Variational Algorithms for Approximate Bayesian Inference: An Introduction

Prof. Nicholas Zabaras
Center for Informatics and Computational Science
https://cics.nd.edu/
University of Notre Dame
Notre Dame, IN, USA

Email: nzabaras@gmail.com
URL: https://www.zabaras.com/

March 21, 2018

Contents

- ❑ Introduction

- ❑ Outline of variational inference, Factorized distributions, Mean Field Approximation, Approximating a Gaussian using a factorized Gaussian.

- ❑ Alternative forms of KL divergence, Gaussian Approximation of *p(x) by KL(p||q) minimization*, Alpha family of divergences, Variational optimization and the EM algorithm

- ❑ Mean Field for the Ising Model, Structured Mean Field

- ❑ Variational Inference for the univariate Gaussian, Variational optimization and model selection.

Following:

- Pattern Recognition and Machine Learning, Christopher M. Bishop, Chapter 10
- Machine Learning: A Probabilistic Perspective, Kevin Murphy, Chapter 21.

Introduction

Many models of interest are based around Bayes' theorem:

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{\boldsymbol{P}(\boldsymbol{X}|\boldsymbol{Z})\boldsymbol{P}(\boldsymbol{Z})}{\boldsymbol{P}(\boldsymbol{X})}$$

where ***Z*** denotes latent variables and ***X*** denotes observed data

We are often interested in computing the posterior $P(\boldsymbol{Z}|\boldsymbol{X})$ or expectations wrt to it (as in the EM algorithm)

For many models this can be difficult. For example:

- Posterior is not analytically tractable
- Computationally expensive:
  - High dimensionality or complexity of the problem

We can use approximation schemes to overcome this issue

Introduction

Two classes of approximation schemes:

Deterministic
e.g. Variational Inference (Variational Bayes)

Analytical approximations to $P(\boldsymbol{Z}|\boldsymbol{X})$

Assume factorization;

Assume solution takes specific form;

Global approximation vs. local approximation (e.g. as in Laplace approximation)

Do not generate exact solutions

Schemes can give fast convergence when by other means not possible

Stochastic
e.g. Markov Chain Monte Carlo

Can generate exact results with infinite resource

Slow convergence, by CLT

High variance (can trade for bias)

Need to generate i.i.d. samples

Must fully explore state space

Outline of Variational Inference

Suppose we have a fully Bayesian model based around N i.i.d observations $\boldsymbol{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ with latent parameters $\boldsymbol{Z} = \{\mathrm{z}_1, \dots, z_N\}$. Our model tells us the joint distribution $p(\boldsymbol{X}, \boldsymbol{Z})$ but **we want to approximate the posterior $\mathbf{p}(\boldsymbol{Z}|\boldsymbol{X})$ and model evidence $p(\boldsymbol{X})$.** We denote by $q(\boldsymbol{Z})$ this approximation.

Variational Inference is based on the following decomposition of the log marginal probability:

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p)$$

where we have defined the lower bound and KL divergence respectively as follows:

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$$

$$KL(q||p) = -\int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \int q(\boldsymbol{Z}) \ln\left\{\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})}\right\} d\boldsymbol{Z}$$

Outline of Variational Inference

Derivation:

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + KL(q||p)$$

where we have defined the lower bound and KL divergence respectively as follows:

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z}$$

$$KL(q||p) = -\int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} = \int q(\boldsymbol{Z}) \ln \left\{ \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} \right\} d\boldsymbol{Z}$$

Proof: Expanding $\mathcal{L}(q)$ and using the mornalization of $q(\boldsymbol{Z})$ leads to this result:

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})p(\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} = \int q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{Z}|\boldsymbol{X}) + \ln p(\boldsymbol{X})\right) d\boldsymbol{Z}$$
$$- \int q(\boldsymbol{Z})\, lnq(\boldsymbol{Z})\, d\boldsymbol{Z}$$
$$= \ln p(\boldsymbol{X}) \int q(\boldsymbol{Z})\, d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z} = \ln p(\boldsymbol{X}) + \int q(\boldsymbol{Z}) \ln \left\{ \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} \right\} d\boldsymbol{Z}$$

$$\mathcal{L}(q) = \ln p(\boldsymbol{X}) - KL(q||p)$$

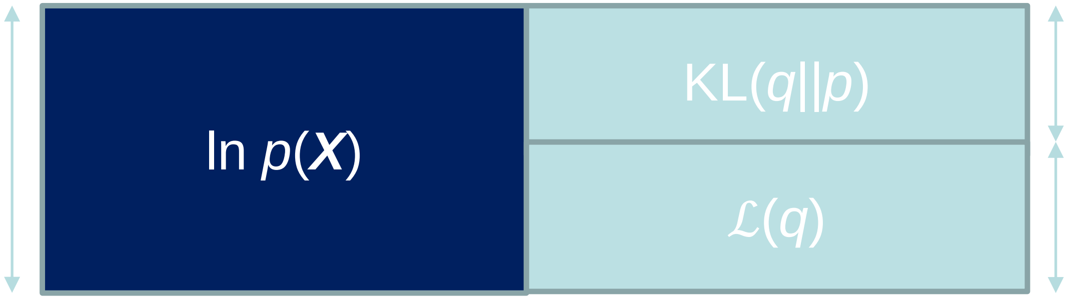
Outline of Variational Inference

The KL-divergence in the context of VI can be thought of as a measure of how good the approximation is.

KL divergence satisfies $KL(q||p) \geq 0$ with equality iff $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$

It therefore follows that $\mathcal{L}(q) \leq \ln p(\boldsymbol{X})$ (since $KL(q||p) = \ln p(x) - \mathcal{L}(q) \geq 0$)

Minimizing the KL divergence $KL(q||p)$ is therefore equivalent to maximizing the lower bound $\mathcal{L}(q)$.



Interpretations of the Variational Objective

If we look our objective as minimizing $-\mathcal{L}(q)$, we can write:

$$-\mathcal{L}(q) = -\int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = -\boldsymbol{H}[\boldsymbol{q}] - \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X},\boldsymbol{Z}) d\boldsymbol{Z} = -\boldsymbol{H}[q] + \mathbb{E}_q[E(\boldsymbol{X},\boldsymbol{Z})]$$

This expected $\mathbb{E}_q[E(\boldsymbol{X},\boldsymbol{Z})]$ value of the energy $E(\boldsymbol{X},\boldsymbol{Z}) = -\ln p(\mathbf{X},\mathbf{Z})$ minus the entropy of the system $\boldsymbol{H}[q]$ is called the *Helmholtz (or variational) free energy*. **Minimizing this free energy is equivalent to maximizing $\boldsymbol{\mathcal{L}(q)}$.**

We can also write:

$$\begin{aligned} &-\mathcal{L}(q) \\ &= -\int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X},\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = -\int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \mathbb{E}_q[-lnp(\boldsymbol{X}|\boldsymbol{Z})] \\ &+ KL(q(\boldsymbol{Z})||p(\boldsymbol{Z})) \end{aligned}$$

This is the **sum of the negative expected log-likelihood (NLL) plus a penalty** measuring the distance between the approximate posterior q(**Z**) and the exact prior p(**Z**).

Information theoretic interpretations are given in the references below.

- Hinton, G. and D. V. Camp (1993). Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*, pp. 5–13. ACM Press.
- Honkela, A. and H. Valpola (2004). Variational Learning and Bits-Back Coding: An Information-Theoretic View to Bayesian Learning. *IEEE. Trans. on Neural Networks 15*(4).

Outline of Variational Inference

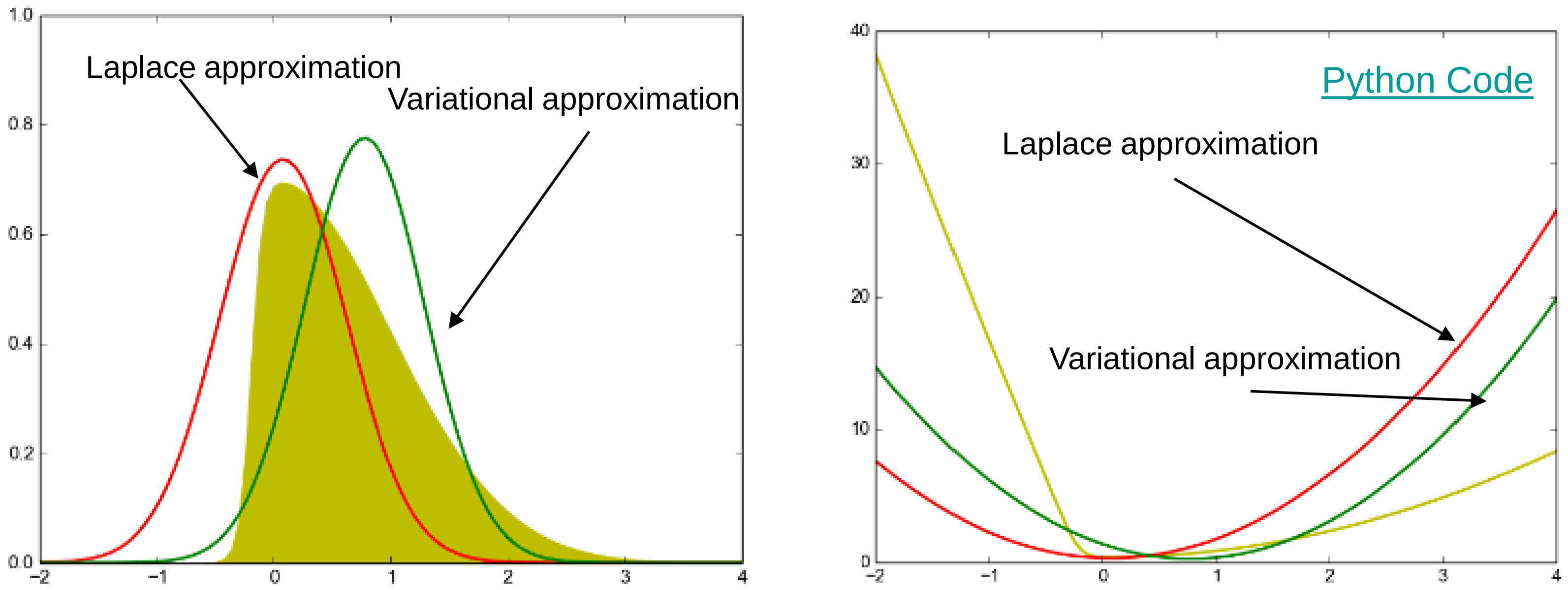
Use factorized approximations of $q(\boldsymbol{Z})$. For example in parametric form use $q(\boldsymbol{Z}|\boldsymbol{\omega})$ and then maximize w.r.t. parameters $\boldsymbol{\omega}$**.**

In the example next, the approximating distribution is taken as a Gaussian with $\boldsymbol{\omega}$ the mean and the variance of the distribution.

- Jordan, M. I., Z. Ghahramani, T. S. Jaakkola, and L. K. Saul (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 105– 162. MIT Press.
- Jaakkola, T. S. and M. I. Jordan (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing 10*, 25–37.
- Jaakkola, T. (2001). Tutorial on variational approximation methods. In M. Opper and D. Saad (Eds.), *Advanced mean field methods*. MIT Press (Presentation)
- Wainwright, M. J. and M. I. Jordan (2008a). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning 1–2*, 1–305.
- Beal, M. (2003). *Variational Algorithms for Approximate Bayesian Inference*. Ph.D. thesis, Gatsby Unit.

