Attacks and Defense on Normal-Form Games and Markov Games

Young Wu

December 8, 2023

Thesis Statement

- There are vulnerabilities in multi-agent systems and attackers can influence the behavior of players of normal-form games or multi-agent reinforcement learners through data or environment poisoning.

Overview

- An attacker minimally modifies the rewards of a normal-form game or a Markov game with the goal of installing a target policy as the unique equilibrium the victims will learn.

1. Planning setting: the victims are directly given the reward matrices.
2. Offline setting: the victims are given a dataset containing realizations of the rewards.
3. Online setting: the victims are given realizations of the rewards during online learning.

List of Projects

- Reward poisoning problems:
- $\boxed{1}$ Mixed-iNash: Planning, zero-sum, stochastic policy target.
- $\boxed{2}$ iNash: Offline, zero-sum, deterministic policy target.
- $\boxed{3}$ iDSE: Offline, general-sum, deterministic policy target.
- $\boxed{4}$ Online-iDSE: Online, general-sum deterministic policy target.

Mixed Nash Attack

- $\boxed{1}$ Joint work ($\approx 75\%$ contribution) with Jeremy McMahan, Yiding Chen, Yudong Chen, Jerry Zhu, Qiaomin Xie.
- Victim setting:

1. The victims are given a normal-form or Markov game, and solves for the (Markov perfect) Nash equilibrium policy, possibly a stochastic (behavioral) policy.

Attacker Setting

- Attacker Setting:

1. The attacker wants the victims to learn a target (possibly stochastic) policy $\pi^{\dagger}$ as the unique (Markov perfect) Nash equilibrium.
2. The attacker can modify the rewards from R^{o} to $R^{\dagger}$.
3. The attacker minimizes the reward modification cost $C\left(R^{\dagger}, R^{o}\right)$, convex in $R^{\dagger}$.

iNash Formulation

- The attack can be formulated as

$$\min_{R^\dagger} C\left(R^\dagger, R^o\right)$$
$$\text{s.t.} R^\dagger \in \text{ iNash } \left(\pi^\dagger\right),$$

where iNash (π) is the inverse Nash set of reward matrices such that π is the unique Nash equilibrium. (The Q functions can be used in place of R for Markov games.)

Characterizing Nash Uniqueness

- For a normal-form game $(R, \mathcal{A})$ and a strategy profile $\pi = (p, q)$ with support $(\mathcal{I}, \mathcal{J})$, define conditions:

Condition

SIISOW (Switch-In Indifferent, Switch-Out Worse):

$$e_{\mathcal{I}}^T Rq = p^T Rq = p^T Re_{\mathcal{J}},$$
$$e_{\mathcal{A}\setminus\mathcal{I}}^T Rq < p^T Rq < p^T Re_{\mathcal{A}\setminus\mathcal{J}}.$$

Condition

INV (Invertibility): $\begin{bmatrix} R_{\mathcal{I}\mathcal{J}} & -1_{|\mathcal{I}|} \\ 1_{|\mathcal{J}|}^T & 0 \end{bmatrix}$ is invertible.

iNash Set

Theorem

SIISOW and INV are sufficient and necessary conditions for a zero-sum game $(R, \mathcal{A})$ *to have a unique Nash* $\pi = (p, q)$ *with support* $(\mathcal{I}, \mathcal{J})$.

- Proof sketch: a zero-sum game can be solved as a linear program, and the uniqueness of its optimal solution can be characterized by strict complementarity (SIISOW) and basic feasibility (INV).

Attacker's Problem

- The attacker's problem can be converted to:

$$\begin{aligned} \min_{R^\dagger} & C\left(R^\dagger, R^o\right) \\ \text{s.t.} & R^\dagger \text{ satisfies SIISOW} \\ & R^\dagger \text{ satisfies INV} \end{aligned}$$

- Other constraints can be added, for example:

1. Reward entry bounds: $R^\dagger \in [-b, b]$,
2. Target range of Nash values: $p^T R^\dagger q \in [\underline{v}, \overline{v}]$.

Feasibility

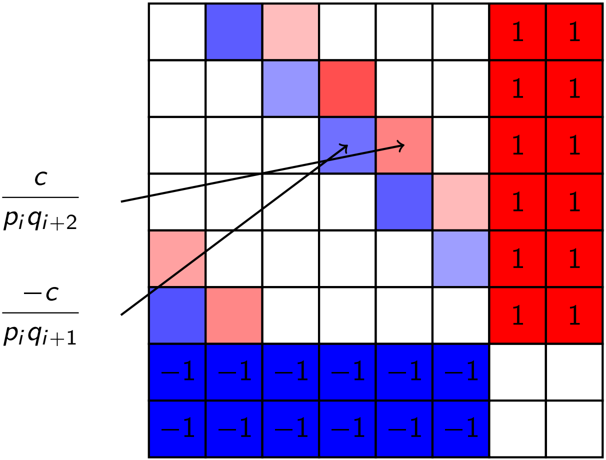
Lemma

A feasible attack $R^\dagger$ exists if and only if $\pi^\dagger$ has equal support sizes, that is $|\mathcal{I}| = |\mathcal{J}|$ and $[\underline{v}, \overline{v}] \cap (-Hb, Hb)$ is non-empty.

- Proof sketch: INV implies equal support sizes, and a translation and scaling of a class of extended rock-paper-scissors games R^{eRPS} guarantees the existence of a feasible attack as long as the target value is in the interior of $[-b, b]$.

Extened Rock Paper Scissors

- A typical R^{eRPS} for $\pi^\dagger = (p, q)$ looks like:



Special Examples

- Special examples of eRPS games include

1. If $p = q = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, then $R^{\text{eRPS}} = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}$ has the unique Nash $\pi^\dagger = (p, q)$.
2. If $p = q = (1, 0, 0)$, then $R^{\text{eRPS}} = \begin{bmatrix} 0 & 1 & 1 \\ -1 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}$ has the unique Nash $\pi^\dagger = (p, q)$.

