Variational Algorithms: Ising Model, Univariate Gaussian, Linear Regression, Model Selection

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 29, 2018

Contents

- Mean Field for the Ising Model, Structured Mean Field
- Variational Inference for the univariate Gaussian, Variational optimization and model selection.
- Variational Linear Regression, Predictive Distribution, Lower Bound, Selection of the order of the polynomial
- Variational Linear Regression with Gam($\beta|c_0, d_0$)

Following:

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

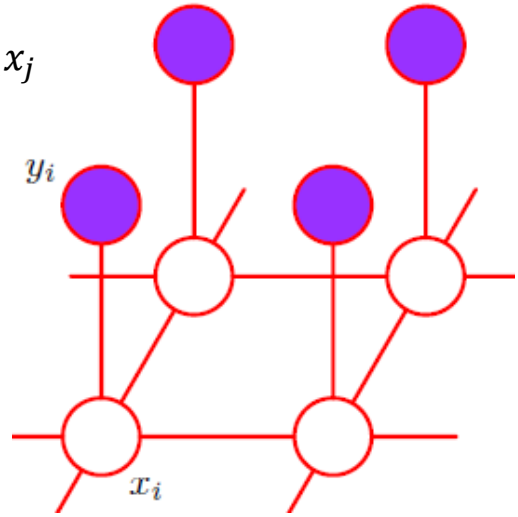
Mean Field for the Ising Model

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:

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

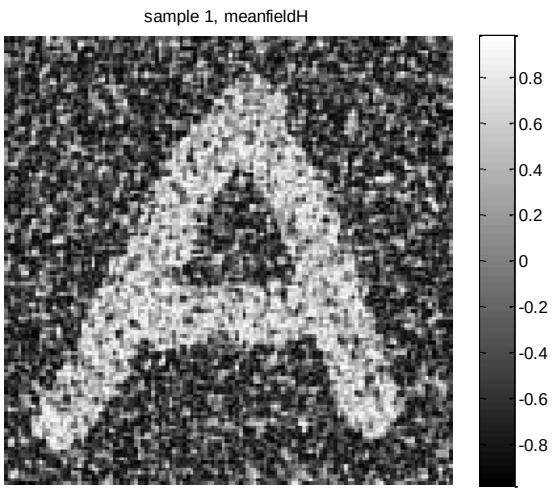
To avoid checkerboarding effects, damped updates are often used as follows:

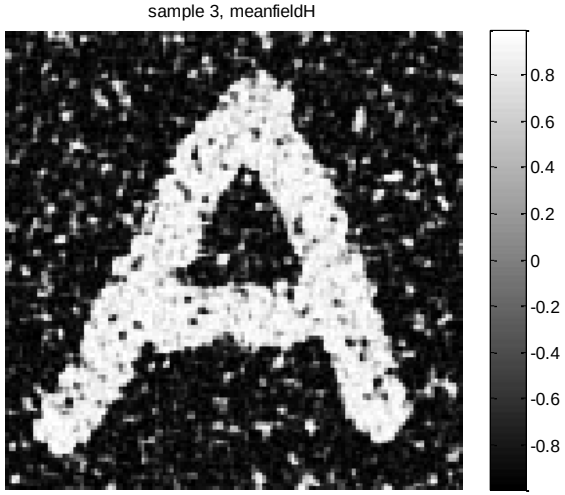
$$\mu_i^t = (1-\lambda)\mu_i^{t-1} + \lambda \tanh\left(\textstyle\sum_{j \in nbr_i} W_{ij}\mu_i^{t-1} + 0.5(L_i^+ - L_i^-)\right),\ 0 < \lambda < 1.$$

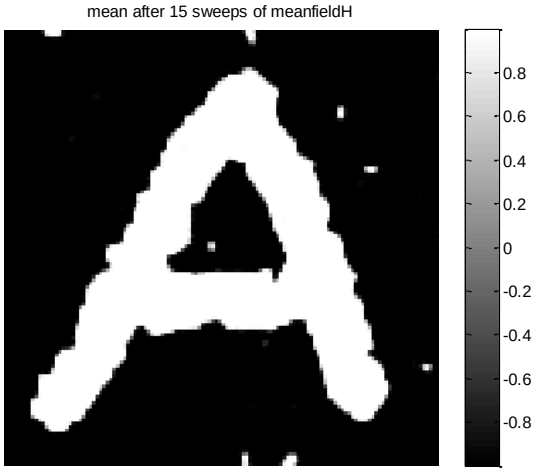
The nodes can be updated in parallel or asynchronously.

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.

