Appendix:
Learning and Planning in Feature Deception Games

A Deferred Algorithms

We show the MILP formulation for the mathematical program $\mathcal{MP}1$. We use $M_c \subseteq M$ to denote the set of continuous features, and $M_d = M - M_c$ denotes the set of discrete features. For discrete feature $k \in M_d$, we assume that η_{ik} and budget B have been processed such that Constraint (3) has been modified to $\sum_{i\in N}\left(\sum_{k\in M_c}\eta_{ik}|x_{ik}-\hat{x}_{ik}| + \sum_{k\in M_d}\eta_{ik}x_{ik}\right) \leq B$. This transformation based on $\hat{x}_{ik} \in \{0,1\}$ simplifies our presentation below.

$$\max_{b,d,g,h,q,s,t,v,y} \quad \sum_{i\in N} t_i \tag{11}$$

$$s.t. \quad t_i = ve^{-2W} + \sum_l \gamma_l(v\epsilon - s_{il}) \tag{12}$$

$$\sum_{k\in M_c} w_k q_{ik} + \sum_{k\in M_d} w_k b_{ik} - Wv = -\sum_l s_{il} \tag{13}$$

$$h_{ik} \geq q_{ik} - \hat{x}_{ik}v, h_{ik} \geq \hat{x}_{ik}v - q_{ik} \qquad \forall k \in M_c \tag{14}$$

$$\sum_{i\in N}\left(\sum_{k\in M_d}\eta_{ik}b_{ik} + \sum_{k\in M_c}\eta_{ik}h_{ik}\right) \leq Bv \tag{15}$$

$$\epsilon g_{il} \leq s_{il}, s_{i(l+1)} \leq \epsilon g_{il} \qquad \forall l \tag{16}$$

$$s_{il} \leq v\epsilon \qquad \forall l \tag{17}$$

$$g_{il} \leq v, g_{il} \leq Zy_{il}, g_{il} \geq v - Z(1-y_{il}) \qquad \forall l \tag{18}$$

$$b_{ik} \leq v, b_{ik} \leq Zd_{ik}, b_{il} \geq v - Z(1-d_{ik}) \qquad \forall k \in M_d \tag{19}$$

$$q_{ik} \in [(\hat{x}_{ik} - \tau_{ik})v, (\hat{x}_{ik} + \tau_{ik})v] \cap [0,1] \qquad \forall k \in M_c \tag{20}$$

$$\sum_{i\in N} u_i t_i = 1 \tag{21}$$

$$\text{Categorical constraints} \tag{22}$$

$$t_i, v, s_{il}, q_{ik}, h_{ik}, g_{il} \geq 0, y_{il} \in \{0,1\} \qquad \forall k \in M_c, \forall l \tag{23}$$

$$b_{ik} \geq 0, d_{ik} \in \{0,1\} \qquad \forall k \in M_d \tag{24}$$

We establish the variables in the MILP above with the FDG variables as below.

$$t_i = \frac{f_i}{\sum_{i\in N} f_i u_i}, \qquad v = \frac{1}{\sum_{i\in N} f_i u_i} \tag{25}$$

$$h_{ik} = \frac{|x_{ik} - \hat{x}_{ik}|}{\sum_{i\in N} f_i u_i}, \qquad q_{ik} = \frac{x_{ik}}{\sum_{i\in N} f_i u_i}, \qquad \forall k \in M_c \tag{26}$$

$$d_{ik} = x_{ik}, \qquad b_{ik} = \frac{x_{ik}}{\sum_{i\in N} f_i u_i}, \qquad \forall k \in M_d \tag{27}$$

$$s_{il} = \frac{z_{il}}{\sum_{i\in N} f_i u_i}, \qquad g_{il} = \frac{y_{il}}{\sum_{i\in N} f_i u_i}, \qquad \forall l \tag{28}$$

$$\tag{29}$$

All equations above involving index i without summation should be interpreted as applying to all $i \in N$.

