Group Meeting - August 14, 2020

Paper review & Research progress

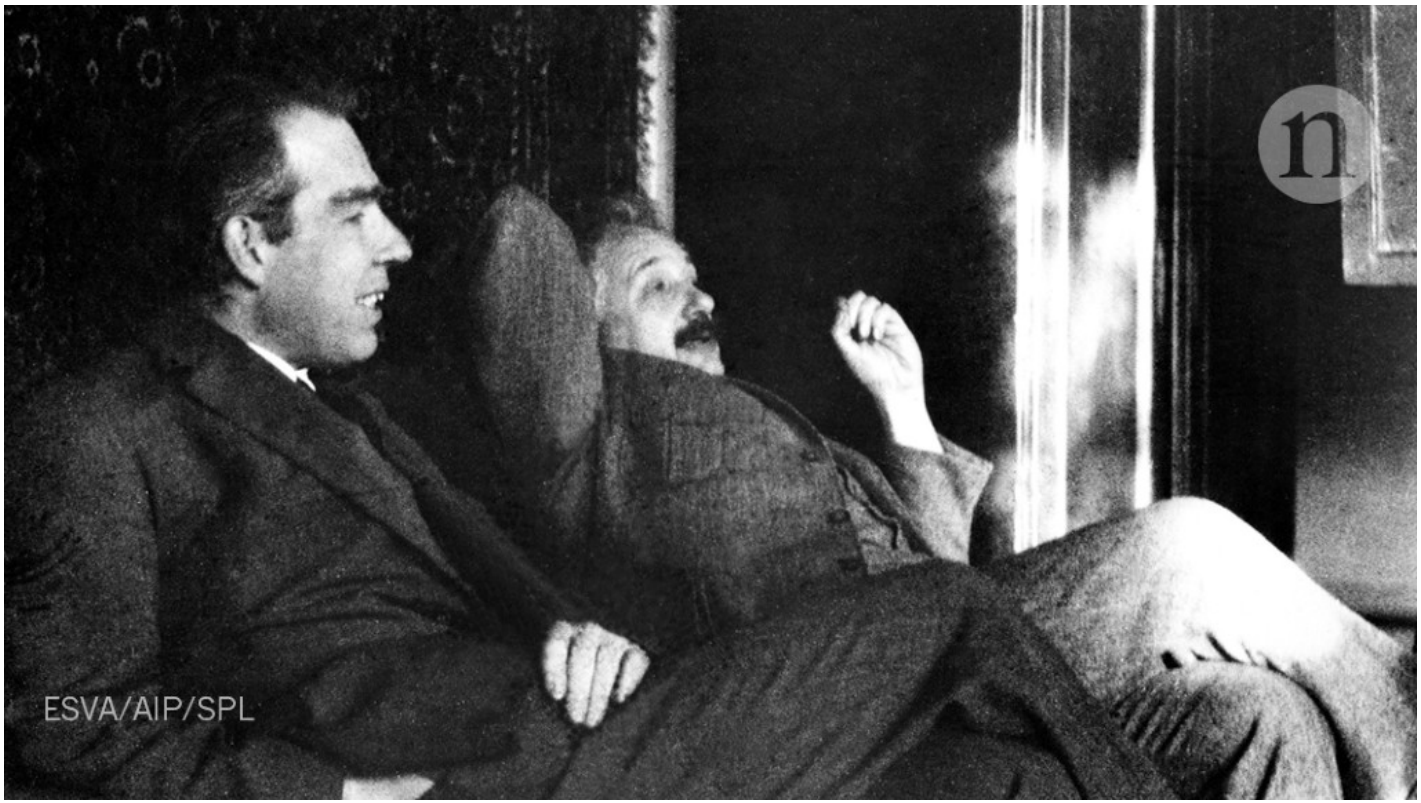
Truong Son Hy *

*Department of Computer Science
The University of Chicago

Ryerson Physical Lab

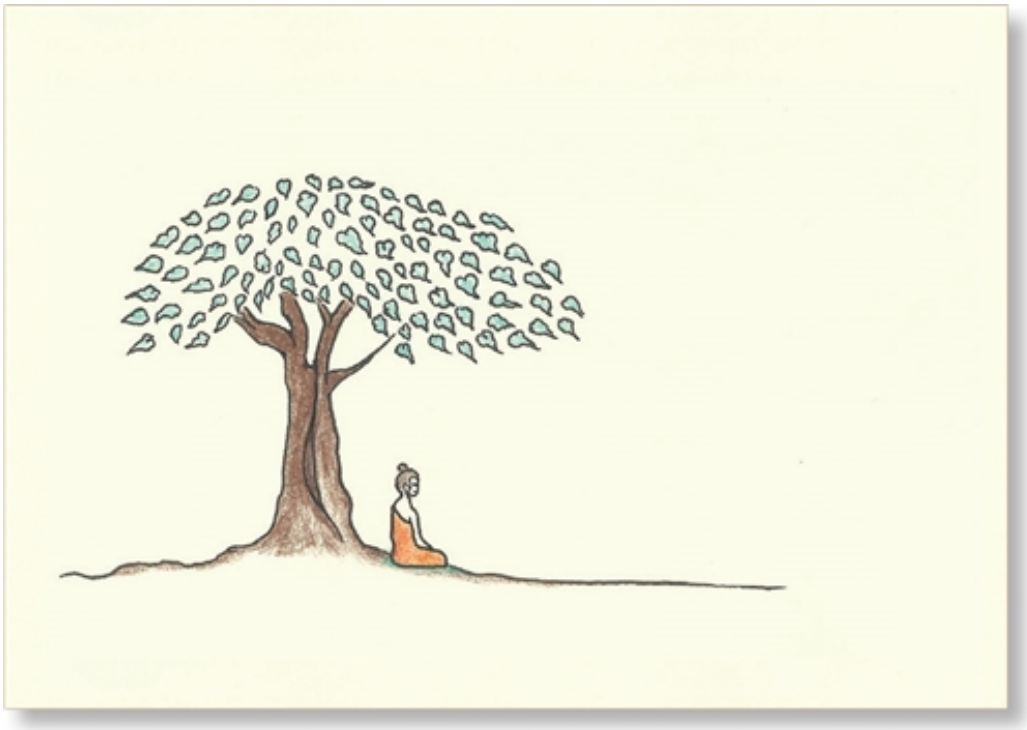
Niels Bohr

Everything we call real is made of things that cannot be regarded as real.



Buddha

1. We live in illusion and the appearance of things. There is a reality. We are that reality. When you understand this, you see that you are nothing, and being nothing, you are everything.
2. With our thoughts, we make the world.



Paper

1. **A Variational Principle for Graphical Models**, Martin J. Wainwright, Michael I. Jordan (2005)
2. **Graphical Models, Exponential Families, and Variational Inference**, Martin J. Wainwright, Michael I. Jordan (2008)

This got 3788 citations. Prof. Jordan Peterson called it the Pareto (distribution) effect of human societies.

Introduction (1)

Statistical models have long been formulated in terms of graphs, and algorithms for computing basic statistical quantities such as likelihoods and marginal probabilities have often been expressed in terms of **recursions** operating on these graphs:

- Hidden Markov Models
- Markov Random Fields
- Kalman filtering
- etc.

These recursive algorithms belong to a group of algorithms called **junction tree algorithm**, that takes advantage of factorization properties of the joint probability distribution.

Introduction (2)

Junction tree algorithm:

- For computing likelihoods and other statistical quantities associated with a graphical model, in suitably **sparse graphs**.
- Without sparsity, computationally expensive.

Alternative:

- **Markov chain Monte Carlo** (MCMC) framework: this is still slow!
- **Variational methods**: we go with this!

Introduction (3)

The class of **variational methods** provides an alternative approach to computing approximate marginal probabilities and expectations in graphical models:

1. Cast a quantity of interest (e.g. a likelihood) as the solution to an optimization problem.
2. Solve a pertubed version of this optimization problem.