Outline of Variational Inference: Example

Example – approximating $\mathrm{f(Z)} = \mathrm{A exp}\left(-\frac{\mathrm{Z}^2}{2}\right)\sigma(20Z+4)$ where *σ*(*Z*) is the logistic sigmoid function defined by *σ*(*Z*) = (1+*e*$^{-Z}$)$^{-1}$. Have restricted $q(Z) \sim \mathcal{N}(\mu, \sigma^2)$ and performed numerical optimization to find values of μ and σ^2 s.t. $KL(q||p)$ is minimized. The Laplace approximation also shown is centered on the mode *of p*(*Z*).



Yellow: original function; red: Laplace approximation; green: variational approximation. Right plot shows the negative logs of the corresponding curves.

Factorized distributions

The goal is to consider a restricted family of distributions $q(\boldsymbol{Z})$ and then seek to minimize the KL divergence.

In restricting the family of $q(\boldsymbol{Z})$ it can be preferential to only consider tractable distributions.

Concept of overfitting does not apply here – expanding the family of distributions can only mean the KL divergence is minimized further.

One approach in restricting the family of distributions $q(\boldsymbol{Z})$ is to partition the elements of $\boldsymbol{Z}$ into disjoint groups $\boldsymbol{Z}_i$ and assuming $q(\boldsymbol{Z})$ can be factorized as follows:

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

No further assumptions are made. This framework corresponds with *Mean Field Theory.*

- Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley.

Factorized distributions

For $q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)$, we seek the distribution for which $\mathcal{L}(q)$ is maximized.

Substituting $q(\boldsymbol{Z})$ into $\mathcal{L}(q)$ we get:

$$\mathcal{L}(q) = \int \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i)\left\{\ln p(\boldsymbol{X},\boldsymbol{Z}) - \sum_{i} \ln \mathrm{q_i}(\boldsymbol{Z}_i)\right\} d\boldsymbol{Z}$$

$$= \int q_j(\boldsymbol{Z}_j)\left\{\int \ln p(\boldsymbol{X},\boldsymbol{Z}) \prod_{i\neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i\right\} d\boldsymbol{Z}_j - \int \sum_{i\neq j} ln q_i(\boldsymbol{Z}_i) \prod_{i\neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i$$

$$- \int q_j\,(\mathbf{Z}_\mathrm{j}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j$$

$$= \int q_j(\boldsymbol{Z}_j) \ln \tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j) d\boldsymbol{Z}_j - \int q_j\,(\mathbf{Z}_\mathrm{j}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + const$$

$$= -KL(q_j(\boldsymbol{Z}_j)||\tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j))$$

where $\ln \tilde{p}(\boldsymbol{X},\boldsymbol{Z}_j) = \int \ln p(\boldsymbol{X},\boldsymbol{Z}) \prod_{i\neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i = \mathbb{E}_{i\neq j}[\ln p(\boldsymbol{X},\boldsymbol{Z})] + const$
i.e. maximizing $\mathcal{L}(q)$ is the equivalent to minimizing the KL divergence above.

Factorized distributions

$\mathcal{L}(q)$ is therefore maximized when $q_j\left(\boldsymbol{Z}_j\right) = \tilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$

The optimal solution can be written as:

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

Taking the exp and normalizing gives the following

$$q_j^*\left(\boldsymbol{Z}_j\right) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})])}{\int exp\left(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]\ \right) d\boldsymbol{Z}_j}$$

These two equations form the basis for variational inference.

Note that each $q_i(\boldsymbol{Z}_i)$ depends on the other factors $q_i(\boldsymbol{Z}_i)$ for $\mathrm{i} \neq j$. We therefore proceed by initializing all factors and cycle through the factors updating one at a time until some convergence criteria (self-consistency) is satisfied.

Convergence is guaranteed due to the convexity of the bound $\mathcal{L}(q)$ wrt each factor $q_j\left(\boldsymbol{Z}_j\right)$.

- Boyd, S. and L. Vandenberghe (2004). [Convex Optimization](). Cambridge University Press.

The mean field method

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

In graphs, we only need to consider the variables that share a factor with $\boldsymbol{Z}_j$ (jth's Markov blanket).

Since we are replacing all other variables with their mean values, this method is called the mean field approximation.

This is similar to Gibbs sampling but instead of sending sampled values between neighboring nodes, we send the mean values. This is efficient as the mean is a proxy for a large number of samples.

Mean field messages are dense whereas samples are sparse – so sampling is more scalable for large scale problems.

The iterative approach can be accelerated with pattern search and/or parameter expansion.

- Opper, M. and D. Saad (Eds.) (2001). *Advanced mean field methods: theory and practice*. MIT Press.
- Honkela, A., H. Valpola, and J. Karhunen (2003). Accelerating Cyclic Update Algorithms for Parameter Estimation by Pattern Searches. *Neural Processing Letters 17*, 191–203.
- Qi, Y. and T. Jaakkola (2008). Parameter Expanded Variational Bayesian Methods. In *NIPS*.

Variational approximation of a Gaussian

We discuss the problem of approximating a Gaussian $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$, $\mathbf{z} = (z_1,\ z_2)$ using a factorized Gaussian. The mean and precision are

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix},\ \boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}, \Lambda_{12} = \Lambda_{21}$$

Now suppose we wish to approximate this distribution using a factorized Gaussian of the form $q(\mathbf{z}) = q_1(z_1)q_2(z_2)$. We apply the general result to find an expression for the $q_1^*(z_1)$

$$\ln q_1^*(z_1) = \mathbb{E}_{z_2}\left[\ln p(\mathbf{z})\right] + const$$

$$= \mathbb{E}_{z_2}\left[-\frac{1}{2}(z_1-\mu_1)^2\Lambda_{11} - (z_1-\mu_1)\Lambda_{12}(z_2-\mu_2)\right] + const$$

$$= -\frac{1}{2}z_1^2\Lambda_{11} + z_1\mu_1\Lambda_{11} - z_1\Lambda_{12}\left(\mathbb{E}_{z_2}\left[z_2\right] - \mu_2\right) + const$$

Completing the square, we find that $q_1^*(z_1)$ and $q_2^*(z_2)$ are Gaussians (coupled solutions)

$$q_1^*(z_1) = \mathcal{N}\left(z_1 \,|\, m_1, \Lambda_{11}^{-1}\right),\ m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mathbb{E}[z_2] - \mu_2\right) = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(m_2 - \mu_2\right)$$

$$q_2^*(z_2) = \mathcal{N}\left(z_2 \,|\, m_2, \Lambda_{22}^{-1}\right),\ m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(\mathbb{E}[z_1] - \mu_1\right) = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}\left(m_1 - \mu_1\right)$$

We can show that the only solution is obtained (for $|\boldsymbol{\Lambda}| = \Lambda_{11}\Lambda_{22} - \Lambda_{21}\Lambda_{12} \neq 0$) when:

$$\mathbb{E}[z_1] = m_1 = \mu_1,\ \mathbb{E}[z_2] = m_2 = \mu_2,\ q_1^*(z_1) = \mathcal{N}\left(z_1 \,|\, \mu_1, \Lambda_{11}^{-1}\right),\ q_2^*(z_2) = \mathcal{N}\left(z_2 \,|\, \mu_2, \Lambda_{22}^{-1}\right)$$

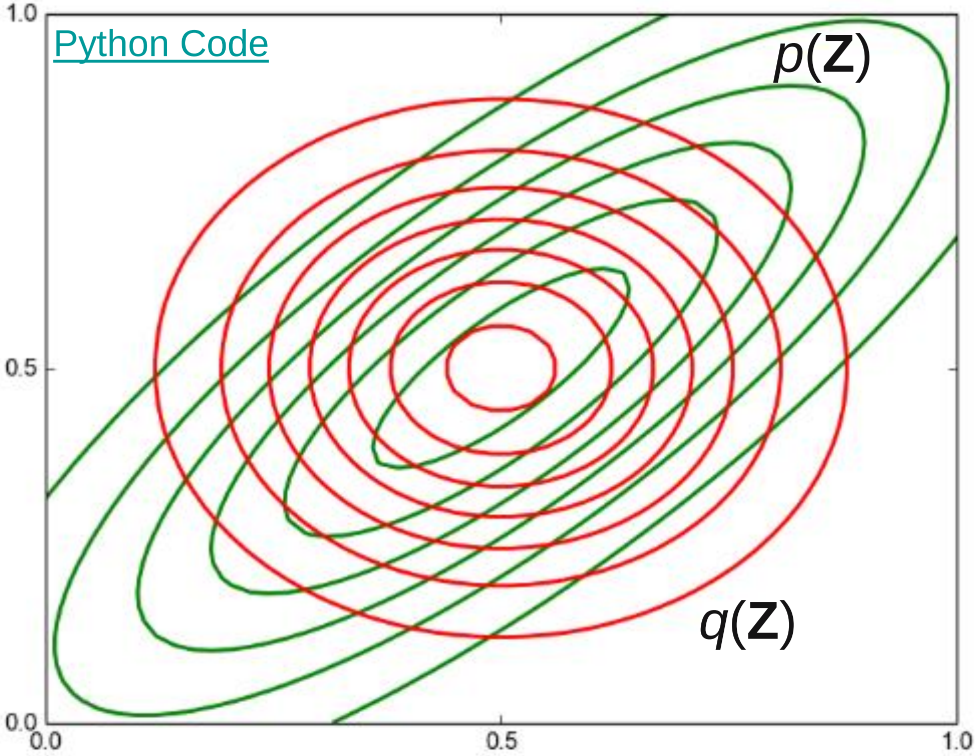
Factorized distributions

Approximating the 2D Gaussian distribution $p(z_1, z_2) = \mathcal{N}(\boldsymbol{Z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ by a factorized distribution $\mathrm{q}(\mathrm{z}_1, \mathrm{z}_2) = q_1(z_1)q_2(z_2)$ gives the closed form solution $\mathbb{E}[z_1] = \mu_1$ and $\mathbb{E}[z_2] = \mu_2$ when minimizing the KL divergence.

The mean is correctly captured but the variance of $q(\mathbf{z})$ is controlled by the direction of smallest variance of $p(\mathbf{z})$, and that the variance along the orthogonal direction is under-estimated.

A factorized variational approximation tends to give *compact approximations* (*overconfident*) to the posterior.

- Turner, R., P. Berkes, M. Sahani, and D. Mackay (2008). Counterexamples to variational free energy compactness folk theorems. Technical report, U. Cambridge.



Alternative Approximate Inference

Whilst variational inference is chiefly concerned with minimizing the KL divergence, we can also consider *alternative forms of KL divergence*.

For example, the reverse KL divergence:

$$KL(p||q) = -\int p(\boldsymbol{Z}) \ln\left\{\frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z})}\right\} d\boldsymbol{Z} = \int p(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$$

Note that in general $KL(p||q) \neq KL(q||p)$ and therefore using the reverse KL divergence would yield different results. Keeping only terms in $q_j(\boldsymbol{Z}_j)$, we can write:

$KL(p||q) = -\int p(\boldsymbol{Z}) \sum_i \ln q_i(\boldsymbol{Z}_i)\, d\boldsymbol{Z} + const = -\int p(\boldsymbol{Z}) \ln q_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z} + const =$
$-\int \ln q_j(\boldsymbol{Z}_j)[\int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_{\boldsymbol{i}}]\, d\boldsymbol{Z}_{\boldsymbol{j}} + const = -\int \ln q_j(\boldsymbol{Z}_j) F_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_{\boldsymbol{j}} + const$, with $F_j(\boldsymbol{Z}_j) =$
$\int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_{\boldsymbol{i}}$.

