T-61.6020 Machine Learning: Basic Principles: Chapter 10: Approximate inference - Part I

Janne Toivola
jatoivol@cis.hut.fi

The figures are from the PRML book
by Christopher M. Bishop

Topics

- Variational inference?
- Factorized distributions
- Variational mixture of Gaussians
- Variational linear regression

Calculus of Variations

- Derivative of a funtion tells how the value changes when a parameter is varied
- Maximum of a function at $\frac{d}{dx}f(x) = 0$
- Similarly, *functional* derivative tells how a functional $H[f]$ changes when the function is varied
- We can find a function that fits the best

Variational inference

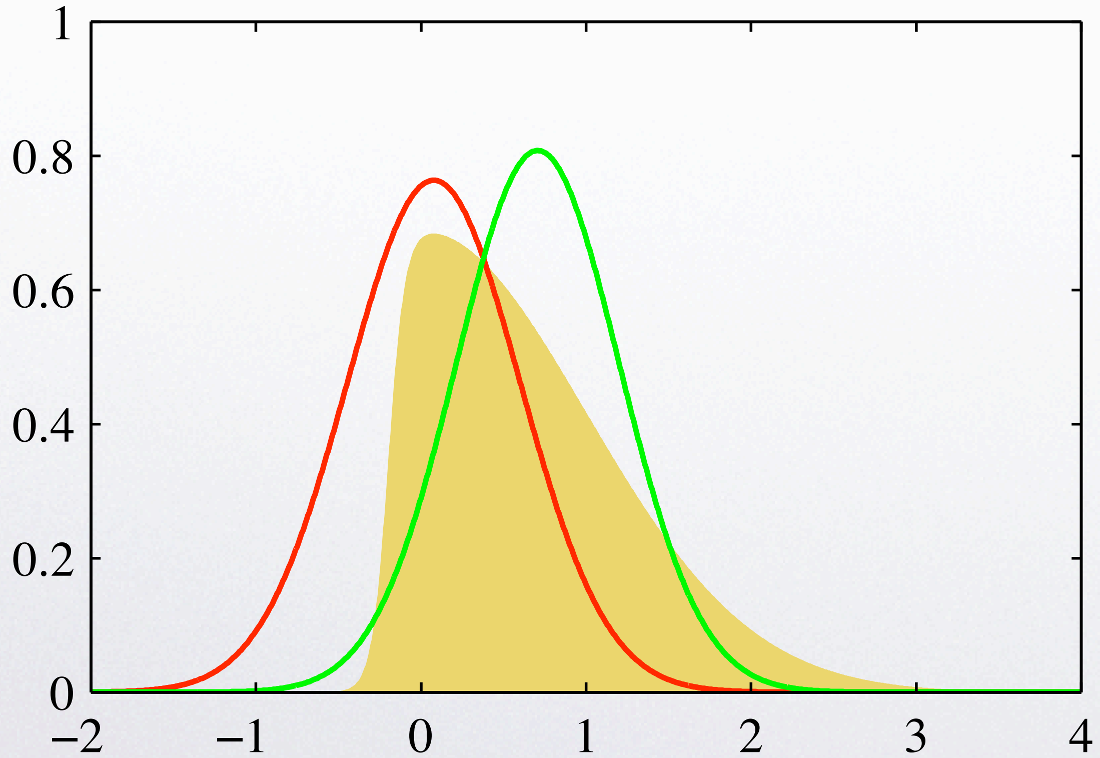
- The true posterior $p(\mathbf{Z}|\mathbf{X})$ considered too complicated to maximize
- We want to fit $q(\mathbf{Z})$ so that it approximates $p(\mathbf{Z}|\mathbf{X})$
- Variational methods allow us to make trade-off between the form of $q(\mathbf{Z})$ and its accuracy