Examples:

- **Belief propagation** or **sum-product** algorithm
- **Mean-field** algorithms

Graphical models (1)

A graphical model consists of a collection of probability distributions that factorize according to the structure of an underlying graph:

- Graph $G = (V, E)$.
- Each vertex $s \in V$ is associated with a random variable x_s taking values from a set $\mathcal{X}_s$ (continuous: $\mathcal{X}_s = \mathbb{R}$, discrete: $\mathcal{X}_s = \{0, 1, .., m-1\}$).
- For $A \subseteq V$, define $x_A = \{x_s | s \in A\}$.

Graphical models (2)

1. **Directed graphical models**:

$$p(\mathbf{x}) = \prod_{s \in V} p(x_s | x_{\pi(s)})$$

where $\pi(s)$ is the set of all parents of given node $s \in V$.

2. **Undirected graphical models**:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(x_C)$$

where $\psi_C : \mathcal{X}^n \to \mathbb{R}_+$ is a compatibility function associated with a clique C, and Z is a constant to ensure $\int p(\mathbf{x}) d\mathbf{x} = 1$. Another name: **Markov Random Fields**.

Graphical models (3)

Inference problems:

1. Compute the likelihood.
2. Compute the marginal distribution $p(x_A)$ where $A \subset V$.
3. Compute the conditional distribution $p(x_A|x_B)$ where $A \cup B \subset V$ and $A \cap B = \emptyset$.
4. Compute a mode of the density:

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})$$

Message passing on trees (1)

The cliques of a tree-structured graph $T = (V, E(T))$ are simply the individual nodes and edges:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E(T)} \psi_{st}(x_s, s_t)$$

Why we need a quantum computer?

By classical computing, to compute the marginal distribution for each node s:

$$\mu_s(x_s) = \sum_{\mathbf{x}' \in \mathcal{X}^n, x'_s = x_s} p(\mathbf{x}')$$

that requires a naive $O(|\mathcal{X}|^n)$ time-complexity algorithm.

Sum-product algorithm

For each substree T_t, we define $x_{V_t} = \{x_u | u \in V_t\}$. Dynamic - programming nature: a subproblem $p(x_{V_t}; T_t)$ for this subtree. The conditional independence properties of a tree alow the computation of the marginal at node s to be broken down as:

$$\mu_s(x_s) \propto \psi_s(x_s) \prod_{t \in \mathcal{N}(s)} M^*_{ts}(x_s)$$

where

$$M^*_{ts}(x_s) = \sum_{\{x'_{T_t} | x'_s = x_s\}} \psi_{st}(x_s, x'_t) p(x'_{T_t}; T_t)$$

Pearl (1988) shows that for tree-structured graphs, the recursion converges to a unique fixed-point $M* = \{M^*_{st}, M^*_{ts}, (s,t) \in E\}$ after a finite number of iterations:

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \left\{ \psi_{st}(x_s, x'_t) \psi_t(x'_t) \prod_{u \in \mathcal{N}(t)/s} M_{ut}(x'_t) \right\}$$

Junction tree representation (1)

