Learning Markov Networks

Presented by:
Mark Berlin, Barak Gross

Introduction

"We shall begin, perhaps…"
Eugene Onegin, Chapter VI

"Off did he take, I followed at his heels."
Inferno, Canto II

Reminder

- Until now we considered Bayesian Networks
  - Intuitive
  - Easily Decomposable → Local

- Markov Networks are a wholly different story

Example

- Consider the simple Markov Network A – B – C
  - Factors: $\phi_1(A,B), \phi_2(B,C)$
  - Partition function: $Z = \sum_{a,b,c} \phi_1(a,b)\phi_2(b,c)$
  - Log-likelihood:

$$l(\boldsymbol{\theta} : \mathcal{D}) = \sum_m [\ln \phi_1(a[m], b[m]) + \ln \phi_2(b[m], c[m]) - \ln Z(\boldsymbol{\theta})]$$

$$= \sum_{a,b} M[a,b] \ln \phi_1(a,b) + \sum_{b,c} M[b,c] \ln \phi_2(b,c) - M \cdot \ln Z(\boldsymbol{\theta})$$

- Z *couples the factors together* and precludes decomposition
- Transition to Bayesian representation is clearly not a panacea, as the MN and BN models are not equivalent

Programme

- Maximum Likelihood
  - Spoiler: Bad news ahead
- Dealing with Missing Data
- Alternative Objectives
  - Pseudo-Likelihood
  - Contrasting Divergence

Max likelihood

“Who controls the past controls the future.”

George Orwell

Log-linear models

Definition 4.15 log-linear model

A distribution P is a log-linear model *over a Markov network $\mathcal{H}$ if it is associated with:*

- *a set of features $\mathcal{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_k(\boldsymbol{D}_k)\}$, where each $\boldsymbol{D}_i$ is a complete subgraph in $\mathcal{H}$,*
- *a set of weights $w_1, \ldots, w_k$,*

such that

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \exp\left[-\sum_{i=1}^{k} w_i f_i(\boldsymbol{D}_i)\right].$$

■

- Let us now focus on the particular case where the features *f* are *indicators*, i.e. of the form $f_{a^0,b^0}(a,b) = \boldsymbol{I}\{a = a^0\} \cdot \boldsymbol{I}\{b = b^0\}$

- The weights w_i become the factor values we look for, of the form θ_{a^0,b^0}

Log likelihood

- The log-likelihood thus becomes:

$$l(\boldsymbol{\theta} : \mathcal{D}) = \sum_i \theta_i \cdot \sum_m f_i(\xi[m]) - M \cdot \ln Z(\boldsymbol{\theta})$$

$$= M \cdot \sum_i \theta_i \cdot \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) - M \cdot \ln Z(\boldsymbol{\theta})$$

  - M is the number of samples
  - $\boldsymbol{d}_i$ is a set of values for the set of variables $\boldsymbol{D}_i$ which is a clique in the Markov Network

Log likelihood, cont.

$$\frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \sum_i \theta_i \cdot \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) - \ln Z(\boldsymbol{\theta})$$

- The first term is linear in θ_i
- What about $\ln Z(\boldsymbol{\theta})$?

Deriving $l(\boldsymbol{\theta} : \mathcal{D})$

$$Z(\theta) = \sum_{\xi} \exp\left\{\sum_{i} \theta_i \cdot f_i(\xi)\right\}$$

$$\frac{\partial}{\partial \theta_i} \ln Z(\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \cdot \sum_{\xi} \frac{\partial}{\partial \theta_i} \exp\left\{\sum_{i} \theta_i \cdot f_i(\xi)\right\}$$

$$= \frac{1}{Z(\boldsymbol{\theta})} \cdot \sum_{\xi} f_i(\xi) \cdot \exp\left\{\sum_{i} \theta_i \cdot f_i(\xi)\right\}$$

$$= \mathbb{E}_{P(\mathcal{X}:\boldsymbol{\theta})} f_i \stackrel{\text{def}}{=} \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

Deriving $l(\boldsymbol{\theta} : \mathcal{D})$, cont.