Algorithm 1: MILP-BS

Initialize $L = -1, U = 1, \delta = 0, \epsilon_{bs}$
while $U - L > \epsilon_{bs}$ **do**
  Solve the MILP $\mathcal{MP}1$ with objective in Eq. (10).
  if *objective value* < 0 **then**
    Let $U = \delta$
  else
    Let $L = \delta$
return U, the MILP solution when U was last updated

Algorithm 2: GREEDY

Use gradient-based method to find $x^{max} \approx \arg\max_x f(x)$ and $x^{min} \approx \arg\min_x f(x)$.
Sort the targets such that $u_1 \leq u_2 \leq \cdots \leq u_n$.
Initialize $i = 1, j = n$.
while $i < j$ *and budget* > 0 **do**
  Let $x_i \leftarrow x^{max}$ if
  if *Cost*$(x_i \leftarrow x^{max}) \leq$ *remaining budget* **then**
    $x_i \leftarrow x^{max}$, decrease the budget, $i = i + 1$.
  if *Cost*$(x_j \leftarrow x^{min}) \leq$ *remaining budget* **then**
    $x_j \leftarrow x^{min}$, decrease the budget, $j = j - 1$.
return feature configuration x

B Deferred Proofs

B.1 Proof of Theorem 1

We require the following lemma.

Lemma 7. *[6] Given observable features $x \in [0,1]^{mn}$, and $\Omega(\frac{1}{\rho\epsilon^2}\log\frac{n}{\delta})$ samples, we have $\frac{1}{1+\epsilon} \leq \frac{\hat{D}^x(t)}{D^x(t)} \leq 1+\epsilon$ with probability $1-\delta$, for all $t \in N$.*

Proof of Theorem 1. Fix $\epsilon, \delta > 0$. Fix two nodes $s \neq t$. For each x^i where $i = 1, 2, \ldots, m$, we have

$$\sum_{j=1}^{m} w_j(x^i_{sj} - x^i_{tj}) = \ln\frac{D^{x^i}(s)}{D^{x^i}(t)}$$

Let

$$b^{st} = (\ln\frac{D^{x^1}(s)}{D^{x^1}(t)}, \ldots, \ln\frac{D^{x^m}(s)}{D^{x^m}(t)}).$$

The system of equations above can be represented by $A^{st}w = b^{st}$. Let $||\cdot||$ be the matrix norm induced by L^1 vector norm, that is,

$$||A^{st}|| = \sup_{x\neq 0}\frac{|A^{st}x|}{|x|}, \quad \text{where } |x| = \sum_{j=1}^{m}|x_j|.$$

It is known that $||A^{st}|| = \max_{1\leq j\leq m}\sum_{i=1}^{m}|a^{st}_{ij}|$. In our case, the feature values are bounded in $[0,1]$ and thus $|a^{st}_{ij}| \leq 1$. This yields $||A^{st}|| \leq m$. Now, choose s, t such that $||(A^{st})^{-1}|| = \alpha$. Suppose A^{st} is invertible.

Let $\epsilon' = \frac{\epsilon}{4\alpha^2 m^2}$ and $\delta' = \frac{\delta}{m}$. Suppose we have $\Omega(\frac{1}{\rho\epsilon'^2}\log\frac{n}{\delta'})$ samples. From Lemma 7, for any node $r \in N$ and any feature configuration x^i where $i = 1, 2, \ldots, m$, $\frac{1}{1+\epsilon'} \leq \frac{\hat{D}^{x^i}(r)}{D^{x^i}(r)} \leq 1+\epsilon'$

with probability $1-\delta'$. The bound holds for all strategies simultaneously with probability at least $1-m\delta' = 1-\delta$, using a union bound argument. In particular, for our chosen nodes s and t, we have

$$\frac{1}{(1+\epsilon')^2} \leq \frac{\hat{D}^{x^i}(s)}{\hat{D}^{x^i}(t)} \frac{D^{x^i}(t)}{D^{x^i}(s)} \leq (1+\epsilon')^2, \quad \forall i = 1,\ldots,m$$

