Exploration Vs. Exploitation in Adaptive Monte Carlo Sampling

Scott C. Schmidler

Department of Statistical Science
Duke University

AdapSkIII Conference
January 4, 2010

Overview of talk

- Mixing times and finite time convergence of adaptive MCMC.
- Combining adaptive strategies.
- Some cautionary notes about MIS kernels.

References

Available from "www.stat.duke.edu/s̃cs":

- SS & Woodard (2010). Lower Bounds on the Convergence Rates of Adaptive MCMC Methods. (Submitted, under revision)
- Wang & SS (2010). Exploration vs Exploitation: Hybrid Strategies for Adaptive MCMC. (in preparation) – see poster
- Ji & SS (2010). Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection. *J. Comp. Graph. Stat.*, (to appear).
- SS & Wiehe (2010). Reservoir Exchange and Adaptive Monte Carlo. (submitted)

Adaptive MCMC

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.

Adaptive MCMC

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.
- Purpose of adaptation is to improve *rate* of convergence.

Adaptive MCMC

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.
- Purpose of adaptation is to improve *rate* of convergence.
- Convergence of MC estimators involves *both* bias *and* variance.

Adaptive MCMC

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.
- Purpose of adaptation is to improve *rate* of convergence.
- Convergence of MC estimators involves *both* bias *and* variance.
- Different adaptation strategies can be understood as improving one or the other.

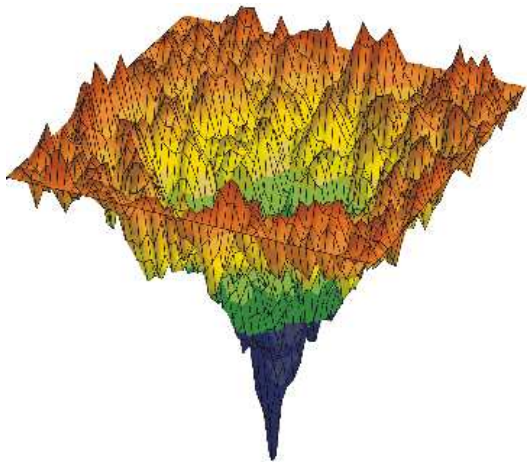
Adaptive MCMC

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.
- Purpose of adaptation is to improve *rate* of convergence.
- Convergence of MC estimators involves *both* bias *and* variance.
- Different adaptation strategies can be understood as improving one or the other.
- Can obtain improved algorithms by combining strategies of different types.

High dimensional, multimodal target distributions

- Molecular simulation
- Bayesian variable selection/model selection
- Mixture models
- Non-linear physics-based models



I. Adaptive Metropolis kernels

Two approaches developed by various authors

Adaptive random-walk proposals

$$q_{n+1}(x, \cdot) = (1 - \alpha)N(x, \hat{\Sigma}_n) + \alpha N(x, \Sigma_0)$$

e.g. Haario *et al*, Roberts & Rosenthal

Adaptive independence proposals (AMIS)

$$q_{n+1}(x, \cdot) = g(\cdot; \hat{\theta}_n) \quad \hat{\theta}_n = \theta(X_1, \ldots, X_n)$$

e.g. Andrieu & Moulines, Ji & Schmidler, etc.

Adaptive Metropolized independence sampler (AMIS) [Ji and Schmidler, 2009]

Finite mixture proposal distribution:

$$q(x) = \lambda N(x; \tilde{\mu}, \tilde{\Sigma}) + (1 - \lambda) \sum_{m=1}^{M} w_m N(x; \mu_m, \Sigma_m)$$

(see also Andrieu & Moulines 2005, others)

Point-mass mixture proposal for variable selection:

$$q(x) = (1 - \lambda) \Big[w_0 \delta(x) + \sum_{m=1}^{M} w_m N(\mu_m, \Sigma_m) \Big] + \lambda N(x; \tilde{\mu}, \tilde{\Sigma})$$

Adapt parameters $\psi = \{w_m, \mu_m, \Sigma_m\}_{m=1}^{M}$ to approximate $\pi(x)$.

Adaptive Metropolized independence sampler (AMIS)

Adaptive strategy: Minimize $\mathcal{D}\left[\pi(x) \parallel q(x;\psi)\right] = \mathbb{E}_\pi\left[\log \frac{\pi(x)}{q(x;\psi)}\right]$

ψ^* obtained as a root of derivative:

$$h(\psi) = -\int \frac{\pi(x)}{q(x;\psi)} \frac{\partial}{\partial\psi} q(x;\psi) = 0$$

Approximate $h(\psi)$ by Monte Carlo integration:

$$h(\psi) \approx \frac{1}{K}\sum_{k=1}^{K} f(X^{(k)},\psi) \quad \text{for} \quad f(x,\psi) = \frac{\partial}{\partial\psi}[\log \frac{\pi(x)}{q(x;\psi)}]$$

where $X^{(k)} \sim \pi(x)$.

$\hat{h}(X^{(1:K)}; \psi)$: estimate of $h(\psi)$ based on sample path $X^{(1:k)}$

Stochastic Approximation algorithm [Robbins and Monro, 1951].

$$\begin{aligned}\psi_{n+1} &= \psi_n + r_{n+1}(h(\psi_n) + \xi_{n+1}) \\ &= \psi_n + r_{n+1}\,\hat{h}(X_n^{(1:K)}; \psi_n)\end{aligned}$$

$\{r_n\}$ decreasing step-sizes satisfying $\sum_n r_n = \infty$ and $\sum_n r_n^2 < \infty$

Resulting chain is non-Markovian, but can be shown to satisfy a WLLN using results of [Roberts and Rosenthal, 2007]

Example: Logistic regression

Bayesian logistic regression model,

$$y_i \mid x_i, \beta \sim \text{Bernoulli}\left(g^{-1}(x_i\beta)\right) \qquad \beta \sim \pi_0(\beta)$$

$y_i \in \{0, 1\}$; $g(u)$ logistic link

Simulated data set:

- 200 observations
- $r = 10$ covariates
- $\beta_{1:10} = [-.01, -1.5, .15, .5, -.15, -.2, -.6, .25, 1.5, -.05]$

Bayesian logistic regression

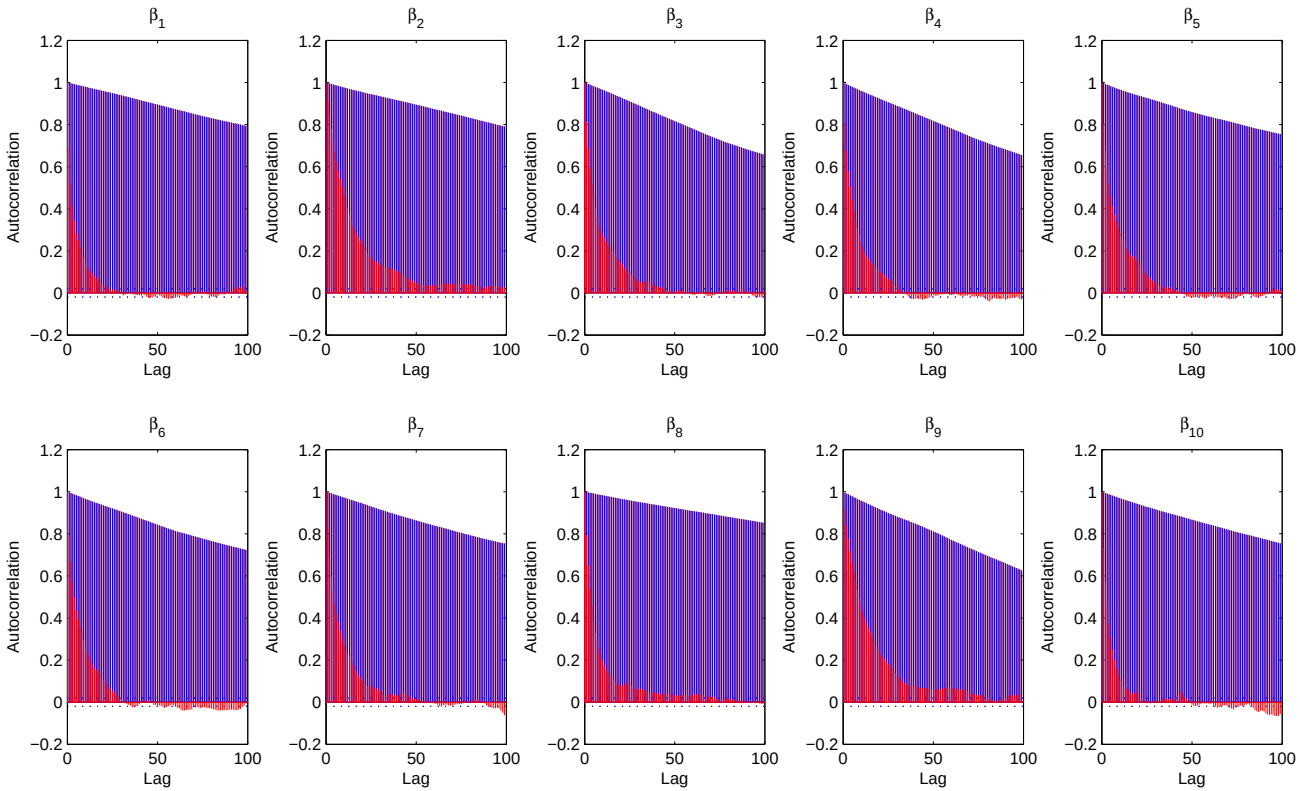


Figure: Autocorrelation of $\beta_{1:10}$ under data-augmentation Gibbs sampler [Holmes and Held, 2006] (blue), and adaptive MCMC algorithm (red).

Equi-Energy Sampler [Kou et al., 2006]

Constructs I processes $X^{(i)}$ with tempered target densities $\pi^{(i)} \propto \pi^{\beta_i}$ for inverse temperatures $1 = \beta_1 > \ldots > \beta_I \geq 0$.
(Also truncation $\pi^{(i)} \propto \pi^{\beta_i} \wedge c_i$)

For each i, bin sample history $(X^{(i)}_{0:n})$ according to energy.