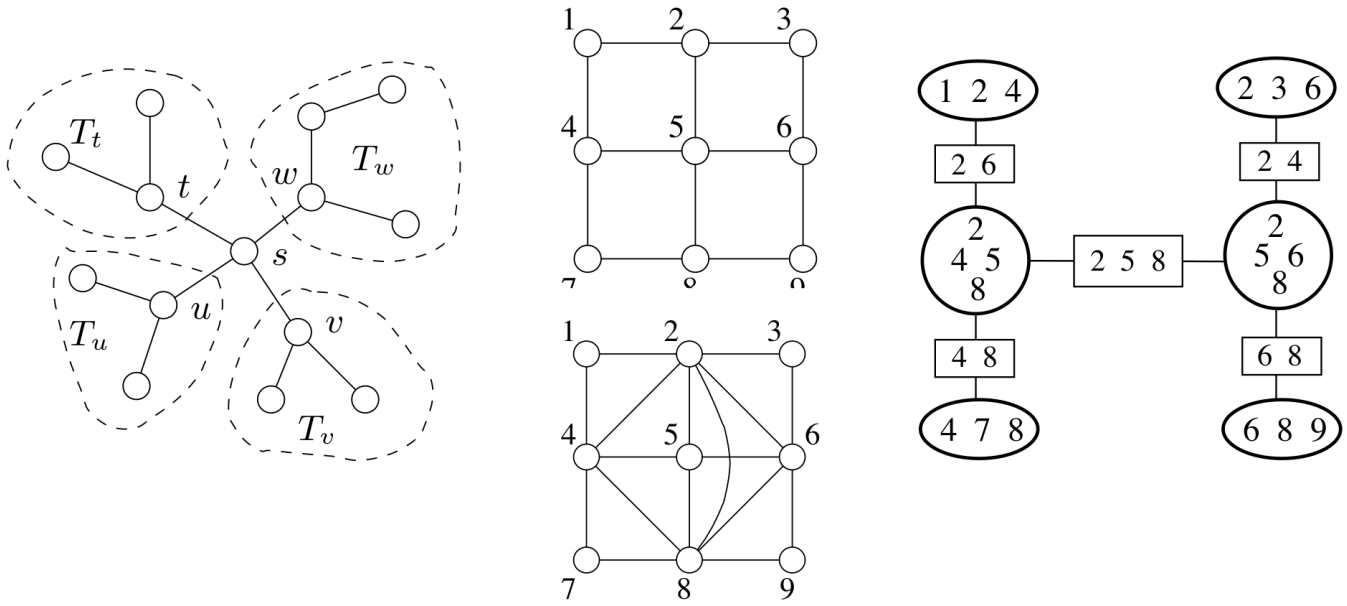
- Given a graph with cycles, cluster its nodes to form a **clique tree** - that is, an acyclic graph whose nodes are formed by cliques of G.
- **Running intersection property**: if for any two clique nodes C_1 and C_2, all nodes on the unique path joining them contain the intersection $C_1 \cap C_2$. Any clique tree with this property is known as a **junction tree**.
- From graph theory: a graph G has a junction tree if and only if it is **triangulated**.

Junction tree representation (2)

Junction tree algorithm [Lauritzen and Spiegelhalter (1988)] for exact inference on arbitrary graphs:

1. Given a graph with cycles G, triangulate it by adding edges as necesasry.
2. From a junction tree associated with the triangulated graph.
3. Run a tree inference algorithm on the junction tree.

Junction tree representation (3)



(a) (b) (c)

Figure 11.1 (a): Decomposition of a tree, rooted at node s, into subtrees. Each neighbor (e.g., u) of node s is the root of a subtree (e.g., T_u). Subtrees T_u and T_v, for $t \neq u$, are disconnected when node s is removed from the graph. (b), (c) Illustration of junction tree construction. Top panel in (b) shows original graph: a 3×3 grid. Bottom panel in (b) shows triangulated version of original graph. Note the two 4-cliques in the middle. (c) Corresponding junction tree for triangulated graph in (b), with maximal cliques depicted within ellipses. The rectangles are separator sets; these are intersections of neighboring cliques.

Junction tree representation (4)

Separator sets = intersections of cliques adjacent in the junction tree. For each separator set $S \in \mathcal{S}$, let $d(S)$ denote the number of maximal cliques to which it is adjacent. The junction tree framework guarantees that the distribution $p(.)$ factorizes in the form:

$$p(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} \mu_C(x_C)}{\prod_{S \in \mathcal{S}} [\mu_S(x_S)]^{d(S)-1}}$$

$\rightarrow$ **Tony Jebara's paper of correlated VAEs!**

Approximate inference

Variational inference algorithms:

1. Belief propagation
2. Naive mean field algorithm

For example, the naive mean field algorithm for the Ising model is a message passing algorithm on the graph:

$$\mu_s \leftarrow \left\{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st}\mu_t)]\right\}^{-1}$$

where θ_s is the observation weight of node s, θ_{st} is the weight for a pair of adjacent nodes s and t.

