Exercises 6

Variational inference & Expectation propagation

Exercise 6.1 (adapted from [2])
Consider the Kullback-Leibler divergence

$$\mathrm{KL}(p \parallel q) = -\int p(x) \ln \frac{q(x)}{p(x)} dx$$

Evaluate $\mathrm{KL}(p(x) \parallel q(x))$ where $p(x)$ and $q(x)$ are ....

a) ... two scalar Gaussians

$$p(x) = \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) \quad \text{and} \quad q(x) = \mathcal{N}\left(x;\ m,\ s^2\right)$$

b) ... two multivariate Gaussians

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma}\right) \quad \text{and} \quad q(\mathbf{x}) = \mathcal{N}\left(\mathbf{x};\ \mathbf{m},\ \mathbf{S}\right).$$

Exercise 6.2
Consider a distribution $p(x)$ which we want to approximate with a scalar Gaussian

$$q_{\mu,\sigma}(x) = \mathcal{N}\left(x;\ \mu,\ \sigma^2\right)$$

by minimizing the Kullback-Leibler divergence

$$\hat{\mu}, \hat{\sigma} = \arg\min_{\mu,\sigma} \mathrm{KL}(p(x) \parallel q_{\mu,\sigma}(x))$$

Show that this results in moment matching

$$\begin{aligned} \hat{\mu} &= \mathbb{E}_p[x] \\ \hat{\sigma}^2 &= \mathbb{E}_p[(x-\mu)^2] \end{aligned}$$

where $\mathbb{E}_p[f(x)] = \int f(x)p(x)dx$ i.e. that $\hat{\mu}$ and $\hat{\sigma}^2$ will be the mean and variance of $p(x)$.

Exercise 6.3
Consider two random variables x and y which are related as

$$t = x + v, \qquad v \sim \mathcal{N}(0,\ 1)$$
$$y = \begin{cases} 1 & \text{if} \quad t > 0, \\ 0 & \text{otherwise} \end{cases}$$

We also have prior information that $x \sim \mathcal{N}(0,\ 1)$ and we receive one measurement $y = 1$.

a) Draw a factor graph of the model

b) Use message passing with moment matching with pen and paper to compute $p(x|y = 1)$.
Hint: The mean and variance for the half-normal distribution

$$\mathcal{HN}(x; \sigma^2) = 2\mathcal{N}\left(x;\ 0,\ \sigma^2\right), \quad x > 0$$

are

$$\mathbb{E}[x] = \frac{\sigma\sqrt{2}}{\sqrt{\pi}}, \qquad \mathrm{Var}[x] = \sigma^2\left(1 - \frac{2}{\pi}\right)$$

c) Write a program to verify the calculated distribution using importance sampling.

Exercise 6.4 **a)** Consider the variational linear regression example from the lecture,

$$p(\mathbf{y} \mid w, \beta) = \mathcal{N}\left(\mathbf{y};\ \mathbf{x}w,\ \beta^{-1}\mathbf{I}_N\right)$$

where we in this problem also treat β as a random variable and use the following prior[1]

$$p(w, \alpha, \beta) = p(w|\alpha)p(\alpha)p(\beta) = \mathcal{N}\left(w;\ 0,\ \alpha^{-1}\right)\mathrm{Gam}\left(\alpha; a_0,\ b_0\right)\mathrm{Gam}\left(\beta; c_0,\ d_0\right)$$

where

$$\mathrm{Gam}\left(\alpha; a,\ b\right) = \frac{1}{\Gamma(a)}b^a\alpha^{a-1}e^{-b\alpha}, \qquad \alpha \in [0, \infty)$$

Assume the factorized variational distribution $q(w, \alpha, \beta) = q(w)q(\alpha)q(\beta)$.

Use variational inference to derive the equations for updating the variational distribution $q(w, \alpha, \beta)$ approximating the posterior $p(w, \alpha, \beta|\mathbf{y})$.

b) Consider the case where $\mathbf{w}$ is a vector, i.e. the standard linear regression setting where we consider the likelihood

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}\left(\mathbf{y};\ \mathbf{X}\mathbf{w},\ \beta^{-1}\mathbf{I}_N\right)$$

Consider a diagonal prior on $\mathbf{w}$ and as in **(a)** Gamma prior on α and β

$$p(\mathbf{w}, \alpha, \beta) = p(\mathbf{w}|\alpha)p(\alpha)p(\beta) = \mathcal{N}\left(\mathbf{w};\ \mathbf{0},\ \alpha^{-1}\mathbf{I}\right)\mathrm{Gam}\left(\alpha; a_0,\ b_0\right)\mathrm{Gam}\left(\beta; c_0,\ d_0\right)$$

Consider the same factorization of the variational distribution $q(\mathbf{w}, \alpha, \beta) = q(\mathbf{w})q(\alpha)q(\beta)$.
Hint: For a multivariate Gaussian distribution $\mathbf{x} \sim \mathcal{N}(\mathbf{x};\ \boldsymbol{\mu},\ \boldsymbol{\Sigma})$ *we have*

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\mathsf{T}] = \boldsymbol{\mu}\boldsymbol{\mu}^\mathsf{T} + \boldsymbol{\Sigma}$$

You may also want to use that

$$\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$$

where $\mathbf{A}$ *and* $\mathbf{B}$ *are two matrices and where* $\mathrm{Tr}(\cdot)$ *is the trace operator.*

[1] In the conjugate prior for linear regression, we need to have $a_0 = c_0$. We are thus treating a more general problem, where the true posterior may not have a closed-form expression.

