3. Graphical models on directed acyclic graphs

- $S = \{S_i \mid i \in V\}$ - collection of K-valued random variables
- (V, E) - directed acyclic graph (DAG)

Definition 1 S is a Bayesian network (aka belief network) w.r.t. the DAG (V, E) if its joint distribution is

$$p(s) = \prod_{i \in V} p(s_i \mid s_{pa(i)}),$$

where $pa(i) \subset V$ denotes the parents of node i. ■

Let $de(i) \subset V$ denote all descendants of a node $i \in V$. Then S_i conditioned on $S_{pa(i)}$ is independent of all S_j, $j \in V \setminus de(i)$, i.e.

$$S_i \perp\!\!\!\perp S_{V \setminus de(i)} \mid S_{pa(i)}.$$

Remark 1 Notice that BNs do not imply causality. For example, a Markov model on a chain is both, a Markov model on an undirected graph (chain) and a BN on the directed chain. ■

4. Stochastic neural networks with binary units

- $X = \{X_i \mid i \in V\}$ a collection of ± 1 valued r.v.
- (V, E) - M-partite DAG with subsets $V_0, V_1, \dots, V_M$

Denote $X^m = \{X_i \mid i \in V_m\}$ and consider the conditional Bayesian network

$$p(x^M, x^{M-1}, \dots, x^1 \mid x^0) = p(x^M \mid x^{M-1}) \cdot \ldots \cdot p(x^1 \mid x^0),$$

where

$$p(X^m \mid X^{m-1}) = \prod_{i \in V_m} p(X_i^m \mid X^{m-1})$$

and

$$p(X_i^m \mid X^{m-1}) = \frac{e^{X_i^m \langle W_i^m, X^{m-1} \rangle}}{2\,\mathrm{ch} \langle W_i^m, X^{m-1} \rangle}$$

This is a sigmoid belief network

Remark 2 A sigmoid belief network on a M-partite DAG is not a MRF on the corresponding undirected graph. ■

Remark 3 Computing the probabilities of the output nodes X^M given the input X^0, requires to marginalise over all hidden (latent) layers $X^1, \dots, X^{M-1}$ and is hard. ■

Remark 4 Sampling a realisation $X^1, \dots, X^M$ given X^0 is easy (linear in model size). ■

1.B. Stochastic neural networks with Gaussian units

- $Z = \{Z_i \mid i \in V\}$ a collection of real valued r.v.
- (V, E) a M-partite DAG with subsets $V_0, V_1, \dots, V_M$

Denote $Z^m = \{Z_i \mid i \in V_m\}$ and consider the conditional BN

$$p(Z^M, Z^{M-1}, \dots, Z^1 \mid Z^0) = p(Z^M \mid Z^{M-1}) \cdot \ldots \cdot p(Z^1 \mid Z^0),$$

where

$$p(z^m | z^{m-1}) \sim \mathcal{N}(\mu(\theta, z^{m-1}), C(\theta, z^{m-1}))$$

with a diagonal covariance matrix C

<u>Remark 5</u> To compute $p(z^M | z^0)$, we need to solve the integrals

$$p(z^M | z^0) = \int dz^{M-1} \ldots \int dz^1 \, p(z^M | z^{M-1}) \cdot \ldots \cdot p(z^1 | z^0).$$

This is hard.

Sampling a realisation $z^1, \ldots, z^M$ given z^0 is easy (linear in model size). Derivatives w.r.t. to model parameters can be computed by sampling and using the identities

$$\nabla_{\mu_i} \mathbb{E}_{\mathcal{N}(\mu, C)} f(z) = \mathbb{E}_{\mathcal{N}(\mu, C)} \frac{\partial f(z)}{\partial z_i}$$

$$\nabla_{C_{ij}} \mathbb{E}_{\mathcal{N}(\mu, C)} f(z) = \frac{1}{2} \mathbb{E}_{\mathcal{N}(\mu, C)} \frac{\partial^2 f(z)}{\partial z_i \partial z_j}$$