Stat 535 C - Statistical Computing & Monte Carlo Methods

Lecture 19 - 21st March 2006

Arnaud Doucet

Email: arnaud@cs.ubc.ca

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 π.

• These moves can be trans-dimensonal 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

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

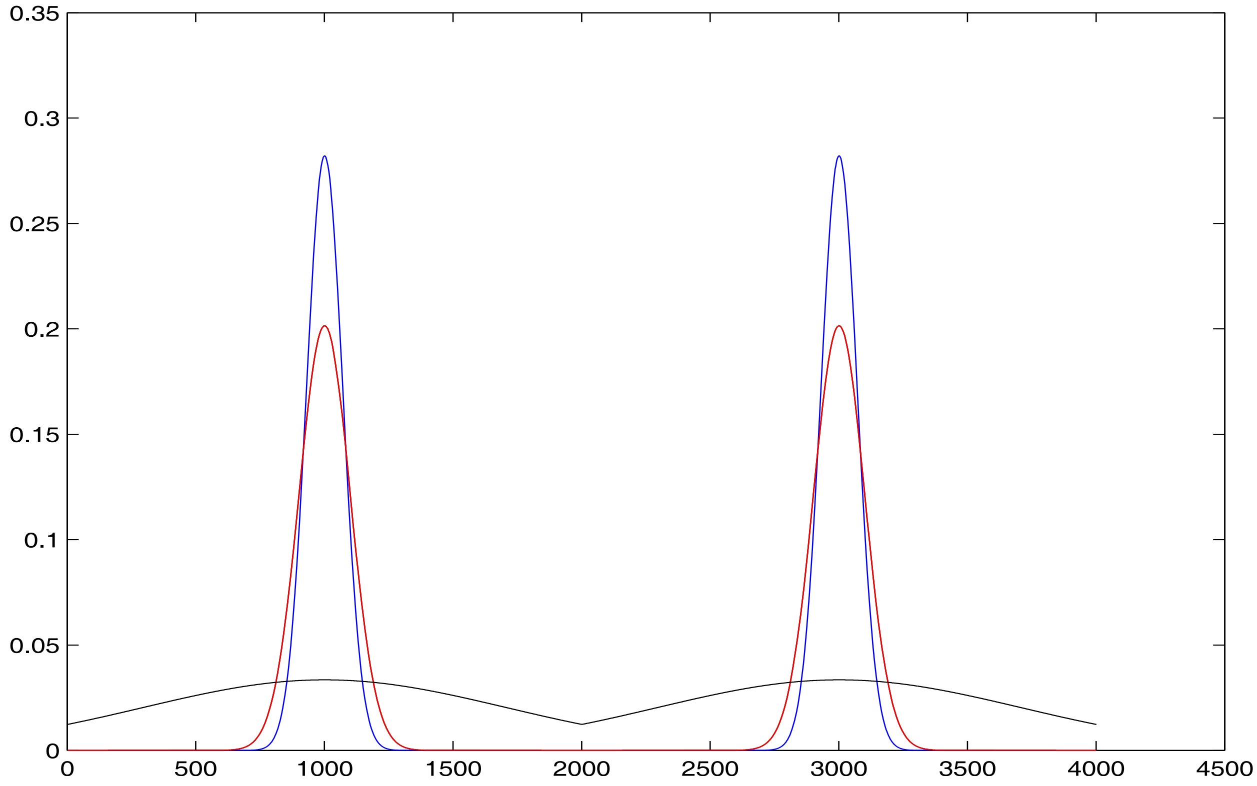
where $\gamma < 1$.

• 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), $\overline{\pi}^{0.5}(x)$ (red) and $\overline{\pi}^{0.01}(x)$ (black)



3.3– Example: Gaussian distribution

- Consider $\pi(x) = \mathcal{N}\left(x; m, \sigma^2\right)$ then $\overline{\pi}^{\gamma}(x) = \mathcal{N}\left(x; m, \sigma^2/\gamma\right)$.

- In one considers a simple random walk MH step then

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

and the acceptance ratio

$$\left(\frac{\pi(x')}{\pi(x)}\right)^{\gamma} \to 1 \text{ as } \gamma \to 0.$$

3.4– Example: Discrete distribution

- Consider a discrete distribution $\pi(x)$ on $\mathcal{X} = \{1, ..., M\}$ then

