Metropolis Hasting Chain and Simulated Annealing

Hongda Li

UBC Okanagan

December 4, 2022

ToC

The MHC

MHC: Metropolis Hasting Chain

Input: $X^{(t)}$

$Y^{(t)} \sim q(\cdot|x^{(t)})$

$$\rho(x,y) := \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$$

$$X^{(t+1)} := \begin{cases} Y^{(t)} & \text{w.p} : \rho(X^{(t)}, Y^{(t)}) \\ X^{(t)} & \text{else} \end{cases}$$

1. $q(x|y)$ is doubly stochastic.
2. $f(x)$ is a distribution function up to a constant.
3. Must have $f(X^{(t)}) > 0$.

What it does

Let $X^{(t)}$ be a sequence of observations sampled from the MHC, then $X^{(t)}$ will approximate f.

1. No integrals are needed.
2. It works very well for distribution functions in a very high dimension.
3. It is not hard to implement it on a computer.

Primary questions

We have questions:

- Is f a stationary distribution for the MHC?
  - Yes.
- Does it converge to the stationary distributions f?
  - Sometimes. We need some regularity conditions.

To converge to f the MHC must satisfy the following:

1. f is a stationary distribution of the MHC.
2. All states in $\text{supp}(f)$ can be commuted with each other. (f-Irreducible)
3. All states are aperiodic.

The transition kernel

The transition kernel for MHC

$$K(x,y) = \rho(x,y)q(y|x) + \left(1 - \underbrace{\sum_{z \in S \setminus \{y\}} \rho(x,z)q(z|x)}_{=:r(x)}\right) \mathbb{1}\{y = x\}.$$

1. When q is doubly stochastic, f satisfies detail balance. (f-stationary)
2. We have $K(x,x) > 0$ for all x such that $f(x) > 0$. It is aperiodic.
3. It is f-irreducible if we assume $q(x|y)$ is non-negative for all $x, y \in \text{supp}(f)$.
4. See Robert and Casella's book [1] for the case where state space is continuous.

Regularity conditions

1. $X^{(t)}$ needs to be able to travel to all states in $x \in \text{supp}(f)$.
2. And this is possible if $q(x|y) > 0 \; \forall x, y \in S$.
3. Weaker conditions exist, and we might have to do that in a case-by-case basis. We need to have:

$$\forall x, y \in S \times S \; \exists n < \infty : K^n(x, y) > 0.$$

Sampling

The function we are sampling is:

$$D := \{(x_1, x_2) : -\sin(4\pi x_1) + 2\sin(2\pi x_2)^2 > 1.5\}$$
$$f(x) := \mathbb{1}_D(\sin(x_1 4\pi) + \cos(x_2 4\pi) + 2),$$

on $[0,1] \times [0,1]$, and we are considering two choices of base chain:

1. A uniform random base chain, where it is just a random jump, is not a Markov chain.
2. A wrapped Guassian random walks, where $Y^{(t+1)} \sim \text{WrappedNormal}(X^{(t)}, 0.1)$. (Or equivalently, Guassian random walks but with periodic boundary conditions.)

