NPFL108 – Bayesian inference

Approximate Inference

Variational Inference

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek

Version: 21/03/2014

Outline

- Variational inference
- Unknown Mean and Variance of a normal dist.

Variational Inference: Introduction

- Based on the calculus of variations, i.e., a generalization of standard calculus.
- Deals with functionals, functions and derivatives of functionals rather than functions, variables and derivatives.
- Similar rules apply.
- Can be applied to models of either continuous or discrete random variables.
- Approximates both
  - the posterior distribution: $p(w|D)$
  - its normalization constant (model evidence): $p(D)$
    - D: evidence – data
    - w: unknown parameters

Variational Inference

- It is based on the following decomposition:

$$\log p(D) = L(q) + KL(q||p)$$

- where

$$L(q) = \int q(w) \log \left\{ \frac{p(w, D)}{q(w)} \right\} dw$$

lowerbound

$$KL(q||p) = \int q(w) \log \left\{ \frac{q(w)}{p(w|D)} \right\} dw$$

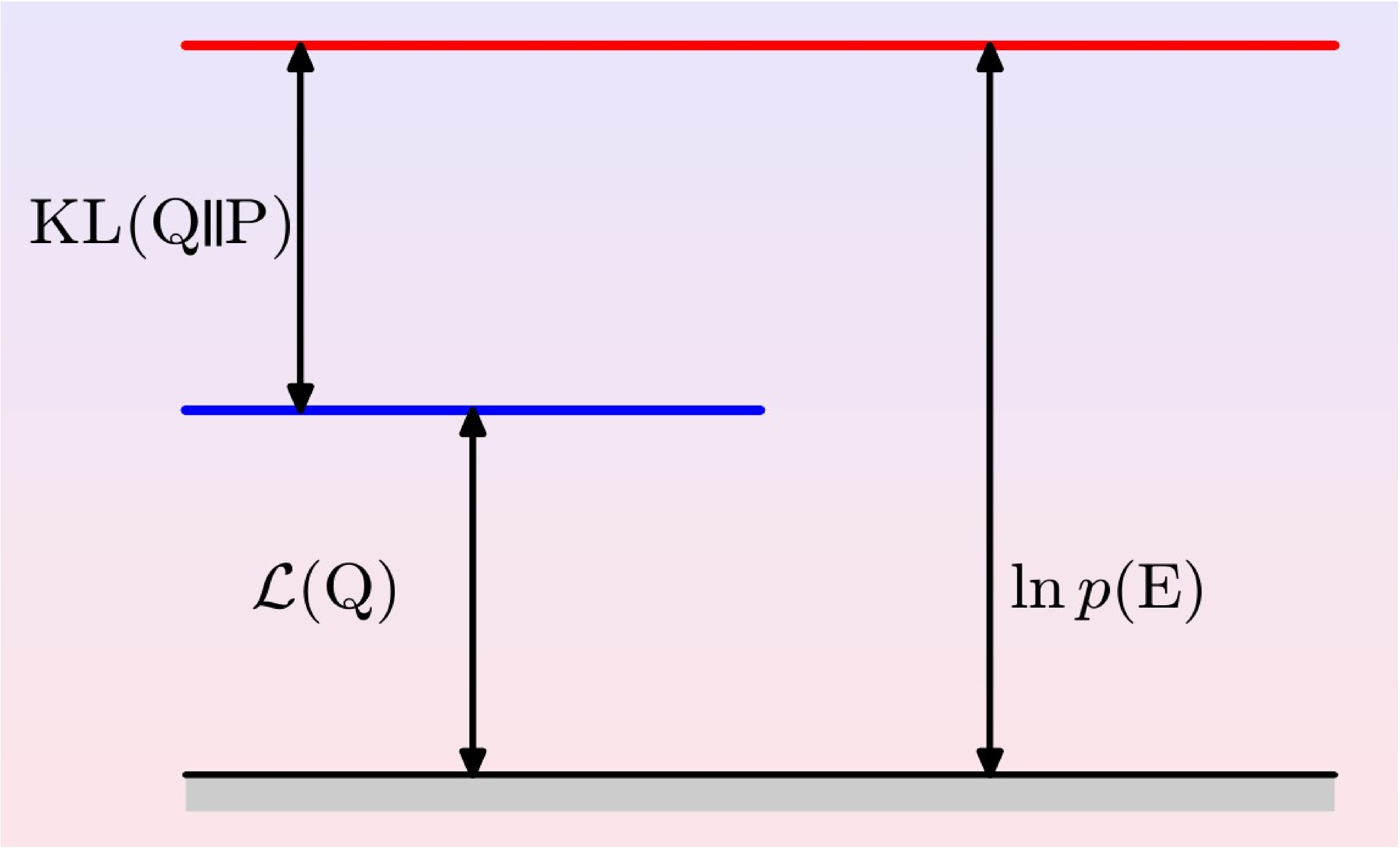
KL-divergence

- L(q) approximates log p(D)
  - We want to maximise
- The Kullback-Leibler divergence measures the **fit** of q(w) to p(w|D)
  - We want to minimise

