COMP9418 – Advanced Topics in Statistical Machine Learning

W4 – Variational Inference (Part I)

Instructor: Edwin V. Bonilla

School of Computer Science and Engineering

Last update: 22/8/17 11:10:04 pm

Time for Feedback

- **Lecture**
  - Anonymous forum
- **Tutorials**
  - Theory
    - » Would be nice to have more time
    - » Know before hand priority and make explicit minimum
  - Practical part too hard
    - » Shorter practical
    - » More step by step
    - » Very good in retrospective
- **Quiz**

Acknowledgements

Material derived from:

- [Bishop, PRML, 2006] Pattern Recognition and Machine Learning, Christopher Bishop, 2006
  - Almost all figures in today's lecture are from this book
- [Murphy, MLaPP, 2012] Machine Learning: A Probabilistic Perspective, Kevin P. Murphy, 2012

Aims

This lecture will allow you to understand variational inference methods for posterior estimation in graphical models. Following it you should be able to:

- Understand the expectation-maximisation (EM) algorithm from a more general perspective, identifying the objective function it is maximising and the relations between its constituents
- Understand and apply properties of factorised distributions for posterior approximation
- Exploit conjugacy properties for obtaining closed-form updates in variational algorithms
- Identify advantages and disadvantages of variational methods for approximate Bayesian inference

Outline

I. EM Revisited

II. Variational Inference

III. Factorised Distributions

IV. Bayesian GMMs

V. Remarks and Conclusions

I. EM Revisited

The EM Algorithm Revisited (1)

$\mathbf{x}^{(i)}$: Observed variable for data point *i* → $\mathbf{X} = \{\mathbf{x}^{(i)}\}$

$\mathbf{z}^{(i)}$: Hidden or missing variable → $\mathbf{Z} = \{\mathbf{z}^{(i)}\}$

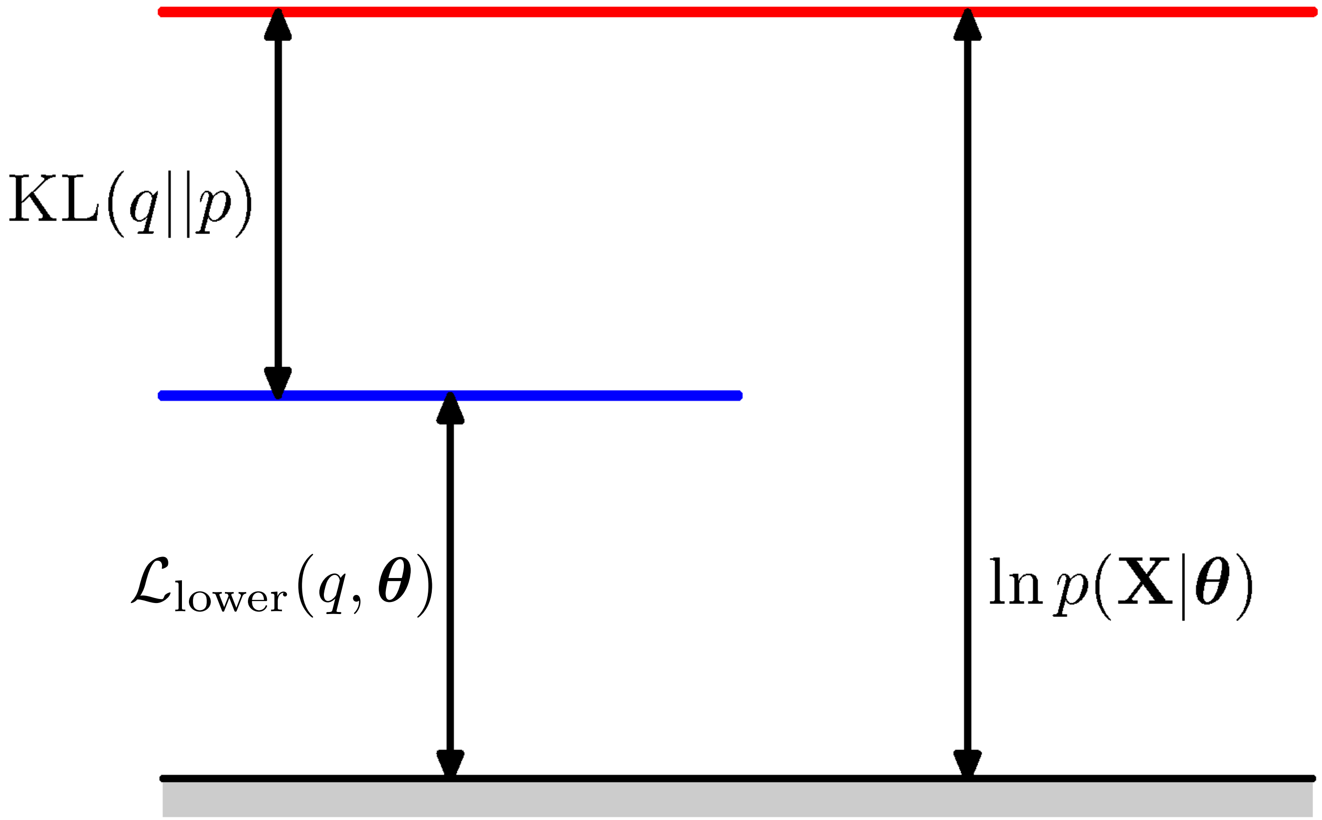
- Direct maximisation of the data log likelihood $\log p(\mathbf{X}|\boldsymbol{\theta})$ is hard
  - Optimization of the expected complete data log likelihood is significantly easier
- We can exploit the decomposition:

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathrm{KL}(q(\mathbf{Z}|\mathbf{X}) \| p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})) + \mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$$

- Where: $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z}|\mathbf{X})}\right]$
  - Is a lower bound on the (marginal) data log likelihood

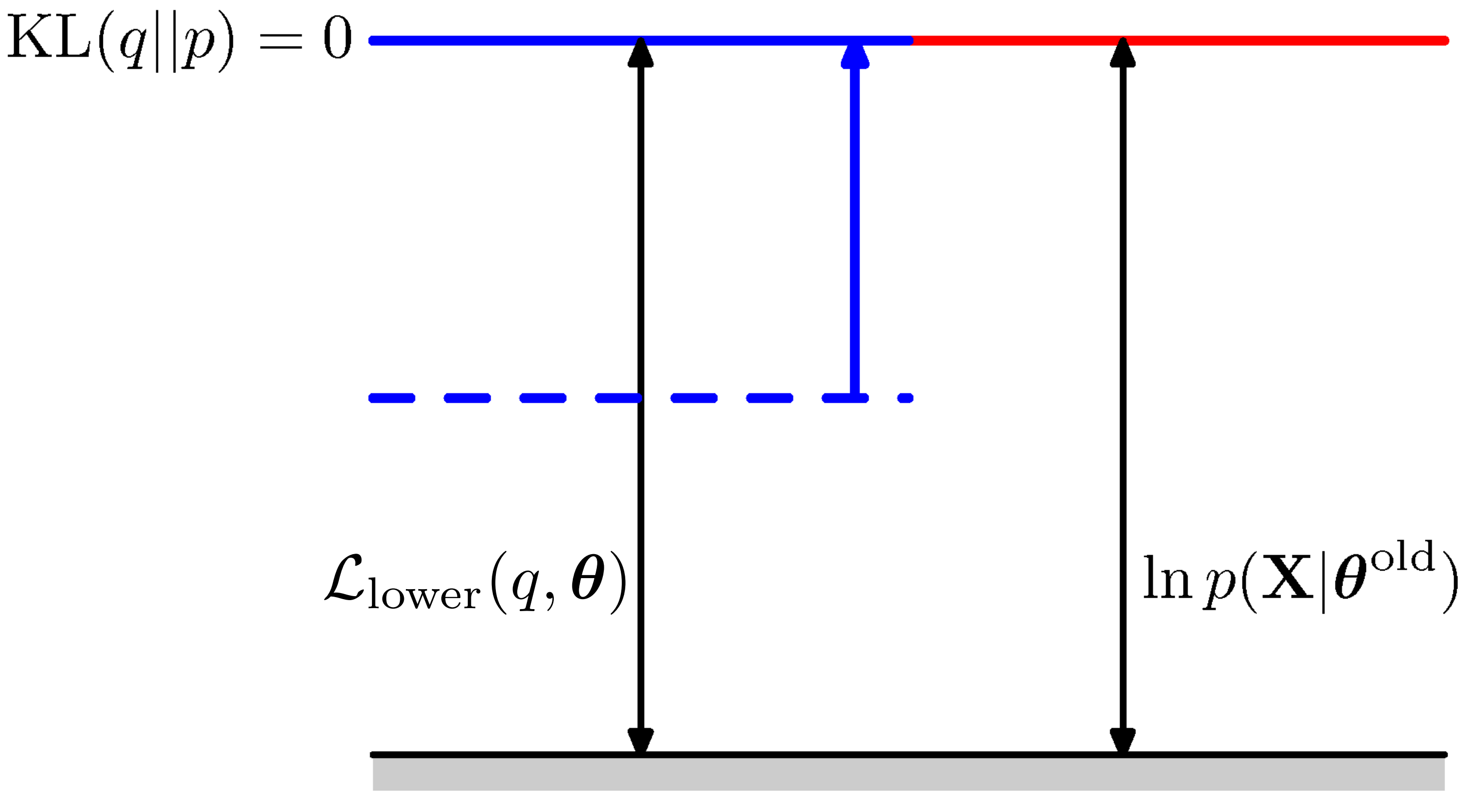
The EM Algorithm Revisited (2)

