Appendix I
Proof of Proposition 2.1

Fix policies g^L, f^L, g^A, f^A. We claim that there is a decoding function $\tilde{f}(x|I^L)$ such that

$$\gamma_T^A(g^A, f^A, \hat{g}^L(g^L, g^A)) = \gamma_T^L(g^L, \tilde{f}^L, \hat{g}^A(g^L, g^A)).$$

Indeed, we write the error probability of agent A using marginalization as follows:

$$\gamma_T^A(g^A, f^A, \hat{g}^L) = 1 - \sum_x \rho_1^A(x) \sum_{I_T^A} P_x(I_T^A) f^A(x|I_T^A) =$$

$$1 - \sum_x \rho_1^A(x) \sum_{I_T^A, w_T^A} P_x(I_T) f^A(x|I_T^A) =$$
$$1 - \sum_x \rho_1(x) \sum_{I_T} P_x(I_T) f^A(x|I_T^A) =$$
$$1 - \sum_x \rho_1(x) \sum_{I_T^L, w_T^L} P_x(I_T^L) P_x[w_T^L|I_T^L] f^A(x|I_T^A) =$$
$$1 - \sum_x \rho_1(x) \sum_{I_T^L} P_x(I_T^L) \tilde{f}^L(x|I_T^L).$$

where

$$\tilde{f}^L(x|I_T^L) = \sum_{w_T^L} P_x[w_T^L|I_T^L] f^A(x|I_T^A) = \sum_{w_T^L} \frac{P[I_T]}{P[I_T^L]} f^A(x|I_T^A).$$

Next we observe that

$$\gamma_T^A(g^A, f^A, \hat{g}^L(g^L, g^A)) \le \max_{g^A, f^A} \gamma_T^A \triangleq \gamma_T^A(g^L).$$

Moreover the claim proved above shows that

$$\gamma_T^A(g^A, f^A, \hat{g}^L(g^L, g^A)) = \gamma_T^L(g^L, \tilde{f}^L, \hat{g}^A(g^L, g^A)).$$

Therefore $\gamma_T^A(g^A, f^A, \hat{g}^L(g^L, g^A))$ is bounded below as

$$\ge \min_{g^L, f^L} \gamma_T^L(g^L, f^L, \hat{g}^A(g^L, g^A)). \triangleq \gamma^L(g^A)$$

We conclude that the maximum of $\gamma^L(g^A)$ with respect to g^A is lower than the minimum of $\gamma^A(g^L)$ with respect to g^L and the proof is complete.

Appendix II
Proof of Proposition 2.2

Let $x^*(I_T) = \arg\max_{x'} \rho_T^L(x')$. Let f be any decoding strategy. Then

$$\gamma_T^L(f^L, g^L, \hat{g}^A) = 1 - E_{I_T^L} \sum_{x \in \mathcal{X}} \rho_T^L(x) f^L(x|I_T^L)$$
$$\ge 1 - E_{I_T^L} \sum_{x \in \mathcal{X}} \rho_T^L(x^*(I_T)) f^L(x|I_T)$$
$$= 1 - E_{I_T^L} \rho_T^L(x^*(I_T)) \sum_{x \in \mathcal{X}} f^L(x|I_T),$$

or

$$\gamma_T^L(f^L, g^L, \hat{g}^A) \ge 1 - E_{I_T} \rho_T^L(x^*(I_T))(1 - f^L(e|I_T))$$
$$= 1 - E_{I_T^L} \rho_T^L(x^*(I_T)) + E_{I_T^L} \rho_T(x^*(I_T)) f(e|I_T).$$

The last term is non-negative. Hence

$$\gamma_T^L(f^L, g^L, \hat{g}^A) \ge 1 - E_{I_T^L} \max_x \rho_T^L(x).$$

The lower bound in the above equation is attained by the MAP decoder and the first claim is proved.

Next, we establish the equality of the lower error values. We omit the subscript T for simplicity. Moreover since $\hat{g}^A$ is a function of g^L and g^A we omit its presence in the error (for the same reason we omit $\hat{g}^L$). Direct application of the definitions gives

$$\begin{aligned}\gamma_{MAP}^l &= \max_{g^A} \min_{g^L} \gamma_{MAP}^L(g^L, g^A) = \\ &\max_{g^A} \min_{g^L} \gamma^L(f_{MAP}^L(g^L, g^A), g^L, g^A) = \\ &\max_{g^A} \min_{g^L} \min_{f^L} \gamma^L(f^L, g^L, g^A) = \\ &\max_{g^A} \min_{g^L, f^L} \gamma^L(f^L, g^L, g^A) = \gamma_l. \end{aligned} \tag{96}$$

We finally consider the relation between upper values. Note first that for each g^L g^A, the definition of the MAP error implies

$$\gamma_{MAP}^A(g^A, g^A) \le \gamma^A(f^A, g^L, g^A), \qquad \forall f^A.$$

The above relationship implies that for any f^A

$$\max_{g^A} \gamma_{MAP}^A(g^L, g^A) \le \max_{g^A} \gamma^A(f^A, g^L, g^A),$$

and

$$\min_{g^L} \max_{g^A} \gamma_{MAP}^A(g^L, g^A) \le \min_{g^L} \max_{g^A} \gamma^A(f^A, g^L, g^A).$$

Since this holds for all f^A, we obtain

$$\begin{aligned}\gamma_{MAP,u} &\le \min_{g^L} \max_{g^A} \max_{f^A} \gamma^A(f^A, g^L, g^A) = \\ &\min_{g^L} \max_{g^A, f^A} \gamma^A(f^A, g^L, g^A) = \gamma_u.\end{aligned}$$

To prove the converse, let f^{A*}, g^{L*}, g^{A*} be a tuple of policies that achieve the upper value γ^u. Then

$$\begin{aligned}\gamma_u = \gamma^A(f^{A*}, g^L*, g^{A*}) &= \min_{g^L} \max_{g^A, f^A} \gamma^A(f^A, g^L, g^A) \\ &\le \gamma^A(f^A, g^L, g^{A*}), \quad \forall f^A, g^L.\end{aligned}$$

In particular, the above holds for all g^L and $f^A = f_{MAP}(g^L, g^{A*})$. Therefore

$$\gamma_u \le \gamma^A(f_{MAP}(g^L, g^{A*}), g^L, g^{A*}) = \gamma_{MAP}^A(g^L, g^{A*}).$$

Now for all g^L it holds

$$\gamma_{MAP}^A(g^L, g^{A*}) \le \max_{g^A} \gamma_{MAP}^A(g^L, g^A).$$

Therefore

$$\gamma^u \le \min_{g^L} \max_{g^A} \gamma_{MAP}^A(g^L, g^A) = \gamma_{MAP}^u.$$

and the proof is complete.

Appendix III
Proof of Proposition 3.1

Using the definition of conditional entropy we have

$$H(X|I_T^L) = E_{I_T^L}(H(\rho_T^L)) = E[E[H(\rho_T^L)|I_{T-1}^L]. \tag{97}$$

Now

$$\begin{aligned}E[H(\rho_T^L)|I_{T-1}^L] &= -\sum_{x \in \mathcal{X}} E[\rho_T^L(x) \log \rho_T^L(x)|I_{T-1}^L] \\ &= -\sum_{x \in X} \sum_{z_T^L} P[z_T^L|I_{T-1}^L] \rho_T^L(x) \log \rho_T^L(x).\end{aligned} \tag{98}$$

Using marginalisation over the possible states we obtain

$$P[z_T^L|I_{T-1}^L] = \sum_x P[z_T^L|I_{T-1}^L, x]P[x|I_{T-1}^L]$$
$$= g^L(a_t|I_{T-1}^L)\sum_x K_x^L(\tilde{z}_T^L|I_{T-1}^L, a_T)\rho_{t-1}^L(x),$$

or

$$P[z_T^L|I_{T-1}^L] = g^L(a_T|I_{T-1}^L)\sigma_T^L(\tilde{z}_T^L|a_T). \quad (99)$$