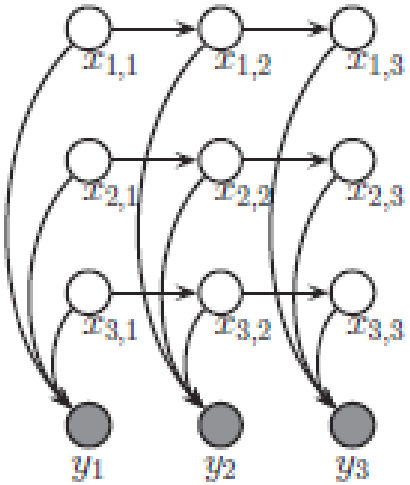
If we can perform efficient inference in each q_i, the method is tractable overall.

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



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

where $p(x_{tm} = k|x_{t-1,m} = j) = A_{mjk}$ 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 $D \times K$ 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:

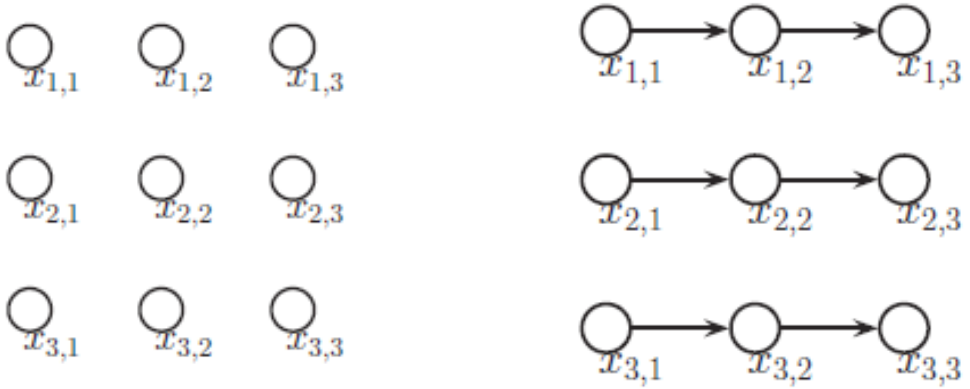
$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{1}{Z}\exp(-E(\boldsymbol{x},\boldsymbol{y}))$$

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

where pointwise $\tilde{\boldsymbol{A}}_m = ln\boldsymbol{A}_m$, and $\tilde{\boldsymbol{\pi}}_m = ln\boldsymbol{\pi}_m$. $\boldsymbol{x}_{tm}$ 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:

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

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

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

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

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

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

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

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:

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

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

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

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

Factorial HMM

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

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

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

we can write:

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

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

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

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

Factorial HMM

The ξ_{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 $\mathcal{D} = \{x_1, \ldots, x_n\}$ 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

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

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

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

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

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

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

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

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

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

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

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

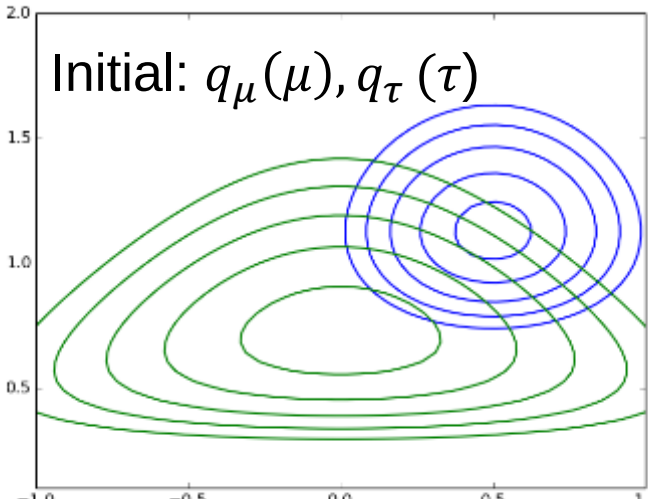
$$\frac{1}{\mathbb{E}[\tau]} = \frac{1}{N+1}\mathbb{E}\left[\sum_{n=1}^{N}(x_n - \mu)^2\right] = \frac{N}{N+1}\left(\overline{x^2} - 2\bar{x}^2 + \underbrace{\mathbb{E}[\mu^2]}_{\bar{x}^2 + \frac{1}{N\mathbb{E}[\tau]}}\right) \Rightarrow \frac{1}{\mathbb{E}[\tau]} = \overline{x^2} - \bar{x}^2$$