Decomposition of the marginal likelihood



The EM Algorithm Revisited (3)

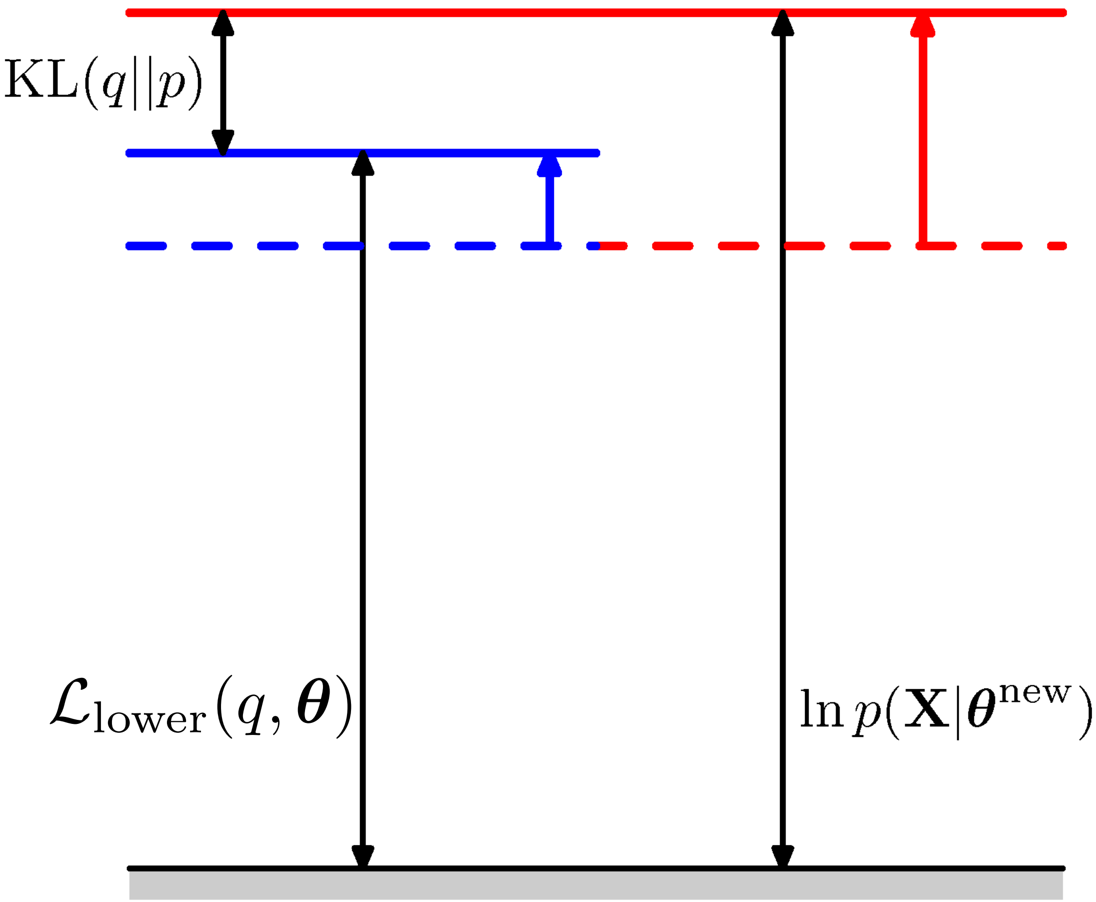
- E-step: Maximisation of $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$ wrt q(**Z**|**X**)
  - Optimal setting is the true posterior q(**Z**|**X**) = p(**Z**|**X**,θ)



The EM Algorithm Revisited (4)

