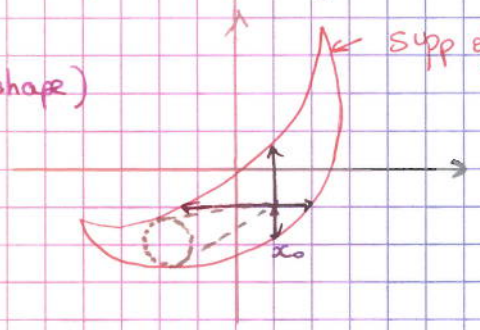
Cours 3
M2NVA

The Gibbs Sampler:

The idea here is to create a sampler which is not a rejection type one and most particularly which will substitute the simulation of our r.v $x \in \mathbb{R}^d$ to sampling in smaller dimension.

The intuition of this is easy to understand in $\mathbb{R}^2$. Consider the following



→ small angle in $\mathbb{R}^2$ which will not make the sample be rejected with a HM sampler for eg.

→ larger possibilities in 1D and iterate.

1) Algorithm:

Assumptions: • $\mathbb{X}$ can be decomposed into $\mathbb{X}_1 \dots \mathbb{X}_k$ so that $\forall x \in \mathbb{X}$, $\exists (x^1, \dots, x^k)$ in $\mathbb{X}_1 \times \dots \mathbb{X}_k$ such that $x = (x^1, \dots, x^k)$. When $k = d$, each $x^i \in \mathbb{R}$.

We denote x^i the element of the i^{th} bloc and $x^{(-i)}$ le vector x where we have removed the i^{th} bloc.

• We know how to simulate from the conditional distributions of the bloc i given the other when targeting π. That is to say that $X \sim \pi$ is not tractable directly BUT: $\forall i \quad X^i \sim \pi_i(x^i \mid x^{(-i)}) = \dfrac{\pi(x)}{\pi_{(-i)}(x)}$

is manageable directly. where $\pi_{(-i)}(x)$ is the marginal distribution

We will see examples in the following where this clearly appears.

Pseudo code: X_0 given ; $n = 0$
then iterate:
for $j = 1 : k$

$$X_{n+1}^j \sim \pi_j^i(\cdot \mid X_{n+1}^1, \dots X_{n+1}^{j-1}, X_n^{j+1}, \dots X_n^k)$$

end ; $n = n+1$.
end

2) Transition kernel and invariant measure:

We will focus on measure which are absolutely continuous w.r.t. the Lebesgue measure and will still denote π the associated pdf.

the associated kernel is the composition of k kernels, each for one bloc. This writes

$$P(x,y) = \prod_{i=1}^{k} \pi_i(y_i \mid y_1, \dots y_{i-1}, x_{i+1}, \dots, x_k)$$

Now of course we expect that π will be the invariant measure. And this holds.

Prop: $\pi P = \pi$

Proof: For sake of simplicity and clarity, we only consider the case where $k=2$, although the proof generalises to higher k.

Let $B \subset \mathcal{X}$

$$\int_{\mathcal{X}\times B} \pi(x) P(x,y)\, dx\, dy = \int_{\mathcal{X}\times B} \pi(x_1,x_2)\, \pi_1(y_1 \mid x_2)\, \pi_2(y_2 \mid y_1)\, dx\, dy$$

$$= \int_{\mathcal{X}\times B} \pi(x_1,x_2) \frac{\pi(y_1,x_2)}{\int \pi(z,x_2)dz} \frac{\pi(y_2,y_1)}{\int \pi(y_1,z)dz}\, dx\, dy$$

$$= \int_{\mathcal{X}\times B} \frac{\pi(x_1,x_2)}{\int \pi(z,x_2)dz} \frac{\pi(y_1,x_2)}{\int \pi(y_1,z)dz} \pi(y_1,y_2)\, dx\, dy$$

$$= \int_{\mathcal{X}\times B} \pi_1(x_1 \mid x_2)\, \pi_2(x_2 \mid y_1)\, \pi(y)\, dx\, dy$$

