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$^1$:

$$p(\omega) = \prod_{i=1}^{L} p(W_i) = \prod_{i=1}^{L} \mathcal{MN}(W_i; 0, I/\ell_i^2, I).$$

We set the approximating distribution to be $q_\theta(\omega) = \int q_\theta(\omega|\epsilon)p(\epsilon)d\epsilon$ where $q_\theta(\omega|\epsilon) = \delta(\omega - g(\theta, \epsilon))$, with $g(\theta, \epsilon) = \{\text{diag}(\epsilon_1)M_1, \text{diag}(\epsilon_2)M_2, b\}$, $\theta = \{M_1, M_2, b\}$, and $p(\epsilon_i)$ defined as a product of Bernoulli distributions ($\epsilon_i$ is a vector of draws from the Bernoulli distribution). Since we assumed $q_\theta(\omega)$ to factorise over the layers and over the rows of each weight matrix, we have

$$\text{KL}(q_\theta(\omega)||p(\omega)) = \sum_{i,k} \text{KL}(q_{\theta_{i,k}}(w_{i,k})||p(w_{i,k}))$$

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

We approximate each $q_{\theta_{i,k}}(w_{i,k}|\epsilon) = \delta(w_{i,k} - g(\theta_{i,k}, \epsilon_{i,k}))$ as a narrow Gaussian with a small standard deviation $\Sigma = \sigma^2 I$. This means that marginally $q_{\theta_{i,k}}(w_{i,k})$ 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 $p = (p_1, \ldots, p_L)$, and $\Sigma_i \in \mathbb{R}^{K \times K}$ diagonal positive-definite for $i = 1, \ldots, L$, with the elements of each $\Sigma_i$ not dependent on

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

\[ q(x) = \sum_{i=1}^{L} p_i \mathcal{N}(x; \mu_i, \Sigma_i) \]

be a mixture of Gaussians with \( L \) components and \( \mu_i \in \mathbb{R}^K \), let \( p(x) = \mathcal{N}(0, I_K) \), and further assume that \( \mu_i - \mu_j \sim \mathcal{N}(0, I) \) for all \( i, j \).

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

\[
\text{KL}(q(x) || p(x)) \approx \sum_{i=1}^{L} p_i \left( \mu_i^T \mu_i + \text{tr}(\Sigma_i) - K(1 + \log 2\pi) - \log |\Sigma_i| \right) - \mathcal{H}(p) \quad (A.1)
\]

with \( \mathcal{H}(p) := -\sum_{i=1}^{L} p_i \log p_i \) 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

\[
p(\omega) = \prod_{i=1}^{L} p(W_i) = \prod_{i=1}^{L} \mathcal{N}(W_i; 0, I/l_i^2, I)
\]

and the approximating distribution to be a dropout variational distribution.

**Proof.**

\[
\frac{\partial}{\partial m_{i,k}} \text{KL}(q_0(\omega)||p(\omega)) = \frac{\partial}{\partial m_{i,k}} \text{KL}(q_{\theta_{i,k}}(w_{i,k})||p(w_{i,k})) \\
\approx \frac{(1 - p_i)}{2} \frac{l_i^2}{\partial m_{i,k} m_{i,k}^T} \frac{\partial}{\partial m_{i,k}} \eta_1 \lambda_1 ||M_1||^2 + \lambda_2 ||M_2||^2 + \lambda_3 ||b||^2
\]

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

Next we prove proposition 4.

**Proof.** We have

\[
\text{KL}(q(x)||p(x)) = \int q(x) \log \frac{q(x)}{p(x)} dx \\
= \int q(x) \log q(x) dx - \int q(x) \log p(x) dx
\]
\[-H(q(x)) - \int q(x) \log p(x) dx \tag{A.2}\]

—a sum of the entropy of \( q(x) \) \((H(q(x))) \) and the expected log probability of \( x \). 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

\[
H(q(x)) = - \sum_{i=1}^{L} p_i \int \mathcal{N}(x; \mu_i, \Sigma_i) \log q(x) dx
\]

\[
= - \sum_{i=1}^{L} p_i \int \mathcal{N}(\epsilon_i; 0, I) \log q(\mu_i + L_i \epsilon_i) d\epsilon_i
\]

using a change of variables \( x = \mu_i + L_i \epsilon_i \), where \( L_i L_i^T = \Sigma_i \) and \( \epsilon_i \sim \mathcal{N}(0, I) \).

