Color test

I just want to be

sure

that all my favourite colors

are

being displayed correctly on this

new

device. If not I'll modify them.

How to unefficiently solve any problem

A tutorial on Bayesian numerical methods

Rémi Bardenet

LAL, LRI, Univ. Paris-Sud XI

May 30th, 2012

Summary

1. **A very generic problem**
2. **First sampling methods**
3. **MCMC algorithms**
4. **A taste of a *monster* MCMC sampler for Auger**

When

- a model $p(\text{data}|\theta)$ of an experiment has been written,
- a prior $p(\theta)$ has been set on the parameters,

then by Bayes' theorem:

$$\pi(\theta) := p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{\int_\Theta p(\text{data}|\theta)p(\theta)d\theta}.$$

When

- a model $p(\text{data}|\theta)$ of an experiment has been written,
- a prior $p(\theta)$ has been set on the parameters,

then by Bayes' theorem:

$$\pi(\theta) := p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{\int_\Theta p(\text{data}|\theta)p(\theta)d\theta}.$$

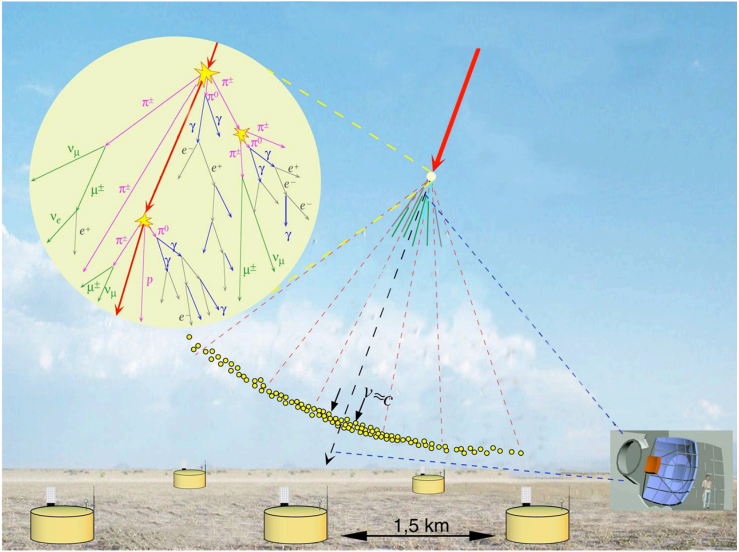
When

- a model $p(\text{data}|\theta)$ of an experiment has been written,
- a prior $p(\theta)$ has been set on the parameters,

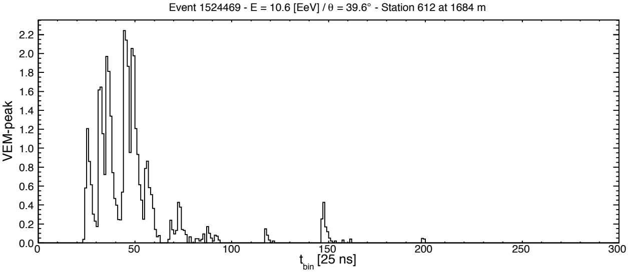
then by Bayes' theorem:

$$\pi(\theta) := p(\theta|\text{data}) = \frac{p(\text{data}|\theta)p(\theta)}{\int_\Theta p(\text{data}|\theta)p(\theta)d\theta}.$$

The generative process of an air shower



Looking at Auger tank signals



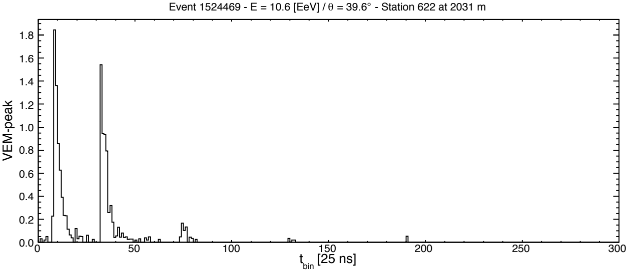
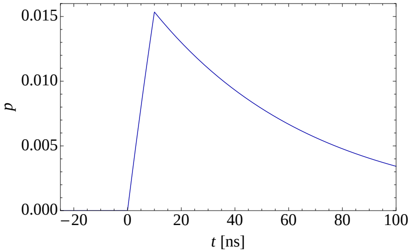
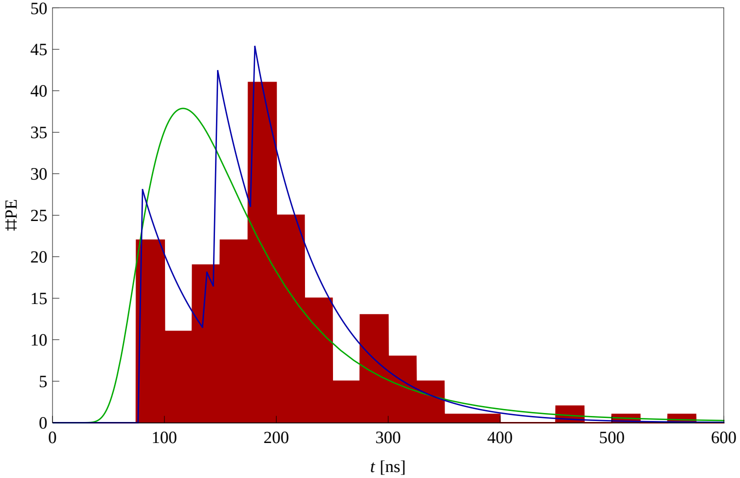


Figure: Muonic time response model p_{τ,t_d}



Mean number of Photo-electrons per bin

$$\bar{n}_i(A_\mu, t_\mu) = A_\mu \int_{t_{i-1}}^{t_i} p_{\tau,t_d}(t - t_\mu)dt.$$



$$n_i \text{ Poisson with mean } \bar{n}_i(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}}) = \sum_{j=1}^{N_\mu} \bar{n}_i(A_{\boldsymbol{\mu}_j}, t_{\boldsymbol{\mu}_j}),$$

So what is what in the cas $N_\mu = 1$?

