Exploiting Human-AI Dependency for Learning to Defer

Anonymous Authors[1]

In our paper, we proposed a surrogate loss called DCE for the vanilla Learning to Defer (L2D) setting. Here, we further provide an extension for the proposed DCE loss to the setting of L2D with deferral costs and L2D with multiple experts.

1. L2D with Deferral Costs

In this part, we would like to introduce the extension of DCE with deferral **c**ost (DCEc). In L2D with deferral cost $c(\boldsymbol{x})$. The performance of an L2D system could be formulated as follows:

$$L_{01c}^{\perp}(f(\boldsymbol{x}), y, m) = \mathbb{I}_{f(\boldsymbol{x})\neq y}\mathbb{I}_{f(\boldsymbol{x})\neq\perp} + (\mathbb{I}_{m\neq y} + c(\boldsymbol{x}))\mathbb{I}_{f(\boldsymbol{x})=\perp},$$

Previous studies showed that the Bayes optimal classifier for L2D with deferral cost $c(\boldsymbol{x})$ can be formulated as follows:

$$f^*(x) = \begin{cases} \perp, & \mathbb{P}(Y = M|\boldsymbol{x}) > c(\boldsymbol{x}) + \max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

Then, the generalized version of DCE for L2D with deferral costs could be defined as:

$$\begin{aligned} L_{\mathrm{DCEc}}^{\perp}(g(\boldsymbol{x}), y, m) = &-\mathbb{I}_{m\neq y}\log(\psi_{\mathcal{Y}^\perp}^{y}(g(\boldsymbol{x}))) \\ &- \mathbb{I}_{m=y}\big(\log(\psi_{\mathcal{Y}}^{y}(g(\boldsymbol{x}))) + \log(\psi_{\mathcal{Y}^\perp/q}^{\perp}(g(\boldsymbol{x})))\big) \\ &- c(\boldsymbol{x})\log(\psi_{\mathcal{Y}^\perp}^{q}(g(\boldsymbol{x}))), \end{aligned}$$

Where $q = \arg\max_{y\in\mathcal{Y}} g_y(\boldsymbol{x})$ denotes the predicted class of the model.

The consistency proof for DCEc is provided in Appendix A.

2. L2D with Multiple Experts

In this part, we show how to extend the DCE loss to L2D with multiple experts. We call this extended version DCEm.

L2D with multiple experts could be formulated as minimizing the following loss function

$$\begin{aligned} L_{01m}^{\perp}(f(\boldsymbol{x}), y, m_1, m_2, \ldots, m_J) = &\mathbb{I}_{f(\boldsymbol{x})\neq y}\mathbb{I}_{f(\boldsymbol{x})\in\mathcal{Y}} \\ &+ \sum_{i=1}^{J}(\mathbb{I}_{m_i\neq y})\mathbb{I}_{f(\boldsymbol{x})=\perp,i}, \end{aligned}$$

Where J is used to denote the number of experts. m_i denotes the prediction made by the i-th expert. When $f(\boldsymbol{x}) = \perp, i$, the L2D system would defer the prediction to the i-th expert. Previous studies have shown that Bayes optimal classifier for L2D with multiple experts should meet the following condition:

$$f^*(x) = \begin{cases} \perp, i, & \mathbb{P}(Y = M_t|\boldsymbol{x}) > \max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise.} \end{cases}$$

Where $t = \arg\max_{0\leq i\leq J} \mathbb{P}(Y = M_i|\boldsymbol{x})$

Then the extended DCE for L2D with multiple experts could be formulated as:

$$\begin{aligned} L_{DCEm}^{\perp}(g(\boldsymbol{x}), y, m_1, m_2, \ldots, m_J) = &\\ &- \log(\psi_{\mathcal{Y}}^{y}(g(\boldsymbol{x}))) - \log(\psi_{\mathcal{Y}\cup\{i|\mathbb{I}[m_i\neq y]\}}^{\mathcal{Y}}(g(\boldsymbol{x}))) \\ - \sum_{i=1}^{J}\mathbb{I}[m_i = y]&\log(\psi_{\mathcal{Y}\cup\perp,i/q}^{\perp,i}g(\boldsymbol{x})) \end{aligned}$$

It is noteworthy that when $J = 1$ and $m \neq y$.

$$\begin{aligned} L_{DCE}^{\perp}(g(x), y, m) &= -\log(\psi_{\mathcal{Y}}^{y}(g(x))) - \log(\psi_{\mathcal{Y}\cup\perp}^{\mathcal{Y}}(g(x))) \\ &= -\log(\psi_{\mathcal{Y}\cup\perp}^{y}(g(x))), \end{aligned}$$

which is identical to the single expert version.

