EN.601.783: Vision as Bayesian Inference
Homework 2

Ha Bui
hbui13@jhu.edu

Spring 2023

Gibbs Distributions and Markov Random Fields: 30 points

1. *What is semantic segmentation? How can Deep Networks be used for it? How can it be improved by using Markov Random Fields?*
- Semantic segmentation is a task that assigns a class label to all pixels in the image.
- Deep Networks for Semantic Segmentation (e.g., DeepLab) give estimates for the class labels of each pixel.
- It can be improved by using Markov Random Fields to model an undirected graph and ignore the spatial context (or temporal context) by using Markov conditions, the core idea is that neighboring pixels are likely to have the same labels.

2. *What is a Gibbs distribution? What is the energy function? What is the normalization constant?*
- The Gibbs distribution $p_i \propto e^{-\epsilon_i/kT}$ is a probability distribution that gives the probability that a system will be in a certain state i as a function of that state's energy ϵ_i and the temperature of the system T with k is a Boltzmann constant. In the setting of MRF, the posterior probability distribution $\mathbb{P}(X|Z)$ is a Gibbs distribution

$$\mathbb{P}(X|Z) = \frac{e^{-E(X,Z)}}{Z},$$

where $Z = \sum_X e^{-E(X,Z)}$ specified by an energy function $E(X, Z)$, $X = \{X_i : i \in \mathcal{V}\}$ are random variables specified at each node of the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, the set of nodes $\mathcal{V}$ is the set of image pixels $\mathcal{D}$, and the edges $\mathcal{E}$ are between neighbouring pixels.
- The energy function therefore is

$$E(X, Z) = \sum_{i \in \mathcal{V}} \phi(X_i, Z) + \sum_{ij \in \mathcal{E}} \psi_{ij}(X_i, X_j),$$

where $\phi(X_i, Z)$ is unary evidence for pixel i to have label X_i and $\psi_{ij}(X_i, X_j)$ is pairwise potentials context terms.
- The normalization constant is the marginal likelihood, i.e., the denominator term of posterior $\mathbb{P}(X|Z)$ which is $Z = \sum_X e^{-E(X,Z)}$ specified by an energy function $E(X, Z)$.

3. *Describe a Markov Random Field (MRF). What is the Markov condition?*
- The Markov Random Field (MRF) is a set of random variables having a Markov property described by an undirected graph.
- The Markov condition is a set of properties that make the set of $\mathcal{V}$ form an MRF with respect to $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The Markov conditions include:

- Pairwise Markov condition: For any $i, j \in \mathcal{V}$ not equal or adjacent, $X_i \perp\!\!\!\perp X_j \mid X_{\mathcal{V}\setminus\{i,j\}}$.
- Local Markov condition: For any $i \in \mathcal{V}$ and $J \subset \mathcal{V}$ not containing or adjacent to i, $X_i \perp\!\!\!\perp X_j \mid X_{\mathcal{V}} \setminus (\{i\} \cup J)$.
- Global Markov condition: For any $I, J \subset \mathcal{V}$ not intersecting or adjacent, $X_I \perp\!\!\!\perp X_J \mid X_{\mathcal{V}\setminus\{I\cup J\}}$.

4. *Give two examples of MRFs from the course.*
Example 1: The MRF is the set of discrete-valued images pixel random variable $X_{ij} = \{0, 1\}$ in binary image X described by the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 1, the set of nodes $\mathcal{V}$ is the set of $n \times n$ image pixels, and the edges $\mathcal{E}$ are between neighboring pixels.

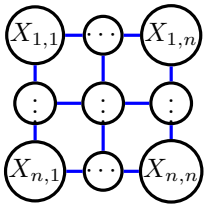

Figure 1: Example of Markov Random Field with Binary image.

Example 2: The MRF is the set of the discrete-valued audio signal random variable $X_i \in \mathbb{R}^{20k} Hz$ in an audio chain X described by the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in Figure 2, the set of nodes $\mathcal{V}$ is the set of n audio signals, and the edges $\mathcal{E}$ are between neighboring signals.
.

Figure 2: Example of Markov Random Field with Audio.

