1.1– Outline

- Tempering.
- Annealing.
- Slice sampling.
2.1– What have we done so far?

- Let the target distribution \( \pi(x) \) be defined on \( \mathcal{X} \) then practical MCMC algorithms consist of designing a collection of MH moves invariant with respect to \( \pi \).

- These moves can be trans-dimensional and typically only update a subset of variables.

- Every heuristic idea can be “Metropolized” to become theoretically valid.
2.2– Limitations of current approaches

- For complex target distributions, it can be very difficult to design efficient algorithms.

- It will always be difficult to explore a multimodal target if nothing is known beforehand about the structure of this distribution.

- We would like to have generic mechanisms to help us improving the performance of MCMC algorithms.
3.1– Introducing auxiliary distributions

• The key is to notice that although it might be difficult to sample from \( \pi(x) \), it could be easier to sample from related distributions.

• In particular, it should be easier to sample from

\[
\overline{\pi}^\gamma(x) = \frac{[\pi(x)]^\gamma}{\int [\pi(x)]^\gamma \, dx} \propto [\pi(x)]^\gamma
\]

where \( \gamma < 1 \).

• For \( \gamma < 1 \) the target \( \overline{\pi}^\gamma(x) \) is flatter than \( \pi(x) \), hence easier to sample from.

• This is called tempering.
3.2– Graphical illustration

Representation of $\pi(x)$ (blue), $\pi^{0.5}(x)$ (red) and $\pi^{0.01}(x)$ (black)
3.3– Example: Gaussian distribution

• Consider \( \pi (x) = \mathcal{N} (x; m, \sigma^2) \) then \( \pi^{\gamma} (x) = \mathcal{N} (x; m, \sigma^2 / \gamma) \).

• In one considers a simple random walk MH step then

\[
\alpha (x, x') = \min \left( 1, \frac{\pi^\gamma (x')} {\pi^\gamma (x)} \right) = \min \left( 1, \left( \frac{\pi (x')} {\pi (x)} \right)^\gamma \right)
\]

and the acceptance ratio

\[
\left( \frac{\pi (x')} {\pi (x)} \right)^\gamma \rightarrow 1 \text{ as } \gamma \rightarrow 0.
\]
3.4– Example: Discrete distribution

• Consider a discrete distribution \( \pi(x) \) on \( \mathcal{X} = \{1, \ldots, M\} \) then

\[
\overline{\pi}^\gamma(x) = \frac{\pi^\gamma(x)}{\sum_{i=1}^{M} \pi^\gamma(i)}
\]

and clearly

\[
\overline{\pi}^\gamma(x) \to \frac{1}{M}
\]
as \( \gamma \to 0 \).

• It is trivial to sample from a uniform distribution
3.5– Graphical illustration

Representation of \( \pi(x) \) (blue), \( \pi^{0.5}(x) \) (red) and \( \pi^{0.01}(x) \) (black)
3.6– Sequence of Tempered Distributions

• Instead of using only one auxiliary distribution \( \pi^\gamma (x) \), we will use a sequence of distribution defined as

\[
\pi_k (x) \propto [\pi (x)]^{\gamma k}
\]

where \( \gamma_1 = 1 \) and \( \gamma_k < \gamma_{k-1} \).

• In this case \( \pi_1 (x) = \pi (x) \) and \( \pi_k (x) \) is a sequence of distributions increasingly simpler to sample.
3.7– How to reuse samples?

- Assume we run an MCMC algorithm to sample from \( \pi_k(x) \), how to use these samples to approximate \( \pi(x) \).

