Echantillonage par chaîne de Markov adaptative : une approche incrémentale

Florian Maire, University College Dublin (UCD)

joint work with: Nial Friel (UCD)
Adrian Raftery (University of Washington)
Antonietta Mira (University of Lugano)

Séminaire de probabilités et mathématiques financières,
Université d'Evry Val d'Essonne,

23 Avril 2015

Outlines

Problématiques : simulation/intégration numérique

Soit π une mesure de probabilité définie sur l'espace mesurable $(\mathsf{X}, \mathcal{X})$, $\mathsf{X} \subseteq \mathbb{R}^d$

But : obtenir des échantillons de π $(X_1, \ldots, X_n) \sim \pi$

Pour (entre autres) : estimer des espérances $\mathbb{E}_\pi\{h(X)\}$ $-$ $\forall\, h \in \mathcal{L}^2(\pi)$

On s'intéresse à des mesures π un peu "compliquées" :

- dimension élevée – (*e.g* en stat. Bayesienne)
- lois multimodales / de mélange – classification
- lois à queue épaisse, etc.

mais π est supposée connue (au moins à une constante près)...

Terminologie

- π n'est pas simulable
- Hypothèse : existence d'une loi *instrumentale* Q sur $(\mathrm{X}, \mathcal{X})$ (simulable) t.q. $Q \approx \pi$?

Méthodes par chaîne de Markov

$\Rightarrow$ chaîne de Markov $\{X_k,\ k \in \mathbb{N}\}$ avec comme loi de proposition

$$\tilde{X} \sim Q$$

accepté ($i.e\ X_{k+1} = \tilde{X}$) avec proba.

$$\alpha(X_k, \tilde{X}) = 1 \wedge \frac{\pi(\tilde{X})Q(X_k)}{\pi(X_k)Q(\tilde{X})}$$

et rejetée ($i.e\ X_{k+1} = X_k$) sinon

Méthodes particulaires

$\Rightarrow$ des *particules i.i.d.* réalisations de Q

$$(X_1, \ldots, X_n) \sim_{i.i.d.} Q$$

et pondérées par des poids d'importance

$$W_k \propto W(X_k) = \frac{\pi(X_k)}{Q(X_k)}$$

Propriétés asymptotiques

Idéalement, on veut avoir des guaranties asymptotiques *e.g* CLT:

$$\sqrt{n}\left(\hat{\mu}_h(X_1,\ldots,X_n) - \mathbb{E}_\pi\left[h(X)\right]\right) \xrightarrow{P} \mathcal{N}(0, \hat{\sigma}_h^2(X_1,\ldots,X_n))$$

Méthodes par chaîne de Markov

$$\hat{\mu}_h(X_1,\ldots,X_n) = \frac{1}{n}\sum_{k=1}^{n} h(X_k)$$

de variance :

$$\hat{\sigma}_h^2(X_1,\ldots,X_n) = \mathrm{Var}_\pi[h(X)]\left(1 + 2\sum_{k=1}^{\infty}\rho_k\right)$$

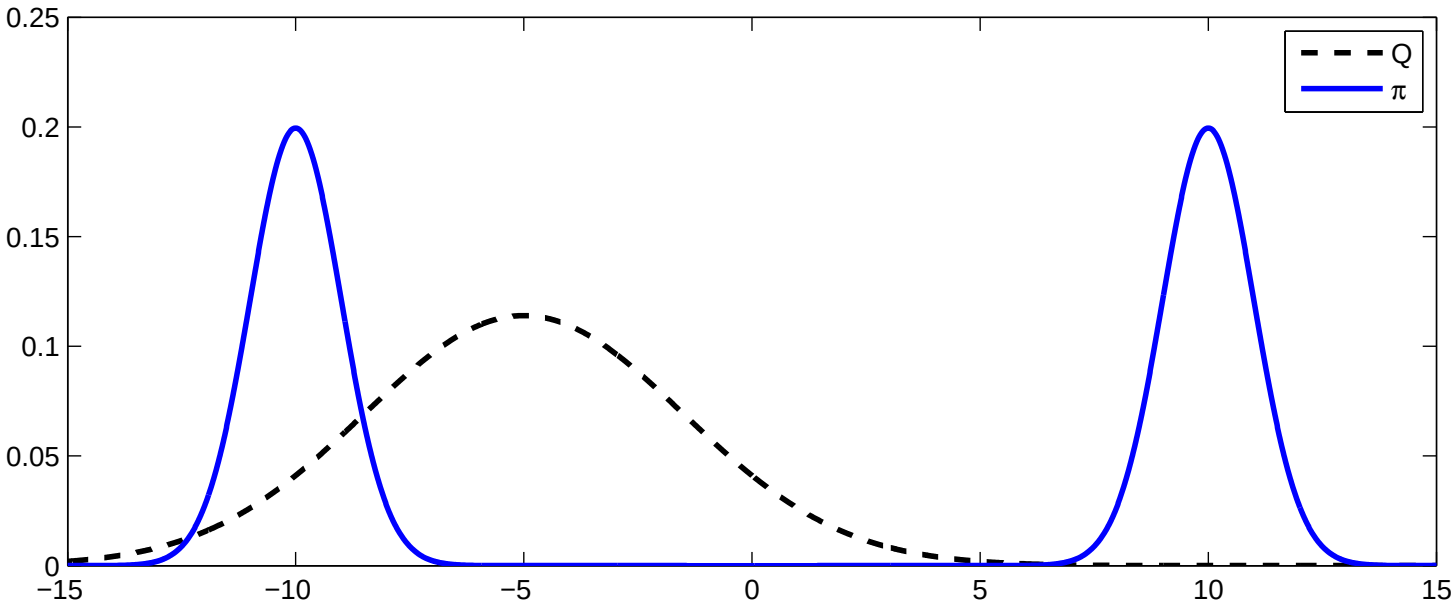
Méthodes particulaires

$$\hat{\mu}_h(X_1,\ldots,X_n) = \frac{\sum_{k=1}^{n} W(X_k)h(X_k)}{\sum_{k=1}^{n} W(X_k)}$$

de variance :

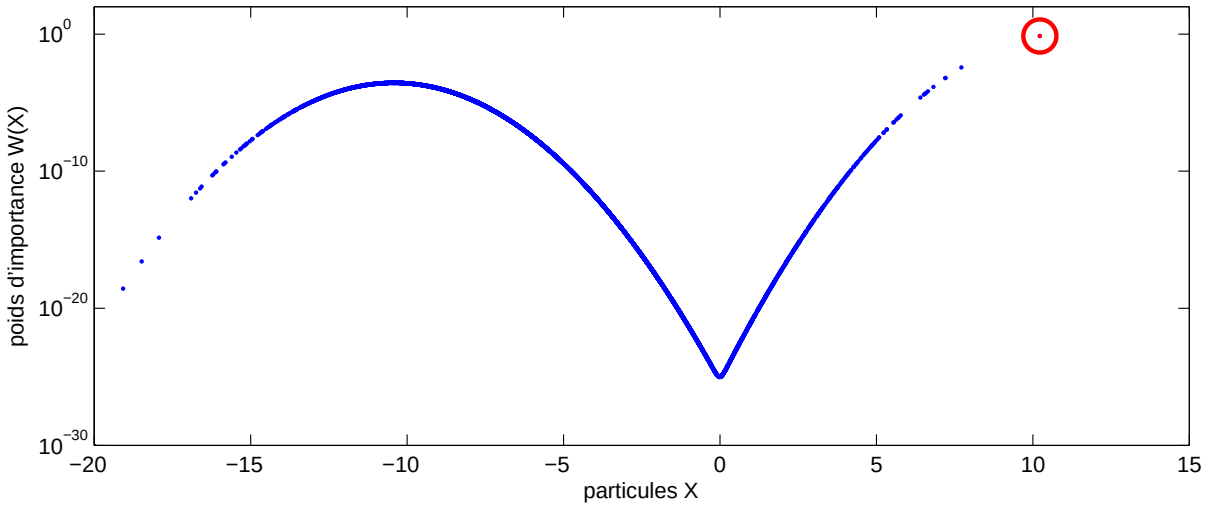
$$\hat{\sigma}_h^2(X_1,\ldots,X_n) = \mathbb{E}_\pi\left[W(X)(h(X) - \mathbb{E}_\pi\left[h(X)\right])^2\right]$$

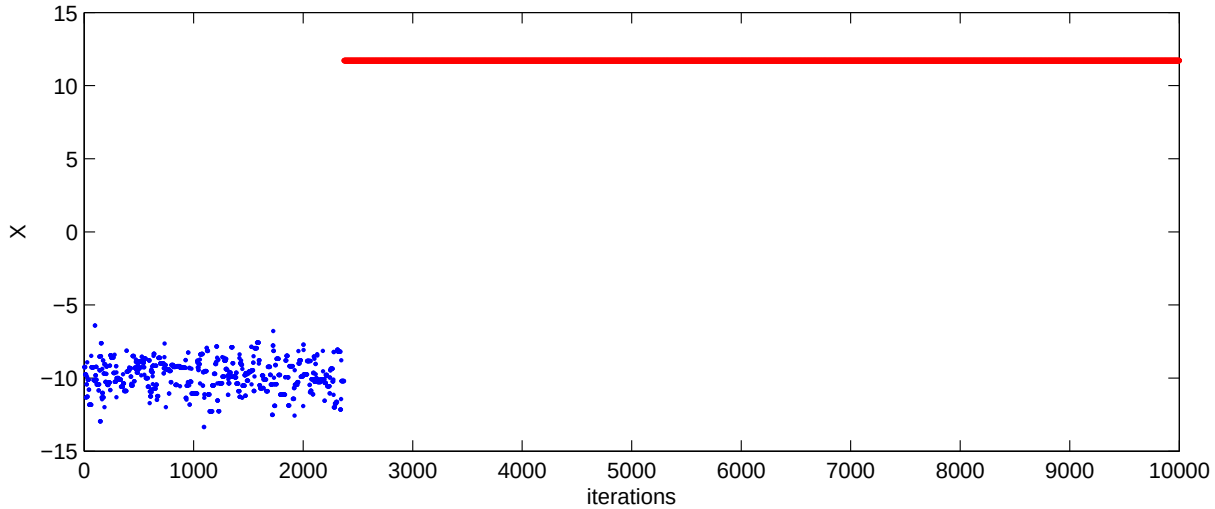
Problèmes liés à ces méthodes (illustration)



Loi cible : $\pi = (1/2)\mathcal{N}(-10, 1) + (1/2)\mathcal{N}(10, 1)$
Loi instrumentale : $Q = \mathcal{N}(-5, 4)$

Problèmes liés à ces méthodes (illustration)