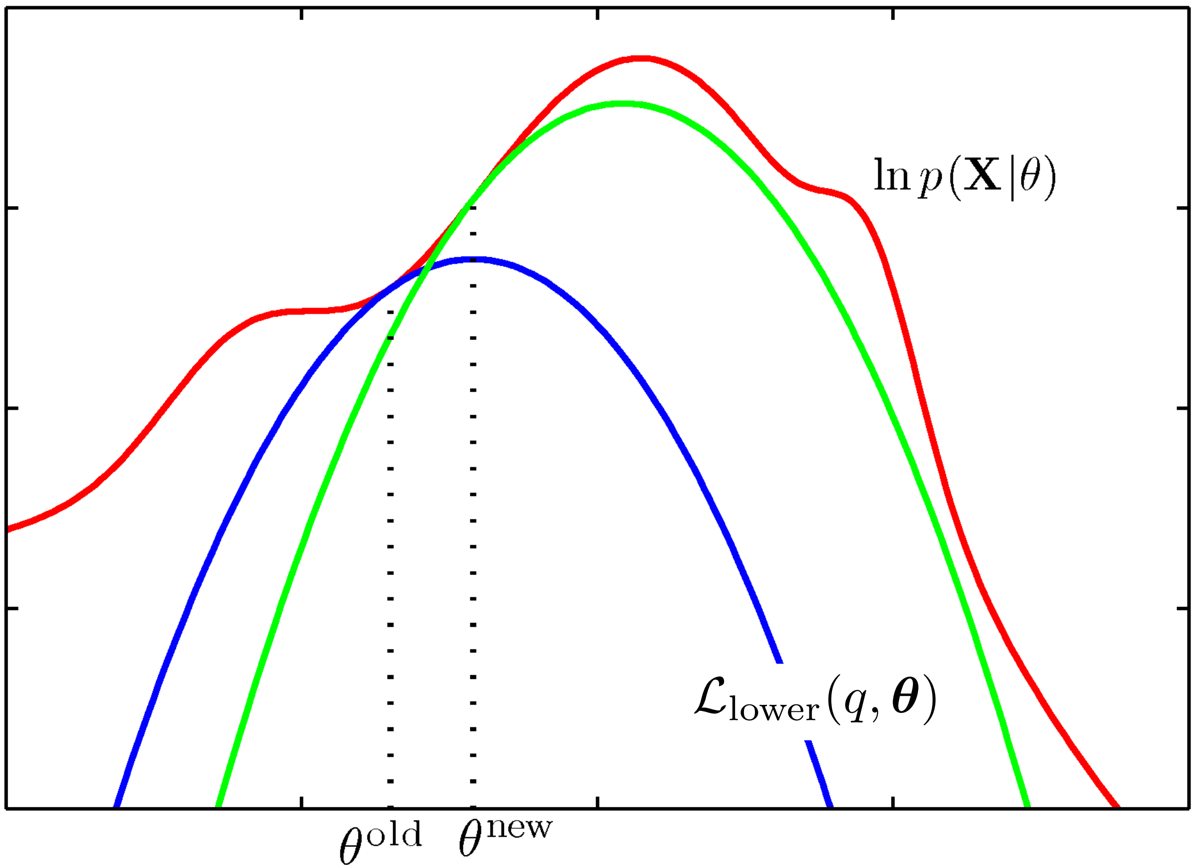
- M-step: Maximisation of $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$ wrt θ
  - Since q(**Z**|**X**) is fixed, moving θ also increases the KL
  - Objective optimised is in fact Q(θ,θ^{old}) as $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta}) = Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \boxed{\mathbb{H}(q)}$

entropy



The EM Algorithm Revisited (5)

- EM in parameter space
  - [E] Updating q → tight bound (blue curve touches red curve)
  - [M] Updating θ → leaves a gap (max of blue curve at θ^{new})
  - [E] green curve



Generalisations of the EM Algorithm

- Vanilla EM indirectly maximises the likelihood
  - E-step and M-step are sometimes simple (e.g. GMMs)
  - However, they may be intractable for complex models
- Generalized EM
  - Deals with intractable M-step by partially optimizing $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$ wrt θ
    - » As before, each cycle will improve the likelihood (not get worse)
    - » Can use standard gradient-based optimization techniques
    - » Expectation conditional maximisation
      - Alternating optimisation of parameter subsets
- Intractable E-step
  - Similarly, we can partially optimize $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$ wrt q(**Z**|**X**)
    - » Any local maximum of $\mathcal{L}_{\text{lower}}(q, \boldsymbol{\theta})$ will also be a local maximum $\log p(\mathbf{X}|\boldsymbol{\theta})$