Mean Field Theory: 30 points

1. *What is the core idea of mean field theory (MFT)? How can MFT be used to convert a discrete optimization problem into a continuous one? What is the Kullback-Leibler divergence and how is it used in mean field theory?*
- The main idea of Mean Field Theory (MFT) is instead of directly estimating $\hat{X} = \arg\max_X \mathbb{P}(X \mid Z)$ which includes an intractable normalizing constant, try to find a distribution $\mathbb{Q}(X)$ which approximates $\mathbb{P}(X \mid Z)$, and, from which, $\hat{X} = \arg\max_X \mathbb{Q}(X)$ can be estimated $\mathbb{Q}(X) = \prod_{X_i \in \mathcal{V}} q_i(x_i)$.
- The MFT converts a discrete optimization problem to a continuous one by replacing discrete variables $\hat{X} = \arg\min_X E(X, Z)$ to continuous probability distributions $\hat{\mathbb{Q}} = \arg\min_{\mathbb{Q}} \sum_X \mathbb{Q}(X) \log\left(\frac{\mathbb{Q}(X)}{\mathbb{P}(X|Z)}\right)$.
- The Kullback-Leibler (KL) divergence $D_{KL}(\mathbb{P}||\mathbb{Q})$ is a statiscal distance to measure how probability distribution $\mathbb{P}$ differs with $\mathbb{Q}$. In MFT, is used to measure the distance between $\mathbb{Q}(X)$ and $\mathbb{P}(X \mid Z)$. As a consequence, the optimization problems become

$$\begin{aligned}\hat{\mathbb{Q}} &= \arg\min_{\mathbb{Q}} D_{KL}(\mathbb{P}(X \mid Z)||\mathbb{Q}(X)) = \arg\min_{\mathbb{Q}} \sum_X \mathbb{Q}(X) \log\left(\frac{\mathbb{Q}(X)}{\mathbb{P}(X \mid Z)}\right) \\ &= \arg\min_{\mathbb{Q}} \sum_X \mathbb{Q}(X) \log(\mathbb{Q}(X)) - \sum_X \mathbb{Q}(X) \log(\mathbb{P}(X|Z)).\end{aligned}$$

2. *What is the MFT free energy? Is it convex or not? What strategy can be used to improve performance if it is not convex?*
- The MFT energy is the objective function

$$\mathcal{F}_{MFT}(\mathbb{Q}) = \sum_{ij \in \mathcal{E}} \sum_{x_i, x_j} q_i(x_i) q_j(x_j) \psi_{ij}(x_i, x_j) + \sum_{i \in \mathcal{V}} \sum_{x_i} q_i(x_i) \phi_i(x_i, Z) + \sum_{i \in \mathcal{V}} \sum_{x_i} q_i(x_i) \log(q_i(x_i)).$$

- In general, there is no guarantee $\mathcal{F}_{MFT}(\mathbb{Q})$ is convex. However, if $\mathbb{P}(X \mid Z) = \frac{e^{-E(X,Z)}}{Z}$ and $\mathbb{Q}(X) = \prod_{i\in\mathcal{V}} q_i(x_i)$, then $F_{MFT}(\mathbb{Q})$ is a convex function because

$$\begin{aligned} D_{KL}(\mathbb{P}(X \mid Z)||\mathbb{Q}(X)) &= \sum_X \mathbb{Q}(X)E(X) + \sum_X \mathbb{Q}(X)\log(\mathbb{Q}(X)) + \log(Z) \\ &= \mathcal{F}_{MFT}(\mathbb{Q}) + \log(Z). \end{aligned}$$

Due to $D_{KL}(\mathbb{P}(X \mid Z)||\mathbb{Q}(X))$ is convex and $\log(Z)$ is concave, we obtain $\mathcal{F}_{MFT}(\mathbb{Q})$ is a convex function.
- If $\mathcal{F}_{MFT}(\mathbb{Q})$ is not convex, then we need to adjust the thermodynamic temperature T in the Gibbs distribution $\mathbb{P}(X|Z) \propto e^{-\epsilon_i/kT}$. Because if $kT \to 1$, we will obtain $\mathbb{P}(X \mid Z) \to \frac{e^{-E(X,Z)}}{Z}$ and $\mathcal{F}_{MFT}(\mathbb{Q})$ will become more convex.