- Parameters to infer are $\theta = (t_\mu, A_\mu)$.
- The likelihood is fixed by the model:

$$p(\text{data}|\theta) = \prod_{i=1}^{N_{\text{bins}}} \frac{\bar{n}_i(\theta)^{n_i}}{n_i!} e^{-\bar{n}_i(\theta)}.$$

- Choose e.g. a uniform prior:

$$p(\theta) = \frac{1}{C}\frac{1}{D}\,\chi_{[0,C]}(t_\mu)\,\chi_{[0,D]}(A_\mu),$$

- Then the posterior reads

$$p(\theta|\text{data}) \propto \chi_{[0,C]}(t_\mu)\,\chi_{[0,D]}(A_\mu)\prod_{i=1}^{N_{\text{bins}}} \bar{n}_i(\theta)^{n_i} e^{-\bar{n}_i(\theta)}.$$

Bayesian inference requires computing integrals

- MEP estimate (aka Bayes'): compute

$$\hat{\theta}_{\text{MEP}} := \int_{\Theta} \theta p(\theta|\text{data}) d\theta,$$

- credible intervals: find I such that

$$\int_I p(\theta|\text{data}) d\theta \geq 1 - \alpha.$$

- Bayes' factors: require evidence computations

$$p(\text{data}|M) = \int p(\text{data}|\theta, M) p(\theta|M) d\theta.$$

The MAP estimate

$$\theta_{\text{MAP}} := \arg\max p(\theta|\text{data})$$

with "Hessian" credible interval does not require integrals.

Averaging whenever necessary

Most Bayesian inference tasks require integrals wrt. the posterior

$$\int h(\theta)p(\theta|\text{data})d\theta.$$

The Monte Carlo principle

$$I := \int h(x)\pi(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} h(x_i) =: \hat{I}_N,$$

where $X_{1:N} \sim \pi \quad i.i.d.$

- $\hat{I}_N$ is unbiased:

$$\mathbb{E}\hat{I}_N = I.$$

- Error bars shrink at speed $1/\sqrt{N}$:

$$\text{Var}(\hat{I}_N) = \frac{\text{Var}(X)}{N}.$$

Most Bayesian inference tasks require integrals wrt. the posterior

$$\int h(\theta) p(\theta|\text{data}) d\theta.$$

The Monte Carlo principle

$$I := \int h(x)\pi(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} h(x_i) =: \hat{I}_N,$$

where $X_{1:N} \sim \pi \quad i.i.d.$

- $\hat{I}_N$ is unbiased:

$$\mathbb{E}\hat{I}_N = I.$$

- Error bars shrink at speed $1/\sqrt{N}$:

$$\mathrm{Var}(\hat{I}_N) = \frac{\mathrm{Var}(X)}{N}.$$

Most Bayesian inference tasks require integrals wrt. the posterior

$$\int h(\theta)p(\theta|\text{data})d\theta.$$

The Monte Carlo principle

$$I := \int h(x)\pi(x)dx \approx \frac{1}{N}\sum_{i=1}^{N} h(x_i) =: \hat{I}_N,$$

where $X_{1:N} \sim \pi \quad i.i.d.$

- $\hat{I}_N$ is unbiased:

$$\mathbb{E}\hat{I}_N = I.$$

- Error bars shrink at speed $1/\sqrt{N}$:

$$\mathrm{Var}(\hat{I}_N) = \frac{\mathrm{Var}(X)}{N}.$$

Pros & Cons

Pros

- Numerical integration (grid methods) is too costly and imprecise when $d \geq 6$.
- MC should concentrate the effort on places where π is big.
- Many, many applications !

Cons

One must be able to sample from π !

$\rightarrow$ Need for generic clever sampling methods !

Pros & Cons

Pros

- Numerical integration (grid methods) is too costly and imprecise when $d \geq 6$.
- MC should concentrate the effort on places where π is big.
- Many, many applications !

Cons

One must be able to sample from π !

$\rightarrow$ Need for generic clever sampling methods !

Pros

- Numerical integration (grid methods) is too costly and imprecise when $d \geq 6$.
- MC should concentrate the effort on places where π is big.
- Many, many applications !

Cons

One must be able to sample from π !

$\rightarrow$ Need for generic clever sampling methods !

Pros

- Numerical integration (grid methods) is too costly and imprecise when $d \geq 6$.
- MC should concentrate the effort on places where π is big.
- Many, many applications !

Cons

One must be able to sample from π !

$\longrightarrow$ Need for generic clever sampling methods !

"A complete Monte Carlo hybrid simulation has been performed to study the trigger efficiency and the detector performance. The simulation sample consists of about 6000 proton and 3000 iron CORSIKA [19] showers"

from "The exposure of the hybrid detector of the P. Auger Observatory", the Auger Collab., Astroparticle Phys., 2010.

Non-Bayesian applications of sampling algorithms: EPOS

"The difficulty with Monte Carlo generation of interaction configurations arises from the fact that the configuration space is huge and rather nontrivial [...] the only way to proceed amounts to employing dynamical MC methods"

from "Parton-based Gribov-Regge theory", Drescher et al., Phys.Rept., 2008.

Summary

Principle

If $U \sim \mathcal{U}_{(0,1)}$ and F is a cdf, then

$$F^{-1}(U) \sim F.$$

- It assumes we know how to sample from $\mathcal{U}(0,1)$!
- It needs a known and convenient cdf.

Principle

If $U \sim \mathcal{U}_{(0,1)}$ and F is a cdf, then

$$F^{-1}(U) \sim F.$$

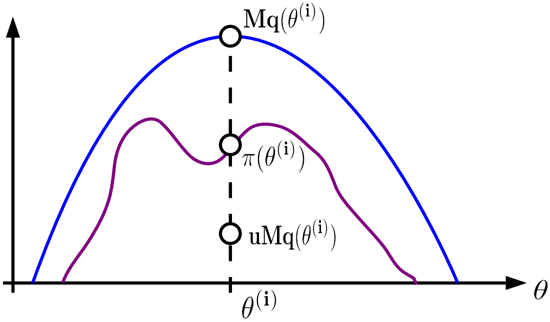
- It assumes we know how to sample from $\mathcal{U}(0,1)$!
- It needs a known and convenient cdf.

Rejection Sampling 1/2

Assumptions

- Target π is known up to a multiplicative constant,
- an easy-to-sample distribution q s.t. $\pi \leq Mq$ is known.

```
RejectionSampling(π, q, M, N)
    for i ← 1 to N,
        Sample θ^(i) ~ q and u ~ U_(0,1).
        Form the acceptance ratio ρ = π(θ^(i)) / (M q(θ^(i))).
        if u < ρ, then accept θ^(i).
        else reject.
```



Remarks & Drawbacks

- One needs to know good q and M,
- Lots of wasted samples.

