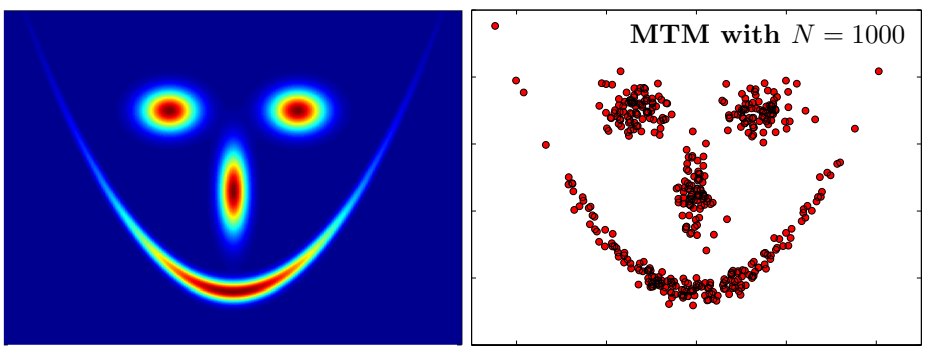
A brief journey through the MCMC world

Luca Martino

Machine Learning Group



March 11, 2013

INTRODUCTION

Introduction: framework

- In many applications, we receive several observations $\mathbf{y}$ and we are interested to obtain information to a related variable $\mathbf{z}$. Assuming a model (the likelihood $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$) and a prior density $p(\mathbf{z},\boldsymbol{\theta})$ over the unknown variables, in Bayesian inference we desire to study the posterior pdf

$$p(\mathbf{z},\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z},\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})p(\mathbf{z},\boldsymbol{\theta}), \tag{1}$$

where $\boldsymbol{\theta}$ are parameters of the model. We are interested to inference $\mathbf{x} = [\mathbf{z},\boldsymbol{\theta}]^T$ computing the mean/mode of the posterior or confidence intervals, for instance.

Introduction: framework

- In some cases, the parameters $\boldsymbol{\theta}$ can be integrated out

$$p(\mathbf{z}|\mathbf{y}) \propto \int_{\Theta} p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}, \boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{2}$$

  Namely, they could be removed from the analysis, $\mathbf{x} = \mathbf{z}$.

- The amount

$$p(\mathbf{y}) = \int_{\mathcal{X}} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = p(\mathbf{y}|model)$$

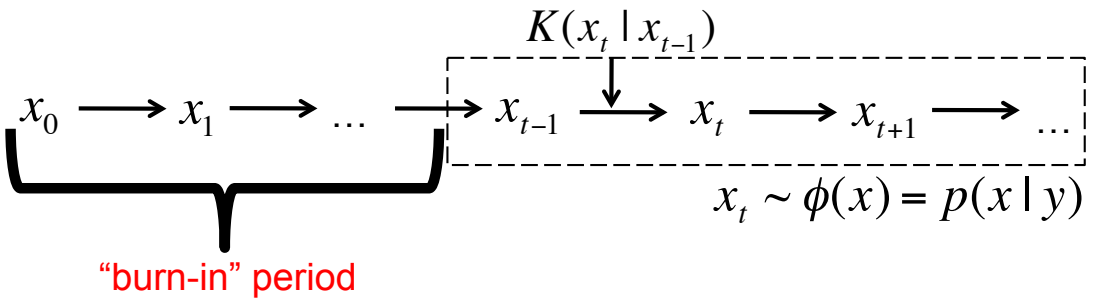
  is called Bayesian evidence and is really important for *model selection*.

Monte Carlo approach

- In general, the posterior pdf $p(\mathbf{x}|\mathbf{y})$ is very complicated and it is impossible to calculate analytically any moments, modes or confidence intervals.
- Monte Carlo (MC) approach: we can draw samples $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ from $\phi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$ in order to approximate mean, variance, probabilities etc. The density $\phi(\mathbf{x})$ is usually called target pdf.
- Problem: in general, we are not able to draw from the target $\phi(x)$.
- MC sampling method: any mechanism/procedure that converts a sample $\boldsymbol{\xi}' \sim \pi(\boldsymbol{\xi})$ (proposal pdf) to a sample $\mathbf{x}'$ distributed according to the target $\phi(\mathbf{x})$.

MCMC

- Within Monte Carlo: Markov Chain Monte Carlo (MCMC) techniques are very powerful methods to produce samples from general target pdf. They generate a Markov chain with *stationary* density our *target* density.
- We denote as $K(\mathbf{x}_t|\mathbf{x}_{t-1})$ the kernel of the chain (probability to obtain a new state $\mathbf{x}_t$ given the previous one, $\mathbf{x}_{t-1}$).



- Design a (standard) MCMC method $\equiv$ Design a kernel $K(\mathbf{x}_t|\mathbf{x}_{t-1})$ such that $\phi(\mathbf{x})$ is the invariant/stationary distribution.
- ($K(\mathbf{x}_t|\mathbf{x}_{t-1})$ summarizes the steps of the corresponding algorithm)

Range of applications

Applications of MCMC:

- *Bayesian inference* (drawing from complicated posterior distributions)
- *stochastic optimizations* (with some little variations in the corresponding sampling methods; for instance, Metropolis-Hastings $\Longleftrightarrow$ Simulated annealing)
- (MCMC/Monte Carlo approach can give some theoretical support to some heuristic optimization methods.)

Advantages of MCMC

- Advantages of the MCMC techniques:
  1. They can be applied to virtually any kind of target pdf (at any dimension $\mathbf{x} \in \mathbb{R}^n$). In general, we just to be able to evaluate the target $p(\mathbf{x}|\mathbf{y})$, that can be known up a normalizing constant.
  2. There are also MCMC techniques to draw samples from target pdfs that cannot be "completely" evaluated! (in this talk you will see an example).

Drawbacks of MCMC

- The drawbacks that we want to avoid/ improve:
  1. The generated samples are correlated. We want to decrease the correlation. Independent samples provide more statistical informations.
  2. Convergence – "burn in" period. We want to speed up the convergence.
  3. Due to the previous two factors, the chain can be trapped in some region "of high probability" (like around a mode). We want to avoid it.
- Actually, all the previous factors are provoked by the correlation.

Invariant/Stationary distribution

- Given a random vector $\mathbf{X}_t \in \mathcal{D} \subseteq \mathbb{R}^m$ and transition probability (*kernel*) function $K(\mathbf{x}_t|\mathbf{x}_{t-1})$, a stationary (invariant) probability density function (pdf) $p_s(\mathbf{x}_t)$ fulfills the following condition

$$\int_{\mathcal{D}} K(\mathbf{x}_t|\mathbf{x}_{t-1}) p_s(\mathbf{x}_{t-1}) d\mathbf{x}_{t-1} = p_s(\mathbf{x}_t). \tag{3}$$

- Note it is the same to build a joint pdf $p(\mathbf{x}_{t-1}, \mathbf{x}_t) = K(\mathbf{x}_t|\mathbf{x}_{t-1}) p_s(\mathbf{x}_{t-1})$ where the marginal pdfs $p_s(\cdot)$ exactly coincide.
- We want to design $K(\mathbf{x}_t|\mathbf{x}_{t-1})$ such that the invariant pdf $p_s(\mathbf{x})$ is exactly the target pdf $\phi(\mathbf{x})$, i.e.,

$$p_s(\mathbf{x}) = \phi(\mathbf{x}) \tag{4}$$

- This problem is related to the search of *eigenvalues* and *eigenfunctions* in the equation

$$\int_{\mathcal{D}} K(\mathbf{y}|\mathbf{x}) q(\mathbf{x}) d\mathbf{x} = \mu q(\mathbf{y}) \tag{5}$$

  where μ is an eigenvalue and $q(\cdot)$ is an eigenfunction (corresponding to μ).
- We are interested in eigenfunctions corresponding to $\mu = 1$ (actually, our problem is different: given a eigenfunction $q(\mathbf{x})$ with eigenvalue $\mu = 1$, we want to find a suitable kernel $K(\mathbf{y}|\mathbf{x})$).
- We have $1 = \mu_1 > |\mu_2| \geq |\mu_3| \geq |\mu_4| \geq \ldots$
- The eigenvalue μ_2 determines the order of the convergence speed to the eigenfunction corresponding to $\mu_1 = 1$.

