Suppose p is a p.m.f. (p.d.f.) on $\mathbb{X}$, and denote its i:th conditional with respect to other variables as

$$p_{i|-i}(x^{(i)} \mid x^{(-i)}) = \frac{p(x)}{p_{-i}(x^{(-i)})},$$

whenever the marginal $p_{-i}(x^{(-i)}) > 0$, where

$$p_{-i}(x^{(-i)}) := \sum_{z \in \mathbb{Z}} p(x^{(1)}, \ldots, x^{(i-1)}, z, x^{(i+1)}, \ldots, x^{(d)})$$
$$\left(p_{-i}(x^{(-i)}) := \int p(x^{(1)}, \ldots, x^{(i-1)}, z, x^{(i+1)}, \ldots, x^{(d)}) \mathrm{d}z \right).$$

Algorithm 6.33 ((Random scan) Metropolis-within-Gibbs)**.** Suppose that $q_i(x^{(i)}, \cdot \mid x^{(-i)})$ determines a p.m.f. (p.d.f.) on $\mathbb{X}_1$ for each $x \in \mathbb{X}$ and for all $i = 1, \ldots, d$. Choose some $X_0 \equiv x_0$ with $p(x_0) > 0$ and iterate for $k = 1, \ldots, n$

(a) Draw random coordinate index $I_k \sim \mathcal{U}\{1, \ldots, d\}$.

(b) Set $X_k^{(-I_k)} = X_{k-1}^{(-I_k)}$.

(c) Simulate $Y_k^{(I_k)} \sim q_{I_k}(X_{k-1}^{(I_k)}, \cdot \mid X_{k-1}^{(-I_k)})$

(d) With probability $\alpha_{I_k}(X_{k-1}^{(I_k)}, Y_k^{(I_k)} \mid X_{k-1}^{(-I_k)})$ accept and set $X_k^{(I_k)} = Y_k^{(I_k)}$, otherwise set $X_k^{(I_k)} = X_{k-1}^{(I_k)}$, where

$$\alpha_i(x, y \mid z^{(-i)}) := \min \left\{ 1, \frac{p_{i|-i}(y \mid z^{(-i)})}{p_{i|-i}(x \mid z^{(-i)})} \frac{q_i(y, x \mid z^{(-i)})}{q_i(x, y \mid z^{(-i)})} \right\}.$$

NB: In practice, we calculate the ratio of conditionals as

$$\frac{p_{i|i-1}(y \mid z^{(-i)})}{p_{i|i-1}(x \mid z^{(-i)})} = \frac{p_u(z^{(1)}, \ldots, z^{(i-1)}, y, z^{(i+1)}, \ldots, z^{(d)})}{p_u(z^{(1)}, \ldots, z^{(i-1)}, x, z^{(i+1)}, \ldots, z^{(d)})},$$

and in case $p(x)$ is defined as a product of terms, of which only few depend on the i:th coordinate, the ratio simplifies...

Proposition 6.34. *Algorithm 6.33 is reversible with respect to p.*

Proof. (Discrete case) We may write the Markov transition in Algorithm 6.33 as follows

$$\begin{aligned} K(x, y) &= \sum_{i=1}^{d} \mathbb{P}(X_k = y \mid X_{k-1} = x, I_k = i) \mathbb{P}(I_k = i \mid X_{k-1} = x) \\ &= \frac{1}{d} \sum_{i=1}^{d} K_i(x, y), \end{aligned}$$

where $K_i(x, y) = \mathbb{P}(X_k = y \mid X_{k-1} = x, I_k = i)$ are Markov transition probabilities, which correspond to the steps (b), (c) and (d) of Algorithm 6.33.

