The Binomial likelihood equation of the HLSM is:

$$L = \prod_k \prod_i \prod_{j<i} \sigma(\eta_{ijk})^{y_{ijk}} (1 - \sigma(\eta_{ijk}))^{1-y_{ijk}}$$

where,

$$\eta_{ijk} = \alpha_k - \|z_{ik} - z_{jk}\|^2 = \alpha_k - \|b_i + \epsilon_{ik} - b_j - \epsilon_{jk}\|^2$$

$$z_{ik} = b_i + \epsilon_{ik}$$

$$\sigma(\eta_{ijk}) = \frac{1}{1 + \exp(-\eta_{ijk})}$$

Then, the log likelihood is

$$\begin{aligned}
\ell &= \sum_k \sum_i \sum_{j<i} y_{ijk} \ln \sigma(\eta_{ijk}) + (1 - y_{ijk}) \ln(1 - \sigma(\eta_{ijk})) \\
&= \sum_k \sum_i \sum_{j<i} -y_{ijk} \ln(1 + \exp(-\eta_{ijk})) + (1 - y_{ijk}) \ln\left(\frac{\exp(-\eta_{ijk})}{1 + \exp(-\eta_{ijk})}\right) \\
&= \sum_k \sum_i \sum_{j<i} -y_{ijk} \ln(1 + \exp(-\eta_{ijk})) + (1 - y_{ijk})[-\eta_{ijk} - \ln(1 + \exp(-\eta_{ijk}))] \\
&= \sum_k \sum_i \sum_{j<i} -y_{ijk} \ln(1 + \exp(-\eta_{ijk})) - \eta_{ijk} - \ln(1 + \exp(-\eta_{ijk})) + y_{ijk}\eta_{ijk} + y_{ijk} \ln(1 + \exp(-\eta_{ijk}))] \\
&= \sum_k \sum_i \sum_{j<i} (y_{ijk} - 1)\eta_{ijk} - \ln(1 + \exp(-\eta_{ijk}))
\end{aligned}$$

We define our decision variables as: $\{b_i\}\ \forall\ i$ and $\{\epsilon_{ik}\}\ \forall\ i, k$.
Then, the gradients of the η_{ijk} are:

$$\frac{\partial \eta_{ijk}}{\partial b_i} = -2(b_i + \epsilon_{ik} - b_j - \epsilon_{jk})$$

$$\frac{\partial \eta_{ijk}}{\partial b_j} = 2(b_i + \epsilon_{ik} - b_j - \epsilon_{jk})$$

$$\frac{\partial \eta_{ijk}}{\partial \epsilon_{ik}} = -2(b_i + \epsilon_{ik} - b_j - \epsilon_{jk})$$

$$\frac{\partial \eta_{ijk}}{\partial \epsilon_{jk}} = 2(b_i + \epsilon_{ik} - b_j - \epsilon_{jk})$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial b_m} &= \sum_k \sum_i \sum_{j<i} (y_{ijk}-1)\frac{\partial \eta_{ijk}}{\partial b_m} - \frac{\partial \ln(1+\exp(-\eta_{ijk}))}{\partial b_m} \\
&= \sum_k \left[\sum_{j<m} (y_{mjk}-1)\frac{\partial \eta_{mjk}}{\partial b_m} - \frac{\partial \ln(1+\exp(-\eta_{mjk}))}{\partial b_m} + \sum_{i>m} (y_{imk}-1)\frac{\partial \eta_{imk}}{\partial b_m} - \frac{\partial \ln(1+\exp(-\eta_{imk}))}{\partial b_m} \right] \\
&= \sum_k \Bigg[-2 \sum_{j<m} \left[(y_{mjk}-1)(b_m + \epsilon_{mk} - b_j + \epsilon_{jk}) + \frac{\exp(-\eta_{mjk})}{1+\exp(-\eta_{mjk})}(b_m + \epsilon_{mk} - b_j - \epsilon_{jk}) \right] \\
&\qquad + 2 \sum_{i>m} \left[(y_{imk}-1)(b_i + \epsilon_{ik} - b_m - \epsilon_{mk}) + \frac{\exp(-\eta_{ijk})}{1+\exp(-\eta_{mjk})}(b_i + \epsilon_{ik} - b_m - \epsilon_{mk}) \right] \Bigg] \\
&= 2 \sum_k \Bigg[- \sum_{j<m} \left(y_{mjk} - 1 + \frac{\exp(-\eta_{mjk})}{1+\exp(-\eta_{mjk})} \right)(z_{mk} - z_{jk}) \\
&\qquad + \sum_{i>m} \left(y_{imk} - 1 + \frac{\exp(-\eta_{ijk})}{1+\exp(-\eta_{mjk})} \right)(z_{ik} - z_{mk}) \Bigg]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \ell}{\partial \epsilon_{mk}} = 2 \Bigg[&- \sum_{j<m} \left(y_{mjk} - 1 + \frac{\exp(-\eta_{mjk})}{1+\exp(-\eta_{mjk})} \right)(z_{mk} - z_{jk}) \\
&+ \sum_{i>m} \left(y_{imk} - 1 + \frac{\exp(-\eta_{ijk})}{1+\exp(-\eta_{mjk})} \right)(z_{ik} - z_{mk}) \Bigg]
\end{aligned}
$$

Finally, our complete objective function is to minimize the sum between the negative log likelihood and a lasso penalty on the deviations ϵ_{ik}.

$$
\min_{\epsilon, b} \sum_k \sum_i \sum_{j<i} [(1 - y_{ijk})\eta_{ijk} + \ln(1 + \exp(\eta_{ijk}))] + \lambda \sum_k \sum_i \|\epsilon_{i,k}\|
$$

where $\|.\|$ represents the L2 norm. Since this is not differentiable, we could use proximal gradient descent. If the gradients for the ML function derived above are correct, we can directly use the code we implemented in HW3 (problem 4) to implement the proximal operator for this problem.

The proximal operator for the penalty term using the L2 norm[1] is:

$$
\text{prox}_{\|.\|,t}(\epsilon_{i,k}) = \begin{cases} \frac{\|\epsilon_{i,k}\| - \lambda t}{\|\epsilon_{i,k}\|}\epsilon_{i,k}, & \|\epsilon_{i,k}\| \geq \lambda t \\ 0, & \|\epsilon_{i,k}\| < \lambda t \end{cases}
$$

[1] According to the results on Homework 3

$$\text{prox}_{\|.\|,t}(b_i) = b_i$$

Therefore, let β be the vector with all parameters ϵ and b. The Proximal Gradient Descent method in this case is:

$$\beta^{+} = \text{prox}_{\|.\|,t}(\beta - t\nabla \ell(\beta))$$