- The *balance (reversibility)* condition

$$K(\mathbf{x}_t|\mathbf{x}_{t-1})p_s(\mathbf{x}_{t-1}) = K(\mathbf{x}_{t-1}|\mathbf{x}_t)p_s(\mathbf{x}_t) \tag{6}$$

  is a *sufficient* condition for the Markov chain to have a *unique* stationary distribution.

- Indeed, integrating both sides w.r.t. $\mathbf{x}_{t-1}$

$$\int_{\mathcal{D}} K(\mathbf{x}_t|\mathbf{x}_{t-1})p_s(\mathbf{x}_{t-1})d\mathbf{x}_{t-1} = \int_{\mathcal{D}} K(\mathbf{x}_{t-1}|\mathbf{x}_t)p_s(\mathbf{x}_t)d\mathbf{x}_{t-1},$$

$$\int_{\mathcal{D}} K(\mathbf{x}_t|\mathbf{x}_{t-1})p_s(\mathbf{x}_{t-1})d\mathbf{x}_{t-1} = p_s(\mathbf{x}_t)\underbrace{\int_{\mathcal{D}} K(\mathbf{x}_{t-1}|\mathbf{x}_t)d\mathbf{x}_{t-1}}_{1},$$

$$\int_{\mathcal{D}} K(\mathbf{x}_t|\mathbf{x}_{t-1})p_s(\mathbf{x}_{t-1})d\mathbf{x}_{t-1} = p_s(\mathbf{x}_t).$$

- If a pdf satisfies the balance condition, then it is invariant w.r.t. the kernel $K(\mathbf{x}_t|\mathbf{x}_{t-1})$.
- In this case, the chain is said *reversible*.

- Metropolis-Hastings (MH) method

Metropolis-Hastings (MH) algorithm

- The most famous MCMC technique is the Metropolis-Hastings (MH) method.
- Given an *unnormalized* target pdf $\phi(\mathbf{x})$ ($\mathbf{x} \in \mathbb{R}^n$), and a proposal pdf $\pi(\mathbf{x}_t|\mathbf{x}_{t-1})$ (easy to draw from), the algorithm is the following:
  1. For $t = 0$, choose arbitrarily $\mathbf{x}_0$.
  2. Draw $\mathbf{x}^*$ from $\pi(\mathbf{x}_t|\mathbf{x}_{t-1})$.
  3. Accept the sample (movement) $\mathbf{x}_t = \mathbf{x}^*$ with probability

$$\alpha(\mathbf{x}_{t-1}, \mathbf{x}^*) = \min\left[1, \frac{\pi(\mathbf{x}_{t-1}|\mathbf{x}^*)\phi(\mathbf{x}^*)}{\pi(\mathbf{x}^*|\mathbf{x}_{t-1})\phi(\mathbf{x}_{t-1})}\right]. \tag{7}$$

  4. Otherwise, $\mathbf{x}_t = \mathbf{x}_{t-1}$.
  5. Set $t = t + 1$ and come back to step 2.

Random Walk/independent proposal

- There are two *typical* kind/class of the proposal pdf.
- random walk proposal

$$X_t = X_{t-1} + \epsilon$$

for instance,

$$\pi(x_t|x_{t-1}) \propto \exp\{-(x_t - x_{t-1})^2/2\}.$$

(it is "shifted/moved" according to the previous state)

- independent proposal for instance,

$$\pi(x_t|x_{t-1}) = \pi(x_t) \propto \exp\{-x_t^2/2\}.$$

(independent from the previous state, it is "fix")

Symmetric proposal

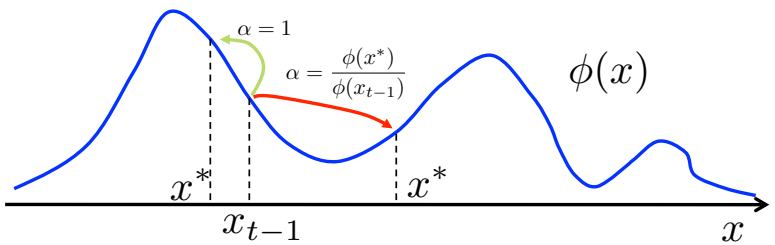
- If we choose a proposal such that

$$\pi(x_t|x_{t-1}) = \pi(x_{t-1}|x_t), \quad \text{(for instance } \propto \exp\{-(x_t - x_{t-1})^2/2\})$$

then the acceptance probability α is simplified. Indeed, in this case,

$$\alpha(\mathbf{x}_{t-1}, \mathbf{x}^*) = \min\left[1, \frac{\phi(\mathbf{x}^*)}{\phi(\mathbf{x}_{t-1})}\right].$$

- This shows a clearly connection with the optimization methods.
  if $\phi(\mathbf{x}^*) \geq \phi(\mathbf{x}_{t-1}) \rightarrow \alpha = 1$,
  if $\phi(\mathbf{x}^*) < \phi(\mathbf{x}_{t-1}) \rightarrow \alpha = \frac{\phi(\mathbf{x}^*)}{\phi(\mathbf{x}_{t-1})}$.



Hence, when we go up the movement is always accept, whereas we go down with probability $\alpha = \frac{\phi(\mathbf{x}^*)}{\phi(\mathbf{x}_{t-1})}$. In the *Simulated Annealing*, this probability vanishes to zero when $t \rightarrow +\infty$.

Trivial case

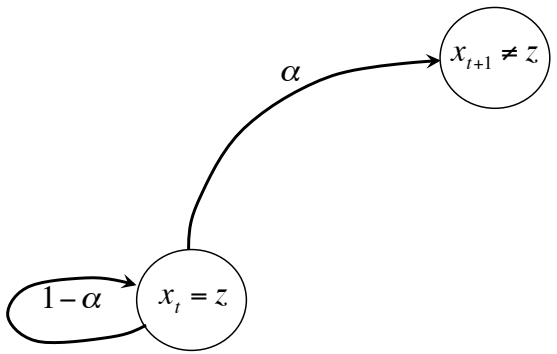
- If $\pi(x) = \phi(x)$ (we draw directly from the target), then

$$\alpha(\mathbf{x}_{t-1}, \mathbf{x}^*) = \min\left[1, \frac{\pi(\mathbf{x}_{t-1})\phi(\mathbf{x}^*)}{\pi(\mathbf{x}^*)\phi(\mathbf{x}_{t-1})}\right] = \min\left[1, \frac{\phi(\mathbf{x}_{t-1})\phi(\mathbf{x}^*)}{\phi(\mathbf{x}^*)\phi(\mathbf{x}_{t-1})}\right] = 1.$$

  clearly, we always accept the movement (this is the best and impossible situation, in general).

- In general, if the proposal is similar/close to the target then α is close to 1....but α can be ≈ 1 also in other cases!! Therefore, $\alpha \approx 1$ is not always a "good situation".

Important consideration



- From the figure, you can think that $\alpha \approx 1$ always diminishes the correlation. But it is not true, in general.
- With a proposal with very small variance $\longrightarrow \alpha \approx 1 \longrightarrow$ very poor performance.
- With a proposal with huge variance $\longrightarrow \alpha \to 0 \longrightarrow$ very poor performance.
- There are several papers that study the optimal value of α with different proposals, targets and situations. For instance, it is suggested $\alpha \approx 0.45$ for low dim. and $\alpha \approx 0.23$ for high dim. problems [3, 1, 2].
- (optimal variance $\Longleftrightarrow$ optimal α)

Kernel of the MH

- The kernel $K(\mathbf{y}|\mathbf{x})$ (probability to go from $\mathbf{x}$ to $\mathbf{y}$) of the MH method is

$$K(\mathbf{y}|\mathbf{x}) = \pi(\mathbf{y}|\mathbf{x})\alpha(\mathbf{x},\mathbf{y}) + \delta(\mathbf{y}-\mathbf{x})\overbrace{\left(1 - \int_{\mathcal{D}} \pi(\mathbf{y}'|\mathbf{x})\alpha(\mathbf{x},\mathbf{y}')d\mathbf{y}'\right)}^{\text{Prob. of discarding a proposed sample } \mathbf{y}'}, \quad (8)$$

  where $\alpha(\mathbf{x},\mathbf{y}) = \min\left[1, \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}\right]$.