3. *Specify an MFT algorithm. What conditions guarantee that an MFT algorithm converges to a local minimum of the free energy?*
- Example of MFT algorithm with binary image $X_i = \{0, 1\}$, then we have $E(X, Z) = \sum_{i\in\mathcal{V}} x_i\psi_i + \sum_{i\in\mathcal{V}, j\in N(i)} x_i x_j \psi_{ij}$, $\psi(x_i) = x_i\psi(x_i = 1) + (1 - x_i)\psi(x_i = 0)$. Let $q_i(x_i = 1) = q_i$, $q_i(x_i = 0) = 1 - q_i$, we have

$$\begin{aligned} F(\mathbb{Q}) &= \sum_X \mathbb{Q}(X)\log(\mathbb{Q}(X)) - \sum_X \mathbb{Q}(X)\log(\mathbb{P}(X|Z)) \\ &= \underbrace{\sum_i [q_i\log(q_i) + (1 - q_i)\log(1 - q_i)]}_{E_{concavce}} \underbrace{- \sum_i q_i\psi_i - \sum_{i,j} q_i q_j \psi_{ij}}_{E_{convex}} - \log(Z). \end{aligned}$$

Since $F(\mathbb{Q})$ is convex, we can apply the Discrete optimization algorithm

$$q_i^{t+1} = \frac{\exp(-\psi_i - \sum_j q_j^t \psi_{ij})}{1 + \exp(-\psi_i - \sum_j q_j^t \psi_{ij})}.$$

- Convergence can be guaranteed by Concave-Convex procedure (CCCP): take the derivative of $dF(\mathbb{Q})$, we get

$$dF(\mathbb{Q}) = \log\frac{q_i}{1 - q_i} - \psi_i - \sum_j q_j\psi_{ij}.$$

Then the MFT algorithm converges to a local minimum of the free energy when $dF(\mathbb{Q}) = 0$, i.e., $\log\frac{q_i}{1-q_i} = \psi_i + \sum_j q_j\psi_{ij}$, i.e,

$$\frac{\partial E_{concavce}}{\partial q_i^t} = \frac{-\partial E_{convex}}{\partial q_i^t}.$$

4. *When can MFT be applied to MRF's with long range interactions efficiently?*
For a long-range interaction efficiently, we can multiply $\frac{\partial \mathcal{F}_{MFT}}{\partial q_i}$ by a positive function to increase the speed of decreasing energy when

$$\frac{d\mathcal{F}_{MFT}}{dt} = \sum_i \frac{\partial \mathcal{F}_{MFT}}{\partial q_i}\frac{dq_i}{dt} = -\sum_i \left(\frac{\partial \mathcal{F}_{MFT}}{\partial q_i}\right)^2$$

is too small.

5. *What is Deterministic Annealing? What is its justification? Is it guaranteed to converge to a global optimum?*
- Deterministic Annealing is a heuristic technique to avoid local minima of the free energy function

$$F(\mathbb{Q}, T) = \sum_X \mathbb{Q}(X)\log(\mathbb{Q}(X)) - \frac{1}{T}\sum_X \mathbb{Q}(X)E(X) - \log(Z),$$

where T is the temperature of probability distribution $\mathbb{P}(X; T) = \frac{e^{-E(X)/T}}{Z(T)}$.
- To do so, it justify temperature T by finds a minimum of $F(\mathbb{Q}, T)$ for large T and gives it initial conditions for minimizing $F(\mathbb{Q}, T)$ at smaller T.
- There is no guarantee that this algorithm converges to the global minimum. However, when T is large enough, $F(\mathbb{Q}, T)$ will become a more convex function.

