Bayesian Statistics, Monte Carlo Methods

June 7, 2017

- 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.

Introducing auxiliary distributions

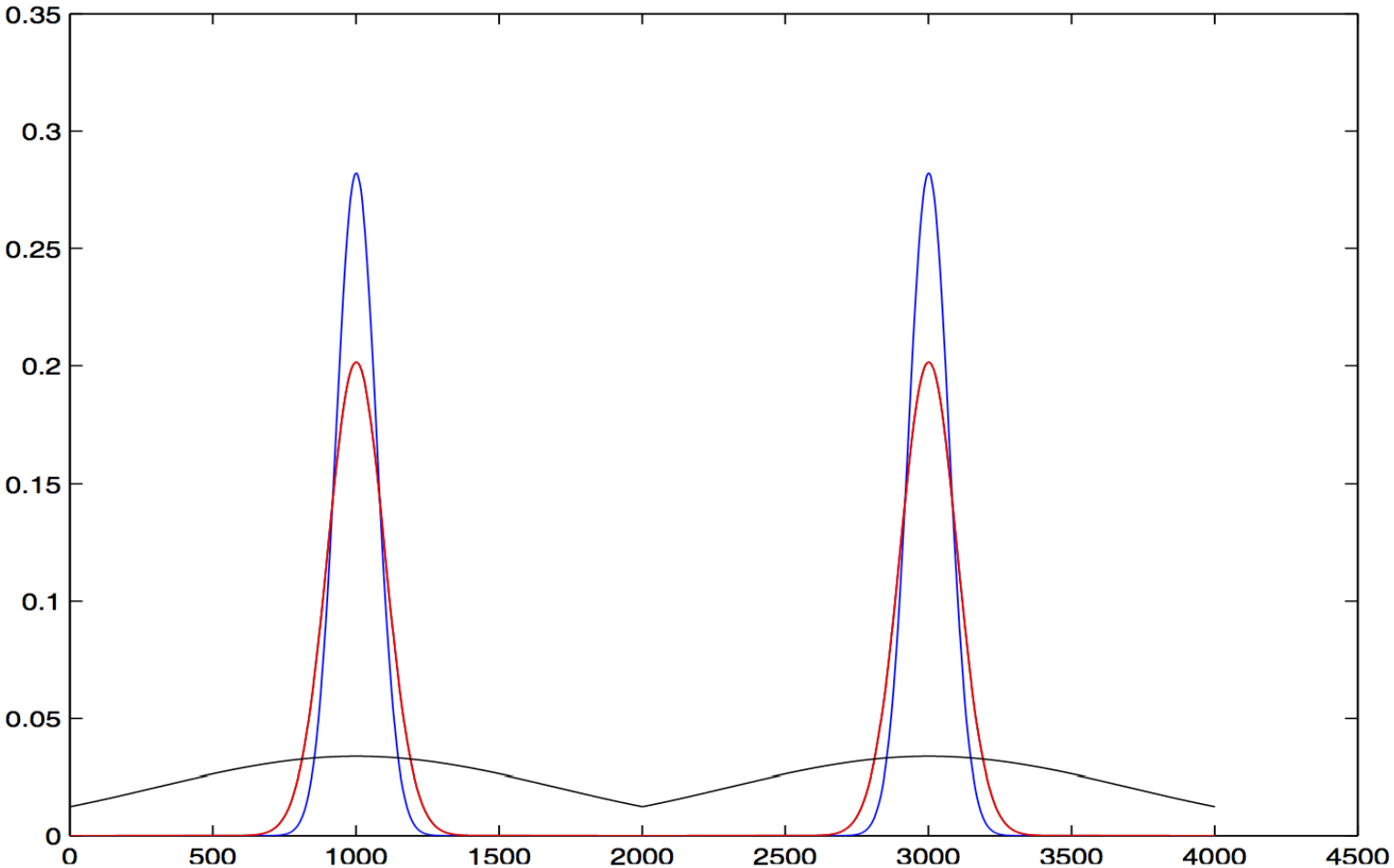
- The key is to notice that although it might be difficult to sample from $\pi(x)$, it could be easier to sample from related distributions.
- In particular, it should be easier to sample from

$$\bar{\pi}^{\gamma}(x) = \frac{(\pi(x))^{\gamma}}{\int (\pi(x))^{\gamma} dx} \propto (\pi(x))^{\gamma}$$

  where $\gamma < 1$.
- For $\gamma < 1$ the target $\bar{\pi}^{\gamma}(x)$ is flatter than $\pi(x)$, hence easier to sample from.
- This is called tempering.