In fact, given $I_k = i$ and $X_{k-1}^{(-i)} = z^{(-i)}$, (c) and (d) correspond a Metropolis-Hastings algorithm targetting $p_{i|-i}(\,\cdot \mid z^{(-i)})$ with proposals $q_i(x, y \mid z^{(-i)})$. If we denote its transition probability $\hat{K}_i(x, y \mid z^{(-i)})$, we have

$$K_i(x, y) = \hat{K}_i(x^{(i)}, y^{(i)} \mid x^{(-i)}) \mathbf{1}\left(y^{(-i)} = x^{(-i)}\right)$$

and then

$$\begin{aligned} p(x)K_i(x, y) &= p_{-i}(x^{(-i)}) p_{i|-i}(x^{(i)} \mid x^{(-i)}) \hat{K}_i(x^{(i)}, y^{(i)} \mid x^{(-i)}) \mathbf{1}\left(y^{(-i)} = x^{(-i)}\right) \\ &= p_{-i}(x^{(-i)}) p_{i|-i}(y^{(i)} \mid x^{(-i)}) \hat{K}_i(y^{(i)}, x^{(i)} \mid x^{(-i)}) \mathbf{1}\left(y^{(-i)} = x^{(-i)}\right) \\ &= p_{-i}(y^{(-i)}) p_{i|-i}(y^{(i)} \mid y^{(-i)}) \hat{K}_i(y^{(i)}, x^{(i)} \mid y^{(-i)}) \mathbf{1}\left(x^{(-i)} = y^{(-i)}\right) \\ &= p(y)K_i(y, x), \end{aligned}$$

where we first use reversibility of $\hat{K}_i(\,\cdot\,, \cdot \mid x^{(-i)})$ with respect to $p_{i|-i}(\,\cdot \mid x^{(-i)})$ and then the fact that the expression is non-zero with $x^{(-i)} = y^{(-i)}$.

The p-reversibility of K follows now easily:

$$p(x)K(x, y) = \frac{1}{d}\sum_{i=1}^{d} p(x)K_i(x, y) = \frac{1}{d}\sum_{i=1}^{d} p(y)K_i(y, x) = p(y)K(y, x). \qquad \square$$

Remark 6.35. In fact, the proof of Proposition 6.34 suggests that we may use multiple possible MCMC transitions, which we use at random. The mixture transition probability is reversible as long as the component transition probabilities are. And the mixing weights need not be uniform.

For instance, we could have K_1 being an independence sampler transition and K_2 a random-walk Metropolis transition, and choose randomly which update we follow.

Definition 6.36. *Gibbs sampling* is a specific instance of Metropolis-within-Gibbs, where the proposal distributions are the conditional distributions,

$$q_i(x, y \mid z^{(-i)}) = p_{i|-i}(y \mid z^{(-i)}).$$

Note that in Gibbs sampling, the acceptance probability $\alpha_i(x, y \mid z^{(-i)}) \equiv 1$.

Remark 6.37 (*). Algorithm 6.33 is valid also in the continuous case $\mathbb{X} = \mathbb{R}^d$. We cannot use Proposition 6.24 directly to verify reversibiity, but we need to check that if $X_0 \sim p$, then $(X_0, X_1) \stackrel{d}{=} (X_1, X_0)$. The proof follows similarly as in the discrete case

$$\begin{aligned} &\mathbb{P}(X_0 \in A, X_1 \in B) \\ &= \int_A \left[\int_B p_{-i}(x^{(-i)}) p_{i|-i}(x^{(i)} \mid x^{(-i)}) \hat{K}_i(x^{(i)}, y^{(i)} \mid x^{(-i)}) \mathbf{1}\left(y^{(-i)} = x^{(-i)}\right) \mathrm{d}x \right] \mathrm{d}y \\ &= \mathbb{P}(X_0 \in B, X_1 \in A). \end{aligned}$$

Example 6.38 (Ising model). Let $\mathbb{X} = \{0, 1\}^{\ell \times m}$ the set of all $\ell \times m$ binary matrices. We can think them as 'images' $x \in \mathbb{X}$ where $x^{(i,j)} = 0$ or 1 corresponds to (i, j):th pixel being black or white, respectively.

