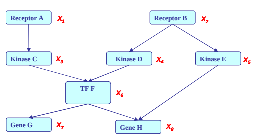
School of Computer Science
Carnegie Mellon

Learning Completely Observed Undirected Graphical Models

Probabilistic Graphical Models (10-708)

Lecture 21, Nov 28, 2007



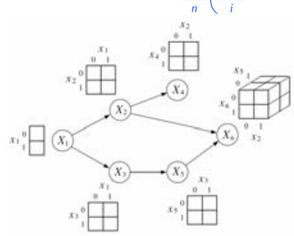
Eric Xing

Reading: J-Chap. 9,19,20; KF-Chap. 18

Recap: MLE for BNs

- Assuming the parameters for each CPD are globally independent, and all nodes are fully observed, then the log-likelihood function decomposes into a sum of local terms, one per node:

$$\ell(\theta; D) = \log p(D \mid \theta) = \log \prod_n \left(\prod_i p(x_{n,i} \mid \mathbf{x}_{\pi_i}, \theta_i) \right) = \sum_i \left(\sum_n \log p(x_{n,i} \mid \mathbf{x}_{\pi_i}, \theta_i) \right)$$



$$\theta_{ijk}^{ML} = \frac{n_{ijk}}{\sum_{i,j',k} n_{ij'k}}$$

- When some variables are not observed …

$$\left\langle \ell_c(\theta; x, z) \right\rangle_q \overset{\text{def}}{=} \sum_z q(z \mid x, \theta) \log p(x, z \mid \theta)$$

MLE for undirected graphical models

- For directed graphical models, the log-likelihood decomposes into a sum of terms, one per family (node plus parents).
- For undirected graphical models, the log-likelihood does not decompose, because the normalization constant Z is a function of **all** the parameters

$$P(x_1,\ldots,x_n) = \frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{x}_c) \qquad Z = \sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- In general, we will need to do inference (i.e., marginalization) to learn parameters for undirected models, even in the fully observed case.

Log Likelihood for UGMs with tabular clique potentials

- Sufficient statistics: for a UGM (V,E), the number of times that a configuration $\mathbf{x}$ (i.e., $\mathbf{X}_V=\mathbf{x}$) is observed in a dataset $D=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ can be represented as follows:

$$m(\mathbf{x}) \stackrel{\text{def}}{=} \sum_n \delta(\mathbf{x},\mathbf{x}_n) \quad \text{(total count)}, \quad \text{and} \quad m(\mathbf{x}_c) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{V\setminus c}} m(\mathbf{x}) \quad \text{(clique count)}$$

- In terms of the counts, the log likelihood is given by:

$$p(D|\theta) = \prod_n \prod_{\mathbf{x}} p(\mathbf{x}\mid\theta)^{\delta(\mathbf{x},\mathbf{x}_n)}$$

$$\log p(D|\theta) = \sum_n \sum_{\mathbf{x}} \delta(\mathbf{x},\mathbf{x}_n)\log p(\mathbf{x}\mid\theta) = \sum_{\mathbf{x}}\sum_n \delta(\mathbf{x},\mathbf{x}_n)\log p(\mathbf{x}\mid\theta)$$

$$\ell = \sum_{\mathbf{x}} m(\mathbf{x})\log\left(\frac{1}{Z}\prod_c \psi_c(\mathbf{x}_c)\right)$$

$$= \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c)\log\psi_c(\mathbf{x}_c) - N\log Z$$

- There is a nasty log Z in the likelihood

Derivative of log Likelihood

- Log-likelihood: $\ell = \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$
- First term: $\dfrac{\partial \ell_1}{\partial \psi_c(\mathbf{x}_c)} = m(\mathbf{x}_c) \Big/ \psi_c(\mathbf{x}_c)$
- Second term:

$$\frac{\partial \log Z}{\partial \psi_c(\mathbf{x}_c)} = \frac{1}{Z} \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\sum_{\tilde{\mathbf{x}}} \prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

Set the value of variables to $\tilde{\mathbf{x}}$

$$= \frac{1}{Z} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$$

$$= \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{1}{\psi_c(\tilde{\mathbf{x}}_c)} \frac{1}{Z} \prod_d \psi_d(\tilde{\mathbf{x}}_d)$$

$$= \frac{1}{\psi_c(\mathbf{x}_c)} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) p(\tilde{\mathbf{x}}) = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