We substitute (99) and the belief update (43) into the entropy expression (98) to obtain

$$E[H(\rho_T^L)|I_{T-1}^L]$$
$$= -\sum_{x\in\mathcal{X}}\sum_{z_T^L}\sigma_T^L(\tilde{z}_T^L|a_T)g_T^L(a_T|I_{T-1}^L)\times$$
$$\rho_{T-1}^L(x)\frac{K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T)}{\sigma_T^L(\tilde{z}_T^L|a_T)}\log\rho_{T-1}^L(x)\frac{K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T)}{\sigma_T^L(\tilde{z}_T^L|a_T)}$$
$$= -\sum_{x\in\mathcal{X}}\sum_{z_T^L}g_T^L(a_T|I_{T-1}^L)\rho_{T-1}^L(x)\times$$
$$K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T)\log\rho_{T-1}^L(x)\frac{K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T)}{\sigma_T^L(\tilde{z}_T^L|a_T)}. \quad (100)$$

We decompose the logarithm into two terms. The first involves $\rho_{T-1}^L(i)$ and is written as

$$-\sum_{a_T\in\mathcal{A}}g^L(a_T|I_{T-1}^L)\sum_{x\in\mathcal{X}}\rho_{T-1}^L(x)\times$$
$$\log\rho_{T-1}^L(x)\sum_{\tilde{z}_T^L}K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T),$$

which due to

$$\sum_{a_T\in\mathcal{A}}g^L(a_T|I_{T-1}^L)\sum_{\tilde{z}_T^L}g^L(a_T|I_{T-1})K_x^L(\tilde{z}_T^L|I_{t-1}^L, a_T) = 1,$$

simplifies to

$$-\sum_{a_T\in\mathcal{A}}g^L(a_T|I_{T-1})\sum_{x\in\mathcal{X}}\rho_{T-1}^L(x)\log\rho_{T-1}^L(x)$$
$$= -\sum_{x\in\mathcal{X}}\rho_{T-1}^L(x)\log\rho_{T-1}^L(x) = H(\rho_{T-1}^L). \quad (101)$$

It is easy to see that the second term is expressed as

$$\sum_{a_T\in\mathcal{A}}g^L(a_T|I_{T-1})\sum_{x\in\mathcal{X}}\rho_{T-1}^L(x)D(K_{xT}^L(a_T)||\sigma_T^L(a_T)).$$

The proof is complete.

Appendix IV
Proof of proposition 3.2

Recall $F^{-1}(H(X|I_T))$ is increasing for $\gamma < \gamma^*$. Fix g^A and let $g^L(g^A)$ be the best response. We write for simplicity $H(g^L, g^A)$ for $H(x|I_T^L)$ and its functional dependence on g^L, g^A. Then $H(g^L(g^A), g^A) = \min_{g^L} H(g^L, g^A) \le H(g^L, q^A)$, $\forall g^L$. Thus $F^{-1}(H(g^L(g^A), g^A)) \le F^{-1}(H(g^L, g^A))$, $\forall g^L$ and so $F^{-1}(H(g^L(g^A), g^L)) = \min_{g^L\in\mathcal{G}} F^{-1}(H(g^L, g^A))$. From this it follows

$$F^{-1}(\max_{g^A}\min_{g^L}H(g^L, g^A)) = \max_{g^A}\min_{g^L}F^{-1}(H(g^L, g^A))$$
$$\le \max_{g^A}\min_{g^L}(\gamma(g^L, g^A)) = \gamma_l. \quad (102)$$

We seek for lower bounds on

$$\max_{g^A}\min_{g^L}H(g^A, g^L) = \max_{g^A}\min_{g^L}EH(\rho_T^L). \quad (103)$$

We invoke proposition 3.1 to write

$$\max_{g^A}\min_{g^L}EH(\rho_T^L) =$$
$$\max_{g^A}\min_{g^L}(-\sum_{t=1}^{T-1}g_{t+1}^L(a)\sum_{x\in\mathcal{X}}\rho_t^L(x)D(K_{xt}^L(a)||\sigma_t^L(a)) + H(\rho_1^L)).$$

We recognize the JS divergence in the first term and we rewrite the above as

$$\max_{g^A}\min_{g^L}EH(\rho_T) = H(\rho_1^L) +$$
$$\max_{g^A}\min_{g^L}(-\sum_{t=1}^{T-1}\sum_{a,u}g_{t+1}^L(a)g_{t+1}^A(u)\times$$
$$JS(\rho_t^L, K_{1t}^L(a), \ldots, K_{Mt}^L(a)).$$

Similarly

$$\max_{g^A}\min_{g^L}EH(\rho_T) \ge H(\rho_1^L) +$$
$$\max_{g^A}\min_{g^L}(-\sum_{t=1}^{T-1}\sum_{a,u}g_{t+1}^A(a)g_{t+1}^L(u)\times$$
$$EJS(\rho_t^L, K_{1t}^L(a), \ldots, K_{Mt}^L(a)).$$

We get a time invariant bound if we define the quantity

$$EJS^* =$$
$$\min_{g^A\in\Delta(\mathcal{U})}\max_{g^L\in\Delta(\mathcal{A})}\max_{\rho}\sum_{a,u\in\mathcal{A},\mathcal{U}}g^L(a)g^A(u)\times$$
$$EJS(\rho, P_1[\cdot|a,u], \ldots, P_M[\cdot|a,u]). \quad (104)$$

Then

$$\max_{g^A}\min_{g^L}E(H(\rho_T^L)) \ge$$
$$H(\rho_1^L) - \min_{g^A}\max_{g^L}\sum_{t=1}^{T-1}\sum_{a,u}g_{t+1}^L(a)g_{t+1}^A(u)\cdot$$
$$\cdot EJS(\rho_t^L, K_{1t}^L(a), \ldots, K_{Mt}^L(a))$$
$$\ge H(\rho_1^L) - TESJ^*. \quad (105)$$

Note that the bounds are only concerned with the fully informed case. Therefore $K_{xt}(a) = P_x[\cdot|a,u]$. A similar somewhat stronger result is obtained if EJS is replaced by JS. Furthermore a simpler bound is obtained when the convexity of the KL divergence is exploited in conjunction with the Jensen's inequality.

$$D(K_{xt}^L(a)||\sigma_t^L(a)) = D(K_{xt}^L(a)||\sum_{x'\in\mathcal{X}}\rho_t(x')K_{x't}^L(a))$$
$$\le \sum_{x'\in\mathcal{X}}\rho_t^L(x)D(K_{xt}^L(a)||K_{x't}^L(a)).$$

Thus

$$\max_{g^A}\min_{g^L}E(H(\rho_T^L)) \ge$$
$$\max_{g^A}\min_{g^L}(-\sum_{t=1}^{T-1}g_{t+1}^L(a)g_{t+1}^A(u)\sum_{x\in\mathcal{X}}\rho_t(x)\sum_{x'\in\mathcal{X}/\{x\}}\rho_t(x')\cdot$$
$$\cdot D(K_{xt}^L(a)||K_{x't}^L(a)) + H(\rho_1^L)). \quad (106)$$

Therefore (106) implies

$$\max_{g^A}\min_{g^L} E(H(\rho_T^L)) \geq$$
$$H(\rho_1^L) - \min_{g^A}\max_{g^L} \sum_{t=1}^{T-1}\sum_{a,u} g_{t+1}^L(a) g_{t+1}^A(u) \cdot$$
$$\cdot \max_{x,x'} D(K_{xt}^L(a)||K_{x't}^L(a)) \geq$$
$$H(\rho_1^L) - T \max_{g^L\in\Delta(\mathcal{A})}\min_{g^A\in\Delta(\mathcal{U})}\max_{x,x'}$$
$$\sum_{a,u} g^L(a) g^A(u) D(P_x[\cdot|a,u]||P_{x'}[\cdot|a,u]). \quad (107)$$

and the theorem is proved.