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

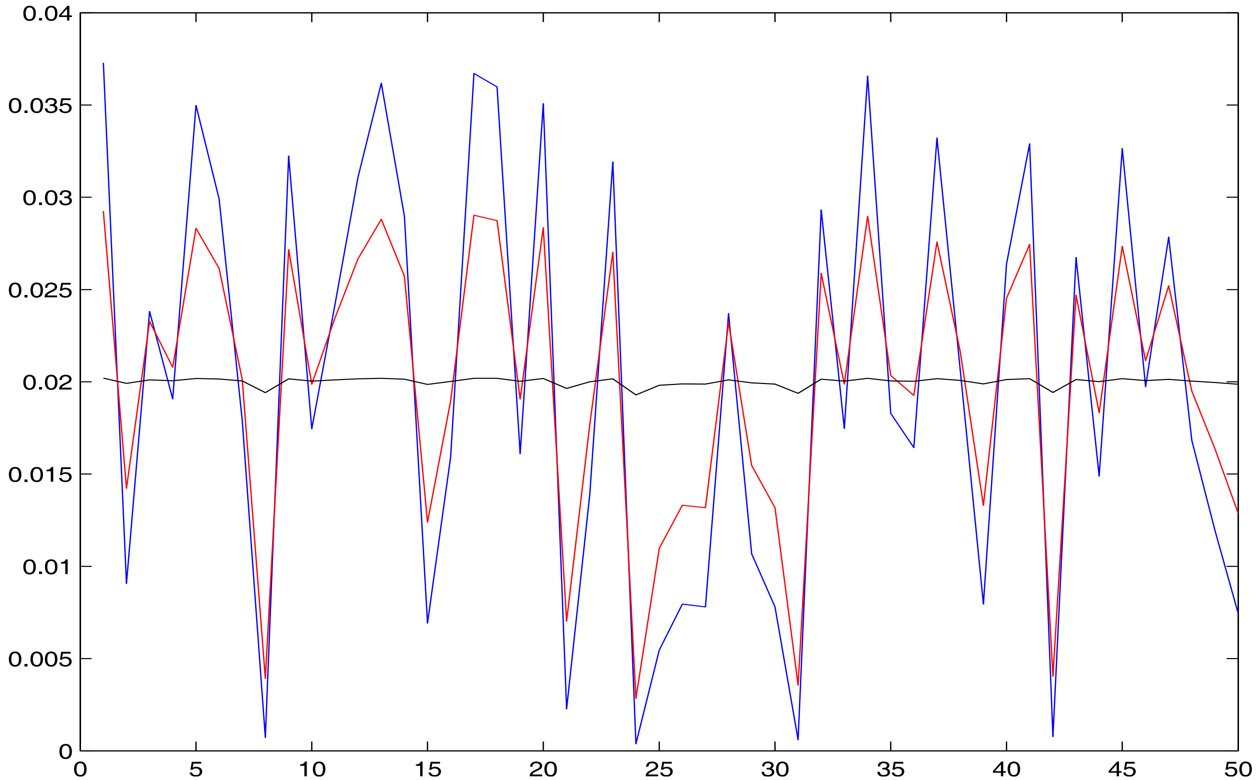
and clearly

$$\overline{\pi}^{\gamma}(x) \to \frac{1}{M}$$

as $\gamma \to 0$.

- It is trivial to sample from a uniform distribution

Representation of $\pi(x)$ (blue), $\overline{\pi}^{0.5}(x)$ (red) and $\overline{\pi}^{0.01}(x)$ (black)



3.6– Sequence of Tempered Distributions

- Instead of using only one auxiliary distribution $\overline{\pi}^{\gamma}(x)$, we will use a sequence of P 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 γ_k is close to 1.

3.8– Simulated tempering

• Alternatively, we could build a target distribution on $\{1, ..., p\} \times \mathcal{X}$ 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

$$\pi(k, x) \propto [f(x)]^{\gamma_k} \text{ where } \pi(x) \propto f(x)$$

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

$$\overline{\pi}\left(x_1,...,x_P\right)=\pi_1\left(x_1\right)\times...\times\pi_P\left(x_P\right)$$

• This seems to be a more difficult problem as the dimension of the new target is higher and includes $\pi_1\left(x_1\right)=\pi\left(x_1\right)$ 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 $\left(X_1^{(i)}, ..., X_P^{(i)}\right)$ independently using

$$K\left(x_{1:P}, x'_{1:P}\right) = \prod_{k=1}^{P} K_i\left(x_i, x'_i\right)$$

where K_i is an MCMC kernel of invariant distribution π_i.

• We can pick two chains associated to π_i and π_j and propose to swap their components, i.e. we propose

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

This is accepted to

$$\alpha\left(x_{1:P}, x'_{1:P}\right) = \min\left(1, \frac{\overline{\pi}\left(x'_{1:P}\right)}{\overline{\pi}\left(x_{1:P}\right)}\right) = \min\left(1, \frac{\pi_i\left(x_j\right)\pi_j\left(x_i\right)}{\pi_i\left(x_i\right)\pi_j\left(x_j\right)}\right).$$

3.11– Alternative Up-and-Down Strategy

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

$$\begin{aligned} X_1' &\sim K_1 \left(X_0', \cdot \right), \; X_2' \sim K_2 \left(X_1', \cdot \right), ..., \; X_P' \sim K_P \left(X_{P-1}', \cdot \right) \\ X_{P-1}^* &\sim K_P \left(X_P', \cdot \right), \; X_{P-2}^* \sim K_{P-1} \left(X_{P-1}^*, \cdot \right), ..., X_0^* \sim K_1 \left(X_1^*, \cdot \right) \end{aligned}$$

where we assume here that K_i is π_i-reversible.