Conditions on Clique Marginals

- Derivative of log-likelihood

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

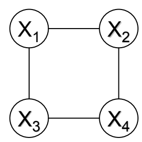
- Hence, for the maximum likelihood parameters, we know that:

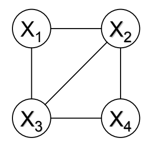
$$p^*_{MLE}(\mathbf{x}_c) = \frac{m(\mathbf{x}_c)}{N} \stackrel{\text{def}}{=} \tilde{p}(\mathbf{x}_c)$$

- In other words, at the maximum likelihood setting of the parameters, for each clique, the model marginals must be equal to the observed marginals (empirical counts).
- This doesn't tell us how to get the ML parameters, it just gives us a condition that must be satisfied when we have them.

MLE for undirected graphical models

- Is the graph decomposable (triangulated)?
- Are all the clique potentials defined on maximal cliques (not sub-cliques)? e.g., ψ_{123}, ψ_{234} not ψ_{12}, ψ_{23}, …





- Are the clique potentials full tables (or Gaussians), or parameterized more compactly, e.g. $\psi_c(\mathbf{x}_c) = \exp\left(\sum_c \theta_k f_k(\mathbf{x}_c)\right)$?

Decomposable?	Max clique?	Tabular?	Method
√	√	√	Direct
-	-	√	IPF
-	-	-	Gradient
-	-	-	GIS

MLE for decomposable undirected models

- Decomposable models:
  - G is decomposable ⇔ G is triangulated ⇔ G has a junction tree
  - Potential based representation: $p(\mathbf{x}) = \dfrac{\prod_c \psi_c(\mathbf{x}_c)}{\prod_s \varphi_s(\mathbf{x}_s)}$
- Consider a chain $X_1 - X_2 - X_3$. The cliques are (X_1, X_2) and (X_2, X_3); the separator is X_2
  - The empirical marginals must equal the model marginals.
- Let us guess that $\hat{p}_{MLE}(x_1, x_2, x_3) = \frac{\tilde{p}(x_1,x_2)\tilde{p}(x_2,x_3)}{\tilde{p}(x_2)}$
  - We can verify that such a guess satisfies the conditions:

    $$\hat{p}_{MLE}(x_1, x_2) = \sum_{x_3} \hat{p}_{MLE}(x_1, x_2, x_3) = \tilde{p}(x_1 \mid x_2) \sum_{x_3} \tilde{p}(x_2, x_3) = \tilde{p}(x_1, x_2)$$

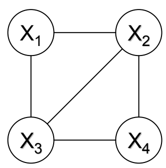
    and similarly $\hat{p}_{MLE}(x_2, x_3) = \tilde{p}(x_2, x_3)$

MLE for decomposable undirected models (cont.)

- Let us guess that $\hat{p}_{MLE}(x_1, x_2, x_3) = \frac{\tilde{p}(x_1, x_2)\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)}$
- To compute the clique potentials, just equate them to the empirical marginals (or conditionals), i.e., the separator must be divided into one of its neighbors. Then Z = 1.

$$\hat{\psi}_{12}^{MLE}(x_1, x_2) = \tilde{p}(x_1, x_2) \qquad \hat{\psi}_{23}^{MLE}(x_2, x_3) = \frac{\tilde{p}(x_2, x_3)}{\tilde{p}(x_2)} = \tilde{p}(x_2 \mid x_3)$$

- One more example:



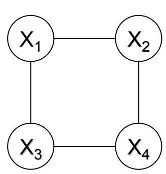
$$\hat{p}_{MLE}(x_1, x_2, x_3, x_4) = \frac{\tilde{p}(x_1, x_2, x_3)\tilde{p}(x_2, x_3, x_4)}{\tilde{p}(x_2, x_3)}$$

$$\hat{\psi}_{123}^{MLE}(x_2, x_3) = \frac{\tilde{p}(x_1, x_2, x_3)}{\tilde{p}(x_2, x_3)} = \tilde{p}(x_1 \mid x_2, x_3)$$

$$\hat{\psi}_{234}^{MLE}(x_2, x_3, x_4) = \tilde{p}(x_2, x_3, x_4)$$

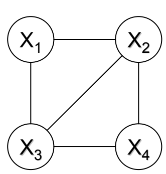
Non-decomposable and/or with non-maximal clique potentials

- If the graph is non-decomposable, and or the potentials are defined on non-maximal cliques (e.g., ψ_{12}, ψ_{34}), we could not equate empirical marginals (or conditionals) to MLE of cliques potentials.



$$p(x_1, x_2, x_3, x_4) = \prod_{\{i,j\}} \psi_{ij}(x_i, x_j)$$

$$\exists (i,j) \quad \text{s.t.} \quad \psi_{ij}^{\text{MLE}}(x_i, x_j) \neq \begin{cases} \tilde{p}(x_i, x_j) \\ \tilde{p}(x_i, x_j) / \tilde{p}(x_i) \\ \tilde{p}(x_i, x_j) / \tilde{p}(x_j) \end{cases}$$



Homework!