Maximum entropy (1)

Given a collection of functions $\phi_\alpha : \mathcal{X}^n \to \mathbb{R}$. Suppose we have observed their expected values:

$$\mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha \qquad \forall \alpha \in \mathcal{I}$$

where $\mu = \{\mu_\alpha | \alpha \in \mathcal{I}\} \in \mathbb{R}^d$, and $d = |\mathcal{I}|$ is the size of the index set.

Maximum entropy constrained optimization:

$$p_{ME} \leftarrow \arg\max_{p \in \mathcal{P}} H(p) = \arg\max_{p \in \mathcal{P}} -\sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \log p(\mathbf{x})$$

subject to:

$$\mathbb{E}_p[\phi_\alpha(\mathbf{x})] = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \phi_\alpha(\mathbf{x}) = \mu_\alpha$$

Maximum entropy (2)

Lagrangian formulation:

$$p(\mathbf{x};\theta) \propto \exp\left\{\sum_{\alpha\in\mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x})\right\}$$

that corresponds to a distribution in exponential form.

$\theta \in \mathbb{R}^d$ is called as **canonical parameter**, θ_α is the Lagrange multiplier associated with the constraint $\mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha$. The collection of functions $\phi = \{\phi_\alpha | \alpha \in \mathcal{I}\}$ is called **sufficient statistics**.

Exponential families

The exponential family associated with ϕ consists of the following parameterized collection of density functions:

$$p(\mathbf{x};\theta) = \exp\left\{\langle\theta, \phi(\mathbf{x})\rangle - A(\theta)\right\}$$

where $A(\theta)$ is called **log partition** or **cumulant generating** function:

$$A(\theta) = \log \int_{\mathcal{X}^n} \exp\langle\theta, \phi(\mathbf{x})\rangle \nu(d\mathbf{x})$$

The canonical parameters θ of interest belong to the set:

$$\Theta = \{\theta \in \mathbb{R}^d | A(\theta) < \infty\}$$

First derivatives of A:

$$\frac{\partial A}{\partial \theta_\alpha} = \int_{\mathcal{X}^n} \phi(\mathbf{x}) p(\mathbf{x};\theta)\nu(d\mathbf{x}) = \mathbb{E}_{p_\theta}[\phi_\alpha(\mathbf{x})]$$

Example: Multinomial MRF

A multinomial MRF with pairwise interactions can be written in exponential form as:

$$p(\mathbf{x};\theta) = \exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t) - A(\theta)\right\}$$

where the cumulant generating function is given by:

$$A(\theta) = \log\sum_{\mathbf{x}\in\mathcal{X}^n}\exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\right\}$$

In the special case $\mathcal{X}_s = \{0,1\}$ for all $s \in V$, the family is known as the **Ising model**.