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

$$\min(1, \frac{\pi_1 \left(X_1' \right)}{\pi_0 \left(X_0' \right)} \times \cdots \times \frac{\pi_P \left(X_{P-1}' \right)}{\pi_{P-1} \left(X_{P-1}' \right)} \times \frac{\pi_{P-1} \left(X_{P-1}^* \right)}{\pi_P \left(X_{P-1}^* \right)} \times \cdots \times \frac{\pi_0 \left(X_0^* \right)}{\pi_1 \left(X_0^* \right)})$$

3.11– Alternative Up-and-Down Strategy

• The proof of validity relies on the fact that π-reversibility can easily be checked.Let's write $X_P^* = X'_{P-1}$ then the proposal distribution is

$$
\begin{aligned}
& \pi_0\left(X'_0\right)\prod_{k=1}^{P} K_k\left(X'_{k-1}, X'_k\right)\prod_{k=1}^{P} K_k\left(X_k^*, X_{k-1}^*\right) \\
= \; & \pi_0\left(X'_0\right)\prod_{k=1}^{P} \frac{\pi_k\left(X'_k\right)}{\pi_k\left(X'_{k-1}\right)} K_k\left(X'_k, X'_{k-1}\right)\prod_{k=1}^{P} \frac{\pi_k\left(X_{k-1}^*\right)}{\pi_k\left(X_k^*\right)} K_k\left(X_{k-1}^*, X_k^*\right) \\
= \; & \pi_0\left(X_0^*\right)\prod_{k=1}^{P} K_k\left(X_{k-1}^*, X_k^*\right)\prod_{k=1}^{P} K_k\left(X'_k, X'_{k-1}\right) \\
& \times \frac{\pi_0\left(X'_0\right)}{\pi_1\left(X'_0\right)} \times \cdots \times \frac{\pi_{P-1}\left(X'_{P-1}\right)}{\pi_P\left(X'_{P-1}\right)} \frac{\pi_P\left(X'_{P-1}\right)}{\pi_{P-1}\left(X'_{P-1}\right)} \times \cdots \times \frac{\pi_1\left(X_0^*\right)}{\pi_0\left(X_0^*\right)}
\end{aligned}
$$

- Multiplying by the acceptance probability we have

$$\pi_0\left(X_0'\right)\prod_{k=1}^{P}K_k\left(X_{k-1}',X_k'\right)\prod_{k=1}^{P}K_k\left(X_k^*,X_{k-1}^*\right)$$

$$\times\min(1,\frac{\pi_1\left(X_1'\right)}{\pi_0\left(X_0'\right)}\times\cdots\times\frac{\pi_P\left(X_{P-1}'\right)}{\pi_{P-1}\left(X_{P-1}'\right)}\times\frac{\pi_{P-1}\left(X_{P-1}^*\right)}{\pi_P\left(X_{P-1}^*\right)}\times\cdots\times\frac{\pi_0\left(X_0^*\right)}{\pi_1\left(X_0^*\right)})$$

$$=\pi_0\left(X_0^*\right)\prod_{k=1}^{P}K_k\left(X_{k-1}^*,X_k^*\right)\prod_{k=1}^{P}K_k\left(X_k',X_{k-1}'\right)$$