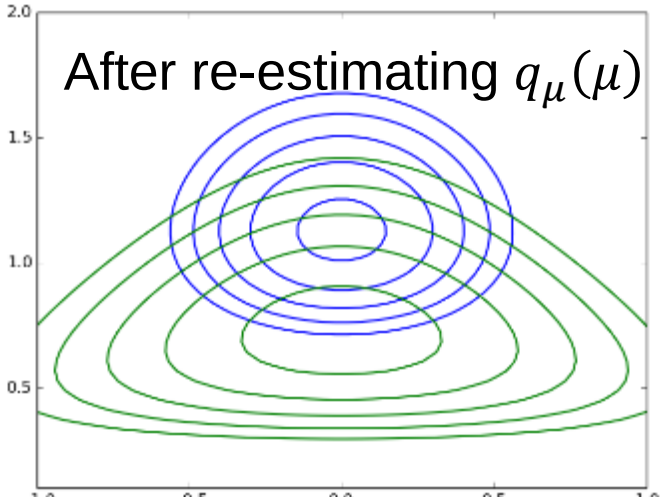
$$\frac{1}{\mathbb{E}[\tau]} = \overline{x^2} - \bar{x}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 \text{ (sample variance)}$$

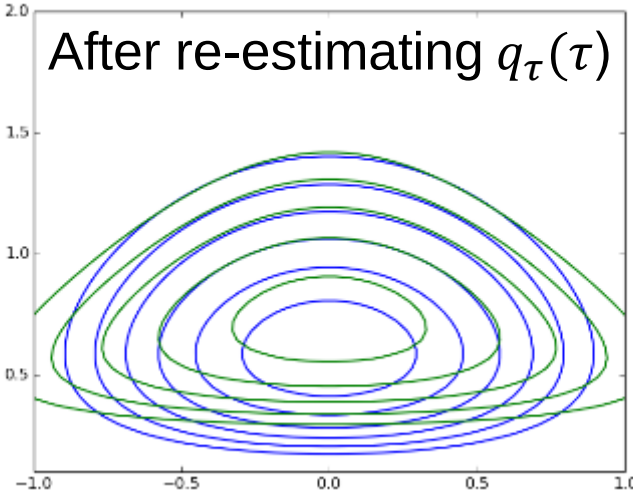
For a Bayesian treatment for the Gaussian including a comparison with MLE see the reference below.

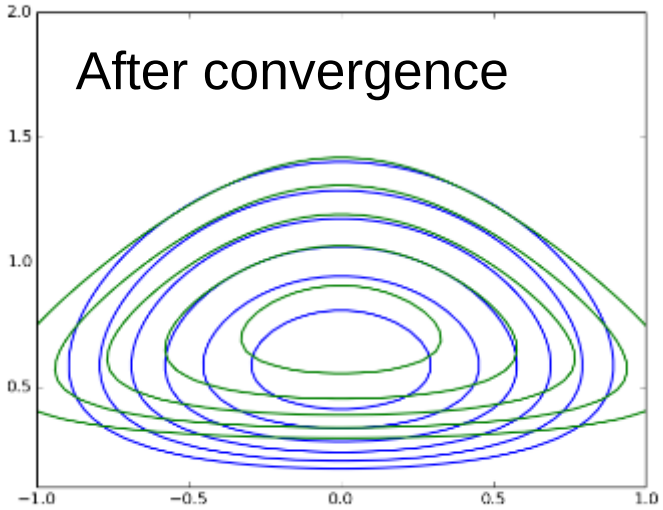
- Thomas P. Minka, Bayesian inference of a uniform distribution, 2001 (Msft Research)

The univariate Gaussian: Example









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.

$-\mathbb{E}[lnq(\mu)] = -\frac{1}{2} ln\lambda_N + \frac{1}{2}\left(1 + ln(2\pi)\right)$ (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 $\text{Gam}(\tau|\, a_N, b_N)$)

$\mathbb{E}_{q(\tau)}[lnp(\tau)] = \mathbb{E}_{q(\tau)}[a_0 lnb_0 - ln\Gamma(a_0) + (a_0 - 1)ln\tau - b_0\tau] = a_0 lnb_0 - ln\Gamma(a_0) + (a_0 - 1)(\psi(a_N) - lnb_N) - b_0 \frac{a_N}{b_N}$ (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:

$$\mathcal{L}(q) =$$
$$-\frac{N}{2}ln(2\pi) + \frac{N}{2}(\psi(a_N) - lnb_N) - \frac{N}{2}\frac{a_N}{b_N}\left(\overline{x^2} - 2\bar{x}\mu_N + \mu_N^2 + \frac{1}{\lambda_N}\right)$$
$$+ \frac{1}{2}ln\left(\frac{\lambda_0}{2\pi}\right) + \frac{1}{2}(\psi(a_N) - lnb_N) - \frac{\lambda_0}{2}\frac{a_N}{b_N}\left((\mu_N - \mu_0)^2 + \frac{1}{\lambda_N}\right)$$
$$+ a_0 lnb_0 - ln\Gamma(a_0) + (a_0 - 1)(\psi(a_N) - lnb_N) - b_0\frac{a_N}{b_N}$$
$$+ln\Gamma(a_N) - (a_N - 1)\psi(a_N) - lnb_N + a_N$$
$$-\frac{1}{2}ln\lambda_N + \frac{1}{2}(1 + ln(2\pi))$$

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:

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

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 $p(m|\mathbf{X})$, where $\mathbf{X}$ is the observed data.