Exercise 6.5
Consider the dynamical model

$$\begin{aligned} x_{n+1} &= \gamma x_n + v_n, \qquad & v_n &\sim \mathcal{N}\left(0,\ \beta_v^2\right) \\ y_n &= \frac{1}{2} x_n + e_n, \qquad & e_n &\sim \mathcal{N}\left(0,\ \sigma_e^2\right) \end{aligned}$$

with the initial state and prior

$$\begin{aligned} x_0 &\sim \mathcal{N}\left(\bar{\mathbf{x}}_0,\ \mathbf{\Sigma}_0\right) \\ \gamma &\sim \mathcal{N}\left(0,\ \sigma_\gamma^2\right) \end{aligned}$$

We observe $y_1, \dots, y_N$ and consider $x_1, \dots, x_N$ and γ as our latent variables.

a) Derive an expression of the joint distribution

$$p(y_{1:N}, x_{1:N}, \gamma)$$

b) Consider the variational approximation

$$q(\gamma, x_1, \dots, x_N) = q(\gamma) \prod_{n=0}^{N} q(x_n)$$

Use Variational inference to derive the equations for estimating the posterior $q(\gamma) \approx p(\gamma | y_1, \dots, y_N)$.

c) We can also solve this problem without factorizing the terms x_n, i.e by considering

$$q(\gamma, x_1, \dots, x_N) = q(\gamma) q(x_{0:n})$$

Use Variational inference to estimate the posterior of $q(\gamma) \approx p(\gamma | y_1, \dots, y_N)$ using this variational approximation.

Solutions 6

Variational inference & Expectation propagation

Solution to Exercise 6.1 a)

$$\mathrm{KL}(p \parallel q) = -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx$$

The first term can be written as

$$\begin{aligned}
-\int p(x) \ln q(x) dx &= \int \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) \frac{1}{2} \left(\ln(2\pi s^2) + \frac{(x-m)^2}{s^2} \right) dx \\
&= \int \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) \frac{1}{2} \left(\ln(2\pi s^2) + \frac{x^2 - 2xm + m^2}{s^2} \right) dx \\
&= \int \mathcal{N}\left(x;\ \mu,\ \sigma^2\right) \frac{1}{2} \left(\ln(2\pi s^2) + \frac{(x-\mu)^2 + 2x(\mu - m) + m^2 - \mu^2}{s^2} \right) dx \\
&= \frac{1}{2} \left(\ln(2\pi s^2) + \frac{\sigma^2 + 2\mu(\mu - m) + m^2 - \mu^2}{s^2} \right) \\
&= \frac{1}{2} \left(\ln(2\pi s^2) + \frac{\sigma^2 + (\mu - m)^2}{s^2} \right) \qquad (18)
\end{aligned}$$

For the second term we get (by replacing m and s with μ and σ in (18)) that

$$\int p(x) \ln p(x) dx = -\frac{1}{2} \left(\ln(2\pi\sigma^2) + 1 \right)$$

which gives

$$\mathrm{KL}(p \parallel q) = \frac{1}{2} \left(\ln\left(\frac{s^2}{\sigma^2}\right) + \frac{\sigma^2 + (\mu - m)^2}{s^2} - 1 \right)$$

Solution to Exercise 6.2

$$\begin{aligned}
\hat{\mu}, \hat{\sigma} &= \arg\min_{\mu,\sigma} \mathrm{KL}(p(x) \parallel q_{\mu,\sigma}(x)) \\
&= \arg\min_{\mu,\sigma} - \int p(x) \ln q_{\mu,\sigma}(x) dx + \int p(x) \ln p(x) dx \\
&= \arg\min_{\mu,\sigma} \int p(x) \left(\ln \sigma + \frac{1}{2\sigma^2}(x-\mu)^2 \right) dx \\
&= \arg\min_{\mu,\sigma} \left(\ln \sigma + \frac{\sigma_p^2 + (\mu - \mu_p)^2}{2\sigma^2} \right)
\end{aligned}$$

For μ this is minimized by $\hat{\mu} = \mu_p$. What remains for σ is

$$\hat{\sigma} = \arg\min_{\sigma} \underbrace{\left(\ln \sigma + \frac{\sigma_p^2}{2\sigma^2} \right)}_{f(\sigma)}$$

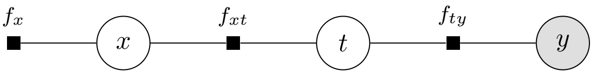
Setting the derivative equal to zero gives

$$0 = \frac{d}{d\sigma} f(\sigma) = \frac{1}{\sigma} - \frac{\sigma_p^2}{\sigma^3} \Rightarrow \sigma^2 = \sigma_p^2$$

This is a minimum point since

$$\frac{d^2}{d\sigma^2} f(\sigma)\Big|_{\sigma=\sigma_p} = -\frac{1}{\sigma_p^2} + \frac{3\sigma_p^2}{\sigma_p^4} = \frac{2}{\sigma_p^2} > 0$$