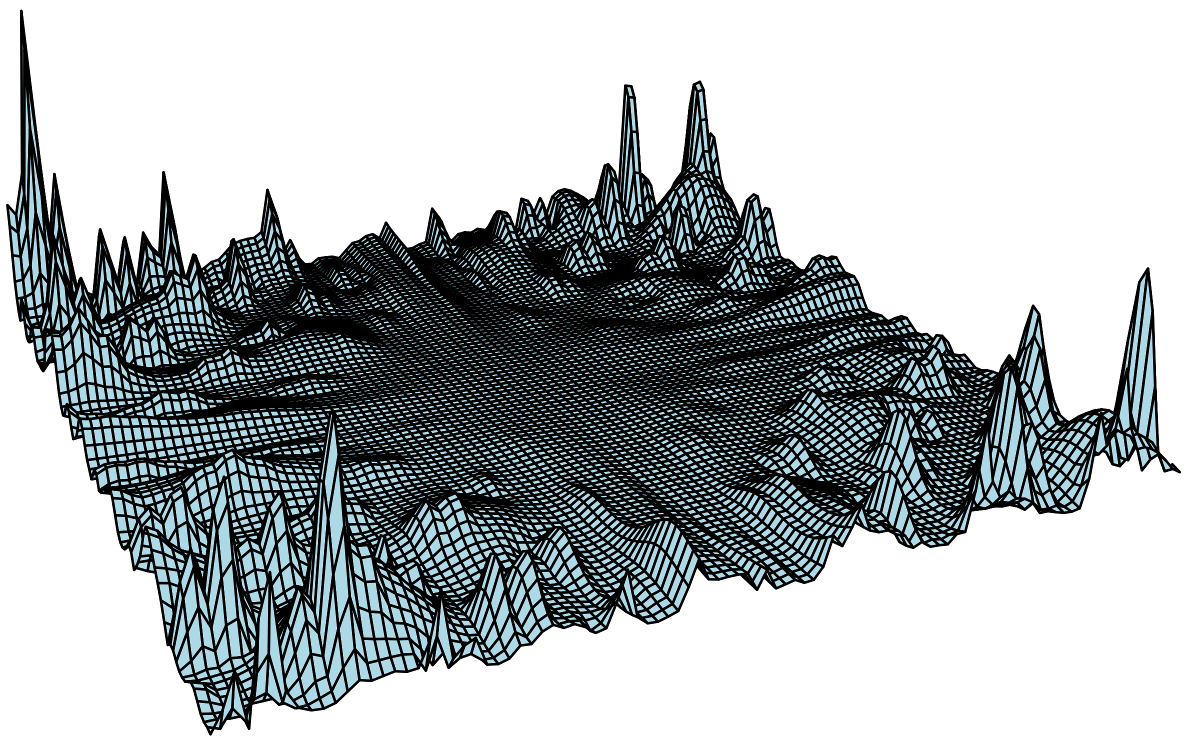
$$\times\frac{\pi_0\left(X_0'\right)}{\pi_1\left(X_0'\right)}\times\cdots\times\frac{\pi_{P-1}\left(X_{P-1}'\right)}{\pi_P\left(X_{P-1}'\right)}\frac{\pi_P\left(X_{P-1}'\right)}{\pi_{P-1}\left(X_{P-1}'\right)}\times\cdots\times\frac{\pi_1\left(X_0^*\right)}{\pi_0\left(X_0^*\right)}$$

$$\times\min(1,\frac{\pi_1\left(X_1'\right)}{\pi_0\left(X_0'\right)}\times\cdots\times\frac{\pi_P\left(X_{P-1}'\right)}{\pi_{P-1}\left(X_{P-1}'\right)}\times\frac{\pi_{P-1}\left(X_{P-1}^*\right)}{\pi_P\left(X_{P-1}^*\right)}\times\cdots\times\frac{\pi_0\left(X_0^*\right)}{\pi_1\left(X_0^*\right)})$$

$$=\pi_0\left(X_0^*\right)\prod_{k=1}^{P}K_k\left(X_{k-1}^*,X_k^*\right)\prod_{k=1}^{P}K_k\left(X_k',X_{k-1}'\right)$$

