Algorithmic Foundations of the Metropolis Algorithm and the Markov-Chain Monte Carlo Method

Sorin Istrail

Department of Computer Science
Brown University, Providence
sorin@cs.brown.edu

June 27, 2016

Outline

Markov Chains

Definition Given a Markov Chain $MC = (\varphi, P)$. We say that a row verctor $\varphi = (\varphi_1, \varphi_2, ..., \varphi_k)$ is said to be a **stationary distribution** for the MC if it satisfies:

- $\varphi_i \geq 0, 1 \leq i \leq k, \sum_{i=1}^{k} \varphi_i = 1$
- $\varphi P = \varphi$, i.e., $\sum_{i=1}^{k} \varphi_i P_{i,j} = \varphi_j, 1 \leq j \leq k$

Stationary Distribution

Theorem (Existence and Uniqueness to Stationary Distributions)
For any irreducible and aperiodic MC there is a unique stationary distribution.

—

Definition of Total Variance Distance
If $v^1 = (v_1^1, ...v_k^1)$ and $v^2 = (v_1^2, ..., v_k^2)$ are probability discributions on $S = \{s_1, ..., s_k\}$ then we define the **total variation distance** between v^1 and v^2 as

$$d_{TV}(v^1, v^2 = \frac{1}{2} \sum_{1}^{k} \mid v_i^1 - v_i^2 \mid$$

Total Variation Distance

Properties of Total Variation Distance

1. If $d_{TV}(v^1, v^2) = 0$ then $v^1 = v^2$
2. If $d_{TV}(v^1, v^2) = 1$ then v^1 and v^2 are "disjoint" in the sense that $S = S^1 \cup S^2$ and v^1 puts its probability on S^1 and v^2 puts its probability on S^2.
3. the Total Variation Distance has also the equivalent natural interpretation:

$$d_{TV}(v^1, v^2) = MAX_{A \subset S} \mid v^1(A) - v^2(A) \mid$$

i.e., the maximal difference between the probabilities that the two distributions assign to any event

Convergence to Equilibrium

Theorem. (Convergence)
Let $(X_1, X_2, ...)$ be an irreducible aperiodic MC with state space $S = \{s_1, ..., s_k\}$ and transition matrix P and an arbitrary initial distribution π^0 . Then for any distribution φ which is stationary for P we have:

$$\pi^0 \rightarrow^{TV} \varphi$$

We say in this case that the MC is approaching **equilibrium** as $n \rightarrow \infty$

Reversible Markov Chains

Definition Let $(X_0, X_1, ...)$ be a MC with state space $S = \{s_1, s_2, ..., s_k\}$ and transition probability P. A probability distribution π is **reversible** for the chain if for all $i, j \in \{1, 2, ...k\}$ we have

$$\pi_i P_{i,j} = \pi_j P_{j,i}$$

A MC is **reversible** if there is a reversible distribution for it.

—

Theorem (A strong form of equilibrium)
If π is a reversible distribution for the MC, then it is a stationary distribution for the MC.

Random Walks on Graphs

An example. Let us consider a graph G that is a triangle with vertices v_0, v_1, v_2. Let us take a random walk on the G. Suppose that we are at node v_i. Flip a fair coin. If we get H then we move to $v_{(i+1(mod3))}$ and if we get T then we move to $v_{(i-1(mod3))}$. Suppose now that we start at v_0. Let with X_n denotes the index of the vertex at the walk at time n. We obtain the chain $(X-0, X_1, ...)$ Then:

- $Pr(X_1 = 1) = \frac{1}{2}$
- $Pr(X_2 = 2) = \frac{1}{2}$

...

Random Walks on Graphs

Definition A **graph** $G = (V, E)$ consists of vertices $V = \{v_1, ..., v_k\}$ and edges $E = \{e_1, ..., e_l\}$. Two vertices are **adjacent** if they share an edge.
A **random walk on a graph** $G = (V, E)$ is a Markov Chain with state space $V = \{v_1, ..., v_n\}$ and the following transition mechanism : If at vertex v_i at time n it moves at time $n+!$ to one of the neighbours of v_i chosen at random with equal probability for each neighbour. The degree of a verted v_i is the number of neighbours d_i of it.

