# The Variational Approximation for Bayesian Inference

## Life after the EM Algorithm

Dimitris G. Tzikas, Aristidis C. Likas, and Nikolaos P. Galatsanos

CMI Lab Journal Club, 11/22/2019, Andrew Van

This subslide runs a bunch of scripts for this presentation.

# Motivation

- Variational Methods?
    - **How does Variational Expectation-Maximization work?**
    - This applies to anything with the word "Variational" (e.g. Variational Autoencoder)
- Succinctly...
    - Variational methods allow for the approximation of the posterior of latent variables
    - Derives a lower bound for the marginal likelihood of the observed data

# Introduction

- Expectation Maximization (EM) is awesome!
    - Applications: Joint problems in image reconstruction, segmentation, registration, etc.
    - We use it to find estimates of model parameters that rely on unobserved latent variables
    - However, limited applicability when encountering complex models
- Variational approximation relaxes requirements of EM
    - Can be applied to a wider range of models
    - EM can be viewed as a special case of Variational Bayesian Inference
        - a.k.a Varitational EM (VEM)

# Expectation-Maximization

We want to find:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, p(x; \theta)$$

where $x$ are observations, and $\theta$ are parameters.

In EM, we introduce unobserved latent variables (by marginalizing the latent variables):

$$p(x; \theta) = \int p(x, z; \theta) dz$$

This expression is known as the **marginal likelihood**.

# Why it's called Bayesian Inference

Using the Marginal Likelihood and Bayes Rule, we can do inference on posterior of latent variables, $p(z|x;\theta)$

Marginal Likelihood:

$$p(x;\theta) = \int p(x,z;\theta)dz = \int p(x|z;\theta)p(z;\theta)dz$$

Bayes Rule:

$$p(z|x;\theta) = \frac{p(x|z;\theta)p(z;\theta)}{p(x;\theta)}$$

- However, it's difficult to solve the integral in the Marginal Likelihood term
    - Effort in developing techniques to get around or approximate integral
        - Numerical Sampling (e.g Monte Carlo) or Deterministic Approximation(Variational Methods)

# Bayesian Inference vs. MAP Estimation

- What is the difference between MAP estimation and Bayesian Inference?
- Both are Bayesian because they place priors on parameters, $\theta$ right?

MAP Estimation:

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(x|\theta)p(\theta)$$

Bayesian Inference:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta)p(\theta)d\theta)}$$

- They are distinctly different in objective!
  - MAP generates a point estimate on $\theta$ (the mode of the posterior) while Bayesian Inference calculates the full posterior distribution.
  - MAP is also known as "Poor Man's Bayesian Inference"

# Evidence Lower Bound (ELBO)

The marginal likelihood can be rewritten as:

$$\ln p(x;\theta) = F(q,\theta) + KL(q||p)$$

with:

$$F(q,\theta) = \int q(z) \ln \Big( \frac{p(x,z;\theta)}{q(z)} \Big) dz$$

- Since the KL divergence must be $\geq 0$, it follows that $\ln p(x;\theta) \geq F(q,\theta)$
- $F(q,\theta)$ is a lower bound of the log-likelihood (known as the Evidence Lower Bound (ELBO))

# ELBO Derivation

# Maximizing the ELBO

We want to maximize the ELBO:

$$F(q, \theta) = \int q(z) \ln\left(\frac{p(x, z; \theta)}{q(z)}\right) dz$$

- We use a two-step process to maximize the lower bound (given starting parameters $\theta^{OLD}$)
  - Step 1: Maximize $F(q, \theta^{OLD})$ w/ respect to $q(z)$
  - Step 2: Maximize $F(q, \theta)$ w/ respect to $\theta$ to generate $\theta^{NEW}$
- The lower bound is the same as maximizing the log likelihood when $KL(q||p) = 0$
  - This implies that $q(z) = p(z|x; \theta^{OLD})$

