Approximate Inference

Sargur N. Srihari
srihari@cedar.buffalo.edu

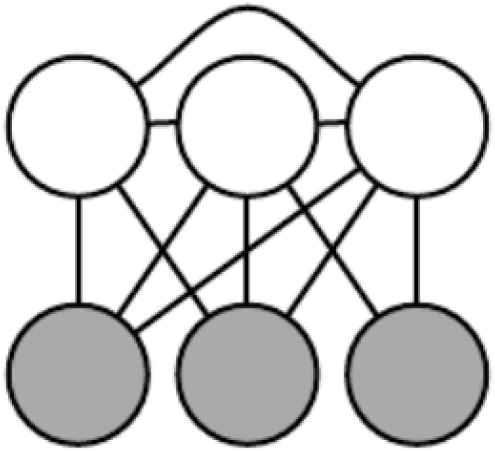
Topics

- Intractability in Inference
- Inference as Optimization
- Expectation Maximization
- MAP Inference and Sparse Coding
- Variational Inference and Learning
- Learned Approximate Inference

Intractable inference in deep learning

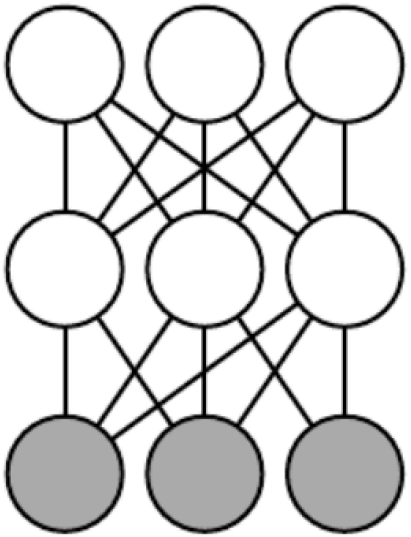
- Intractability arises due to interactions of latent variables in a PGM
  - Undirected PGMs
    - Direct interactions
  - Directed PGMs
    - "Explaining away" interactions between mutual ancestors of the same visible
- Examples are shown next

3 Intractable Inference Problems



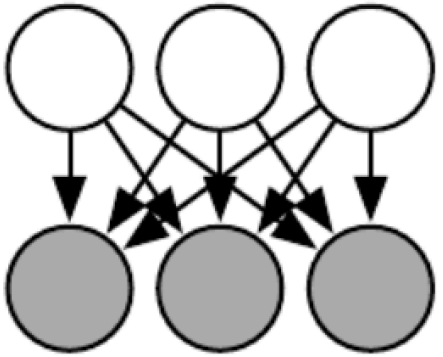
A semi-restricted Boltzmann machine

- With connection Between hidden units Posterior distribution Is intractable due to Large cliques of latent variables



A deep Boltzmann machine

- without intra-layer connections but still is intractable due to between layer connections



A directed model

- with interactions between latent variables when visible units observed

1. Inference as Optimization

- Exact inference can be described as an optimization problem
  - Approximate inference algorithms can be derived by approximating the optimization
- Constructing the optimization problem
  - Assume observed variables $\boldsymbol{v}$ and latent variables $\boldsymbol{h}$
  - We would like to calculate the log probability of the observed data $\log p(\boldsymbol{v} ; \boldsymbol{\theta})$
    - It may be too costly to marginalize $\boldsymbol{h}$
    - Instead we compute a lower bound $\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q)$ on $p(\boldsymbol{v} ; \boldsymbol{\theta})$
    - This is called the evidence lower bound (ELBO)

Evidence Lower Bound

- Evidence Lower Bound (ELBO) defined to be

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \log p(\boldsymbol{v}\,; \boldsymbol{\theta}) - D_{KL}(q(\boldsymbol{h}|\boldsymbol{v}) \,||\, p(\boldsymbol{h}|\boldsymbol{v}\,; \boldsymbol{\theta}))$$

- where q is an arbitrary prob. distribution over h
  - Since KL divergence is always non-negative, $\mathcal{L}$ always has at most the same value as the desired log probability
  - The two are equal if and only if q is the same distribution as $p(\boldsymbol{h}|\boldsymbol{v})$
- Surprisingly $\mathcal{L}$ can be easy to compute for some

