Online Appendix for Strategic Disinformation Generation and Detection

A.7 Proof of Lemma 6

We start with two general lemmas to establish the optimization principle, then apply it to $J(\beta, \alpha)$.

Lemma 9. *Let $f_1, f_2, f_3 : \mathcal{F} \to \mathbb{R}$ be real-valued functions on the feasible set $\mathcal{F} \subset \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are non-empty sets, and $\mathcal{F}$ is defined by constraints coupling $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. For each $\beta \in \mathcal{B}$, let $A(\beta) = \{\alpha \mid (\beta, \alpha) \in \mathcal{F}\}$ be non-empty, and assume there exists a common $\alpha^*(\beta) \in A(\beta)$ such that:*

$$f_i(\beta, \alpha^*(\beta)) = \max_{\alpha \in A(\beta)} f_i(\beta, \alpha) \quad \textit{for } i = 1, 2, 3,$$

with maxima attained. Define the weighted function $f(\beta, \alpha) = w_1 f_1(\beta, \alpha) + w_2 f_2(\beta, \alpha) + w_3 f_3(\beta, \alpha)$, where $w_1, w_2, w_3 \geq 0$. Then:

$$\max_{(\beta,\alpha) \in \mathcal{F}} f(\beta, \alpha) = \max_{\beta \in \mathcal{B}} f(\beta, \alpha^*(\beta)).$$

Proof. Define $g : \mathcal{B} \to \mathbb{R}$ by $g(\beta) = f(\beta, \alpha^*(\beta)) = w_1 f_1(\beta, \alpha^*(\beta)) + w_2 f_2(\beta, \alpha^*(\beta)) + w_3 f_3(\beta, \alpha^*(\beta))$, where $(\beta, \alpha^*(\beta)) \in \mathcal{F}$. We prove $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) = \max_{\beta\in\mathcal{B}} g(\beta)$.

- $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \geq \max_{\beta\in\mathcal{B}} g(\beta)$. For any $\beta \in \mathcal{B}$, since $(\beta, \alpha^*(\beta)) \in \mathcal{F}$, we have $f(\beta, \alpha^*(\beta)) = g(\beta) \leq \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$. Thus, $\max_{\beta\in\mathcal{B}} g(\beta) \leq \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$.
- $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \leq \max_{\beta\in\mathcal{B}} g(\beta)$. For any $(\beta, \alpha) \in \mathcal{F}$, since $\alpha \in A(\beta)$ and $\alpha^*(\beta)$ maximizes each $f_i(\cdot, \beta)$ over $A(\beta)$, we have $f_i(\beta, \alpha) \leq f_i(\beta, \alpha^*(\beta))$ for $i = 1, 2, 3$. Since $w_i \geq 0$, it follows that:
$$f(\beta, \alpha) = \sum_{i=1}^{3} w_i f_i(\beta, \alpha) \leq \sum_{i=1}^{3} w_i f_i(\beta, \alpha^*(\beta)) = f(\beta, \alpha^*(\beta)) = g(\beta).$$
Since $g(\beta) \leq \max_{\beta'\in\mathcal{B}} g(\beta')$, we have $f(\beta, \alpha) \leq \max_{\beta'\in\mathcal{B}} g(\beta')$. This holds for all $(\beta, \alpha) \in \mathcal{F}$, so $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) \leq \max_{\beta\in\mathcal{B}} g(\beta)$.

Thus, $\max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha) = \max_{\beta\in\mathcal{B}} g(\beta)$. □

Lemma 10. *Let $f_1, f_2, f_3 : \mathcal{F} \to \mathbb{R}$ be real-valued functions on the feasible set $\mathcal{F} \subset \mathcal{A} \times \mathcal{B}$, where $\mathcal{A}$ and $\mathcal{B}$ are non-empty sets, and $\mathcal{F}$ is defined by constraints coupling $\alpha \in \mathcal{A}$ and $\beta \in \mathcal{B}$. For each $\beta \in \mathcal{B}$, let $A(\beta) = \{\alpha \mid (\beta, \alpha) \in \mathcal{F}\}$ be non-empty, and assume there exists a common $\alpha^*(\beta) \in A(\beta)$ such that:*

$$f_i(\beta, \alpha^*(\beta)) = \max_{\alpha \in A(\beta)} f_i(\beta, \alpha) \quad \textit{for } i = 1, 2, 3,$$

with maxima attained. Define the weighted function $f(\beta, \alpha) = w_1 f_1(\beta, \alpha) + w_2 f_2(\beta, \alpha) + w_3 f_3(\beta, \alpha)$, where $w_1, w_2, w_3 \geq 0$, and assume that for each $\beta \in \mathcal{B}$, there exists a unique $\alpha \in A(\beta)$ that maximizes $f(\beta, \alpha)$. Then:

$$\left\{(\beta, \alpha^*(\beta)) \mid \beta \in \mathcal{B},\ f(\beta, \alpha^*(\beta)) = \max_{\beta' \in \mathcal{B}} f(\alpha^*(\beta'), \beta')\right\} = \left\{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{F},\ f(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}} f(\beta', \alpha')\right\}.$$

Proof. Define $M = \max_{(\beta,\alpha)\in\mathcal{F}} f(\beta, \alpha)$ and $g(\beta) = f(\beta, \alpha^*(\beta))$, where $(\beta, \alpha^*(\beta)) \in \mathcal{F}$. By Lemma 9, $M = \max_{\beta\in\mathcal{B}} g(\beta)$. Let:

- $S_1 = \{(\beta, \alpha^*(\beta)) \mid \beta \in \mathcal{B},\ g(\beta) = \max_{\beta'\in\mathcal{B}} g(\beta')\}$,

- $S_2 = \{(\beta, \alpha) \mid (\beta, \alpha) \in \mathcal{F},\ f(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}} f(\beta', \alpha')\}$.

We prove $S_1 = S_2$.

Step 1: $S_1 \subseteq S_2$**.** For $(\beta, \alpha^*(\beta)) \in S_1$, we have $g(\beta) = f(\beta, \alpha^*(\beta)) = \max_{\beta' \in \mathcal{B}} g(\beta') = M$. Since $(\beta, \alpha^*(\beta)) \in \mathcal{F}$ and $f(\beta, \alpha^*(\beta)) = M$, it follows that $(\beta, \alpha^*(\beta)) \in S_2$.

Step 2: $S_2 \subseteq S_1$**.** For $(\beta, \alpha) \in S_2$, we have $(\beta, \alpha) \in \mathcal{F}$ and $f(\beta, \alpha) = M$. Since $\alpha^*(\beta)$ maximizes each $f_i(\cdot, \beta)$ over $A(\beta)$, and $w_i \geq 0$, we have:

$$f(\beta, \alpha) = \sum_{i=1}^{3} w_i f_i(\beta, \alpha) \leq \sum_{i=1}^{3} w_i f_i(\beta, \alpha^*(\beta)) = f(\beta, \alpha^*(\beta)).$$

Since $f(\beta, \alpha) = M$ and $f(\beta, \alpha^*(\beta)) \leq M$, we get $f(\beta, \alpha^*(\beta)) = M$. By the uniqueness of the maximizer of $f(\cdot, \beta)$ over $A(\beta)$, $\alpha = \alpha^*(\beta)$. Thus, $f(\beta, \alpha^*(\beta)) = M = \max_{\beta' \in \mathcal{B}} g(\beta')$, so $(\beta, \alpha) = (\beta, \alpha^*(\beta)) \in S_1$.

Since $S_1 \subseteq S_2$ and $S_2 \subseteq S_1$, we conclude $S_1 = S_2$. □

Application to the Optimal Design Problem

Consider the following optimal design problem:

$$\begin{aligned} \max_{(\beta,\alpha)} \quad & J(\beta, \alpha) = \omega_R \mathbb{E}U^R(\beta, \alpha) + \omega_H \rho\, \mathbb{E}U_H^S(\beta, \alpha) + \omega_L (1 - \rho)\, \mathbb{E}U_L^S(\beta, \alpha) \\ \text{s.t.} \quad & (\beta, \alpha) \in \mathcal{F}(\phi), \end{aligned}$$

where $\omega_R, \omega_H, \omega_L \geq 0$ and $\omega_R + \omega_H + \omega_L = 1$. In the notation of Lemmas 9 and 10, let $f_1(\beta, \alpha) = \mathbb{E}U^R(\beta, \alpha)$, $f_2(\beta, \alpha) = \mathbb{E}U_H^S(\beta, \alpha)$, $f_3(\beta, \alpha) = \mathbb{E}U_L^S(\beta, \alpha)$, with weights $w_1 = \omega_R, w_2 = \omega_H \rho, w_3 = \omega_L(1 - \rho)$. Since $w_1, w_2, w_3 \geq 0$, the two general lemmas apply directly.