Assumptions & principle

- Target π is fully known,
- an easy-to-sample distribution q is known s.t. $\mathrm{Supp}(\pi) \subset \mathrm{Supp}(q)$.

Then

$$\hat{\mathcal{I}}_N := \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}) \frac{\pi(\theta^{(i)})}{q(\theta^{(i)})} \to \int_\Theta h(\theta)\pi(\theta)d\theta, \quad \theta^{(i)} \sim q \ i.i.d.$$

ImportanceSampling(π, q, N)

1 Sample independent $\theta^{(i)} \sim q, i = 1..N$,

2 Form the weights $w_i = \frac{\pi(\theta^{(i)})}{q(\theta^{(i)})}$.

3 π is approximated by $\frac{1}{N}\sum_{i=1}^{N} w_i \delta_{\theta^{(i)}}$.

Importance Sampling 2/2

- If $|h| \leq M$, then

$$\mathrm{Var}(\hat{\mathcal{I}}_N) \leq \frac{M^2}{N}\Big(\int \frac{(\pi - q)^2}{q} d\theta + 1\Big),$$

thus q has to be close to π and have heavier tails than π.

- Further: the q achieving the smallest variance is

$$q = \frac{|h|\pi}{\int |h|\pi d\theta}.$$

- Better and achievable: For smaller variance, or if π is known up to a constant, use

$$\tilde{\mathcal{I}}_N := \frac{\sum_{i=1}^N h(\theta^{(i)})\pi(\theta^{(i)})/q(\theta^{(i)})}{\sum_{i=1}^N \pi(\theta^{(i)})/q(\theta^{(i)})} \quad \text{(biased)}.$$

- One needs to know a good, heavy-tailed q.
- Adaptive strategies for tuning q are possible [WKB+09].

- If $|h| \leq M$, then
$$\mathrm{Var}(\hat{\mathcal{I}}_N) \leq \frac{M^2}{N}\Big(\int \frac{(\pi - q)^2}{q} d\theta + 1\Big),$$
thus q has to be close to π and have heavier tails than π.
- Further: the q achieving the smallest variance is
$$q = \frac{|h|\pi}{\int |h|\pi d\theta}.$$
- Better and achievable: For smaller variance, or if π is known up to a constant, use
$$\tilde{\mathcal{I}}_N := \frac{\sum_{i=1}^N h(\theta^{(i)})\pi(\theta^{(i)})/q(\theta^{(i)})}{\sum_{i=1}^N \pi(\theta^{(i)})/q(\theta^{(i)})} \quad \text{(biased)}.$$
- One needs to know a good, heavy-tailed q.
- Adaptive strategies for tuning q are possible [WKB+09].

- If $|h| \leq M$, then

$$\mathrm{Var}(\hat{\mathcal{I}}_N) \leq \frac{M^2}{N}\Big(\int \frac{(\pi - q)^2}{q} d\theta + 1\Big),$$

thus q has to be close to π and have heavier tails than π.

- Further: the q achieving the smallest variance is

$$q = \frac{|h|\pi}{\int |h|\pi d\theta}.$$

- Better and achievable: For smaller variance, or if π is known up to a constant, use

$$\tilde{\mathcal{I}}_N := \frac{\sum_{i=1}^N h(\theta^{(i)})\pi(\theta^{(i)})/q(\theta^{(i)})}{\sum_{i=1}^N \pi(\theta^{(i)})/q(\theta^{(i)})} \quad \text{(biased)}.$$

- One needs to know a good, heavy-tailed q.
- Adaptive strategies for tuning q are possible [WKB$^+$09].

Importance Sampling 2/2

- If $|h| \leq M$, then

$$\mathrm{Var}(\hat{\mathcal{I}}_N) \leq \frac{M^2}{N}\Big(\int \frac{(\pi - q)^2}{q} d\theta + 1\Big),$$

thus q has to be close to π and have heavier tails than π.

- Further: the q achieving the smallest variance is

$$q = \frac{|h|\pi}{\int |h|\pi d\theta}.$$

- Better and achievable: For smaller variance, or if π is known up to a constant, use

$$\tilde{\mathcal{I}}_N := \frac{\sum_{i=1}^N h(\theta^{(i)})\pi(\theta^{(i)})/q(\theta^{(i)})}{\sum_{i=1}^N \pi(\theta^{(i)})/q(\theta^{(i)})} \quad \text{(biased)}.$$

- One needs to know a good, heavy-tailed q.
- Adaptive strategies for tuning q are possible [WKB+09].

Benchmark

Let $\pi(\theta) = \mathcal{N}(0, I_d)$ and $q(\theta) = \mathcal{N}(0, \sigma I_d)$.

- Rejection sampling
  - needs $\sigma \geq 1$.
  - Fraction of accepted proposals goes as σ^{-d}.
- Importance sampling
  - yields infinite variance when $\sigma \leq 1/\sqrt{2}$,
  - variance of the weights goes as

$$\left(\frac{\sigma^2}{2 - 1/\sigma^2}\right)^{d/2}.$$

Benchmark

Let $\pi(\theta) = \mathcal{N}(0, I_d)$ and $q(\theta) = \mathcal{N}(0, \sigma I_d)$.

- Rejection sampling
  - needs $\sigma \geq 1$.
  - Fraction of accepted proposals goes as σ^{-d}.
- Importance sampling
  - yields infinite variance when $\sigma \leq 1/\sqrt{2}$,
  - variance of the weights goes as

$$\left(\frac{\sigma^2}{2 - 1/\sigma^2}\right)^{d/2}.$$

Benchmark

Let $\pi(\theta) = \mathcal{N}(0, I_d)$ and $q(\theta) = \mathcal{N}(0, \sigma I_d)$.

- Rejection sampling
  - needs $\sigma \geq 1$.
  - Fraction of accepted proposals goes as σ^{-d}.
- Importance sampling
  - yields infinite variance when $\sigma \leq 1/\sqrt{2}$,
  - variance of the weights goes as

$$\left(\frac{\sigma^2}{2 - 1/\sigma^2}\right)^{d/2}.$$

We have seen that

- Sums & integrals are ubiquitous in Statistics.
- Numerical integration is limited to small dimensions.
- "Basic" sampling is limited to full-information easy cases.
- RS and IS, well-tuned, are efficient and versatile,

Now what do we do when

- it is not possible to sample from π directly, but only evaluate it pointwise, possibly up to a multiplicative constant:

$$\pi(\theta) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta)p(\theta)d\theta}.$$

- We don't know a good approximation q of π.
- Θ is high-dimensional.

