

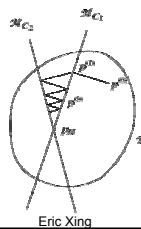
Advanced Machine Learning

Learning Graphical Models Maximum likelihood learning of undirected GM

Eric Xing

Lecture 15, August 13, 2009

Reading:



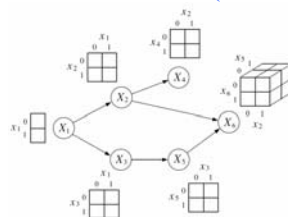
© Eric Xing @ CMU, 2006-2009

1

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:

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



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

- When some variables are not observed ...

$$\langle \mathcal{L}_c(\theta; x, z) \rangle_q \stackrel{\text{def}}{=} \sum_z q(z | x, \theta) \log p(x, z | \theta)$$

Eric Xing

© Eric Xing @ CMU, 2006-2009

2

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

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

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

3

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 $D = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ can be represented as follows:

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

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

$$\begin{aligned} p(D|\theta) &= \prod_n \prod_{\mathbf{x}} p(\mathbf{x} | \theta)^{\delta(\mathbf{x}, \mathbf{x}_n)} \\ \log p(D|\theta) &= \sum_n \sum_{\mathbf{x}} \delta(\mathbf{x}, \mathbf{x}_n) \log p(\mathbf{x} | \theta) = \sum_{\mathbf{x}} \sum_n \delta(\mathbf{x}, \mathbf{x}_n) \log p(\mathbf{x} | \theta) \\ \ell &= \sum_{\mathbf{x}} m(\mathbf{x}) \log \left(\frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c) \right) \\ &= \sum_c \sum_{\mathbf{x}_c} m(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z \end{aligned}$$

- There is a nasty $\log Z$ in the likelihood

Eric Xing

© Eric Xing @ CMU, 2006-2009

4

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: $\frac{\partial \ell_1}{\partial \psi_c(\mathbf{x}_c)} = \frac{m(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$
- Second term: $\frac{\partial \log Z}{\partial \psi_c(\mathbf{x}_c)} = \frac{1}{Z} \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\sum_{\tilde{\mathbf{x}}} \prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$
 $= \frac{1}{Z} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\prod_d \psi_d(\tilde{\mathbf{x}}_d) \right)$
 $= \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) \frac{1}{\psi_c(\tilde{\mathbf{x}}_c)} \frac{1}{Z} \prod_d \psi_d(\tilde{\mathbf{x}}_d)$
 $= \frac{1}{\psi_c(\mathbf{x}_c)} \sum_{\tilde{\mathbf{x}}} \delta(\tilde{\mathbf{x}}_c, \mathbf{x}_c) p(\tilde{\mathbf{x}}) = \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

5

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.

Eric Xing

© Eric Xing @ CMU, 2006-2009

6

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. $\psi_c(\mathbf{x}_c) = \exp(\sum_k \theta_k f_k(\mathbf{x}_c))$?

Eric Xing

© Eric Xing @ CMU, 2006-2009

7

Properties on MLE of clique potentials



- For decomposable models, where potentials are defined on maximal cliques, the MLE of clique potentials equate to the empirical marginals (or conditionals) of the corresponding clique. **Thus the MLE can be solved by inspection!!**
- 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 MLE of cliques potentials to empirical marginals (or conditionals).
 - Potential expressed as a tabular form: IPF
 - Feature-based potentials: GIS

Eric Xing

© Eric Xing @ CMU, 2006-2009

8

MLE for decomposable undirected models



- Decomposable models:
 - G is decomposable $\Leftrightarrow G$ is triangulated $\Leftrightarrow G$ has a junction tree
 - Potential based representation:
$$p(\mathbf{x}) = \frac{\prod_c \psi_c(\mathbf{x}_c)}{\prod_s \phi_s(\mathbf{x}_s)}$$
- 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:

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

9

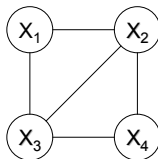
MLE for decomposable undirected models (cont.)



- 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 $Z = 1$.

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

- One more example:



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

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

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

Eric Xing

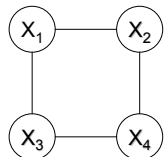
© Eric Xing @ CMU, 2006-2009

10

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

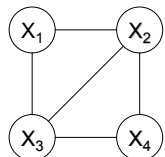


- If the graph is non-decomposable, and/or the potentials are defined on non-maximal cliques (e.g., ψ_{12} , ψ_{34}), we could not equate empirical marginals (or conditionals) to MLE of cliques potentials.



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

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



Homework!