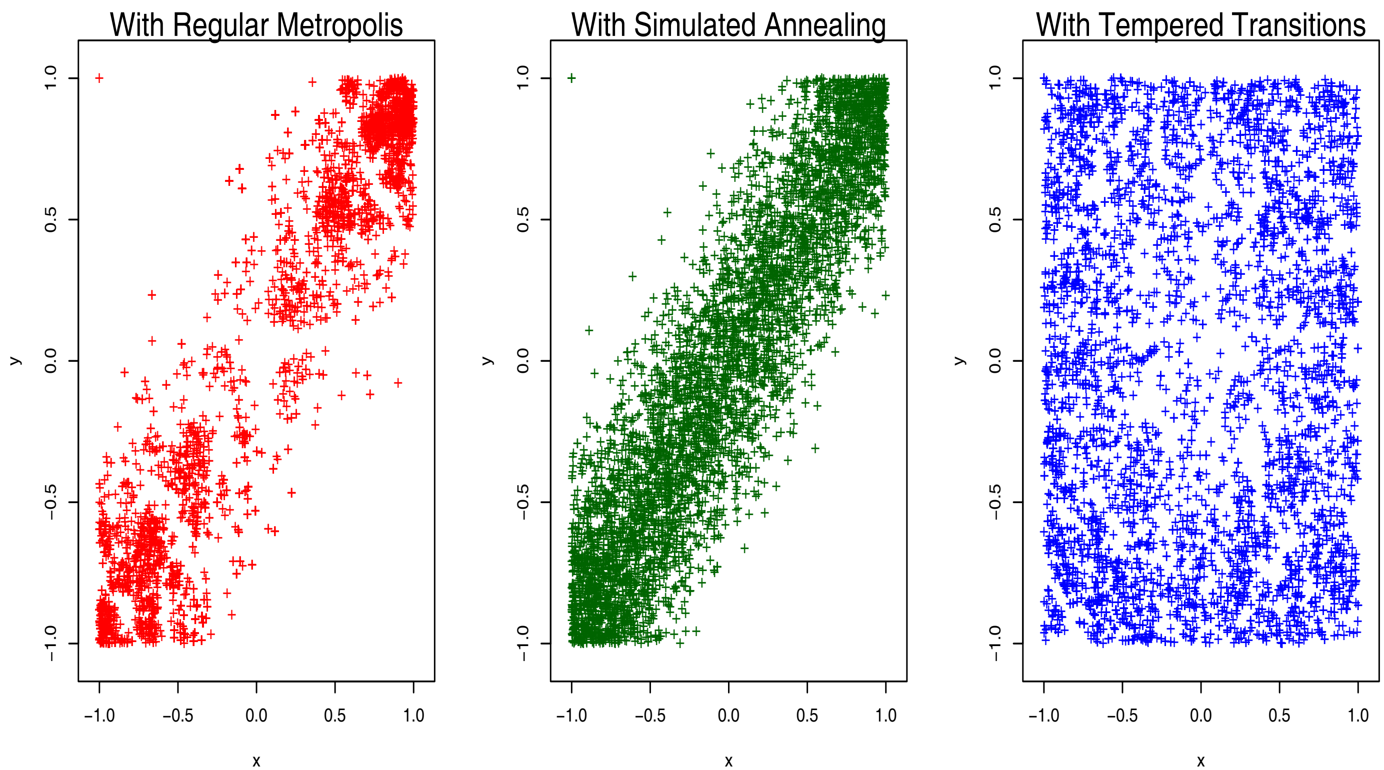
$$\times\min(1,\frac{\pi_0\left(X_0'\right)}{\pi_1\left(X_0'\right)}\times\cdots\times\frac{\pi_{P-1}\left(X_{P-1}'\right)}{\pi_P\left(X_{P-1}'\right)}\frac{\pi_P\left(X_{P-1}'\right)}{\pi_{P-1}\left(X_{P-1}'\right)}\times\cdots\times\frac{\pi_1\left(X_0^*\right)}{\pi_0\left(X_0^*\right)})$$

3.12– Application to A Complex Toy Example



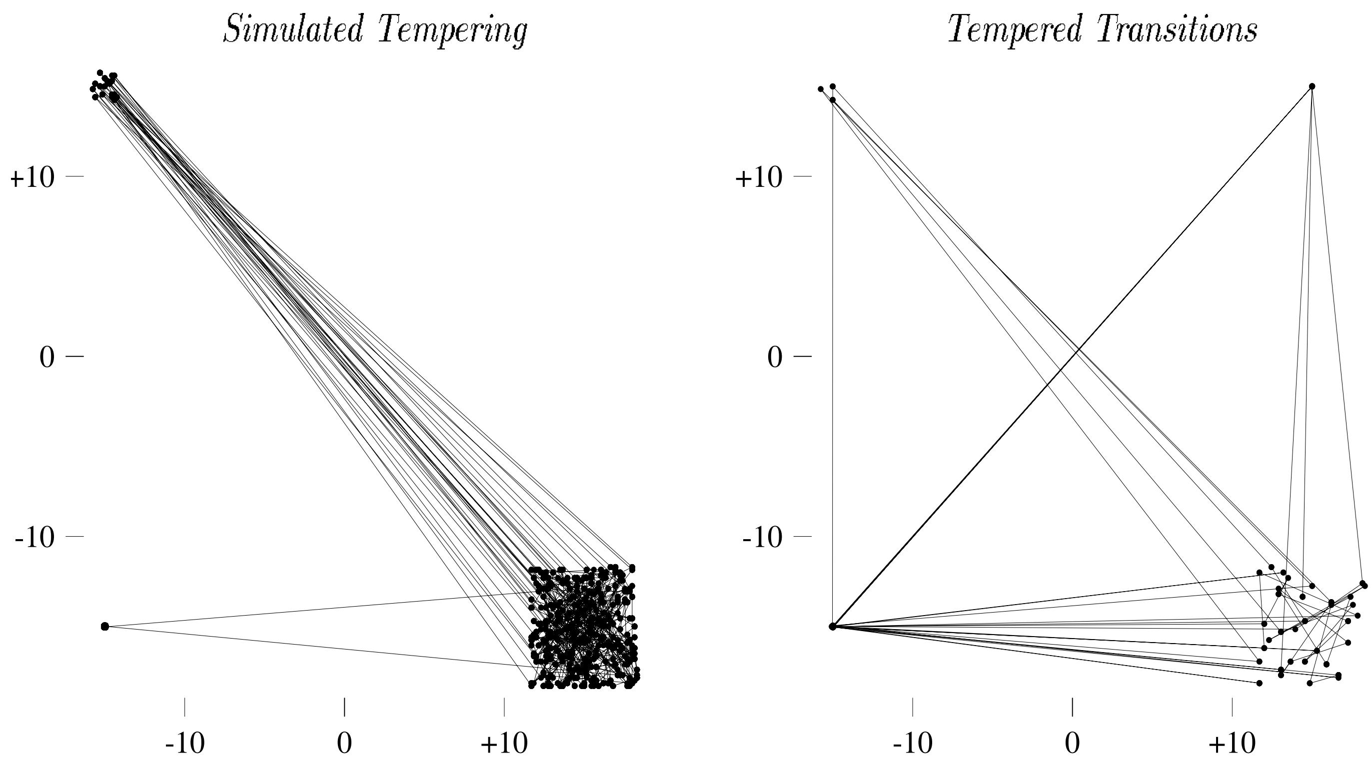
Artificial Target Distribution on $(-1, 1) \times (-1, 1)$