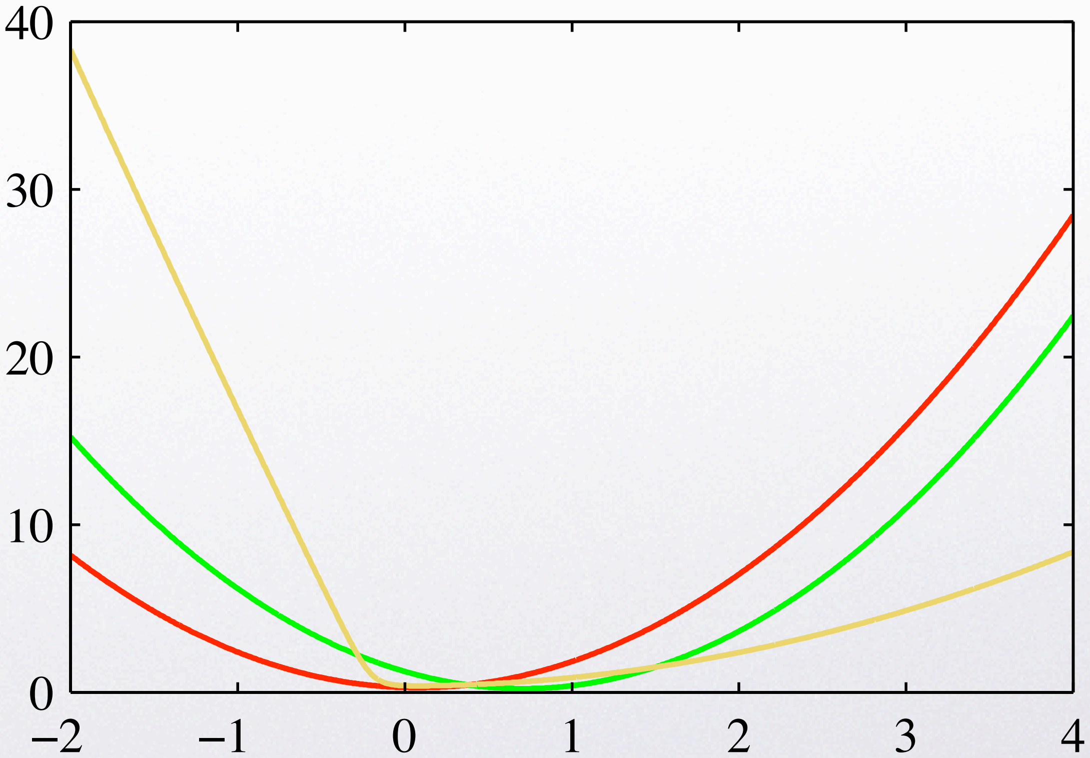
Variational inference

- As with EM, the objective is to maximize the likelihood of evidence $p(\mathbf{X})$
- We end up with $\ln p(\mathbf{X}) = \mathcal{L}(q) + KL(q||p)$
- Improving the lower bound $\mathcal{L}(q)$ is equivalent with minimizing the KL divergence

Example

- A distribution approximated by a Gaussian
- Laplace (red) vs. variational (green)