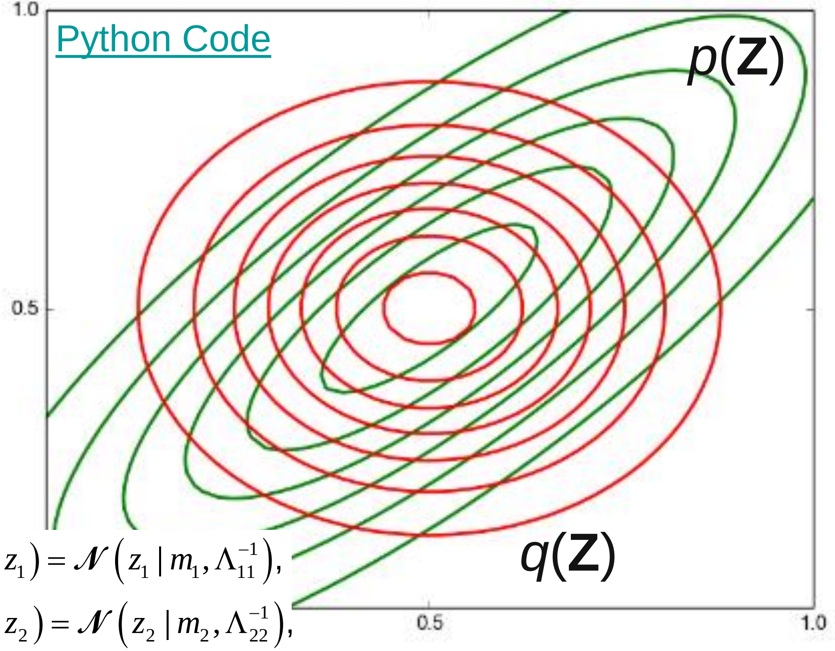
Optimize wrt $q_j(\boldsymbol{Z}_j)$:$-\int \ln q_j(\boldsymbol{Z}_j) F_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_{\boldsymbol{j}} + \lambda(\int q_j(\boldsymbol{Z}_j)\, d\boldsymbol{Z}_j - 1)$.

This gives: $\lambda = \frac{F_j(\boldsymbol{Z}_j)}{q_j(\boldsymbol{Z}_j)}$. Integrating over $\boldsymbol{Z}_j$ gives λ=1. Thus we conclude:

$$q_j^*(\boldsymbol{Z}_j) = F_j(\boldsymbol{Z}_j) = \int p(\boldsymbol{Z}) \prod_{i\neq j} d\boldsymbol{Z}_i = p(\boldsymbol{Z}_j) \text{ (marginal, no iteration needed)}$$

Comparison of the two forms for the KL divergence

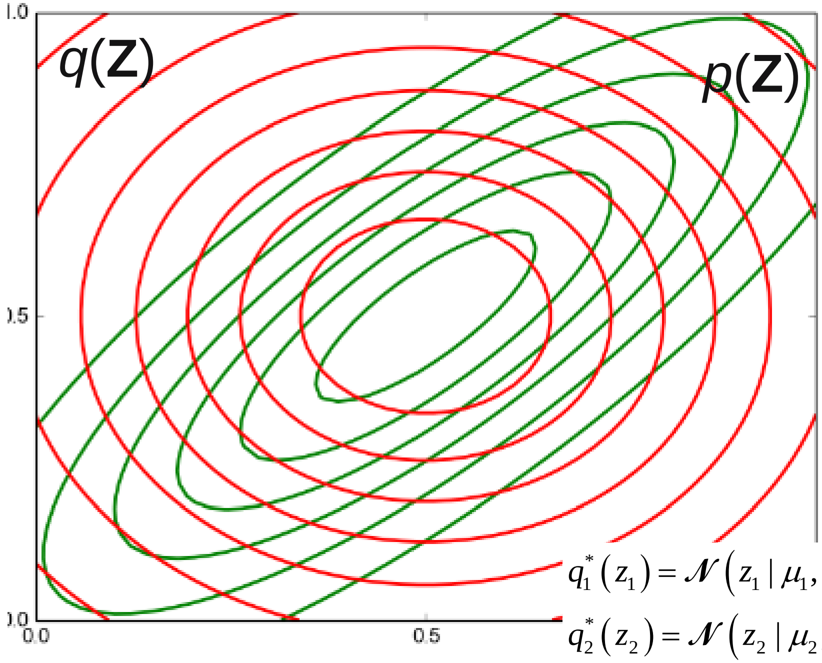
Consider a correlated Gaussian $p(\boldsymbol{Z})$ over z_1 and z_2 and the approximating $q(\mathbf{Z})$ over the same variables given by the product of two independent univariate Gaussian distributions. We obtain different results using alternate KL measures.


Python Code

$q_1^*(z_1) = \mathcal{N}\left(z_1 \mid m_1, \Lambda_{11}^{-1}\right),$

$q_2^*(z_2) = \mathcal{N}\left(z_2 \mid m_2, \Lambda_{22}^{-1}\right),$

Mininimization of $KL(q||p)$



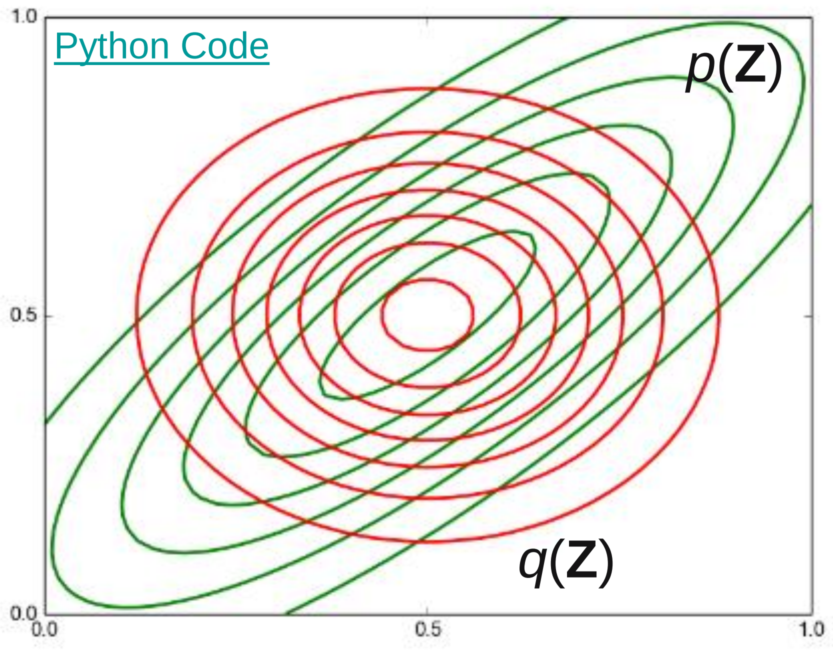
$q_1^*(z_1) = \mathcal{N}\left(z_1 \mid \mu_1, \Sigma_{11}\right),$

$q_2^*(z_2) = \mathcal{N}\left(z_2 \mid \mu_2, \Sigma_{22}\right),$

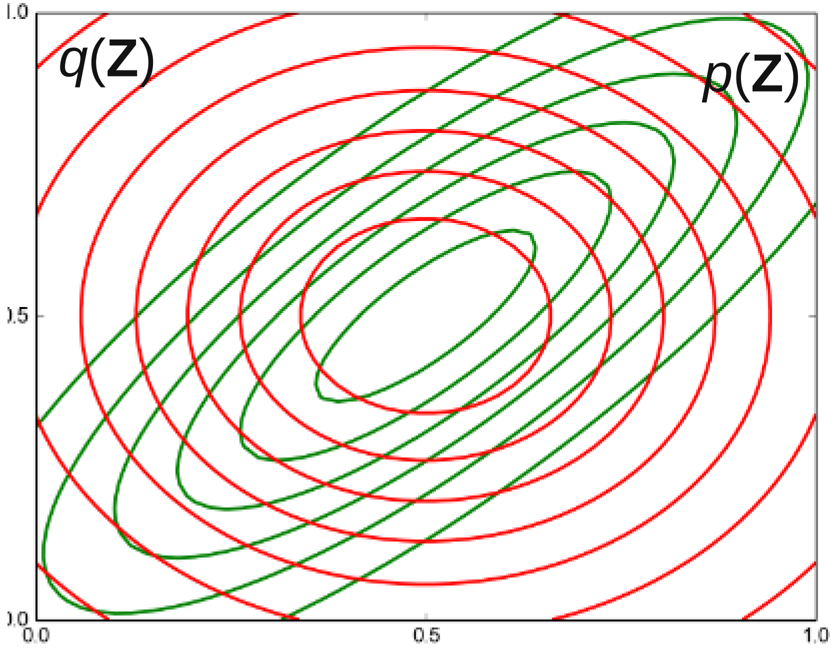
Mininimization of $KL(p||q)$

There is a large positive contribution to $KL(q||p) = -\int q(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$ when p($\boldsymbol{Z}|\boldsymbol{X}$) is close to 0 but q($\boldsymbol{Z}$) is not. *q(**Z**)* avoids regions where *p(**Z**)* is small (*zero forcing* for *q*(**Z**)). *q(**Z**)* underestimates in the direction of the highest variance

Comparison of the two forms for the KL divergence



Minimization of $KL(q||p)$



Minimization of $KL(p||q)$

On the other hand $KL(p||q) = \int p(\boldsymbol{Z}) \ln\left\{\frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})}\right\} d\boldsymbol{Z}$ is infinite if *q*(**Z**)=0 when *p*(**Z**)>0. So it is minimized by *q(**Z**)>0* in regions where *p(**Z**)*>0 (*zero-avoiding*)

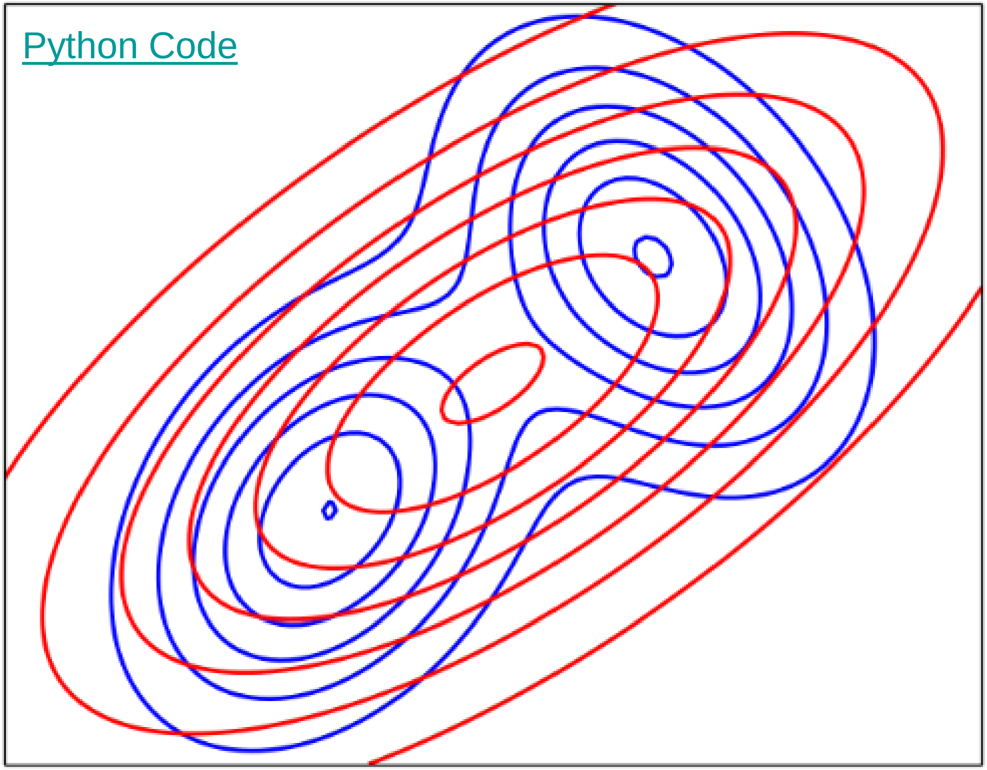
*q(**Z**)* puts non-zero mass on regions of low-probability of *p(**Z**)*. $KL(p||q)$ *overestimates the support of p(**Z**).*

The two forms of the KL divergence for a mixture of Gaussians

Example – approximating a mixture of Gaussian distributions using KL divergence

- ❑ If we were to minimize KL(*p*||*q*), the resulting approximations would average across all of the modes and, in the context of the mixture model, would lead to poor predictive distributions.

