Alternatives to MLE
Alternatives to MLE

• Exact MLE estimation is intractable
  • To compute the gradient of the log-likelihood, we need to compute marginals of the model

• Alternatives include
  • Pseudolikelihood approximation to the MLE problem that relies on computing only local probabilities
  • For structured prediction problems, we could avoid likelihoods entirely by minimizing a loss function that measures our prediction error
### Pseudolikelihood

- Consider a log-linear MRF \( p(x|\theta) = \frac{1}{Z(\theta)} \prod_c \exp(\theta, f_c(x_c)) \)

- By the chain rule, the joint distribution factorizes as

\[
p(x|\theta) = \prod_i p(x_i|x_1, \ldots, x_{i-1}, \theta)
\]

- This quantity can be approximated by conditioning on all of the other variables (called the pseudolikelihood)

\[
p(x|\theta) \approx \prod_i p(x_i|x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, \theta)
\]
Pseudolikelihood

- Using the independence relations from the MRF
  \[ p(x|\theta) \approx \prod_{i} p(x_i|x_{N(i)}, \theta) \]

- Only requires computing local probability distributions (typically much easier)
  - Does not require knowing \( Z(\theta) \)
  - Why not?
Pseudolikelihood

- For samples $x^1, \ldots, x^M$

\[
\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x^m_i | x^m_{N(i)}, \theta)
\]

- This approximation is called the pseudolikelihood

  - If the data is generated from a model of this form, then in the limit of infinite data, maximizing the pseudolikelihood recovers the true model parameters

  - Can be much more efficient to compute than the log likelihood
Pseudolikelihood

\[
\log \ell_{PL}(\theta) = \sum_{m} \sum_{i} \log p(x_i^m | x_{N(i)}^m, \theta)
\]

\[
= \sum_{m} \sum_{i} \log \frac{p(x_i^m, x_{N(i)}^m | \theta)}{\sum_{x_i'} p(x_i', x_{N(i)}^m | \theta)}
\]

\[
= \sum_{m} \sum_{i} \left[ \log p(x_i^m, x_{N(i)}^m | \theta) - \log \sum_{x_i'} p(x_i', x_{N(i)}^m | \theta) \right]
\]

\[
= \sum_{m} \sum_{i} \left[ \left( \theta, \sum_{c \supseteq i} f_C(x_c^m) \right) - \log \sum_{x_i'} \exp \left( \theta, \sum_{c \supseteq i} f_C(x_i', x_{C\setminus i}^m) \right) \right]
\]
Pseudolikelihood

\[
\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x_i^m | x_N(i), \theta)
\]

\[
= \sum_m \sum_i \log \frac{p(x_i^m, x_N(i) | \theta)}{\sum_{x_i'} p(x_i', x_N(i) | \theta)}
\]

\[
= \sum_m \sum_i \left[ \log p(x_i^m, x_N(i) | \theta) - \log \sum_{x_i'} p(x_i', x_N(i) | \theta) \right]
\]

\[
= \sum_m \sum_i \left[ \log \left\{ \theta, \sum_{C \ni i} f_C(x_C^m) \right\} - \log \sum_{x_i'} \exp \left\{ \theta, \sum_{C \ni i} f_C(x_i', x_C^m) \right\} \right]
\]

Only involves summing over \(x_i\)!
Pseudolikelihood

\[
\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x_i^m | x_N^m, \theta)
\]

\[
= \sum_m \sum_i \log \frac{p(x_i^m, x_N^m | \theta)}{\sum_{x_i'} p(x_i', x_N^m | \theta)}
\]

\[
= \sum_m \sum_i \left[ \log p(x_i^m, x_N^m | \theta) - \log \sum_{x_i'} p(x_i', x_N^m | \theta) \right]
\]

\[
= \sum_m \sum_i \left[ \theta, \sum_{C \ni i} f_C(x_C^m) \right] - \log \sum_{x_i'} \exp \left\{ \theta, \sum_{C \ni i} f_C(x_i', x_C^m) \right\}
\]

Concave in \( \theta \)!(proof?)
Consistency of Pseudolikelihood

- Pseudolikelihood is a consistent estimator
  - That is, in the limit of large data, it is exact if the true model belongs to the family of distributions being modeled

\[
\nabla_{\theta} \ell_{PL} = \sum_{m} \sum_{i} \left[ \sum_{C \supset i} f_{C}(x_{C}^{m}) - \frac{\sum_{x_{i}}' \exp\langle \theta, \sum_{C \supset i} f_{C}(x_{i}', x_{C \setminus i}^{m}) \rangle \sum_{C \supset i} f_{C}(x_{i}', x_{C \setminus i}^{m})}{\sum_{x_{i}}' \exp\langle \theta, \sum_{C \supset i} f_{C}(x_{i}', x_{C \setminus i}^{m}) \rangle} \right]
\]

\[
= \sum_{m} \sum_{i} \left[ \sum_{C \supset i} f_{C}(x_{C}^{m}) - \sum_{x_{i}}' p(x_{i}'|x_{N(i)}^{m}, \theta) \sum_{C \supset i} f_{C}(x_{i}', x_{C \setminus i}^{m}) \right]
\]

Can check that the gradient is zero in the limit of large data if \( \theta = \theta^* \)
Structured Prediction

- Suppose we have, \( p(x|y, \theta) = \frac{1}{Z(\theta, y)} \prod_C \exp(\langle \theta, f_C(x_C, y) \rangle) \)

- If goal is \( \arg\max_x p(x|y) \), then MLE may be overkill
  
  - We only care about classification error, not about learning the correct marginal distributions as well

- Recall that the classification error is simply the expected number of incorrect predictions made by the learned model on samples from the true distribution

- Instead of maximizing the likelihood, we could minimize the classification error over the training set
Structured Prediction

- For samples \((x^1, y^1), \ldots, (x^M, y^M)\), the (unnormalized) classification error is

\[
\sum_m 1\{x^m \in \text{argmax}_x p(x|y^m, \theta)\}
\]

- The classification error is zero when \(p(x^m|y^m, \theta) \geq p(x|y^m, \theta)\) for all \(x\) and \(m\) or equivalently

\[
\left\langle \theta, \sum_C f_C(x^m_C, y^m) \right\rangle \geq \left\langle \theta, \sum_C f_C(x_C, y^m) \right\rangle
\]
Structured Prediction

• In the exact case, this can be thought of as having a linear constraint for each possible $x$ and each $y^1, \ldots, y^M$

$$\left\langle \theta, \sum_C [f_C(x^m_C, y^m) - f_C(x_C, y^m)] \right\rangle \geq 0$$

• Any $\theta$ that simultaneously satisfies each of these constraints will guarantee that the classification error is zero

• As there are exponentially many constraints, finding such a $\theta$ (if one even exists) is still a challenging problem

• If such a $\theta$ exists, we say that the problem is separable
Structured Perceptron Algorithm

• In the separable case, a straightforward algorithm can be designed to for this task

• Choose an initial $\theta$

• Iterate until convergence
  
  • For each $m$

  • Choose $x' \in \text{argmax}_x p(x|y^m, \theta)$

  • Set $\theta = \theta + \sum_C [f_C(x^m_C, y^m) - f_C(x'_C, y^m)]$