The sampling results

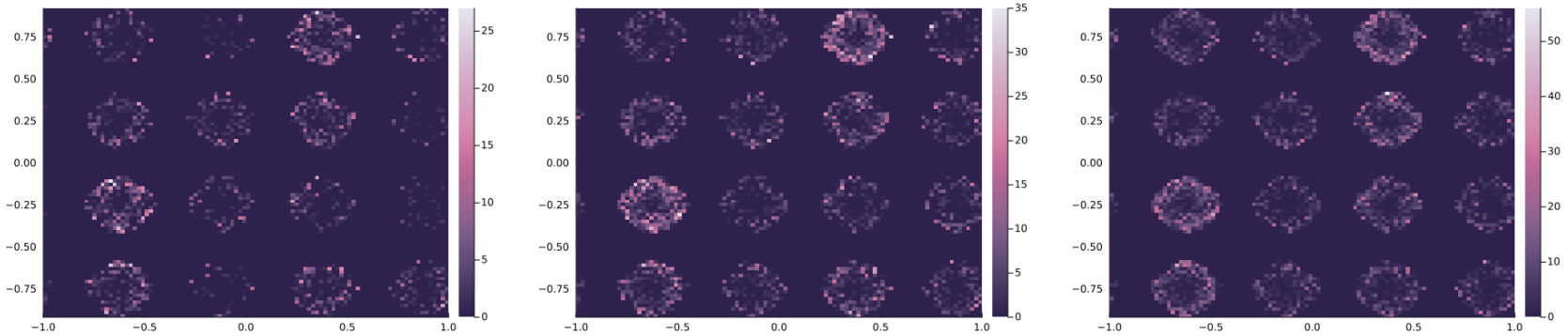

Figure: Snapshots of accumulated samples when a wrapped Gaussian random walk base chain. The sampling is not quite even.

The sampling results

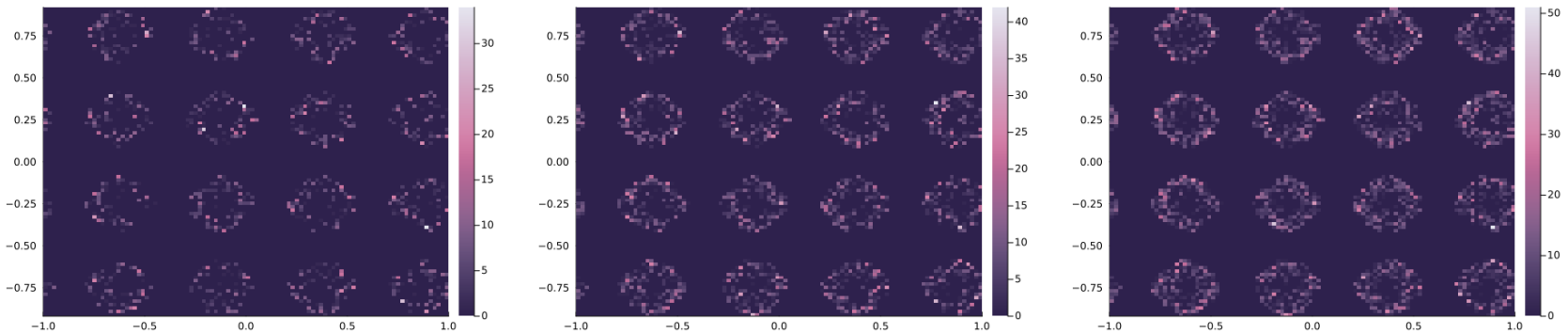

Figure: Snapshots of the accumulated samples when a uniform random base chain over $[0, 1] \times [0, 1]$. The sampling is quite even.

Brief summary

1. It converges slowly if the base chain can not make $X^{(t)}$ travel from anywhere to everywhere.
2. The uniform base chain with an indicator function is the same Monte Carlo method.
4. If $\text{supp}(f)$ are disconnected, we need a base chain q that is "aggressive" enough to make the jump, keeping MHC f-irreducible.
5. A combination of both is the key to efficiency. (In practice, we add these 2 base chains together, weighted by positive constants)

Simulated Annealing

MHC can be used to maximizes $g : S \mapsto \mathbb{R}_+ \cup \{-\infty\}$ through $f(x) = \exp(g(x)/T_i)$. T_i is called a temperature. The lowering temperature accentuates the maximums of the function. Let X^* denotes he set of maximizers for f then as temperature goes to 0, f has:

$$\begin{aligned}
\lim_{i\to\infty} f(x) &= \lim_{i\to\infty} \frac{\exp(g(x)/T_i)}{\sum_{y\in S}\exp(g(y)/T_i)} \\
&= \lim_{i\to\infty} \frac{1}{\sum_{y\in S}\exp(g(y)-g(x)/T_i)} \\
&= \lim_{i\to\infty} \frac{1}{\exp(0)+\sum_{y\in S\setminus\{x\}}\exp(g(y)-g(x)/T_i)} \\
&= \lim_{T\to 0} \frac{1}{1+\sum_{y\in S\setminus\{x\}}\exp(g(y)-g(x)/T)} \\
&= \mathbb{1}\{X^*\}.
\end{aligned}$$