$$= \int_B \pi(y) \left[\int_{\mathcal{X}} \pi_1(x_1 \mid x_2)\, \pi_2(x_2 \mid y_1)\, dx \right] dy$$

$$= \int_B \pi(y) \underbrace{\left[\int_{\mathcal{X}_1} \underbrace{\left\{ \int_{\mathcal{X}_2} \pi_1(x_1 \mid x_2)\, dx_1 \right\}}_{=1} \pi(x_2 \mid y_1)\, dx_2 \right]}_{=1} dy$$

$$= \int_B \pi(y)\, dy = \pi(B)$$

⚠ This kernel is not reversible! By construction, it is easy to see that the order the blocs are updated is important!

3) Examples:

a) Hierarchical models:

Let $(Y_j)_{1\le j\le n}$ iid $\sim \mathcal{N}(\mu, \sigma^2)$. Let $\delta = 1/\sigma^2$. $(Y_j)_j$ are n observations and the goal is to estimate μ and δ given these observations.

In order to do this we will introduce a prior on the parameters (μ, τ) so that we work in the Bayesian framework.

For example : $p(\mu, \tau) = \tau^{-1/2} \mathbb{1}_{\{\tau > 0\}}$

this prior is non informative on μ [it is a degenerated uniform distribution on $\mathbb{R}$] and forces τ to be positive and behave like $\frac{1}{\sqrt{\tau}}$ in probability

Rq: these priors are not probability measures as they do not sum to 1

this seems a problem at first sight, however, as long as the posterior distribution $q(\tau, \mu | y)$ is well defined you are allowed to introduce degenerated priors.

Our target distribution is here $q(\tau, \mu | (y_i)) = \pi(\mu, \tau)$. This writes:

$$\pi(\mu, \tau) \propto \tau^{n/2} \exp\left(-\frac{1}{2} \tau \sum_{i=1}^{n} (y_i - \mu)^2 \right) \tau^{-1/2} \mathbb{1}_{\{\tau > 0\}}$$

this probability measure is quite hard to sample from!

But looking at $\pi(\mu | \tau)$ we have:

$$\pi(\mu | \tau) \propto \exp\left(-\frac{1}{2} \tau n \mu^2 + \mu \tau n \left[\frac{1}{n} \sum_{j=1}^{n} y_j \right] \right) \quad \text{where we will denote } \frac{1}{n} \sum_{j=1}^{n} y_j = \bar{y}$$

$$\propto \exp\left(-\frac{1}{2 \times \left(\frac{1}{n\tau}\right)} \left(\mu - \bar{y} \right)^2 \right)$$

We get that $\pi(\mu | \tau) \sim \mathcal{N}\left(\bar{y} ; \frac{1}{n\tau} \right)$ easy to sample from.

In the same way: $\pi(\tau | \mu) \propto \tau^{\frac{n-1}{2}} \exp\left(-\tau \times \frac{1}{2} \sum_{j=1}^{n} (y_j - \mu)^2 \right)$

Denoting $B = \frac{1}{2} \sum_{j=1}^{n} (y_j - \mu)^2$ we recognise that $\pi(\tau | \mu) \sim \Gamma\left(\frac{n+1}{2}, B \right)$

This produces a very simple Gibbs sampler: Given (μ_j, τ_j)

- $\mu_{j+1} \sim \mathcal{N}\left(\bar{y} ; (n \tau_j)^{-1} \right)$
- $\sigma_{j+1}^2 = \tau_{j+1}^{-1} \sim \Gamma\left(\frac{n+1}{2}, B_{j+1} \right)$ where B_{j+1} is B where μ has been substituted by μ_{j+1}.

→ Rq sampling from a Γ distribution is easy in the case where $\Gamma(a, b)$ $a \in \mathbb{N}^*$ as it can be done by summing iid $\mathrm{Exp}(b)$

Otherwise use a rejection method with $\Gamma[a], b)$ as proposal.

b) the Ising model for image processing:

Let V be a set of connected vertices. the connections between vertices are called edges. this forms a graph. think as an example of the pixels in a grid of an image:

the Ising model is a distribution which assigns to each vertex a r.v in $\{0,1\}$ or $\{-1,1\}$ with additional constraints: the value at a vertex is influenced by its immediate neighbours.

