Probabilistic Graphical Models for Image Analysis - Lecture 4

Stefan Bauer

12 October 2018

Max Planck ETH Center for Learning Systems

Overview

Repetition

Framework

x observations, Z hidden variables, α additional parameters

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

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".

Mean-field

Assumes that each variable is independent:

$$q(z_1, \cdots, z_k) = \prod_{i=1}^{k} q(z_i) \tag{2}$$

- 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.
- Use coordinate ascent updates for each $q(z_k)$.
- Here we only specified factorization but **not** the form of the $q(z_k)$.

Using Jensen's inequality to obtain a lower bound

$$\begin{aligned}\log p(x) &= \log \int_Z p(z,x) = \\ &= \log \int_Z p(z,x)\frac{q(z)}{q(z)} \\ &= \log \mathbb{E}_q\left(\frac{p(x,Z)}{q(Z)}\right) \\ &\geq \mathbb{E}_q\left(\log p(x,Z)\right) - \mathbb{E}_q\left(\log q(Z)\right)\end{aligned}$$

Proposal: Choose/design variational family Q such that the expectations are easily computable.

Relation with KL

$$
\begin{aligned}
\mathrm{KL}[q(z) \mid\mid p(z \mid y)] &= \mathbb{E}_q\left[\log \frac{q(Z)}{p(Z \mid y)}\right] \\
&= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z \mid y)] \\
&= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z, y)] + \log p(y) \\
&= -\left(\mathbb{E}_q[\log p(Z, y)] - \mathbb{E}_q[\log q(Z)]\right) + \log p(y)
\end{aligned}
$$

Difference between KL and ELBO is precisely the log normalizer, which does not depend on q and is bounded by the ELBO.

α-Divergence

α-Divergence

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} dx \tag{3}$$

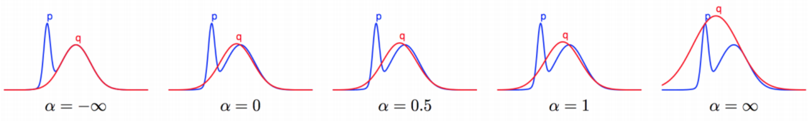
with $\alpha \in (-\infty, \infty)$.

Properties:

- $D_\alpha(p||q)$ is convex with respect to both q and p.
- $D_\alpha(p||q) \geq 0$
- $D_\alpha(p||q) = 0$ when $q = p$
- $\lim_{\alpha\to 0} D_\alpha(p||q) = \mathrm{KL}(q||p)$
- $\lim_{\alpha\to 1} D_\alpha(p||q) = \mathrm{KL}(p||q)$

Inclusive v.s Exclusive

Take again Gaussian q approximating two mode Gaussian p from before:



- For $\alpha \geq 1$: **inclusive** since it prefers to stretch across p.
- For $\alpha \leq 0$: **exclusive**, q seeks area of largest total mass, mode seeking.

Expectation Propagation

Instead of minimizing with respect to $\mathrm{KL}(q||p)$ we minimize wrt. $\mathrm{KL}(p||q)$.

Assume q is a member of the exponential family i.e.:

$$q(\mathbf{x}|\eta) = h(\mathbf{x})g(\eta)\exp(\eta^\intercal \mathbf{u}(\mathbf{x}))$$

Then

$$\mathrm{KL}(p||q)(\eta) = -\log g(\eta) - \eta^\intercal \mathbb{E}_{p(z)}[\mathbf{u}(\mathbf{z})] + const.$$

Minimize KL by setting gradient to zero:

$$-\nabla g(\eta)! = \mathbb{E}_{p(z)}[\mathbf{u}(\mathbf{z})]$$

from previous slide $-\nabla \log g(\eta) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$, we can then conclude (moment matching)

$$\mathbb{E}_{q(z)}[\mathbf{u}(\mathbf{z})] = \mathbb{E}_{p(z)}[\mathbf{u}(\mathbf{z})]$$

Expectation Propagation - Summary

- The reversed KL is harder to optimized. If the true posterior p factorizes, then we can update single factors iteratively by moment matching.
- Factors are in the exponential family.
- There is no guarantee that the iterations will converge.
- Expectation Propagation aims to preserve the marginals!
- Restriction to exponential family in EP implies: Any product and any division between distributions stays in parametric family and can be done analytically.
- Main applications involve Gaussian Processes and logistic regression (less well suited for GMM).