Iterative Proportional Fitting (IPF)

- From the derivative of the likelihood:

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

- we can derive another relationship:

$$\frac{\tilde{p}(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

  in which ψ_c appears implicitly in the model marginal $p(\mathbf{x}_c)$.

- This is therefore a fixed-point equation for ψ_c.
  - Solving ψ_c in closed-form is hard, because it appears on both sides of this implicit nonlinear equation.
- The idea of IPF is to hold ψ_c fixed on the right hand side (both in the numerator and denominator) and solve for it on the left hand side. We cycle through all cliques, then iterate:

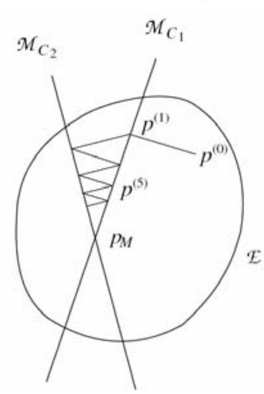
$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

Need to do inference here

Eric Xing 11

Properties of IPF Updates

- IPF iterates a set of fixed-point equations.
- However, we can prove it is also a coordinate ascent algorithm (coordinates = parameters of clique potentials).
- Hence at each step, it will increase the log-likelihood, and it will converge to a global maximum.
- I-projection: finding a distribution with the correct marginals that has the maximal entropy



Eric Xing 12

KL Divergence View

- IPF can be seen as coordinate ascent in the likelihood using the way of expressing likelihoods using KL divergences.
- Recall that we have shown maximizing the log likelihood is equivalent to minimizing the KL divergence (cross entropy) from the observed distribution to the model distribution:

$$\max \ell \Leftrightarrow \min KL\left(\tilde{p}(x) \,\|\, p(x \mid \theta)\right) = \sum_x \tilde{p}(x) \log \frac{\tilde{p}(x)}{p(x \mid \theta)}$$

- Using a property of KL divergence based on the conditional chain rule: $p(x) = p(x_a)p(x_b|x_a)$:

$$\begin{aligned}KL\left(q(x_a, x_b) \,\|\, p(x_a, x_b)\right) &= \sum_{x_a, x_b} q(x_a) q(x_b \mid x_a) \log \frac{q(x_a) q(x_b \mid x_a)}{p(x_a) p(x_b \mid x_a)} \\ &= \sum_{x_a, x_b} q(x_a) q(x_b \mid x_a) \log \frac{q(x_a)}{p(x_a)} + \sum_{x_a, x_b} q(x_a) q(x_b \mid x_a) \log \frac{q(x_b \mid x_a)}{p(x_b \mid x_a)} \\ &= KL\left(q(x_a) \,\|\, p(x_a)\right) + \sum_{x_a} q(x_a) KL\left(q(x_b \mid x_a) \,\|\, p(x_b \mid x_a)\right)\end{aligned}$$

IPF minimizes KL divergence

- Putting things together, we have

$$\begin{aligned}KL\left(\tilde{p}(\mathbf{x}) \,\|\, p(\mathbf{x} \mid \theta)\right) = {} & KL\left(\tilde{p}(\mathbf{x}_c) \,\|\, p(\mathbf{x}_c \mid \theta)\right) + \\ & \sum_{x_a} \tilde{p}(\mathbf{x}_c) KL\left(\tilde{p}(\mathbf{x}_{-c} \mid \mathbf{x}_c) \,\|\, p(\mathbf{x}_{-c} \mid \mathbf{x}_c)\right)\end{aligned}$$

It can be shown that changing the clique potential ψ_c has no effect on the conditional distribution, so the second term in unaffected.