Different models may have different structure and different dimensionality for the hidden variables $\mathbf{Z}$. 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:

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

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

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

Variational Optimization and Model Selection

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

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

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

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

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

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

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

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

Variational Linear Regression

Consider the Bayesian linear regression model.

Recall that the likelihood function is given by

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

where $\boldsymbol{\phi}_n = \boldsymbol{\phi}(\boldsymbol{x}_n)$.
We use the following conjugate prior distributions (i.e. Gaussian-Gamma):

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}\boldsymbol{I})$$
$$p(\alpha) = \text{Gam}(\alpha|a_0, b_0)$$

To simplify discussion the noise precision parameter β is fixed and assumed to be known. The framework can be extended with this assumption being relaxed.

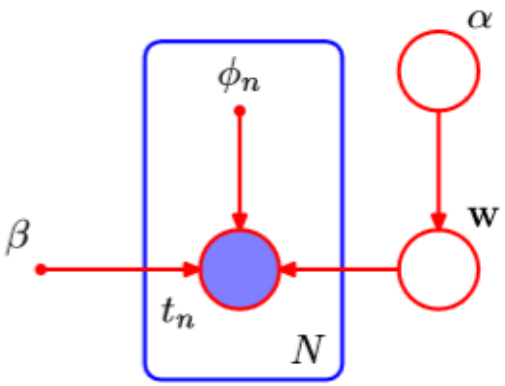
- Drugowitsch, J. (2008). Bayesian linear regression. Technical report, U. Rochester.

Variational Linear Regression

Collectively we have the joint distribution

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

which can be represented by the following graphical model



We assume a variational posterior of the following form to obtain the unknown hyperparameters

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

Variational Linear Regression

Using the same framework as previously discussed, we obtain for $q(\alpha)$:

$$\ln q^*(\alpha) = \ln p(\alpha) + \mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{w}|\alpha)] + const$$
$$= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}] + const$$

As expected this gives a Gamma distribution

$$q^*(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N)$$

with

$$a_N = a_0 + \frac{M}{2}$$
$$b_N = b_0 + \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

Variational Linear Regression

Similarly we obtain for $q(\boldsymbol{w})$:

$$\ln q^*(\boldsymbol{w}) = \ln p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_\alpha[\ln p(\boldsymbol{w}|\alpha)] + const$$

$$= -\frac{\beta}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\boldsymbol{w}^T\boldsymbol{w} + const$$

$$= -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{w} + \beta\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t} + const$$

As expected this gives a normal distribution

$$q^*(\boldsymbol{w}) = \mathcal{N}(w|\boldsymbol{m}_N, \boldsymbol{S}_N)$$

with

$$\boldsymbol{m}_N = \beta\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}$$

$$\boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

Recall that $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = (\boldsymbol{\phi}_1 \quad .. \quad \boldsymbol{\phi}_N)\begin{pmatrix}\boldsymbol{\phi}_1^T \\ .. \\ \boldsymbol{\phi}_N^T\end{pmatrix} = \sum_{n=1}^{N}\boldsymbol{\phi}_n\boldsymbol{\phi}_n^T$

These are the same formulas obtained when α is treated as constant. In the VI approach, α is replaced by $\mathbb{E}[\alpha]$.

Variational Linear Regression

By standard properties for Gaussian and Gamma distributions we finally have

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$$
$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N$$

which collectively give

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N} = \frac{a_0 + M/2}{b_0 + \mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]/2}$$

For the case of a_0=b_0=0 (infinitely broad prior over a): $\mathbb{E}[\alpha] = \frac{M}{\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{tr}(\boldsymbol{S}_N)}$. This is similar to the result obtained using the model evidence approximation.

It then remains a task to cycle between computing a_N, b_N, $\boldsymbol{m}_N$ and $\boldsymbol{S}_N$ until some convergence criterion is met.

These results are consistent with those obtained by maximizing the evidence using EM (except that the point estimate of α is now replaced by $\mathbb{E}[\alpha]$) and give identical results in the case of an infinitely broad prior for which $\mathbb{E}[\alpha] = \frac{M}{\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{tr}(\boldsymbol{S}_N)}$.