Verification: Workout

$$
\begin{aligned}
\log p(D) &= L(q) + KL(q||p) \\
&= \int q(w) \log\left\{\frac{p(w,D)}{q(w)}\right\} dw + \int q(w) \log\left\{\frac{q(w)}{p(w|D)}\right\} dw \\
&= \int q(w) \left\{\log\left\{\frac{p(w,D)}{q(w)}\right\} + \log\left\{\frac{q(w)}{p(w|D)}\right\}\right\} dw \\
&= \int q(w) \log\left\{\frac{p(w,D)}{q(w)} \frac{q(w)}{p(w|D)}\right\} dw \\
&= \int q(w) \log\left\{\frac{p(w,D)}{p(w|D)}\right\} dw \\
&= \int q(w) \log\left\{\frac{p(w|D)p(D)}{p(w|D)}\right\} dw \\
&= \int q(w) \log p(D) dw \\
&= \log p(D) \int q(w) dw \\
&= \log p(D) \cdot 1 \\
&= \log p(D)
\end{aligned}
$$

Decomposition of the Marginal Likelihood



Source Bishop, 2006

Choosing the approximation q

- One can use a gradient ascend on L(q)
  - Hill climbing
- One selects **q** to be a parametric distribution
  - **q(z|θ)** for which L(q) can be computed analytically
- The lower bound then becomes a function of θ and can be optimized

Alternative approximation of q

- An alternative is to assume that **q** factorizes with respect to a partition of **w** into M disjoint groups w_i,
  - with i = 1, . . . , M:

$$q(w) = \prod_{i}^{M} q_i(w_i)$$

- no further assumptions are made about **q**

- This approach is known in the literature as variational mean field or global variation inference

Variational Inference

- Substituting **q** in KL(q||p) and looking for the dependence with respect to $\mathbf{q}_j$
  - Similar to coordinate ascend

$$q(w) = \prod_{i}^{M} q_i(w_i) = q_1(w_1)q_2(w_2)\ldots q_M(w_M)$$

  - Optimising
    - KL(q(w) || p(w|D))

Derivation 1

$$
\begin{aligned}
KL(q|p) &= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log \left\{ \frac{\prod_{k=1}^{M} q_k(\mathbf{w}_k)}{p(\mathbf{w}|D)} \right\} d\mathbf{w} \\
&= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \left\{ \sum_{k=1}^{M} \log q_k(\mathbf{w}_k) - \log p(\mathbf{w}|D) \right\} d\mathbf{w} \\
&= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \left\{ \sum_{k=1}^{M} \log q_k(\mathbf{w}_k) - \log p(\mathbf{w}, D) + \log p(D) \right\} d\mathbf{w} \\
&= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \left\{ \sum_{k=1}^{M} \log q_k(\mathbf{w}_k) - \log p(\mathbf{w}, D) \right\} d\mathbf{w} + C_1 \\
&= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \left\{ \sum_{k=1}^{M} \log q_k(\mathbf{w}_k) \right\} d\mathbf{w} - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \left\{ \log p(\mathbf{w}, D) \right\} d\mathbf{w} + C_1 \\
&= \sum_{k=1}^{M} \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log q_k(\mathbf{w}_k) d\mathbf{w} - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log p(\mathbf{w}, D) d\mathbf{w} + C_1 \\
&= \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log q_j(\mathbf{w}_j) d\mathbf{w} - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log p(\mathbf{w}, D) d\mathbf{w} + C_2 \\
&= \int q_j(\mathbf{w}_j) \log q_j(\mathbf{w}_j) \prod_{i=1; i \neq j}^{M} q_i(\mathbf{w}_i) d\mathbf{w} - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log p(\mathbf{w}, D) d\mathbf{w} + C_2 \\
&= \int q_j(\mathbf{w}_j) \log q_j(\mathbf{w}_j) d\mathbf{w}_j - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i) \log p(\mathbf{w}, D) d\mathbf{w} + C_2
\end{aligned}
$$