- To minimize the first term, we set the marginal to the observed marginal, just as in IPF.
- We can interpret IPF updates as retaining the "old" conditional probabilities $p^{(t)}(\mathbf{x}_{-c}|\mathbf{x}_c)$ while replacing the "old" marginal probability $p^{(t)}(\mathbf{x}_c)$ with the observed marginal $\tilde{p}(\mathbf{x}_c)$.

Feature-based Clique Potentials

- So far we have discussed the most general form of an undirected graphical model in which cliques are parameterized by general potential functions $\psi_c(\mathbf{x}_c)$.
- But for large cliques these general potentials are exponentially costly for inference and have exponential numbers of parameters that we must learn from limited data.
- One solution: change the graphical model to make cliques smaller. But this changes the dependencies, and may force us to make more independence assumptions than we would like.
- Another solution: keep the same graphical model, but use a less general parameterization of the clique potentials.
- This is the idea behind feature-based models.

Features

- Consider a clique $\mathbf{x}_c$ of random variables in a UGM, e.g. three consecutive characters $c_1c_2c_3$ in a string of English text.
- How would we build a model of $p(c_1c_2c_3)$?
  - If we use a single clique function over $c_1c_2c_3$, the full joint clique potential would be huge: 263−1 parameters.
  - However, we often know that some particular joint settings of the variables in a clique are quite likely or quite unlikely. e.g. ing, ate, ion, ?ed, qu?, jkx, zzz,...
- A "feature" is a function which is vacuous over all joint settings except a few particular ones on which it is high or low.
  - For example, we might have $f_{ing}(c_1c_2c_3)$ which is 1 if the string is 'ing' and 0 otherwise, and similar features for '?ed', etc.
- We can also define features when the inputs are continuous. Then the idea of a cell on which it is active disappears, but we might still have a compact parameterization of the feature.

Features as Micropotentials

- By exponentiating them, each feature function can be made into a "micropotential". We can multiply these micropotentials together to get a **clique potential**.
- Example: a clique potential $\psi(c_1 c_2 c_3)$ could be expressed as:

$$\psi_c(c_1,c_2,c_3) = e^{\theta_{\text{ing}} f_{\text{ing}}} \times e^{\theta_{\text{?ed}} f_{\text{?ed}}} \times \ldots$$
$$= \exp\left\{\sum_{k=1}^{K} \theta_k f_k(c_1,c_2,c_3)\right\}$$

- This is still a potential over 26^3 possible settings, but only uses K parameters if there are K features.
  - By having one indicator function per combination of $\mathbf{x}_c$, we recover the standard tabular potential.

Combining Features

- Each feature has a weight θ_k which represents the numerical strength of the feature and whether it increases or decreases the probability of the clique.
- The marginal over the clique is a generalized exponential family distribution, actually, a GLIM:

$$p(c_1,c_2,c_3) \propto \exp\left\{\begin{matrix}\theta_{\text{ing}} f_{\text{ing}}(c_1,c_2,c_3) + \theta_{\text{?ed}} f_{\text{?ed}}(c_1,c_2,c_3) + \\ \theta_{\text{qu?}} f_{\text{qu?}}(c_1,c_2,c_3) + \theta_{\text{zzz}} f_{\text{zzz}}(c_1,c_2,c_3) + \cdots\end{matrix}\right\}$$

- In general, the features may be overlapping, unconstrained indicators or any function of any subset of the clique variables:

$$\psi_c(\mathbf{x}_c) \stackrel{\text{def}}{=} \exp\left\{\sum_{i \in I_c} \theta_k f_k(\mathbf{x}_{c_i})\right\}$$

- How can we combine feature into a probability model?

Feature Based Model

- We can multiply these clique potentials as usual:

$$p(\mathbf{x}) = \frac{1}{Z(\theta)} \prod_c \psi_c(\mathbf{x}_c) = \frac{1}{Z(\theta)} \exp\left\{ \sum_c \sum_{i \in I_c} \theta_k f_k(\mathbf{x}_{c_i}) \right\}$$

- However, in general we can forget about associating features with cliques and just use a simplified form:

$$p(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(\mathbf{x}_{c_i}) \right\}$$

- This is just our friend the exponential family model, with the features as sufficient statistics!

