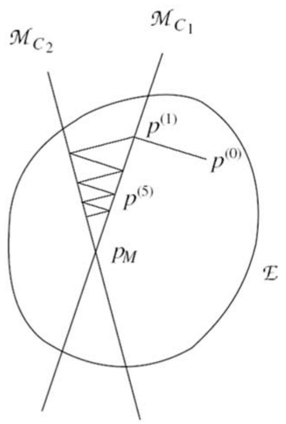
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Maximum likelihood learning of fully-observed undirected GM



Eric Xing

Lecture 7, February 8, 2017

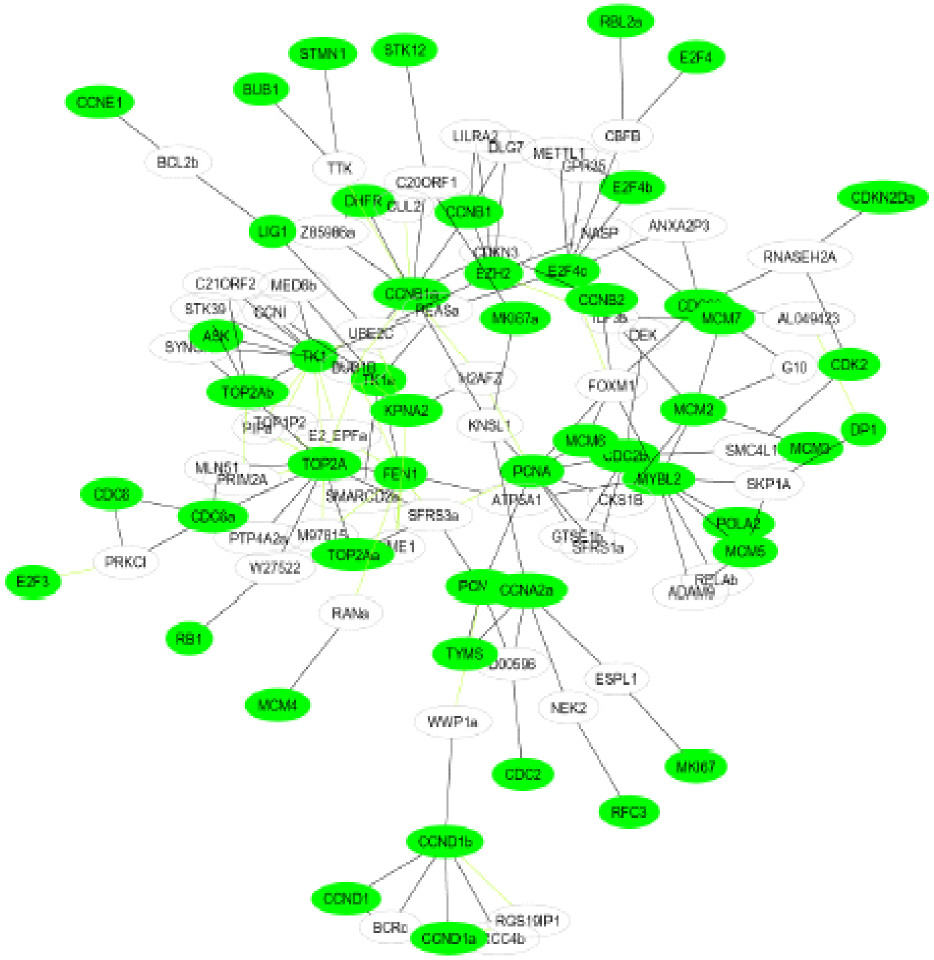
Reading: MJ Chap 9, and 11

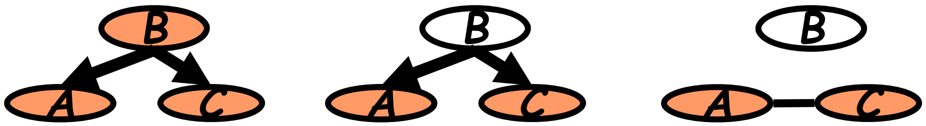
Undirected Graphical Models

- **Why?**

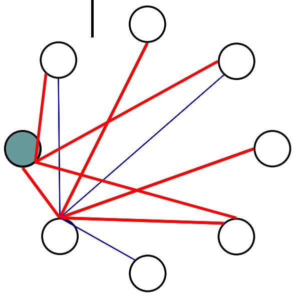
  Sometimes an UNDIRECTED association graph makes more sense and/or is more informative