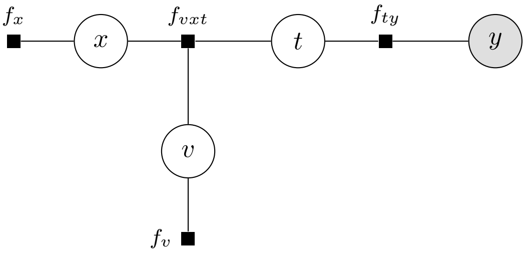
Solution to Exercise 6.3 a) The factor graph can be written as



where

$$\begin{aligned}
f_x(x) &= \mathcal{N}(x;\ 0,\ 1) \\
f_{xt}(x,t) &= \mathcal{N}(t;\ x,\ 1) \\
f_{ty}(t,y) &= \delta(y - \mathrm{sign}(t))
\end{aligned}$$

(Another equivalent factor graph is



with

$$\begin{aligned}
f_x(x) &= \mathcal{N}(x;\ 0,\ 1) \\
f_v(v) &= \mathcal{N}(v;\ 0,\ 1) \\
f_{vxt}(v,x,t) &= \delta(t - v - x) \\
f_{ty}(t,y) &= \delta(y - \mathrm{sign}(t)).
\end{aligned}$$

Since we are not interested in v itself (we do not want to infer $p(v \mid y)$, for instance), we opt for the other formulation, since it is more compact and therefore requires fewer messages in the message passing. Both formulations will result in exactly the same exspression for $p(x \mid y)$.)

b)

$$\begin{aligned} \mu_{y\to f_{ty}}(y) &= \mathbf{1}_{[y=1]} \\ \mu_{f_{ty}\to t}(t) &= \sum_y \delta(y - \text{sign}(t))\,\mathbf{1}_{[y=1]} \\ &= \mathbf{1}_{[t>0]} \end{aligned}$$

The message $\mu_{f_{ty}\to t}(t)$ will result in a non-Gaussian marginal $p(t) = \mu_{f_{ty}\to t}(t)\mu_{f_{xt}\to t}(t)$ and we need to do moment-matching at this node. For this we need all incoming messages. To get $\mu_{f_{xt}\to t}(t)$ we start from the left in the graph

$$\begin{aligned} \mu_{f_x\to x}(x) &= \mathcal{N}\left(x;\ 0,\ 1\right) \\ \mu_{x\to f_{xt}}(x) &= \mu_{f_x\to x}(x) = \mathcal{N}\left(x;\ 0,\ 1\right) \\ \mu_{f_{xt}\to t}(t) &= \int f_{xt}(x,t)\mu_{x\to f_{xt}}(x)dx = \int \mathcal{N}\left(t;\ x,\ 1\right)\mathcal{N}\left(x;\ 0,\ 1\right)dx = \mathcal{N}\left(t;\ 0,\ 2\right) \end{aligned}$$

So we have

$$p(t|y=1) \propto \mu_{f_{ty}\to t}(t)\mu_{f_{xt}\to t}(t) = \mathbf{1}_{[t>0]}\mathcal{N}\left(t;\ 0,\ 2\right)$$

Following the hint, the mean and variance of this truncated Gaussian is

$$\begin{aligned} \mu_t &= \frac{\sqrt{2}\sqrt{2}}{\sqrt{\pi}} = \frac{2}{\sqrt{\pi}} \\ \sigma_t^2 &= 2\left(1 - \frac{2}{\pi}\right) \end{aligned}$$

and with moment matching the approximated marginal $\hat{p}(t)$ is

$$\hat{p}(t|y=1) = \mathcal{N}\left(t;\ \mu_t,\ \sigma_t^2\right)$$

$$\begin{aligned} \hat{\mu}_{f_{ty}\to t}(t) &= \frac{\hat{p}(t|y=1)}{\mu_{f_{xt}\to t}(t)} \quad \Rightarrow \\ \mathcal{N}\left(t;\ \hat{\mu}_{ty},\ \hat{\sigma}_{ty}^2\right) &= \frac{\mathcal{N}\left(t;\ \mu_t,\ \sigma_t^2\right)}{\mathcal{N}\left(t;\ \mu_{xt},\ \sigma_{xt}^2\right)} \end{aligned}$$

where we from division of Gaussians get

$$\begin{aligned} \frac{1}{\hat{\sigma}_{ty}^2} &= \frac{1}{\sigma_t^2} - \frac{1}{\sigma_{xt}^2} = \frac{1}{2\left(1-\frac{2}{\pi}\right)} - \frac{1}{2} = \frac{1}{2\left(1-\frac{2}{\pi}\right)} - \frac{\left(1-\frac{2}{\pi}\right)}{2\left(1-\frac{2}{\pi}\right)} = \frac{1}{\pi - 2} \\ \hat{\mu}_{ty} &= \hat{\sigma}_{ty}^2\left(\frac{\mu_t}{\sigma_t^2} - \frac{\mu_{xt}}{\sigma_{xt}^2}\right) = (\pi - 2)\left(\frac{\frac{2}{\sqrt{\pi}}}{2\left(1-\frac{2}{\pi}\right)} - \frac{0}{2}\right) = \sqrt{\pi} \end{aligned}$$

which gives the new approximated incomming message to t as

$$\hat{\mu}_{f_{ty}}(t) = \mathcal{N}\left(t;\ \sqrt{\pi},\ \pi - 2\right)$$

Now we can proceed propagating back to the node x

$$\begin{aligned} \hat{\mu}_{t\to f_{xt}}(t) &= \hat{\mu}_{f_{ty}\to t}(t) = \mathcal{N}\left(t;\ \sqrt{\pi},\ \pi - 2\right) \\ \hat{\mu}_{f_{xt}\to x}(x) &= \int f_{xt}(x,t)\hat{\mu}_{t\to f_{xt}\to t}(t)dt = \int \mathcal{N}\left(t;\ x,\ 1\right)\mathcal{N}\left(t;\ \sqrt{\pi},\ \pi - 2\right)dt = \mathcal{N}\left(x;\ \sqrt{\pi},\ \pi - 1\right) \end{aligned}$$

