Uncertainty in Deep Learning

Enbo lyu

February 2024

Contents

1 Preliminary

1.1 Bayes Law

$$P(W|X,Y) = \frac{P(Y|W,X)P(W)}{P(Y|X)}$$

$$\text{Posterier} = \frac{\text{Likelihood} \cdot \text{Prior}}{\text{Evidence}}$$

1.2 Laws of probability

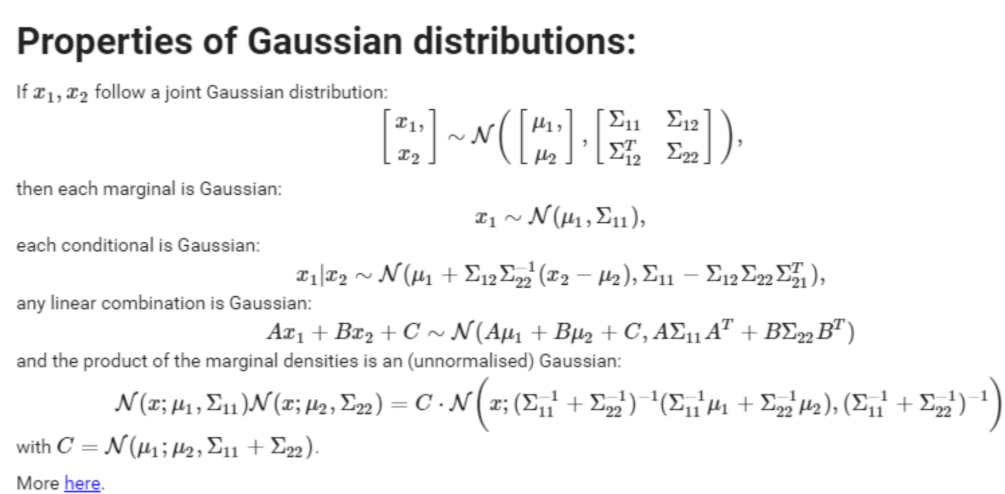
1. Sum rule
$$p(X = x) = \sum_y p(X = x, Y = y) = \int p(X = x, Y = y)dy$$

2. Product rule
$$p(X = x, Y = y) = p(X = x|Y = y)p(Y = y)$$

3. Bayes rule
$$P(W|X,Y) = \frac{P(Y|W,X)P(W)}{P(Y|X)}$$

1.3 Properties of Gaussian distributions

1. Products, ratios, marginals, and conditionals of Gaussians are Gaussian.



Properties of Gaussian distributions:

If x_1, x_2 follow a joint Gaussian distribution:

$$\begin{bmatrix} x_1, \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1, \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}\right),$$

then each marginal is Gaussian:

$$x_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}),$$

each conditional is Gaussian:

$$x_1|x_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{21}^T),$$

any linear combination is Gaussian:

$$Ax_1 + Bx_2 + C \sim \mathcal{N}(A\mu_1 + B\mu_2 + C, A\Sigma_{11}A^T + B\Sigma_{22}B^T)$$

and the product of the marginal densities is an (unnormalised) Gaussian:

$$\mathcal{N}(x; \mu_1, \Sigma_{11})\mathcal{N}(x; \mu_2, \Sigma_{22}) = C \cdot \mathcal{N}\left(x; (\Sigma_{11}^{-1} + \Sigma_{22}^{-1})^{-1}(\Sigma_{11}^{-1}\mu_1 + \Sigma_{22}^{-1}\mu_2), (\Sigma_{11}^{-1} + \Sigma_{22}^{-1})^{-1}\right)$$

with $C = \mathcal{N}(\mu_1; \mu_2, \Sigma_{11} + \Sigma_{22})$.

More here.