Graphical illustration

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



Example: Gaussian distribution

- Consider $\pi(x) = \mathcal{N}(x; m, \sigma^2)$, then $\bar{\pi}^\gamma(x) = \mathcal{N}(x; m, \sigma^2/\gamma)$.
- If one considers a simple random walk MH step then

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

and the acceptance ratio $(\frac{\bar{\pi}(x')}{\bar{\pi}(x)})^\gamma \to 1$, as $\gamma \to 0$.

Example: Discrete distribution

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

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

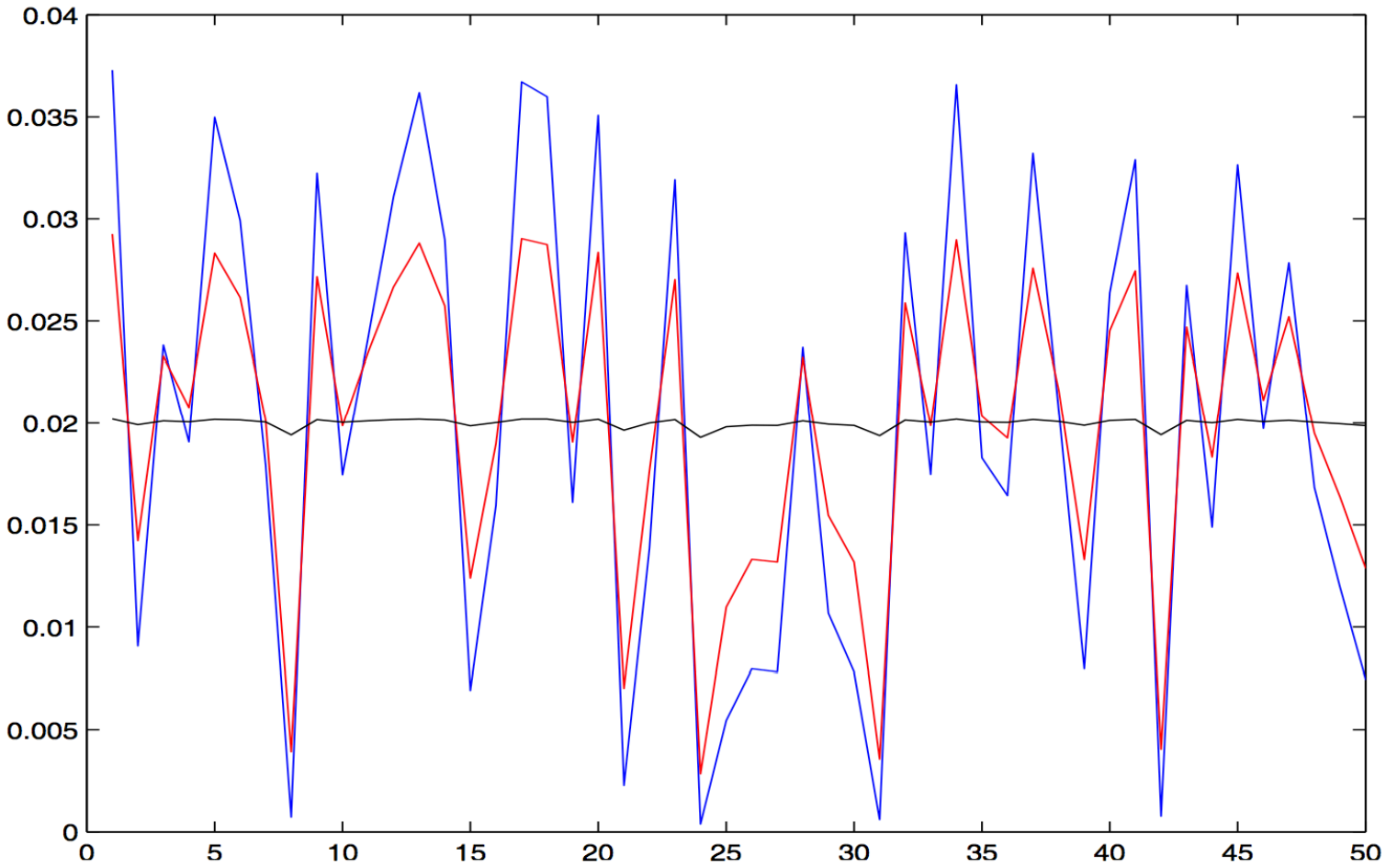
  and clearly,

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

  as $\gamma \to 0$.
- It is trivial to sample from a uniform distribution.