Canonical Definition of ELBO

- Simple algebra yields

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) &= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - D_{\mathrm{KL}}(q(\boldsymbol{h} \mid \boldsymbol{v}) \| p(\boldsymbol{h} \mid \boldsymbol{v}; \boldsymbol{\theta})) \\
&= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h} \sim q} \log \frac{q(\boldsymbol{h} \mid \boldsymbol{v})}{p(\boldsymbol{h} \mid \boldsymbol{v})} \\
&= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h} \sim q} \log \frac{q(\boldsymbol{h} \mid \boldsymbol{v})}{\frac{p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta})}{p(\boldsymbol{v}; \boldsymbol{\theta})}} \\
&= \log p(\boldsymbol{v}; \boldsymbol{\theta}) - \mathbb{E}_{\mathbf{h} \sim q} \left[\log q(\boldsymbol{h} \mid \boldsymbol{v}) - \log p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta}) + \log p(\boldsymbol{v}; \boldsymbol{\theta}) \right] \\
&= -\mathbb{E}_{\mathbf{h} \sim q} \left[\log q(\boldsymbol{h} \mid \boldsymbol{v}) - \log p(\boldsymbol{h}, \boldsymbol{v}; \boldsymbol{\theta}) \right] .
\end{aligned}
$$

- This leads to the canonical definition of ELBO

$$\mathcal{L}(\boldsymbol{v}, \boldsymbol{\theta}, q) = \mathrm{E}_{h \sim \mathrm{q}} \left[\log p(\boldsymbol{h}, \boldsymbol{v}) \right] + H(q)$$

- Thus inference is a procedure for finding the q

Inference, Learning and ELBO

- Evidence Lower Bound $\boldsymbol{\mathcal{L}}(\boldsymbol{v}, \boldsymbol{\theta}, q)$ is a lower bound on $\log p(\boldsymbol{h}, \boldsymbol{v})$
- Inference can be viewed as maximizing $\boldsymbol{\mathcal{L}}$ wrt $\boldsymbol{h}$
- Learning can be viewed as maximizing $\boldsymbol{\mathcal{L}}$ wrt $\boldsymbol{\theta}$

Approximate Inference Methods

- Different approximate inference methods
  - different ways of approximate optimization to find q
    - Restrict family of distributions q
    - Perfect optimization procedure using a restricted family of q distributions
    - Imperfect approximation procedure that may not completely maximize $\mathcal{L}$
- Methods
  1. Expectation Maximization
  2. MAP inference and sparse coding
  3. Variational Inference and Learning

Summary of Approx Inference Methods

1. Expectation Maximization
   - A method for learning rather than inference
     - E-step: maximize $\mathcal{L}$ wrt q
     - M-step: maximize $\mathcal{L}$ wrt $\boldsymbol{\theta}$
     - Allows to make large learning steps wrt fixed q
2. MAP inference and sparse coding
   - Learn using a point estimate of $\log p(\boldsymbol{h},\boldsymbol{v})$ rather than inferring the entire distribution
3. Variational Inference and Learning

4. Variational Inference

- Core idea: maximize $\mathcal{L}$ over family of distributions q
  - Family chosen so easy to compute $E_q \log p(\boldsymbol{h}, \boldsymbol{v})$
  - Typical: assumptions over how q factorizes

