



CS 6347

Lecture 14

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

- Consider a log-linear MRF $p(x|\theta) = \frac{1}{Z(\theta)} \prod_C \exp\langle \theta, f_c(x_c) \rangle$
- By the chain rule, the joint distribution factorizes as

$$p(x|\theta) = \prod_i p(x_i | x_1, \dots, x_{i-1}, \theta)$$

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

$$p(x|\theta) \approx \prod_i p(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n, \theta)$$

- 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?

- For samples x^1, \dots, x^M

$$\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x_i^m | x_{N(i)}^m, \theta)$$

- 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

$$\begin{aligned}\log \ell_{PL}(\theta) &= \sum_m \sum_i \log p(x_i^m | x_{N(i)}^m, \theta) \\ &= \sum_m \sum_i \log \frac{p(x_i^m, x_{N(i)}^m | \theta)}{\sum_{x'_i} p(x'_i, x_{N(i)}^m | \theta)} \\ &= \sum_m \sum_i \left[\log p(x_i^m, x_{N(i)}^m | \theta) - \log \sum_{x'_i} p(x'_i, x_{N(i)}^m | \theta) \right] \\ &= \sum_m \sum_i \left[\left\langle \theta, \sum_{C \ni i} f_C(x_C^m) \right\rangle - \log \sum_{x'_i} \exp \left\langle \theta, \sum_{C \ni i} f_C(x'_i, x_{C \setminus i}^m) \right\rangle \right]\end{aligned}$$

$$\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x_i^m | x_{N(i)}^m, \theta)$$

$$= \sum_m \sum_i \log \frac{p(x_i^m, x_{N(i)}^m | \theta)}{\sum_{x'_i} p(x'_i, x_{N(i)}^m | \theta)}$$

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

$$= \sum_m \sum_i \left[\left\langle \theta, \sum_{C \ni i} f_C(x_C^m) \right\rangle - \log \sum_{x'_i} \exp \left\langle \theta, \sum_{C \ni i} f_C(x'_i, x_{C \setminus i}^m) \right\rangle \right]$$

Only involves summing over x_i !

$$\log \ell_{PL}(\theta) = \sum_m \sum_i \log p(x_i^m | x_{N(i)}^m, \theta)$$

$$= \sum_m \sum_i \log \frac{p(x_i^m, x_{N(i)}^m | \theta)}{\sum_{x'_i} p(x'_i, x_{N(i)}^m | \theta)}$$

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

$$= \sum_m \sum_i \left[\left\langle \theta, \sum_{C \ni i} f_C(x_C^m) \right\rangle - \log \sum_{x'_i} \exp \left\langle \theta, \sum_{C \ni i} f_C(x'_i, x_{C \setminus i}^m) \right\rangle \right]$$

Concave in θ ! (*proof?*)

- 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

$$\begin{aligned}\nabla_{\theta} \ell_{PL} &= \sum_m \sum_i \left[\sum_{C \supset i} f_C(x_C^m) - \frac{\sum_{x'_i} \exp\langle \theta, \sum_{C \supset i} f_C(x'_i, x_{C \setminus i}^m) \rangle \sum_{C \supset i} f_C(x'_i, x_{C \setminus i}^m)}{\sum_{x'_i} \exp\langle \theta, \sum_{C \supset i} f_C(x'_i, x_{C \setminus i}^m) \rangle} \right] \\ &= \sum_m \sum_i \left[\sum_{C \supset i} f_C(x_C^m) - \sum_{x'_i} p(x'_i | x_{N(i)}^m, \theta) \sum_{C \supset i} f_C(x'_i, x_{C \setminus i}^m) \right]\end{aligned}$$

Can check that the gradient is zero in the limit of large data if $\theta = \theta^*$

- 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 $\operatorname{argmax}_x p(x|y)$, 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

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

$$\sum_m 1_{\{x^m \in \operatorname{argmax}_x p(x|y^m, \theta)\}}$$

- 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

$$\left\langle \theta, \sum_c f_c(x_c^m, y^m) \right\rangle \geq \left\langle \theta, \sum_c f_c(x_c, y^m) \right\rangle$$

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

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

- Any θ that simultaneously satisfies each of these constraints will guarantee that the classification error is zero
 - As there are exponentially many constraints, finding such a θ (if one even exists) is still a challenging problem
 - If such a θ 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 θ
- Iterate until convergence
 - For each m
 - Choose $x' \in \operatorname{argmax}_x p(x|y^m, \theta)$
 - Set $\theta = \theta + \sum_c [f_c(x_c^m, y^m) - f_c(x'_c, y^m)]$