Temperature illustrations

Let's make $g(x) := \text{sinc}(x)$ and plot with different temperatures:

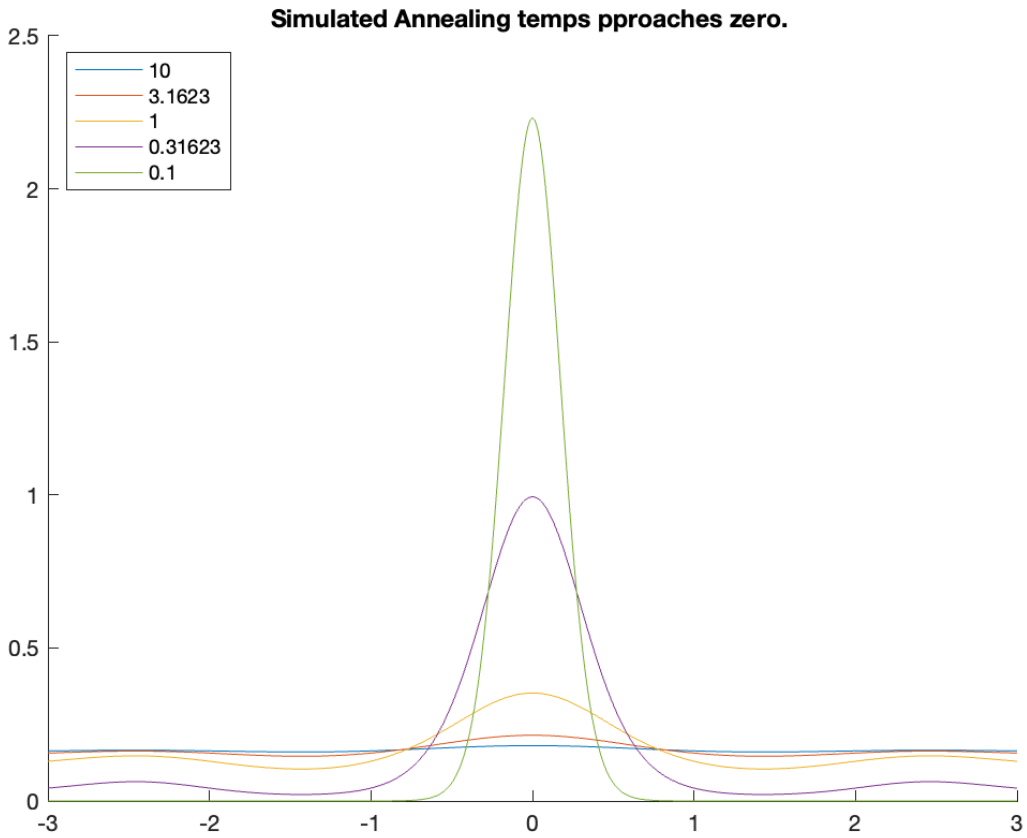

Figure: Different temperature when $g(x) = \text{sinc}(x)$, and $f := \exp(g(x)/T_i)$.

Temperatures and questions

1. Temperature helps with the convergence of Simulated annealing. As it gets lower, the MHC is more and more likely to reject sub-optimal solutions. Making it converges to a local optimal.
2. How would one choose the cooling schedule?
3. On top of all these, the choice of base chain is still a concern.

Bin packing problem

Bin packing problem finds subsets from a given set such that it sums up as close to 1 as possible while still having it less than 1. It is phrased as:

$$g(x) := \begin{cases} \langle w, x \rangle & \text{if } \langle w, x \rangle \leq 1, \\ -\infty & \text{else.} \end{cases}$$
$$f(x) := \exp(g(x)/T_i)$$
$$S := \{0, 1\}^n.$$

The first 100 elements of w sum up to 1 exactly, but the next 100 elements are sampled from $\sim \text{Uniform}(1, 2)$.