Finally, $\hat{p}(x|y=1)$ is computed by multiplying the incomming messages

$$\begin{aligned}\hat{p}(x|y=1) &= \mu_{f_x\to x}(x)\hat{\mu}_{f_{xt}\to x}(x)\\ &= \mathcal{N}\left(x;\ 0,\ 1\right)\mathcal{N}\left(t;\ \sqrt{\pi},\ \pi-1\right)\\ &= \mathcal{N}\left(x;\ \mu_x,\ \sigma_x^2\right)\end{aligned}$$

where

$$\begin{aligned}\frac{1}{\sigma_x^2} &= \frac{1}{1}+\frac{1}{\pi-1} = \frac{\pi}{\pi-1}\\ \mu_x &= \sigma_x^2\left(\frac{0}{1}+\frac{\sqrt{\pi}}{\pi-1}\right)\\ &= \frac{1}{\sqrt{\pi}}\end{aligned}$$

Thus, we have

$$\hat{p}(x|y=1) = \mathcal{N}\left(x;\ \frac{1}{\sqrt{\pi}},\ 1-\frac{1}{\pi}\right)$$

c)

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import truncnorm
from scipy.stats import norm

m0 = 0 # Mean of p(x)
s0 = 1 # Variance of p(x)
s = 1 # Variance of p(t|x)
y0 = 1 # Measurement

# Analytical answer for x
px_m = 1/np.sqrt(np.pi)
px_s = 1-1/np.pi

# Analytical answer t
pt_m = 2/np.sqrt(np.pi)
pt_s = 2*(1-2/np.pi)

# Importance sampler
L = 100000 # number of samples
x = np.random.normal(size=L)*np.sqrt(s0)+m0 #draw from p(x)
t = np.random.normal(size=L)*np.sqrt(s)+x #draw from p(t|x)
y = np.sign(t)
w = (y==y0)

w = L*w/np.sum(w)

# plot a weighted histogram of x
plt.hist(x,weights=w,bins=150,density=True,label="Importance sampling")
xv = np.linspace(-4,4,1000)
plt.plot(xv,norm.pdf(xv,px_m,px_s),label="Moment matching")
plt.xlim((-4,4))
plt.xlabel("x")
plt.legend()
plt.show()

# plot a weighted histogram of t
plt.hist(t,weights=w,bins=150,density=True,label="Importance sampling")
xv = np.linspace(-4,4,1000)
plt.plot(xv,norm.pdf(xv,pt_m,pt_s),label="Moment matching")
plt.xlim((-4,4))
plt.xlabel("t")
plt.legend()
plt.show()