  - **gene expressions may be influenced by unobserved factor that are post-transcriptionally regulated**

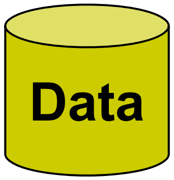




  - **The unavailability of the state of B results in a constrain over A and C**



ML Structural Learning via Neighborhood Selection for completely observed MRF



$$(x_1^{(1)},\dots,x_n^{(1)})$$
$$(x_1^{(2)},\dots,x_n^{(2)})$$
$$\dots$$
$$(x_1^{(M)},\dots,x_n^{(M)})$$

Gaussian Graphical Models

- Multivariate Gaussian density:

$$p(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

- WOLG: let $\mu = 0 \quad Q = \Sigma^{-1}$

$$p(x_1, x_2, \cdots, x_p \mid \mu = 0, Q) = \frac{|Q|^{1/2}}{(2\pi)^{n/2}} \exp\left\{-\tfrac{1}{2}\sum_i q_{ii}(x_i)^2 - \sum_{i<j} q_{ij} x_i x_j\right\}$$

- We can view this as a continuous Markov Random Field with potentials defined on every node and edge:

Pairwise MRF (e.g., Ising Model)

- Assuming the nodes are discrete, and edges are weighted, then for a sample x$_d$, we have

$$P(\mathbf{x}_d|\Theta) \quad = \quad \exp\left(\sum_{i\in V} \theta^t_{ii} x_{d,i} + \sum_{(i,j)\in E} \theta_{ij} x_{d,i} x_{d,j} - A(\Theta) \right)$$

The covariance and the precision matrices

- Covariance matrix Σ

$$\Sigma_{i,j} = 0 \quad \Rightarrow \quad X_i \perp X_j \quad \text{or} \quad p(X_i, X_j) = p(X_i)p(X_j)$$

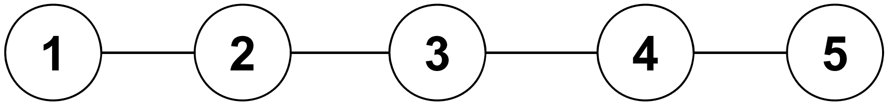
  - Graphical model interpretation?

- Precision matrix $Q = \Sigma^{-1}$

$$Q_{i,j} = 0 \quad \Rightarrow \quad X_i \perp X_j | \mathbf{X}_{-ij} \quad \text{or} \quad p(X_i, X_j | \mathbf{X}_{-ij}) = p(X_i | \mathbf{X}_{-ij}) p(X_j | \mathbf{X}_{-ij})$$

  - Graphical model interpretation?

Sparse precision vs. sparse covariance in GGM



$$\Sigma^{-1} = \begin{pmatrix} 1 & 6 & 0 & 0 & 0 \\ 6 & 2 & 7 & 0 & 0 \\ 0 & 7 & 3 & 8 & 0 \\ 0 & 0 & 8 & 4 & 9 \\ 0 & 0 & 0 & 9 & 5 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 0.10 & 0.15 & -0.13 & -0.08 & 0.15 \\ 0.15 & -0.03 & 0.02 & 0.01 & -0.03 \\ -0.13 & 0.02 & 0.10 & 0.07 & -0.12 \\ -0.08 & 0.01 & 0.07 & -0.04 & 0.07 \\ 0.15 & -0.03 & -0.12 & 0.07 & 0.08 \end{pmatrix}$$

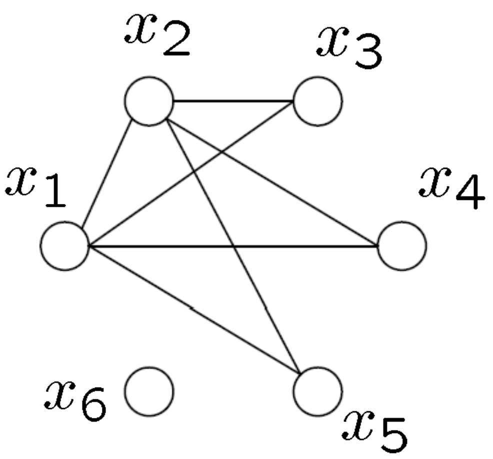
$$\Sigma^{-1}_{15} = 0 \Leftrightarrow X_1 \perp X_5 \,\big|\, X_{nbrs(1)\text{ or }nbrs(5)}$$

