Markov chains and Algorithmic Applications
Mini project: Signal recovery using MCMC

Stefano Cubeddu, Michele Lupini, Stefano Viel

December 2023

1 Optimizing over the binary hypercube

Assume $\Theta = \{0,1\}^d$, thus we have $\text{card}(\Theta) = 2^d < +\infty$, and define the Hamming weight as $\|\cdot\|_0 : \Theta \to \mathbb{N}$ such that $\|\theta\|_0$ is the number of 1's in θ for each $\theta \in \Theta$ or equivalently the Hamming distance of θ from the all-zero vector.

1. If $\sigma = 0$, then $m = 1$ is the minimum number of measurements required to recover θ with probability 1. Indeed, in this case the measurement is defined as $y = X\theta + \xi \in \mathbb{R}$, where $X \in \mathbb{R}^d$, $\theta \in \Theta$ such that $X_i \sim \mathcal{N}(0,1)$ i.i.d. and $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ independent of X. Therefore, we have that $y = X\theta$. Moreover, given (X, y) we can recover θ by finding the components of X whose sum is equal to y, which can be translated as the problem
$$\text{Find } J \subseteq \{1, \dots, d\} \quad \text{s.t.} \quad \sum_{j \in J} X_j = y. \tag{1}$$
Notice that X is a finite vector, whose components are sampled i.i.d. from a standard normal distribution, being $y = X\theta$, then there exist almost surely one and only one subset of indices $J \subseteq \{1, \dots, d\}$ that solves the problem (1). Indeed, assuming for the sake of argument that there is another set J' which solves (1), one can prove that $J = J'$ almost surely, i.e. $\mathbb{P}(J \neq J') = 0$. The following recursive function can be used to solve the problem, by comparing with y the sum of the components of X associated to any subset of indices, with a computational complexity of $\mathcal{O}(2^d)$.

Algorithm 1 Subset Sum Algorithm

```
Input: $X, y$
  SubsetSum($X, y$, currentIndex, currentSubset)
if sum(currentSubset) $= y$ then
    return currentSubset
end if
if currentIndex $=$ length($X$) then
    return None
end if
includeSubset $\leftarrow$ SubsetSum($X, y$, currentIndex $+ 1$, currentSubset $+ [X$[currentIndex]$]$)
if includeSubset $\neq$ None then
    return includeSubset
end if
return SubsetSum($X, y$, currentIndex $+ 1$, currentSubset)
```

2. In order to compute $\mathbb{P}(y|\theta, X)$, we define $Y = h(\theta, X, \xi) = X\theta + \xi$. Notice that Y can be seen as a measurable function of random variables (θ, X, ξ) and thus as a random variable itself with a certain probability density $f_Y(y)$ for any $y \in \mathbb{R}^m$. Moreover, we can consider the random variable $Y|(\theta, X)$ with conditional probability density function $f_{Y|(\theta,X)}(y)$, that turns out to be distributed as $\mathcal{N}(X\theta, \sigma^2 I_m)$, by using a well known result on the affine transformation of a multivariate normal random vector,

being $\xi \sim \mathcal{N}(0, \sigma^2 I^m)$. To summarize, defining $\mathbb{P}(y|\theta, X) = f_{Y|\theta,X}(y)$ for the sake of simplicity, it holds that

$$\mathbb{P}(y|\theta, X) = \frac{1}{\sqrt{(2\pi)^m \sigma^{2m}}} e^{-\frac{\|y-X\theta\|_2^2}{2\sigma^2}}.$$

Since the logarithm is a concave monotonic and non-decreasing function, we can reformulate the maximization problem to be solved to find the MLE in terms of the above result as

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log \mathbb{P}(y|\theta, X) = \arg\max_{\theta \in \Theta} \left\{ -\frac{m}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|y - X\theta\|_2^2 \right\} \tag{2}$$

Finally, notice that the problem is insensitive to constants and there exists a solution, as the state space Θ is finite. Thus, we can rewrite as the minimization problem

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \|y - X\theta\|_2^2. \tag{3}$$

3. Let β be the inverse temperature, a probability distribution π_β that concentrates on MLE as $\beta \to \infty$ is given by

$$\pi_\beta(\theta) = \frac{e^{-\beta f(\theta)}}{Z_\beta}, \tag{4}$$

where $f(\theta) = \frac{1}{m}\|y - X\theta\|_2^2$ is the function to be minimized and $Z_\beta = \sum_{\theta \in \Theta} e^{-\beta f(\theta)}$ is the partition function. We divided the loss function by m to have a normalized value regardless of how many measures we have. By doing so the same value of β will work for any number of measures m.