- The target pdf $\phi(\mathbf{x})$ is invariant w.r.t. this kernel.
- (just notation: observe that $\mathbf{y}$ here denotes a possible state, it is not a observation/measurement!)

- Indeed, the kernel $K(\mathbf{y}|\mathbf{x})$ of MH fulfills the balance condition

$$K(\mathbf{y}|\mathbf{x})\phi(\mathbf{x}) = K(\mathbf{x}|\mathbf{y})\phi(\mathbf{y}) \tag{9}$$

- For $\mathbf{x} = \mathbf{y}$ is trivial, we obtain $\phi(\mathbf{x}) = \phi(\mathbf{y})$.
- For $\mathbf{x} \neq \mathbf{y}$, we have to verify that

$$\pi(\mathbf{y}|\mathbf{x})\alpha(\mathbf{x},\mathbf{y})\phi(\mathbf{x}) = \pi(\mathbf{x}|\mathbf{y})\alpha(\mathbf{y},\mathbf{x})p(\mathbf{y}), \tag{10}$$

$$\pi(\mathbf{y}|\mathbf{x}) \min\left[1, \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}\right]\phi(\mathbf{x}) = \pi(\mathbf{x}|\mathbf{y}) \min\left[1, \frac{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}\right]p(\mathbf{y}), \tag{11}$$

- hence finally we obtain

$$\min\left[\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x}), \pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})\right] = \min\left[\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y}), \pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})\right], \tag{12}$$

that, since the function $\min[\cdot,\cdot]$ is symmetric, Eq. (9) is true !!! □

Other possible $\alpha(\mathbf{x}, \mathbf{y})$

- The acceptance function $\alpha(\mathbf{x}, \mathbf{y}) = \min\left[1, \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}\right]$, is not the only possible choice to satisfy the balance condition.
- Other possible functions $\alpha(\mathbf{x}, \mathbf{y}) : \mathcal{D} \times \mathcal{D} \to [0, 1]$ are, for instance,

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{\lambda(\mathbf{x}, \mathbf{y})}{1 + \frac{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}} \qquad \text{(Hastings gen.; } \lambda(\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{y}, \mathbf{x})\text{)}.$$

With $\lambda(\mathbf{x}, \mathbf{y}) = 1$, we have

$$\alpha(\mathbf{x}, \mathbf{y}) = \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y}) + \pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})} \qquad \text{(Barker)}$$

- and other examples

$$\begin{aligned} \alpha(\mathbf{x}, \mathbf{y}) &= \frac{\lambda(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})} \qquad \text{(Stein–1; } \lambda(\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{y}, \mathbf{x})\text{)} \\ \alpha(\mathbf{x}, \mathbf{y}) &= \frac{\phi(\mathbf{y})\lambda(\mathbf{x}, \mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})} \qquad \text{(Stein–2; } \lambda(\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{y}, \mathbf{x})\text{)} \end{aligned} \tag{13}$$

with $\lambda(\mathbf{x}, \mathbf{y})$ such that $\alpha(\mathbf{x}, \mathbf{y}) : \mathcal{D} \times \mathcal{D} \rightarrow [0, 1]$.

- Moreover, given a function

$$r(\mathbf{x}, \mathbf{y}) \triangleq \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})},$$

  and a function $F(z)$ such that

$$F(z) = zF(1/z),$$

- it is possible to build other suitable acceptance functions in this way

$$\alpha(\mathbf{x}, \mathbf{y}) \triangleq F \circ r(\mathbf{x}, \mathbf{y}) = F\left(r(\mathbf{x}, \mathbf{y})\right). \tag{14}$$

- When $F(z) = \min[1, z]$ we obtain the standard acceptance function. With $F(z) = \frac{1}{1+\frac{1}{z}} = \frac{z}{1+z}$ we obtain the Barker acceptance function.

- It is possible to show, for all α that generates a reversible MH kernel,

$$\alpha(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{y}, \mathbf{x}) r(\mathbf{x}, \mathbf{y}). \tag{15}$$

- For instance for the Hastings' class, since $\alpha(\mathbf{x}, \mathbf{y}) = \frac{\lambda(\mathbf{x},\mathbf{y})}{1+\frac{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}} = \frac{\lambda(\mathbf{x},\mathbf{y})}{1+r(\mathbf{y},\mathbf{x})}$, $\lambda(\mathbf{x}, \mathbf{y}) = \lambda(\mathbf{y}, \mathbf{x})$ and $r(\mathbf{x}, \mathbf{y}) = \frac{1}{r(\mathbf{y},\mathbf{x})}$, then

$$\alpha(\mathbf{y}, \mathbf{x}) r(\mathbf{x}, \mathbf{y}) = \frac{\lambda(\mathbf{y}, \mathbf{x}) r(\mathbf{x}, \mathbf{y})}{1 + r(\mathbf{x}, \mathbf{y})} = \frac{\lambda(\mathbf{x}, \mathbf{y}) \frac{1}{r(\mathbf{y},\mathbf{x})}}{1 + \frac{1}{r(\mathbf{y},\mathbf{x})}} = \frac{\lambda(\mathbf{x}, \mathbf{y})}{1 + r(\mathbf{y}, \mathbf{x})} = \alpha(\mathbf{x}, \mathbf{y}).$$

- (Peskun (1973), Tierney (1998)) The best α for the MH is $\alpha_{MH}(\mathbf{x}, \mathbf{y}) = \min\left[1, \frac{\pi(\mathbf{x}|\mathbf{y})\phi(\mathbf{y})}{\pi(\mathbf{y}|\mathbf{x})\phi(\mathbf{x})}\right]$. Indeed,

$$\alpha(\mathbf{x}, \mathbf{y}) = \alpha(\mathbf{y}, \mathbf{x}) r(\mathbf{x}, \mathbf{y}) \leq \min\left[1, r(\mathbf{x}, \mathbf{y})\right] = \alpha_{MH}(\mathbf{x}, \mathbf{y}). \tag{16}$$

- (Basic idea) with the same proposal and target, with α_{MH} we have: more jumps, less correlation, less variance in the estimation.

Computational troubles

- How long is the "Burn in" period? when does the chain converge?
- Another problem is to determine the total sample size or run length required for accurate estimates.
- There are several works about these issues [2, 3].
- However, advanced MCMC samplers improve the estimation and reduce the "burn-in" period.

