Appendix A

KL condition

We show that in the dropout case, the KL condition (eq. (3.12)) holds for a large enough number of hidden units when we specify the model prior to be a product of uncorrelated Gaussian distributions over each weight[1]:

$$p(\boldsymbol{\omega}) = \prod_{i=1}^{L} p(\mathbf{W}_i) = \prod_{i=1}^{L} \mathcal{MN}(\mathbf{W}_i; 0, \mathbf{I}/l_i^2, \mathbf{I}).$$

We set the approximating distribution to be $q_\theta(\boldsymbol{\omega}) = \int q_\theta(\boldsymbol{\omega}|\boldsymbol{\epsilon})p(\boldsymbol{\epsilon})\mathrm{d}\boldsymbol{\epsilon}$ where $q_\theta(\boldsymbol{\omega}|\boldsymbol{\epsilon}) = \delta(\boldsymbol{\omega} - g(\theta, \boldsymbol{\epsilon}))$, with $g(\theta, \boldsymbol{\epsilon}) = \{\mathrm{diag}(\boldsymbol{\epsilon}_1)\mathbf{M}_1, \mathrm{diag}(\boldsymbol{\epsilon}_2)\mathbf{M}_2, \mathbf{b}\}$, $\theta = \{\mathbf{M}_1, \mathbf{M}_2, \mathbf{b}\}$, and $p(\boldsymbol{\epsilon}_i)$ defined as a product of Bernoulli distributions ($\boldsymbol{\epsilon}_i$ is a vector of draws from the Bernoulli distribution). Since we assumed $q_\theta(\boldsymbol{\omega})$ to factorise over the layers and over the rows of each weight matrix, we have

$$\mathrm{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) = \sum_{i,k} \mathrm{KL}(q_{\theta_{i,k}}(\mathbf{w}_{i,k})||p(\mathbf{w}_{i,k}))$$

with i summing over the layers and k summing over the rows in each layers' weight matrix.

We approximate each $q_{\theta_{i,k}}(\mathbf{w}_{i,k}|\boldsymbol{\epsilon}) = \delta(\mathbf{w}_{i,k} - g(\theta_{i,k}, \epsilon_{i,k}))$ as a narrow Gaussian with a small standard deviation $\Sigma = \sigma^2 I$. This means that marginally $q_{\theta_{i,k}}(\mathbf{w}_{i,k})$ is a mixture of two Gaussians with small standard deviations, and one component fixed at zero. For large enough models, the KL condition follows from this general proposition:

Proposition 4. *Fix $K, L \in \mathbb{N}$, a probability vector $\mathbf{p} = (p_1, ..., p_L)$, and $\boldsymbol{\Sigma}_i \in \mathbb{R}^{K \times K}$ diagonal positive-definite for $i = 1, ..., L$, with the elements of each $\boldsymbol{\Sigma}_i$ not dependent on*

[1] Here $\mathcal{MN}(0, \mathbf{I}, \mathbf{I})$ is the standard matrix Gaussian distribution.

K. Let

$$q(\mathbf{x}) = \sum_{i=1}^{L} p_i \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

be a mixture of Gaussians with L components and $\boldsymbol{\mu}_i \in \mathbb{R}^K$, let $p(\mathbf{x}) = \mathcal{N}(0, \mathbf{I}_K)$, and further assume that $\boldsymbol{\mu}_i - \boldsymbol{\mu}_j \sim \mathcal{N}(0, I)$ for all i, j.

The KL divergence between $q(\mathbf{x})$ and $p(\mathbf{x})$ can be approximated as:

$$KL(q(\mathbf{x})||p(\mathbf{x})) \approx \sum_{i=1}^{L} \frac{p_i}{2} \Big(\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + tr(\boldsymbol{\Sigma}_i) - K(1 + \log 2\pi) - \log |\boldsymbol{\Sigma}_i| \Big) - \mathcal{H}(\mathbf{p}) \tag{A.1}$$

with $\mathcal{H}(\mathbf{p}) := -\sum_{i=1}^{L} p_i \log p_i$ for large enough K.

