GRAPHICAL MODELS

WITHIN DL, GMs ARE A 'MODELING TOOL'. **TASKS**: DENSITY ESTIMATION, DENOISING, IMPUTATION, SAMPLING

TABULAR, FULL CASE NOT VIABLE: k^n PARAMS, STORAGE COST, CURSE OF DIMENSIONALITY, COST OF SAMPLING, COST OF INFERENCE

- **DIRECTED MODELS** → CONDITIONAL DISTRIBUTIONS

  $p(x) = \prod_i p(x_i \mid \text{Pa}(x))$ • $O(k^n) \xrightarrow[\text{DAG}]{\text{TAB}} O(k^m)$, m IS MAX NO VARS APPEARING IN A CONDITIONAL

- **UNDIRECTED MODELS** → FACTORS, CLIQUE POTENTIALS $\phi(\mathbb{C})$

- $\tilde{p}(x) = \prod_{\mathbb{C} \in \mathcal{G}} \phi(\mathbb{C})$ UNNORMALIZED DISTRIBUTION • **PARTITION FCN** $Z = \int \tilde{p}(x)\,dx = 1$ → NORMALIZED DISTRIBUTION $p(x) = \frac{1}{Z}\tilde{p}(x)$

- **ENERGY FORMULATION**

  $\tilde{p}(x) = \exp(-E(x))$ ALLOW NOT TO EXPLICITLY CONSTRAIN POTENTIAL/FACTOR FCN → **BOLTZMANN DISTRIBUTION**, HENCE 'BOLTZMANN MACHINES'

  • MAKES EASY COMPUTATIONS: $\exp(a)\exp(b) = \exp(a+b)$

- **D-SEPARATION & MORALIZATION**

- **FACTOR GRAPHS**

  UNDIRECTED MODELS WITH 'FACTOR' NODES EXPLICITLY REPRESENTING SCOPES OF EACH ϕ. FAILS OF UNNORMALIZED DISTRIBUTION, RESOLVE AMBIGUITY

- **LATENT VARIABLES VS STRUCTURE LEARNING**

  USING LATENTS TO MODEL STRUCTURE AVOIDS THE NEED TO PERFORM DISCRETE SEARCHES AND MULTIPLE ROUNDS OF TRAINING, IMPOSITION OF INTERACTIONS BETWEEN VISIBLE VARS.
  → CAN ALSO USE LATENT VARS FOR FEATURE LEARNING

- **APPROXIMATE INFERENCE**

  AKA VARIATIONAL OR SAMPLING APPROXIMATIONS **REPARAMETRIZATION TRICK** ALLOWS TO COMPUTE GRADIENTS THROUGH hs AND DO BACKPROP IS BACKPROPAGATION THROUGH SAMPLING. → $h \sim p(h,\theta) \rightarrow h = f(\theta,\eta)$ f CONTINUOUS, η NOISE. → $\frac{\partial}{\partial\theta}\int L(h)\,p(h\mid\theta)\,dh = \frac{\partial}{\partial\theta}\int L(f(\theta,\eta))\,p(\eta)\,d\eta$

→ $\hat{g} = \frac{\partial L(f(\theta,\eta))}{\partial\theta}$

GRAPHICAL MODELS IN DEEP LEARNING

EMPHASIS ON LATENT VARS, LOTS OF LATENT VARS. $|h| > |x|$. LATENTS NOT DESIGNED TO HAVE SPECIFIC MEANING. TRAINING SORTS IT OUT.
ALSO, IN DL WE RARELY CARE TO OBTAIN EXACT INFERENCE. WE USE MIN AMOUNT OF INFO AND APPROXIMATE IT AS QUICKLY AS POSSIBLE.
MODEL POWER / CAPACITY RAISED UNTIL IT 'BREAKS'

RESTRICTED BOLTZMANN MACHINE

ENERGY MODEL WITH BINARY VISIBLES AND HIDDENS $E(v,h) = -b^T v - c^T v - v^T W h$. NO DIRECT INTERACTIONS BETWEEN v_i OR h_i

$p(h\mid v) = \prod_i p(h_i \mid v)$ $\quad p(v\mid h) = \prod_i p(v_i \mid h)$ $\quad p(h_i = 1 \mid v) = \sigma(v^T W_{\cdot i} + b_i)$ → **EFFICIENT BLOCK GIBBS SAMPLING** h OR v SIMULTANEOUS

→ CAN TRAIN W/ APPROX OF $\nabla_\theta \log Z$ → $\frac{\partial}{\partial W_{ij}} \mathbb{E}_{v,h}[E(v,h)] = -v_i h_j$

MONTE CARLO METHODS

FOR SAMPLING FROM GRAPHICAL/ENERGY MODELS.

- **ANCESTRAL SAMPLING** SAMPLE IN TOPOLOGICAL ORDER, CONDITIONING ON PARENTS
- **MARKOV CHAINS** EQUILIBRIUM DISTRIBUTION, DETAILED BALANCE. π IS EIGENVECTOR OF T WITH $\lambda = 1$ AND IT'S THE LARGEST AND ONLY ONE WITH $\lambda = 1$
  → BURNIN, MIXING-RATE