Avoiding curse of dimensionality

Using the base chain that mutates precisely one bit and flips that bit in x, we can avoid the curse of dimensionality. We consider the following three types of temperature schedules for our experiment:

1. The temperature drops quadratically from 1 to 0.001 every 1000 iterations, 10k iterations in total.
2. The temperature is $exp(-k)$ where k goes from 0 to 10. It decreases every 1000 iterations in a total of 10k iterations.
3. The temperature is $1/k^2$; it drops whenever the algorithm discovers a solution lower than all previous solutions for the objective function.

Numerical experiments results

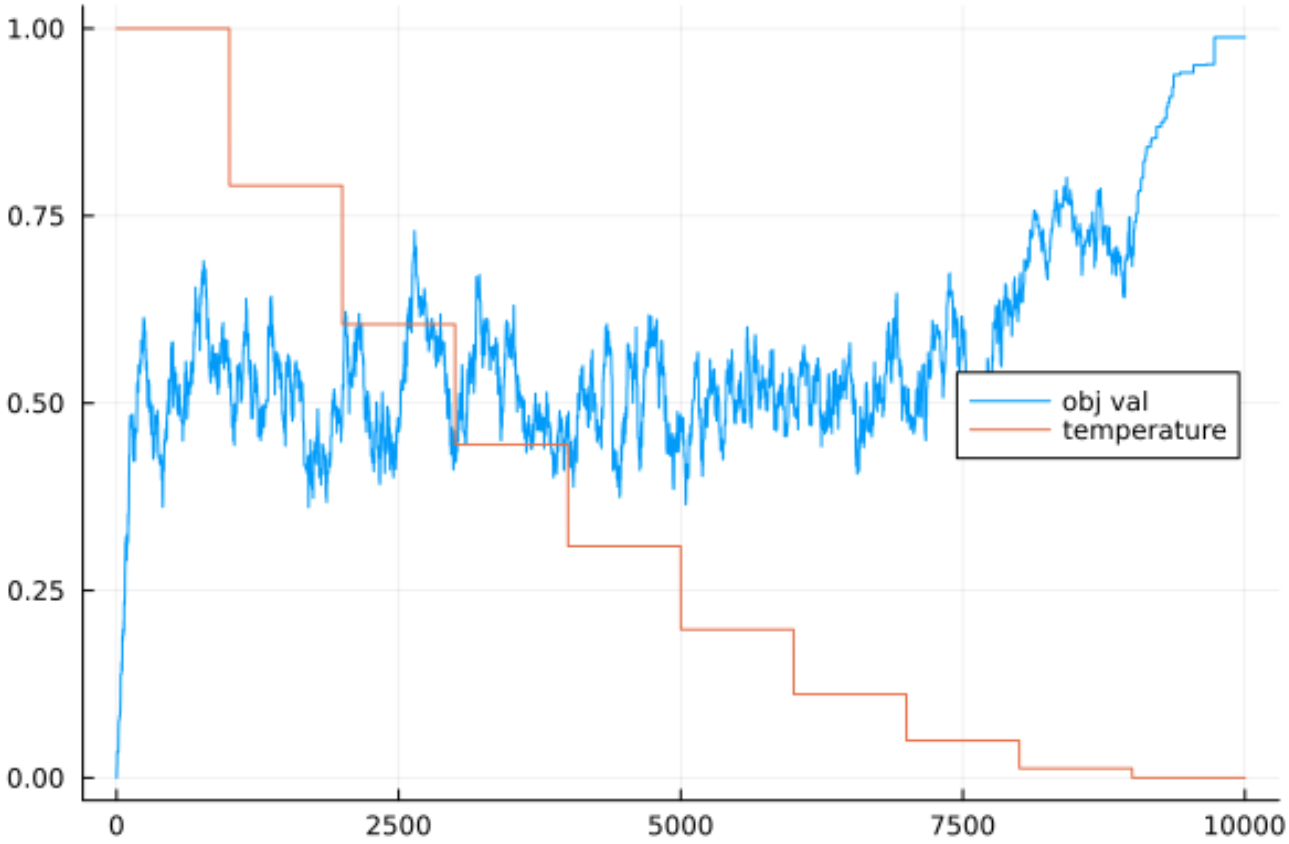

Figure: Quadratic temperature schedule.