Well, MCMC is bringing both answers and new problems !

We have seen that

- Sums & integrals are ubiquitous in Statistics.
- Numerical integration is limited to small dimensions.
- "Basic" sampling is limited to full-information easy cases.
- RS and IS, well-tuned, are efficient and versatile,

Now what do we do when

- it is not possible to sample from π directly, but only evaluate it pointwise, possibly up to a multiplicative constant:

$$\pi(\theta) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta)p(\theta)d\theta}.$$

- We don't know a good approximation q of π.
- Θ is high-dimensional.

Well, MCMC is bringing both answers and new problems !

We have seen that

- Sums & integrals are ubiquitous in Statistics.
- Numerical integration is limited to small dimensions.
- "Basic" sampling is limited to full-information easy cases.
- RS and IS, well-tuned, are efficient and versatile,

Now what do we do when

- it is not possible to sample from π directly, but only evaluate it pointwise, possibly up to a multiplicative constant:

$$\pi(\theta) = \frac{p(x|\theta)p(\theta)}{\int_\Theta p(x|\theta)p(\theta)d\theta}.$$

- We don't know a good approximation q of π.
- Θ is high-dimensional.

Well, MCMC is bringing both answers and new problems !

Summary

Goal is to explore Θ, spending more time in places where π is high.

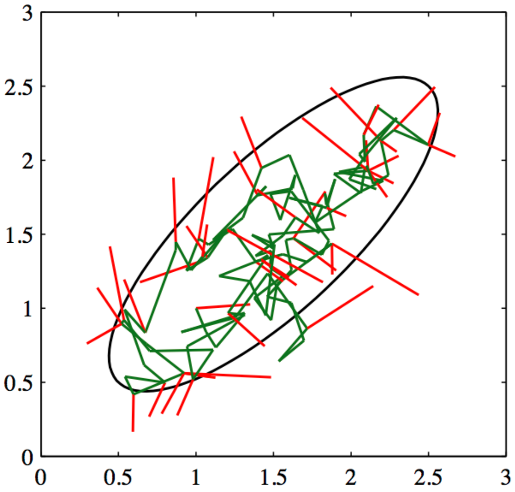


Figure: Taken from [Bis06]

- Metropolis' algorithm builds a random walk with symmetric steps q, that mimics independent draws from π.
- Symmetricity means $q(x|y) = q(y|x)$, e.g. $q(y|x) = \mathcal{N}(x, \sigma)$.

$\textsc{MetropolisSampler}\left(\pi, q, T, \theta^{(0)}\right)$

$\mathcal{S} \leftarrow \emptyset$.

for $t \leftarrow 1$ **to** T,

Sample $\theta^* \sim q(.|\theta^{(t-1)})$ and $u \sim \mathcal{U}_{(0,1)}$.

Form the acceptance ratio

$$\rho = 1 \wedge \frac{\pi(\theta^*)}{\pi(\theta^{(t-1)})}.$$

if $u < \rho$, **then** $\theta^{(t)} \leftarrow \theta^*$ **else** $\theta^{(t)} \leftarrow \theta^{(t-1)}$.

$\mathcal{S} \leftarrow \mathcal{S} \cup \{\theta^{(t)}\}$.

The Metropolis-Hastings algorithm

- When no symmetricity is assumed, change acceptance to

$$\rho(x,y) = 1 \wedge \frac{\pi(y)}{q(y|x)}\frac{q(x|y)}{\pi(x)}.$$

- Remark the exploration/exploitation trade-off.

MetropolisHastingsSampler$(\pi, q, T, \theta^{(0)})$

$\mathcal{S} \leftarrow \emptyset$.

for $t \leftarrow 1$ **to** T,

Sample $\theta^* \sim q(.|\theta^{(t-1)})$ and $u \sim \mathcal{U}_{(0,1)}$.

Form the acceptance ratio

$$\rho = 1 \wedge \frac{\pi(\theta^*)}{q(\theta^*|\theta^{(t-1)})}\frac{q(\theta^{(t-1)}|\theta^*)}{\pi(\theta^{(t-1)})}.$$

if $u < \rho$, **then** $\theta^{(t)} \leftarrow \theta^*$ **else** $\theta^{(t)} \leftarrow \theta^{(t-1)}$.

$\mathcal{S} \leftarrow \mathcal{S} :: \{\theta^{(t)}\}$.

- A (stationary) Markov chain $(\theta^{(t)})$ is defined through a kernel:

$$\mathbb{P}(\theta^{(t+1)} \in dy|\theta^{(t)} = x) = K(x, dy).$$

- If $\theta^{(0)} = x$, then

$$\begin{aligned}\mathbb{P}(\theta^{(t)} \in dy|\theta^{(0)} = x) &= \int\int K(x, dx_1)\dots K(dx_{t-1}, dy) \\ &=: K^t(x, dy).\end{aligned}$$

- Under technical conditions, the MH kernel K satisfies

$$\|\pi - K^t(x,.)\| \to 0, \forall x.$$

- This justifies a burn-in period after which samples are discarded.
- Further results like a Law of Large Numbers guarantee that

$$\frac{1}{T+1}\sum_{t=0}^{T} h(\theta^{(t)}) \to \int h(\theta)\pi(\theta)d(\theta)$$

  for h bounded.
- Central Limit Theorem-type results also exist, see Section 6.7 of [RC04].

Autocorrelation in equilibrium

- For a scalar chain, define the integrated autocorrelation time

$$\tau_{\text{int}} = 1 + 2\sum_{k>0} \text{Corr}\,(\theta^{(t)}, \theta^{(t+k)}).$$

- One can show that

$$\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})\right) = \frac{\tau_{\text{int}}}{T}\,\text{Var}\left(h(\theta^{(0)})\right).$$

Rule of thumb for the proposal

Optimizing similar criteria leads to choosing $q(.|x) = \mathcal{N}(x, \sigma^2)$ s.t.

- acceptance rate is ≈ 0.5 for $d = 1, 2$.
- acceptance rate is ≈ 0.25 for $d \geq 3$.