Derivation 2

$$KL(q|p) = \int q_j(\mathbf{w}_j)\log q_j(\mathbf{w}_j)d\mathbf{w}_j - \int \prod_{i=1}^{M} q_i(\mathbf{w}_i)\log p(\mathbf{w},D)d\mathbf{w} + C_2$$

$$= \int q_j(\mathbf{w}_j)\log q_j(\mathbf{w}_j)d\mathbf{w}_j - \int q_j(\mathbf{w}_j)\int \prod_{i=1;i\neq j}^{M} q_i(\mathbf{w}_i)\log p(\mathbf{w},D)d\mathbf{w} + C_2$$

$$= \int q_j(\mathbf{w}_j)\log q_j(\mathbf{w}_j)d\mathbf{w}_j - \int q_j(\mathbf{w}_j)\log\left(\exp\left\{\int \prod_{i=1;i\neq j}^{M} q_i(\mathbf{w}_i)\log p(\mathbf{w},D)d\mathbf{w}_{\setminus j}\right\}\right)d\mathbf{w}_j + C_2$$

$$= \int q_j(\mathbf{w}_j)\log\frac{q_j(\mathbf{w}_j)}{\exp\left\{\int \prod_{i=1;i\neq j}^{M} q_i(\mathbf{w}_i)\log p(\mathbf{w},D)d\mathbf{w}_{\setminus j}\right\}}d\mathbf{w}_j + C_2$$

$$= KL\left(q_j(\mathbf{w}_j)||\exp\left\{\int \prod_{i=1;i\neq j}^{M} q_i(\mathbf{w}_i)\log p(\mathbf{w},D)d\mathbf{w}_{\setminus j}\right\}d\mathbf{w}_j\right) + C_2$$

Variational Inference: Variational Mean-Field

- The KL(q || p) is minimised when both q = p
- The optimal q_j given that the other factors are kept fixed is:

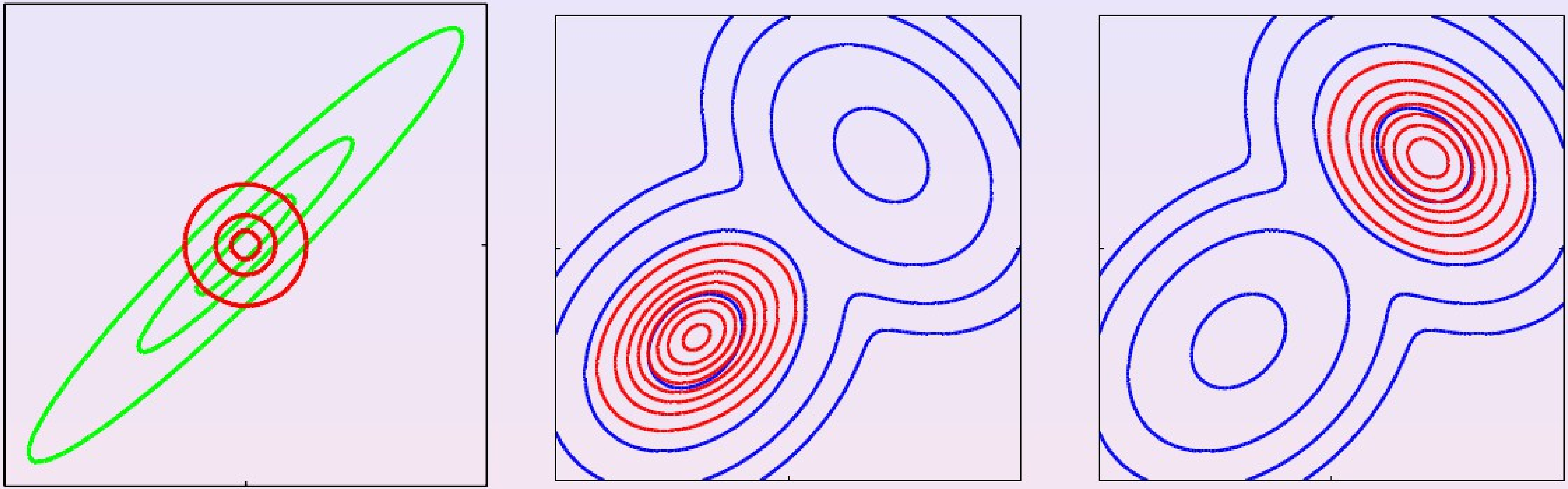
$$q_j(\mathbf{w}_j) \propto \exp\left\{ \int \prod_{i=1; i\neq j}^{M} q_i(\mathbf{w}_i) \log p(\mathbf{w}, D) d\mathbf{w}_{\setminus j} \right\}$$

$$\propto E_{q_{i\neq j}} \left[\log p(\mathbf{w}, \mathbf{D}) \right]$$

- Iteratively
  - Compute this for all $\mathbf{q}_j$ - multiple times

- This is a "coordinate" optimisation over factors $\mathbf{q}_j$ with respect to others.