$$Z(\theta) = \sum_{\xi} \exp\left\{\sum_{i} \theta_i \cdot f_i(\xi)\right\}$$

$$\frac{\partial^2}{\partial\theta_i\theta_j} \ln Z(\boldsymbol{\theta})$$

$$= \frac{\partial}{\partial\theta_j}\left[\frac{1}{Z(\boldsymbol{\theta})} \cdot \sum_{\xi} f_i(\xi) \cdot \exp\left\{\sum_{i} \theta_i \cdot f_i(\xi)\right\}\right]$$

$$= \mathbb{E}_{\boldsymbol{\theta}}[f_i \cdot f_j] - \mathbb{E}_{\boldsymbol{\theta}}[f_i] \cdot \mathbb{E}_{\boldsymbol{\theta}}[f_j] \stackrel{\text{def}}{=} \mathbb{C}ov_{\boldsymbol{\theta}}[f_i, f_j]$$

Log likelihood, cont.

- We have derived the Hessian of $\ln Z(\boldsymbol{\theta})$ has elements of the form $\frac{\partial^2}{\partial \theta_i \theta_j} \ln Z(\boldsymbol{\theta}) = \mathbb{C}ov_{\boldsymbol{\theta}}[f_i, f_j]$
  - A covariance matrix is always positive semidefinite
  - Therefore, $\ln Z(\boldsymbol{\theta})$ is convex

Log likelihood, cont.

$$\frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \sum_i \theta_i \cdot \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) - \ln Z(\boldsymbol{\theta})$$

- The log-likelihood function is a sum of a linear element and a concave element → it is **concave**
- Therefore, it has no local optima, only a global maximum (albeit a non-unique one)

Max likelihood

$$\frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \sum_i \theta_i \cdot \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) - \ln Z(\boldsymbol{\theta})$$

- Apparently, what remains is to compute the gradient and find its zeroes

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- Meaning the maximum is attained when $\forall i \; \mathbb{E}_{\mathcal{D}} f_i(\boldsymbol{d}_i) = \mathbb{E}_{\boldsymbol{\theta}}[f_i]$

Max likelihood, cont.

Theorem 20.1

Let $\mathcal{F}$ be a set of features. Then, $\boldsymbol{\theta}$ is a maximum-likelihood parameter assignment if and only if $\boldsymbol{E}_{\mathcal{D}}[f_i(\mathcal{X})] = \boldsymbol{E}_{\hat{\boldsymbol{\theta}}}[f_i]$ for all i.

Max likelihood, cont.

- Good news:
  - The maximum is unique
  - It is attained when the empirical expectancy of all features matches their a priori expectancy
- Bad news: impossible to compute analytically
  - Gradient ascent to the rescue

Max likelihood, cont.

- Gradient ascent to the rescue
  - Guaranteed to converge to the maximum

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[f_i] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- How to compute?

Max likelihood – Gradient ascent

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[f_i] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- The first element, for indicator features, is the empirical frequency of the relevant event

Max likelihood – Gradient ascent

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[f_i] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- The second element, for indicator features, is of the form $P_{\boldsymbol{\theta}}(a^0, b^0)$
  - Computed using inference on the graph
  - Single inference pass yields all probabilities

Max likelihood – Gradient ascent

$$\frac{\partial}{\partial \theta_i} \frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \mathbb{E}_{\mathcal{D}}[f_i] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- The second element, for indicator features, is of the form $P_{\boldsymbol{\theta}}(a^0, b^0)$
  - Computed using inference on the graph
  - But such computation is to be performed afresh for <u>each</u> step of the ascent!

Max likelihood = Max Entropy

- Idea: we want to re-construct the distribution corresponding to the given empirical expectations of the features
- We demand maximal entropy as a sign that we add no other constraints (no more information)

Max likelihood = Max Entropy

- Formally:
  - Find $Q(\mathcal{X})$ maximizing $\mathbb{H}_Q(\mathcal{X})$
  - Subject to $\forall i, \mathbb{E}_Q[f_i] = \mathbb{E}_{\mathcal{D}}[f_i]$

- Turns out that…

Theorem 20.2

The distribution Q^* is the maximum entropy distribution satisfying equation (20.10) if and only if $Q^* = P_{\hat{\boldsymbol{\theta}}}$, where

$$P_{\hat{\boldsymbol{\theta}}}(\mathcal{X}) = \frac{1}{Z(\hat{\boldsymbol{\theta}})} \exp\left\{\sum_i \hat{\theta}_i f_i(\mathcal{X})\right\}$$

and $\hat{\boldsymbol{\theta}}$ is the maximum likelihood parameterization relative to $\mathcal{D}$.