Figure 1: Gaussian Properties

1.4 Feature vector

ϕ_k are the basis functions, input $\mathbf{x}$ are fed through K non-linear transformations, then linear regression are done with $\phi(\mathbf{x})$ vector instead of $\mathbf{x}$ itself.

$$\phi(\mathbf{x}) = [\phi_1(\mathbf{x}), \ldots, \phi_K(\mathbf{x})] \in \mathbb{R}^K$$

Feature vector = basis functions' outputs = inputs to linear transformations

1.5 Feature matrix

$$\Phi(\mathbf{X}) = [\phi^T(\mathbf{x}_1), \ldots, \phi^T(\mathbf{x}_N)] \in \mathbb{R}^{N \times K}$$

1.6 Layer

For the moment we only look at the last layer. Denote W to be the weight matrix of the last layer and b the bias of the last layer.

$$W \in \mathbb{R}^{K \times 1}$$

For now assume $b = 0$, then

$$f^W(\mathbf{x}) = \sum w_k \phi_k(\mathbf{x}) = W^T \phi(x) \in \mathbb{R}^{1 \times 1}$$

1.7 Generative story

1. Nature chose W which defines a function: $f^W(x) := W^T \phi(x)$
2. Generated function values f_n with inputs $x_1, \ldots, x_N : f_n := f^W(x_n)$
3. Corrupted function values with noise:

$$y_n := f_n + \epsilon_n, \epsilon_n \sim N(0, \sigma^2)$$

1.8 Model

1. Prior distribution over parameters W

$$p(w_k) = N(w_k; 0, s^2), k \in [1, \ldots, K]$$

2. Likelihood: conditioned on W generate observations by adding gaussian noise

$$p(y|W, x) = N(y; W^T \phi(x), \sigma^2)$$

3. Then by the property of Gaussian, the posterior over W is Gaussian as well.

$$p(W|X,Y) = N(W;\mu',\Sigma')$$

$$\Sigma' = (\sigma^{-2}\sum_n(\phi(x_n)\phi^T(x_n) + s^{-2}I_K)^{-1} = (\sigma^{-2}\Phi^T(\mathbf{X})\Phi(\mathbf{X}) + s^{-2}I_K)^{-1} \in \mathbb{R}^{K\times K}$$

$$\mu' = \Sigma'\sigma^{-2}\sum_n(y_n\phi(x_n)) = \Sigma'\sigma^{-2}\Phi^T(\mathbf{X})Y \in \mathbb{R}^{N\times 1}$$

1.9 Multivariate Bayesian basis function regression

This means input X and Y are now vectors, some dimensions:

1. $\mathbf{X} \in \mathbb{R}^{N\times Q}, \mathbf{Y} \in \mathbb{R}^{N\times D}$
2. $X_n \in R^{1\times Q}$, $Y_n \in R^{1\times D}$
3. $W \in R^{K\times D}$ (transfer dimension from the common dimension of X and Y (K) to the other dimension of Y(D))
4. $\phi(\cdot) \in R^K, \phi(\mathbf{X}) \in R^{K\times N}, \phi(X_n) \in R^{K\times 1}$

1. Prior:

$$p(w_{k,d}) = N(w_{k,d};0,s^2); W \in \mathbb{R}^{K\times D}$$

2. Likelihood:

$$p(\mathbf{Y}|\mathbf{X},W) = \prod_n N(Y_n;f^W(X_n),\sigma^2 I_D); f^W(X_n) = W^T\phi(X) \in \mathbb{R}^{D\times 1}$$

2 Lecture 3 4

2.1 P47

Predictive distribution $p(y^*|x^*, X, Y)$

$$\begin{aligned} p(y^*|x^*, X, Y) &= \int p(y^*, W|x^*, X, Y)dW \\ &= \int p(y^*|x^*, W)p(W|X, Y)dW \end{aligned}$$

2.2 P50, 51 Predictive mean and variance

Q: $p(y^*|x^*, D) \sim N(\mu^*, \Sigma^*)$, what are μ^*, Σ^*?