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)

3.14– Discussion

- 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.

- An idea closely related to tempering is annealing.
- We have seen that

$$\overline{\pi}^{\gamma}(x) \propto [\pi(x)]^{\gamma}$$

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

- On the contrary, $\overline{\pi}^{\gamma}(x)$ is a peakened version of the target as γ increases.

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

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

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

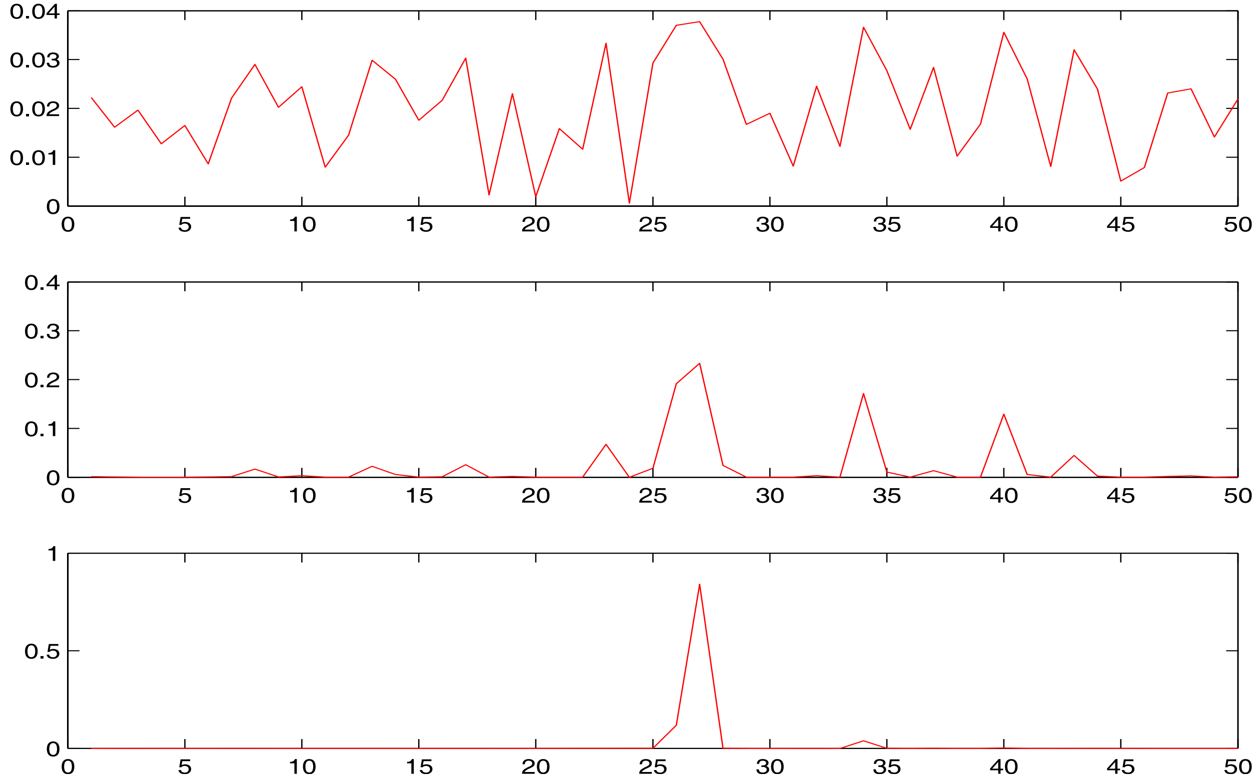
then

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

as for any $x \neq x^{*}$

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

Representation of $\pi(x)$ (top), $\overline{\pi}^{10}(x)$ (middle) and $\overline{\pi}^{100}(x)$ (bottom)



