A. Rhodes

TAT GIO : March 2nd, 2017

Intro To Graphical Models

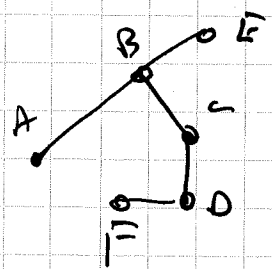
Why deploy graphical models in ML-related domains? A: (1) They are often interpretable and give a principled, model-based framework of a phenomenon of study (compare to a "black Box" NN)

(2) We can easily encode domain/structural knowledge into a GM.

(3) They offer computational savings, since they compactly encode: $p(x|\theta)$.

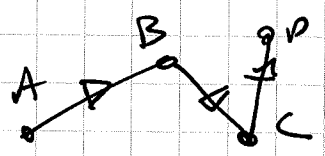
In general, there are (2) paradigms for GMs: (J. Pearl) 1988

(1) MRFs (undirected GMs)



vs.

(2) Bayes Nets ("Causal" Models) (Directed) (Belief Nets)



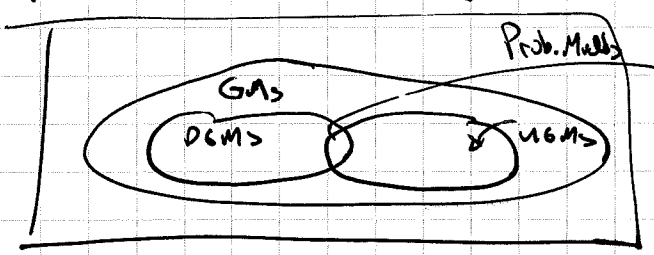
NB: MRFs & Bayes Nets are not equivalent!

More generally, we may ask: What is the expressive power of an MRF vs. a Bayes Net?

We say: G (a graph) is an I-map of a distribution p if $I(G) \subseteq I(p)$, which is to say G does not make any conditional independence statements that are not true of p itself (Here: $I(p)$ stands for the collection of independence assumptions encoded by p).

If $I(G) = I(p)$ we say G is a "perfect" map of p.

It turns out that the set of all probability distributions that admit of an MRF representation is different from the set of distributions that can be represented by DGMs (i.e. Bayes Nets)



overlap are called "chordal" graphs.

Let's further draw out the distinction b/w UGMs & DGMs for deeper understanding.

UGMs (MRFs)

Advantages of UGMs over DGMs:

1. UGMs are "symmetric" & thus "natural" in certain domains (e.g. vision)
2. They can be used discriminatively (so they can be fine-tuned using labeled data)

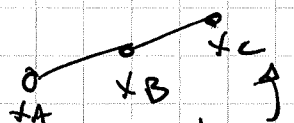
Disadvantages:

1. Parameters can be less interpretable
2. (P) Parameter estimation is computationally burdensome.

Key Properties of MRFs

3. Equivalent Notions of CI (conditional indep.)
   1. Global Markov Property:
      $X_A \perp X_B \mid X_C$ iff Nodes in C separate A & B in graph.

   2. Local Markov Property:
      $X_A \perp X_B \mid N(X_A)$, $X_B \notin N(X_A)$



(3) Pairwise Markov Property

$$X_A \perp X_B \mid V(G) \setminus \{X_A, X_B\} \longleftrightarrow G_{AB} = 0$$

This says X_A & X_B are conditionally independent given all other vertices in the graph iff $X_A \not\perp X_B$ in G.

Recall: (1) $\longleftrightarrow$ (2) $\longleftarrow$ (3)

More succinctly. Define the **Markov Blanket** of a node: x_A as the smallest set of nodes in G that renders x_A conditionally independent of all other nodes in the graph.

(*)

Notationally: $$x_A \perp V(G) \setminus \text{closure}(x_A) \mid MB(x_A)$$

where $\text{closure}(x_A) \triangleq N(x_A) \cup \{x_A\}$ &

$MB(x_A) \triangleq$ Markov Blanket of x_A.

NB: $MB(x_A) \triangleq N(x_A)$ for an MRF,

so (*) is equivalent to (2), the Local Markov Property.

Conditional Independence Properties of DGMs (Bonus)

Note that CI properties are less "intuitive" for DGMs vs. UGMs.

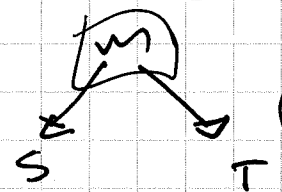
CI properties for DGMs are codified as the notion of **d-separation**.