To introduce things formally, we will need some definitions:

def: Let $\mathcal{S}$ be the set of vertices (of finite number)
$\mathcal{X}_s$ is the state space at vertex $s \in \mathcal{S}$ (typically for images $\mathcal{X}_s = [\![0, 255]\!]$)
we assume the random variables take only a finite number of values. ie $|\mathcal{X}_s| < \infty$.

Let also $\mathcal{X} = \prod_{s \in \mathcal{S}} \mathcal{X}_s$ the space of all possible configurations

We denote $\mathcal{X}_A = \prod_{s \in A} \mathcal{X}_s$ and $X_A : x \mapsto x_s|_{s \in A}$ the canonical projection

def: We call random field any measure π on $\mathcal{X}$ s.t $\forall x \in \mathcal{X}$ $\pi(x) > 0$

This means that all configurations in $\mathcal{X}$ are possible with a non zero proba.

def: Let $A \subset \mathcal{S}$, $x_A \in \mathcal{X}_A$, $x_{\mathcal{S}\setminus A} \in \mathcal{X}_{\mathcal{S}\setminus A}$ and π be a random field. We call local characteristic the conditional probability $\pi(X_A = x_A \mid X_{\mathcal{S}\setminus A} = x_{\mathcal{S}\setminus A})$

Rq: As $\forall x \in \mathcal{X}$ $\pi(x) > 0$ this is always well-defined.

def: A neighbour system $\mathcal{V} = \{\mathcal{V}_s ; s \in \mathcal{S}\}$ is a collection of sub sets of $\mathcal{S}$ s.t.

(i) $\forall s \in \mathcal{S}$, $s \notin \mathcal{V}_s$
(ii) $s \in \mathcal{V}_t \iff t \in \mathcal{V}_s$

def: A clique is a subset of $\mathcal{S}$ s.t. all its elements are neighbours from each other

def: A markov random field π with respect to $\mathcal{V}$ is a random field such that

$\forall s \in \mathcal{S}$: $\pi(X_s = x_s \mid X_r = x_r, \forall r \neq s) = \pi(X_s = x_s \mid X_r = x_r, \forall r \in \mathcal{V}_s)$

Proof: $\pi = \pi(X_A = x_A \mid X_{S\setminus A} = x_{S\setminus A}) = \dfrac{\pi(X_A = x_A, X_{S\setminus A})}{\pi(X_{S\setminus A} = x_{S\setminus A})}$ (Bayes' rule)

$\pi(X_{S\setminus A} = x_{S\setminus A})$ ← marginal on x_A

$$= \frac{\exp\left[-\sum_{c\in\mathcal{C}} U_c(x_A x_{S\setminus A})\right]}{\sum_{y_A \in \mathcal{X}_A} \exp\left[-\sum_{c\in\mathcal{C}} U_c(y_A x_{S\setminus A})\right]}$$

Let $\mathcal{C} = \mathcal{C}_1 \sqcup \mathcal{C}_2$ (disjoint union) where:

$\mathcal{C}_1 = \{c \in \mathcal{C} \text{ s.t. } c \cap A = \emptyset\}$ and $\mathcal{C}_2 = \{c \in \mathcal{C} \text{ s.t. } c \cap A = \emptyset\}$

Let $R = S \setminus (A \cup \partial(A))$ where $\partial(A) = \{$neighbours of elements in $A\} \setminus A$

We introduce a neutral point o (a landmark)

If $C \in \mathcal{C}_2$: $U_c(z_A z_{\partial A} z_R) = U_c(o_A z_{\partial(A)} z_R)$ because as $C \cap A = \emptyset$ the value of U_c will only depend on the projection of z on C therefore we do not care about what happens on A.

