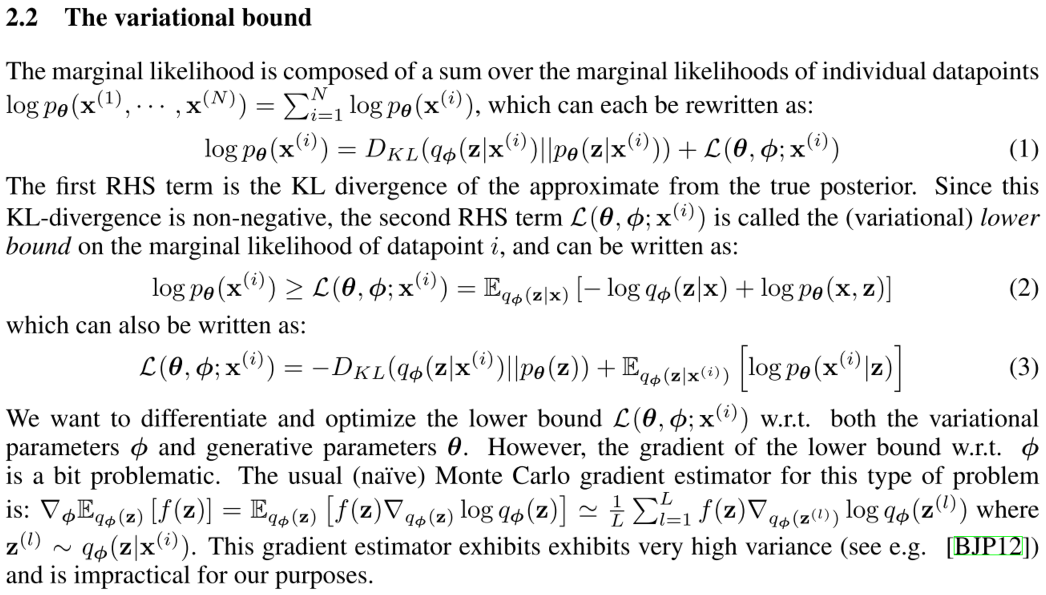
Generalized Evidence Lower Bound (ELBO) using Coupled Entropy

The purpose of this notebook is derive the generalization of the Evidence Lower Bound (ELBO), also known as the Variational Lower Bound, using two forms of the Coupled Entropy, with and without a root term for the coupled logarithm. Furthermore, the notebook will show that the ELBO can be translated into the probability domain, so that the ELBO for an individual sample is the same for all generalizations, while the averages correspond to different generalized means of the individuals.

ELBO or Variational Bound using Shannon Entropy

From (Kingma, Welling, 2014 and (Kingma, Welling, 2019) we have the following derivation and justification for the ELBO or Variational Bound.



2.2 The variational bound

The marginal likelihood is composed of a sum over the marginal likelihoods of individual datapoints $\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$, which can each be rewritten as:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \tag{1}$$

The first RHS term is the KL divergence of the approximate from the true posterior. Since this KL-divergence is non-negative, the second RHS term $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ is called the (variational) *lower bound* on the marginal likelihood of datapoint i, and can be written as:

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right] \tag{2}$$

which can also be written as:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \tag{3}$$

We want to differentiate and optimize the lower bound $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$ w.r.t. both the variational parameters $\boldsymbol{\phi}$ and generative parameters $\boldsymbol{\theta}$. However, the gradient of the lower bound w.r.t. $\boldsymbol{\phi}$ is a bit problematic. The usual (naïve) Monte Carlo gradient estimator for this type of problem is: $\nabla_{\boldsymbol{\phi}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[f(\mathbf{z})\right] = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[f(\mathbf{z})\nabla_{q_{\boldsymbol{\phi}}(\mathbf{z})} \log q_{\boldsymbol{\phi}}(\mathbf{z})\right] \simeq \frac{1}{L}\sum_{l=1}^{L} f(\mathbf{z})\nabla_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(l)})} \log q_{\boldsymbol{\phi}}(\mathbf{z}^{(l)})$ where $\mathbf{z}^{(l)} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})$. This gradient estimator exhibits exhibits very high variance (see e.g. [BJP12]) and is impractical for our purposes.