Process $X^{(i)}$ occasionally proposes to move to a state previously visited by $X^{(i+1)}$ lying in same energy bin.

These "equi-energy" moves can be non-local in the state space, potentially enabling transitions between distinct modes of π.

MRAM Processes

Let $X^{(1)}, \ldots, X^{(I)}$ discrete time stochastic processes on $\mathcal{X}$.
So $X^{(i)} = X_0^{(i)}, X_1^{(i)}, \ldots$

Generated by time-inhomogeneous sequences of transition kernels:

$$K_{i,n} = \alpha\, T_i + (1 - \alpha) R_{i,n}$$

with $\alpha \in [0, 1]$, T_i an ergodic time-homogeneous Markov $\pi^{(i)}$-reversible transition kernel, and $R_{i,n}$ is a *resampling* kernel with proposal:

$$Q_{i,n}(X_{n-1}^{(i)}, y) = \sum_{i'=1}^{I} \sum_{j=0}^{n-1} w_{i'j} \delta(y - X_j^{(i')})$$

(Proposes new state from the set of previous samples $X_{0:n-1}^{(1:I)}$.)

MRAM Algorithms

Mulitchain resampling adaptive Metropolis (MRAM):

- Equi-Energy Sampler
- Importance-Resampling from the Past (Atchadé)
- Gelfand-Sahu

IAMC Processes

Let $\{T_\theta\}_{\theta \in \Theta^{(i)}}$ be a set of ergodic, $\pi^{(i)}$-reversible Markov kernels.

$T_{\theta_{i,n}}$ time-inhomogeneous *but* $\pi^{(i)}$-*invariant* transition kernels

$\theta_{i,n} = g_i(X_{0:n-1}^{(1:I)})$

Examples: Haario algorithm and similar variants; multi-chain algorithm of Rosenthal *et al*.

Convergence theorems

$X_1, \ldots, X_n$ no longer a Markov chain.

Under what conditions does $\hat{f}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)$ converge?

- Haario *et al* 2001: WLLN, using "mixingales"
- Andrieu & Robert (2001): SA interpretation of Haario algorithm
- Andrieu & Moulines (2005), Atchadé & Rosenthal (2005): generalizations to other algorithms (and a CLT)
- Roberts & Rosenthal (2007): Simplified conditions, coupling

Adaptive MCMC theory

Nearly all theory to date deals with *ergodicity* (LLN).
A few give conditions for CLTs (e.g Andrieu & Moulines (2005)).

This was needed and a major breakthrough IMO. But all *asymptotic* theory.

Adaptation is only interesting if it improves *rates*!

Efficiency of MCMC

Statistical Efficiency: $var(\hat{f})$

Under reasonably weak conditions*, for any function f with $var_\pi(f) \leq \infty$, we obtain a CLT:

$$\sqrt{n}(\bar{f}_n - \mu_f) \rightarrow N(0, \sigma^2_{\bar{f}_n})$$

for $\sigma^2_{\bar{f}_n} = \sigma^2_f(1 + 2\sum_{j=1}^{n}(1 - \frac{j}{n})\rho_j)$ and

$$\rho_j = \frac{1}{\sigma^2_f} E\left((f(X^{(n)}) - \mu_f)(f(X^{(n+j)}) - \mu_f)\right)$$

lag-j autocorrelation.

Efficiency revisited

Asymptotic efficiency:
Relies on CLT, asymptotic variance = integrated autocorrelation

Finite sample efficiency:
Convergence as well as autocorrelation

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

Efficiency revisited

Asymptotic efficiency:
Relies on CLT, asymptotic variance = integrated autocorrelation

Finite sample efficiency:
Convergence as well as autocorrelation

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

For multimodal targets, bias can dominate in MCMC.

Efficiency revisited

Asymptotic efficiency:
Relies on CLT, asymptotic variance = integrated autocorrelation

Finite sample efficiency:
Convergence as well as autocorrelation

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

For multimodal targets, bias can dominate in MCMC.
For good adaptive MCMC algorithms, bias *will* dominate.

MCMC Theory

- Ergodicity: SLLN under usual conditions (ϕ-irred, aper, π-invariant)
- Geometric: $\exists \lambda \in [0,1)$ and $M(x) < \infty$ $(\pi - \text{a.e.} x \in \mathcal{X})$ s.t.

$$\|\mu K^n - \pi\| \leq M(x)\lambda^n$$

  Requires minorization, drift conditions. Implies CLT.
- Uniform: $M(x) \equiv M$
- Rapid mixing: λ grows at most polynomially in d (Note G.E. requires only $\lambda^* > 0$; e.g.holds for any $|\mathcal{X}| < \infty$)
- Quantitative: e.g. Rosenthal 1995

Rapidly mixing Markov chains

Let $(\mathcal{X}^{(d)}, \mathcal{F}^{(d)}, \lambda^{(d)})$ a sequence of measure spaces, and $\pi^{(d)}$ densities wrt $\lambda^{(d)}$ for $d \in \mathbb{N}$ the *problem size*.

Mixing time

$$\tau_\epsilon = \sup_{\pi_0} \min\{n : \|\pi_{n'} - \pi\|_{\mathrm{TV}} < \epsilon \quad \forall n' \geq n\}.$$

where

$$\|\pi_n - \pi\|_{\mathrm{TV}} = \sup_{A \subset \mathcal{X}} |\pi_n(A) - \pi(A)|$$

P is *rapidly mixing* if τ_ϵ is bounded above by a polynomial in d.

P is *torpidly mixing* if τ_ϵ is bounded below by an exponential in d.

Rapidly mixing Markov chains

For $\mathcal{X}$ finite or compact, we have

A sequence of transition kernels $P^{(d)}$ is *rapidly mixing* if $\mathbf{Gap}(P^{(d)})$ decreases at most polynomially in d.

$P^{(d)}$ is *torpidly (or slowly) mixing* if $\mathbf{Gap}(P^{(d)})$ decreases exponentially in d.

Compare to *geometric ergodicity*, which requires only $\mathbf{Gap}(P^{(d)}) > 0$. (true for *any* $|\mathcal{X}| < \infty$.)

Spectral bounds for reversible Markov chains

Convergence bounds for Markov chains:

- Spectral bounds (reversibilize if needed); or operator norm
  - E.g. conductance and canonical paths
- Coupling (minorization/drift)

Adaptive algorithms *aren't Markov chains!*

Produce non-Markovian, time-inhomogeneous, irreversible stochastic processes.

How to obtain bounds?

Bounding convergence of Adaptive MCMC processes

We obtain lower bounds on mixing times via the *hitting time* for subsets $A \subset \mathcal{X}$:

$$H_A = \min_i H_A^{(i)} \qquad\qquad H_A^{(i)} = \min\{n : X_n^{(i)} \in A\}$$

and involving the familiar *conductance* of a π-reversible Markov kernel T:

$$\Phi_T = \inf_{\substack{A \subset \mathcal{X}: \\ 0 < \pi(A) < 1}} \Phi_T(A) \qquad\qquad \Phi_T(A) = \frac{\int_A \pi(dv)\, T(v, A^c)}{\pi(A)\pi(A^c)}$$

$\Phi_T(A)$ captures the probability of moving between A and A^c
Φ_T quantifies the worst "bottleneck".

Hitting times and mixing times

Key idea:

$$\Pr(H_A \leq n) \leq \pi(A) - \epsilon \quad \Rightarrow \quad \|\pi_n - \pi\|_{\mathrm{TV}} \geq \epsilon$$
$$\Rightarrow \quad \tau_\epsilon > n$$

$\Rightarrow$ We can lower bound mixing times via bounds on hitting times.

Lower bounds on MRAM mixing

Theorem (SW09)

For any $\epsilon > 0$ and any $A \subset \mathcal{X}$ such that $0 < \pi^{(i)}(A) < 1$ for all i, the mixing time τ_ϵ^* of the MRAM satisfies:

$$\tau_\epsilon^* \geq (\pi(A) - \epsilon) \left[cI \max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1}.$$

Here $\gamma(A, i) = \min\{1, \pi^{(i)}(A)/\pi(A)\}$ is the *persistence* defined by Woodard,Schmidler,Huber (2007).

Single chain

Note appearance of the conductance:

Corollary

For any $0 < \epsilon < 1/4$, the mixing time τ_ϵ^* of an adaptive sampler based on T, with $I = 1$, satisfies:

$$\tau_\epsilon^* \geq \frac{1}{4\Phi_T}.$$

Corollary

Slow mixing of the Markov chain with transition kernel T implies slow mixing of any MRAM process based on T that has $I = 1$.

Note also the similarity of the bound for MRAM processes:

$$\tau_\epsilon^* \geq (\pi(A) - \epsilon) \left[cI \max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1}.$$

to the bound obtained by Woodard, Schmidler, Huber (2007) for non-adaptive swapping processes:

$$\tau_\epsilon^* \geq 2^{-8} \ln(2\epsilon)^{-1} \left[\max_i \gamma(A, i) \Phi_{T_i}(A) \right]^{-1/2}.$$

Normal mixtures

Mixtures of normals

$$\pi(z) = \frac{1}{2} N_M(z; -1_M, \sigma_1^2 \mathrm{I}_M) + \frac{1}{2} N_M(z; 1_M, \sigma_2^2 \mathrm{I}_M)$$

Theorem (WSH07a): Tempering is rapidly mixing for $\sigma_1 = \sigma_2$.
Theorem (WSH07b): Tempering is torpidly mixing for $\sigma_1 \neq \sigma_2$.

Theorem (SW09): EES is torpidly mixing for $\sigma_1 \neq \sigma_2$.

IAMC Processes

Similar (hitting time) argument gives:

Theorem (SW09)

Haario and multi-chain samplers are torpidly mixing on mixture-of-normals problem.

Efficiency revisited

Asymptotic efficiency:
Relies on CLT, asymptotic variance = integrated autocorrelation

Finite sample efficiency:
Convergence as well as autocorrelation

$$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

$\Rightarrow$ MRAM and IAMC sampling can only improve autocorrelation piece!