• Similarly in the continuous case, one can show that

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

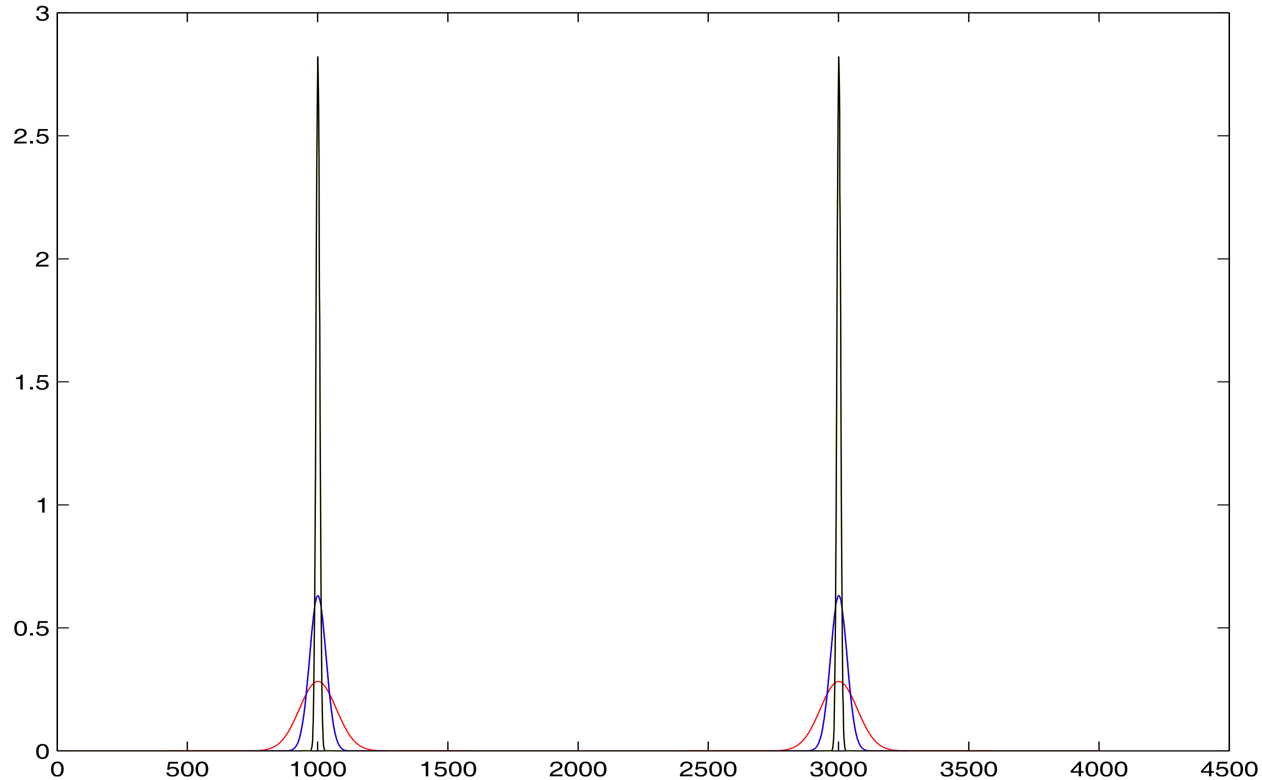
• If one could sample from $\overline{\pi}^{\gamma}(x)$ for large γ (asymptotically $\gamma \to \infty$) then we could solve any global optimization problem! Indeed maximizing any function $f : \mathcal{X} \to \mathbb{R}$ would be equivalent to sample

$$\overline{\pi}^{\gamma}(x) \propto [\exp(f(x))]^{\gamma}$$

where we have $\gamma \to \infty$.

• As γ increases, sampling from $\overline{\pi}^{\gamma}(x)$ is becoming harder. If it was simple, global optimization problem could be solved easily.

Representation of $\pi(x)$ (red), $\overline{\pi}^{10}(x)$ (blue) and $\overline{\pi}^{100}(x)$ (black)



4.4– Graphical illustration

• To sample from $\overline{\pi}^{\gamma}(x)$ for a large γ, 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 π_k.

4.4– Graphical illustration

- 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 π_∞ as $k \to \infty$ you need to have conditions such as

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

- The second condition is not realistic, γ_k increases too slowly and in practice we select γ_k growing polynomially.

5.1– Hybrid Monte Carlo

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

- *Hybrid Monte Carlo*: Define

$$\pi(x, y) \propto \pi(x) \exp\left(-\beta y^{\mathsf{T}} y\right)$$

- *Basis*: It is possible to move approximately on the manifold defined by $\pi(x, y)$ =cst. See tutorial paper by Stoltz & al.

- 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.

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

$$\begin{aligned}\overline{\pi}(x|u) &= \mathcal{U}(\{x : u \leq f(x)\}),\\ \overline{\pi}(u|x) &= \mathcal{U}(\{u : u \leq f(x)\}).\end{aligned}$$

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

- MH step can be used to sample from $\overline{\pi}(u|x)$.