$$\not\Rightarrow$$

$$X_1 \perp X_5 \Leftrightarrow \Sigma_{15} = 0$$

Another example

$$Q = \begin{pmatrix} * & * & * & * & * & 0 \\ * & * & * & * & * & 0 \\ * & * & * & 0 & 0 & 0 \\ * & * & 0 & * & 0 & 0 \\ * & * & 0 & 0 & * & 0 \\ 0 & 0 & 0 & 0 & 0 & * \end{pmatrix}$$



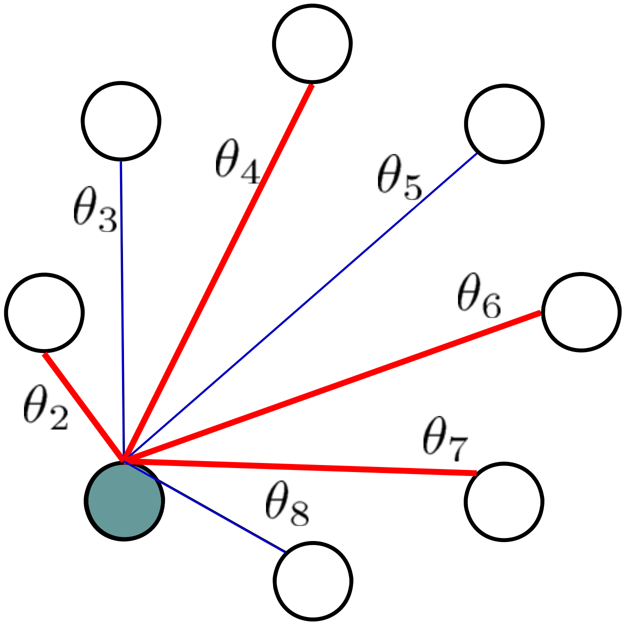
- How to estimate this MRF?
- What if $p >> n$
  - MLE does not exist in general!
  - What about only learning a “sparse” graphical model?
    - This is possible when $s=o(n)$
    - Very often it is the structure of the GM that is more interesting …

Recall lasso

$$\hat{\theta}_i \quad = \quad \arg\min_{\theta_i} l(\theta_i) + \lambda_1 || \ \theta_i \ ||_1$$

$$\text{where} \quad l(\theta_i) = \log \ P\big(y_i | \mathbf{x}_i, \theta_i\big).$$