4. In order to construct a Metropolis-Hastings (MH) algorithm to draw samples form π_β, firstly define an easy to simulate base chain ψ on the state space $\Theta = \{0, 1\}^d$. For the sake of simplicity, let us consider the symmetric random walk on the hypercube such that for each $x, y \in \Theta$ we have

$$\psi_{xy} = \begin{cases} \frac{1}{d} & \text{if } \|y - x\|_0 = 1 \\ 0 & \text{otherwise} \end{cases}$$

Thus, the Markov chain with transition probability matrix ψ is irreducible and such that $\psi_{xy} > 0$ if and only if $\psi_{yx} > 0$. Because of the MH theorem a convenient design for the acceptance probabilities is the following

$$a_{xy} = \min\left\{1, \frac{\pi_\beta(y)\psi_{yx}}{\pi_\beta(x)\psi_{xy}}\right\} = \min\left\{1, \frac{\pi_\beta(y)}{\pi_\beta(x)}\right\} = \min\left\{1, e^{-\beta(f(y)-f(x))}\right\},$$

for any $x, y \in \Theta$ such that $\|y - x\|_0 = 1$, which can be also expressed as

$$a_{xy} = \begin{cases} e^{-\beta(f(y)-f(x))} & \text{if } f(y) > f(x) \\ 1 & \text{otherwise} \end{cases}$$

Therefore, the transition probabilities of the Metropolis chain are given by

$$p_{xy} = \begin{cases} \frac{1}{d} \min\left\{1, e^{-\beta(f(y)-f(x))}\right\} & \text{if } \|y - x\|_0 = 1 \\ \frac{1}{d} \sum_{z:\|z-x\|=1} (1 - \min\left\{1, e^{-\beta(f(z)-f(x))}\right\}) & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

The Metropolis-Hastings algorithm's acceptance probability computation hinges on evaluating $\pi_\beta(\theta)$. This involves two key steps: computing the loss function $f(\theta)$, dominated by a $\mathcal{O}(md)$ complexity matrix-vector multiplication $X\theta$, and applying the exponential function to $-\beta \cdot L(\theta)$ with constant time complexity. Consequently, each step's overall computational complexity is $\mathcal{O}(md)$. We simplified the complexity by observing that, after each chain step, only one element (two for subsequent questions)

Algorithm 2 Metropolis-Hastings for Sampling from π_β

```
Input: Matrix $X$, vector $y$, inverse temperature $\beta$, number of iterations $N$
Output: A sample from the distribution $\pi_\beta$
Initialize $\theta$ as a random state from $\Theta$
for iteration = 1 to $N$ do
    Propose $\theta'$ by randomly flipping one component of $\theta$
    Calculate $A(\theta, \theta') = \min\left(1, \frac{\pi_\beta(\theta')}{\pi_\beta(\theta)}\right)$
    Draw a random number $r$ from a uniform distribution in $[0, 1]$
    if $r < A(\theta, \theta')$ then
        $\theta = \theta'$ {Accept the new state}
    end if
end for
return $\theta$ {Final sample obtained after Metropolis-Hastings iterations}
```

of the vector θ was altered. Thus, to calculate $X\theta$, we only needed to consider the column of X corresponding to the modified components of θ and adjust them based on the new θ value to the previous $X\theta$ value. The final computational complexity is $\mathcal{O}(m + d)$, considering the cost of adding a column of X of size m to the previously computed $X\theta$ and the cost of identifying which components of θ underwent changes.

5. In the preliminary phase of our investigation, we systematically explored various fixed values for the parameter β, specifically considering $\beta = [3, 5, 10, 20, 50, 100]$. These initial experiments were conducted with parameters set at $m = 2000$, $d = 2000$, and `iterations` $= 2e5$. The primary objective of these initial trials was to establish a high-level range for the β values. Our evaluation criterion involved assessing the performance based on $\frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$, where the expected value was computed repeating each experiment 15 times. In instances where the value reached zero, we recorded the number of iterations required for the Metropolis-Hastings algorithm to achieve $\frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] = 0$ for the first time.

