Probabilistic Graphical Models for Image Analysis - Lecture 3

Stefan Bauer

5 October 2018

Max Planck ETH Center for Learning Systems

Overview

1. Variational Inference - Recall
2. α-Divergence

Variational Inference - Recall

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

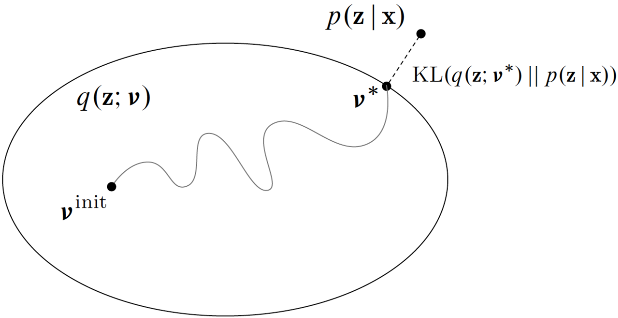
$$p(z \mid x, \theta) = \frac{p(z, x \mid \theta)}{\int p(z, x \mid \theta)} \tag{1}$$

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 - Overview



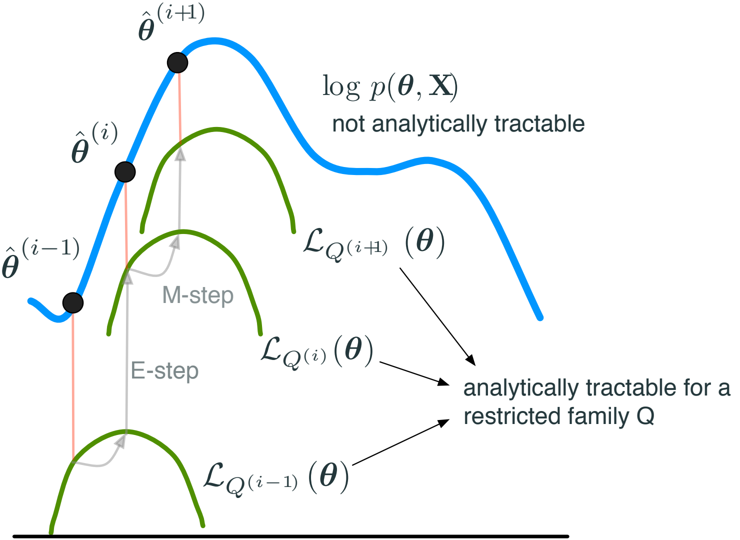
*

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

Illustration Mean Field Approximations



Concept

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

Note: We can not calculate $\mathrm{KL}[q(z) \mid\mid p(z \mid x)]$, since we do not know $p(z \mid x)$ but we can see that maximizing ELBO is equivalent to minimizing the KL divergence between the posteriors.

Summary

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

where:

$$\mathrm{ELBO}(q, \theta) = \int q(z) \log \left(\frac{p(x, z, \theta) q(z)}{p(z|x, \theta) q(z)} \right) dz$$
$$\mathrm{KL}(q(z) || p(z|x, \theta)) = \int q(z) \log \left(\frac{p(z|x, \theta)}{q(z)} \right) dz$$

Kullback-Leibler Divergence

Properties

- $\mathrm{KL}(q||p) \geq 0, \forall q, p$
- $\mathrm{KL}(q||p)0$ if and only if $q = p$

Note: $\mathrm{KL}(q||p) \neq \mathrm{KL}(p||q)$ i.e. the KL is not a distance but a so called *divergence*.

If we use $\mathrm{KL}(q||p)$ we have the following characteristics of the divergence:

- If $q == p$: distributions are equal.
- If q is high and p is high -> captures what we want!
- If q is high but p is low -> problematic
- If q is low, then the Expectation is zero.

Question: Why do we choose $\mathrm{KL}(q||p)$ over $\mathrm{KL}(p||q)$?

