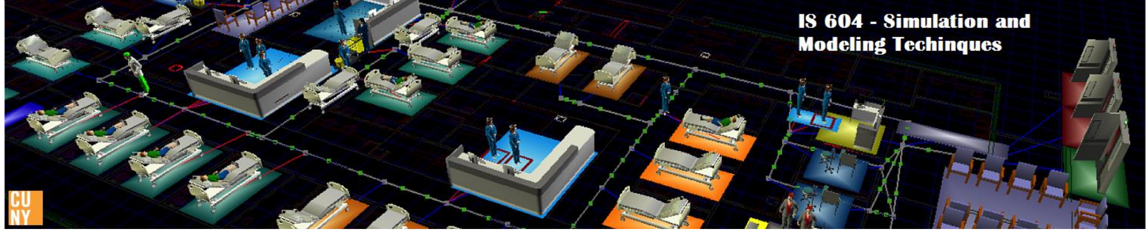
CUNY School of Professional Studies



Week #9:

Markov Chains / Markov Chain Monte Carlo

Graduate Program in Data Analytics (MSDA)
CUNY School of Professional Studies
The City University of New York

IS 604 – Simulation and Modeling Techniques

Assignment

- **Reading:** Ch. 9 (SCR), Supplemental Handouts
- **Activity:** Week #9 Quiz, Discussion #9

Learning Outcomes

- Understand the basic concepts of Markov chains.

- Understand the basic algorithms of Markov Chain Monte Carlo, such as Metropolis-Hastings and Gibbs Sampler.

Stochastic Processes

- Probability Space $(\Omega, \mathcal{H}, \mathbb{P})$

  Ω: Sample space, collection of all possible outcomes

  $\mathcal{H}$: Collection of all possible events (subsets of Ω)

  also called the σ-algebra

  $\mathbb{P}$: Probability measure, maps from each event to $[0,1]$
- The outcome from a random experiment with a probability measure is called a **random variable**
- A vector of random variables $\mathrm{X} = (X_1, X_2, \ldots, X_n)$ is a **random vector**
- A collection of random variables $\{X_t, t \in \mathfrak{I}\}$ is a **stochastic process**

Stochastic Processes II

- Stochastic Process:
  - A Sequence of Random Variables
  - State space: possible outcomes of each r.v. in the sequence
  - Index set: defines the elements in the sequence
    - Often (not always) index refers to time
- State Space of Stochastic Process
  - Discrete
  - Continuous
- Index Set of Stochastic Process
  - Countable/discrete
  - Continuous

Examples of Stochastic Processes

- Markov Chains
- Markov Jump Processes
- Gaussian Processes
- Poisson Processes
- Weiner Process (Brownian Motion)
- Stochastic Differential Equations (SDEs) and Diffusion Processes
- Time Series

Markov Property

- A "Markov" process:
  - A process with no memory
  - All information needed is in current state
  - Given a stochastic process with discrete state space

$$X = \{X_n; n = 0,1,\ldots\}$$

  - Then *X* is a Markov Chain if

$$\Pr\{X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n\}$$
$$= \Pr\{X_{n+1} = j \mid X_n = i_n\} \qquad \forall i_k$$

More on Markov

- Discrete Time Steps ⇒ **Markov *Chain***
- Continuous Time ⇒ **Markov *Process***
- Continuous Time, Discrete State Space

  Markov *Jump Process*
- Transition Probabilities

$$P(i, j) = \Pr\{X_{n+1} = j \mid X_n = i\}$$

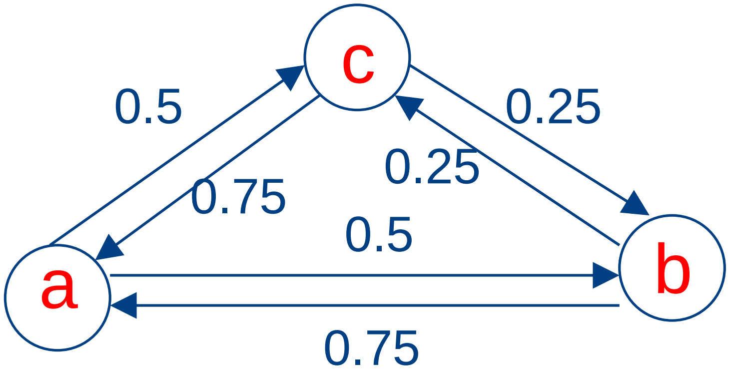
- Stationary Transition Probabilities

$$\Pr\{X_1 = j \mid X_0 = i\} = \Pr\{X_{n+1} = j \mid X_n = i\}$$

  - Discrete State Space: Transition Matrix
  - Continuous State Space: Transition Kernel