Max Likelihood = Max Entropy

- Formally:
  - Find $Q(\mathcal{X})$ maximizing $\mathbb{H}_Q(\mathcal{X})$
  - Subject to $\forall i, \mathbb{E}_Q[f_i] = \mathbb{E}_{\mathcal{D}}[f_i]$

- The distribution given by the MLE clearly satisfies the constraints
  - Now we need to prove it maximizes the entropy
  - Point of notation: we denote $P \stackrel{\text{def}}{=} P_{\widehat{\boldsymbol{\theta}}}$

Max Likelihood = Max Entropy

- Let Q be another distribution adhering to the same constraints. Then:

$$\mathbb{E}_P \ln P = \mathbb{E}_P \left(\sum_i \hat{\theta}_i \cdot f_i - \ln Z(\hat{\boldsymbol{\theta}}) \right) = \sum_i \hat{\theta}_i \cdot \mathbb{E}_P[f_i] - \ln Z(\hat{\boldsymbol{\theta}})$$

$$= \sum_i \hat{\theta}_i \cdot \mathbb{E}_Q[f_i] - \ln Z(\hat{\boldsymbol{\theta}}) = \mathbb{E}_P \left(\sum_i \hat{\theta}_i \cdot f_i - \ln Z(\hat{\boldsymbol{\theta}}) \right) = \mathbb{E}_Q \ln P$$

Max Likelihood = Max Entropy

- Let Q be another distribution adhering to the same constraints. Then:

$$\mathbb{H}_P - \mathbb{H}_Q = -\mathbb{E}_P \ln P + \mathbb{E}_Q \ln Q$$
$$= -\mathbb{E}_Q \ln P + \mathbb{E}_Q \ln Q = D(Q \parallel P) \geq 0$$

- Meaning $\forall Q \neq P, \mathbb{H}_P > \mathbb{H}_Q$, Q.E.D.

MLE Prior

- Reminder: MLE by itself is prone to overfitting
- Therefore, a prior distribution is taken in order to bias the solution toward a prior model

MLE Prior

- Two priors:
  - Gaussian (L_2): $P(\boldsymbol{\theta}) = \prod_i \frac{1}{\sqrt{2\pi}\sigma_i} \cdot \exp\left(-\frac{\theta_i^2}{2\sigma_i^2}\right)$
  - Laplacian (L_1): $P(\boldsymbol{\theta}) = \prod_i \frac{1}{2\beta_i} \cdot \exp\left(-\frac{|\theta_i|}{\beta_i}\right)$

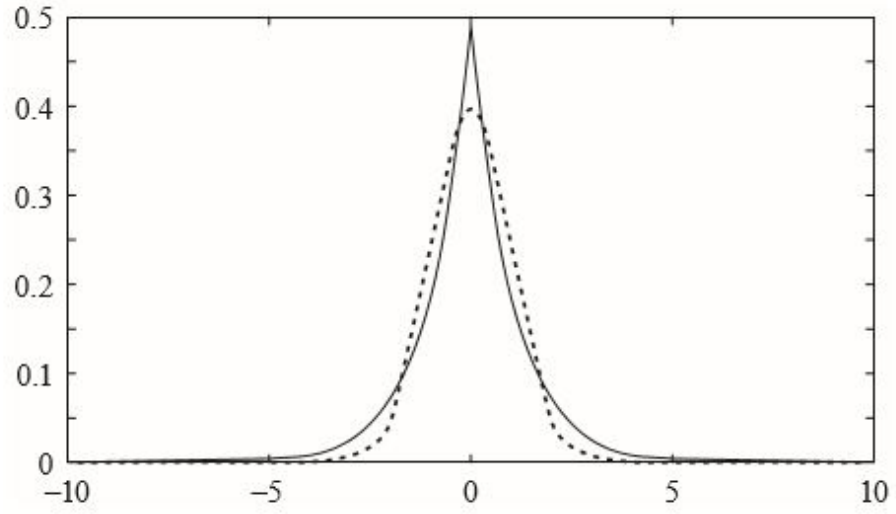

Figure 20.3 Laplacian distribution ($\beta = 1$) and Gaussian distribution ($\sigma^2 = 1$)