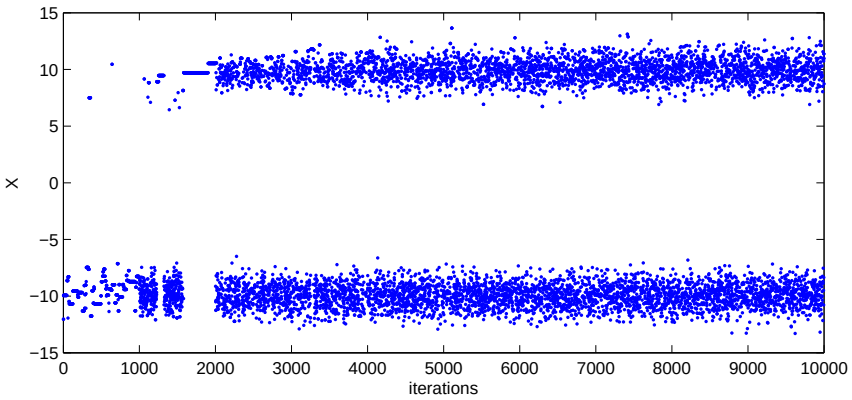


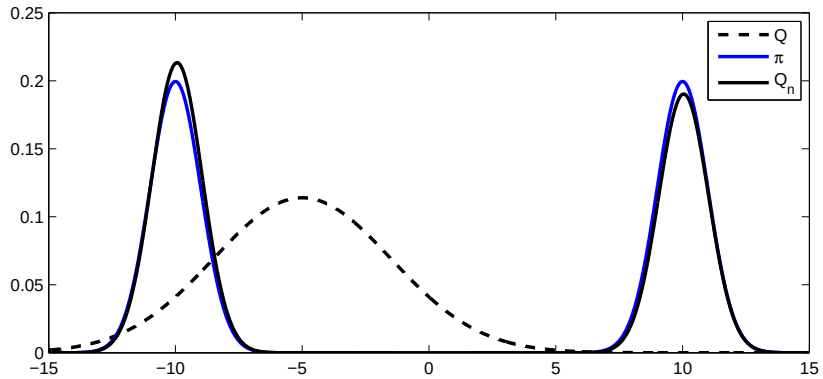


Méthodes adaptatives

Idée : Tirer profit des états précédents de la chaîne pour améliorer la loi instrumentale $Q \equiv Q_0 \to Q_1 \to \cdots \to Q_n$

Exemple : estimer un modèle de mélange pour les états passés





Questions de recherche...

- Quel mécanisme pour construire une "bonne" loi Q?
  $\Rightarrow$ pour une loi π "compliquée"
  $\Rightarrow$ avec une connaissance minimale de π (pas d'étape préliminaire!)
- Que pouvons nous-dire de la séquence de v.a. $X_1, X_2, \ldots$
  $\Rightarrow$ ergodicité? vitesse de convergence? TCL?

On veut aussi une méthode qui soit:

- facile à paramétrer
- largement appliquable
- réaliste en terme de temps CPU

Outlines

AIMM: the rationale

AIMM is an Adaptive Independent Metropolis algorithm:

- starts from an initial (naive) proposal distribution $Q_0 = \phi_0$
- develops a collection of Gaussian kernels $\{\phi_1, \phi_2, \ldots\}$ s.t. at iteration n, it features M_n components ($M_n \leq n$)

$$Q_n := \omega_n \phi_0 + (1 - \omega_n) \frac{1}{1 \vee M_n} \sum_{\ell=1}^{M_n} \phi_\ell \,, \qquad (\omega_n \in (0,1))$$

- a new component is added to the mixture if the chain unveils an area which is not well supported by the current proposal...

$\Rightarrow$ in such a case, ϕ_{M_n+1} should provide a coverage of this area
$\Rightarrow$ ϕ_{M_n+1} allows to reduce the local discrepancy between π and Q_n

AIMM: algorithmic picture

AIMM produces a Markov chain $\{X_k, k \in \mathbb{N}\}$ with transition $X_n \to X_{n+1}$

(1) Propose a new state

$$\tilde{X} \sim Q_n := \omega_n \phi_0 + (1-\omega_n) \frac{1}{1 \vee M_n} \sum_{\ell=1}^{M_n} \phi_\ell \,,$$

(2) Set $\tilde{X}_{n+1} = \tilde{X}$ with probability

$$\alpha_n(X_n, \tilde{X}) = 1 \wedge \frac{W_n(\tilde{X})}{W_n(X_n)} \,, \qquad W_n = \pi / Q_n \,,$$

(3) Improve the proposal if

$$\left\{ W_n(\tilde{X}) > W^* \right\} \,,$$

increment the mixture with the kernel $\phi_{M_n+1} = \mathcal{N}(\tilde{X}, \widehat{\Sigma}(\tilde{X}; \mathfrak{N}(\tilde{X})))$ and set $M_{n+1} = M_n + 1$.

AIMM: illustration

$$\pi = .2\mathcal{N}(-3; .5) + .1\mathcal{N}(0; .5) + .3\mathcal{N}(2, 2) + .1\exp(10) + .3\mathcal{N}(25, 1)$$

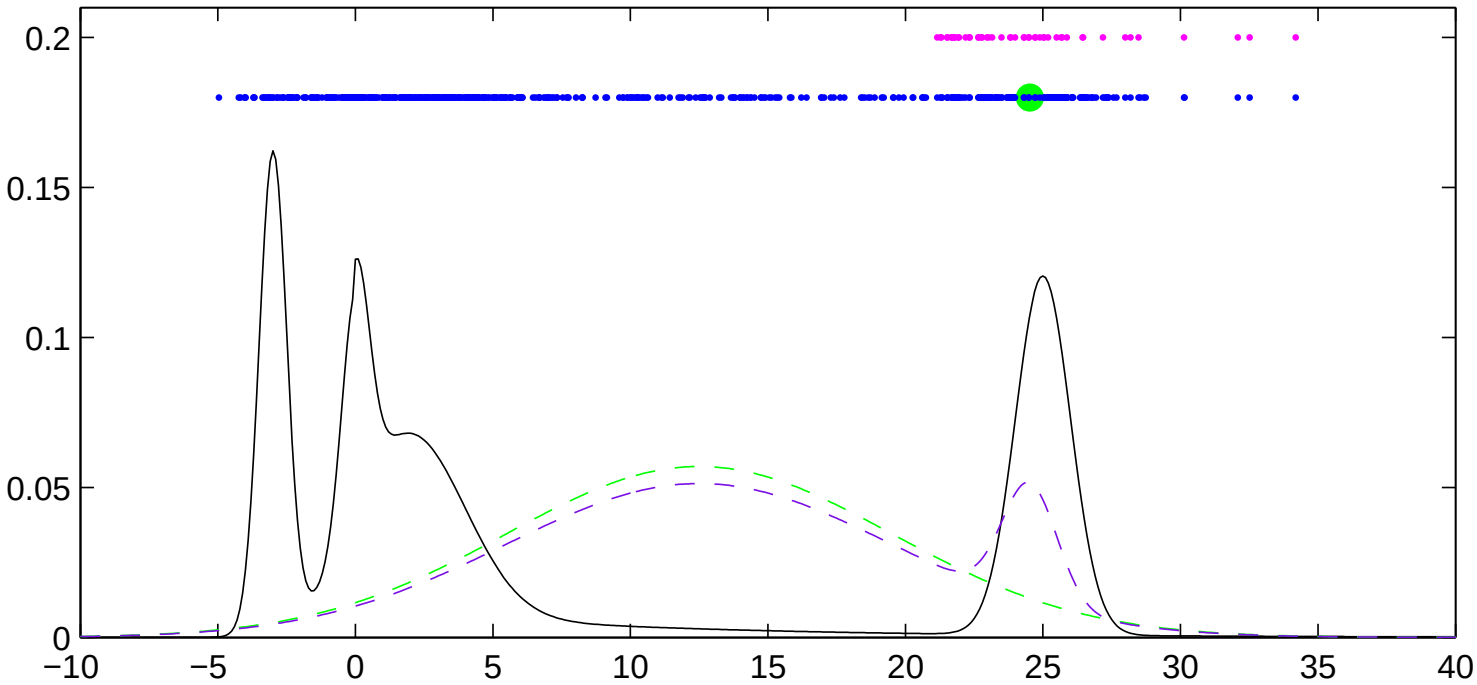


Figure: black: π, green: ϕ_0, magenta: Q_{M_1} (first update)

