Convex Optimization CMU-10725

Simulated Annealing

Barnabás Póczos & Ryan Tibshirani

MACHINE LEARNING DEPARTMENT

Carnegie Mellon.
School of Computer Science

Andrey Markov

Markov Chains

Markov Chains

Markov chain:

$$P(X_{t+1}|X_t, \ldots, X_1) = P(X_{t+1}|X_t)$$

Homogen Markov chain:

$P(X_{t+1}|X_t)$ is invariant for all t.

Markov Chains

- Assume that the state space is finite:

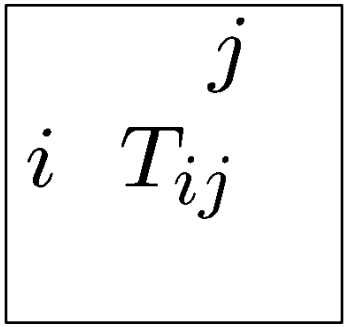
$$\mathcal{X} = \{1, \ldots, k\}.$$

- 1-Step state transition matrix:

$$T_{ij} = P(X_{t+1} = j | X_t = i)$$

Lemma: The state transition matrix is stochastic:

$$\sum_j T_{ij} = 1 \ \forall i$$



- t-Step state transition matrix:

$$Q_{ij} \doteq P(X_{k+t} = j | X_k = i)$$

Lemma:

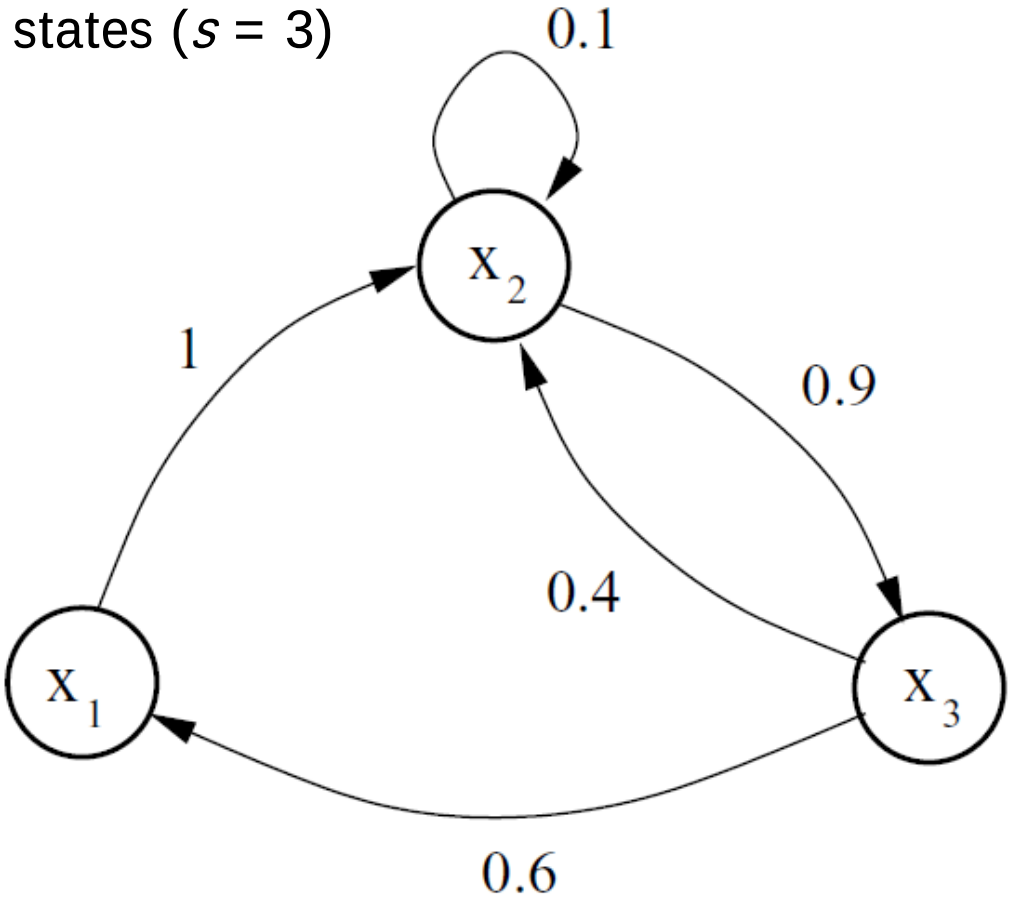
$$P(X_{k+t} = j | X_k = i) = Q_{ij} = [T^t]_{ij}, \ \forall (k, i, j)$$

Markov Chains Example

Markov chain with three states (s = 3)

$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

Transition matrix



Transition graph

Markov Chains

$$T = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0.1 & 0.9 \\ 0.6 & 0.4 & 0 \end{bmatrix}$$

If the probability vector for the initial state is $\mu(x^{(1)}) = (0.5, 0.2, 0.3)$

it follows that $\mu(x^{(1)})T = (0.2, 0.6, 0.2)$

and, after several iterations (multiplications by T)

$\mu(x^{(1)})T^t \rightarrow p(x) = (0.22, 0.41, 0.37)$ **stationary distribution**

no matter what initial distribution $^1(x^1)$ was.

$$T^\infty = \begin{bmatrix} 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \\ 0.22 & 0.41 & 0.37 \end{bmatrix}$$

The chain has forgotten its past.