Such generalizations naturally lead us to variational inference

The Need for Approximate Inference

- Computation of posterior (or expectations over it) is crucial:

$$p(\mathbf{Z}|\mathbf{X}) = \frac{\overbrace{p(\mathbf{X}|\mathbf{Z})}^{\text{Likelihood}} \ \overbrace{p(\mathbf{Z})}^{\text{prior}}}{\underbrace{p(\mathbf{X})}_{\text{marginal}}}$$

  - Exact inference through conjugate priors for simple models
  - JTA in discrete case → exponential in tree width
  - EM requires expectations over posterior for parameter estimation
  - For many models exact computation is unfeasible
    - » Intractability of marginal likelihood (sums or integrals in high dimensions)

Stochastic approaches

- Sampling
- Exact results given infinite computation
- Usually demanding, non-scalable

Deterministic approaches

- Assumptions on (approximate posterior)
- Almost never exact
- Usually scalable
- *Variational inference*

II. Variational Inference

Variational Inference and Calculus of Variations

Standard calculus

- Function
  - Input: variable
  - Output: Scalar
- Derivatives of functions
  - Changes in the output wrt infinitesimal changes in the input

Calculus of variations

- Functionals
  - Input: Function
  - Output: scalar

$$\mathbb{H}[q(\mathbf{x})] = -\int p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x}$$

- Functional derivatives
  - Changes in the output wrt infinitesimal changes in the input
    - » Similar rules to std calculus

Main idea: Approximate inference as an optimisation problem

- Objective is a functional (input is the posterior distribution)
  - » Constrain the posterior to a suitable family of functions
  - » Optimise wrt (approximate) posterior

The Variational Objective

$\mathbf{X}$: Observed variables

$\mathbf{Z}$: Hidden or missing variables

- Goal: given prior P(**Z**) and conditional likelihood p(**X**|**Z**) → approximate the posterior p(**Z**|**X**) with q(**Z**|**X**)
  - Omitting θ as we can include them in **Z** as random variables
- We have seen that

Approximate posterior

True posterior

$$\log p(\mathbf{X}) = \mathrm{KL}(q(\mathbf{Z}|\mathbf{X}) \| p(\mathbf{Z}|\mathbf{X})) + \mathcal{L}_{\text{lower}}(q)$$

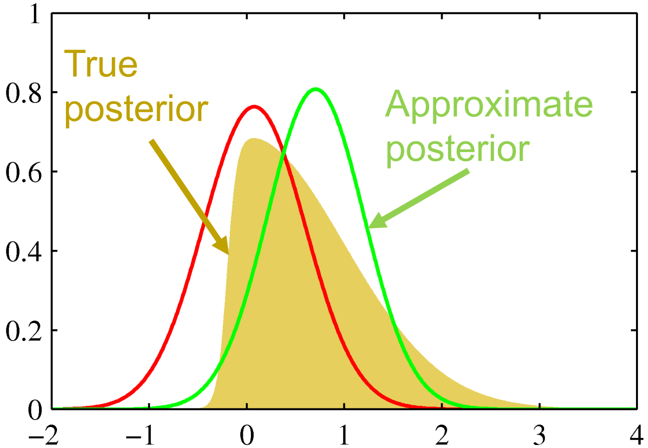
- Where: $$\mathcal{L}_{\text{lower}}(q) \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z}|\mathbf{X})}\right]$$

This is our variational objective (functional)

  - We will attempt to maximise $\mathcal{L}_{\text{lower}}(q)$ wrt q(**Z**|**X**)
  - This is indeed equivalent to minimizing KL(q(**Z**|**X**) || p(**Z**|**X**))

What exactly is q(Z|X)?

- Free-form $q(\mathbf{Z}|\mathbf{X})$: optimisation of the functional $\mathcal{L}_{\text{lower}}(q)$
  - Would give us the right answer as the KL vanishes at the true posterior
  - However, need to solve normalization, which was our initial problem!
- Fixed-form $q(\mathbf{Z}|\mathbf{X})$: Consider a restricted family of distributions
  - Minimize the objective wrt members of this family
  - E.g. Use factorised distribution
  - E.g. Use a parametrized distribution $q(\mathbf{Z}|\mathbf{X},\lambda)$
    - » Optimisation via standard calculus



What family of distributions?