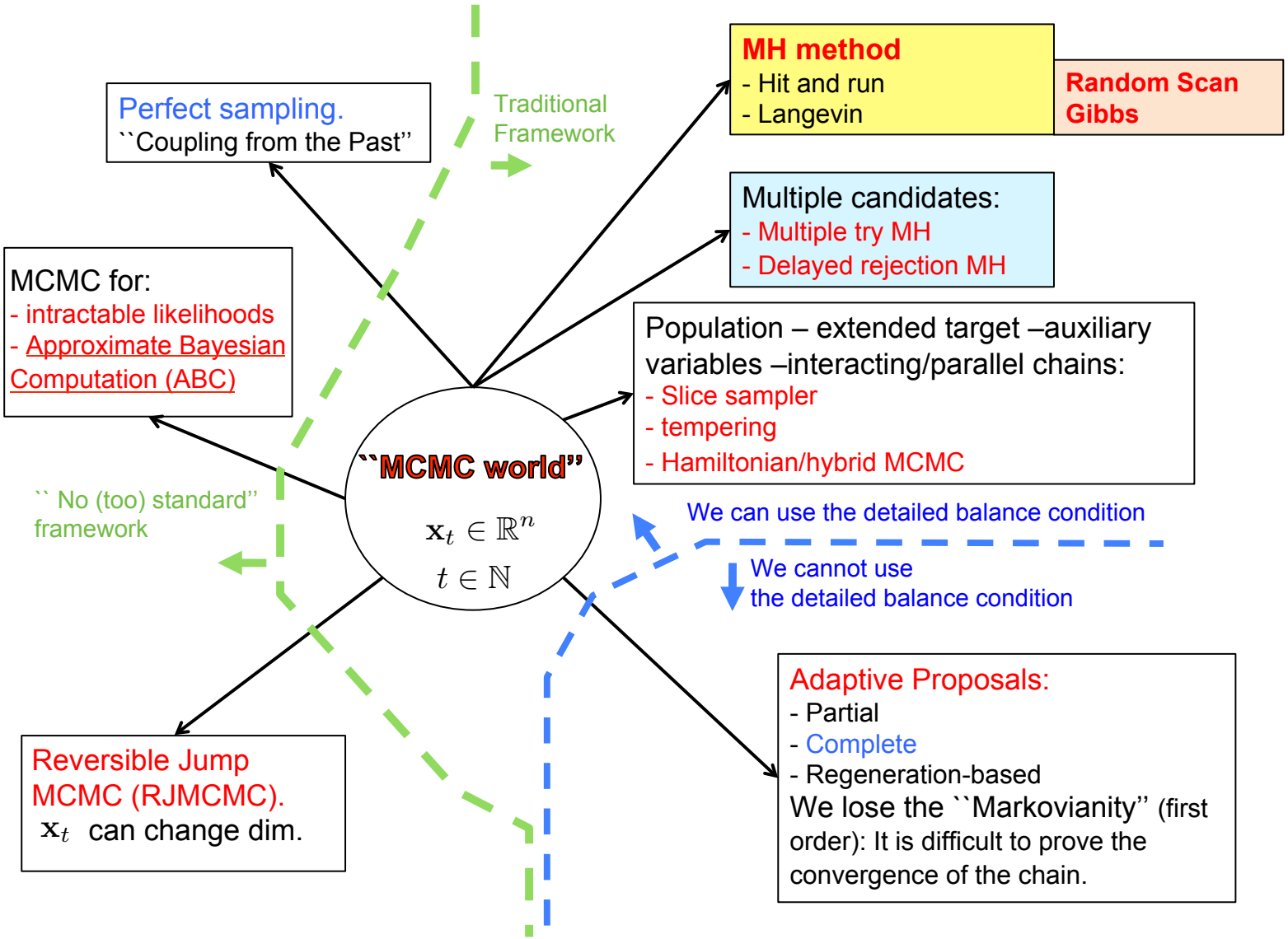
Important and general considerations

- Speed up the convergence (reduce the "burn-in" period) $\Longrightarrow$ reduce the correlation (The eigenvalue μ_2 depends on the correlation function)
- Improve the estimation $\Longrightarrow$ reduce the correlation
  (better with independent samples)
- There are different strategies to reduce the correlation $\Longrightarrow$ advanced MCMC techniques.
- However, the most important idea (in my opinion) is
  to reduce the correlation $\Longrightarrow$ diminish the discrepancy in the shape between proposal and target. Choose (or build) a good proposal.
- We desire to stay as close as possible to the "independent samples" case.
- We also desire to design *black-box* algorithms: the parameters of the MCMC method are adaptively tuned, for different targets.

how improve the MH ?

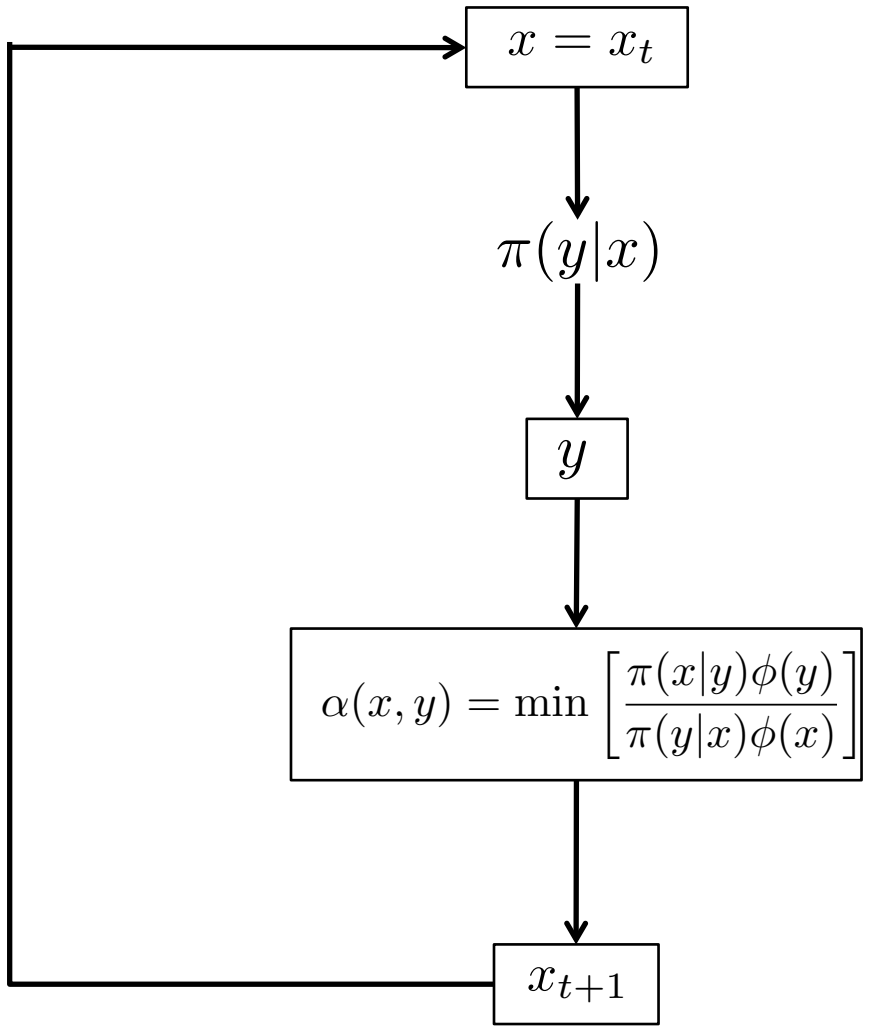
Components in MH that can be changed/improved:

- proposal $\Longrightarrow$ specific and sophisticated choices, adaptive proposals (that change with t)
- $\alpha \Longrightarrow$ different schemes, for instance, with multiple candidates (Multiple Try Metropolis).
- target $\Longrightarrow$ extended target in higher dim., data augmentation, auxiliary variables etc..