Graphical illustration

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



Sequence of Tempered Distributions

- Instead of using only one auxiliary distribution $\bar{\pi}^{\gamma}(x)$, we will use a sequence of P distribution defined as

$$\pi^{k}(x) \propto [\pi(x)]^{\gamma_k}$$

  where $\gamma_1 = 1$, $\gamma_k < \gamma_{k-1}$.
- In this case $\pi_1(x) = \pi(x)$ and $\pi_k(x)$ is a sequence of distributions increasingly simpler to sample.

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
$\pi^N(x) = \sum_{i=1}^N W_k^{(i)} \delta_{X_k^{(i)}}$, where $W_k^{(i)} \propto (\pi(X_k^{(i)}))^{(1-\gamma_k)}$.

- This idea is simple and will work properly if γ_k is close to 1.

Simulated tempering

- Suppose $\pi(x)$ is a mixture distribution of p distributions.
- Alternatively, we could build a target distribution on $\{1, \cdots, p\} \times \mathcal{X}$ defined as

$$\pi(k, x) = c_k \pi_k(x)$$

- Then we could propose deterministic moves like jumping from dimension k to 1 accepted with probability

$$\min(1, \frac{\pi(1, x)}{\pi(k, x)})$$

- 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

$$c_k \propto \int [f(x)]^{\gamma_k} dx$$

and you might biase unnecessarily the time spent in high temperatures.

Parallel tempering

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

$$\bar{\pi}(x_1, \cdots, x_P) = \pi_1(x_1) \times \cdots \times \pi_P(x_P)$$

- This seems to be a more difficult problem as the dimension of the new target is higher.
- The advantage is that we can design clever moves and use sample from “hot” chains to feed the “cold” chain.

Swap moves

- We can have a simple update kernel which updates each component of the Markov chain $(X_1^{(i)}, \cdots, X_P^{(i)})$ independently using

$$K(x_{1,\cdots,P}, x'_{1,\cdots,P}) = \prod_{i=1}^{P} K_i(x_i, x'_i)$$

  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

$$x'_{-(i,j)} = x_{-(i,j)}, x'_i = x_j, x'_j = x_i$$

  This is accepted to

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

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

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

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

  where we assume here that K_i is π_i-reversible.
- The acceptance rate for the candidate X'_{2P-1} is given by

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

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

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

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

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

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

Alternative Up-and-Down Strategy

- Multiplying by the acceptance probability we have

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

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

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

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

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

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

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

Simulated Annealing

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

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

  is a flattened version of $\pi(x)$ when $\gamma < 0$.
- On the contrary, $\bar{\pi}^{\gamma}(x)$ is a peakened version of the target as γ increases.

Simulated Annealing

- Under regularity conditions, it can be shown that the support of $\bar{\pi}^{\gamma}(x)$ concentrates itself on the set of global maxima of $\pi(x)$.
- In the discrete case, let us write the unique maximum