Define $\hat{b}^{st}$ similarly as b^{st} but using empirical distribution $\hat{D}$ instead of true distribution D. Let $e = \hat{b}^{st} - b^{st}$. Then, for each $i = 1,\ldots,m$, we have

$$-2\epsilon' \leq 2\ln\frac{1}{1+\epsilon'} \leq e_i = \ln\frac{\hat{D}^{x^i}(s)D^{x^i}(t)}{\hat{D}^{x^i}(t)D^{x^i}(s)} \leq 2\ln(1+\epsilon') \leq 2\epsilon'$$

Therefore, we have $|e| \leq 2\epsilon' m$. Let $\hat{w}$ be such that $A^{st}\hat{w} = \hat{b}^{st}$, i.e. $\hat{w} - w = (A^{st})^{-1}e$. Observe that

$$\begin{aligned}
\frac{|(A^{st})^{-1}e|/|(A^{st})^{-1}b^{st}|}{|e|/|b^{st}|} &\leq \max_{\tilde{e},\tilde{b}^{st}\neq 0} \frac{|(A^{st})^{-1}\tilde{e}|/|(A^{st})^{-1}\tilde{b}^{st}|}{|\tilde{e}|/|\tilde{b}^{st}|} \\
&= \max_{\tilde{e}\neq 0} \frac{|(A^{st})^{-1}\tilde{e}|}{|\tilde{e}|} \max_{\tilde{b}^{st}\neq 0} \frac{|\tilde{b}^{st}|}{|(A^{st})^{-1}\tilde{b}^{st}|} \\
&= \max_{\tilde{e}\neq 0} \frac{|(A^{st})^{-1}\tilde{e}|}{|\tilde{e}|} \max_{y\neq 0} \frac{|A^{st}y|}{|y|} \\
&= ||(A^{st})^{-1}|| \cdot ||A^{st}||
\end{aligned}$$

This leads to

$$\begin{aligned}
|(A^{st})^{-1}e| &\leq ||(A^{st})^{-1}|| \cdot ||A^{st}|| \cdot |e| \cdot \frac{|(A^{st})^{-1}b^{st}|}{|b^{st}|} \\
&\leq ||(A^{st})^{-1}|| \cdot ||A^{st}|| \cdot |e| \cdot \max_{\tilde{b}^{st}\neq 0} \frac{|(A^{st})^{-1}\tilde{b}^{st}|}{|\tilde{b}^{st}|} \\
&= ||(A^{st})^{-1}||^2 \cdot ||A^{st}|| \cdot |e| \\
&\leq \alpha^2 m(2\epsilon' m)
\end{aligned}$$

For any observable feature configuration x,

$$\begin{aligned}
&\left| \left(\sum_{j=1}^m w_j x_{ij}\right) - \left(\sum_{j=1}^m \hat{w}_j x_{ij}\right) \right| \leq \sum_{j=1}^m |\hat{w}_j - w_j| \\
&= |(A^{st})^{-1}e| \leq \alpha^2 m(2\epsilon' m) = \frac{\epsilon}{2}
\end{aligned}$$

Therefore,

$$\frac{1}{1+\epsilon} \leq \frac{f(x_i)}{\hat{f}(x_i)} \leq 1+\epsilon.$$

□

It is easy to see that we do not have to use the same pair of targets (s,t) for every feature configuration. In fact, this result can be easily adapted to allow for each feature configuration being implemented on a different system with a different set and number of targets. Instead of defining A^{st} and b^{st}, we could define A and b, where row i of A and i-th entry of b correspond to feature configuration x^i and targets (s^i, t^i). If feature configuration x^i is implemented on a system with n_i targets, we need $\Omega(\frac{1}{\rho\epsilon'^2}\log\frac{n_i}{\delta'})$ samples from this system, and then the argument above still holds.

