SML SS22 | L10

Computing marginal probabilities for GRFs

$S = \{S_i \in K \mid i \in V\}$ is a K-valued Gibbs random field w.r.t. the undirected graph (V, E)

$$p_u(s) = \frac{1}{Z(u)} \exp \sum_{ij \in E} u_{ij}(s_i, s_j)$$

Task compute its marginal distributions $p(s_i)$, $i \in V$ for vertices and $p(s_i, s_j)$, $\{i,j\} \in E$ for edges

We know that for exponential families

$$p_u(s) = \frac{1}{Z(u)} \exp \langle \varphi(s), u \rangle$$

implies $\mathbb{E}_u \varphi = \nabla_u \log Z(u)$, but this does not help.

Remark 1 The task can be solved in polynomial time if (V, E) is a tree.

□

A. Gibbs Sampler

Let $F : K^V \to \mathbb{R}$ be a random vector on the field $S = \{S_i \in K \mid i \in V\}$
We can estimate its expectation by

(1) generating an i.i.d. sample $s^j \sim p_u(s)$, $j = 1, \dots, \ell$

(2) approximating $\mathbb{E}_u[F] \approx \frac{1}{\ell} \sum_{j=1}^{\ell} F(s^j)$

Remark 2 If F is bounded, we can use the Hoeffding inequality to estimate the sample size ℓ required for a given confidence interval.

□

Gibbs sampler: define a homogeneous Markov chain with transition probability $T(s \mid s')$, $s, s' \in K^V$ s.t.

- the chain is irreducible and a-periodic
- its limiting distribution is $p_u(s)$

Remark 3 A stronger and sometimes easier to prove condition is detailed balance

$$T(s|s')p_u(s') = T(s'|s)p_u(s),$$

which ensures that the reverse chain is identical with the direct one. □

Practically: Construct simple, „atomic" samplers $B_i(s|s')$, $i \in V$ by

$$B_i(s|s') = \begin{cases} p_u(s_i|s_{V\setminus i}) & \text{if } s_{V\setminus i} = s'_{V\setminus i} \\ 0 & \text{otherwise} \end{cases}$$

Notice that

$$p_u(s_i|s_{V\setminus i}) = p_u(s_i|s_{N_i}) = \frac{1}{Z_i(u)} \exp \sum_{j \in N_i} u_{ij}(s_i, s_j)$$

is easy to compute for a GRF. Each of the samplers B_i has $p_u(s)$ as stationary (but not limiting) distribution. Finally, construct T by $T = \prod_{i \in V} B_i$ or by $T = \sum_{i \in V} \alpha_i B_i$. It is easy to prove that T is irreducible and a-periodic.

Gibbs samplers are easy to implement, but slow. The successive realisations are correlated:

$$C_F(t) = \operatorname{cov}(F_{t_0}, F_{t_0+t}) =$$

$$= \sum_{s,s'} p_u(s)\, F(s)\, T^t(s'|s)\, F(s') - \mathbb{E}_u^2[F]$$

and $$\rho_F(t) = \frac{C_F(t)}{C_F(0)} \sim e^{-t/\tau} \quad \text{with mixing time } \tau.$$

Side step: Sampling & stochastic gradient estimators

Let us consider a deep stochastic Gaussian network with layer activations $z^k \in \mathbb{R}^{n_k}$. Its conditional distribution $p(z_m, z_{m-1}, .., z_1 | z_0)$ is given by

$$p(z_m, z_{m-1}, .., z_1 | z_0) = \prod_{k=1}^{m} p(z_k | z_{k-1})$$

with $p(z_k|z_{k-1}) = \mathcal{N}(\mu_k, \mathrm{diag}^{-1}(\sigma_k))$, where

$$\mu_k = W^k z_{k-1}, \quad \sigma_k = f(\tilde{W}^k z_{k-1})$$

and $f(u) = e^u$ or $f(u) = \log(1+e^u)$.

We fix a loss function $L(z_0, z_m)$ that is differentiable in z_m.