This generalized version of surrogate loss holds consistency for L2D with multiple experts. The proof is provided in Appendix B.

References

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

A. Consistency Proof of DCEc

To begin with proof, let us define

$$C^{\perp}_{\mathrm{DCEc}}(g(\boldsymbol{x})) = \sum_{y\in\mathcal{Y}} \Big[-\mathbb{P}(Y=y, M=y|\boldsymbol{x})\big(\log(\psi^{y}_{\mathcal{Y}}(g(\boldsymbol{x}))) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/q}(g(\boldsymbol{x})))\big) \\ -(\mathbb{P}(Y=y, M\neq y|\boldsymbol{x}) + c(\boldsymbol{x}))\log(\psi^{y}_{\mathcal{Y}}(g(\boldsymbol{x})))\Big]$$

The conditional surrogate risk *w.r.t.* $\boldsymbol{x}$.

Let $g^{\star}$ be the optimal scoring function for DCEc.

We first proof that $\arg\max_{\tilde{y}\in\mathcal{Y}} g^{*}_{\tilde{y}}(\boldsymbol{x}) = \arg\max_{\tilde{y}\in\mathcal{Y}} \eta_{\tilde{y}}(\boldsymbol{x})$ by contradiction. Let us use $r = \arg\max_{\tilde{y}\in\mathcal{Y}} \eta_{\tilde{y}}(\boldsymbol{x})$, $q = \arg\max_{\tilde{y}\in\mathcal{Y}} g^{*}_{\tilde{y}}(\boldsymbol{x})$ to denote the index of maximum dimension in posterior distribution and scoring function respectively. For simplicity in notation, we represent the score vector outputted by the scoring function as $\boldsymbol{s} = g(\boldsymbol{x})$, and $\boldsymbol{s}^{*} = g^{*}(\boldsymbol{x})$

Suppose $r \neq q$, i.e. $\boldsymbol{s}^{*}_{q} > \boldsymbol{s}^{*}_{r}$. Then we show that we could obtain a lower value of conditional surrogate risk by switching the value between $\boldsymbol{s}^{*}_{q}$ and $\boldsymbol{s}^{*}_{r}$. Let us define $\tilde{\boldsymbol{s}}$ the score vector obtained by switching the value between $\boldsymbol{s}^{*}_{q}$ and $\boldsymbol{s}^{*}_{r}$, i.e. $\tilde{\boldsymbol{s}}_{r} = s^{*}_{q}$, $\tilde{\boldsymbol{s}}_{q} = s^{*}_{r}$ and $\tilde{\boldsymbol{s}}_{i} = s^{*}_{i}$ for $i \in \mathcal{Y}^{\perp}, i \neq r, q$. Thus $\arg\max_{\tilde{y}\in\mathcal{Y}} \tilde{\boldsymbol{s}} = r$ Then $C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}^{*}) - C^{\perp}_{\mathrm{DCEc}}(\tilde{\boldsymbol{s}})$ could be expressed as:

$$\begin{aligned}
&C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}^{*}) - C^{\perp}_{\mathrm{DCEc}}(\tilde{\boldsymbol{s}}) \\
&= \mathbb{P}(Y=q, M=q|\boldsymbol{x})\big(\log(\psi^{q}_{\mathcal{Y}}(\tilde{\boldsymbol{s}})) - \log(\psi^{q}_{\mathcal{Y}}(\boldsymbol{s}^{*})) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\tilde{\boldsymbol{s}})) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/q}(\boldsymbol{s}^{*}))\big) \\
&+ \mathbb{P}(Y=q, M\neq q|\boldsymbol{x})\big(\log(\psi^{q}_{\mathcal{Y}^{\perp}}(\tilde{\boldsymbol{s}})) - \log(\psi^{q}_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^{*}))\big) \\
&+ \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\psi^{r}_{\mathcal{Y}}(\tilde{\boldsymbol{s}})) - \log(\psi^{r}_{\mathcal{Y}}(\boldsymbol{s}^{*})) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/}(\tilde{\boldsymbol{s}})) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/q}(\boldsymbol{s}^{*}))\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\psi^{r}_{\mathcal{Y}^{\perp}}(\tilde{\boldsymbol{s}})) - \log(\psi^{r}_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^{*}))\big) \\
&\overset{(a)}{=} \mathbb{P}(Y=q, M=q|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^{*}_{r})}{\exp(\boldsymbol{s}^{*}_{q})}) + \log(\frac{\exp(\boldsymbol{s}^{*}_{r}) + \sum_{i\neq r,q}\exp(\boldsymbol{s}^{*}_{i})}{\exp(\boldsymbol{s}^{*}_{r}) + \sum_{i\neq r,q}\exp(\boldsymbol{s}^{*}_{i}))}\big) \\
&+ \mathbb{P}(Y=q, M\neq q|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^{*}_{r})}{\exp(\boldsymbol{s}^{*}_{q})})\big) \\
&+ \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^{*}_{q})}{\exp(\boldsymbol{s}^{*}_{r})}) + \log(\frac{\exp(\boldsymbol{s}^{*}_{r}) + \sum_{i\neq r,q}\exp(\boldsymbol{s}^{*}_{i})}{\exp(\boldsymbol{s}^{*}_{r}) + \sum_{i\neq r,q}\exp(\boldsymbol{s}^{*}_{i})}\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^{*}_{q})}{\exp(\boldsymbol{s}^{*}_{r})})\big) \\
&= (\mathbb{P}(Y=r|\boldsymbol{x}) - \mathbb{P}(Y=q|\boldsymbol{x}))\log(\frac{\exp(\boldsymbol{s}^{*}_{q})}{\exp(\boldsymbol{s}^{*}_{r})}) \\
&> 0
\end{aligned}$$

The equation (a) holds since $\sum_{i\in\mathcal{Y}}\exp(\boldsymbol{s}^{*}_{i}) = \sum_{i\in\mathcal{Y}}\exp(\tilde{\boldsymbol{s}}_{i})$ and $\sum_{i\in\mathcal{Y}^{\perp}}\exp(\boldsymbol{s}^{*}_{i}) = \sum_{i\in\mathcal{Y}^{\perp}}\exp(\tilde{\boldsymbol{s}}_{i})$. Because $\mathbb{P}(Y = r|\boldsymbol{x}) > \mathbb{P}(Y=q|\boldsymbol{x})$ and $\boldsymbol{s}^{*}_{q} > \boldsymbol{s}^{*}_{r}$, then $(\mathbb{P}(Y=r|\boldsymbol{x}) - \mathbb{P}(Y=q|\boldsymbol{x}))\log(\frac{\exp(\boldsymbol{s}^{*}_{q})}{\exp(\boldsymbol{s}^{*}_{r})}) > 0$,$C^{\perp}_{\mathrm{DCE}}(\boldsymbol{s}^{*}) > C^{\perp}_{\mathrm{DCE}}(\tilde{\boldsymbol{s}})$. This contradicts to $\boldsymbol{s}^{*}$ is the minimizer of $C^{\perp}_{\mathrm{DCE}}$, thus $r = q$. Then we can conclude the proof of $r = q$.

Then we prove that $\boldsymbol{s}^{*}_{r} > \boldsymbol{s}^{*}_{\perp}$ when $\mathbb{P}(Y=r, Y\neq M|\boldsymbol{x}) + c(\boldsymbol{x}) > \mathbb{P}(Y\neq r, Y=M|\boldsymbol{x})$. Suppose $\boldsymbol{s}^{*}_{\perp} > \boldsymbol{s}^{*}_{r}$ when $\mathbb{P}(Y=r, M\neq r|\boldsymbol{x}) + c(\boldsymbol{x}) > \mathbb{P}(Y\neq r, Y=M|\boldsymbol{x})$ and we use $\boldsymbol{s}'$ to represent the score vector obtained by switching the value between $\boldsymbol{s}^{*}_{r}$ and $\boldsymbol{s}^{*}_{\perp}$, i.e. $\boldsymbol{s}'_{r} = \boldsymbol{s}^{*}_{\perp}$, $\boldsymbol{s}'_{\perp} = \boldsymbol{s}^{*}_{r}$ and $\boldsymbol{s}'_{i} = \boldsymbol{s}^{*}_{i}$ for all $i \in \mathcal{Y}, i \neq r$. We can directly derive that