Example: Ising (MRF), HMM, GMM

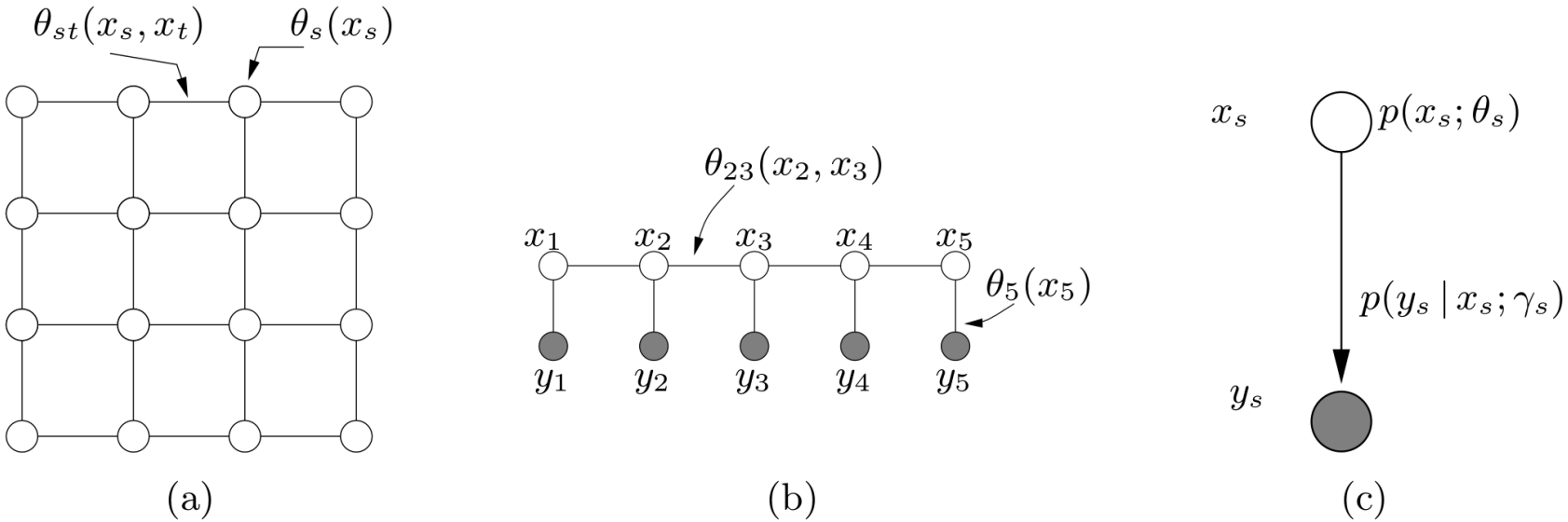


Figure 11.3 (a) A multinomial MRF on a 2-D lattice model. (b) A hidden Markov model (HMM) is a special case of a multinomial MRF for a chain-structured graph. (c) The graphical representation of a scalar Gaussian mixture model: the multinomial x_s indexes components in the mixture, and y_s is conditionally Gaussian (with exponential parameters γ_s) given the mixture component x_s.

Exact variational principle for inference

Inference problems such as:

1. computing the cumulant generating function $A(\theta)$
2. computing the vector of mean parameters $\mu = \mathbb{E}_{p_\theta}[\phi(\mathbf{x})]$

can be represented **variationally** as the solution of an **optimization problem**.

Conjugate duality (1)

Associated with any convex function $f : \mathbb{R}^d \to \mathbb{R}_*$ ($\mathbb{R}_* = \mathbb{R} \cup \{+\infty\}$) is a conjugate dual function $f^* : \mathbb{R}^d \to \mathbb{R}_*$:

$$f^*(y) = \sup_{x \in \mathbb{R}^d} \left\{ \langle y, x \rangle - f(x) \right\}$$

- **Variational definition**: $f^*(y)$ is specified as the solution of an optimization problem parameterized by y.
- **Duality**:

$$f(x) = \sup_{y \in \mathbb{R}^d} \left\{ \langle x, y \rangle - f^*(y) \right\}$$

$$A^*(\mu) = \sup_{\theta \in \Theta} \left\{ \langle \theta, \mu \rangle - A(\theta) \right\}$$

Conjugate duality (2)