Appendix V
Proof of proposition 3.3

The ML error is written as

$$\gamma_{ML,T}(g^L,g^A) = \gamma(f_{ML},g^L,g^A) = P[\hat{X}_{ML} \neq X]$$
$$= \sum_{x\in\mathcal{X}} \rho_1^L(x) \sum_{x'\in\mathcal{X}/\{x\}} P[\hat{X}_{ML}=x'|X=x]. \quad (108)$$

For each $x' \in \mathcal{X}$, let

$$A_{x'} = \{I_T^L : \hat{X}_{ML}(I_T^L) = x'\}, \quad (109)$$

and $1_{A_{x'}}$ be the indicator function of $A_{x'}$. Then

$$\gamma_{ML,T}(g^L,g^A) = \sum_{x\neq x',x,x'\in\mathcal{X}} \sum_{I_T^L} 1_{A_{x'}}(I_T^L) P[I_T^L|X=x]\rho_1^L(x). \quad (110)$$

By definition of the ML decoder

$$1_{A_{x'}}(I_T^L) \leq \left(\frac{P[I_T^L|X=x']}{P[I_T^L|X=x]}\right)^s, \qquad \forall s>0.$$

Therefore:

$$\gamma_{ML,T}(g^L,g^A) \leq \sum_{x\in\mathcal{X}} \rho_1^L(x) \sum_{I_T^L} P[I_T|x] \times$$
$$\sum_{x'\in\mathcal{X}/\{x\}} \left(\frac{P[I_T^L|x]}{P[I_T^L|x]}\right)^s,$$

which proves the first part of (69).

Next we convert the above into an expression involving the belief vectors using Bayes' rule.

$$P[I_T^L|X=x] = \frac{\rho_T^L(x) P[I_T^L]}{\rho_1^L(x)}$$

Then

$$\left(\frac{P[I_t^L|x']}{P[I_t^L|x]}\right)^s = \left(\frac{\rho_T^L(x')}{\rho_1^L(x)}\right)^s \left(\frac{\rho_1^L(x)}{\rho_1^L(x')}\right)^s.$$

Therefore

$$\gamma_{ML,T} \leq \sum_{x\in\mathcal{X}} \rho_1^L(x) \sum_{I_T^L} \frac{\rho_T(x)P[I_T^L]}{\rho_1^L(x)} \times$$
$$\sum_{x'\in\mathcal{X}/\{x\}} \left(\frac{\rho_T^L(x')}{\rho_1^L(x)}\right)^s \left(\frac{\rho_1^L(x)}{\rho_1^L(x')}\right)^s$$
$$= \sum_{I_T^L} P[I_T^L] \sum_{x\in\mathcal{X}} \rho_T^L(x) \sum_{x'\in\mathcal{X}/\{x\}} e^{-s(C_{x'}^x(\rho_T^L) - C_{x'}^x(\rho_1^L))}. \quad (111)$$

This proves the second part of (69).

Appendix VI
Proof of proposition 3.4

The proof follows the line of reasoning developed in [7] and [8]. The next expression follows from the updating equation of the belief vectors. Recall that we focus on the fully informed case $z_t = [a_t, u_t, y_t]$, $\tilde{z}_t^L = [u_t, y_t]$.

$$c_{x'}^x(\rho_T^L) - c_{x'}^x(\rho_1^L) = \sum_{t=1}^T \Lambda_{x'}^x(a_t, y_t, u_t). \quad (112)$$

where the log likelihood ratio is defined as

$$\Lambda_{x'}^x(\tilde{z}_t^L) = \ln\frac{K_x^L(\tilde{z}_t^L|I_{t-1},a_t)}{K_{x'}^L(\tilde{z}_t^L|I_{t-1},a_t)}. \quad (113)$$

In the fully informed case this becomes

$$\Lambda_{x'}^x(a,y,u) = \ln\frac{P_x[y|a,u]}{P_{x'}[y|a,u]}. \quad (114)$$

We make the standard assumption (see for instance [9]) that the likelihood ratios are bounded i.e:

$$|\ln\frac{P[y|a,u,X=x]}{P[y|a,u,X=x']}| < B, \qquad \forall a\in\mathcal{A}, u\in\mathcal{U}, x,x'\in\mathcal{X}. \quad (115)$$

For any $a,u \in \mathcal{A},\mathcal{U}$ the *moment generating function of the adversarial likelihood ratio* is

$$\mu(s,a,u) = E_y e^{-s\Lambda_{x'}^x(a,y,u)} = \sum_{y\in\mathcal{Y}} P_x[y|a,u] e^{-s\Lambda_{x'}^x(a,y,u)}. \quad (116)$$

So the expectation is taken with respect to $P_x[y|a,u]$ on y.

We write (111) accordingly as

$$\gamma_{ML,T} \leq E[\sum_{x\in\mathcal{X}} \rho_T^L(x) \sum_{x'\in\mathcal{X}/\{x\}} e^{-s\sum_{t=1}^T \Lambda_{x'}^x(a_t,y_t,u_t)}]. \quad (117)$$

We use the law of total expectations to write the right hand side as

$$E[E[\sum_{x\in\mathcal{X}} \rho_T^L(x) \sum_{x'\in\mathcal{X}/\{x\}} e^{-s\sum_{t=1}^T \Lambda_{x'}^x(A_t,Y_t,U_t)} | I_{T-1}^L]], \quad (118)$$

which is further written as

$$= E[\sum_{x\in\mathcal{X}} \rho_{T-1}^L(x) \sum_{x'\in\mathcal{X}/\{x\}} e^{-s\sum_{t=1}^{T-1} \Lambda_{x'}^x(A_T,Y_T,U_T)}.$$
$$\cdot E[\frac{K_{xT}^L(A_T)}{\sigma_T^L(A_T)} e^{-s\Lambda_{x'}^x(A_T,Y_T,U_T)} | I_{T-1}^L]]. \quad (119)$$

The right most conditional expectation is written as

$$\sum_{a_T,u_T,y_T} P[a_T,u_T,y_T|I_{T-1}^L] \frac{K_{xT}^L(A_T)}{\sigma_T^L(A_T)} e^{-s\Lambda_{x'}^x(A_T,Y_T,U_T)}.$$

But

$$P[a_T,u_T,y_T|I_{T-1}^L] =$$
$$\sum_x P_x[y_T|a_T,u_T]\rho_{T-1}(x) g^L(a_T|I_{T-1}^L) g^A(u_T|I_{T-1}^A) =$$
$$\sigma_T^L(A_T) g^L(a_T|I_{T-1}^L) g^A(u_T|I_{T-1}^A).$$

Therefore $\sigma_T^L(A_T)$ drops out and (119) is written as

$$E[\sum_{x\in\mathcal{X}} \rho_{T-1}^L(x) \sum_{x'\in\mathcal{X}/\{x\}} e^{-s\sum_{t=1}^{T-1} \Lambda_{x'}^x(A_t,Y_t,U_t)} \times$$
$$E_{a\sim g^L(\cdot|I_{T-1}^L), y\sim K_x(a,\cdot), u\sim g^A(\cdot|I_{T-1}^A)}[e^{-s\Lambda_{x'}^x(A_T,Y_T,U_T)} | I_{T-1}^L].$$

The conditional expectation inside the brackets can be tackled using Hoeffding's lemma, which states that if a random variable X is bounded in the interval $[a, b]$, then it is subgaussian with moment generating function bounded as

$$Ee^{-sX} \leq e^{-sEX+s^2\frac{(b-a)^2}{8}}.$$

For fixed x and $x' \neq x$, the random variable $\Lambda^x_{x'}(A_T, Y_T, U_T)$ is bounded in the interval $[-B, B]$. Thus

$$E[e^{-s\Lambda^x_{x'}(A_T,Y_T,U_T)}|I^L_{T-1}] \leq \tag{120}$$

$$e^{-sE[\Lambda^x_{x'}(A_T,Y_T,U_T)|I^L_{T-1}]+\frac{s^2B^2}{2}}. \tag{121}$$

The expectation in the right hand side is calculated as follows.