Properties of Variational Approximations



- The KL divergence KL(q||p) favours solutions that take high probability where **p** takes high probability, but can ignore important regions.
- The optimization problem is not convex and can have multiple local optima.
- Though, convergence is guaranteed

Example: Unknown Mean and Variance of a normal dist. #1

- Goal: infer the posterior distribution of the mean **μ** and precision **τ** of a normal distribution given a dataset D = {x_1, . . . , x_N} of independent samples.
- The log likelihood of μ and τ is:

$$\log p(\mathcal{D}|\mu,\tau) = -\frac{N}{2}\log 2\pi\tau^{-1} - \frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2$$
$$= \frac{N}{2}\log\tau - \frac{\tau}{2}\left[N(\mu-\overline{x})^2 + S\right] + \text{const}\,,$$

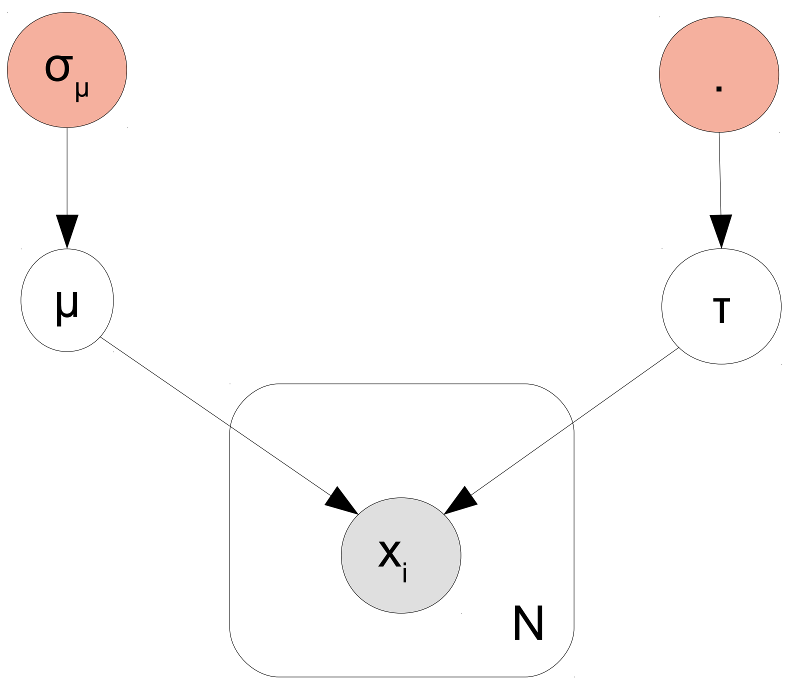
$\mathrm{S} = \sum_n (x_n - \overline{x})^2$ and $\overline{x}$ is empirical mean

Example: cont.

- The priors for μ and τ are uniform and conjugate :

$$p(\mu) = 1/\sigma_\mu \,, \qquad\qquad p(\tau) = 1/\tau$$

- These are improper priors!



Example: cont.

- We enforce that the posterior approximation factorizes $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$ and solve for the optimal factors

$$\log q_\mu(\mu) = E_{q_\tau}\left[\log p(D, \mu, \tau)\right]$$
$$\log q_\tau(\tau) = E_{q_\mu}\left[\log p(D, \mu, \tau)\right]$$

- This gives the following optimal factors given that the other factor is fixed

$$q_\mu(\mu) = \mathcal{N}(\mu|\overline{x}, \lambda^{-1})$$
$$q_\tau(\tau) = \mathrm{Gamma}(\tau|a, b) = b^a \frac{1}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\}$$

Example: cont.