# Maximizing the ELBO (continued)

Setting the latent posterior ($p(z|x;\theta^{OLD})$) to be $q(z)$, we derive the familiar form for the EM algorithm:

$$\begin{aligned} F(q,\theta) &= \int p(z|x;\theta^{OLD}) \ln p(x,z;\theta) dz - \int p(z|x;\theta^{OLD}) \ln p(z|x;\theta^{OLD}) dz \\ &= \langle \ln p(x,z;\theta) \rangle_{p(z|x;\theta^{OLD})} + constant \\ &= Q(\theta,\theta^{OLD}) + constant \end{aligned}$$

- The EM algorithm
  - E-step: Compute $Q(\theta,\theta^{OLD})$
  - M-step: Find $\theta$ that maximizes $Q(\theta,\theta^{OLD})$
- The trick is that we must explicilty know $p(z|x;\theta)$ to compute $Q(\theta,\theta^{OLD})$ (or $F(q,\theta)$)
  - Not always applicable in all problems, and we can't use EM :(

## Variational Approximation

- We can bypass knowing $p(z|x;\theta)$ exactly by assuming $q(z)$ has a specific form and optimize over $F(q,\theta)$ using calculus of variations
    - Thus the name "Variational Approximation"
- Common form used: Mean Field approximation:

$$q(z) = \prod_{i=1}^{M} q_i(z_i)$$

# Variational Approximation (Continued)

Applying the form $q(z) = \prod_{i=1}^{M} q_i(z_i)$ specified by the mean field approximation, we get the optimal $q(z)$ that maximizes the lower bound:

$$q_j^*(z_j) = \frac{\exp(\langle \ln p(x, z; \theta) \rangle_{i \neq j})}{\int \exp(\langle \ln p(x, z; \theta) \rangle_{i \neq j}) dz_j}$$

for each latent variable $j = 1, \ldots, M$

# Mean Field Solution Derivation

$$F(q,\theta) = \int q(z) \ln\Big(\frac{p(x,z;\theta)}{q(z)}\Big) dz$$

$$= \int \prod_i q(z_i) \ln p(x,z;\theta) dz - \sum_i \int q(z_i) \ln q(z_i) dz_i$$

$$= \int q(z_j) \int \Big(\prod_{i\neq j} q(z_i) \ln p(x,z;\theta)\Big) \prod_{i\neq j} dz_i dz_j - \int q(z_j) \ln q(z_j) dz_j - \sum_{i\neq j} \int q(z_i) \ln q(z_i) dz_i$$

$$= \int q(z_j) \ln\Big(\frac{\exp(\langle \ln p(x,z;\theta)\rangle_{i\neq j})}{q(z_j)}\Big) dz_j - \sum_{i\neq j} \int q(z_i) \ln q(z_i) dz_i$$

# Mean Field Solution Derivation (Continued)

$$= \int q(z_j) \ln\Big(\frac{\tilde{p}(x, z_j; \theta)}{q(z_j)}\Big) dz_j - \sum_{i \neq j} \int q(z_i) \ln q(z_i) dz_i$$

$$= -KL(q_j || \tilde{p}) - \sum_{i \neq j} \int q(z_i) \ln q(z_i) dz_i$$

where:

$$\ln \tilde{p}(x, z_j; \theta) = \langle \ln p(x, z; \theta) \rangle_{i \neq j} = \int \ln p(x, z; \theta) \prod_{i \neq j} (q_i dz_i$$

## Mean Field Solution Derivation (Continued 2)

Like before, $F(q, \theta)$ is maximized when the KL divergence is 0, which occurs when $q_j(z_j) = \tilde{p}(x, z_j; \theta)$, so:

$$\ln q_j^*(z_j) = \langle \ln p(x, z; \theta) \rangle_{i \neq j} + constant$$

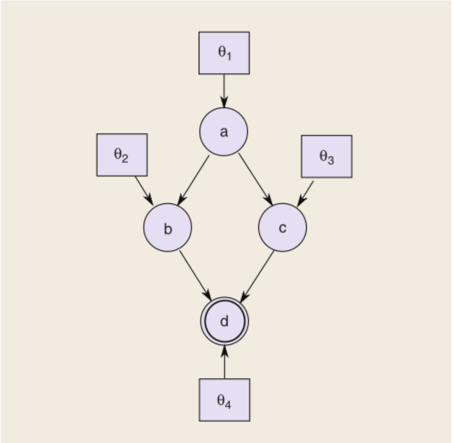
The additive constant can be obtained through normalization, so the final solution form is:

$$q_j^*(z_j) = \frac{\exp(\langle \ln p(x, z; \theta) \rangle_{i \neq j})}{\int \exp(\langle \ln p(x, z; \theta) \rangle_{i \neq j}) dz_j}$$