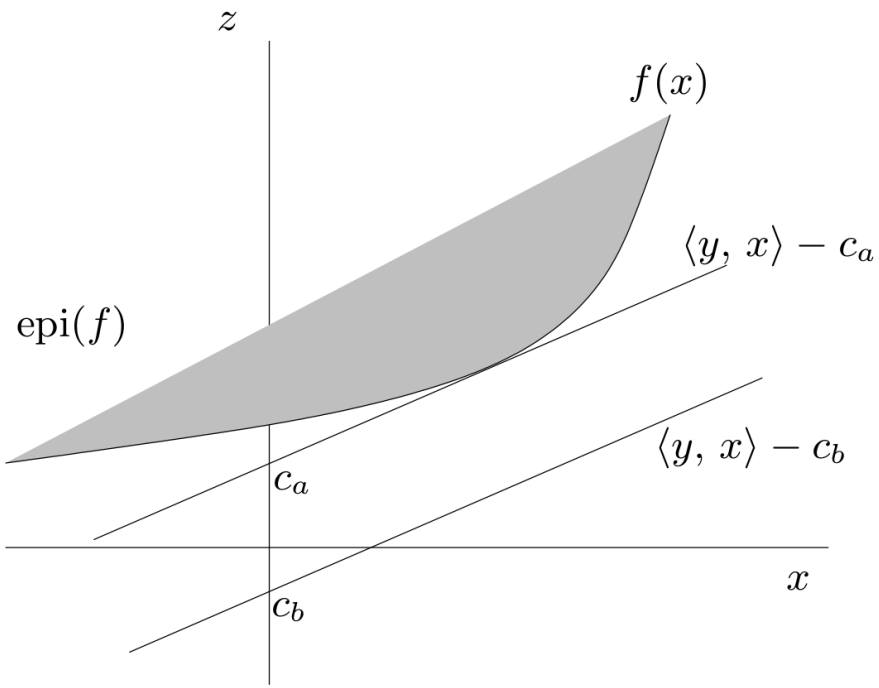


Figure 11.4 Interpretation of conjugate duality in terms of supporting hyperplanes to the epigraph of f, defined as $\text{epi}(f) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq y\}$. The dual function is obtained by translating the family of hyperplane with normal y and intercept $-c$ until it just supports the epigraph of f (the shaded region).

Duality for maximum entropy optimization (1)

Let $\mathcal{M}$ be the domain of the conjugate dual function A^*. We have the zero-gradient condition:

$$\mu = \nabla A(\theta) = \mathbb{E}_\theta[\phi(\mathbf{x})]$$

The domain $\mathcal{M}$ is defined as:

$$\mathcal{M} = \{\mu \in \mathbb{R}^d | \exists p(.) : \int \phi(\mathbf{x}) p(\mathbf{x}) \nu(d\mathbf{x}) = \mu\}$$

It is proven that: if μ is in the interior of $\mathcal{M}$, then there exists an canonical parameter $\theta(\mu) \in \Theta$ such that:

$$\mathbb{E}_{\theta(\mu)}[\phi(\mathbf{x})] = \mu$$

Duality for maximum entropy optimization (2)

Duality A^* becomes:

$$A^*(\mu) = \langle \mu, \theta(\mu) \rangle - A(\theta(\mu)) = \mathbb{E}_{\theta(\mu)}[\log p(\mathbf{x}; \theta(\mu))]$$

The right hand side is nothing else bu the negative entropy $-H(p(\mathbf{x}; \theta(\mu)))$ where:

$$H(p) = -\int_{\mathcal{X}^n} p(\mathbf{x}) \log[p(\mathbf{x})] \nu(d\mathbf{x}) = -\mathbb{E}_p[\log p(\mathbf{x})]$$

Maximum entropy problem:

$$A^*(\mu) = \max_{p \in \mathcal{P}} H(p)$$

such that:

$$\mathbb{E}_p[\phi_\alpha(\mathbf{x})] = \mu_\alpha \qquad (\forall \alpha \in \mathcal{I})$$

Exact variational principle

Exact variational principle

When it is too hard to compute any thing, remember to write it in a dual form and solve the corresponding optimization problem:

$$A(\theta) = \sup_{\mu \in \mathcal{M}} \{\langle \theta, \mu \rangle - A^*(\mu)\}$$

Approximate inference in variational form (1)

Approximation?

Duality saves our days, but optimization is still expensive computationally (and hard to solve). Therefore, we need approximate algorithms. The candidate here is **mean field methods**.

Approximate inference in variational form (1)

Approximation?

Duality saves our days, but optimization is still expensive computationally (and hard to solve). Therefore, we need approximate algorithms. The candidate here is **mean field methods**.