$$\begin{aligned} \mu^* &= E_{p(y^*|x^*,D)}[y^*] \\ &= \int y^* p(y^*|x^*, D)dy \\ &= \int y^* \int p(y^*, W|x^*, D)dW dy \\ &= \int y^* \int p(y^*|x^*, W)p(W|X, Y)dW dy \\ &= \int \int y^* p(y^*|x^*, W)dy p(W|X, Y)dW \\ &= \int E_{p(y^*|x^*,W)}[y^*]p(W|X, Y)dW \\ &= \int W^T \phi(x^*)p(W|X, Y)dW \\ &= \int W^T p(W|X, Y)dW \phi(x^*) \\ &= E_{p(W|X,Y)}[W^T]\phi(x^*) \\ &= \mu^T \phi(x^*) \end{aligned}$$

$$\Sigma^* = E_{p(y^*|x^*,D)}[{y^*}^T y^*] - E_{p(y^*|x^*,D)}[{y^*}^T]E_{p(y^*|x^*,D)}[y^*]$$

$$
\begin{aligned}
E_{p(y^*|x^*,D)}[y^{*T}y^*] &= \int y^{*T}y^* p(y^*|x^*,D)dy \\
&= \int y^{*T}y^* \int p(y^*|x^*,W)p(W|X,Y)dWdy \\
&= \int E_{p(y^*|x^*,W)}[y^{*T}y^*]p(W|X,Y)dW \\
&= \int (\sigma^2 + \phi(x^*)^T WW^T\phi(x^*))p(W|X,Y)dW \\
&= \sigma^2 + \int \phi(x^*)^T WW^T\phi(x^*))p(W|X,Y)dW \\
&= \sigma^2 + \phi(x^*)^T \int WW^T p(W|X,Y)dW\phi(x^*) \\
&= \sigma^2 + \phi(x^*)^T E_{p(W|X,Y)}[WW^T]\phi(x^*) \\
&= \sigma^2 + \phi(x^*)^T[\Sigma^T + \mu\mu^T]\phi(x^*)
\end{aligned}
$$

$$
\begin{aligned}
\Sigma^* &= E_{p(y^*|x^*,D)}[y^{*T}y^*] - E_{p(y^*|x^*,D)}[y^{*T}]E_{p(y^*|x^*,D)}[y^*] \\
&= \sigma^2 + \phi(x^*)^T[\Sigma^T + \mu\mu^T]\phi(x^*) - \phi(x^*)^T\mu\mu^T\phi(x^*) \\
&= \sigma^2 + \phi(x^*)^T\Sigma^T\phi(x^*)
\end{aligned}
$$

3 Lecture 5 6

3.1 P11

Q: Show that for the new generative story

$$
\begin{gathered}
f_n \mid x_n, W \sim \delta\left(f_n = W^T\phi\left(x_n\right)\right) \\
y_n \mid f_n \sim \mathcal{N}\left(y_n; f_n, \sigma^2\right)
\end{gathered}
$$

we have

$$
\operatorname{Var}_{p(y^*|f^*,X,Y)}\left[y^*\right] = \sigma^2
$$

and

$$
\operatorname{Var}_{p(f^*|x^*,X,Y)}\left[f^*\right] = \phi\left(x^*\right)^T\Sigma'\phi\left(x^*\right)
$$

(hint: use the identity $\int g(X)\delta(X=a)dX = g(a)$ and $\operatorname{Var}(z) = E\left[z^Tz\right] - E[z]^TE[z]$ with simple manipulations)

A:

$$\because y^*|f^*, D \sim N(y_n; f_n, \sigma^2)$$
$$\therefore Var_{p(y^*|f^*,X,Y)}[y^*] = \sigma^2$$

$$Var_{p(f^*|x^*,X,Y)}[f^*] = E[f^{*T} f^*] - E[f^{*T}]E[f^*]$$

$$\begin{aligned} E[f^*] &= \int f^* p(f^*|x^*, D) df^* \\ &= \int f^* \int p(f^*|x^*, D) p(w|D) dW df^* \\ &= \int f^* p(f^*|x^*, D) df^* \int p(W|D) dW \end{aligned}$$