   Upon analysis, we determined that the most effective value for β was found to be 10, which, on average, achieved $\frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] = 0$ for the first time with 39986 iterations.

   Subsequently, we adopted a more versatile approach involving simulated annealing. This method facilitated a more extensive exploration of the hyperparameter space, encompassing the initial β value, the iteration interval for its increase, and the multiplicative magnitude of each increment of β. We systematically tested all possible combinations of the following parameters:

   - β : $[0.01, 0.05, 0.1, 0.5, 1, 5, 10, 20, 30, 100]$
   - increase frequency: $[2, 5, 10, 100, 1000, 1500, 2000, 3000]$
   - multiplicative increase: $[1.00001, 1.0001, 1.0005, 1.001, 1.01, 1.1, 1.2, 1.3]$

   The optimal parameters were identified as starting with $\beta = 10$, increasing every 1500 iterations, and incorporating a multiplicative increase of 1.00001. These settings facilitated convergence to $\frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] = 0$ within 36404 iterations. The improvements in simulated annealing (measured as the number of iterations necessary for convergence) are not substantial compared to the result with fixed $\beta = 10$ and the multiplicative increase is very small. Thus, we can conclude that simulated annealing is not useful in this case.

6. In Figure 1, we report the MSE $= \frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$ computed at the $2e5$-th iteration, rerunning the experiment 15 times, with respect to the number of samples m. The following parameters have been used: $d = 2000$, $\beta = 10$, without simulated annealing. As expected with a higher number of measures m, the Markov chain converges more quickly to the stationary distribution. Which makes $\frac{2}{d}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$

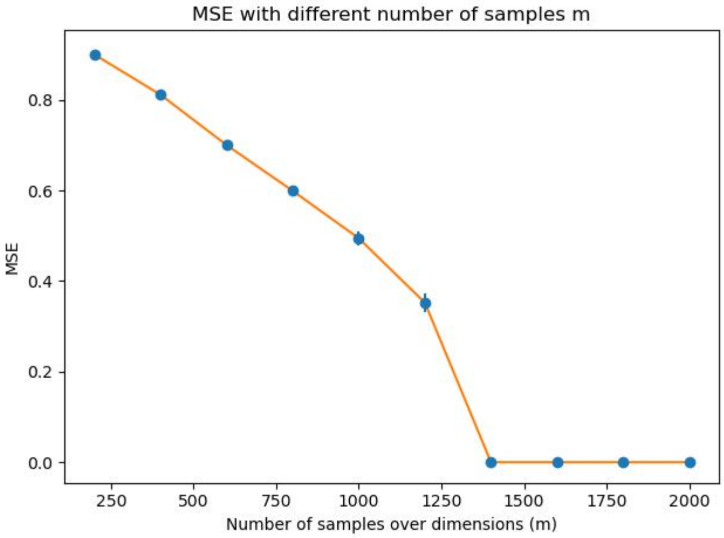

Figure 1: Recovered signal MSE over the binary hypercube. The MSE is computed on the final state of the sampling procedure, with parameters: $d = 2000$, $\beta = 10$. The plot reports the mean and the 95% confidence intervals of the MSE for 15 different runs of the same experiment.

go to zero with a smaller number of iterations. In Figure 1, we can also see that for $m > 1500$ the signal is always recovered with less than $2e5$ iterations. So, in order to reliably recover θ we need $\frac{m}{d} = 0.7$

2 Recovering a sparse, binary signal

Let $s \in \{0, \dots, d\}$ be a fixed parameter such that $s \ll d$, and now consider $\Theta = \{\theta \in \{0,1\}^d : \|\theta\|_0 = s\}$