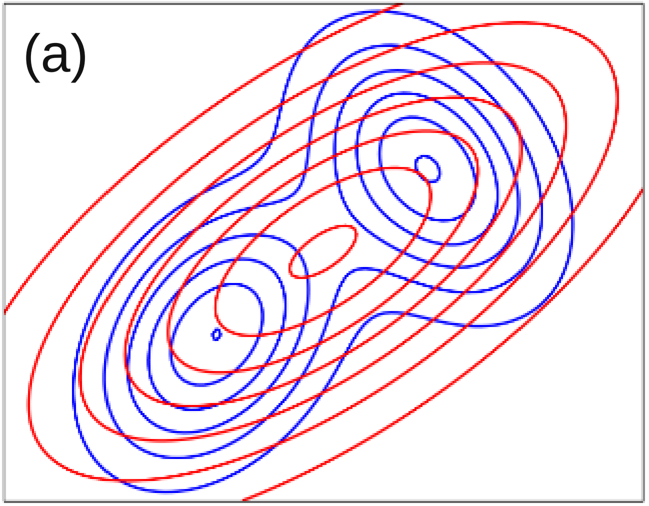
- ❑ It is possible to make use of KL(*p*||*q*) to define a useful inference procedure, but this requires a rather different approach (expectation propagation).

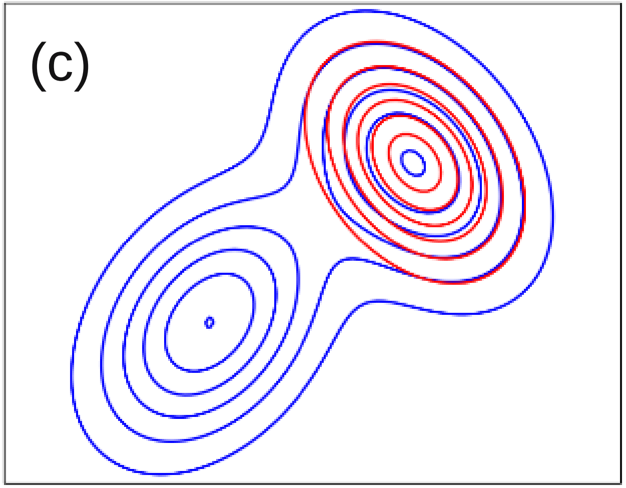


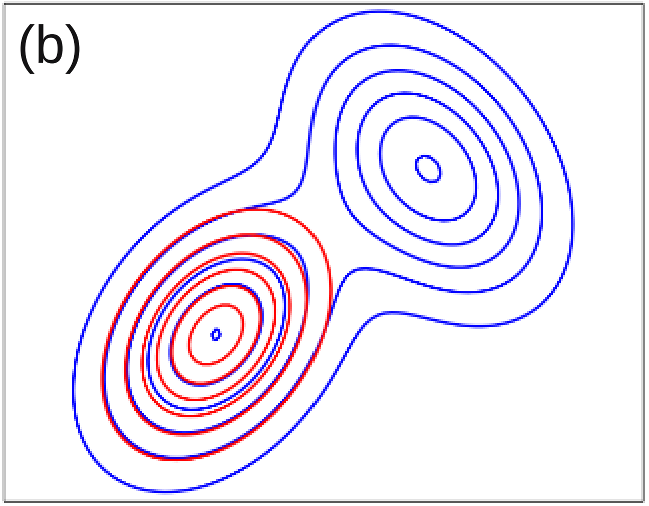
Minimization of $KL(p||q)$

The two forms of the KL divergence for a mixture of Gaussians

Consider $p(\mathbf{Z})$ a mixture of two Gaussians.
(a)The red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing KL($p||q$).
(b,c) The red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of KL($q||p$). Two local minima of the KL distance are shown.





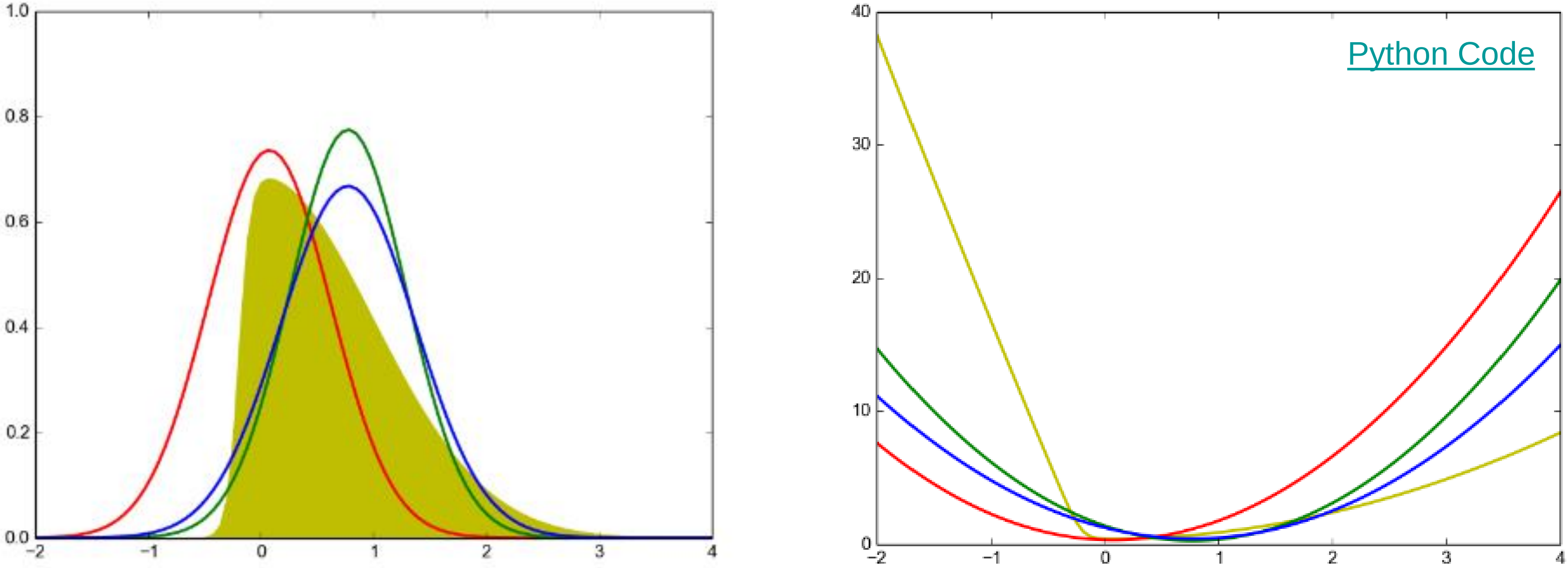


Python Code

Alternative Approximate Inference

Example – returning to approximating $\mathrm{f(Z)} = \mathrm{A}\exp\left(-\frac{\mathrm{Z}^2}{2}\right)\sigma(20Z+4)$

Restrict $q(Z) \sim \mathcal{N}(\mu, \sigma^2)$ and perform optimization to find values of μ and σ^2 s.t. $KL(q||p)$ is minimized. The results of Expectation-Propagation are also shown (to be discussed later on). The EP distribution is broader than the VI due to the different KL distance used.



Python Code

Yellow: original distribution; red: Laplace approximation; green: variational approximation, blue: expectation propagation

Gaussian Approximation $q(\boldsymbol{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\Sigma)$ of $p(\boldsymbol{x})$

Consider minimizing the distance *KL(p||q).*

$$KL(p\,\|\,q) = -\int p(\boldsymbol{x})\ln q(\boldsymbol{x})d\boldsymbol{x} + \int p(\boldsymbol{x})\ln p(\boldsymbol{x})d\boldsymbol{x}$$

$$= -\int p(\boldsymbol{x})\left(-\frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)d\boldsymbol{x} + const$$

$$= \frac{1}{2}\left(\ln|\boldsymbol{\Sigma}| + tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right)\right)\right) + const.$$

$$= \frac{1}{2}\left(\ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) + tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)\right) + const.$$

We take derivatives wrt $\boldsymbol{\mu}$ and then $\boldsymbol{\Sigma}^{-1}$. We will use the following identities:

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{a}\right) = \boldsymbol{a},\ \frac{\partial}{\partial \boldsymbol{A}}\left(\boldsymbol{a}^T\boldsymbol{A}\boldsymbol{b}\right) = \boldsymbol{a}\boldsymbol{b}^T,\ \frac{\partial}{\partial \boldsymbol{A}}tr\left(\boldsymbol{A}^T\boldsymbol{B}\right) = \boldsymbol{B},\ \frac{\partial}{\partial \boldsymbol{A}}\ln|\boldsymbol{A}| = \left(\boldsymbol{A}^{-1}\right)^T$$

$$\frac{\partial KL(p\,\|\,q)}{\partial \boldsymbol{\mu}} = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}}\left(\ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) + tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)\right) = \frac{1}{2}\left(2\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x})\right) = 0 \Rightarrow \boldsymbol{\mu} = \mathbb{E}(\boldsymbol{x})$$

$$\frac{\partial KL(p\,\|\,q)}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}^{-1}}\left(\ln|\boldsymbol{\Sigma}| + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\mathbb{E}(\boldsymbol{x}) + tr\left(\boldsymbol{\Sigma}^{-1}\mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right)\right) = \frac{1}{2}\left(-\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right)\right) = 0 \Rightarrow$$

$$\boldsymbol{\Sigma} = \mathbb{E}\left(\boldsymbol{x}\boldsymbol{x}^T\right) - \mathbb{E}(\boldsymbol{x})\mathbb{E}(\boldsymbol{x})^T = \mathrm{cov}(\boldsymbol{x})$$

Alpha Divergence

But why restrict ourselves to KL divergence? The *alpha family* of divergence is defined as follows:

$$D_\alpha(p||q) = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{\frac{1+\alpha}{2}} q(x)^{\frac{1-\alpha}{2}} dx\right), -\infty < \alpha < \infty\text{: } D_\alpha(p||q) \geq 0,\ D_\alpha(p||q) = 0\ iff\ p = q$$

This is not symmetric and thus not a distance. Lets now introduce the following notation:

$$\gamma_p = \frac{1+\alpha}{2} = 1-\gamma_q, \gamma_q = \frac{1-\alpha}{2}$$

Note that: $\gamma_p \underset{\alpha\to 1}{\to} 1, \gamma_q \underset{\alpha\to 1}{\to} 0$. We can now simplify $D_\alpha(p||q)$ using Taylor series expansions of the two terms inside the integral:

$$q^{\gamma_q} = \exp\left(\gamma_q \ln q\right) = 1 + \gamma_q \ln q + \mathrm{O}\left(\gamma_q^2\right)$$

$$p^{\gamma_p} = \exp\left(\gamma_p \ln p\right) = \exp\left(\left(1-\gamma_q\right)\ln p\right) = p\exp\left(-\gamma_q \ln p\right) = p - \gamma_q p \ln p + \mathrm{O}\left(\gamma_q^2\right)$$

$$\int q^{\gamma_q} p^{\gamma_p} dx = \int\left(1+\gamma_q \ln q + \mathrm{O}\left(\gamma_q^2\right)\right)\left(p - \gamma_q p \ln p + \mathrm{O}\left(\gamma_q^2\right)\right)dx = \int\left(p + \gamma_q\left(p \ln q - p \ln p\right)\right)dx + \mathrm{O}\left(\gamma_q^2\right) = 1 + \int \gamma_q\left(p \ln q - p \ln p\right)dx + \mathrm{O}\left(\gamma_q^2\right)$$

- Ali, S. M. and S. D. Silvey (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, B* **28**(1), 131–142.
- Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer.
- Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Msft Research.