- As flexible as possible
- Tractability is the main constraint
- No risk of overfitting
  - The more flexible the better the approximation to the true posterior

Understanding the Variational Objective

The lower bound $\mathcal{L}_{\text{lower}}(q) \stackrel{\text{def}}{=} \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}\left[\log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z}|\mathbf{X})}\right]$ can be written as:

$$\mathcal{L}_{\text{lower}}(q) = \underbrace{\mathbb{E}_{q(\mathbf{Z}|\mathbf{X})}[\log p(\mathbf{X}|\mathbf{Z})]}_{\text{Expected log likelihood (ELL)}} - \underbrace{\mathrm{KL}(q(\mathbf{Z}|\mathbf{X})\|p(\mathbf{Z}))}_{\text{KL (approx. posterior || prior)}}$$

- ELL term is a model fit: How well the (samples from the) posterior explains the observations
- KL is a penalty term: Keep posterior close to prior beliefs
- Also known as
  - Variational free energy
  - Evidence lower bound (ELBO)

III. Factorised Distributions

Factorised Distributions

Mean field approximation

- For notational simplicity, make $q(\mathbf{Z}) \stackrel{\text{def}}{=} q(\mathbf{Z}|\mathbf{X})$ and assume that it factorises over M disjoint groups:

$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- In this case, we can optimise the variational objective in "free-form"
  - No additional assumptions on the functional forms of $q_i(\mathbf{Z}_i)$

Optimal solution

$$\log q_j^\star(\mathbf{Z}_j) = \boxed{\mathbb{E}_{i \neq j}[\log p(\mathbf{X}, \mathbf{Z})]} + \boxed{\text{const.}}$$

Expectation over all q($\mathbf{Z}_i$) except $\mathbf{Z}_j$ of the log joint

Additive constant set by normalisation

  - General consistency conditions
    - » Need to iterate
    - » Guaranteed convergence as bound is convex wrt to each factor

Properties of Factorised Distributions (1)

- Consider a 2-dimensional Gaussian $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$

Precision matrix

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \qquad \Lambda_{21} = \Lambda_{12}$$

- Goal: Approximate p($\mathbf{z}$) with a q($\mathbf{z}$) = $q_1(z_1)\ q_2(z_2)$
  - No assumptions on the functional form of the approximating distributions
- Using the general mean-field update we obtain:

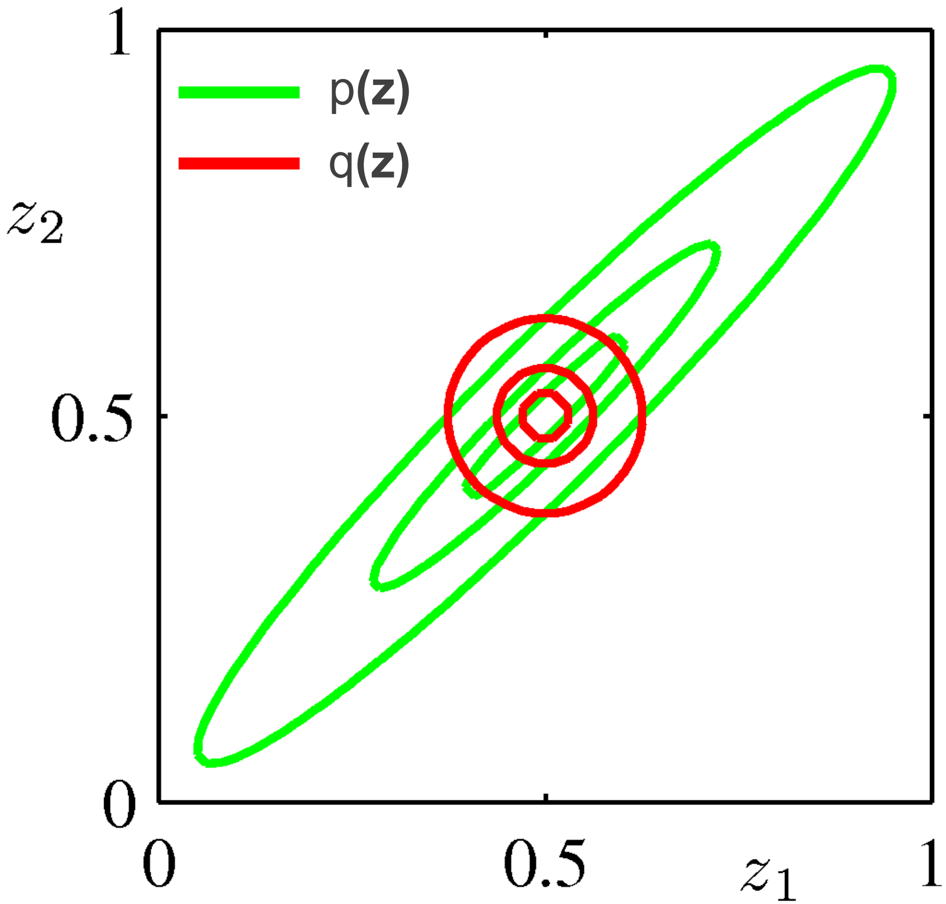
$$\log q_1^\star(z_1) = \mathbb{E}_{z_2}[\log p(\mathbf{z})] + \text{const.}$$

  - and similarly for q_2
- In general, we need to iterate but here we have a closed-form solution

$$q_i^\star(z_i) = \mathcal{N}(z_i|\mu_i, \Lambda_{ii}^{-1})$$

- Correct mean
- Variance?