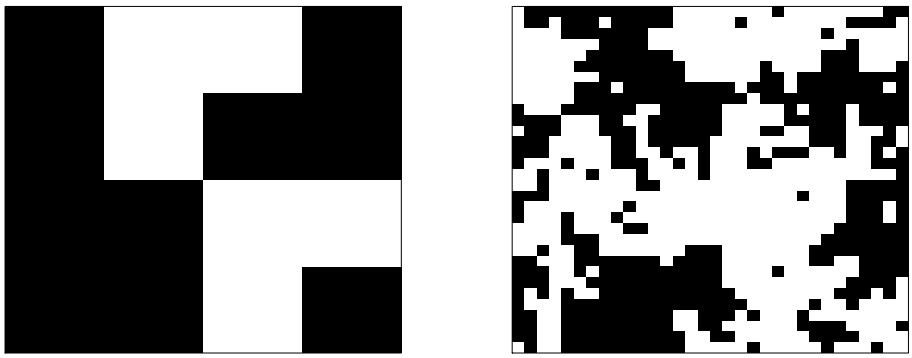

Figure 13: Left: Example 4-by-4 configuration with $\#x = 12$; right: realisation of the Ising model in $m = 32$, $\theta = 0.8$.

For $x \in \mathbb{X}$, denote $\#x$ for the number of disagreeing neighbours in x, which we may calculate by

$$\#x = \sum_{i=1}^{\ell} \sum_{j=1}^{m-1} \mathbf{1}\left(x^{(i,j)} \neq x^{(i,j+1)}\right) + \sum_{j=1}^{m} \sum_{i=1}^{\ell-1} \mathbf{1}\left(x^{(i,j)} \neq x^{(i+1,j)}\right).$$

The *Ising model* is defined as the following distribution on $\mathbb{X}$:

$$p(x) \propto \exp(-\theta \#x),$$

where $\theta > 0$ is a 'smoothing' parameter.

Example 6.39 (MCMC for the Ising model). Let $X_0^{(i,j)} \overset{\text{i.i.d.}}{\sim} \mathcal{U}\{0,1\}$, and do

(a) Draw random indices $I_k \sim \mathcal{U}\{1,\dots,\ell\}$, $J_k \sim U\{1,\dots,m\}$.
(b) Set $Y_k^{(I_k,J_k)} = 1 - X_{k-1}^{(I_k,J_k)}$.
(c) Set $X_k^{(i,j)} = X_{k-1}^{(i,j)}$ for all $(i,j) \neq (I_k, J_k)$.
(d) With probability $\alpha_{I_k,J_k}(X_{k-1}^{(I_k,J_k)}, Y_k^{(I_k,J_k)} \mid X_{k-1}^{(-(I_k,J_k))})$ set $X_k^{(I_k,J_k)} = Y_k^{(I_k,J_k)}$; otherwise set $X_k^{(I_k,J_k)} = X_{k-1}^{(I_k,J_k)}$, where

$$\alpha_{i,j}(x, y \mid z^{(-(i,j))}) = \min\left\{1, \exp\left[-\theta\big(\#(y, z^{(-(i,j))}) - \#(x, z^{(-(i,j))})\big)\right]\right\},$$

where $(x, z^{(-(i,j))})$ stands for the image where the (i,j):th pixel equals x and the rest are defined by $z^{(-(i,j))}$.

Remark 6.40. Note that $q_{i,j}$ here corresponds to a deterministic 'flip' of the (i,j):th pixel value. In fact, we shall see later that this choice of $q_{i,j}$ is the most efficient in terms of the *asymptotic variance*.

Remark 6.41. Note that in practice one should not re-calculate $\#(y, z^{(-(i,j))})$ and $\#(x, z^{(-(i,j))})$, but only their difference.