All told, there are (3) distinct cases we need to consider:

We say an undirected path P is **d-separated** by a set of nodes "E" (containing evidence/observation) iff at least one of the following holds:

(1) P contains a chain: $S \to M \to T$ or $T \to M \to S$, where $M \in E$

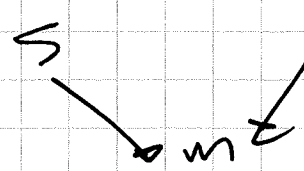
(2) P contains a fork:



where $M \in E$

(*)

(3) P contains a v-structure:



Note: (where $M \notin E$ nor is any descendant of E)

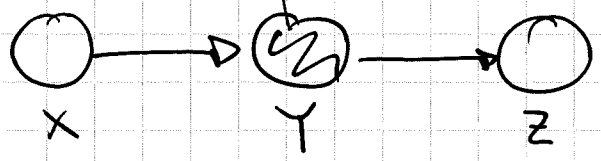
The CI properties of a DGM (Technically: an acyclic DGM) are defined Thus:

$$X_A \perp X_B \mid X_E \quad \longleftrightarrow \quad A \text{ is } \underline{\text{d-separated}} \text{ from } B \text{ given } E$$

Remark: It is helpful to think of d-separation as "stopping the flow of information" from one set of nodes to another. ("Bayes Ball Theorem")

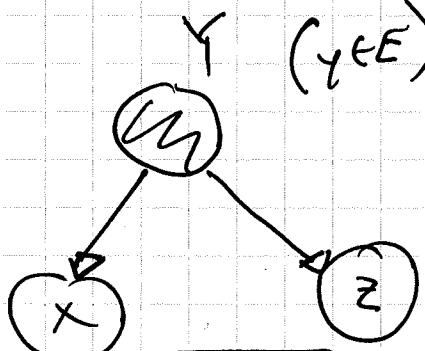
Examples

1)



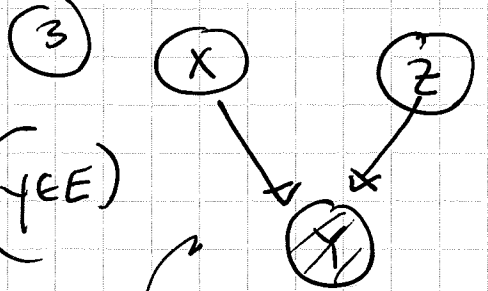
$X \perp Z \mid Y$

2)



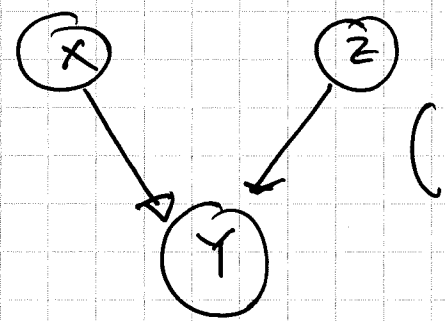
$X \perp Z \mid Y$

3)



$X \not\perp Z \mid Y$

Note ↑



$X \perp Z \mid Y$

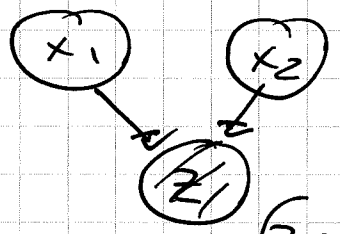
Remark: Don't be nonplussed by case (3).

The v-structure case is simply the effect known as "explaining away" evidence (Berkson's Paradox).

Ex. For easy illustration: where $X_1 \perp X_2$

Suppose $X_1, X_2 \sim Ber(.5)$ $S = \{0, 1\}$

Define $Z = X_1 + X_2$

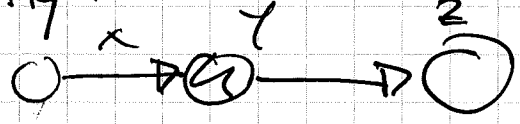


Then $P(X_1 = 0 \mid Z = 0) = 1$, ($Z \in E$)

$P(X_2 = 1 \mid Z = 0) = 0$, etc.

so, $X_1 \not\perp X_2 \mid Z$, as promised.

In more detail, we can prove the CI/d-separation properties easily:

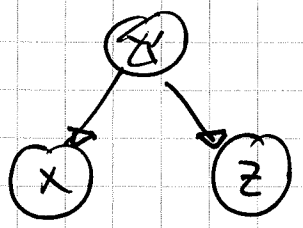


(1)

$$p(x,y,z) = p(x)p(y|x)p(z|y)$$ (By "chain rule of Prob.")