- The first simple idea consists of using importance sampling, i.e.

\[
\pi(x) = \frac{(\pi(x) / \pi_k(x)) \pi_k(x)}{\int (\pi(x) / \pi_k(x)) \pi_k(x) \, dx}
\]

that is

\[
\pi^N(x) = \sum_{i=1}^{N} W_k^{(i)} \delta_{X_k^{(i)}}(x) \text{ where } W_k^{(i)} \propto \left( \pi\left( X_k^{(i)} \right) \right)^{1-\gamma_k}.
\]

- This idea is simple and will work properly if \( \gamma_k \) is close to 1.
3.8– Simulated tempering

- Alternatively, we could build a target distribution on \(\{1, ..., p\} \times X\) defined as
  \[
  \pi(k, x) = \pi(k) \pi_k(x)
  \]

- Then we could proposed deterministic moves like jumping from dimension \(k\) to 1 accepted with probability
  \[
  \min\left(1, \frac{\pi(1, x)}{\pi(k, x)}\right)
  \]

- Unfortunately, we don’t know the normalizing constants of \(\pi_k(x)\)!
  For example, if we were selecting
  \[
  \pi(k, x) \propto [f(x)]^{\gamma_k}
  \]
  where \(\pi(x) \propto f(x)\)
  then it means that
  \[
  \pi(k) \propto \int [f(x)]^{\gamma_k} dx.
  \]
  and you might biased unnecessarily the time spent in high temperatures.
3.9– Parallel tempering

- A more computationally intensive consists of building an MCMC on $\mathcal{X}^P$ of invariant distribution

$$\overline{\pi} (x_1, \ldots, x_P) = \pi_1 (x_1) \times \ldots \times \pi_P (x_P)$$

- This seems to be a more difficult problem as the dimension of the new target is higher and includes $\pi_1 (x_1) = \pi (x_1)$ as a marginal.

- The advantage is that we can design clever moves and use sample from “hot” chains to feed the “cold” chain.
3.10– Swap moves

• We can have a simple update kernel which updates each component of the Markov chain \( \left( X_1^{(i)}, \ldots, X_P^{(i)} \right) \) independently using

\[
K(x_{1:P}, x'_{1:P}) = \prod_{k=1}^{P} K_i(x_k, x'_k)
\]

where \( K_i \) is an MCMC kernel of invariant distribution \( \pi_i \).

• We can pick two chains associated to \( \pi_i \) and \( \pi_j \) and propose to swap their components, i.e. we propose

\[
x'_{-i,j} = x_{-i,j}, \quad x'_i = x_j \quad \text{and} \quad x'_j = x_i.
\]

This is accepted to

\[
\alpha(x_{1:P}, x'_{1:P}) = \min \left( 1, \frac{\pi(x'_{1:P})}{\pi(x_{1:P})} \right) = \min \left( 1, \frac{\pi_i(x_j) \pi_j(x_i)}{\pi_i(x_i) \pi_j(x_j)} \right).
\]
3.11– Alternative Up-and-Down Strategy

• The idea is to propose to sample from $\pi$ by using the following MCMC move of invariant distribution $\pi = \pi_0$ (Neal, 1996). The proposal is given by first tempering and then annealing

$$X'_1 \sim K_1(X'_0, \cdot), \quad X'_2 \sim K_2(X'_1, \cdot), \ldots, \quad X'_P \sim K_P(X'_{P-1}, \cdot)$$

$$X^*_P \sim K_P(X'_P, \cdot), \quad X^*_P-2 \sim K_{P-1}(X^*_{P-1}, \cdot), \ldots, \quad X^*_0 \sim K_1(X^*_1, \cdot)$$

where we assume here that $K_i$ is $\pi_i$–reversible.

• The acceptance rate for the candidate $X'_{2P-1}$ is given by

$$\min(1, \frac{\pi_1(X'_1)}{\pi_0(X'_0)} \times \ldots \times \frac{\pi_P(X'_{P-1})}{\pi_{P-1}(X'_{P-1})} \times \frac{\pi_{P-1}(X^*_{P-1})}{\pi_P(X^*_{P-1})} \times \ldots \times \frac{\pi_0(X^*_0)}{\pi_1(X^*_0)})$$
The proof of validity relies on the fact that $\pi$-reversibility can easily be checked. Let’s write $X^*_P = X'_{P-1}$ then the proposal distribution is