Variational Inference

Framework

We use variational inference to approximate the posterior distribution

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

$$\log p(x, \theta) \geq \mathbb{E}_q[\log p(Z, x)] - \mathbb{E}_q[\log q(Z)]$$

To optimize the lower bound, we can use coordinate ascent!

Problems:

- In each iteration we go over all the data!
- Computing the gradient of the expectations above.

Coordinate Ascent

Given the independence assumption, we can decompose the ELBO as a function of individual $q(z_k)$:

$$\mathrm{ELBO}_j = \int q(z_j)\mathbb{E}_{i\neq j} \log p(z_j|z_{-j}, x, \theta) dz_j - \int q(z_j) \log q(z_j) dz_j$$

Optimality condition $\frac{d\mathrm{ELBO}}{dq(z_j)} = 0$

Exercise (Lagrange multipliers):

$$q^\star(z_j) \propto \exp \mathbb{E}_{-j}[\log p(z_j|Z_{-j}, x)]$$

Coordinate Ascent: Iteratively update each $q(z_\cdot)$.

Conditionals

Last slide: $q^\star(z_j) \propto \exp \mathbb{E}_{-j}[\log p(z_j|Z_{-j}, x)]$

Assume conditional is in the exponential family i.e.

$$p(z_j|z_{-j}, x) = h(z_j) \exp\left(\eta(z_{-j}, x)^\intercal t(z_j) - a(\eta(z_{-j}, x))\right)$$

Note: We will see examples in the following Lectures!

Mean-field for exponential family

- Compute log of the conditional

$$\log p(z_j|z_{-j}, x) = \log(h(z_j)) + \eta(z_{-j}, x)^\intercal t(z_j) - a(\eta(z_{-j}, x))$$

- Compute expectation with respect to $q(z_{-j})$

$$\mathbb{E}[\log p(z_j|z_{-j}, x)] = \log(h(z_j)) + \mathbb{E}[\eta(z_{-j}, x)^\intercal]t(z_j) - \mathbb{E}[a(\eta(z_{-j}, x))]$$

- Thus $q^\star(z_j) \propto h(z_j) \exp(\mathbb{E}[\eta(z_{-j}, x)^\intercal]t(z_j))$

Conditionals

Note: In the case of an exponential family, the optimal $q(z_j)$ is in the same family as the conditional.

Coordinate Ascent

- Assuming variational parameter ν.

$$q(z_1, \cdots, z_k|\nu) = \prod_{i=1}^{k} q(z_i|\nu_i) \tag{4}$$

- Then each natural variational parameter is set equal to the expectation of the natural conditional parameter given all the other variables and the observations:

$$\nu_j^\star = \mathbb{E}[\eta(z_{-j}, x)] \tag{5}$$

Exponential Families

Exponential Families

The exponential family of distributions over x, given parameters η, is defined to be the set of distribution of the form

$$p(x|\eta) = h(x)\exp(\eta^\intercal t(x) - a(\eta)) \tag{6}$$

where:

- η are the so called natural parameters
- $t(x)$ sufficient statistic
- $a(\eta)$ log normalizer
- $\mathbb{E}[t(x)] = \frac{d}{d\eta}a(\eta)$
- Higher moments are next derivatives.

Examples: Bernoulli (last week), Gaussian, Binomial, Multinomial, Poisson, Dirichlet, Beta, Gamma etc.