Alpha Divergence

$D_\alpha(p||q)$ for the case $\alpha \to 1$ becomes:

$$For\ \alpha \to 1: D_\alpha(p\|q) = -\frac{4\gamma_q}{1-\alpha^2}\int\left(p\ln q - p\ln p\right)dx + \mathrm{O}\left(\gamma_q^2\right) = \frac{2}{1+\alpha}KL(p\|q) + \mathrm{O}\left(\gamma_q^2\right) \underset{\alpha\to 1}{\to} KL(p\|q)$$

Similar result can be shown for KL(q||p). Thus

$$KL(p||q) = \lim_{\alpha\to 1} D_\alpha(p||q)$$
$$KL(q||p) = \lim_{\alpha\to -1} D_\alpha(p||q)$$

In general:

$\alpha \leq -1$ the divergence is zero forcing, *q(x)=0* when *p(x)=0* (under-estimated support, seek for mode with largest mass)

$\alpha \geq 1$ the divergence is zero avoiding, *q(x)>0* when *p(x)>0* (over-estimated support)

$\alpha = 0$ gives a distance linearly related to the *Hellinger Distance* ($\sqrt{D_H(p||q)}$ is a valid distance metric, i.e. symmetric, non-negative and satisfies the triangle inequality)

$$D_H(p||q) = \int\left(p(x)^{\frac{1}{2}} - q(x)^{\frac{1}{2}}\right)^2 dx$$

- Ali, S. M. and S. D. Silvey (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, B* **28**(1), 131–142.
- Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer.
- Minka, T. (2005). Divergence measures and message passing. Technical Report MSR-TR-2005-173, Msft Research.

Variational Bayes EM

Consider a set of hidden variables ***Z*** that includes latent variables ***z*** and parameters θ. Assume a factorization of the posterior in the form q(**Z**)=q(**z**)q(θ), where q(θ)=δ(θ-θ_0). Minimizing the distance *KL(q||p)* leads to the following:

$$KL(q\,||\,p) = -\iint q(\boldsymbol{\theta})q(\mathbf{z})\ln\frac{p(\mathbf{z},\boldsymbol{\theta}\,|\,\boldsymbol{X})}{q(\mathbf{z})q(\boldsymbol{\theta})}d\mathbf{z}d\boldsymbol{\theta}$$

Start with the particular choice of q(θ) for a given θ_0, KL(q||p) simplifies to:

$$KL(q\,||\,p) = -\int q(\mathbf{z})\ln\frac{p(\mathbf{z},\boldsymbol{\theta}_0\,|\,\boldsymbol{X})}{q(\mathbf{z})}d\mathbf{z} + const = -\int q(\mathbf{z})\ln\frac{p(\mathbf{z}\,|\,\boldsymbol{\theta}_0,\boldsymbol{X})p(\boldsymbol{\theta}_0\,|\,\boldsymbol{X})}{q(\mathbf{z})}d\mathbf{z} + const$$

$$= -\int q(\mathbf{z})\ln\frac{p(\mathbf{z}\,|\,\boldsymbol{\theta}_0,\boldsymbol{X})}{q(\mathbf{z})}d\mathbf{z} + const$$

This is maximized for $q(\mathbf{z}) = p(\mathbf{z}\,|\,\boldsymbol{\theta}_0,\boldsymbol{X})$. This is the E-step of the EM algorithm. Now with the updated q(***z***), we can compute q(θ) (new estimate of θ_0) by maximizing:

$$\int q(\boldsymbol{\theta})\int q(\mathbf{z})\ln\frac{p(\boldsymbol{X},\boldsymbol{\theta},\mathbf{z})}{q(\mathbf{z})q(\boldsymbol{\theta})}d\mathbf{z}d\boldsymbol{\theta} + const = \int q(\boldsymbol{\theta})\mathbb{E}_{q(\mathbf{z})}\left[\ln p(\boldsymbol{X},\boldsymbol{\theta},\mathbf{z})\right]d\boldsymbol{\theta} - \int q(\boldsymbol{\theta})\ln q(\boldsymbol{\theta})d\boldsymbol{\theta} + const$$

For the particular form of q(θ) this is maximized by taking $\max_{\theta_0}\mathbb{E}_{q(\mathbf{z})}\left[\ln p(\boldsymbol{X},\boldsymbol{\theta}_0,\mathbf{z})\right]$
This is exactly the M-step of the EM algorithm.
Note: The entropy $-\int q(\boldsymbol{\theta})\ln q(\boldsymbol{\theta})d\boldsymbol{\theta}$ is constant independent of θ_0 (formally diverging).

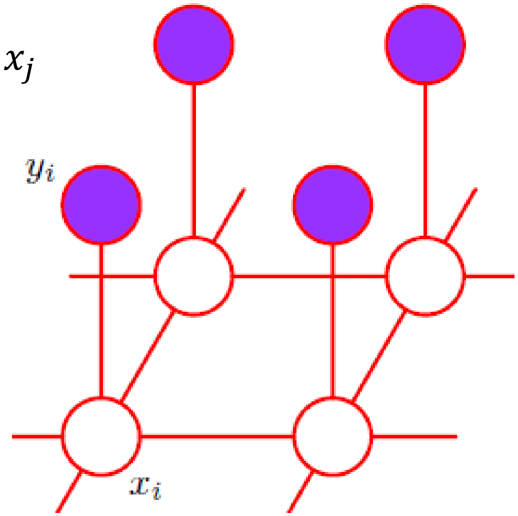
Mean Field for the Ising Model

Consider image denoising, where $x_i \in \{-1,+1\}$ are the hidden pixel values of the clean image. We have a joint model of the form

$$p(\boldsymbol{x},\boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})$$

where the prior has the form

$$p(\boldsymbol{x}) = \frac{1}{Z_0}\exp\big(-E_0(\boldsymbol{x})\big), \qquad E_0(\boldsymbol{x}) = -\sum_{i=1}^{D}\sum_{j\in nbr_i} W_{ij}x_i x_j$$



and the likelihood has the form

$$p(\boldsymbol{y}|\boldsymbol{x}) = \prod_i p(y_i|x_i) \; = exp\left(\sum L_i\,(x_i)\right)$$

Therefore the posterior has the form

$$p(\boldsymbol{x}|\boldsymbol{y}) = 1/Z\,\exp(-E(\boldsymbol{x})),\, E(\boldsymbol{x}) {=} E_0(\boldsymbol{x}) - \textstyle\sum_i \; L_i\,(x_i)$$

We will now approximate this by a fully factored approximation

$$q(\mathbf{x}) = \textstyle\prod_i \; q(x_i,\mu_i)$$

where μ_i is the mean value of node *i*. To derive the update for the variational parameter μ_i, we first write out $ln\tilde{p}(\boldsymbol{x}) = lnp(\boldsymbol{x},\boldsymbol{y}) = -E(\boldsymbol{x})$, and dropping terms that do not involve x_i:

$$ln\,\tilde{p}(\boldsymbol{x}) = x_i \textstyle\sum_{j\in nbr_i} W_{ij}x_j + L_i\,(x_i) + \text{const}$$

Mean Field for the Ising Model

$$ln\,\tilde{p}(\boldsymbol{x}) = x_i \sum_{j \in nbr_i} W_{ij} x_j + L_i\,(x_i) + const$$

This only depends on the states of the neighboring nodes. Now we take expectations of this wrt $\prod_{j \neq i} q_j(x_j)$ to get

$$q_i(x_i) \propto \exp\left(x_i \sum_{j \in nbr_i} W_{ij}\mu_j + L_i\,(x_i) \right)$$

Thus we replace the states of the neighbors by their average values. Let $m_i = \sum_{j \in nbr_i} W_{ij}\mu_j$ be the mean field influence on node *i*. Also let $L_i^+ \equiv L_i(+1), L_i^- \equiv L_i(-1)$. The approximate marginal posterior is given by

$$q_i(x_i = 1) = \frac{e^{m_i\ + L_i^+}}{e^{m_i\ + L_i^+} + e^{-m_i\ + L_i^-}} = \frac{1}{1 + e^{-2m_i\ + L_i^- - L_i^+}} = sigm(2a_i),\ a_i \equiv m_i\ \ + 0.5(L_i^+ - L_i^-)$$

Similarly $q_i(x_i = -1) = sigm(-2a_i)$ and thus:$\mu_i = \mathbb{E}_{q_i}[x_i] = q_i(x_i = 1)(+1) + q_i(x_i =$

Mean Field for the Ising Model

$$\mu_i = \tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_j + 0.5(L_i^+ - L_i^-)\right)$$

We can turn the above Eqs in to a fixed point algorithm by writing:

$$\mu_i^t = \tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_i^{t-1} + 0.5(L_i^+ - L_i^-)\right)$$

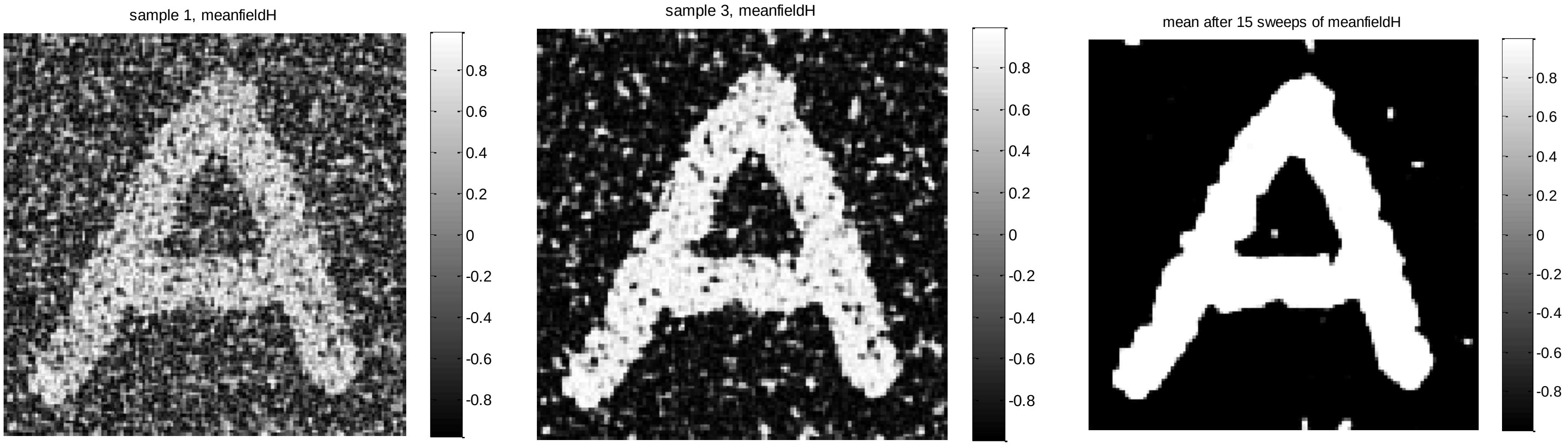
To avoid checkerboarding effects, damped updates are often used as follows:

$$\mu_i^t = (1-\lambda)\mu_i^{t-1} + \lambda\tanh\left(\sum_{j\in nbr_i} W_{ij}\mu_i^{t-1} + 0.5(L_i^+ - L_i^-)\right),\ 0 < \lambda < 1.$$

The nodes can be updated in parallel or asynchronously.