Graph Regression

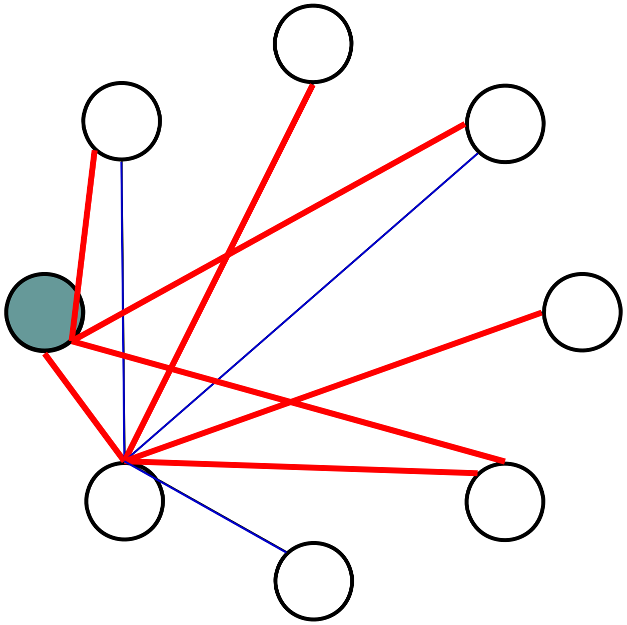


Neighborhood selection

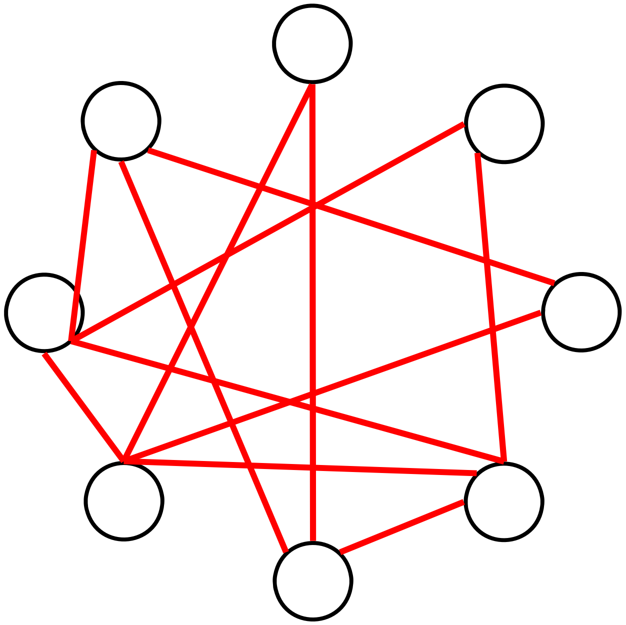
Lasso:

$$\hat{\theta} \;=\; \arg\min_{\theta} \sum_{t=1}^{T} l(\theta) + \lambda_1 \| \, \theta \, \|_1$$

Graph Regression



Graph Regression



It can be shown that:
given *iid* samples, and under several technical conditions (e.g., "irrepresentable"), the recovered structured is "sparsistent" even when p >> n

Learning Ising Model (i.e. pairwise MRF)

- Assuming the nodes are discrete, and edges are weighted, then for a sample x$_d$, we have

$$P(\mathbf{x}_d|\Theta) \;=\; \exp\left(\sum_{i\in V}\theta^t_{ii}x_{d,i} + \sum_{(i,j)\in E}\theta_{ij}x_{d,i}x_{d,j} - A(\Theta)\right)$$

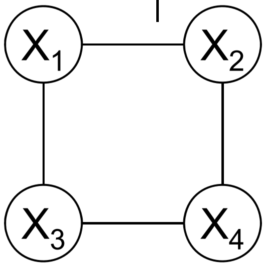
- It can be shown following the same logic that we can use L_1 regularized logistic regression to obtain a sparse estimate of the neighborhood of each variable in the discrete case.

Consistency

- **Theorem**: for the graphical regression algorithm, under certain verifiable conditions (omitted here for simplicity):

$$\mathbb{P}\left[\hat{G}(\lambda_n) \neq G\right] = \mathcal{O}\left(\exp\left(-Cn^{\epsilon}\right)\right) \to 0$$

Note the from this theorem one should see that the regularizer is not actually used to introduce an “artificial” sparsity bias, but a devise to ensure consistency under finite data and high dimension condition.



ML Parameter Est. for completely observed MRFs of given structure

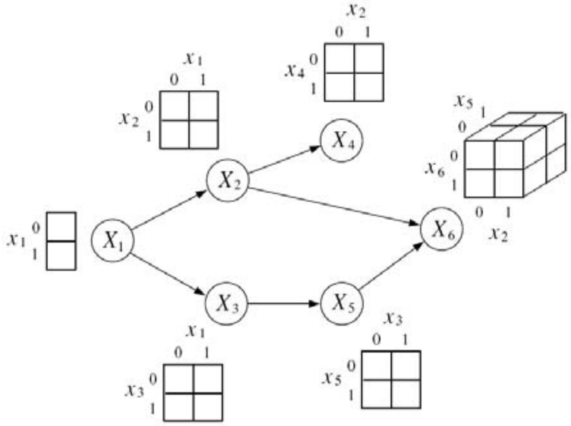
- The data:

$$\{ (z_1,x_1), (z_2,x_2), (z_3,x_3), \ldots (z_N,x_N)\}$$

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:

$$\ell(\theta;D)=\log p(D\,|\,\theta)=\log\prod_{n}\left(\prod_{i} p(x_{n,i}\,|\,\mathbf{x}_{\pi_i},\theta_i)\right)=\sum_{i}\left(\sum_{n}\log p(x_{n,i}\,|\,\mathbf{x}_{\pi_i},\theta_i)\right)$$



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

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

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

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

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:

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

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

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

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

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

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

- There is a nasty log Z in the likelihood

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:

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

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

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

- 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: $$\frac{\partial \ell_1}{\partial \psi_c(\mathbf{x}_c)} = m(\mathbf{x}_c) \Big/ \psi_c(\mathbf{x}_c)$$

