Inference in Computer Science and Systems Biology Part I

Alfredo Braunstein

June 11, 2012

Outline

Transcriptional Gene Regulation

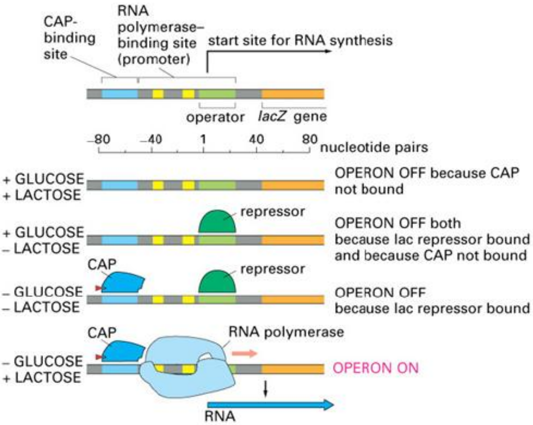


Figure 8-9 Essential Cell Biology, 2/e. (© 2004 Garland Science)

Expression Data

- Identifying each precise regulation mechanism by experiments is very costly and time consuming: too many genes, way too many possible interactions!

- Hope to *infer* regulatory mechanisms from whole genome-scale experiments: *microarrays*

6152 genes × 172 stress conditions

	172 stress conditions		
YAL001C	1.53	-0.06	⋯
YAL002W	-0.01	-0.30	⋯
YAL004W	0.24	0.76	⋯
⋮	⋮	⋮	

Yeast Dataset from: *Grasch, Spellman, Mol. Biol. Cell (2000)*

- Log-ratios of expression data: overexpression, underexpression.

Inference of the gene-regulatory network

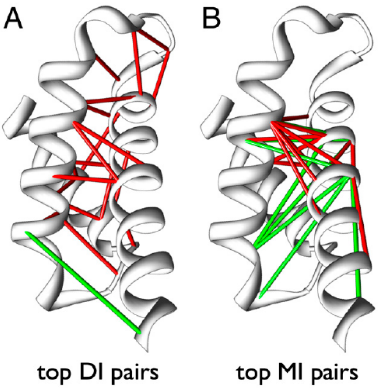
Two main goals:

- Inference of topology: Who regulates who?
- Inference of behaviour: predict the expresssion of a gene given the expression of other genes
- These are method of *inverse* inference: infer the model from the data

Inference of topology

- One way to do this is using *coexpression networks.*
  - Compute the Pearson correlation coefficient C_{ij} for every pair i, j of genes
  - Potential regulators of a gene are most correlated inputs
  - Build the network of links for which $|C_{ij}|$ is above a certain treshold.
- But we can do better!

Inference of protein structure from protein families sequences



F. Morcos, A.Pagnani et al, 2011

Conditional probability

Conditional probability: restriction of a probability distribution to a subspace B:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$$

"Probability of A given B"

Example

What is the probability of the output of a die to be ≥ 2 given that the result is odd?

$$P(d \geq 2 | d \text{ odd}) = P(d \geq 2, d \text{ odd}) / P(d \text{ odd}) = \frac{2}{6} / \frac{1}{2} = \frac{2}{3}$$

Bayes

$$P(A|B) = \frac{P(A,B)}{P(A)} = P(B|A)\,\frac{P(A)}{P(B)}$$

Example

You are tested for an illness that is very rare (about 1:100000) with a fairly precise test (99% accuracy in both cases). You come up positive, yuck! Probability of illness? (a) 99% (b) 90% (c) 10% (d) 1% (e) 0.1%

$$\begin{aligned}
P(I|+) &= P(+|I)\,P(I)\,P(+)^{-1} \approx 0.99 \cdot 10^{-5}\,(0.01)^{-1} \approx 10^{-3}! \\
P(+) &= P(+,I) + P(+,\text{not } I) \\
&= P(+|I)\,P(I) + P(+|\text{not } I)\,P(\text{not } I) \\
&= 0.99 \times 10^{-5} + 0.01 \times \left(1 - 10^{-5}\right) \approx 0.01
\end{aligned}$$

Bayes' rule in inference

- D =data, S =stochastic "machine", $P(D|S)$ =stochastic rule, $P(S)$ prior information about S

A double stochastic process:

1. S is extracted from $P(S)$
2. D is extracted from $P(D|S)$

We observe only D. What can we guess about S?

$$\overbrace{P(S|D)}^{\text{posterior}} = \frac{P(D|S)\,P(S)}{P(D)} \propto \overbrace{P(D|S)}^{\text{likelihood}}\,\overbrace{P(S)}^{\text{prior}}$$

Just the maths of common sense!

Bayes' rule iterated

Suppose we have the following multiple stochastic process:

1. S is extracted from $P(S)$
2. $D^1, \ldots, D^M$ are extracted i.i.d from $P(D|S)$

$$P\left(S|D^1, \ldots, D^M\right) = \frac{P\left(D^1, \ldots, D^M|S\right)}{P(D)} P(S) \propto P(S) \prod_{\mu=1}^{M} P\left(D^\mu|S\right)$$

Sometimes it is written in *update* form:

$$\begin{aligned} P\left(S|D^1, \ldots, D^M\right) &\propto P\left(D^M|S\right) \left(P(S) \prod_{\mu=1}^{M-1} P\left(D^\mu|S\right)\right) \\ &= P\left(D^M|S\right) P\left(S|D^1, \ldots, D^{M-1}\right) \end{aligned}$$

MAP vs. Max likelihood

$$\overbrace{P(S|D)}^{\text{posterior}} \propto \overbrace{P(S)}^{\text{prior}} \overbrace{P(D|S)}^{\text{likelihood}}$$

- *Maximum A Posteriori* (**MAP**):

$$(\arg)\max_S P(S|D)$$

- *Maximum Likelihood* (**ML**):

$$(\arg)\max_S P(D|S)$$

- ML=MAP for *uniform* prior, when it makes sense
- Two “Schools of thought”

Example: biased coins

I have two coins with head probabilities $p_1 = 0.5$ and $p_2 = 0.2$.

1. I choose one at random with $P(1) = 0.6, P(2) = 0.4$.
2. I flip the coin and the output is tail.

Can we say something about the coin?

$$\begin{aligned}
P(1|\text{tail}) &\propto P(\text{tail}|1)\,P(1) = 0.5 \times 0.6 = 0.30 \\
P(2|\text{tail}) &\propto P(\text{tail}|2)\,P(2) = 0.8 \times 0.4 = 0.32 \\
P(1|\text{tail}) &= 0.30/(0.30 + 0.32) = 0.484 \\
P(2|\text{tail}) &= 0.32/(0.30 + 0.32) = 0.516
\end{aligned}$$

Not much!

Binomial distribution

Consider the outcome of n p-biased coins. The probability of k heads is

$$P(k|p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Uniform prior

If $P(p) = \mathit{uniform}$, likelihood=posterior!

$$P(k_1, \ldots, k_M|p) \propto p^{\sum_{\mu=1}^{M} k_\mu} (1-p)^{\sum_{\mu=1}^{M} n-k_\mu} = \left(p^{\tilde{k}} (1-p)^{n-\tilde{k}}\right)^M$$

Binomial distribution

$$P(k_1, \ldots, k_M | p) \propto \left(p^{\tilde{k}} (1-p)^{n-\tilde{k}} \right)^M$$

with $\tilde{k} = \frac{1}{M} \sum_{\mu=1}^{M} k_\mu$ heads, the ML is attained at the max of

$$\mathscr{L} = \tilde{k} \log p + \left(n - \tilde{k} \right) \log (1-p)$$

Let us find critical points:

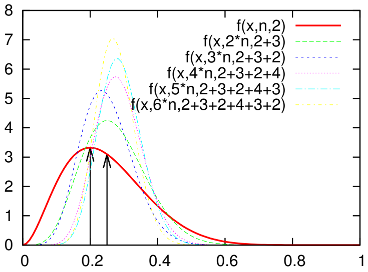
$$0 = \frac{\partial \mathscr{L}}{\partial p} = \frac{\tilde{k}}{p} - \frac{\left(n - \tilde{k} \right)}{1-p}$$

So $\frac{n-\tilde{k}}{\tilde{k}} = \frac{1-p}{p}$, i.e. $p = \frac{\tilde{k}}{n}$.