1. Let us consider the base chain such that, given $\theta \in \Theta$ the current state of the MH algorithm, choose $i, j \in \{1, \dots, d\}$ independently and uniformly at random and swap the components θ_i, θ_j of θ to obtain the next state. This can equivalently be stated as choosing the base chain with transition probabilities

$$\psi_{xy} = \begin{cases} \frac{2}{d^2} & \text{if } \|y - x\|_0 = 2 \\ \frac{d^2 - 2ds + 2s^2}{d^2} & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

for any $x, y \in \Theta$. Thus we can design the acceptance probabilities for any $x, y \in \Theta$ such that $\|y - x\|_0 = 2$ or $y = x$ as

$$a_{xy} = \min\left\{1, \frac{\pi_\beta(y)\psi_{yx}}{\pi_\beta(x)\psi_{xy}}\right\} = \min\left\{1, e^{-\beta(f(y) - f(x))}\right\} \tag{5}$$

where π_β is defined as in the (4) on the new state space Θ. Therefore the transition probabilities of the Metropolis chain are given by

$$p_{xy} = \begin{cases} \frac{2}{d^2} \min\left\{1, e^{-\beta(f(y) - f(x))}\right\} & \text{if } \|y - x\|_0 = 2 \\ 1 - \sum_{z : \|z - x\|_0 = 2} \frac{2}{d^2} \min\left\{1, e^{-\beta(f(y) - f(x))}\right\} & \text{if } y = x \\ 0 & \text{otherwise} \end{cases}$$

The proposed base chain may not be suitable for the current problem due to the sparsity of the signal. With $s \ll d$, the chance of selecting two components with different values to swap is low, indeed notice that $\psi_{xx} \approx 1$ in the regime of d high, which means that the algorithm would often propose states that are identical to the current state, leading to a lack of exploration in the state space. To summarize the chain behaviour is close to the one of a reducible chain and the the convergence will be slow.

2. To improve the convergence of the MH algorithm for sparse binary signals, we can modify the selection process of the next state as follows. Given $\theta \in \Theta$ current state, find $A, B \subseteq \{0, \ldots, d\}$ sets of the indices of the zeros and ones components of θ, respectively, i.e. $i \in A$ if and only if $\theta_i = 0$, and $j \in B$ if and only if $\theta_j = 1$. Choose $i \in A,\ j \in B$ independently and uniformly at random, then swap the components θ_i, θ_j of θ to obtain the next state. This modification aims to improve the efficiency of the convergence, facilitating the exploration of the state space by making proposals different from the current state more likely to be accepted. Equivalently, we propose the base chain with transition probabilities

$$\psi_{xy} = \begin{cases} \frac{1}{s(d-s)} & \text{if } \|y - x\|_0 = 2 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

for any $x, y \in \Theta$. Similarly to the previous points, we design the acceptance probabilities according to the MH theorem as in (5) for any $x, y \in \Theta$ such that $\|y - x\|_0 = 2$. Thus the transition probabilities of the Metropolis chain are given by

$$p_{xy} = \begin{cases} \frac{1}{s(d-s)} \min\left\{1, e^{-\beta(f(y)-f(x))}\right\} & \text{if } \|y - x\|_0 = 2 \\ 1 - \sum_{z:\|z-x\|_0=2} \frac{1}{s(d-s)} \min\left\{1, e^{-\beta(f(y)-f(x))}\right\} & \text{if } y = x \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Similarly to the exploration done in Point 5 of Section 1 we tried both fixed β and simulated annealing. Simulated annealing did not lead to substantial improvement and we determined that it was better to have β fixed. More specifically $\beta = 10$. The results, visually depicted in Figure 2, highlight our ability to successfully recover the signal, achieving $\frac{1}{2s}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right] = 0$ with 170 or more measures. Since we were considering $d = 2000$, $s = 20$, and $2e5$ chain steps, our approach satisfies the conditions for signal recovery with $m = \mathcal{O}(s \log(d))$. In fact, in our case $s \log(d) = 66$ which only differs from 170 by a factor of 2.57.

3. As in Point 6 of Section 1, we run the experiments with different numbers of measures. In Figure 2, we report the MSE $= \frac{1}{2s}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$ computed as before at the $2e5$-th iteration, rerunning the experiment 15 times. The following parameters have been used: $d = 2000,\ \beta = 10$. Figure 2 illustrates that with $m \geq 170$ it was always possible to recover the signal. As in the previous question $\frac{1}{2s}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$ was computed on 15 experiments. The signal was recovered with fewer measures than before, this can be justified by the fact that the state space is smaller than before, as we are only considering the signals with Hamming weight fixed and equal to s which is much less than d. Thus, we can expect to recover it with fewer measures than before. To conclude, the minimum m needed to recover the signal is 170 which translates into $\frac{m}{d} = 0.085$.