MLE Prior

- Idea: both priors penalize too large θ_i
  - Since we do not want to assume too much dependence on a single feature
- Gaussian Prior:
  - penalizes large values more
  - but there is no incentive to get to 0
- Laplacian Prior: exactly the opposite
  - Resulting in more sparse constructions

MLE Prior

- Note: in log form, both priors are concave
- Therefore, they can be added to $l(\boldsymbol{\theta} : \mathcal{D})$ with no need to change the algorithm

- The parameters regulating the width of the prior(s) reflect how important it is for us to drive them to 0
  - Method for selection – Cross-Validation: select values, run on part of the data, check vs. the remaining data

MLE Conjugate Prior

$$P(\boldsymbol{\theta}|\mathcal{D}) \propto P(\boldsymbol{\theta}) \cdot P(\mathcal{D}|\boldsymbol{\theta})$$
$$= P(\boldsymbol{\theta})$$
$$\cdot \exp\left[\sum_i M \cdot \mathbb{E}_{\mathcal{D}}[f_i] \cdot \theta_i - M \cdot \ln Z(\boldsymbol{\theta})\right]$$

- In order for the posterior to be of the same form, the prior has to be

$$P(\boldsymbol{\theta}) \propto \exp\left[\sum_i M_0 \cdot \alpha_i \cdot \theta_i - M_0 \cdot \ln Z(\boldsymbol{\theta})\right]$$

  - Which might be construed as a prior data of size M_0 yielding α_i as the expected value per feature f_i

Missing Data

“All the business of war… is to endeavour to find out what you don't know by what you do…”

The Duke of Wellington

MLE with missing data

- As previously, let us denote:
  - **o**[m]: vector of **o**bserved values in the mth sample
  - **h**[m]: vector of **h**idden values in the mth sample

- The log likelihood then becomes:

$$\frac{1}{M} \cdot l(\boldsymbol{\theta} : \mathcal{D}) = \frac{1}{M} \sum_{m} \ln \left(\sum_{\boldsymbol{h}[m]} P(\boldsymbol{o}[m], \boldsymbol{h}[m] | \boldsymbol{\theta}) \right)$$

$$= \frac{1}{M} \sum_{m} \ln \left(\sum_{\boldsymbol{h}[m]} \tilde{P}(\boldsymbol{o}[m], \boldsymbol{h}[m] | \boldsymbol{\theta}) \right) - \ln Z(\boldsymbol{\theta})$$

MLE with missing data, cont.

- Let us have a closer look at the term $\sum_{\boldsymbol{h}[m]} \tilde{P}(\boldsymbol{o}[m], \boldsymbol{h}[m]|\boldsymbol{\theta})$

- It has the form of a partition function
  - It is the partition function on the reduction of the original network by the observation $\boldsymbol{o}[m]$
  - and thus adheres to the derivation of $\ln Z(\boldsymbol{\theta})$ presented previously

MLE with missing data, cont.

- The derivand is a partition function, and thus:

$$\frac{\partial}{\partial \theta_i} \ln \sum_{\boldsymbol{h}[m]} \tilde{P}(\boldsymbol{h}[m], \boldsymbol{o}[m] \mid \boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{h}[m] \sim P(\mathcal{H}[m] \mid \boldsymbol{o}[m], \boldsymbol{\theta})}[f_i]$$

- Leading to the conclusion that the gradient of the log-likelihood is

$$\frac{\partial}{\partial \theta_i} \left[\frac{1}{M} l(\boldsymbol{\theta} : \mathcal{D}) \right]$$

$$= \frac{1}{M} \left[\sum_m \mathbb{E}_{\boldsymbol{h}[m] \sim P(\mathcal{H}[m] \mid \boldsymbol{o}[m], \boldsymbol{\theta})}[f_i] \right] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

MLE with missing data, cont.

