Prove true parameters of ML2P model are a minimum of VAE loss

We define the true and predicted probability of student j answering item i correctly with P_{ij} and $\hat{P}_{ij}$, respectively. The former comes from the ML2P model, and the latter is the output of a neural network. We assume the more simple case, where $\theta \sim \mathcal{N}(0, I)$, rather than $\mathcal{N}(\mu, \Sigma)$.

$$P_{ij} = \frac{1}{1 + \exp\left(-\sum_{k=1}^{K} a_{ik}\theta_{jk} + b_i\right)} \tag{1}$$

$$\hat{P}_{ij} = \frac{1}{1 + \exp\left(-\sum_{k=1}^{K} \hat{a}_{ik}(\hat{\theta}_{jk} + \varepsilon_k) + \hat{b}_i\right)} \tag{2}$$

$$\tag{3}$$

I think this should be $(\hat{\theta}_{jk} + \varepsilon_k\hat{\sigma}_k)$ throughout

Note that P_{ij} is unknown, and we instead have a response sequence $u_{:j} = (u_{1j}, \ldots, u_{nj})^\top$ with $u_{ij} = \text{Bern}(P_{ij})$. This also means that $\mathbb{E}[u_{ij}] = P_{ij}$. The variables $\hat{a}_{ik}$, $\hat{b}_i$, $\hat{\theta}_{ik}$, and are parameter estimates from the VAE. The first two are parameters in the neural network, and the ability estimates are taken from feeding responses to the encoder, i.e., $\hat{\theta}_j = \text{Encoder}(u_{:j})$. The noise $\varepsilon = (\varepsilon_k)_{1 \le k \le K} \sim \mathcal{N}(0, I)$ is introduced by the sampling operation in the VAE.

The loss function for a VAE is given by

$$\mathcal{L}(u_{:j}) = -\sum_{i=1}^{n}\left[u_{ij}\log(\hat{P}_{ij}) + (1 - u_{ij})\log(1 - \hat{P}_{ij})\right] + \mathbb{E}_{q_\alpha(\hat{\theta}|u_j)}\log\left(\frac{q_\alpha(\hat{\theta}|u_j)}{p(\theta)}\right) \tag{4}$$

$$= \mathcal{L}_{\text{REC}} + \mathcal{L}_{\text{KL}} \tag{5}$$

We write $P_{ij} = \mathbb{E}(u_{ij})$, and similarly define a new "expected" loss function:

$$\mathcal{L}_{\mathbb{E}}(P_j) = \mathbb{E}_{u_j}[\mathcal{L}(u_{:j})] \tag{6}$$

$$= -\sum_{i=1}^{n}(P_{ij}\log(\hat{P}_{ij}) + (1 - P_{ij})\log(1 - \hat{P}_{ij}) + \mathbb{E}_{q_\alpha(\hat{\theta}|u_j)}\log\left(\frac{q_\alpha(\hat{\theta}|u_j)}{p(\theta)}\right) \tag{7}$$

$$= \mathcal{L}_{\mathbb{E}[\text{REC}]} + \mathcal{L}_{\mathbb{E}[\text{KL}]} \tag{8}$$

Notice that calculation of this "expected loss" requires the unknown $P_{:j}$. But when we have large amounts of data, we can think of P_{ij} as the average value of the response u_{ij}, so using this unknown value here is justified.

Define $z_i = a_{i:} \cdot \theta_{j:} - b_i$ and $\hat{z}_i = \hat{a}_{i:} \cdot (\hat{\theta}_{j:} + \varepsilon) - \hat{b}_i$. Note that z_i is fixed, dependent on the data, and does not depend on any parameters of the neural network. $\hat{z}_:$ is the input to the final layer of the decoder, and the VAE output is $\hat{P}_{ij} = \sigma(\hat{z}_i)$, where $\sigma(\cdot)$ is the sigmoidal activation function. We compute derivatives of the expected loss function, looking individually at the reconstruction and KL terms.

$$\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i} = \frac{-1}{1 + e^{-z_i}} \cdot \frac{1}{1 + e^{\hat{z}_i}} - \frac{1}{1 + e^{z_i}} \cdot \frac{-1}{1 + e^{-\hat{z}_i}} \tag{9}$$

$$= \frac{-1}{(1 + e^{-z_i})(1 + e^{\hat{z}_i})} + \frac{1}{(1 + e^{z_i})(1 + e^{-\hat{z}_i})} \tag{10}$$