## Variational EM

- With the mean field form of $q(z)$, the Variation EM (VEM) algorithm can be summarized as:
  - E-step: Evaluate $q^{NEW}(z)$ to maximize $F(q, \theta^{OLD})$
  - M-step: Find $\theta^{NEW}$ that maximizes $F(q^{NEW}, \theta)$
- A common case is where the model only contains latent variables and no parameters
  - In this case, the VEM algorithm only has a Expectation step and no Maximization step

# A quick background on graphical models

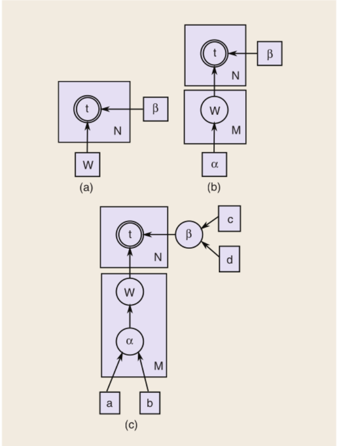



[FIG2] Example of directed graphical model. Nodes denoted with circles correspond to random variables, while nodes denoted with squares correspond to parameters of the model. Doubly circled nodes represent observed random variables, while single circled nodes represent hidden random variables.

- Graphical models represent dependencies among random variables and parameters
  - Double Circle: Observations
  - Circle: Latent Variables
  - Squares: Parameters
- Example (left):
  - Each graph node $(s)$ has a set of parents $\pi(s)$ they are conditioned on: $p(x_s|x_{\pi(s)})$
  - Joint distribution can be defined by: $p(x;\theta) = \prod_s p(x_s|x_{\pi(s)})$
  - So the joint distribution of the left graph can be defined as: $p(a,b,c,d;\theta) = p(a;\theta_1)p(b|a;\theta_2)p(c|a;\theta_3)p(d|b,c;\theta_4)$
- In VEM, do the E-step on circles (Latent Variables) and M-step on squares (parameters)

# Example 1 (Linear Regression)

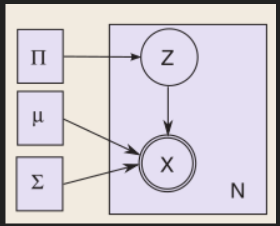
- [GitHub (https://github.com/vanandrew/variational_em_demo/blob/master/regression-example.ipynb)](https://github.com/vanandrew/variational_em_demo/blob/master/regression-example.ipynb)

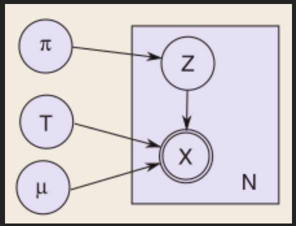



[FIG3] Graphical models for linear regression solved using (a) direct ML estimation (model without prior), (b) EM (model with stationary prior), and (c) variational EM (model with hierarchical prior).

# Example 2 (Gaussian Mixture Model)

- [GitHub (https://github.com/vanandrew/variational_em_demo/blob/master/gmm-example.ipynb)](https://github.com/vanandrew/variational_em_demo/blob/master/gmm-example.ipynb)




GMM Model using EM    Full Bayesian GMM Model using VEM

## Useful Links and References

This paper: https://ieeexplore.ieee.org/document/4644060 (https://ieeexplore.ieee.org/document/4644060)

Derivations for linear regression solutions: https://arxiv.org/abs/1301.3838 (https://arxiv.org/abs/1301.3838)

Latex derivations: https://chrischoy.github.io/research/Expectation-Maximization-and-Variational-Inference/ (https://chrischoy.github.io/research/Expectation-Maximization-and-Variational-Inference/)