Conditioning on y: $$p(x,z|y) = \frac{p(x)p(y|x)p(z|y)}{p(y)} = \frac{p(x,y)p(z|y)}{p(y)} = p(x|y)p(z|y)$$

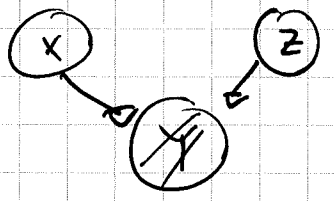
2



$$p(x,y,z) = p(y)p(x|y)p(z|y)$$

$$\rightarrow p(x,z|y) = \frac{\cancel{p(y)}p(x|y)p(z|y)}{\cancel{p(y)}} = p(x|y)p(z|y). \quad \square$$

3



$$p(x,y,z) = p(x)p(z)p(y|x,z)$$

Suppose $y \in E \rightarrow p(x,z|y) = \dfrac{p(x)p(z)p(y|x,z)}{p(y)}$

$$\neq p(x|y) \cdot p(z|y)$$

clearly, $\boxed{x \not\perp z \mid y}$

However, $p(x,z) = p(x) \cdot p(z)$ for the unconditional distribution.

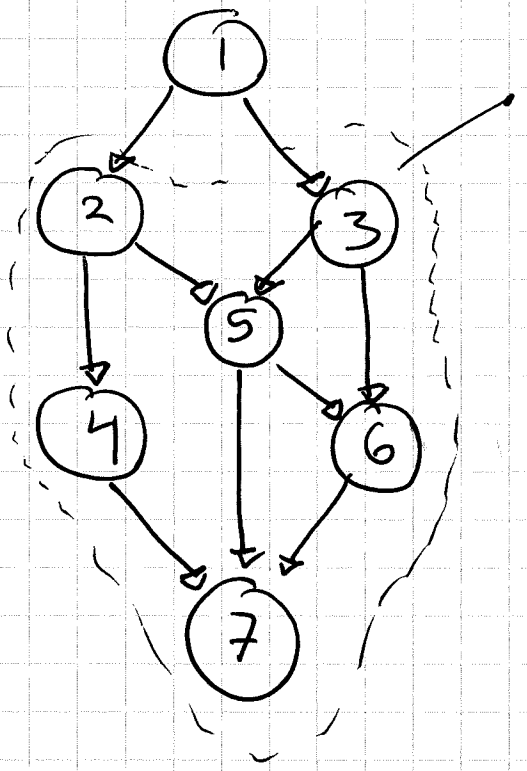
By analogy with MRFs, we can define the

Markov Blanket for DGMs.

Recall that the M.B. of a node in a graph effectively renders that node conditionally independent of all other nodes in the graph.

For a DGM, $MB(x_A) \triangleq$ Children $(x_A) \cup$ Parents (x_A) $\cup$ (co-parents (x_A)

Example:



$MB(x_5) =$

$\{x_6, x_7, x_2, x_3, x_4\}$

Children, parents, co-parent

You are welcome to verify, for instance, that:

$$p(x_5 | x_{-5}) \propto p(x_5 | x_2, x_3)\, p(x_6 | x_3, x_5)\, p(x_7 | x_4, x_3, x_6)$$

$x_{-5} \triangleq V \setminus \{x_5\}$

Returning now to our first discussion point — remember that DGMs $\neq$ UGMs are perfect maps for different sets of distributions.

(However neither regime is more "powerful" than the other)

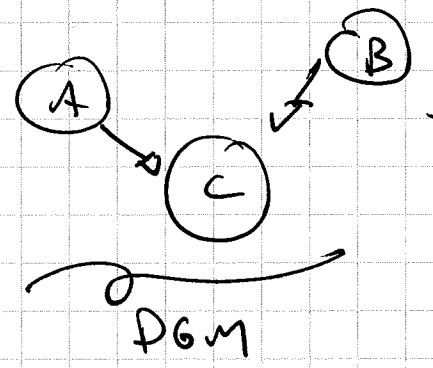
We are now equipped to see this distinction directly.

Q: Can we simply convert, say, a DGM → UGM?

A: NOT quite. In practice, one can "moralize" an UGM (by connecting co-parents) but this process unfortunately loses some CI information.

(Also: converting UGM → DOM is problematic because of normalization constraints).