The feasible set $\mathcal{F}(\phi)$ is defined as:

$$\mathcal{F}(\phi) = \left\{ (\beta, \alpha) \left| \begin{array}{l} \beta = \phi(s_L | \theta = L)\lambda_L + \phi(s_H | \theta = L)\lambda_H, \\ \alpha = \phi(s_L | \theta = H)\lambda_L + \phi(s_H | \theta = H)\lambda_H, \\ \lambda_L, \lambda_H \in [0, 1] \end{array} \right. \right\}.$$

By Proposition 2, for each $\beta \in \mathcal{B}$, there exists $\alpha^*(\beta; \phi) \in A(\beta)$ such that:

$$\begin{aligned} \mathbb{E}U^R(\beta, \alpha^*(\beta; \phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U^R(\beta, \alpha), \\ \mathbb{E}U_H^S(\beta, \alpha^*(\beta; \phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U_H^S(\beta, \alpha), \\ \mathbb{E}U_L^S(\beta, \alpha^*(\beta; \phi)) &= \max_{\alpha \in A(\beta)} \mathbb{E}U_L^S(\beta, \alpha). \end{aligned}$$

We select the Pareto-optimal equilibrium for those cases with multiple equilibria, thus for each $\beta \in \mathcal{B}$, the maximizer $\alpha^*(\beta) \in A(\beta)$ of $J(\beta, \alpha)$ is unique.

Step 1: Maximum Value. Applying Lemma 9 with $f_1 = \mathbb{E}U^R$, $f_2 = \mathbb{E}U_H^S$, $f_3 = \mathbb{E}U_L^S$, and weights $w_R, w_H, w_L \geq 0$, we have $\max_{(\beta,\alpha) \in \mathcal{F}(\phi)} J(\beta, \alpha) = \max_{\beta \in \mathcal{B}} J(\beta, \alpha^*(\beta; \phi))$.

Step 2: Maximizers. Applying Lemma 10, we obtain the set of maximizers as:

$$\begin{aligned} & \left\{ (\beta, \alpha^*(\beta; \phi)) \left| \beta \in \mathcal{B},\ J(\beta, \alpha^*(\beta; \phi)) = \max_{\beta' \in \mathcal{B}} J(\beta', \alpha^*(\beta')) \right. \right\} \\ = \quad & \left\{ (\beta, \alpha) \left| (\beta, \alpha) \in \mathcal{F}(\phi),\ J(\beta, \alpha) = \max_{(\beta', \alpha') \in \mathcal{F}(\phi)} J(\beta', \alpha') \right. \right\}. \end{aligned}$$

A.8 Proof of Proposition 3

For simplicity, we slightly abuse the notation by denoting $U^R(\beta, \alpha^*(\beta;\phi))$ by $U^R(\beta)$, $U^S_H(\beta, \alpha^*(\beta;\phi))$ by $U^S_H(\beta)$, and $U^S_L(\beta, \alpha^*(\beta;\phi))$ by $U^S_L(\beta)$.

We first characterize the sender's equilibrium strategy for a given detector $\{\beta, \alpha^*(\beta;\phi)\}$.

Lemma 11. *For a classifier with a high capacity, the equilibrium is*

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L}, \sigma^R_{na} = \frac{C}{(1-\beta)\Delta^S_L}, \sigma^R_a = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma^R_{na} = 1, \sigma^R_a = 0, & \beta \in \left[\beta_1, \hat{\beta}\right), \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta^R_H}{\beta(1-\rho)\Delta^R_L}, \sigma^R_{na} = 1, \sigma^R_a = \frac{C}{\beta\Delta^S_L} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

where $\beta_1 := [\rho\Delta^R_H - (1-\rho)\Delta^R_L]/[(\phi(s_L|\theta = H)/\phi(s_L|\theta = L))\rho\Delta^R_H - (1-\rho)\Delta^R_L] \leq \hat{\beta}$*.*

For a classifier with a low capacity, the equilibrium is

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L}, \sigma^R_{na} = \frac{C}{(1-\beta)\Delta^S_L}, \sigma^R_a = 0, & \beta \in \left[0, \hat{\beta}\right) \\ \sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta^R_H}{\beta(1-\rho)\Delta^R_L}, \sigma^R_{na} = 1, \sigma^R_a = \frac{C}{\beta\Delta^S_L} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Proof. The equilibrium when $\beta \geq \hat{\beta}$ follows directly from Proposition 1. Now consider $\beta \in [0, \hat{\beta})$, Proposition 1 implies that the equilibrium is $\{\sigma^S = 1, \sigma^R_{na} = 1, \sigma^R_a = 0\}$ if $(1-\alpha^*(\beta;\phi))/(1-\beta) \geq ((1-\rho)\Delta^R_L)/(\rho\Delta^R_H)$ and is $\{\sigma^S = [(1-\alpha^*(\beta;\phi))\rho\Delta^R_H]/[(1-\beta)(1-\rho)\Delta^R_L], \sigma^R_{na} = C/((1-\beta)\Delta^S_L), \sigma^R_a = 0\}$ if $(1-\alpha^*(\beta;\phi))/(1-\beta) < [(1-\rho)\Delta^R_L]/(\rho\Delta^R_H)$.

According to Lemma 5,

$$\begin{aligned} \frac{1-\alpha^*(\beta;\phi)}{1-\beta} &= \begin{cases} -\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)} + \frac{1}{1-\beta}, & if\ \beta \leq \phi(s_L|\theta = L) \\ \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} & if\ \beta > \phi(s_L|\theta = L), \end{cases} \\ &= \min\left\{-\frac{\beta}{1-\beta}\frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)} + \frac{1}{1-\beta}, \frac{\phi(s_H|\theta = H)}{\phi(s_H|\theta = L)}\right\} \in \left[1, \frac{\phi(s_H|\theta = H)}{\phi(s_H|\theta = L)}\right] \end{aligned}$$

Let $g(\beta) := -[\beta/(1-\beta)][\phi(s_L|\theta = H)/\phi(s_L|\theta = L)] + 1/(1-\beta)$, which increases in β. One can see that $g(\hat{\beta}) = -[(\Delta^S_L - C)/C][\phi(s_L|\theta = H)/\phi(s_L|\theta = L)] + \Delta^S_L/C$.

- If the classifier has a low capacity, $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ or $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) < (\Delta^S_L - C)\rho\Delta^R_H/[\Delta^S_L\rho\Delta^R_H - (1-\rho)\Delta^R_L C]$(*i.e.* $g(\hat{\beta}) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$), then $(1-\alpha^*(\beta;\phi))/(1-\beta) < (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ and the equilibrium is $\{\sigma^S = [(1-\alpha^*(\beta;\phi))\rho\Delta^R_H]/[(1-\beta)(1-\rho)\Delta^R_L], \sigma^R_{na} = C/[(1-\beta)\Delta^S_L], \sigma^R_a = 0\}$ for all $\beta < \hat{\beta}$.

- If the classifier has a high capacity, $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$ and $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq (\Delta^S_L - C)\rho\Delta^R_H/[\Delta^S_L\rho\Delta^R_H - (1-\rho)\Delta^R_L C]$(*i.e.* $g(\hat{\beta}) \geq (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$), then $\beta_1 \in \left(0, \hat{\beta}\right]$, $g(\beta_1) = (1-\rho)\Delta^R_L/(\rho\Delta^R_H)$, and the equilibrium is

$$\begin{cases} \sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta^R_H}{(1-\beta)(1-\rho)\Delta^R_L}, \sigma^R_{na} = \frac{C}{(1-\beta)\Delta^S_L}, \sigma^R_a = 0, & \beta \in [0, \beta_1) \\ \sigma^S = 1, \sigma^R_{na} = 1, \sigma^R_a = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \end{cases}$$

□

According to Lemma 11 and Table 3, $\mathbb{E}U^R(\beta)$ weakly decreases in β when $\beta \in \left[\hat{\beta}, 1\right)$.

1. If the classifier has a low capacity, $\mathbb{E}U^R(\beta) = 0$ when $\beta < \hat{\beta}$. When $\beta \geq \hat{\beta}$, $\mathbb{E}U^R(\beta) = [1 - \alpha^*(\beta;\phi)/\beta]\rho\Delta_H^R$, which is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ and strictly decreases in β for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$. So, the optimal true-positive rate of the detector that maximizes the receiver's expected payoff is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$.

