208891 Probabilistic Graphical Models **Spring 2019**

Lecture 16: February 28

Lecturer: Donlapark Pornnopparath

16.1 Parameter estimation of Markov random fields

We focus on the MRF with the *log-linear* model:

$$p(x|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp\{\boldsymbol{\theta}^T f(x)\}.$$

Given data D, we want to estimate $\boldsymbol{\theta}$ by maximizing the log-likelihood function:

$$\frac{1}{|D|} \log p(D|\boldsymbol{\theta}) = \frac{1}{|D|} \sum_{x \in D} \boldsymbol{\theta}^T f(x) - \log Z(\boldsymbol{\theta}). \tag{16.1}$$

To find a maximum with respect to $\boldsymbol{\theta}$, we compute the gradient.

$$\begin{aligned}\frac{1}{|D|} \nabla_{\boldsymbol{\theta}} \log p(D|\boldsymbol{\theta}) &= \frac{1}{|D|} \sum_{x \in D} f(x) - \frac{1}{Z(\boldsymbol{\theta})} \sum_{x \in D} f(x) \exp\{\boldsymbol{\theta}^T f(x)\} \\ &= \frac{1}{|D|} \sum_{x \in D} f(x) - \mathbb{E}_{x \sim p}[f(x)],\end{aligned}$$

which is equal zero when

$$\frac{1}{|D|} \sum_{x \in D} f(x) = \mathbb{E}_{x \sim p}[f(x)]. \tag{16.2}$$

In other words, at the optimum, the sufficient statistics of p and the empirical distribution match. The method of directly solving this equation is called *moment matching*.

16.2 Approximation inference for MLE

Since analytical solution of (16.2) might not exist, we have to resort to iterative methods to obtain and approximated solution. Fortunately, the right-hand side of (16.1) has a nice property; notice that the first term is linear and the second term is convex, which can be seen from the Hölder inequality; for any $\alpha \in (0, 1)$,

$$\begin{aligned}\log Z(\alpha\boldsymbol{\theta}_1 + (1-\alpha)\boldsymbol{\theta}_2) &= \log \sum_{x \in D} \exp\left\{\alpha \boldsymbol{\theta}_1^T f(x)\right\} \exp\left\{(1-\alpha)\boldsymbol{\theta}_2^T f(x)\right\} \\ &\leq \log \left[\Big(\sum_{x \in D} \exp\{\boldsymbol{\theta}_1^T f(x)\} \Big)^{\alpha} \Big(\sum_{x \in D} \exp\{\boldsymbol{\theta}_2^T f(x)\} \Big)^{1-\alpha} \right] \\ &= \alpha \log Z(\boldsymbol{\theta}_1) + (1-\alpha) \log Z(\boldsymbol{\theta}_2).\end{aligned}$$

Thus, the log-likelihood function (16.1) is concave, which is good news for us since gradient ascent is guaranteed to find the global maximum. However, computing $\mathbb{E}_{x \sim p}[f(x)]$ is an inference problem, which is often intractable in higher dimensions. There are several methods that we can use to approximate this term.

16.2.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a sampling-based method that approximates expectations of the form

$$\mathbb{E}_{x\sim p}[f(x)] = \int f(x)p(x)\ dx.$$

which is often impossible to perform analytically. The main idea of this method is to sample $x^1, \dots x^S$ from p and approximate the expectation using the Monte Carlo principle:

$$\mathbb{E}_{x\sim p}[f(x)] \approx I_S = \frac{1}{S}\sum_{k=1}^{S} f(x^k).$$

To produce samples whose distribution is close to p, we draw these samples from p sequentially as a *Markov chain*.

Definition 16.1. A (discrete-time) Markov chain is a sequence of random variables $X^0, X^1, X^2, \dots$, where X^t has finite *states* $X^t \in \{1, 2, \dots, d\}$, such that the *transition probability*

$$T(i|j) = \Pr(X^{t+1} = i | X^t = j)$$

does not depend on k.

Thus, we have a $d \times d$ matrix $T_{ij} = T(i|j)$. If x^0 is drawn from an initial probability p_0, then the probability of X^t being in each state is given by

$$p_t = T^t p_0.$$

If the limit $\pi = \lim_{t\to\infty} p_t$ exists, we call π a *stationary distribution* of the Markov chain. In particular, we have

$$T\pi = T \lim_{t\to\infty} p_t = \lim_{t\to\infty} p_{t+1} = \pi,$$

and it can be shown directly from the definition that any π satisfying $\pi = T\pi$ is a stationary distribution.

