Modern Discrete Probability

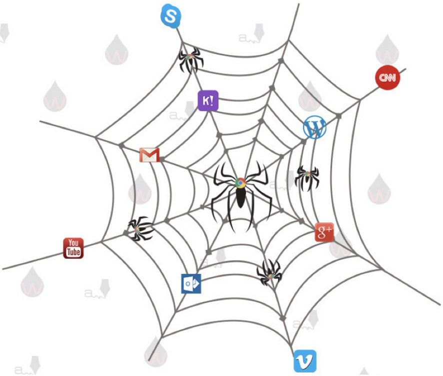
I - Introduction (continued)

Review of Markov chains

Sébastien Roch
UW–Madison
Mathematics

August 31, 2020

Exploring graphs



Random walk on a graph

Definition

Let $G = (V, E)$ be a countable graph where every vertex has finite degree. Let $c : E \to \mathbb{R}_+$ be a positive edge weight function on G. We call $\mathcal{N} = (G, c)$ a *network*. Random walk on $\mathcal{N}$ is the process on V, started at an arbitrary vertex, which at each time picks a neighbor of the current state proportionally to the weight of the corresponding edge.

Questions:

- How often does the walk return to its starting point?
- How long does it take to visit all vertices once or a particular subset of vertices for the first time?
- How fast does it approach equilibrium?

Undirected graphical models I

Definition

Let S be a finite set and let $G = (V, E)$ be a finite graph. Denote by $\mathcal{K}$ the set of all cliques of G. A positive probability measure μ on $\mathcal{X} := S^V$ is called a *Gibbs random field* if there exist *clique potentials* $\phi_K : S^K \to \mathbb{R}$, $K \in \mathcal{K}$, such that

$$\mu(x) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{K \in \mathcal{K}} \phi_K(x_K)\right),$$

where x_K is x restricted to the vertices of K and $\mathcal{Z}$ is a normalizing constant.

Undirected graphical models II

Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature β is the Gibbs random field with $S := \{-1, +1\}$, $\phi_{\{i,j\}}(\sigma_{\{i,j\}}) = \beta\sigma_i\sigma_j$ and $\phi_K \equiv 0$ if $|K| \neq 2$. The function $\mathcal{H}(\sigma) := -\sum_{\{i,j\}\in E} \sigma_i\sigma_j$ is known as the *Hamiltonian*. The normalizing constant $\mathcal{Z} := \mathcal{Z}(\beta)$ is called the *partition function*. The states $(\sigma_i)_{i\in V}$ are referred to as *spins*.

Questions:

- How fast is correlation decaying?
- How to sample efficiently?
- How to reconstruct the graph from samples?

Directed graphs

Definition

A *directed graph* (or digraph for short) is a pair $G = (V, E)$ where V is a set of *vertices* (or nodes, sites) and $E \subseteq V^2$ is a set of *directed edges*.

A *directed path* is a sequence of vertices $x_0, \ldots, x_k$ with $(x_{i-1}, x_i) \in E$ for all $i = 1, \ldots, k$. We write $u \to v$ if there is such a path with $x_0 = u$ and $x_k = v$. We say that $u, v \in V$ *communicate*, denoted by $u \leftrightarrow v$, if $u \to v$ and $v \to u$. The $\leftrightarrow$ relation is clearly an equivalence relation. The equivalence classes of $\leftrightarrow$ are called the *(strongly) connected components* of G.

Markov chains I

Definition (Stochastic matrix)

Let V be a finite or countable space. A *stochastic matrix* on V is a nonnegative matrix $P = (P(i,j))_{i,j \in V}$ satisfying

$$\sum_{j \in V} P(i,j) = 1, \qquad \forall i \in V.$$

Let μ be a probability measure on V. One way to construct a *Markov chain* (X_t) on V with transition matrix P and initial distribution μ is the following. Let $X_0 \sim \mu$ and let $(Y(i,n))_{i \in V, n \geq 1}$ be a mutually independent array with $Y(i,n) \sim P(i,\cdot)$. Set inductively $X_n := Y(X_{n-1}, n)$, $n \geq 1$.