Note!

$$\langle p \rangle = \frac{\int_0^1 p p^{M\tilde{k}} (1-p)^{M\left(n-\tilde{k}\right)} dp}{\int_0^1 p^{M\tilde{k}} (1-p)^{M\left(n-\tilde{k}\right)} dp} = \frac{M\tilde{k}+1}{Mn+2}$$

Binomial



gnuplot code

```
f(p,n,k)=p**k*(1-p)**(n-k)/(k!*(n-k)!/(n+1)!)
pml(n,k)=k*1./n
pav(n,k)=(k+1)*1.0/(n+2)
n=10;k=2;
set arrow from pml(n,k),0 to pml(n,k), f(pml(n,k), n, k)
set arrow from pav(n,k),0 to pav(n,k), f(pav(n,k), n, k)

plot [0:1] f(x,n,2) lw 3, f(x,2*n,2+3), f(x,3*n,2+3+2),

f(x,4*n,2+3+2+4), f(x,5*n,2+3+2+4+3), f(x,6*n,2+3+2+4+3+2)
```

Example: Normal

$$P\left(x|\left(m,\sigma\right)\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-m)^2}$$

Given $x^1,\ldots,x^M$, we have

$$P\left(x^1,\ldots,x^M|\left(m,\sigma\right)\right) \propto e^{-\frac{1}{2\sigma^2}\sum_{\mu=1}^M (x^\mu-m)^2 - M\log\sigma}$$

If we try to maximize the log-likelihood
$\mathscr{L}\left(m,\sigma\right) = -\frac{1}{2\sigma^2}\frac{1}{M}\sum_{\mu=1}^M \left(x^\mu - m\right)^2 - \log\sigma$

$$0 = \frac{\partial\mathscr{L}}{\partial m} = \frac{1}{\sigma^2}\frac{1}{M}\sum_{\mu=1}^M \left(x^\mu - m\right) \qquad 0 = \frac{\partial\mathscr{L}}{\partial\sigma} = \sigma^{-3}\left(\frac{1}{M}\sum_{\mu=1}^M \left(x^\mu - m\right)^2 - \sigma^2\right)$$

I.e. $m = \frac{1}{M}\sum_\mu^M x^\mu, \sigma = \sqrt{\frac{1}{M}\sum_{\mu=1}^M \left(x^\mu - m\right)^2}$

- What is the likelihood of (m,σ) when $M = 1$?

ML and KL divergence

Remember the KL divergence

$$KL(P||Q) = \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{Q(\mathbf{x})}$$

Assume you have a set of sample data $\mathbf{x}^\mu$ for $\mu = 1, \ldots, M$. Then consider the distribution $P(\mathbf{x}) = \frac{1}{M}\sum_{\mu=1}^{M} \delta(\mathbf{x}, \mathbf{x}^\mu)$, and a distribution Q_θ parametrized by θ

$$\begin{aligned} KL(P||Q_\theta) &= \sum_{\mathbf{x}} \sum_{\mu=1}^{M} \delta(\mathbf{x}, \mathbf{x}^\mu) \log \frac{\sum_{\mu'=1}^{M} \delta\left(\mathbf{x}, \mathbf{x}^{\mu'}\right)}{Q_\theta(\mathbf{x})} \\ &= -\log \prod_{\mu=1}^{M} Q_\theta(\mathbf{x}^\mu) \end{aligned}$$

That is, ML is the same as minimizing the KL divergence with $\frac{1}{M}\sum_{\mu=1}^{M} \delta(\mathbf{x}, \mathbf{x}^\mu)$!

Ising model

Suppose given $\sigma^1, \ldots, \sigma^M$ samples, and assume they were generated **independently** by an Ising model

$$P_{\boldsymbol{J},\boldsymbol{h}}(\sigma) = Z_{\boldsymbol{J},\boldsymbol{h}}^{-1} e^{\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i\sigma_i}$$

$$\begin{aligned} P\left(\sigma^1, \ldots, \sigma^M | \boldsymbol{J}, \boldsymbol{h}\right) &= \prod_{\mu=1}^{M} e^{\sum_{i<j} J_{ij}\sigma_i^\mu \sigma_j^\mu + \sum_i h_i \sigma_i^\mu - \log Z_{\boldsymbol{J},\boldsymbol{h}}} \\ &= e^{M\left(\sum_{i<j} J_{ij}\tilde{c}_{ij} + \sum_i h_i \tilde{m}_i - \log Z_{\boldsymbol{J},\boldsymbol{h}}\right)} \end{aligned}$$

- Depends *only* on the experimental first ($\tilde{m}_i$) and second moments ($\tilde{c}_{ij}$) of the data!
- The log-likelihood

$$\mathscr{L}(\boldsymbol{J}, \boldsymbol{h}) = M\left(\sum_{i<j} J_{ij}\tilde{c}_{ij} + \sum_i h_i \tilde{m}_i - \log Z_{\boldsymbol{J},\boldsymbol{h}}\right)$$

How can we find $\boldsymbol{J}, \boldsymbol{h}$ of maximum likelihood?

Ising Likelihood

$$\mathscr{L}(\boldsymbol{J}, \boldsymbol{h}) = M\left(\sum_{i<j} J_{ij}\tilde{c}_{ij} + \sum_i h_i \tilde{m}_i - \log Z_{\boldsymbol{J},\boldsymbol{h}}\right)$$

Lets try to find critical points:

$$0 = \frac{\partial \mathscr{L}}{\partial J_{ij}} = M\left(\tilde{c}_{ij} - \frac{\partial \log Z_{\boldsymbol{J},\boldsymbol{h}}}{\partial J_{ij}}\right) = M\left(\tilde{c}_{ij} - \langle \sigma_i \sigma_j \rangle\right) \qquad 0 = \frac{\partial \mathscr{L}}{\partial h_i} = M\left(\tilde{m}_i - \langle \sigma_i \rangle\right)$$

- Better: $-\log Z_{\boldsymbol{J},\boldsymbol{h}}$ is a concave ($\cap$) function on **J**,**h** (and so is $\mathscr{L}$), so we can use gradient ascent!
- Unfortunately, estimating $\langle \sigma_i \sigma_j \rangle$ and $\langle \sigma_i \rangle$ is computationally **hard**! (NP-Complete). Possibilities:
  1. Exact enumeration (up to $N \approx 30$)
  2. Monte-Carlo methods (slow!)
  3. Mean-field type approximations (e.g. Belief Propagation)

Boltzmann learning

Boltzmann learning algorithm

1. (init) Set $\mathbf{J} = 0$, $\mathbf{h} = 0$
2. (direct inference) **somehow** estimate $\{\langle \sigma_i \sigma_j \rangle\}_{i<j}$ and $\{\langle \sigma_i \rangle\}_i$ from $P_{\mathbf{J},\mathbf{h}}$
3. (delta) Compute $\Delta J_{ij} = \tilde{c}_{ij} - \langle \sigma_i \sigma_j \rangle$, $\Delta h_i = \tilde{m}_i - \langle \sigma_i \rangle$
4. (end?) if $|\Delta J_{ij}| < \varepsilon$ for all $i < j$, $|\Delta h_i| < \varepsilon$ for all i, exit
5. (update) $\mathbf{J} \leftarrow \mathbf{J} + \eta \Delta \mathbf{J}$, $\mathbf{h} \leftarrow \mathbf{h} + \eta \Delta \mathbf{h}$
6. Go to 2

But we need an (approximate) inference method for 2!