Use the identity $\int g(X)\delta(X = a)dX = g(a)$, equation becomes:

$$\begin{aligned} E[f^*] &= \int W^T \phi(x^*) p(W|D) dW \\ &= E_{p(W|D)}[W^T]\phi(x^*) \\ &= \mu'^T \phi(x^*) \end{aligned}$$

By the same trick,

$$\begin{aligned} E[f^{*T} f^*] &= \int\int f^{*T} f^* p(f^*|x^*, W) df^* p(W|D) dW \\ &= \int \phi(x^*)^T W W^T \phi(x^*) p(W|D) dW \\ &= \phi(x^*)^T \int W W^T p(W|D) dW \phi(x^*) \\ &= \phi(x^*)^T E_{p(W|D)}[W W^T] \phi(x^*) \\ &= \phi(x^*)^T E_{p(W|D)}[W W^T] \phi(x^*) \\ &= \phi(x^*)^T (\Sigma' - \mu' \mu'^T) \phi(x^*) \end{aligned}$$

Therefore:

$$\begin{aligned} Var[f^*] &= E[f^{*T} f^*] - E[f^{*T}]E[f^*] \\ &= \phi(x^*)^T (\Sigma' - \mu' \mu'^T) \phi(x^*) - \phi(x^*)^T \mu' \mu'^T \phi(x^*) \\ &= \phi(x^*)^T \Sigma' \phi(x^*) \end{aligned}$$

3.2 $k(x, x)$, inner product of feature vectors P11

1. $k(x^*, x) = \phi^T(x^*)\phi(x)$
2. $k(x^*, x) \approx 0$ if dissimilar, since the two most dissimilar vectors are orthogonal to each other, their dot product is 0.

3.3 Rewrite the predictive mean and variance P23

3.4 KL Properties P36

3.5 P40 1

Q: For $q(x) = \mathcal{N}\left(x; m_0, s_0^2\right), p(x) = \mathcal{N}\left(x; m_1, s_1^2\right)$ we have

$$\mathrm{KL}(q, p) = 1/2\left(s_1^{-2}s_0^2 + s_1^{-2}\left(m_1 - m_0\right)^2 - 1 + \log\left(s_1^2/s_0^2\right)\right)$$

Show this using def of KL (hint: $E_q\left[x^2\right] = s_0^2 + m_0^2$)

A:

$$\begin{aligned}
KL(q, p) &= \int q(x) log\frac{q(x)}{p(x)} \\
&= \int q(x) log(\frac{1/s_0}{1/s_1} \cdot \frac{exp(-(x - m_0)^2/2s_0^2}{exp(-(x - m_1)^2/2s_1^2}dx \\
&= \int q(x)(log\frac{s_1}{s_0} - (x - m_0)^2/2s_0^2 + (x - m_1)^2/2s_1^2)dx \\
&= log\frac{s_1}{s_0} - (\frac{m_0^2}{2s_0^2} - \frac{m_1^2}{2s_1^2}) + \frac{m_0}{s_0^2}E[x] - \frac{m_1}{s_1^2}E[x] - \frac{1}{2s_0^2}E[x^2] + \frac{1}{2s_1^2}E[x^2]
\end{aligned}$$

3.6 P40 2

Q: If X_1, X_2 are independent unde p and q, then

$$KL(q(X_1, X_2), p(X_, X_2)) = KL(q(X_1), p(X_1)) + KL(q(X_2), p(X_2))$$

A:

By definition o KL, and the independence of X_1, X_2:

$$\begin{aligned} KL(q(X_1,X_2),p(X_1,X_2)) &= \int q(X_1,X_2)log\frac{q(X_1,X_2)}{p(X_1,X_2)}dX \\ &= \int\int q_1q_2log\frac{q_1q_2}{p_1p_2}dX_1dX_2 \\ &= \int q_2dX_2\int q_1log\frac{q_1}{p_1}dX_1 + \int q_1dX_1\int q_2log\frac{q_2}{p_2}dX_2 \\ &= \int q_1log\frac{q_1}{p_1}dX_1 + \int q_2log\frac{q_2}{p_2}dX_2 \\ &= KL(q_1,p_1) + KL(q_2,p_2) \end{aligned}$$