Factorized distributions

- One way to restrict the family of distributions

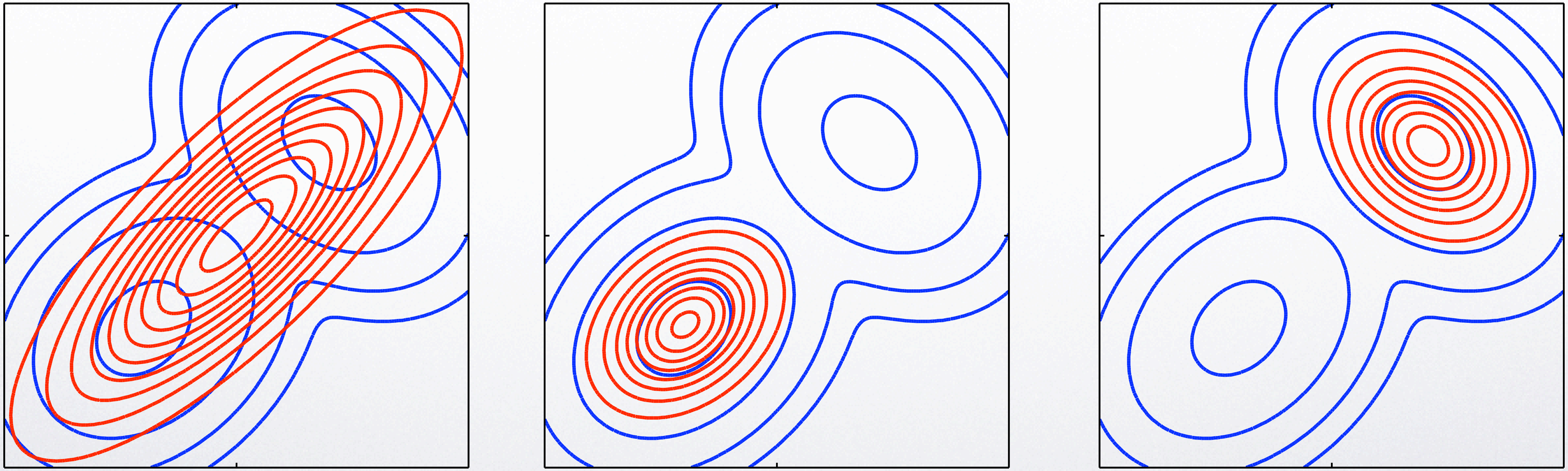
$$q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$$

- The lower bound $\mathcal{L}(q)$ is maximized by iteratively considering each $\mathbf{Z}_j$ in turn