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

then

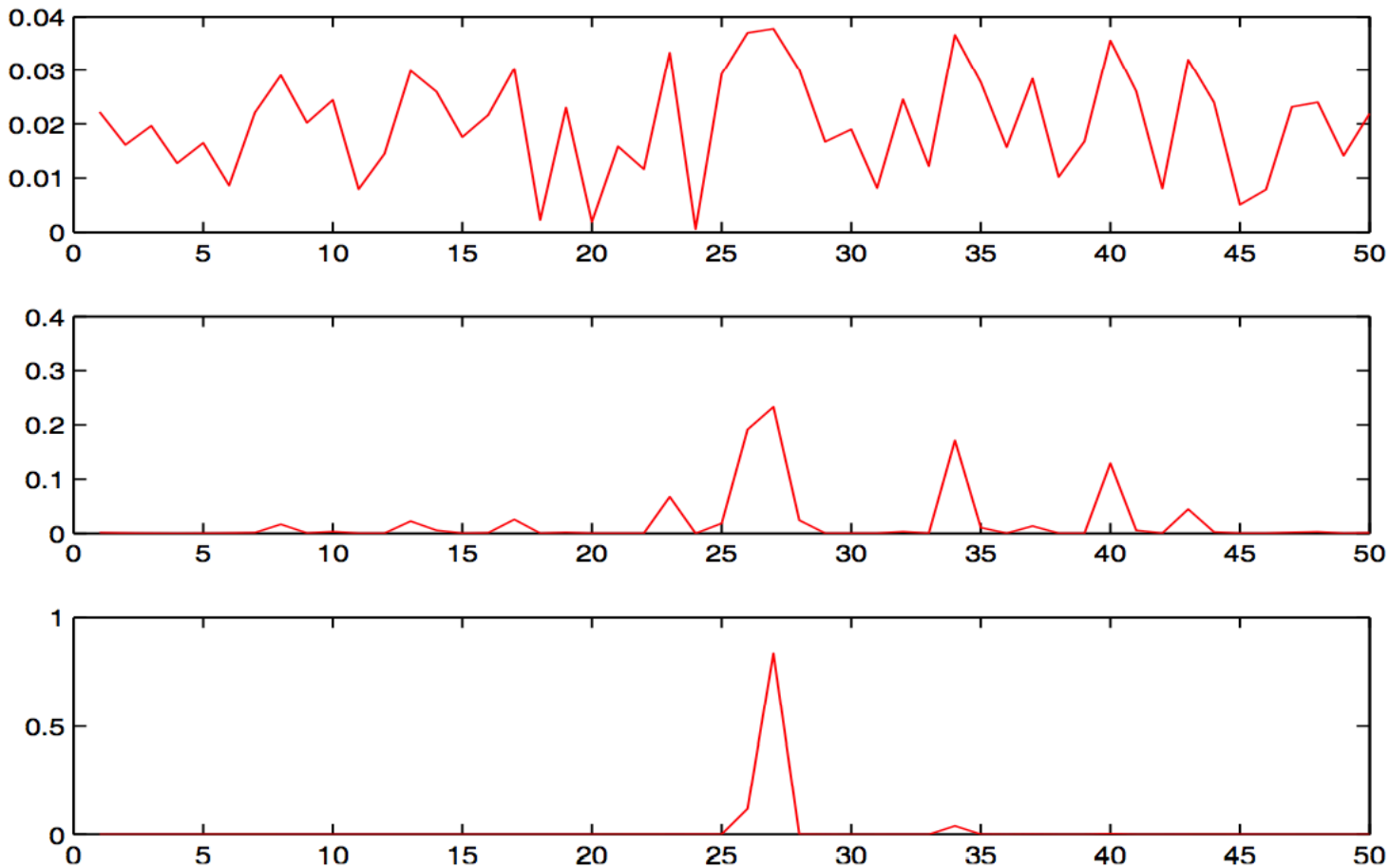
$$\lim_{\gamma \to \infty} \bar{\pi}^{\gamma}(x^*) = 1$$