Conjugacy

Bayesian modeling allows to incorporate priors:

$$\eta \sim F(\cdot|\lambda),$$
$$x_i \sim G(\cdot|\eta), \quad \text{for } i \in \{1, \cdots, n\}$$

The posterior distribution of η given the data $x_{1:n}$ is then given by

$$p(\eta|x, \lambda) \propto F(\eta|\lambda) \prod_{i=1}^{n} G(x_i|\eta)$$

We say F and G are conjugate if the above posterior belongs to the same functional family as F.

Conjugate prior for the Exponential Family

Exponential family members have a conjugate prior:

$$p(x_i|\eta) = h_l(x)\exp(\eta^\intercal t(x) - a_l(\eta))$$
$$p(\eta|\lambda) = h_c(x)\exp(\lambda_1^\intercal \eta + \lambda_2^\intercal(-a_l(\eta)) - a_c(\lambda))$$

- The natural parameter $\lambda = (\lambda_1, \lambda_2)$ has dimension $\dim(\eta) + 1$
- The sufficient statistics are $(\eta, -a(\eta))$

Exercise: Show the above result.

Conditional conjugacy

Let β be a vector of *global latent variables* (with corresponding global parameters α and z be a vector of local latent variables:

$$p(\beta, z, x) = p(\beta) \prod_{i=1}^{n} p(z_i, x_i | \beta)$$

Note: For stochastic variational inference, we make an additional assumption:

$$p(z_i, x_i | \beta) = h(z_i, x_i) \exp(\beta^\intercal t(z_i, x_i) - a(\beta))$$

and take an exponential prior on the global variables as the corresponding conjugate prior:
$p_\alpha(\beta) = h(\beta) \exp(\alpha^\intercal [\beta, -a(\beta)] - a(\alpha))$ with $\alpha = (\alpha_1, \alpha_2)$.

Mean-field for conjugates

Mean-field: $q(z, \beta) = q_\lambda(\beta) \prod_{i=1}^{k} q_{\varphi_i}(z_i)$

- λ global variational parameter
- φ local variational parameter

Local update $\varphi_i \leftarrow \mathbb{E}_\lambda[\eta_l(\beta, x_i)]$

Global update: $\lambda \leftarrow \mathbb{E}_\varphi[\eta_g(x, z)]$

Note: Coordinate ascent iterates between local and global updates.

Algorithm

Algorithm 1 Mean-field with conjugate family assumption

Input:

model p, variational family $q_\varphi(z)$, , $q_\lambda(\beta)$

while ELBO is not converged **do**

  for each data point i **do**

    Update $\varphi_i \leftarrow \mathbb{E}_\lambda[\eta_l(\beta, x_i)]$

  end for

  Update $\lambda \leftarrow \mathbb{E}_\varphi[\eta_g(x, z)]$

end while

Stochastic Variational Inference

Gradient Optimization

$\lambda_{t+1} = \lambda_t + \delta \nabla_\lambda f(\lambda_t)$

or equally

$$\arg\max_{d\lambda} f(\lambda + d\lambda) \;\; \text{st.} \;\; ||d\lambda||^2 \leq \varepsilon \tag{7}$$

Problem: Here it is the euclidean distance, which is not suitable for probability distributions.

Natural Gradient for ELBO:

$$\arg\max_{d\lambda} \mathrm{ELBO}(\lambda + d\lambda) \;\; \text{st.} \;\; D_{\mathrm{KL}}^{\mathrm{sym}}(q_\lambda, q_{\lambda+d\lambda}) \leq \varepsilon \tag{8}$$

where $D_{\mathrm{KL}}^{\mathrm{sym}}(q,p) = \mathrm{KL}(q||p) + \mathrm{KL}(p||q)$

Gradient Optimization

Riemannian metric $G(\lambda)$ sucht that:

$$d\lambda^\intercal G(\lambda) d\lambda = D_{\mathrm{KL}}^{\mathrm{sym}}(q_\lambda(\beta), q_{\lambda+d\lambda}(\beta))$$