2. If the classifier has a high capacity, then

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1 - \alpha^*(\beta;\phi))\rho\Delta_H^R - (1-\beta)(1-\rho)\Delta_L^R, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta_H^R, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Lemma 12. $\phi(s_L|\theta = L) \geq \beta_1$ *if and only if* $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$.

Proof. Because $\mathcal{Z}(x) := (x-1)/[x - \phi(s_L|\theta = H)/\phi(s_L|\theta = L)]$ increases in x, $\phi(s_L|\theta = L) := \mathcal{Z}\left(\phi(s_H|\theta = H)/\phi(s_H|\theta = L)\right) \geq \beta_1 := \mathcal{Z}\left((1-\rho)\Delta_L^R/(\rho\Delta_H^R)\right)$ if and only if $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$. □

(a) If $\hat{\beta} > \phi(s_L|\theta = L)$,

$$\frac{\partial \mathbb{E}U^R(\beta)}{\partial \beta} = \begin{cases} (1-\rho)\Delta_L^R - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\rho\Delta_H^R > 0, & \beta \in [\beta_1, \phi(s_L|\theta = L)) \\ (1-\rho)\Delta_L^R - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\rho\Delta_H^R \leq 0, & \beta \in \left[\phi(s_L|\theta = L), \hat{\beta}\right) \end{cases}$$

So, $\mathbb{E}U^R(\beta)$ is maximized at $\phi(s_L|\theta = L)$ for $\beta \in [\beta_1, \hat{\beta})$. Because $\mathbb{E}U^R(\beta) = 0$ for all $\beta \leq \beta_1$ and $\mathbb{E}U^R(\beta)$ is maximized at $\hat{\beta}$ for $\beta \in [\hat{\beta}, 1)$, the maximizer of $\mathbb{E}U^R(\beta)$ among $\beta \in [0, 1]$ must be $\hat{\beta}$ or $\phi(s_L|\theta = L)$.[1] Thus, we only need to compare $\mathbb{E}U^R(\hat{\beta})$ and $\mathbb{E}U^R(\phi(s_L|\theta = L))$. One can see that $\mathbb{E}U^R(\hat{\beta}) = (1/\hat{\beta} - 1)\left(\phi(s_H|\theta = H)/\phi(s_H|\theta = L) - 1\right)\rho\Delta_H^R$ and $\mathbb{E}U^R(\phi(s_L|\theta = L)) = \phi(s_H|\theta = H)\rho\Delta_H^R - \phi(s_H|\theta = L)(1-\rho)\Delta_L^R = \phi(s_H|\theta = L)[\rho\Delta_H^R\phi(s_H|\theta = H)/\phi(s_H|\theta = L) - (1-\rho)\Delta_L^R]$. So, $\mathbb{E}U^R(\hat{\beta}) \geq \mathbb{E}U^R(\phi(s_L|\theta = L))$ if and only if:

$$\hat{\beta} \leq \frac{\phi(s_H \mid \theta = H) - \phi(s_H \mid \theta = L)}{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L)\right]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H) - \phi(s_H \mid \theta = L)}$$

$$=: \hat{\beta}_{\text{critical}}, \text{ which increases in } \phi(s_L|\theta = H)/\phi(s_L|\theta = L).$$

(b) If $\hat{\beta} \leq \phi(s_L|\theta = L)$,

$$\frac{\partial \mathbb{E}U^R(\beta)}{\partial \beta} = (1-\rho)\Delta_L^R - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\rho\Delta_H^R > 0, \beta \in \left[\beta_1, \hat{\beta}\right)$$

Hence, all $\beta \in \left[\beta_1, \hat{\beta}\right)$ is dominated by $\beta = \hat{\beta}$. We can also find $\mathbb{E}U^R(\beta)$ is constant for $\beta \in [\hat{\beta}, \phi(s_L|\theta = L)]$ and decreasing for $\beta > \phi(s_L|\theta = L)$.

[1] If $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) = (1-\rho)\Delta_L^R/(\rho\Delta_H^R)$, $\phi(s_L|\theta = L) = \beta_1$ and choosing $\beta = \phi(s_L|\theta = L)$ is equivalent to choosing any $\beta < \beta_1$ which makes zero payoff and is dominated by choosing $\beta = \hat{\beta}$.

All in all, if the classifier has a high capacity, the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ if $\hat{\beta} \leq \hat{\beta}_{\text{critical}}$ and is $\beta = \phi(s_L|\theta = L)$ if $\hat{\beta} > \hat{\beta}_{\text{critical}}$.

$$
\begin{aligned}
& \hat{\beta} \leq \hat{\beta}_{\text{critical}} \\
\Leftrightarrow\ & C \geq \hat{C} = \frac{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L)\right]\phi(s_H \mid \theta = L)}{\left[\phi(s_H|\theta = H) - \frac{(1-\rho)\Delta_L^R}{\rho\Delta_H^R}\phi(s_H \mid \theta = L) - 1\right]\phi(s_H \mid \theta = L) + \phi(s_H \mid \theta = H)}\Delta_L^S.
\end{aligned}
$$

Therefore, if the classifier has a high capacity, the optimal true-positive rate for the receiver is any $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ if $C \geq \hat{C}$ and is $\beta = \phi(s_L|\theta = L)$ if $C < \hat{C}$.

A.9 Proof of Proposition 4

1. Classifier with a low capacity

   According to Lemma 11, the equilibria are

$$
\begin{cases}
\sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in \left[0, \hat{\beta}\right) \\
\sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right)
\end{cases}
$$

   So, the payoff of the low-type sender is $\mathbb{E}U_L^S(\beta) = 0$ and the payoff of the high-type sender is

$$
\mathbb{E}U_H^S(\beta) = \begin{cases}
C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta}, & \beta \in \left[0, \hat{\beta}\right) \\
\Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right)
\end{cases}
$$

   For $\beta \in [0, \hat{\beta})$, $\mathbb{E}U_H^S(\beta)$ is weakly increasing and is dominated by $\beta = \hat{\beta}$. $\mathbb{E}U_H^S(\beta)$ is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ and is decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$. So, in this case, $\mathbb{E}U_H^S(\beta)$ is maximized at $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$.

2. Classifier with a high capacity

   According to Lemma 11, the equilibria are

$$
\begin{cases}
\sigma^S = \frac{(1-\alpha^*(\beta;\phi))\rho\Delta_H^R}{(1-\beta)(1-\rho)\Delta_L^R}, \sigma_{na}^R = \frac{C}{(1-\beta)\Delta_L^S}, \sigma_a^R = 0, & \beta \in [0, \beta_1) \\
\sigma^S = 1, \sigma_{na}^R = 1, \sigma_a^R = 0, & \beta \in \left[\beta_1, \hat{\beta}\right) \\
\sigma^S = \frac{\alpha^*(\beta;\phi)\rho\Delta_H^R}{\beta(1-\rho)\Delta_L^R}, \sigma_{na}^R = 1, \sigma_a^R = \frac{C}{\beta\Delta_L^S} - \frac{1-\beta}{\beta}, & \beta \in \left[\hat{\beta}, 1\right)
\end{cases}
$$

   So, the payoff of the low-type sender is

$$
\mathbb{E}U_L^S(\beta) = \begin{cases}
0, & \beta \in [0, \beta_1) \\
(1-\beta)\Delta_L^S - C & \beta \in \left[\beta_1, \hat{\beta}\right) \\
0, & \beta \in \left[\hat{\beta}, 1\right)
\end{cases}
$$

   and the utility of the high-type sender is

$$\mathbb{E}U_H^S(\beta) = \begin{cases} C\frac{\Delta_H^S}{\Delta_L^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0,\beta_1) \\ (1-\alpha^*(\beta;\phi))\Delta_H^S, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \Delta_H^S - (\Delta_L^S - C)\frac{\Delta_H^S}{\Delta_L^S}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Note that we have proved $\phi(s_L|\theta = L) \geq \beta_1$ for a strong classifier in Lemma 12. Hence, for $\beta \in [0, \beta_1)$, $\mathbb{E}U_H^S(\beta)$ is increasing in β. Moreover, $\mathbb{E}U_H^S(\beta)$ is decreasing in $\beta \in \left[\beta_1, \hat{\beta}\right)$, constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$, and decreasing for $\beta > \max\{\hat{\beta}, \phi(s_L|\theta = L)\}$.