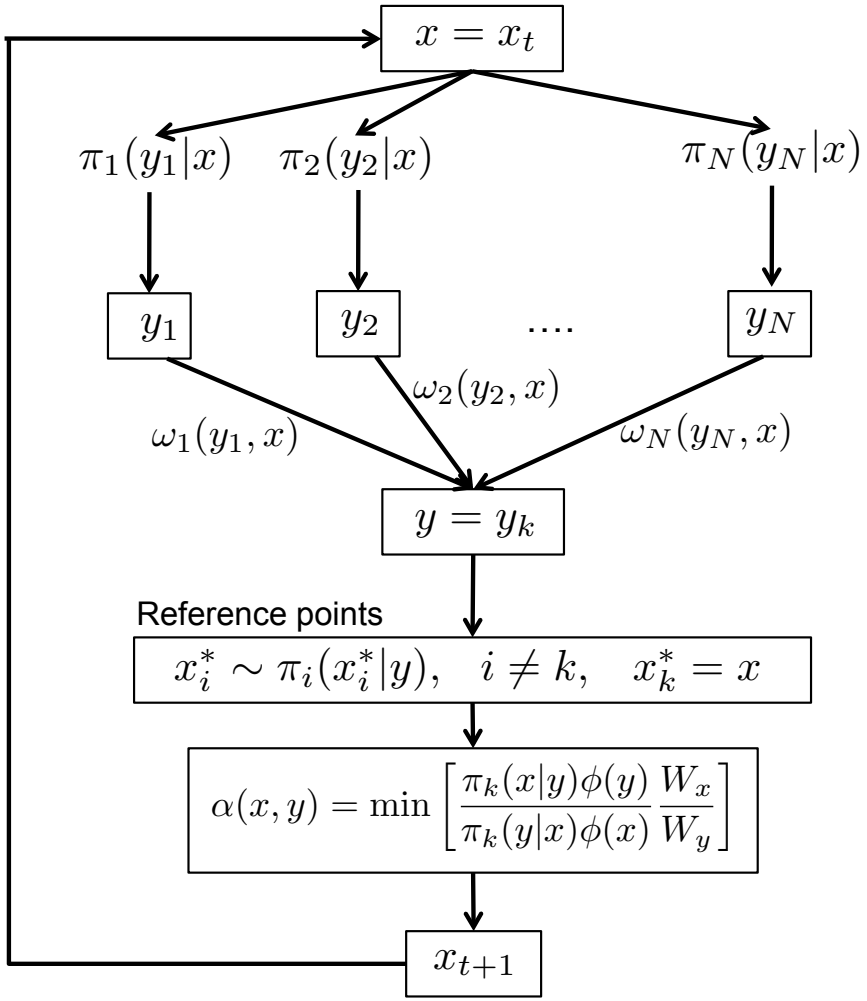

Perfect sampling.
``Coupling from the Past''
Traditional Framework
MH method
- Hit and run
- Langevin
Random Scan Gibbs
Multiple candidates:
- Multiple try MH
- Delayed rejection MH
MCMC for:
- intractable likelihoods
- Approximate Bayesian Computation (ABC)
Population – extended target –auxiliary variables –interacting/parallel chains:
- Slice sampler
- tempering
- Hamiltonian/hybrid MCMC
``MCMC world''
$\mathbf{x}_t \in \mathbb{R}^n$
$t \in \mathbb{N}$
`` No (too) standard'' framework
We can use the detailed balance condition
We cannot use the detailed balance condition
Reversible Jump MCMC (RJMCMC).
$\mathbf{x}_t$ can change dim.
Adaptive Proposals:
- Partial
- Complete
- Regeneration-based
We lose the ``Markovianity'' (first order): It is difficult to prove the convergence of the chain.

- Multiple Try Metropolis (MTM) methods

MH scheme.


$x = x_t$
$\pi(y|x)$
y
$\alpha(x, y) = \min \left[\frac{\pi(x|y)\phi(y)}{\pi(y|x)\phi(x)} \right]$
x_{t+1}

General MTM scheme, with generic weight functions and different proposals.



1. Draw N samples $y_1, y_2 \ldots, y_N$ from $y_i \sim \pi_i(y|x_{t-1})$.
2. Calculate some (bounded and positive) weights $\omega_i(y_i, x_{t-1})$.
3. Choose a sample $y_k \in \{y_1, \ldots, y_N\}$ according to the ω_i, and set

$$W_y = \frac{\omega_k(y_k, x_{t-1})}{\sum_{i=1}^{N} \omega_i(y_i, x_{t-1})}.$$

4. Draw $N-1$ reference samples $x_i^* \sim \pi_i(x|y_k), i \neq k$, and set $x_k^* = x_{t-1}$.
5. Set

$$W_x = \frac{\omega_k(x_{t-1}, y_k)}{\sum_{i=1}^{N} \omega_i(x_i^*, y_k)}.$$

6. Accept $x_t = y_k$ with probability

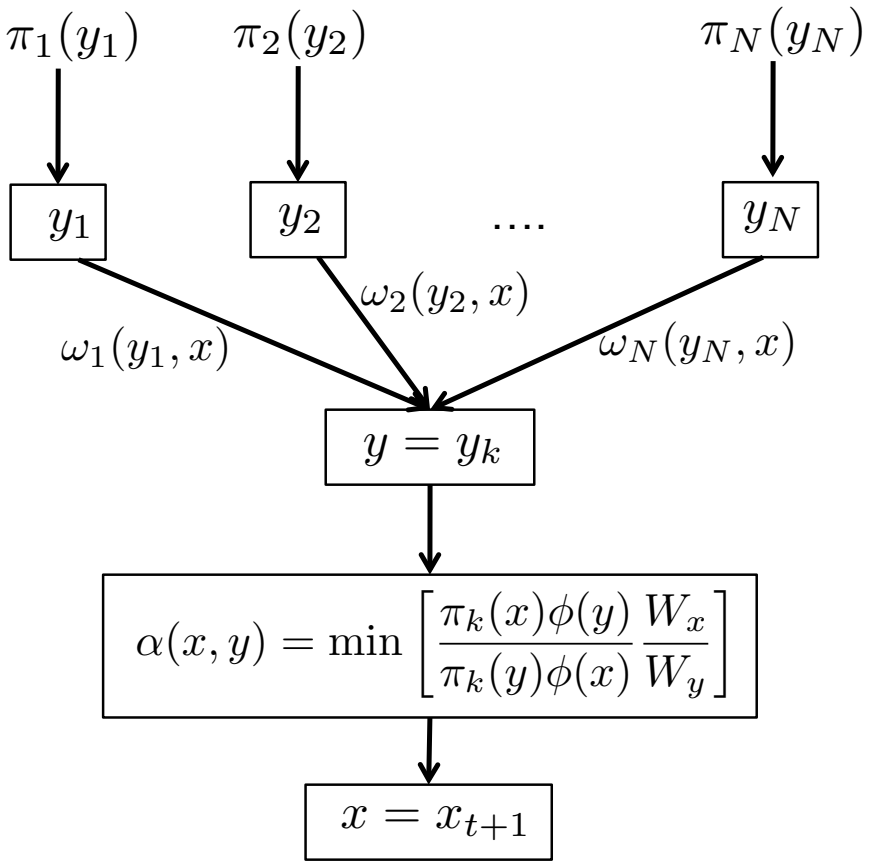
$$\alpha(x_{t-1}, y_k) = \min\left[1, \frac{\pi_k(x_{t-1}|y_k)\phi(y_k)}{\pi_k(y_k|x_{t-1})\phi(x_{t-1})} \frac{W_x}{W_y}\right]. \tag{17}$$

otherwise, set $x_t = x_{t-1}$.

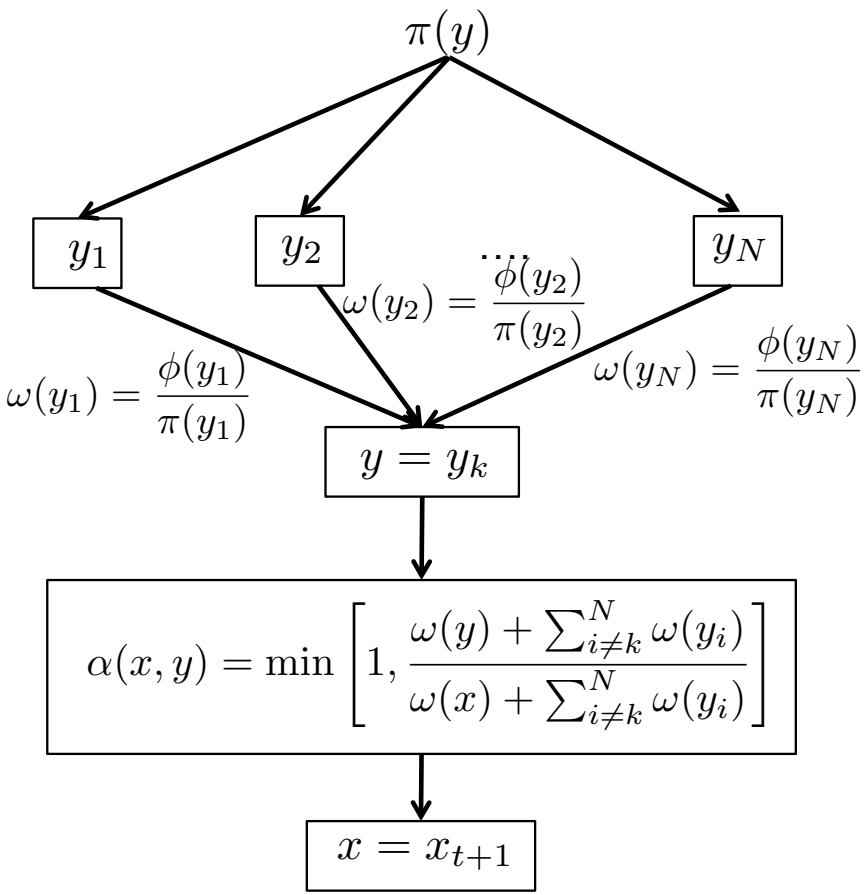
- You can find: the study of the state of art, the most relevant references, proofs, possible choices of the weights, comparisons and further considerations in

  L. Martino, J. Read, "On the flexibility of the design of Multiple Try Metropolis schemes", (submitted to Computational Statistics), arXiv:1201.0646, 2012.