3 Recovering a sparse signal from 1-bit measurements

Let $s \in \{0, \ldots, d\}$ be a fixed parameter such that $s \ll d$, consider $\Theta = \{\theta \in \{0, 1\}^d : \|\theta\|_0 = s\}$ and assume that the measurement is generated as $y = \text{sign}\,(X\theta + \xi)$, where sign is a function that acts component-wise on a vector.

1. In order to formulate the optimization problem to find the MLE, firstly let us compute $\mathbb{P}(y|\theta, X)$. Notice that $\xi \sim \mathcal{N}(0, \sigma^2 I_m)$ and proceeding as in Point 2 of Section 1, we define the random variable $Y = \text{sign}\,(X\theta + \xi)$ taking values in $\{-1, 1\}^m$, with probability mass function f_Y. Thus, the random variable $Y|(\theta, X)$ has probability mass function $f_{Y|(\theta,X)}(y) = \mathbb{P}(Y = y|\theta, X) = \mathbb{P}(y|\theta, X)$. By independence

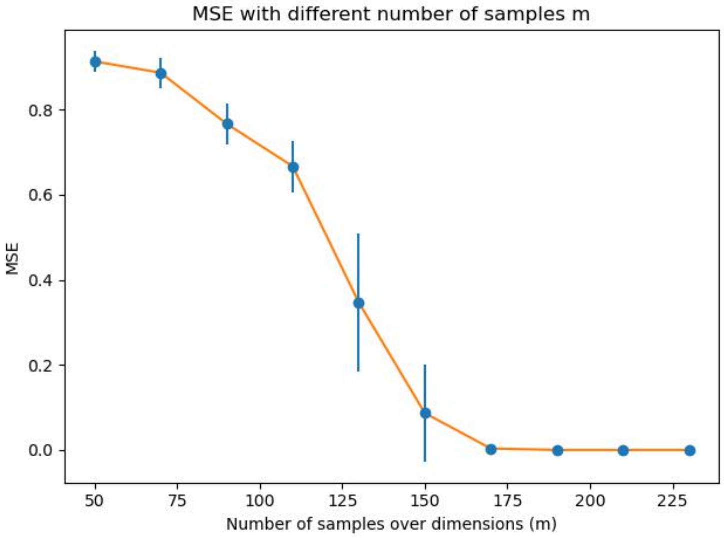

Figure 2: Recovered sparse signal MSE. The MSE is computed on the final state of the sampling procedure, with parameters $d = 2000,\ \beta = 10$. The plot reports the mean and the 95% confidence intervals of the MSE for 15 different runs of the same experiment.

one can see that for any $y \in \{-1, 1\}^m$

$$\mathbb{P}(y|\theta, X) = \mathbb{P}(\text{sign}\,(X\theta + \xi) = y) = \prod_{i=1}^{m} \mathbb{P}\big(\text{sign}\,(X\theta + \xi)_i = y_i\big) =$$

$$= \prod_{i=1}^{m} \left\{ \left[\int_{-\infty}^{\frac{(X\theta)_i}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, \mathrm{d}x \right]^{\frac{y_i+1}{2}} \left[1 - \int_{-\infty}^{\frac{(X\theta)_i}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, \mathrm{d}x \right]^{1-\frac{y_i-1}{2}} \right\}. \tag{8}$$

Indeed, for each fixed $i \in \{1, \dots, m\}$, let us consider the random variable $Y_i|(\theta, X) = (X\theta)_i + \xi_i \in \mathbb{R}$ and notice that $Y_i \sim \mathcal{N}((X\theta)_i, \sigma^2)$. Additionally, let Φ be the cumulative density function of the standard normal distribution $\mathcal{N}(0, 1)$. By using standard results on the affine transformation of a multivariate normal random vector, we have that

$$\mathbb{P}\big(\text{sign}\,(X\theta + \xi)_i = y_i\big) = \mathbb{P}(\text{sign}\,(Y_i) = y_i) = \begin{cases} \mathbb{P}(Y_i \geq 0) & \text{if } y_i = 1 \\ \mathbb{P}(Y_i < 0) & \text{if } y_i = -1 \end{cases} = \begin{cases} \Phi\left(\frac{(X\theta)_i}{\sigma}\right) & \text{if } y_i = 1 \\ 1 - \Phi\left(\frac{(X\theta)_i}{\sigma}\right) & \text{if } y_i = -1 \end{cases}$$