AIMM: illustration

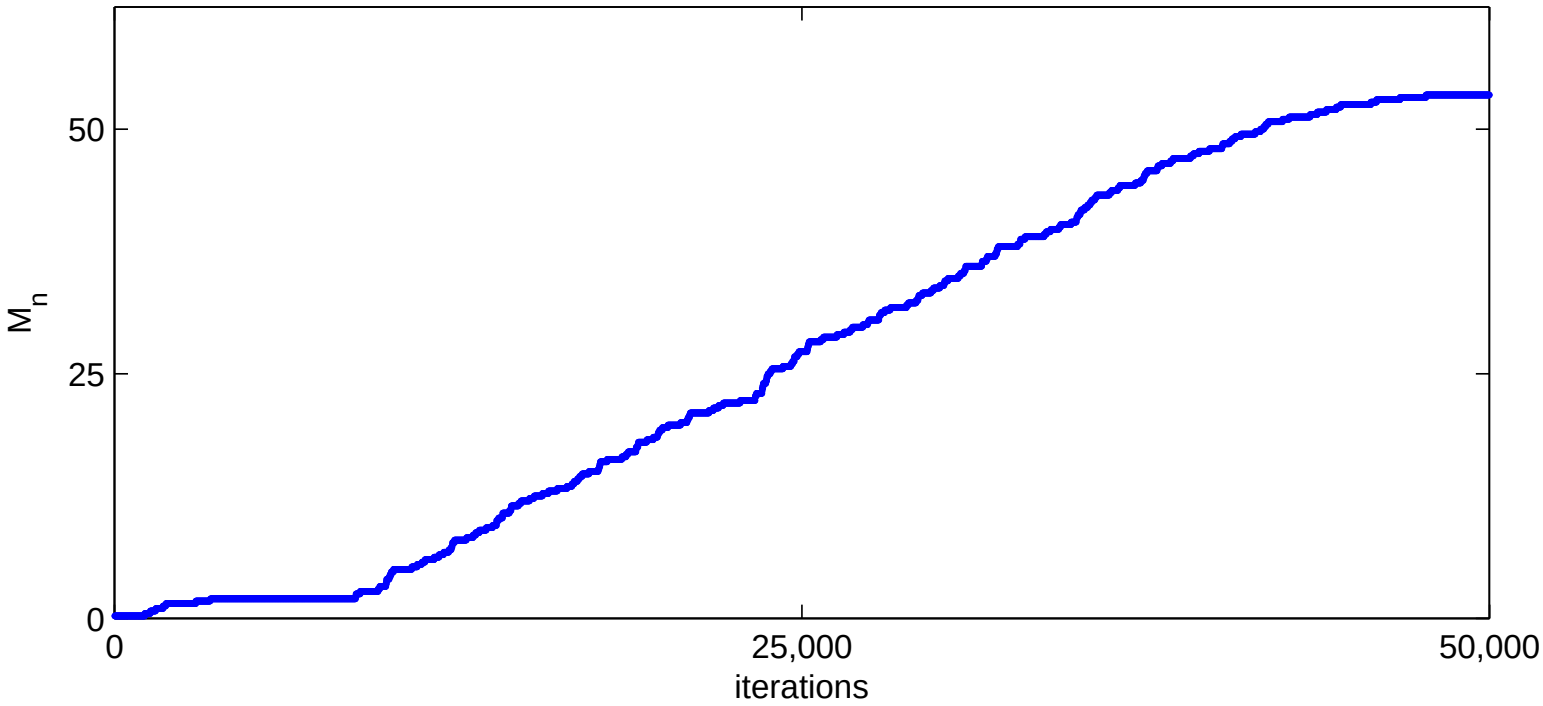


Figure: Evolution of the number of kernels throughout the algorithm

AIMM: illustration

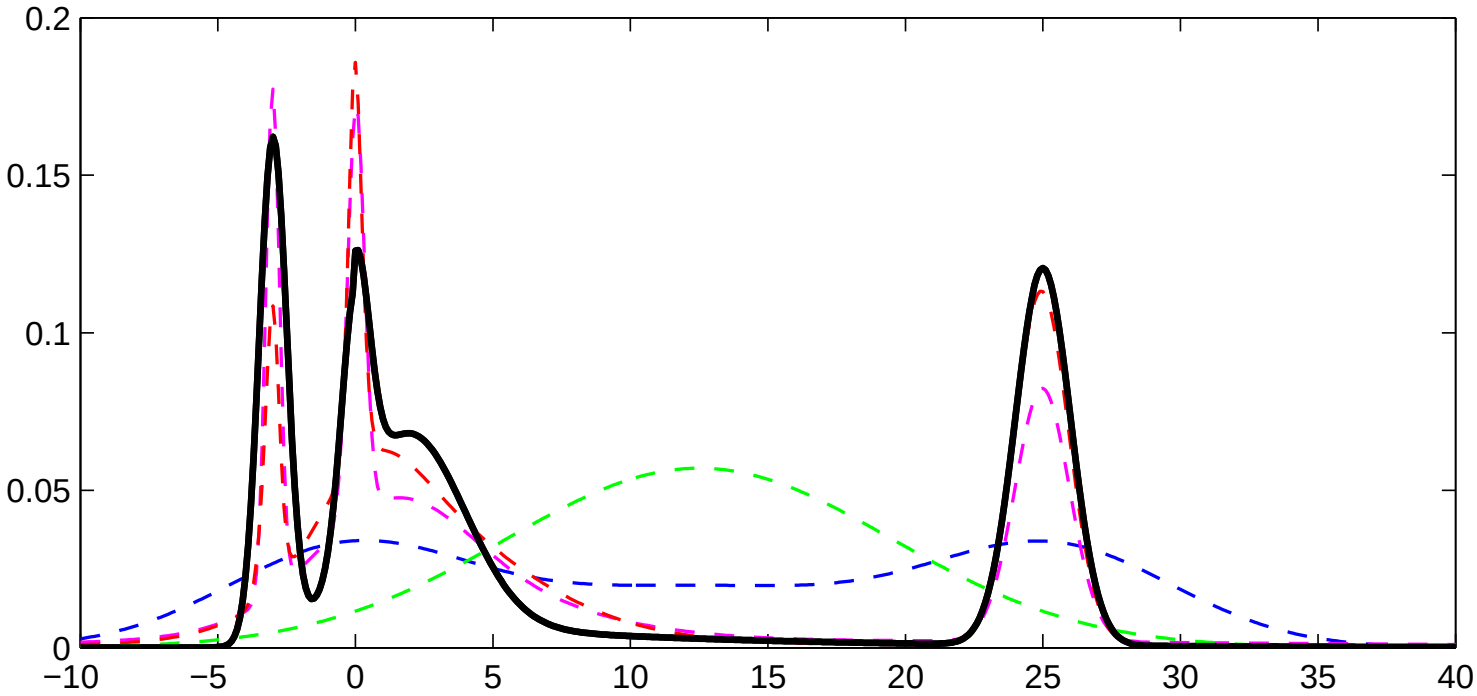


Figure: Evolution of the proposal distribution Q_n, for $M_n = 0$ (green), $M_n = 5$ (blue), $M_n = 25$ (magenta) and $M_n = 50$ (red)

AIMM - observation and improvements

- are 52 kernels really necessary to cover this 1 dimensional distribution?
- is the resulting incremental mixture really satisfactory?

Improvements:

(i) forget progressively the oldest kernels

(ii) and replace them (stochastically) by new ones

(iii) attach a weight to each component:

$$Q_n := \omega_n \phi_0 + (1-\omega_n) \frac{1}{M_n} \sum_{\ell=1}^{M_n} \phi_\ell$$

becomes

$$Q_n := \omega_n \phi_0 + (1-\omega_n) \sum_{\ell=1}^{M_n} \alpha_\ell \phi_\ell / \sum_{\ell=1}^{M_n} \alpha_\ell, \quad \alpha_\ell \propto \pi(\mu_\ell)\,.$$

AIMM - observation and improvements

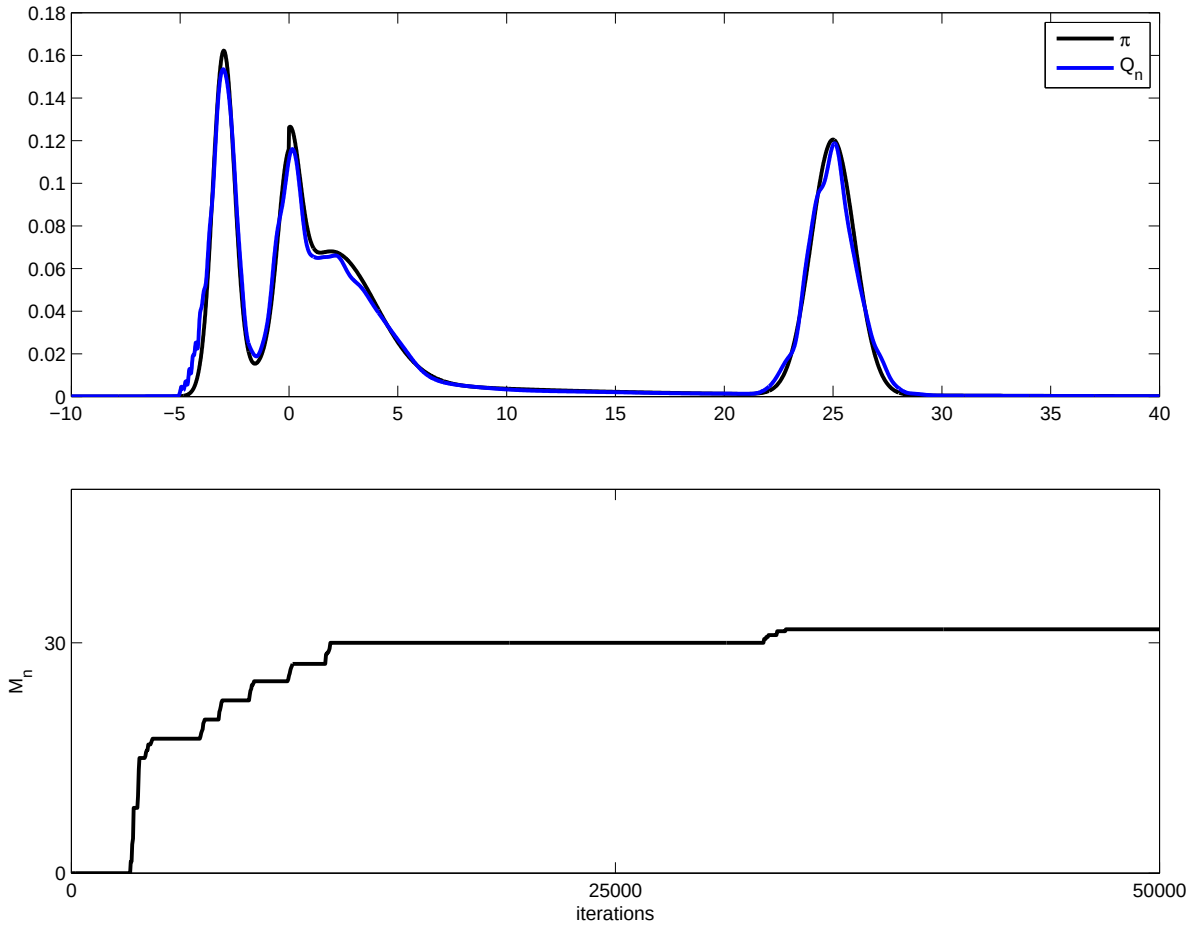


Figure: Target and final incremental Mixture (top) – number of kernels throughout (bottom)

AIMM - observations and improvements

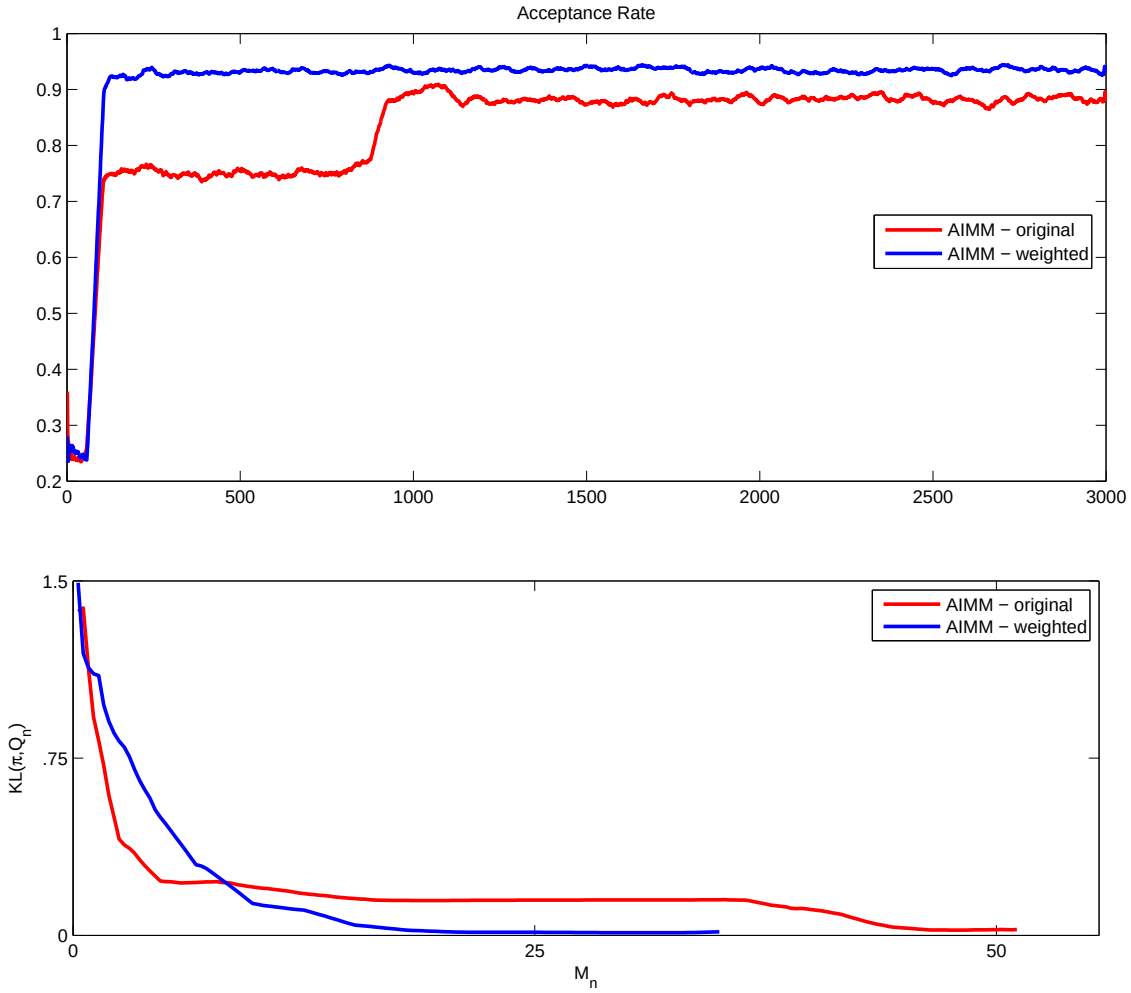


Figure: Acceptance rate (top) – KL divergence $\mathrm{KL}(\pi, Q_n)$ (bottom)

AIMM: what's the novelty?