$r = \arg\max_{\tilde{y}\in\mathcal{Y}} \boldsymbol{s}'$. Then $C^{\perp}_{\mathrm{DCE}}(\boldsymbol{s}^*) - C^{\perp}_{\mathrm{DCE}}(\boldsymbol{s}')$ could be expressed as:

$$
\begin{aligned}
&C^{\perp}_{\mathrm{DCcE}}(\boldsymbol{s}^*) - C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}') \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M\neq i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ c(\boldsymbol{x})(\log(\psi^q_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^q_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^{\star}))) \\
&= \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ c(\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^{\star}_{\perp})}{\exp(\boldsymbol{s}^{\star}_r)}) \\
&= \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\frac{\sum_{j\in\mathcal{Y}, j\neq r}\exp(\boldsymbol{s}^*_j)+\exp(\boldsymbol{s}^*_r)}{\sum_{j\in\mathcal{Y}, j\neq r}\exp(\boldsymbol{s}^*_j)+\exp(\boldsymbol{s}^*_{\perp})}) + \log(\frac{\exp(\boldsymbol{s}^*_r)}{\exp(\boldsymbol{s}^*_{\perp})}\frac{\sum_{j\in\mathcal{Y}, j\neq r}\exp(\boldsymbol{s}^*_j)+\exp(\boldsymbol{s}^*_{\perp})}{\sum_{j\in\mathcal{Y}, j\neq r}\exp(\boldsymbol{s}^*_j)+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ c(\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^{\star}_{\perp})}{\exp(\boldsymbol{s}^{\star}_r)}) \\
&= \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&- \mathbb{P}(Y\neq r, M=Y|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ c(\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^{\star}_{\perp})}{\exp(\boldsymbol{s}^{\star}_r)}) \\
&= \big(\mathbb{P}(Y=r, M\neq r|\boldsymbol{x}) + c(\boldsymbol{x}) - \mathbb{P}(Y\neq r, M=Y|\boldsymbol{x})\big)\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) > 0
\end{aligned}
$$

Thus $C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}^*) > C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}')$, which is contradictory to $\boldsymbol{s}^*$ is a minimizer of $C^{\perp}_{\mathrm{DCEc}}$, then we conclude the proof that $\boldsymbol{s}^*_r > \boldsymbol{s}^*_{\perp}$ if $\mathbb{P}(Y=r, Y\neq M|\boldsymbol{x}) + c(\boldsymbol{x}) > \mathbb{P}(Y\neq r, Y=M|\boldsymbol{x})$.

Lastly, we prove that $\boldsymbol{s}^*_{\perp} > \boldsymbol{s}^*_r$ when $\mathbb{P}(Y\neq r, Y=M|\boldsymbol{x}) > \mathbb{P}(Y=r, Y\neq M|\boldsymbol{x}) + c(\boldsymbol{x})$. Suppose $\boldsymbol{s}^*_r > \boldsymbol{s}^*_{\perp}$ when $\mathbb{P}(Y\neq r, Y=M|\boldsymbol{x}) > \mathbb{P}(Y=r, Y\neq M|\boldsymbol{x}) + c(\boldsymbol{x})$ and we still use $\boldsymbol{s}'$ to represent the score vector obtained by switching the value between $\boldsymbol{s}^*_r$ and $\boldsymbol{s}^*_{\perp}$, i.e. $\boldsymbol{s}'_r = \boldsymbol{s}^*_{\perp}$, $\boldsymbol{s}'_{\perp} = \boldsymbol{s}^*_r$ and $\boldsymbol{s}'_i = \boldsymbol{s}^*_i$ for all $i\in\mathcal{Y}, i\neq r$. Let $t = \arg\max_{\tilde{y}\in\mathcal{Y}} \boldsymbol{s}'_y$. We define $\epsilon = \sum_{i\in\mathcal{Y}, i\neq r}\exp(\boldsymbol{s}^*)$ and $\epsilon' = \sum_{i\in\mathcal{Y}, i\neq r}\exp(\boldsymbol{s}')$. We could derive that $\epsilon \geq \epsilon'$. Then $C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}^*) - C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}')$ could be expressed as:

$$
\begin{aligned}
&C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}^*) - C^{\perp}_{\mathrm{DCEc}}(\boldsymbol{s}') \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/t}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/t}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M\neq i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ c(\boldsymbol{x})(\log(\psi^q_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^q_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^{\star}))) \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})}) - \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon+\exp(\boldsymbol{s}^*_r)}) + \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) - \log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})})\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_i)}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})}) - \log(\frac{\exp(\boldsymbol{s}^*_i)}{\epsilon+\exp(\boldsymbol{s}^*_r)}) + \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) - \log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})})\big) \\
&+ c(\boldsymbol{x})(\log(\frac{\exp(\boldsymbol{s}'_t)}{\exp(\boldsymbol{s}^{\star}_r)} \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ \mathbb{P}(Y\neq r, M=Y|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_r)}{\exp(\boldsymbol{s}^*_{\perp})}) + \log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ c(\boldsymbol{x})(\log(\frac{\exp(\boldsymbol{s}'_t)}{\exp(\boldsymbol{s}^{\star}_r)} \\
&\geq \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ \big(\mathbb{P}(Y\neq r, M=Y|\boldsymbol{x}) - \mathbb{P}(Y=r, M\neq r|\boldsymbol{x}) - c(\boldsymbol{x})\big)\log(\frac{\exp(\boldsymbol{s}^*_r)}{\exp(\boldsymbol{s}^*_{\perp})}) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) > 0
\end{aligned}
$$

Which conclude the proof $\boldsymbol{s}^*_{\perp} > \boldsymbol{s}^*_r$ when $\mathbb{P}(Y\neq r, Y=M|\boldsymbol{x}) > \mathbb{P}(Y=r, Y\neq M|\boldsymbol{x}) + c(\boldsymbol{x})$ by contradiction.

B. Consistency Proof of DCEm

To begin with proof, let us define $L = 2^{\{1,2,\dots J\}}$ the powerset of $\{1,2,3\dots,J\}$, $L_i = \{E\in L | i\in E\}$. Let us use $\mathbb{P}(Y = r, E|x)$ to denote the probability that $Y=r$ and $\forall i\in E, m_i = r, \forall j\notin E, m_j\neq r$, for $E\in L$. $\mathcal{Y}^{\perp,E} = \mathcal{Y}\cup\{\perp, i|i\notin E\}$.

$$
\begin{aligned}
C^{\perp}_{\mathrm{DCEm}}(g(\boldsymbol{x})) = \sum_{y\in\mathcal{Y}}\sum_{E\in L}\Big[&-\mathbb{P}(Y=y, E|\boldsymbol{x})\big(\log(\psi^y_{\mathcal{Y}}(g(\boldsymbol{x}))) + \sum_{i\in E}\log(\psi^{\perp}_{\mathcal{Y}^{\perp,i}/q}(g(\boldsymbol{x})))\big) \\
&+ \log(\psi^{\mathcal{Y}}_{\mathcal{Y}^{\perp,E}})\Big]
\end{aligned}
$$

The conditional surrogate risk *w.r.t.* $\boldsymbol{x}$. We first proof that $\arg\max_{\tilde{y}\in\mathcal{Y}} g^*_{\tilde{y}}(\boldsymbol{x}) = \arg\max_{\tilde{y}\in\mathcal{Y}}\eta_{\tilde{y}}(\boldsymbol{x})$ by contradiction. Let us use $r = \arg\max_{\tilde{y}\in\mathcal{Y}}\eta_{\tilde{y}}(\boldsymbol{x})$, $q = \arg\max_{\tilde{y}\in\mathcal{Y}} g^*_{\tilde{y}}(\boldsymbol{x})$ to denote the index of maximum dimension in posterior distribution and scoring function respectively. For simplicity in notation, we represent the score vector outputted by the scoring function as $\boldsymbol{s} = g(\boldsymbol{x})$, and $\boldsymbol{s}^* = g^*(\boldsymbol{x})$