Before we prove the proposition, we observe that a direct result from it is the following:

Corollary 2. The KL condition (eq. (3.12)) holds for a large enough number of hidden units when we specify the model prior to be

$$p(\boldsymbol{\omega}) = \prod_{i=1}^{L} p(\mathbf{W}_i) = \prod_{i=1}^{L} \mathcal{MN}(\mathbf{W}_i; 0, \mathbf{I}/l_i^2, \mathbf{I})$$

and the approximating distribution to be a *dropout variational distribution.*

Proof.

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{m}_{i,k}} \text{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega})) &= \frac{\partial}{\partial \mathbf{m}_{i,k}} \text{KL}(q_{\theta_{i,k}}(\mathbf{w}_{i,k})||p(\mathbf{w}_{i,k})) \\
&\approx \frac{(1-p_i) l_i^2}{2} \frac{\partial}{\partial \mathbf{m}_{i,k}} \mathbf{m}_{i,k}^T \mathbf{m}_{i,k} \\
&= \frac{\partial}{\partial \mathbf{m}_{i,k}} N\tau(\lambda_1 ||\mathbf{M}_1||^2 + \lambda_2 ||\mathbf{M}_2||^2 + \lambda_3 ||\mathbf{b}||^2)
\end{aligned}$$

for $\lambda_i = \frac{(1-p_i) l_i^2}{2N\tau}$. □

Next we prove proposition 4.

Proof. We have

$$\begin{aligned}
\text{KL}(q(\mathbf{x})||p(\mathbf{x})) &= \int q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} \mathrm{d}\mathbf{x} \\
&= \int q(\mathbf{x}) \log q(\mathbf{x}) \mathrm{d}\mathbf{x} - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x}
\end{aligned}$$

$$= -\mathcal{H}(q(\mathbf{x})) - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x} \tag{A.2}$$

—a sum of the entropy of $q(\mathbf{x})$ ($\mathcal{H}(q(\mathbf{x}))$) and the expected log probability of $\mathbf{x}$. The expected log probability can be evaluated analytically, but the entropy term has to be approximated.

We begin by approximating the entropy term. We write

$$\begin{aligned}
\mathcal{H}(q(\mathbf{x})) &= -\sum_{i=1}^{L} p_i \int \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \log q(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= -\sum_{i=1}^{L} p_i \int \mathcal{N}(\boldsymbol{\epsilon}_i; 0, \mathbf{I}) \log q(\boldsymbol{\mu}_i + \mathbf{L}_i \boldsymbol{\epsilon}_i) \mathrm{d}\boldsymbol{\epsilon}_i
\end{aligned}$$

using a change of variables $\mathbf{x} = \boldsymbol{\mu}_i + \mathbf{L}_i \boldsymbol{\epsilon}_i$ with $\mathbf{L}_i \mathbf{L}_i^T = \boldsymbol{\Sigma}_i$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, I)$.

Now, the term inside the logarithm can be written as

$$\begin{aligned}
q(\boldsymbol{\mu}_i + \mathbf{L}_i \boldsymbol{\epsilon}_i) &= \sum_{j=1}^{L} p_i \mathcal{N}(\boldsymbol{\mu}_i + \mathbf{L}_i \boldsymbol{\epsilon}_i; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\
&= \sum_{j=1}^{L} p_i (2\pi)^{-K/2} |\boldsymbol{\Sigma}_j|^{-1/2} \exp\Big\{ -\frac{1}{2} ||\boldsymbol{\mu}_j - \boldsymbol{\mu}_i - \mathbf{L}_i \boldsymbol{\epsilon}_i||^2_{\boldsymbol{\Sigma}_j} \Big\}
\end{aligned}$$