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

Example 2

- Asymmetry of KL divergence when approximating a mixture of two Gaussians



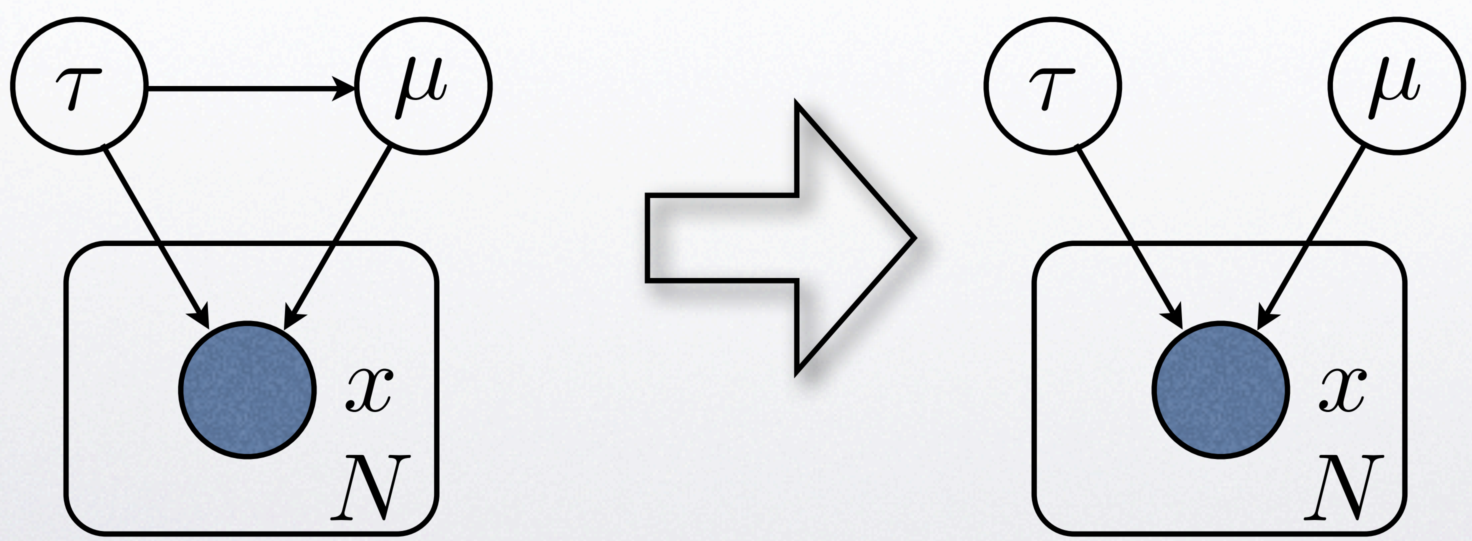
$KL(p||q)$ $KL(q||p)$

Example 3

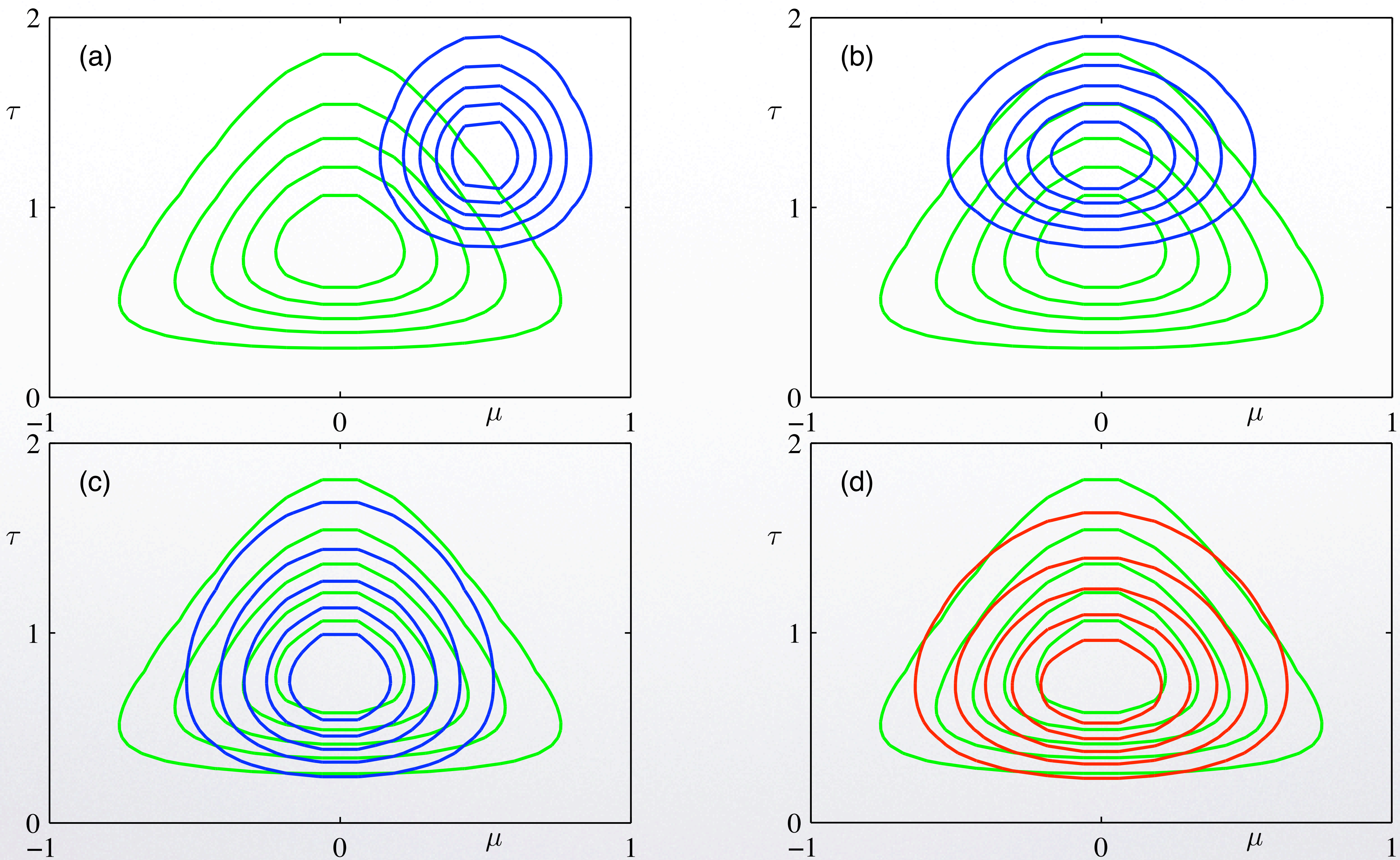
- Univariate Gaussian, true priors are:

$$p(\tau) = Gam(\tau|a_0, b_0) \qquad p(\mu|\tau) = \mathcal{N}(\mu|\mu_0, (\lambda_0\tau)^{-1})$$

- Approximated by: $q(\mu, \tau) = q_\mu(\mu)q_\tau(\tau)$



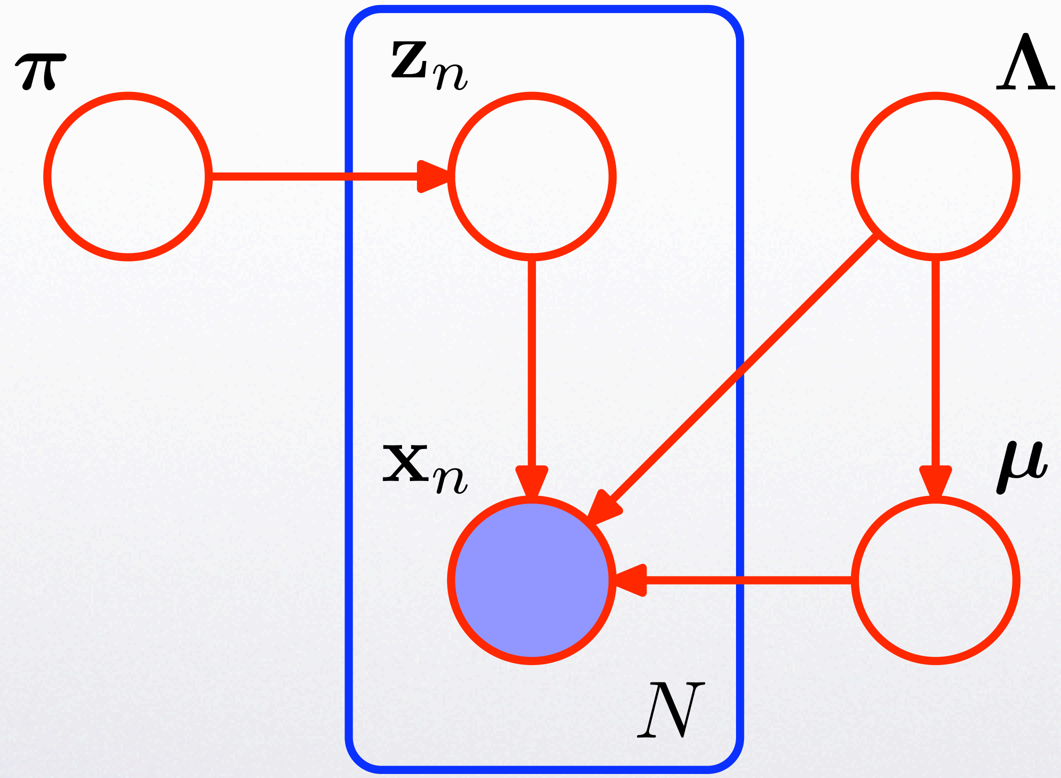
Example 3: posteriors



Example 4

- Variational mixture of Gaussians

$$p(\mathbf{U}) = p(\mathbf{X}|\mathbf{Z}, \mu, \mathbf{\Lambda})p(\mathbf{Z}|\pi)p(\pi)p(\mu|\mathbf{\Lambda})p(\mathbf{\Lambda})$$