$$= \frac{-1}{(1 + e^{-a_{ik}\theta_{jk} + b_i})(1 + e^{\hat{a}_{ik}(\hat{\theta}_{kj} + \varepsilon_k) - \hat{b}_i})} + \frac{1}{(1 + e^{a_{ik}\theta_{jk} - b_i})(1 + e^{-\hat{a}_{ik}(\hat{\theta}_{jk} + \varepsilon_k) + \hat{b}_i})} \tag{11}$$

$$\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{a}_{ik}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}\frac{\partial \hat{z}_i}{\partial \hat{a}_{ik}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}(\hat{\theta}_{jk} + \varepsilon_k) \tag{12}$$

$$\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{b}_{i}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}\frac{\partial \hat{z}_i}{\partial \hat{b}_{i}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}(-1) \tag{13}$$

$$\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{\theta}_{ik}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}\frac{\partial \hat{z}_i}{\partial \hat{\theta}_{ik}} = \frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}(\hat{a}_{ik}) \tag{14}$$

Rather than setting these to zero and solving, we show that the most intuitive solution, $\hat{a}_{ik} = a_{ik}$, $\hat{b}_i = b_i$, and $\hat{\theta}_{jk} = \theta_{jk}$, is in fact a minimum of the expected loss function. But first, we must take another expectation over the random variable $\varepsilon \sim \mathcal{N}(0, I)$. Obviously, we have that $\mathbb{E}[\varepsilon_k] = 0$; this makes our calculations very simple. Notice that we have

$$\mathbb{E}_\varepsilon \left[\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{z}_i}\right]\Big|_{\hat{a}_{ik}=a_{ik},\hat{b}_i=b_i,\hat{\theta}_{jk}=\theta_{jk}} \tag{15}$$

$$= \frac{-1}{(1+e^{-a_{ik}\theta_{jk}+b_i})(1+e^{a_{ik}(\theta_{kj}+0)-b_i})} + \frac{1}{(1+e^{a_{ik}\theta_{jk}-b_i})(1+e^{-a_{ik}(\theta_{jk}+0)+b_i})} \tag{16}$$

$$= 0 \tag{17}$$

Therefore we clearly have

$$\mathbb{E}_\varepsilon \left[\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{a}_{ik}}\right]\Big|_{\hat{a}_{ik}=a_{ik},\hat{b}_i=b_i,\hat{\theta}_{jk}=\theta_{jk}} \tag{18}$$

$$= \mathbb{E}_\varepsilon \left[\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{b}_{i}}\right]\Big|_{\hat{a}_{ik}=a_{ik},\hat{b}_i=b_i,\hat{\theta}_{jk}=\theta_{jk}} \tag{19}$$

$$= \mathbb{E}_\varepsilon \left[\frac{\partial \mathcal{L}_{\mathbb{E}[\text{REC}]}}{\partial \hat{\theta}_{jk}}\right]\Big|_{\hat{a}_{ik}=a_{ik},\hat{b}_i=b_i,\hat{\theta}_{jk}=\theta_{jk}} \tag{20}$$

$$= 0 \tag{21}$$

This proves that the true parameters give a local minimum for the expected reconstruction error in the VAE loss.

We now consider the Kullback-Leibler divergence term in the expected loss function. Again assuming independent latent traits, we have

$$\mathcal{L}_{KL} = \mathbb{E}_{q(\theta|u)}[\log\left(\frac{q(\hat{\theta}|u)}{p(\theta)}\right) = KL(q(\hat{\theta}|u)||p(\theta)) = -\frac{1}{2}\sum_{k=1}^{K}(1 + \log(\hat{\sigma}_k^2) - \hat{\theta}_k^2 - \hat{\sigma}_k^2) \tag{22}$$

It is clear that this regularization term is minimized (and equal to zero) when $\hat{\theta}_{jk} = 0$ and $\hat{\sigma}_{jk} = 1$. But what happens when we plug in the "true" student ability values as before? We have

$$\mathcal{L}_{KL}\Big|_{\hat{\theta}=\theta,\hat{\sigma}=\sigma} = KL(p(\theta|u)||p(\theta)) \tag{23}$$

Notice that this is the KL divergence between the **true posterior** $p(\theta|u)$ and the **prior** $p(\theta)$. This is interpreted as the average difference of number of bits required to encode samples of $p(\theta|u)$ using a code optimized for $p(\theta)$, rather than one optimized for $p(\theta|u)$. We should be okay with accepting this loss, since the true posterior is not actually known, and we are just using the prior as a reference.

****

TODO: try to show that this is a global minimum. Also take derivatives of the KL loss w.r.t $\hat{\theta}_{jk}$. The Q-matrix may help with an identifiability issue (existence of other local minimums) in solving the system $(a_{ik}\theta_{jk} + b_i)_{jk} = z_i$. The Q-matrix *may* make the solution unique.

****