Numerical experiments results

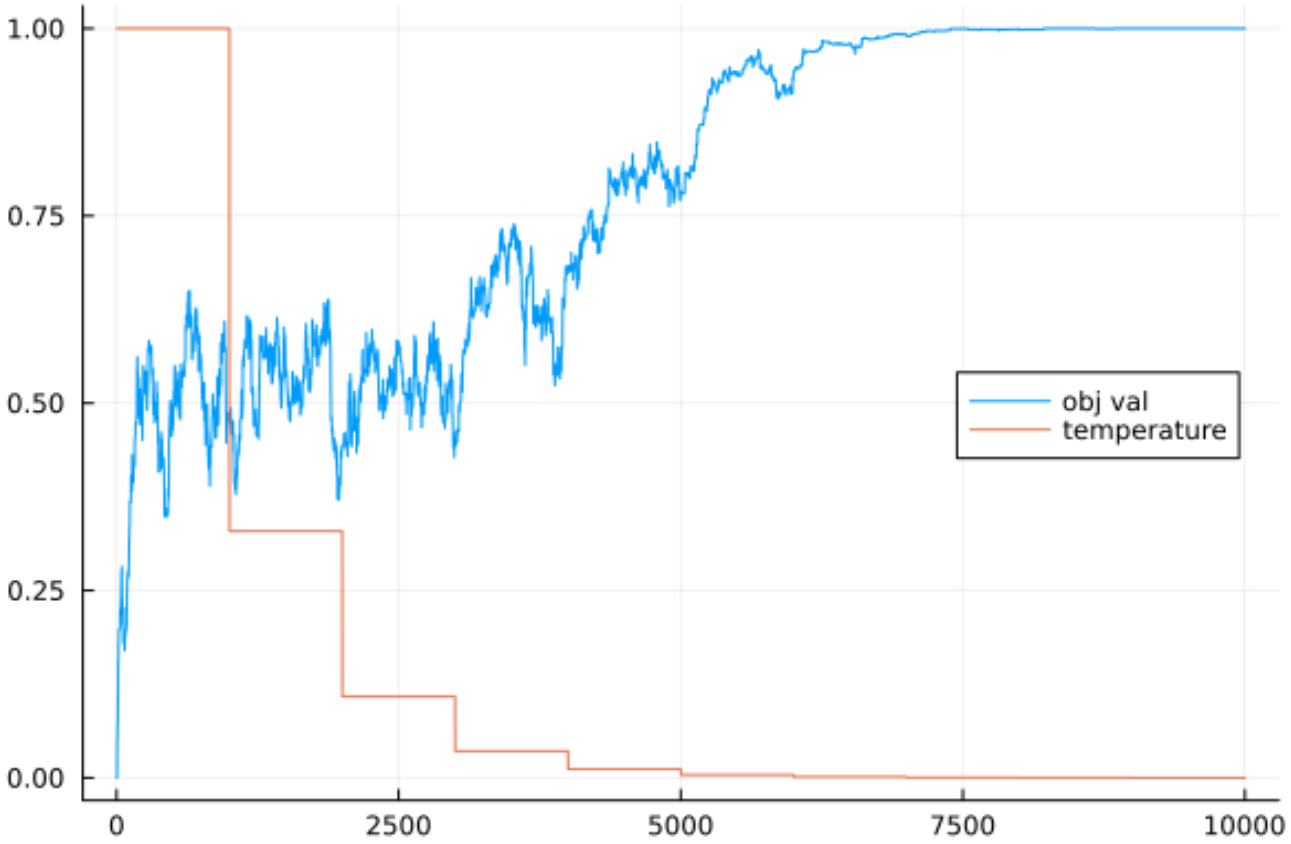

Figure: Exponential temperature schedule.