Properties of Factorised Distributions (2)



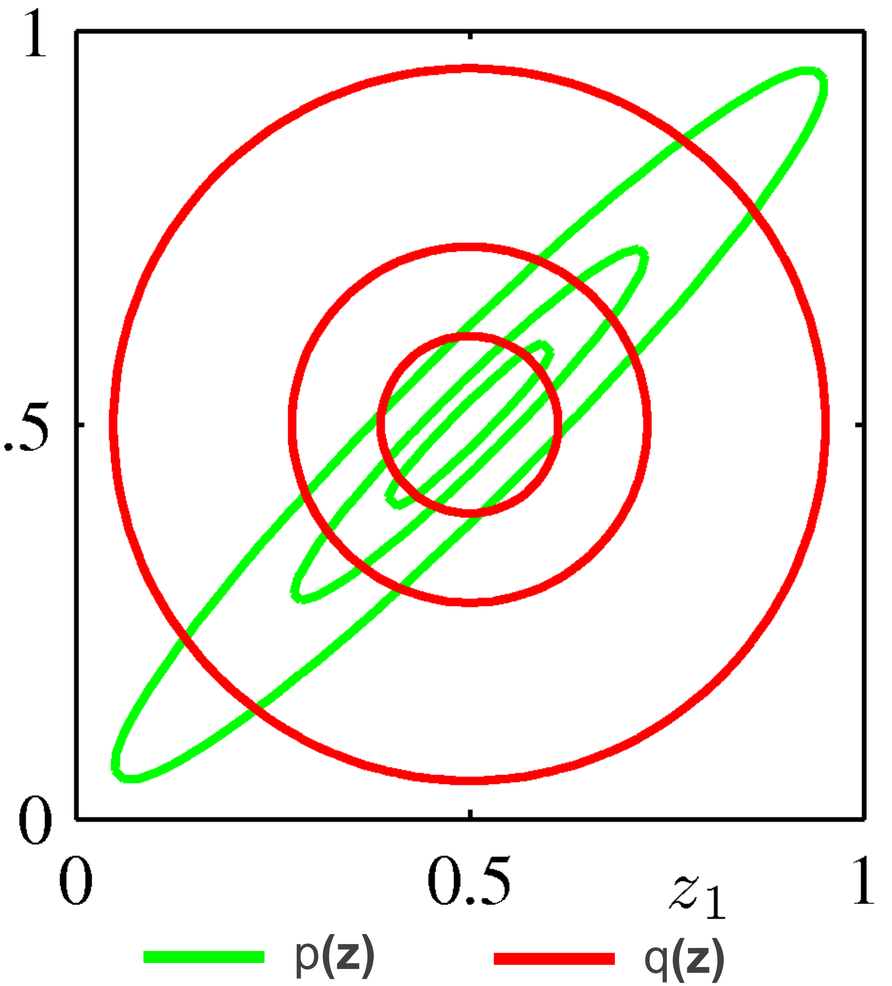
Contours at 1, 2, 3 standard deviations

- Variance controlled by the direction of smallest variance
- Variance on orthogonal direction significantly underestimated
- Factorized variational → too compact posteriors

$$\mathrm{KL}(q(\mathbf{z})\|p(\mathbf{z})) = \mathbb{E}_{q(\mathbf{z})}\left[\log \frac{q(\mathbf{z})}{p(\mathbf{z})}\right]$$

  - Solution will avoid regions where p(**z**) is small

Properties of Factorised Distributions (3)