B.2 Proof of Theorem 2

Fix two nodes s,t. Recall that in Theorem 1, without data poisoning, we learned the weights w by solving the linear equations $A^{st}\tilde{w} = \tilde{b}^{st}$ based on the empirical distribution of attacks, where

$\tilde{b}^{st} = (\ln \frac{\tilde{D}^{x^1}(s)}{\tilde{D}^{x^1}(t)}, \ldots, \ln \frac{\tilde{D}^{x^m}(s)}{\tilde{D}^{x^m}(t)})$[3]. Denote a parallel system of equations $A^{st}\hat{w} = \hat{b}^{st}$ which uses the poisoned data. We are interested in bounding $|\hat{w} - \tilde{w}| = |(A^{st})^{-1}(\hat{b}^{st} - \tilde{b}^{st})|$. Consider the k-th entry in the vector $\hat{b}^{st} - \tilde{b}^{st}$:

$$|(\hat{b}^{st} - \tilde{b}^{st})_k| = \left| \ln \frac{\hat{D}^{x^k}(s)}{\hat{D}^{x^k}(t)} \frac{\tilde{D}^{x^k}(t)}{\tilde{D}^{x^k}(s)} \right|$$

To simplify the notations, we denote $\tilde{D}^{x^k}(t) = \gamma_t^k$ and $\tilde{D}^{x^k}(s) = \gamma_s^k$, and without loss of generality, assume $\gamma_t^k \leq \gamma_s^k$. To find an upper bound of RHS of the above equation, we define function $g(\gamma_1, \gamma_2) = \frac{\gamma_t^k(\gamma_s^k + \gamma_1)}{\gamma_s^k(\gamma_t^k - \gamma_2)}$, and define function $h(\gamma_1, \gamma_2) = |\ln g(\gamma_1, \gamma_2)|$. The constraint that the attacker can only change γ fraction of the points translates into $|\gamma_1|, |\gamma_2|, |\gamma_1 - \gamma_2| \leq \gamma$. Since g is increasing in γ_1 and γ_2, g attains maximum at $(\gamma_1, \gamma_2) = (\gamma, \gamma)$ and minimum at $(\gamma_1, \gamma_2) = (-\gamma, -\gamma)$, which are the only two possible maxima of h. Observe that $g(\gamma, \gamma) \geq 1$ and $g(-\gamma, -\gamma) \leq 1$. It then suffices to compare $g(\gamma, \gamma)$ with $1/g(-\gamma, -\gamma)$:

$$\frac{1/g(-\gamma, -\gamma)}{g(\gamma, \gamma)} = \frac{\gamma_s(\gamma_t + \gamma)}{\gamma_t(\gamma_s - \gamma)} \frac{\gamma_s(\gamma_t - \gamma)}{\gamma_t(\gamma_s + \gamma)} = \frac{\gamma_s^2\gamma_t^2 - \gamma_s^2\gamma^2}{\gamma_t^2\gamma_s^2 - \gamma_t^2\gamma^2} \leq 1$$

Therefore, $h(\gamma_1, \gamma_2)$ is maximized at $(\gamma_1, \gamma_2) = (\gamma, \gamma)$. From here, we obtain

$$|(\hat{b}^{st} - \tilde{b}^{st})_k| \leq \ln \frac{(\gamma_s^k + \gamma)\gamma_t^k}{(\gamma_t^k - \gamma)\gamma_s^k} = \ln\left(\left(1 + \frac{\gamma}{\gamma_s^k}\right)\left(1 + \frac{\gamma}{\gamma_t^k - \gamma}\right)\right) \leq \frac{\gamma}{\gamma_s^k} + \frac{\gamma}{\gamma_t^k - \gamma}.$$

Recall that

$$\frac{\left|(A^{st})^{-1}(\hat{b}^{st} - \tilde{b}^{st})\right|}{\left|\hat{b}^{st} - \tilde{b}^{st}\right|} \leq \sup_{y \neq 0} \frac{\left|(A^{st})^{-1}y\right|}{|y|} = ||(A^{st})^{-1}|| = \alpha$$

Thus, we get

$$|\hat{w} - \tilde{w}| = |(A^{st})^{-1}(\hat{b}^{st} - \tilde{b}^{st})| \leq \alpha \left|\hat{b}^{st} - \tilde{b}^{st}\right| \leq \alpha \sum_{k=1}^{m} \left(\frac{\gamma}{\gamma_s^k} + \frac{\gamma}{\gamma_t^k - \gamma}\right)$$