Predictive Distribution

The predictive distribution of t for a new input x is given as:

$$p(t|\boldsymbol{x},\boldsymbol{t}) = \int p(t|\boldsymbol{x},\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{t})d\boldsymbol{w} \approx \int p(t|\boldsymbol{x},\boldsymbol{w})q(\boldsymbol{w})d\boldsymbol{w}$$

$$= \int \mathcal{N}(t|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}),\beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{m}_N,\boldsymbol{S}_N)\, d\boldsymbol{w} = \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}),\sigma^2(\boldsymbol{x}))$$

Here we used an earlier result for linear Gaussian models. Given a marginal Gaussian distribution for **x** and a conditional Gaussian distribution for **y** given **x** in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Lambda}^{-1})$$
$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b},\mathbf{L}^{-1})$$

the marginal distribution of **y** is given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b},\mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^{\mathsf{T}})$$

Using the above result, the input dependent variance is given as:

$$\sigma^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\,\boldsymbol{\phi}(\boldsymbol{x}), \text{ with } \boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

This is the same form as that obtained when α is treated as constant but now α is replaced by $\mathbb{E}[\alpha]$.

Lower Bound

But how can we select M (degree of polynomial)? We can compute the lower bound

$$\mathcal{L}(q) = \mathbb{E}[\ln p(\boldsymbol{w}, \alpha, \boldsymbol{t})] - \mathbb{E}[\ln q(\boldsymbol{w}, \alpha)]$$

$$= \mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{t}|\boldsymbol{w})] + \mathbb{E}_{\boldsymbol{w},\alpha}[\ln p(\boldsymbol{w}|\alpha)] + \mathbb{E}_{\alpha}[\ln p(\alpha)] - \mathbb{E}_{\boldsymbol{w}}[\ln q(\boldsymbol{w})] - \mathbb{E}_{\alpha}[\ln q(\alpha)]$$

where

$$\mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{t}|\boldsymbol{w})] = -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) - \frac{\beta}{2}\mathbb{E}\left[\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2\right]$$

$$= -\frac{N}{2}\ln(2\pi) + \frac{N}{2}\ln(\beta) - \frac{\beta}{2}\{\boldsymbol{t}^T\boldsymbol{t} - 2\mathbb{E}[\boldsymbol{w}^T]\boldsymbol{\Phi}^T\boldsymbol{t} + Tr(\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T]\boldsymbol{\Phi}^T\boldsymbol{\Phi})\}$$

$$= \frac{N}{2}\ln\frac{\beta}{2\pi} - \frac{\beta}{2}\boldsymbol{t}^T\boldsymbol{t} + \beta\boldsymbol{m}_N^T\boldsymbol{\Phi}^T\boldsymbol{t} - \frac{\beta}{2}\mathrm{Tr}\,[\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N)]$$

$$\mathbb{E}_{\boldsymbol{w},\alpha}[\ln p(\boldsymbol{w}|\alpha)] = -\frac{M}{2}\ln 2\pi + \frac{M}{2}\mathbb{E}[\ln\alpha] - \frac{\mathbb{E}[\alpha]}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

$$= -\frac{M}{2}\ln 2\pi + \frac{M}{2}(\psi(a_N) - \ln b_N) - \frac{a_N}{2b_N}[\boldsymbol{m}_N^T\boldsymbol{m}_N + \mathrm{Tr}\,(\boldsymbol{S}_N)]$$

Use expectation of the log of a Gamma distributed variable

$$\mathbb{E}_{\alpha}[\ln p(\alpha)] = a_0\ln b_0 + (a_0 - 1)\mathbb{E}[\ln\alpha] - b_0\,\mathbb{E}[\alpha] - \ln\Gamma(a_0)$$

$$= a_0\ln b_0 + (a_0 - 1)[\psi(a_N) - \ln b_N] - \frac{b_0 a_N}{b_N} - \ln\Gamma(a_0)$$

Lower Bound

The final two terms in $\mathcal{L}(q)$ represent the entropies of the Gaussian and Gamma distributions:

$$-\mathbb{E}_{\boldsymbol{w}}[\ln q(\boldsymbol{w})] \quad = \frac{1}{2}\ln|\boldsymbol{S}_N| + \frac{M}{2}[1 + \ln 2\pi]$$

$$-\mathbb{E}_{\alpha}[\ln q(\alpha)] = \ln \Gamma(a_N) - (a_N - 1)\psi(a_N) - \ln b_N + a_N$$

We substitute in the Eqs. above the following expressions for the moments:

$$\mathbb{E}[\boldsymbol{w}] = \boldsymbol{m}_N$$

$$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N$$

$$\mathbb{E}[\alpha] = \frac{a_N}{b_N}$$

$$\mathbb{E}[ln\alpha] = \psi(a_N) - \ln b_N$$

to obtain the final expression for $\mathcal{L}(q)$:

$$\mathcal{L}(q) = \mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{t}|\boldsymbol{w})] + \mathbb{E}_{\boldsymbol{w},\alpha}[\ln p(\boldsymbol{w}|\alpha)] + \mathbb{E}_{\alpha}[\ln p(\alpha)] \; - \mathbb{E}_{\boldsymbol{w}}[\ln q(\boldsymbol{w})] - \mathbb{E}[\ln q(\alpha)]$$