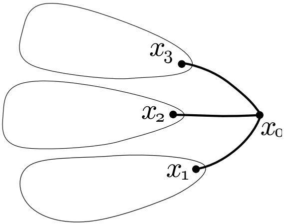
Example: 3-Coloring (Potts)

- Given a (finite) undirected graph $G = (V, E)$
- A proper 3−coloring is $\sigma_i \in \{\textcolor{red}{\bullet}, \textcolor{green}{\bullet}, \textcolor{blue}{\bullet}\}$ for $i \in V$ such that $\sigma_i \neq \sigma_j$ if $(i,j) \in E$

$$P(\sigma) = \frac{1}{Z} \prod_{(ij) \in E} \left(1 - \delta(\sigma_i, \sigma_j)\right)$$

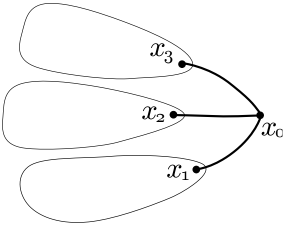
- Hard computational problems (NP-Complete):
  - Finding a proper coloring
  - Estimating $P(\sigma_i, \sigma_j)$
  - Counting proper colorings
  - Deciding if there is at least one proper coloring!

Belief Propagation



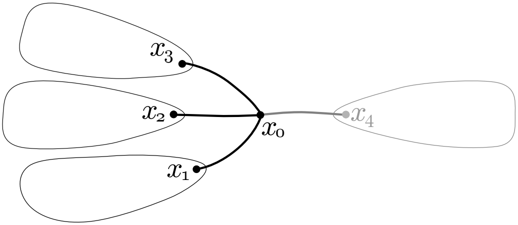
$$
\begin{aligned}
N_0\left(\textcolor{red}{\bullet}\right) &= N^{(0)}\left(\textcolor{blue}{\bullet}\textcolor{blue}{\bullet}\textcolor{blue}{\bullet}\right) + N^{(0)}\left(\textcolor{blue}{\bullet}\textcolor{blue}{\bullet}\textcolor{green}{\bullet}\right) + N^{(0)}\left(\textcolor{blue}{\bullet}\textcolor{green}{\bullet}\textcolor{blue}{\bullet}\right) + N^{(0)}\left(\textcolor{green}{\bullet}\textcolor{blue}{\bullet}\textcolor{blue}{\bullet}\right) + \cdots \\
&= N_1^{(0)}\left(\textcolor{blue}{\bullet}\right) N_2^{(0)}\left(\textcolor{blue}{\bullet}\right) N_3^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_1^{(0)}\left(\textcolor{blue}{\bullet}\right) N_2^{(0)}\left(\textcolor{blue}{\bullet}\right) N_3^{(0)}\left(\textcolor{green}{\bullet}\right) + \cdots \\
&= \left(N_1^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_1^{(0)}\left(\textcolor{green}{\bullet}\right)\right) \left(N_2^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_2^{(0)}\left(\textcolor{green}{\bullet}\right)\right) \left(N_3^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_3^{(0)}\left(\textcolor{green}{\bullet}\right)\right) \\
N_0\left(\textcolor{green}{\bullet}\right) &= \left(N_1^{(0)}\left(\textcolor{blue}{\bullet}\right) + P_1^{(0)}\left(\textcolor{red}{\bullet}\right)\right) \left(N_2^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_2^{(0)}\left(\textcolor{red}{\bullet}\right)\right) \left(N_3^{(0)}\left(\textcolor{blue}{\bullet}\right) + N_3^{(0)}\left(\textcolor{red}{\bullet}\right)\right) \\
N_0\left(\textcolor{blue}{\bullet}\right) &= \left(N_1^{(0)}\left(\textcolor{red}{\bullet}\right) + N_1^{(0)}\left(\textcolor{green}{\bullet}\right)\right) \left(N_2^{(0)}\left(\textcolor{red}{\bullet}\right) + N_2^{(0)}\left(\textcolor{green}{\bullet}\right)\right) \left(N_3^{(0)}\left(\textcolor{red}{\bullet}\right) + N_3^{(0)}\left(\textcolor{green}{\bullet}\right)\right)
\end{aligned}
$$

Belief Propagation



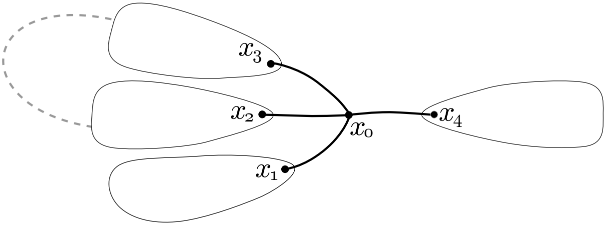
$$
\begin{aligned}
P_0(\textcolor{red}{\bullet}) &\propto P^{(0)}(\textcolor{blue}{\bullet\bullet\bullet}) + P^{(0)}(\textcolor{blue}{\bullet\bullet}\textcolor{green}{\bullet}) + P^{(0)}(\textcolor{blue}{\bullet}\textcolor{green}{\bullet}\textcolor{blue}{\bullet}) + P^{(0)}(\textcolor{green}{\bullet}\textcolor{blue}{\bullet\bullet}) + \cdots \\
&= P_1^{(0)}(\textcolor{blue}{\bullet})\,P_2^{(0)}(\textcolor{blue}{\bullet})\,P_3^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{blue}{\bullet})\,P_2^{(0)}(\textcolor{blue}{\bullet})\,P_3^{(0)}(\textcolor{green}{\bullet}) + \cdots \\
&= \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right) \\
P_0(\textcolor{green}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{red}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{red}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{red}{\bullet})\right) \\
P_0(\textcolor{blue}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{red}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{red}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{red}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right)
\end{aligned}
$$

Belief Propagation



$$
\begin{aligned}
P_0^{(4)}(\textcolor{red}{\bullet}) &\propto P^{(0)}(\textcolor{blue}{\bullet\bullet\bullet}) + P^{(0)}(\textcolor{blue}{\bullet\bullet}\textcolor{green}{\bullet}) + P^{(0)}(\textcolor{blue}{\bullet}\textcolor{green}{\bullet}\textcolor{blue}{\bullet}) + P^{(0)}(\textcolor{green}{\bullet}\textcolor{blue}{\bullet\bullet}) + \cdots \\
&= P_1^{(0)}(\textcolor{blue}{\bullet})\,P_2^{(0)}(\textcolor{blue}{\bullet})\,P_3^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{blue}{\bullet})\,P_2^{(0)}(\textcolor{blue}{\bullet})\,P_3^{(0)}(\textcolor{green}{\bullet}) + \cdots \\
&= \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right) \\
P_0^{(4)}(\textcolor{green}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{red}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{red}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{red}{\bullet})\right) \\
P_0^{(4)}(\textcolor{blue}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{red}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{red}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{red}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right)
\end{aligned}
$$

Belief Propagation



$$
\begin{aligned}
P_0^{(4)}(\textcolor{red}{\bullet}) &\propto P^{(0)}(\textcolor{blue}{\bullet\bullet\bullet}) + P^{(0)}(\textcolor{blue}{\bullet\bullet}\textcolor{green}{\bullet}) + P^{(0)}(\textcolor{blue}{\bullet}\textcolor{green}{\bullet}\textcolor{blue}{\bullet}) + P^{(0)}(\textcolor{green}{\bullet}\textcolor{blue}{\bullet\bullet}) + \cdots \\
&\textcolor{magenta}{\simeq} P_1^{(0)}(\textcolor{blue}{\bullet})\, P_2^{(0)}(\textcolor{blue}{\bullet})\, P_3^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{blue}{\bullet})\, P_2^{(0)}(\textcolor{blue}{\bullet})\, P_3^{(0)}(\textcolor{green}{\bullet}) + \cdots \\
&= \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right) \\
P_0^{(4)}(\textcolor{green}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{blue}{\bullet}) + P_1^{(0)}(\textcolor{red}{\bullet})\right)\left(P_2^{(0)}(\textcolor{blue}{\bullet}) + P_2^{(0)}(\textcolor{red}{\bullet})\right)\left(P_3^{(0)}(\textcolor{blue}{\bullet}) + P_3^{(0)}(\textcolor{red}{\bullet})\right) \\
P_0^{(4)}(\textcolor{blue}{\bullet}) &\propto \left(P_1^{(0)}(\textcolor{red}{\bullet}) + P_1^{(0)}(\textcolor{green}{\bullet})\right)\left(P_2^{(0)}(\textcolor{red}{\bullet}) + P_2^{(0)}(\textcolor{green}{\bullet})\right)\left(P_3^{(0)}(\textcolor{red}{\bullet}) + P_3^{(0)}(\textcolor{green}{\bullet})\right)
\end{aligned}
$$