State Diagrams

- Diagram of States and Transition Probabilities
- Example: Traveling Salesman
  - Lives in town *a*, covers {*a*,*b*,*c*}
  - From home, flips coin to decide *b* or *c*
  - From *b* or *c*, flips 2 coins: both heads other town else home



Transition Matrix

- Expresses Conditional Probability of Moving Between States
- Ex: If state space $X = \{1,2,3\}$ then

Next State

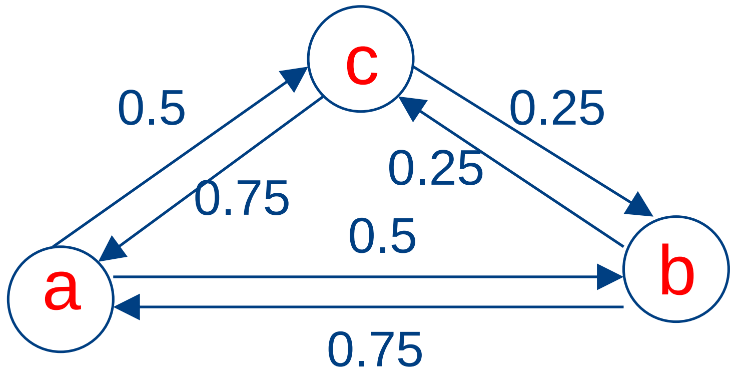
$$
\begin{array}{c|ccc}
 & 1 & 2 & 3 \\
\hline
1 & 0.25 & 0.25 & 0.5 \\
2 & 0.5 & 0.25 & 0.25 \\
3 & 0.25 & 0.25 & 0.5
\end{array}
\qquad
\begin{array}{c}
 \\
1 \\
2 \\
3
\end{array}
\begin{pmatrix}
0.25 & 0.25 & 0.5 \\
0.5 & 0.25 & 0.25 \\
0.25 & 0.25 & 0.5
\end{pmatrix}
$$

Current State

Rows must sum to 1

Columns Need not Sum to 1

Traveling Salesman Example



Write out the
Transition Matrix

$$P = \begin{array}{c c} & \begin{array}{ccc} a & b & c \end{array} \\ \begin{array}{c} a \\ b \\ c \end{array} & \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.75 & 0 & 0.25 \\ 0.75 & 0.25 & 0 \end{pmatrix} \end{array}$$

Questions to Ask of a Markov Chain

- What state will the chain be in in *n* steps?
- What is the probability that the chain is in state *i* in *n* steps?
- What percent of the time is chain in state *i*?
- Can the chain every get from state *i* to state *j*?
- If there is a reward for each state, what is the long-run average reward?

Multi-Step Transitions

- What is the probability of moving from state *b* to state *a* in two steps?

$$\Pr\{X_2 = a \mid X_0 = b\} = P(b,a)P(a,a) + P(b,b)P(b,a) + P(b,c)P(c,a)$$

- This is just matrix multiplication:

$$\Pr\{X_2 = a \mid X_0 = b\} = P^2(b,a)$$

- Example: 2 Steps in Traveling Salesman

$$P^2 = \begin{array}{c} a \\ b \\ c \end{array} \begin{pmatrix} 0.75 & 0.125 & 0.125 \\ 0.1875 & 0.4375 & 0.375 \\ 0.1875 & 0.325 & 0.4375 \end{pmatrix}$$