- This gives the following optimal factors given that the other factor is fixed

$$q_\mu(\mu) = \mathcal{N}(\mu|\overline{x}, \lambda^{-1})$$

$$q_\tau(\tau) = \mathrm{Gamma}(\tau|a,b) = b^a \frac{1}{\Gamma(a)} \tau^{a-1} \exp\{-b\tau\}$$

- Where

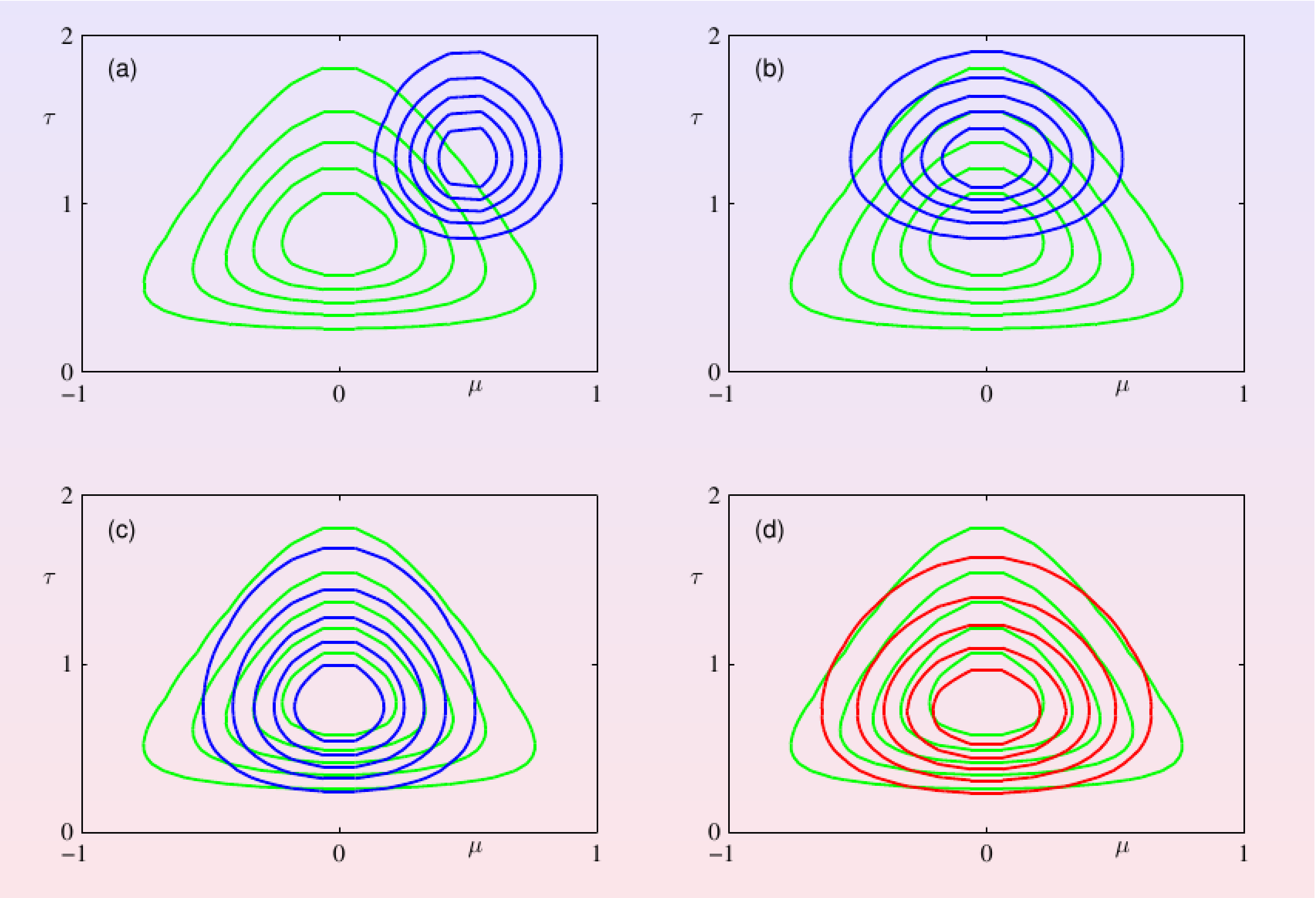
$$\lambda = N E_{q_\tau}[\tau] = Na/b$$

$$a = N/2$$

$$b = N/2(\lambda^{-1} + S)$$

- We iteratively optimize q_μ and q_τ until convergence

Mean Field: Unknown Mean and Variance of a Gaussian



Source Bishop, 2006

Thank you!

Filip Jurčíček

Institute of Formal and Applied Linguistics
Charles University in Prague
Czech Republic

Home page: http://ufal.mff.cuni.cz/~jurcicek