Non Linear Constraints

- SIISOW and other value constraints are linear and INV constraints are non-linear:

$$\begin{aligned} &\min_{R^\dagger} C\left(R^\dagger, R^o\right) \\ &\text{s.t.}\, R^\dagger \text{ satisfies SIISOW} \\ &\quad R^\dagger \text{ satisfies INV} \\ &\quad R^\dagger \in [-b, b] \text{ and } p^T R^\dagger q \in [\underline{v}, \overline{v}]\,. \end{aligned}$$

Relax and Perturb

- Relax And Perturb (RAP) algorithm:

1. Solve the relaxed problem:

$$\begin{aligned} &\min_{R'} C\left(R', R^{o}\right) \\ &\text{s.t.}\, R' \text{ satisfies SIISOW} \\ &\qquad R' \in (-b, b) \text{ and } p^{T} R' q \in [\underline{v}, \overline{v}] . \end{aligned}$$

2. Perturb the solution:

$$R^{\dagger} = R' + \varepsilon R^{\text{eRPS}}, \text{ with } \varepsilon \sim \text{Unif} .$$

Feasibility of RAP

Proposition

RAP produces a solution $R^{\dagger} = R' + \varepsilon R^{\ eRPS}$ feasible to the original problem with probability 1.

Intuition: if matrix A is not invertible and has eigenvalues λ some of which are 0, then A $+ \varepsilon I$ has eigenvalues $\lambda + \varepsilon$, and if $\varepsilon \sim$ Unif , the eigenvalues are all non-zero with probability 1.

Optimality of RAP

Proposition

RAP produces a near-optimal solution $R^{\dagger} = R' + \varepsilon R^{\,eRPS}$ for sufficiently small relaxation and perturbation parameters and assuming the cost function is Lipschitz.

Intuition: the perturbation preserves SIISOW and other value constraints, so $R^{\dagger}$ is feasible for the original problem, and for small relaxation and perturbation parameters, $R^{\dagger}$ is close to R', implying a near-optimal cost due to the Lipschitz assumption.

Pure Nash Attack

- $\boxed{2}$ Joint work ($\approx 75\%$ contribution) with Jeremy McMahan, Jerry Zhu, Qiaomin Xie. (Thanks: Yudong Chen)
- The victims are uncertainty-aware offline learner of a zero-sum normal-form or Markov game, and estimates a set of plausible games.
- The attacker has to ensure that all games in the set of plausible games have the target policy as unique (Markov perfect) Nash equilibrium.

The Attack Problem

- The attack can be formulated as

$$\min_{r^\dagger} C\left(r^\dagger, r^o\right)$$
$$\text{s.t.} \hat{R}\left(r^\dagger; \rho\right) \subseteq \text{ iNash } \left(\pi^\dagger\right),$$

where $\hat{R}$ represents the set of plausible games given the data $r^\dagger$ with some confidence parameter ρ.

Reduction from $\boxed{1}$

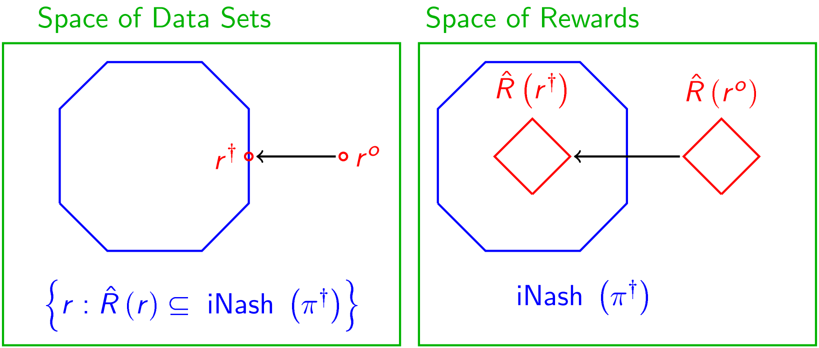
- When $\pi^\dagger = (i, j)$ is deterministic, INV is always satisfied, and SIISOW can be reduced to strict Nash condition:

$$R_{ij'} < R_{ij} < R_{i'j}, \forall\; i' \neq i, j' \neq j.$$

- The attacker's problem can be written as:

$$\begin{aligned} &\min_{r^\dagger} C\left(r^\dagger, r^o\right) \\ &\text{s.t. UCB } \left(\hat{R}_{ij'}\left(r^\dagger\right)\right) < \text{ LCB } \left(\hat{R}_{ij}\left(r^\dagger\right)\right) \\ &\qquad < \text{ UCB } \left(\hat{R}_{ij}\left(r^\dagger\right)\right) < \text{ LCB } \left(\hat{R}_{i'j}\left(r^\dagger\right)\right). \end{aligned}$$

iNash Diagram



Feasibility and Optimization

- The attack problem is feasible if the set of plausible games are sufficiently small, and the problem can be converted into a convex program with linear constraints.

Comparison between $\boxed{1}$ and $\boxed{2}$

- In general, it is impossible to install a stochastic policy $\pi^\dagger$ in the offline data poisoning setting, since $\text{iNash}(\pi^\dagger)$ for stochastic $\pi^\dagger$ is a measure-zero set, but the set of plausible games is usually not.

Dominant Strategy Offline Attack

- $\boxed{3}$ Joint work ($\approx$ 50% contribution) with Jeremy McMahan, Jerry Zhu, Qiaomin Xie. (Thanks: Yudong Chen)
- The settings are similar to the Nash Attack $\boxed{2}$, except there are n victims learning general-sum normal-form or Markov games.

Feasibility and Optimization

- iDSE with (Markov perfect) dominant strategy equilibrium is used in place of iNash: the feasibility conditions are similar, and the attack can also be converted into a convex program with linear constraints.