$$
\pi_0 (X'_0) \prod_{k=1}^{P} K_k \left( X'_{k-1}, X'_k \right) \prod_{k=1}^{P} K_k \left( X^*_k, X^*_k - 1 \right)
$$

$$
= \pi_0 (X'_0) \prod_{k=1}^{P} \frac{\pi_k (X'_k)}{\pi_k (X'_{k-1})} K_k \left( X'_k, X'_{k-1} \right) \prod_{k=1}^{P} \frac{\pi_k (X^*_k - 1)}{\pi_k (X^*_k)} K_k \left( X^*_k - 1, X^*_k \right)
$$

$$
= \pi_0 (X^*_0) \prod_{k=1}^{P} K_k \left( X^*_{k-1}, X^*_k \right) \prod_{k=1}^{P} K_k \left( X'_k, X'_{k-1} \right)
$$

$$
\times \frac{\pi_0 (X'_0)}{\pi_1 (X'_0)} \times \cdots \times \frac{\pi_{P-1} (X'_{P-1})}{\pi_P (X'_{P-1})} \frac{\pi_P (X'_{P-1})}{\pi_{P-1} (X'_{P-1})} \times \cdots \times \frac{\pi_1 (X^*_0)}{\pi_0 (X^*_0)}
$$
• Multiplying by the acceptance probability we have

$$
\pi_0 (X'_0) \prod_{k=1}^{P} K_k \left( X'_{k-1}, X'_k \right) \prod_{k=1}^{P} K_k \left( X^*_k, X^*_{k-1} \right)
$$

$$
\times \min(1, \frac{\pi_1(X'_1)}{\pi_0(X'_0)}) \times \cdots \times \frac{\pi_P(X'_{P-1})}{\pi_{P-1}(X'_{P-1})} \times \frac{\pi_{P-1}(X^*_P)}{\pi_P(X^*_P)} \times \cdots \times \frac{\pi_0(X^*_0)}{\pi_1(X^*_0)}
$$

$$
= \pi_0 (X^*_0) \prod_{k=1}^{P} K_k \left( X^*_{k-1}, X^*_k \right) \prod_{k=1}^{P} K_k \left( X'_k, X'_{k-1} \right)
$$

$$
\times \frac{\pi_0(X'_0)}{\pi_1(X'_0)} \times \cdots \times \frac{\pi_{P-1}(X'_{P-1})}{\pi_P(X'_{P-1})} \times \frac{\pi_P(X^*_P)}{\pi_{P-1}(X^*_P)} \times \cdots \times \frac{\pi_0(X^*_0)}{\pi_1(X^*_0)}
$$

$$
= \pi_0 (X^*_0) \prod_{k=1}^{P} K_k \left( X^*_{k-1}, X^*_k \right) \prod_{k=1}^{P} K_k \left( X'_k, X'_{k-1} \right)
$$

$$
\times \min(1, \frac{\pi_1(X'_1)}{\pi_0(X'_0)}) \times \cdots \times \frac{\pi_P(X'_{P-1})}{\pi_{P-1}(X'_{P-1})} \times \frac{\pi_{P-1}(X^*_P)}{\pi_P(X^*_P)} \times \cdots \times \frac{\pi_0(X^*_0)}{\pi_1(X^*_0)}
$$
Artificial Target Distribution on $(-1, 1) \times (-1, 1)$
3.12– Application to A Complex Toy Example

MH (left), Parallel Tempering (center) and Tempered transitions (right)
3.13– Application to Mixture of Distributions

Mixture of 4 Gaussians (Neal, 1996)
Parallel tempering and Tempered transitions are generic and powerful methods for sampling in complex problems.

Selection of the number $P$ of proposals and $\{\gamma_k\}$ is complex.

Various rules of thumb have been derived and preliminary runs are also often used.
4.1– Simulated Annealing

• An idea closely related to tempering is annealing.