$$\frac{\partial}{\partial \theta_i}\left[\frac{1}{M} l(\boldsymbol{\theta} : \mathcal{D})\right] = \frac{1}{M}\left[\sum_m \mathbb{E}_{\boldsymbol{h}[m] \sim P(\mathcal{H}[m] | \boldsymbol{o}[m], \boldsymbol{\theta})}[f_i]\right] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- The second term, as before, requires an inference computation
- The first term requires a computation of inference on the reduced network
  - per instance of $\boldsymbol{o}[m]$
  - for a single iteration of the ascent…

Expectation Maximization

- EM: an alternative approach
  - Efficient for BNs (previous lecture)

- Main idea: bootstrapping
  - Start with some initial $\boldsymbol{\theta}^0$
  - Compute corresponding distribution for missing data $\mathcal{H}$
  - Based on the full data $< \mathcal{D}, \mathcal{H} >$, compute a new $\boldsymbol{\theta}^1$
  - etc until convergence

Expectation Maximization, cont.

- For BNs:
  - Assess probabilities for each $\overline{M_t}(x_i, \boldsymbol{u}_i)$
  - Based on these probabilities, compute:

$$\boldsymbol{\theta}_{x_i|\boldsymbol{u}_i}^{t+1} = \frac{\overline{M_t}(x_i, \boldsymbol{u}_i)}{\overline{M_t}(\boldsymbol{u}_i)}$$

Expectation Maximization, cont.

- For MNs:
  - Assess probabilities for each feature (E-step):

$$\overline{M_t}[f_i] = \frac{1}{M}\sum_m \mathbb{E}_{\boldsymbol{h}[m]\sim P(\mathcal{H}[m]|\boldsymbol{o}[m],\boldsymbol{\theta}_t)}[f_i]$$

  - Done using inference per each $\boldsymbol{o}[m]$ ☹
  - Compute next $\boldsymbol{\theta}$ based on that… how?

Expectation Maximization, cont.

- Computing an optimal $\boldsymbol{\theta}$ from a set of full data in MN is done using gradient ascent
  - involving running inference… multiple times
  - for a single iteration of the EM

Missing Data: GA vs. EM

- In both methods, we need inference
  - In GA: M+1 times per each step
  - In EM: (M times + 1 time per step of the GA) per step of the algorithm

Alternative Objectives

“A clever man goes not over a mountain, but rather around it.”

A Russian proverb

Log likelihood revisited

- Reminder: $l(\boldsymbol{\theta} : \xi) = \ln \tilde{P}(\xi|\boldsymbol{\theta}) - \ln Z(\boldsymbol{\theta})$
- Interpretation:
  - We strive to increase the difference between the log-measures of the data and the aggregate of all instances
  - Problem: the aggregate (second term, $Z(\boldsymbol{\theta})$) is exponential
- Idea: define a simpler objective

Pseudolikelihood

- Consider the probability of a single instance

$$P(\xi) = P(X_1 = x_1, \dots X_n = x_n)$$
$$= P(x_n | x_1, \dots x_{n-1}) \cdot P(x_1, \dots x_{n-1})$$

$$P(\xi) = \prod_{j=1}^{n} P\left(x_j | x_1, \dots x_{j-1}\right) \approx \prod_{j=1}^{n} P\left(x_j | \boldsymbol{x}_{-j}\right)$$

Pseudolikelihood

$$P(\xi) = \prod_{j=1}^{n} P\left(x_j | x_1, \ldots x_{j-1}\right) \approx \prod_{j=1}^{n} P\left(x_j | \boldsymbol{x}_{-j}\right)$$

- From here we can derive the ***pseudo-likelihood***:

$$l_{PL}(\theta : \mathcal{D}) = \frac{1}{M} \sum_{m} \sum_{j} \ln P\left(x_j[m] | \boldsymbol{x}_{-j}[m], \boldsymbol{\theta}\right)$$

Pseudolikelihood

$$l_{PL}(\theta : \mathcal{D}) = \frac{1}{M}\sum_{m}\sum_{j} \ln P\left(x_j[m] | \boldsymbol{x}_{-j}[m], \boldsymbol{\theta}\right)$$