Because $\mathbb{E}U_H^S(\hat{\beta}) = \Delta_H^S - (\Delta_L^S - C)(\Delta_H^S/\Delta_L^S)(\alpha^*(\hat{\beta};\phi)/\hat{\beta}) = (1 - \alpha^*(\hat{\beta};\phi))\Delta_H^S \leq (1 - \alpha^*(\beta_1;\phi))\Delta_H^S = \mathbb{E}U_H^S(\beta_1)$, $\mathbb{E}U_H^S(\beta)$ is maximized at β_1. One can see that $\mathbb{E}U_L^S(\beta)$ is also maximized at β_1.

To show that $\beta_1 = [\rho\Delta_H^R - (1-\rho)\Delta_L^R]/[(\phi(s_L|\theta = H)/\phi(s_L|\theta = L))\rho\Delta_H^R - (1-\rho)\Delta_L^R]$ decreases in $\phi(s_L|\theta = L)/\phi(s_L|\theta = H)$, one just needs to observe that the numerator of β_1 is negative, $(1-\rho)\Delta_L^R > 0$, and $\rho\Delta_H^R > 0$.

A.10 Proof of Proposition 5

We begin by establishing that $\mathbb{E}\widetilde{W}(\beta)$ is equivalent to $\mathbb{E}W(\beta)$ with the following parameter rescaling:

$$\Delta_H'^S = w_H\Delta_H^S, \quad \Delta_L'^S = w_L\Delta_L^S, \quad C' = w_L C, \quad \hat{C}' = w_L\hat{C}.$$

The expected social welfare function can be expressed as:

$$\begin{aligned} \mathbb{E}\widetilde{W}(\beta) &= \mathbb{E}U^R(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta) + \rho w_H \mathbb{E}U_H^S(\beta) \\ &= \begin{cases} \rho C'\frac{\Delta_H'^S}{\Delta_L'^S}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0,\beta_1) \\ \rho(1-\alpha^*(\beta;\phi))(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right] & \beta \in \left[\beta_1,\hat{\beta}\right) \\ \rho\left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)(\Delta_H'^S + \Delta_H^R) + \rho C'\frac{\Delta_H'^S}{\Delta_L'^S}\frac{\alpha^*(\beta;\phi)}{\beta} & \beta \in \left[\hat{\beta},1\right) \end{cases} \end{aligned}$$

The proof proceeds by analyzing two distinct cases based on the classifier's capacity.

Case 1: Low-capacity classifier. According to Propositions 3 and 4, $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$ constitutes the optimal detector for both the receiver and the sender. Consequently, any β in this interval maximizes social welfare when the classifier has low capacity.

Case 2: High-capacity classifier. From Propositions 3 and 4, we obtain the following optimal lie detector configurations:

- *Receiver's optimization:* $\{\beta^* \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}], \alpha^* = \alpha^*(\beta^*;\phi)\}$ when $\hat{\beta} \leq \hat{\beta}_{\text{critical}}$, and $\{\beta^* = \phi(s_L \mid \theta = L), \alpha^* = \alpha^*(\beta^*;\phi)\}$ when $\hat{\beta} > \hat{\beta}_{\text{critical}}$.
- *Sender's optimization:* $\{\beta^* = \beta_1, \alpha^* = \alpha^*(\beta^*;\phi)\}$.

This yields the following expression for expected social welfare:

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \underbrace{\mathbb{E}U^R(\beta)}_{\substack{\text{maximized at any} \\ \beta\in[\hat{\beta},\max\{\hat{\beta},\phi(s_L|\theta=L)\}]}} + \underbrace{w_L(1-\rho)\mathbb{E}U_L^S(\beta) + w_H\rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at } \beta=\beta_1}, & \text{if } \hat{\beta} \leq \hat{\beta}_{\text{critical}} \\ \underbrace{\mathbb{E}U^R(\beta)}_{\text{maximized at } \beta=\phi(s_L|\theta=L)} + \underbrace{w_L(1-\rho)\mathbb{E}U_L^S(\beta) + w_H\rho\mathbb{E}U_H^S(\beta)}_{\text{maximized at } \beta=\beta_1}, & \text{if } \hat{\beta} > \hat{\beta}_{\text{critical}} \end{cases}$$

First, a high-capacity classifier satisfies $\phi(s_L|\theta = L)/\phi(s_L|\theta = H) \geq (\Delta_L^S - C)\rho\Delta_H^R/[\Delta_L^S\rho\Delta_H^R - (1-\rho)\Delta_L^R C]$, which is equivalent to $C \leq \Delta_L^S(1-\beta_1)$.

To characterize the optimal true-positive rate, we analyze the following three scenarios based on the lying cost parameter:

1. **Low lying cost regime** ($C < \hat{C}$): This condition is equivalent to $\hat{\beta} > \hat{\beta}_{\text{critical}}$. In this case, the optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$.
2. **Intermediate lying cost regime** ($C \in [\hat{C}, \Delta_L^S\phi(s_H \mid \theta = L))$): This corresponds to $\hat{\beta} \in (\phi(s_L \mid \theta = L), \hat{\beta}_{\text{critical}}]$. The optimal β lies in the interval $[\beta_1, \hat{\beta}]$.
3. **High lying cost regime** ($C \in [\Delta_L^S\phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)]$): This corresponds to $\hat{\beta} \leq \phi(s_L \mid \theta = L)$. The optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$.

The inequality $\hat{C} < \Delta_L^S\phi(s_H \mid \theta = L)$ implies $\hat{\beta}_{\text{critical}} > \phi(s_L \mid \theta = L)$.

A.10.1 Low lying cost regime: $C < \hat{C}$

In this regime, we have $\hat{\beta} > \hat{\beta}_{\text{critical}} > \phi(s_L \mid \theta = L)$, which implies that we need only consider $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$. For this interval, the expected social welfare function takes the following form:

$$\mathbb{E}\widetilde{W}(\beta) = \rho\left(1 - \frac{\phi(s_L|\theta = H)}{\phi(s_L|\theta = L)}\beta\right)(\Delta'^S_H + \Delta^R_H) + (1-\rho)\left[(1-\beta)(\Delta'^S_L - \Delta^R_L) - C'\right]$$

Because $\mathbb{E}\widetilde{W}(\beta)$ is linear in β over the interval $[\beta_1, \phi(s_L \mid \theta = L)]$, the optimal true-positive rate β^* is characterized by:

$$\beta^* = \begin{cases} \phi(s_L \mid \theta = L) & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta'^S_H + \Delta^R_H) + (1-\rho)(\Delta'^S_L - \Delta^R_L) < 0 \\ \beta_1 & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta'^S_H + \Delta^R_H) + (1-\rho)(\Delta'^S_L - \Delta^R_L) > 0 \\ \text{any value in } [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}(\Delta'^S_H + \Delta^R_H) + (1-\rho)(\Delta'^S_L - \Delta^R_L) = 0 \end{cases}$$

To simplify the characterization of the optimal β, we define two parameters:

$$n_0 := \frac{\rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta^S_H}{(1-\rho)\Delta^R_L - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta^R_H} > 0, \;\; l_0 \;\; := \frac{(1-\rho)\Delta^S_L}{(1-\rho)\Delta^R_L - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta^R_H} > 0.$$

Using these parameters, we can express the set of optimal true-positive rates more concisely as:

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

and the maximum expected welfare can be written as:

$$\max_{\beta\in[\beta_1, \phi(s_L|\theta=L)]} \mathbb{E}\widetilde{W}(\beta) = \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) + (n_0 w_H + l_0 w_L - 1)^+ \left[(1-\rho)\Delta^R_L - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta^R_H\right](\phi(s_L \mid \theta = L) - \beta_1)$$

A.10.2 Intermediate lying cost regime: $C \in [\hat{C}, \Delta_L^S \phi(s_H \mid \theta = L))$

In this regime, we have $\hat{\beta} \in (\phi(s_L \mid \theta = L), \hat{\beta}_{\text{critical}}]$, which implies that we need only consider $\beta \in [\beta_1, \hat{\beta}]$. For this interval, the expected social welfare function takes the following form for $\beta \in [\beta_1, \hat{\beta}]$:

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \rho(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta)(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right] & \beta \in [\beta_1, \phi(s_L \mid \theta = L)) \\ \rho(1-\beta)\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}(\Delta_H'^S + \Delta_H^R) + (1-\rho)\left[(1-\beta)(\Delta_L'^S - \Delta_L^R) - C'\right] & \beta \in \Big[\phi(s_L \mid \theta = L), \hat{\beta}\Big) \\ -\rho\frac{C}{\Delta_L^S - C}\left(1 - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\right)\Delta_H^R + \rho C \frac{\Delta_H'^S}{\Delta_L^S}\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}, & \beta = \hat{\beta} \end{cases}$$

1. For $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$, the pattern of $\mathbb{E}\widetilde{W}(\beta)$ is the same as the low lying cost regime. That is,

$$\underset{\beta \in [\beta_1, \phi(s_L|\theta=L)]}{\text{argmax}} \quad \mathbb{E}\widetilde{W}(\beta) = \mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

2. For $\beta \in \Big[\phi(s_L \mid \theta = L), \hat{\beta}\Big)$, we have

$$\begin{aligned} \frac{\partial \mathbb{E}\widetilde{W}(\beta)}{\partial \beta} &= -\rho \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}(\Delta_H'^S + \Delta_H^R) - (1-\rho)(\Delta_L'^S - \Delta_L^R) \\ &= \underbrace{(1-\rho)\Delta_L^R - \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\rho\Delta_H^R}_{<0, \text{ by high capacity condition}} - w_H\,\rho\,\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\Delta_H^S - w_L(1-\rho)\Delta_L^S < 0 \end{aligned}$$

Thus, $\mathbb{E}\widetilde{W}(\beta)$ is decreasing in β for $\beta \in [\phi(s_L \mid \theta = L), \hat{\beta})$.