Markov chains II

So in particular:

$$\mathbb{P}[X_0 = x_0, \ldots, X_t = x_t] = \mu(x_0)P(x_0, x_1)\cdots P(x_{t-1}, x_t).$$

We use the notation $\mathbb{P}_x, \mathbb{E}_x$ for the probability distribution and expectation under the chain started at x. Similarly for $\mathbb{P}_\mu, \mathbb{E}_\mu$ where μ is a probability measure.

Example (Simple random walk)

Let $G = (V, E)$ be a finite or countable, locally finite graph. *Simple random walk* on G is the Markov chain on V, started at an arbitrary vertex, which at each time picks a uniformly chosen neighbor of the current state.

Markov chains III

The *transition graph* of a chain is the directed graph on V whose edges are the transitions with nonzero probabilities.

Definition (Irreducibility)

A chain is *irreducible* if V is the unique connected component of its transition graph, i.e., if all pairs of states communicate.

Example

Simple random walk on G is irreducible if and only if G is connected.

Aperiodicity

Definition (Aperiodicity)

A chain is said to be *aperiodic* if for all $x \in V$

$$\gcd\{t \,:\, P^t(x,x) > 0\} = 1.$$

Example (Lazy walk)

A *lazy, simple random walk* on G is a Markov chain such that, at each time, it stays put with probability $1/2$ or chooses a uniformly random neighbor of the current state otherwise. Such a walk is aperiodic.

Stationary distribution I

Definition (Stationary distribution)

Let (X_t) be a Markov chain with transition matrix P. A *stationary measure* π is a measure such that

$$\sum_{x \in V} \pi(x) P(x,y) = \pi(y), \qquad \forall y \in V,$$

or in matrix form $\pi = \pi P$. We say that π is a *stationary distribution* if in addition π is a probability measure.

Example

The measure $\pi \equiv 1$ is stationary for simple random walk on $\mathbb{L}^d$.

Stationary distribution II

Theorem (Existence and uniqueness: finite case)

If P is irreducible and has a finite state space, then it has a unique stationary distribution.

Definition (Reversible chain)

A transition matrix P is *reversible* w.r.t. a measure η if $\eta(x)P(x,y) = \eta(y)P(y,x)$ for all $x, y \in V$. By summing over y, such a measure is necessarily stationary.

By induction, if (X_t) is reversible w.r.t. a stationary distribution π

$$\mathbb{P}_\pi[X_0 = x_0, \ldots, X_t = x_t] = \mathbb{P}_\pi[X_0 = x_t, \ldots, X_t = x_0].$$

Stationary distribution III

Example

Let (X_t) be simple random walk on a connected graph G. Then (X_t) is reversible w.r.t. $\eta(v) := \delta(v)$.

Example

The Metropolis algorithm modifies a given irreducible symmetric chain Q to produce a new chain P with the same transition graph and a prescribed positive stationary distribution π. The definition of the new chain is:

$$P(x,y) := \begin{cases} Q(x,y)\left[\frac{\pi(y)}{\pi(x)} \wedge 1\right], & \text{if } x \neq y, \\ 1 - \sum_{z \neq x} P(x,z), & \text{otherwise.} \end{cases}$$

Convergence

Theorem (Convergence to stationarity)

Suppose P is irreducible, aperiodic and has stationary distribution π. Then, for all x, y, $P^t(x, y) \to \pi(y)$ as $t \to +\infty$.

For probability measures μ, ν on V, let their *total variation distance* be $\|\mu - \nu\|_{\mathrm{TV}} := \sup_{A \subseteq V} |\mu(A) - \nu(A)|$.

Definition (Mixing time)

The *mixing time* is

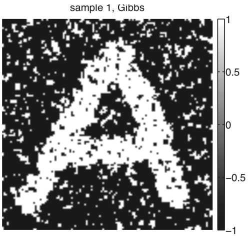
$$\mathrm{t_{mix}}(\varepsilon) := \min\{t \geq 0 \,:\, d(t) \leq \varepsilon\},$$

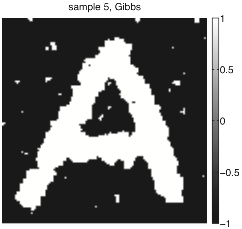
where $d(t) := \max_{x \in V} \|P^t(x, \cdot) - \pi(\cdot)\|_{\mathrm{TV}}$.