Markov Chains

Our goal is

to **find conditions** under which the Markov chain forgets its past, and independently from its starting distribution, the state distributions converge to a **stationary distribution**.

Definition: [stationary distribution, invariant distribution]

The distribution $\pi = (\pi_1, \ldots, \pi_k)$ is **stationary** distribution if $\pi_i \geq 0 \ \forall i$, $\sum_{i=1}^{T} \pi_i = 1$, and $\pi \mathbf{T} = \pi$.

This is a necessary condition for having limit behavior of the Markov chain.

Markov Chains

Theorem:

For any starting point, the chain will convergence to the unique **invariant distribution** $p(x)$, as long as

1. **T** is a stochastic transition matrix
2. **T** is irreducible
3. **T** is aperiodic

Limit Theorem of Markov Chains

More formally:

If the Markov chain is **Irreducible** and **Aperiodic**, then:

1. $\exists! \pi = (\pi_1, \ldots \pi_n)$ stationary distribution
2. $\lim_{t \to \infty} \frac{E(\textit{ number of chain visits state } i \textit{ in } t \textit{ steps})}{t} = \pi_i$
3. $\lim_{t \to \infty} \Pr(X_t = i) = \pi_i \ \forall i$
4. $\lim_{t \to \infty} \mathbf{v}\mathbf{T}^t = \pi \ \forall \mathbf{v}$, that is, the Markov chain forgets its past.

Markov Chains

Definition

Irreducibility:

For each pairs of states *(i,j)*, there is a positive probability, starting in state *i*, that the process will ever enter state *j*.

= The matrix *T* cannot be reduced to separate smaller matrices

= Transition graph is connected.

It is possible to get to any state from any state.

Markov Chains

Definition

Aperiodicity: The chain cannot get trapped in cycles.

Definition

A state *i* has period *k* if any return to state *i,* must occur in multiples of *k* time steps. Formally, the period of a state *i* is defined as

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

(where "gcd" is the greatest common divisor)

For example, suppose it is possible to return to the state in {6,8,10,12,...} time steps. Then k=2

Markov Chains

Definition

If k = 1, then the state is said to be aperiodic:

returns to state i can occur at irregular times.

In other words,

a state i is aperiodic if there exists n such that for all n' ≥ n,

$$\Pr(X_{n'} = i | X_0 = i) > 0$$

Definition

A Markov chain is aperiodic if every state is aperiodic.

Markov Chains

Theorem:

An irreducible Markov chain only needs one aperiodic state to imply all states are aperiodic.

Corollary:

An irreducible chain is said to be aperiodic, if for some n ¸ 0 and some state j $\Pr(X_n = j | X_0 = j) > 0$ and $\Pr(X_{n+1} = j | X_0 = j) > 0$

Example for periodic Markov chain:

Let $\mathrm{T} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$

In this case $(1/2, 1/2)T = (1/2, 1/2)$

However, if we start the chain from (1,0), or (0,1), then the chain get traps into a cycle, it doesn't forget its past.

Periodic Markov chains don't forget their past.

Reversible Markov chains
Detailed Balance Property

Definition: reversibility /detailed balance condition:

$$\pi_i P_{ij} = \pi_j P_{ji}, \ \forall (i,j)$$

Theorem:

A sufficient, but not necessary, condition to ensure that a particular ¼ is the desired invariant distribution of the Markov chain is the detailed balance condition.

How fast can Markov chains forget the past?

MCMC samplers are

- irreducible and aperiodic Markov chains
- have the target distribution as the invariant distribution.
- the detailed balance condition is satisfied.

It is also important to design samplers that converge quickly.

Spectral properties

Theorem: If

$$\pi \mathbf{T} = \pi, \text{ then}$$

- ¼ is the left eigenvector of the matrix *T* with eigenvalue 1.
- The Perron-Frobenius theorem from linear algebra tells us that the remaining eigenvalues have absolute value less than 1.
- The second largest eigenvalue, therefore, determines the rate of convergence of the chain, and should be as small as possible.

The Hastings-Metropolis Algorithm

The Hastings-Metropolis Algorithm

Let $b_1, \ldots, b_m > 0$, and $B = \sum_{j=1}^{m} b_j$

Assume that m is so big, that it is difficult to calculate B.

Our goal:

Generate samples from the following **discrete** distribution:

$$P(X = j) = \pi_j = \frac{b_j}{B}$$

We don't know *B* !

The main idea is to construct a time-reversible Markov chain with ($\pi_1, \ldots, \pi_m$) limit distributions

Later we will discuss what to do when the distribution is continuous

The Hastings-Metropolis Algorithm

Let {1,2,...,m} be the state space of a Markov chain that we can simulate.