Modification from $\boxed{2}$

- A strict DSE is defined by:

$$R_{i'j'} < R_{ij'}, \forall\, j' \text{ and } i' \neq i,$$
$$R_{i'j} < R_{i'j'}, \forall\, i' \text{ and } j' \neq j.$$

- The attacker's problem can be written as:

$$\begin{aligned} &\min_{r^\dagger} C\left(r^\dagger, r^o\right) \\ &\text{s.t. } \mathrm{UCB}\ \left(R_{i'j'}\left(r^\dagger\right)\right) < \ \mathrm{LCB}\ \left(R_{i'j}\left(r^\dagger\right)\right), \\ &\qquad \mathrm{UCB}\ \left(R_{ij'}\left(r^\dagger\right)\right) < \ \mathrm{LCB}\ \left(R_{i'j'}\left(r^\dagger\right)\right). \end{aligned}$$

Comparison between $\boxed{2}$ and $\boxed{3}$

- Characterization of general-sum games with a unique Nash equilibrium is difficult, but general-sum games with a unique dominant strategy equilibrium can be characterized by a set of linear constraints, which can be used in the attacker's problem.

Dominant Strategy Online Attack

- $\boxed{4}$ Joint work ($\approx 15\%$ contribution) with Yuzhe Ma (main author), and Jerry Zhu.
- The victims are learning the equilibrium policy of a general-sum bandit game using online no-regret algorithms.
- iDSE is also used, the problem is always feasible, and the attack costs can be sub-linear.

Modification from $\boxed{3}$

- The online victims get bandit feedback, so the attacker only incurs cost when the an action is used.
- Since the online victims use no-regret learning algorithms, the target action profile will be use in all but sub-linear number of rounds.
- Cost minimization in this setting can be further simplified to not changing or minimally changing the rewards from the target action profile.

Comparison between $\boxed{3}$ and $\boxed{4}$

- In $\boxed{1}$, $\boxed{2}$, $\boxed{3}$, game value at the boundary of $[-b, b]$ is not feasible.
- In $\boxed{4}$, game value at the boundary $\{-b, b\}$ is possible due to repeated interactions between the attacker and the victims: the first few iterations can be used to mislead the no-regret victims to choose the target action profile.

Summary

- The attacker installs $\pi^\dagger$ as the unique ...:

Pure $\pi^\dagger$	Zero-sum	General-sum
Planning	NE $\boxed{1}$	DSE $\boxed{3}$
Offline	NE $\boxed{2}$	DSE $\boxed{3}$
Online	NE	DSE $\boxed{4}$

Mixed $\pi^\dagger$	Zero-sum	General-sum
Planning	NE $\boxed{1}$	?
Offline	impossible $\boxed{2}$	impossible
Online	?	?

Thank you!

Game Redesign

- Victim setting:

1. The victims are no-regret online learners with $O\left(T^{\alpha}\right)$ regret, e.g. EXP3.P.
2. The victims participate in an n-player general-sum bandit game with original reward $r^{o}\left(a\right) \in \left[-1, 1\right]^{n}$ for action profile $a = \left(a_1, a_2, \ldots, a_n\right)$.

Attacker Setting

- Attacker setting:

1. The attacker wants the victims to take a target (deterministic) policy $\pi^{\dagger} = \left(\pi_1^{\dagger}, \pi_2^{\dagger}, ..., \pi_n^{\dagger}\right)$ as often as possible, i.e. maximize $\sum_{t=1}^{T} \mathbb{1}_{\left(a_t = \pi^{\dagger}\right)}$.
2. The attacker can modify the rewards that the victims see from $r^o(a)$ to $r^{\dagger}(a)$.
3. The attacker wants sublinear design cost $\sum_{t=1}^{T} \left\| r^o(a_t) - r_t^{\dagger}(a_t) \right\|_p$.

Interior Design Example

- Suppose $\pi^\dagger = (1,1)$, the attacker can redesign the game r^o to $r^\dagger$,

$$r^o = \begin{bmatrix} (0,0) & \left(-1, \boxed{1}\right) & \left(\boxed{1}, -1\right) \\ \left(\boxed{1}, -1\right) & (0,0) & \left(-1, \boxed{1}\right) \\ \left(-1, \boxed{1}\right) & \left(\boxed{1}, -1\right) & (0,0) \end{bmatrix},$$

$$r_1^\dagger = r_2^\dagger = \ldots = \begin{bmatrix} \left(\boxed{0}, \boxed{0}\right) & \left(\boxed{0.1}, -0.1\right) & \left(\boxed{0.1}, -0.1\right) \\ \left(-0.1, \boxed{0.1}\right) & (0,0) & (0,0) \\ \left(-0.1, \boxed{0.1}\right) & (0,0) & (0,0) \end{bmatrix}.$$

Interior Design Algorithm

- Given $r^o(a) \in [-1, 1]$, first consider the interior case when $r^o(\pi^\dagger) > -1$.
- Assumption: $r^o(\pi^\dagger) \geqslant -1 + \rho$, for some $\rho > 0$.
- Attack: $r^\dagger_{i,t}(a) = \begin{cases} r^o_i(\pi^\dagger) + \left(1 - \frac{d(a_t)}{n}\right)\rho & \text{if } a_{i,t} = \pi^\dagger_i \\ r^o_i(\pi^\dagger) - \frac{d(a_t)}{n}\rho & \text{if } a_{i,t} \neq \pi^\dagger_i \end{cases},$

  where $d(a_t) = \sum_{i=1}^{n} \mathbb{1}_{\left\{a_{i,t} = \pi^\dagger_i\right\}}$.

Interior Design Result

Theorem

Using the interior design, $\pi^{\dagger}$ is used $T - O\left(nT^{\alpha}\right)$ times while incurring design cost $O\left(n^{1+1/p}T^{\alpha}\right)$.

