Gaussian Mixture Kullback-Leibler Divergence

This is a reference article to help me with building the KL divergence for Gaussian mixtures. First, we demonstrate the closed-form solution tot he KL divergence between two Gaussians distributions, $p_0(x) = \mathcal{N}(x; \boldsymbol{\mu}_0, \Sigma_1)$ and $p_1(x) = \mathcal{N}(x; \boldsymbol{\mu}_1, \Sigma_1)$. Note that,

$$\mathrm{KL}(p_0 \| p_1) = \mathbb{E}\left[\log \frac{p_0(x)}{p_1(x)}\right], \tag{1}$$

where the expectation is w.r.t. x drawn from p_0. Recall that for the Gaussian distribution,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{2}$$

$$\log \frac{p_0}{p_1} = \log\left[\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right)^{1/2} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right)\right] \tag{3}$$

$$= \frac{1}{2}\left[\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + (\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) - (\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right]. \tag{4}$$

We focus on the third term within the expectation and apply the trace trick,

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0)\right] &= \mathbb{E}\left[\mathrm{tr}\left((\mathbf{x}-\boldsymbol{\mu}_0)(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_0^{-1}\right)\right] && (5)\\
&= \mathrm{tr}\left(\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_0)(\mathbf{x}-\boldsymbol{\mu}_0)^\top\right] \Sigma_0^{-1}\right) && (6)\\
&= \mathrm{tr}\left(\Sigma_0 \Sigma_0^{-1}\right) && (7)\\
&= d. && (8)
\end{aligned}$$

We now focus on the second term within the expectation (and the trace trick),

$$\begin{aligned}
\mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)\right] &= \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_0+\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_0+\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\right] && (9)\\
&= \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + (\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\right] \\
&\quad + \mathbb{E}\left[(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\right] && (10)\\
&= \mathbb{E}\left[(\mathbf{x}-\boldsymbol{\mu}_0)^\top \Sigma_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)\right] && (11)\\
&= \Sigma_0 \Sigma_1^{-1} + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1). && (12)
\end{aligned}$$

Combined, the KL divergence for two gaussians is,

$$\mathrm{KL}(p_0 \| p_1) = \frac{1}{2}\left[\log\left(\frac{|\Sigma_1|}{|\Sigma_0|}\right) + \Sigma_0 \Sigma_1^{-1} + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^\top \Sigma_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) - d\right]. \tag{13}$$

If the two distributions have diagonal covariances and $\rho = \log \sigma^2$ then,

$$\mathrm{KL}(p_0 \| p_1) = \frac{1}{2}\mathbf{1}^\top\left[\boldsymbol{\rho}_1 - \boldsymbol{\rho}_0 + \exp\left(\boldsymbol{\rho}_0 - \boldsymbol{\rho}_1\right) + (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^2 \circ \exp\left(-\boldsymbol{\rho}_1\right) - \mathbf{1}\right] \tag{14}$$

The graidents w.r.t. μ and ρ are as follows,

$$\begin{aligned}
\nabla_{\mu_0} \mathrm{KL}(p_0 \| p_1) &= (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) \circ \exp\left(-\boldsymbol{\rho}_1\right) && (15)\\
\nabla_{\boldsymbol{\rho}_0} \mathrm{KL}(p_0 \| p_1) &= \frac{1}{2}\left[-\mathbf{1} + \exp\left(\boldsymbol{\rho}_0 - \boldsymbol{\rho}_1\right)\right] && (16)\\
\nabla_{\mu_1} \mathrm{KL}(p_0 \| p_1) &= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0) \circ \exp\left(-\boldsymbol{\rho}_1\right) && (17)\\
\nabla_{\boldsymbol{\rho}_1} \mathrm{KL}(p_0 \| p_1) &= \frac{1}{2}\left[\mathbf{1} - \exp\left(\boldsymbol{\rho}_0 - \boldsymbol{\rho}_1\right) - (\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^2 \circ \exp\left(-\boldsymbol{\rho}_1\right)\right] && (18)\\
& && (19)
\end{aligned}$$

For a gaussian mixture, we deal with the very specific case where p_1 is gaussian by p_1 is a gaussian mixture. We use the variational approximation,

$$\mathrm{KL}(p_0 \| p_{1:m}) = -\log \sum_{j=1}^{m} \pi \exp(-\mathrm{KL}(p_0 \| p_j)) \tag{20}$$

$$= -\log \sum_{j=1}^{m} \pi \exp(-\mathrm{KL}_j). \tag{21}$$

To prevent underflow/overflow, we do the following,

$$\mathrm{KL}(p_0 \| p_{1:m}) = -\log \frac{\exp(\mathrm{KL}_{min})}{\exp(\mathrm{KL}_{min})} \sum_{j=1}^{m} \pi \exp(-\mathrm{KL}_j) \tag{22}$$

$$= -\left[\log \sum_{j=1}^{m} \pi \exp(\mathrm{KL}_{min} - \mathrm{KL}_j) - \log \exp(\mathrm{KL}_{min})\right] \tag{23}$$

$$= \mathrm{KL}_{min} - \log \sum_{j=1}^{m} \pi \exp(\mathrm{KL}_{min} - \mathrm{KL}_j) \tag{24}$$

We then compute gradients w.r.t. KL_j,

$$\nabla_j \mathrm{KL}(p_0 \| p_{1:m}) = \frac{\exp(-\mathrm{KL}_j)}{\sum \exp(-\mathrm{KL}_j)} \tag{25}$$

$$= \frac{\exp(\mathrm{KL}_{min} - \mathrm{KL}_j)}{\sum \exp(\mathrm{KL}_{min} - \mathrm{KL}_j)} \tag{26}$$

The gradients w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$ are as follows for $j \neq 0$,

$$\nabla_{\boldsymbol{\mu}_j} \mathrm{KL} = (\nabla_j \mathrm{KL}) \, (\boldsymbol{\mu}_j - \boldsymbol{\mu}_0) \circ \exp\left(-\boldsymbol{\rho}_j\right) \tag{27}$$

$$\nabla_{\boldsymbol{\rho}_j} \mathrm{KL} = (\nabla_j \mathrm{KL}) \, \frac{1}{2} \left[\mathbf{1} - \exp\left(\boldsymbol{\rho}_0 - \boldsymbol{\rho}_j\right) - (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_j)^2 \circ \exp\left(-\boldsymbol{\rho}_j\right)\right]. \tag{28}$$

The gradients w.r.t. $\boldsymbol{\mu}_0$ and $\boldsymbol{\rho}_0$ are,

$$\nabla_{\boldsymbol{\mu}_0} \mathrm{KL} = \sum_j (\nabla_j \mathrm{KL}) \, (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_j) \circ \exp\left(-\boldsymbol{\rho}_j\right) \tag{29}$$

$$\nabla_{\boldsymbol{\rho}_0} \mathrm{KL} = \sum_j (\nabla_j \mathrm{KL}) \, \frac{1}{2} \left[-\mathbf{1} + \exp\left(\boldsymbol{\rho}_0 - \boldsymbol{\rho}_j\right)\right] \tag{30}$$