Let $q(i,j) = p(j|i)$

Let $\{X_0, X_1, \ldots X_n, \ldots\}$ Markov chain be defined as follows:

$$\Pr(X_n = j \,|X_{n-1} = i) =$$

1., from state i go to state j with prob. $q(i,j)$

2., $\begin{cases} \text{with prob } 1-\alpha(i,j) \text{ go back to state } i, \\ \text{with prob } \alpha(i,j) \text{ stay in state } j. \end{cases}$

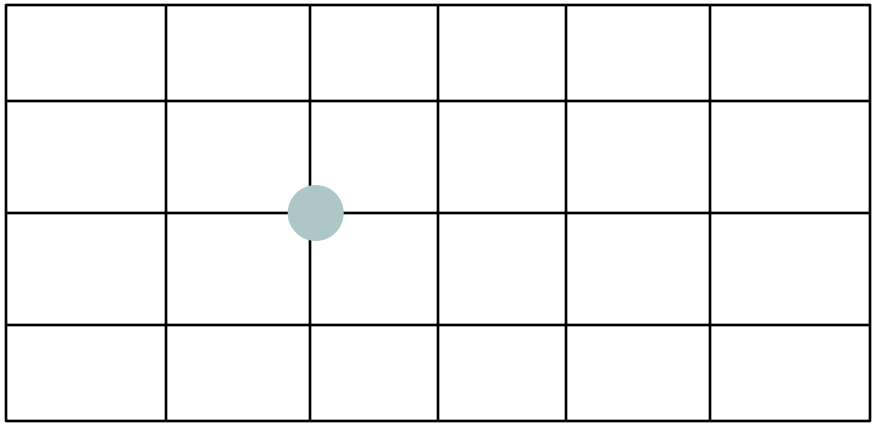
No rejection: we use all X_1, X_2,... X_n, ...

Example for Large State Space

Let {1,2,...,m} be the state space of a Markov chain that we can simulate. Let $q(i,j) = p(j|i)$

d-dimensional grid:

- Max 2d possible movements at each grid point (linear in d)
- Exponentially large state space in dimension d



The Hastings-Metropolis Algorithm

$$\Pr(X_n = j \,|X_{n-1} = i) =$$

1., from state i go to state j with prob. $q(i,j)$

2., $\begin{cases} \text{with prob } 1-\alpha(i,j) \text{ go back to state } i, \\ \text{with prob } \alpha(i,j) \text{ stay in state } j. \end{cases}$

Theorem

$$\begin{aligned} P(X_{n+1} = j | X_n = i) &= q(i,j)\alpha(i,j) \quad \forall j \neq i \\ P(X_{n+1} = i | X_n = i) &= q(i,i) + \sum_{k \neq i} q(i,k)(1-\alpha(i,k)) \end{aligned}$$

Proof

We can go to state i from state i and also from other states $k \neq i$.

The Hastings-Metropolis Algorithm

Observation

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall j \neq i \Leftrightarrow \pi_i q(i,j)\alpha(i,j) = \pi_j q(j,i)\alpha(j,i) \quad \forall j \neq i \quad (*)$$

Proof: $P_{ij} = P(X_{n+1} = j | X_n = i) = q(i,j)\alpha(i,j) \ \forall j \neq i$

Corollary

$$\Rightarrow \begin{cases} X_0, X_1, \ldots X_n, \ldots \text{ time reversible Markov chain} \\ \exists \pi_1, \ldots \pi_m \text{ stationary distribution} \end{cases}$$

Theorem

$$\text{If } \alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right) \Rightarrow$$

$$\Rightarrow \ (*) \text{ holds}$$

$$\Rightarrow \ (\pi_1, \ldots \pi_m) \text{ stationary distribution}$$

The Hastings-Metropolis Algorithm

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall j \neq i \Leftrightarrow \pi_i q(i,j)\alpha(i,j) = \pi_j q(j,i)\alpha(j,i) \quad \forall j \neq i \quad (*)$$

Theorem

$$\text{If } \alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{b(j)q(j,i)}{b(i)q(i,j)}, 1\right) \Rightarrow$$

$\Rightarrow$ (*) holds

$\Rightarrow$ $(\pi_1, \ldots \pi_m)$ stationary distribution

Proof:

$$\text{If } \alpha(i,j) = \frac{\pi_j q(j,i)}{\pi_i q(i,j)} \Leftrightarrow = \alpha(j,i) = 1$$

Note: To calculate $\alpha(i,j)$ we didn't need to use $B = \sum_{j=1}^{m} b(j)$.

The Hastings-Metropolis Algorithm

1) Let Q be a Markov chain with $q(i,j) = P(j|i)$ state transition probabilites. Assume that we can sample from $P(i|j)$.

2) Let $1 \leq k \leq m$ arbitrary, $n = 0$, and $X_0 = k$.