Condition 1: $n_0 w_H + l_0 w_L \leq 1$ Given $n_0 w_H + l_0 w_L \leq 1$, the highest expected welfare for $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$ is achieved at $\phi(s_L \mid \theta = L)$, which is given by:

$$\begin{aligned} \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) = &\left[-(1-\rho)\Delta_L^R \phi(s_H \mid \theta = L) + \rho\Delta_H^R \phi(s_H|\theta = H)\right] \\ &+ w_H \rho \Delta_H^S \phi(s_H|\theta = H) + w_L(1-\rho)\left(\Delta_L^S \phi(s_H \mid \theta = L) - C\right) \end{aligned}$$

The expected welfare at $\hat{\beta}$ is given by:

$$\mathbb{E}\widetilde{W}(\hat{\beta}) = \rho\Delta_H^R\left(\frac{\phi(s_H \mid \theta = H)}{\phi(s_H \mid \theta = L)} - 1\right)\left(\frac{C}{\Delta_L^S - C}\right) + w_H \rho \Delta_H^S \frac{\phi(s_H \mid \theta = H)}{\phi(s_H \mid \theta = L)}\left(\frac{C}{\Delta_L^S}\right)$$

The difference in expected welfare is:

$$\begin{aligned} \mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) = &\left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right][m_1(C) - n_1(C)w_H - l_1(C)w_L] \\ &\propto m_1(C) - n_1(C)w_H - l_1(C)w_L, \end{aligned}$$

where

$$
\begin{aligned}
m_1(C) :&= \frac{\mathbb{E}U^R\left(\hat{\beta}\right) - \mathbb{E}U^R\left(\phi(s_L \mid \theta = L)\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R} \\
&= \frac{\rho\Delta_H^R\left(\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} - 1\right)\left(\frac{C}{\Delta_L^S - C}\right) + (1-\rho)\Delta_L^R\phi(s_H \mid \theta = L) - \rho\Delta_H^R\phi(s_H|\theta = H)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R} \\
n_1(C) :&= \frac{\rho\frac{\Delta_H^S}{\Delta_L^S}\frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)}\left(\Delta_L^S\phi(s_H \mid \theta = L) - C\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R},\ l_1(C) := \frac{(1-\rho)\left(\Delta_L^S\phi(s_H \mid \theta = L) - C\right)}{(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R}
\end{aligned}
$$

By the definition of $\hat{C}$, the zero point of $m_1(C)$ is $\hat{C}$. As $m_1(C)$ is increasing in C, we have $m_1(C) \geq m_1(\hat{C}) = 0$ for $C \in [\hat{C}, \Delta_L^S\phi(s_H \mid \theta = L))$. Because $\hat{\beta} := 1 - C/\Delta_L^S > \phi(s_L \mid \theta = L)$, we have $C < \Delta_L^S\phi(s_H \mid \theta = L)$. Thus, $n_1(C) > 0$ and $l_1(C) > 0$.

Condition 2: $n_0 w_H + l_0 w_L > 1$ Given $n_0 w_H + l_0 w_L > 1$, the highest expected welfare for $\beta \in [\beta_1, \phi(s_L \mid \theta = L)]$ is achieved at β_1, which is given by:

$$
\begin{aligned}
\mathbb{E}\widetilde{W}(\beta_1) &= \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) \\
&\quad + (n_0 w_H + l_0 w_L - 1)\left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right](\phi(s_L \mid \theta = L) - \beta_1)
\end{aligned}
$$

The difference in expected welfare of β_1 and $\hat{\beta}$ is:

$$
\begin{aligned}
\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) &= \mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\phi(s_L \mid \theta = L)) \\
&\quad - (n_0 w_H + l_0 w_L - 1)\left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right](\phi(s_L \mid \theta = L) - \beta_1) \\
&= \left[(1-\rho)\Delta_L^R - \rho\frac{\phi(s_L \mid \theta = H)}{\phi(s_L \mid \theta = L)}\Delta_H^R\right][m_2(C) - w_H n_2(C) - w_L l_2(C)]
\end{aligned}
$$

where

$$
\begin{aligned}
m_2(C) &= m_1(C) + (\phi(s_L \mid \theta = L) - \beta_1) \\
n_2(C) &= n_1(C) - n_0(\phi(s_L \mid \theta = L) - \beta_1) \\
l_2(C) &= l_1(C) - l_0(\phi(s_L \mid \theta = L) - \beta_1)
\end{aligned}
$$

Then, $m_2(C) - w_H n_2(C) - w_L l_2(C) = m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)(\phi(s_L \mid \theta = L) - \beta_1)$. Then,

$$
\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) \propto m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)(\phi(s_L \mid \theta = L) - \beta_1)
$$

Optimal β: Based on the derivation of Condition 1 and 2,

$$
\mathbb{E}\widetilde{W}(\hat{\beta}) - \max_{\beta \in [\beta_1, \phi(s_L|\theta=L)]} \mathbb{E}\widetilde{W}(\beta) \propto M(C; w_H, w_L)
$$

where

$$
M(C; w_H, w_L) := m_1(C) - n_1(C)w_H - l_1(C)w_L - (n_0 w_H + l_0 w_L - 1)^+(\phi(s_L \mid \theta = L) - \beta_1),
$$

and it is increasing in C and decreasing in w_H and w_L, with $M(\hat{C};w_H,w_L) = -n_1(\hat{C})w_H - l_1(\hat{C})w_L - (n_0 w_H + l_0 w_L - 1)^+(\phi(s_L \mid \theta = L) - \beta_1) < 0$ for any w_H and w_L.

Because $m_1(C) - n_1(C)w_H - l_1(C)w_L$ is increasing in C and

$$m_1(C) - n_1(C)w_H - l_1(C)w_L\Big|_{C=\Delta_L^S\phi(s_H|\theta=L)} = m_1(\Delta_L^S\phi(s_H \mid \theta = L)) = \phi(s_H \mid \theta = L) > 0$$

The $M(C;w_H,w_L) < 0$ holds for any $C \in [\hat{C}, \Delta_L^S\phi(s_H \mid \theta = L))$ if and only if

$$n_0 w_H + l_0 w_L \geq \frac{m_1(\Delta_L^S\phi(s_H \mid \theta = L))}{\phi(s_L \mid \theta = L) - \beta_1} + 1 = \frac{1-\beta_1}{\phi(s_L \mid \theta = L) - \beta_1} > 1$$

The set of the optimal true-positive rate β^* is given as follows:

1. If $n_0 w_H + l_0 w_L < (1-\beta_1)/[\phi(s_L \mid \theta = L) - \beta_1]$, the zero point of $M(C;w_H,w_L)$ in the interval $[\hat{C}, \Delta_L^S\phi(s_H \mid \theta = L))$ is denoted as $C_1^*(w_H,w_L)$. The set of the optimal true-positive rate β^* is given by

$$\begin{cases} \mathcal{B}(w_H,w_L) & \text{if } C \in [\hat{C}, C_1^*(w_H,w_L)) \\ \mathcal{B}(w_H,w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H,w_L) \\ \{\hat{\beta}\} & \text{if } C \in (C_1^*(w_H,w_L), \Delta_L^S\phi(s_H \mid \theta = L)) \end{cases}$$

By implicit function theorem,

$$\frac{\partial C_1^*(w_H,w_L)}{\partial w_H} = -\frac{\frac{\partial M(C;w_H,w_L)}{\partial w_H}\Big|_{C=C_1^*(w_H,w_L)}}{\frac{\partial M(C;w_H,w_L)}{\partial C}\Big|_{C=C_1^*(w_H,w_L)}} > 0, \frac{\partial C_1^*(w_H,w_L)}{\partial w_L} = -\frac{\frac{\partial M(C;w_H,w_L)}{\partial w_L}\Big|_{C=C_1^*(w_H,w_L)}}{\frac{\partial M(C;w_H,w_L)}{\partial C}\Big|_{C=C_1^*(w_H,w_L)}} > 0.$$

2. If $n_0 w_H + l_0 w_L \geq (1-\beta_1)/[\phi(s_L \mid \theta = L) - \beta_1]$, the set of the optimal true-positive rate β^* is $\{\beta_1\}$.