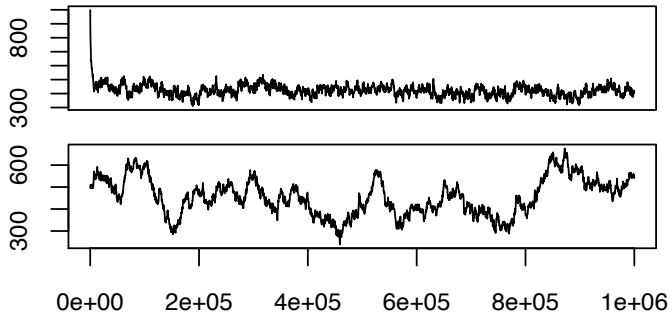

Figure 14: Trajectories of $f(X_k) = \#X_k$ (top) and $w(X_k)$ (bottom).

```
m = 32; n = 32; theta = 0.8; n = 100_000
function hashdiff(y, X, i, j, m, n)
    hash_y = 0; hash_x = 0; x = X[i,j]
    function check_ind!(i_, j_)
        hash_y += (y != X[i_,j_]);  hash_x += (x != X[i_,j_])
    end
    if i>1 check_ind!(i-1,j) end
    if i<m check_ind!(i+1,j) end
    if j>1 check_ind!(i,j-1) end
    if j<n check_ind!(i,j+1) end
    hash_y - hash_x
end
X = [rand(0:1) for i=1:m, j=1:m] # Independent random initialisation
for k = 1:n
  i = rand(1:m); j = rand(1:m) # Pick random index
  y = 1-X[i,j]                 # Propose swap 0<->1
  if rand() < exp(-theta*hashdiff(y, X, i, j, m, m))
      X[i,j] = y
  end
end
```

What would be good indicators to monitor the convergence of the Ising model simulation? We could look at:

- the function $f(x) = \#x$,
- the function $w(x) = \sum_{i,j} \mathbf{1}\left(x^{(i,j)} = 1\right)$, that is, the total number of white pixels.

Example 6.42 (Bayesian image recovery). Let X be an unknown true image,

$$X \sim \text{Ising}(\theta),$$

with θ known. Denote $p_0(x) = \mathbb{P}(X = x)$.

Suppose we do not observe X directly, but through a 'noisy channel'. At pixel $i, j = 1, 2, \ldots, m$ we observe

$$O^{(i,j)} = X^{(i,j)} + \epsilon^{(i,j)}, \qquad \text{with} \qquad \epsilon^{(i,j)} \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

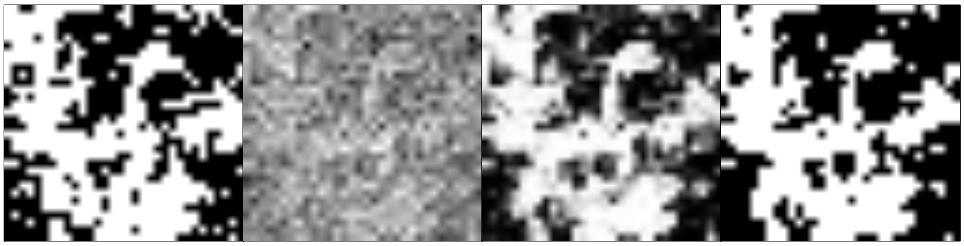

Figure 15: From the left: Simulation of the Ising model X with $\theta = 0.8$; the noisy observations O of X with $\sigma^2 = 1$; the posterior mean approximation by MCMC; the MAP approximation by MCMC (estiamtes with ten million samples).

and with σ known. The likelihood for $x^{(i,j)}$ is $L(x^{(i,j)}; o^{(i,j)}) = N(o^{(i,j)}; x^{(i,j)}, \sigma^2)$ so

$$L(x; o) \propto \prod_{i,j=1}^{m} \exp\left(- \frac{(x^{(i,j)} - o^{(i,j)})^2}{2\sigma^2} \right).$$

If we observe $O = o$ we are interested in the *posterior distribution* of X given $O = o$,

$$p(x) = \mathbb{P}(X = x \mid O = o) \propto L(x; o) p_0(x),$$

so we have

$$\log p_u(x) = -\frac{|x - o|^2}{2\sigma^2} - \theta \# x \qquad \text{where} \qquad |x - o|^2 = \sum_{i,j=1}^{m} (x^{(i,j)} - o^{(i,j)})^2.$$