3) Sample X^* according to $P(X^* = j) = q(X_n, j)$, $j = 1, \ldots, m$ distribution.
(Go from X_n to state j using Markov chain Q)

4) Let $u \sim U_{[0,1]}$ (With prob $\alpha(i,j)$ stay in $X^* = j$)

5) If $u < \dfrac{b(X)q(X, X_n)}{b(X_n)q(X_n, X)} \Rightarrow X_{n+1} = X^*$
else $\Rightarrow X_{n+1} = X_n$ (Otherwise go back)

6) $n = n + 1$

7) Back to 3

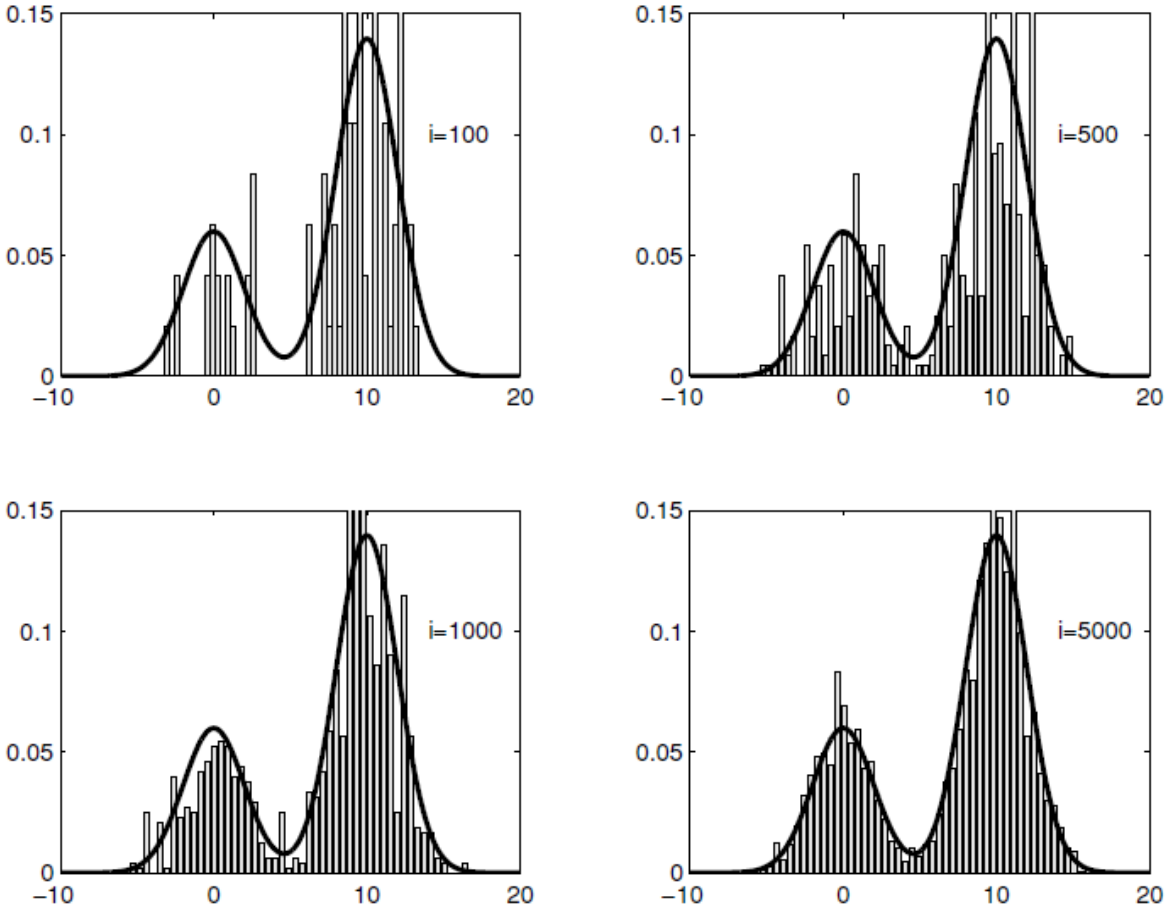
It is not rejection sampling, we use all the samples!

Continuous Distributions

- The same algorithm can be used for continuous distributions as well.
- In this case, the state space is continuous.