Overview of talk

- Mixing times and finite time convergence of adaptive MCMC.
- Combining adaptive strategies.
- Some cautionary notes about MIS kernels.

Types of MCMC adaptation

Suggests considering alternative "adaptation" strategies.

I: IAMC (adaptive random walks, AMIS)

II MRAM (equi-energy)

Types of MCMC adaptation

Suggests considering alternative "adaptation" strategies.

I: IAMC (adaptive random walks, AMIS)

II MRAM (equi-energy)

(III) Modifying the stationary distribution

- Wang-Landau, and generalizations (Atchade & Liu, Liang)
- Multi-canonical
- Metadynamics (Parisi et al)

Have received much interest in physics literature; recently adopted for statistical problems. (Liang, Atchade & Liu).

Generalized Wang-Landau (Atchade & Liu, 2009)

Partition state space $\mathcal{X} = \mathcal{X}_0 \cup \ldots \cup \mathcal{X}_k$ according to predefined energy levels $-\infty \leq e_0 < e_1 < \cdots < e_k \leq \infty$.

Goal: Sample from $\tilde{\pi}(x) = \sum_{i=1}^{k} \frac{\pi(x)}{\pi(\mathcal{X}_i)} \mathbf{1}_{\mathcal{X}_i}(x)$ *uniform energy*

Algorithm: Adaptively estimate $\hat{\pi}_n(i) \approx \pi(\mathcal{X}_i)$ by SA:
$\{\gamma_n\}$ a sequence of decreasing positive numbers.
Initialize $\phi_0(i) > 0$ for $i = 1, \ldots, k$, and $\hat{\pi}_0(i) = \frac{\phi_0(i)}{\sum_j \phi_0(j)}$

(i) Sample $X_{n+1} \sim \sum_{i=1}^{k} \frac{\pi(x)}{\hat{\pi}_n(i)} \mathbf{1}_{\mathcal{X}_i}(x)$ by MH.

(ii) Set $\phi_{n+1}(i) = \phi_n(i) \left(1 + \gamma_{a_n} \mathbf{1}_{\{X_{n+1} \in \mathcal{X}_i\}}\right)$; $\hat{\pi}_{n+1}(i) = \frac{\phi_{n+1}(i)}{\sum_j \phi_{n+1}(j)}$.

(iii) If $\max_i \left| v_{\kappa,n+1}(i) - \frac{1}{k} \right| \leq \frac{c}{k}$ where $v_{\kappa,n}(i) = \frac{1}{n-\kappa} \sum_{j=\kappa+1}^{n} \mathbf{1}_{\{X_j \in \mathcal{X}_i\}}$
then set $\kappa = n+1$ and $a_{n+1} = a_n + 1$, otherwise $a_{n+1} = a_n$.

Types of MCMC adaptation

These ways of adapting address fundamentally different problems:

I & II: Improve mixing of chain among regions of target distribution *already visited*

- Improves autocorrelation of chain
- In general cannot help in exploring previously unseen regions

Call these *Exploitation* methods.

III: Tries to push chain away from points "like" those already seen.

- *Can* help in finding new regions; improve *mixing time*.
- May suffer from high autocorrelation.

Call these *Exploration* methods.

Caveats

Note:

- Not rigorous statement for *all* IAMC methods on *all* targets; depends on form of kernel. But clear that method's power is essentially limited by these choices.

Caveats

Note:

- Not rigorous statement for *all* IAMC methods on *all* targets; depends on form of kernel. But clear that method's power is essentially limited by these choices.
- Some authors (Craiu *et al*, Heaton & Schmidler) use multiple parallel chains to aid exploration in IAMC. Can help in practice but ultimately limited by ability to initialize well.

Caveats

Note:

- Not rigorous statement for *all* IAMC methods on *all* targets; depends on form of kernel. But clear that method's power is essentially limited by these choices.
- Some authors (Craiu *et al*, Heaton & Schmidler) use multiple parallel chains to aid exploration in IAMC. Can help in practice but ultimately limited by ability to initialize well.
- No method will work for *all* problems - some are provably hard (see e.g Schmidler & Woodard, in prep). Can hope for improved behavior on practical problems.

Hybrid adaptation strategies

Can we combine types to achieve best of both?
Yes but requires some care.

One approach: Mixture kernels

$$K_{\text{adapt}} = \alpha K_{\text{exploit}} + (1 - \alpha) K_{\text{explore}}$$

Suffers problems in multimodal examples (Wiehe & Schmidler, 2010).

Alternative approach:

Run exploration chain independently in parallel, but use samples to augment AMIS approximation.

Exploration/Exploitation Algorithm (Wang & SS, 2010)

1. Run two chains in parallel: X^{WL} and $X^{\text{AMIS+}}$
2. Every N_c iterations, update the proposal distribution for $X^{\text{AMIS+}}$.
3. At iteration $n = m * N_c$, let E_n be the energy ring of $X_{n-1}^{\text{AMIS+}}$. Form KDE $\hat{f}$ by adding the samples $\{X_1^{\text{WL}}, \ldots, X_n^{\text{WL}}\}$ to those in E_n.
4. Propose $X_n^{\text{AMIS+}}$ from $\hat{f}_c$.
5. At other iterations, run the two chains independently.

Improving on (generalized) Wang-Landau

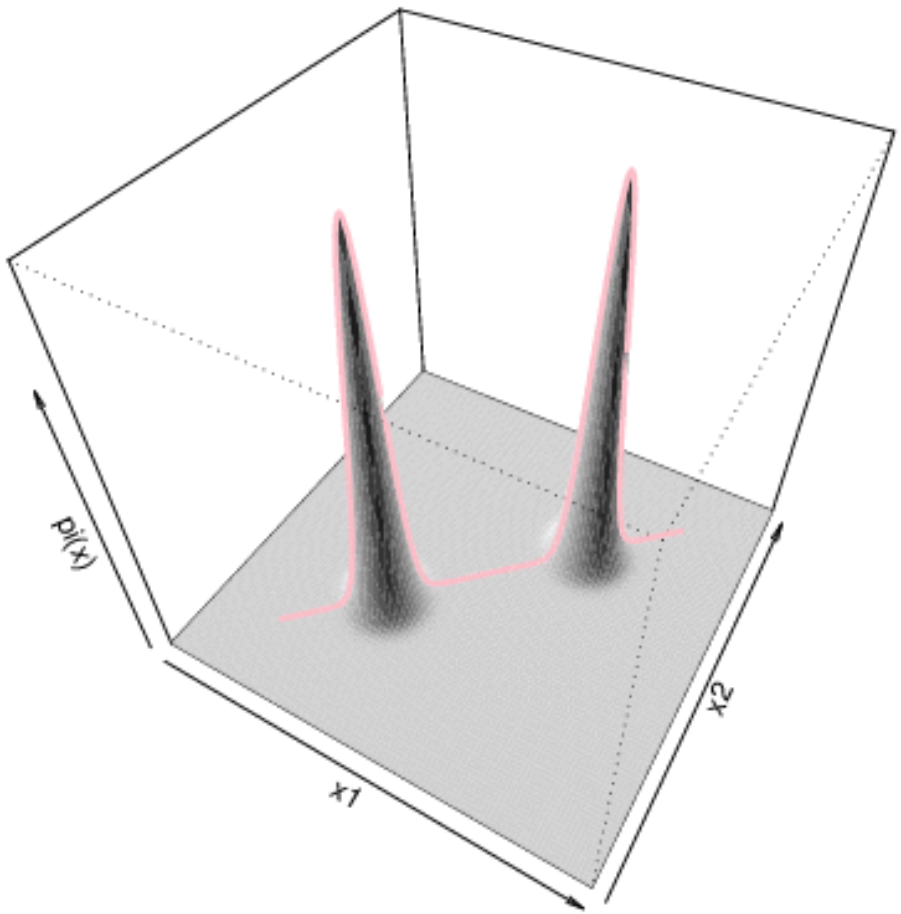
But . . .

Problem 1: Performance of the WL algorithm depends heavily on a good choice of the energy rings $E_0, \ldots, E_k$: number, spacing, max.

Recommended heuristics:
Estimate highest energy, lowest, form geometric progression.