- Second term: $$\frac{\partial \log Z}{\partial \psi_c(\mathbf{x}_c)} = \frac{1}{Z} \frac{\partial}{\partial \psi_c(\mathbf{x}_c)} \left(\sum_{\tilde{\mathbf{x}}} \prod_{d} \psi_d(\tilde{\mathbf{x}}_d) \right)$$

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

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

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

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

© Eric Xing @ CMU, 2005-2017

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:

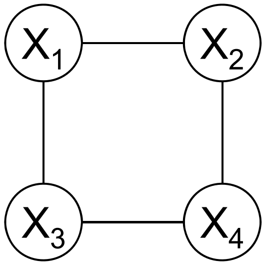
$$p^*_{MLE}(\mathbf{x}_c) = \frac{m(\mathbf{x}_c)}{N} \overset{\text{def}}{=} \tilde{p}(\mathbf{x}_c)$$

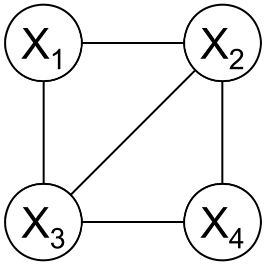
- In other words, at the maximum likelihood setting of the parameters, for each clique, the model marginals must be equal to the observed marginals (empirical counts).

- 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. $\psi_c(\mathbf{x}_c) = \exp\left(\sum_c \theta_k f_k(\mathbf{x}_c)\right)$?

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

MLE for decomposable undirected models

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

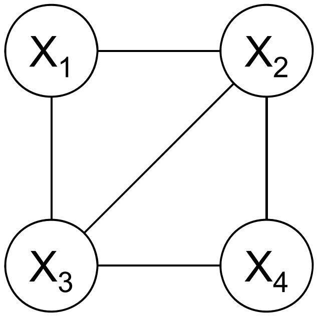
$$\hat{p}_{MLE}(x_2, x_3) = \tilde{p}(x_2, x_3)$$

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:



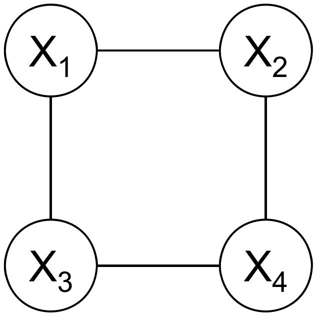
$$\hat{p}_{MLE}(x_1,x_2,x_3,x_4) = \frac{\tilde{p}(x_1,x_2,x_3)\tilde{p}(x_2,x_3,x_4)}{\tilde{p}(x_2,x_3)}$$

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

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

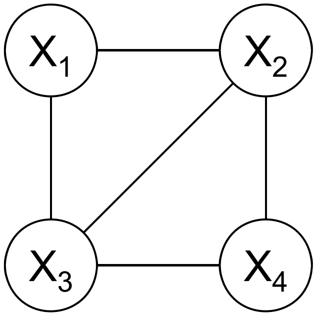
Non-decomposable and/or with non-maximal clique potentials

- 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.



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

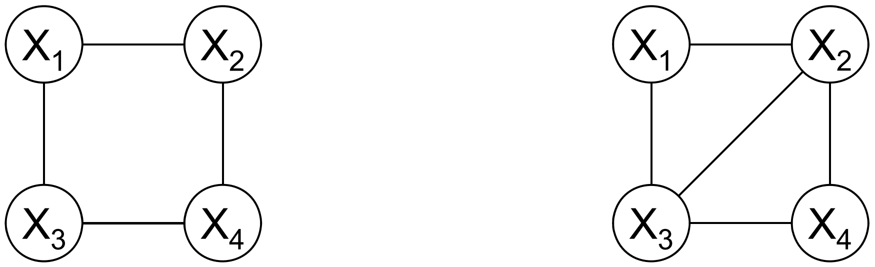
$$\exists(i,j) \quad \text{s.t.} \quad \psi_{ij}^{\mathrm{MLE}}(x_i,x_j)\neq\begin{cases}\tilde{p}(x_i,x_j)\\ \tilde{p}(x_i,x_j)/\tilde{p}(x_i)\\ \tilde{p}(x_i,x_j)/\tilde{p}(x_j)\end{cases}$$



Homework!