- The main advantage: it is easier to compute, as there is less coupling and summation
  - Since $P\left(x_j | \boldsymbol{x}_{-j}\right) = \frac{P\left(j, \boldsymbol{x}_{-j}\right)}{P\left(\boldsymbol{x}_{-j}\right)} = \frac{\tilde{P}\left(x_j, \boldsymbol{x}_{-j}\right)}{\tilde{P}\left(\boldsymbol{x}_{-j}\right)} = \frac{\tilde{P}\left(x_j, \boldsymbol{x}_{-j}\right)}{\sum_{x'_j} \tilde{P}\left(x'_j, \boldsymbol{x}_{-j}\right)}$
  - The number of elements to sum is $M \cdot \sum_j \left|Val\left(X_j\right)\right|$, which is clearly sub-exponential

Pseudolikelihood

- Let us now further analyze the summands:

$$\ln P\left(x_j | \boldsymbol{x}_{-j}\right) = \ln \tilde{P}\left(x_j, \boldsymbol{x}_{-j}\right) - \ln \sum_{x'_j} \tilde{P}\left(x'_j, \boldsymbol{x}_{-j}\right)$$

$$= \sum_{i : X_j \in Scope[f_i]} \theta_i \cdot f_i\left(x_j, \boldsymbol{x}_{-j}\right)$$

$$- \ln \sum_{x'_j} \exp\left(\sum_{i : X_j \in Scope[f_i]} \theta_i \cdot f_i\left(x'_j, \boldsymbol{x}_{-j}\right) \right)$$

Pseudolikelihood

$$\ln P(x_j|\boldsymbol{x}_{-j}) = \sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x_j, \boldsymbol{x}_{-j}) - \ln \sum_{x'_j} \exp\left(\sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x'_j, \boldsymbol{x}_{-j}) \right)$$

- The expression above is the log-likelihood of a MN over X_j conditioned on the rest
  - Meaning it is concave…
  - And the pseudo-likelihood, being the sum of such terms, is concave as well!
- Gradient Ascent again…

Pseudolikelihood

$$\ln P(x_j|\boldsymbol{x}_{-j}) = \sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x_j, \boldsymbol{x}_{-j}) - \ln \sum_{x'_j} \exp\left(\sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x'_j, \boldsymbol{x}_{-j}) \right)$$

$$\begin{aligned}
&\frac{\partial}{\partial \theta_i} \ln P(x_j|\boldsymbol{x}_{-j}) \\
&\quad = f_i(x_j, \boldsymbol{x}_{-j}) - \frac{\sum_{x'_j} f_i(x'_j, \boldsymbol{x}_{-j}) \cdot \exp\left(\sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x'_j, \boldsymbol{x}_{-j}) \right)}{\sum_{x'_j} \exp\left(\sum_{i:X_j \in \text{Scope}[f_i]} \theta_i \cdot f_i(x'_j, \boldsymbol{x}_{-j}) \right)} \\
&\quad = f_i(x_j, \boldsymbol{x}_{-j}) - \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j|\boldsymbol{x}_{-j})}\left[f_i(x'_j, \boldsymbol{x}_{-j}) \right]
\end{aligned}$$

- Note: when $X_j \notin \text{Scope}[f_i]$, it does not affect the value of the feature, and so the expression becomes 0

Pseudolikelihood

$$\frac{\partial}{\partial \theta_i} \ln P(x_j | \boldsymbol{x}_{-j}) = f_i(x_j, \boldsymbol{x}_{-j}) - \mathbb{E}_{x'_j \sim P_\theta(X_j | \boldsymbol{x}_{-j})}\left[f_i(x'_j, \boldsymbol{x}_{-j})\right]$$

$$\frac{\partial}{\partial \theta_i} l_{PL}(\boldsymbol{\theta} : \mathcal{D}) = \sum_{j : X_j \in \text{Scope}(f_i)} \left[\frac{1}{M} \sum_m f_i(\xi[m]) - \mathbb{E}_{x'_j \sim P_\theta(X_j | \boldsymbol{x}_{-j}[m])}\left[f_i(x'_j, \boldsymbol{x}_{-j}[m])\right] \right]$$

- The computation is much simpler
  - All expectations (summations) are local
  - Finding the maximum of the PL *is* tractable

- But… does it do us any good?

