Probabilistic Graphical Models for Image Analysis - Lecture 2

Stefan Bauer

28 September 2018

Max Planck ETH Center for Learning Systems

Overview

1. Expectation Maximization
2. Variational Inference

Expectation Maximization

Motivation

Probabilistic Models are often quite complex, thus inference is often challenging or even infeasible -> thus we often approximate solutions to the inference problem using

- **sampling** or
- **variational**

based methods.

Problem: In practical applications we do not observe everything; on the contrary, we are often interested in unobserved variables, which we can not measure!

Today: Learning in latent variable models using variational inference.

Latent Variable Model

A latent variable model is a probability distribution over observed and unobserved variables $p(x, z; \theta)$, where as before $\mathcal{D} = (x^1, \dots, x^n) \in X^n$ are our observations and K variables z_i are unobserved.

Example: **Gaussian Mixture Models** -> allow to model subpopulations in the data (e.g. in Object Tracking, Speech, etc.) The joint distribution is $p(x, z) = p(x|z)p(z)$, where cluster membership assignment is a random variable z_i with $p(x|z = k) \sim \mathcal{N}(\mu_k, \sigma_k)$.

$$p(x) = \sum_{k=1}^{K} p(x|z = k)p(z = k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \sigma_k)$$

The EM Algorithm: Maximum Likelihood Learning with Hidden Variables

Need to maximize

$$\log p(\mathcal{D}) = \sum_{x \in D} \log p(x) = \sum_{x \in D} \log \left(\sum_{z} p(x|z) p(z) \right)$$

Problem: Only x is observed but we have parameters θ and latent variables z

The Expectation Maximization (EM) algorithm:

- **Expectation**: Assign values to hidden/missing variables i.e. compute $p(z|x;\theta_t)$
- **Maximization**: Maximize parameter log likelihood via $\theta_{t+1} = \arg\max_{\theta} \sum_{x \in D} \mathbb{E}_{z \sim p(z|x,\theta_t)} \log p(x, z, \theta)$
- Repeat until convergence for $t = 1, 2, \cdots$, starting with θ_0

Example: EM for Gaussian Mixtures

E-Step: $p(z_j|x;\theta_t) = \frac{p(z_j,x,\theta_t)}{p(x,\theta_t)} = \frac{p(x|z_j,\theta_t)p(z_j,\theta_t)}{\sum_{k=1}^{K} p(x|z_k,\theta_t)p(z_k,\theta_t)} =: \omega_j(x)$

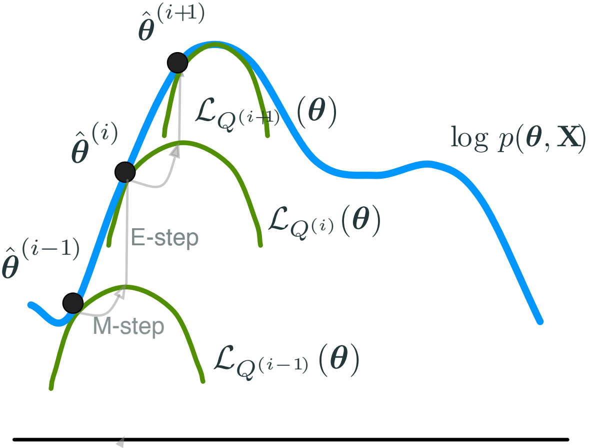
Interpretation: z is indicator for cluster assignment of and in the E-step we thus calculate a membership "weight" ω_j of belonging to the j-th cluster for each data point x.

M-Step:

$$\begin{aligned}\theta_{t+1} &= \arg\max_\theta \sum_{x\in D} \mathbb{E}_{z\sim p(z|x,\theta_t)} \log p(x,z,\theta) \\ &= \arg\max_\theta \sum_{k=1}^{K}\sum_{x\in D} p(z_k|x,\theta_t)\log p(x|z_k,\theta) \\ &\quad + \sum_{k=1}^{K}\sum_{x\in D} p(z_k|x,\theta_t)\log p(z_k,\theta)\end{aligned}$$

Exercise: Derive the precise equations for the M-Step.

Illustration EM



Summary EM

- EM is a general framework for partially observable data
- Idea of maximizing the log-likelihood given the "expected complete" dataset.
- Various extensions: Stochastic EM, Hard EM, Neural EM
- Local optima: initialization often important
- The marginal likelihood increases after each EM cycle!

Question: Why does it work?

Variational Inference

Motivation and Recall

- A probabilistic model is a joint distribution of hidden variables z and observed variables x:

$$p(z, x)$$

- Inference about the unknowns is through the *posterior*, the conditional distribution of the hidden variables given the observations

$$p(z \mid x) = \frac{p(z, x)}{p(x)}$$

- For most interesting models, the denominator is not tractable.