- For a scalar chain, define the integrated autocorrelation time

$$\tau_{\text{int}} = 1 + 2\sum_{k>0} \text{Corr}\left(\theta^{(t)}, \theta^{(t+k)}\right).$$

- One can show that

$$\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})\right) = \frac{\tau_{\text{int}}}{T}\,\text{Var}\left(h(\theta^{(0)})\right).$$

Rule of thumb for the proposal

Optimizing similar criteria leads to choosing $q(.|x) = \mathcal{N}(x, \sigma^2)$ s.t.

- acceptance rate is ≈ 0.5 for $d = 1, 2$.
- acceptance rate is ≈ 0.25 for $d \geq 3$.

Autocorrelation in equilibrium

- For a scalar chain, define the integrated autocorrelation time

$$\tau_{\text{int}} = 1 + 2 \sum_{k>0} \text{Corr}\,(\theta^{(t)}, \theta^{(t+k)}).$$

- One can show that

$$\text{Var}\left(\frac{1}{T}\sum_{t=1}^{T} h(\theta^{(t)})\right) = \frac{\tau_{\text{int}}}{T}\ \text{Var}\left(h(\theta^{(0)})\right).$$

Rule of thumb for the proposal

Optimizing similar criteria leads to choosing $q(.|x) = \mathcal{N}(x, \sigma^2)$ s.t.

- acceptance rate is ≈ 0.5 for $d = 1, 2$.
- acceptance rate is ≈ 0.25 for $d \geq 3$.

A practical wrap-up

We have justified the acceptance ratio, the burn-in period, and the optimization of the proposal.

MetropolisHastingsSampler$(\pi, q, T, \theta^{(0)}, \Sigma_0)$

$\mathcal{S} \leftarrow \emptyset.$

for $t \leftarrow 1$ **to** T,

Sample $\theta^* \sim \mathcal{N}(.|\theta^{(t-1)}, \sigma\Sigma_0)$ and $u \sim \mathcal{U}_{(0,1)}$.

Form the acceptance ratio

$$\rho = 1 \wedge \frac{\pi(\theta^*)}{q(\theta^*|\theta^{(t-1)})} \frac{q(\theta^{(t-1)}|\theta^*)}{\pi(\theta^{(t-1)})}.$$

if $u < \rho$, **then** $\theta^{(t)} \leftarrow \theta^*$ **else** $\theta^{(t)} \leftarrow \theta^{(t-1)}$.

if $t \leq T_b$,

$\sigma \leftarrow \sigma + \frac{1}{t^{0.6}}(\text{acc. rate} - 0.25/0.50)$.

else if $t > T_b$, **then** $\mathcal{S} \leftarrow \mathcal{S} \cup \theta^{(t)}$.

Feel free to experiment with Laird Breyer's applet on

http://www.lbreyer.com/classic.html.

- Consider again our muonic signal reconstruction task, with

$$p(\text{data}|\theta) = \prod_{i=1}^{N_{\text{bins}}} \frac{\bar{n}_i(\theta)^{n_i}}{n_i!} e^{-\bar{n}_i(\theta)}.$$

- The model (the physics) suggest using specific independent priors for A_μ and t_μ.

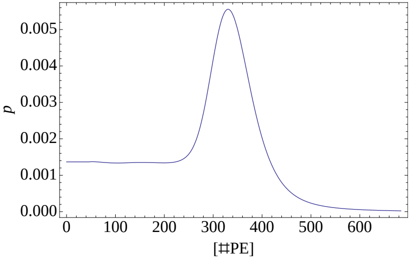
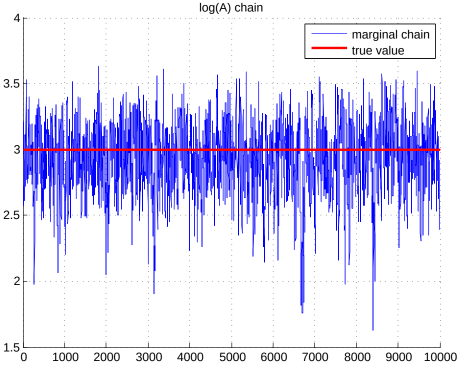
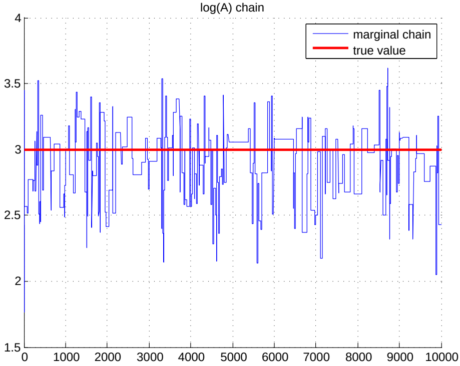


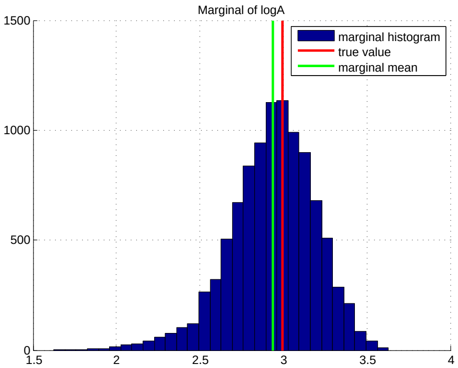
Figure: Prior on A_μ for a given zenith angle

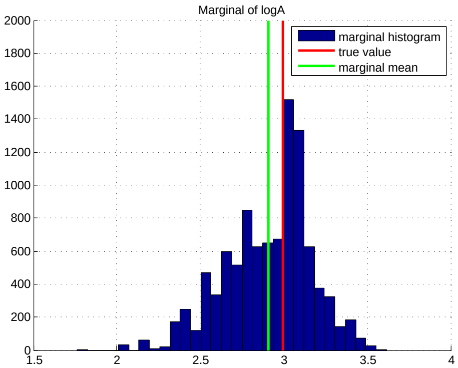




- One can often detect bad mixing by eye.
- Acceptance for the left panel chain is only 3%.