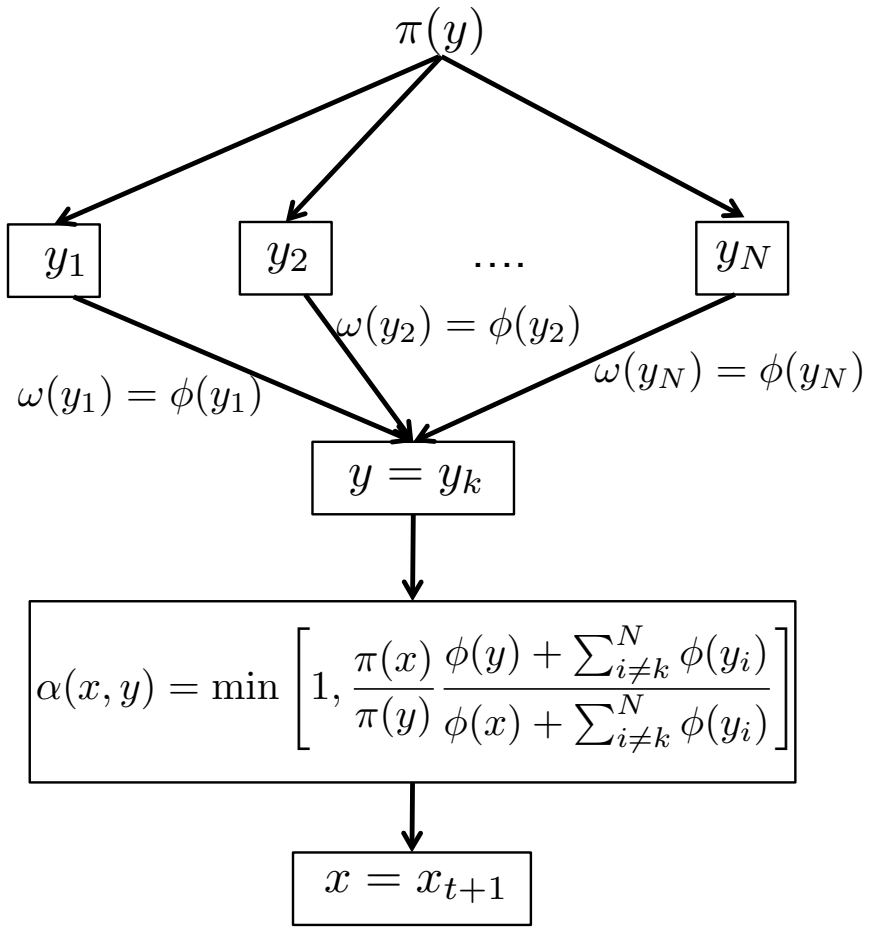
MTM scheme with generic weights and different independent proposals

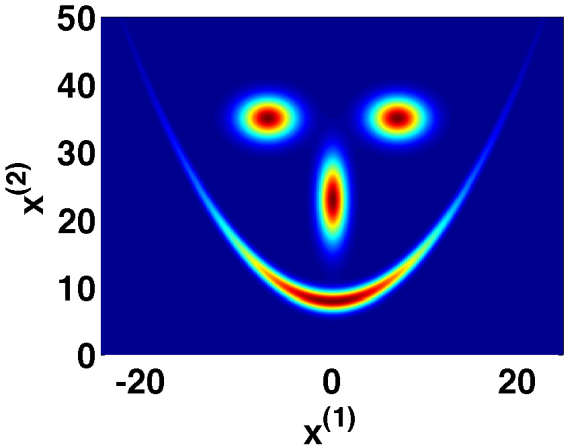


The simplest MTM schemes: one independent proposal and importance weights

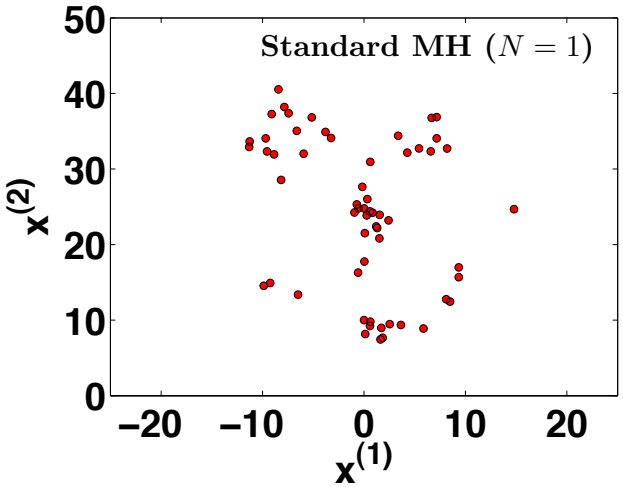


The simplest MTM schemes: one independent proposal and weights $\propto$ to the target

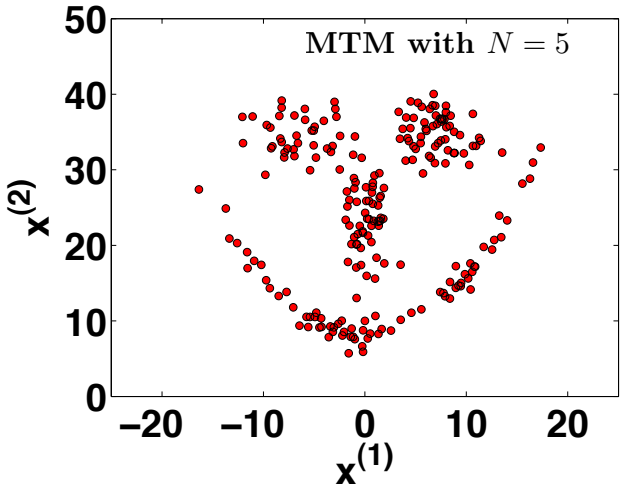



50
40
30
20
10
0
-20
0
20
x(2)
x(1)

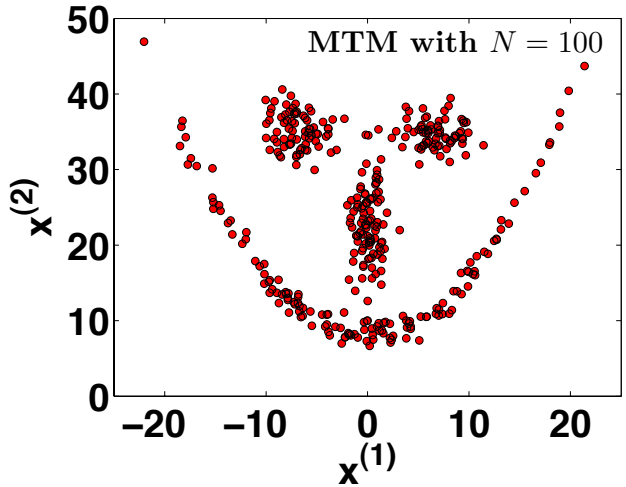
(a)


Standard MH (N = 1)
50
40
30
20
10
0
−20
−10
0
10
20
x(2)
x(1)