In the Markov chain part of the MCMC, we will construct a Markov chain whose transition probability is equal to the model probability p. The main question, then, is whether a stationary distribution exists for this chain. It turns out that this is the case under two conditions:

- *Irreducibility*: There is a positive probability that state i is transferred to state j in a finite number of steps for all i and j.
- *Aperiodicity*: It is always possible to return to the initial state after a fixed period of time i.e. there exists t such that for all i and all $t' > t$, $\mathbb{P}(X^{t'} = i | X^0 = i) > 0$.

The following proposition gives us a sufficient condition that pi is a stationary distribution.

Proposition 16.2. In a Markov chain with transition matrix T, any distribution π satisfying *detailed balance*

$$\pi(i)T(j|i) = \pi(j)T(i|j) \quad \text{for all } i, j$$

is a stationary distribution.

Proof: By summing both sides over j, we have that

$$\pi(i) = \sum_j \pi(j)T(i|j) = T\pi(i).$$

This is true for all i, and so $\pi = T\pi$. ■

The following algorithms use this principle to generate a Markov chain.

Metropolis-Hasting algorithm

The main idea of Metropolis-Hasting (MH) algorithm is to decompose the transition operator $T(x'|x)$ as follows:

$$T(x'|x) = q(x'|x)A(x'|x),$$

where

- $q(x'|x)$ is an arbitrary distribution chosen by the user—usually a Gaussian centered at x .
- An acceptance probability

$$A(x'|x) = \min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right).$$

The algorithm starts with an initial state x^0 then iterate the following procedure: at the current point x in the chain, we pick the next point x' according to q. Then, we sample $u \in \text{unif}[0,1]$ which determines the next term in the chain according to the following rule:

- $u \leq A(x'|x)$: move to x'.
- $u > A(x'|x)$: keep x.

Thus, we are more likely to move to x' if it is assigned with high probability (high $p(x')$). On the other hands, if q suggests us low probability moves (high $q(x'|x)$ but low $p(x')$), we will follow only some of those moves.

To see that p is the stationary distribution of this chain, we will use Proposition 16.2. Since either $A(x'|x) < 1$ or $A(x|x') < 1$, we can assume without loss in generality that $A(x'|x) < 1$, which yields $A(x|x') = 1$. It then follows that

$$\begin{aligned}
\frac{A(x'|x)}{A(x|x')} &= \frac{p(x')q(x|x')}{p(x)q(x'|x)} \\
p(x)q(x'|x)A(x'|x) &= p(x')q(x|x')A(x|x') \\
p(x)T(x'|x) &= p(x')T(x|x'),
\end{aligned}$$

which is the detailed balance condition.

Gibbs sampling

Suppose that we want to take samples of $\boldsymbol{X} = (x_1, x_2, \ldots, x_n)$ from a distribution $p(x_1, x_2, \ldots, x_n)$. We can use Gibbs sampling which is a special case of MH algorithm where only one variable is changed in the transition

$$(\ldots, x_{j-1}^{i+1}, x_j^i, x_{j+1}^i, \ldots) \longrightarrow (\ldots, x_{j-1}^{i+1}, x_j^{i+1}, x_{j+1}^i, \ldots).$$

Thus, if we want k samples, there will be nk states in the Markov chain. The transition kernel q has a specific form:

$$q(x'|x) = p(x_j'|x_{-j}),$$

where $x_{-j} = (x'_1, \ldots, x'_{j-1}, x_{j+1}, \ldots, x_n)$. If we denote x'_{-j} to be the Consequently,

$$p(x') = q(x'_j|x)p(x_{-j})$$
$$p(x) = q(x_j|x')p(x_{-j})$$

Therefore, the acceptance probability $A(x'|x)$ in the MH algorithm is simplified to one, leading to the following algorithm of Gibbs sampling:

- Start an initial sample $x^0 = (x_1^0, \ldots, x_n^0)$.
- For $t = 1, 2, \ldots$, repeat until convergence:
  1. Sample $x_i^{t+1} \sim p(x_i|x_{-i}^t)$
  2. Update $x^{t+1} \leftarrow (x_1^{t+1}, \ldots, x_i^{t+1}, x_{i+1}^t, \ldots x_n^t)$
  3. For $i = 1, 2, \ldots$, repeat the previous steps until $i = n$.

Here, we denote $x_{-i}^t = (x_1^{t+1}, \ldots, x_{i-1}^{t+1}, x_{i+1}^t, \ldots, x_n^t)$.