Other useful random walk quantities

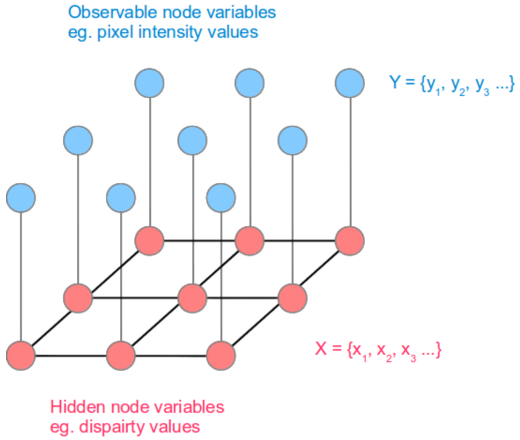
- Hitting times
- Cover times
- Heat kernels

Application: Bayesian image analysis I





Bayesian image analysis II



Recall: Undirected graphical models I

Definition

Let S be a finite set and let $G = (V, E)$ be a finite graph. Denote by $\mathcal{K}$ the set of all cliques of G. A positive probability measure μ on $\mathcal{X} := S^V$ is called a *Gibbs random field* if there exist *clique potentials* $\phi_K : S^K \to \mathbb{R}$, $K \in \mathcal{K}$, such that

$$\mu(x) = \frac{1}{\mathcal{Z}} \exp\left(\sum_{K \in \mathcal{K}} \phi_K(x_K)\right),$$

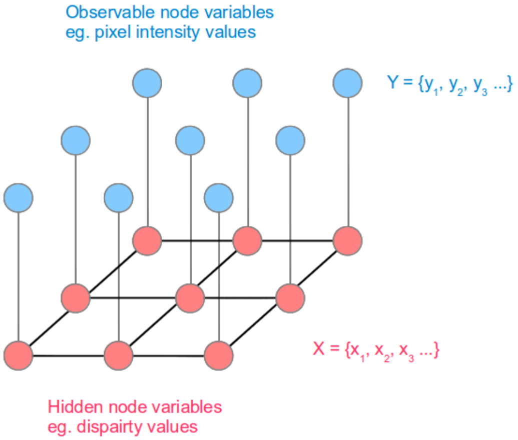
where x_K is x restricted to the vertices of K and $\mathcal{Z}$ is a normalizing constant.

Recall: Undirected graphical models II

Example

For $\beta > 0$, the *ferromagnetic Ising model* with inverse temperature β is the Gibbs random field with $S := \{-1, +1\}$, $\phi_{\{i,j\}}(\sigma_{\{i,j\}}) = \beta\sigma_i\sigma_j$ and $\phi_K \equiv 0$ if $|K| \neq 2$. The function $\mathcal{H}(\sigma) := -\sum_{\{i,j\}\in E} \sigma_i\sigma_j$ is known as the *Hamiltonian*. The normalizing constant $\mathcal{Z} := \mathcal{Z}(\beta)$ is called the *partition function*. The states $(\sigma_i)_{i\in V}$ are referred to as *spins*.

Back to Bayesian image analysis I



Back to Bayesian image analysis II

We assume the prior (i.e. distribution of hidden variables) is an Ising model $\mu_\beta(\sigma)$ on the $L \times L$ grid $G = (V, E)$. The observed variables τ are independent flips of the corresponding hidden variables with flip probability $q \in (0, 1/2)$, i.e.,

$$
\begin{aligned}
\mathbb{P}[\tau \mid \sigma] &= \prod_{i \in V} (1-q)^{\mathbb{1}_{\tau_i = \sigma_i}} q^{\mathbb{1}_{\tau_i \neq \sigma_i}} \\
&= \exp\left(\sum_{i \in V} \left\{ \log(1-q) \frac{1 + \sigma_i \tau_i}{2} + \log(q) \frac{1 - \sigma_i \tau_i}{2} \right\} \right) \\
&= \exp\left(\sum_{i \in V} \sigma_i \frac{\tau_i}{2} \log \frac{1-q}{q} + \mathcal{Y}(q) \right).
\end{aligned}
$$