- Learning: recall that in IPF, we have $\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c) \dfrac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$
  - Not obvious how to update the weights and features individually

MLE of Feature Based UGMs

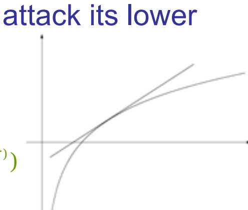
- Scaled likelihood function

$$\tilde{\ell}(\theta; D) = \ell(\theta; D)/N = \frac{1}{N} \sum_n \log p(x_n \mid \theta)$$
$$= \sum_x \tilde{p}(x) \log p(x \mid \theta)$$
$$= \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta)$$

- Instead of optimizing this objective directly, we attack its lower bound
  - The logarithm has a linear upper bound …

$$\log Z(\theta) \le \mu Z(\theta) - \log \mu - 1$$

  - This bound holds for all μ, in particular, for $\mu = Z^{-1}(\theta^{(t)})$
  - Thus we have



$$\tilde{\ell}(\theta; D) \ge \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

Generalized Iterative Scaling (GIS)

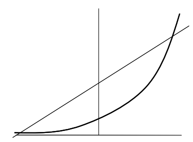
- Lower bound of scaled loglikelihood

$$\tilde{\ell}(\theta;D) \geq \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

- Define $\Delta\theta_i^{(t)} \overset{\text{def}}{=} \theta_i - \theta_i^{(t)}$

$$\begin{aligned}
\tilde{\ell}(\theta;D) &\geq \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{1}{Z(\theta^{(t)})} \sum_x \exp\left\{\sum_i \theta_i f_i(x)\right\} - \log Z(\theta^{(t)}) + 1 \\
&= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \frac{1}{Z(\theta^{(t)})} \sum_x \exp\left\{\sum_i \theta_i^{(t)} f_i(x)\right\} \exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1 \\
&= \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)}) \exp\left\{\sum_i \Delta\theta_i^{(t)} f_i(x)\right\} - \log Z(\theta^{(t)}) + 1
\end{aligned}$$

- Relax again
  - Assume $f_i(x) \geq 0, \quad \sum_i f_i(x) = 1$
  - Convexity of exponential: $\exp\left(\sum_i \pi_i x_i\right) \leq \sum_i \pi_i \exp(x_i)$



- We have:

$$\tilde{\ell}(\theta;D) \geq \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)}) \sum_i f_i(x) \exp\left(\Delta\theta_i^{(t)}\right) - \log Z(\theta^{(t)}) + 1 \overset{\text{def}}{=} \Lambda(\theta)$$

GIS

- Lower bound of scaled loglikelihood

$$\tilde{\ell}(\theta;D) \geq \sum_i \theta_i \sum_x \tilde{p}(x) f_i(x) - \sum_x p(x \mid \theta^{(t)}) \sum_i f_i(x) \exp\left(\Delta\theta_i^{(t)}\right) - \log Z(\theta^{(t)}) + 1 \overset{\text{def}}{=} \Lambda(\theta)$$

- Take derivative: $\dfrac{\partial \Lambda}{\partial \theta_i} = \sum_x \tilde{p}(x) f_i(x) - \exp\left(\Delta\theta_i^{(t)}\right) \sum_x p(x \mid \theta^{(t)}) f_i(x)$
- Set to zero

$$e^{\Delta\theta_i^{(t)}} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p(x \mid \theta^{(t)}) f_i(x)} = \frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)})$$

  - where $p^{(t)}(x)$ is the unnormalized version of $p(x|\theta^{(t)})$

- Update $\quad \theta_i^{(t+1)} = \theta_i^{(t)} + \Delta\theta_i^{(t)} \Rightarrow p^{(t+1)}(x) = p^{(t)}(x) e^{\Delta\theta_i^{(t)} f_i(x)}$

$$\Rightarrow \quad \begin{aligned}
p^{(t+1)}(x) &= \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} Z(\theta^{(t)}) \right)^{f_i(x)} \\
&= \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)} \left(Z(\theta^{(t)})\right)^{\sum_i f_i(x)} \\
&= p^{(t)}(x) \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)}
\end{aligned}$$

Where does the exponential form come from?

- Review: Maximum Likelihood for exponential family

$$\ell(\theta;D) = \sum_x m(x)\log p(x\mid\theta)$$
$$= \sum_x m(x)\left(\sum_i \theta_i f_i(x) - \log Z(\theta)\right)$$
$$= \sum_x m(x)\sum_i \theta_i f_i(x) - N\log Z(\theta)$$
$$\frac{\partial}{\partial\theta_i}\ell(\theta;D) = \sum_x m(x) f_i(x) - N\frac{\partial}{\partial\theta_i}\log Z(\theta)$$
$$= \sum_x m(x) f_i(x) - N\sum_x p(x\mid\theta) f_i(x)$$
$$\Rightarrow \quad \sum_x p(x\mid\theta) f_i(x) = \sum_x \frac{m(x)}{N} f_i(x) = \sum_x \tilde{p}(x\mid\theta) f_i(x)$$

- i.e., At ML estimate, the expectations of the sufficient statistics under the model must match empirical feature average.

