14. Computing marginal probabilities for GRFs

Let us consider a GRF for pairs (x,s) on a graph (V,E), where $x: V \to F$ is a field of features and $s: V \to K$ is a field of hidden states

$$p_u(x,s) = \frac{1}{Z(u)} \exp\Big[\sum_{i\in V} u_i(x_i, s_i) + \sum_{ij\in E} u_{ij}(s_i, s_j)\Big].$$

Computing its marginal probabilities on nodes and edges like

$$p_u(s_i),\quad p_u(s_i|x),\quad p_u(s_i,s_j),\quad p_u(s_i,s_j|x)$$

for $i,j\in V$, $\{i,j\}\in E$ is needed for

(1) Inference with locally additive loss functions (e.g. Hamming dist.)
$\Rightarrow$ the optimal prediction is based on $p(s_i|x)$

(2) Learning model parameters u_i, u_{ij}. This requires to compute node an edge marginals from u-s and vice versa (see next section).

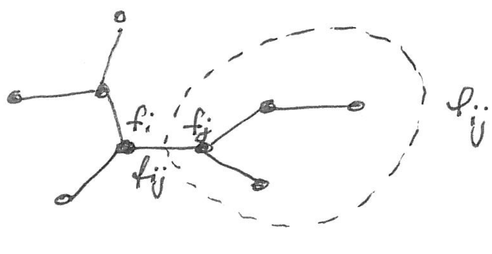
Computing the partition function $Z(u)$ and marginal prob's for general GRFs is NP-hard. We have to rely on approximations.

A. Belief propagation

Let us start from computing marginal prob's of a Markov model on a tree (see Sec. 10)

$$p(s) = \frac{1}{Z} \prod_{i\in V} f_i(s_i) \prod_{ij} f_{ij}(s_i, s_j) =$$

$$= \prod_{i\in V} p(s_i) \prod_{ij\in E} \frac{p(s_i,s_j)}{p(s_i)\,p(s_j)}$$



Hence, the marginals can be written as

$$p(s_i) \propto f_i(s_i) \prod_{j\in N_i} \varphi_{ij}(s_i)$$

$$p(s_i,s_j) \propto f_i(s_i) f_{ij}(s_i,s_j) f_j(s_j) \prod_{l\in N_i\setminus j} \varphi_{il}(s_i) \prod_{m\in N_j\setminus i} \varphi_{jm}(s_j)$$

It follows that

$$\frac{p(s_i, s_j)}{p(s_i)\, p(s_j)} = \frac{f_{ij}(s_i, s_j)}{\varphi_{ij}(s_i)\, \varphi_{ji}(s_j)}$$

Remark 1

- Recall that φ-s are defined on oriented edges
- The formula above is a reparametrisation in the $(+, \times)$- domain. □

The recursive definition of the φ-s is

$$\varphi_{ij}(s_i) = \sum_{s_j \in K} f_{ij}(s_i, s_j)\, f_j(s_j) \prod_{\ell \in \mathcal{N}_j \setminus i} \varphi_{j\ell}(s_j)$$

Belief propagation for random fields on general graphs (aka message passing)

(1) repeatedly recompute φ-s using the formula above, until (hopefully) a fixpoint φ^* is reached.

(2) estimate marginals from

$$p(s_i) \propto f_i(s_i) \prod_{j \in \mathcal{N}_i} \varphi^*_{ij}(s_i)$$

$$p(s_i, s_j) \propto f_{ij}(s_i, s_j) \frac{p(s_i)\, p(s_j)}{\varphi^*_{ij}(s_i)\, \varphi^*_{ji}(s_j)}$$

Remark 2

- Quite often BP gives reasonable estimates for unary marginals. However, it inherently fails to estimate pairwise marginals.
- In the log-domain, i.e replacing $(+, \times)$ by $(\min, +)$, this gives an approximation algorithm for solving $(\min, +)$-problems.

B. Sampling

Let $s = \{s_i \in K \mid i \in V\}$ be a K-valued random field with joint p.d. $p(s)$ and let $F: K^V \to \mathbb{R}$ be a function. How can we estimate its expectation $\mathbb{E}_p(F) = \sum_{s \in K^V} p(s) F(s)$?

- generate an i.i.d. sample $\{s^j \in K^V \mid j = 1, \dots, \ell\}$ of realisations from $p(s)$
- estimate the expectation by $\mathbb{E}_p(F) \approx \frac{1}{\ell} \sum_{j=1}^{\ell} F(s^j)$.

How to sample from $p(s)$? Theorem 1, Sec. 1 $\Rightarrow$ design a homogeneous Markov chain with transition probability $T(s \mid s')$, $s, s' \in K^V$ s.t.

(a) the chain is irreducible and a-periodic,

(b) its unique limiting distribution is $p(s)$.

In practice:

- Design a set of simple (sparse) transition prob. matrices B_m, $m \in M$ s.t. $p(s)$ is stationary for all of them
- Compose T by

$$T = \prod_{m \in M} B_m \quad \text{or} \quad T = \sum_{m \in M} \alpha_m B_m \quad \text{with } \alpha_m \geq 0, \ \sum_{m \in M} \alpha_m = 1$$

- Prove that T is irreducible and a-periodic.

Gibbs Sampler for GRFs

Define B_i, $i \in V$ by

$$B_i(s \mid s') = \begin{cases} p(s_i \mid s'_{V \setminus i}) \overset{!}{=} p(s_i \mid s'_{\mathcal{N}_i}) & \text{if } s_{V \setminus i} \equiv s'_{V \setminus i} \\ 0 & \text{otherwise} \end{cases}$$

Stationarity of $p(s)$:

$$\sum_{s' \in K^V} B_i(s \mid s') p(s') = \sum_{k \in K} p(s_i \mid s_{V \setminus i}) \, p(s'_i = k \mid s_{V \setminus i}) =$$

$$= p(s_i \mid s_{V \setminus i}) \, p(s_{V \setminus i}) = p(s)$$

It is easy to see that $T = \prod_{i \in V} B_i$ and $T = \sum_{i \in V} \alpha_i B_i$ are irreducible and a-periodic if $p(s)$ is strictly positive.

Remark 3

- Gibbs samplers are easy to implement
- Gibbs samplers are very slow: long „burn-in" time and slow mixing.

C. Mean field approximation

If only the unary marginals of a GRF are needed $\Rightarrow$ approximate $p(s)$ by a factorising distribution $q(s) = \prod_{i \in V} q_i(s_i)$ with smallest KL-divergence from p:

$$D_{KL}(q \| p) = \sum_{s \in K^V} q(s) \log \frac{q(s)}{p(s)} \to \min_q$$

For a GRF on a graph we get

$$\sum_{i \in V} \sum_{s_i \in K} q_i(s_i) \log q_i(s_i) - \sum_{i \in V} \sum_{s_i \in K} q_i(s_i) u_i(s_i) -$$

$$- \sum_{ij \in E} \sum_{s_i, s_j \in K} q_i(s_i) q_j(s_j) u_{ij}(s_i, s_j) \to \min_{q \geq 0}$$

s.t. $\sum_{s_i \in K} q_i(s_i) = 1 \quad \forall\, i \in V$

This can be solved approximately e.g. by block-coordinate descent.