Adaptive (independent) MCMC: many papers and good ideas...

- Most frequent approach:
  - the proposal distribution Q_k belongs to a parameterized family:

$$Q_k \equiv Q_{\theta_k}\,, \qquad (\theta_k \in \Theta \subseteq \mathbb{R}^m)$$

  - samples of the chain are recursively used to build a sequence of parameter $\{\hat{\theta}_k,\, k \in \mathbb{N}\}$ *optimizing* a criterion (acceptance rate, moment matching with π, Kullback-Leibler minimization w.r.t. π...)
  - connected with the EM literature (optimization approach, losing the MCMC spirit...)

issue $\Rightarrow$ needs some *apriori* knowledge of π to chose a reasonable Q_θ

$\Rightarrow$ constrain (sometimes dramatically) the adaption

AIMM: reviving the importance weight

Consider an Adaptive Independent M–H, at iteration n, state X_n,
Let $\tilde{X}_1 \sim Q$ and $\tilde{X}_2 \sim Q$ be two proposed states s.t.

$$W_n(\tilde{X}_1) > W_n(X_n)\,, \qquad W_n(\tilde{X}_2) \gg W_n(X_n)\,.$$

Both of them would be accepted w.p. 1.

- Thus both of them would have the same "weight" in the final MCMC estimate ($\tilde{X}_1$ and $\tilde{X}_2$ are processed in the same way)
- Yet, they do not convey the same information...
- In particular, $\{W_n(\tilde{X}_2) \gg W_n(X_n)\}$ is lost because of the M–H threshold

AIMM aims at retrieving this information by adding a component in the case of $\tilde{X}_2$ and not $\tilde{X}_1$

Outlines

About the convergence of AIMM

- Instead of Roberts and Rosenthal assumptions,
  - (i) diminishing adaption: transition kernels tend to become closer and closer in probability
  - (ii) containment: each transition kernel is a finite time step away from an ϵ-ball centered on the target (relaxation of the simultaneous ergodicity)
- Holden's proof of geometric convergence for adaptive chains with independent proposals, seems more straightforward in our case,
  - (i) main assumption: a strong Dœblin condition – it exists a function $\gamma_n : \mathsf{X}^n \to [1, +\infty($ such that for all $x \in \mathsf{X}^2$

$$\pi(x) \leq \gamma_n(\tilde{y}^n) Q_n(x)$$

where $\tilde{y}^n \in \mathsf{Y}^n$ is a history dependent vector.

Strong Doeblin assumption

Given the proposal, it is hard to satisfy a Doeblin condition:

- if π is a posterior distribution, we have:

$$\frac{\pi(X)}{Q_n(X)} \leq \frac{L(Y_{\text{obs}} \,|\, \hat{X})}{\omega_n \int L(Y_{\text{obs}} \,|\, x)\phi_0(\mathrm{d}x)}, \qquad \hat{X} = \arg\max_x L(Y_{\text{obs}} \,|\, x)$$

- we are rather interested in having something like

$$\exists \gamma_n(\tilde{y}^n), \text{ s.t. } \quad \pi\Big\{\pi(x) \leq \gamma_n(\tilde{y}^n) Q_n(x)\Big\} > 1 - \epsilon$$

- which is related to the KL divergence being arbitrarily small

$$\mathrm{KL}_n(\pi, Q_n) = \int \pi(\mathrm{d}X) \log \frac{\pi(X)}{Q_n(X)}$$

$\mathrm{KL}(\pi, Q_n) > \mathrm{KL}(\pi, Q_{n+1})$?

We consider here only the KL div. between π and incremental part of the mixture:

$$\mathrm{KL}_n = \int_{\mathsf{X}} \pi(\mathrm{d}x) \log \frac{\pi(x)}{\sum_{\ell=1}^{n} \alpha_\ell \phi_\ell(x)/\boldsymbol{\alpha}_n}, \qquad \boldsymbol{\alpha}_n = \sum_{\ell=1}^{n} \alpha_\ell \,.$$

The idea is to split the state space between the region E_λ where adding the component ϕ_{n+1} yields a local KL reduction between Q_n and Q_{n+1}:

$$\text{for} \qquad \lambda > 0\,, \qquad \mathsf{E}_\lambda := \left\{ x \in \mathsf{X}, \quad (1+\lambda) \frac{\pi(x)}{Q_{n+1}(x)} < \frac{\pi(x)}{Q_n(x)} \right\}$$

and $\bar{\mathsf{E}}_\lambda$. The existence of a region E_λ (as narrow as it can be), is related to regularity assumption on π. If π admits a density w.r.t. Lebesgue this is a mild assumption.

$\mathrm{KL}(\pi, Q_n) > \mathrm{KL}(\pi, Q_{n+1})$?

- Denoting for all $A \in \mathcal{X}$, $\mathrm{KL}_n(A) = \int_A \pi(\mathrm{d}x) \log(\pi(x)/Q_n(x))$, we have:

$$KL_{n+1}(\mathrm{E}_\lambda) < \log(1+\lambda)\pi(\mathrm{E}_\lambda) + KL_n(\mathrm{E}_\lambda)\,. \tag{1}$$

- Now for the rest of the state space: we can bound the "possibly" increase in KL which results from adding a kernel and therefore diminishes the weight of any other component ϕ_ℓ: $\alpha_\ell/\boldsymbol{\alpha}_n \to \alpha_\ell/\boldsymbol{\alpha}_{n+1}$. This gives:

$$\mathrm{KL}_{n+1}(\bar{\mathrm{E}}_\lambda) - \mathrm{KL}_n(\bar{\mathrm{E}}_\lambda) \leq \log\left(1 + \frac{\alpha_{n+1}}{\boldsymbol{\alpha}_n}\right)\pi(\bar{\mathrm{E}}_\lambda)\,. \tag{2}$$

$\mathrm{KL}(\pi, Q_n) > \mathrm{KL}(\pi, Q_{n+1})$?

- At a global scale, (1) and (2) combines as:

$$\mathrm{KL}_{n+1} \leq \mathrm{KL}_n + \underbrace{\log\left(1 + \frac{\alpha_{n+1}}{\boldsymbol{\alpha}_n}\right)\pi(\bar{\mathrm{E}}_\lambda) - \log(1+\lambda)\pi(\mathrm{E}_\lambda)}_{\Delta_n}$$

- clearly if $\Delta_n < 0$ this yields a reduction in KL after adding ϕ_{n+1}

$$\{\Delta_n < 0\} \Leftrightarrow \left\{\alpha_{n+1} \leq \boldsymbol{\alpha}_n \underbrace{\left\{(1+\lambda)\exp\left(\frac{\pi(\mathrm{E}_\lambda)}{1-\pi(\mathrm{E}_\lambda)}\right) - 1\right\}}_{>0}\right\} \quad (3)$$

- the right hand side being positive makes $\{\Delta_n < 0\}$ an event with non zero probability as, $\alpha_{n+1} > 0$.

$\mathrm{KL}(\pi, Q_n) > \mathrm{KL}(\pi, Q_{n+1})$?

$$\{\Delta_n < 0\} \Leftrightarrow \left\{ \alpha_{n+1} \leq \boldsymbol{\alpha_n} \underbrace{\left\{ (1+\lambda) \exp\left(\frac{\pi(\mathrm{E}_\lambda)}{1-\pi(\mathrm{E}_\lambda)} \right) - 1 \right\}}_{>0} \right\}$$

- Intuitively, this means that if ϕ_{n+1} is located in a low region area (eg $\pi(X_n) \ll 1$ but $Q_n(X_n) \approx 0$), the right hand side will be probably low ($\pi(\mathrm{E}_\lambda) \ll \pi(\bar{\mathrm{E}}_\lambda)$) but so will be α_{n+1} as a fct of $\pi(X_n)$... so (3) might hold with a reasonable probability.
- This would not be as likely, choosing $\alpha_{n+1} \equiv 1$ as in the original AIMM (see Slide 11 for an illustration of this).
- Conversely, if ϕ_{n+1} is located in a high region area, α_{n+1} will be larger but so will be the right hand side.

Outlines

Adaptive Metropolis, AM (Haario et al.,2001)

An adaptive Random walk Metropolis algorithm with proposal

$$Q_n(X_n, \cdot) = \phi_0(\cdot)\, \mathbb{1}_{n \leq N_0} + \psi_n(X_n, \cdot)\, \mathbb{1}_{n > N_0}$$

- ϕ_0 is the "naive" proposal or prior
- ψ_n is a Gaussian with mean X_n and covariance matrix

$$\Sigma_n = s_d\, \Gamma_n + s_d\, \epsilon I_d$$

where $\Gamma_n = \mathrm{cov}(X_1, \ldots, X_{n-1})$, $s_d = (2.4)^2/d$ (d is the dimension of the state) and $\epsilon \ll d$ a constant parameter (allowing to have Σ_n positive definite)

Adaptive Gaussian Mixture M–H, AGM (Kohn and Giordani, 2010) and (Luengo and Martino, 2013)

Considere an independent proposal kernel of the form

$$Q = \sum_{k=1}^{M} \omega_k \phi_k, \qquad \phi_k \equiv \phi_{\theta_k}, \qquad \left(\omega_k \in (0,1), \sum_{k=1}^{M} \omega_k = 1 \right)$$

where ϕ_k is a Gaussian kernel with mean and covariance $\theta_k = (\mu_k, \Sigma_k)$.