Idea

X observations, Z hidden variables, θ additional parameters

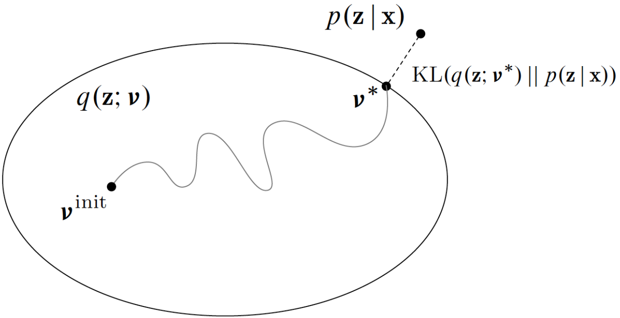
$$p(z \mid x, \alpha) = \frac{p(z, x \mid \theta)}{\int p(z, x \mid \theta)} \tag{1}$$

Idea: Pick family of distributions over latent variables with its own variational parameter

$$q(z \mid \nu) = \ldots?$$

and find variational parameters ν such that q and p are "close".

Variational Inference - Concept



*

Variational Inference

- VI turns inference into optimization.
- Place a variational family of distributions over latent variables.
- Fit the variational parameters to be close (in KL)

*Figure from Blei et.al, Variational Inference Tutorial, Nips 2016

Convexity

Definition

Let f be a real valued function defined on an interval $I = [a, b]$, then f is said to be convex on I if $\forall x_1, x_2 \in I$ and $lambda \in [0, 1]$, we have:

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \tag{2}$$

A function f is concave if $-f$ is convex.

Intuition of Convexity: The function is never above the straight line from points $(x_1, f(x_1))$ to $(x_2, f(x_2))$.

Jensen's Inequality

Theorem

Let f be a convex function defined on an interval I. If $x_1, x_2, \cdots, x_n \in I$ and $\lambda_1, \lambda_2, \cdots, \lambda_n \geq 0$ with $\sum_{i=1}^{n} = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f(x_i) \tag{3}$$

Proof: Induction; n=1 trivial, n=2 definition of convexity, for n+1 (Black Board).

Exercise: $-\log(x)$ is a convex function on $(0, \infty)$.

Derivation

Let $q(z)$ be some probability distribution on z. Then

$$
\begin{aligned}
\log p(x,\theta) &= \int q(z) \log p(x,\theta) dz = \\
&= \int q(z) \log\left(\frac{p(x,\theta)p(z|x,\theta)}{p(z|x,\theta)}\right) dz \\
&= \int q(z) \log\left(\frac{p(x,z,\theta)}{p(z|x,\theta)}\right) dz \\
&= \int q(z) \log\left(\frac{p(x,z,\theta)q(z)}{p(z|x,\theta)q(z)}\right) dz \\
&= \int q(z) \log\left(\frac{p(x,z,\theta)}{q(z)}\right) dz - \int q(z) \log\left(\frac{p(z|x,\theta)}{q(z)}\right) dz \\
&=: \mathrm{ELBO}(q,\theta) + \mathrm{KL}(q(z)||p(z|x,\theta))
\end{aligned}
$$

By Jensen's inequality the KL divergence is non-negative and thus the first term is a lower bound (so called Evidence Lower Bound). -> What is $q(z)$?

Revisiting Expectation Maximization

If $p(z|x,\theta_t)$ can be analytically calculated, we can substitute $q(z) := p(z|x,\theta_t)$:

$$\begin{aligned}
\mathrm{ELBO}(q,\theta) &= \int q(z)\log\left(\frac{p(x,z,\theta)}{q(z)}\right)dz \\
&= \int q(z)\log p(x,z,\theta)dz - \int q(z)\log q(z)dz \\
&= \int p(z|x,\theta_t)\log p(x,z,\theta)dz \\
&- \int p(z|x,\theta_t)\log p(z|x,\theta_t)dz \\
&= \mathcal{Q}(\theta,\theta_t) + \mathcal{H}(z|x)
\end{aligned}$$

The second term $\mathcal{H}(z|x)$ is called the entropy of z. Note: It is just a function of θ_t not θ.

Revisiting Expectation Maximization

The Expectation Maximization (EM) algorithm maximizes the evidence lower bound

$$\mathrm{ELBO} = \int q(z) \log \left(\frac{p(x, z, \theta)}{q(z)} \right) dz = \mathbb{E}_q[\log p(x, z, \theta) - \log q(z)] \tag{4}$$

instead of directly optimizing

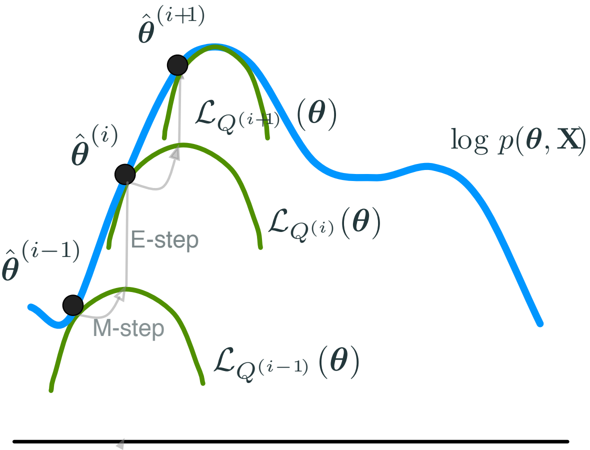
$$\log p(x, \theta) = \mathrm{ELBO}(q, \theta) + \mathrm{KL}(q(z)||p(z|x, \theta))$$

Note: The KL is non-negative thus the ELBO is maximal when $q = p(z|x, \theta)$ -> so called tight lower bound.