BP Equations (coloring)

$$q_{ij}(\sigma_i) \propto \psi_i(\sigma_i) \prod_{k \in \partial i \setminus j} \sum_{\sigma_k \neq \sigma_i} q_{ki}(\sigma_k)$$

This system is a fixed point $\mathbf{F}(\mathbf{q}) = \mathbf{q}$ equation for $\mathbf{q} = \{q_{ij}, q_{ji}\}_{(ij) \in E} \in [0,1]^{2|E|}$ and is solved normally by iteration:

$$\mathbf{q}_\infty = \lim_{k \to \infty} \mathbf{F}^{(k)}(\mathbf{q}_0)$$

On a fixed point, we can compute

$$p_i(\sigma_i) \propto \psi_i(\sigma_i) \prod_{k \in \partial i} \sum_{\sigma_k \neq \sigma_i} q_{ki}(\sigma_k)$$

$$p_{ij}(\sigma_i, \sigma_j) \propto q_{ij}(\sigma_i)\, q_{ji}(\sigma_j)\, (1 - \delta(\sigma_i, \sigma_j))$$

Belief Propagation (pairwise models)

Given a distribution:

$$P(\sigma) = \frac{1}{Z} \prod_{(ij)\in E} \psi_{ij}(\sigma_i, \sigma_j) \prod_i \psi_i(\sigma_i) = \frac{1}{Z} e^{-\left(\sum_{(ij)\in E} -\log \psi_{ij}(\sigma_i,\sigma_j) + \sum_i -\log \psi_i(\sigma_i)\right)}$$

BP Equations, pairwise potentials

$$\begin{aligned}
q_{ij}(\sigma_i) &\propto \psi_i(\sigma_i) \prod_{k\in\partial i\setminus j} \sum_{\sigma_k} q_{ki}(\sigma_k)\,\psi_{ki}(\sigma_k,\sigma_i) \ \text{(message)} \\
p_i(\sigma_i) &\propto \psi_i(\sigma_i) \prod_{k\in\partial i} \sum_{\sigma_k} q_{ki}(\sigma_k)\,\psi_{ki}(\sigma_k,\sigma_i) \ \text{(marginal)} \\
p_{ij}(\sigma_i,\sigma_j) &\propto \psi_{ij}(\sigma_i,\sigma_j)\, q_{ij}(\sigma_i)\, q_{ji}(\sigma_j) \ \text{(marginal)}
\end{aligned}$$

BP for crosswords

- English dictionary D (set of english words)
- ***Indices***: a set X of letters coordinates, one for each non-black square, a set H of horizontal words *indices*, one for each horizontal blank sequence, a set V of vertical word *indices*, one for each vertical blank sequence,
- ***Variables***: $h_s \in D$ for each $s \in H$, $v_t \in D$ for each $t \in V$, $x_{ij} \in \{a, \ldots, z\}$ for each $ij \in X$
- For each non-black square ij,
  - $s(ij) \in H$=crossing horizontal word, $p(ij)$= position of ij within,
  - $t(ij) \in V$=crossing vertical word, $q(ij)$ position of ij within
- ***Constraints***: For each non black position ij: the following two conditions have to be ensured: $\left(h_{s(ij)}\right)_{p(ij)} = x_{ij}$ and $\left(v_{t(ij)}\right)_{q(ij)} = x_{ij}$
- In summary: $|H| + |V| + |X|$ variable nodes, $2|X|$ constraints

$$P(\mathbf{h}, \mathbf{v}, \mathbf{x}) = \frac{1}{Z} \prod_{ij \in X} \delta\left(\left(h_{s(ij)}\right)_{p(ij)}; x_{ij}\right) \delta\left(\left(v_{t(ij)}\right)_{q(ij)}; x_{ij}\right)$$

Exact inference on trees

Let $T = (V, E)$ be a tree, and assume P a *T-factorized* distribution, i.e. $P(\sigma) = \frac{1}{Z}\prod_{(ij)\in E}\psi_{ij}(\sigma_i, \sigma_j)$. Then:

$$P(\sigma) = \prod_{(ij)\in E}\frac{P(\sigma_i, \sigma_j)}{P(\sigma_i)\,P(\sigma_j)}\prod_i P(\sigma_i)$$

For a general graph G, it is only an approximation!

- It is called the *Bethe* approximation.

Entropy of a tree distribution

If P is T-factorized, then

$$\begin{aligned}
-S(P) &= \sum_{\sigma} P(\sigma) \ln P(\sigma) \\
&= \sum_{(ij)\in E} KL(P(\sigma_i,\sigma_j) || P(\sigma_i) P(\sigma_j)) - \sum_i S(P(\sigma_i)) \\
&= \sum_{(ij)\in E} M_{ij} - \sum_i H_i
\end{aligned}$$

Average Energy and Free Energy

For every $G = (V, E)$-factorized Ising model,

$$\begin{aligned} -\langle H \rangle &= \sum_{(ij)\in E} J_{ij} \langle \sigma_i \sigma_j \rangle + \sum_i h_i \langle \sigma_i \rangle \\ -\log Z_{\boldsymbol{J},\boldsymbol{h}} &= \langle H \rangle - S \\ &= \langle H \rangle + \sum_\sigma P(\sigma) \log P(\sigma) \end{aligned}$$

If P is T-factorized, then

$$-\log Z_{\boldsymbol{J},\boldsymbol{h}} = \langle H \rangle + \sum_{(ij)\in E} M_{ij} - \sum_i H_i$$

These expressions for S and $\log Z$ are exact for trees, just approximations for general graphs!

Mutual Information

Mutual Information is a measure of **correlation**:

$$MI(x,y) = \sum_x P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

- In terms of the KL divergence:
$$MI(x,y) = KL(P(x,y)||P(x)P(y))$$
- It can be also thought as "information gain": how much information about x is gained (in average) by knowing the value of y:
$$\begin{aligned} MI(x,y) &= S(P(x)) - \sum_y P(y) S(P(x|y)) \\ &= S(P(y)) - \sum_x P(x) S(P(y|x)) \end{aligned}$$
- $MI(x,y) \leq S(P(x))$
- If $x = y$ (i.e. $P(x,y) = \delta(x,y)P(x)$), $MI(x,y) = S(P(x))$.
- If $P(x,y) = P(x)P(y)$, $MI(x,y) = 0$

Inference of trees

- Suppose that we are told that some tree-factorized Ising model produced a set of samples:

$$\sigma^1, \ldots, \sigma^M \sim P(\sigma) = \frac{1}{Z_{\mathbf{J},\mathbf{h}}} e^{\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i \sigma_i}$$

- How do we find the tree $T = (V, E)$ and the T-factorized $\mathbf{J}, \mathbf{h}$ (i.e. such that $J_{ij} \neq 0 \implies (ij) \in E$) of ML?