- $P_{i,j} = \frac{1}{d_i}$ if i, and j are neighbours; and
- $P_{i,j} = 0$ otherwise.

Random Walks on Graphs

Theorem

The stationary distribution for this Markov Chian is

$$\varphi = (\frac{d_1}{d}, ..., \frac{d_k}{d})$$

where $d = \sum_{i=1}^{k} d_i$.

It is easy to see that φ is a reversible distribution for the Markov Chain.

Proof

- If v_i and v_j are neighbours (adjacent) then

$$\varphi_i P_{i,j} = \frac{d_i}{d} * \frac{1}{d} = \frac{1}{d} = \frac{d_j}{d} * \frac{1}{d} = \varphi_j P_{j,i}$$

- If If v_i and v_j are not neighbours (adjacent) then

The Metropolis Algorithm

- We want to simulate a given probability distribution $\varphi = (\varphi_1, ..., \varphi_k)$ on a set $S = (s_1, ..., s_k)$.
- The first step is to construct a graph G with vertex set S.
- We construct edges in G such that
  - The graph must be connected to assure irreducibility of the resuting chain
  - Each vertex should not be with high degrees as such a Markov chain is "heavy" - observe that in the stationary distribution on the standard random walk on a graph that the time visiting a certain vertex is proportional to its degree.

The Metropolis Algorithm

The following is the Metropolis Markov Chaint probability transition matrix corresponding to the graph $G = (V, E)$

- If $(s_i, s_j) \in E$ then

$$P_{i,j} = \frac{1}{d} MIN\{\frac{\varphi_j d_i}{\varphi_i d_j}, 1\}$$

- If $(s_i, s_j) \not\in E$ (s_i is not adjacent ot s_j) then

$$P_{i,j} = 0$$

- If $i = j$ then

$$P_{i,j} = 1 - \sum_{(s_l, s_i) \in E} \frac{1}{d} MIN\{\frac{\varphi_l d_i}{\varphi_i d_l}, 1\}$$

The Metropolis Algorithm

The transition $P_{i,j}$ corresponds to the following mechanism.

- Suppose $X_n = s_i$
- First pick a state s_j according to uniform distribution to the set of neighbours of s_i, so each neighbour is chosen with probability $\frac{1}{d_i}$
  Then
  - $X_{n+1} = s_j$ with probability $MIN\{\frac{\varphi_j d_i}{\varphi_i d_j}, 1\}$ (move to a neighbour state) or
  - $X_{n+1} = s_i$ with probability $1 - MIN\{\frac{\varphi_j d_i}{\varphi_i d_j}, 1\}$ (remains in the same state)

The Metropolis Algorithm

To show that this mechanism has φ as its stationary distribution, it is enough to verify that the reversibility condition

$$\varphi_i P_{i,j} = \varphi_j P_{j,i}$$

for all i, j. We prove this in three steps.

- For $i = j$ we have
$$\varphi_i P_{i,i} = \varphi_i P_{i,i}$$
- For $i \neq j$ and $(s_i, s_j) \notin E$ both sides are equal to zero as $P_{i,j} = 0$
- For $i \neq j$ and $(s_i, s_j) \in E$ we have two cases to consider.
  1. CASE A. $\frac{\varphi_j d_i}{\varphi_i d_j} \geq 1$
  2. CASE B. $\frac{\varphi_j d_i}{\varphi_i d_j} < 1$