Numerical experiments results

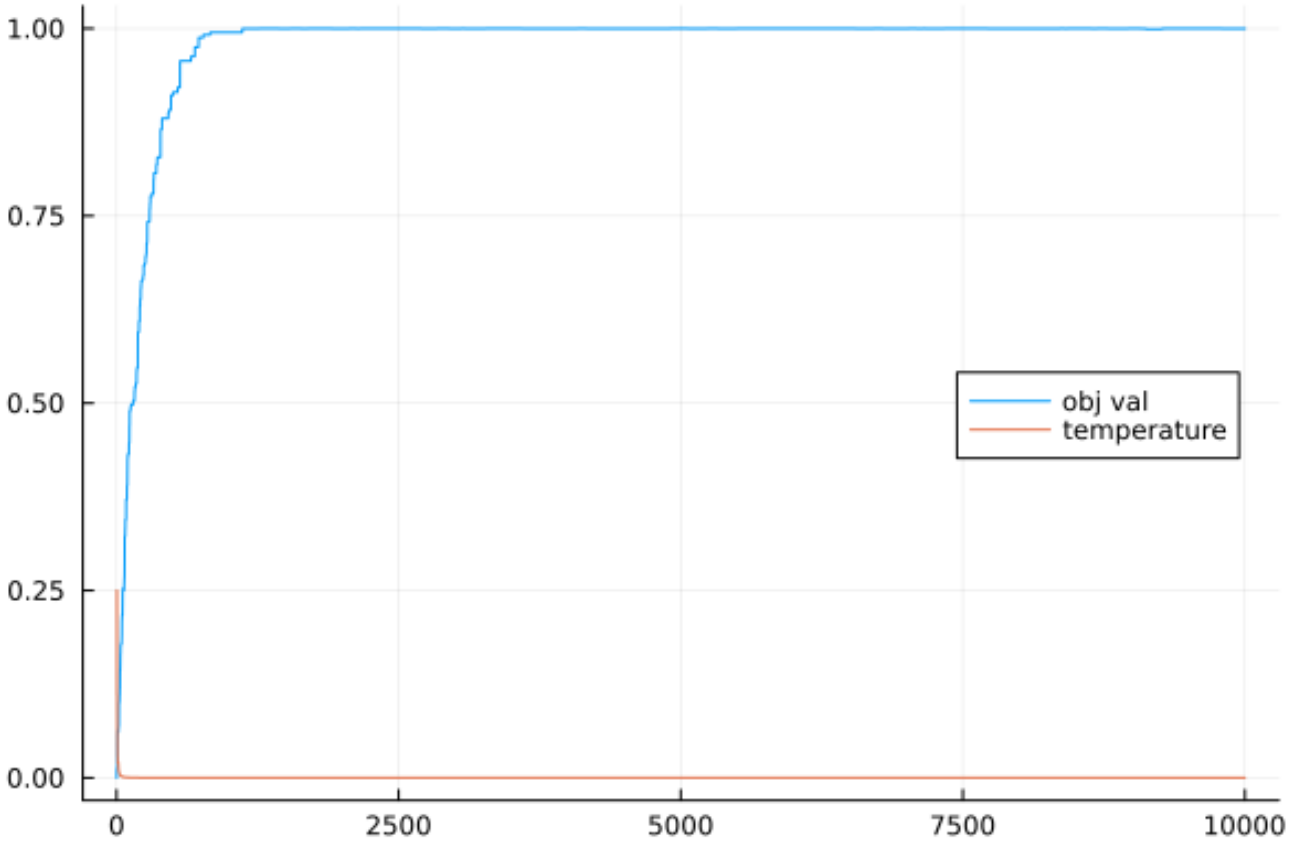

Figure: Automatic temperature schedule.

Conclusion

- Simulated Annealing makes a few assumptions, but at the same time, it is finicky, and the efficiency depends on many factors. These factors includes:
  - underlying structure of the problem.
  - properties of the feasible region (Disconnected? Measurable?).
  - temperature schedule.
  - base chain (is it going to be cursed by dimensionality? does it achieve f-irreducibility?).
- It is relatively easy to implement and is trivial to parallelize on a modern computer.
- In general, it can find the local optimal, but it might not be able to find the global optimal.

References

[1] Christian P. Robert and George Casella, *Monte carlo statistical methods*, Springer, 2005.