• We have seen that

\[ \pi^\gamma (x) \propto [\pi (x)]^\gamma \]

is a flattened version of \( \pi (x) \) when \( \gamma < 0 \).

• On the contrary, \( \pi^\gamma (x) \) is a peakened version of the target as \( \gamma \) increases.
4.2– Simulated Annealing

- Under regularity conditions, it can be shown that the support of $\pi^\gamma(x)$ concentrates itself on the set of global maxima of $\pi(x)$.

- In the discrete case, let us write the unique maximum

$$x^* = \arg \max \{ \pi(x) : x \in X \}$$

then

$$\lim_{\gamma \to \infty} \pi^\gamma(x^*) = 1$$

as for any $x \neq x^*$

$$\lim_{\gamma \to \infty} \frac{\pi^\gamma(x)}{\pi^\gamma(x^*)} = \lim_{\gamma \to \infty} \left( \frac{\pi(x)}{\pi(x^*)} \right)^\gamma = 0.$$
4.3– Graphical illustration

Representation of $\pi(x)$ (top), $\pi^{10}(x)$ (middle) and $\pi^{100}(x)$ (bottom)
4.3– Graphical illustration

• Similarly in the continuous case, one can show that

\[ \lim_{\gamma \to \infty} \frac{\pi^\gamma (x)}{\sum_{x^* \in \mathcal{X}^*} \left| -\frac{\partial^2 \log \pi (x)}{\partial x_i \partial x_j} \right|_{x^*}^{-1/2}} \delta (x) \]

• If one could sample from \( \pi^\gamma (x) \) for large \( \gamma \) (asymptotically \( \gamma \to \infty \)) then we could solve any global optimization problem! Indeed maximizing any function \( f : \mathcal{X} \to \mathbb{R} \) would be equivalent to sample

\[ \pi^\gamma (x) \propto \left[ \exp (f (x)) \right]^{\gamma} \]

where we have \( \gamma \to \infty \).

• As \( \gamma \) increases, sampling from \( \pi^\gamma (x) \) is becoming harder. If it was simple, global optimization problem could be solved easily.
4.4– Graphical illustration

Representation of $\pi(x)$ (red), $\pi^{10}(x)$ (blue) and $\pi^{100}(x)$ (black)
• To sample from $\bar{\pi}^{\gamma} (x)$ for a large $\gamma$, we could use the same idea as parallel tempering where we would consider a sequence of distribution $\pi_k (x)$ with a decreasing sequence $\{\gamma_k\}$ such that $\gamma_1 >> 1$.

• However, this could be very expensive so an alternative simpler technique is used known as simulated annealing (highly popular method proposed in 1982)

• **Basic idea:** Sample an nonhomogeneous Markov chain at each time $k$ with transition kernel $K_k (x, x')$ of invariant distribution $\pi_k$. 
Any MCMC algorithm can be modified straightforwardly to perform global optimization! Just consider now a sequence of nonhomogeneous targets.

To ensure that this nonhomogeneous Markov chain converges towards $\pi_{\infty}$ as $k \to \infty$ you need to have conditions such as

$$K_k (x, x') \geq \delta^k \mu_k (x') \quad \text{and} \quad \gamma_k = C \log (k + k_0).$$

The second condition is not realistic, $\gamma_k$ increases too slowly and in practice we select $\gamma_k$ growing polynomially.

- Annealing
5.1– Hybrid Monte Carlo

• Alternative approaches consists of increasing the target distributions with auxiliary variables.