Likelihood of a tree

Given samples $\sigma^1, \ldots, \sigma^M$, consider $\boldsymbol{J}^*, \boldsymbol{h}^*$ the $T = (V, E)$-factorized ML couplings (T **tree**), then

$$
\begin{aligned}
\mathscr{L}(\boldsymbol{J}^*, \boldsymbol{h}^*) &= \sum_{(ij)\in E} J^*_{ij} \tilde{c}_{ij} + \sum_i h^*_i \tilde{m}_i - \log Z_{\boldsymbol{J}^*,\boldsymbol{h}^*} \\
&= -\langle \tilde{H}^* \rangle - \log Z_{\boldsymbol{J}^*,\boldsymbol{h}^*} \\
&= -\langle \tilde{H}^* \rangle + \langle H^* \rangle - S^* \\
&= -\langle \tilde{H}^* \rangle + \langle H^* \rangle + \sum_{(ij)\in E} M^*_{ij} + \sum_i S^*_i
\end{aligned}
$$

Chow-Liu (1968)

$$\mathscr{L}(\boldsymbol{J}^*, \boldsymbol{h}^*) = -\langle \tilde{H}^* \rangle + \langle H^* \rangle + \sum_{(ij)\in E} M^*_{ij} + \sum_i S^*_i$$

Two key observations:

1. We have seen that $\mathbf{P}_{\mathbf{J}^*,\mathbf{h}^*}$ must reproduce the first ($\tilde{m}_i$) and second ($\tilde{c}_{ij}$) moments of the data over T (so $\langle \tilde{H}^* \rangle = \langle H^* \rangle$). Then it must reproduce also $\tilde{P}(\sigma_i, \sigma_j) = \frac{1}{4}(\tilde{c}_{ij}\sigma_i\sigma_j + \tilde{m}_i\sigma_i + \tilde{m}_j\sigma_j + 1)$. In particular, $M^*_{ij} = \tilde{M}_{ij}$ and $S^*_i = \tilde{S}_i$.
2. The term $\tilde{S}_i$ does not depend on T

$$\mathscr{L}(\boldsymbol{J}^*, \boldsymbol{h}^*) = \sum_{(ij)\in E} \tilde{M}_{ij} + \text{const.}$$

And we want to maximize with respect to T (topology)

Maximum Spanning Tree (Kruskal 1956)

Given a connected graph $G = (V, E)$ and weights $M : E \to \mathbb{R}_+$, finding the maximum spanning tree can be done as follows:

Kruskal's algorithm

1. Order edges so as to have $M_{e_1} \geq M_{e_2} \geq \cdots M_{e_{|E|}}$
2. **Set** $E' \leftarrow \emptyset$
3. **For** $s = 1, \ldots, |E|$:
   - **If** $(V, E' \cup \{e_s\})$ has no loop:
     $E' \leftarrow E' \cup \{e_s\}$

At the end, (V, E') is a maximum spanning tree, i.e. a tree that maximizes $\sum_{e \in E'} M_e$

Putting all the bits toghether

1. Compute M_{ij} for $i < j$
1. Use Kruskal to compute T the MST for the M_{ij}
1. $\tilde{P}(\sigma_i, \sigma_j) = e^{J_{ij}\sigma_i\sigma_j + a_{ij}\sigma_i + b_{ij}\sigma_j + f_{ij}}$ $\quad \tilde{P}(\sigma_i) = e^{h'_i\sigma_i + f_i}$

$$\begin{aligned} P(\sigma) &= \prod_{(ij)\in T} \tilde{P}(\sigma_i, \sigma_j) \prod_i \tilde{P}(\sigma_i)^{1-d_i} \\ &\propto e^{\sum_{(ij)\in T} J_{ij}\sigma_i\sigma_j + a_{ij}\sigma_i + b_{ij}\sigma_j + \sum_i h'_i\sigma_i(1-d_i)} \\ &\propto e^{\sum_{(ij)\in T} J_{ij}\sigma_i\sigma_j + \sum_i \sigma_i h_i} \end{aligned}$$

where h_i is computed by collecting all coefficients of σ_i.

Maximum Spanning Tree

Proof by induction on t: $E' \subseteq E''$ for some MST E'' in every step t of Kruskal (assume that for some step $t \geq 0$, E' *is* included in an MST E'' and prove that $E' \cup \{e_t\}$ is also included in some MST)

1. If $E' \cup \{e_t\}$ is also included in E'', **done**. Otherwise:
2. $E'' \cup \{e_t\}$ has a loop p (E'' is a tree) .
3. Take any edge f in $p \setminus (E' \cup \{e_t\})$ (such an edge must exist, otherwise $p \subseteq E' \cup \{e_t\}$).
4. We have $M_{e_t} \geq M_f$ (otherwise f would have been added before e_t).
5. $E''' = E'' \setminus \{f\} \cup \{e_t\}$ is a tree, $\sum_{(ij) \in E'''} M_{ij} \geq \sum_{(ij) \in E''} M_{ij}$, so E''' MST, and $E' \cup \{e_t\} \subseteq E'''$ **done**

Example

$N = 5,\ M = 6,$ Data:

σ_1	σ_2	σ_3	σ_4	σ_5
1	1	1	−1	1
1	−1	1	−1	1
1	1	1	−1	1
−1	−1	1	1	1
−1	1	−1	1	−1
−1	1	1	−1	−1

Marginals:
$P_1 = \frac{1}{6}\begin{pmatrix} 3 \\ 3 \end{pmatrix}, P_2 = \frac{1}{6}\begin{pmatrix} 2 \\ 4 \end{pmatrix}, P_3 = \frac{1}{6}\begin{pmatrix} 1 \\ 5 \end{pmatrix}, P_4 = \frac{1}{6}\begin{pmatrix} 4 \\ 2 \end{pmatrix}, P_5 = \frac{1}{6}\begin{pmatrix} 3 \\ 3 \end{pmatrix}$ and
$P_{12} = \frac{1}{6}\begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}, P_{13} = \frac{1}{6}\begin{pmatrix} 1 & 2 \\ 0 & 3 \end{pmatrix}, P_{14} = \frac{1}{6}\begin{pmatrix} 1 & 2 \\ 3 & 0 \end{pmatrix}, P_{15} =$
$\frac{1}{6}\begin{pmatrix} 2 & 1 \\ 0 & 3 \end{pmatrix}, P_{23} = \frac{1}{6}\begin{pmatrix} 0 & 2 \\ 1 & 3 \end{pmatrix}, P_{24} = \frac{1}{6}\begin{pmatrix} 1 & 1 \\ 3 & 1 \end{pmatrix}, P_{25} = \frac{1}{6}\begin{pmatrix} 0 & 2 \\ 2 & 2 \end{pmatrix}, P_{34} =$
$\frac{1}{6}\begin{pmatrix} 0 & 1 \\ 4 & 1 \end{pmatrix}, P_{35} = \frac{1}{6}\begin{pmatrix} 1 & 0 \\ 1 & 4 \end{pmatrix}, P_{45} = \frac{1}{6}\begin{pmatrix} 1 & 3 \\ 1 & 1 \end{pmatrix}$
Mutual information: $M_{15} = 0.459, M_{45} = 0.459, M_{35} = 0.317, M_{34} = 0.317, M_{25} =$
$0.252, M_{13} = 0.191, M_{23} = 0.109, M_{45} = 0.044, M_{24} = 0.044, M_{12} = 0$
Kruskal edges: $(15), (45), (35), (34), (25)$

Independent pairs

Assume the Bethe expression for trees to be valid for the complete graph:

$$P(\sigma|\boldsymbol{J},\boldsymbol{h}) = \prod_{i<j} \frac{P(\sigma_i,\sigma_j)}{P(\sigma_i)P(\sigma_i)} \prod_i P(\sigma_i)$$

we parametrize

$$P(\sigma_i,\sigma_j) = e^{J'_{ij}\sigma_i\sigma_j + a_{ij}\sigma_i + b_{ij}\sigma_j + f_{ij}} \quad P(\sigma_i) = e^{h''_i\sigma_i + f_i}$$

But then,

$$P(\sigma|\boldsymbol{J},\boldsymbol{h}) = e^{\sum_{i<j} J'_{ij}\sigma_i\sigma_j + \sum_i (1-d_i)\left(\sum_{j>i}\left(a_{ij}+b_{ji}\right)+h''_i\right)\sigma_i} \implies J'_{ij} = J_{ij}$$

but we know that on the point of ML, $P(\sigma_i\sigma_j) = \tilde{P}(\sigma_i,\sigma_j)$ so we can get J_{ij} directly from the data as in the two-spin system:

$$J_{ij} = \log \frac{\tilde{p}_{++}\tilde{p}_{--}}{\tilde{p}_{+-}\tilde{p}_{-+}}$$

This exactly the same as if we consider each link separately (a single link **is** a tree!). This is called the **independent pairs** approximation.