as for any $x \neq x^*$

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

Graphical illustration

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

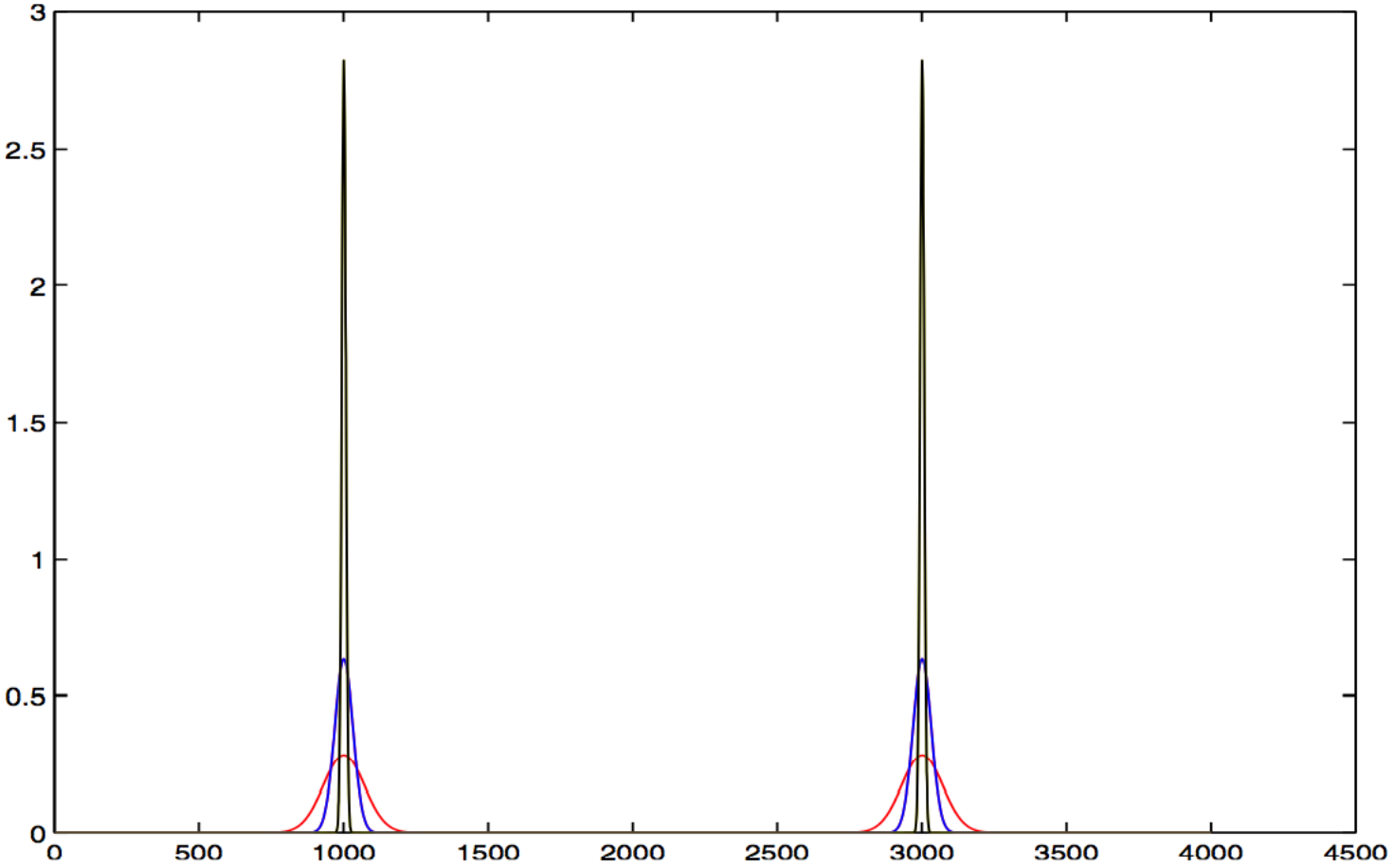


Graphical illustration

- If one could sample from $\bar{\pi}^{\gamma}(x)$ for large γ (asymptotically $\gamma \to \infty$) then we could solve any global optimization problem! Indeed maximizing any function $f : \mathcal{X} \to R$ would be equivalent to sample $\bar{\pi}^{\gamma}(x) \propto [\exp(f(x))]^{\gamma}$ where we have $\gamma \to \infty$.
- As γ increases, sampling from $\bar{\pi}^{\gamma}(x)$ is becoming harder. If it was simple, global optimization problem could be solved easily.

Graphical illustration

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



Graphical illustration

- To sample from $\bar{\pi}^{\gamma}(x)$ 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 $\gamma_1 \gg 1$.
- 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.

Simulated annealing

Use a more complex Evaluation Function:

- Do sometimes accept candidates with higher cost to escape from local optimum
- Adapt the parameters of this Evaluation Function during execution
- Based upon the analogy with the simulation of the annealing of solids

Analogy

- Slowly cool down a heated solid, so that all particles arrange in the ground energy state
- At each temperature wait until the solid reaches its thermal equilibrium
- Probability of being in a state with energy E:

$$Pr\{\mathbf{E} = E\} = 1/Z(T)\exp(-E/k_B \cdot T)$$

E: Energy; T:Temperature;
k_B: Boltzmann constant; $Z(T)$: normalizing factor

Simulation of cooling:

- At a fixed temperature T :
- Perturb (randomly) the current state to a new state
- ΔE is the difference in energy between current and new state
- If $\Delta E < 0$ (new state is lower), accept new state as current state
- If $\Delta E \geq 0$, accept new state with probability

$$Pr(accepted) = \exp(-\Delta E / k_B \cdot T)$$

- Eventually the systems evolves into thermal equilibrium at temperature T; then the formula mentioned before holds
- When equilibrium is reached, temperature T can be lowered and the process can be repeated

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

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

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

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 $\pi(x, y) = \textit{constant}$. See tutorial paper by Stoltz & al.

Slice Sampling

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

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

- By construction, we have

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

- Note that the same representation was implicitly used in Rejection sampling.

Slice Sampling

- To sample from $\bar{\pi}(x, u)$ 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 $\bar{\pi}(u|x)$.