Markov Chains with Rewards

- If X is a Markov Chain with transition matrix P and profit or reward function f, then the expected profit at step n is:

$$E\left[f(X_n) \mid X_0 = i\right] = P^n f(i)$$

- If Markov Chain X has initial probability vector μ, then the probability of being in state j after n steps is:

$$\Pr_{\mu}\left\{X_n = j\right\} = \mu P^n(j)$$

- Expected profit after n steps is therefore:

$$E\left[f(X_n)\right] = \mu P^n f$$

Traveling Salesman Example

- Suppose Profit by Town is:

$$f = \begin{bmatrix} 1000 \\ 1200 \\ 1250 \end{bmatrix}$$

- Suppose Salesman Starts in Town A: $\mu = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$

- What is the expected profit after 3 weeks?

$$E\left[f(X_n)\right] = \mu P^3 f \cong \$1183$$

Classifying States

- Types of states in a MC:
  - Absorbing: "Once you enter, you'll never leave"
    (Hotel California state? Roach Motel state?)
  - Transient: "Once you leave, you'll never come back"
    (Bad restaurant state?)
  - Recurrent: "You will be back again eventually"
    (MacArthur state?)
- Example: Which states are transient and recurrent?

Classifying Sets of States

- Two states ***communicate*** if you get from *i* to *j* and from *j* to *i* with non-zero probability.
- A ***closed*** set is a set of states where every state communicates with every other state in the set.

$$\sum_{j \in C} P(i,j) = 1 \quad \text{for all } j \in C$$

- An ***irreducible*** set of states is a closed set that does not contain any proper subsets that are closed.
- If *C* is an irreducible set of states and the number states is finite, then every state in *C* is recurrent.

Exercise

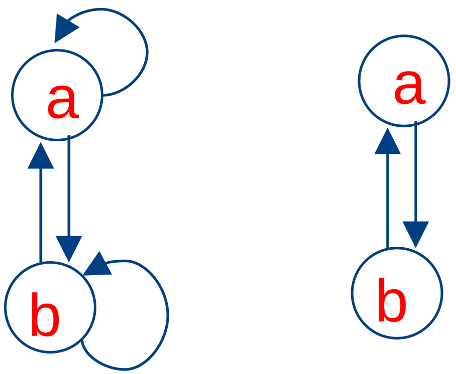
- What are **all** the subsets of states for this MC?
- Which of these are **closed**?
- Which of those closed sets are **irreducible**?

Irreducible Recurrent Markov Chains

- Why do we care?
  - Because IF the entire state space of the chain is an ***irreducible recurrent*** chain…
  - THEN the steady-state probabilities are ***independent*** of initial conditions
- One more property: periodicity

$$k = \gcd\left\{n : \Pr\left(X_n = i \mid X_0 = i\right) > 0\right\}$$

  - If k=1, state is aperiodic
  - If k>1, state is periodic with period k



aperiodic   periodic

Ergodicity

- A finite state Markov Chain which is
  - Irreducible,
  - Positive Recurrent, and
  - Aperiodic

  … is ***ergodic***.
- Meaning: as number of steps approaches infinity, the probability distribution π over states reaches steady-state (does not change).

$$\pi(j) = \lim_{n\to\infty} \Pr\{X_n = j \mid X_0 = i\}$$

$$\pi P = \pi$$

Why We Care

If we can construct a Markov Chain that is ergodic (reaches steady-state), we can:

- Simulate from complex joint probabilty distributions (Markov Chain Monte Carlo)

AND, if you construct a chain that is NOT ergodic, and use it anyway -> YOU GET GARBAGE!

Finding the Stationary Distribution

$$\pi = \pi P$$

- The stationary distribution π is the left eigenvector of the transition matrix P associated with the eigenvalue 1.
- As $k \rightarrow \infty$, P^k converges to a rank 1 matrix where every row is π

$$Q = \lim_{k \to \infty} P^k \qquad QP = Q \qquad Q(P - I_n) = 0_{n,n}$$

- Define $f(.)$ as replacing last column with "1"s

$$Q = f(0_{n,n})[f(P - I_n)]^{-1}$$

Exercise

- What is Traveling Salesman's expected long-run average profit?