- For example, EXP3.P with L_1 cost can achieve $\pi^{\dagger}$ being used $T - O\left(n\sqrt{T}\right)$ times with cost $O\left(n^2\sqrt{T}\right)$.

Interior Design Proof Sketch

- Under this attack, we have,

$$r_{i,t}^{\dagger}(a) = \begin{cases} r_i^o\left(\pi^{\dagger}\right) + \left(1 - \dfrac{d\left(a_t\right)}{n}\right)\rho & \text{if } a_{i,t} = \pi_i^{\dagger} \\ r_i^o\left(\pi^{\dagger}\right) - \dfrac{d\left(a_t\right)}{n}\rho & \text{if } a_{i,t} \neq \pi_i^{\dagger} \end{cases}.$$

1. $\pi^{\dagger}$ is strictly dominant:
   $r_{i,t}^{\dagger}\left(\pi_{i,t}^{\dagger}, a_{-i,t}\right) = r_{i,t}^{\dagger}\left(a_{i,t}, a_{-i,t}\right) + \left(1 - \frac{1}{n}\right)\rho, \forall\, a_{i,t} \neq \pi_{i,t}^{\dagger}$.
2. $\pi^{\dagger}$ rewards remain unchanged: $r_{i,t}^{\dagger}\left(\pi^{\dagger}\right) = r_i^o\left(\pi^{\dagger}\right)$.

- No-regret learners will use the optimal action profile $\pi^{\dagger}$ in all but $O\left(T^{\alpha}\right)$ rounds while incurring $O\left(T^{\alpha}\right)$ design cost.

Boundary Design Example

- When $r^o\left(\pi^\dagger\right) = -1$, it is impossible to decrease other entries below -1: another design is needed.
- Suppose again $\pi^\dagger = (1,1)$, then,

$$r^o = \begin{bmatrix} (-1,-1) & \left(-1,\boxed{1}\right) & \left(\boxed{1},-1\right) \\ \left(\boxed{1},-1\right) & (-1,-1) & \left(-1,\boxed{1}\right) \\ \left(-1,\boxed{1}\right) & \left(\boxed{1},-1\right) & (-1,-1) \end{bmatrix},$$

$$r_1^\dagger \approx \begin{bmatrix} \left(\boxed{-0.8},\boxed{-0.8}\right) & \left(\boxed{-0.7},-0.9\right) & \left(\boxed{-0.7},-0.9\right) \\ \left(-0.9,\boxed{-0.7}\right) & (-1,-1) & (-1,-1) \\ \left(-0.9,\boxed{-0.7}\right) & (-1,-1) & (-1,-1) \end{bmatrix},$$

Boundary Design Example Limit

$$r_1^\dagger \approx \begin{bmatrix} \left(\boxed{-0.8}, \boxed{-0.8}\right) & \left(\boxed{-0.7}, -0.9\right) & \left(\boxed{-0.7}, -0.9\right) \\ \left(-0.9, \boxed{-0.7}\right) & (-1,-1) & (-1,-1) \\ \left(-0.9, \boxed{-0.7}\right) & (-1,-1) & (-1,-1) \end{bmatrix},$$

$$r_2^\dagger \approx \begin{bmatrix} \left(\boxed{-0.9}, \boxed{-0.9}\right) & \left(\boxed{-0.85}, -0.95\right) & \left(\boxed{-0.85}, -0.95\right) \\ \left(-0.95, \boxed{-0.85}\right) & (-1,-1) & (-1,-1) \\ \left(-0.95, \boxed{-0.85}\right) & (-1,-1) & (-1,-1) \end{bmatrix},$$

$$\lim_{t\to\infty} r_t^\dagger = \begin{bmatrix} (-1,-1) & (-1,-1) & (-1,-1) \\ (-1,-1) & (-1,-1) & (-1,-1) \\ (-1,-1) & (-1,-1) & (-1,-1) \end{bmatrix}.$$

Boundary Design Algorithm

- Assumption: $r^o\left(\pi^\dagger\right) = -1$.
- Attack: $r^\dagger_{i,t}(a) = w_t r^\dagger_{i,\text{ interior}}(a) + (1 - w_t)\, r^o\left(\pi^\dagger\right)$, where $w_t = t^{\alpha+\varepsilon-1}$, for some $\varepsilon \in (0, 1-\alpha]$.

Boundary Design Result

Theorem

Using the boundary deisng with $\varepsilon = \dfrac{1-\alpha}{2}, \pi^{\dagger}$ *is used* $T - O\left(nT^{(1+\alpha)/2}\right)$ *times while incurring design cost* $O\left(n^{1/p}\left(1+n\right)T^{(1+\alpha)/2}\right)$.

Boundary Design Proof Sketch

- Under this attack, we have,

$$r_{i,t}^{\dagger}(a) = w_t r_{i,\text{ interior}}^{\dagger}(a) + (1 - w_t)\, r^o\left(\pi^{\dagger}\right), \text{ where } w_t = t^{\alpha+\varepsilon-1}.$$

1. $\pi^{\dagger}$ is strictly dominant:
   $$r_{i,t}^{\dagger}\left(\pi_{i,t}^{\dagger}, a_{-i,t}\right) = r_{i,t}^{\dagger}\left(a_{i,t}, a_{-i,t}\right) - \left(1 - \frac{1}{n}\right)\rho w_t, \forall\ a_{i,t} \neq \pi_{i,t}^{\dagger}.$$
2. $\pi^{\dagger}$ rewards are almost unchanged:
   $$\left\| r_{i,t}^{\dagger}\left(\pi^{\dagger}\right) - r_i^o\left(\pi^{\dagger}\right)\right\|_p \leqslant 2bn^{1/p} w_t.$$

- No-regret learners will use the optimal action profile $\pi^{\dagger}$ in all but $O\left(T^{(1+\alpha)/2}\right)$ rounds while incurring $O\left(T^{(1+\alpha)/2}\right)$ design cost.