$$\begin{aligned}
&E[\Lambda^x_{x'}(A_T, Y_T, U_T)|I^L_{T-1}] \\
&= \sum_{a,u\in\mathcal{A},\mathcal{U}} g^L(a|I^L_{T-1})g^A(u|I^A_{t-1})\times \\
&\qquad\qquad \sum_{y\in\mathcal{Y}} P_x[y|a,u]\log(\frac{P_x[y|a,u]}{P_{x'}[y|a,u]}) \\
&= \sum_{a,u\in\mathcal{A},\mathcal{U}} g^L(a|I^L_{T-1})g^A(u|I^A_{t-1})D(P_x[\cdot|a,u])||P_{x'}[\cdot|a,u])
\end{aligned}$$

Putting the latter expression back into (117) we obtain

$$\gamma_{ML} \leq E[\sum_{x\in\mathcal{X}} \rho^L_{T-1}(x)e^{-s\sum_{t=1}^{T-1}\Lambda^x_{x'}(A_t,Y_t,U_t)}].$$

$$\cdot\, e^{-s\sum_{a,u}g^L(a|I^L_{T-1})g^A(u|I^A_{T-1})D(P_x[\cdot|a,u])||P_{x'}[\cdot|a,u])+\frac{s^2B^2}{2}}. \tag{122}$$

Let $g^{A*}(g^L)$ be the best response of the adversary to the legitimate encoder's policy g^L. Then

$$\min_{g^L}\max_{g^A}\gamma^L_T(g^L, g^A) = \min_{g^L}\gamma^L_T(g^L, g^{A*}(g^L)).$$

Consider a one shot deviation of $g^L, \tilde{g}^L$, which coincides with g^L at each stage but deviates at the terminal time T, where it becomes $\tilde{g}^L_T(a|I_{T-1}) = g^{L*}(a)$, where g^{L*} is a solution to the min-max problem defining $\tilde{D}$ (see (71)): Then

$$\gamma_u = \min_{g^L}\gamma^L_T(g^L, g^{A*}(g^L)) \leq \gamma^L_T(\tilde{g}^L, g^{A*}(\tilde{g}^L)) \leq$$

$$E[\sum_{x\in\mathcal{X}}\rho^L_{T-1}(x)\sum_{x'\in\mathcal{X}/\{x\}} e^{-s\sum_{t=1}^{T-1}\Lambda^x_{x'}(A_t,Y_t,U_t)}].$$

$$\cdot\, e^{-s\sum_{a,u\in\mathcal{A},\mathcal{U}}g^{L*}(a)g^{A*}(\tilde{g}^L)(u)D(P_x[\cdot|a,u]||P_{x'}[\cdot|a,u])+\frac{s^2B^2}{2}}.$$

Note that

$$e^{-s\sum_{a,u}g^{L*}(a)g^{A*}(\tilde{g}^L)(u)D(P_x[\cdot|a,u]||P_{x'}[\cdot|a,u])} \leq$$

$$e^{-s\sum_{a\in\mathcal{A}}g^{L*}(a)\min_{g^A\in\Delta(\mathcal{U})}\max_{x\neq x'}g^A(u)D(P_x[\cdot|a,u]||P_{x'}[\cdot|a,u])}.$$

The latter term is constant and does not depend on I_{T-1}. Hence it can be taken outside the expectation. Now, the expectation term coincides with the expression we started with, except that it involves $T-1$ instead of T. Straightforward induction proves the theorem.

Appendix VII
Calculating decay rates

In this section, we will calculate the decay rates of propositions 3.2 and 3.4 for both finite horizon problems discussed in section V.

A. Small example

We will begin the discussion by calculating the approximation of D^* of eq (66) (which is equal to $\tilde{D}$). Recall that

$$D^* \leq \min_{g^A}\max_{g^L}\sum_{a,u} g^L(a)g^A(u)\max_{x,x'} D(K^L_{xt}(a)||K^L_{x't}(a)) =$$

$$\min_{g^A}\max_{g^L}\sum_{a,u} g^L(a)g^A(u)\max_{x,x'} D(P_x[\cdot|a,u]||P_{x'}[\cdot|a,u]).$$

If the legitimate agent selects action $a = 1$, or if the adversary selects action $u = 0$, the maximum KL Divergence is

$$0.8log0.8/0.2 + 0.2log0.2/0.8 = 0.6log0.8/0.2 = 1.2$$

On the other hand, if the legitimate agent selects $a = 0$ and the adversary selects $a = 1$, the maximum KL divergence is

$$(0.8-s)log\frac{(0.8-s)}{(0.2+s)} + (0.2+s)log\frac{(0.2+s)}{(0.8-s)} =$$

$$(0.6-2s)log\frac{(0.8-s)}{(0.2+s)}.$$

Which is smaller than 1.2 because the function $(0.6-2s)log(0.8-s)/(0.2+s)$ is decreasing for the values of s we consider. Therefore, it makes sense for the legitimate agent to always select action $a = 1$ (deterministically). In that case $D^* \leq \tilde{D} = 1.2$.

The bounds do not depend on the strength parameter s. A plot of the bounds for different horizons is available at fig 2.

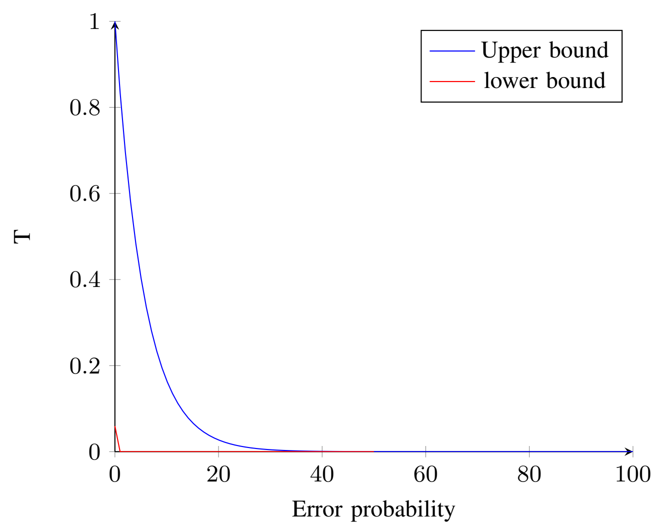


Fig. 2: Information theoretic bounds for the small example.

B. Large example

Assume the agent selects any action a. It is easy to see that the best response for the adversary is to attack the same sensor, in which case the maximum KL Divergence is $(0.6-2s)log\frac{(0.8-s)}{(0.2+s)}$. If the legit agent selects a non deterministic policy g^L, the same KL divergence is obtained by setting the response, $g^A(g^L) = g^L$.

