Introduction of graphical model, mainly challenge will be how we find an appropriate modeling of variables.

The challenge of Unstructured modeling

- Some applications of probabilistic modeling: Density estimation, denoising, missing value imputation, sampling.
- Modeling all possible structures of input variable (table-based) is not feasible.
- Find a way modeling variables based on their relationships and dependencies, we can get graphical modeling.

Graphical Modeling

- Directed Models (belief network, Bayesian network)

  The definition of direction of edges are dependencies. $P(a|b)$: $b \rightarrow a$

  - Local conditional probability contributions

    $P(X_i \mid Pa_G(X_i))$ $P(X_i|X)$ depends on its parents, not on others.

    ↳ $P(X) = \prod_i P(X_i \mid Pa_G(X_i))$

  - Conditional Independence means that: we can speculate conditional independence relationships from graphical models.

- Undirected models (Markov Random Field, Markov Networks)

  when we are not clear about the causal relationships, sometimes a mutual influence.

  - Clique: The sub-model, aims to model maximum local interactions. for variables within a clique, a potential function $\phi(c)$ is defined, then for the joint probability $\tilde{P}(X)$ (unnormalized)

$$\tilde{P}(X) = \prod_{c \in C_G} \phi(c)$$

Product of all clique potentials

- The Partition Function

  When we get $\tilde{p}(x)$, how we normalize it to be a probability distribution.

  $$p(x) = \frac{1}{Z}\tilde{p}(x), \text{ where } Z = \int \tilde{p}(x)\,dx$$

  - Sometimes the partition function is intractable, e.g. $Z = \int x^2\,dx$, where potential function is a quadratic one. (Domain of x is important)
  - For directed graph, modeling is based on probalistic dependence, thus partition function often do not engage, but for undirected graph, we model interaction of variables using potential function, thus partition function is useful.

- Energy-Based Models

  An assumption $\tilde{p}(x) \geq 0, \forall x$ is important, a way of ensuring positive definite of $\tilde{p}(x)$ is energy-based models:

  $$\tilde{p}(x) = \exp(-E(x))$$

  Energy function to design

  Ensure positive and convinient of calculating log-likelihood.

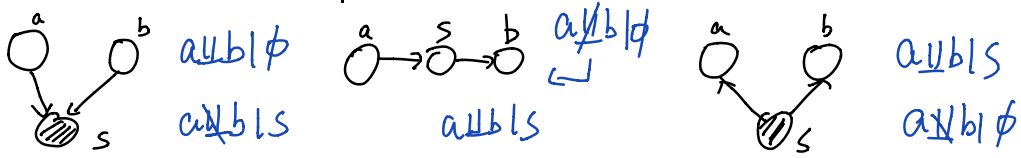
  - A term: Bolzman machine: Considering undirected graph with latent variables.

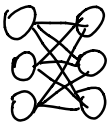
    Markov Random field: " without latent variables.

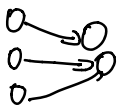
  - Product of cliques can be seen as combining knowledge of experts. each expert may be a clique.

- Separation and D-Separation

  - Undirected graphs: all the paths from a to b is obsered (inactive), we can say $a \perp b \mid$ paths.
  - Directed graphs: D-separation



- Converting between directed and undirected models
  - Boltzmann machine:  Undirected. latent variable.

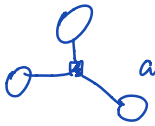
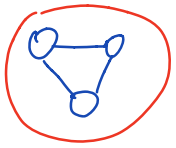
    Sparse coding:  Directed, coding sparsely
  - Why we need undirected / directed graph respectively?
    - How we choose: Which one express the most independence with lower representation cost?
    - How we use: Directed graph → Drawing samples

      Undirected graph → Approximate Inference.
    - Converting between them.

      Directed → Undirected ⇒ Dealing with "Immorality" Connections 

      Undirected → Directed ⇒ Adding chord, dealing with loop.
- Factor Graphs:

  Solving ambiguity of how the variables actually interact with each other.

   and  and  Undirected, no factor

Sampling from Graphical Models

- Ancestral Sampling: Sampling according to the dependence relationship, first of parents and then sampling according to $P(x|P_G(x))$ (Direct Graphical Models)
- Gibbs Sampling: For each time step, sampling only $P(x_i|x_{-i})$ (Depending only on neighbours), then iteratively doing sampling till convergence.

Advantage of Structured modeling and learning about dependencies / Inference

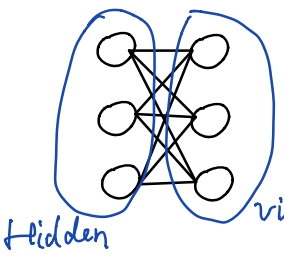
- Advantage: Separation of knowledge representation and inferencing from knowledge
- Learning dependencies taking use of latent variables: some relationship (dependence) or representations can be modeled efficiently using latent / hidden variables.
- Inference concept: Speculating probalistic distribution of $P(X \mid \text{given/observed})$

Deep learning vs. Structured Model

- Differences:
  - Concept of depth: Computational iterations vs. Hierachical Dependencies
  - Representation: Non-explainable, higher dimension of latent variables vs. Explainable, lower dimension of latent variables.
  - Inference: Distributed representation from $x \to h$ (approx. inference) vs. Intrinsic dependency modeling of data itself.

SGD ← Computational feasibility consideration (sampling, approx. inference, etc.)

- Relation:
  - Tolerance of unknown: Deep learning can be seen as models that have high tolerance of unknown, we focus on representation and gradient.
  - Concepts within them: Graphical models ⇒ Latent variables, observed
    Deep learning ⇒ Hidden Units, input/output.
- An example: RBM seen in two views.



$P(h \mid v) = \prod_i P(h_i \mid v)$

$P(v \mid h) = \prod_i P(v_i \mid h_j)$ Due to restriction property.

{ Gibbs Sampling: Undirected graph to inference.

Derivatives: $\frac{\partial}{\partial w_{ij}} \log(v,h)$

Such as deep learning.

RBM (parametric): learning mapping rules $P(h) = P^*(\text{something})$

Sparse coding (Non-parametric): learning optimal value with sparse prior.