Pseudolikelihood

Theorem (20.3):

Assume the data is generated by a log-linear model $P_{\boldsymbol{\theta}^*}$ of the form described previously.

Then, as M goes to infinity, $\boldsymbol{\theta}^*$ is the argument for the PL global optimum, with probability approaching 1.

Idea of proof:

Show the gradient is 0 at $\boldsymbol{\theta}^*$ (why is that sufficient?)

Pseudolikelihood

$$\frac{\partial}{\partial \theta_i} l_{PL}(\boldsymbol{\theta} : \mathcal{D})$$

$$= \sum_{j : X_j \in \text{Scope}(f_i)} \left[\frac{1}{M} \sum_m f_i(\xi[m]) - \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right] \right]$$

- The first term is the empirical expectancy $\mathbb{E}_{\mathcal{D}}[f_i]$, which as $M \to \infty$, goes to $\mathbb{E}_{P_{\boldsymbol{\theta}^*}}[f_i]$

Pseudolikelihood

$$\frac{\partial}{\partial \theta_i} l_{PL}(\boldsymbol{\theta} : \mathcal{D}) = \sum_{j : X_j \in \text{Scope}(f_i)} \left[\frac{1}{M} \sum_m f_i(\xi[m]) - \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right] \right]$$

- The second term is:

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right]$$

$$= \frac{1}{M} \sum_{m=1}^{M} \sum_{x'_j} P_{\boldsymbol{\theta}}\left(x'_j \middle| \boldsymbol{x}_{-j}[m]\right) \cdot f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right)$$

$$= \sum_{\boldsymbol{x}_{-j}} P_{\mathcal{D}}\left(\boldsymbol{x}_{-j}\right) \cdot \sum_{x'_j} P_{\boldsymbol{\theta}}\left(x'_j \middle| \boldsymbol{x}_{-j}\right) \cdot f_i\left(x'_j, \boldsymbol{x}_{-j}\right)$$

Pseudolikelihood

$$\frac{\partial}{\partial \theta_i} l_{PL}(\boldsymbol{\theta} : \mathcal{D}) = \sum_{j : X_j \in \text{Scope}(f_i)} \left[\frac{1}{M} \sum_m f_i(\xi[m]) - \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right] \right]$$

$$\frac{1}{M} \sum_{m=1}^{M} \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right]$$

$$= \sum_{\boldsymbol{x}_{-j}} P_{\mathcal{D}}\left(\boldsymbol{x}_{-j}\right) \cdot \sum_{x'_j} P_{\boldsymbol{\theta}}\left(x'_j \middle| \boldsymbol{x}_{-j}\right) \cdot f_i\left(x'_j, \boldsymbol{x}_{-j}\right)$$

$$\xrightarrow[\boldsymbol{\theta} = \boldsymbol{\theta}^*]{M \to \infty} \sum_{\boldsymbol{x}_{-j}} P_{\boldsymbol{\theta}^*}\left(\boldsymbol{x}_{-j}\right) \cdot \sum_{x'_j} P_{\boldsymbol{\theta}^*}\left(x'_j \middle| \boldsymbol{x}_{-j}\right) \cdot f_i\left(x'_j, \boldsymbol{x}_{-j}\right) = \mathbb{E}_{P_{\boldsymbol{\theta}^*}}[f_i]$$

Pseudolikelihood

$$\frac{\partial}{\partial \theta_i} l_{PL}(\boldsymbol{\theta} : \mathcal{D})$$

$$= \sum_{j : X_j \in \text{Scope}(f_i)} \left[\frac{1}{M} \sum_m f_i(\xi[m]) - \mathbb{E}_{x'_j \sim P_{\boldsymbol{\theta}}(X_j | \boldsymbol{x}_{-j}[m])} \left[f_i\left(x'_j, \boldsymbol{x}_{-j}[m]\right) \right] \right]$$

- The first term, as $M \to \infty$, goes to $\mathbb{E}_{P_{\boldsymbol{\theta}^*}}[f_i]$
- The second term, as $M \to \infty$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, goes to $\mathbb{E}_{P_{\boldsymbol{\theta}^*}}[f_i]$
- Ergo: as $M \to \infty$, the gradient at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is 0, QED