Mean Field for the Ising Model

- Example of image denoising using mean field
- Parallel updates and λ =0.5.
- Ising prior with W_{ij} = 1 and a Gaussian noise model with σ = 2.
- Results are shown after 1, 3 and 15 iterations.



Run isingImageDenoiseDemo in the
PMTK3 toolbox

Structured Mean Field Approach

Assuming that all the variables are independent in the posterior is a very strong assumption that can lead to poor results.

Sometimes we can exploit **tractable substructure** in our problem to efficiently handle particular dependencies. This is called the **structured mean field approach**.

We group sets of variables together, and we update them simultaneously. We treat the variables in the *i*'th group as a single "mega-variable" and we follow the earlier variational derivation.

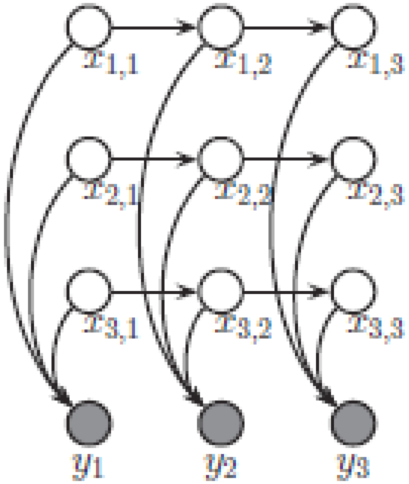
If we can perform efficient inference in each q_i, the method is tractable overall.

A factorial HMM example is discussed next.

- Saul, L. and M. Jordan (1995). Exploiting tractable substructures in intractable networks. In *NIPS*, Volume 8.
- Ghahramani, Z. and M. Jordan (1997). Factorial hidden Markov models. *Machine Learning 29*, 245–273.
- Bouchard-Cote, A. and M. Jordan (2009). Optimization of structured mean field objectives. In *UAI*.

Factorial HMM

Consider the factorial HMM model. Suppose there are *M* chains, each of length *T*, and suppose each hidden node has *K* states. The model is defined as follows



$$p(\boldsymbol{x},\boldsymbol{y}) = \prod_m \prod_t p\big(x_{tm}|x_{t-1,m}\big)\, p(\boldsymbol{y}_t|x_{tm})$$

where $p\big(x_{tm} = k|x_{t-1,m} = j\big)= A_{mjk}$ is an entry in the transition matrix for chain *m*, $p(x_{1m} = k|x_{0m}) = p(x_{1m} = k) = \pi_{mk}$, is the initial state distribution for chain *m*, and

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{y}_t|\sum_{m=1}^{M} \boldsymbol{W}_m \boldsymbol{x}_{tm}, \boldsymbol{\Sigma}\right)$$

is the observation model, where $\boldsymbol{x}_{tm}$ is a 1-of-*K* encoding of x_{tm} and $\boldsymbol{W}_m$ is a *D* × *K* matrix (assuming $\boldsymbol{y_t} \in \mathbb{R}^D$).

Even though each chain is a priori independent, they become coupled in the posterior due to having an observed common child, $\mathbf{y}_t$. The junction tree algorithm applied to this graph takes $\mathcal{O}(TMK^{M+1})$ time.

The structured mean field algorithm discussed next takes $\mathcal{O}(TMK^2I)$ time, where *I* is the number of mean field iterations (I ~ 10 for good performance).

Factorial HMM

We can write the exact posterior as:

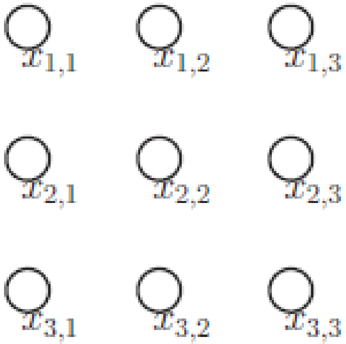
$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z}\exp(\,-E(\boldsymbol{x},\boldsymbol{y})\,)$$

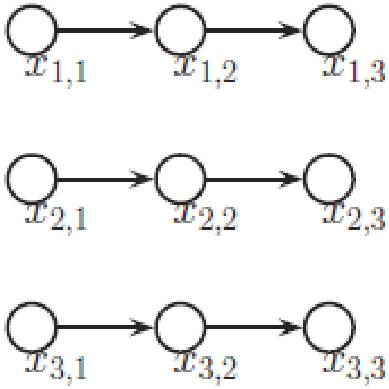
$$E(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{y}_t - \sum_m \boldsymbol{W}_m\boldsymbol{x}_{tm}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \sum_m \boldsymbol{W}_m\boldsymbol{x}_{tm}\right) - \sum_m \boldsymbol{x}_{1m}^T\widetilde{\boldsymbol{\pi}}_m - \sum_{t=2}^{T}\sum_m \boldsymbol{x}_{tm}^T\widetilde{\boldsymbol{A}}_m\,\boldsymbol{x}_{t-1,m}^T$$

where pointwise $\widetilde{\boldsymbol{A}}_m = ln\boldsymbol{A}_m$, and $\widetilde{\boldsymbol{\pi}}_m$=$ln\boldsymbol{\pi}_m$. $\boldsymbol{x}_{tm}$ is a vector with 1 in one location and zero elsewhere else.

We can approximate the posterior as a product of marginals (see Fig on the left)

But a better approximation is to use a product of chains (Fig on the right) . Each chain can be tractably updated individually using the forwards-backwards algorithm.





Factorial HMM

We assume:

$$q(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z_q}\prod_{m=1}^{M}\left(q(\boldsymbol{x}_{1m}|\boldsymbol{\xi}_{1m})\prod_{t=2}^{T} q\left(\boldsymbol{x}_{tm}|\boldsymbol{x}_{t-1,m},\boldsymbol{\xi}_{tm}\right)\right)$$

$$q(\boldsymbol{x}_{1m}|\boldsymbol{\xi}_{1m}) = \prod_{k=1}^{K}(\xi_{1mk}\pi_{mk})^{x_{1mk}}$$

$$q\left(\boldsymbol{x}_{tm}|\boldsymbol{x}_{t-1,m},\boldsymbol{\xi}_{tm}\right)=\prod_{k=1}^{K}\left(\xi_{tmk}\prod_{j=1}^{K}\left(A_{mjk}\right)^{x_{t-1,m,j}}\right)^{x_{tmk}}$$

We see that the parameters $\boldsymbol{\xi}_{1m}$ play the role of an approximate local evidence averaging out the effects of the other chains. Comparing with $p(\boldsymbol{x},\boldsymbol{y}) = \prod_m\prod_t\ p\left(\boldsymbol{x}_{tm}|\boldsymbol{x}_{t-1,m}\right)p(\boldsymbol{y}_t|\boldsymbol{x}_{tm})$, the $\mathrm{K}\times 1$ vector $\boldsymbol{\xi}_{tm}$ plays the role of the probability of an observation $p(\boldsymbol{y}_t|\boldsymbol{x}_{tm})$ for each of the K settings of $\boldsymbol{x}_{tm}$.

This is in contrast to the exact local evidence which coupled all the chains together. We can rewrite the approximate posterior as $q(\boldsymbol{x}) = \frac{1}{Z_q}\exp\left(-E_q(\boldsymbol{x})\right)$ where

$$E_q(\boldsymbol{x})= -\sum_{t=1}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm} - \sum_{m=1}^{M}\boldsymbol{x}_{1m}^T\tilde{\boldsymbol{\pi}}_m - \sum_{t=2}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^T\tilde{\boldsymbol{A}}_m\boldsymbol{x}_{t-1,m},\ \tilde{\boldsymbol{\xi}}_{tm}=\ln\boldsymbol{\xi}_{tm}$$

Factorial HMM

$$q(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z_q}\prod_{m=1}^{M} q(\boldsymbol{x}_{1m}|\boldsymbol{\xi}_{1m})\prod_{t=2}^{T} q\left(\boldsymbol{x}_{tm}|\boldsymbol{x}_{t-1,m},\boldsymbol{\xi}_{tm}\right), \qquad q(\boldsymbol{x}_{1m}|\boldsymbol{\xi}_{1m}) = \prod_{k=1}^{K}(\xi_{1mk}\pi_{mk})^{x_{1mk}}$$

$$q\left(\boldsymbol{x}_{tm}|\boldsymbol{x}_{t-1,m},\boldsymbol{\xi}_{tm}\right)=\prod_{k=1}^{K}\left(\xi_{tmk}\prod_{j=1}^{K}\left(A_{mjk}\right)^{x_{t-1,mj}}\right)^{x_{tmk}}$$

$$E_q(\boldsymbol{x})= -\sum_{t=1}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^{T}\tilde{\boldsymbol{\xi}}_{tm} - \sum_{m=1}^{M}\boldsymbol{x}_{1m}^{T}\tilde{\boldsymbol{\pi}}_{m} - \sum_{t=2}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^{T}\widetilde{\boldsymbol{A}}_{m}\boldsymbol{x}_{t-1,m}, \ \ \tilde{\boldsymbol{\xi}}_{tm} = \ln\boldsymbol{\xi}_{tm}$$

This has the same temporal factors as the exact posterior but the local evidence is different. The objective function $KL(q||p) = -\int q ln\frac{p}{q}d\boldsymbol{x}$ is given by:

$$KL(q||p) = \int qlnqd\boldsymbol{x} - \int qlnpd\boldsymbol{x} = \int qln\frac{e^{-E_q}}{Z_q}d\boldsymbol{x} - \int qln\frac{e^{-E}}{Z}d\boldsymbol{x} = \mathbb{E}_q[E] - \mathbb{E}_q\left[E_q\right] - lnZ_q + lnZ$$

Let $\overline{\boldsymbol{x}}_{tm} = \mathbb{E}_q[\boldsymbol{x}_{tm}]$. Using, $E(\boldsymbol{x},\boldsymbol{y}) = \frac{1}{2}\sum_{t=1}^{T}\left(\boldsymbol{y}_t - \sum_m \boldsymbol{W}_m\boldsymbol{x}_{tm}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \sum_m \boldsymbol{W}_m\boldsymbol{x}_{tm}\right) - \sum_m \boldsymbol{x}_{1m}^{T}\tilde{\boldsymbol{\pi}}_m - \sum_{t=2}^{T}\sum_m \boldsymbol{x}_{1m}^{T}\widetilde{\boldsymbol{A}}_m\,\boldsymbol{x}_{t-1,m}^{T}$, we can write:

$$KL(q||p) = \sum_{m=1}^{M}\sum_{t=1}^{T}\overline{\boldsymbol{x}}_{tm}^{T}\tilde{\boldsymbol{\xi}}_{tm} + \frac{1}{2}\sum_{t=1}^{T}\Big[\boldsymbol{y}_t^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - 2\sum_{m=1}^{M}\boldsymbol{y}_t^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m\overline{\boldsymbol{x}}_{tm} + \sum_{m=1}^{M}\sum_{n\neq m}^{M} tr\left(\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_n\overline{\boldsymbol{x}}_{tn}\overline{\boldsymbol{x}}_{tm}^{T}\right) + \sum_{m=1}^{M} tr\left(\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m diag(\overline{\boldsymbol{x}}_{tm})\right)\Big] - lnZ_q + lnZ$$

$diag(.)$ is an operator that takes a vector and returns a square matrix with the elements of the vector along its diagonal.