- the samples of the chain helps to construct a sequence of parameters

$$\left\{ \hat{\Xi}_n = \left((\hat{\omega}_{1,n} \hat{\theta}_{1,n}), \ldots, (\hat{\omega}_{M,n}, \hat{\theta}_{M,n}) \right), \ n \in \mathbb{N} \right\}$$

- the proposal kernel uses of the current $\hat{\Xi}_n$ ($Q_n \equiv Q_{\hat{\Xi}_n}$)
- a new state $X_{n+1} \sim K_n(X_n, \cdot) \equiv K_{\hat{\Xi}_n}(X_n, \cdot)$ is used to define the next parameter $\hat{\Xi}_{n+1} = F(X_{n+1}, \hat{\Xi}_n)$ (F being a deterministic mapping)

Challenging target: banana–shaped distribution

We consider $\mathsf{X} = \mathbb{R}^d$

$$\pi_d(\mathrm{d}x) = \Phi_d(f_b(x), \mu, \Sigma)\mathrm{d}x\,,$$

where

- $x \to \Phi_d(x, \mu, \Sigma)$ is the d-dimensional Gaussian density function with mean μ and covariance matrix Σ
- $f_b : \mathbb{R}^d \to \mathbb{R}^d$ is the mapping defined by

$$f_b : \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \to \begin{pmatrix} x_1 \\ x_2 + bx_1^2 - 100b \\ \vdots \\ x_d \end{pmatrix}$$

We have used $b = 0.1$, $\mu = [0, 0, \cdots, 0]$ and $\Sigma = \mathrm{diag}([100, 1, \cdots, 1])$

Example in two dimension

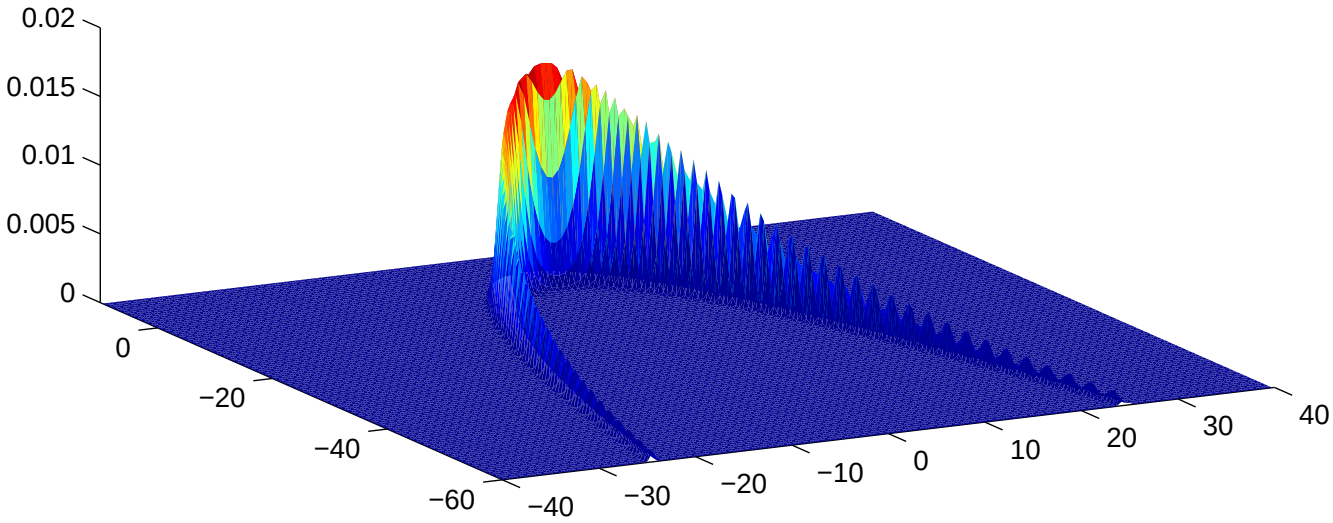


Figure: representation of π_2

Interests: benchmark "challenging" distribution, long tail, narrow support...

Comparison AIMM-AM-AGM

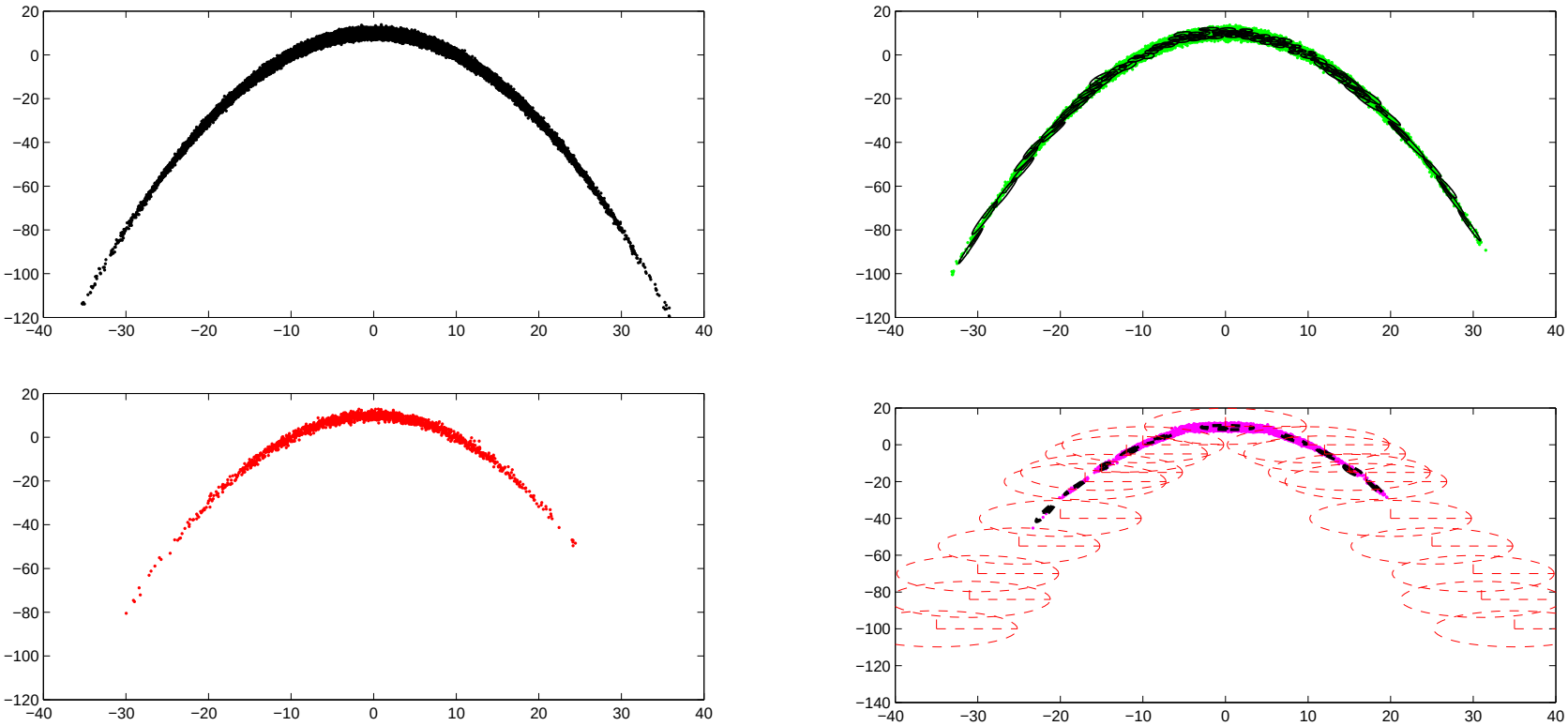
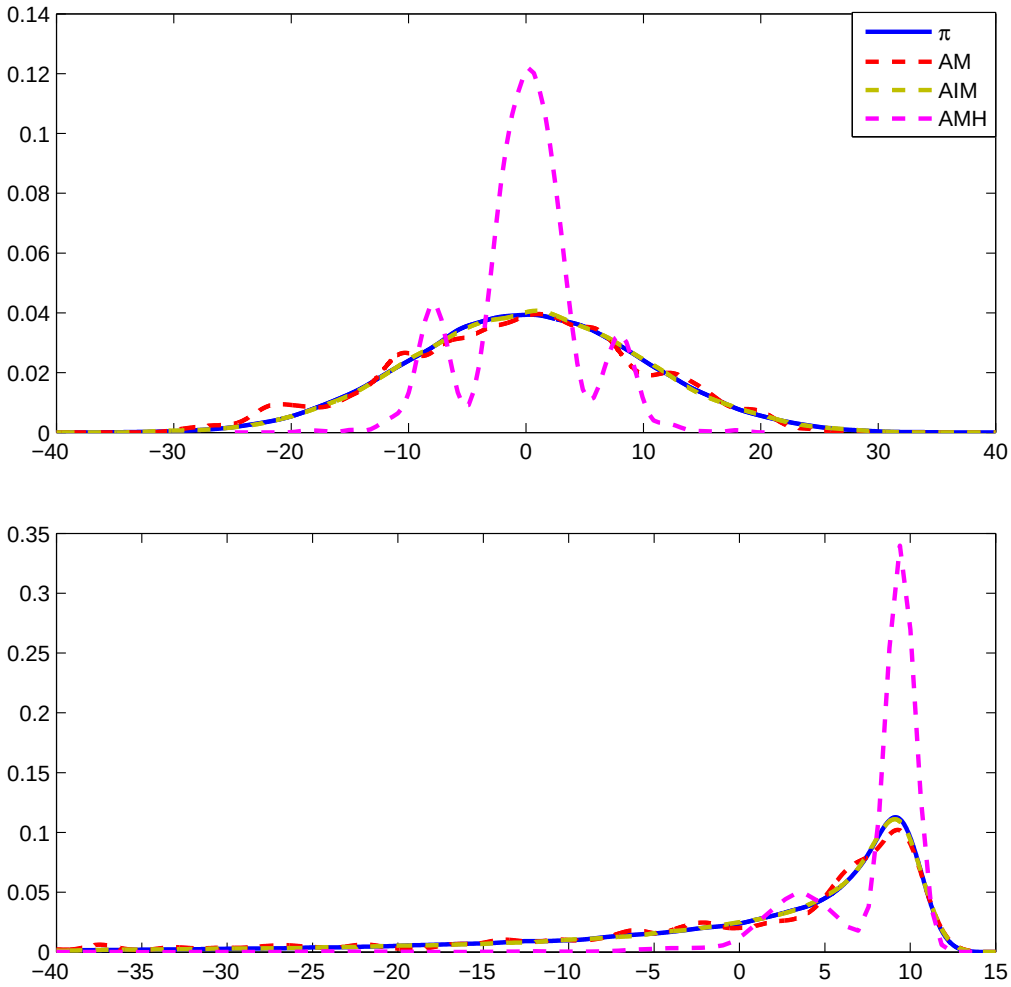


Figure: draws from π (top left), from AIMM (top right), from AM (bottom left), from AGM (bottom right)

Comparison: Kernel density approximation



Why AGM is failing?

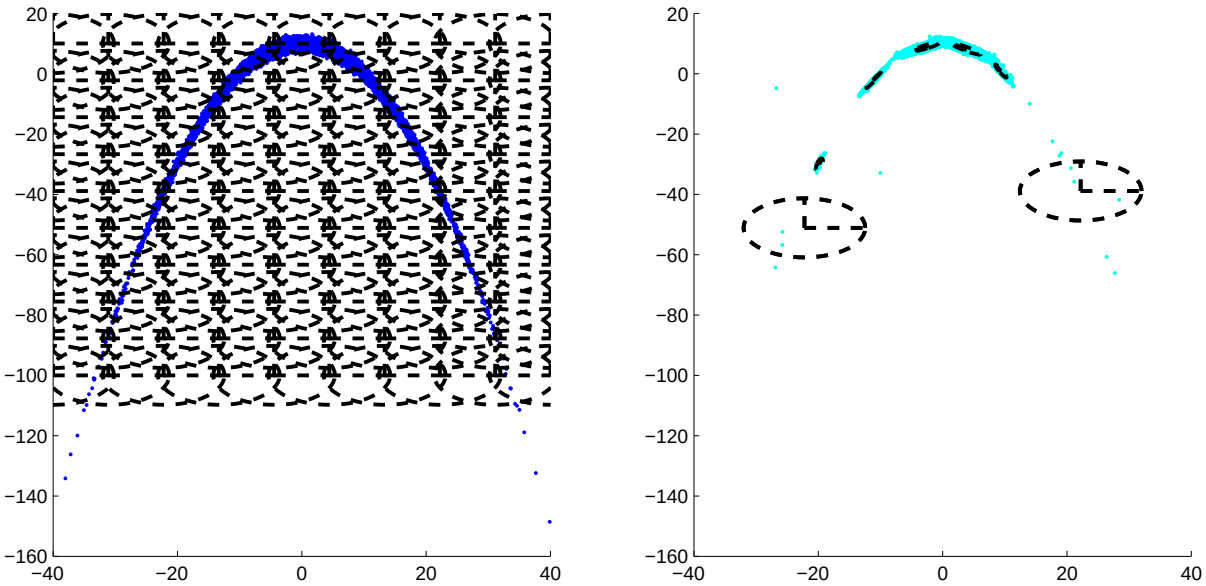


Figure: Setup with $M = 100$ kernels and $L = 50,000$ it. – left: samples from the target and initial set of Gaussian kernels – right: samples from the Markov chain in blue and final set of Gaussian kernels (display only kernels ϕ_k s.t. $\hat{\omega}_{k,L} > .001$)

$\Rightarrow$ Because it does not account for discrepancy between π and Q_n

How about AM?

