Lecture 4

Variable Elimination

Prof. Dahua Lin
dhlin@ie.cuhk.edu.hk

Probabilistic Inference

- Given a *probabilistic model*, **inference** is to answer **queries** based on
  - a learned/estimated **model**
  - other observed **evidences**

- In a probabilistic context, inference can be formalized as computing *conditional probabilities*.

$$P(\underbrace{Y}_{query} \mid \underbrace{X = x}_{evidences}; \underbrace{\theta}_{model})$$

Conditional Inference

- **Conditional inference** often refers to the inference *conditioned on* certain *evidences*:

$$P(y|x;\theta) = \frac{P(y,x|\theta)}{P(x|\theta)}$$

with:

$$P(y,x) = \sum_z P(y,x,z|\theta)$$
$$P(x) = \sum_y P(y,x|\theta)$$

- Here, z indicates those variables that are not directly queried or observed, but can influence the computation. They are often referred to as **latent variables** or **hidden variables**.

Marginal Inference

- Sometimes, one may only concern about *marginal probabilities* (those not conditioned on any evidences):

$$P(y) = \sum_z P(y, z)$$

- On graphical models, *conditional inference* can be done via two marginal inference steps:

$$P(y, z, x) \implies P(y, x) \implies P(y|x)$$

  The first step here can usually be simplified via evidence absorption.

Evidence Absorption: Motivating Example

- Consider a Markov network over (X, Y, Z):

$$p(x, y, z) \propto \psi_x(x)\psi_y(y)\psi_z(z)\phi_{xy}(x, y)\phi_{xz}(x, z)\phi_{yz}(y, z)$$

- The conditional probabilities for $p(Y, Z|X = x)$ can be derived as

$$p(y, z|x) \propto \psi_y(y)\psi_z(z)\phi_{y|x}(y)\phi_{z|x}(z)\phi_{yz}(y, z)$$

where $\phi_{y|x}(y) = \phi_{xy}(x, y)$ and $\phi_{z|x}(z) = \phi_{xz}(x, z)$.

- Therefore, for $p(Y|X = x)$, we have

$$p(y|x) = \sum_z p(y, z|x).$$

Evidence Absorption: Generic Procedure

- The procedure of **evidence absorption** can be summarized as:
  - Factors depend purely on known variables: *remove*
  - Factors depend partly on known variables: *reduce*
  - Factors depend purely on unknowns: *retain*
- As conditional inference can be reduced/decomposed into a series of marginal inference. In following discussion, we primarily focus on marginal inference.

Complexity Analysis

- Given a joint distribution $p(Y, Z)$:
  - Y: the *queried variables*.
  - Z: the variables to be marginalized out.
- $P(Y)$ is given by

$$p(y) = \sum_{z \in \mathcal{Z}} p(y, z)$$

  - Need to compute $|\mathcal{Y}|$ values.
  - Each value sums over $|\mathcal{Z}|$ terms.
  - Overall complexity: $|\mathcal{Y}| \cdot |\mathcal{Z}|$, the size of the entire sample space.
  - Grows exponentially as the number of variables increases.

Basic Ideas to Reduce Complexity

- For Markov networks, computation can be restructured into sum of subexpressions, where each *subexpression* depends on a small number of variables.
- *Subexpressions* are reused. By computing these expressions once and caching the results, we can avoid generating them exponentially many times.

Example

- Formulation:

$$p(x, y, z, w) = \frac{1}{Z} f(x, y) g(y, z) h(y, w)$$

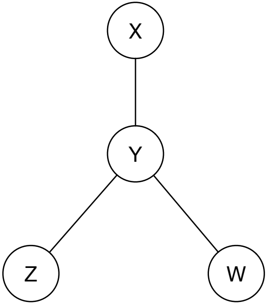
- Naive computation:

$$\tilde{p}(x) = \sum_y \sum_z \sum_w f(x, y) g(y, z) h(y, w)$$

$$Z = \sum_x \tilde{p}(x)$$

$$p(x) = \frac{1}{Z} \tilde{p}(x)$$

- Overall complexity: $O(m_x m_y m_z m_w)$.



Restructured Computation

- Push sums to the right:

$$\tilde{p}(x) = \sum_y f(x,y) \sum_z g(y,z) \sum_w h(y,w)$$

- Detailed analysis:

$$g_{\backslash z}(y) = \sum_z g(y,z) \implies \text{complexity } O(m_y m_z)$$

$$h_{\backslash w}(y) = \sum_w h(y,w) \implies \text{complexity } O(m_y m_w)$$

$$\tilde{p}(x) = \sum_y f(x,y) g_{\backslash z}(y) h_{\backslash w}(y) \implies \text{complexity } O(m_x m_y)$$

- This can be generalized into a systematic procedure to compute marginals – <u>variable elimination</u>.