$$E\left[f(X_\infty)\right] = \mu P^\infty f = \pi f$$

$$f = \begin{bmatrix} 1000 \\ 1200 \\ 1250 \end{bmatrix}$$

$$\pi = \begin{bmatrix} 0.4286 & 0.2857 & 0.2857 \end{bmatrix}$$

$$\pi f \cong \$1128.57$$

Markov Chain Monte Carlo

- Monte Carlo Integration: $$E_f\left[h(x)\right] \cong \frac{1}{n}\sum_{i=1}^{n} h(x)$$

  … If you can sample from $f(x)$.
- What if it is not feasible to sample from $f(x)$?
- Alternative Strategy:
  - Construct a Markov Chain by choosing a transition kernel $q(x_{t+1}|x_t)$
  - If the chain is ergodic, then after a large number of samples (burn-in) from the chain will approximate samples from the stationary distribution $\pi(x)=f(x)$.

Sampling via MCMC

- Traditional Random Sampling
  - Samples are independent (+)
  - Can use all samples (+) [assuming good RNG]
  - Cannot sample all distributions (-)
- Markov Chain Monte Carlo Sampling
  - Samples are ***dependent***! (-?)
  - Requires burn-in period (to converge) (-)
  - Difficult to verify convergence (---)
  - Can sample any complex distribution (+)
  - Other applications beyond just sampling (+)

History of MCMC

- Almost as old as Monte Carlo itself
- Metropolis et al (1953)
  - Original article
- Hastings (1970)
  - Generalized Metropolis algorithm and demonstrated its uses
- Geman and Geman (1984)
  - First Gibbs sampling

MCMC Samplers

- Metropolis-Hastings
  - Most general: all MCMC are special cases of M-H
  - Basic Samplers: Random Walk or Independence
- Gibbs Sampling
  - Samples from conditional distributions
  - Especially useful in Bayesian applications
- Other Variations
  - Hit-and-Run sampler
  - Slice Sampler
  - Reversible Jump Sampler

Typical Applications of MCMC

- Sampling from Complex (Nasty) Joint Distributions
- Bayesian Statistics
- Data Assimilation
- Monte Carlo Optimization
  - Simulated Annealing

Metropolis-Hastings Algorithm

- Given the current state (sample) X_t
- Sample a candidate point Y from a proposal distribution $q(Y|X_t)$
- Accept Y as the next sample with probability:

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X\,|\,Y)}{\pi(X)q(Y\,|\,X)}\right)$$

- If accepted, $X_{t+1} = Y$, else $X_{t+1} = X_t$.

Why Metropolis-Hastings Works

- For any proposal distribution $q(.|.)$, the stationary distribution will be $\pi(.)$.
- The transition kernel for M-H is:

$$P(X_{t+1} \mid X_t) = q(X_{t+1} \mid X_t)\alpha(X_t, X_{t+1}) + I(X_{t+1} = X_t)\left[1 - \int q(Y \mid X_t)\alpha(X_t, Y)dY\right]$$

- Using Symmetry:

$$\pi(X_t)q(X_{t+1} \mid X_t)\alpha(X_t, X_{t+1}) = \pi(X_{t+1})q(X_t \mid X_{t+1})\alpha(X_{t+1}, X_t)$$

- Detailed Balance Equation:

$$\pi(X_t)P(X_{t+1} \mid X_t) = \pi(X_{t+1})P(X_t \mid X_{t+1})$$

Why Metropolis-Hastings Works II

$$\pi(X_t)P(X_{t+1} \mid X_t) = \pi(X_{t+1})P(X_t \mid X_{t+1})$$

- Integrate both sides w.r.t. X_t:

$$\int \pi(X_t)P(X_{t+1} \mid X_t)dX_t = \pi(X_{t+1})$$

- If X_t is from $\pi(.)$, then X_{t+1} will be from $\pi(.)$

M-H Samplers

- What defines a particular form of M-H?
  - The choice of $q(.|.)$! (and α)

- Some common M-H samplers:
  - Independence Sampler: $q(Y|X) = q(Y)$

$$\alpha(X,Y) = \min\left(1, \frac{w(Y)}{w(X)}\right) \qquad w(X) = \frac{\pi(X)}{q(X)}$$