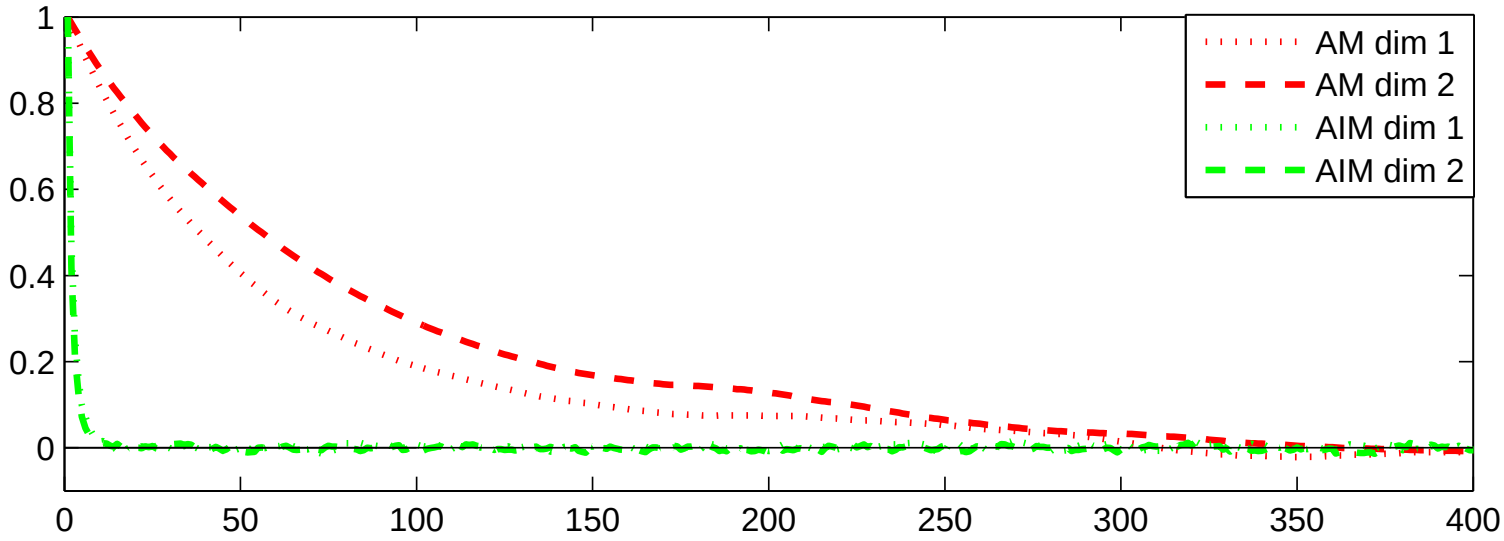


Figure: Autocorrelation for AM(red) and AIMM (green)

⇒ AM suffers very bad mixing (AIMM benefits the independent sampler)

AIMM – convergence of the proposal

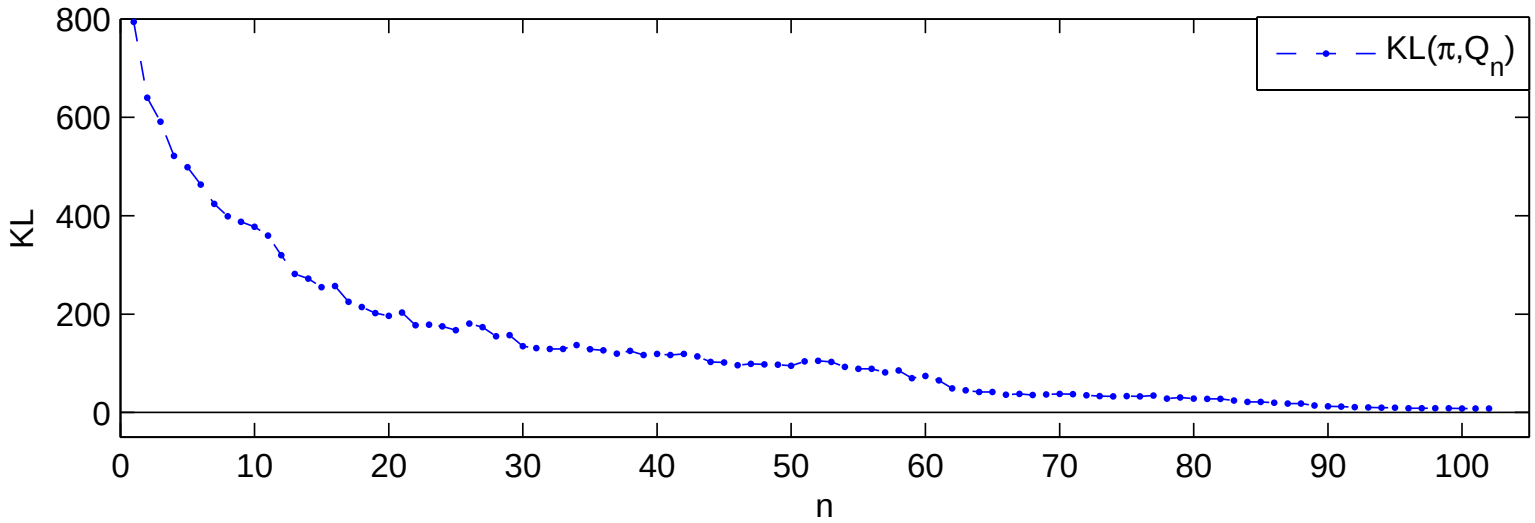


Figure: Kullback–Leibler divergence between π_2 and Q_n the sequence of proposal distributions created by AIMM, ($Q_n \equiv Q_{m_n}$)

Robustness of AIMM to "bad prior" ϕ_0

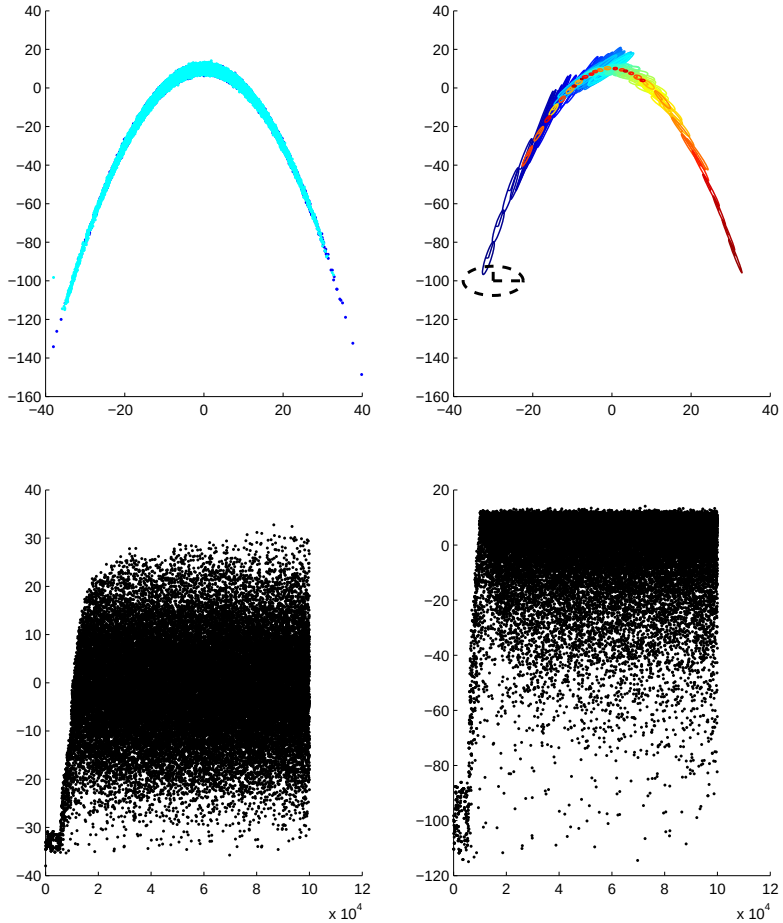


Figure: sample path of one Markov chain with "bad" initial proposal displayed by the thick ellipse (upper right hand side) – note the sequence of kernels created by AIMM in rainbow style

Robustness of AIMM in case of "bad prior" ϕ_0 (even worst)