Lower Bound Vs the Order of the Polynomial

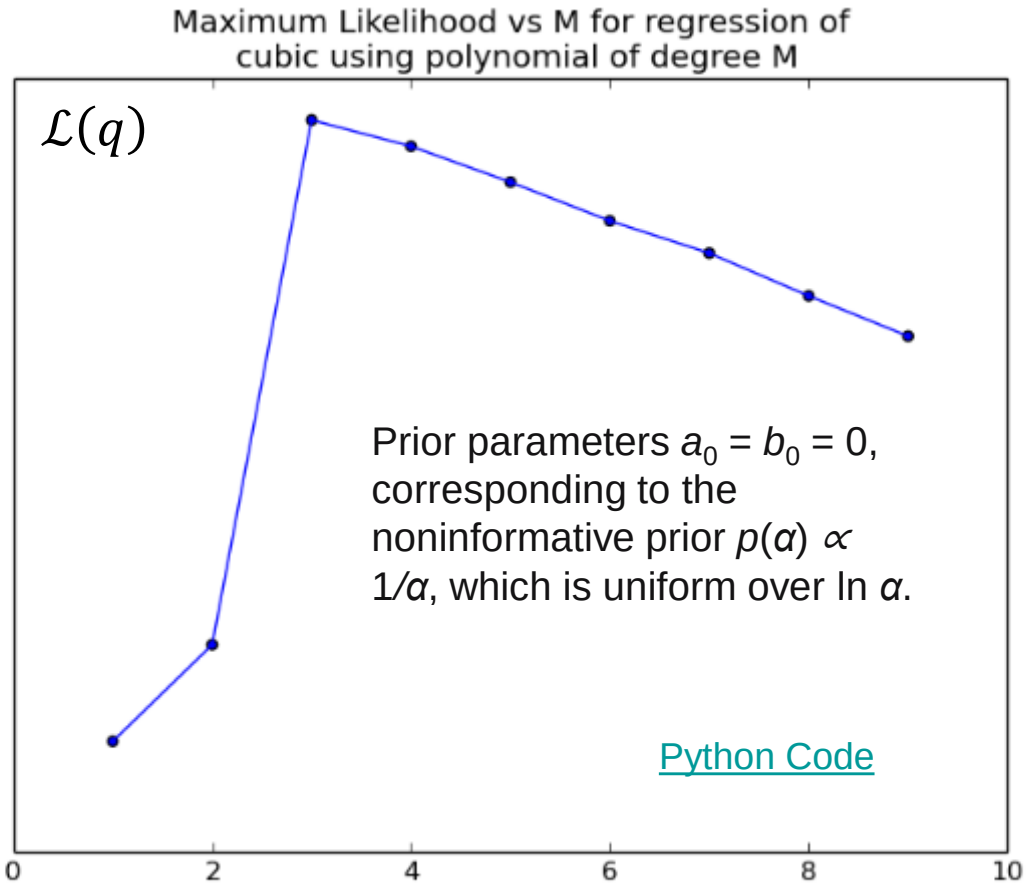
10 data points have been generated from a polynomial of degree 3 over [-5, 5] with additive Gaussian noise. The lower bound is maximized for M=3 corresponding to the true model form which the data was generated.

$\mathcal{L}$ represents lower bound on the log marginal likelihood $\ln p(\mathbf{t}|M)$ for the model.

If we assign equal prior probabilities $p(M)$ to the different values of M, then we can interpret $\mathcal{L}$ as an approximation to $p(M|\mathbf{t})$.

Thus the variational framework assigns the highest probability to the model with $M = 3$.