Generalized ELBO with Coupled Entropy without root

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \kappa, \alpha, d; \boldsymbol{x}^{(i)}\right) &= - D_{\kappa}\left(q_{\boldsymbol{\phi}}^{\alpha/(1+d\,\kappa)}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) || \, p_{\boldsymbol{\theta}}^{\alpha/(1+d\,\kappa)}(\boldsymbol{z})\right) + \mathbb{E}_{q_{\boldsymbol{\phi}}^{(\alpha,d,\kappa)}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}\left[\log_{\kappa} p_{\boldsymbol{\theta}}^{\alpha/(1+d\,\kappa)}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right] \\ &= \frac{-1}{2\,\kappa} \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\,\kappa}{1+d\,\kappa}} \left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{\frac{-\alpha\,\kappa}{1+d\,\kappa}} - p_{\boldsymbol{\theta}}(\boldsymbol{z})^{\frac{-\alpha\,\kappa}{1+d\,\kappa}}\right) d z \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\,\kappa}{1+d\,\kappa}} d z \, + \\ &\quad \frac{-1}{2\,\kappa} \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\,\kappa}{1+d\,\kappa}} \left(p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)^{\frac{-\alpha\,\kappa}{1+d\,\kappa}} - 1\right) d z \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\,\kappa}{1+d\,\kappa}} d z\end{aligned}$$

Care is needed with the negative signs in this expression for the ELBO. In particular the relationship $\log_{\kappa,\alpha,d} p(\boldsymbol{x}) = \frac{1}{\alpha\,\kappa}\left(p(\boldsymbol{x})^{\frac{\alpha\,\kappa}{1+d\,\kappa}} - 1\right) = \frac{-1}{\alpha\,\kappa}\left(p(\boldsymbol{x})^{\frac{-\alpha\,\kappa}{1+d\,\kappa}} - 1\right)$. The cross-entropy expression from which the coupled entropy and coupled divergence can be derived has the form

$H_{\kappa,\alpha,d}(p, q) = -\mathbb{E}_{p,\kappa,\alpha,d}[\log_{\kappa,\alpha,d} q] = \mathbb{E}_{p,\kappa,\alpha,d}[\log_{\kappa,\alpha,d} q^{-1}]$. To facilitate further analysis the expressions for the generalized ELBO were written so that $\frac{-1}{2\kappa}$is the multiple for both terms and the probabilities are raised to the power $\frac{-\alpha\kappa}{1+d\kappa}$.

The generalized ELBO is going to be optimized for different values of κ with $\alpha = 2$ and $d = 1$. To facilitate a comparison across κ, the single sample expression can be translated to the probability domain by applying the inverse of the coupled logarithm expression.
Given $x = \frac{1}{\alpha\kappa}\left(y^{\frac{\alpha\kappa}{1+d\kappa}} - 1\right)$, the inverse is then $y = (1 + \alpha\kappa x)^{\frac{1+d\kappa}{\alpha\kappa}}$, which is expressed as $(\exp_\kappa(\alpha x))^{\frac{1+d\kappa}{\alpha}}$. Completing this translation to the probability domain and substituting for alpha and d gives

$$\left(\exp_\kappa\left(2\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; \kappa, 2, 1; \boldsymbol{x}^{(i)}\right)\right)\right)^{\frac{1+\kappa}{-2\kappa}} =$$