1. *Mean field approach*: q is a factorial distribution

$$q(\boldsymbol{h}|\boldsymbol{v}) = \prod_i q(h_i|\boldsymbol{v})$$

2. *Structured variational approach*:
   - Impose any PGM on q

Variational: Discrete vs Continuous

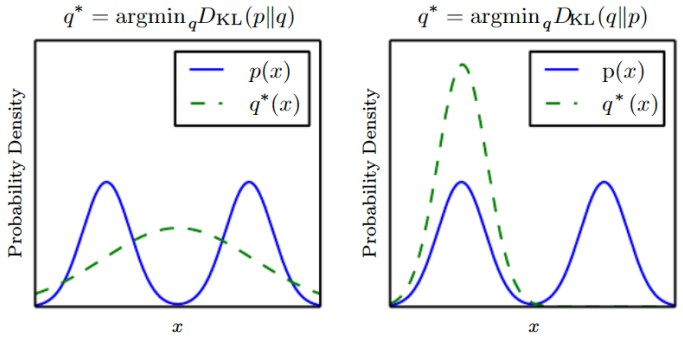
- Specifying how to factorize means:
  - Discrete case:
    - Use traditional optimization techniques to optimize a finite no. of variables describing the q distributions
  - Continuous case:
    - Use calculus of variations over a space of functions
    - Determine which function should be used to represent q
  - Although calculus of variations is not used in discrete case, it is still called as variational
- Calculus of variations is powerful
  - Only need specify factorization, not design of q

Maximizing $\mathcal{L}$ = Minimizing D_{KL}

- Maximizing $\mathcal{L}$ wrt q is equivalent to minimizing $D_{KL}(q(\boldsymbol{h}|\boldsymbol{v})||P(\boldsymbol{h}|\boldsymbol{v})$
  - This is because $\mathcal{L}(\boldsymbol{v},\theta,q) = \log p(\boldsymbol{v};\theta) - D_{KL}(q(\boldsymbol{h}|\boldsymbol{v})||p(\boldsymbol{h}|\boldsymbol{v})$
- Maximum likelihood encourages the model to have high probability everywhere that the data has high probability
- Optimization-based inference encourages q to have low probability everywhere the true posterior has low probability
  - Illustrated in figure next

Asymmetry of KL divergence

- We wish to approximate p with q
  - We have a choice of using $D_{KL}(p||q)$ or $D_{KL}(q||p)$
  - Two Gaussians for p and one Gaussian for q



  - (a) Effect of minimizing $D_{KL}(p||q)$
    - select q that has high probability where p has high probability
    - q chooses to blur the two modes together
  - (b) Effect of minimizing $D_{KL}(q||p)$
    - select q that has low probability where p has low probability
    - Avoids putting probability mass in low probability areas between modes

Topics in Variational Inference

1. Discrete Latent Variables
2. Calculus of Variations
3. Continuous Latent Variables
4. Interactions between Learning and Inference

- We discuss a subset of the above topics here

Calculus of Variations

- A function of a function f is known as a functional $J[f]$
- We can take *functional derivatives* wrt to individual values of of the function $f(\boldsymbol{x})$ at any specific value of $\boldsymbol{x}$
- Denoted by $\frac{\delta}{\delta f(\boldsymbol{x})} J$

Functional Derivative Identity

- For differentiable functions $f(\boldsymbol{x})$ and differentiable functions $g(y, \boldsymbol{x})$ with continuous derivatives

$$\frac{\delta}{\delta f(\boldsymbol{x})} \int g(f(\boldsymbol{x}), \boldsymbol{x})\, d\boldsymbol{x} = \frac{\partial}{\partial y} g(f(\boldsymbol{x}), \boldsymbol{x})$$

- Intuition: $f(\boldsymbol{x})$ is an infinite vector indexed by $\boldsymbol{x}$
  - Identity is same as for a vector $\theta \in R^n$ indexed by positive integers

$$\frac{\partial}{\partial \theta_i} \sum_j g(\theta_j, j) = \frac{\partial}{\partial \theta_i} g(\theta_i, i)$$

- More general is the Euler-Lagrange equation
  - Allows g to depend on derivatives of f as well as value of f, but not needed for deep learning

Optimization wrt a vector

- Procedure to optimize a function wrt a vector
  - Take the gradient of the function wrt the vector
  - Solve for the point where every element of the gradient is equal to zero
- Procedure to optimize a functional
  - Solve for the function where the functional derivative at every point is equal to zero

Example of Optimizing a Functional

- Find probability distribution over $x \in R$ that has the maximum entropy
- Entropy of probability distribution $p(x)$ is defined as $H[p] = -E_x \log p(x)$
- For continuous values the expectation is an integral $H[p] = \int p(x) \log p(x) \, dx$