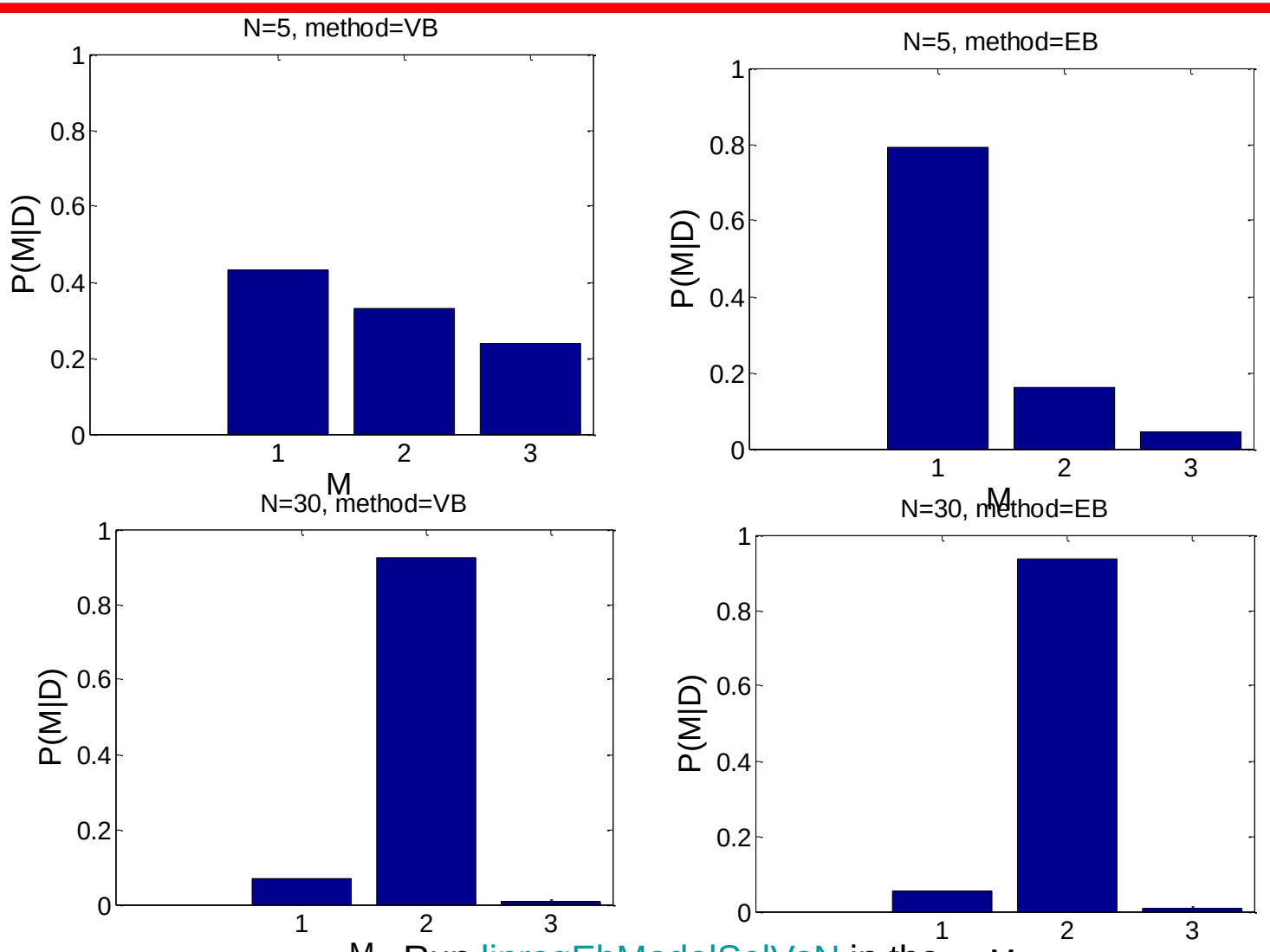
This should be contrasted with the MLE result, which assigns ever smaller residual error to models of increasing complexity until the residual error is driven to zero, causing MLE to over-fitted models.



Prior parameters $a_0 = b_0 = 0$, corresponding to the noninformative prior $p(\alpha) \propto 1/\alpha$, which is uniform over $\ln \alpha$.

Python Code

Lower Bound Vs the Order of the Polynomial



Run linregEbModelSelVsN in the PMTK3 toolbox

Variational Linear Regression with Gam($\beta|c_0, d_0$)

We now extend the variational treatment of Bayesian linear regression to include a gamma hyperprior Gam($\beta|c_0, d_0$) over β and solve variationally, by assuming a factorized variational distribution of the form $q(\mathbf{w})q(\alpha)q(\beta)$.

We modify the joint distribution of all variables as:

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

The formulae for $p(\alpha)$ remain the same:

$$q^*(\alpha) = \mathrm{Gam}(\alpha|a_N, b_N)\,, a_N = a_0 + \frac{M}{2},\ b_N = b_0 + \frac{1}{2}\mathbb{E}[\boldsymbol{w}^T\boldsymbol{w}]$$

For $q^*(\boldsymbol{w})$ we have:

$$\ln q^*(\boldsymbol{w}) = \ln p(\boldsymbol{t}|\boldsymbol{w},\beta) + \mathbb{E}_\alpha[\ln p(\boldsymbol{w}|\alpha)] + const$$
$$= -\frac{\mathbb{E}[\beta]}{2}\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2 - \frac{1}{2}\mathbb{E}[\alpha]\boldsymbol{w}^T\boldsymbol{w} + const$$
$$= -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}[\alpha]\boldsymbol{I} + \mathbb{E}[\beta]\boldsymbol{\Phi}^T\boldsymbol{\Phi})\boldsymbol{w} + \mathbb{E}[\beta]\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{t} + const$$