# Estimate mean and variance
est_mean = np.sum(x*w)/L
```

```
est_var = np.sum(w*(est_mean-x)**2)/L

print(est_mean) # Output: 0.559303292236
print(px_m) # Output: 0.564189583548

print(est_var) # Output: 0.68086688604
print(px_s) # Output: 0.6816901138162093
```

Solution to Exercise 6.4 a) Variational approximation

$$q(w, \alpha, \beta) = q(w)q(\alpha)q(\beta)$$

The two equations we will iterate are

$$\begin{aligned}\ln \hat{q}(w) &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}, w, \alpha)] + \textit{const.}\\ \ln \hat{q}(\alpha) &= \mathbb{E}_{\hat{q}(w),\hat{q}(\beta)}[\ln p(\mathbf{y}, w, \alpha)] + \textit{const.}\\ \ln \hat{q}(\beta) &= \mathbb{E}_{\hat{q}(w),\hat{q}(\alpha)}[\ln p(\mathbf{y}, w, \beta)] + \textit{const.}\end{aligned}$$

The joint of distribution of $\mathbf{y}$, w and α is

$$\begin{aligned}p(\mathbf{y}, w, \alpha, \beta) &= p(\mathbf{y}|w, \beta)p(w|\alpha)p(\alpha)p(\beta) \quad \Rightarrow\\ \ln p(\mathbf{y}, w, \alpha, \beta) &= \ln p(\mathbf{y}|w, \beta) + \ln p(w|\alpha) + \ln p(\alpha) + \ln p(\beta)\end{aligned}$$

where

$$\begin{aligned}\ln p(\mathbf{y}|w, \beta) &= \frac{N}{2}\ln\beta - \frac{\beta}{2}(w\mathbf{x} - \mathbf{y})^{\mathsf{T}}(w\mathbf{x} - \mathbf{y}) + \textit{const.}\\ \ln p(w|\alpha) &= \frac{1}{2}\ln\alpha - \frac{\alpha}{2}w^2 + \textit{const.}\\ \ln p(\alpha) &= (a_0 - 1)\ln\alpha - b_0\alpha + \textit{const.}\\ \ln p(\beta) &= (c_0 - 1)\ln\beta - d_0\beta + \textit{const.}\end{aligned}$$

where we have included everything in the constant terms that neither depends on w, α nor β.

We start with $\hat{q}(\alpha)$. These will be the same as in the lecture

$$\begin{aligned}\ln \hat{q}(\alpha) &= \mathbb{E}_{\hat{q}(w),\hat{q}(\beta)}[\ln p(\mathbf{y}, w, \alpha, \beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(w),\hat{q}(\beta)}[\ln p(\mathbf{y}|w, \beta) + \ln p(w|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \ln p(\alpha) + \mathbb{E}_{\hat{q}(w)}[\ln p(w|\alpha)] + \textit{const.}\\ &= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{1}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\hat{q}(w)}[w^2] + \textit{const.}\end{aligned}$$

We recognize this as a Gamma distribution

$$\ln \hat{q}(\alpha) = \ln \mathrm{Gam}\left(\alpha; a_N,\, b_N\right) = (a_N - 1)\ln\alpha - b_N\alpha$$

with

$$\begin{aligned}a_N &= a_0 + \frac{1}{2},\\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_{\hat{q}(w)}[w^2]\end{aligned}$$

Now we proceed with $\hat{q}(w)$.

$$\begin{aligned}\ln \hat{q}(w) &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}, w, \alpha, \beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}|w, \beta) + \ln p(w|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\beta)}[\ln p(\mathbf{y}|w, \beta)] + \mathbb{E}_{\hat{q}(\alpha)}[\ln p(w|\alpha)] + \textit{const.}\\ &= -\frac{1}{2}\mathbb{E}_{\hat{q}(\beta)}[\beta](w\mathbf{x} - \mathbf{y})^{\mathsf{T}}(w\mathbf{x} - \mathbf{y}) - \frac{1}{2}\mathbb{E}[\alpha]w^2 + \textit{const.}\\ &= -\frac{1}{2}(\mathbb{E}_{\hat{q}(\alpha)}[\alpha] + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{x}^{\mathsf{T}}\mathbf{x})w^2 + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{x}^{\mathsf{T}}\mathbf{y}w + \textit{const.}\\ &= -\frac{(w - m_N)^2}{2\sigma_N^2} + \textit{const.}\end{aligned}$$

where

$$\begin{aligned}\sigma_N^2 &= (\mathbb{E}_{\hat{q}(\alpha)}[\alpha] + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{x}^\mathsf{T}\mathbf{x})^{-1}\\ m_N &= \mathbb{E}_{\hat{q}(\beta)}[\beta]\sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{y}\end{aligned}$$

So we have $\hat{q}(w) = \mathcal{N}\left(w;\ m_N,\ \sigma_N^2\right)$

We proceed with $q(\beta)$

$$\begin{aligned}\ln \hat{q}(\beta) &= \mathbb{E}_{\hat{q}(w),\hat{q}(\alpha)}[\ln p(\mathbf{y}, w, \alpha, \beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(w),\hat{q}(\alpha)}[\ln p(\mathbf{y}|w,\beta) + \ln p(w|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \ln p(\beta) + \mathbb{E}_{\hat{q}(w)}[\ln p(\mathbf{y}|w,\beta)] + \textit{const.}\\ &= (c_0-1)\ln\beta - d_0\beta + \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}_{\hat{q}(w)}[(w\mathbf{x}-\mathbf{y})^\mathsf{T}(w\mathbf{x}-\mathbf{y})] + \textit{const.}\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{\hat{q}(w)}[(w\mathbf{x}-\mathbf{y})^\mathsf{T}(w\mathbf{x}-\mathbf{y})] &= \mathbb{E}_{\hat{q}(w)}[w^2]\mathbf{x}^\mathsf{T}\mathbf{x} - 2\mathbb{E}_{\hat{q}(w)}[w]\mathbf{x}^\mathsf{T}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= (m_N^2 + \sigma_N^2)\mathbf{x}^\mathsf{T}\mathbf{x} - 2m_N\mathbf{x}^\mathsf{T}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= \|\mathbf{y} - m_N\mathbf{x}\|^2 + \sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{x}\end{aligned}$$