Exponential Families

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

$$p(\mathbf{x}|\eta) = h(\mathbf{x})g(\eta)\exp(\eta^\intercal \mathbf{u}(\mathbf{x})) \tag{2}$$

where:

- η are the so called *natural parameters*
- $g(\eta)$ can be interpreted as a normalization i.e. ensuring $g(\eta)\int h(\mathbf{x})\exp(\eta^\intercal \mathbf{u}(\mathbf{x}))dx = 1$

Example: Many! e.g. Bernoulli $p(x|\mu) = \mu^x(1-\mu)^{1-x}$

Maximum Likelihood for Exponential Family

Taking the gradient of the normalization condition on both sides:

$$\nabla g(\eta) \int h(\mathbf{x}) \exp(\eta^\intercal \mathbf{u}(\mathbf{x})) dx + g(\eta) \int h(\mathbf{x}) \exp(\eta^\intercal \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) dx = 0$$

Rearranging and using normalization condition:

$$-\frac{1}{g(\eta)} \nabla g(\eta) = g(\eta) \int h(\mathbf{x}) \exp(\eta^\intercal \mathbf{u}(\mathbf{x})) \mathbf{u}(\mathbf{x}) dx = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Summarizing:

$$-\nabla \log g(\eta) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Exponential Families Continued

In case of set of iid. data $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ we have the likelihood as:

$$p(\mathbf{X}|\eta) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\eta)^N \exp\left(\eta^\intercal \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right) \tag{3}$$

Thus from previous slide:

$$-\nabla \log g(\eta_{ML}) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) \tag{4}$$

Note: Solution for ML estimator depends only on data through $\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$; so called *sufficient statistic* since it is enough to store the sufficient statistic instead of the whole data!

Example: For Bernoulli $\mathbf{u}(x) = x$ and we store only the sum of all data points but not all data itself.

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

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

Expectation Propagation - Moment Matching

Previous slide:

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

Note: Optimal solutions implies matching of expected sufficient statistics!

Example: If $q(z)$ is Gaussian $\mathcal{N}(z|\mu, \Sigma)$, then we minimize the KL by setting μ equal to mean of $p(z)$ and Σ equal to covariance of p (so called *moment matching*).

Expectation Propagation - Factorized posteriors

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

Problem: Optimizing $\mathrm{KL}(q||p)$ requires computing expectations wrt. q, while $\mathrm{KL}(p||q)$ requires expectations wrt. p, which is typically intractable.

Assume the case where the true distribution p factorizes in a product of factors $p(\mathcal{D}, \theta) = \prod_{i=1}^{N} f_i(\theta)$!

Assume q is from exponential family and factorized $q(\theta) = \frac{1}{Z_{EP}} \prod_{n=1}^{N} \tilde{f}_i(\theta)$

Then our aim is to minimize

$$\mathrm{KL}\left(\frac{1}{p(\mathcal{D})} \prod_{n=1}^{N} f_n(\theta) || \frac{1}{Z_{EP}} \prod_{n=1}^{N} \tilde{f}_n(\theta)\right) \tag{5}$$

Expectation Propagation - Factorized posteriors

Our aim is to minimize

$$\mathrm{KL}\left(\frac{1}{p(\mathcal{D})}\prod_{n=1}^{N} f_n(\theta) || \frac{1}{Z_{EP}}\prod_{n=1}^{N}\tilde{f}_n(\theta)\right) \tag{6}$$

Problem: In general intractable!

Idea: Update single factors iteratively and if the factors belong to the exponential family this can simply be done by moment matching!

Expectation Propagation Algorithm

Algorithm 1 Minimizing $\mathrm{KL}(p||q)$ for factorized distributions

Input:
Initialisations of approximations $\tilde{f}_i(\cdot)$

Until Convergence:
for each factor i **do**
Delete factor i from approximation

$$q^{\setminus i} = \frac{q(\theta)}{\tilde{f}_i(\theta)} = \prod_{n \neq i} \tilde{f}_n(\theta)$$

Projection $\tilde{f}_i^{\mathrm{new}} = \arg\min_{f_i'} \mathrm{KL}\left(f_i(\theta) q^{\setminus i}(\theta) || f_i'(\theta) q^{\setminus i}(\theta)\right)$
Update $q = \tilde{f}_i^{\mathrm{new}}(\theta) q^{\setminus i}(\theta)$
end for
Return: After convergence one has $p(D) \approx \int \prod_n \tilde{f}_n(\theta) d\theta$

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 (compare with last lecture).
- 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 (less well suited for GMM).