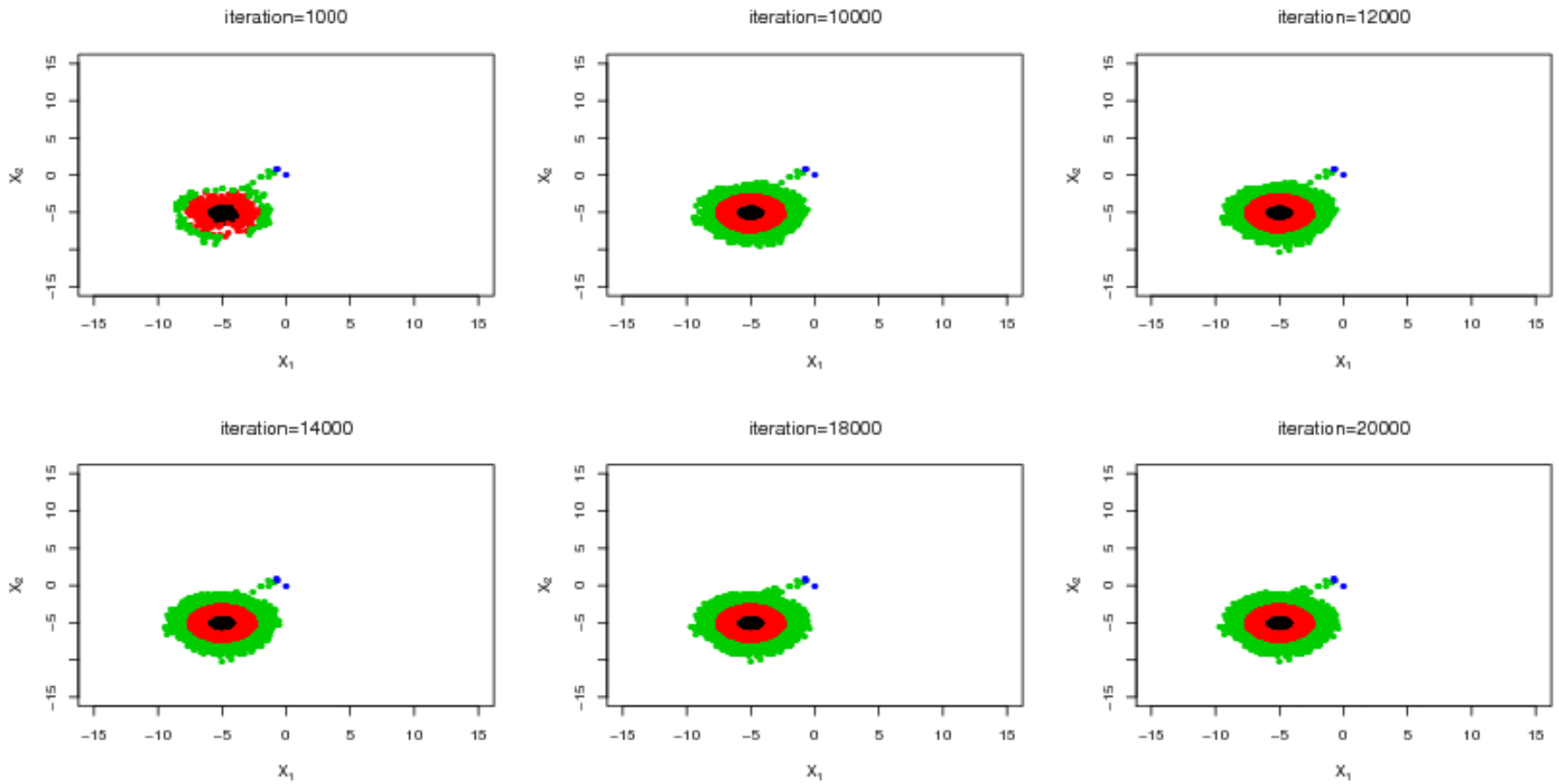
Example

Figure: Normal mixture with modes at (-5,-5) and (5,5)

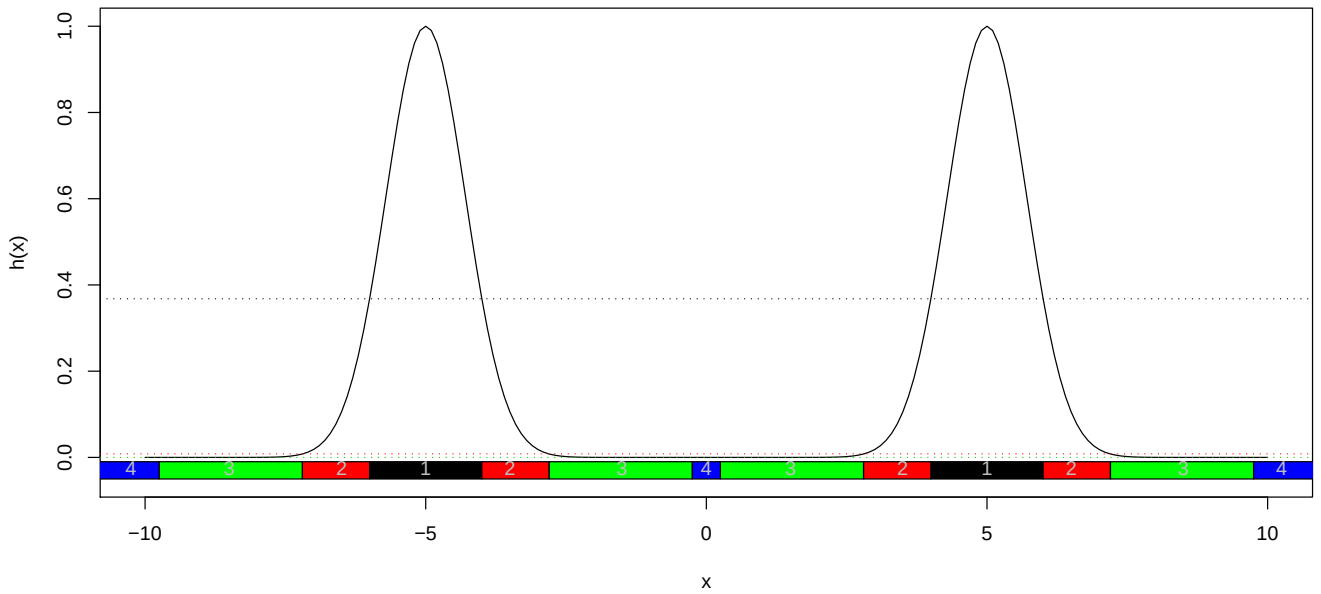


Example

Figure: Example 2, modes at (-5,-5) and (5,5)



Example



(b) $d = 4$, fixed energy levels

Conductance argument yields *provably* slow mixing.

Energy level adaptation scheme

Performance of the WL algorithm depends heavily on a good choice of the energy rings $E_0, \ldots, E_k$.

We introduce an adaptive scheme to make updating energy levels fully automatic:

1. Initialize by a geometric progression:

   $$e_0 = \inf_x E(x) = 0,\ e_1 = 1,\ e_2 = r_e, \ldots, E_{k-1} = r_e^{k-2}, E_k = \infty.$$

2. Every n_{split} iterations: if any $|\log(\phi_i) - \log(\phi_{i+1})| > E$, divide the i-th energy ring by adding a new $e^*_{i+1} = e_i \times \sqrt{\frac{e_{i+1}}{e_i}}$, again using geometric progression. Set $\log(\phi^*_{i+1}) = 0$.
3. Also update the second largest e_i;

   $$E^*_{k-1} = \frac{E^2_{k-1}}{E_k}$$

   Set $\log(\phi^*_k) = 0$.

Adaptive Energy Generalized Wang-Landau (AE-GWL)

Algorithm: Adaptively estimate $\hat{\pi}_n(i) \approx \pi(\mathcal{X}_i)$ by SA:
$\{\gamma_n\}$ a sequence of decreasing positive numbers.
Initialize $\phi_0(i) > 0$ for $i = 1, \ldots, k$, and $\hat{\pi}_0(i) = \frac{\phi_0(i)}{\sum_j \phi_0(j)}$

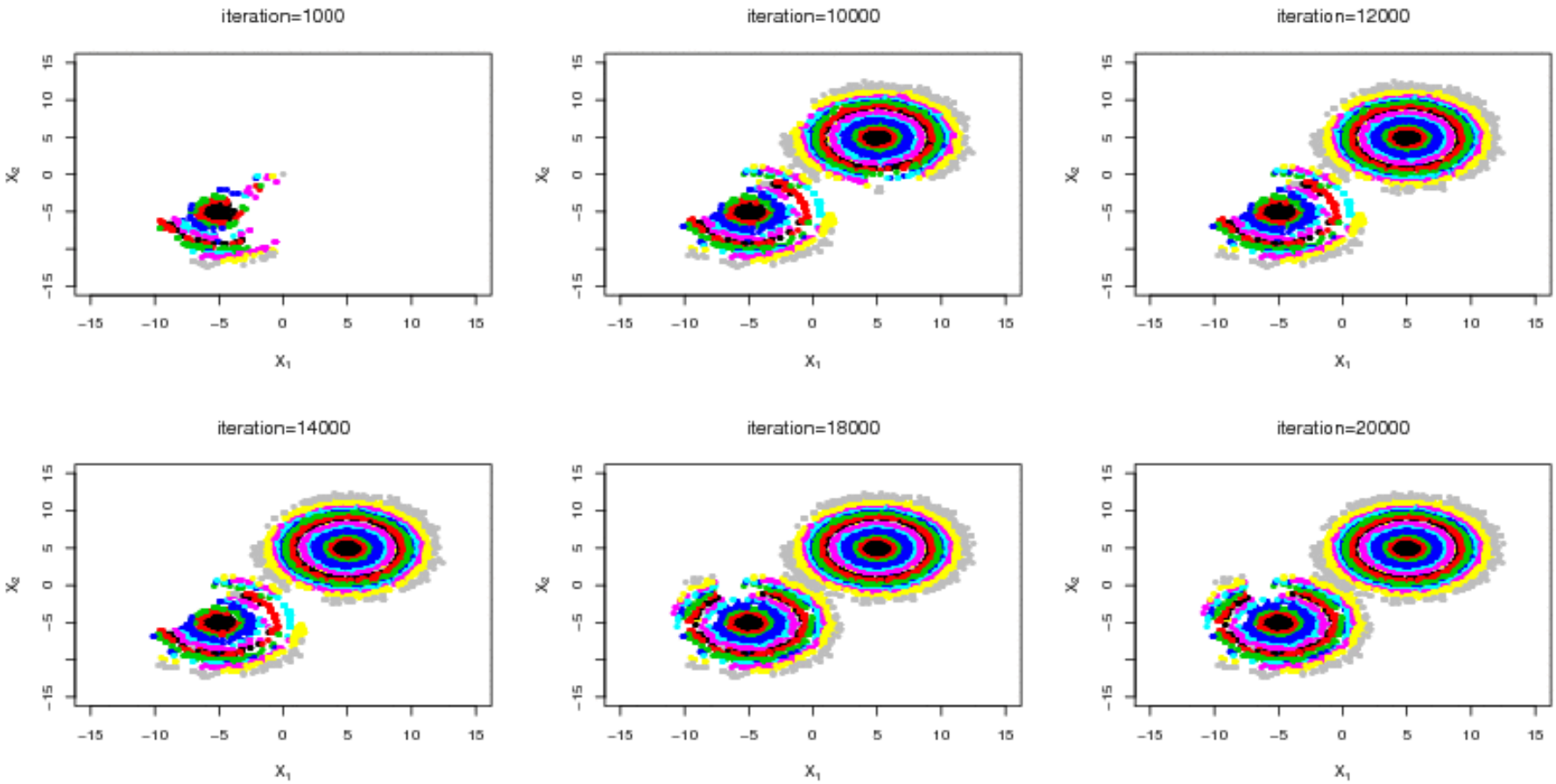
(i) Sample $X_{n+1} \sim \sum_{i=1}^{k} \frac{\pi(x)}{\hat{\pi}_n(i)} \mathbf{1}_{\mathcal{X}_i}(x)$ by MH.