As an example of some CI relationships that can't be modeled perfectly by a UGM — but can be modeled by a DGM consider:



→ implicitly models (2) CI claims:

(1) $A \perp B$

(2) $A \not\perp B \mid C$

You are welcome to confirm that no UGM can represent precisely (& only) these (2) CI statements!

Inference in GMs

Most often, we are interested in computing $p(x|\theta)$ for some density of interest. As we show, sampling can furnish us with a mechanism to approximate (or even exactly solve) problems of high complexity (e.g. NP-hard TSP).

In an explicit optimization setting, we can use a sampling-based procedure called simulated annealing.

Remember that GMs provide many useful advantages as a ML tool - least of not which are their (potential) computational savings.

In the main, if $x_i \in \{1, .., K\}$ then the full specification of $p(x)$, for $x \in \mathbb{R}^n$ $x \in \{1, .., K\}^D$ is $\mathcal{O}(|K|^D)$ - which is infeasible in a general setting.

By encoding various CI conditions, a GM greatly reduces this computation.

In domains for which "causality"/directionality attributions are appropriate (e.g. Medical Diagnosis), we would use a DGM. Here, however, we focus on inference for UGM, with specific applications to Gibbs sampling & image segmentation in Computer vision. (Our GMs are thus more conducive to spatial - as opposed to causal - structures).

Hammersley-Clifford Theorem

Since UGM have no inherent Topological ordering, we can't use the "chain rule" to represent $p(x|\theta)$. Instead, we associate potential functions with Maximal cliques in the graph; H-M asserts that $p(x|\theta)$ factors as a product of potentials over maximal cliques. (if $I(\theta) \subseteq I(\varphi)$, then we can write $p(x|\theta)$ iff

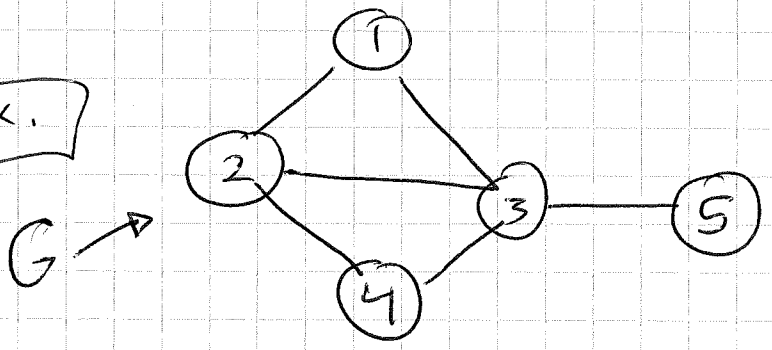
H-M:

$$p(x|\theta) = \frac{1}{Z(\theta)} \prod_{c \in C} \psi_c(x_c|\theta_c)$$

$$Z(\theta) \triangleq \sum_x \prod_{c \in C} \psi_c(x_c|\theta_c)$$ partition function

→ see Koller for proof

Ex.



$$p(x|\theta) = \frac{1}{Z(\theta)} \psi_{123}(x_1, x_2, x_3)\, \psi_{234}(x_2, x_3, x_4)\, \psi_{35}(x_3, x_5)$$

In fact, as a consequence of H-C, one can show that $p(x|\theta)$ can, equivalently, be parameterized according to the edges (only), instead of maximal cliques. (Called a Pairwise MRF)

Returning to the example above, we have:

$$p(x|\theta) \propto \prod_{s \sim T} \psi_{ST}(x_s, x_T)$$

$$= \psi_{12}(x_1, x_2)\, \psi_{13}(x_1, x_3)\, \psi_{23}(x_2, x_3)$$

$$\cdot \psi_{24}(x_2, x_4)\, \psi_{34}(x_3, x_4)\, \psi_{35}(x_3, x_5)$$

Broadly speaking, inference for GMs is commonly performed using "message-passing" (Pearl) algorithms (e.g. Belief Propagation / sum-product).

Importantly, exact inference ($\mathcal{O}(|V(G)|)$) is possible for tree-structures (we simply propagate "messages" from the root). However, in most practical applications only approximate inference is possible (using an iterative, "loopy belief" propagation.) - convergence (or approximate convergence), is often not guaranteed & difficult, moreover, to ascertain.

We content ourselves to consider a classical, Gibbs model, inspired by statistical physics.

Define the potential (over a clique/edge):

$$\psi_c(x_c|\theta_c) \triangleq \exp(-E(x_c|\theta_c)), \text{ for}$$

some energy function: $E(x_c) > 0$.