For $s = 0.1$ we have

$$D^* = \tilde{D} = 0.4log(0.7/0.3) \approx 0.49$$

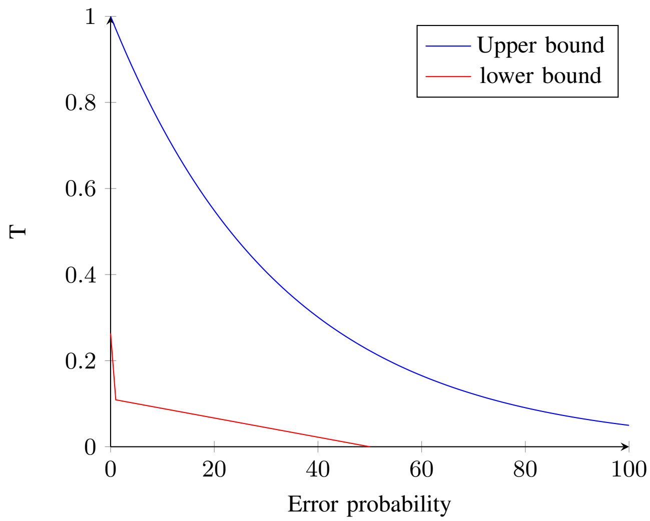
For $s = 0.2$ we have

$$D^* = \tilde{D} = 0.2log(0.6/0.4) = 0.4 \approx 0.117$$

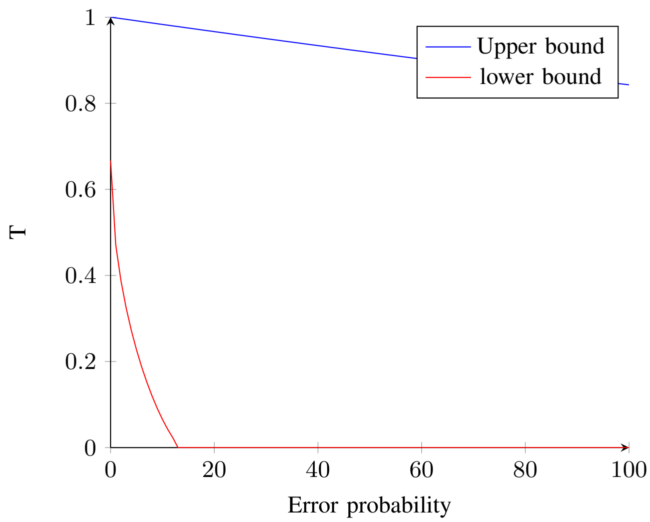
For $s = 0.25$ we have

$$D^* = \tilde{D} = 0.1log(0.55/0.45) \approx 0.029$$

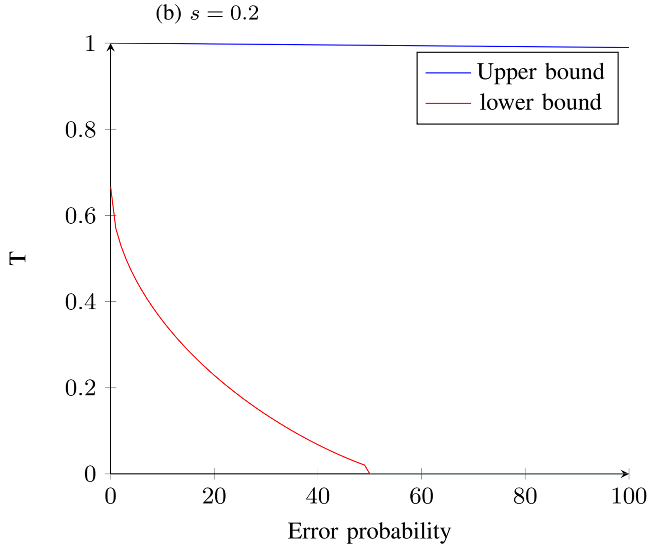
Plots of the bounds can be seen in fig. 3.



(a) $s = 0.1$



(b) $s = 0.2$



(c) $s = 0.25$

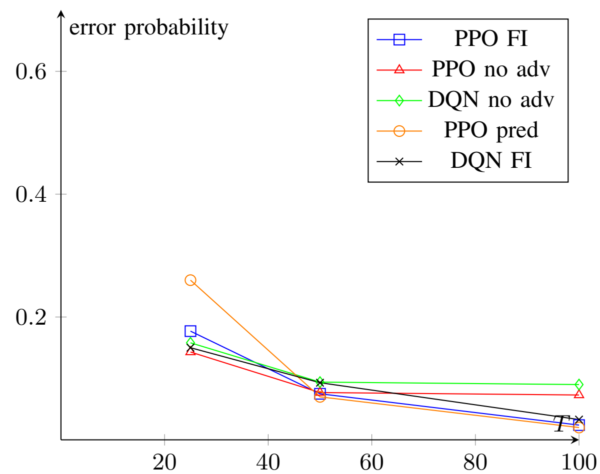
Fig. 3: Information theoretic bounds for the larger example

Appendix VIII
Finite horizon experiments against fully informed adversaries in the large environment

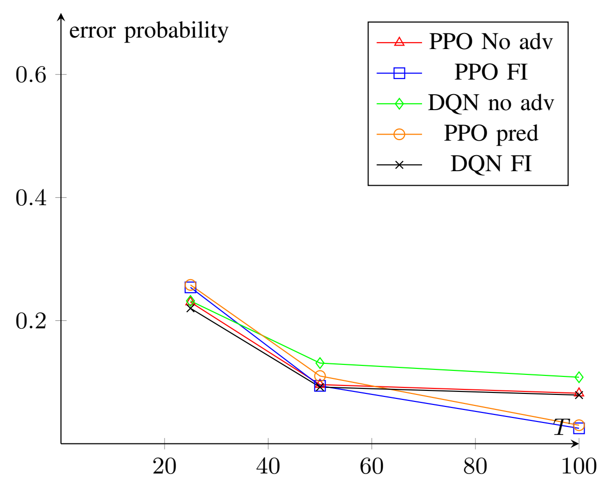
The results can be seen in figures 4 5 6.

Fig. 4: Error probabilities for the finite horizon problem: $s = 0.1$

(a) Error adversary



(b) Confidence adversary



(c) Entropy adversary

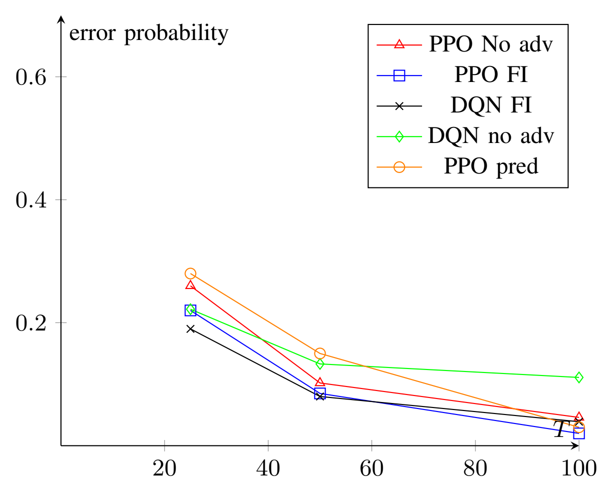
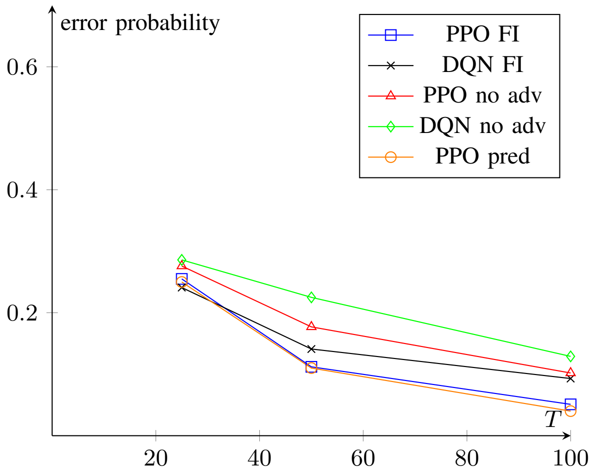
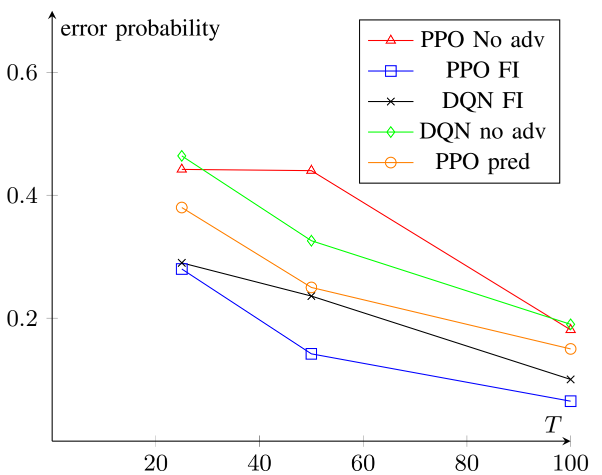