BP on the Ising model

With the change of variables

$$h_{ij} = \frac{1}{2} \log \frac{q_{ij}(+1)}{q_{ij}(-1)}$$

The BP equations for the Ising model

$$q_{ij}(\sigma_i) \propto e^{h_i \sigma_i} \prod_{k \in \partial i \setminus j} \sum_{\sigma_k} q_{ki}(\sigma_k) \, e^{J_{ki} \sigma_k \sigma_i}$$

become:

$$\begin{aligned} h_{ij} &= h_i + \sum_{l \in \partial i \setminus j} \tanh^{-1}(\tanh J_{li} \tanh h_{li}) \\ m_i &= \tanh\left(h_i + \sum_{l \in \partial i} \tanh^{-1}(\tanh J_{li} \tanh h_{li}) \right) \end{aligned}$$

Susceptibility Propagation

If we define

$$g_{ijk} = \frac{\partial h_{ij}}{\partial h_k}$$

Taking derivatives of the BP equations we obtain Susceptibility Propagation Equations (Mézard & Mora 2007):

$$g_{ijk} \quad = \quad \delta_{ik} + \sum_{l \in \partial i \setminus j} g_{lik} \tanh J_{li} \frac{1 - \tanh^2 h_{li}}{1 - \tanh^2 J_{li} \tanh^2 h_{li}}$$

This gives a much better approximation for the susceptibility $\chi_{ij} = c_{ij} - m_i m_j = \frac{\partial m_i}{\partial h_j}$:

$$\chi_{ij} = \left(\frac{\tanh J_{ij} + \tanh h_{ij} \tanh h_{ji}}{1 + \tanh J_{ij} \tanh h_{ij} \tanh h_{ji}} - m_i m_j \right) g_{jij} + g_{ijj} \left(1 - m_i^2\right)$$

that can be employed for gradient ascent or on a coordinated $h_{ij}, g_{ijk}, J_{ij}, h_i$ updating scheme.

Example

We will deal with partial observation of extractions from a distribution over $X = \{1, \dots, n\}$.

- Suppose you see that over M samples, n_3 samples were the number 3. In the remaining $M - n_3$, you just don't know.
- You need to point out **one** plausible distribution for the data.
- Would your guess be e.g. $P(k) = \frac{n_3}{M}\delta(k,3) + \frac{M-n_3}{M}\delta(k,2)$? This one is compatible with the observations!
- Or would you rather guess $P(k) = \frac{n_3}{M}\delta(k,3) + \frac{M-n_3}{M}(1-\delta(k,3))$, i.e. completely flat in the unobserved part?

Another example

Same setup as before.

- Suppose you only observe that over M samples, n_{23} samples were either 2 or 3, and n_{34} samples were either 3 or 4.
- How do we find the *flattest* possible distribution given the observations?

General case: making predictions from partial observations

- How can we come up with reasonable models?
- Suppose we have a distribution $\mathbf{P} : \mathbf{X} \to [0, 1]$ and we are given an observable for a variable f:

$$\bar{f} = \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{x})$$

How does one compute

$$\bar{g} = \sum_{\mathbf{x}} g(\mathbf{x}) P(\mathbf{x})?$$

But!

(very) undertermined system ($|\mathbf{X}|$ unknows, 2 equations)!

Maximum Entropy

- Let us find the distribution $\mathbf{P}$ that satisfies $\bar{f} = \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{x})$ and $S(\mathbf{P}) = -\sum_{\mathbf{x}} P(\mathbf{x}) \ln P(\mathbf{x})$ is maximum (Jaynes 1957)
- This is the "less constrained / flattest distribution" compatible with the observation
- Using Lagrange multipliers...

$$\Gamma(\lambda, \mu, \mathbf{P}) = S(\mathbf{P}) + \mu \left(\bar{f} - \sum_{\mathbf{x}} f(\mathbf{x}) P(\mathbf{x}) \right) + \lambda \left(1 - \sum_{\mathbf{x}} P(\mathbf{x}) \right)$$

- And we need to find an unconstrained maximum for $\max_{\lambda,\mu,\mathbf{P}} \Gamma(\lambda, \mu, \mathbf{P})$. Taking derivative w.r.t $P(\mathbf{x})$

$$0 = \frac{\partial \Gamma}{\partial P(\mathbf{x})} = -\ln P(\mathbf{x}) - P(\mathbf{x}) / P(\mathbf{x}) - \mu f(\mathbf{x}) - \lambda$$

$$P(\mathbf{x}) = e^{-\mu f(\mathbf{x}) - (1+\lambda)} \propto e^{-\mu f(\mathbf{x})}$$

- (A Boltzmann / exponential distribution!)

Many observations

In general for many simultaneous observations $f_1, \ldots, f_m$,

$$\max_{\lambda, \mu_1, \ldots, \mu_m, \mathbf{P}} S(\mathbf{P}) + \sum_{a=1}^{m} \mu_a \left(\bar{f}_a - \sum_{\mathbf{x}} f_a(\mathbf{x}) P(\mathbf{x}) \right) + \lambda \left(1 - \sum_{\mathbf{x}} P(\mathbf{x}) \right)$$

$0 = \frac{\partial \Gamma}{\partial P(\mathbf{x})} = -\ln P(\mathbf{x}) - P(\mathbf{x}) / P(\mathbf{x}) - \sum_{a=1}^{m} \mu_a f_a(\mathbf{x}) - \lambda$, so

$$P(\mathbf{x}) \propto e^{-\sum_{a=1}^{m} \mu_a f_a(\mathbf{x})}$$

Fact!

$\sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{\alpha} = -S - \log \alpha$, i.e. $\min KL(P, \text{uniform}) = \max S$

Going back to our n_{23}, n_{34} example

- Our observables were $f_1(i) = \delta(i,2) + \delta(i,3)$ and $f_2(i) = \delta(i,3) + \delta(i,4)$, and $\langle f_1 \rangle = \frac{n_{23}}{M} = p_{23}, \langle f_2 \rangle = \frac{n_{34}}{M} = p_{34}$
- Maximum entropy says:

$$P(i) \propto e^{-\mu_1(\delta(i,2)+\delta(i,3))-\mu_2(\delta(i,3)+\delta(i,4))}$$

- i.e, defining $r = e^{-\mu_1}$ and $s = e^{-\mu_2}$ we get $P(i) = \frac{1}{Z} r^{\delta(i,2)+\delta(i,3)} s^{\delta(i,4)+\delta(i,3)}$

$$\begin{aligned}
Z &= \sum_{i=1}^{n} r^{\delta(i,2)+\delta(i,3)} s^{\delta(i,4)+\delta(i,3)} = r + rs + s + (n-3) \\
p_{23} &= \frac{1}{Z} \sum_{i=1}^{n} (\delta(i,2) + \delta(i,3)) r^{\delta(i,2)+\delta(i,3)} s^{\delta(i,4)+\delta(i,3)} = \frac{1}{Z}(r + rs) \\
p_{34} &= \frac{1}{Z} \sum_{i=1}^{n} (\delta(i,3) + \delta(i,4)) r^{\delta(i,2)+\delta(i,3)} s^{\delta(i,4)+\delta(i,3)} = \frac{1}{Z}(rs + s)
\end{aligned}$$

a 3×3 system (solve it!)

Example: ME distribution on $\mathbb{N}_0$ with fixed mean

Let P be the distribution of maximum entropy on $\{0, 1, \ldots\}$ with mean $m \geq 0$ (that is $m = \sum_i iP(i)$).

$$P(i) = \frac{1}{Z} e^{-\mu i} = \frac{1}{Z} \left(e^{-\mu}\right)^i$$

Denote $r = e^{-\mu}$.

$$1 = \sum_{i=0}^{\infty} P(i) = \frac{1}{Z} \sum_{i=0}^{\infty} r^i$$

So $Z = \frac{1}{1-r}$, i.e. $P(i) = r^i(1-r)$. This is called the *geometric* distribution.

$$\begin{aligned} m &= \sum_{i=0}^{\infty} iP(i) = (1-r)\, r \sum_{i=0}^{\infty} ir^{i-1} = (1-r)\, r \frac{\partial}{\partial r} \left(\sum_{i=0}^{\infty} r^i \right) \\ &= (1-r)\, r \frac{1}{(1-r)^2} = \frac{r}{1-r} = \frac{1}{1-r} - 1 \end{aligned}$$

So $r = 1 - \frac{1}{m+1}$.