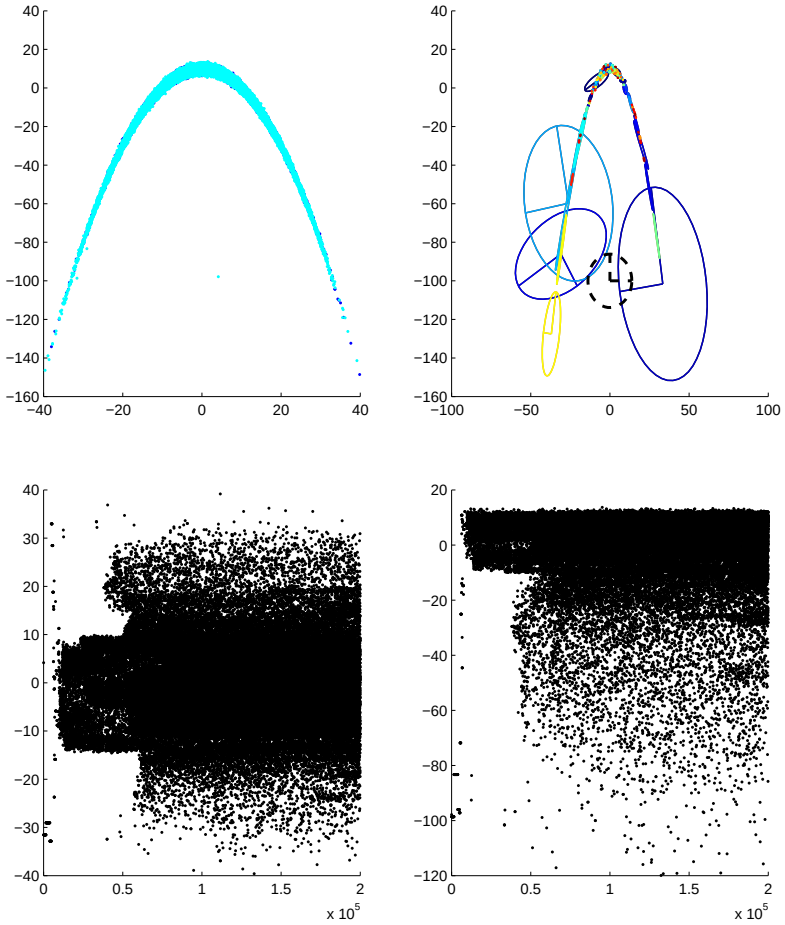
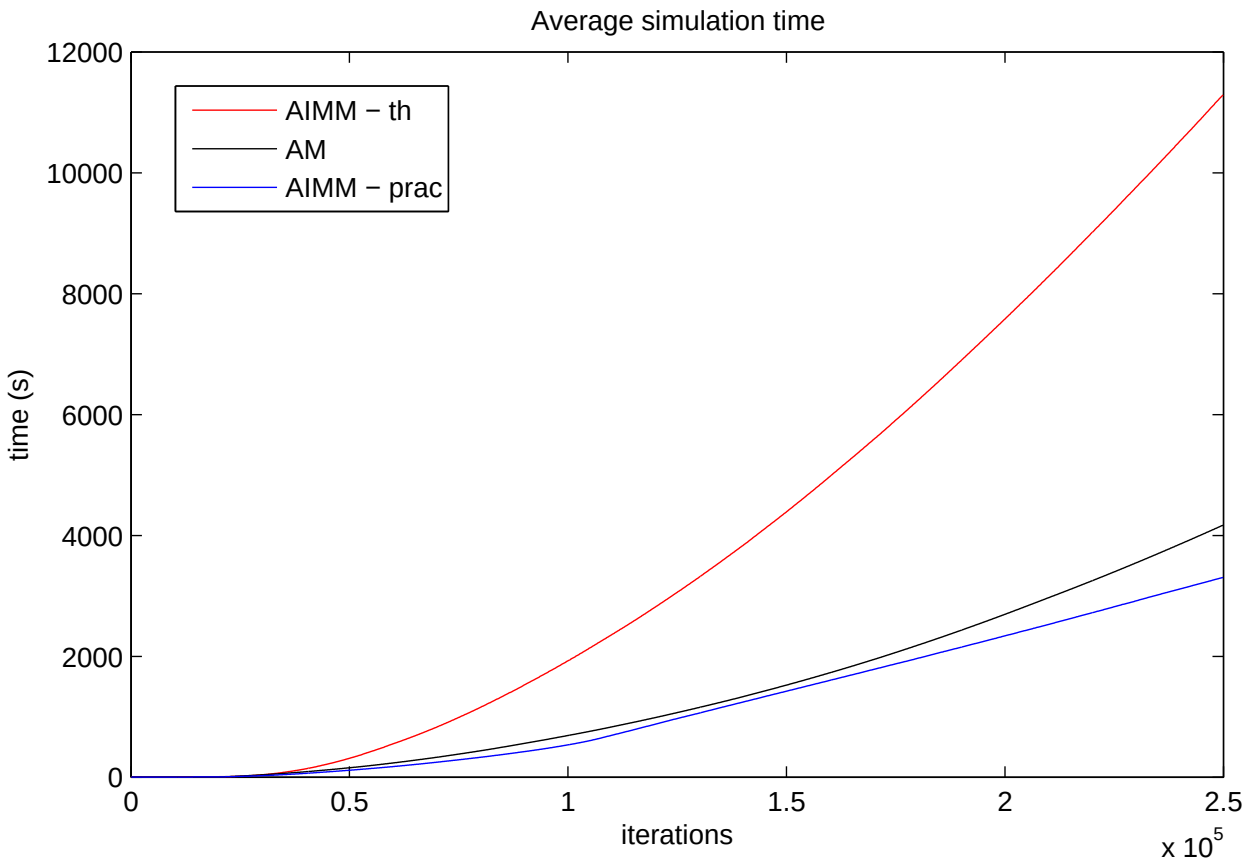


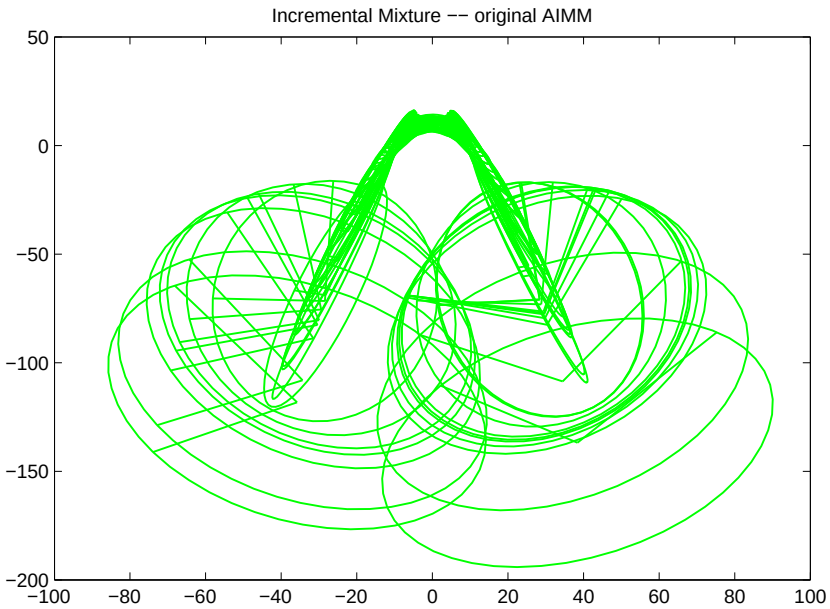
Figure: sample path of one Markov chain with "bad" initial proposal displayed by the thick ellipse (upper right hand side) – note the sequence of kernels created by AIMM in rainbow style

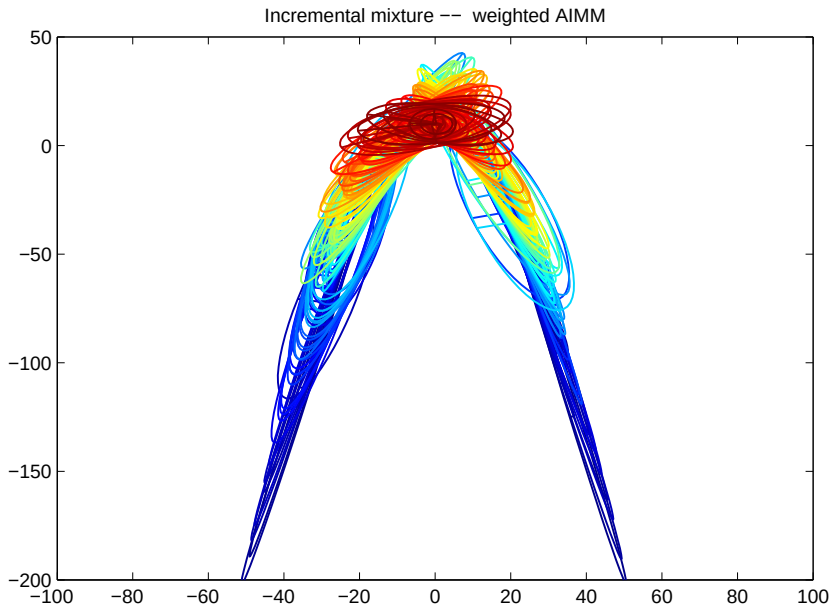
Average simulation time



- We work hereafter with chain of length $250,000$ so that both AM and AIMM-prac are time normalized (conservative)
- Clearly we see here that AIMM-th is exponential in CPU time (while by construction, AIMM-prac is linear)...

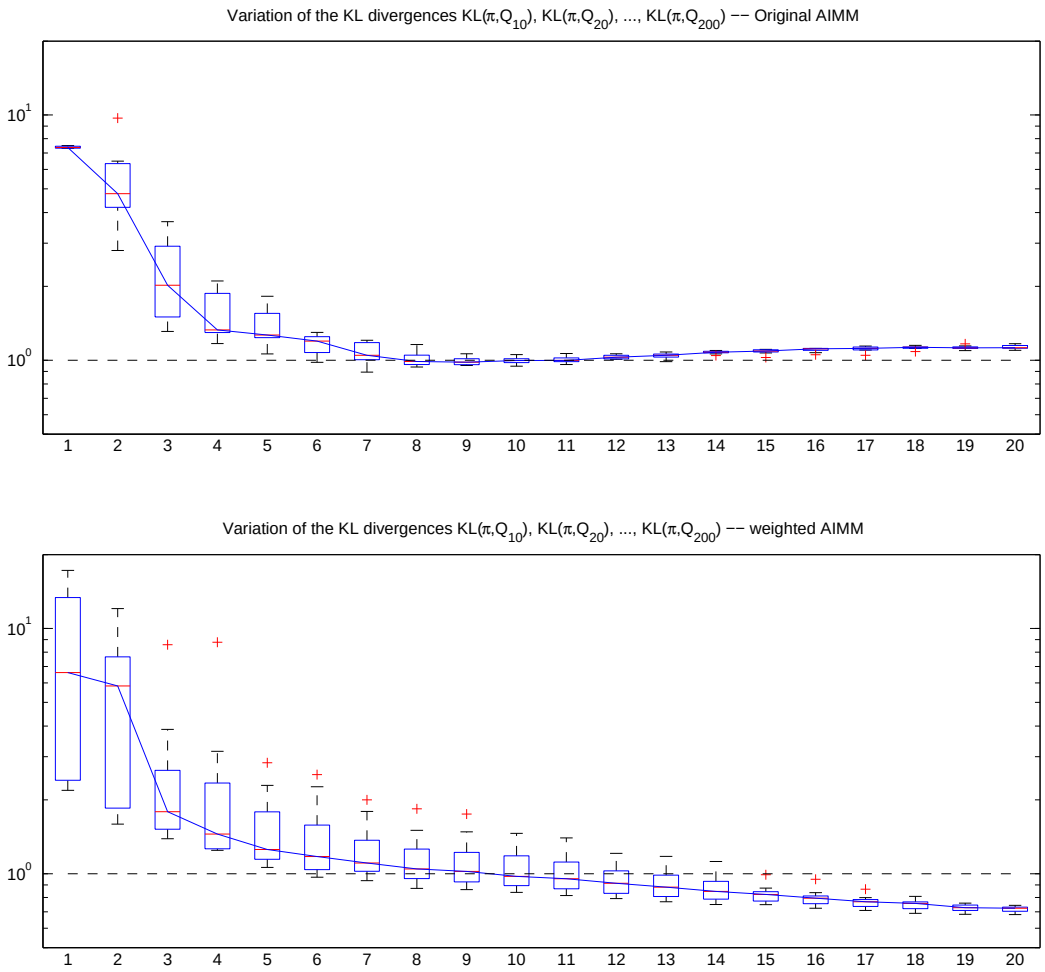
A word about the two AIMM methods





The components located in the tail represents a waste of mass: they reduce the acceptance rate and increase the ACF of the chain