A.10.3 High lying cost regime: $C \in [\Delta_L^S\phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)]$

This corresponds to $\hat{\beta} \leq \phi(s_L \mid \theta = L)$. The optimal β lies in the interval $[\beta_1, \phi(s_L \mid \theta = L)]$. The expected welfare is given by

$$\mathbb{E}\widetilde{W}(\beta) = \begin{cases} \rho(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta)(\Delta'^S_H + \Delta^R_H) + (1-\rho)\left[(1-\beta)(\Delta'^S_L - \Delta^R_L) - C'\right], & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \rho\left(1 - \frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right)(\Delta'^S_H + \Delta^R_H) + \rho C' \frac{\Delta'^S_H}{\Delta'^S_L}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}, & \beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)] \end{cases}$$

The $\mathbb{E}\widetilde{W}(\beta)$ is constant for $\beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)]$. For $\beta \in \left[\beta_1, \hat{\beta}\right)$, the form of $\mathbb{E}\widetilde{W}(\beta)$ is the same as it within the low lying cost regime. Thus, the $\partial\mathbb{E}\widetilde{W}(\beta)/\partial\beta$ is proportional to $1 - (n_0 w_H + l_0 w_L)$. Thus, if $n_0 w_H + l_0 w_L \leq 1$, all $\beta \in \left[\beta_1, \hat{\beta}\right)$ is dominated by any $\beta \in [\hat{\beta}, \phi(s_L \mid \theta = L)]$.

If $n_0 w_H + l_0 w_L > 1$, we need to compare $\mathbb{E}\widetilde{W}(\beta_1)$ and $\mathbb{E}\widetilde{W}(\hat{\beta})$, where

$$\begin{aligned}
\mathbb{E}\widetilde{W}(\beta_1) &= \rho\Delta_H^R(1-\beta_1\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}) - (1-\beta_1)(1-\rho)\Delta_L^R \\
&\quad + w_L(1-\rho)[(1-\beta_1)\Delta_L^S - C] + w_H\rho\Delta_H^S(1-\beta_1\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}) \\
\mathbb{E}\widetilde{W}(\hat{\beta}) &= \rho\left(1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right)(w_H\Delta_H^S + \Delta_H^R) + w_H\rho C\frac{\Delta_H^S}{\Delta_L^S}\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}
\end{aligned}$$

The difference in expected welfare between $\hat{\beta}$ and β_1 is:

$$\begin{aligned}
\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) = &\underbrace{(\beta_1-1)\left[-(1-\rho)\Delta_L^R + \rho\frac{\phi(s_L\mid\theta=H)}{\phi(s_L\mid\theta=L)}\Delta_H^R\right]}_{>0} \\
&+ w_H\rho\frac{\phi(s_L\mid\theta=H)}{\phi(s_L\mid\theta=L)}\Delta_H^S(\beta_1-\hat{\beta}) \\
&+ w_L(1-\rho)\Delta_L^S(\beta_1-\hat{\beta})
\end{aligned}$$

Then, $\mathbb{E}\widetilde{W}(\hat{\beta}) - \mathbb{E}\widetilde{W}(\beta_1) \le 0$ can be given by $m_3 w_H + n_3 w_L \ge 1$, where

$$\begin{aligned}
m_3 &:= \frac{\rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^S(\hat{\beta}-\beta_1)}{(\beta_1-1)\left[-(1-\rho)\Delta_L^R + \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R\right]} = \frac{\hat{\beta}-\beta_1}{1-\beta_1}n_0 > 0, \\
n_3 &:= \frac{(1-\rho)\Delta_L^S(\hat{\beta}-\beta_1)}{(\beta_1-1)\left[-(1-\rho)\Delta_L^R + \rho\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\Delta_H^R\right]} = \frac{\hat{\beta}-\beta_1}{1-\beta_1}l_0 > 0
\end{aligned}$$

Thus, $m_3 w_H + n_3 w_L \ge 1$ is equivalent to $n_0 w_H + l_0 w_L \ge (1-\beta_1)/(\hat{\beta}-\beta_1)$, which is equivalent to $C \le (1-\beta_1)[1 - 1/(n_0 w_H + l_0 w_L)]\Delta_L^S$. Note that $\hat{\beta} \le \phi(s_L \mid \theta = L)$, we have $n_0 w_H + l_0 w_L \le (1-\beta_1)/(\hat{\beta}-\beta_1)$ holds for all $C \in [\Delta_L^S\phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)]$ if and only if $n_0 w_H + l_0 w_L \le (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$.

The set of the optimal true-positive rate β^* is given as follows:

1. If $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of the optimal true-positive rate β^* is $[\hat{\beta}, \phi(s_L \mid \theta = L)]$.

2. If $n_0 w_H + l_0 w_L \ge (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of the optimal true-positive rate β^* is

$$\begin{cases}
\{\beta_1\} & \text{if } C \in \left[\Delta_L^S\phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right) \\
\{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\
[\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1-\beta_1)\right]
\end{cases}$$

A.10.4 Summary

In summary, the set of the optimal true-positive rate β^* is given by

1. **Case 1** $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$**:** the set of the optimal true-positive rate β^*

is

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < C_1^*(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H, w_L) \\ \{\hat{\beta}\} & \text{if } C \in (C_1^*(w_H, w_L), \Delta_L^S \phi(s_H \mid \theta = L)) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S(1 - \beta_1)] \end{cases}$$

$$= \begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < C_1^*(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup \{\hat{\beta}\} & \text{if } C = C_1^*(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C \in (C_1^*(w_H, w_L), \Delta_L^S(1 - \beta_1)] \end{cases}$$

2. **Case 2** $n_0 w_H + l_0 w_L \geq (1 - \beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$**:** the set of the optimal true-positive rate β^* is

$$\begin{cases} \{\beta_1\} & \text{if } C < (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1 - \beta_1)\right] \end{cases}$$

where

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \end{cases}$$

By the definition of $C_1^*(w_H, w_L)$, we can define a continuous function $\tilde{C}(w_H, w_L)$ as follows:

$$\tilde{C}(w_H, w_L) := \begin{cases} C_1^*(w_H, w_L), & \text{if } n_0 w_H + l_0 w_L < \frac{1-\beta_1}{\phi(s_L|\theta=L)-\beta_1} \\ (1 - \beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, & \text{if } n_0 w_H + l_0 w_L \geq \frac{1-\beta_1}{\phi(s_L|\theta=L)-\beta_1} \end{cases},$$

which is increasing in w_H and w_L. Then, the set of the optimal true-positive rate β^* can be written as

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \tilde{C}(w_H, w_L) \\ \mathcal{B}(w_H, w_L) \cup [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C = \tilde{C}(w_H, w_L) \\ [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}] & \text{if } C \in (\tilde{C}(w_H, w_L), \Delta_L^S(1 - \beta_1)] \end{cases}$$

A.11 Proof of Proposition 6

According to Lemma 5, we must have the constraints on the detector with $\{\beta, \alpha^*(\beta, \phi)\}$, where $\beta \leq \phi(s_L \mid \theta = L)$ and $\alpha^*(\beta, \phi) = (\phi(s_L|\theta = H)/\phi(s_L|\theta = L))\beta$. That is, compared to the original model, the only difference in the extension is the constraint on the range of β.

Case 1: Low-capacity classifier According to the proof of Proposition 3 in Appendix A.8, if the classifier has a low capacity, we have $\mathbb{E}U^R(\beta) = 0$ when $\beta < \hat{\beta}$. When $\beta \geq \hat{\beta}$, we have $\mathbb{E}U^R(\beta) = [1 - \alpha^*(\beta; \phi)/\beta]\rho\Delta_H^R$, which is constant for $\beta \in [\hat{\beta}, \max\{\hat{\beta}, \phi(s_L|\theta = L)\}]$.