Nash Attack

- Victim setting:

1. The victims are uncertainty-aware offline learners that use additive bonus terms β when estimating the Q function, i.e. $Q = \hat{R} - \beta + \mathbb{E}_{\hat{P}}\left[V'\right]$.
2. The victims learn a two-player zero-sum Markov game from a training set $\left\{\left(\left(s_t^{(k)}, a_t^{(k)}, r_t^{(k)}\right)_{t=1}^{T}\right)\right\}_{k=1}^{K}$, with $r_t^{(k)} \in [0, 1]$.

Attacker Setting

- Attacker setting:

1. The attacker wants the victims to learn a target (deterministic) policy $\pi^\dagger$ as the unique Markov perfect (Nash) equilibrium.
2. The attacker can modify the rewards in the training set from r^o to $r^\dagger$.
3. The attacker minimizes the reward modification cost $\left\| r^\dagger - r^o \right\|$, e.g. $\sum_{k=1}^{K} \sum_{t=1}^{T} \left\| r_t^{\dagger,(k)} - r_t^{o,(k)} \right\|_1$.
4. The attacker does not know $\hat{R}$ and $\hat{P}$, but assumes $\left| \hat{R} - R^{(\mathrm{MLE})} \right| < \rho^{(R)}$ and $\left\| \hat{P} - P^{(\mathrm{MLE})} \right\|_1 < \rho^{(P)}$.

iNash Formulation

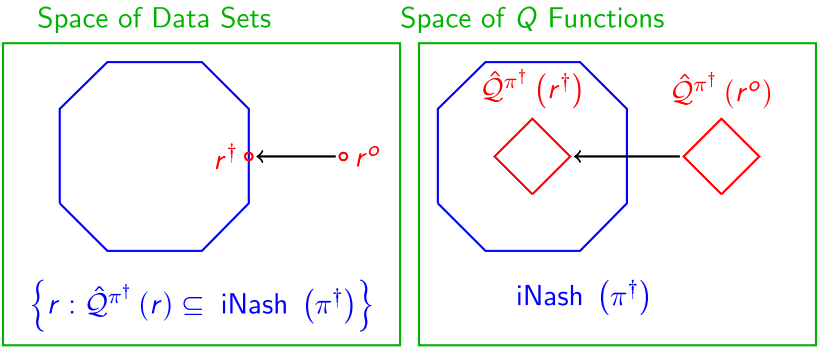
- The attack can be formulated as

$$\min_{r^\dagger} \left\| r^\dagger - r^o \right\|$$
$$\text{s.t.} \hat{\mathcal{Q}}^{\pi^\dagger}\left(r^\dagger; \rho^{(R)}, \rho^{(P)}\right) \subseteq \text{ iNash }\left(\pi^\dagger\right),$$

where,

1. $\hat{\mathcal{Q}}^{\pi}(r)$ is the set of plausible Q functions computed based on r evaluated on π,
2. iNash (π) is the inverse Nash polytope of Q functions such that π is the strict Markov perfect (Nash) equilibrium.

iNash Diagram



Feasibility

Theorem

The attack is feasible if $\rho_t^{(R)}(s,a) + |\beta_t(s,a)| < \frac{1}{4T}, \forall\ t, s,$ *and actions* a *such that* $a_1 = \pi_{1,t}^{\dagger}(s)$ *or* $a_2 = \pi_{2,t}^{\dagger}(s)$*.*

- For example, if $\rho^{(R)} = 0$ and $\beta = \frac{c}{\sqrt{N_t(s,a)}}$, then the condition is a data coverage condition, $N_t(s,a) > 16cT^2$ for actions profiles in the same row or column as $\pi^{\dagger}$ in the stage game matrices.

Feasible Example

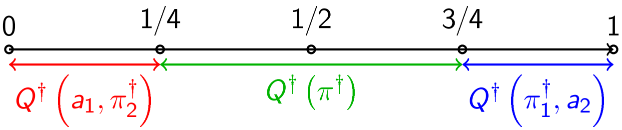
- Suppose $\pi^\dagger = (1, 1)$ in a stage game, then the following attack is feasible under the previous feasibility condition,

$a_1 \backslash a_2$	1	2	3	4
1	0.5	1	1	1
2	0	-	-	-
3	0	-	-	-
4	0	-	-	-

- Unspecified cells' corresponding rewards do not need to be poisoned.

Feasibility Proof Sketch

- The condition $\rho_t^{(R)}(s,a) + |\beta_t(s,a)| < 1/(4T)$ implies that the cumulated confidence interval width for R and P in the future periods is bounded by $1/4$.
- In period t, state s, for every $a_1 \neq \pi_1^\dagger$ and $a_2 \neq \pi_2^\dagger$, the Q values have the following relationship.



- Therefore, $\pi_t^\dagger(s)$ is the strict, thus unique, Nash equilibrium in every stage game (t, s).

Linear Program Formulation

- The attacker's problem is given by,

$$\min_{r^\dagger} \sum_{k=1}^{K} \sum_{t=1}^{T} \left\| r_t^{\dagger,(k)} - r_t^{o,(k)} \right\|_1$$

s.t. for every t, s, and $Q_t^\dagger \in \hat{\mathcal{Q}}^{\pi^\dagger}\left(r^\dagger\right)$,

$$Q_t^\dagger\left(s, \pi_t^\dagger(s)\right) > Q_t^\dagger\left(s, \left(a_1, \pi_{t,2}^\dagger(s)\right)\right), \forall\, a_1 \neq \pi_{t,1}^\dagger(s),$$
$$Q_t^\dagger\left(s, \pi_t^\dagger(s)\right) < Q_t^\dagger\left(s, \left(\pi_{t,1}^\dagger(s), a_2\right)\right), \forall\, a_2 \neq \pi_{t,2}^\dagger(s).$$

- Since $\hat{\mathcal{Q}}^\pi(r)$ are polytopes, this problem can be formulated as a linear program and solved efficiently.