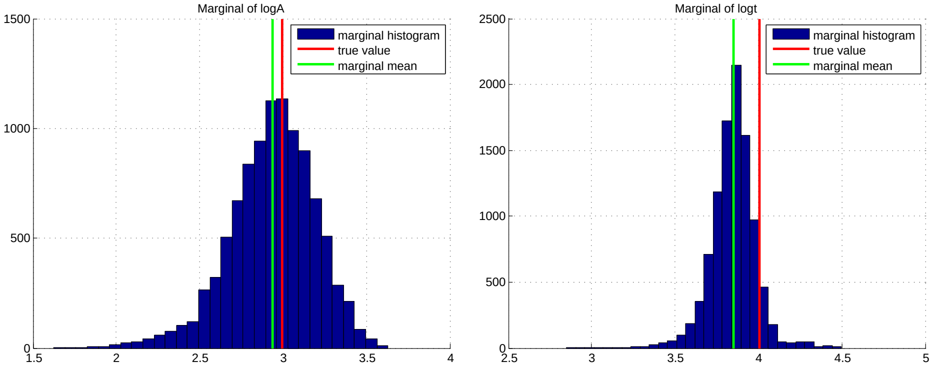
Case study: simulation results





- Marginals are simply component histograms !
- Even the marginals look ugly when mixing is bad.

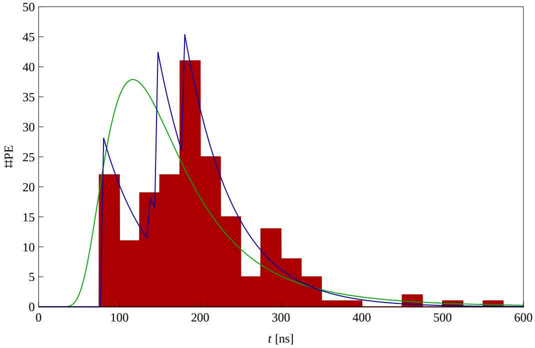
Case study: simulation results



- From now on, we only consider the good mixing case.
- Try to reproduce your marginals with different starting values!
- Producing the chain was the hard part. Now everything is easy: estimation, credible intervals, ...

- Recall

$$\bar{n}_i(A_{\boldsymbol{\mu}}, t_{\boldsymbol{\mu}}) = A_\mu \int_{t_{i-1}}^{t_i} p_{\tau,t_d}(t - t_\mu) dt.$$



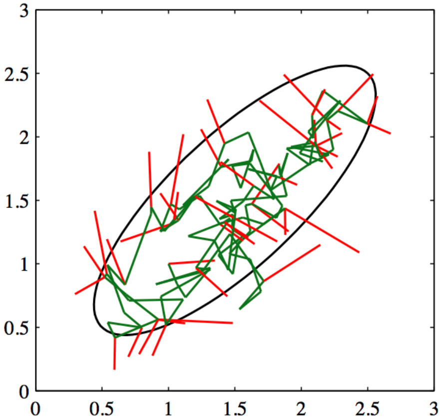
- Now

$$n_i \text{ Poisson with mean } \bar{n}_i(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}}) = \sum_{j=1}^{N_\mu} \bar{n}_i(A_{\boldsymbol{\mu}_j}, t_{\boldsymbol{\mu}_j}),$$

- Bayesian inference: obtain

$$\pi(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}}) = p(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}} | \text{signal}) \propto p(\text{signal} | \mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}}) p(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}}).$$

- MCMC methods sample from π.



Metropolis (B) and adaptive Metropolis (B+G) algorithms

$\mathrm{M}(\pi(x), X_0, \Sigma_0, T,)$

$\mathcal{S} \leftarrow \emptyset,\ \Sigma \leftarrow \Sigma_0$

for $t \leftarrow 1$ **to** T

$\quad \widetilde{X} \sim \mathcal{N}(\cdot \,|X_{t-1}, \Sigma)$ ▷ *proposal*

$\quad$ **if** $\dfrac{\pi(\widetilde{X})}{\pi(X_{t-1})} > \mathcal{U}[0,1]$ **then**

$\quad\quad X_t \leftarrow X$ ▷ *accept*

$\quad$ **else**

$\quad\quad X_t \leftarrow X_{t-1}$ ▷ *reject*

$\quad \mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$ ▷ *update posterior sample*

return $\mathcal{S}$

AM$(\pi(x), X_0, \Sigma_0, T, \mu_0, c)$

$\mathcal{S} \leftarrow \emptyset$,

for $t \leftarrow 1$ **to** T

  $\Sigma \leftarrow c\Sigma_{t-1}$ ▷ *scaled adaptive covariance*

  $\widetilde{X} \sim \mathcal{N}(\cdot \,|X_{t-1}, \Sigma)$ ▷ *proposal*

  if $\dfrac{\pi(\widetilde{X})}{\pi(X_{t-1})} > \mathcal{U}[0,1]$ **then**

    $X_t \leftarrow X$ ▷ *accept*

  else

    $X_t \leftarrow X_{t-1}$ ▷ *reject*

  $\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$ ▷ *update posterior sample*

  $\mu_t \leftarrow \mu_{t-1} + \dfrac{1}{t}(X_t - \mu_{t-1})$ ▷ *update running mean and covariance*

  $\Sigma_t \leftarrow \Sigma_{t-1} + \dfrac{1}{t}\left((X_t - \mu_{t-1})(X_t - \mu_{t-1})^\intercal - \Sigma_{t-1}\right)$

  $c \leftarrow c + \frac{1}{t^{0.6}}(\text{acc. rate} - 0.25/0.50)$.

return $\mathcal{S}$

- Possibly high dimensions but also highly correlated model.
  - Use adaptive proposals.
- The number of muons N_μ is unknown.
  - Use a nonparametric prior [Nea00] or
  - use a Reversible Jump sampler [Gre95].
- Likelihood $p(\mathbf{n}|\mathbf{A}_\mu, \mathbf{t}_\mu)$ is permutation invariant.
  - Indeed, if $N_\mu = 2$,

$$p(\mathbf{n}|\ A_1, A_2,\ t_1, t_2) = p(\mathbf{n}|\ A_2, A_1,\ t_2, t_1).$$

  - Marginals are useless, a problem known as label-switching [Ste00].