Example: spins, first moments

Suppose $\sigma_i \in \{-1, 1\}$ for $i = 1, \ldots, N$, and we are given the N observables $m_i = \langle \sigma_i \rangle$ for $i = 1, \ldots, N$. Then the maximum entropy distribution is

$$P(\sigma) \propto e^{-\sum_i \mu_i \sigma_i} = \prod_i e^{-\mu_i \sigma_i}$$

As $m_i = \sum_\sigma P(\sigma) \sigma_i$,

$$\begin{aligned} m_i &= \frac{\sum_\sigma \sigma_i \prod_j e^{-\mu_j \sigma_j}}{\sum_\sigma \prod_j e^{-\mu_j \sigma_j}} \\ &= \frac{\sum_{\sigma^{-i}} \prod_{j \neq i} e^{-\mu_j \sigma_j} \sum_{\sigma_i} \sigma_i e^{-\mu_i \sigma_i}}{\sum_{\sigma^{-i}} \prod_{j \neq i} e^{-\mu_j \sigma_j} \sum_{\sigma_i} e^{-\mu_i \sigma_i}} \\ &= \frac{\sum_{\sigma_i} \sigma_i e^{-\mu_i \sigma_i}}{\sum_{\sigma_i} e^{-\mu_i \sigma_i}} = \tanh(-\mu_i) \end{aligned}$$

So $\mu_i = -\tanh^{-1}(m_i)$.

Example: spins, first two moments

Suppose $\sigma_i \in \{-1,1\}$ for $i = 1,\ldots,N$, and we are given the $\frac{1}{2}N(N-1)+N$ observables $c_{ij} = \langle \sigma_i \sigma_j \rangle$ for $1 \leq i < j \leq N$ and $m_i = \langle \sigma_i \rangle$ for $i = 1,\ldots,N$. Then the maximum entropy distribution is

$$P(\sigma) \propto e^{\sum_{i<j} J_{ij}\sigma_i\sigma_j + \sum_i h_i \sigma_i}$$

i.e. an Ising model!
This model has further restrictions on couplings and fields:
$\sum_\sigma P(\sigma)\sigma_i\sigma_j = c_{ij}$, $\sum_\sigma P(\sigma)\sigma_i = m_i$

- We know how to find J_{ij} and h_i in the case of a tree prior...

Maximum Likelihood and Maximum Entropy

For an Ising model, we have seen that

$$\mathscr{L}(\mathbf{J},\mathbf{h}) = \sum_{i<j} \tilde{c}_{ij} J_{ij} + \sum_i \tilde{m}_i h_i - \log Z_{\mathbf{J},\mathbf{h}}$$

But also that on the point of ML, $\tilde{c}_{ij} = c^*_{ij} = \langle \sigma_i \sigma_j \rangle$ and $\tilde{m}_i = m^*_i = \langle \sigma_i \rangle$. So

$$\mathscr{L}(\mathbf{J}^*,\mathbf{h}^*) = -\langle E \rangle_{\mathbf{J}^*,\mathbf{h}^*} - \log Z_{\mathbf{J}^*,\mathbf{h}^*} = S\left(P_{\mathbf{J}^*,\mathbf{h}^*}\right)$$

ML=ME

The $\mathbf{J},\mathbf{h}$ of ML describe the distribution of ME that reproduce $\tilde{m}_i, \tilde{c}_{ij}$

ARACNE

Data Processing inequality: If $P(x, y|z) = P(x|z)P(y|z)$ then $M_{xy} \leq \min\{M_{xz}, M_{yz}\}$

- This can be used for reconstruction (Califano & al, 2006): for every triplet i, j, k consider M_{ij}, M_{ik}, M_{jk} and eliminate the smallest one.
- The resulting graph contains the Chow-Liu tree.
- Running time $\sim N^3$

Reconstruction using independence

Observation

If $j \notin \partial i \cup \{i\}$

$$P(x_i, x_j | \mathbf{x}_{\partial i}) = P(x_i | \mathbf{x}_{\partial i}) P(x_j | \mathbf{x}_{\partial i})$$

and this can be used to identify $\mathbf{x}_{\partial i}$.

Reconstruction algorithm (Bresler, Mossel & Sly 2010)

For each i, check $\binom{N}{d}$ candidate neighborhoods ∂i. For each candidate ∂i , check condition on the remaining $N - d - 1$ nodes j

Running time: $\sim N^{d+1}$

The binary perceptron

- The *perceptron* is an stylized model of a neuron and the simplest example of neural network (NN). The binary perceptron receives $x_1, \ldots, x_N$ (real valued) inputs and produces a binary output

$$\sigma = \operatorname{sign}\left(\sum_{i=1}^{N} w_i x_i\right) = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x})$$

- A perceptron is capable of *learning:* let's suppose we are given $\mathbf{x}^1, \ldots, \mathbf{x}^M$ patterns together with desired classification labels $\sigma^1, \ldots, \sigma^M$. The learning procedure consists in finding $\mathbf{w}$ such that $\sigma^\mu = \operatorname{sign}(\mathbf{w} \cdot \mathbf{x}^\mu)$ for $\mu = 1, \ldots, M$
- This can be thought as the problem of finding the *separating plane*

The perceptron: generalizations and simplifications

- A slightly more general rule $\sigma = \text{sign}\left(\sum_{i=1}^{N} w_i x_i - \theta\right)$ can be simply implemented as an extra dummy output $x_{N+1} = -1$
- We can assume $\sigma^\tau = +1$ for all τ! Multiplying by σ^τ we get $1 = \sigma^\tau \sigma^\tau = \text{sign}\left(\mathbf{w} \cdot (\sigma^\tau \mathbf{x})\right)$
- We will be interested in the following cases: $\mathbf{w} \in \mathbb{R}^N$ and $\mathbf{w} \in \{-1, 1\}^N$ and $\mathbf{w} \in \{-q, \ldots, 0, \ldots, q\}^N$
- We can assume that $\|\mathbf{x}^\tau\| = 1$ since normalization doesn't affect classification

The online perceptron algorithm

Perceptron Algorithm

1. $\mathbf{w}_0 = 0$
2. done $= 0$
3. **while** done $= 0$:
   - done $= 1$
   - **for** $\tau = 1, \dots M$:
     - **if** $\mathbf{x}^\tau \cdot \mathbf{w}_t \leq 0$ (**mistake**):

       $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{x}^\tau$
       done $= 0$
       $t \leftarrow t + 1$

i.e. on any mistake, the algorithm greedily "helps" the classification of the missclassified pattern $\mathbf{x}^\tau$, because $\mathbf{w}_{t+1} \cdot \mathbf{x}^\tau = (\mathbf{w}_t + \mathbf{x}^\tau) \cdot \mathbf{x}^\tau = \mathbf{w}_t \cdot \mathbf{x}^\tau + 1$

The Perceptron algorithm (analysis)

Assume there exists a classifier $\mathbf{w}^*$, i.e. $\mathbf{w}^* \cdot \mathbf{x}^\tau > 0$ for $\tau = 1, \ldots, M$. Then the number of (mistake) events t must satisfy $t < \gamma^{-2}$

$$\gamma = \min_{\tau=1,\ldots,M} \mathbf{x}^\tau \cdot \mathbf{w}^*$$

i.e. the algorithm must terminate in less than γ^{-2} iterations.