  - Random Walk Sampler: $q(Y|X) = q(|X - Y|)$

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)}{\pi(X)}\right)$$

Gibbs Sampler

- First used for image processing (Geman and Geman), based on Gibbs distribution from Statistical Mechanics:

$$f(x_1,\ldots x_n) \propto \exp\left[-\frac{1}{kT}E(x_1,\ldots,x_n)\right]$$

- General approach
  - Start with initial guess for all parameters
  - Sample from conditional distribution of one parameter, given all other parameters
  - Update each parameter (component in some order)
  - Repeat many times
- Special case of M-H where $\alpha = 1$ (always accept)

Conditional Distributions

- Given a joint distribution $f(x,y,z)$ for random variables x, y, and z

- You can write the conditional of any one in terms of the others
  - E.g. $f(x|y,z)$

- Take the equation for the joint pdf, and drop all terms except those that contain x.

Example: Bivariate Normal

- Suppose $f(x_1,x_2) \sim \text{Normal}(\mu_1, \sigma_1, \mu_2, \sigma_2, \rho)$

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{(1-\rho^2)}\sigma_1\sigma_2} \times$$

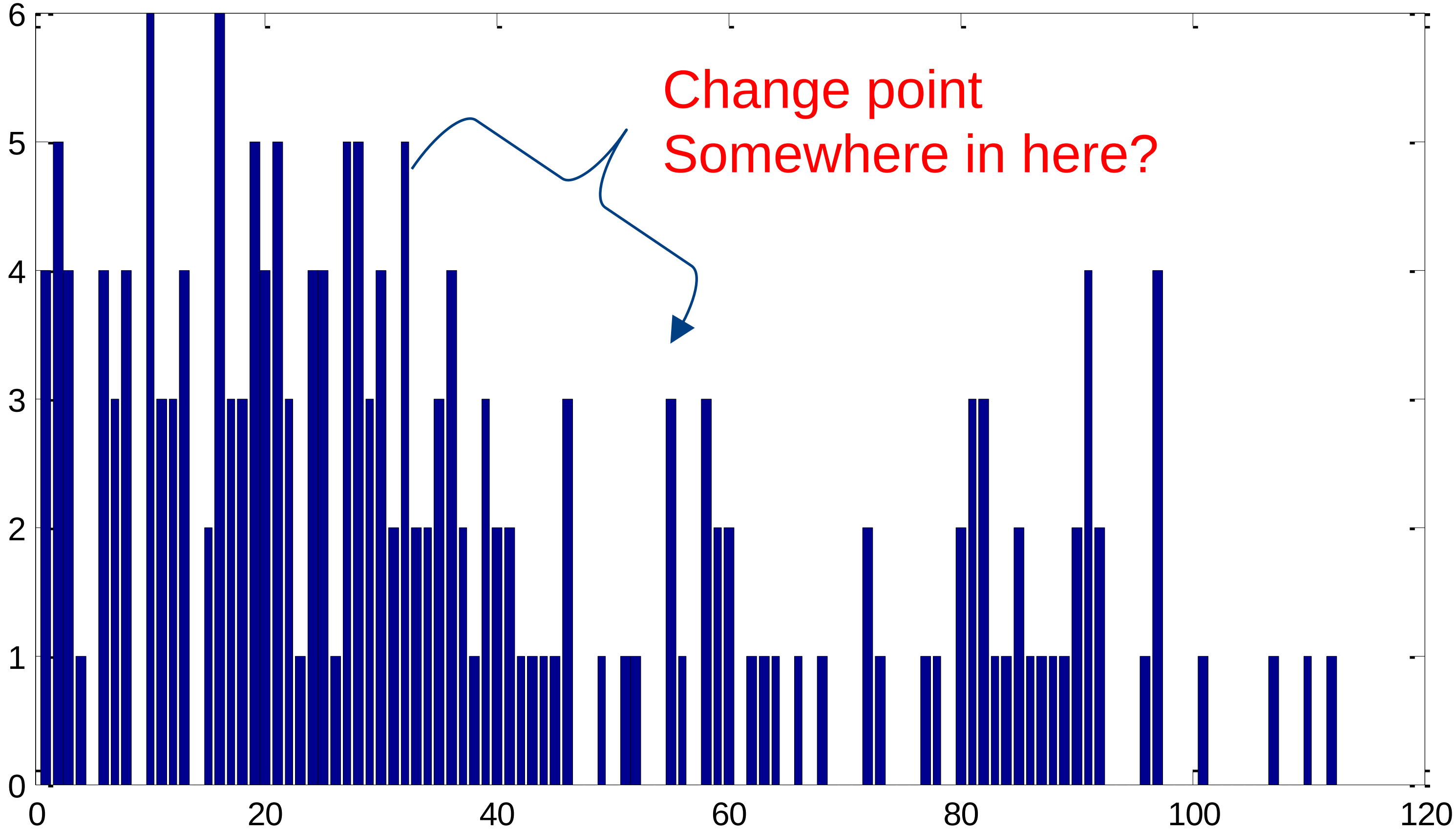
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]\right\}$$