Maximum Entropy

- We can approach the modeling problem from an entirely different point of view. Begin with some fixed feature expectations:
$$\sum_x p(x) f_i(x) = \alpha_i$$
- Assuming expectations are consistent, there may exist many distributions which satisfy them. Which one should we select?
  - The most uncertain or flexible one, i.e., the one with maximum entropy.
- This yields a new optimization problem:

$$\max_p \ \mathrm{H}(p(x)) = -\sum_x p(x)\log p(x)$$
$$\text{s.t.} \quad \sum_x p(x) f_i(x) = \alpha_i$$
$$\sum_x p(x) = 1$$

This is a **variational** definition of a distribution!

Solution to the MaxEnt Problem

- To solve the MaxEnt problem, we use Lagrange multipliers:

$$L = -\sum_x p(x)\log p(x) - \sum_i \theta_i \left(\sum_x p(x) f_i(x) - \alpha_i \right) - \mu \left(\sum_x p(x) - 1 \right)$$

$$\frac{\partial L}{\partial p(x)} = 1 + \log p(x) - \sum_i \theta_i f_i(x) - \mu$$

$$p^*(x) = e^{\mu-1} \exp\left\{ \sum_i \theta_i f_i(x) \right\}$$

$$Z(\theta) = e^{\mu-1} = \sum_x \exp\left\{ \sum_i \theta_i f_i(x) \right\} \qquad (\text{since } \textstyle\sum_x p^*(x) = 1)$$

$$p(x|\theta) = \frac{1}{Z(\theta)} \exp\left\{ \sum_i \theta_i f_i(x) \right\}$$

- So feature constraints + MaxEnt $\Rightarrow$ exponential family.
- Problem is strictly convex w.r.t. p, so solution is unique.

A more general MaxEnt problem

$$\begin{aligned}
\min_p \quad & \mathrm{KL}\big(p(x) \,\|\, h(x)\big) \\
& \stackrel{\text{def}}{=} \sum_x p(x) \log \frac{p(x)}{h(x)} = -\mathrm{H}(p) - \sum_x p(x) \log h(x) \\
\text{s.t.} \quad & \sum_x p(x) f_i(x) = \alpha_i \\
& \sum_x p(x) = 1
\end{aligned}$$

$$\Rightarrow \qquad p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\}$$

Constraints from Data

- Where do the constraints α_i come from?
- Just as before, measure the empirical counts on the training data:

$$\alpha_i = \sum_x \tfrac{m(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_x \tilde{p}(\mathbf{x}) f_i(\mathbf{x})$$

- This also ensures consistency automatically.
- Known as the “method of moments”. (c.f. law of large numbers)
- We have seen a case of convex duality:
  - In one case, we assume exponential family and show that ML implies model expectations must match empirical expectations.
  - In the other case, we assume model expectations must match empirical feature counts and show that MaxEnt implies exponential family distribution.
  - No duality gap $\Rightarrow$ yield the same value of the objective

Geometric interpretation

- All exponential family distribution:

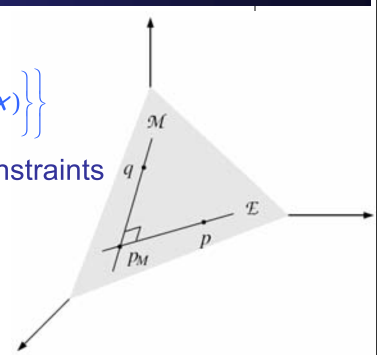
$$\mathcal{E} = \left\{ p(x) \, : \, p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\} \right\}$$

- All distributions satisfying moment constraints

$$\mathcal{M} = \left\{ p(x) \, : \, \sum_x p(x) f_i(x) = \sum_x \tilde{p}(x) f_i(x) \right\}$$

- Pythagorean theorem

$$\mathrm{KL}(q \,\|\, p) = \mathrm{KL}(q \,\|\, p_M) + \mathrm{KL}(p_M \,\|\, p)$$



