PGM lecture notes: pseudo-likelihood

Amir Globerson (modified by David Sontag)

Consider a pairwise Markov random field and data $\{\boldsymbol{x}^{(m)}\}_{m=1\ldots M}$:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} e^{\sum_{ij}\theta_{ij}(x_i,x_j)} \tag{1}$$

As we have seen earlier, the gradient w.r.t. $\boldsymbol{\theta}$ of the likelihood is $p_D(x_i,x_j) - p(x_i,x_j;\boldsymbol{\theta})$. When inference in the model is hard (e.g., because of high tree width) it is not feasible to calculate the gradient, and likelihood maximization is typically intractable. One way to address this is to use approximate inference methods to approximate the gradient (e.g., loopy BP, sampling etc).

The pseudo-likelihood method (Besag 1971) offers a different approach to this problem, which surprisingly yields an exact solution if the data is generated by a model $p(\boldsymbol{x};\boldsymbol{\theta}^*)$ and $n \to \infty$ (i.e., it is consistent).

The goal is to replace the likelihood by a more tractable objective. To do this, we note that:

$$p(\boldsymbol{x};\boldsymbol{\theta}) = \prod_i p(x_i|x_1,\ldots,x_{i-1}) \tag{2}$$

via the chain rule. We consider the following approximation:

$$p(\boldsymbol{x};\boldsymbol{\theta}) \approx \prod_i p(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n;\boldsymbol{\theta}) = \prod_i p(x_i|x_{-i};\boldsymbol{\theta}) \tag{3}$$

where we have added conditioning over additional variables. In an undirected model, the above has a particularly simple form:

$$\begin{aligned} p(x_i|x_{-i};\boldsymbol{\theta}) &= p(x_i|x_{N(i)};\boldsymbol{\theta}) = \frac{p(x_i,x_{-i};\boldsymbol{\theta})}{\sum_{\hat{x}_i} p(\hat{x}_i,x_{-i};\boldsymbol{\theta})} \\ &= \frac{e^{\sum_{j\in N(i)}\theta(x_j,x_i)}}{\sum_{\hat{x}_i} e^{\sum_{j\in N(i)}\theta(x_j,\hat{x}_i)}} \\ &= \frac{1}{Z(x_{N(i)};\boldsymbol{\theta})} e^{\sum_{j\in N(i)}\theta(x_j,x_i)} \end{aligned}$$

So we have the approximation:

$$p(\boldsymbol{x};\boldsymbol{\theta}) \approx \prod_i p(x_i|x_{N(i)};\boldsymbol{\theta}) \tag{4}$$

or:

$$\log p(\boldsymbol{x};\boldsymbol{\theta}) \approx \sum_i \log p(x_i|x_{N(i)};\boldsymbol{\theta}) \tag{5}$$

The pseudolikelihood is defined as the following function of $\boldsymbol{\theta}$:

$$\ell_{\text{PL}}(\boldsymbol{\theta}) = \frac{1}{M}\sum_m \left[\sum_{i=1}^n \log p(x_i^{(m)}|\boldsymbol{x}_{N(i)}^{(m)};\boldsymbol{\theta})\right] \tag{6}$$

Where:

$$p(\boldsymbol{x}_i^{(m)}|\boldsymbol{x}_{N(i)}^{(m)};\boldsymbol{\theta}) = \frac{e^{\sum_{j\in N(i)}\theta(\boldsymbol{x}_j^{(m)},\boldsymbol{x}_i^{(m)})}}{\sum_{\hat{x}_i} e^{\sum_{j\in N(i)}\theta(\boldsymbol{x}_j^{(m)},\hat{x}_i)}} \tag{7}$$

What this implies is that we want the expression in the numerator to obtain its highest value when we set it to the observed value of $\boldsymbol{x}_i^{(m)}$. This is a so-called *contrastive* that seeks to contrast between the values we observed and the other values (in standard likelihood maximization we contrast between $\boldsymbol{x}^{(m)}$ and all other values.