1. $\mathbf{w}_{t+1} \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$.
   Because $\mathbf{w}_{t+1} \cdot \mathbf{w}^* = \mathbf{w}_t \cdot \mathbf{w}^* + \mathbf{x}^\tau \cdot \mathbf{w}^* \geq \mathbf{w}_t \cdot \mathbf{w}^* + \gamma$
2. $\|\mathbf{w}_{t+1}\|^2 \leq \|\mathbf{w}_t\|^2 + 1$
   Because $\|\mathbf{w}_{t+1}\|^2 = \mathbf{w}_t \cdot \mathbf{w}_t + 2\mathbf{x}^\tau \cdot \mathbf{w}_t + \mathbf{x}^\tau \cdot \mathbf{x}^\tau \leq \|\mathbf{w}_t\|^2 + 1$. This implies $\|\mathbf{w}_{t+1}\| \leq \sqrt{t}$

Now after t mistakes, $t\gamma \leq \mathbf{w}_{t+1} \cdot \mathbf{w}^* \leq \|\mathbf{w}_{t+1}\| \leq \sqrt{t}$, thus $t \leq \gamma^{-2}$

I/O association as Bayesian Inference

We will just use the bayesian framework assuming that

- Data samples are formed by both input and output $D = I, O$
- The stochastic machine defines a stochastic rule $P(O|S, I)$
- S and I are independent

Then we can use Bayes:

$$\begin{aligned} P(S|I,O) &= P(S,I,O)P(I,O)^{-1} = P(O|S,I)P(S)P(I)P(I,O)^{-1} \\ &\propto P(O|S,I)P(S) \end{aligned}$$

Similarly for $I^1, O^1, \ldots, I^M, O^M$ (assuming $I^1, \ldots, I^M, S$ independent):

$$P\left(S|I^1, O^1, \ldots, I^M, O^M\right) \propto \prod_{\mu=!}^{M} P(O^\mu|S, I^\mu)P(S)$$

Posterior distribution of binary perceptrons

Suppose we are given $I^1 = \mathbf{x}^1, O^1 = \boldsymbol{\sigma}^1, \ldots, I^M = \mathbf{x}^M, O^M = \boldsymbol{\sigma}^M$ and we want to describe the posterior distribution for the binary perceptron $S = \mathbf{w}$

$$P\left(\mathbf{w}|\mathbf{x}^1, \sigma^1, \ldots, \mathbf{x}^M, \sigma^M\right) \propto \prod_{\mu=!}^{M} P\left(\sigma^\mu|\mathbf{w}, \mathbf{x}^\mu\right) P\left(\mathbf{w}\right)$$

The rule can be e.g. for $\sigma^\mu \in \{-1, 1\}$ and $\mathbf{w}, \mathbf{x}^\mu \in \mathbb{R}^N$:

$$P\left(\sigma^\mu|\mathbf{w}, \mathbf{x}^\mu\right) = \delta\left(\sigma^\mu; \mathrm{sign}\left(\mathbf{w} \cdot \mathbf{x}^\mu\right)\right)$$

or more in general

$$P\left(\sigma^\mu|\mathbf{w}, \mathbf{x}^\mu\right) = f\left(\sigma^\mu; \mathbf{w} \cdot \mathbf{x}^\mu\right)$$

- Normally much easier to sample from $\mathbf{I}, \mathbf{O}$, given S than from a generic Boltzmann weight!

Posterior distribution as constraint satisfaction

- $P(S)$ can be set to favour diluted classifiers S, e.g. $P(S) \propto \prod_i e^{\mu \delta(S_i, 0)}$
- In fact, $P\left(S | I^1, O^1, \ldots, I^M, O^M\right)$ can be thought as a *direct* model:

$$P(S|\mathbf{I}, \mathbf{O}) \propto \prod_{\mu=1}^{M} \delta\left(O^\mu; \operatorname{sign}(S \cdot I^\mu)\right) \prod_i e^{\mu \delta(S_i, 0)}$$

- And solved with mean-field approximations (e.g. Belief Propagation)
- Particularly simple if e.g. $S_i \in \{-q, \ldots, 0, \ldots, q\}$

Recurrent network

Suppose we have a binary network $\sigma_i \in \{-1, 1\}$, and $w_{ij} \in \{-q, \ldots, 0, \ldots, q\}$. Consider

$$P(\sigma|\mathbf{w}) \propto \prod_i \delta\left(\sigma_i; \mathrm{sign}\left(\sum_{j\neq i} w_{ji}\sigma_j\right)\right)$$

and dilution prior $P(\mathbf{w}) = \prod_{i\neq j} e^{\mu\delta\left(w_{ij},0\right)}$

$$P\left(\mathbf{w}|\sigma^1, \ldots, \sigma^\mu\right) \propto \prod_i \left(\prod_{\mu=1}^{M} \delta\left(\sigma_i; \mathrm{sign}\left(\sum_{j\neq i} w_{ji}\sigma_j\right)\right) \prod_{j\neq i} e^{\mu\delta\left(w_{ij},0\right)}\right)$$

That is, the posterior distribution factorizes! N separate inference problems

For each i, BP can be used to find posterior statistics of the w_{ji}.

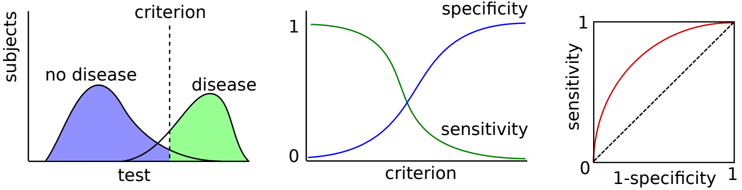
Insufficient data

- What to do if data is insufficient to infer a good model?
- ML/MAP are too risky: maybe the point of ML/MAP is not representative at all!
- In general we will be happier to get a small amount of sure information (e.g. a number of interactions that are present with high confidence) than a complete model with no poor confidence.
- How to measure performance (at least when we know the answer)? Something finer than correct/incorrect: ROC curves!

ROC curves

- ROC curves are thoroughly used in diagnostics
- Suppose we have a *test* which gives a scalar value $0 \leq \alpha \leq 1$ giving confidence of a certain disease.

Then depending on a given criterion value α_0, we will predict P (disease) if $\alpha \geq \alpha_0$ and N (no disease) if $\alpha < \alpha_0$. How good is the test?



- Sensitivity=TP/(TP+FN)=TP/disease
- Specificity=TN/(TN+FP)=TN/no disease
- Area below the curve: discrimination. Probability for a random subject with disease to have α larger than that of a random subject without disease.

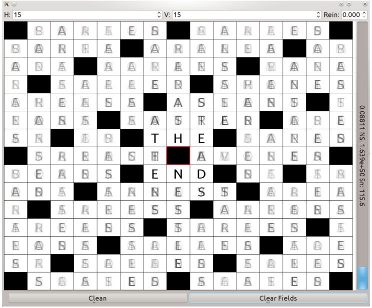
ROC curves for network inference

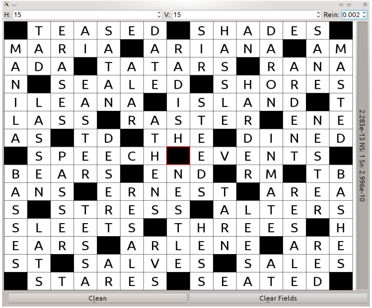
- Subject = edge
- With disease = present link, i.e. $e \in E$, $J_{ij} \neq 0$
- Without disease = absent link, i.e. $e \notin E$, $J_{ij} = 0$
- Criterion: e.g. M_{ij}, inferred $\left|J_{ij}^{ML}\right|$, inferred $\left|J_{ij}^{MAP}\right|$, $P\left(J_{ij} \neq 0 | \text{data}\right)$
- What criterion do we choose to have the best possible ROC curve?
- The best is to use $P\left(J_{ij} \neq 0 | \text{data}\right)$ as criterion! Better expected ROC curve than MAP or ML estimate.

Things to read

- Yedidia, Weiss & Freeman, Belief propagation and its generalizations + variational interpretations
- David MacKay's book "Information Theory, Inference and Learning Algorithms"
- Jaynes paper on Maximum entropy
- Chow-Liu paper on inference on trees
- Mézard & Montanari's book
- Mezard & Mora's Susceptibility Propagation

The End


H: 15
V: 15
Rein: 0.000
0.08811 NS: 1.639e+50 Sn: 115.6
Clean
Clear Fields


H: 15
V: 15
Rein: 0.002
T E A S E D S H A D E S
M A R I A A R I A N A A M
A D A T A T A R S R A N A
N S E A L E D S H O R E S
I L E A N A I S L A N D T
L A S S R A S T E R E N E
A S T D T H E D I N E D
S P E E C H E V E N T S
B E A R S E N D R M T B
A N S E R N E S T A R E A
S S T R E S S A L T E R S
S L E E T S T H R E E S H
E A R S A R L E N E A R E
S T S A L V E S S A L E S
S T A R E S S E A T E D
2.283e-13 NS: 1 Sn: 2.996e-10
Clean
Clear Fields