Three problems, many "solutions"

- Possibly high dimensions but also highly correlated model.
  - Use adaptive proposals.
- The number of muons N_μ is unknown.
  - Use a nonparametric prior [Nea00] or
  - use a Reversible Jump sampler [Gre95].
- Likelihood $p(\mathbf{n}|\mathbf{A}_\mu, \mathbf{t}_\mu)$ is permutation invariant.
  - Indeed, if $N_\mu = 2$,

$$p(\mathbf{n}|\ A_1, A_2,\ t_1, t_2) = p(\mathbf{n}|\ A_2, A_1,\ t_2, t_1).$$

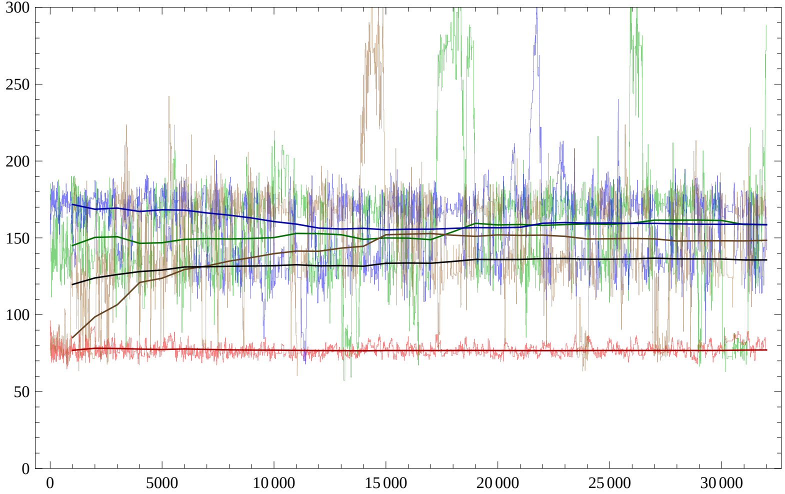
  - Marginals are useless, a problem known as label-switching [Ste00].

- Possibly high dimensions but also highly correlated model.
  - Use adaptive proposals.
- The number of muons N_μ is unknown.
  - Use a nonparametric prior [Nea00] or
  - use a Reversible Jump sampler [Gre95].
- Likelihood $\mathcal{P}(\mathbf{n}|\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}})$ is permutation invariant.
  - Indeed, if $N_\mu = 2$,

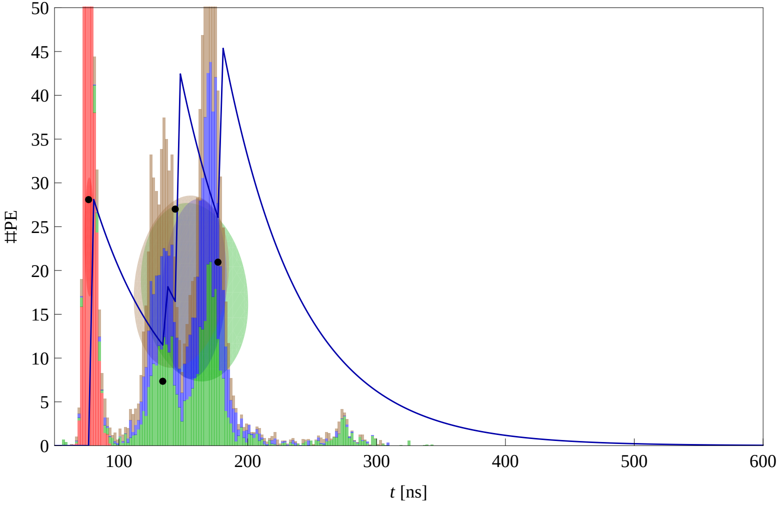
$$p(\mathbf{n}|\ A_1, A_2,\ t_1, t_2) = p(\mathbf{n}|\ A_2, A_1,\ t_2, t_1).$$

  - Marginals are useless, a problem known as label-switching [Ste00].

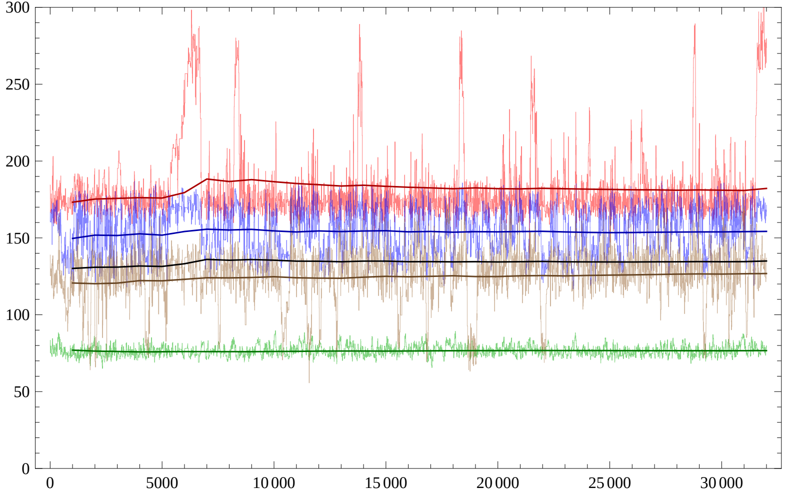
- If the prior is also permutation invariant, then so is $\pi(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}})$.



- If the prior is also permutation invariant, then so is $\pi(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}})$.



- If the prior is also permutation invariant, then so is $\pi(\mathbf{A}_{\boldsymbol{\mu}}, \mathbf{t}_{\boldsymbol{\mu}})$.



AMOR$(\pi(x), X_0, T, \mu_0, \Sigma_0, c)$

$\mathcal{S} \leftarrow \emptyset$

for $t \leftarrow 1$ **to** T

$\Sigma \leftarrow c\Sigma_{t-1}$ $\triangleright$ *scaled adaptive covariance*

$\widetilde{X} \sim \mathcal{N}(\,\cdot\,|X_{t-1}, \Sigma)$ $\triangleright$ *proposal*

$\widetilde{P} \sim \underset{P \in \mathfrak{P}}{\arg\min}\, L_{(\mu_{t-1}, \Sigma_{t-1})}(P\widetilde{X})$ $\triangleright$ *pick an optimal permutation*

$\widetilde{X} \leftarrow \widetilde{P}\widetilde{X}$ $\triangleright$ *permute*

if $\dfrac{\pi(\widetilde{X})\sum_{P \in \mathfrak{P}} \mathcal{N}(PX_{t-1}|X, \Sigma)}{\pi(X_{t-1})\sum_{P \in \mathfrak{P}} \mathcal{N}(PX|X_{t-1}, \Sigma)} > \mathcal{U}[0, 1]$ **then**

$X_t \leftarrow X$ $\triangleright$ *accept*

else

$X_t \leftarrow X_{t-1}$ $\triangleright$ *reject*

$\mathcal{S} \leftarrow \mathcal{S} \cup \{X_t\}$ $\triangleright$ *update posterior sample*