(ii) Set $\phi_{n+1}(i) = \phi_n(i)\left(1 + \gamma_{a_n} \mathbf{1}_{\{X_{n+1} \in \mathcal{X}_i\}}\right)$ and
$\hat{\pi}_{n+1}(i) = \frac{\phi_{n+1}(i)}{\sum_j \phi_{n+1}(j)}$.

(iii) If $\max_i \left| v_{\kappa,n+1}(i) - \frac{1}{k} \right| \leq \frac{c}{k}$ where $v_{\kappa,n}(i) = \frac{1}{n-\kappa} \sum_{j=\kappa+1}^{n} \mathbf{1}_{\{X_j \in \mathcal{X}_i\}}$
then set $\kappa = n+1$ and $a_{n+1} = a_n + 1$, otherwise $a_{n+1} = a_n$.

(iv)* **For every n_{split} iterations, adaptively update $E = \{E_i\}$.**

Example



(c) $d = 4$, update internal energy levels

Improving on (generalized) Wang-Landau

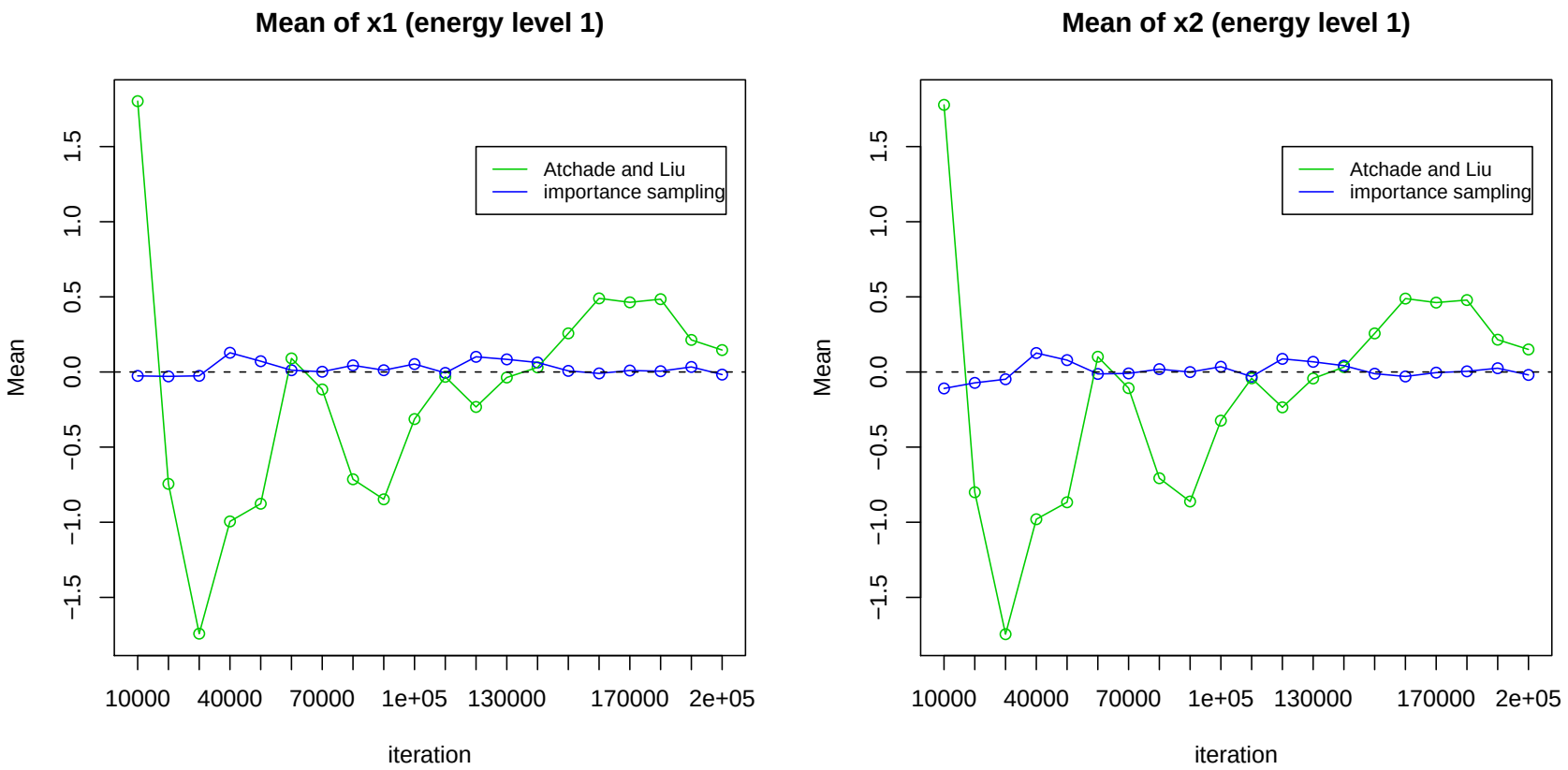
Problem 2: Wang-Landau for MC integration: converges in limit, but can be slow. WL inefficient for MC integration.

Reweighting complicated due to WL process.

We use importance sampling:

1. Use all samples to find a kernel density estimate $\hat{f}$.
2. Importance resampling: compute importance weights $w_i = \frac{h(x_i)}{\hat{f}(x_i)}$ and resample $x_1, \ldots, x_m$.
3. Form kernel density estimate $\hat{\pi}$ from $x_1, \ldots, x_m$.

Importance sampling vs ergodic averaging



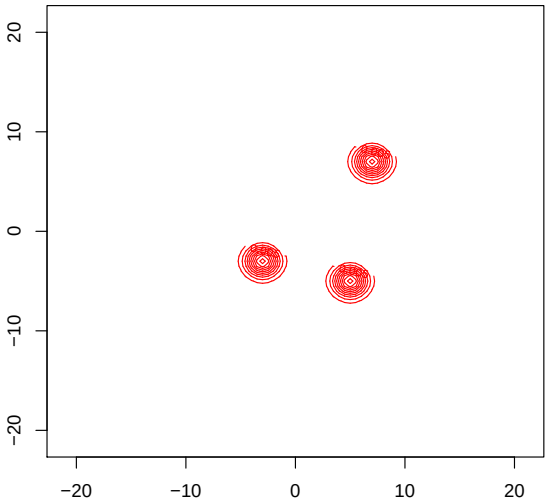
(d)

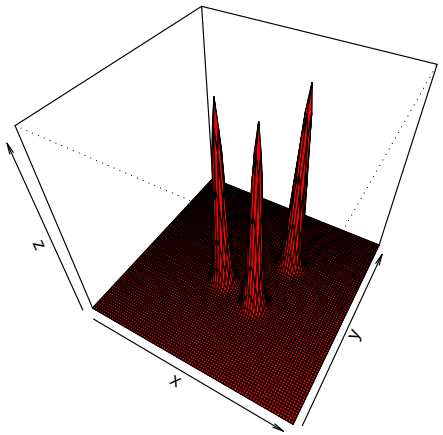
Exploration/Exploitation Algorithm

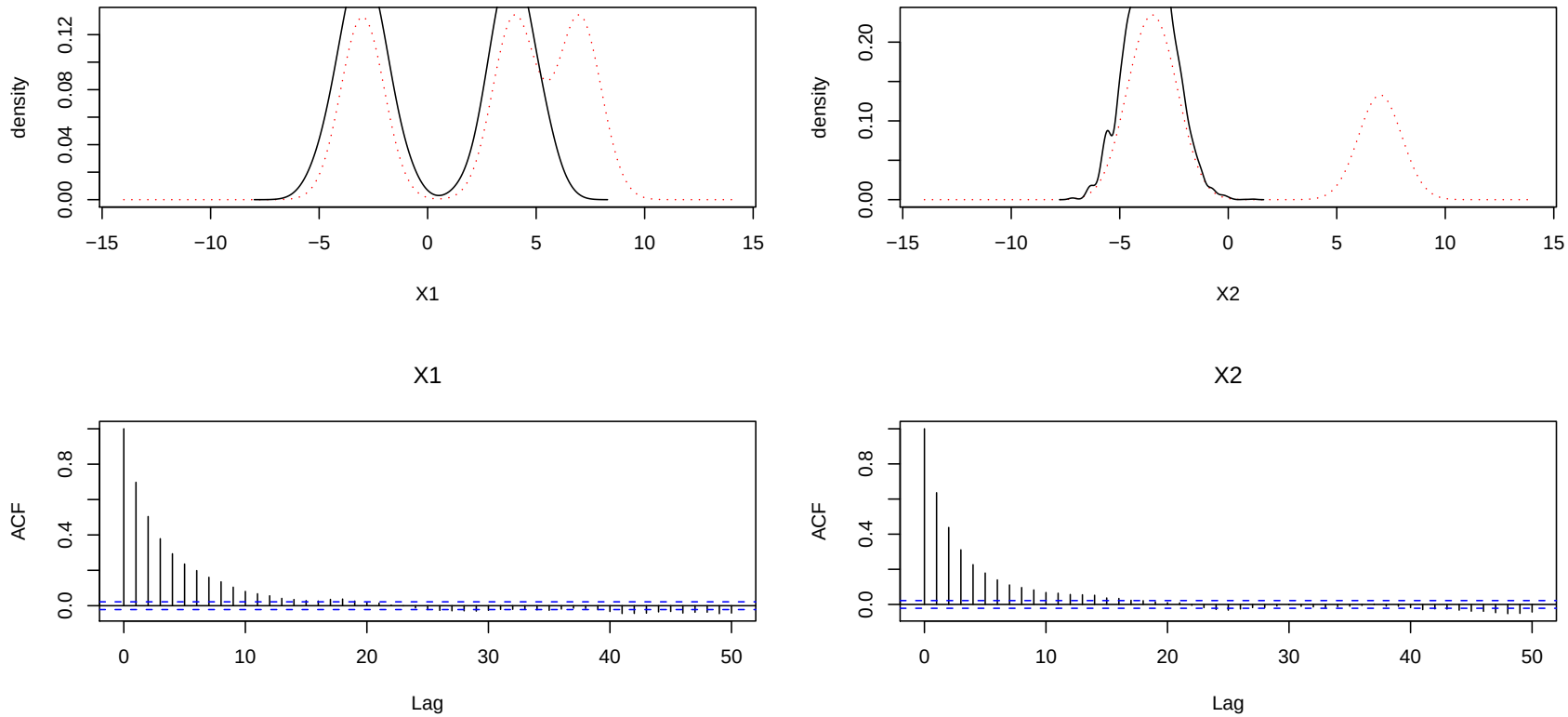
1. Run two chains in parallel: $X^{\text{AE-WL}}$ and $X^{\text{AMIS+}}$
2. Every N_c iterations, update the proposal distribution for $X^{\text{AMIS+}}$.
3. At iteration $n = m * N_c$, let E_n be the energy ring of $X_{n-1}^{\text{AMIS+}}$. Form KDE $\hat{f}$ by adding the samples $\{X_1^{\text{AE-WL}}, \ldots, X_n^{\text{AE-WL}}\}$ to those in E_n.
4. Propose $X_n^{\text{AMIS+}}$ from $\hat{f}_c$.
5. At other iterations, run the two chains independently.