Eric Xing

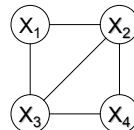
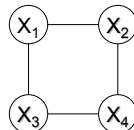
© Eric Xing @ CMU, 2006-2009

11

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. $\psi_c(\mathbf{x}_c) = \exp(\sum_k \theta_k f_k(\mathbf{x}_c))$?

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

12

Iterative Proportional Fitting (IPF)

- From the derivative of the likelihood:

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

- 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:

$$\psi_c^{(r+1)}(\mathbf{x}_c) = \psi_c^{(r)}(\mathbf{x}_c) \frac{\tilde{p}(\mathbf{x}_c)}{p^{(r)}(\mathbf{x}_c)} \leftarrow \text{Need to do inference here}$$

Eric Xing

© Eric Xing @ CMU, 2006-2009

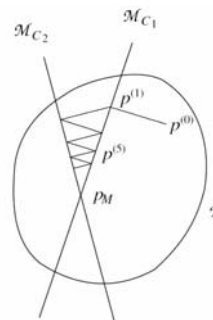
13

Properties of IPF Updates

- IPF iterates a set of fixed-point equations:

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

- 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

© Eric Xing @ CMU, 2006-2009

14

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:

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

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

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

15

IPF minimizes KL divergence



- Putting things together, we have

$$KL(\tilde{p}(x) \| p(x | \theta)) = KL(\tilde{p}(x_c) \| p(x_c | \theta)) + \sum_{x_a} \tilde{p}(x_c) KL(\tilde{p}(x_{-c} | x_c) \| p(x_{-c} | x_c))$$

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

- To minimize the first term, we **set the marginal to the observed marginal**, just as in IPF.
 - Note that this is only good when the model is decomposable !
- We can interpret IPF updates as retaining the “old” conditional probabilities $p^{(t)}(x_{-c} | x_c)$ while replacing the “old” marginal probability $p^{(t)}(x_c)$ with the observed marginal $\tilde{p}(x_c)$.

Eric Xing

© Eric Xing @ CMU, 2006-2009

16

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. $\psi_c(\mathbf{x}_c) = \exp(\sum_k \theta_k f_k(\mathbf{x}_c))$?

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

17

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 “**tabular**” 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.

Eric Xing

© Eric Xing @ CMU, 2006-2009

18

Features



- Consider a clique \mathbf{x}_c of random variables in a UGM, e.g. three consecutive characters $c_1 c_2 c_3$ in a string of English text.
- How would we build a model of $p(c_1 c_2 c_3)$?
 - If we use a single clique function over $c_1 c_2 c_3$, 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?**, **jkk**, **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 $f_{\text{ing}}(c_1 c_2 c_3)$ 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:

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

- This is still a potential over 26³ 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:

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

- 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?

Eric Xing

© Eric Xing @ CMU, 2006-2009

21

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 $\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c) \frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$
 - Not obvious how to use this rule to update the weights and features individually !!!

Eric Xing

© Eric Xing @ CMU, 2006-2009

22

MLE of Feature Based UGMs



- Scaled likelihood function

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

- Instead of optimizing this objective directly, we attack its lower bound

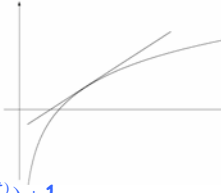
- The logarithm has a linear upper bound ...

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

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

- Thus we have

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



Eric Xing

© Eric Xing @ CMU, 2006-2009

23

Generalized Iterative Scaling (GIS)



- Lower bound of scaled loglikelihood

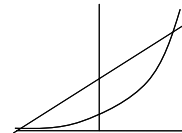
$$\tilde{\ell}(\theta; D) \geq \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \frac{Z(\theta)}{Z(\theta^{(t)})} - \log Z(\theta^{(t)}) + 1$$

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

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

- Relax again

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



- We have:

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

24

GIS



- Lower bound of scaled loglikelihood

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

- Take derivative: $\frac{\partial \Lambda}{\partial \theta_i} = \sum_x \tilde{p}(x) f_i(x) - \exp(\Delta \theta_i^{(t)}) \sum_x p(x | \theta^{(t)}) f_i(x)$

- Set to zero

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

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

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

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

Recall IPF:

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

25

Summary



- IPF is a general algorithm for finding MLE of UGMs.

- a **fixed-point equation** for ψ_c over single cliques, coordinate ascent
- I-projection in the clique marginal space
- Requires the potential to be fully parameterized
- The clique described by the potentials do not have to be max-clique
- For fully decomposable model, reduces to a single step iteration

- GIS

- Iterative scaling on general UGM with feature-based potentials
- IPF is a special case of GIS which the clique potential is built on features defined as an indicator function of clique configurations.

GIS:

$$p^{(t+1)}(x) = p^{(t)}(x) \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)}$$

IPF:

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

$$\theta_i^{(t+1)} = \theta_i^{(t)} + \log \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)$$

Eric Xing

© Eric Xing @ CMU, 2006-2009

26

Where does the exponential form come from?



