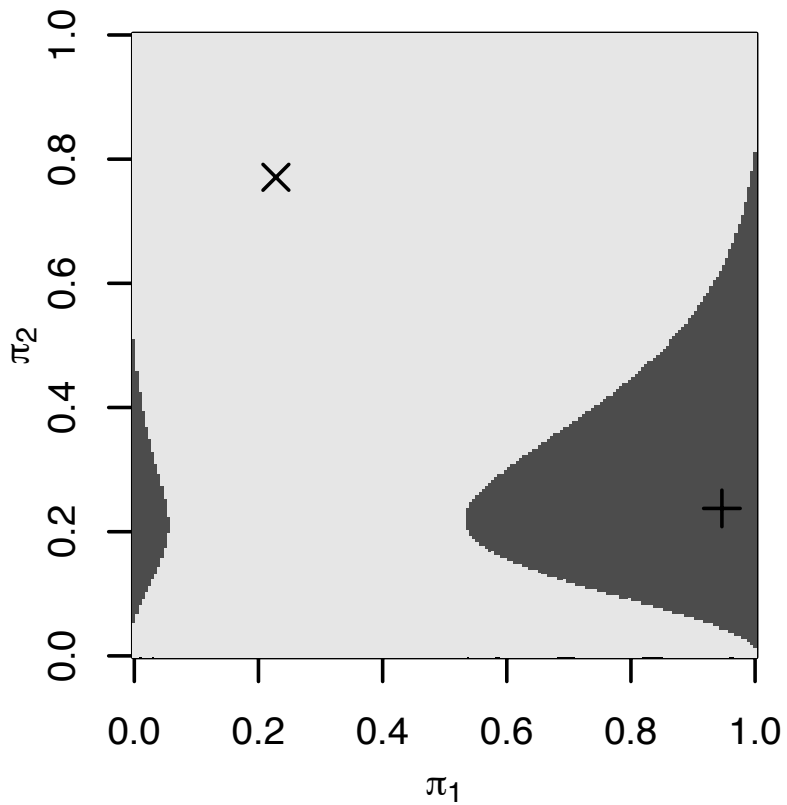
- Cancer cells suffer a number of genetic changes during disease progression, one of which is *loss of heterozygosity (LOH)*.
- Chromosome regions with high rates of LOH are hypothesized to contain genes which regulate cell behaviour and may be of interest in cancer studies.
- We consider 40 measures of frequencies of the event of interest (LOH) with their associated sample sizes. The model adopted for those frequencies is a mixture model

$$X_i \sim \eta \,\text{Binomial}(N_i, \pi_1) + (1-\eta)\,\text{Beta-Binomial}(N_i, \pi_2, \gamma).$$

LOH Analysis



Fixed Regions



Adaptive Regions

LOH Analysis - RAPT

	$\mathcal{S}^1$	$\mathcal{S}^2$	$\mathcal{S}$
η	0.897	0.079	0.838
π_1	0.229	0.863	0.275
π_2	0.714	0.237	0.679
γ	15.661	-14.796	13.435

Theory for AMCMC

- Consider an adaptive MCMC procedure, i.e. a collection of transition kernels $\{T_\gamma\}_{\gamma\in\Gamma}$ each of which has π as a stationary distribution. One can think of γ as being the *adaptation parameter*.
- At iteration n we use the transition kernel $T_{\gamma_n}(X_n,\cdot)$ for the adaptive chain.
- Let $D_n = \sup_{X\in\mathcal{X}} ||T_{\gamma_{n+1}}(X,\cdot) - T_{\gamma_n}(X,\cdot)||_{TV}$ $(||\mu-\nu||_{TV} = \sup_{A\in\mathcal{X}} |\mu(A)-\nu(A)|)$.
- Let M_ϵ be the ϵ-*convergence time function* $M_\epsilon : \mathcal{X}\times\Gamma \to \mathbf{N}$

$$M_\epsilon(x,\gamma) = \inf\{n \geq 1 : ||P^n_\gamma(x,\cdot) - \pi(\cdot)||_{TV} \leq \epsilon\}$$

Two sufficient conditions

- Diminishing Adaptation:

$$\lim_{n\to\infty} D_n = 0 \text{ in probability.}$$

  This is an intuitive condition, relatively easy to verify or ensure.

- Containment Condition:
  For any $X_0 = x_0 \in \mathcal{X}$ and $\Gamma_0 = \gamma_0 \in \mathcal{Y}$, $\epsilon > 0$ the stochastic process $M(X_n, \Gamma_n)$ is bounded in probability $\mathbf{P}_{(x_0,\gamma_0)}$.
- CC is hard to ensure and prove/disprove.