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

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

and

$$u \leq \frac{1}{2}\exp\left(-\sqrt{x}\right) \Leftrightarrow 0 \leq x \leq \left[\log\left(2u\right)\right]^2$$

then

$$X|\,u \sim \mathcal{U}\left(0, \left[\log\left(2u\right)\right]^2\right)$$

• 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

$$\int \cdots \int \overline{\pi}(x, u_{1:k})\, du_{1:k} = \pi(x).$$

- In this case the Gibbs sampler satisfies

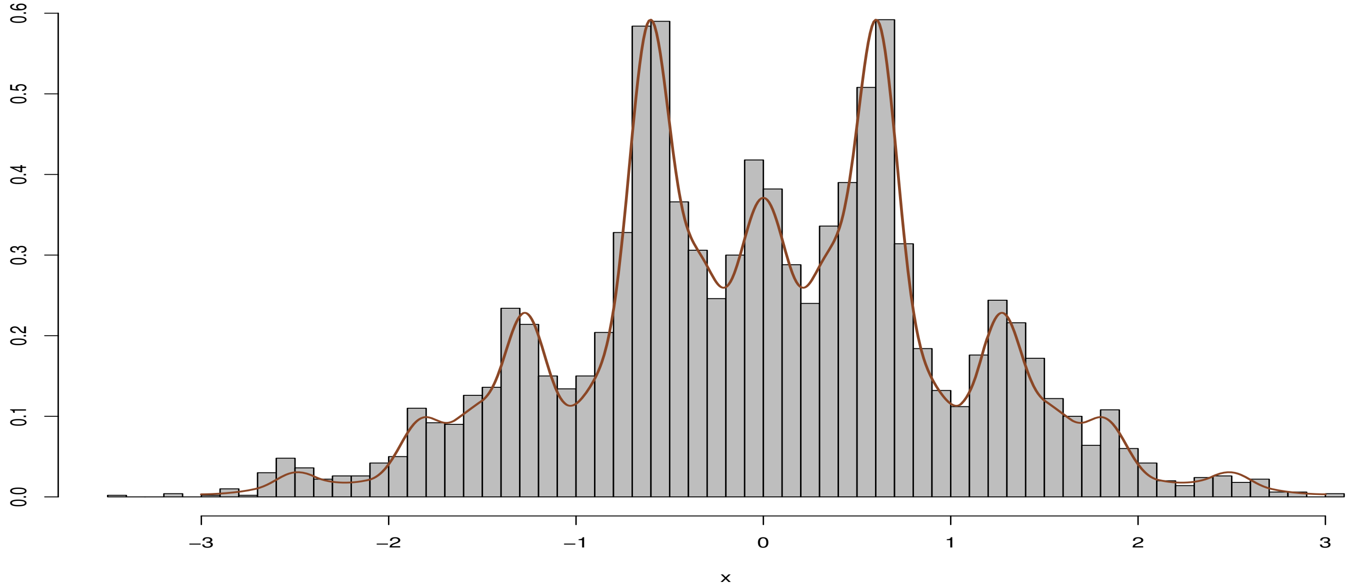
$$\begin{aligned}\overline{\pi}\left(u_{1:k}\right|x) &= \prod_{i=1}^{k}\mathcal{U}\left(\{u_i : u_i \leq f\left(x\right)\}\right) \\ \overline{\pi}\left(x\right|u) &= \mathcal{U}\left(\{x : u_1 \leq f_1\left(x\right), ..., u_k \leq f_k\left(x\right)\}\right).\end{aligned}$$

- *Example*: Sample from

$$\pi\left(x\right) \propto \underbrace{\left(1+\sin^2\left(3x\right)\right)}_{f_1(x)}\underbrace{\left(1+\cos^4\left(5x\right)\right)}_{f_2(x)}\underbrace{\exp\left(-\frac{x^2}{2}\right)}_{f_3(x)}$$

- We need to sample uniformly from the set

$$\left\{x : \sin^2(3x) \geq 1 - u_1\right\} \cap \left\{x : \cos^4(5x) \geq 1 - u_2\right\} \cap \left\{x : |x| \leq \sqrt{-2\log u_3}\right\}$$



5.3– Poisson-log-normal model

- Suppose we have $X \sim \mathcal{N}(0,1)$ and
$$Y|\,X \sim P\text{oisson}\,(\exp(X))$$

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

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

which yields

$$\begin{aligned}
\overline{\pi}(u|\,x) &\propto \exp(-u)\,\mathbb{I}(u > \exp(x)),\\
\overline{\pi}(u,x) &\propto \exp\left(-0.5\left(x^2 - 2yx\right)\right)\mathbb{I}(x < \log u).
\end{aligned}$$

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.