3.7 ELBO P44

Q: Show $\mathrm{KL}\left(q_\theta(W), p(W \mid X, Y)\right) = \log p(Y \mid X) - \int q_\theta(W)\log p(Y \mid X, W)\mathrm{d}W + \mathrm{KL}\left(q_\theta(W), p(W)\right)$

A:

$$\begin{aligned} \mathrm{KL}\left(q_\theta(W), p(W \mid X, Y)\right) &= \int q(W)log\frac{q(W)}{p(W|X,Y)}dW \\ &= \int q(W)log\frac{q(W)}{\frac{p(Y|X,W)p(W)}{p(Y|X)}}dW \\ &= logP(Y|X) + \int q(W)log\frac{q(W)}{p(W)}dW + \int q(W)log\frac{1}{p(Y|W,X)}dW \\ &= logP(Y|X) - \int q(W)logp(Y|W,X)dW + KL(q(W),p(W)) \end{aligned}$$

3.8 ELBO from a different way P52

3.8.1 Preliminary

1. Jensen's inequality with log and $\mathbb{E}$ (based on the convexity of $-log$)

$$log(E[f(x)]) \geq E[log(f(x))]$$

2. Useful trick to change the base of expectation

$$
\begin{aligned}
E_{p(x)}[f(x)] &= \int p(X) f(X) dX \\
&= \int p(X) \frac{q(X)}{q(X)} f(X) dX \\
&= \int q(X) \frac{p(X)}{q(X)} f(X) dX \\
&= E_{q(x)}[\frac{p(X)}{q(X)} f(x)]
\end{aligned}
$$

3.8.2 ELBO

$$
\begin{aligned}
logp(Y|X) &= log \int p(Y, W|X) dW \\
&= log \int p(Y|W, X) p(W) dW \\
&= log(E_{p(W)}[p(Y|W, X)]) \\
&= log(E_{q(W)}[\frac{p(W)}{q(W)} p(Y|W, X)]) \\
&\geq E_{q(W)}[log(\frac{p(W)}{q(W)} p(Y|W, X))] \\
&= E_{q(W)}[log(p(Y|W, X)) + log(-\frac{q(W)}{p(W)})] \\
&= \int q(W) log(p(Y|W, X)) dW - KL(q(W), p(W))
\end{aligned}
$$

3.9 ELBO in matrix P55

Q: write the ELBO in terms of s, σ, M, S only

1. prior $p\left(w_{kd}\right) = \mathcal{N}\left(w_{kd}; 0, s^2\right)$
2. $f^W(\mathbf{x}) = W^T \phi(\mathbf{x})$
3. likelihood $p\left(Y_n \mid X_n, W\right) = \mathcal{N}\left(Y_n; f^W\left(X_n\right), \sigma^2 I_D\right)$
4. approx post $q_{m,\sigma}\left(w_{kd}\right) = \mathcal{N}\left(w_{kd}; m_{kd}, \sigma_{kd}^2\right), M = \left[m_{kd}\right], S = \left[\sigma_{kd}\right]$
5. Reminder: $\mathcal{N}(X \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^K |\Sigma|}} e^{-\frac{1}{2}(X-\mu)^T \Sigma^{-1}(X-\mu)}$

A:

$$
ELBO = \int q(W) log(p(Y|W, X)) dW - KL(q(W), p(W))
$$

3.9.1 $\int q(W)log(p(Y|W,X))dW$

$$
\begin{aligned}
log(p(Y|W,X)) &= log((2\pi\sigma^2)^{-\frac{1}{2}}exp[-\frac{1}{2}\sigma^{-2}\sum_n \|y_n - f^W(x_n)\|_2^2] \\
&= log((2\pi\sigma^2)^{-\frac{1}{2}}) - \frac{1}{2}\sigma^{-2}\sum_n \|y_n - f^W(x_n)\|_2^2 \\
&= log((2\pi\sigma^2)^{-\frac{1}{2}}) - \frac{1}{2}\sigma^{-2}\sum_n (y_n^T y_n + f^T f - 2y_n f) \\
&= log((2\pi\sigma^2)^{-\frac{1}{2}}) - \frac{1}{2}\sigma^{-2}\sum_n y_n^T y_n + \sum_n f^T f - \sum_n 2y_n^T f
\end{aligned}
$$

