$$p(Y, \boldsymbol{y}|X) = p(Y|X)p(\boldsymbol{y}|X, Y)$$

$$\begin{aligned} p(\boldsymbol{y}|X, Y) &= p(\boldsymbol{y}|X) \text{ assuming } \boldsymbol{y} \text{ is conditionally independent of } Y \text{ given } X \\ &= p(y_1, y_2, ..., y_N|X) \\ &= p(y_1|X)p(y_2|X, y_1)...p(y_N|X, y_1, y_2, ..., y_N) \end{aligned}$$

In general, we can assume that y_j is conditionally independent of $\{x_i\}_{i \neq j}$ and $\{y_i\}_{i \neq j}$ given x_j

$$p(y_j|X, y_1, ..., y_{j-1}) = p(y_j|x_j)$$

Then, the probability $p(\boldsymbol{y}|X, Y)$ can be written as:

$$\begin{aligned} p(\boldsymbol{y}|X, Y) &= p(y_1|x_1)p(y_2|x_2)...p(y_N|x_N) \\ &= \prod_{i=1}^{N} p(y_i|x_i) \end{aligned}$$

Therefore the joint probability $p(Y, \boldsymbol{y}|X)$ can be written as:

$$p(Y, \boldsymbol{y}|X) = p(Y|X) \prod_{i=1}^{N} p(y_i|x_i)$$

With this formulation, our objective is to maximize the likelihood (is this correct?):

$$\max_{\theta} p(Y|X) \prod_{i=1}^{N} p(y_i|x_i)$$

Assuming a uniform distribution over the instance labels, i.e. $p(y_i|x_i) \sim \text{uniform}\{0, 1\}$. This implies $p(y_i = 1|x_i) = p(y_i = 0|x_i) = 0.5$, then

$$\begin{aligned} p(Y, \boldsymbol{y}|X) &= p(Y|X) \prod_{i=1}^{N} p(y_i|x_i) \quad (1) \\ &= p(Y|X) \prod_{i=1}^{N} 0.5 \\ &= p(Y|X) 0.5^N \\ &= 0.5^N * p(Y|X) \end{aligned}$$

Thus, maximising the likelihood with such an assumed instance-level distribution amounts to simply maximising the following:

$$p(Y, \boldsymbol{y}|X) = p(Y|X)$$

This is actually what one does in a standard MIL setup, including plain DeepRC.

On the other hand, if one has some information about individual $p(y_i|x_i)$'s for the different instances or a subset thereof, then one can include it in equation 1.

In this case, maximising the likelihood 1 can be done by minimising the negative logarithm of it as follows:

$$
\begin{aligned}
&\max_{\Theta} p(Y, \boldsymbol{y}|X) \\
&\equiv \max_{\Theta} p(Y|X) \prod_{i=1}^{N} p(y_i|x_i) \\
&\equiv \min_{\Theta} \; -\log[p(Y|X) \prod_{i=1}^{N} p(y_i|x_i)] \\
&\equiv \min_{\Theta} \; -\left[\log p(Y|X) + \log \prod_{i=1}^{N} p(y_i|x_i)\right] \\
&\equiv \min_{\Theta} \; -\left[\log p_{\Theta}(Y|X) + \sum_{i=1}^{N} \log p(z_i|m_i)\right]
\end{aligned}
$$

Now, assuming both conditional probabilities $p(Y|X)$ and $p(y|x_i)$ are binomial distributions with some labels, then one could apply a CE loss on each, which is basically what we have done until now.