MaxEnt:
$\min_p \ \mathrm{KL}(q \,\|\, h)$
s.t. $\quad q \in \mathcal{M}$
$\mathrm{KL}(q \,\|\, h) = \mathrm{KL}(q \,\|\, p_M) + \cancel{\mathrm{KL}(p_M \,\|\, h)}$

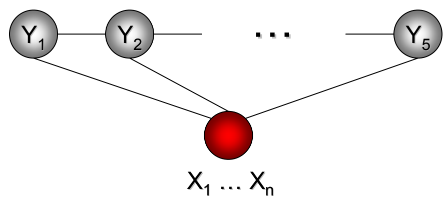
MaxLik:
$\min_p \ \mathrm{KL}(\tilde{p} \,\|\, p)$
s.t. $\quad q \in \mathcal{E}$
$\mathrm{KL}(\tilde{p} \,\|\, p) = \cancel{\mathrm{KL}(\tilde{p} \,\|\, p_M)} + \mathrm{KL}(p_M \,\|\, p)$

Conditional Random Fields

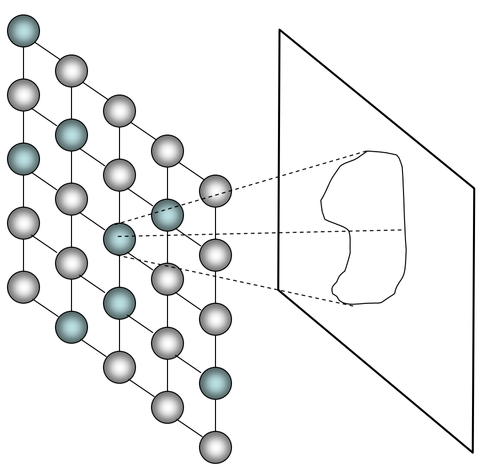
- So far we have focused on maxent models for density estimation.
- We can also formulate such models for classification and regression (conditional density estimation).

$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- The model above is like doing logistic regression on the features. Now features can be very complex, nonlinear functions of the data.



Conditional Random Fields



$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

- Allow arbitrary dependencies on input
- Clique dependencies on labels
- Use approximate inference for general graphs

Alternative Learning Strategy

- Recall that in CRF
  - We predict based on:

$$y^* \mid x = \arg\max_y p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{\sum_c \theta_c f_c(x, y_c)\right\}$$

  - And we learn based on:

$$\theta_c^* \mid \{y_n, x_n\} = \arg\max_{\theta_c} \prod_n p_\theta(y_n \mid x_n) = \prod_n \frac{1}{Z(\theta, x_n)} \exp\left\{\sum_c \theta_c f_c(x_n, y_{n,c})\right\}$$

- MaxMargin:
  - We predict based on:

$$y^* \mid x = \arg\max_y \sum_c \theta_c f_c(x, y_c) = \arg\max_y w^T F(x, y)$$

  - And we learn based on:

$$w^* \mid \{y_n, x_n\} = \arg\max_w \left(\max_{y'_n \neq y_n, \forall n} w^T \left(F(y_n, x_n) - F(y'_n, x_n)\right)\right)$$

Max-Margin Learning

$$\max \quad \frac{1}{2}\|w\| - \sum_n \xi_n$$

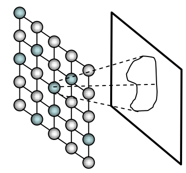
$$\text{s.t.} \quad w^T\left(F(y_n, x_n) - F(y'_n, x_n)\right) \geq \xi_n + \Delta(y'_n, y_n) \qquad \forall n, y'_n \in \mathcal{Y}_n \setminus y_n$$

$$\xi_n \geq 0$$

- Solutions:
  - Convex optimization (akin to SVM) with exponentially many constrains
  - Many algorithms and heuristics exist
    - Interior-point methods
    - Iterative active-support elimination
    - Inference based on GM
    - …

Open Problems

- Unsupervised CRF learning and MaxMargin Learning
  - We want to recognize a pattern that is maximally different from the rest!



  - What does margin or conditional likelihood mean in these cases? Given only $\{X_n\}$, how can we define the cost function?

$$p_\theta(y \mid x) = \frac{1}{Z(\theta, x)} \exp\left\{ \sum_c \theta_c f_c(x, y_c) \right\}$$

$$\text{margin} = w^T \left(F(y_n, x_n) - F(y'_n, x_n) \right)$$

  - Algorithmic challenge