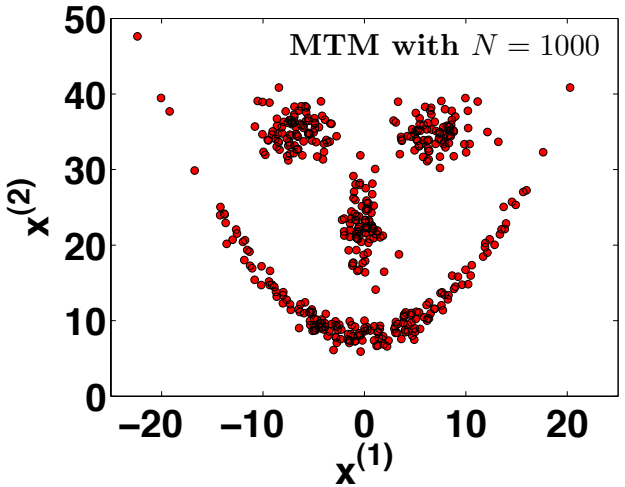
(b)


MTM with N = 5
50
40
30
20
10
0
−20
−10
0
10
20
x(2)
x(1)

(c)


MTM with N = 100
50
40
30
20
10
0
−20
−10
0
10
20
x(2)
x(1)

(d)


MTM with N = 1000
50
40
30
20
10
0
−20
−10
0
10
20
x(2)
x(1)

(e)

Important feature

- In a suitable MTM scheme (e.g., good choice weights, proposal "no too ugly" etc.), the probability of accepting the new state approaches 1, $\alpha \longrightarrow 1$, when the number of candidates grows, $N \longrightarrow +\infty$.
- In this case, $\alpha \approx 1$ it is a good news. The performance is not poor (is excellent !!), since we are comparing several (with a huge N) candidates that can be arbitrarily far or close to the current state, and then choosing the best one.
- Clearly, when $N \longrightarrow +\infty$ we are increasing the computational cost.

- Approximate Bayesian Computation (ABC)

Approximate Bayesian Computation (ABC)

- We desire to draw samples from the posterior pdf

$$\phi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}') \propto \underbrace{p(\mathbf{y}'|\mathbf{x})p(\mathbf{x})}_{p(\mathbf{x},\mathbf{y})}, \tag{18}$$

  where $\mathbf{y}'$ is a given vector of observations.

- If we can evaluate both prior, $p(\mathbf{x})$, and likelihood, $p(\mathbf{y}|\mathbf{x})$, we can apply standard MCMC techniques (we does not need to know the normalization constant, $p(\mathbf{y}')$ - the evidence).

- Assume that we can evaluate the prior $p(\mathbf{x})$ (clearly) but we do not know nothing about the analytic form of the likelihood (or it is extremely complicated to evaluate the likelihood). Assuming also that we can just simulate from the model, i.e., we can draw "observations"

$$\mathbf{y}^{(i)} \sim p(\mathbf{y}|\mathbf{x}), \quad i = 1, 2, 3.... \tag{19}$$

- *How can we draw from $\phi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y})$? this is a very "hot" problem... this is Approximate Bayesian Computation (ABC) problem.*

Main applications

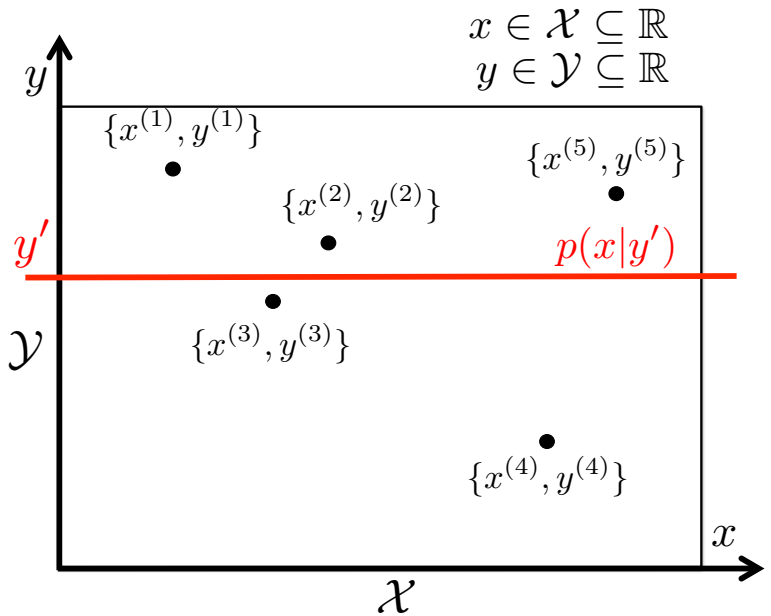
- Molecular population genetic - evolutionary biology, where the likelihood function is not available in a closed form. In this kind of application, for instance, they can "*draw samples (molecular variation) obtaining a discrete variation of the data set. Inference and estimation for population parameters of interest such as mutation rates, recombination rates, migration rates, and demographic parameters are then based on a stochastic model (denoting the "likelihood").*"
- ABC techniques could be also used to find maximum likelihood, without assuming a specific model (a likelihood), in applications where "artificial", "fictitious" observations can be generated (something like a "training" sequence of observations/data....to give an idea).
- Jorge Plata's Example: $\mathbf{y} = f(\mathbf{x}) + \mathbf{r}$, where $\mathbf{r} \sim R_1 + ... + R_N$ and each $R_i \sim p_i(\mathbf{r})$ is easy to draw from, but we do not know analytically the probability of the sum.

The ABC problem

- Clearly, one possible approach is a "two-steps" procedure **(1)** to draw several samples $\mathbf{y}^{(i)} \sim p(\mathbf{y}|\mathbf{x})$, for different $\mathbf{x}$, and approximate the likelihood function using these samples, obtaining $\hat{p}(\mathbf{y}|\mathbf{x}) \approx p(\mathbf{y}|\mathbf{x})$. **(2)** Then, we could use a standard MCMC to draw from $\hat{\phi}(\mathbf{x}) = \hat{p}(\mathbf{x}|\mathbf{y}') \propto \hat{p}(\mathbf{y}'|\mathbf{x})p(\mathbf{x})$.
- Note that, in general we have $\mathbf{y} \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^n$, we have to approximate a function of $m + n$ variables. It is complicated... however it is a possible strategy.
- Another possibility is to design "ad-hoc" Monte Carlo methods to obtain ("approximately") samples from $\phi(\mathbf{x})$ directly while we draw samples $\mathbf{y}^{(i)}$ from the model, without dividing the procedure in two steps.

The ABC problem

- Since the previous assumptions, we can easily draw samples from the joint pdf $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \sim p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$, $i = 1, ..., N$. Indeed, we can first draw $\mathbf{x}^{(i)} \sim p(\mathbf{x})$ and then $\mathbf{y}^{(i)} \sim p(\mathbf{y}|\mathbf{x}^{(i)})$.
- [Note that if $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \sim p(\mathbf{x}, \mathbf{y})$, $\mathbf{x}^{(i)} \sim p(\mathbf{x})$ and $\mathbf{y}^{(i)} \sim g(\mathbf{y})$, where $g(\mathbf{y}) = \int p(\mathbf{x}, \mathbf{y}) d\mathbf{x}$ is the other marginal pdf.]



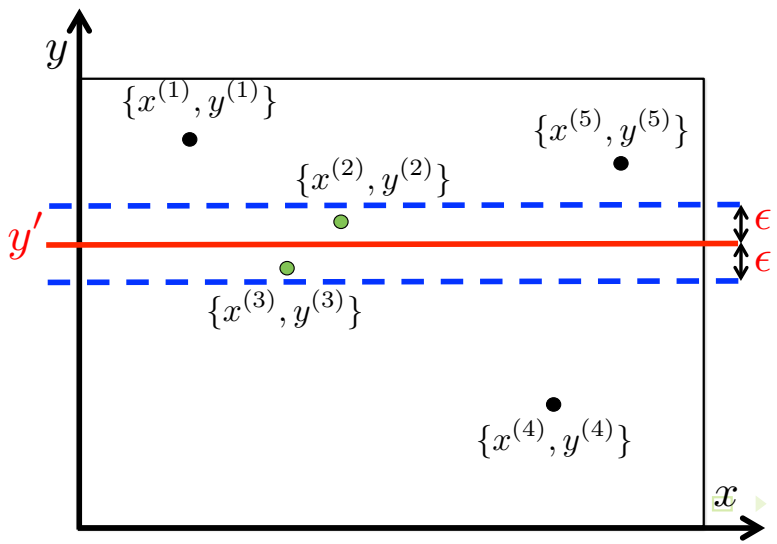
- Clearly, if $\mathbf{y} \in \mathcal{Y} \subseteq \mathbb{R}^m$ (continuous variable), $\text{Prob}\{\mathbf{y}^{(i)} = \mathbf{y}'\} = 0$. If $\mathbf{y}$ would be a discrete variable we could easily use a rejection scheme....

