6.3.5 Inference problem

Lemma 9 *Estimating maximally probable decisions y, given attributes x and parameters θ, i.e.*

$$\operatorname*{argmax}_{y \in \{0,1\}^S} \quad p_{\mathcal{Y}|\mathcal{X},\Theta}(x, y, \theta) \tag{6.22}$$

is identical to the structured inference problem with $\hat{H}(x, y) = H_\theta(x, y)$.

Exercise 5 *Prove Lemma 9.*

6.3.6 Learning algorithm

On the on hand, the supervised structured learning problem can be solved exactly by means of the steepest descent algorithm, due to its convexity (Lemma 8).

Algorithm 1 Steepest descent with tolerance parameter $\epsilon \in \mathbb{R}_0^+$

```
θ := 0
repeat
    d := ∇_θ L(H_θ(x, ·), y)
    η := argmin_{η'∈ℝ} L(H_{θ−η'd}(x, ·), y)    (line search)
    θ := θ − ηd
    if ‖d‖ < ε
        return θ
```

On the other hand, the time complexity of computing the gradient is $O(2^{|S|})$, due to the summations involved in computing the partition function $Z(x, \theta)$ and expectation values (6.19). More specifically, computing a derivative

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} \ln Z &= -\mathbb{E}_{y' \sim p_{\mathcal{Y}|\mathcal{X},\Theta}}(\xi_j(x, y')) \\
&= -\frac{1}{Z(x,\theta)} \sum_{y' \in \{0,1\}^S} \xi_j(x, y')\, e^{-\langle \theta, \xi(x,y') \rangle} \\
&= \frac{1}{Z(x,\theta)} \sum_{y' \in \{0,1\}^S} \sum_{f \in F} \varphi_{fj}(x_f, y'_{S_f})\, e^{-\langle \theta, \xi(x,y') \rangle} \\
&= \frac{1}{Z(x,\theta)} \sum_{f \in F} \sum_{y'_{S(f)} \in \{0,1\}^{S(f)}} \sum_{y'_{S \setminus S(f)} \in \{0,1\}^{S \setminus S(f)}} \varphi_{fj}(x_f, y'_{S(f)})\, e^{-\langle \theta, \xi(x,y') \rangle} \\
&= \sum_{f \in F} \sum_{y'_{S(f)} \in \{0,1\}^{S(f)}} \varphi_{fj}(x_f, y'_{S(f)}) \frac{1}{Z(x,\theta)} \sum_{y'_{S \setminus S(f)} \in \{0,1\}^{S \setminus S(f)}} e^{-\langle \theta, \xi(x,y') \rangle} && (6.23) \\
&= \sum_{f \in F} \sum_{y'_{S(f)} \in \{0,1\}^{S(f)}} \varphi_{fj}(x_f, y'_{S(f)})\, p_{\mathcal{Y}_{S(f)}|\mathcal{X},\Theta}(y'_{S(f)} \mid x, \theta) && (6.24) \\
&= \sum_{f \in F} \mathbb{E}_{y'_{S(f)} \sim p_{\mathcal{Y}_{S(f)}|\mathcal{X},\Theta}}(\varphi_{fj}(x_f, y'_{S(f)})) && (6.25)
\end{aligned}$$

requires computing

- the partition function

$$Z(x, \theta) = \sum_{y' \in \{0,1\}^S} e^{-\langle \theta, \xi(x,y') \rangle} \tag{6.26}$$

- for every factor $f \in F$, the so-called *factor marginal*

$$p_{\mathcal{Y}_{S(f)}|\mathcal{X},\Theta}(y'_{S(f)} \mid x, \theta) = \frac{1}{Z(x,\theta)} \sum_{y'_{S \setminus S(f)} \in \{0,1\}^{S \setminus S(f)}} e^{-\langle \theta, \xi(x,y') \rangle} \tag{6.27}$$

- for every factor $f \in F$, the expectation value

$$\sum_{y'_{S(f)} \in \{0,1\}^{S(f)}} \varphi_{fj}(x_f, y'_{S(f)}) \, p_{\mathcal{Y}_{S(f)}|\mathcal{X},\Theta}(y'_{S(f)} \mid x, \theta) \ . \tag{6.28}$$