Experiment with HM

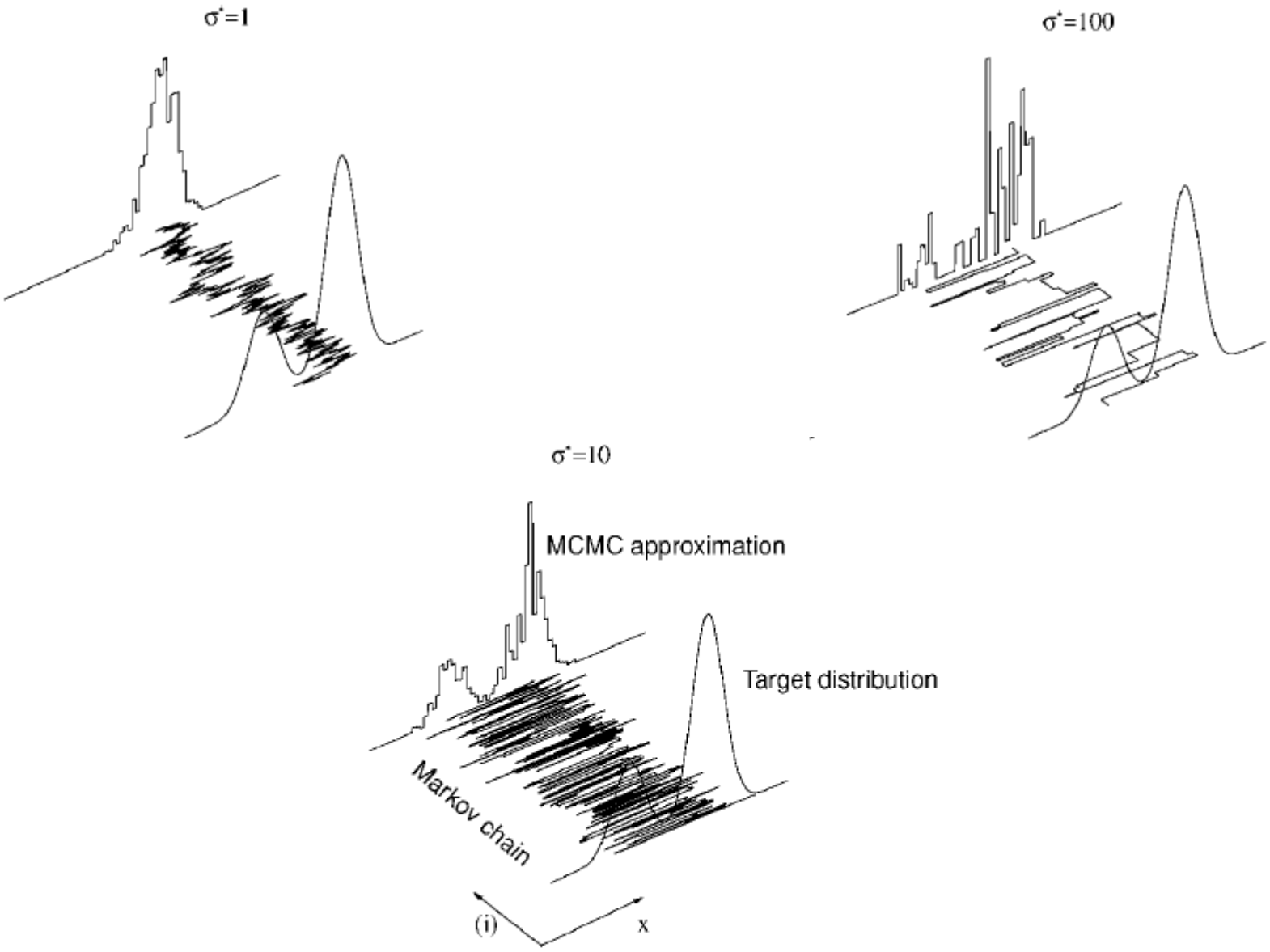
An application for continuous distributions



Bimodal target distribution: $p(x) \propto 0.3 \exp(-0.2x^2) + 0.7 \exp(-0.2(x-10)^2)$

$q(x \mid x^{(i)}) = N(x^{(i)}, 100)$, 5000 iterations

Good proposal distrib. is important



HM on Combinatorial Sets

Let $\mathcal{P} = \{x_1, \ldots x_n \,|\, x_1, \ldots x_n$ is a permuation of $(1, \ldots, n)$ such that, $\sum_{j=1}^{n} j x_j > a\}$, where a is a given constant.

Generate **uniformly distributed** samples from the set of permutations $\mathcal{P}$.

Let n=3, and a=12:

{1,2,3}: 1+4+9=14

{1,3,2}: 1+6+6=13

{2,3,1}: 2+6+3=11

{2,1,3}: 2+2+9=13

{3,1,2}: 3+2+6=11

{3,2,1}: 3+4+3=10

HM on Combinatorial Sets

To define a simple Markov chain on $\mathcal{P}$, we need the concept of **neighboring elements** (permutations):

Definition: Two permutations are **neighbors**, if one results from the interchange of two of the positions of the other:

(1,2,3,4) and (1,2,4,3) are neighbors.

(1,2,3,4) and (1,3,4,2) are not neighbors.

HM on Combinatorial Sets

Let $N(i)$ be the number set of state i.

Let $q(i,j) = P(j|i) = \begin{cases} \frac{1}{|N(i)|} & \text{if } j \in N(i). \\ 0 & \text{Otherwise} \end{cases}$

$$\alpha(i,j) = \min\left(\frac{\pi_j q(j,i)}{\pi_i q(i,j)}, 1\right) = \min\left(\frac{1\frac{1}{N(j)}}{1\frac{1}{N(i)}}, 1\right) = \min\left(\frac{N(i)}{N(j)}, 1\right)$$

$\Rightarrow$ the limit distribution of the Markov chain is uniform over $\mathcal{P}$ with probabilities $\frac{1}{|\mathcal{P}|}$

That is what we wanted!

Gibbs Sampling: The Problem

Let $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$

Let $p(x_1, \ldots, x_n) \geq 0$ be a non-normalized distribution ($\int p(x) \neq 1$, $p(x) \geq 0$), and let A be a complicated set.