Note: Both *KL* can be embedded into a wider framework of α-divergences.

α-Divergence

α-Divergence

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} dx \tag{7}$$

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 $q = p$

α-Divergence Special cases

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} dx \tag{8}$$

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

Special cases

- $\lim_{\alpha \to 0} D_\alpha(p||q) = \mathrm{KL}(q||p)$
- $\lim_{\alpha \to 1} D_\alpha(p||q) = \mathrm{KL}(p||q)$
- $D_{-1}(p||q) = \frac{1}{2} \int \frac{(q(x)-p(x))^2}{p(x)} dx$
- $D_2(p||q) = \frac{1}{2} \int \frac{(q(x)-p(x))^2}{q(x)} dx$
- $D_{\frac{1}{2}}(p||q) = 2 \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 dx$

Illustration - Intuition for different α

$$D_\alpha(p||q) = \frac{\int \alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha}}{\alpha(1-\alpha)} dx \qquad (9)$$

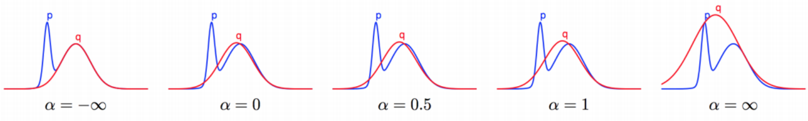
To understand how the choice of α might affect the result of approximate inference, consider the problem of approximating a complicated distribution p with a tractable Gaussian distribution q by minimizing $D_\alpha[p||q]$.

- α is large positive number: q covers all modes of p
- α is large negative number: q covers mode with highest probability mass
- Optimal α hard to choose, probably depending on learning task.

Example: If the true distribution p has many modes, a global approximation might be bad by placing substantial probability mass in the area where the true posterior does not.

Visual interpretation of α

Gaussian *q* approximating two mode Gaussian *p*.



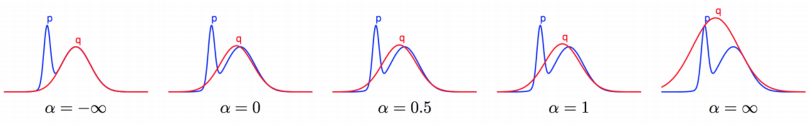
If the goal is to compute marginal distributions, using a fully-factorized approximation, then the best choice (among α-divergences) is inclusive KL $(\alpha = 1)$, because it is the only α which strives to preserve the marginals *

---

*Picture and Quote from: Thomas Minka, Divergence measures and message passing, Technical report 2005.

Explanation of "Mode"-seeking

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

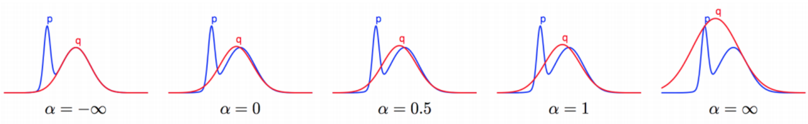


For $D_{-1}(p||q) = \frac{1}{2}\int \frac{(q(x)-p(x))^2}{p(x)}dx$ i.e. $\alpha = -1$, a small $p(x)$ forces the optimal q to be small, too (**zero-forcing**) i.e. false-positives are avoided under the cost of excluding some parts of p. The cost of excluding an x i.e. setting $q(x) := 0$ is equal to $\frac{p(x)}{1-\alpha}$; Thus q will seek area of largest total mass (**mode-seeking**).

Problem: Underestimating the variance for $\alpha << 0$

Inclusive v.s Exclusive

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



For $\alpha \geq 1$ it requires that $q > 0$ whenever $p > 0$ i.e. avoiding "false-negatives". The divergence is **inclusive** since it prefers to stretch across p.

Next week

Plan for next week

- So far: Scaling with variables but not data.
- Stochastic and Black Box Variational Inference
- Summary Variational Inference

Questions?