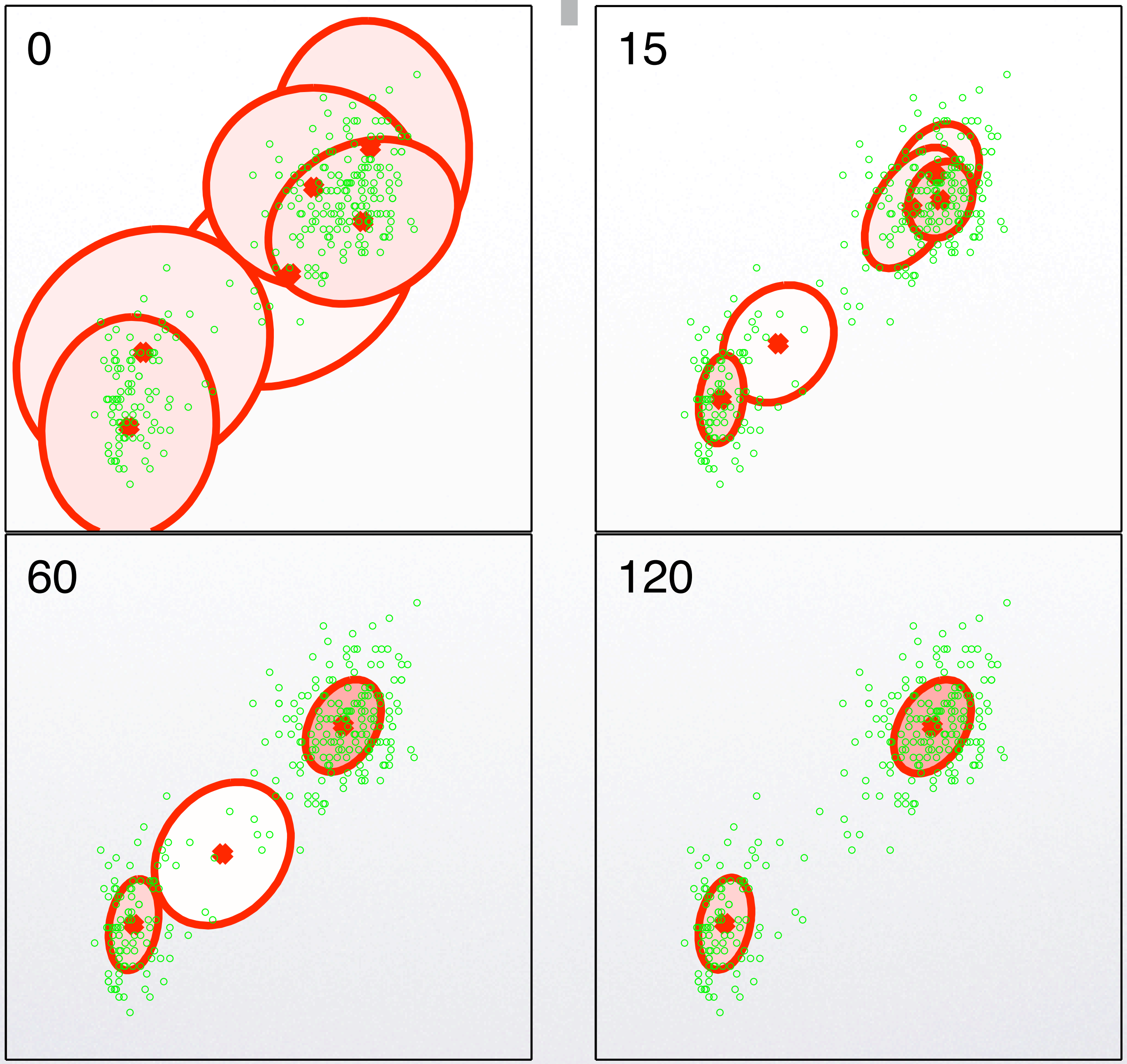


Example 4

- Make a wild assumption about the factorization: $q(\mathbf{Z}, \pi, \mu, \mathbf{\Lambda}) = q(\mathbf{Z})q(\pi, \mu, \mathbf{\Lambda})$
- Additionally it turns out that
$$q(\pi, \mu, \mathbf{\Lambda}) = q(\pi) \prod_{k=1}^{K} q(\mu_k, \mathbf{\Lambda}_k)$$
- Solve update equations for

$q^\star(\pi)$, $q^\star(\mu_k, \mathbf{\Lambda}_k)$ and $q^\star(\mathbf{Z})$

Example 4



Example 5

- Variational linear regression

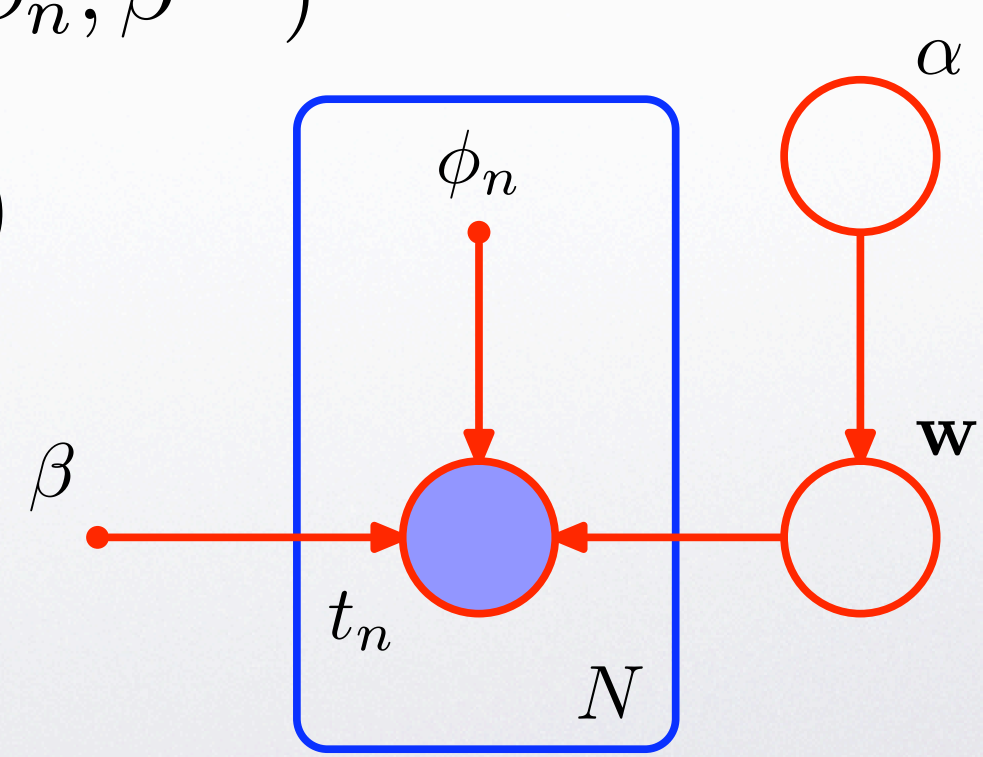
$$p(\mathbf{t}, \mathbf{w}, \alpha) = p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\alpha)p(\alpha)$$

$$p(\mathbf{t}|\mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^T\phi_n, \beta^{-1})$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$p(\alpha) = Gam(\alpha|a_0, b_0)$$

$$q(\mathbf{w}, \alpha) = q(\mathbf{w})q(\alpha)$$



More questions?

- Should I really know how to derive equations for probability distributions?-)
- How can we delegate that to a computer?