ABC - rejection

1. Draw $\mathbf{x}^* \sim p(\mathbf{x})$ (from the prior).
2. Draw $\mathbf{y}^* \sim p(\mathbf{y}|\mathbf{x}^*)$ (from the model).
3. Accept $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) = (\mathbf{x}^*, \mathbf{y}^*)$ if $\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon$, with $\epsilon > 0$ (for instance $\rho(\mathbf{y}^*, \mathbf{y}') = |\mathbf{y}^* - \mathbf{y}'|$), and set $i = i + 1$. Otherwise, reject $(\mathbf{x}^*, \mathbf{y}^*)$.
4. Repeat until obtaining the desired number of samples.

The generated samples are distributed as $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)}) \sim p(\mathbf{x}|\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon)$, for this reason it is *Approximate* Bayesian Computation. Clearly,

$$\lim_{\epsilon \to 0} p(\mathbf{x}|\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon) = p(\mathbf{x}|\mathbf{y}'), \quad \left(\lim_{\epsilon \to +\infty} p(\mathbf{x}|\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon) = p(\mathbf{x}) \right).$$



ABC - rejection

- Drawbacks: the acceptance rate can be very low. For instance, it occurs when there is significant discrepancy between the prior and likelihood.
- The best Acceptance Rate: when the prior pdf is

$$p(\mathbf{x}) = \delta(\mathbf{x} - \hat{\mathbf{x}}_M),$$

where

$$\hat{\mathbf{x}}_M \triangleq \arg\max p(\mathbf{y}'|\mathbf{x}).$$

since, in this case, we have the *highest probability* to draw a sample $\mathbf{y}^*$ (always from $p(\mathbf{y}|\hat{\mathbf{x}}_M)$) close to the given observation $\mathbf{y}'$.

- However, **(1)** in general, we do not know $\hat{\mathbf{x}}_M$, **(2)** we do not want to change the prior (we also change the posterior!! in this case, $p(\hat{\mathbf{x}}_M|\mathbf{y})$ is a delta!!) and, **(3)** even if we use this prior, the acceptance rate can be small (depending on the variance of the pdf $p(\mathbf{y}|\hat{\mathbf{x}}_M)$).
- Then, we try to overcome at least the first two points above (without varying the prior).

MCMC-ABC (Metropolis-Hastings ABC)

1. For $t = 0$, choose arbitrarily $\mathbf{x}_0$.
2. Draw $\mathbf{x}^* \sim \pi(\mathbf{x}_t|\mathbf{x}_{t-1})$ (from a generic proposal) and $\mathbf{y}^* \sim p(\mathbf{y}|\mathbf{x}^*)$ (from the model).
3. if $\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon$:
   Accept the sample (movement) $\mathbf{x}_t = \mathbf{x}^*$ with probability

$$\alpha(\mathbf{x}_{t-1}, \mathbf{x}^*) = \min\left[1, \frac{\pi(\mathbf{x}_{t-1}|\mathbf{x}^*)p(\mathbf{x}^*)}{\pi(\mathbf{x}^*|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}\right], \tag{20}$$

   otherwise, $\mathbf{x}_t = \mathbf{x}_{t-1}$ (with prob. $1 - \alpha$).
4. if $\rho(\mathbf{y}^*, \mathbf{y}') > \epsilon$: $\mathbf{x}_t = \mathbf{x}_{t-1}$.
5. Set $t = t + 1$ and come back to step 2. [Also in this case, the invariant distribution is $p(\mathbf{x}|\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon)$.]

Note that the prior pdf $p(\mathbf{x})$ plays the role of the "target" in $\alpha(\mathbf{x}_{t-1}, \mathbf{x}^*)$!!! Indeed, since $\mathbf{x}^*$ is drawn from a generic proposal (not from the prior), we need a "mechanism" to undo this effect and finally draw from the *approximate* true posterior pdf.

MCMC-ABC: advantages and balance condition

- Advantages: a suitable choice of the $\pi(\mathbf{x}_t|\mathbf{x}_{t-1})$ (or $\pi(\mathbf{x}_t)$) can increase the acceptance rate in $\rho(\mathbf{y}^*, \mathbf{y}') \leq \epsilon$. The proposal can have a less discrepancy with the likelihood function (w.r.t. the prior).
- The price to pay is that we obtained correlated samples.
- Let me recall the detailed balance condition

$$K(\mathbf{x}_t|\mathbf{x}_{t-1})\phi(\mathbf{x}_{t-1}) = K(\mathbf{x}_{t-1}|\mathbf{x}_t)\phi(\mathbf{x}_t). \tag{21}$$

(it is sufficient condition for $\phi(\mathbf{x})$ to be stationary/invariant w.r.t. the kernel $K(\mathbf{x}_t|\mathbf{x}_{t-1})$)

MH-ABC: proof (maybe *partial*)

- Now we assume that the $\mathbf{y}$ is discrete variable (then we can use $\epsilon = 0$).
- Hence, the kernel (for the case $\mathbf{x}_t \neq \mathbf{x}_{t-1}$) of the method is the following (when $\mathbf{y}$ takes values in a discrete space)

$$K(\mathbf{x}_t|\mathbf{x}_{t-1}) = \underbrace{\pi(\mathbf{x}_t|\mathbf{x}_{t-1})}_{\text{(1) propose;}} \quad \underbrace{p(\mathbf{y}'|\mathbf{x}_t)}_{\text{(2) accept } (\epsilon = 0);} \quad \underbrace{\alpha(\mathbf{x}_{t-1}, \mathbf{x}_t)}_{\text{(3) accept, MH-}\alpha;} , \qquad \text{when } \mathbf{x}_t \neq \mathbf{x}_{t-1}.$$

- Since our target is $\phi(\mathbf{x}) = p(\mathbf{x}|\mathbf{y}')$, we can write

$$K(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}') = \pi(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}'|\mathbf{x}_t)\alpha(\mathbf{x}_{t-1}, \mathbf{x}_t) \cdot \frac{p(\mathbf{y}'|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\mathbf{y}')},$$

$$= \pi(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{y}'|\mathbf{x}_t) \min\left[1, \frac{\pi(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)}{\pi(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}\right] \cdot \frac{p(\mathbf{y}'|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1})}{p(\mathbf{y}')},$$

$$= \frac{p(\mathbf{y}'|\mathbf{x}_t)p(\mathbf{y}'|\mathbf{x}_{t-1})}{p(\mathbf{y}')} \min\left[\pi(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}), \pi(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)\right].$$

MH-ABC: proof (maybe *partial*)

- We have obtained

$$K(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}') = \frac{p(\mathbf{y}'|\mathbf{x}_t)p(\mathbf{y}'|\mathbf{x}_{t-1})}{p(\mathbf{y}')} \min\left[\pi(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}), \pi(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t)\right],$$

  and note that replacing $\mathbf{x}_t$ with $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t-1}$ with $\mathbf{x}_t$, the expression remains the same!! therefore we can write

$$K(\mathbf{x}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\mathbf{y}') = K(\mathbf{x}_{t-1}|\mathbf{x}_t)p(\mathbf{x}_t|\mathbf{y}'),$$

  that is the detailed balance condition!! :)

- Hence $p(\mathbf{x}_{t-1}|\mathbf{y}')$ is the stationary distribution w.r.t. kernel $K(\mathbf{x}_t|\mathbf{x}_{t-1})$ (when $\mathbf{y}$ is a discrete variable, so that we can use $\epsilon = 0$).

- Thank you very much!
- Any questions?

[1] D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 1997.

[2] F. Liang, C. Liu, and R. Caroll. *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*. Wiley Series in Computational Statistics, England, 2010.

[3] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer, 2004.