Suppose that we can generate samples from

$$P(X_i = x | X_j = x_j, \forall j \neq i)$$

e.g. $P(X_3 = x_3 | X_1 = x_1, X_2 = x_2, X_4 = x_4, X_5 = x_5)$

Our goal is to generate samples from

$$f(x_1, \ldots, x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x} \in A)} & \text{if } \mathbf{x} \in A \end{cases}$$

Gibbs Sampling: Pseudo Code

1. We are in $\mathbf{x} = (x_1, \ldots, x_n) \in A$

2. Draw a random state $i \in \{1, \ldots, n\}$ with prob. $1/n$.

3. Sample x from $x \sim P(X_i = x | X_j = x_j, \forall j \neq i)$.

4. Let $\mathbf{y} = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$

5. If

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \in \mathbf{A} \Rightarrow x_i = x$, accept this new state

   $(x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n) \notin \mathbf{A} \Rightarrow x_i$ stays in the old x_i

6. New sample point: $(x_1, \ldots, x_n)$. Go back to 2

Gibbs Sampling: Theory

Let

$$\begin{aligned} q(\mathbf{x},\mathbf{y}) &= q(\overbrace{(x_1,\ldots,x_n)}^{\mathbf{x}},\overbrace{(x_1,\ldots,x_{i-1},x,x_{i+1},x_n)}^{\mathbf{y}}) \\ &\doteq \frac{1}{n}P(X_i = x | X_j = x_j, \forall j \neq i) \\ &= \frac{1}{n}\frac{P(\mathbf{y})}{P(X_j = x_j, \forall j \neq i)} \end{aligned}$$

and let $\alpha(\mathbf{x},\mathbf{y}) = \min\left(\frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})}, 1\right)$

Observation: By construction, this **HM sampler** would sample from

$$f(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x}\in A)} & \text{if } \mathbf{x} \in A \end{cases}$$

Gibbs Sampling is a Special HM

Theorem: The Gibbs sampling is a special case of HM with

$$\alpha(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A, \mathbf{y} \in A \\ 0 & \text{if } \mathbf{x} \in A, \mathbf{y} \notin A \end{cases}$$

Proof:

By definition: $f(x_1,\ldots,x_n) = \begin{cases} 0 & \text{if } \mathbf{x} \notin A \\ \frac{p(\mathbf{x})}{p(\mathbf{x}\in A)} & \text{if } \mathbf{x} \in A \end{cases}$

If $\mathbf{x} \in A, \mathbf{y} \in A \Rightarrow \dfrac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \dfrac{\frac{p(\mathbf{y})}{p(\mathbf{y}\in A)}q(\mathbf{y},\mathbf{x})}{\frac{p(\mathbf{x})}{p(\mathbf{x}\in A)}q(\mathbf{x},\mathbf{y})} =$

$$= \frac{p(\mathbf{y})q(\mathbf{y},\mathbf{x})}{p(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \frac{p(\mathbf{y})\frac{1}{n}\frac{P(\mathbf{x})}{P(Y_j=y_j, j\neq i)}}{p(\mathbf{x})\frac{1}{n}\frac{P(\mathbf{y})}{P(X_j=x_j, j\neq i)}} = \frac{p(\mathbf{y})p(\mathbf{x})}{p(\mathbf{x})p(\mathbf{y})} = 1$$