Factorial HMM

$$KL(q||p) = \sum_{m=1}^{M}\sum_{t=1}^{T}\overline{\boldsymbol{x}}_{tm}^{T}\tilde{\boldsymbol{\xi}}_{tm} + \frac{1}{2}\sum_{t=1}^{T}\Big[\boldsymbol{y}_t^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t - 2\sum_{m=1}^{M}\boldsymbol{y}_t^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m\overline{\boldsymbol{x}}_{tm} +$$

$$\sum_{m=1}^{M}\sum_{n\neq m}^{M} tr\left(\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_n\overline{\boldsymbol{x}}_{tn}\overline{\boldsymbol{x}}_{tm}^T\right) + \sum_{m=1}^{M} tr\left(\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m diag(\overline{\boldsymbol{x}}_{tm})\right)\Big] - lnZ_q + lnZ$$

Using $E_q(\boldsymbol{x})$=$\sum_{t=1}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^T\tilde{\boldsymbol{\xi}}_{tm} - \sum_{m=1}^{M}\boldsymbol{x}_{1m}^T\tilde{\boldsymbol{\pi}}_m - \sum_{t=2}^{T}\sum_{m=1}^{M}\boldsymbol{x}_{tm}^T\tilde{\boldsymbol{A}}_m\boldsymbol{x}_{t-1,m}$, and since

$$\frac{\partial Z_q}{\partial\tilde{\boldsymbol{\xi}}_{\tau n}} = \int -e^{-E_q}\frac{\partial E_q}{\partial\tilde{\boldsymbol{\xi}}_{\tau n}}d\boldsymbol{x} = \int e^{-E_q}\boldsymbol{x}_{\tau n}d\boldsymbol{x} = Z_q\overline{\boldsymbol{x}}_{\tau n}, \quad \frac{\partial lnZ_q}{\partial\tilde{\boldsymbol{\xi}}_{\tau n}} = \overline{\boldsymbol{x}}_{\tau n}$$

we can write:

$$\frac{\partial KL}{\partial\tilde{\boldsymbol{\xi}}_{\tau n}} = \overline{\boldsymbol{x}}_{\tau n} + \left(\frac{\partial\overline{\boldsymbol{x}}_{tm}}{\partial\tilde{\boldsymbol{\xi}}_{\tau n}}\right)^T\sum_{t=1}^{T}\sum_{m=1}^{M}\left[\tilde{\boldsymbol{\xi}}_{tm} - \boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{y}_t + \sum_{\ell\neq m}^{M}\boldsymbol{W}_{\boldsymbol{m}}^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_\ell\overline{\boldsymbol{x}}_{t\ell} + \frac{1}{2}\delta_m\right] - \overline{\boldsymbol{x}}_{\tau n} = 0$$

Thus the term in parenthesis is zero and the update Eqs are as follows:

$$\xi_{tm} = exp\left(\boldsymbol{W}_m^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{y}}_{tm} - \frac{1}{2}\delta_m\right),\ \delta_m = diag\left(\boldsymbol{W}_m^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m\right),\ \tilde{\boldsymbol{y}}_{tm} = \boldsymbol{y}_t - \sum_{\ell\neq m}^{M}\boldsymbol{W}_\ell\,\overline{\boldsymbol{x}}_{t\ell}$$

δ_m is the vector of diagonal elements of $\boldsymbol{W}_m^T\boldsymbol{\Sigma}^{-1}\boldsymbol{W}_m$.

Factorial HMM

The $\boldsymbol{\xi}_{tm}$ parameter plays the role of the local evidence, averaging over the neighboring chains.

Having computed this for each chain, we can perform forwards-backwards in parallel, using these approximate local evidence terms to compute $q(\mathbf{x}_{t,m}|\mathbf{y}_{1:T})$ for each m and t.

The update cost is $\mathcal{O}(TMK^2)$ for a full “sweep” over all the variational parameters, since we have to run forwards-backwards M times, for each chain independently.

This is the same cost as a fully factorized approximation, but is much more accurate.

Variational Bayes (VB) for the Univariate Gaussian

Suppose our goal is to infer the posterior distribution for the mean μ and precision τ given data $\mathcal{D}= \{\mathrm{x}_1, \dots, \mathrm{x}_\mathrm{n}\}$ assumed to be drawn independently from the Gaussian.

The likelihood function is given by:

$$p(\mathcal{D}|\mu,\tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2\right\}$$

We introduce conjugate prior distributions for μ and τ given by

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

Since we chose conjugate priors, this can be solved analytically. However, we consider a factorized approximation of the form:

$$q(\mu,\tau) = q_\mu(\mu)q_\tau(\tau)$$

- MacKay, D. J. C. (2003). *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.
- MacKay, D. (1995a). Developments in probabilistic modeling with neural networks — ensemble learning. In *Proc. 3rd Ann. Symp. Neural Networks*.
- Attias, H. (2000). A variational Bayesian framework for graphical models. In *NIPS-12*.
- Beal, M. and Z. Ghahramani (2006). Variational Bayesian Learning of Directed Graphical Models with Hidden Variables. *Bayesian Analysis 1*(4).
- Smidl, V. and A. Quinn (2005). *The Variational Bayes Method in Signal Processing*. Springer.

Computing $q_\mu^*(\mu)$

Computing $\ln q_j^*(Z_j) = \mathbb{E}_{i\neq j}[\ln p(X,Z)] + const$, we get:

$$\ln q_\mu^*(\mu) = \mathbb{E}_\tau[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)] + const$$
$$= -\frac{\mathbb{E}[\tau]}{2}\{\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^N (x_n-\mu)^2\} + const$$

By completing the square over μ we see that $q_\mu(\mu) = \mathcal{N}(\mu|\mu_N,\lambda_N^{-1})$ with

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}$$ (no need updating)

$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau]$$

Note that as $N \to \infty$ the MLE estimate is recovered and the precision is infinite:

$$\mu_N = \frac{\lambda_0\ \mu_0 + N\bar{x}}{\lambda_0 + N} \xrightarrow[N\to\infty]{} \bar{x}$$
$$\lambda_N = (\lambda_0 + N)\mathbb{E}[\tau] \xrightarrow[N\to\infty]{} \infty$$

Computing $q_\tau^*(\tau)$

Similarly:

$$\ln q_\tau^*(\tau) = \mathbb{E}_\mu[\ln p(\mathcal{D}|\mu,\tau) + \ln p(\mu|\tau)] + \ln p(\tau) + const$$

$$= (a_0 - 1)\ln\tau - b_0\tau + \frac{1}{2}\ln\tau + \frac{N}{2}\ln\tau - \frac{\tau}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + const$$

It can then be shown that $q_\tau(\tau) = \text{Gam}\,(\tau|\, a_N, b_N)$ with

$$a_N = a_0 + \frac{N+1}{2}, \qquad b_N = b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right]$$

Using the expressions for the mean and variance of the Gamma distribution, we note:

$$\mathbb{E}[\tau] = \frac{a_N}{b_N} = \frac{2a_0 + N + 1}{2b_0 + \mathbb{E}\left[\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right]} \xrightarrow[N\to\infty]{} \frac{N}{\mathbb{E}\left[\sum_{n=1}^{N}(x_n-\mu)^2\right]}$$

$$\text{var}[\tau] = \frac{a_N}{b_N^2} = \frac{\mathbb{E}[\tau]}{b_0 + \frac{1}{2}\mathbb{E}\left[\lambda_0(\mu-\mu_0)^2 + \sum_{n=1}^{N}(x_n-\mu)^2\right]} \xrightarrow[N\to\infty]{} 0$$

Self-consistent iterative process

$$q_\mu(\mu) = \mathcal{N}(\mu|\mu_N, \lambda_N^{-1}), \qquad \mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}, \qquad \lambda_N = (\lambda_0 + N)\mathbb{E}[\tau]$$

$$q_\tau(\tau) = \mathrm{Gam}\,(\tau|\, a_N, b_N),$$

$$a_N = a_0 + \frac{N+1}{2}, \qquad b_N = b_0 + \frac{1}{2}\mathbb{E}_\mu\left[\sum_{n=1}^{N}(x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]$$

The eqs for the two optimal distributions $q_\mu(\mu), q_\tau\,(\tau)$ can be iterated until convergence.

For example after computing $q_\mu(\mu)$, we use the moments $\mathbb{E}_\mu[\mu]$, $\mathbb{E}_\mu[\mu^2]$ for updating $q_\tau(\tau)$, etc.

Using non-informative priors

Assume non-informative priors: $\mu_0 = a_0 = b_0 = \lambda_0 = 0$. The mean and the precision of the optimal $q_\mu(\mu)$ are simplified leading to:

$$\mathbb{E}[\mu] = \bar{x}, \qquad \mathbb{E}[\mu^2] = \bar{x}^2 + \frac{1}{N\mathbb{E}[\tau]}$$

Substituting these equations in the expression for $\mathbb{E}[\tau]$ gives

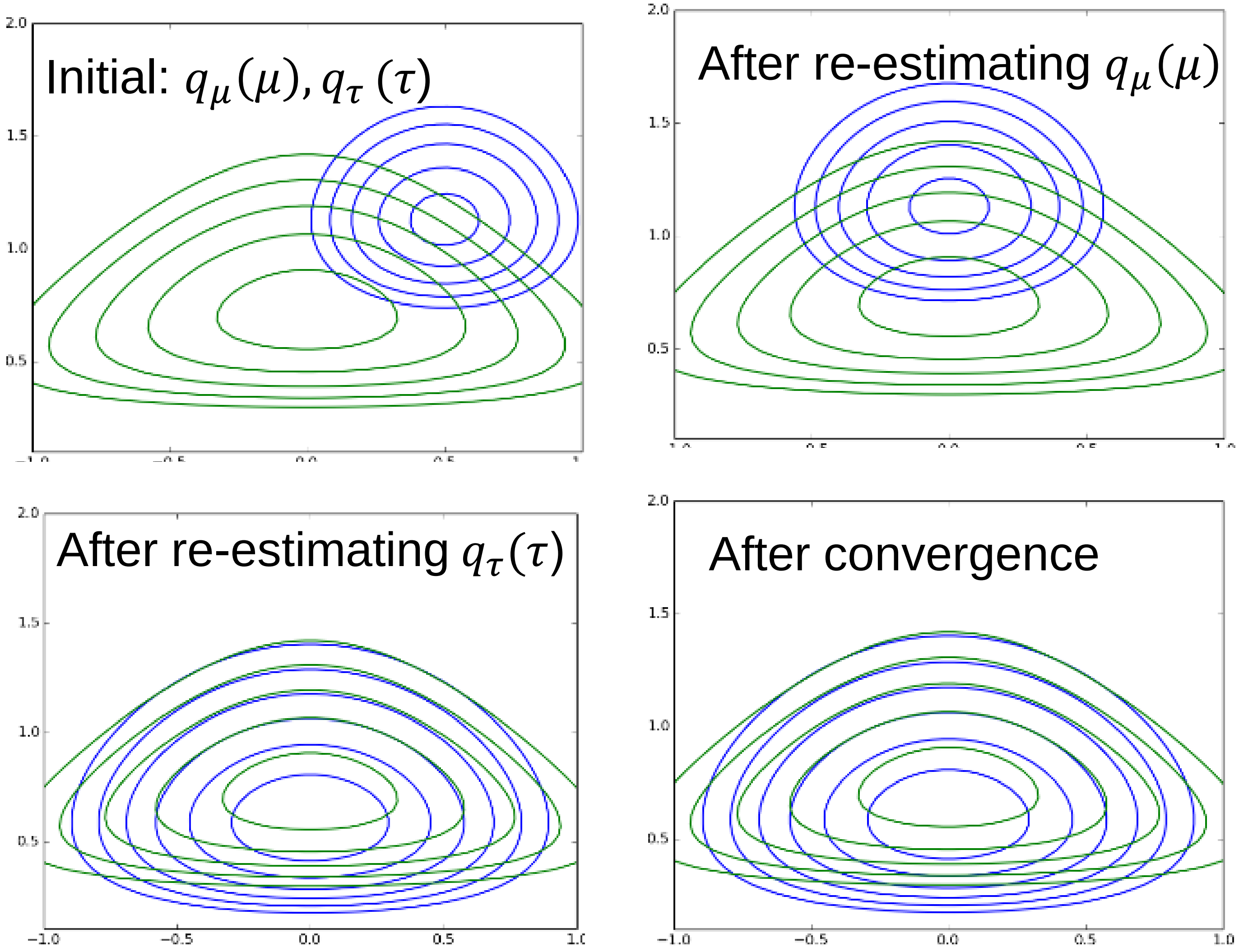
$$\frac{1}{\mathbb{E}[\tau]} = \frac{1}{N+1}\mathbb{E}\left[\sum_{n=1}^{N}(x_n-\mu)^2\right] = \frac{N}{N+1}\left(\overline{x^2} - 2\bar{x}^2 + \underbrace{\mathbb{E}\left[\mu^2\right]}_{\bar{x}^2+\frac{1}{N\mathbb{E}[\tau]}}\right) \Rightarrow \frac{1}{\mathbb{E}[\tau]} = \overline{x^2} - \bar{x}^2$$