Therefore, the set of optimal true-positive rates that maximize the receiver's expected payoff is:

- $[0, \phi(s_L|\theta = L)]$ if $\phi(s_L|\theta = L) < \hat{\beta}$ (i.e., $C < \Delta_L^S \phi(s_H \mid \theta = L)$), which gives zero payoff to the receiver.

- $[\hat{\beta}, \phi(s_L|\theta = L)]$ if $\phi(s_L|\theta = L) \geq \hat{\beta}$ (i.e., $C \geq \Delta^S_L \phi(s_H \mid \theta = L)$), which gives a constant payoff $[1 - \phi(s_L|\theta = H)/\phi(s_L|\theta = L)]\rho\Delta^R_H$ to the receiver.

By the proof of Proposition 4 in Appendix A.9, if the classifier has a low capacity, the payoff of the low-type sender is $\mathbb{E}U^S_L(\beta) = 0$ and the payoff of the high-type sender is

$$\mathbb{E}U^S_H(\beta) = \begin{cases} C\frac{\Delta^S_H}{\Delta^S_L}\frac{1-\alpha^*(\beta;\phi)}{1-\beta}, & \beta \in \left[0, \hat{\beta}\right) \\ \Delta^S_H - (\Delta^S_L - C)\frac{\Delta^S_H}{\Delta^S_L}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

Thus, the set of optimal true-positive rates that maximize the high-type sender's payoff is:

- $\{\phi(s_L|\theta = L)\}$ if $\phi(s_L|\theta = L) < \hat{\beta}$ (i.e., $C < \Delta^S_L \phi(s_H \mid \theta = L)$)
- $[\hat{\beta}, \phi(s_L|\theta = L)]$ if $\phi(s_L|\theta = L) \geq \hat{\beta}$ (i.e., $C \geq \Delta^S_L \phi(s_H \mid \theta = L)$)

For the weighted sum of the receiver's and the sender's payoffs, $\mathbb{E}\widetilde{W}(\beta) = \mathbb{E}U^R(\beta) + w_L(1 - \rho)\mathbb{E}U^S_L(\beta) + \rho w_H \mathbb{E}U^S_H(\beta)$, the set of optimal true-positive rates is:

- $\{\phi(s_L|\theta = L)\}$ if $\phi(s_L|\theta = L) < \hat{\beta}$ (i.e., $C < \Delta^S_L \phi(s_H \mid \theta = L)$)
- $[\hat{\beta}, \phi(s_L|\theta = L)]$ if $\phi(s_L|\theta = L) \geq \hat{\beta}$ (i.e., $C \geq \Delta^S_L \phi(s_H \mid \theta = L)$)

Case 2: High-capacity classifier: $\phi(s_H|\theta = H)/\phi(s_H|\theta = L) \geq [(1 - \rho)\Delta^R_L]/(\rho\Delta^R_H)$ **and** $C \leq (1 - \beta_1)\Delta^S_L$ According to the proof of Proposition 3 in Appendix A.8, if the classifier has a high capacity, the receiver's payoff is

$$\mathbb{E}U^R(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1 - \alpha^*(\beta;\phi))\rho\Delta^R_H - (1 - \beta)(1 - \rho)\Delta^R_L, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \left(1 - \frac{\alpha^*(\beta;\phi)}{\beta}\right)\rho\Delta^R_H, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

By Lemma 12 that $\phi(s_L|\theta = L) \geq \beta_1$, the set of optimal true-positive rates that maximize the receiver's payoff is

$$\begin{cases} [0, \phi(s_L|\theta = L)] & \text{if } \phi(s_L|\theta = L) < \hat{\beta} \text{ and } \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} = \frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H} \\ \{\phi(s_L|\theta = L)\} & \text{if } \phi(s_L|\theta = L) < \hat{\beta} \text{ and } \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} > \frac{(1-\rho)\Delta^R_L}{\rho\Delta^R_H} \\ [\hat{\beta}, \phi(s_L|\theta = L)] & \text{if } \phi(s_L|\theta = L) \geq \hat{\beta} \end{cases}$$

By the proof of Proposition 4 in Appendix A.9, if the classifier has a high capacity, the low-type sender's payoff is

$$\mathbb{E}U^S_L(\beta) = \begin{cases} 0, & \beta \in [0, \beta_1) \\ (1 - \beta)\Delta^S_L - C, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ 0, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

and the high-type sender's payoff is

$$\mathbb{E}U^S_H(\beta) = \begin{cases} C\frac{\Delta^S_H}{\Delta^S_L}\frac{1-\alpha^*(\beta;\phi)}{1-\beta} & \beta \in [0, \beta_1) \\ (1 - \alpha^*(\beta;\phi))\Delta^S_H, & \beta \in \left[\beta_1, \hat{\beta}\right) \\ \Delta^S_H - (\Delta^S_L - C)\frac{\Delta^S_H}{\Delta^S_L}\frac{\alpha^*(\beta;\phi)}{\beta}, & \beta \in \left[\hat{\beta}, 1\right) \end{cases}$$

which is maximized at $\beta = \beta_1$ by Proposition 4. Since $\phi(s_L|\theta = L) \geq \beta_1$ by Lemma 12, $\mathbb{E}U_H^S(\beta)$ is also maximized at $\beta = \beta_1$ in this case.

For the weighted sum of payoffs $\mathbb{E}\widetilde{W}(\beta) = \mathbb{E}U^R(\beta) + w_L(1-\rho)\mathbb{E}U_L^S(\beta) + \rho w_H \mathbb{E}U_H^S(\beta)$, according to the proof of Proposition 5 in Appendix A.10, we have the following results:

For $C < \Delta_L^S \phi(s_H \mid \theta = L)$ (i.e., $\phi(s_L|\theta = L) < \hat{\beta}$), the set of optimal true-positive rates is given by

$$\mathcal{B}(w_H, w_L) := \begin{cases} \{\phi(s_L \mid \theta = L)\} & \text{if } n_0 w_H + l_0 w_L < 1 \\ \{\beta_1\} & \text{if } n_0 w_H + l_0 w_L > 1 \\ [\beta_1, \phi(s_L \mid \theta = L)] & \text{if } n_0 w_H + l_0 w_L = 1 \end{cases}$$

For $C \geq \Delta_L^S \phi(s_H \mid \theta = L)$ (i.e., $\phi(s_L|\theta = L) \geq \hat{\beta}$), the analysis is more complex:

1. If $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of optimal true-positive rates is $[\hat{\beta}, \phi(s_L \mid \theta = L)]$.
2. If $n_0 w_H + l_0 w_L \geq (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$, the set of optimal true-positive rates is

$$\begin{cases} \{\beta_1\} & \text{if } C \in \left[\Delta_L^S \phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right) \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1-\beta_1)\right] \end{cases}$$

Summary: The set of optimal true-positive rates β^* is characterized as follows:

1. **Case 1:** $n_0 w_H + l_0 w_L < (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \Delta_L^S \phi(s_H \mid \theta = L) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in [\Delta_L^S \phi(s_H \mid \theta = L), \Delta_L^S(1-\beta_1)] \end{cases}$$

2. **Case 2:** $n_0 w_H + l_0 w_L \geq (1-\beta_1)/(\phi(s_L \mid \theta = L) - \beta_1)$

$$\begin{cases} \{\beta_1\} & \text{if } C < (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C \in \left((1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S, \Delta_L^S(1-\beta_1)\right] \end{cases}$$

Combining the results from the two cases, the set of optimal true-positive rates β^* is characterized as follows:

$$\begin{cases} \mathcal{B}(w_H, w_L) & \text{if } C < \max\left\{\Delta_L^S \phi(s_H \mid \theta = L), (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S\right\} \\ \{\beta_1\} \cup [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{if } C = (1-\beta_1)\left(1 - \frac{1}{n_0 w_H + l_0 w_L}\right)\Delta_L^S > \Delta_L^S \phi(s_H \mid \theta = L) \\ [\hat{\beta}, \phi(s_L \mid \theta = L)] & \text{Otherwise} \end{cases}$$

A.12 Proof of Proposition 7

Equilibrium First, if $f > 1 - u_H/P$, high-quality sellers do not enter the market because the maximum profit from entering, $(1-f)P$, is lower than the reservation utility, u_H. In this case, low-quality sellers also do not enter the market because they cannot induce any transactions when consumers know there are no high-quality sellers. The market breaks down.