Finally, we can rewrite this results as

$$\mathbb{P}\big(\text{sign}\,(X\theta + \xi)_i = y_i\big) = \Phi\left(\frac{(X\theta)_i}{\sigma}\right)^{\frac{y_i+1}{2}} \left[1 - \Phi\left(\frac{(X\theta)_i}{\sigma}\right)\right]^{1-\frac{y_i+1}{2}}.$$

and we can easily see that the expression (8) holds. At this point, it is easy to write the optimization problem to find the MLE in terms of the log-likelihood

$$\log \mathbb{P}(y|\theta, X) = \sum_{i=1}^{m} \left\{ \frac{y_i + 1}{2} \log\left[\Phi\left(\frac{(X\theta)_i}{\sigma}\right)\right] + \left(1 - \frac{y_i + 1}{2}\right) \log\left[1 - \Phi\left(\frac{(X\theta)_i}{\sigma}\right)\right] \right\}.$$

In fact, the minimization problem to be solved to find the MLE is the following

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min}\, [-\log \mathbb{P}(y, |\theta, X)]\,.$$

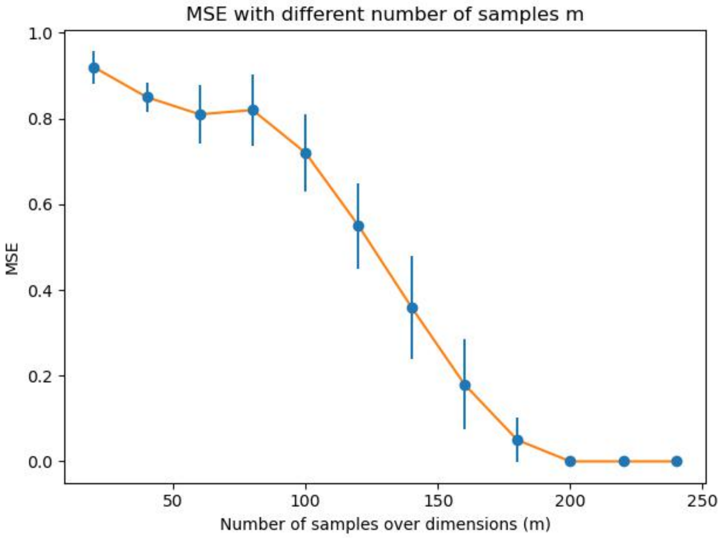

Figure 3: Recovered sparse signal MSE from 1-bit measurements. The MSE is computed on the final state of the sampling procedure, with parameters $d = 500,\ \beta = 5$. Simulated annealing was performed increasing the value of β every 1000 steps by a multiplicative factor of 1.01. The plot reports the mean and the 95% confidence intervals of the MSE for 15 different runs of the same experiment.

Finally, proceeding as in Point 3 of Section 1, we can define a probability distribution π_β that concentrates on the MLE as $\beta \to \infty$

$$\pi_\beta(\theta) = \frac{e^{-\beta f(\theta)}}{Z_\beta},$$

where $f(\theta) = -\log \mathbb{P}(y, |\theta, X)$ is the function to be minimized and $Z_\beta = \sum_{\theta \in \Theta} e^{-\beta f(\theta)}$ is the partition function.

2. In order to design a MH algorithm to estimate the MLE, we can proceed as in Point 4 of Section 2. For the sake of simplicity we can take the same base chain (6) on the state space Θ, then the transition probabilities of the Metropolis chain will be given by (7), where f is defined in Point 1. The algorithm is similar to Algorithm 2 designed for Point 4 of Section 1. The only difference is in line 3, which will be the following: Propose θ' by randomly switching two components with different values of θ.

   In Figure 3, we report the MSE $= \frac{1}{2s}\mathbb{E}\left[\|\hat{\theta} - \theta\|^2\right]$ with respect to the number of samples m. All the experiments were run with $d = 500,\ s = 5$ and `iterations` $= 50000$ as the maximum number of iterations. In this case, simulated annealing leads to better results. The optimal values we found are $\beta = 5$ which gets increased every 1000 steps by a multiplicative increase of 1.01.