De

Lagrangian contraints for maxent

1. To ensure that the result is a distribution we add constraint that $p(x)$ integrates to 1
2. Since entropy increases unbounded with variance, seek distribution with variance σ^2
3. Since problem is undetermined
   – as distribution can be arbitrarily shifted without changing entropy we fix the mean to be μ

- Lagrangian functional is

$$L[p] = \lambda_1 \left(\int p(x)\,dx - 1 \right) + \lambda_2 (E[x] - \mu) + \lambda_3 (E[(x-\mu)^2] - \sigma^2) + H[p]$$

$$L[p] = \int (\lambda_1 p(x)\,dx + \lambda_2 p(x) x + \lambda_3 p(x)(x - \mu^2) - p(x) \log p(x))\,dx - \lambda_1 - \mu\lambda_2 - \sigma^2\lambda_3$$

Functional derivative of Lagrangian

- To minimize Lagrangian wrt p, we set the functional derivatives equal to 0:

$$L[p] = \int (\lambda_1 p(x)\,dx + \lambda_2 p(x)x + \lambda_3 p(x)(x-\mu^2) - p(x)\log p(x))\,dx - \lambda_1 - \mu\lambda_2 - \sigma^2\lambda_3$$

$$\boxed{\frac{\delta}{\delta f(\boldsymbol{x})}\int g(f(\boldsymbol{x}),\boldsymbol{x})\,d\boldsymbol{x} = \frac{\partial}{\partial y} g(f(\boldsymbol{x}),\boldsymbol{x})}$$

$y=p(x)$

$g_1(y,\boldsymbol{x})=y$
$g_2(y,x)=yx$
$g_3(y,x)=y(x-\boldsymbol{\mu})^2$
$g_4(y,x)=y\log y$

$$\frac{d}{dy}(y\log y) = 1+\log y$$

$$\forall x, \frac{\delta}{\delta p(x)} L = \lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1 - \log p(x) = 0$$

- This condition tells us functional form of $p(x)$
  - By algebraic rearrangement

$$p(x) = \exp(\lambda_1 + \lambda_2 x + \lambda_3 (x-\mu)^2 - 1)$$

  - We never assumed directly that $p(x)$ has this form

Choosing λ values

- We must choose λ values so that all constraints are satisfied
- We may set $\lambda_1 = 1 - \log \sigma\sqrt{2\pi}$. $\lambda_2 = 0$, $\lambda_3 = -1/2\sigma^2$ to obtain $p(x) = \mathrm{N}(x; \mu, \sigma^2)$
- Because the normal distribution has maximum entropy, we impose the least possible structure by using maximum entropy

Continuous Latent Variables

- When our graphical model contains continuous latent variables, we can perform variational inference and learning by maximizing $\mathcal{L}$
- We must now use calculus of variations when maximizing $\mathcal{L}$ with respect to $q(\boldsymbol{h}|\boldsymbol{v})$
- Not necessary for practitioners to solve calculus of variations problems
- Instead there is a general equation for mean-field fixed point updates

Optimal $q(h_i|\boldsymbol{v})$ for Mean Field

- If we make the mean field approximation
$$q(\boldsymbol{h}|\boldsymbol{v}) = \prod_i q(h_i|\boldsymbol{v})$$
- The optimal $q(h_i|\boldsymbol{v})$, if we fix $q(h_j|\boldsymbol{v})$ for all $j \neq i$, is obtained by normalizing the equation $\boxed{q(h_i \mid \boldsymbol{v}) = \exp(E_{hi \sim q(\boldsymbol{h}_{-i}|\boldsymbol{v})} \log p(\boldsymbol{v}, \boldsymbol{h}))}$

So long as p does not assign 0 probability to any joint configuration of variables.
Carrying out expectation yields correct functional form of $q(h_i|\boldsymbol{v})$

Designed to be iteratively applied for each value of i until convergence

Ex: Latent $\boldsymbol{h} \in \mathbb{R}^2$ and one visible variable v