Suppose $r \neq q$, i.e. $\boldsymbol{s}^*_q > \boldsymbol{s}^*_r$. Then we show that we could obtain a lower value of conditional surrogate risk by switching the value between $\boldsymbol{s}^*_q$ and $\boldsymbol{s}^*_r$. Let us define $\tilde{\boldsymbol{s}}$ the score vector obtained by switching the value between $\boldsymbol{s}^*_q$ and $\boldsymbol{s}^*_r$, i.e. $\tilde{\boldsymbol{s}}_r = s^*_q$, $\tilde{\boldsymbol{s}}_q = s^*_r$ and $\tilde{\boldsymbol{s}}_i = s^*_i$ for $i \in \mathcal{Y}^\perp, i \neq r, q$. Thus $\arg\max_{\tilde{y} \in \mathcal{Y}} \tilde{\boldsymbol{s}} = r$ Then $C^\perp_{\mathrm{DCE}}(\boldsymbol{s}^*) - C^\perp_{\mathrm{DCE}}(\tilde{\boldsymbol{s}})$ could be expressed as:

$$
\begin{aligned}
&C^\perp_{\mathrm{DCEm}}(\boldsymbol{s}^*) - C^\perp_{\mathrm{DCEm}}(\tilde{\boldsymbol{s}}) \\
&= \sum_{E \in L_i} \mathbb{P}(Y = r, E|x)\big(\log(\psi^y_{\mathcal{Y}}(g(\boldsymbol{x}))) + \sum_{i \in E} \log(\psi^\perp_{\mathcal{Y}^{\perp,i}/q}(g(\boldsymbol{x})))\big) + \log(\psi^{\mathcal{Y}}_{\mathcal{Y}^{\perp,E}}) \\
&> 0
\end{aligned}
$$

Then we can conclude the proof of $r = q$ by contradiction.

Then we prove that $\boldsymbol{s}^*_r > \boldsymbol{s}^*_{\perp,i}$ when $\mathbb{P}(Y = r, Y \neq M_i|\boldsymbol{x}) > \mathbb{P}(Y \neq r, Y = M_i|\boldsymbol{x})$. Suppose $\boldsymbol{s}^*_{\perp,i} > \boldsymbol{s}^*_r$ when $\mathcal{Y}(Y = r, M_i \neq r)(\boldsymbol{x}) > \mathbb{P}(Y \neq r, Y = M_i|\boldsymbol{x})$ and we use $\boldsymbol{s}'$ to represent the score vector obtained by switching the value between $\boldsymbol{s}^*_r$ and $\boldsymbol{s}^*_{\perp,i}$, i.e. $\boldsymbol{s}'_r = \boldsymbol{s}^*_{\perp,i}$, $\boldsymbol{s}'_{\perp,i} = \boldsymbol{s}^*_r$ and $\boldsymbol{s}'_j = \boldsymbol{s}^*_j$ for all $j \in \mathcal{Y}, j \neq r$. We can directly derive that $r = \arg\max_{\tilde{y} \in \mathcal{Y}} \boldsymbol{s}'$. Then $C^\perp_{\mathrm{DCEm}}(\boldsymbol{s}^*) - C^\perp_{\mathrm{DCEm}}(\boldsymbol{s}')$ could be expressed as:

$$
\begin{aligned}
&C^\perp_{\mathrm{DCEm}}(\boldsymbol{s}^*) - C^\perp_{\mathrm{DCEm}}(\boldsymbol{s}') \\
&= \sum_{E \in L} -\mathbb{P}\Big[(Y = y, E|\boldsymbol{x})\big(\log(\psi^y_{\mathcal{Y}}(g(\boldsymbol{x}))) + \sum_{i \in E} \log(\psi^\perp_{\mathcal{Y}^{\perp,i}/q}(g(\boldsymbol{x})))\big) \\
&+ \log(\psi^{\mathcal{Y}}_{\mathcal{Y}^{\perp,E}})\Big]
\end{aligned}
$$

Thus $C^\perp_{\mathrm{DCE}}(\boldsymbol{s}^*) > C^\perp_{\mathrm{DCE}}(\boldsymbol{s}')$, which is contradictory to $\boldsymbol{s}^*$ is a minimizer of $C^\perp_{\mathrm{DCE}}$, then we conclude the proof that $\boldsymbol{s}^*_r > \boldsymbol{s}^*_\perp$ if $\mathbb{P}(Y = r, Y \neq M|\boldsymbol{x}) > \mathbb{P}(Y \neq r, Y = M|\boldsymbol{x})$.