- Review: Maximum Likelihood for exponential family

$$\begin{aligned}
 \ell(\theta; \mathcal{D}) &= \sum_{\mathbf{x}} m(\mathbf{x}) \log p(\mathbf{x} | \theta) \\
 &= \sum_{\mathbf{x}} m(\mathbf{x}) \left(\sum_i \theta_i f_i(\mathbf{x}) - \log Z(\theta) \right) \\
 &= \sum_{\mathbf{x}} m(\mathbf{x}) \sum_i \theta_i f_i(\mathbf{x}) - N \log Z(\theta) \\
 \frac{\partial}{\partial \theta_i} \ell(\theta; \mathcal{D}) &= \sum_{\mathbf{x}} m(\mathbf{x}) f_i(\mathbf{x}) - N \frac{\partial}{\partial \theta_i} \log Z(\theta) \\
 &= \sum_{\mathbf{x}} m(\mathbf{x}) f_i(\mathbf{x}) - N \sum_{\mathbf{x}} p(\mathbf{x} | \theta) f_i(\mathbf{x})
 \end{aligned}$$

$$\Rightarrow \sum_{\mathbf{x}} p(\mathbf{x} | \theta) f_i(\mathbf{x}) = \sum_{\mathbf{x}} \frac{m(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x} | \theta) f_i(\mathbf{x})$$

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

27

Maximum Entropy



- We can approach the modeling problem from an entirely different point of view. Begin with some fixed feature expectations:

$$\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) = \alpha_i$$

- 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:

$$\max_p H(p(\mathbf{x})) = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x})$$

$$\text{s.t. } \sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) = \alpha_i$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

28

Solution to the MaxEnt Problem



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

$$L = -\sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) - \sum_i \theta_i \left(\sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) - \alpha_i \right) - \mu \left(\sum_{\mathbf{x}} p(\mathbf{x}) - 1 \right)$$

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

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

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

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

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

29

A more general MaxEnt problem



$$\min_p \quad \text{KL}(p(\mathbf{x}) \| h(\mathbf{x}))$$

$$\stackrel{\text{def}}{=} \sum_{\mathbf{x}} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{h(\mathbf{x})} = -H(p) - \sum_{\mathbf{x}} p(\mathbf{x}) \log h(\mathbf{x})$$

$$\text{s.t.} \quad \sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) = \alpha_i$$

$$\sum_{\mathbf{x}} p(\mathbf{x}) = 1$$

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

Eric Xing

© Eric Xing @ CMU, 2006-2009

30

Constraints from Data

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

$$\alpha_i = \sum_{\mathbf{x}} \frac{m(\mathbf{x})}{N} f_i(\mathbf{x}) = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) f_i(\mathbf{x})$$

- 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

Eric Xing

© Eric Xing @ CMU, 2006-2009

31

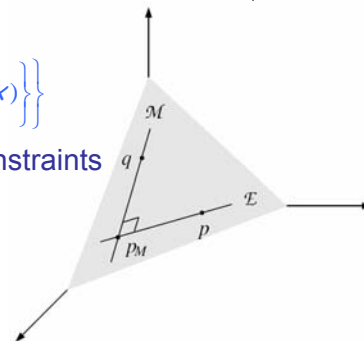
Geometric interpretation

- All exponential family distribution:

$$\mathcal{E} = \left\{ p(\mathbf{x}) : p(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} h(\mathbf{x}) \exp \left\{ \sum_i \theta_i f_i(\mathbf{x}) \right\} \right\}$$
- All distributions satisfying moment constraints

$$\mathcal{M} = \left\{ p(\mathbf{x}) : \sum_{\mathbf{x}} p(\mathbf{x}) f_i(\mathbf{x}) = \sum_{\mathbf{x}} \tilde{p}(\mathbf{x}) f_i(\mathbf{x}) \right\}$$
- Pythagorean theorem

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



MaxEnt :

$$\min_p \text{KL}(q \| h)$$

s.t. $q \in \mathcal{M}$

$$\text{KL}(q \| h) = \text{KL}(q \| p_M) + \text{KL}(p_M \| h)$$

MaxLik :

$$\min_p \text{KL}(\tilde{p} \| p)$$

s.t. $q \in \mathcal{E}$

$$\text{KL}(\tilde{p} \| p) = \text{KL}(\tilde{p} \| p_M) + \text{KL}(p_M \| p)$$

Eric Xing

© Eric Xing @ CMU, 2006-2009

32

Summary



- Exponential family distribution can be viewed as the solution to an variational expression --- the maximum entropy!
- The max-entropy principle to parameterization offers a dual perspective to the MLE.