If $C \in \mathcal{C}_1$ $U_c(z_A z_{\partial(A)} z_R) = U_c(z_A z_{\partial(A)} o_R)$. Again what is important is the projection of the cliques which contains elements of A. This includes A and $\partial(A)$. Then,

$$\pi = \frac{\exp\left[-\sum_{c\in\mathcal{C}_1} U_c(x_A x_{\partial(A)} x_R)\right] \exp\left[-\sum_{c\in\mathcal{C}_2} U_c(\overset{o_A \text{ on } \mathcal{C}_2}{\overbrace{x_A}} x_{\partial(A)} x_R)\right]}{\left(\sum_{y_A \in \mathcal{X}_A} \exp\left[-\sum_{c\in\mathcal{C}_2} U_c(y_A x_{\partial(A)} x_R)\right]\right) \exp\left[-\sum_{c\in\mathcal{C}_2} U_c(\underset{= o_A \text{ on } \mathcal{C}_2}{\underbrace{y_A}} x_{\partial(A)} x_R)\right]}$$

$$= \frac{\exp\left[-\sum_{c\in\mathcal{C}_2} U_c(x_A x_{\partial(A)} x_R)\right]}{\sum_{y_A \in \mathcal{X}_A}\left[\exp\left(-\sum_{c\in\mathcal{C}_1} U_c(y_A x_{\partial(A)} x_R)\right)\right]}$$ which proves (*)

Moreover,

$$\pi = \frac{\exp\left(-\sum_{c\in\mathcal{C}_1} U_c(x_A x_{\partial(A)} \overset{o_R = y_R \text{ on } \mathcal{C}_1}{\overbrace{x_R}})\right) \times \left[\sum_{y_R \in \mathcal{X}_R} \exp\left(-\sum_{c\in\mathcal{C}_2} U_c(\overset{x_A \text{ on } \mathcal{C}_2}{\overbrace{o_A}} x_{\partial(A)} y_R)\right)\right]}{\left[\sum_{y_A \in \mathcal{X}_A} \exp\left(-\sum_{c\in\mathcal{C}_1} U_c(y_A x_{\partial(A)} \underset{= o_R = y_R \text{ on } \mathcal{C}_1}{\underbrace{x_R}})\right)\right] \times \left[\sum_{y_R \in \mathcal{X}_R} \exp\left(-\sum_{c\in\mathcal{C}_2} U_c(\underset{= y_A \text{ on } \mathcal{C}_2}{\underbrace{o_A}} x_{\partial(A)} y_R)\right)\right]}$$

$$= \frac{\sum_{y_R \in \mathcal{X}_R} \exp\left[-\sum_{c\in\mathcal{C}_1} U_c(x_A x_{\partial(A)} y_R) - \sum_{c\in\mathcal{C}_2} U_c(x_A x_{\partial(A)} y_R)\right]}{\sum_{y_A y_R \in \mathcal{X}_A \mathcal{X}_R} \exp\left[-\sum_{c\in\mathcal{C}_1} U_c(y_A x_{\partial(A)} y_R) - \sum_{c\in\mathcal{C}_2} U_c(y_A x_{\partial(A)} y_A)\right]} \to \mathcal{L}(x_A x_{\partial(A)})$$

$= \pi(X_A = x_A \mid X_{\partial(A)} = x_{\partial(A)})$ and for $A = \{s\}$ we get the Markov property.

def: A Gibbs random field induced by an energy function H is defined as

$$\pi(x) = \frac{\exp(-H(x))}{\sum_{y \in \mathcal{X}} \exp(-H(y))}$$

def: • A potential is a family of functions $U = \{U_A, A \subset \mathcal{S}\}$ on $\mathcal{X}$ s.t

(i) $U_\emptyset = 0$

(ii) $U_A(x) = U_A(y)$ if $x_A(x) = x_A(y)$

• We can define an energy related to a potential U as $H_U = \sum_{A \in \mathcal{S}} U_A$

(We some times denote $U = H_U$ to make things simpler)

