1 General Variational Inference

The general form of the ELBO is:

$$\mathbb{E}_{q(w_l)}[\sum_{i=1}^{n} p(y_i|f(x_i))] - D_{KL}(q(w_l)\|p(w_l))$$

1.1 Likelihood for categorical data

Assuming $f_w(x)$ outputs a probability distribution over the classes, we use a categorial distribution for the likelihood

$$p(y_i\|f_w(x_i)) = \prod_{j=1}^{k} f_w(x_{i_j})^{y_{i_j}}$$

If we have just one true class l and one-hot encoded, we have:

$$= f_w(x_{i_l})$$

2 Predictions

$$\begin{aligned} p(y|f_w(x)) &= \mathbb{E}_{p(w)}[p(y|f_w(x))] \\ &\approx E_{q(w)}[p(y|f_w(x))] \\ &\approx \frac{1}{M}\sum_{i=1}^{m} P(y|f_{w_i}(x)) \end{aligned}$$

2.1 ELBO scaling

2.1.1 Batch size

If we have B equally sized batches with size b, then we need to scale the ELBO by

$$\mathbb{E}_{q(w_l)}[\sum_{i=1}^{b} p(y_i|f(x_i))] - \frac{1}{B} D_{KL}(q(w_l)\|p(w_l))$$

for each parameter update according to: "Practical Variational Inference for Neural Networks" or "Bayesian Learning via Stochastic Gradient Langevin Dynamics" (if we assume equally sized batches)

2.1.2 Mean instead of sum

If we use a sum for our likelihood-loss, the values could get too big, that's why we need to scale the KL div also by the batch size b which results in

$$\mathbb{E}_{q(w_l)}[\frac{1}{b}\sum_{i=1}^{b} p(y_i|f(x_i))] - \frac{1}{B*b} D_{KL}(q(w_l)\|p(w_l))$$
$$= \mathbb{E}_{q(w_l)}[\frac{1}{b}\sum_{i=1}^{b} p(y_i|f(x_i))] - \frac{1}{N} D_{KL}(q(w_l)\|p(w_l))$$

where N is the total number of samples-

3 Different Kernels for Gaussians

3.1 KL-Divergence Gaussian

The KL-Div between two Gaussians is:

$$\begin{aligned}
D_{KL}(q(w_l)\|p(w_l)) &= \mathbb{E}_{q(w_l)}[log(\frac{p(w_l)}{q(w_l)})] \\
&= \mathbb{E}_{q(w_l)}[log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_q)}} exp(-\frac{1}{2}(w_l-\mu_q)^T\Sigma_q^{-1}(w_l-\mu_q))) \\
&- log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_p)}} exp(-\frac{1}{2}(w_l-\mu_p)^T\Sigma_p^{-1}(w_l-\mu_p)))] \\
&= \mathbb{E}_{q(w_l)}[log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_q)}}) - \frac{1}{2}(w_l-\mu_q)^T\Sigma_q^{-1}(w_l-\mu_q) \\
&- log(\frac{1}{\sqrt{(2\pi)^{\frac{n}{2}} det(\Sigma_p)}}) + \frac{1}{2}(w_l-\mu_p)^T\Sigma_p^{-1}(w_l-\mu_p)] \\
&= \mathbb{E}_{q(w_l)}[\frac{1}{2}log(\frac{det(\Sigma_p)}{det(\Sigma_q)}) - \frac{1}{2}(w_l-\mu_q)^T\Sigma_q^{-1}(w_l-\mu_q) \\
&+ \frac{1}{2}(w_l-\mu_p)^T\Sigma_p^{-1}(w_l-\mu_p)] \\
&= \frac{1}{2}(log(\frac{det(\Sigma_p)}{det(\Sigma_q)}) - n + tr(\Sigma_p^{-1}\Sigma_q) + (\mu_p-\mu_q)^T\Sigma_p^{-1}(\mu_p-\mu_q))
\end{aligned}$$

3.2 Diagonal

We set the prior $p(w_l)$ as well as q to a diagonal gaussian, this means the Kl-Divergence can be further simplified to:

$$= \frac{1}{2}\sum_{i=1}^{n} log(\frac{\Sigma_{p_i}}{\Sigma_{q_i}}) - n + \frac{\Sigma_{q_i}}{\Sigma_{p_i}} + (\mu_{p_i}-\mu_{q_i})^2\Sigma_{p_i}^{-1} \tag{1}$$

If we use a prior of $\mu_p = 0$ and $\Sigma_p = Id$, then

$$= \frac{1}{2} \sum_{i=1}^{n} -log(\Sigma_{q_i}) - n + \Sigma_{q_i} + \mu_{q_i}^2 \tag{2}$$

3.3 K-Fac

First K-Fac just has covariances within each layer, however as we have just one layer we have a full covariance matrix. Next K-Fac approximates each block covariance C with a Kronecker-Product of two Matrices. $C = A \otimes B$

3.3.1 Determinant

If $A, n \times n$ and $B, m \times m$ are square, the determinant of a the Kronecker Product is

$$det(A \otimes B) = det(A)^n det(B)^m \tag{3}$$

3.3.2 Cholesky decomposition

Because the covariance C needs to be psd, each factor A, B needs to be psd according to 3. We can achieve this by parametrizing them with a Cholesky decomposition. To do that, decompose A, B in lower triangular matrices L_A, L_B with positive diagonal elements, such that: $A = L_A L_A^T$, $B = L_B L_B^T$. Moreover, if we have the Choleksy decomposition then we can rewrite the Kronecker product as:

$$A \otimes B = L_A L_A^T \otimes L_B L_B^T = (L_A \otimes L_B)(L_A \otimes L_B)^T \tag{4}$$

3.3.3 Determinant together with Cholesky

The logdeterminant for the KL divergence with 3 is:

$$\begin{aligned}
\log\det(A \otimes B) &= \log(\det(A)^n \det(B)^m) && (5)\\
&= n \cdot \log(\det(A)) + m \cdot \log(\det(B)^m) && (6)\\
&= 2n \cdot \sum_i \log(A_{ii}) + 2m \cdot \sum_i \log(B_{ii}) && (7)
\end{aligned}$$

where we used the fact that $\det(L) = \prod_i L_{ii}$ for triangular matrices.

However with equation 4 we can simplify it easier:

$$\begin{aligned}
\log\det(A \otimes B) &= \log\det(L_A L_A^T \otimes L_B L_B^T) && (8)\\
&= \log\det((L_A \otimes L_B)(L_A \otimes L_B)^T) && (9)\\
&= \log(\det(L_A \otimes L_B)\det((L_A \otimes L_B)^T))) && (10)\\
&= \log((\det(L_A \otimes L_B))^2) && (11)\\
&= 2\log(\det(L_A \otimes L_B)) && (12)\\
&= 2\sum_i \log(L_A \otimes L_B)_{ii} && (13)\\
& && (14)
\end{aligned}$$