which inserted in (19) gives

$$\ln \hat{q}(\beta) = (c_0-1)\ln\beta - d_0\beta + \frac{N}{2}\ln\beta - \frac{\beta}{2}\left(\|\mathbf{y} - m_N\mathbf{x}\|^2 + \sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{x}\right) + \textit{const.}$$

We recognize this also as a Gamma distribution

$$\ln \hat{q}(\beta) = \ln \text{Gam}\,(\beta; c_N,\, d_N) = (c_N-1)\ln\beta - d_N\beta$$

with

$$\begin{aligned}c_N &= c_0 + \frac{N}{2},\\ d_N &= d_0 + \frac{1}{2}\left(\|\mathbf{y} - m_N\mathbf{x}\|^2 + \sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{x}\right)\end{aligned}$$

Since

$$\hat{q}(\alpha) = \text{Gam}\,(\alpha; a_N,\, b_N)\,, \qquad \hat{q}(\beta) = \text{Gam}\,(\beta; c_N,\, d_N)\,, \qquad \hat{q}(w) = \mathcal{N}\left(w;\ m_N,\ \sigma_N^2\right)$$

we can compute

$$\mathbb{E}_{\hat{q}(\alpha)}[\alpha] = \frac{a_N}{b_N}, \qquad \mathbb{E}_{\hat{q}(\beta)}[\beta] = \frac{c_N}{d_N}, \qquad \mathbb{E}_{\hat{q}(w)}[w^2] = m_N^2 + \sigma_N^2$$

Now we can state the equations we need to iterate.

Solution: Iterate the following three steps until convergence

- Compute

$$\begin{aligned}a_N &= a_0 + \frac{1}{2},\\ b_N &= b_0 + \frac{1}{2}(m_N^2 + \sigma_N^2)\end{aligned}$$

- Compute

$$\begin{aligned}c_N &= c_0 + \frac{N}{2},\\ d_N &= d_0 + \frac{1}{2}\left(\|\mathbf{y} - m_N\mathbf{x}\|^2 + \sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{x}\right)\end{aligned}$$

- Compute

$$\sigma_N^2 = \left(\frac{a_N}{b_N} + \frac{c_N}{d_N}\mathbf{x}^\mathsf{T}\mathbf{x}\right)^{-1}$$
$$m_N = \frac{a_N}{b_N}\sigma_N^2\mathbf{x}^\mathsf{T}\mathbf{y}$$

b) Variational approximation

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

The two equations we will iterate are

$$\begin{aligned}\ln \hat{q}(\mathbf{w}) &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}, \mathbf{w}, \alpha)] + \textit{const.}\\ \ln \hat{q}(\alpha) &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\beta)}[\ln p(\mathbf{y}, \mathbf{w}, \alpha)] + \textit{const.}\\ \ln \hat{q}(\beta) &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\alpha)}[\ln p(\mathbf{y}, \mathbf{w}, \beta)] + \textit{const.}\end{aligned}$$

The joint of distribution of $\mathbf{y}$, $\mathbf{w}$ and α is

$$\begin{aligned}p(\mathbf{y}, \mathbf{w}, \alpha, \beta) &= p(\mathbf{y}|\mathbf{w}, \beta)p(\mathbf{w}|\alpha)p(\alpha)p(\beta) \quad \Rightarrow\\ \ln p(\mathbf{y}, \mathbf{w}, \alpha, \beta) &= \ln p(\mathbf{y}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha) + \ln p(\alpha) + \ln p(\beta)\end{aligned}$$

where

$$\begin{aligned}\ln p(\mathbf{y}|\mathbf{w}, \beta) &= \frac{N}{2}\ln\beta - \frac{\beta}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) + \textit{const.}\\ \ln p(\mathbf{w}|\alpha) &= \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbf{w}^\mathsf{T}\mathbf{w} + \textit{const.}\\ \ln p(\alpha) &= (a_0 - 1)\ln\alpha - b_0\alpha + \textit{const.}\\ \ln p(\beta) &= (c_0 - 1)\ln\beta - d_0\beta + \textit{const.}\end{aligned}$$

where M is the dimension of $\mathbf{w}$ and where we have included everything in the constant terms that neither depends on $\mathbf{w}$, α nor β.

We start with $\hat{q}(\alpha)$. These will be the same as in the lecture (but now multivariate)

$$\begin{aligned}\ln \hat{q}(\alpha) &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\beta)}[\ln p(\mathbf{y}, \mathbf{w}, \alpha, \beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\beta)}[\ln p(\mathbf{y}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \ln p(\alpha) + \mathbb{E}_{\hat{q}(\mathbf{w})}[\ln p(\mathbf{w}|\alpha)] + \textit{const.}\\ &= (a_0 - 1)\ln\alpha - b_0\alpha + \frac{M}{2}\ln\alpha - \frac{\alpha}{2}\mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}^\mathsf{T}\mathbf{w}] + \textit{const.}\end{aligned}$$

We recognize this as a Gamma distribution

$$\ln \hat{q}(\alpha) = \ln \mathrm{Gam}\left(\alpha; a_N,\, b_N\right) = (a_N - 1)\ln\alpha - b_N\alpha$$

with