From information geometry, we know how to calculate the natural gradient:

$$\widehat{\nabla}_\lambda \mathrm{ELBO} = G^{-1}(\lambda) \nabla_\lambda \mathrm{ELBO}$$

where

$$G(\lambda) = \mathbb{E}[(\nabla_\lambda \log q_\lambda(\beta))(\nabla_\lambda \log q_\lambda(\beta))^\intercal] \tag{9}$$

Gradient Optimization for conjugate models

For our model class, we have:

$$\nabla_\lambda \log q_\lambda(\beta) = t(\beta) - \mathbb{E}[t(\beta)] \tag{10}$$

Thus

$$G(\lambda) = \nabla^2_\lambda a(\lambda) \tag{11}$$

Recall

$$\nabla_\lambda \mathrm{ELBO} = \nabla^2_\lambda a(\lambda)(\mathbb{E}[\eta(x,z)] - \lambda) \tag{12}$$

then

$$\begin{aligned}
\widehat{\nabla}_\lambda \mathrm{ELBO} &= G^{-1}(\lambda)\nabla_\lambda \mathrm{ELBO} \\
&= G^{-1}(\lambda)\nabla^2_\lambda a(\lambda)(\mathbb{E}[\eta(x,z)] - \lambda) \\
&= (\mathbb{E}[\eta(x,z)] - \lambda)
\end{aligned}$$

Gradient Step

In each iteration

$$\lambda_t = \lambda_{t-1} + \delta_t \nabla_{\lambda_{t-1}} \mathrm{ELBO}$$

where δ_t is the step size.

Then substituting the above

$$\lambda_t = (1 - \delta_t)\lambda_{t-1} + \delta_t \mathbb{E}[\eta(x, z)]$$

Algorithm

Algorithm 2 Mean-field with natural gradient

Input:
  model p, variational family $q_\varphi(z)$, $, q_\lambda(\beta)$

while ELBO is not converged **do**
  for each data point i **do**
    Update $\varphi_i \leftarrow \mathbb{E}_\lambda[\eta_l(\beta, x_i)]$
  end for
  Update $\lambda \leftarrow \lambda + \delta \widehat{\nabla}_\lambda \mathrm{ELBO} = (1-\delta)\lambda + \delta \mathbb{E}_{q(\varphi)}[\eta_g(x, z)]$
end while

Stochastic Optimization

Idea: Maximize a function f using noisy gradients H of that function.

- Noisy gradient H: $\mathbb{E}[H] = \nabla f$
- Step size δ_t
- $x_{t+1} \leftarrow x_t + \delta_t H(x_t)$

Convergence to a local optimum is guaranteed, when:

$$\sum_{t=1}^{\infty} \delta_t = \infty$$

$$\sum_{t=1}^{\infty} \delta_t^2 < \infty$$

Stochastic Gradient Step

Recall

$$\begin{aligned}\widehat{\nabla}_\lambda \mathrm{ELBO} &= (\mathbb{E}[\eta(x, z)] - \lambda) \\ &= \left(\alpha_1 + \sum_{i=1}^{n} \mathbb{E}_q[t(z_i, x_i)], n + \alpha_2\right) - \lambda\end{aligned}$$

Idea: Construct a noisy natural gradient by sampling

- Sample index $j \sim \mathrm{Uniform}(1, \cdots, n)$
- Rescale

$$\begin{aligned}\widehat{\nabla}_\lambda \mathrm{ELBO} &= (\mathbb{E}[\eta(x_j^{(n)}, z_j^{(n)})] - \lambda) \\ &= (\alpha_1 + n\mathbb{E}_q[t(z_j, x_j)], 1 + \alpha_2) - \lambda \\ &=: \widehat{\lambda} - \lambda\end{aligned}$$

Summary gradient step $\lambda_t = (1 - \delta_t)\lambda_{t-1} + \delta_t \widehat{\lambda}$