where $||\cdot||_{\boldsymbol{\Sigma}}$ is the Mahalanobis distance. Since $\boldsymbol{\mu}_i, \boldsymbol{\mu}_j$ are assumed to be normally distributed, the quantity $\boldsymbol{\mu}_j - \boldsymbol{\mu}_i - \mathbf{L}_i \boldsymbol{\epsilon}_i$ is also normally distributed[2]. Since the expectation of a generalised χ^2 distribution with K degrees of freedom increases with K, we have that[3] $K \gg 0$ implies that $||\boldsymbol{\mu}_j - \boldsymbol{\mu}_i - \mathbf{L}_i \boldsymbol{\epsilon}_i||^2_{\boldsymbol{\Sigma}_j} \gg 0$ for $i \neq j$ (since the elements of $\boldsymbol{\Sigma}_j$ do not depend on K). Finally, we have for $i = j$ that $||\boldsymbol{\mu}_i - \boldsymbol{\mu}_i - \mathbf{L}_i \boldsymbol{\epsilon}_i||^2_{\boldsymbol{\Sigma}_i} = \boldsymbol{\epsilon}_i^T \mathbf{L}_i^T \mathbf{L}_i^{-T} \mathbf{L}_i^{-1} \mathbf{L}_i \boldsymbol{\epsilon}_i = \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$. Therefore the last equation can be approximated as

$$q(\boldsymbol{\mu}_i + \mathbf{L}_i \boldsymbol{\epsilon}_i) \approx p_i (2\pi)^{-K/2} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\Big\{ -\frac{1}{2} \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i \Big\}.$$

I.e., in high dimensions the mixture components will not overlap. This gives us

$$\begin{aligned}
\mathcal{H}(q(\mathbf{x})) &\approx -\sum_{i=1}^{L} p_i \int \mathcal{N}(\boldsymbol{\epsilon}_i; 0, \mathbf{I}) \log \Big(p_i (2\pi)^{-K/2} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\Big\{ -\frac{1}{2} \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i \Big\} \Big) \mathrm{d}\boldsymbol{\epsilon}_i \\
&= \sum_{i=1}^{L} \frac{p_i}{2} \Big(\log |\boldsymbol{\Sigma}_i| + \int \mathcal{N}(\boldsymbol{\epsilon}_i; 0, \mathbf{I}) \boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i \mathrm{d}\boldsymbol{\epsilon}_i + K \log 2\pi \Big) + \mathcal{H}(\mathbf{p})
\end{aligned}$$

[2] With mean zero and variance $\mathrm{Var}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i - \mathbf{L}_i \boldsymbol{\epsilon}_i) = 2I + \boldsymbol{\Sigma}_i$.

[3] To be exact, for diagonal matrices Λ, Δ and $\mathbf{v} \sim \mathcal{N}(0, \Lambda)$, we have $\mathbb{E}[||\mathbf{v}||_\Delta] = \mathbb{E}[\mathbf{v}^T \Delta^{-1} \mathbf{v}] = \sum_{k=1}^{K} \mathbb{E}[\Delta_k^{-1} v_k^2] = \sum_{k=1}^{K} \Delta_k^{-1} \Lambda_k$.

where $\mathcal{H}(\mathbf{p}) := -\sum_{i=1}^{L} p_i \log p_i$. Since $\boldsymbol{\epsilon}_i^T \boldsymbol{\epsilon}_i$ distributes according to a χ^2 distribution, its expectation is K, and the entropy can be approximated as

$$\mathcal{H}(q(\mathbf{x})) \approx \sum_{i=1}^{L} \frac{p_i}{2} \Big(\log |\boldsymbol{\Sigma}_i| + K(1 + \log 2\pi) \Big) + \mathcal{H}(\mathbf{p}). \tag{A.3}$$

Next, evaluating the expected log probability term of the KL divergence we get

$$\int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x} = \sum_{i=1}^{L} p_i \int \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \log p(\mathbf{x}) \mathrm{d}\mathbf{x}$$

for $p(\mathbf{x}) = \mathcal{N}(0, \mathbf{I}_K)$ it is easy to show that

$$\int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x} = -\frac{1}{2} \sum_{i=1}^{L} p_i \Big(\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \mathrm{tr}(\boldsymbol{\Sigma}_i) \Big). \tag{A.4}$$

Finally, combining eq. (A.3) and eq. (A.4) as in (A.2) we get:

$$\mathrm{KL}(q(\mathbf{x}) || p(\mathbf{x})) \approx \sum_{i=1}^{L} \frac{p_i}{2} \Big(\boldsymbol{\mu}_i^T \boldsymbol{\mu}_i + \mathrm{tr}(\boldsymbol{\Sigma}_i) - K(1 + \log 2\pi) - \log |\boldsymbol{\Sigma}_i| \Big) - \mathcal{H}(\mathbf{p}),$$

as required to show. $\square$