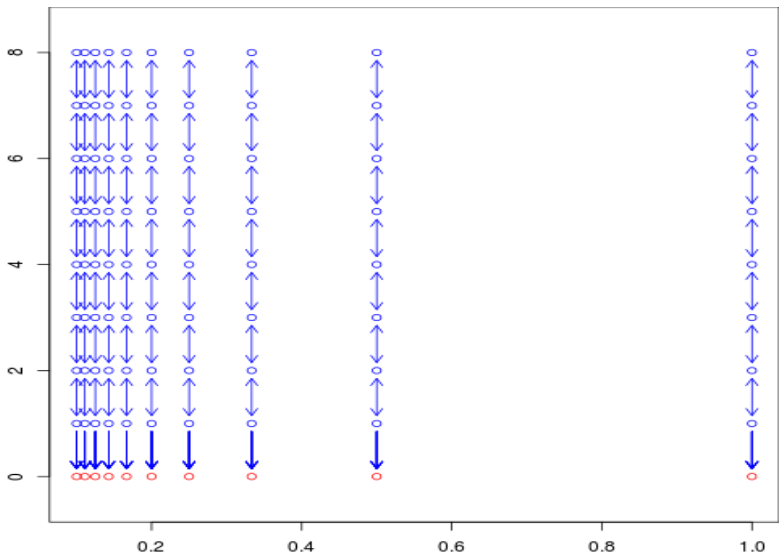
An apparently simple question

- CC can be avoided if $\mathcal{X}$ is compact.
- Q1: Can we restrict adaptation on a compact in $\mathcal{X}$ so CC is no longer needed?.
- Let $\mathcal{X}$ be the sample space endowed with a metric η.
- Let T be a fixed transition kernel on $\mathcal{X}$ with stationary distribution π.
- The chain's jumps are bounded by $D < \infty$, i.e.

$$T(x, \{y \in \mathcal{X} : \ \eta(x, y) \leq D\}) = 1, \ \forall x \in \mathcal{X}.$$

- Q2: Is it possible to modify the transition kernel inside a compact K such that the process is no longer bounded in probability (while maintaining the bounded jump constraint)?

A counterexample



- $\mathcal{X} = \{(\frac{1}{i}, j) : i \in \mathbb{N}^*, j = 0, 1, \ldots\}$

$$\pi(\frac{1}{i}, j) = 2^{-i} \left(\frac{1}{i}\right) \left(1 - \frac{1}{i}\right)^j$$

- π restricted to each $\mathcal{X}_i$ is geometric with mean i
- $K = \cup_i \{(\frac{1}{i}, 0)\}$ is a bounded set.
- The larger the column number i, the higher is the conditional mean of π on $\mathcal{X}_i$.

A counterexample

- Fixed chain:
  - Outside K the MC has a ± 1 MH kernel reversible w.r.t π.
  - Inside K the MC is irreducible and reversible with respect to π.
- Modified chain:
  - The modified chain proceeds within K using the rule: if $X_n \in K$, then $X_{n+1} = (\frac{1}{n},\ 1)$
  - Given an arbitrary $L > 0$ it can be shown that $\underline{\lim}_{n \to \infty} \Pr(X_{n,2} \geq L) \geq 1/2$
  - Details and other examples in Craiu et al. (2015, Ann. Appl. Prob).

Adaption made ~~easy~~ easier

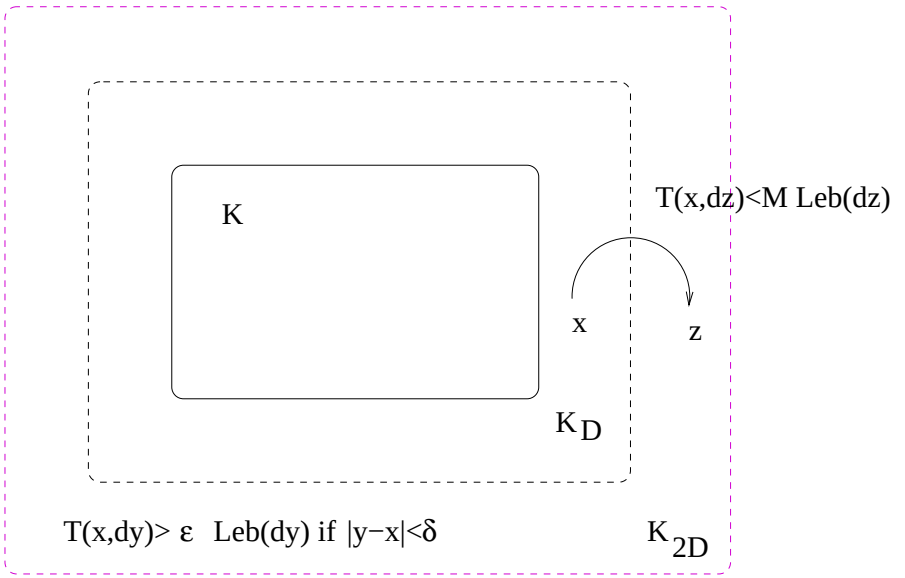
- The process $\{X_n\}$ has adaptive transition kernel P_{Γ_n},