In the special case where the degree $\max_{f \in F} S(f)$ of the conditional graphical model is bounded by a constant, computing (6.28) from the factor marginal takes constant time. The challenge in (6.27) or all (6.26) is to sum the function

$$\psi_\theta(x, y') := e^{-\langle \theta, \xi(x,y') \rangle} \tag{6.29}$$

over assignments to some (6.27) or all (6.26) variables y'. Defining

$$\psi_{f\theta}(x_f, y'_{S(f)}) = e^{-\langle \theta, \varphi_f(x_f, y'_{S(f)}) \rangle} \tag{6.30}$$

and exploiting factorization (6.6), we obtain

$$\begin{aligned}
& e^{-\langle \theta, \xi(x,y') \rangle} \\
&= e^{-\sum_{f \in F} \langle \theta, \varphi_f(x_f, y_{S(f)}) \rangle} && (6.31) \\
&= \prod_{f \in F} e^{-\langle \theta, \varphi_f(x_f, y_{S(f)}) \rangle} && (6.32) \\
&= \prod_{f \in F} \psi_{f\theta}(x_f, y_{S(f)}) \ . && (6.33)
\end{aligned}$$

Thus, the challenge in (6.27) and (6.26) is to compute a sum of a product of functions. Specifically:

$$Z(x,\theta) = \sum_{y' \in \{0,1\}^S} \prod_{f \in F} \psi_{f\theta}(x_f, y_{S(f)}) \tag{6.34}$$

$$p_{\mathcal{Y}_{S(f)}|\mathcal{X},\Theta}(y'_{S(f)} \mid x, \theta) = \frac{1}{Z(x,\theta)} \sum_{y'_{S \setminus S(f)} \in \{0,1\}^{S \setminus S(f)}} \prod_{f \in F} \psi_{f\theta}(x_f, y_{S(f)}) \tag{6.35}$$

One approach to tackle this problem is to sum over variables recursively. In order to avoid redundant computation, Kschischang et al. (2001) define partial sums:

Definition 13 (Kschischang et al. (2001)) For any variable node $s \in S$ and any factor node $f \in F$, the functions

$$m_{s \to f}, m_{f \to s} \colon \{0,1\} \to \mathbb{R} \ , \tag{6.36}$$

called *messages*, are defined such that for all $y_s \in \{0,1\}$:

$$m_{s \to f}(y_s) = \prod_{f' \in F(s) \setminus \{f\}} m_{f' \to s}(y_s) \tag{6.37}$$

$$m_{f \to s}(y_s) = \sum_{y_{S(f) \setminus \{s\}}} \psi_{f\theta}(x_f, y_{S(f)}) \prod_{s' \in S(f) \setminus \{s\}} m_{s' \to f}(y_{s'}) \tag{6.38}$$

Lemma 10 *If the factor graph is acyclic, messages are defined recursively by* (6.37) *and* (6.38)*, beginning with the messages from leaves. Moreover, for any $s \in S$ and any $f \in F$:*

$$Z(x, \theta) = \sum_{y_s \in \{0,1\}} \prod_{f' \in F(s)} m_{f' \to s}(y_s) \tag{6.39}$$

$$p_{\mathcal{Y}_{S(f)} | \mathcal{X}, \Theta}(y'_{S(f)} \mid x, \theta) = \frac{1}{Z(x, \theta)} \psi_{f\theta}(x_f, y_{S(f)}) \prod_{s' \in S(f)} m_{s' \to f}(y_{s'}) \tag{6.40}$$

Exercise 6 *Prove Lemma 10.*

The recursive computation of messages is known as *message passing*.

If the factor graph is acylic, the supervised structured learning problem can be solved efficiently by means of the steepest descent algorithm and message passing, by Lemma 8 and Lemma 10.

If the factor graph is cyclic, the definition of messages by (6.37) and (6.38) is cyclic as well. The partition function and marginals cannot be computed by message passing in general. A heuristic without guarantee of correctness or even convergence is to initialize all messages as normalized constant functions and update messages according to some schedule, e.g., synchronously. This heuristic is known as *loopy belief propagation* and has proven suitable for some applications.