• A potential U is a neighbour potential w.r.t. $\mathcal{P}$ if $U_A \equiv 0$ for all $A \subset \mathcal{S}$ which are NOT cliques for $\mathcal{P}$.

ex: Ising: $\mathcal{X} = \{-1, 1\}$ $\mathcal{S} = \mathbb{Z}^2$ or a sub grid of $\mathbb{Z}^2$. Let $c = (s,t)$

$\mathcal{P}$ = 4-connectivity

$$U_c(x_s, x_t) = \begin{cases} -\beta & \text{if } x_s = x_t \\ +\beta & \text{if } x_s \neq x_t \end{cases} = -\beta x_s x_t$$

And $U_{c=s}(x_s) = -\beta x_s$. This defines an energy as: $\forall x \in \mathcal{X}$

$$U(x) = -\sum_{c=(s,t)} \beta x_s x_t - \sum_{s \in \mathcal{S}} B x_s$$

We will prove the Hammersley-Clifford theorem which says that if we have a Gibbs random field w.r.t. a potential coming from a neighbour system $\mathcal{P}$, then it is a Markov random field and we can easily calculate the local characteristics.

Theorem: Let π a random field given by a neighbour potential U w.r.t a neighbour system $\mathcal{P}$ is:

$$\pi(x) = \frac{\exp\left(-\sum_{c \in \mathcal{C}} U_c(x)\right)}{\sum_{y \in \mathcal{X}} \exp\left(-\sum_{c \in \mathcal{C}} U_c(y)\right)}$$ where $\mathcal{C}$ is the set of cliques for $\mathcal{P}$.

Then the local characteristics are:

$$\pi\left(X_s = x_s, \forall s \in A \mid X_s = x_s, s \in \mathcal{S} \setminus A\right) = \frac{\exp\left[-\sum_{c \in \mathcal{C}, c \cap A \neq \emptyset} U_c(x)\right]}{\sum_{y_A \in \mathcal{X}} \exp\left[-\sum_{c \in \mathcal{C}, c \cap A \neq \emptyset} U_c(y_A x_{\mathcal{S} \setminus A})\right]} \quad (*)$$

where $x_y = x = x_A x_{\mathcal{S} \setminus A}$

Moreover $\pi(X_s = x_s, s \in A \mid X_s = x_s, s \in \mathcal{S} \setminus A) = \pi\left(X_s = x_s, s \in A \mid X_s = x_s, s \in \mathcal{P}(A)\right)$

the Ising model is very often used for many applications as it is a simple Markov random field model. It is a very interesting choice as prior law on images as they are given a natural neighbour system thanks to pixels. This is particularly appealing for segmentation and it is the choice that may be used in our 1st example of estimation of x^* as our "cost" on images. It favour the homogeneous areas as we tried to.

the Ising model which has only 2 classes can be extended to more and is then called the Potts model.

As the local characteristics are very easy to compute, it is a very interesting choice for a Gibbs sampler.

4) Around the traditional Gibbs sampler:

the previous algorithm is sometimes called the deterministic updating GS (DUGS). And we have seen that it is not reversible. Some other versions of the sampler have been proposed to face this problem.

a) the reversible Gibbs sampler

Pseudo code: Given $x_n \in \mathbb{R}^k$

$$\text{2k-1 simulations} \begin{cases} \tilde{x}^1 \sim \pi(\cdot \mid x_n^2, \dots, x_n^k) \\ \vdots \\ \tilde{x}^k \sim \pi(\cdot \mid \tilde{x}^1, \dots, \tilde{x}^{k-1}) \quad \rightarrow \tilde{x}_n^k = x_{n+1}^k \\ x_{n+1}^{k-1} \sim \pi(\cdot \mid \tilde{x}^1, \dots, \tilde{x}^{k-2}, x_{n+1}^k) \\ \vdots \\ x_{n+1}^1 \sim \pi(\cdot \mid x_{n+1}^2, \dots, x_{n+1}^k) \end{cases}$$

As expected it is reversible:

Proof: Again for sake of simplicity, we prove it for $k=2$ but it easily generalises to any $k \geq 2$