Now, the term inside the logarithm can be written as

\[
q(\mu_i + L_i \epsilon_i) = \sum_{j=1}^{L} p_i \mathcal{N}(\mu_i + L_i \epsilon_i; \mu_j, \Sigma_j)
\]

\[
= \sum_{j=1}^{L} p_i (2\pi)^{-K/2} |\Sigma_j|^{-1/2} \exp \left\{ - \frac{1}{2} ||\mu_j - \mu_i - L_i \epsilon_i||_{\Sigma_j}^2 \right\}
\]

where \( || \cdot ||_\Sigma \) is the Mahalanobis distance. Since \( \mu_i, \mu_j \) are assumed to be normally distributed, the quantity \( ||\mu_j - \mu_i - L_i \epsilon_i||_{\Sigma_j}^2 \) is also normally distributed. Since the expectation of a generalised \( \chi^2 \) distribution with \( K \) degrees of freedom increases with \( K \), we have that \( K \gg 0 \) implies that \( ||\mu_j - \mu_i - L_i \epsilon_i||_{\Sigma_j}^2 \gg 0 \) for \( i \neq j \) (since the elements of \( \Sigma_j \) do not depend on \( K \)). Finally, we have for \( i = j \) that \( ||\mu_i - \mu_i - L_i \epsilon_i||_{\Sigma_i}^2 = \epsilon_i^T L_i^T L_i L_i^{-1} L_i \epsilon_i = \epsilon_i^T \epsilon_i \). Therefore the last equation can be approximated as

\[
q(\mu_i + L_i \epsilon_i) \approx p_i (2\pi)^{-K/2} |\Sigma_i|^{-1/2} \exp \left\{ - \frac{1}{2} \epsilon_i^T \epsilon_i \right\}
\]

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

\[
H(q(x)) \approx - \sum_{i=1}^{L} p_i \int \mathcal{N}(\epsilon_i; 0, I) \log \left( p_i (2\pi)^{-K/2} |\Sigma_i|^{-1/2} \exp \left\{ - \frac{1}{2} \epsilon_i^T \epsilon_i \right\} \right) d\epsilon_i
\]

\[
= \sum_{i=1}^{L} \frac{p_i}{2} \left( \log |\Sigma_i| + \int \mathcal{N}(\epsilon_i; 0, I) \epsilon_i^T \epsilon_i d\epsilon_i + K \log 2\pi \right) + H(p)
\]

2With mean zero and variance \( \text{Var}(\mu_i - \mu_i - L_i \epsilon_i) = 2I + \Sigma_i \).

3To be exact, for diagonal matrices \( \Lambda, \Delta \) and \( v \sim \mathcal{N}(0, \Lambda) \), we have \( \mathbb{E}[||v||_\Delta] = \mathbb{E}[v^T \Delta^{-1} v] = \sum_{k=1}^{K} \mathbb{E}[\Delta_k^{-1} v_k^2] = \sum_{k=1}^{K} \Delta_k^{-1} \Lambda_k \).
where $\mathcal{H}(p) := -\sum_{i=1}^{L} p_i \log p_i$. Since $\mathbf{e}_i^T \mathbf{e}_i$ distributes according to a $\chi^2$ distribution, its expectation is $K$, and the entropy can be approximated as

$$\mathcal{H}(q(x)) \approx \sum_{i=1}^{L} \frac{p_i}{2} \left( \log |\Sigma_i| + K(1 + \log 2\pi) \right) + \mathcal{H}(p). \quad (A.3)$$

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

$$\int q(x) \log p(x) dx = \sum_{i=1}^{L} p_i \int \mathcal{N}(x; \mu_i, \Sigma_i) \log p(x) dx$$

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

$$\int q(x) \log p(x) dx = -\frac{1}{2} \sum_{i=1}^{L} p_i \left( \mu_i^T \mu_i + \text{tr}(\Sigma_i) \right). \quad (A.4)$$

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

$$\text{KL}(q(x)||p(x)) \approx \sum_{i=1}^{L} \frac{p_i}{2} \left( \mu_i^T \mu_i + \text{tr}(\Sigma_i) - K(1 + \log 2\pi) - \log |\Sigma_i| \right) - \mathcal{H}(p),$$

as required to show. \qed