Slice Sampling

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

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

and

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

then

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

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

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

  which satisfies

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

Slice Sampling

- In this case the Gibbs sampler satisfies

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

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

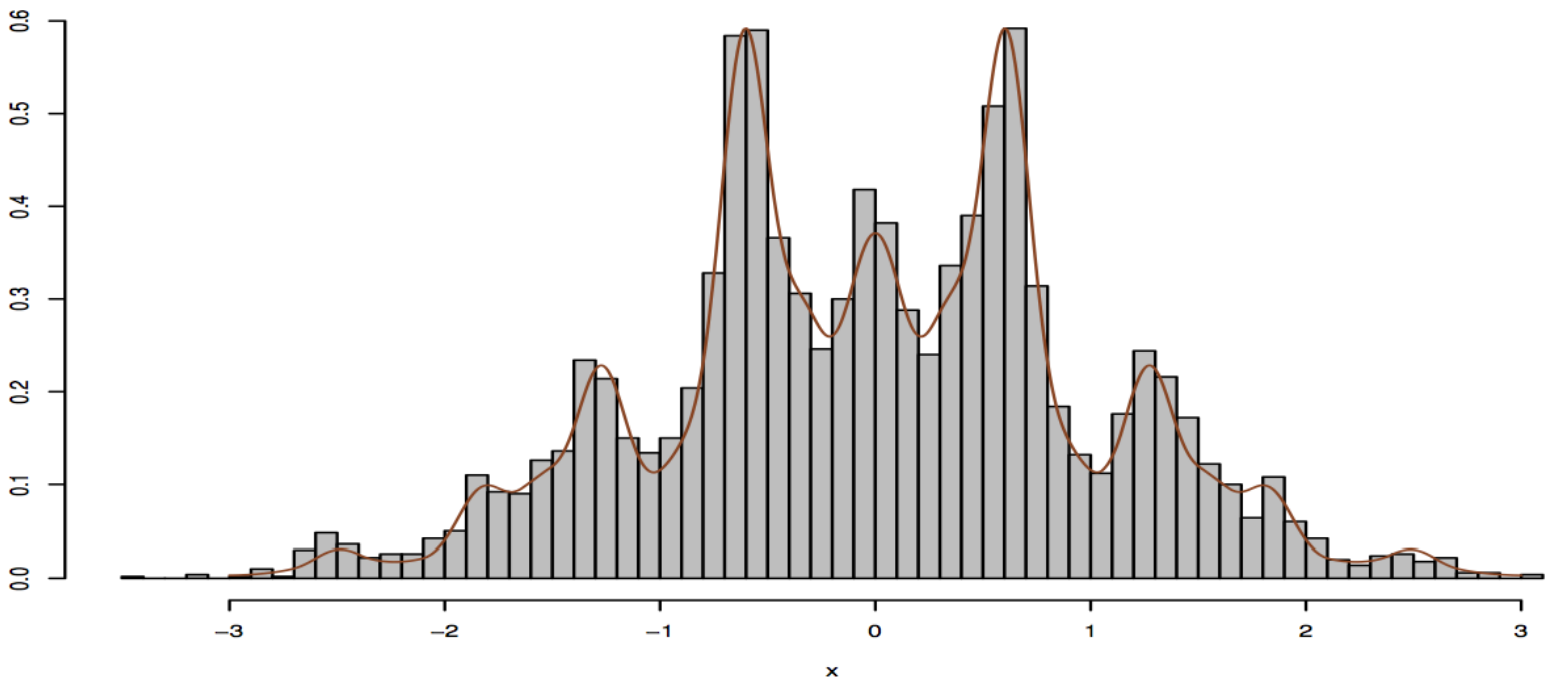
- Example: Sample from

$$\pi(x) \propto (1 + sin^2(3x))(1 + cos^4(5x)) \exp(-\frac{x^2}{2})$$

Slice Sampling

- We need to sample uniformly from the set

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



Poisson-log-normal model

- Suppose we have $X \sim \mathcal{N}(0, 1)$ and

$$Y|X \sim Poisson(\exp(X))$$

- The posterior is

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

- We introduce the following joint density where $u \in (0, \infty)$,

$$\bar{\pi}(x, u) \propto \exp(-u) I(u > \exp(x)) \exp(-0.5x^2 - 2yx))$$

which yields

$$\bar{\pi}(u|x) \propto \exp(-u) I(u > \exp(x))$$

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