Expanding $\ell_{\text{PL}}(\boldsymbol{\theta})$ we obtain:

$$\begin{aligned}
\ell_{\text{PL}}(\boldsymbol{\theta}) &= \frac{1}{M}\sum_m\sum_i\sum_{j\in N(i)}\theta(\boldsymbol{x}_j^{(m)},\boldsymbol{x}_i^{(m)}) - \frac{1}{M}\sum_m\sum_i \log Z(x_{N(i)};\boldsymbol{\theta}) \\
&= \frac{1}{M}\sum_i\sum_{j\in N(i)}\sum_m\theta(\boldsymbol{x}_j^{(m)},\boldsymbol{x}_i^{(m)}) - \frac{1}{M}\sum_i\sum_m \log Z(x_{N(i)};\boldsymbol{\theta}) \\
&= \sum_i\sum_{j\in N(i)} p_D(x_i,x_j)\theta(x_i,x_j) - \sum_i\sum_{x_{N(i)}} p_D(x_{N(i)})\log Z(x_{N(i)};\boldsymbol{\theta})
\end{aligned}$$

Some good things about this function:

- It only involves summation over x_i and is thus tractable
- It is concave in $\boldsymbol{\theta}$ and hence has no local minima. We shall furthermore focus on cases in which it is strictly concave (i.e., it has a single maximizer).

Next we will show the following important property: Assume the data is generated IID by a distribution $p(\boldsymbol{x};\boldsymbol{\theta}^*)$ for some $\boldsymbol{\theta}^*$. Then as $n\to\infty$ we have $\ell_{\text{PL}}(\boldsymbol{\theta})\to\ell_{\text{PL}}^{\infty}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}^*$ maximizes $\ell_{\text{PL}}^{\infty}(\boldsymbol{\theta})$. In other words as $n\to\infty$ the true parameter will be the solution to maximizing the pseudolikelihood function.

To take derivatives we need the following property (for $j\in N(i)$)

$$\frac{\partial}{\partial\theta_{ij}(x_i,x_j)}\log Z(x_{N(i)};\boldsymbol{\theta}) = \frac{1}{Z(x_{N(i)};\boldsymbol{\theta})}e^{\sum_{k\in N(i)}\theta_{ki}(x_k,x_i)} = p(x_i|x_{x_{N(i)}};\boldsymbol{\theta}) \tag{8}$$

Take the derivative of $\ell(\boldsymbol{\theta})$ w.r.t. $\theta_{ij}(x_i,x_j)$ to obtain:

$$\frac{\partial\ell_{\text{PL}}}{\partial\theta_{ij}(x_i,x_j)} = 2p_D(x_i,x_j) - \sum_{x_{N(i)}} p_D(x_{N(i)})p(x_i|x_{x_{N(i)}};\boldsymbol{\theta}) - \sum_{x_{N(j)}} p_D(x_{N(j)})p(x_j|x_{N(j)};\boldsymbol{\theta}) \tag{9}$$

(the factor of 2 is because $\theta_{ij}(x_i,x_j)$ shows up in two terms, one for node i and another for node j.)

As $n\to\infty$ we have $p_D(\boldsymbol{x})\to p(\boldsymbol{x};\boldsymbol{\theta}^*)$. So the gradient becomes:

$$2p(x_i,x_j;\boldsymbol{\theta}^*) - \sum_{x_{N(i)\setminus j}} p(x_{N(i)};\boldsymbol{\theta}^*)p(x_i|x_{N(i)};\boldsymbol{\theta}) - \sum_{x_{N(j)\setminus i}} p(x_{N(j)};\boldsymbol{\theta}^*)p(x_j|x_{x_{N(j)}};\boldsymbol{\theta}) \tag{10}$$

We now want to argue that the above is zero when $\boldsymbol{\theta}=\boldsymbol{\theta}^*$. The second term becomes:

$$\sum_{x_{N(i)\setminus j}} p(x_{N(i)};\boldsymbol{\theta}^*)p(x_i|x_{N(i)};\boldsymbol{\theta}^*) = \sum_{x_{N(i)\setminus j}} p(x_i,x_{N(i)};\boldsymbol{\theta}^*) = p(x_i,x_j;\boldsymbol{\theta}^*) \tag{11}$$

and the same for the third term in Eq. 10 so that the gradient is indeed zero and ℓ_{PL} is (uniquely) maximized by the true parameter value. The uniqueness comes from ℓ_{PL} being *strictly* concave.

In terms of convergence time we need the marginals over the neighbors to converge to their true value. If there are many neighbors this could take a long time. The more problematic assumption of pseudolikelihood is that the data is generated by a distribution in the class, which is rarely the case.