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

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:

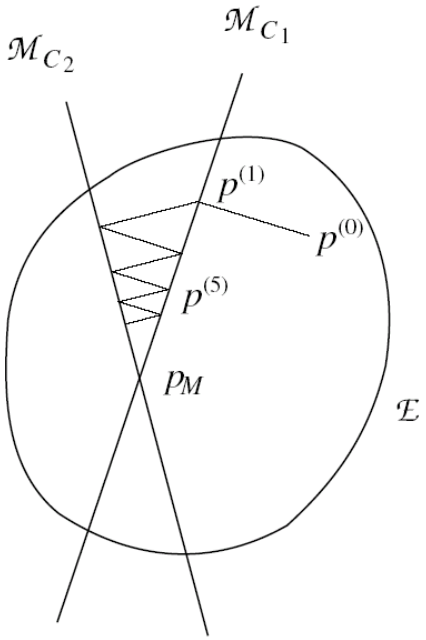
$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c)\frac{\tilde{p}(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)} \longleftarrow \text{Need to do inference here}$$

Properties of IPF Updates

- IPF iterates a set of fixed-point equations:

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

- 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



KL Divergence View

- IPF can be seen as coordinate ascent in the likelihood using the way of expressing likelihoods using KL divergences.
- We can show that 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: $p(x) = p(x_a)p(x_b|x_a)$:

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

IPF minimizes KL divergence

- Putting things together, we have

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

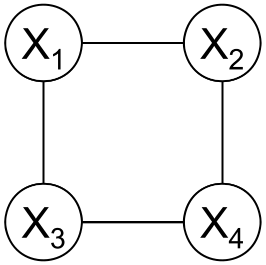
  It can be shown that changing the clique potential ψ_c has no effect on the conditional distribution, so the second term in unaffected.

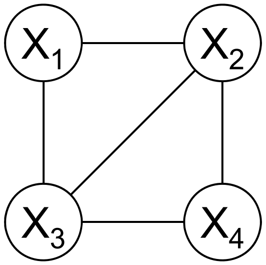
- To minimize the first term, we set the marginal to the observed marginal, just as in IPF.
  - Note that this is only good when the model is decomposable !

- 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 $\tilde{p}(\mathbf{x}_c)$.

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

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.

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: 26^3-1 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 $f_{\text{ing}}(c_1c_2c_3)$ 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:

$$\psi_c(c_1, c_2, c_3) = e^{\theta_{\mathrm{ing}} f_{\mathrm{ing}}} \times e^{\theta_{\mathrm{?ed}} f_{\mathrm{?ed}}} \times \ldots$$

$$= \exp\left\{ \sum_{k=1}^{K} \theta_k f_k(c_1, c_2, c_3) \right\}$$

- 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:

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

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

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

- 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 use this rule to update the weights and features individually !!!**

MLE of Feature Based UGMs

- Scaled likelihood function

$$\tilde{\ell}(\theta;D) = \ell(\theta;D)/N = \frac{1}{N}\sum_n \log p(x_n \mid \theta)$$

$$= \sum_x \tilde{p}(x) \log p(x \mid \theta)$$

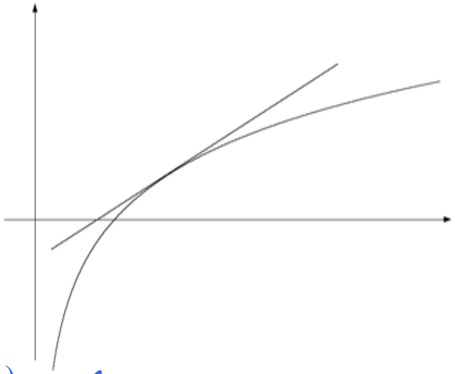
$$= \sum_x \tilde{p}(x) \sum_i \theta_i f_i(x) - \log Z(\theta)$$

- 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)

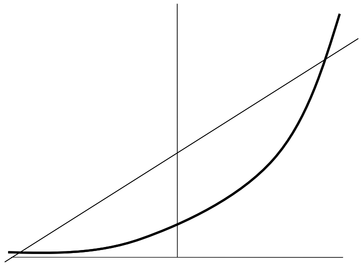
- Lower bound of scaled loglikelihood

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

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

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

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



- We have:

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

GIS

- Lower bound of scaled loglikelihood

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

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

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

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

$$\Rightarrow \quad = \frac{p^{(t)}(x)}{Z(\theta^{(t)})} \prod_i \left(\frac{\sum_x \tilde{p}(x) f_i(x)}{\sum_x p^{(t)}(x) f_i(x)} \right)^{f_i(x)} \left(Z(\theta^{(t)}) \right)^{\sum_i f_i(x)}$$

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

Recall IPF:

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

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:

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

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

IPF:

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

Where does the exponential form come from?