We will simulate $X_1, \ldots, X_n \sim p$ with MCMC and use the samples to approximate the posterior mean and pixel-wise maximum a Posteriori (MAP) estimates $i = 1, \ldots, m^2$

$$\bar{X}^{(i)} := \frac{1}{n} \sum_{k=1}^{n} X_k^{(i)} \approx \mathbb{E}[X^{(i)} \mid O = o]$$
$$\mathbf{1}\left(\bar{X}^{(i)} > 1/2\right) \approx \arg\max_{x \in \{0,1\}} \mathbb{P}(X^{(i)} = x \mid O = o].$$

In order to implement the MCMC, we can recycle the implementation in Example 6.39 only modifying the acceptance probability $\alpha(y \mid x)$ to incorporate $-|x - o|^2/(2\sigma^2)$ factor.

Variants of Metropolis-within-Gibbs

Algorithm 6.33 introduced earlier is only variant of (Metropolis-within-)Gibbs sampling, in terms how $(I_k)_{k \geq 1}$ are chosen.

Random scan means we choose I_k at random, as in Algorithm 6.33. It is customary to take $I_k \sim \mathcal{U}\{1, \ldots, d\}$, but I_k can be chosen also from a non-uniform distribution over $\{1, \ldots, d\}$.

Deterministic scan version of the algorithm means I_k are not random, but deterministic. The common choice is to sweep $I_k = (k-1 \mod d) + 1$. Unlike the random scan version, the deterministic scan algorithm is *time-inhomogeneous*, but the 'skeleton' chain $(X_{dk})_{k\geq 0}$, is homogeneous, with composition of transition probabilities

$$\mathbb{P}(X_{dk} = y \mid X_{d(k-1)} = x) = (K_1 K_2 \cdots K_d)(x, y)$$

This transition probability is *not reversible* wrt. p in general, but is still p-invariant.

Random sweep is a hybrid of the two above: Simulate a random permutation of $\{1, \ldots, d\}$, and sweep through once in the corresponding order; simulate a new random permutation etc.

Remark 6.43. Metropolis-within-Gibbs moves can update a 'block' of coordinates instead of a single coordinate. The blocks need not be fixed size, and there can be moves with overlapping blocks (sharing same variables).

Convergence of Metropolis-within-Gibbs

Theorem 6.44. *Suppose that the Metropolis-within-Gibbs chain is p-irreducible and that starting from any $x \in \operatorname{supp}(p)$, there is a positive probability of accepting at least one move in each coordinate direction. Then, the strong law of large numbers holds (see Theorem 6.26).*

Proof. Theorem 12 of [23] shows that the chain is Harris recurrent[12], and the SLLN is implied by [14, Theorem 17.0.1 (i)]. □

Remark 6.45 (*). Theorem 6.44 adds one natural (and practically non-restrictive) condition over the irreducibility condition of Theorem 6.26, which only avoids some pathological scenarios (like if $x \in \operatorname{supp}(p)$ but the conditionals are well-defined...).

Because all moves in the Gibbs sampler are accepted, we have:

Corollary 6.46. *Any p-irreducible Gibbs sampler satisfies the SLLN.*

We give next natural sufficient conditions which enable to check the p-irreducibility of a (Metropolis-within-)Gibbs.

Definition 6.47 (Positivity of p)**.** The distribution p satisfies the *positivity condition* if the marginal distributions $p_i(x)$ satisfy for all $x \in \mathbb{R}$

$$\operatorname{supp}(p) = \operatorname{supp}(p_1) \times \cdots \times \operatorname{supp}(p_d).$$

In other words, $p_i(x^{(i)}) > 0$ for all $i = 1, \ldots, d$ if and only if $p(x^{(1)}, \ldots, x^{(d)}) > 0$.

Proposition 6.48. *If p satisfies the positivity condition, then the conditional densities $p_{i|-i}$ are well-defined everywhere on the support of p and the Gibbs sampling Markov chain is p-irreducible.*

---

12. *From any initial point $x \in \operatorname{supp}(p)$, the chain will visit each set $A \subset \mathbb{X}$ such that $\int_A p(x)\mathrm{d}x > 0$ with probability one.