Fig. 5: Error probabilities for the finite horizon problem: $s = 0.2$

(a) Error adversary



(b) Confidence adversary



(c) Entropy adversary

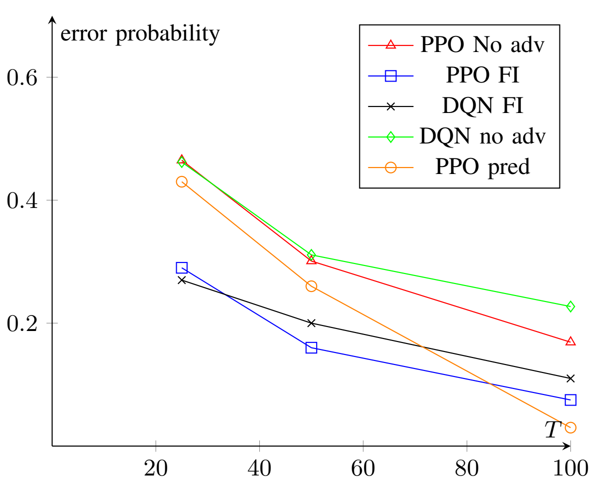
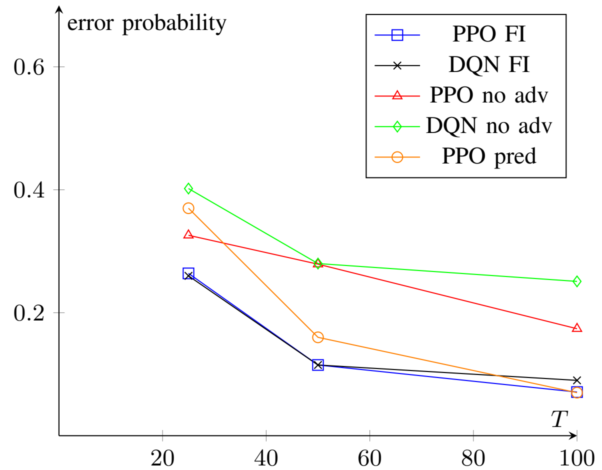
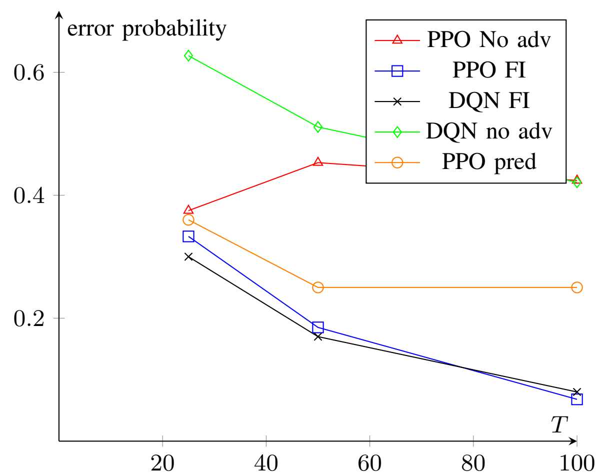


Fig. 6: Error probabilities for the finite horizon problem: $s = 0.25$

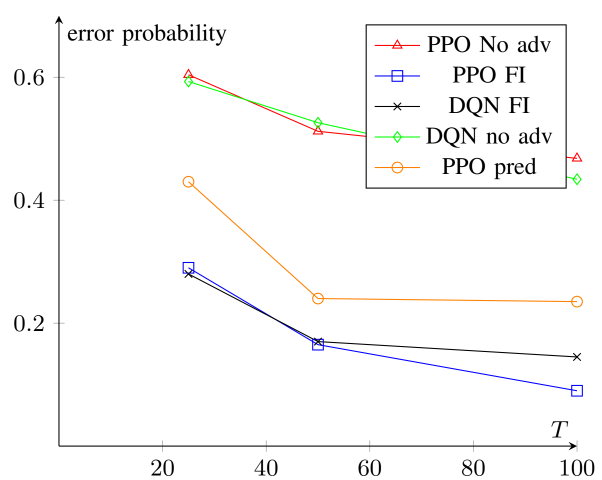
(a) Error adversary



(b) Confidence adversary



(c) Entropy adversary



APPENDIX IX
EXPERIMENTS AGAINST PARTIALLY INFORMED ADVERSARIES

We consider the finite horizon experiments of V-B. We limit our study to $s = 0.2$. The adversary has access to the past actions of the legitimate agent only $a_1, \cdots, a_{t-1}$. Using this information he seeks to predict what actions the legitimate agent will choose and attack the corresponding sensor. We test our algorithms against four adversarial heuristics.

1) Naive forecast adversary (NF): At each time t, the adversary forecasts the future legitimate action a_t through a naive forecast method, assuming $a_t = a_{t-1}$. Then he attacks the predicted sensor. At time $t = 0$ he attacks randomly.
2) Average forecast adversary (AF): At time $t > 0$ the adversary assigns a probability to each legitimate action a as $\hat{g}^L(a|a_1, \cdots, a_{t-1}) = \sum_{j=1}^{t-1} 1_{a_j=a}/(t-1)$ and predicts the action with the largest probability estimate. He then attacks the corresponding sensor.
3) Average forecast K latest (AF-K): Same as the AF adversary only he uses a window of the K latest actions. We set $K = 5$
4) Average forecast ϵ (AF-E): Same as the AF adversary but he randomly attacks with a probability ϵ. We set $\epsilon = 0.1$.

The results are illustrated in fig 7.

APPENDIX X
INFINITE HORIZON EXPERIMENTS AGAINST FULLY INFORMED ADVERSARIES.

We will limit the study to $s = 0.2$. We set $nSteps = 10$ for $tol = 0.1, 0.15$ and $nSteps = 20$ for $tol = 0.05$. The average stopping time and the empirical error probability can be seen in figures 9 and 8 respectively.

APPENDIX XI
PSEUDOCODES

Algorithms 3 4 and 5 present a high level overview of our DQN-FI, PPO-FI and PPO-pred algorithms.