Let H represent a subgraph of G over which it feasible to perform exact calculations (e.g. A^* is easy to compute on H). We call H as a **tractable subgraph**. Let $\mathcal{I}(H)$ denote the subset of indices associated with cliques in H. The set of canonical parameters restricted to H:

$$\mathcal{E}(H) = \{\theta \in \Theta | \theta_\alpha = 0 \quad (\forall \alpha \in \mathcal{I} - \mathcal{I}(H))\}$$

For example, if T is a spanning forest of G:

$$\mathcal{E}(T) = \{\theta \in \Theta | \theta_{st} = 0 \quad (\forall (s,t) \notin E(T))\}$$

Approximate inference in variational form (2)

For a given subgraph H, define a restricted domain:

$$\mathcal{M}_{tract}(G;H) = \{\mu \in \mathbb{R}^d | \exists \theta \in \mathcal{E}(H) : \mu = \mathbb{E}_\theta[\phi(\mathbf{x})]\}$$

$\mathcal{M}_{tract}$ is an **inner approximation** to the set $\mathcal{M}$ of realizable mean parameters: $\mathcal{M}_{tract}(G;H) \subseteq \mathcal{M}(G)$. Based on Jensen's inequality, from the variational principle:

$$A(\theta) \geq \langle \theta, \mu \rangle - A^*(\mu) \qquad (\forall \mu \in \mathcal{M})$$

Mean-field method:

$$\mu^{MF} = \sup_{\mu \in \mathcal{M}_{tract}(G;H)} \{\langle \mu, \theta \rangle - A^*_H(\mu)\}$$

that is a relaxation of the exact variational principle.

Approximate inference in variational form (3)

Clearly, we see that the solution of mean-field algorithm μ^{MF} is a lower bound of $A(\theta)$. Given two densities p and q, the KL divergence is given by:

$$D(p||q) = \int_{\mathcal{X}^n} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x}) \nu(d\mathbf{x})$$

With a bit of algebra:

$$D(\mu||\theta) = A(\theta) + A_H^*(\mu) - \langle \mu, \theta \rangle$$

Approximate inference in variational form (3)

Clearly, we see that the solution of mean-field algorithm μ^{MF} is a lower bound of $A(\theta)$. Given two densities p and q, the KL divergence is given by:

$$D(p||q) = \int_{\mathcal{X}^n} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x}) \nu(d\mathbf{x})$$

With a bit of algebra:

$$D(\mu||\theta) = A(\theta) + A_H^*(\mu) - \langle \mu, \theta \rangle$$

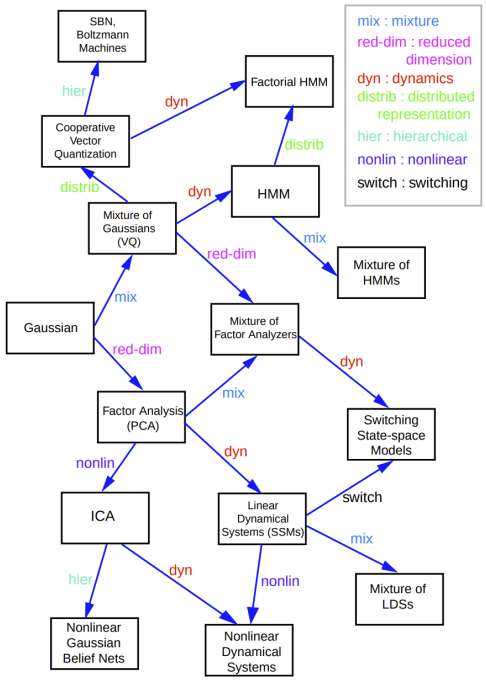
Conclusion

Therefore, solving the mean field variational problem is equivalent to minimizing KL divergence subject to the constraint that μ belongs to tractable set of mean parameters, or equivalently that p is a tractable distribution.

What more?

1. **Variational EM**: Might be useful?
2. **Bethe entropy approximation**: Prof. Tony Jebara has 7 papers in this topic. I am not sure if this is related to our work, but it seems to be important in statistics.

A Generative Model for Generative Models



Source: `http://www.cs.cmu.edu/~tom/10-702/Zoubin-702.pdf`