Belief Propagation: 30 points

1. *Describe the belief propagation algorithm? What are the messages and how do they relate to the marginal probabilities? Under what conditions is it guaranteed to converge to the correct solution?*
- The Belief Propagation (BP) algorithm includes (1) proceeds by passing messages $m_{ij}(x_j : t)$ between the graph nodes i passes to node j to affect state x_j. (2) The messages get updated as follows:

$$m_{ij}(x_j : t+1) = \sum_{x_i} \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k \neq j} m_{ki}(x_i : t). \tag{1}$$

Alternative: the max-product

$$m_{ij}(x_j : t+1) = \max_{x_i} \left\{ \psi_{ij}(x_i, x_j)\psi_i(x_i) \prod_{k \neq j} m_{ki}(x_i : t) \right\}. \tag{2}$$

- The messages is $m_{ij}(x_j : t)$ and if the algorithm converges, then we compute approximations to the marginals:

$$q_i(x_i) \propto \psi_i(x_i) \prod_k m_{ki}(x_i)$$
$$q_{ij}(x_i, x_j) \propto \psi_i(x_i)\psi_{ij}(x_i, x_j)\psi_j(x_j) \prod_{k \neq j} m_{ki}(x_i) \prod_{l \neq i} m_{lj}(x_j).$$

- BP will often converge to a good approximation to the marginals for graphs that do not have closed loops.

2. *What is the Bethe free energy and how does it relate to belief propagation? How can the messages in belief propagation be justified?*
- Bethe free energy is an energy function that has the following form

$$F(\mathbb{Q}) = \sum_{ij} \sum_{x_i, x_j} q_{ij}(x_i, x_j) \ln \frac{q_{ij}(x_i, x_j)}{\psi_i(x_i)\psi_j(x_j)\psi_{ij}(x_i, x_j)} - \sum_i (n_i - 1) \sum_{x_i} q_i(x_i) \ln \frac{q_i(x_i)}{\psi_i(x_i)}.$$

The extreme of this Bethe free energy corresponds to the fixed point of BP.
- The message $m_{ij}(x_j : t)$ in BP can be justified by Equation 1 or Equation 2.

3. *Under what conditions does the Bethe free energy reduce to the MFT free energy?*
- The Bethe free energy reduces to the MFT free energy when $q_{ij}(x_i, x_j) \to q_i(x_i)q_j(x_j)$ because if so, we have

$$\begin{aligned} F(\mathbb{Q}) &= \sum_{ij} \sum_{x_i, x_j} q_{ij}(x_i, x_j) \ln \frac{q_{ij}(x_i, x_j)}{\psi_i(x_i)\psi_j(x_j)\psi_{ij}(x_i, x_j)} - \sum_i (n_i - 1) \sum_{x_i} q_i(x_i) \ln \frac{q_i(x_i)}{\psi_i(x_i)} \\ &= \sum_{ij} \sum_{x_i, x_j} q_i(x_i)q_j(x_j) \ln \frac{q_i(x_i)q_j(x_j)}{\psi_i(x_i)\psi_j(x_j)\psi_{ij}(x_i, x_j)} - \sum_i (n_i - 1) \sum_{x_i} q_i(x_i) \ln \frac{q_i(x_i)}{\psi_i(x_i)} = \mathcal{F}_{MFT}(\mathbb{Q}). \end{aligned}$$

Probability Distributions on Graphs: 30 points

1. *What are the advantages of formulating probability distributions in terms of graphs?*
- Advantages of formulating probability distributions in terms of graphs include:

- It helps represent probability distributions compactly by exploiting the dependencies between variables.
- It helps understand the structure of the data described by the distributions, enabling transferring distributions from one domain to another and reducing the number of parameters to describe the distribution.
- It helps perform inference and to do learning easier.

2. *What is the dynamic programming algorithm? For what class of problems does dynamic programming apply? For these problems, how does it reduce the complexity of computation?*
- Dynamic Programming (DP) is an algorithm that exploits the linear structure to break problems down into sub-components.
- DP can be applied to any graph without closed loops. It can also be extended to some graphs with closed loops – junction trees. It also can be modified to compute other quantities of interest.
- For these problems, assuming we want to minimize

$$\varphi(x_1, \cdots, x_N) = \varphi_{12}(x_1, x_2) + \varphi_{23}(x_2, x_3) + \cdots + \varphi_{N-1N}(x_{N-1}, x_N), x_i \in \{1, \cdots, k\}.$$

DP reduces the complexity of computation from k^N possibles states of $\varphi(x_1, \cdots, x_N)$ to $\sim k^2 N$ by solving

$$\hat{x}_N = \arg\min h_N(x_N) \cdots \hat{x}_{N-1} = \arg\min h_{N-1}(x_{N-1} + \varphi_{N,N-1}(\hat{x}_N))$$

to recovers the states $x_N, x_{N-1}, ..., x_1$.