Recall (Exercise): $p(z|x; \theta_t)$ can be analytically calculated for the Gaussian Mixture Model.

- **E-Step**: compute posterior $p(z|x; \theta_t)$ and evaluate ELBO for $q = p(z|x, \theta)$ (tight ELBO).
- **M-Step**: $\theta_{t+1} = \arg\max_\theta \int p(z|x; \theta_t) \log p(x, z, \theta) dz$

Illustration EM



Mean-field Variational Inference

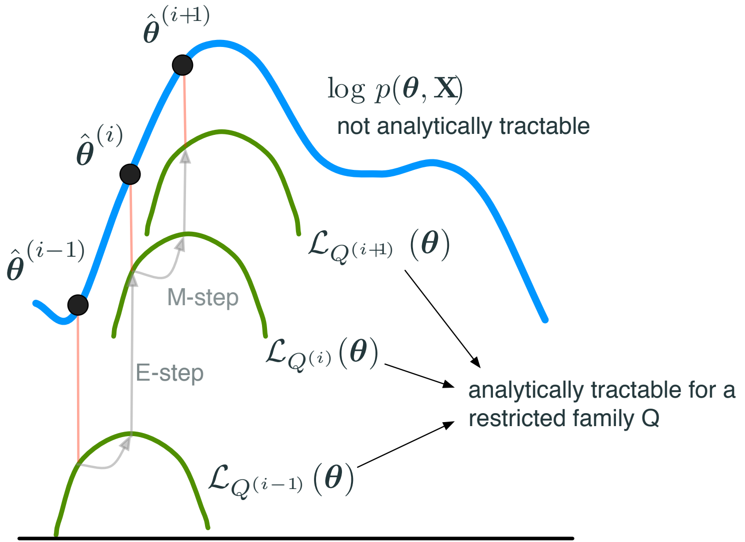
Problem for EM: What can we do if we can not find a closed form for $p(z|x;\theta_t)$?

Idea: Choose/design variational family Q such that the expectations are easily computable!

$$q(z_1, \cdots, z_k) = \prod_{i=1}^{k} q(z_i) \tag{5}$$

- It does not contain the true posterior since the variables are dependent which can now not be captured by q.
- Offers the possibility to group variables together.

Illustration Mean Field Approximations



ELBO for mean field approximation

$$
\begin{aligned}
\mathrm{ELBO}(q,\theta) &= \int q(z)\log\left(\frac{p(x,z,\theta)}{q(z)}\right)dz \\
&= \int \prod_i q(z_i)\log p(x,z,\theta)dz - \sum_i \int q(z_i)\log q(z_i)dz \\
&= \int q(z_j)\int \prod_{i\neq j} q(z_i)\log p(x,z,\theta)\prod_{i\neq j} dz_i dz_j \\
&\quad - \int q(z_j)\log q(z_j)dz_j - \sum_{i\neq j}\int q(z_i)\log q(z_i)dz_i \\
&= \int q(z_j)\log\left(\frac{\exp(\mathbb{E}_{i\neq j}\log p(x,z,\theta))}{q(z_j)}\right)dz_j \\
&\quad - \sum_{i\neq j}\int q(z_i)\log q(z_i)dz_i =: -\mathrm{KL}(q_j||\tilde{p}_{i\neq j}) + \mathcal{H}(z_{i\neq j}) + c
\end{aligned}
$$

c normalization constant.

Coordinate Ascent

Again: KL-divergence is non-negative and thus the ELBO is maximal when

$$q(z_j) = \tilde{p}_{i \neq j} = \frac{1}{Z} \mathbb{E}_{i \neq j}(\log p(x, z, \theta)) \tag{6}$$

Finally, once again:

- **E-Step**: $\forall j$ evaluate $q^\star(z_j) = \frac{1}{Z} \mathbb{E}_{i \neq j}(\log p(x, z, \theta))$ and set $q^{t+1} = \prod_i q_i^\star$
- **M-Step**:Find $\theta_{t+1} = \arg\max_\theta \mathrm{ELBO}(q^{t+1}, \theta)$

Example Bayesian Mixture of Gaussians

Exercise: Gaussian Mixture Model with Dirichlet prior on the weights.

Summary

- Choose variational family q
- Derive the ELBO
- Coordinate ascent of each q_i
- Repeat until convergence

Deterministic and fast (unlike MCMC), often works well in practice, multiple (parallel) initializations needed (local minima), the ELBO is not always "easy" to derive (Exercise).

Key idea: Bounding by convexity!

Next week & Open Master's thesis topics

Open master's thesis topics

- MuJoCo Simulator for a Robotic Platform
- Online Outlier Detection
- Sleep Classification

Plan for next week

- Black Box Variational Inference
- Stochastic Variational Inference
- Belief Propagation and Expectation Propagation

Questions?