  $$\Pr(X_{n+1} \in A \,|\, X_n = x,\, \Gamma_n = \gamma,\, X_0, \ldots, X_{n-1}, \Gamma_{1:(n-1)}) \;=\; T_\gamma(x, A)\,.$$

- The transition probabilities $T_\gamma(x, dy)$ have densities that are continuous wrt to x and y (subsequently further relaxed by Yang and Rosenthal, 2015).
- $T_\gamma\left(x,\ \{y \in \mathcal{X} : |y - x| \le D\}\right) \;=\; 1.$
- $T_\gamma(x, A) \;=\; T(x, A)\,, \quad x \in \mathcal{X} \setminus K\,,$

Adaption made ~~easy~~ easier

- The fixed kernel used outside K must satisfy:
  - $\exists\ M < \infty \quad s.t. \quad T(x, dz) \leq M\text{Leb}(dz)$, $x \in K_D \setminus K,\ z \in K_{2D} \setminus K_D$.
  - $T(x, dy) \geq \epsilon\text{Leb}(dy)$ whenever $x, y \in J$ with $|y - x| < \delta$, $K_{2D} \setminus K_D \subseteq J \subseteq \mathcal{X}$.



These are conditions that are within reach through careful selection of the fixed kernel T.

A regime change algorithm (RCA)

- Run a fixed sampler during the initialization period.
- This sampler may now be ideal but will provide some information about the target.
- Given a compact K perform a regime change to an adaptive sampler inside K
- Example: Gibbs to Independent Metropolis.

Lupus Data

Table: *The number of latent membranous lupus nephritis cases (numerator), and the total number of cases (denominator).*

	IgA				
ΔIgG	0	0.5	1	1.5	2
-3.0	0/ 1	-	-	-	-
-2.5	0/ 3	-	-	-	-
-2.0	0/ 7	-	-	-	0/ 1
-1.5	0/ 6	0/ 1	-	-	-
-1.0	0/ 6	0/ 1	0/ 1	-	0/ 1
-0.5	0/ 4	-	-	1/ 1	-
0	0/ 3	-	0/ 1	1/ 1	-
0.5	3/ 4	-	1/ 1	1/ 1	1/ 1
1.0	1/ 1	-	1/ 1	1/ 1	4/ 4
1.5	1/ 1	-	-	2/ 2	-

RCA for Probit Regression

- For each patient $i = 1, \ldots, 55$, we model

$$Y_i \sim \text{Bernoulli}(\Phi(x_i^T \beta)),$$

  and $p(\beta) \propto 1$.
- The posterior is thus

$$\begin{aligned} \pi_{PR}(\vec{\beta} | \vec{Y}, \vec{IgA}, \vec{\Delta IgG}) &\propto \prod_{i=1}^{55} \Big[\Phi(\beta_0 + \Delta IgG_i \beta_1 + IgA_i \beta_2)^{Y_i} \times \\ &\times (1 - \Phi(\beta_0 + \Delta IgG_i \beta_1 + IgA_i \beta_2))^{(1-Y_i)} \Big] \end{aligned}$$

RCA for Probit Regression

- State of the art: PX-DA algorithm of Van Dyk and Meng (JCGS, 2001).
- Draw

$$\phi_i^{(t+1)} \sim \begin{cases} N_+(x_i^T \beta^{(t)}, 1), & \text{if } Y_i = 1 \\ N_-(x_i^T \beta^{(t)}, 1), & \text{if } Y_i = 0 \end{cases},$$

  Set $\phi^{(t+1)} = (\phi_1^{(t+1)}, \ldots, \phi_n^{(t+1)})$.
- Let $\tilde{\beta}^{t+1} = (X^T X)^{-1} X^T \phi^{(t+1)}$ and define $R^{(t+1)} = \sum_{i=1}^n (\phi_i^{(t+1)} - x_i^T \tilde{\beta}^{(t+1)})^2$
- Sample $Z \sim N(0,1)$, $W \sim \chi_n^2$ and set $\beta^{(t+1)} = \sqrt{\frac{W}{R^{(t+1)}}} \tilde{\beta}^{(t+1)} + \text{Chol}[(X^T X)^{-1}] Z$