$$\frac{1}{\mathbb{E}[\tau]} = \overline{x^2} - \bar{x}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \text{ (sample variance)}$$

For a Bayesian treatment for the Gaussian including a comparison with MLE see the reference below.

- Thomas P. Minka, Bayesian inference of a uniform distribution, 2001 (Msft Research)

The univariate Gaussian: Example



Initial: $q_\mu(\mu), q_\tau\,(\tau)$

After re-estimating $q_\mu(\mu)$

After re-estimating $q_\tau(\tau)$

After convergence

Python Code

The Univariate Gaussian: Lower Bound

It is a good practice to compute the lower bound and monitor its values during the iterations (it should always monotonically increase).

$$\mathcal{L}(q) = \iint q(\mu,\tau)ln\frac{p(\mathcal{D},\mu,\tau)}{q(\mu,\tau)}d\mu d\tau = \mathbb{E}[lnp(\mathcal{D}|\mu,\tau)] + \mathbb{E}[lnp(\mu|\tau)] + \mathbb{E}[lnp(\tau)] - \mathbb{E}[lnq(\mu)] - \mathbb{E}[lnq(\tau)]$$

We compute the various terms below using the posteriors we arrived at with the variational approach.

$-\mathbb{E}[lnq(\mu)] = -\frac{1}{2}ln\lambda_N + \frac{1}{2}\big(1+ln(2\pi)\big)$ (entropy of $\mathcal{N}(\mu|\mu_N,\lambda_N^{-1})$)

$-\mathbb{E}[lnq(\tau)] = ln\Gamma(a_N) - (a_N-1)\psi(a_N) - lnb_N + a_N$ (entropy of Gam $(\tau|\, a_N, b_N)$)

$\mathbb{E}_{q(\tau)}[lnp(\tau)] = \mathbb{E}_{q(\tau)}[a_0lnb_0 - ln\Gamma(a_0) + (a_0-1)ln\tau - b_0\tau] = a_0lnb_0 - ln\Gamma(a_0) +$
$(a_0-1)(\psi(a_N) - lnb_N) - b_0\frac{a_N}{b_N}$ (recall for $\tau\sim\mathcal{Gamma}$(a,b): $\mathbb{E}_{q(\tau)}[ln\tau] = \psi(a) - lnb, \mathbb{E}_{q(\tau)}[\tau] = \frac{a}{b}$).

$\mathbb{E}_{q(\mu,\tau)}[lnp(\mathcal{D}|\mu,\tau)] = \mathbb{E}_{q(\mu,\tau)}\left[-\frac{N}{2}ln(2\pi) + \frac{N}{2}ln\tau - \frac{1}{2}\tau\sum_{n=1}^{N}(x_n-\mu)^2\right] = -\frac{N}{2}ln(2\pi) + \frac{N}{2}\mathbb{E}[ln\tau] -$
$\frac{1}{2}\mathbb{E}[\tau]\mathbb{E}[\sum_{n=1}^{N}(x_n-\mu)^2] = -\frac{N}{2}ln(2\pi) + \frac{N}{2}(\psi(a_N) - lnb_N) - \frac{N}{2}\frac{a_N}{b_N}\left(\overline{x^2} - 2\bar{x}\mathbb{E}[\mu] + \mathbb{E}[\mu^2]\right) =$
$-\frac{N}{2}ln(2\pi) + \frac{N}{2}(\psi(a_N) - lnb_N) - \frac{N}{2}\frac{a_N}{b_N}\left(\overline{x^2} - 2\bar{x}\mu_N + \mu_N^2 + \frac{1}{\lambda_N}\right)$

$\mathbb{E}_{q(\mu,\tau)}[lnp(\mu|\tau)] = \mathbb{E}_{q(\mu,\tau)}\left[\frac{1}{2}ln\left(\frac{\lambda_0}{2\pi}\right) + \frac{1}{2}ln\tau - \frac{1}{2}\lambda_0\tau(\mu-\mu_0)^2\right] = \frac{1}{2}ln\left(\frac{\lambda_0}{2\pi}\right) + \frac{1}{2}(\psi(a_N) - lnb_N) -$
$\frac{\lambda_0}{2}\frac{a_N}{b_N}\left((\mu_N-\mu_0)^2 + \frac{1}{\lambda_N}\right)$

The Univariate Gaussian: Lower Bound

Finally:

$$\mathcal{L}(q) =$$

$$\begin{gathered}
-\frac{N}{2}ln(2\pi)+\frac{N}{2}(\psi(a_N)-lnb_N)-\frac{N}{2}\frac{a_N}{b_N}\left(\overline{x^2}-2\bar{x}\mu_N+\mu_N^2+\frac{1}{\lambda_N}\right)\\
+\frac{1}{2}ln\left(\frac{\lambda_0}{2\pi}\right)+\frac{1}{2}(\psi(a_N)-lnb_N)-\frac{\lambda_0}{2}\frac{a_N}{b_N}\left((\mu_N-\mu_0)^2+\frac{1}{\lambda_N}\right)\\
+a_0lnb_0-ln\Gamma(a_0)+(a_0-1)(\psi(a_N)-lnb_N)-b_0\frac{a_N}{b_N}\\
+ln\Gamma(a_N)-(a_N-1)\psi(a_N)-lnb_N+a_N\\
-\tfrac{1}{2}ln\lambda_N+\tfrac{1}{2}(1+ln(2\pi))
\end{gathered}$$

Light blue terms are obviously constants.

The terms in black when combined and using , $b_N=b_0+\frac{1}{2}\mathbb{E}_{\mu}[\sum_{n=1}^{N}(x_n-\mu)^2+\lambda_0(\mu-\mu_0)^2]=b_0+\frac{1}{2}\left(\overline{x^2}N-2\bar{x}N\mu_N+N\mu_N^2+\frac{N}{\lambda_N}+\lambda_0(\mu_N-\mu_0)^2+\frac{\lambda_0}{\lambda_N}\right)$ lead to a constant.

Red terms contribute to the final expression below using $a_N=a_0+\frac{N+1}{2}$. We conclude:

$$\mathcal{L}(q)=-\frac{1}{2}ln\lambda_N+ln\Gamma(a_N)-a_Nlnb_N+const$$

Variational Optimization and Model Selection

Consider comparing a set of candidate models, labelled by *m*, and having prior probabilities *p*(*m*).

Our goal is then to approximate the posterior probabilities *p*(*m*|**X**), where **X** is the observed data.

Different models may have different structure and different dimensionality for the hidden variables **Z**. We factorize the joint posterior as follows:

$$q(\mathbf{Z},m) = q(\mathbf{Z}|m)q(m)$$

Starting with the definition of the lower bound (see below) $\mathcal{L}$, we can show:

$$\mathcal{L} = \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m)q(m)\ln\frac{p(\mathbf{Z},\mathbf{X},m)}{q(\mathbf{Z}\,|\,m)q(m)} = \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m)q(m)\ln\frac{p(\mathbf{Z},m\,|\,\mathbf{X})p(\mathbf{X})}{q(\mathbf{Z}\,|\,m)q(m)}$$

$$= \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m)q(m)\ln\frac{p(\mathbf{Z},m\,|\,\mathbf{X})}{q(\mathbf{Z}\,|\,m)q(m)} + \ln p(\mathbf{X})$$

$$\ln p(\mathbf{X}) = \mathcal{L} - \sum_m \sum_{\mathbf{Z}} q(\mathbf{Z}\,|\,m)q(m)\ln\frac{p(\mathbf{Z},m\,|\,\mathbf{X})}{q(\mathbf{Z}\,|\,m)q(m)}$$

Variational Optimization and Model Selection

We first optimize each p(**Z**|m) by optimization of the lower bound of $\mathcal{L}_m = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|m) \ln \frac{p(\boldsymbol{Z},\boldsymbol{X}|m)}{q(\boldsymbol{Z}|m)}$ for each model m. Then note that the overall lower bound $\mathcal{L}$ is:

$$\mathcal{L} = \sum_m \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}\,|\,m) q(m) \ln \frac{p(\boldsymbol{Z},\boldsymbol{X},m)}{q(\boldsymbol{Z}\,|\,m)q(m)} = \sum_m \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}\,|\,m) q(m) \left(\ln p(\boldsymbol{Z},\boldsymbol{X}\,|\,m) + \ln p(m) - \ln q(\boldsymbol{Z}\,|\,m) - \ln q(m) \right)$$

$$= \sum_m q(m) \left(\ln p(m) - \ln q(m) + \underbrace{\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}\,|\,m) \left(\ln p(\boldsymbol{Z},\boldsymbol{X}\,|\,m) - \ln q(\boldsymbol{Z}\,|\,m) \right)}_{\mathcal{L}_m} \right) =$$

$$= \sum_m q(m) \left(\ln p(m) - \ln q(m) + \ln e^{\mathcal{L}_m} \right) = \sum_m q(m) \left(\ln \frac{p(m) e^{\mathcal{L}_m}}{q(m)} \right)$$

This is as the distance $-KL\left(q(m), p(m)e^{\mathcal{L}_m}\right)$ which is maximized when

$$q(m) \propto p(m) \exp\left(\mathcal{L}_m\right)$$

$$where: \mathcal{L}_m = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}\,|\,m) \ln \frac{p(\boldsymbol{Z},\boldsymbol{X}\,|\,m)}{q(\boldsymbol{Z}\,|\,m)}$$ (lower bound for model m)

After normalization, we can use q(m) for model selection or model averaging.