Let $(A, B) \in \mathcal{X}^2$:

$$\int_{A\times B} \pi(y_1,y_2)\left[\int_{\mathcal{X}} \pi(\omega|y_2)\,\pi(y_2'|\omega)\,d\omega\right]\pi(y_1'|y_2')\,dy\,dy'$$

$$= \int_{A\times B} \pi(y_1|y_2)\,\pi(y_2)\left[\int_{\mathcal{X}} \frac{\pi(\omega,y_2)}{\pi(y_2)}\,\pi(y_2'|\omega)\,d\omega\right]\frac{\pi(y_1',y_2')}{\pi(y_2')}\,dy\,dy'$$

$$= \int_{A\times B} \pi(y_1',y_2')\left[\int_{\mathcal{X}} \pi(\omega,y_2)\,\frac{\pi(y_2'|\omega)}{\pi(y_2')}\,d\omega\right]\pi(y_1|y_2)\,dy\,dy'$$

$$= \int_{A\times B} \pi(y_1',y_2')\left[\int_{\mathcal{X}} \pi(\omega,y_2)\,\frac{\pi(\omega|y_2')}{\pi(\omega)}\,d\omega\right]\pi(y_1|y_2)\,dy\,dy'$$

$$= \int_{A\times B} \pi(y_1',y_2')\left[\int_{\mathcal{X}} \pi(y_2|\omega)\,\pi(\omega|y_2')\,d\omega\right]\pi(y_1|y_2)\,dy\,dy'$$

b) Random Scan gibbs Sampler:

the simulations of each component is done in a random order. But not one at a time. We need to scan all coordinate each time otherwise it fails to be reversible.

Pseudo code: 1) Sample a permutation $\sigma \in \mathfrak{S}_k$

2) Sample $y_{\sigma_1}^{n+1} \sim \pi_{\sigma_1}(y_{\sigma_1} | y^{(-\sigma_1)^n})$

$\vdots$

$$y_{\sigma_k}^{n+1} \sim \pi_{\sigma_k}(y_{\sigma_k} | y^{(-\sigma_k)^{n+1}})$$

Proof of reversibility: Σ over all permutations in $\mathfrak{S}_k$

Rq: It is better than the previous one as one does not samples coordinates twice. All samples are used!

5) The Gibbs Sampler as an MH sampler:

We said that the GS is not a rejection based sampler. However it can be seen that it is a particular case of the MH algorithm.

Let consider for $1 \le i \le k$ the following transition kernel: It only updates the i^{th} coordinate and $P_i(x,y) = \pi_i(y^i | x^{(-i)})\,\mathbb{1}_{\{y^{(-i)} = x^{(-i)}\}}$

If we use this proposal in a MH it appears that $\alpha(x,y) = 1$ and it is therefore always accepted.

By iterating over the coordinates we get that the GS is an iteration of MH with proposal evolving with $1 \leq i \leq k$

8) Metropolis - within - Gibbs sampler:

the GS assumes we are able to sample from $\Pi_i(x_i \mid x^{(-i)})$. There are unfortunately many models where this is not as trivial as the previous examples.

To face this, we can use a MH algorithm at each step of the GS to sample each coordinate. Let $q(x_i; x^{(-i)}, y_i; x^{(-i)})$ be a proposal; a candidate y_i is proposed. One computes the acceptation ratio α_i and accept or not with probability α_i. Then, we change the coordinate.

The transition kernel is a little more complicated:

$$P_j(x, dy) = \bigotimes_{m \neq j} \delta_{x_m}(dy^m) \times \Big[q_j(dy^j \mid x^{-j}) \alpha_j(x^j, y^j) + \delta_{x^j}(dy_j) \int (1 - \alpha_j(x^j, b)) q(x^j, b) db \Big]$$

9) Which algo for high dimensional random variable?

Gibbs Sampler designed for this But loops over k

Tala is a good alternative

Tradeoff between k and the calculation on $\nabla \log \Pi$

k α to compute

$1 \times$ only.