$$\begin{aligned}a_N &= a_0 + \frac{M}{2},\\ b_N &= b_0 + \frac{1}{2}\mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}^\mathsf{T}\mathbf{w}]\end{aligned}$$

Now we proceed with $\hat{q}(\mathbf{w})$.

$$\begin{aligned}\ln \hat{q}(\mathbf{w}) &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}, \mathbf{w}, \alpha, \beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\alpha),\hat{q}(\beta)}[\ln p(\mathbf{y}|\mathbf{w}, \beta) + \ln p(\mathbf{w}|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\beta)}[\ln p(\mathbf{y}|\mathbf{w}, \beta)] + \mathbb{E}_{\hat{q}(\alpha)}[\ln p(\mathbf{w}|\alpha)] + \textit{const.}\\ &= -\frac{1}{2}\mathbb{E}_{\hat{q}(\beta)}[\beta](\mathbf{X}\mathbf{w} - \mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w} - \mathbf{y}) - \frac{1}{2}\mathbb{E}[\alpha]\mathbf{w}^\mathsf{T}\mathbf{w} + \textit{const.}\\ &= -\frac{1}{2}\mathbf{w}^\mathsf{T}(\mathbb{E}_{\hat{q}(\alpha)}[\alpha]\mathbf{I}_M + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{w} + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} + \textit{const.}\\ &= -\frac{1}{2}(\mathbf{w} - \mathbf{m}_N)^\mathsf{T}\mathbf{S}_N^{-1}(\mathbf{w} - \mathbf{m}_N) + \textit{const.}\end{aligned}$$

This results in:

$$\begin{aligned}\mathbf{S}_N &= (\mathbb{E}_{\hat{q}(\alpha)}[\alpha]\mathbf{I}_M + \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\\ \mathbf{m}_N &= \mathbb{E}_{\hat{q}(\beta)}[\beta]\mathbf{S}_N\mathbf{X}^\mathsf{T}\mathbf{y}\end{aligned}$$

So we have $\hat{q}(\mathbf{w}) = \mathcal{N}(\mathbf{w};\ \mathbf{m}_N,\ \mathbf{S}_N)$

We proceed with $q(\beta)$

$$\begin{aligned}\ln\hat{q}(\beta) &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\alpha)}[\ln p(\mathbf{y},\mathbf{w},\alpha,\beta)] + \textit{const.}\\ &= \mathbb{E}_{\hat{q}(\mathbf{w}),\hat{q}(\alpha)}[\ln p(\mathbf{y}|\mathbf{w},\beta) + \ln p(\mathbf{w}|\alpha) + \ln p(\alpha) + \ln p(\beta)] + \textit{const.}\\ &= \ln p(\beta) + \mathbb{E}_{\hat{q}(\mathbf{w})}[\ln p(\mathbf{y}|\mathbf{w},\beta)] + \textit{const.}\\ &= (c_0-1)\ln\beta - d_0\beta + \frac{N}{2}\ln\beta - \frac{\beta}{2}\mathbb{E}_{\hat{q}(\mathbf{w})}[(\mathbf{X}\mathbf{w}-\mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w}-\mathbf{y})] + \textit{const.} \qquad (19)\end{aligned}$$

where

$$\begin{aligned}\mathbb{E}_{\hat{q}(\mathbf{w})}[(\mathbf{X}\mathbf{w}-\mathbf{y})^\mathsf{T}(\mathbf{X}\mathbf{w}-\mathbf{y})] &= \mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w} - 2\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{y}]\\ &= \mathbb{E}_{\hat{q}(\mathbf{w})}[\mathrm{Tr}(\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}) - 2\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y}] + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= \mathbb{E}_{\hat{q}(\mathbf{w})}[\mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{w}\mathbf{w}^\mathsf{T}) - 2\mathbf{w}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y}]\\ &= \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}\mathbf{w}^\mathsf{T}]) - 2\mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}^\mathsf{T}]\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{m}_N\mathbf{m}_N^\mathsf{T} + \mathbf{S}_N) - 2\mathbf{m}_N^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= \mathbf{m}_N\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{m}_N^\mathsf{T} - 2\mathbf{m}_N^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} + \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{S}_N) + \mathbf{y}^\mathsf{T}\mathbf{y}\\ &= \|\mathbf{y} - \mathbf{X}\mathbf{m}_N\|^2 + \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{S}_N)\end{aligned}$$

where we have used

$$\mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}\mathbf{w}^\mathsf{T}] = \mathbf{m}_N\mathbf{m}_N^\mathsf{T} + \mathbf{S}_N$$

which inserted in (19) gives

$$\ln\hat{q}(\beta) = (c_0-1)\ln\beta - d_0\beta + \frac{N}{2}\ln\beta - \frac{\beta}{2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_N\|^2 + \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{S}_N)\Big) + \textit{const.}$$

We recognize this also as a Gamma distribution

$$\ln\hat{q}(\beta) = \ln\mathrm{Gam}(\beta; c_N,\ d_N) = (c_N-1)\ln\beta - d_N\beta$$

with

$$\begin{aligned}c_N &= c_0 + \frac{N}{2},\\ d_N &= d_0 + \frac{1}{2}\Big(\|\mathbf{y} - \mathbf{X}\mathbf{m}_N\|^2 + \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{S}_N)\Big)\end{aligned}$$