since $P(X_j = x_j, j \neq i) = P(Y_j = y_j, j \neq i)$

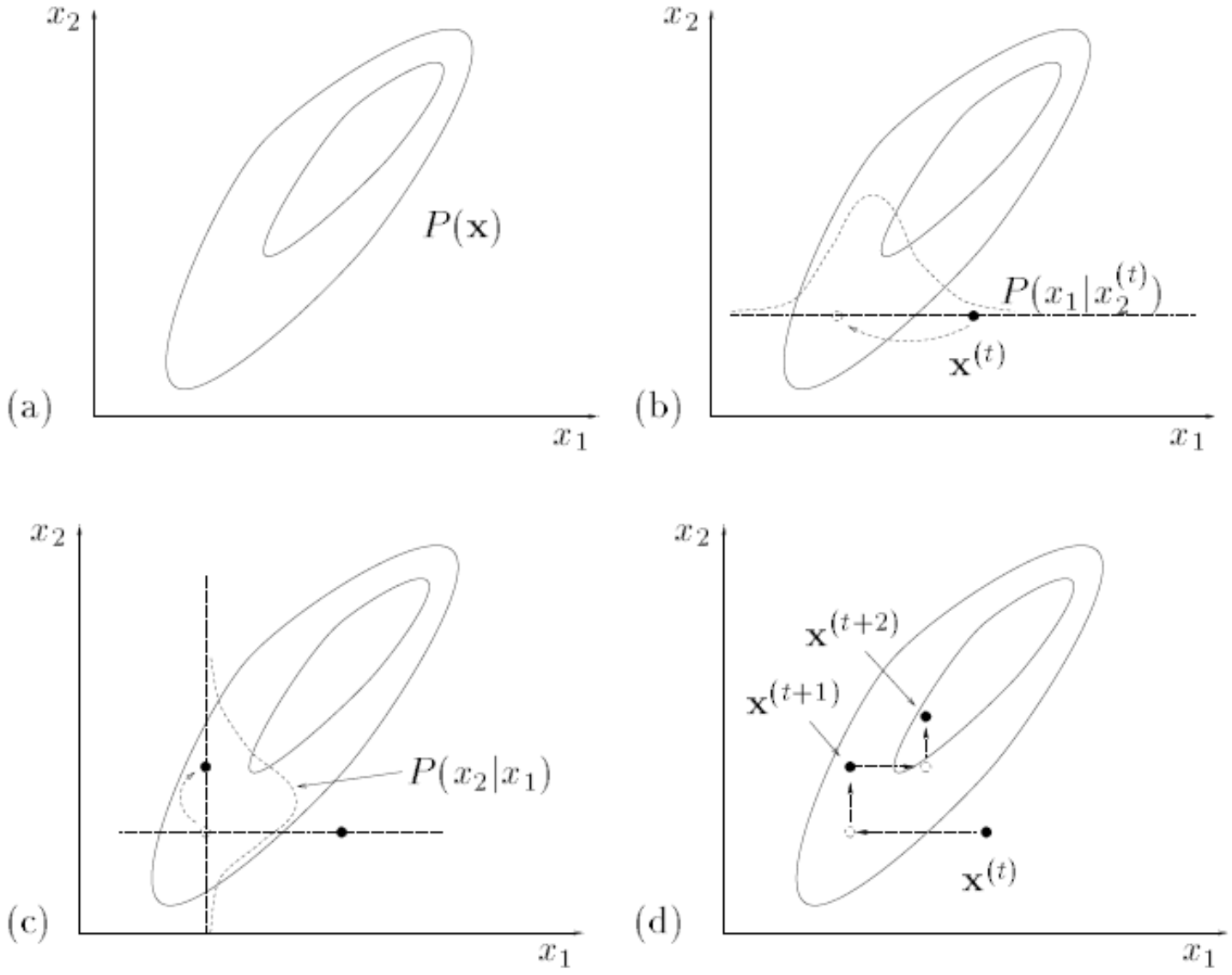
Gibbs Sampling is a Special HM

Proof:

$$\text{If} \quad \mathbf{x} \in A, \mathbf{y} \notin A \Rightarrow \frac{f(\mathbf{y})q(\mathbf{y},\mathbf{x})}{f(\mathbf{x})q(\mathbf{x},\mathbf{y})} = \frac{0q(\mathbf{y},\mathbf{x})}{\frac{p(\mathbf{x})}{p(\mathbf{x}\in A)}q(\mathbf{x},\mathbf{y})} = 0$$

$$\text{since } \alpha(\mathbf{x},\mathbf{y}) = \begin{cases} 1 & \text{if } \mathbf{x} \in A, \mathbf{y} \in A \\ 0 & \text{if } \mathbf{x} \in A, \mathbf{y} \notin A \end{cases}$$

Gibbs Sampling in Practice



Simulated Annealing

Simulated Annealing

Let $\mathcal{A}$ be a huge finite set of vectors.

Let $V : \mathcal{A} \to \mathbb{R}_+$.

V can have lots of local maximum.

Goal: Find $$V^* = \max_{x \in \mathcal{A}} V(x)$$

Let $\mathcal{M} = \{x \in \mathcal{A} : V(\mathbf{x}) = V^*\}$, set of maximum points

Simulated Annealing

Let $\lambda > 0$ $\quad P_\lambda(\mathbf{x}) = \dfrac{\exp(\lambda V(x))}{\sum_{\mathbf{x}\in\mathcal{A}} \exp(\lambda V(x))}$

Theorem: $\quad P_\lambda(x) \xrightarrow{\lambda\to\infty} \dfrac{\delta(x,\mathcal{M})}{|\mathcal{M}|}$

where $\delta(x,\mathcal{M}) = 1$, if $x \in \mathcal{M}$, and 0 otherwise.

Proof: $\quad P_\lambda(x) = \dfrac{\exp(\lambda(V(x) - V^*))}{|\mathcal{M}| + \sum_{\substack{\mathbf{x}\in\mathcal{A} \\ \mathbf{x}\notin\mathcal{M}}} \exp(\lambda(V(x) - V^*))}$

$(V(x) - V^*) \leq 0, \ \forall x$

If $(V(x) - V^*) < 0$, then $\exp(\lambda(V(x) - V^*)) \xrightarrow{\lambda\to\infty} = 0$

Simulated Annealing

Main idea

- Let ¸ be big.
- Generate a Markov chain with limit distribution P¸(x).
- In long run, the Markov chain will jump among the maximum points of P¸(x).

Introduce the relationship of **neighboring vectors:**

For example, let $\mathbf{x} \in \mathcal{A}$, and $\mathbf{y} \in \mathcal{A}$ be neighbors, if they only differ in one coordinate.

Let $N(\mathbf{x})$ be the set of neighbors of $\mathbf{x}$

Simulated Annealing

Let $q(\mathbf{x}, \mathbf{y}) = P(\mathbf{y}|\mathbf{x}) = \frac{1}{|N(\mathbf{x})|}$ Uniform distribution

We want $\pi(\mathbf{x}) = \exp(\lambda V(\mathbf{x}))$ limit distribution.

Use the Hastings- Metropolis sampling:

$$\alpha(\mathbf{x}, \mathbf{y}) = \min\left(\frac{\pi(\mathbf{y})q(\mathbf{y}, \mathbf{x})}{\pi(\mathbf{x})q(\mathbf{x}, \mathbf{y})}, 1\right)$$

$$= \min\left(1, \frac{\exp(\lambda V(\mathbf{y}))|N(\mathbf{y})|}{\exp(\lambda V(\mathbf{x}))|N(\mathbf{x})|}\right)$$

Simulated Annealing: Pseudo Code

1) At iteration t we are in $\mathbf{x}(t) \in \mathcal{A}$
2) Let $\mathbf{z} \in \mathcal{A}$ be a neighbor of $\mathbf{x}(t)$.

   drawn from uniform distribution $(\frac{1}{|N(x(t))|})$.