Multi Attacker

- Incomplete, joint work ($\approx 75\%$ contribution) with Elliot Pickens, Jin-Yi Cai, and Jerry Zhu.
- Victim setting:

1. Single or multiple identical victims that estimates the mean $\hat{\mu}$ of a data set, based on a training provided by the attackers.

Attacker Setting, Direction, Continuous

- Attacker Setting 1.1:

1. Each of K attackers has a target direction $x_k^\dagger$ with the goal of minimizing $\left(x_k^\dagger\right)^T \hat{\mu}$.
2. Each attacker creates a training set with X_k from a convex and compact domain X, and the (disjoint) union of $\{X_k\}_{k=1}^K$ is given to the victim.

- The game has a (weakly) dominant strategy equilibrium, in which the attackers choose the most extreme points in X in the $x^\dagger$ directions.

Attacker Setting, Direction, Discrete

- Attacker Setting 1.2:

1. Each of K attackers has a target direction $x_k^\dagger$ with the goal of minimizing $\left(x_k^\dagger\right)^T \hat{\mu}$.
2. Each attacker creates a training set with X_k from n existing points X, and the (disjoint) union of $\{X_k\}_{k=1}^{K}$ is given to the victim.

- The game has a (weakly) dominant strategy equilibrium, in which the attackers choose the most extreme points in X in the $x^\dagger$ directions.

Attacker Setting, Point, Continuous

- Attacker Setting 2.1:

1. Each of K attackers has a target point $x_k^\dagger$ with the goal of minimizing $\left\|x_k^\dagger - \hat{\mu}\right\|^2$.
2. Each attacker creates a training set with X_k from a convex and compact domain X, and the (disjoint) union of $\{X_k\}_{k=1}^K$ is given to the victim.

- The game has at least one pure strategy Nash equilibrium, and it can be found using

1. Best response dynamics,
2. Maximizing a (weakly) concave potential function on convex and compact X.

Attacker Setting, Point, Discrete

- Attacker Setting 2.2:

1. Each of K attackers has a target point $x_k^\dagger$ with the goal of minimizing $\left\| x_k^\dagger - \hat{\mu} \right\|^2$.
2. Each attacker creates a training set with X_k from n existing points X, and the (disjoint) union of $\{X_k\}_{k=1}^K$ is given to the victim.

- The game has at least one pure strategy Nash equilibrium, and it can be found using

1. Best response dynamics,
2. Maximizing a potential function on finite X.

Potential Function Formulation

- The payoff to attacker k can be written as

$$-\left\| x_k^\dagger - \left(x_0 + \sum_{k=1}^{K} x_k \right) \right\|^2,$$

where x_k is the centroid of the points provided by attacker k.

- The potential function is

$$-\sum_{k=1}^{K} \| w_k x_k \|^2$$
$$-2 \sum_{i \neq j} \left(x_i^\dagger - \left(x_0 + \sum_{k \neq i} w_k x_k \right) \right) \left(x_j^\dagger - \left(x_0 + \sum_{k \neq j} w_k x_k \right) \right).$$

References I

[1] Pieter Abbeel and Andrew Y Ng.
Apprenticeship learning via inverse reinforcement learning.
In *Proceedings of the twenty-first international conference on Machine learning*, page 1, 2004.

[2] Natalia Akchurina.
Multiagent reinforcement learning: algorithm converging to nash equilibrium in general-sum discounted stochastic games.
In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 725–732, 2009.

References II

[3] Ashton Anderson, Yoav Shoham, and Alon Altman.
Internal implementation.
In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*, pages 191–198. Citeseer, 2010.

[4] Gautam Appa.
On the uniqueness of solutions to linear programs.
Journal of the Operational Research Society, 53:1127–1132, 2002.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire.
The nonstochastic multiarmed bandit problem.
SIAM journal on computing, 32(1):48–77, 2002.

References III

[6] Avraham Bab and Ronen I Brafman.
Multi-agent reinforcement learning in common interest and fixed sum stochastic games: An experimental study.
Journal of Machine Learning Research, 9(12), 2008.

[7] Kiarash Banihashem, Adish Singla, and Goran Radanovic.
Defense against reward poisoning attacks in reinforcement learning.
arXiv preprint arXiv:2102.05776, 2021.

[8] Kiarash Banihashem, Adish Singla, Jiarui Gan, and Goran Radanovic.
Admissible policy teaching through reward design.
arXiv preprint arXiv:2201.02185, 2022.

References IV

[9] Vahid Behzadan and Arslan Munir.
Vulnerability of deep reinforcement learning to policy induction attacks.
In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer, 2017.

[10] Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Dkebiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al.
Dota 2 with large scale deep reinforcement learning.
arXiv preprint arXiv:1912.06680, 2019.

References V

[11] Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett.
Stochastic linear bandits robust to adversarial attacks.
In *International Conference on Artificial Intelligence and Statistics*, pages 991–999. PMLR, 2021.

[12] Michael Bowling.
Convergence problems of general-sum multiagent reinforcement learning.
In *ICML*, pages 89–94, 2000.

[13] Michael Bowling and Manuela Veloso.
Rational and convergent learning in stochastic games.
In *International joint conference on artificial intelligence*, volume 17, pages 1021–1026. Lawrence Erlbaum Associates Ltd, 2001.

References VI

[14] David Brandfonbrener, Will Whitney, Rajesh Ranganath, and Joan Bruna.
Offline rl without off-policy evaluation.
Advances in Neural Information Processing Systems, 34:4933–4946, 2021.

[15] Noam Brown, Tuomas Sandholm, and Strategic Machine.
Libratus: The superhuman ai for no-limit poker.
In *IJCAI*, pages 5226–5228, 2017.