Since

$$\hat{q}(\alpha) = \mathrm{Gam}(\alpha; a_N,\ b_N), \qquad \hat{q}(\beta) = \mathrm{Gam}(\beta; c_N,\ d_N), \qquad \hat{q}(\mathbf{w}) = \mathcal{N}(\mathbf{w};\ \mathbf{m}_N,\ \mathbf{S}_N)$$

we can compute

$$\mathbb{E}_{\hat{q}(\alpha)}[\alpha] = \frac{a_N}{b_N}, \qquad \mathbb{E}_{\hat{q}(\beta)}[\beta] = \frac{c_N}{d_N}, \qquad \mathbb{E}_{\hat{q}(\mathbf{w})}[\mathbf{w}^\mathsf{T}\mathbf{w}] = \mathbf{m}_N^\mathsf{T}\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N)$$

Now we can state the equations we need to iterate.

Solution: Iterate the following three steps until convergence

- Compute

$$\begin{aligned}a_N &= a_0 + \frac{M}{2},\\ b_N &= b_0 + \frac{1}{2}(\mathbf{m}_N^\mathsf{T}\mathbf{m}_N + \mathrm{Tr}(\mathbf{S}_N))\end{aligned}$$

- Compute

$$\begin{aligned} c_N &= c_0 + \frac{N}{2}, \\ d_N &= d_0 + \frac{1}{2}\left(\|\mathbf{y} - \mathbf{X}\mathbf{m}_N\|^2 + \mathrm{Tr}(\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{S}_N)\right) \end{aligned}$$

- Compute

$$\begin{aligned} \mathbf{S}_N &= \left(\frac{a_N}{b_N}\mathbf{I}_N + \frac{c_N}{d_N}\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} \\ \mathbf{m}_N &= \frac{a_N}{b_N}\mathbf{S}_N\mathbf{X}^\mathsf{T}\mathbf{y} \end{aligned}$$

Solution to Exercise 6.5 a)

$$p(y_{1:N}, x_{1:N}, \gamma) = p(\gamma)p(x_0)\prod_{n=1}^{N} p(x_n|x_{n-1}, \gamma)p(y_n|x_n)$$

where

$$\begin{aligned} p(x_0) &= \mathcal{N}\left(x_0;\ \mu_0,\ \beta_0^{-1}\right) \\ p(\gamma) &= \mathcal{N}\left(\gamma;\ 0,\ \beta_{\gamma 0}^{-1}\right) \\ p(x_n|x_{n-1}, \gamma) &= \mathcal{N}\left(x_n;\ \gamma x_{n-1},\ \beta_v^{-1}\right) \\ p(y_n|x_n) &= \mathcal{N}\left(y_n;\ x_n,\ \beta_e^{-1}\right) \end{aligned}$$

b) First specify the mean field equations

$$\begin{aligned} \ln q(\gamma) &= \mathbb{E}_{\{q_n\}_{n=1}^N}[\ln p(y_{1:N}, x_{0:N}, \gamma)] + \textit{const.} \\ \ln q(x_m) &= \mathbb{E}_{\gamma, \{q_n\}_{n\neq m}}[\ln p(y_{1:N}, x_{0:N}, \gamma)] + \textit{const.} \end{aligned}$$

First, start with $q(\gamma)$

$$\begin{aligned} \ln q(\gamma) &= \mathbb{E}_{\{q_n\}_{n=0}^N}[\ln p(y_{1:N}, x_{0:N}, \gamma)] + \textit{const.} \\ &= \ln p(\gamma) - \frac{\beta_v}{2}\mathbb{E}_{\{q_n\}_{n=0}^N}\left[\sum_{n=1}^{N}\ln p(x_n|x_{n-1}, \gamma)\right] + \textit{const.} \\ &= \ln p(\gamma) - \frac{\beta_v}{2}\sum_{n=1}^{N}\mathbb{E}_{q_n, q_{n-1}}\left[(x_n - x_{n-1}\gamma)^2\right] + \textit{const.} \\ &= \ln p(\gamma) - \frac{\beta_v}{2}\sum_{n=1}^{N}\mathbb{E}_{q_n, q_{n-1}}\left[-2x_n x_{n-1}\gamma + x_{n-1}^2\gamma^2\right] + \textit{const.} \\ &= \ln p(\gamma) - \frac{\beta_v}{2}\sum_{n=1}^{N}\left(-2\bar{x}_n\bar{x}_{n-1}\gamma + \overline{x_{n-1}^2}\gamma^2\right) + \textit{const.} \\ &= \ln p(\gamma) - \frac{\beta_v}{2}\sum_{n=1}^{N}\overline{x_{n-1}^2}\left(\gamma^2 - \frac{\bar{x}_n\bar{x}_{n-1}}{\overline{x_{n-1}^2}}\right)^2 + \textit{const.} \\ &= \ln p(\gamma) + \sum_{n=1}^{N}\ln\mathcal{N}\left(\gamma;\ \frac{\bar{x}_n\bar{x}_{n-1}}{\overline{x_{n-1}^2}},\ (\beta_v\overline{x_{n-1}^2})^{-1}\right) + \textit{const.} \end{aligned}$$

where

$$\bar{x}_n = \mathbb{E}_{q_n}[x_n], \quad \overline{x_n^2} = \mathbb{E}_{q_n}[x_n^2]$$

This gives

$$q(\gamma) = \mathcal{N}\left(\gamma;\ \bar{\gamma},\ \beta_\gamma^{-1}\right)$$

where

$$\beta_\gamma = \beta_{\gamma 0} + \beta_v \sum_{n=1}^{N} \overline{x_{n-1}^2}$$

$$\bar{\gamma} = \frac{1}{\beta_\gamma}\left(\sum_{n=1}^{N} \frac{\bar{x}_n \bar{x}_{n-1}}{\overline{x_{n-1}^2}} \beta_v \overline{x_{n-1}^2}\right) = \frac{\beta_v}{\beta_\gamma}\left(\sum_{n=1}^{N} \bar{x}_n \bar{x}_{n-1}\right)$$

Proceed similarly to derive the update equations for $\ln q(x_m)$.

c) The update equations for $q(\gamma)$ will be the same except that we need to replace $\bar{x}_n \bar{x}_{n-1}$ with $\overline{x_n x_{n-1}} = \mathbb{E}_{\{q_n q_{n-1}\}}[x_n x_{n-1}]$.

The update equations for $\ln q(x_m)$ will result in a Kalman smoother (not in this course, but you might familiar with it if you know some time-series modeling or control theory). By considering an augmented system $\tilde{x}_n = [x_n,\ x_{n-1}]$ to run the Kalman smoother on, we can get the expressions required above.

Bibliography

[1] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.

[2] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.