Second, if $f \in (1-(C+u_L)/P, 1-u_H/P]$, low-quality sellers do not enter the market because the maximum profit from entering, $-C+(1-f)P$, is less than the reservation utility, u_L, and only high-quality sellers may enter the market. Thus, the optimal commission fee above $1-(C+u_L)/P$ is the highest possible f that incentivizes high-quality sellers to enter the market, $1-u_H/P$, which generates a profit of $\Pi = P(1-u_H/P)\rho$.

Third, we consider the case where $f \le \min\{1-(C+u_L)/P, 1-u_H/P\}$. In such cases, both types of sellers may enter the market. There are four possible (pure strategy) entry decisions:

1. Only low-quality sellers enter the market. This cannot be an equilibrium because consumers will never purchase a product.

2. Only high-quality sellers enter the market. The platform's profit is $\Pi = Pf\rho$.

3. None of the sellers enter the market. This cannot be an equilibrium because high-quality sellers can profitably deviate by entering the market.

4. Both types of sellers enter the market. Upon entry, the game coincides with that in the main model where $\Delta_H^S = \Delta_L^S = (1-f)P$, $\Delta_H^R = 1-P$, and $\Delta_L^R = P-v$. In this case, each high-type seller's sales are $\mathbb{E}U_H^S/(1-f)$ and each low-type seller's sales are $(\mathbb{E}U_L^S + C\sigma^S)/(1-f)$. The platform earns f fractions of the total sales. Thus, its profit is $\Pi = f[\rho\mathbb{E}U_H^S/(1-f) + (1-\rho)(\mathbb{E}U_L^S + C\sigma^S)/(1-f)] = [f/(1-f)][\rho\mathbb{E}U_H^S + (1-\rho)(\mathbb{E}U_L^S + C\sigma^S)]$. Table 4 presents the equilibrium payoffs of the platform and senders for a given commission rate f and detector (β, α).

α Range β Range	$\alpha \le 1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta)$	$\alpha \in \left(1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta), \beta\right]$
$[1-C/[(1-f)P], 1)$	$\Pi = \frac{f}{1-f}\rho\left[(1-f)P\left(1-\frac{\alpha}{\beta}\right) + C\frac{1-v}{P-v}\frac{\alpha}{\beta}\right]$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = (1-f)P\left(1-\frac{\alpha}{\beta}\right) + C\frac{\alpha}{\beta}$	
$(0, 1-C/[(1-f)P])$	$\Pi = f[\rho(1-\alpha) + (1-\rho)(1-\beta)]P$, $\mathbb{E}U_L^S = (1-\beta)(1-f)P - C$, $\mathbb{E}U_H^S = (1-\alpha)(1-f)P$	$\Pi = \frac{f}{1-f}\rho C\frac{1-v}{P-v}\frac{1-\alpha}{1-\beta}$, $\mathbb{E}U_L^S = 0$, $\mathbb{E}U_H^S = C\frac{1-\alpha}{1-\beta}$

Table 4: Equilibrium Payoffs for a Given Commission Rate and Detector

It is an equilibrium for both types of sellers to enter the market if and only if the following conditions hold:

1. A low-quality seller's equilibrium payoff exceeds his reservation utility, $\mathbb{E}U_L^S \ge u_L$.

2. A high-quality seller's equilibrium payoff exceeds his reservation utility, $\mathbb{E}U_H^S \ge u_H$.

According to Table 4, the above conditions are equivalent to $\alpha \le 1-(1-\beta)(1-\rho)(P-v)/[\rho(1-P)]$, $\alpha \le 1-u_H/[(1-f)P]$, and $\beta \le 1-(C+u_L)/[(1-f)P]$. Following Section 5.1, we analyze the case of $\lambda_H = 0$, where $\beta \le \phi(s_L|\theta = L)$ and $\alpha = \alpha^*(\beta,\phi) = \beta\phi(s_L|\theta = H)/\phi(s_L|\theta = L)$. The designer's optimization problem is:

$$\max_{f,\beta,\alpha} f[\rho(1-\alpha)+(1-\rho)(1-\beta)]P \tag{3}$$

$$\text{s.t. } \beta \le \phi(s_L|\theta = L), \alpha = \alpha^*(\beta,\phi),$$

$$\alpha \le 1 - \frac{(1-\rho)(P-v)}{\rho(1-P)}(1-\beta), \alpha \le 1 - \frac{u_H}{(1-f)P}, \text{ and } \beta \in \left(0, 1-\frac{C+u_L}{(1-f)P}\right]$$

$$\Leftrightarrow \max_{f,\beta} f\left[\rho\left(1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\beta\right)+(1-\rho)(1-\beta)\right]P$$

$$\text{s.t. } \eta \le \beta \le \phi(s_L|\theta=L),\ \beta \le \frac{\phi(s_L|\theta=L)}{\phi(s_L|\theta=H)}\left[1-\frac{u_H}{(1-f)P}\right],\ \text{and } \beta \le 1-\frac{C+u_L}{(1-f)P},$$

$$\text{where } \eta := \left[\frac{(1-\rho)(P-v)}{\rho(1-P)}-1\right]\Big/\left[\frac{(1-\rho)(P-v)}{\rho(1-P)}-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\right].$$

Notice that the upper bounds on β decrease in f. So, $f=0$ corresponds to the most relaxed constraints on β. The constraints of the designer's optimization problem can be satisfied if and only if there exists a feasible β when $f=0$, which is equivalent to

$$\eta \le \phi(s_L|\theta=L)\ \left(\Leftrightarrow \frac{\phi(s_H|\theta=H)}{\phi(s_H|\theta=L)} \ge \frac{(1-\rho)(P-v)}{\rho(1-P)}\right) \tag{4}$$

$$\text{and } \eta \le \min\left\{\frac{\phi(s_L|\theta=L)}{\phi(s_L|\theta=H)}(1-\frac{u_H}{P}), 1-\frac{C+u_L}{P}\right\} \tag{5}$$

Condition (5) is more likely to be satisfied when the classifier has higher capacity (i.e., when $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$ is larger) because η decreases in $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$ whereas the right-hand side of the inequality increases in $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$. In the limiting case where $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)\to+\infty$, the condition reduces to $(C+u_L)/[(1-\bar{\eta})P]\le 1$, where $\bar{\eta}:=1-\rho(1-P)/[(1-\rho)(P-v)]$. When $(C+u_L)/[(1-\bar{\eta})P]<1$, there exists a threshold ϕ_1^* such that (5) holds if and only if $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)\ge\phi_1^*$.

When both (4) and (5) hold, the optimal detector inducing both types of sellers to enter the market is

$$\beta^*=\eta, \alpha^*=\alpha^*(\eta,\phi)=\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\eta,\ \text{and } f^*=1-\max\left\{\frac{u_H}{1-\frac{\phi(s_L|\theta=H)}{\phi(s_L|\theta=L)}\eta}, \frac{C+u_L}{1-\eta}\right\}\frac{1}{P}.$$

The optimal true-positive rate $\beta^*=\eta$ decreases in $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$. The corresponding platform's profit is $\Pi_2^*=f^*\cdot P\cdot\{\rho[1-\eta\phi(s_L|\theta=H)/\phi(s_L|\theta=L)]+(1-\rho)(1-\eta)\}$, which is strictly increasing in $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)$.

Optimal Commission Rate and Detector We just need to compare the profit under the optimal commission rate and detector that induces only high-quality sellers to enter the market with the profit under the optimal commission rate and detector that induces both types of sellers to enter the market.

The platform obtains the highest profit when only high-quality sellers enter the market, $\Pi_1^*=P(1-u_H/P)\rho$, by setting $f=1-u_H/P$. In this case, one can verify that conditions (4) and (5) cannot be simultaneously satisfied. Thus, only high-quality sellers enter the market given any detector. The choice of the detector does not affect the platform's profits.

When $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)\to+\infty$, the platform's profit becomes

$$\left[1-\max\left\{\frac{u_H}{P}, \frac{C+u_L}{(1-\bar{\eta})P}\right\}\right][\rho+(1-\rho)(1-\bar{\eta})]P.$$

This exceeds Π_1^* if and only if $(C+u_L)/[(1-\bar{\eta})P]\le 1-(1-u_H/P)\rho/[\rho+(1-\rho)(1-\bar{\eta})] \Leftrightarrow C<\rho(1-P)P[1-(1-u_H/P)(P-v)/(1-v)]/[(1-\rho)(P-v)]-u_L$. Under the above condition, there exists ϕ_2^* such that Π_2^* exceeds Π_1^* if and only if $\phi(s_L|\theta=L)/\phi(s_L|\theta=H)>\phi_2^*$.

We conclude the proof by defining $\bar{\phi}:=\max\{\phi_1^*,\phi_2^*\}$.