Back to Bayesian image analysis III

By Bayes' rule, the posterior is then given by

$$\begin{aligned}\mathbb{P}[\sigma \mid \tau] &= \frac{\mathbb{P}[\tau \mid \sigma]\mu_\beta(\sigma)}{\sum_\sigma \mathbb{P}[\tau \mid \sigma]\mu_\beta(\sigma)} \\ &= \frac{1}{\mathcal{Z}(\beta, q)} \exp\left(\beta \sum_{i\sim j} \sigma_i\sigma_j + \sum_i h_i\sigma_i\right),\end{aligned}$$

where $h_i = \frac{\tau_i}{2} \log \frac{1-q}{q}$.

Gibbs sampling I

Definition

Let μ_β be the Ising model with inverse temperature $\beta > 0$ on a graph $G = (V, E)$. The *(single-site) Glauber dynamics* is the Markov chain on $\mathcal{X} := \{-1, +1\}^V$ which at each time:

- selects a site $i \in V$ uniformly at random, and
- updates the spin at i according to μ_β conditioned on agreeing with the current state at all sites in $V \backslash \{i\}$.

Gibbs sampling II

Specifically, for $\gamma \in \{-1, +1\}$, $i \in \Lambda$, and $\sigma \in \mathcal{X}$, let $\sigma^{i,\gamma}$ be the configuration σ with the spin at i being set to γ. Let $n = |V|$ and $S_i(\sigma) := \sum_{j \sim i} \sigma_j$. Then

$$
\begin{aligned}
Q_\beta(\sigma, \sigma^{i,\gamma}) &:= \frac{1}{n} \frac{\frac{1}{\mathcal{Z}(\beta)} \exp\left(\beta \sum_{j \sim k} \sigma_j^{i,\gamma} \sigma_k^{i,\gamma}\right)}{\sum_{i' = -,+} \frac{1}{\mathcal{Z}(\beta)} \exp\left(\beta \sum_{j \sim k} \sigma_j^{i',\gamma} \sigma_k^{i',\gamma}\right)} \\
&= \frac{1}{n} \cdot \frac{e^{\gamma \beta S_i(\sigma)}}{e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}}.
\end{aligned}
$$

The Glauber dynamics is reversible w.r.t. μ_β. How quickly does the chain approach μ_β?

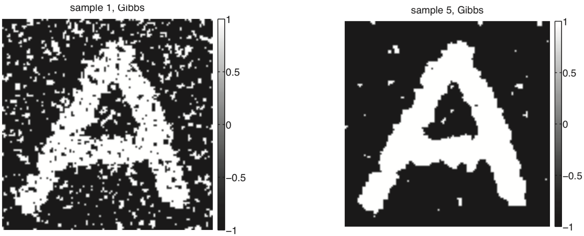
Gibbs sampling III

Proof of reversibility: This chain is clearly irreducible. For all $\sigma \in \mathcal{X}$ and $i \in V$, let $S_{\neq i}(\sigma) := \mathcal{H}(\sigma^{i,+}) + S_i(\sigma) = \mathcal{H}(\sigma^{i,-}) - S_i(\sigma)$. We have

$$\begin{aligned}
\mu_\beta(\sigma^{i,-})\, Q_\beta(\sigma^{i,-}, \sigma^{i,+}) &= \frac{e^{-\beta S_{\neq i}(\sigma)} e^{-\beta S_i(\sigma)}}{\mathcal{Z}(\beta)} \cdot \frac{e^{\beta S_i(\sigma)}}{n[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \frac{e^{-\beta S_{\neq i}(\sigma)}}{n\mathcal{Z}(\beta)[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \frac{e^{-\beta S_{\neq i}(\sigma)} e^{\beta S_i(\sigma)}}{\mathcal{Z}(\beta)} \cdot \frac{e^{-\beta S_i(\sigma)}}{n[e^{-\beta S_i(\sigma)} + e^{\beta S_i(\sigma)}]} \\
&= \mu_\beta(\sigma^{i,+})\, Q_\beta(\sigma^{i,+}, \sigma^{i,-}).
\end{aligned}$$

■

Back to Bayesian image analysis



Go deeper

More details at:

`http://www.math.wisc.edu/~roch/mdp/`