A word about the two AIMM methods



estimated KL between π and the incremental mixture

Comparing AIMM and AM in dimension $d = 10$

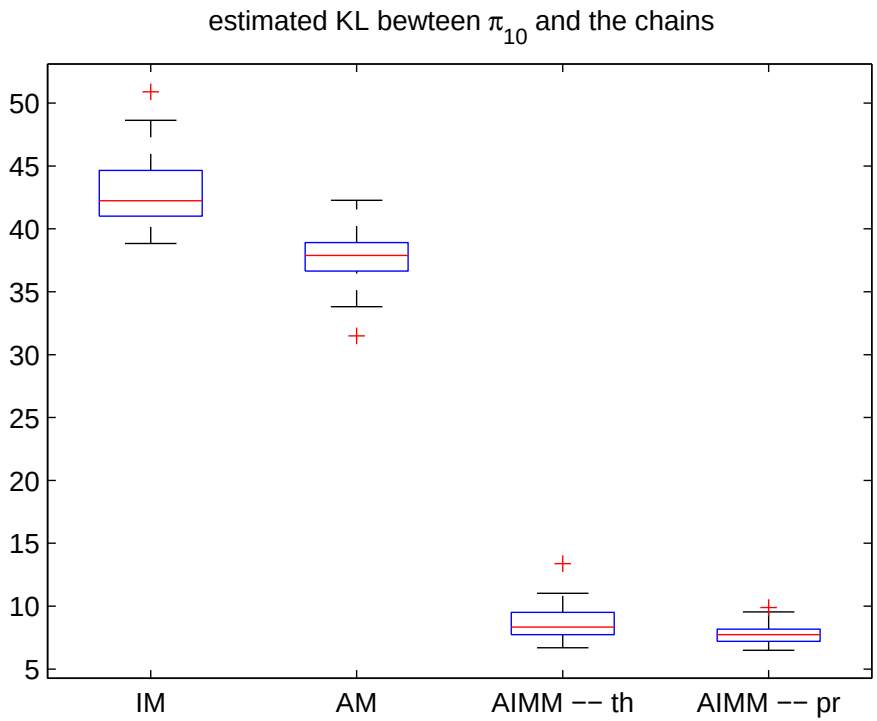
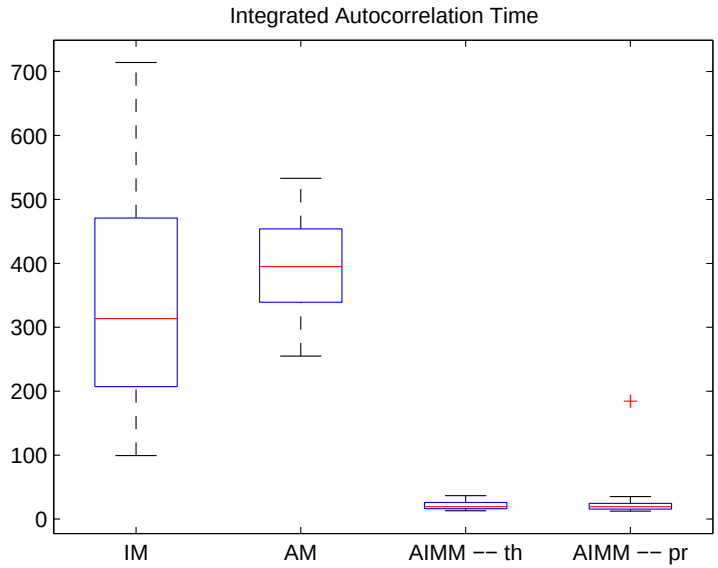
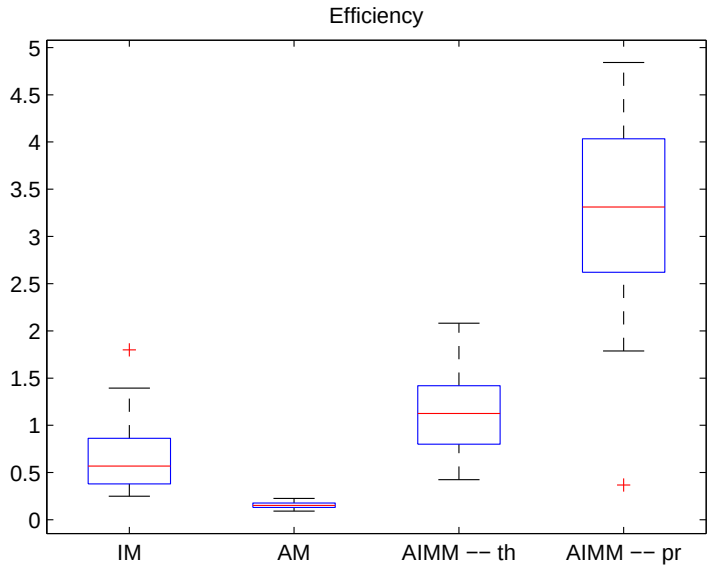


Figure: KL between π_{10} and the kernel density approximation of the 4 chains at stationarity (50 runs)

Integrated Autocorrelation Time and Efficiency





No surprise as acceptance rate in AM is on average .1 while it is .3 – .4 for both AIMMs

$$EFF = N/(\mathrm{T_{CPU}} \times \mathrm{IAT})$$

Tail exploration

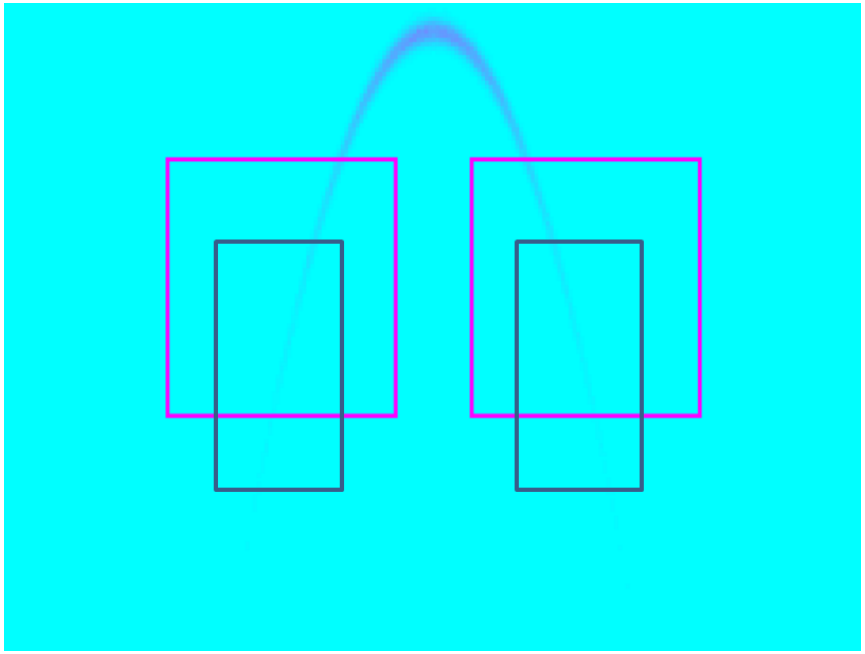
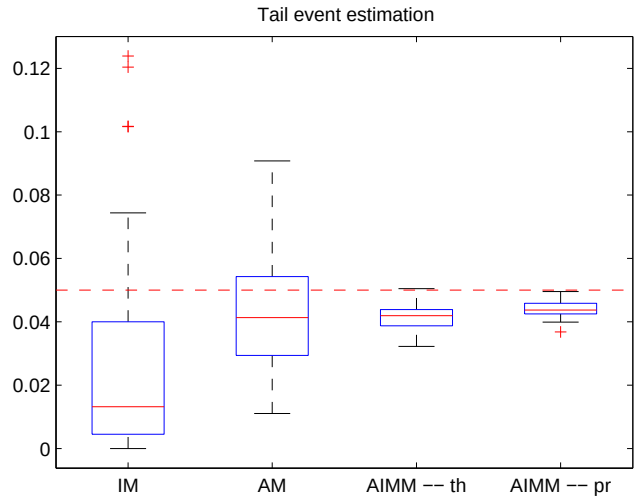
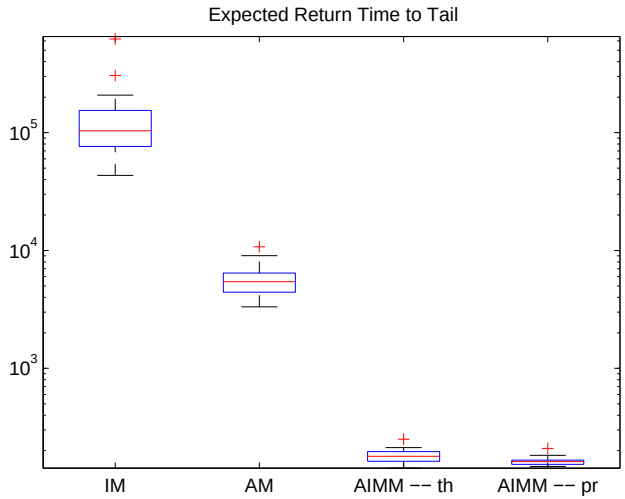
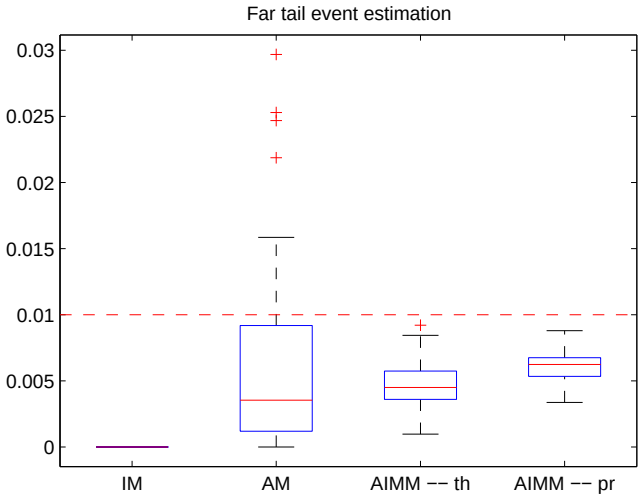


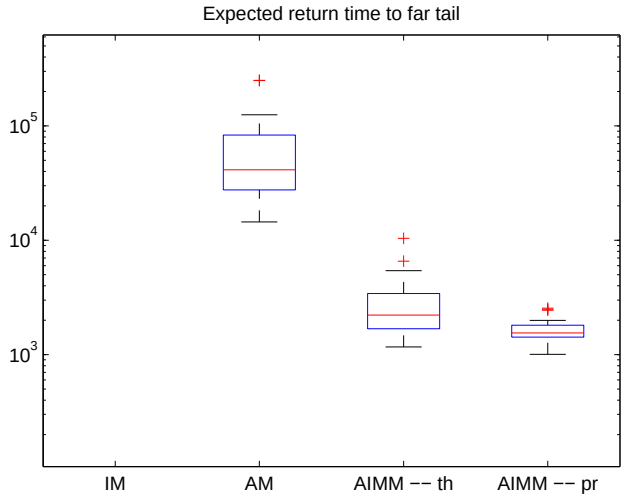
Figure: Projection of π_{10} in the two first dimensions with two events: blue, with Pr. .05 and gray, with Pr. .01

Tail estimation and Expected Returning Time to tail









Discussion / Perspectives

- AIMM seems to be a decent alternative to existing Adaptive MCMC
- Works well in situation where few initial knowledge is available
- Easy to implement...

Some theoretical properties would be nice to prove:

- Can we improve the proof of

$$\mathrm{KL}(\pi \| Q_{n+1}) \leq \mathrm{KL}(\pi \| Q_n) \ ?$$

  Or with some probability close to one...
- A better constant for the Doeblin condition $\Rightarrow$ the bound function $x \to \gamma(x)$ in the Doeblin condition needs to be improved!