Pseudolikelihood, concluded

- An alternative objective to the MLE
  - Tractable
  - The same as MLE as the data size increases

- But! A large data sample is required for the PL to reflect on the real MLE

Contrasting Divergence

- Main idea: increase difference between the log-probability of the observed data and some other value, representing “the world”
  - Global Partition Function (MLE)
  - Single-Variable Partition Function (PL)
  - Log-Probability of Perturbed Data (CD)

Contrasting Divergence

- CD is about the difference between the original data set and a perturbed data set

- Formally:

$$l_{CD}(\theta : \mathcal{D} \parallel \mathcal{D}^{-}) = \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}}}[\ln \tilde{P}_{\boldsymbol{\theta}}(\xi)] - \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}^{-}}}[\ln \tilde{P}_{\boldsymbol{\theta}}(\xi)]$$

  - The difference between the empirical expectations on the log-probability

Contrasting Divergence

- CD is about the difference between the original data set and a perturbed data set
  - How to choose this data set?

- The contrasting data set $\mathcal{D}^-$ needs to represent a data sample characteristic of the current $\boldsymbol{\theta}$
  - So that we strive to increase the probability of the original sampled data $\mathcal{D}$ relative to the current result, which serves as the contrast in this iterative process

Contrasting Divergence

- The contrasting data set $\mathcal{D}^-$ needs to represent a data sample characteristic of the current $\boldsymbol{\theta}$
  - How?

- Gibbs Sampling starting from $\mathcal{D}$
  - "Long" sampling until convergence too expensive
  - Short chain is good enough and yields better convergence

Contrasting Divergence

- How do we compute the optimum?
  - Gradient Ascent yet again

$$l_{CD}(\boldsymbol{\theta} : \mathcal{D} \parallel \mathcal{D}^{-}) = \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}}}[\ln \tilde{P}_{\boldsymbol{\theta}}(\xi)] - \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}^{-}}}[\ln \tilde{P}_{\boldsymbol{\theta}}(\xi)]$$

$$= \sum_{\xi} \left[\left(\hat{P}_{\mathcal{D}}(\xi) - \hat{P}_{\mathcal{D}^{-}}(\xi) \right) \cdot \sum_{i} \theta_i f_i(\xi) \right]$$

$$\frac{\partial}{\partial \theta_i} l_{CD} = \sum_{\xi} \left[\left(\hat{P}_{\mathcal{D}}(\xi) - \hat{P}_{\mathcal{D}^{-}}(\xi) \right) \cdot f_i(\xi) \right]$$

$$= \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}}}[f_i(\xi)] - \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}^{-}}}[f_i(\xi)]$$

Contrasting Divergence

$$\frac{\partial}{\partial \theta_i} l_{CD}(\boldsymbol{\theta} : \mathcal{D} \parallel \mathcal{D}^{-})$$

$$= \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}}}[f_i(\xi)] - \mathbb{E}_{\xi \sim \hat{P}_{\mathcal{D}^{-}}}[f_i(\xi)]$$

$$\xrightarrow[\mathcal{D}^{-} \to \boldsymbol{\theta}]{} \mathbb{E}_{\mathcal{D}}[f_i] - \mathbb{E}_{\boldsymbol{\theta}}[f_i]$$

- Note: as $\mathcal{D}^{-} \to \boldsymbol{\theta}$, the elements of the gradient converge to the gradient of the max log likelihood
  - At the limit of the Markov chain, the CD converges to the actual MLE

Checkpoint

“Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.”
Winston Churchill

Checkpoint

- Maximum likelihood for MN
  - Not easily decomposable due to the Partition Function
  - Soluble using Gradient Ascent… but requires running inference on the MN for each step
- Priors:
  - Gaussian vs. Laplacian
    - Both aim to reduce too strong dependency of overall probability on single feature
  - Conjugate Prior

Checkpoint

- MLE with missing data
  - GA: requires running inference per observation per step
  - EM: reduces to Gradient Ascent requiring slightly less runs of inference
- Alternative goals
  - Pseudo-likelihood: tractable, requires sufficiently large data sample
  - Contrasting Divergence: tractable, does not require much sampling