3) Let $u \sim U_{[0,1]}$
4) If $u < \alpha(\mathbf{x}(t), \mathbf{z}) \Rightarrow \mathbf{x}(t+1) = \mathbf{z}$

   With prob. ® accept the new state

   If $u \geq \alpha(\mathbf{x}(t), \mathbf{z}) \Rightarrow \mathbf{x}(t+1) = \mathbf{x}(t)$

   with prob. (1-®) don't accept and stay

5) Back to 2

Simulated Annealing: Special case

$$\text{If } |N(\mathbf{z})| = |N(\mathbf{x})| \ \forall \mathbf{x}, \mathbf{z} \in \mathcal{A}$$
$$\Rightarrow \alpha(\mathbf{x}, \mathbf{z}) = \min\left\{1, \exp(\lambda(V(\mathbf{z}) - V(\mathbf{x}))\right\}$$

In this special case:

4)

$$\text{If } V(\mathbf{z}) \geq V(\mathbf{x}(t)) \Rightarrow \mathbf{x}(t+1) = \mathbf{z}$$

With prob. ®=1 accept the new state since we increased V

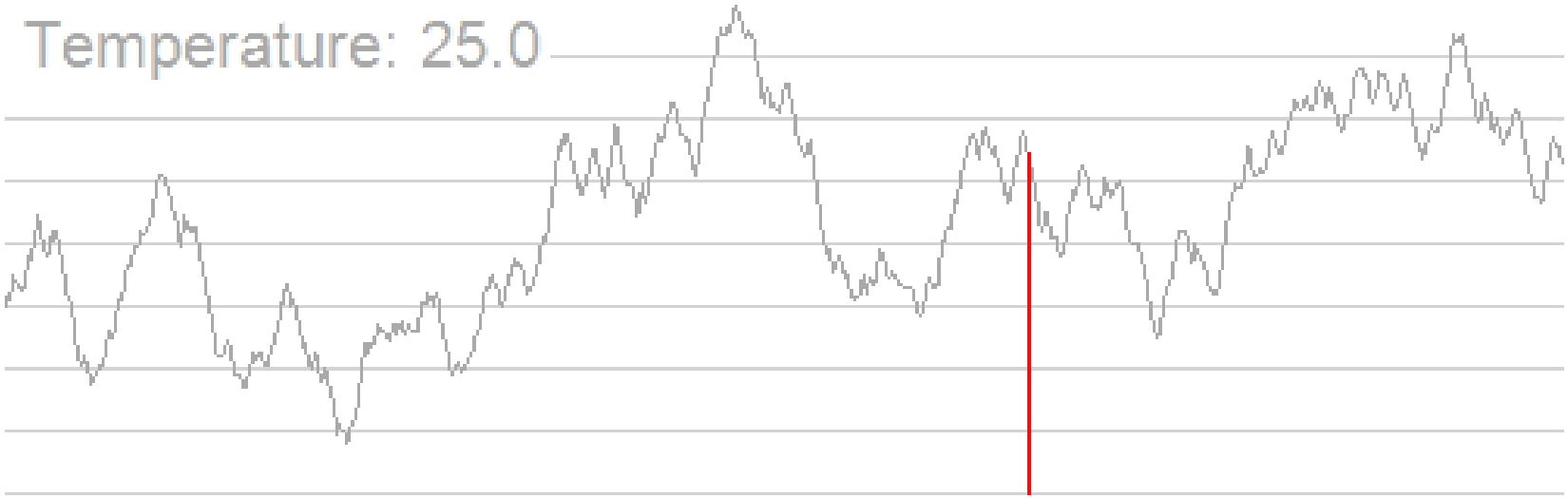
$$\text{If } V(\mathbf{z}) < V(\mathbf{x}(t)) \Rightarrow \begin{cases} \exp\{\lambda[V(\mathbf{z}) - V(\mathbf{x}(t))]\} = \alpha < 1 \\ \text{with prob. } 1-\alpha\text{: } \mathbf{x}(t+1) = \mathbf{x}(t) \\ \text{with prob. } \alpha\text{: accept } \mathbf{z}, \ \mathbf{x}(t+1) = \mathbf{z} \end{cases}$$

Simulated Annealing: Problems

- If $V(\mathbf{z}) < V(\mathbf{x}(t))$, then the probability to move to $\mathbf{z}$ is exp small.
- If λ is big and $\mathbf{x}(t)$ is a local maximum, then it might take for a looong time to get to a new $\mathbf{z}$ from $\mathbf{x}(t)$.
- Nonetheless, we need big λ to find the set $\mathcal{M}$.
- Solution: Increase λ slowly, e.g. $\lambda_t = c \log(1+t)$, $c > 0$

Simulated Annealing

Temperature = 1/ ¸



Simulated Annealing