Example: Trimodal distribution in $d = 2$

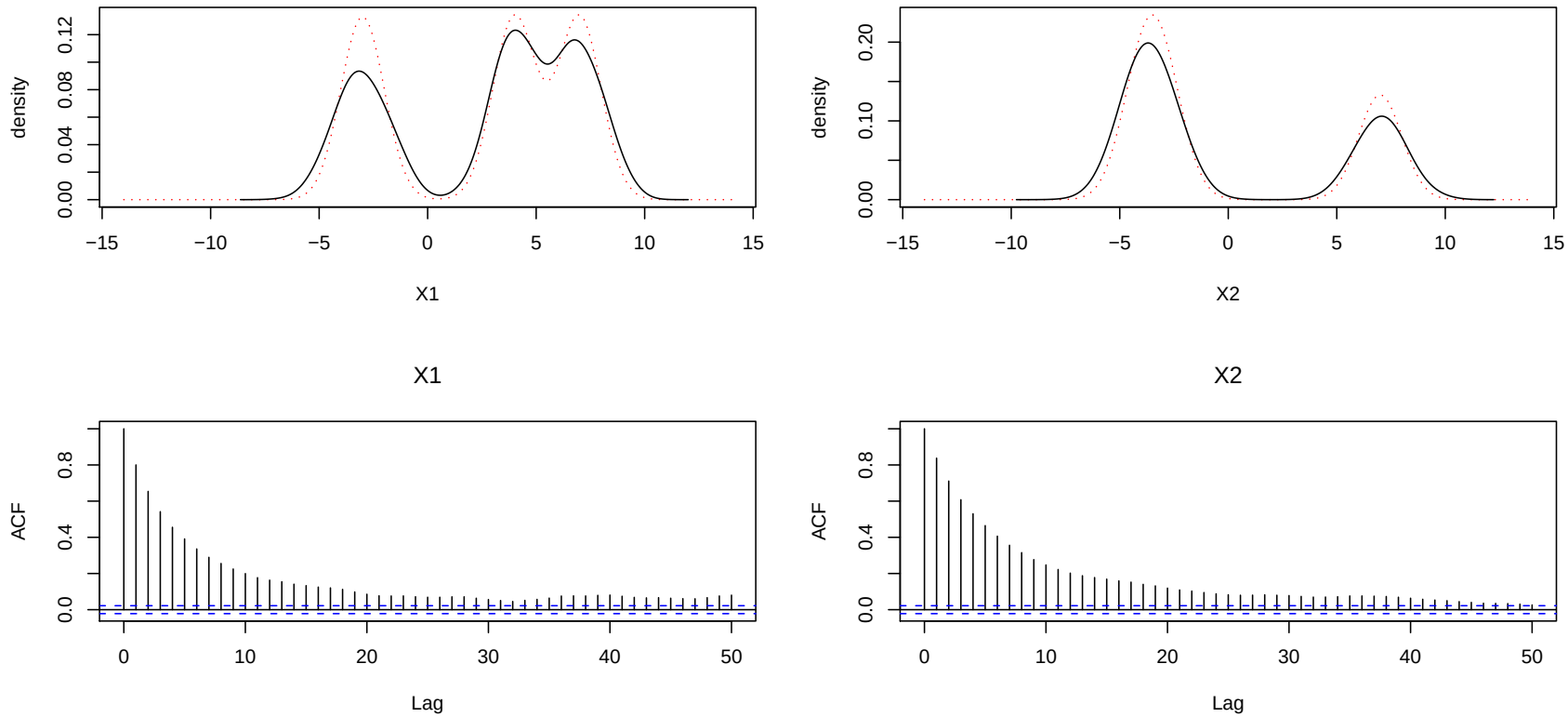
$$\pi(x) = \frac{1}{3}N(x; [-3, -3]^T, I) + \frac{1}{3}N(x; [7, 7]^T, I) + \frac{1}{3}N(x; [5, -5]^T)$$



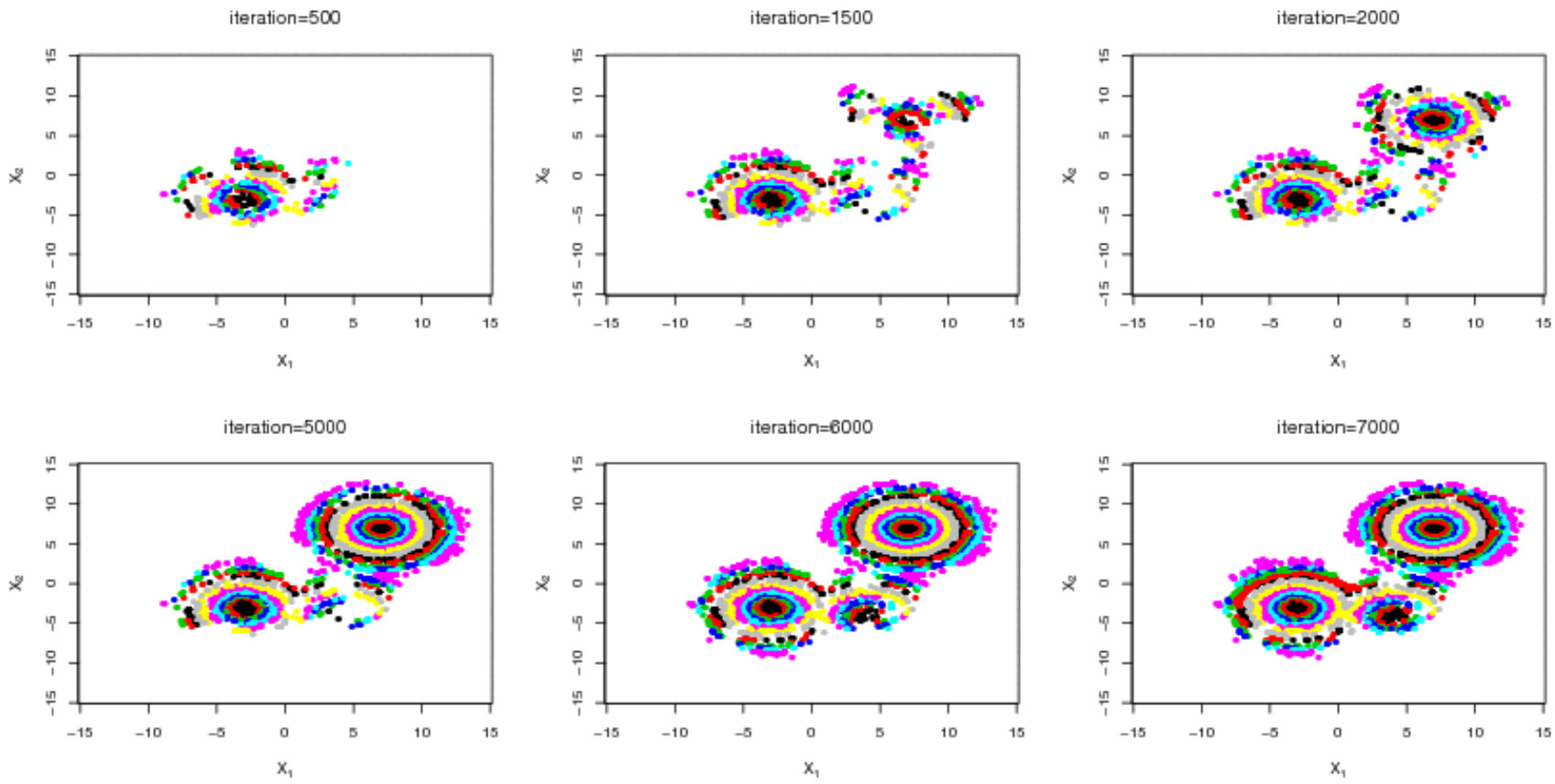



density
0.00
0.04
0.08
0.12
−15
−10
−5
0
5
10
15
X1
density
0.00
0.10
0.20
−15
−10
−5
0
5
10
15
X2
X1
ACF
0.0
0.4
0.8
0
10
20
30
40
50
Lag
X2
ACF
0.0
0.4
0.8
0
10
20
30
40
50
Lag