$$\left(1 + 2\kappa\left(\frac{-1}{2\kappa}\int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{\frac{-\alpha\kappa}{1+d\kappa}} - p_{\boldsymbol{\theta}}(\boldsymbol{z})^{\frac{-\alpha\kappa}{1+d\kappa}}\right)dz \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}} dz + \frac{-1}{2\kappa}\int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}\left(p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)^{\frac{-\alpha\kappa}{1+d\kappa}} - 1\right)dz \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}} dz\right)\right)^{\frac{1+\kappa}{-2\kappa}}$$

$$= \left(1 - \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{\frac{-\alpha\kappa}{1+d\kappa}} - p_{\boldsymbol{\theta}}(\boldsymbol{z})^{\frac{-\alpha\kappa}{1+d\kappa}}\right)dz \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}} dz - \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}}\left(p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)^{\frac{-\alpha\kappa}{1+d\kappa}} - 1\right)dz \Big/ \int_{\boldsymbol{z}\in\mathbb{R}^d} q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)^{1+\frac{\alpha\kappa}{1+d\kappa}} dz\right)^{\frac{1+\kappa}{-2\kappa}}$$

Since this translation has some difficulties first examine the case with $\kappa = 0$

$$\mathrm{Exp}\left[\mathcal{L}\left(\boldsymbol{\theta}, \boldsymbol{\phi}; 0, 2, 1; \boldsymbol{x}^{(i)}\right)\right] = \mathrm{Exp}\left[-D\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}(\boldsymbol{z})\right) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right]\right]$$

$$= \mathrm{Exp}\left[-D\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right) \| p_{\boldsymbol{\theta}}(\boldsymbol{z})\right)\right]\mathrm{Exp}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right]\right]$$

While this could be expressed as a product of two geometric means in continuous domain or equivalently log-averages, its not clear that a) this further simplifies and b) that the same expression would be achieved for different values of the κ. While each data sample $\boldsymbol{x}^{(i)}$is being considered individually, there are still full distributions of q and p under examination.

Putting aside the expectation over $q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)$for each term, the expression can be arranged as
$\mathrm{Exp}\left[\log\left(p_{\boldsymbol{\theta}}(\boldsymbol{z}) / q_{\boldsymbol{\phi}}\left(\boldsymbol{z} \mid \boldsymbol{x}^{(i)}\right)\right)\right]\mathrm{Exp}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right] = \frac{p_{\boldsymbol{\theta}}(\boldsymbol{z})\, p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} \mid \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}$
$p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)$: I believe this probability is what we plotted for our evaluation
Replacing the expectation but as a power, the integral is
$\mathrm{Exp}\left[\int_{-\infty}^{\infty} \mathrm{Log}\left(\frac{p_{\boldsymbol{\theta}}(\boldsymbol{z})\, p_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} \mid \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}\right)^{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})} dz\right]$
Examining each component
$\mathrm{Exp}\left[\int_{-\infty}^{\infty} \mathrm{Log}\left(\frac{1}{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})}\right)^{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})} dz\right]$: Density average (entropy in density domain) of the latent distribution given a sample $\boldsymbol{x}^{(i)}$
$\mathrm{Exp}\left[\int_{-\infty}^{\infty} \mathrm{Log}(p_{\boldsymbol{\theta}}(\boldsymbol{z}))^{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})} dz\right]$: Inverse cross density average of the prior latent distribution over the latent distribution given a sample $\boldsymbol{x}^{(i)}$
$\mathrm{Exp}\left[\int_{-\infty}^{\infty} \mathrm{Log}\left(p_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}\right)\right)^{q_{\boldsymbol{\phi}}(\boldsymbol{z} \mid \boldsymbol{x}^{(i)})} dz\right]$: Inverse cross density average of the sample given the latent vari-

able over the latent distribution given a sample $\boldsymbol{x}^{(i)}$

If this expression can be computed for each sample, then there is the possibility of also computing the respective averages for different values of κ. The relevant values of κ would be a) those that form the -2/3rd, zero, and 1 geometric means and b) the average formed by the value of κ which was used for the optimization.