Note that by Lemma 7, we have $\gamma_t^k \geq \frac{\rho}{1+\epsilon'} \geq \frac{\rho}{2}$. Since we assumed that $\gamma \leq \frac{\epsilon\rho}{4\alpha m} \leq \frac{\epsilon\rho}{4}$, we know that $\gamma \leq \gamma_t/2$. Thus, we get

$$|\hat{w} - \tilde{w}| \leq \alpha \sum_{k=1}^{m} \left(\frac{\gamma}{\gamma_s^k} + \frac{2\gamma}{\gamma_t^k}\right) \leq \frac{3\epsilon(1+\epsilon')}{4} \leq \frac{3}{4}\epsilon\left(1 + \frac{1}{4}\epsilon\right)$$

From here, using the triangle inequality, we have

$$|\hat{w} - w| \leq |\hat{w} - \tilde{w}| + |\tilde{w} - w| \leq \frac{3}{4}\epsilon\left(1 + \frac{1}{4}\epsilon\right) + \frac{\epsilon}{2} \leq \frac{3}{2}\epsilon$$

Thus, in the end, we get

$$\frac{1}{1+3\epsilon} \leq \frac{f(x_i)}{\hat{f}(x_i)} \leq 1 + 3\epsilon.$$

$\square$

B.3 Proof of Theorem 3

Let $\hat{f}(x_i) = \exp(\sum_k \hat{w}_k x_{ik})$ and $f(x_i) = \exp(\sum_k w_k x_{ik})$. Since

$$\frac{1}{1+\epsilon} < \frac{\hat{f}(x_i)}{f(x_i)} < 1 + \epsilon,$$

we get

$$-\epsilon \leq -\ln(1+\epsilon) < \sum_k (\hat{w}_k - w_k)x_{ik} = \ln \frac{\hat{f}(x_i)}{f(x_i)} < \ln(1+\epsilon) \leq \epsilon.$$

[3]Refer to Appendix B.1 for the notations used.

That is, $|\sum_k(\hat{w}_k - w_k)x_{ik}| < \epsilon$. The proof of Theorem 3.7 in [6] now follows if we redefine their $u_i(p_i)$ as $\sum_{k\in M} w_k x_{ik}$ and $\hat{u}_i(p_i)$ as $\sum_{k\in M} \hat{w}_k x_{ik}$. For completeness, we adapt their proof below using our notations.

Let $\bar{D}^x(t) = \frac{\hat{f}(x_t)}{\sum_i \hat{f}(x_i)}$. Then, we have

$$\begin{aligned}
\left|\ln \frac{\bar{D}^x(t)}{D^x(t)}\right| &= \left|\left(\sum_k (\hat{w}_k - w_k)x_{tk}\right) - \ln \frac{\sum_i \exp\{\sum_k \hat{w}_k x_{ik}\}}{\sum_i \exp\{\sum_k w_k x_{ik}\}}\right| \\
&\leq \left|\sum_k (\hat{w}_k - w_k)x_{tk}\right| + \\
&\left|\ln \frac{\sum_i \exp\{\sum_k w_k x_{ik}\}\exp\{\sum_k(\hat{w}_k - w_k)x_{ik}\}}{\sum_i \exp\{\sum_k w_k x_{ik}\}}\right| \\
&< \epsilon + \max_i \left|\ln \exp\{\sum_k (\hat{w}_k - w_k)x_{ik}\}\right| \\
&< 2\epsilon
\end{aligned}$$

Using a few inequalities we can bound $\left|\frac{\bar{D}^x(t)}{D^x(t)} - 1\right| \leq 4\epsilon$. Finally,

$$\begin{aligned}
|\hat{U}(x) - U(x)| &= \left|\sum_{i\in N} (\bar{D}^x(i) - D^x(i))u_i\right| \\
&\leq \sum_{i\in N} \left|\bar{D}^x(i) - D^x(i)\right| |u_i| \\
&= \sum_{i\in N} \left|\frac{\bar{D}^x(i)}{D^x(i)} - 1\right| |u_i|\, D^x(i) \\
&\leq 4\epsilon \sum_{i\in N} |u_i|\, D^x(i) \\
&\leq 4\epsilon \max_{i\in N} |u_i| \\
&\leq 4\epsilon
\end{aligned}$$

Let $x^* = \arg\min_x U(x)$ be the true optimal feature configuration and $x' = \arg\min_x \hat{U}(x)$ be the optimal configuration using the learned score function $\hat{f}$. Thus, we have $U(x') \leq \hat{U}(x') + 4\epsilon \leq \hat{U}(x^*) + 4\epsilon \leq U(x^*) + 8\epsilon$.