The Metropolis Algorithm

1. CASE A. If $\frac{\varphi_j d_i}{\varphi_i d_j} \geq 1$ then $\frac{\varphi_i d_j}{\varphi_j d_i} <= 1$
   Then $\varphi_i P_{i,j} = \varphi_i * \frac{1}{d_i} * 1 = \frac{\varphi_i}{d_i}$
   Also $\varphi_j P_{j,i} = \varphi_j * \frac{1}{d_j} * \frac{\varphi_i d_j}{\varphi_j d_i} = \frac{\varphi_i}{d_i}$
   In conclusion: $\varphi_i P_{i,j} = \varphi_j P_{j,i}$ (reversibility)
2. CASE B. If $\frac{\varphi_j d_i}{\varphi_i d_j} < 1$ then $\frac{\varphi_i d_j}{\varphi_j d_i} > 1$
   Then $\varphi_i P_{i,j} = \varphi_i * \frac{1}{d_i} * \frac{\varphi_j d_i}{\varphi_i d_j} = \frac{\varphi_j}{d_j}$
   Also we have
   $\varphi_j P_{j,i} = \varphi_j * \frac{1}{d_j} * (MIN\{\frac{\varphi_i d_j}{\varphi_j d_i}, 1\}) = \varphi_j * \frac{1}{d_j} * 1 = \frac{\varphi_j}{d_j} * 1 = \frac{\varphi_j}{d_j}$
   In conclusion: $\varphi_i P_{i,j} = \varphi_j P_{j,i}$ (reversibility).

The Hard-Core Model - in statistical physics

Consider a graph $G = (V, E), V = \{v_1, ..., v_n\}, E = \{e_1, ..., e_l\}$
Randomly assign value 0 and 1 on each vertex, such that no two adjacent vertices (endpoints of an edge) both take value 1.
An assignment of 0's and 1's to the vertices is called a **configuration** $C : V \to \{0, 1\}$.
The set of all configurations is $\{0, 1\}^V$. A configuration is **feasible** if no two 1s are adjacent.

—

This is a statistical mechanics model, called "Hard-Core" as it tries to capture some of the behavior of gas molecules where particles have non-negative radii and cannot overlap;
1s represent a particles, and 0s empty spaces.

The Hard-Core Model - in statistical physics

We assign equal probaility to each configuration.
Consider μ_G a probability distribution on $\{0, 1\}^V$ defined as follows.

- If ξ is feasible then $mu_G(\xi) = \frac{1}{Z}$
- If ξ is not feasible then $mu_G(\xi) = 0$

Z is the number of feasible configurations.

—-

NATURAL QUESTION:
What is the expected number of 1s in a random configuration chosen according to μ_G?

The Hard-Core Model - in statistical physics

If we write $n(\xi)=$ the number of 1s in configuration ξ and we denote by X a random configuration chosen according to $\mu_G(\xi)$ then:

$$E[n(\xi)] = \sum_{\xi \in \{0,1\}^V} n(\xi)\mu_G(\xi) = \frac{1}{Z_G} \sum_{\xi \in \{0,1\}^V} n(\xi) I_{[\xi \, feasible]}$$

Hard to compute and $Z_G =$ the total number of feasible configurations on graph G is hard to compute as well.

The Hard-Core Model - in statistical physics

- To evaluate this sum is infeasible unless the graph is very small. For an 8x8 grid there are $2^64 = 10^19$ configurations.
- Most terms are zero but the number of non-zero terms grows exponential as well.
- **When we cannot compute $E[n(X)]$ we go to simulations!**

The Hard-Core Model - in statistical physics

- If we know how to simulate a random configuration X with distribution μ_G, then we can do this many tmes, and estimate $E[n(X)]$ by the average number of 1's in our simulation.
- By the **Law of Large Numbers** this estimate converges to the same true value of $E[n(X)]$ as the number of simultations tends to infinity.

The Hard-Core Model - in statistical physics

- How is it possible to be easier to construct a Markov Chain with the desired property than to construct a random variable with distribution φ directly?
- We typically solve such problems by finding a stronger Markov Chain satisfying the property of reversibility not just stationarity of the distribution.

The Hard-Core Model - in statistical physics

An Markov Chain Monte Carlo Algorithm for the Hard-Core Model on a graph G

We are at time n in configuration X_n. At time $n+1$ we do the following:

1. Pick a vertex $v \in V$ at random uniformly
2. Toss a fair coin
3. If the coin comes up Heads, and all neighbours at V take value 0 in X_n then we let $X_{n+1}(v) = 1$; otherwise $X_{n+1}(v) = 0$
4. For all vertices w other than v leave the value of w unchanged, i.e., $X_{n+1}(w) = X_n(w)$

It is not difficult to verify that this MC is irreducible and aperiodic and μ_G is reversible.