• *Hybrid Monte Carlo*: Define

\[ \pi (x, y) \propto \pi (x) \exp ( -\beta y^T y ) \]

• *Basis*: It is possible to move approximately on the manifold defined by \( \pi (x, y) = \text{cst.} \). See tutorial paper by Stoltz & al.
5.2– Slice Sampling

- Consider the target $\pi(x) \propto f(x)$. We consider the extended target

$$\overline{\pi}(x, u) \propto 1 \{(x, u); 0 \leq u \leq f(x)\}$$

- By construction, we have

$$\int \overline{\pi}(x, u) \, du = \frac{\int 1 \{(x, u); 0 \leq u \leq f(x)\} \, du}{\int \int 1 \{(x, u); 0 \leq u \leq f(x)\} \, dudx} = \frac{f(x)}{\int f(x) \, dx}$$

- Note that the same representation was implicitly used in Rejection sampling.
5.2– Slice Sampling

- To sample from $\pi(x, u)$ hence from $\pi(x)$, we can use Gibbs sampling

  \[
  \bar{\pi}(x|u) = \mathcal{U}(\{x : u \leq f(x)\}) ,
  \]

  \[
  \bar{\pi}(u|x) = \mathcal{U}(\{u : u \leq f(x)\}) .
  \]

- Sampling from $\bar{\pi}(u|x)$ is trivial but $\bar{\pi}(x|u)$ can be complex!

- MH step can be used to sample from $\bar{\pi}(u|x)$.
5.2– Slice Sampling

- Example: \( \pi(x) \propto \frac{1}{2} \exp(-\sqrt{x}) \) can be sampled using

\[
U | x \sim U \left(0, \frac{1}{2} \exp(-\sqrt{x})\right)
\]

and

\[
u \leq \frac{1}{2} \exp(-\sqrt{x}) \iff 0 \leq x \leq [\log(2u)]^2
\]

then

\[
X | u \sim U \left(0, [\log(2u)]^2\right)
\]
5.2– Slice Sampling

- In practice, the slice sampler is not really useful per se but can be straightforwardly extended when

\[ \pi(x) \propto f(x) = \prod_{i=1}^{k} f_i(x) \]

where \( f_i(x) > 0 \).

- We built the extended target

\[ \overline{\pi}(x, u_{1:k}) \propto \prod_{i=1}^{k} 1 \{ (x, u) ; 0 \leq u_i \leq f_i(x) \} \]

which satisfies

\[ \int \cdots \int \overline{\pi}(x, u_{1:k}) du_{1:k} = \pi(x). \]
5.2– Slice Sampling

- In this case the Gibbs sampler satisfies

$$\bar{\pi}(u_{1:k} | x) = \prod_{i=1}^{k} U(\{u_i : u_i \leq f(x)\})$$

$$\bar{\pi}(x | u) = U(\{x : u_1 \leq f_1(x), ..., u_k \leq f_k(x)\}) .$$

- Example: Sample from

$$\pi(x) \propto \left(1 + \sin^2(3x)\right) \left(1 + \cos^4(5x)\right) \exp\left(-\frac{x^2}{2}\right)$$
5.2– Slice Sampling

- We need to sample uniformly from the set

\[
\{x : \sin^2 (3x) \geq 1 - u_1\} \cap \{x : \cos^4 (5x) \geq 1 - u_2\} \cap \{x : |x| \leq \sqrt{-2 \log u_3}\}
\]
5.3– Poisson-log-normal model

- Suppose we have $X \sim \mathcal{N}(0, 1)$ and
  $$Y \mid X \sim \text{Poisson} \left( \exp(X) \right)$$

- The posterior is
  $$\pi(x) \propto \exp(yx - \exp(x)) \exp(-0.5x^2)$$.

- We introduce the following joint density where $u \in (0, \infty)$
  $$\bar{\pi}(x, u) \propto \exp(-u) \mathbb{I}(u > \exp(x)) \exp(-0.5(x^2 - 2yx))$$

  which yields

  $$\bar{\pi}(u \mid x) \propto \exp(-u) \mathbb{I}(u > \exp(x))$$,

  $$\bar{\pi}(u, x) \propto \exp(-0.5(x^2 - 2yx)) \mathbb{I}(x < \log u)$$.
5.4– Discussion

- MCMC is a very active research area with many possibilities and ideas to explore.

- On Thursday, we will discuss another class of methods known as SMC.