Thus $q^*(\boldsymbol{w}) = \mathcal{N}(w|\boldsymbol{m}_N, \boldsymbol{S}_N)$ with

$$\boldsymbol{m}_N = \mathbb{E}[\beta]\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}$$
$$\boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \mathbb{E}[\beta]\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$$

Variational Linear Regression with $Gam(\beta|c_0, d_0)$

For $q^*(\beta)$

$$\ln q^*(\beta) = \mathbb{E}[\ln p(\boldsymbol{t}|\boldsymbol{w}, \beta)] + \ln p(\beta) + const$$

$$= \frac{N}{2} ln\beta - \frac{\beta}{2}\mathbb{E}_{\boldsymbol{w}}\left[\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2\right] + (c_0 - 1)ln\beta - d_0\beta + const$$

We recognize the log of a Gamma distribution with:

$$q^*(\beta) = \mathrm{Gam}(\beta|c_N, d_N), c_N = c_0 + \frac{N}{2},$$

$$d_N = d_0 + \frac{1}{2}\mathbb{E}[\textstyle\sum_{n=1}^{N}\{\boldsymbol{w}^T\boldsymbol{\phi}_n - t_n\}^2] = d_0 + \frac{1}{2}(Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi})\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] + \boldsymbol{t}^T\boldsymbol{t}) - \boldsymbol{t}^T\,\boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}]$$

$$= d_0 + \frac{1}{2}\left(\|\boldsymbol{t} - \boldsymbol{\Phi}\boldsymbol{m}_N\|^2 + Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_N)\right)$$

Where we used:

$\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N$ and

$\mathbb{E}[\boldsymbol{w}]=\boldsymbol{m}_N$, with $\boldsymbol{m}_N = \mathbb{E}[\beta]\boldsymbol{S}_N\boldsymbol{\Phi}^T\boldsymbol{t}, \boldsymbol{S}_N = (\mathbb{E}[\alpha]\boldsymbol{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$

Thus $\mathbb{E}[\beta] = \frac{c_N}{d_N}$

Variational Linear Regression with $Gam(\beta|c_0, d_0)$

The lower bound also needs to be modified. Starting with the modified log-likelihood and using $\mathbb{E}[\boldsymbol{w}]=\boldsymbol{m}_N$, $\mathbb{E}[\beta] = \frac{c_N}{d_N}, \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^T] = \boldsymbol{m}_N\boldsymbol{m}_N^T + \boldsymbol{S}_N$ and $\mathbb{E}[ln\beta] = \psi(c_N) - lnd_N$:

$$\mathbb{E}_\beta\left[\mathbb{E}_{\boldsymbol{w}}[\ln p(\boldsymbol{t}|\boldsymbol{w})]\right] = \frac{N}{2}(\mathbb{E}[\beta] - \ln(2\pi)) - \frac{\mathbb{E}[\beta]}{2} - \mathbb{E}[\|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2]$$
$$= \frac{N}{2}(\psi(c_N) - lnd_N - \ln(2\pi)) - \frac{c_N}{2d_N}\left(\|\boldsymbol{t} - \boldsymbol{\Phi w}\|^2 + Tr(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{S}_N)\right)$$

Next using $\mathbb{E}[\beta] = \frac{c_N}{d_N}$ and $\mathbb{E}[ln\beta] = \psi(c_N) - lnd_N$, we consider the term corresponding to log prior over β:

$$\mathbb{E}[lnp(\beta)] = (c_0 - 1)\mathbb{E}[ln\beta] - d_0\mathbb{E}[\beta] + c_0\ ln\ d_0 - ln\Gamma(c_0)$$
$$= (c_0 - 1)(\psi(c_N) - lnd_N) - \frac{d_0 c_N}{d_N} + c_0\ ln\ d_0 - ln\Gamma(c_0)$$

Finally we compute the negative entropy of the posterior over β:

$$-\mathbb{E}[lnq^*(\beta)] = (c_N - 1)\psi(c_N) + lnd_N - c_N - ln\Gamma(c_N)$$

Finally the predictive distribution is given as:

$$p(t|\boldsymbol{x}, \boldsymbol{t}) \approx \mathcal{N}(t|\boldsymbol{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}), \sigma^2(\boldsymbol{x})),\ \sigma^2(\boldsymbol{x}) = \frac{1}{\mathbb{E}[\beta]} + \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{S}_N\ \boldsymbol{\phi}(\boldsymbol{x})$$