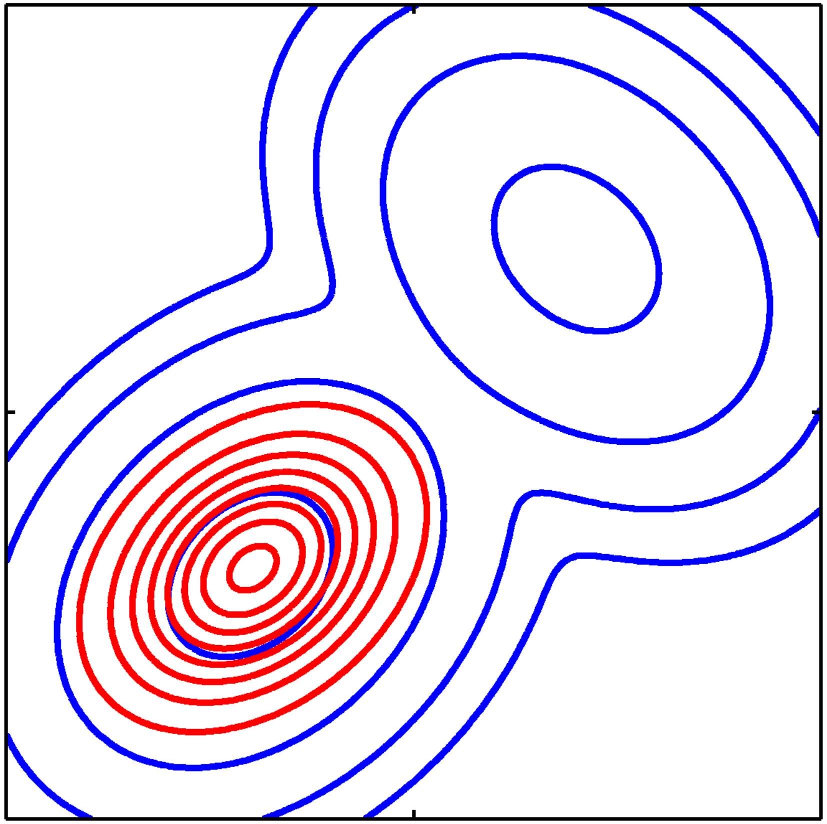


Contours at 1, 2, 3 standard deviations

- Minimizing the reverse KL(p(**z**)||q(**z**))
  - Expectation propagation
- Solution corresponds to the marginal distributions
- Correct means
- Significant mass in regions of low p(**z**)

Properties of Factorised Distributions (4)

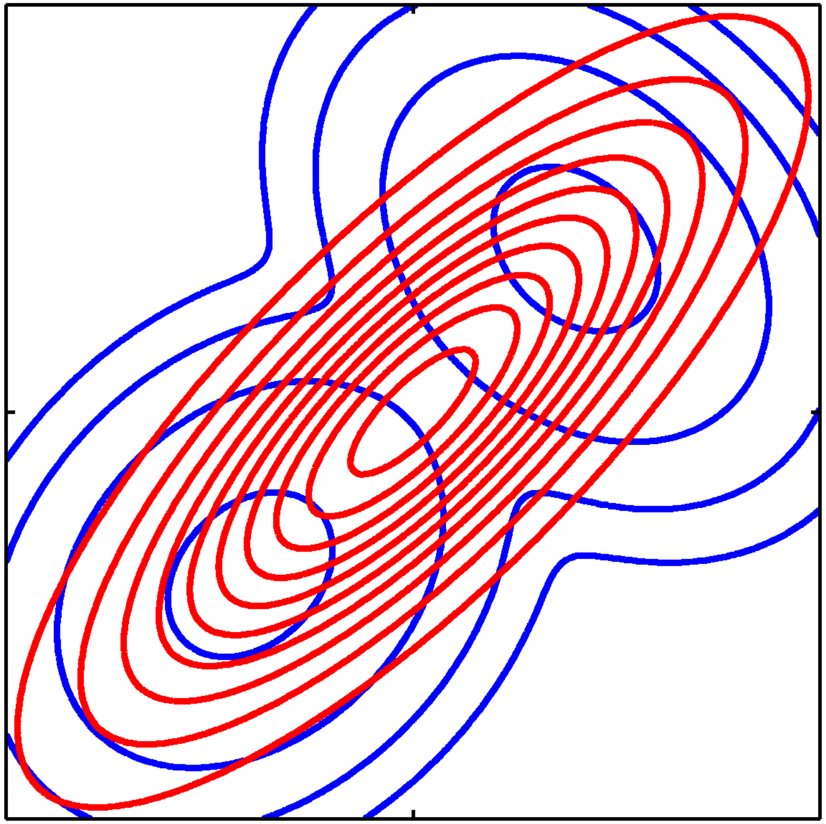
Optimisation of KL(q||p)



True multimodal posterior p(**z**)

Unimodal q(**z**)

Optimisation of KL(p||q)



True multimodal posterior p(**z**)

Unimodal q(**z**)