The Gibbs distribution is defined:

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left(-\sum_c E(x_c|\theta_c)\right)$$

This is an "energy-based" model – whereupon high probability states correspond to low energy configurations.

In the binary case (i.e. $x_i \in \{-1, 1\}$) (e.g. Ising Model)

we can write: ~~$\log p(x|\theta) = \gamma \sum_{s \sim t} x_s x_t$~~

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left[\gamma \sum_{s \sim t} x_s x_t\right]$$

Often times, we consider the presence of an external field using a local energy term.

Finally, this gives us:

$$p(x|\theta) = \frac{1}{Z(\theta)} \left[\sum_{s \in S} \lambda_s x_s + \gamma \sum_{s \sim t} x_s x_t\right]$$

General Version of Gibbs Sampling

We want to sample from: $p(x|\theta)$ $(x \in \mathbb{R}^d)$

Basic Idea: Sample, sequentially, from full conditionals.

Gibbs Sampling

(1) Initialize: $\{x_i : i = 1, \ldots, D\}$

(2) For $t = 1, \ldots, T$

loop

sample: $x_1^{(t+1)} \sim p(x_1 | x_{(-1)}^{(t)})$ "full conditional"

sample: $x_2^{(t+1)} \sim p(x_2 | x_{(-2)}^{(t)})$

$\vdots$

sample: $x_D^{(t+1)} \sim p(x_D | x_{(-D)}^{(t)})$

Complications: (i) How-to "order" samples (can be random, but careful)

(ii) Need full conditionals - can use an approximation or M-H procedure

Under "nice" assumptions (ergodicity), we get $x \sim p(x|\theta)$.

Sampling Algorithms More Generally

Q: How to perform posterior inference more generally?

Oftentimes we can't simply rely on strong parametric assumptions (eg. conjugacy / exp. family structures).

Monte Carlo Approximation / inference can get around this.

Basic Idea: ① Draw samples: $x^s \sim p(x|\theta)$

② Compute quantity of interest, eg.

Marginals: $E[p(x_1|\theta)] = \frac{1}{S}\sum_{i=1}^{S} x_{1,i}$, etc.

In general:

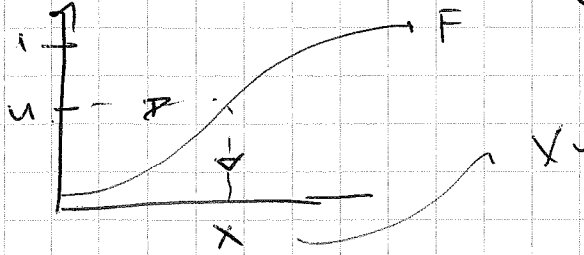
$$E[f(x)] = \int f(x)\, p(x|\theta)\, dx = \frac{1}{S}\sum^{S} f(x^s)$$

Common MC Sampling Techniques:

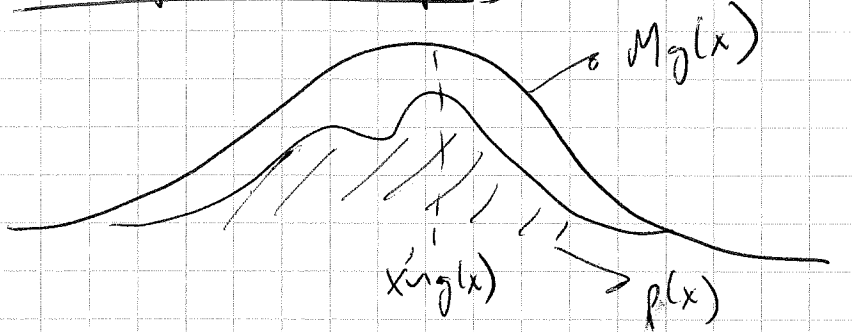
① CDF method:

① sample $U \sim U(0,1)$

② $F^{-1}(U) \sim F$



② Rejection Sampling



Want: $x \sim p(x)$

Idea: ① Sample $x' \sim g(x)$ ← proposal distribution where $\tilde{p} \leq Mg(x)$ $\left(p(x) = \frac{\tilde{p}}{Z}\right)$

② Sample $U \sim U(0,1)$

③ Reject x' if $U > \frac{\tilde{p}(x)}{Mg(x)}$, else accept.

One can show that $x' \sim p(x)$.

Issues: Need "Good" $g(x)$, M value & rejection rate can grow astronomically in high dimensions!

Pros of MC sampling: samples are independent!