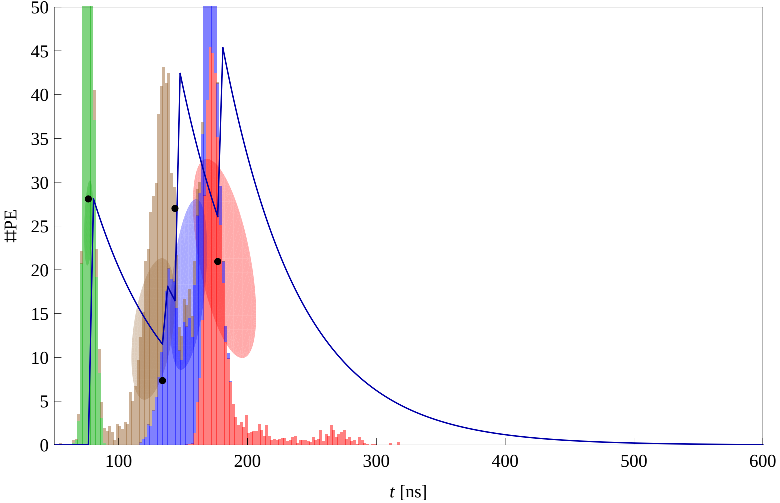
$\mu_t \leftarrow \mu_{t-1} + \dfrac{1}{t}(X_t - \mu_{t-1})$ $\triangleright$ *update running mean and covariance*

$\Sigma_t \leftarrow \Sigma_{t-1} + \dfrac{1}{t}\left((X_t - \mu_{t-1})(X_t - \mu_{t-1})^{\mathsf{T}} - \Sigma_{t-1}\right)$

$c \leftarrow c + \frac{1}{t^{0.6}}(\text{acc. rate} - 0.25/0.50)$.

return $\mathcal{S}$

AMOR results on the previous example



Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,

Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,
2. Quasi-MC,

Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,
2. Quasi-MC,
3. Simple MC, Importance Sampling,

Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,
2. Quasi-MC,
3. Simple MC, Importance Sampling,
4. MCMC (MH, Slice sampling, etc.),

Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,
2. Quasi-MC,
3. Simple MC, Importance Sampling,
4. MCMC (MH, Slice sampling, etc.),
5. Adaptive MCMC [AFMP11], Hybrid MC [Nea10], Tempering methods [GRS96], SMC [DdFG01], particle MCMC [ADH10].

Growing item number means higher complexity and either higher efficiency or wider applicability. Check [RC04] when no further indication is given.

1. Quadrature,
2. Quasi-MC,
3. Simple MC, Importance Sampling,
4. MCMC (MH, Slice sampling, etc.),
5. Adaptive MCMC [AFMP11], Hybrid MC [Nea10], Tempering methods [GRS96], SMC [DdFG01], particle MCMC [ADH10].
6. ABC [FP12].

Conclusions

Take-home message

- MC provides generic integration methods,
- Potential applications in Physics are numerous:
  - in forward sampling (aka "simulation"),
  - in Bayesian inference tasks.
- Producing a good mixing MCMC chain can be difficult
- Higher efficiency can result from:
  - Learning dependencies.
  - Exploiting existing/added structure of the problem.
- Broad range of methods allows to find the level of sophistication required by the difficulty of your problem.

Take-home message

- MC provides generic integration methods,
- Potential applications in Physics are numerous:
  - in forward sampling (aka "simulation"),
  - in Bayesian inference tasks.
- Producing a good mixing MCMC chain can be difficult
- Higher efficiency can result from:
  - Learning dependencies.
  - Exploiting existing/added structure of the problem.
- Broad range of methods allows to find the level of sophistication required by the difficulty of your problem.

“Introductory” references

- Tutorials: Lots on *videoconference.net*.
  - I. Murray's video lessons:

    videolectures.net/mlss09uk_murray_mcmc/
  - A. Sokal's tutorial, MCMC from a physicist's point of view + applications to Statistical Physics [Sok96]
- The Monte Carlo bible: C. Robert & G. Casella's “Monte Carlo Methods” [RC04], and references within.
- For more precise informations, please bug me.

- Christophe Andrieu, Arnaud Doucet, and Roman Holenstein, *Particle Markov chain Monte Carlo methods*, Journal of the Royal Statistical Society B (2010).

- Y. Atchadé, G. Fort, E. Moulines, and P. Priouret, *Bayesian time series models*, ch. Adaptive Markov chain Monte Carlo : Theory and Methods, pp. 33–53, Cambridge Univ. Press, 2011.

- R. Bardenet, O. Cappé, G. Fort, and B. Kégl, *An adaptive Metropolis algorithm with online relabeling*, International Conference on Artificial Intelligence and Statistics (AISTATS), (JMLR workshop and conference proceedings), vol. 22, April 2012, pp. 91–99.

- C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.

- R. Bardenet and B. Kégl, *An adaptive monte-carlo markov chain algorithm for inference from mixture signals*, Proceedings of ACAT'11, Journal of Physics: Conference series, 2012.
- A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo in practice*, Springer, 2001.
- P. Fearnhead and D. Prangle, *Constructing summary statistics for approx- imate bayesian computation; semi-automatic abc*, Journal of the Royal Statistical Society B (2012).
- P. J. Green, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), no. 4, 711–732.
- W.R. Gilks, S. Richardson, and D. Spiegelhalter (eds.), *Markov chain Monte Carlo in practice*, Chapman & Hall, 1996.

- R. M. Neal, *Markov chain sampling methods for Dirichlet process mixture models*, Journal of Computational and Graphical Statistics **9** (2000), 249–265.
- ______, *Handbook of Markov Chain Monte Carlo*, ch. MCMC using Hamiltonian dynamics, Chapman & Hall / CRC Press, 2010.
- C. P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer-Verlag, New York, 2004.
- A.D. Sokal, *Monte Carlo methods in statistical mechanics: Foundations and new algorithms*, 1996, Lecture notes at the Cargèse summer school.
- M. Stephens, *Dealing with label switching in mixture models*, Journal of the Royal Statistical Society, Series B **62** (2000), 795–809.

D. Wraith, M. Kilbinger, K. Benabed, O. Cappé, J.-F. Cardoso, G. Fort, S. Prunet, and C. P. Robert, *Estimation of cosmological parameters using adaptive importance sampling*, Phys. Rev. D **80** (2009), no. 2.

Thanks for your attention