3. *What is the difference between direct and indirect influence?*
- The direct influence contains the directed graph which captures the causal structure of the variables. Formally, $\mathbb{P}(x_1, \cdots, x_N) = \prod_i \mathbb{P}(x_i \mid \mathbb{P}_a(x_i))$, where $\mathbb{P}_a(x_i)$ are he parents of x_i, the nodes which have directed arcs directly into x_i. The indirect influences in constant, differ by containing the undirected graph in which the edges are not directed. Formally, $\mathbb{P}(x_1, \cdots, x_N) = \frac{1}{Z} \prod_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j) \prod_{i\in\mathcal{V}} \psi_i(x_i)$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges in the undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

4. *What are Stochastic Grammars? What are AND/OR graphs?*
- Stochastic Grammar is a variable topology of Graphical Models by using grammar structure to assign the probability to rules with $A, B, C, \cdots$ as non-terminal nodes and $a, b, c, \cdots$ are terminal nodes, and the prediction rules are $A \to (B, C)$ and $A \to a$.
- AND/OR graph also a variable topology of Graphical Models by using AND/OR operation to assign to AND node and OR node, the OR nodes act as 'switch variables' and the graph topology changes when we select the switch.

Binocular Stereo: 30 points

1. *What is the correspondence problem in binocular stereo? What is the disparity and how does it relate to depth?*
- The correspondence problem in binocular stereo determines for each point in one retinal image (from the left eye camera) which point in the other originated (from the right eye camera) from the same part of the same object.
- Disparity $\{d_i\}$ means matchs point i in left image to point $i + d_i$ in right image. This is related to the depth because estimate disparity $\{d_i\}$ determines depth if matching is known and the depth is estimated by trigonometry.

2. *How can be stereo be formulated in terms of a Markov Random Field (MRF)? and what are the properties of that MRF?*
- The stereo can be formulated in terms of MRF with the Energy function

$$E[\{d_i\}] = \sum_i \Phi(d_i, I_L, I_R) + \sum_i \psi(d_i, d_{i+1}).$$

- The properties of this MRF include $\sum_i \Phi(d_i, I_l, I_r)$ is the data cues (e.g, $|F(I_L)_i - F(I_R)_{i+d_i}|$) and weak smoothness constraint $\sum_i \psi(d_i, d_{i+1})$ (e.g, $K(d_i - d_{i+1})$), where $F(I_L)_i$ is image feature, computed on left image at position i and $F(I_R)_{i+d_i}$ is image feature, computed on right image at position $i + d_i$.

3. *What is the epipolar line constraint, and how does it simplify the correspondence problem? What inference algorithm can be used to solve it? What inference algorithm can be used if the epipolar line constraint is not used?*
- The Epipolar Line Constraint is the geometry that means that a point in the left eye camera can only

match points on one line in the right eye camera. It simplifies the problems by only needing to match a point to the line of points instead of each point in correspondence.
- The stereo-matching inference algorithm can be used to solve this by using a 1-D problem if the epipolar geometry is known.
- If the epipolar line constraint is not used, we can use better stereo algorithms to enforce weak smoothness of disparity to resolve matching ambiguities across the epipolar lines.

4. *Why is Belief Propagation a sensible algorithm to use for binocular stereo?*
- The Belief Propagation (BP) is a sensible algorithm to use for binocular stereo because in the 2D problem, the Energy function

$$E[\{d_{i,j}\}] = \sum_{i,j} \Phi(d_{i,j}, I_L, I_R) + A \sum_{i,j} \psi(d_{i,j}, d_{i+1,j}) + B \sum_{i,j} \psi(d_{i,j}, d_{i,j+1}),$$

the second term is "along eipolar line" and the third term is across eipolar line, and the sum of these two terms will be weak smoothness. Therefore, the graph is no closed loop so BP can effectively estimate $\left\{\hat{d}_{i,j}\right\} = \arg\min E[\{d_{i,j}\}]$.