- Then the pdf of x_1 conditional on x_2 is:

$$f(x_1 \mid x_2) \propto \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)\right]\right\}$$

Example: Poisson with Change Point

Counts of coal mining disasters in UK 1851-1962



Example: Poisson with Change Point

- Let observations y_i be samples from a Poisson distribution, $i=1,\ldots,n$
- Let m: $1 \leq m \leq n$ be the change point
- Then $y_i \sim \text{Poisson}(\lambda)$ $i=1,\ldots,m$

  and $y_i \sim \text{Poisson}(\phi)$ $i=m+1,\ldots,n$
- Define priors of unknown parameters

  $\lambda \sim \text{Gamma}(\alpha, \beta)$

  $\phi \sim \text{Gamma}(\gamma, \delta)$

Example: Poisson with Change Point

- The posterior joint distribution for λ, ϕ, and m, given the observations is:

$$\pi(\lambda,\phi,m \mid y_1,\ldots,y_n) \propto f(y_1,\ldots,y_n \mid \lambda,\phi,m)\,p(\lambda,\phi,m)$$

$$= \left[\prod_{i=1}^{m} f(y_i \mid \lambda) \prod_{i=m+1}^{n} f(y_i \mid \phi)\right] f(\lambda \mid \alpha,\beta)\, f(\phi \mid \gamma,\delta)\, \frac{1}{n}$$

$$\propto \left[\prod_{i=1}^{m} e^{-\lambda}\lambda^{y_i} \prod_{i=m+1}^{n} e^{-\phi}\phi^{y_i}\right] \left(\lambda^{\alpha-1}e^{-\beta\lambda}\right)\left(\phi^{\gamma-1}e^{-\delta\phi}\right)$$

$$\propto \lambda^{\alpha+s_m-1}e^{-(\beta+m)\lambda}\phi^{\gamma+s_n-s_m-1}e^{-(\delta+n-m)\phi}$$

where

$$s_m = \sum_{i=1}^{m} y_i$$

$$s_n = \sum_{i=1}^{n} y_i$$

Application: Optimization

- Traditional Methods
  - Linear Programming
  - Non-Linear Programming
  - Mixed Complementarity
  - All will find the LOCAL minimum (Why?)
- Global Optimization Methods
  - Deterministic methods (e.g., Branch and Bound)
  - Stochastic methods (e.g., Simulated Annealing)
  - Heuristic methods (e.g., genetic methods)
  - Response surface methods

Simulated Annealing

- A stochastic global optimization method
- Finds an *approximation* to the optimum
- “Annealing” – refers to a process of heating and cooling in metallurgy to increase crystal size and reduce defects
- Basic idea:
  - Replace current solution with “nearby” solution
  - When temperature is high, changes almost random
  - As temperature cools, increasingly “downhill”
  - Possibility of “uphill” allows for getting unstuck from local minima

Simulated Annealing Algorithm (M-H)

1. Initialize starting state $X_{0,}$ temperature T_0, set t=0

2. Generate a candidate state Y from the symmetric proposal $q(X_t, Y)$

3. If $S(Y)<S(X_t)$, $X_{t+1}=Y$.
   If $S(Y)>S(X_t)$, generate U ~ U(0,1) and
   let $X_t+1=Y$ if $U \leq \exp(-(S(Y)-S(X_t))/T)$

4. Set $T_{t+1}=\beta T$, $t=t+1$, go to step 2 until done.