$$
\begin{aligned}
E_{q(W)}[log(p(Y|W,X))] &= E_{q(W)}[-\frac{1}{2}\sigma^{-2}\sum_n y_n^T y_n + \sum_n f^T f - \sum_n 2y_n f] \\
&= -\frac{1}{2}\sigma^{-2}\sum_n E_{q(W)}[y_n^T y_n] + \sum_n E_{q(W)}[f^T f] - \sum_n E_{q(W)}[2y_n^T f]
\end{aligned}
$$

$$
\sum_n E_{q(W)}[y_n^T y_n] = \sum_n y_n^T y_n
$$

$$
\begin{aligned}
\sum_n E_{q(W)}[2y_n^T f] &= \sum_n 2y_n^T E_{q(W)}[W^T \phi(X_n)] \\
&= \sum_n 2y_n^T E_{q(W)}[W^T]\phi(X_n) \\
&= \sum_n 2y_n^T M_n^T \phi(X_n)(dim(D\times 1)(1\times K)(K\times 1) = (D\times 1))?
\end{aligned}
$$

$$
\begin{aligned}
\sum_n E_{q(W)}[f^T f] &= \sum_n E_{q(W)}[\phi^T(X_n) W W^T \phi(X_n)] \\
&= \sum_n \phi^T(X_n) E[W W^T]\phi(X_n)
\end{aligned}
$$

$$
\begin{aligned}
E[WW^T]_{kk'} &= E[\sum_d w_{kd}w_{k'd}] \\
&= \sum_d E[w_{kd}w_{k'd}]
\end{aligned}
$$

1. if $k = k'$:
$$E[WW^T]_{kk'} = \sum_d M_{kd}M_{kd'} + \sigma_{kd}^2$$
2. Otherwise, covariance:
$$E[WW^T]_{kk'} = \sum_d M_{kd}M_{kd'} + \sigma_{kd}^2 \mathbb{I}_{k=k'} = MM^T + diag(SS^T)$$

Putting altogether, we have:

$$
\begin{aligned}
E_{q(W)}[log(p(Y|W,X))] &= -\frac{1}{2\sigma^2}\sum_n E_{q(W)}[y_n^T y_n] + \sum_n E_{q(W)}[f^T f] - \sum_n E_{q(W)}[2y_n^T f] - \frac{D}{2}log2\pi\sigma^2 \\
&= -\frac{1}{2\sigma^2}(\sum_n Y_n Y_n^T + \phi^T(X_n)(M^T M + diag(SS^T)\phi(X_n) - 2Y_n^T(M^T\phi(X_n))) - \frac{D}{2}log2\pi\sigma^2 \\
&= -\frac{1}{2\sigma^2}\sum_n \|Y_n - M^T\phi(X_n)\|_2^2 + \phi^T(X_n)diag(SS^T)\phi(X_n) - \frac{D}{2}log2\pi\sigma^2 \\
&= -\frac{1}{2\sigma^2}\sum_n (\|Y_n - M^T\phi(X_n)\|_2^2 + \phi^T(X_n)diag(SS^T)\phi(X_n)) - \frac{ND}{2}log2\pi\sigma^2
\end{aligned}
$$