Cons of MC sampling: very inefficient in high dimensions!

Alternatively, we can use MCMC Methods

Pros: Works in H-D

Cons: Dependent samples, convergence can be problematic.

MCMC Sampling Methods

Idea: Build a MC (Markov Chain) - Take a random walk (for a sufficiently long Time - called the "burn in" phase); Then $x^{(T)} \sim p(x|\theta)$.

The All-Time Classic MCMC Method:

Metropolis-Hastings Algorithm (Los Alamos)

Idea: Want sample: $x \sim p(x|\theta)$, where, possibly, $p(x|\theta)$ is high-dimensional.

(Initialize: x)

(1) Define proposal distribution: $q(x'|x)$.

The probability we transition: $x \to x'$.

Usually: $q(x'|x) = N(x'|x, \Sigma)$.

(2) Sample: $x' \sim q(x'|x)$

(3) Compute: (acceptance probability)

$$\alpha = \frac{\tilde{p}(x')\, q(x|x')}{\tilde{p}(x)\, q(x'|x)}$$

(unnormalized $\tilde{p} = \frac{p}{Z}$)

let $r = \min(1, \alpha)$

(4) sample $u \sim U(0,1)$ (5) $x^{(t+1)} = \begin{cases} x' & \text{if } u < r \\ x & \text{if } u \geq r \end{cases}$

Remark: We can show that Gibbs sampling is a special case of M-H (where we only accept).

Gibbs Sampling for Binary Segmentation

Define: $p(x_T \mid x_{(-T)}, \theta) = \prod_{s \in N(T)} \psi_{ST}(x_s, x_T)$

↑ "full conditional" J: coupling strength

Define the edge potentials: $\psi(x_s, x_T) = \exp(J x_s x_T)$

& $x_T \in \{-1, 1\}$.

$$p(x_T = +1 \mid x_{(T)}, \theta) = \frac{\prod_{s \in N(T)} \psi_{ST}(x_T = +1, x_s)}{\prod_{s \in N(T)} \psi(x_T = +1, x_s) + \prod_{s \in N(T)} \psi(x_T = -1, x_s)}$$

$$= \frac{\exp\left[J * \sum_{s \in N(T)} x_s\right]}{\exp\left[J * \sum_{s \in N(T)} x_s\right] + \exp\left[-J * \sum_{s \in N(T)} x_s\right]}$$

$$= \frac{\exp[J \eta_T]}{\exp[J \eta_T] + \exp[-J \eta_T]} = \phi(2 J \eta_T) \quad \left(\eta_T = \sum_{s \in N(T)} x_s\right)$$

(ϕ = sigmoid)

Remark: It is not difficult to see that:

$$\eta_T = x_T(a_T - d_T)$$

a_T: # agrees neighbors

d_T: # disagrees neighbors

Consequently, if, say $a_T = d_T$, we get a uniform conditional.

For image segmentation/denoising, we include a local evidence term: ψ_T. (e.g. $\psi_T(x_T) = N(y_T | x_T, \sigma^2)$)

With local evidence added, then, we have:

$$p(x_T = (+1) \mid x_{(-T)}, y, \theta) = \frac{\exp[J\eta_T]\psi_T(+1)}{\exp[J\eta_T]\psi_T(+1) + \exp[-J\eta_T]\psi_T(-1)}$$

(2: local evidence)

$$= \sigma\left(2J\eta_T - \log\frac{\psi_T(+1)}{\psi_T(-1)}\right)$$

Now the probability of x_T entering each state is determined by both compatibility with its neighbors & compatibility with the data (local likelihood term).

(*) For segmentation, we would execute the above Gibbs sampling technique for many iterations.

Lastly, as an addendum: one can also apply the technique of **simulated annealing** to the preceding procedure.

Simulated Annealing

Idea: A stochastic algorithm used in optimizing of a black-box function.

We assume: $p(x|\theta) \propto \exp\left(-\frac{f(x)}{T}\right)$,

where $f(x)$ is the "energy" of the system & T is the Temperature.

① Given $x^{(t)}$, sample $x' \sim g(x'|x^{(t)})$ ← proposal

② Set $\alpha = \exp\left[\frac{f(x)-f(x')}{T}\right]$

③ Accept new state x' with probability: $\min(1,\alpha)$.

(*) In practice we impose a cooling schedule: $T \to 0$.

(*) NB: This algorithm allows "down-hill" moves, but less frequently as $T \to 0$. If we cool sufficiently slowly, algorithm finds global optimum.