RCA for Probit Regression

- Set $\mu_n := \frac{\ll X_0 \gg + \ll X_1 \gg + \ldots + \ll X_{n-1} \gg}{n}$, and $\Sigma_n := \mathrm{Cov}(\ll X_0 \gg, \ll X_1 \gg, \ldots, \ll X_{n-1} \gg) + \epsilon I_d$, where Cov is the empirical covariance function and $\ll r \gg_i = \max[-L, \min(L, r_i)]$.
- K is the ball centred at μ_M, of radius $\max_{1 \le i \le d} (\Sigma_M)_{ii}^{1/2}$ (i.e., the largest sample standard deviation on the diagonal of Σ_M). And, we let D be any suitably large distance bound (e.g. $D = 20$)
- Note that the compact K is not "large" but is calibrated so that the normal approximation is appropriate.

RCA

- If $X_n \in K^c$, then $X_{n+1} \sim P_{PX}(X_n, \cdot)$.
- If $X_n \in K$ and $d(X_n, K^c) > 1$, then

$$X_{n+1} \sim P_{AD}(X_n, \cdot)$$

where $P_{AD}(X_n, \cdot)$ has density

$$\lambda_{n+1} P_{\mu_n, \Sigma_n}(X_n, \cdot) + (1 - \lambda_{n+1}) P_{PX}(X_n, \cdot),$$

- $\lambda_n = \min[\max(\theta_n, 0.2), 0.8]$, and θ_n is the empirical acceptance rate of the IM proposals.

RCA

- As we approach the boundary of K we need to smoothly transfer regimes: from IM to RWM.
- If $X_n \in K$ and $d(X_n, K^c) = u$ with $0 \leq u \leq 1$, then

$$X_{n+1} \sim u\, P_{AD}(X_n, \cdot) + + (1-u)\, P_{PX}(X_n, \cdot)\,,$$

with λ_n as above.

RCA for Probit Regression

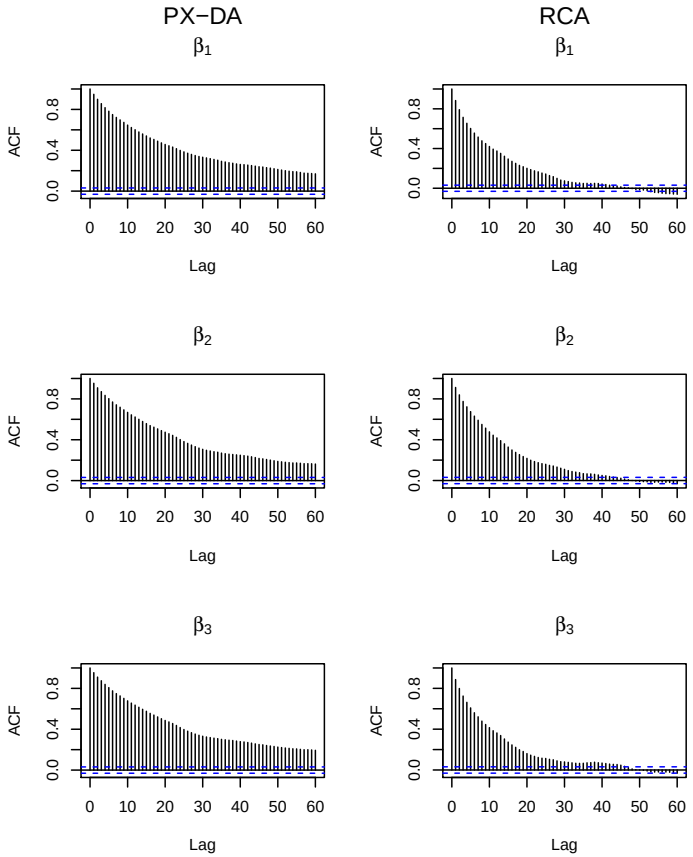


Figure: *Effective sample sizes are increased 300-400%.*

References

- Learn from Thy Neighbor: Parallel-Chain and Regional Adaptive MCMC. *JASA*, 2009
- Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC. *JCGS*, 2011.
- Stability of Adversarial Markov chains, with an Application to Adaptive MCMC Algorithms. *Ann. Appl. Prob.*, 2015.
- Adaptive Strategies for the Multiple-Try Metropolis within Gibbs. Submitted - available on arXiv, 2016.

Web: www.utstat.toronto.edu/craiu/Papers/index.html