Algorithm

Algorithm 3 Mean-field with stochastic gradient

Input:
model p, variational family $q_{\varphi}(z)$, , $q_{\lambda}(\beta)$

while Stopping criteria is not fulfilled **do**
Sample index $j \sim \mathrm{Uniform}(1, \cdots, n)$
Update $\varphi_i \leftarrow \mathbb{E}_{\lambda}[\eta_l(\beta, x_i)]$
Compute global parameter estimate $\widehat{\lambda} = \mathbb{E}_{\varphi}[\eta_g(x_j, z_j)]$
Optimize the global variational parameter $\lambda_{t+1} \leftarrow \lambda_t(1 - \delta_t) + \delta_t \widehat{\lambda}$
Check step size and update if required!
end while

Black Box Variational Inference

Gradient Estimates of the ELBO

$$\mathrm{ELBO} = \mathbb{E}_{q_\nu}[\log p_\theta(z, x)] - \mathbb{E}_q[\log q_\nu(z)]$$

where ν are the parameters of the variational distribution and θ the parameters of the model (as before).

Aim: Maximize the ELBO

Problem: Need unbiased estimates of $\nabla_{\nu,\theta}\mathrm{ELBO}$.

Reparametrization trick *

Simplified notation:

$$\nabla_\nu \mathbb{E}_{q_\nu}[f_\nu(z)]$$

Assume that there exists a fixed reparameterization such that

$$\mathbb{E}_{q_\nu}[f_\nu(z)] = \mathbb{E}_q[f_\nu(g_\nu(\varepsilon))]$$

where the expectation on the right does now not depend on ν. Then

$$\nabla_\nu \mathbb{E}_q[f_\nu(g_\nu(\varepsilon))] = \mathbb{E}_q[\nabla_\nu f_\nu(g_\nu(\varepsilon))]$$

Solution: Obtain unbiased estimates by taking a Monte Carlo estimate of the expectation on the right.

*Will be covered later

Optimizing the ELBO

$\mathbb{E}_{q(\lambda)}[\log p(z,x) - \log q(z)] =: \mathbb{E}[g(z)]$

Exercise:

$\nabla_\lambda \mathrm{ELBO} = \nabla_\lambda \mathbb{E}[g(z)] = \mathbb{E}[g(z) \nabla \log q(z)] + \mathbb{E}[\nabla g(z)]$

where $\nabla \log q(z)$ is called the *score function*.

Note: The expectation of the score function is zero for any q i.e.

$$\mathbb{E}_q[\nabla \log q(z)] = 0$$

Thus, to compute a noisy gradient of the ELBO

- sample from $q(z)$
- evaluate $\nabla \log q(z)$
- evaluate $\log p(x,z)$ and $\log q(z)$

Algorithm

Algorithm 4 Black Box Variational Inference

1: **Input:** data x, model p(x,z), variational family $q_\varphi(z)$,
2: **while** Stopping criteria is not fulfilled **do**
3: Draw L samples $z_l \sim q_\varphi(z)$
4: Update variational paramater using the collected samples

$$\varphi \leftarrow \varphi + \delta_t \frac{1}{L} \sum_{l=1}^{L} \nabla \log q(z_l)(\log p(x, z_l) - \log q(z_l))$$

5: Check step size and update if required!
6: **end while**

Note: Active research area (problem) is the reduction of the variance of the noisy gradient estimator.

Summary

What we have seen

- Expectation Maximization
- Introduced $\mathbb{KL}$ to understand convergence of EM.
- α- Divergence as a more general concept, including Expectation Propagation
- Scalable Stochastic Variational Inference for conjugates in the exponential family
- black box variational inference to overcome exponential family assumption

What is not covered

- Problem for black box is the variance of the noisy gradient
- Extensions black box α-Divergence
- Online inference algorithms

Next week

Plan for next week

- Originally planned: Bayesian Non-parametrics
- Most likely: State Space Models

Questions?