B.4 Proof of Theorem 4

We reduce from the Knapsack problem: given $v \in [0,1]^n$, $\omega \in \mathbb{R}^n_+, \Omega, V \in \mathbb{R}_+$, decide whether there exists $y \in \{0,1\}^n$ such that $\sum_{i=1}^n v_i y_i \geq V$ and $\sum_{i=1}^n \omega_i y_i \leq \Omega$.

We construct an instance of FDG. Let the set of targets be $N = \{1, \ldots, n+1\}$, and let there be a single binary feature, i.e. $M = \{1\}$ and $x_{i1} \in \{0,1\}$ for each $i \in N$. Since there is only one feature, we abuse the notation by using $x_i = x_{i1}$. Suppose each target's hidden value of the feature is $\hat{x}_i = 0$. Consider a score function f such that $f(0) = 1$ and $f(1) = 2$. For each $i \in N$, let $u_i = \frac{1-v_i}{\delta}$ if $i \neq n+1$, and $u_{n+1} = \frac{1+V+\sum_{i=1}^n v_i}{\delta}$. We chose a large enough $\delta \geq 1$ such that $u_{n+1} \leq 1$. In addition, for each $i \in N$, let $\eta_i = \omega_i$ if $i \neq n+1$, and $\eta_{n+1} = 0$. Finally, let the budget $B = \Omega$.

For a solution y to a Knapsack instance, we construct a solution x to the above FDG where $x_i = y_i$ for $i \neq n+1$, and $x_{n+1} = 0$. We know $\sum_{i\in N} \eta_i |x_i - \hat{x}_i| = \sum_{i\in N} \eta_i x_i \leq B$ if and only if $\sum_{i=1}^n \omega_i y_i \leq \Omega$. Since $f(x_i) > 0$ for all x_i, $\frac{\sum_{i\in N} f(x_i)u_i}{\sum_{i\in N} f(x_i)} \leq 1/\delta$ if and only if $\sum_{i\in N}(1 - \delta u_i)f(x_i) \geq 0$. Note that $\sum_{i\in N}(1-\delta u_i) = \sum_{i=1}^n v_i(y_i + 1) - \sum_{i=1}^n v_i - V$. Thus, y is a certificate of Knapsack if and only if x is feasible for FDG and the defender's expected loss is at most $1/\delta$. □

B.5 Proof of Theorem 5

To analyze the approximation bound of this MILP, we first need to analyze the tightness of the linear approximation.

Consider two points s_1, s_2 where $s_2 - s_1 = \epsilon$. The line segment is $t(s) = \frac{1}{\epsilon}(e^{s_2} - e^{s_1})s - \frac{1}{\epsilon}(e^{s_2} - e^{s_1})s_1 + e^{s_1}$. Let $\Delta(s)$ be the ratio between the line and e^s on the interval $[s_1, s_2]$. It is easy to find that $\Delta(s)$ is maximized at

$$s^* = 1 + s_1 - \frac{\epsilon}{e^\epsilon - 1},$$

with

$$\Delta(s^*) = \frac{\frac{e^\epsilon - 1}{\epsilon}}{\exp\{1 - \frac{\epsilon}{e^\epsilon - 1}\}}.$$