Variable Elimination: Problem Setup

- Consider a Markov network over $Y_0, Y_1, \ldots, Y_n$, and we intend to compute $P(Y_0)$.
- Initialize:
  - The set of **active factors**: $\mathcal{F} \leftarrow \{\phi_1, \ldots, \phi_m\}$.
  - The set of **active variables**: $\mathcal{V} = \{X, Y_1, \ldots, Y_n\}$.

Variable Elimination: Skeleton

- Given an order π over $\{1, \ldots, n\}$.
- For $j = 1, \ldots, n$:
    - let $i = \pi(j)$
    - Eliminate the variable Y_i:

$$\mathcal{F}, \mathcal{V} \leftarrow \text{EliminateVar}(\mathcal{F}, \mathcal{V}, Y_i)$$

Variable Elimination: EliminateVar

- Notations:
  - $\mathcal{F}(Y_i)$: the set of *active factors* involving Y_i.
  - $\mathcal{V}(\phi)$: the set of *active variables* involved in ϕ.
  - Neighbors of Y_i: $\mathcal{N}_i = \{V \neq Y_i : \exists \phi \in \mathcal{F}(Y_i)\ V \in \mathcal{V}(\phi)\}$.
- Construct ψ_i on $\mathcal{N}_i$:

$$\psi_i(\mathbf{z}) = \sum_{y \in \mathrm{dom}(Y_i)} \prod_{\phi \in \mathcal{F}(Y_i)} \phi\left(y, \mathbf{z}|_{\mathcal{V}(\phi)}\right), \quad \forall \mathbf{z} \in \bigotimes_{j \in \mathcal{N}_i} \mathcal{Y}_j$$

- Set $\mathcal{V} \leftarrow \mathcal{V} \backslash Y_i$.
- Set $\mathcal{F} \leftarrow (\mathcal{F} \backslash \mathcal{F}(Y_i)) \cup \{\psi_i\}$.

Variable Elimination: Compute $P(Y_0)$

- After variable elimination, every active factor that remains in $\mathcal{F}$ involves *only* Y_0.
- Compute *unnormalized probabilities* on Y_0

$$\tilde{p}(y_0) = \prod_{\phi \in \mathcal{F}} \phi(y_0)$$

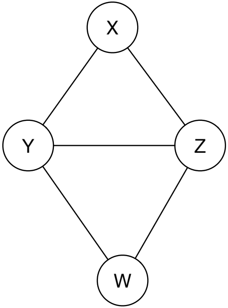
- Compute normalization constant

$$Z_0 = \sum_{y_0} \tilde{p}(y_0)$$

- Normalize the probability values:

$$p(y_0) = \frac{1}{Z_0} \tilde{p}(y_0)$$

Variable Elimination: Example



$$p(x, y, z, w) = \frac{1}{Z}\psi_x(x)\psi_y(y)\psi_z(z)\psi_w(w)$$
$$\phi_{xy}(x, y)\phi_{xz}(x, z)\phi_{yz}(y, z)\phi_{yw}(y, w)\phi_{zw}(z, w)$$

Complexity Analysis

- At each iteration, when we eliminate Y_i, we introduce a new factor on $\mathcal{N}_i$.
- This factor involves $\prod_{j \in \mathcal{N}_i} |\mathcal{Y}_j|$ values, and computing each value requires summing up $|\mathcal{Y}_i|$ product terms.
- The complexity depends on the *maximal cliques* of the induced graphs, which depends strongly on the *elimination order*.

Complexity Analysis (cont'd)

- Finding the *optimal elimination ordering* is in general is *NP-complete*.
- For simple graphs, we can often easily identify a *reasonably good* order of elimination.
- For trees, the *optimal ordering* is to eliminate from *leafs* towards the *root* (*i.e.* the variable of interest).
- *Greedy elimination* often works *reasonably well* in practice.

Questions

Please analyze the complexity of *direct marginal inference* and *variable elimination* for two cases:

- A fully connected Markov network over n discrete variables, each defined on a finite space of cardinality m.

Questions

Please analyze the complexity of *direct marginal inference* and *variable elimination* for two cases:

- A fully connected Markov network over n discrete variables, each defined on a finite space of cardinality m.
- A chain of n discrete variables, each defined on a finite space of cardinality m.

Variable Elimination: Just the first step

- An entire *variable elimination* procedure only computes the probabilities of a single variable (or a small subset of variables).
- Have to re-run the procedure n times if one wants to compute the marginals of all n variables.
- Are there any way to <u>share</u> the computation for all these procedures?