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 follow loss function:

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

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

$$f^*(x) = \begin{cases} \perp, & \mathbb{P}(Y = M|\boldsymbol{x}) > c(\boldsymbol{x}) + \max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), \\ \arg\max\limits_{y\in\mathcal{Y}} \eta_y(\boldsymbol{x}), & \text{otherwise.} \end{cases} \tag{1}$$

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. Then we could train a model by minimize the risk $R_{DCEc}^{\perp}(g) = \mathbb{E}_{p(\boldsymbol{x},y,m)}[L_{01c}^{\perp}(g(\boldsymbol{x}), y, m)]$.

Theorem 1.1. *Let us define $g^{\star}(\boldsymbol{x}) \in \arg\min_g R_{DCEm}^{\perp}(g)$ the optimized scoring function, then $\phi(g^{\star}(\boldsymbol{x}))$ meets the condition in Eq. 1, which further means $L_{DCEc}^{\perp}$ is a consistent loss for L2D with deferral costs.*

The proof of Theorem 1.1 is provided in Appedix A.

[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.

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

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

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

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

where $t = \arg\max_{0\leq j\leq J} \mathbb{P}(Y = M_j|\boldsymbol{x})$.

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

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

Where $\mathcal{Y}^{\perp_j} = \mathcal{Y}\cup\{\perp_j\}$. However, L_{DCEm} is not a consistent loss for L2D with multiple experts. Since DCE considers the dependency patterns between different M_j and Y during training, various M_j are connected through Y. These diverse M_j mutually influence each other during the training process, resulting in increased complexity. How to utilize dependency patterns in scenarios with multiple experts poses an interesting challenge.

A. Proof of Theorem 1.1

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.