Now, let $v = \frac{e^\epsilon - 1}{\epsilon}$. It is known that $v \in [1, 1+\epsilon]$ when $\epsilon < 1.7$. Note that $\delta(x^*) = v \exp\{\frac{1}{v} - 1\} \leq 1 + (v-1)^2/2$, which holds for all $v \geq 1$. Let $\hat{f}(\cdot)$ be the piecewise linear approximation. For any target i and observable feature configuration x_i, we have

$$\frac{\hat{f}(x_i)}{f(x_i)} \leq v \leq 1 + \frac{\epsilon^2}{2}.$$

Let x^* be the optimal observable features against the true score function f, and let x' be the optimal observable features to the above MILP. Let $U(\cdot)$ be the defender's expected loss, and $\hat{U}(\cdot)$ be the approximate defender's expected loss. For any observable feature configuration x, we have

$$\begin{aligned}
|\hat{U}(x) - U(x)| &= \left| \frac{\sum_i \hat{f}(x_i) u_i}{\sum_i \hat{f}(x_i)} - \frac{\sum_i f(x_i) u_i}{\sum_i f(x_i)} \right| \\
&= \left| \frac{\sum_i \hat{f}(x_i) u_i}{\sum_i \hat{f}(x_i)} - \frac{\sum_i \hat{f}(x_i) u_i}{\sum_i f(x_i)} + \frac{\sum_i \hat{f}(x_i) u_i}{\sum_i f(x_i)} - \frac{\sum_i f(x_i) u_i}{\sum_i f(x_i)} \right| \\
&\leq \frac{2}{\sum_i f(x_i)} \left| \sum_i f(x_i) - \sum_i \hat{f}(x_i) \right| = 2 \left(\frac{\sum_i \hat{f}(x_i)}{\sum_i f(x_i)} - 1 \right) \\
&\leq \epsilon^2
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
U(x') - U(x^*) &= U(x') - \hat{U}(x') + \hat{U}(x') - U(x^*) \\
&\leq U(x') - \hat{U}(x') + \hat{U}(x^*) - U(x^*) \\
&\leq 2\epsilon^2
\end{aligned}$$

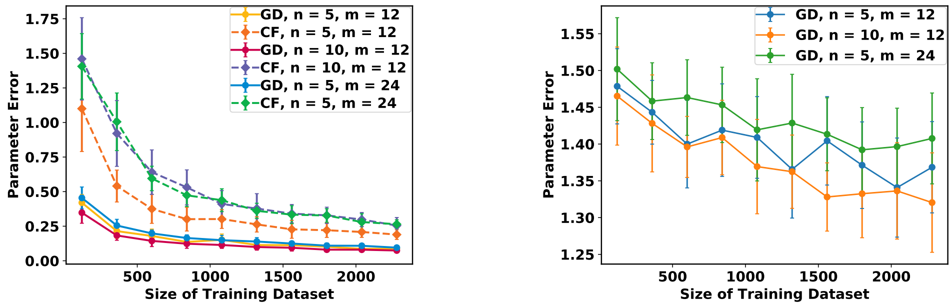
□

B.6 Proof of Theorem 6

Suppose binary search terminates with interval of length $U - L \leq \epsilon_{bs}$, and observable features x^{bs}. Both x^{bs} and the optimal observable features x' to the MILP lie in this interval. This means $\hat{U}(x^{bs}) - \hat{U}(\hat{x}) \leq \epsilon_{bs}$. Recall that x^* is the optimal observable features against the true score function f. Therefore, we have

$$\begin{aligned}
U(x^{bs}) - U(x^*) &= U(x^{bs}) - \hat{U}(x^{bs}) + \hat{U}(x^{bs}) - U(x^*) \\
&\leq U(x^{bs}) - \hat{U}(x^{bs}) + \hat{U}(\hat{x}) + \epsilon_{bs} - U(x^*) \\
&\leq U(x^{bs}) - \hat{U}(x^{bs}) + \hat{U}(x^*) + \epsilon_{bs} - U(x^*) \\
&\leq 2\epsilon^2 + \epsilon_{bs}
\end{aligned}$$

□



(a) Learning adversary's preferences, 1-layer score function

(b) Learning adversary's preferences, 3-layer score function

Figure 2: Experimental results

C Additional Experiments

In addition to the mean total variation distance reported in the main text, we present another metric to measure the performance of learning. We consider $|\hat{\theta} - \theta|$, the L_1 error in the score function parameter θ, which directly relates to the sample complexity bound in Theorem 1. Since the dimension of θ depends on the number of features k and other factors, we consider the L_1 error divided by the number of parameters and report this metric in Fig. 2a and Fig. 2b.