[16] Noam Brown and Tuomas Sandholm.
Superhuman ai for multiplayer poker.
Science, 365(6456):885–890, 2019.

References VII

[17] Sébastien Bubeck and Nicolò Cesa-Bianchi.
Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
Foundations and Trends® in Machine Learning, 5(1):1–122, 2012.

[18] William J. Clancey.
The Engineering of Qualitative Models.
Forthcoming, 2021.

[19] Mathieu Bouville.
Crime and punishment in scientific research, 2008.

[20] NASA.
Pluto: The 'other' red planet, 2015.
Accessed: 2018-12-06.

References VIII

[21] William J. Clancey.
Transfer of Rule-Based Expertise through a Tutorial Dialogue.
Ph.D. diss., Dept. of Computer Science, Stanford Univ., Stanford, Calif., 1979.

[22] William J. Clancey.
Communication, Simulation, and Intelligent Agents: Implications of Personal Intelligent Machines for Medical Education.
In *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83)*, pages 556–560, Menlo Park, Calif, 1983. IJCAI Organization.

References IX

[23] William J. Clancey.
Classification Problem Solving.
In *Proceedings of the Fourth National Conference on Artificial Intelligence*, pages 45–54, Menlo Park, Calif., 1984. AAAI Press.

[24] Qiwen Cui and Simon S Du.
When is offline two-player zero-sum markov game solvable?
arXiv preprint arXiv:2201.03522, 2022.

[25] Qiwen Cui and Simon S Du.
When is offline two-player zero-sum Markov game solvable?
arXiv preprint arXiv:2201.03522, 2022.

[26] George Dantzig.
Linear programming and extensions.
Princeton university press, 1963.

References X

[27] Gregory Dudek, Michael RM Jenkin, Evangelos Milios, and David Wilkes.
A taxonomy for multi-agent robotics.
Autonomous Robots, 3(4):375–397, 1996.

[28] Robert Engelmore and Anthony Morgan, editors.
Blackboard Systems.
Addison-Wesley, Reading, Mass., 1986.

[29] Wikipedia contributors.
Volunteer's dilemma — Wikipedia, the free encyclopedia, 2021.
[Online; accessed 16-September-2021].

References XI

[30] Wei Fu, Chao Yu, Zelai Xu, Jiaqi Yang, and Yi Wu.
Revisiting some common practices in cooperative multi-agent reinforcement learning.
arXiv preprint arXiv:2206.07505, 2022.

[31] Evrard Garcelon, Baptiste Roziere, Laurent Meunier, Olivier Teytaud, Alessandro Lazaric, and Matteo Pirotta.
Adversarial attacks on linear contextual bandits.
arXiv preprint arXiv:2002.03839, 2020.

[32] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell.
Adversarial policies: Attacking deep reinforcement learning.
arXiv preprint arXiv:1905.10615, 2019.

References XII

[33] RA Good.
f-finger morra.
SIAM Review, 7(1):81–87, 1965.

[34] Shixiang Gu, Ethan Holly, Timothy Lillicrap, and Sergey Levine.
Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates.
In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE, 2017.

References XIII

[35] Ziwei Guan, Kaiyi Ji, Donald J Bucci Jr, Timothy Y Hu, Joseph Palombo, Michael Liston, and Yingbin Liang.
Robust stochastic bandit algorithms under probabilistic unbounded adversarial attack.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4036–4043, 2020.

[36] Wenbo Guo, Xian Wu, Sui Huang, and Xinyu Xing.
Adversarial policy learning in two-player competitive games.
In *International Conference on Machine Learning*, pages 3910–3919. PMLR, 2021.

[37] Sergiu Hart and Andreu Mas-Colell.
A simple adaptive procedure leading to correlated equilibrium.
Econometrica, 68(5):1127–1150, 2000.

References XIV

[38] Diane Warner Hasling, William J. Clancey, and Glenn Rennels.
Strategic explanations for a diagnostic consultation system.
International Journal of Man-Machine Studies, 20(1):3–19, 1984.

[39] Diane Warner Hasling, William J. Clancey, Glenn R. Rennels, and Thomas Test.
Strategic Explanations in Consultation—Duplicate.
The International Journal of Man-Machine Studies, 20(1):3–19, 1983.

[40] Johannes Heinrich, Marc Lanctot, and David Silver.
Fictitious self-play in extensive-form games.
In *International conference on machine learning*, pages 805–813. PMLR, 2015.

References XV

[41] Pablo Hernandez-Leal and Michael Kaisers.
Towards a fast detection of opponents in repeated stochastic games.
In *International Conference on Autonomous Agents and Multiagent Systems*, pages 239–257. Springer, 2017.

[42] GA Heuer.
Uniqueness of equilibrium points in bimatrix games.
International Journal of Game Theory, 8:13–25, 1979.

[43] Junling Hu and Michael P Wellman.
Nash q-learning for general-sum stochastic games.
Journal of machine learning research, 4(Nov):1039–1069, 2003.

References XVI

[44] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel.
Adversarial attacks on neural network policies.
arXiv preprint arXiv:1702.02284, 2017.

[45] Yunhan Huang and Quanyan Zhu.
Deceptive reinforcement learning under adversarial manipulations on cost signals.
In *International Conference on Decision and Game Theory for Security*, pages 217–237. Springer, 2019.

References XVII

[46] Lantao Yu, Jiaming Song, and Stefano Ermon.
Multi-agent adversarial inverse reinforcement learning.
In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7194–7201. PMLR, 09–15 Jun 2019.

[47] Haoqi Zhang and David Parkes.
Value-based policy teaching with active indirect elicitation.
In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 1*, AAAI'08, page 208–214. AAAI Press, 2008.

References XVIII

[48] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al.
Human-level performance in 3d multiplayer games with population-based reinforcement learning.
Science, 364(6443):859–865, 2019.