Given training data $(z_0, z_m) \in T$, we want to solve the following optimisation task

$$\frac{1}{|T|} \sum_{(z_0, z_m) \in T} p_W(z_m|z_0)\, L(z_0, z_m) \to \min_W$$

Since $p_{W^m}(z_m|z_{m-1})$ is known in closed form, we can redefine

$$p_W(z_m|z_0)\, L(z_0, z_m) \to p_W(z_{m-1}|z_0)\, L(z_0, z_{m-1}, W^m) = F(W)$$

Computing

$$p_W(z_{m-1}|z_0) = \int\cdots\int dz_1 \cdots dz_{m-2}\; p_W(z_{m-1}, \dots, z_1|z_0)$$

is intractable. However, we can easily sample

$$z_{m-1}, \dots, z_1 \sim p_W(z_{m-1}, \dots, z_1|z_0)$$

in linear time! We still need to compute ∇_W. This can be done by the reparametrisation trick

$$z \sim \mathcal{N}(\mu, \sigma) \Leftrightarrow \varepsilon \sim \mathcal{N}(0,1), \quad z = \sigma\varepsilon + \mu$$

We obtain the following stochastic gradient estimator

$$\varepsilon_k \sim \mathcal{N}(0, \mathbb{I}), \quad z_k = \sigma_k \varepsilon_k + \mu_k \quad \text{with}$$

$$\mu_k = W^k z_{k-1}, \quad \sigma_k = f(\tilde{W}^k z_{k-1})$$

Hence, $z_{m-1} = h(z_0, \varepsilon, W)$ is differentiable in W and we get

$$\nabla_W L(z_0, z_{m-1} = h(z_0, \varepsilon, W), W^m).$$

B. Mean field approximation

Idea: approximate $p_u(s)$ by a simpler distribution $q(s)$, e.g. assuming that the s_i, $i\in V$ are independent. Use the KL-divergence as criterion.

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

$$\text{s.t.}\quad q(s) = \prod_{i\in V} q_i(s_i)$$

We get

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

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

The task is not convex. However, it is convex for a single q_i if all other q_j, $j\neq i$ are fixed. Hence, solve the task by block-coordinate descent. A single step of it reads

$$q_i(s_i) \leftarrow \frac{1}{Z_i}\exp\sum_{j\in N_i}\sum_{s_j\in K} u_{ij}(s_i,s_j)\,q_j(s_j)$$

Example 1 Approximate a multivariate Gaussian $\mathcal{N}(\mu_0, S^{-1})$ by a factorising Gaussian $\mathcal{N}(\mu, \operatorname{diag}^{-1}(\sigma^2))$.

We have

- $$\int dx\, q_i(x)\log q_i(x) = \int dx\, q_i(x)\Big[-\frac{1}{2\sigma_i^2}(x-\mu_i)^2 - \log\sigma_i\Big] = -\log\sigma_i + \text{const}$$

- $$\int dx_1..dx_n\, q(x)\,\langle x-\mu_0, S(x-\mu_0)\rangle \overset{x'=x-\mu_0}{=} \int dx_1..dx_n\, q(x)\,\langle x, Sx\rangle$$

$$= \sum_{i\neq j}\int dx_i dx_j\, q_i(x_i)\,q_j(x_j)\,x_i S_{ij} x_j + \sum_i \int dx_i\, q_i(x_i)\, S_{ii}\, x_i^2$$

$$= \sum_{i\neq j}\mu_i S_{ij}\mu_j + \sum_i S_{ii}(\mu_i^2+\sigma_i^2) = \langle \mu, S\mu\rangle + \sum_i S_{ii}\sigma_i^2$$

Putting all terms together, we get

$$-\sum_i \log \sigma_i + \frac{1}{2}\langle \mu, S\mu \rangle + \frac{1}{2}\sum_i S_{ii}\,\sigma_i^2 \to \min_{\mu,\sigma}$$

Recalling the shift $\mu \to \mu + \mu_0$, we get

$$\mu = \mu_0, \quad \sigma_i^2 = \frac{1}{S_{ii}}$$

□