Lastly, we prove that $\boldsymbol{s}^*_\perp > \boldsymbol{s}^*_r$ when $\mathbb{P}(Y \neq r, Y = M|\boldsymbol{x}) > \mathbb{P}(Y = r, Y \neq M|\boldsymbol{x})$. Suppose $\boldsymbol{s}^*_r > \boldsymbol{s}^*_\perp$ when $\mathbb{P}(Y \neq r, Y = M|\boldsymbol{x}) > \mathbb{P}(Y = r, Y \neq M|\boldsymbol{x})$ and we still use $\boldsymbol{s}'$ to represent the score vector obtained by switching the value between $\boldsymbol{s}^*_r$ and $\boldsymbol{s}^*_\perp$, i.e. $\boldsymbol{s}'_r = \boldsymbol{s}^*_\perp$, $\boldsymbol{s}'_\perp = \boldsymbol{s}^*_r$ and $\boldsymbol{s}'_i = \boldsymbol{s}^*_i$ for all $i \in \mathcal{Y}, i \neq r$. Let $t = \arg\max_{\tilde{y} \in \mathcal{Y}} \boldsymbol{s}'_y$. We define $\epsilon = \sum_{i \in \mathcal{Y}, i \neq r} \exp(\boldsymbol{s}^*)$ and $\epsilon' = \sum_{i \in \mathcal{Y}, i \neq r} \exp(\boldsymbol{s}')$. We could derive that $\epsilon \geq \epsilon'$. Then $C^\perp_{\mathrm{DCE}}(\boldsymbol{s}^*) - C^\perp_{\mathrm{DCE}}(\boldsymbol{s}')$ could be expressed as:

$$
\begin{aligned}
&C^{\perp}_{\mathrm{DCE}}(\boldsymbol{s}^*) - C^{\perp}_{\mathrm{DCE}}(\boldsymbol{s}') \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/t}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big(\log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^r_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}}(\boldsymbol{s}^*)) + \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/t}(\boldsymbol{s}')) - \log(\psi^{\perp}_{\mathcal{Y}^{\perp}/r}(\boldsymbol{s}^*))\big) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M\neq i|\boldsymbol{x})\big(\log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}')) - \log(\psi^i_{\mathcal{Y}^{\perp}}(\boldsymbol{s}^*))\big) \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})}) - \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon+\exp(\boldsymbol{s}^*_r)}) + \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) - \log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})})\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ \sum_{i\in\mathcal{Y}, i\neq r} \mathbb{P}(Y=i, M=i|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_i)}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})}) - \log(\frac{\exp(\boldsymbol{s}^*_i)}{\epsilon+\exp(\boldsymbol{s}^*_r)}) + \log(\frac{\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) - \log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\epsilon+\exp(\boldsymbol{s}^*_{\perp})})\big) \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\exp(\boldsymbol{s}^*_{\perp})}{\exp(\boldsymbol{s}^*_r)}) \\
&+ \mathbb{P}(Y\neq r, M=Y|\boldsymbol{x})\big(\log(\frac{\exp(\boldsymbol{s}^*_r)}{\exp(\boldsymbol{s}^*_{\perp})}) + \log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&= \mathbb{P}(Y=r, M=r|\boldsymbol{x})\big(\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)})\big) \\
&+ \big(\mathbb{P}(Y\neq r, M=Y|\boldsymbol{x}) - \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\big)\log(\frac{\exp(\boldsymbol{s}^*_r)}{\exp(\boldsymbol{s}^*_{\perp})}) \\
&+ \mathbb{P}(Y=r, M\neq r|\boldsymbol{x})\log(\frac{\epsilon+\exp(\boldsymbol{s}^*_r)}{\epsilon'+\exp(\boldsymbol{s}^*_r)}) > 0
\end{aligned}
$$

Which conclude the proof $\boldsymbol{s}^*_{\perp} > \boldsymbol{s}^*_r$ when $\mathbb{P}(Y\neq r, Y=M|\boldsymbol{x}) > \mathbb{P}(Y=r, Y\neq M|\boldsymbol{x})$ by contradiction.