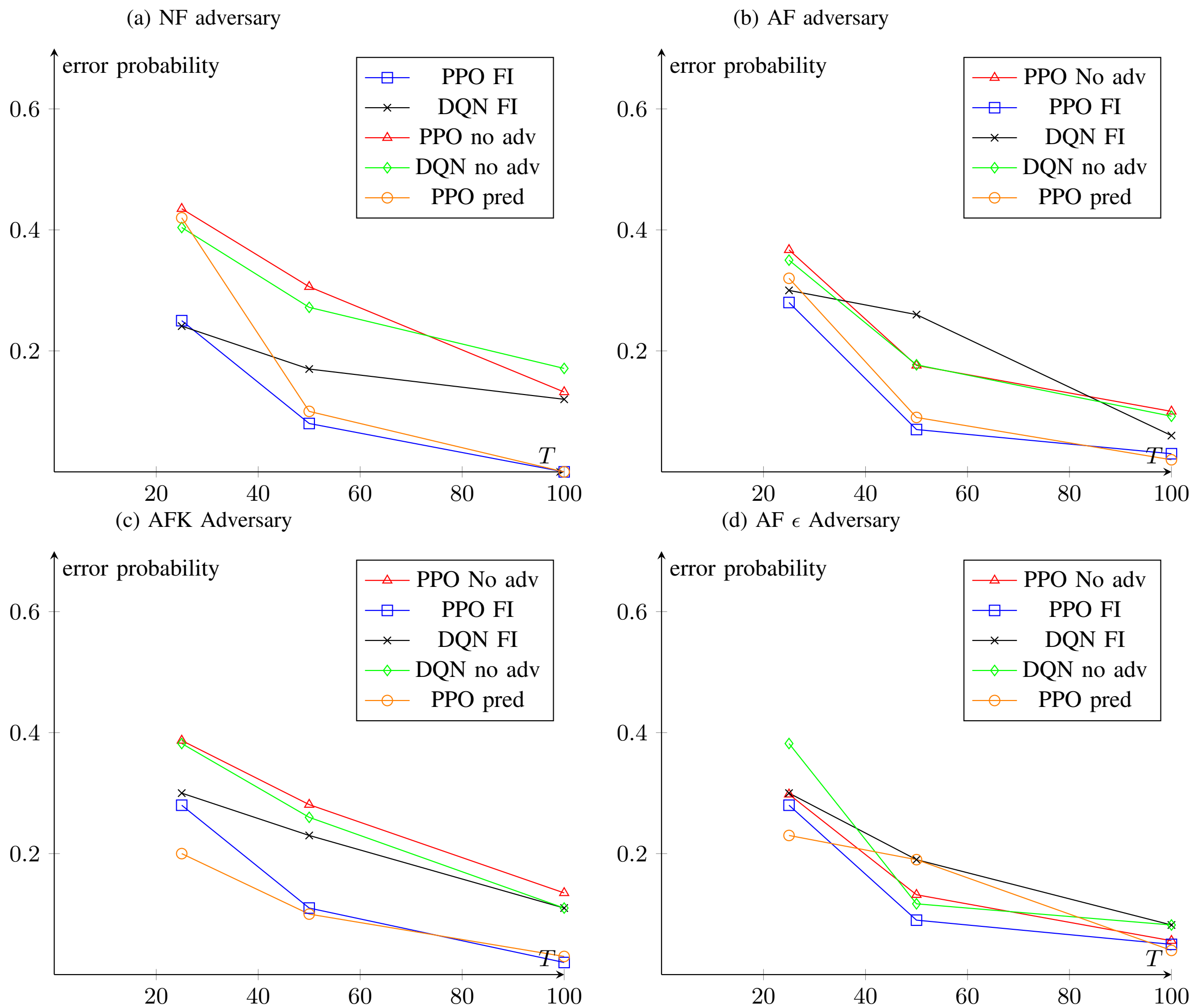
Algorithm 3: The DQN-FI algorithm for AAHT

```
Data: T ≥ 0, trainEpisodes, β, ρ_1
Initialise Q networks φ^L = φ̂^L, φ^A = φ̂^A.
Initialise replay buffers b1, b2.
for training episode in range(training episodes) do
    /*First part :Generate trajectory */
    Sample x ~ ρ_1
    initialise ρ_1^L = ρ_1^A = ρ_1
    for t = 1, 2, ···, T do
        Choose actions a_t and u_t using the neural networks
          φ^L, φ^A with epsilon-greedy exploration.
        Sample y_t ~ p_x^{a_t,u_t}(·)
        Update the legitimate belief according to eq (43),
          using a_t, y_t, u_t.
        Update the adversarial belief according to eq (44),
          using a_t, y_t, u_t.
        Compute legit reward r_t^L = γ_t^L − γ_{t+1}^L and
          adversarial reward r_t^A = −r_t^L.
        Store (ρ_t^L, a_t, ρ_{t+1}^L, r_t^L) to b1.
        Store (ρ_t^A, u_t, ρ_{t+1}^A, r_t^A) to b2.
        /*Second part: Train legit Q network */
        Sample a minibatch 𝓑 from b1.
        for each tuple (ρ, a, ρ', r) ∈ 𝓑 do
            Update φ^L by performing a gradient descent step
              on the loss
              (r + β max_{a'} Q_{φ̂^L}(ρ', a') − Q_{φ^L}(ρ, a))^2.
        end
        /*Third part: Train adversarial Q network */
        Sample a minibatch 𝓑 from b2. for each tuple
          (ρ, u, ρ', r) ∈ 𝓑 do
            Update φ^A by performing a gradient descent step
              on the loss
              (r + β max_{u'} Q_{φ̂^A}(ρ', u') − Q_{φ^A}(ρ, u))^2.
        end
        Update target networks φ̂^L ← φ^L and φ̂^A ← φ^A.
    end
end
```

APPENDIX XII
ALTERNATIVE REWARD STRUCTURES

In this section we demonstrate how the information theoretic bounds on the error probability of section III can be used to train DRL agents. We consider the finite horizon example of V-B and we limit our study to $s = 0.2$. Since the error probability is bounded below by a function of the belief entropy we use as reward the entropy difference $r_t^L = H(\rho_t) - H(\rho_{t+1})$. Similarly, the confidence based

Fig. 7: Error probabilities for games against uninformed adversaries: $s = 0.2$



reward $\rho_t^L = C(\rho_{t+1}) - C(\rho_t)$ is used. We train the PPO-FI agents and compare them with the PPO-FI algorithm of the main text that is trained using the error probability. The results are shown in fig 10. The performance differences are minor.

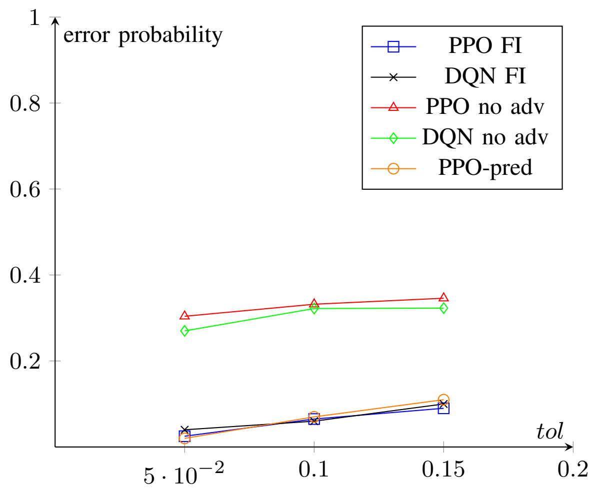
Appendix XIII
Implementation Details

All DRL agents use relatively small neural networks with two hidden layers of 200 units and reLU activation functions. The algorithms were trained for one million time steps. PPO uses $clip = 0.3$ instead of 0.2 which is the default in Stable Baselines 3. The size of the DQN replay buffer is 10000. The exploration parameter is initially set to $\epsilon = 0.9$ and gradually decreased to 0.05. γ is set to 0.999. We used the Adam optimizer with a learning rate of 0.00001. The training process was interrupted for evaluation purposes every 25 episodes.

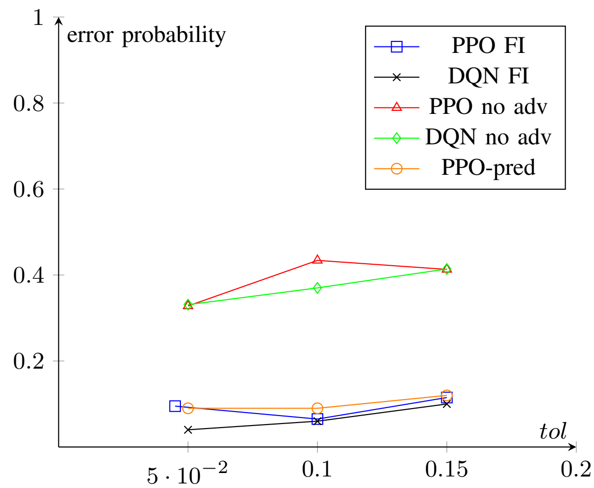
Recurrent neural networks, such as a recurrent variant of DQN, have been applied to POMDP computations and AHT [9] with good results. In this work we make direct use of the belief updates given in section II and we approximately represent policy and/or value functions by feed forward neural networks (see also [19]). In this manner the instability issues associated with the recurrent approximation of state dynamics are avoided.