For a single layer score function, the log-likelihood is concave. Thus GD is expected to find the global maximizer. Indeed, we see that in Fig 2a, the learning error is close to zero, which corroborates this claim. The L_1 error for CF also decreases as the sample size increases, though not as small as GD. According to Theorem 1 we would need much more samples than 2000 to achieve an error of 0.25.

For complex score function, the learning error is larger as shown in Fig. 2b, even though Fig. 1b in the main text shows the total variation distance is small. This suggests that the loss surface for complex score function is, true to its name, more complex. Comparing Fig. 2a- 2b with Fig. 1i- 1k, we can obtain more intuition why the solution gap in Fig. 1k is much larger than that in Fig. 1i.

D Experiment Parameters and Hyper-parameters

Complex score function architecture The 3-layer neural network score function has input layer of size $m \times 24$, second layer 24×12, and third layer 12×1. The first and second layers are followed by a tanh activation, and the last layer is followed by an exponential function. The neural network parameters are initialized uniformly at random in $[-0.5, 0.5]$. We use this network architecture for all of our experiments.

FDG parameters for 1-layer score function We detail in Table 2 the parameter distributions used in the planning and combined learning and planning experiments, when the adversary assumes the single-layer score function. These distributions apply to the results shown in Fig. 1c, 1d, 1i, 1j.

FDG parameters for 3-layer score function We detail in Table 3 the parameter distributions used in the planning and combined learning and planning experiments, when the adversary assumes the 3-layer score function. These distributions apply to the results shown in Fig. 1e,1f, 1g, 1h,1k, 1l.

Hyper-parameters for learning Table 4 shows the hyper-parameters we used in learning the attacker's score function f.

Discrete feature $k \in M_d$		Continuous feature $k \in M_c$	
Variable	Distribution	Variable	Distribution
$\lvert M_d \rvert$	$2m/3$	$\lvert M_c \rvert$	$m/3$
η_{ik}	$U(-3, 3)$	η_{ik}	$U(0, 3)$
τ_{ik}	N/A	τ_{ik}	$U(0, 0.25)$
$\hat{x}_{ik}$	$U\{0, 1\}$	$\hat{x}_{ik}$	$U(0, 1)$
u_i		$U(0, 1)$	
Variable		Distribution	
B		$U(0, 0.2C_{\max})$	
$C_{\max}$		$\sum_{i \in N} \sum_{k \in M_c} \eta_{ik} \min(\hat{x}_{ik}, 1 - \hat{x}_{ik}, \tau_{ik}) + \sum_{k \in M_d} \eta_{ik}$	

Table 2: FDG parameter distributions for experiments on 1-layer attacker score function. Used in Fig. 1c, 1d, 1i, 1j

Variable	Distribution
η_{ik}	$U(0, 1)$
τ_{ik}	1
$\hat{x}_{ik}$	$U(0, 1)$
u_i	$U(0, 1)$
B	$U(0, 0.2nm)$

Table 3: FDG parameter distributions for experiments on 3-layer attacker score function. Used in Fig. 1e,1f, 1g, 1h,1k, 1l

Parameter	Fig 1k ($\lvert D_{train} \rvert > 10000$), 1l	Fig. 1j	All other experiments
Learning rate	{1e-3, 1e-2, 1e-1} $\rightarrow$ 1e-1	{1e-3, 1e-2, 1e-1} $\rightarrow$ 1e-1	{1e-3, 1e-2, 1e-1} $\rightarrow$ 1e-1
Number of epochs	{20, 30, 60} $\rightarrow$ 30	{20, 30, 60} $\rightarrow$ 30	{10, 20, 40} $\rightarrow$ 20
Steps per epoch	{20, 30, 40} $\rightarrow$ 30	12	{10, 20} $\rightarrow$ 10
Batch size	{120, 600, 5000} $\rightarrow$ 5000	{120, 600, 5000} $\rightarrow$ 5000	$\lvert D_{train} \rvert$/Number of epochs

Table 4: Hyper-parameters for the experiments. The values between the braces are the ones we tested. The values after the arrows are the ones we used in generating the results.