3.9.2 $KL(q, p)$

$$
\begin{aligned}
KL(q(W), p(W)) &= \int q(W) log\frac{q(W)}{p(W)} dW \\
&= \mathbb{E}_{q(W)}[log\frac{q(W)}{p(W)}]
\end{aligned}
$$

$$
\begin{aligned}
log\frac{q(W)}{p(W)} &= log(\frac{(\frac{1}{\sqrt{(2\pi)^N \sigma_{kd}^2}} exp(-\frac{1}{2}(w - m_{kd})^T \sigma_{kd}^{-2}(w - m_{kd}))}{(\frac{1}{\sqrt{(2\pi)^N s^2}} exp(-\frac{1}{2}w^T s^{-2} w)} \\
&= log(\frac{s}{\sigma_{kd}}) + \frac{1}{2}w^T s^{-2} w - \frac{1}{2}(w - m_k d)^T \sigma_{kd}^{-2}(w - m_k d)
\end{aligned}
$$

$$\begin{aligned}
\mathbb{E}_{q(W)}[log\frac{q(W)}{p(W)}] &= \sum_{kd} \mathbb{E}_{q(W)}[log(\frac{s}{\sigma_{kd}}) + \frac{1}{2}w^T s^{-2} w - \frac{1}{2}(w - m_k d)^T \sigma_{kd}^{-2}(w - m_k d)] \\
&= \sum_{kd} \frac{1}{2}[log(\frac{s^2}{\sigma_{kd}^2}) + \mathbb{E}_{q(W)}[w^T s^{-2} w] - \mathbb{E}_{q(W)}[(w - m_k d)^T \sigma_{kd}^{-2}(w - m_k d)] \\
&= \sum_{kd} \frac{1}{2}[log(\frac{s^2}{\sigma_{kd}^2}) + s^{-2}\mathbb{E}_{q(W)}[w^T w] - \sigma_{kd}^{-2}\mathbb{E}_{q(W)}[(w - m_k d)^T (w - m_k d)]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{q(W)}[w^T w] &= m_{kd}^2 + \sigma_{kd}^2 \\
\mathbb{E}_{q(W)}[w] &= m_{kd}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{q(W)}[(w - m_k d)^T (w - m_k d)] &= E_{q(W)}[w^T w + m^T m - 2mw] \\
&= E_{q(W)}[w^T w] + m^T m - 2mE_{q(W)}[w]
\end{aligned}$$

Putting altogether:

$$\begin{aligned}
equa &= \sum_{kd} \frac{1}{2}[log(\frac{s^2}{\sigma_{kd}^2}) + (m_{kd}^2 + \sigma_{kd}^2)s^{-2} - \sigma_{kd}^{-2}(m_{kd}^2 + \sigma_{kd}^2) - \frac{m_{kd}^2}{\sigma_{kd}^2} + 2\frac{m_{kd}^2}{\sigma_{kd}^2}] \\
&= \sum_{kd} \frac{1}{2}[log(\frac{s^2}{\sigma_{kd}^2}) + s^{-2}(m_{kd}^2 + \sigma_{kd}^2) - 1]
\end{aligned}$$

3.9.3 ELBO

$$\begin{aligned}
ELBO &= \int q(W) log(p(Y|W,X))dW - KL(q(W), p(W)) \\
&= -\frac{1}{2\sigma^2}\left(\sum_n \|Y_n - M^T\phi(X_n)\|_2^2 + \phi^T(X_n)\mathrm{diag}(SS^T)\phi(X_n)\right) - \frac{ND}{2}\log(2\pi\sigma^2) \\
&\quad - \sum_{kd}\frac{1}{2}\left(s^{-2}(\sigma_{kd}^2 + m_{kd}^2) - 1 + \log\frac{s^2}{\sigma_{kd}^2}\right)
\end{aligned}$$

3.10 Optimal likelihood variance σ^2 P58

Let $a = \sigma^2$ and differentiate wrt a:

$$L = -\frac{1}{2a} * b - \frac{ND}{2} log 2\pi a - c$$
$$\frac{dL}{da} = \frac{1}{2} a^{-2} b - \frac{ND}{2a}$$

$$\frac{da}{d\sigma^2} = 1$$

$$\frac{dL}{d\sigma^2} = \frac{1}{2} a^{-2} b - \frac{ND}{2a} = 0$$
$$\sigma^2 = a = \frac{b}{ND}$$