Fig. 8: Error probability for the infinite horizon problem: $s = 0.2$

(a) error based adversary



(b) confidence based adversary



(c) Entropy based adversary

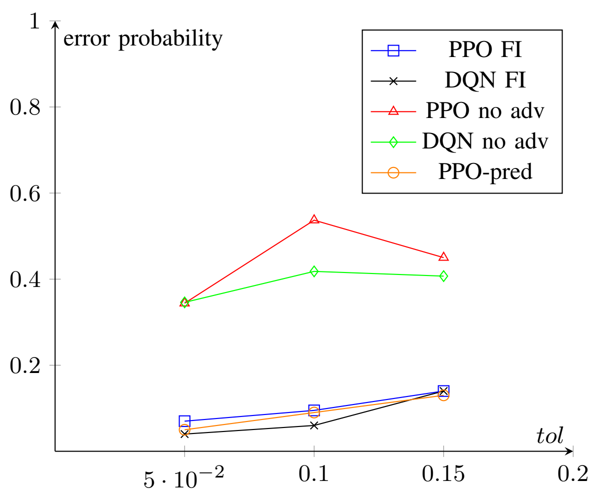
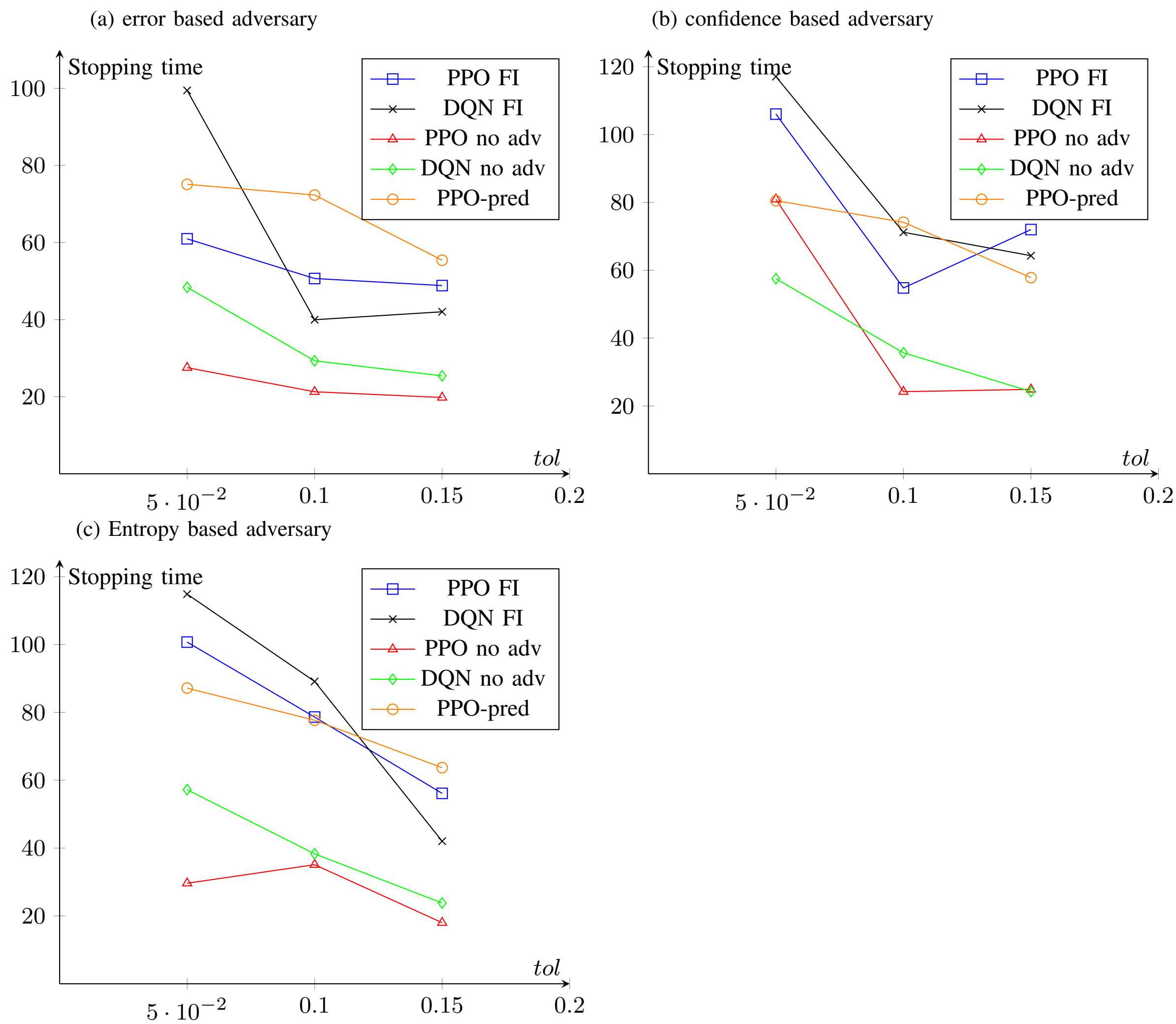


Fig. 9: Stopping time for the infinite horizon problem: $s = 0.2$



Algorithm 4: The PPO-FI algorithm for AAHT

Data: $T \geq 0, trainEpisodes, \rho_1, nTraj$
Initialise networks $\theta_E^L, \theta_E^A, \phi^L, \phi^A$.
Initialise buffers $b1, b2$.
for *training episode in range(training episodes)* **do**
  /∗First part :Generate trajectories ∗/
  Sample $x \sim \rho_1$
  initialise $\rho_1^L = \rho_1^A = \rho_1$
  for $t = 1, 2, \cdots, T$ **do**
    Compute distributions g_t^L, g_t^A using the neural networks θ_E^L and θ_E^A respectively.
    Sample actions $a_t \sim g_t^L$ and $u_t \sim g_t^A$.
    Sample $y_t \sim p_x^{a_t,u_t}(\cdot)$.
    Update the legitimate belief according to eq (43), using a_t, y_t, u_t.
    Update the adversarial belief according to eq (44), using a_t, y_t, u_t.
    Compute legit reward $r_t^L = \gamma_t^L - \gamma_{t+1}^L$ and adversarial reward $r_t^A = -r_t^L$.
    Store $a_t, \rho_t^L, r_t^L, g_t^L$ to $b1$ and $u_t, \rho_t^A, r_t^A, g_t^A$ to $b2$.
  end
  if *training episode* % *nTraj* $\neq 0$ **then**
    continue
  end
  /∗Second part: Train the legit agent ∗/
  Sample data from the trajectory buffer $b1$
  Estimate average rewards to go R_t^L for each time step.
  Estimate the advantages $A(\rho_t^L, a_t)$ using GAE.
  Update the policy network θ_E^L by maximising the objective of eq. (80).
  Update the critic ϕ^L by minimising the loss of eq (83).
  Empty $b1$.
  /∗Third part: Train the adversary ∗/
  Sample data from the trajectory buffer $b2$
  Estimate average rewards to go R_t^A for each time step.
  Estimate the advantages $A(\rho_t^A, u_t)$ using GAE.
  Update the policy θ_E^A by maximising the objective of eq. (84).
  Update the critic ϕ^A by minimising the loss of eq (85).
  Empty $b2$.
end

Algorithm 5: The PPO-pred algorithm for AAHT

Data: $T \geq 0, trainEpisodes, \rho_1, nTraj$
Initialise networks $\theta_E^L, \theta_E^A, \phi^L, \phi^A, \theta^P$.
Initialise buffers $b1, b2, b3$.
for *training episode in range(training episodes)* **do**
  /∗First part :Generate trajectories ∗/
  Sample $x \sim \rho_1$
  initialise $\rho_1^L = \rho_1^A = \rho_1$
  for $t = 1, 2, \cdots, T$ **do**
    Compute distributions $g_t^L, \hat{g}_t^A, g_t^A$ using the neural networks θ_E^L, θ^P, and θ_E^A respectively.
    Sample actions $a_t \sim g_t^L$ and $u_t \sim g_t^A$.
    Sample $y_t \sim p_x^{a_t,u_t}(\cdot)$.
    Update the legitimate belief according to eq (43), using $a_t, y_t, \hat{g}_t^A$.
    Update the adversarial belief according to eq (44), using a_t, y_t, u_t.
    Compute legit reward $r_t^L = \gamma_t^L - \gamma_{t+1}^L$ and adversarial reward $r_t^A = \gamma_{t+1}^A - \gamma_t^A$.
    Store $a_t, \rho_t^L, r_t^L, g_t^L$ to $b1$ and $u_t, \rho_t^A, r_t^A, g_t^A$ to $b2$.
    Store ρ^L, g_t^A to $b3$.
  end
  if *training episode* $\%nTraj \neq 0$ **then**
    continue
  end
  /∗Second part: Train predictor ∗/
  Using the tuples (ρ_t^L, g_t^A) from $b3$, train the predictor network by minimising the loss of eq (88). Use an optimizer like SGD or Adam.
  Empty $b3$.
  /∗Third part: Train the DRL algorithm ∗/
  Train θ_E^L, θ_E^A similarly to algorithm 4.
end

Fig. 10: Error probabilities of the PPO-FI algorithm for different reward structures: $s = 0.2$