- Review: Maximum Likelihood for exponential family

$$\begin{aligned}\ell(\theta;D) &= \sum_x m(x)\log p(x\mid\theta)\\ &= \sum_x m(x)\left(\sum_i \theta_i f_i(x) - \log Z(\theta)\right)\\ &= \sum_x m(x)\sum_i \theta_i f_i(x) - N\log Z(\theta)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\ell(\theta;D) &= \sum_x m(x) f_i(x) - N\frac{\partial}{\partial\theta_i}\log Z(\theta)\\ &= \sum_x m(x) f_i(x) - N\sum_x p(x\mid\theta) f_i(x)\end{aligned}$$

$$\Rightarrow \quad \sum_x p(x\mid\theta)\, f_i(x) = \sum_x \frac{m(x)}{N}\, f_i(x) = \sum_x \tilde{p}(x\mid\theta)\, f_i(x)$$

- 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:

$$\max_p \ \mathrm{H}\big(p(x)\big) = -\sum_x p(x) \log p(x)$$

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

$$\sum_x p(x) = 1$$

This is a **<u>variational</u>** 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 ⇒ exponential family.
- Problem is strictly convex w.r.t. *p*, so solution is unique.

A more general MaxEnt problem

$$\min_p \quad \mathrm{KL}\big(p(x)\,\|\,h(x)\big)$$

$$\overset{\text{def}}{=} \sum_x p(x)\log\frac{p(x)}{h(x)} = -\mathrm{H}(p) - \sum_x p(x)\log h(x)$$

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

$$\sum_x p(x) = 1$$

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

Constraints from Data

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

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

- 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 ⇒ yield the same value of the objective

Geometric interpretation

- All exponential family distribution:

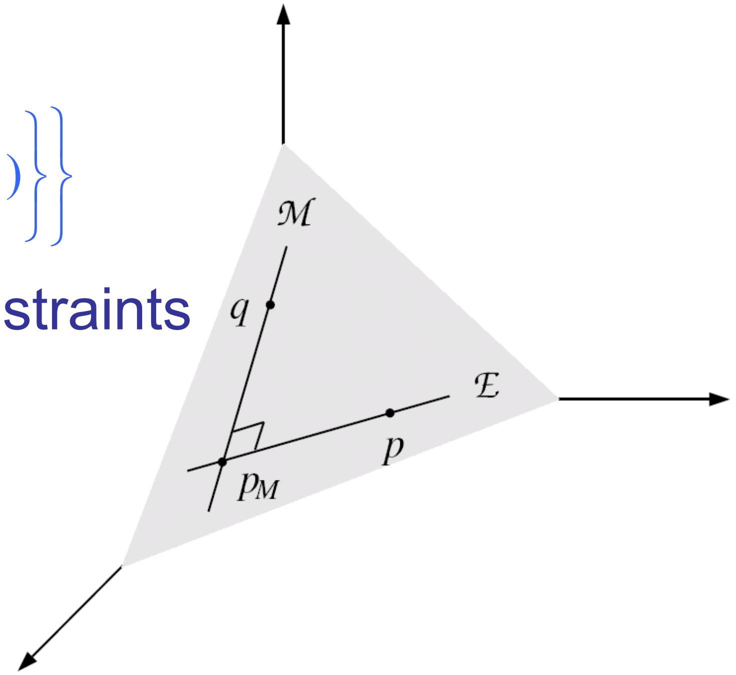
$$\mathcal{E} = \left\{ p(x) \ : \ p(x|\theta) = \frac{1}{Z(\theta)} h(x) \exp\left\{ \sum_i \theta_i f_i(x) \right\} \right\}$$

- 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 :

$$\min_p \quad \mathrm{KL}(q \,\|\, h)$$

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

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

MaxLik :

$$\min_p \quad \mathrm{KL}(\tilde{p} \,\|\, p)$$

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

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

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.