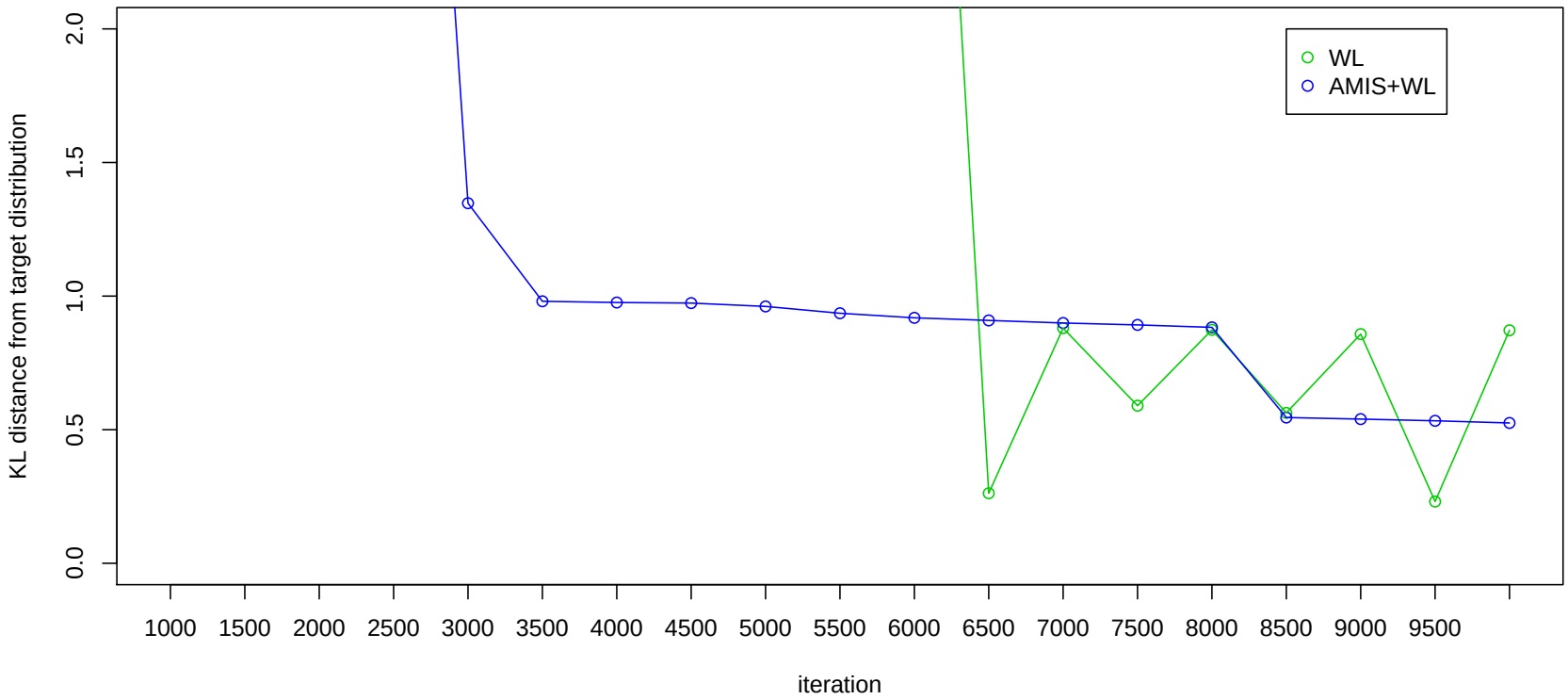
(e) AMIS


density
0.00
0.04
0.08
0.12
−15
−10
−5
0
5
10
15
X1
density
0.00
0.10
0.20
X2
X1
X2
ACF
0.0
0.4
0.8
0
10
20
30
40
50
Lag
ACF
Lag

(f) AMIS + WL


iteration=500
iteration=1500
iteration=2000
iteration=5000
iteration=6000
iteration=7000
X1
X2

(g)

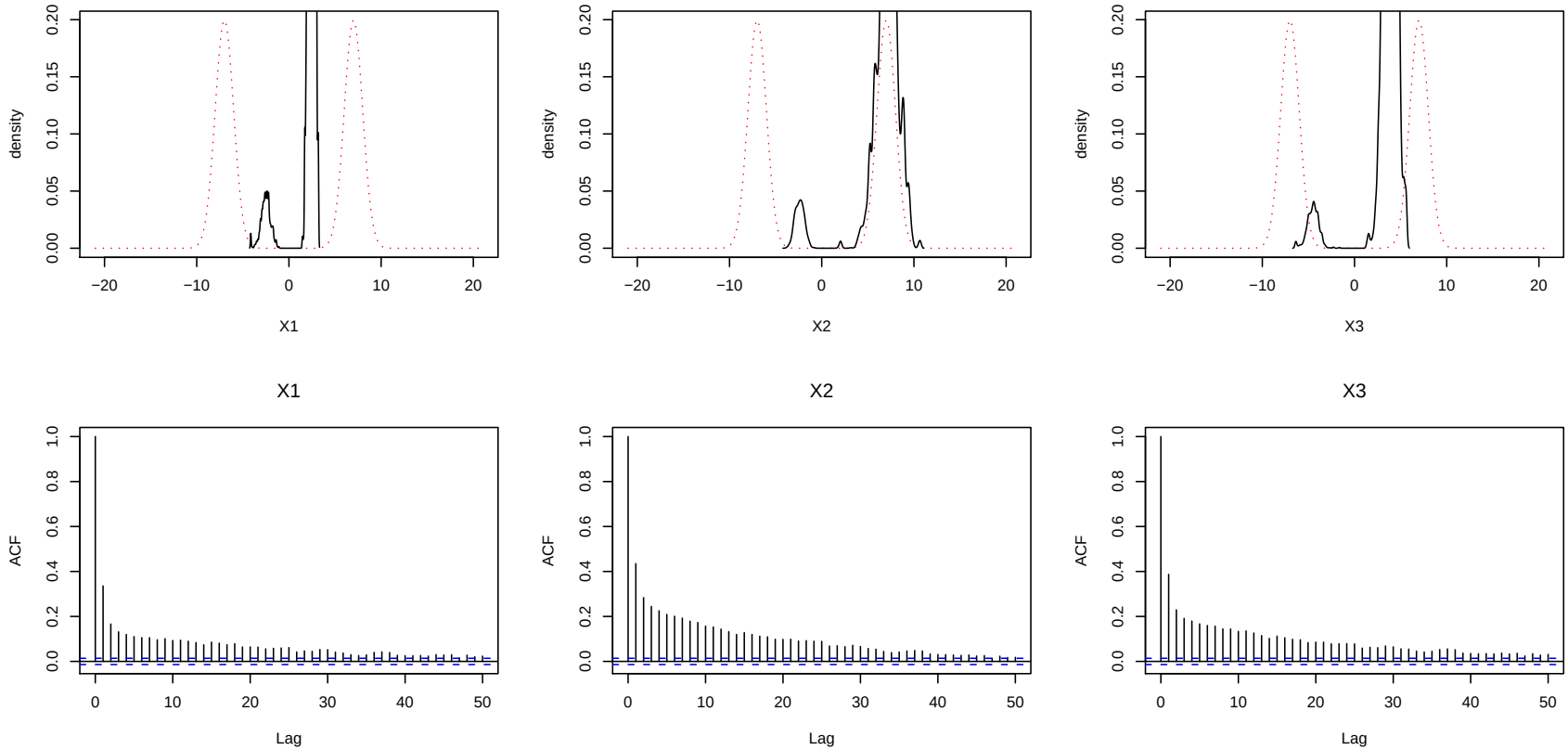

KL distance from target distribution
0.0
0.5
1.0
1.5
2.0
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
6000
6500
7000
7500
8000
8500
9000
9500
iteration
WL
AMIS+WL

(h)

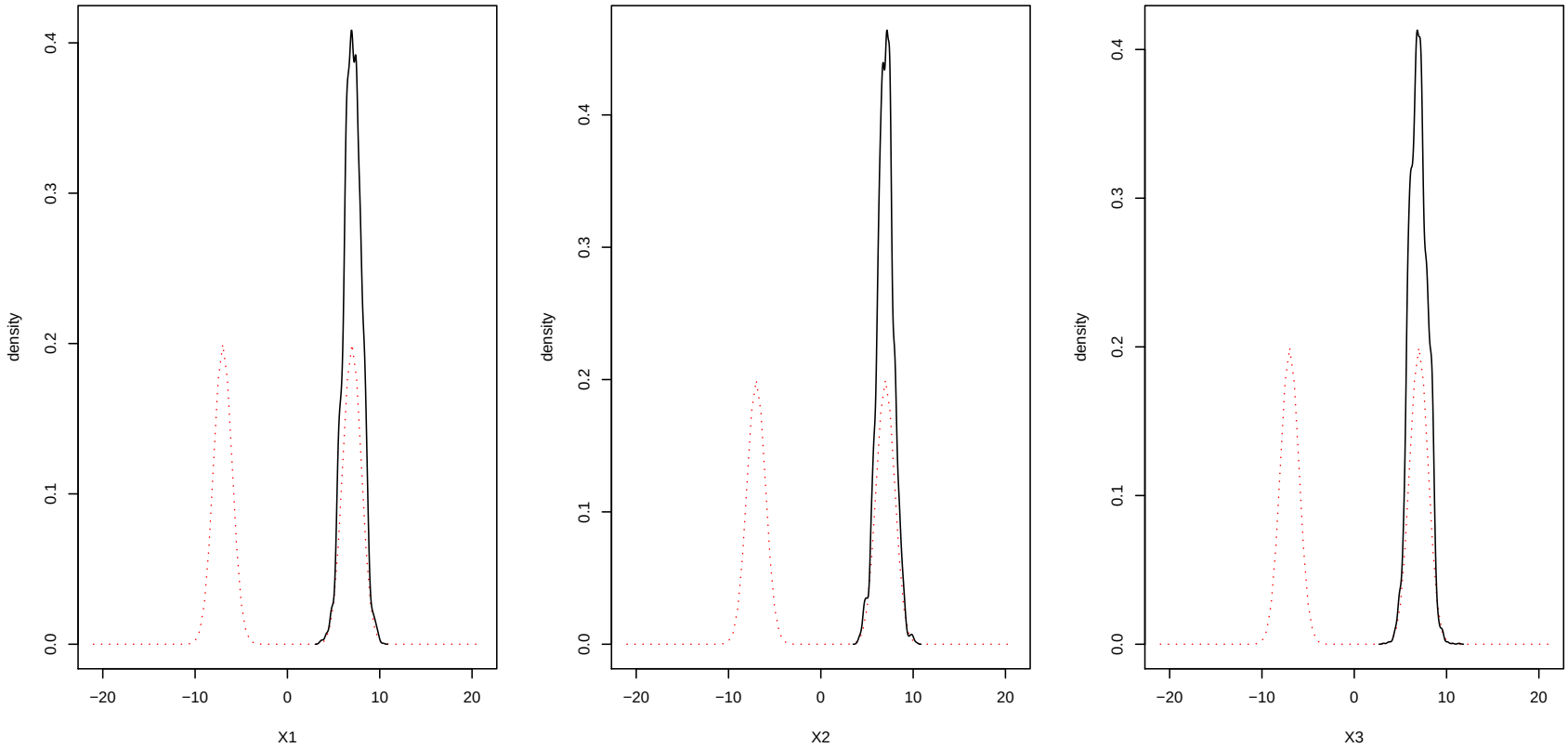
Example

Bimodal distribution in $d = 3$:

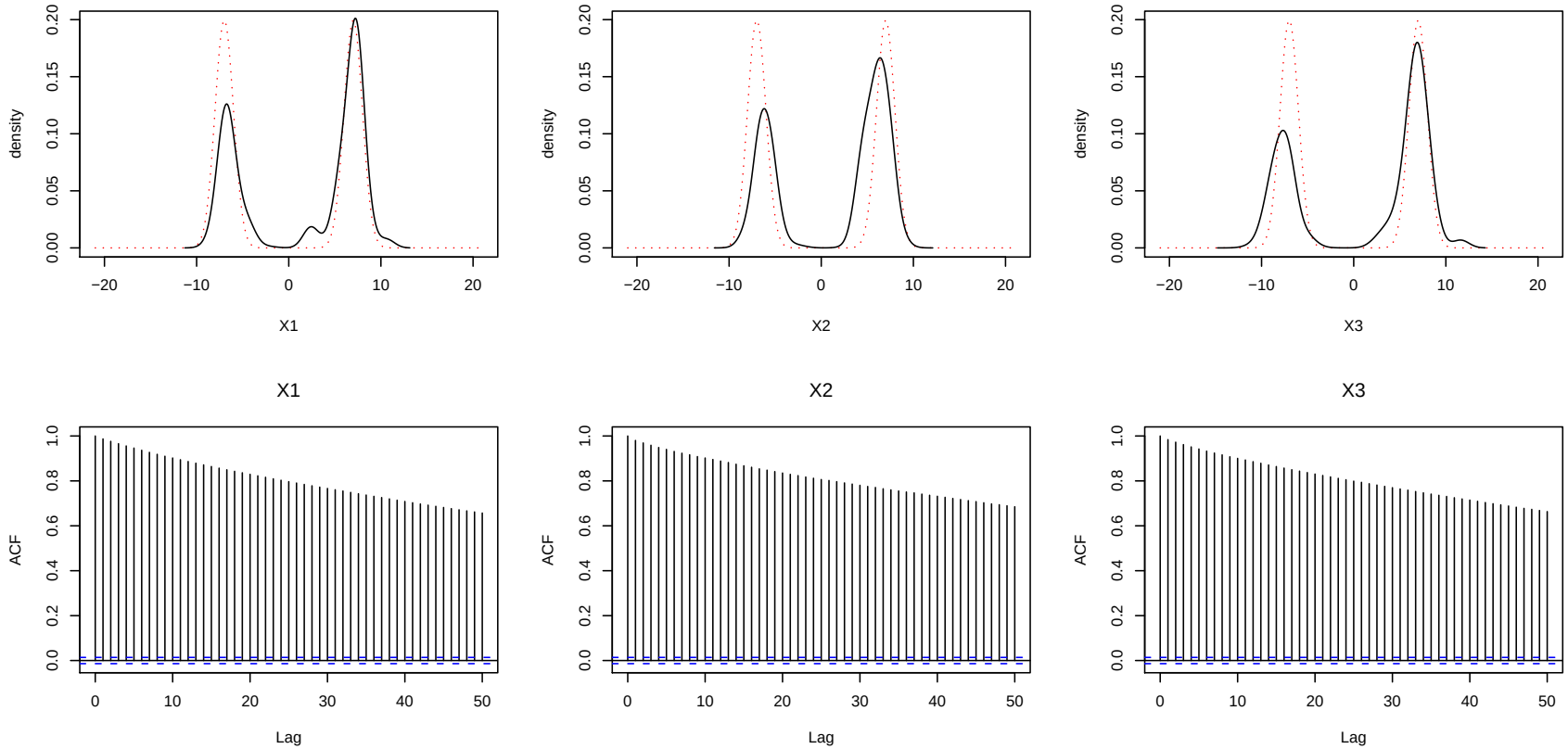
$$\pi(x) = \frac{1}{2}N(x; [-7, -7, -7]^T, I) + \frac{1}{2}N(x; [7, 7, 7]^T, I)$$



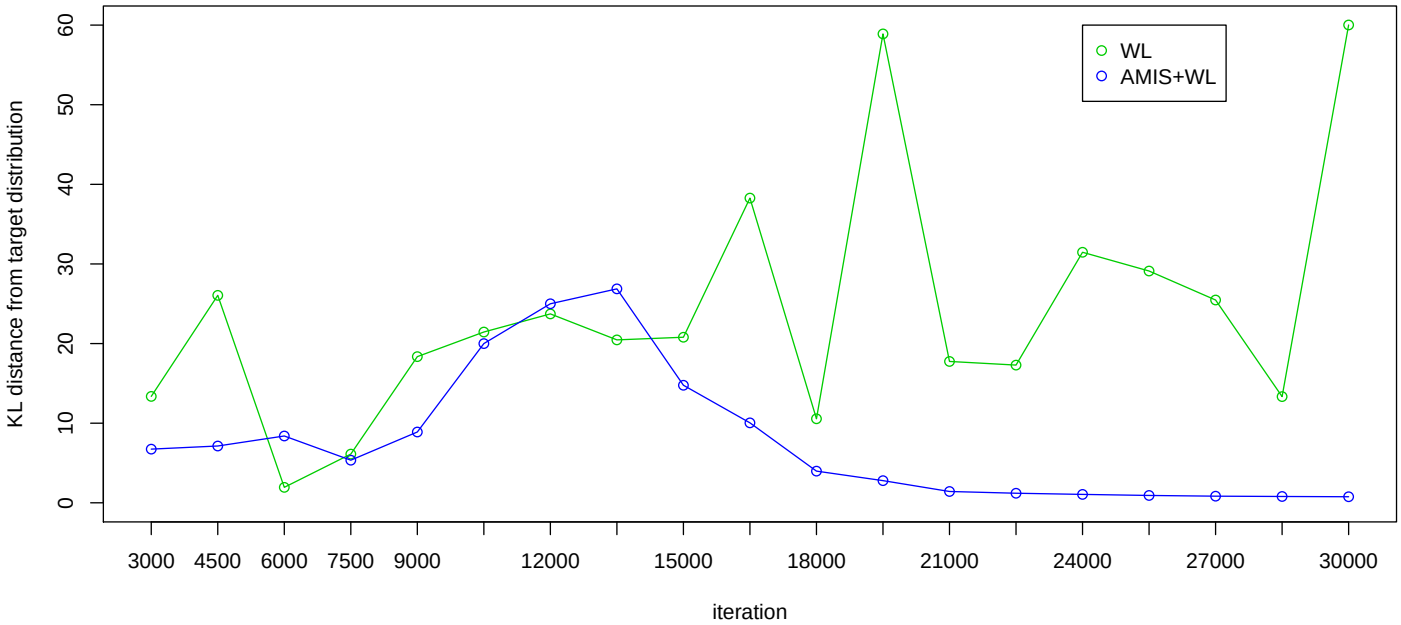
(i) AMIS


density
0.0
0.1
0.2
0.3
0.4
−20
−10
0
10
20
X1
density
0.0
0.1
0.2
0.3
0.4
−20
−10
0
10
20
X2
density
0.0
0.1
0.2
0.3
0.4
−20
−10
0
10
20
X3

(j) WL



(k) AMIS + WL



(l) KL divergence from target

Conclusions

Key ideas:

- Many ways of "adapting" an MCMC algorithm based on sample path exist; many can be shown to satisfy LLNs.
- Purpose of adaptation is to improve *rate* of convergence.
- Convergence of MC estimators involves *both* bias *and* variance.
- Different adaptation strategies can be understood as improving one or the other.
- Can obtain improved algorithms by combining strategies of different types.

Acknowledgements

- Dawn Woodard (Cornell)
- Jianyu Wang (Duke Statistics)
- Chunlin Ji (Kuang-Chi Inst)
- Kevin Wiehe (Duke postdoc)

Atchadé, Y. F. (2009).
Resampling from the past to improve on MCMC algorithms.
Far East Journal of Theoretical Probability, 27:81–99.

Holmes, C. and Held, L. (2006).
Bayesian auxiliary variable models for binary and multinomial regression.
Bayesian Analysis, 1:145–168.

Ji, C. and Schmidler, S. C. (2009).
Adaptive Markov chain Monte Carlo for Bayesian variable selection.
J. Comp. Graph. Stat., (to appear).

Kou, S. C., Zhou, Q., and Wong, W. H. (2006).
Equi-energy sampler with applications in statistical inference and statistical mechanics.
Ann. Statist., 34(4):1581–1619.

Robbins, H. and Monro, S. (1951).
A stochastic approximation method.
Ann. Math. Stat., 22:400–407.

Roberts, G. O. and Rosenthal, J. S. (2007).
Coupling and ergodicity of adaptive MCMC.
J. Appl. Prob., 44:458–475.