[49] Jiechuan Jiang and Zongqing Lu.
Offline decentralized multi-agent reinforcement learning.
arXiv preprint arXiv:2108.01832, 2021.

[50] Ying Jin, Zhuoran Yang, and Zhaoran Wang.
Is pessimism provably efficient for offline rl?
In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

References XIX

[51] Ying Jin, Zhuoran Yang, and Zhaoran Wang.
Is pessimism provably efficient for offline rl?
In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[52] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, and Jerry Zhu.
Adversarial attacks on stochastic bandits.
Advances in Neural Information Processing Systems, 31:3640–3649, 2018.

[53] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li.
Trojdrl: evaluation of backdoor attacks on deep reinforcement learning.
In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pages 1–6. IEEE, 2020.

References XX

[54] Jens Kober, J Andrew Bagnell, and Jan Peters.
Reinforcement learning in robotics: A survey.
The International Journal of Robotics Research, 32(11):1238–1274, 2013.

[55] Jernej Kos and Dawn Song.
Delving into adversarial attacks on deep policies.
arXiv preprint arXiv:1705.06452, 2017.

[56] Erich Kutschinski, Thomas Uthmann, and Daniel Polani.
Learning competitive pricing strategies by multi-agent reinforcement learning.
Journal of Economic Dynamics and Control, 27(11-12):2207–2218, 2003.

References XXI

[57] Jae Won Lee and Jangmin O.
A multi-agent q-learning framework for optimizing stock trading systems.
In *International Conference on Database and Expert Systems Applications*, pages 153–162. Springer, 2002.

[58] Jae Won Lee, Jonghun Park, O Jangmin, Jongwoo Lee, and Euyseok Hong.
A multiagent approach to *q*-learning for daily stock trading.
IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 37(6):864–877, 2007.

References XXII

[59] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel.
Multi-agent reinforcement learning in sequential social dilemmas.
arXiv preprint arXiv:1702.03037, 2017.

[60] Xiaomin Lin, Peter A Beling, and Randy Cogill.
Multiagent inverse reinforcement learning for two-person zero-sum games.
IEEE Transactions on Games, 10(1):56–68, 2017.

[61] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun.
Tactics of adversarial attack on deep reinforcement learning agents.
arXiv preprint arXiv:1703.06748, 2017.

References XXIII

[62] Xiaomin Lin, Stephen C Adams, and Peter A Beling.
Multi-agent inverse reinforcement learning for certain general-sum stochastic games.
Journal of Artificial Intelligence Research, 66:473–502, 2019.

[63] Michael L Littman.
Markov games as a framework for multi-agent reinforcement learning.
In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

[64] Fang Liu and Ness Shroff.
Data poisoning attacks on stochastic bandits.
In *International Conference on Machine Learning*, pages 4042–4050. PMLR, 2019.

References XXIV

[65] Yong Liu, Yujing Hu, Yang Gao, Yingfeng Chen, and Changjie Fan.
Value function transfer for deep multi-agent reinforcement learning based on n-step returns.
In *IJCAI*, pages 457–463, 2019.

[66] Guanlin Liu and Lifeng Lai.
Action-manipulation attacks on stochastic bandits.
In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3112–3116. IEEE, 2020.

References XXV

[67] Guanlin Liu and Lifeng Lai.
Provably efficient black-box action poisoning attacks against reinforcement learning.
Advances in Neural Information Processing Systems, 34, 2021.

[68] Yunlong Lu and Kai Yan.
Algorithms in multi-agent systems: a holistic perspective from reinforcement learning and game theory.
arXiv preprint arXiv:2001.06487, 2020.

[69] Shiyin Lu, Guanghui Wang, and Lijun Zhang.
Stochastic graphical bandits with adversarial corruptions.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8749–8757, 2021.

References XXVI

[70] Thodoris Lykouris, Max Simchowitz, Alex Slivkins, and Wen Sun.
Corruption-robust exploration in episodic reinforcement learning.
In *Conference on Learning Theory*, pages 3242–3245. PMLR, 2021.

[71] Yuzhe Ma, Kwang-Sung Jun, Lihong Li, and Xiaojin Zhu.
Data poisoning attacks in contextual bandits.
In *International Conference on Decision and Game Theory for Security*, pages 186–204. Springer, 2018.

[72] Yuzhe Ma, Xuezhou Zhang, Wen Sun, and Jerry Zhu.
Policy poisoning in batch reinforcement learning and control.
Advances in Neural Information Processing Systems, 32:14570–14580, 2019.

References XXVII

[73] Yuzhe Ma, Young Wu, and Xiaojin Zhu.
Game redesign in no-regret game playing.
arXiv preprint arXiv:2110.11763, 2021.

[74] Liam MacDermed, Charles Isbell, and Lora Weiss.
Markov games of incomplete information for multi-agent reinforcement learning.
In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[75] Olvi Mangasarian.
Uniqueness of solution in linear programming.
Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1978.

References XXVIII

[76] Patrick Mannion, Karl Mason, Sam Devlin, Jim Duggan, and Enda Howley.
Dynamic economic emissions dispatch optimisation using multi-agent reinforcement learning.
In *Proceedings of the Adaptive and Learning Agents workshop (at AAMAS 2016)*, 2016.

[77] Eric Maskin and Jean Tirole.
Markov perfect equilibrium: I. observable actions.
Journal of Economic Theory, 100(2):191–219, 2001.

References XXIX

[78] Linghui Meng, Muning Wen, Yaodong Yang, Chenyang Le, Xiyun Li, Weinan Zhang, Ying Wen, Haifeng Zhang, Jun Wang, and Bo Xu.
Offline pre-trained multi-agent decision transformer: One big sequence model conquers all starcraftii tasks.
arXiv preprint arXiv:2112.02845, 2021.

[79] CB Millham.
Constructing bimatrix games with special properties.
Naval Research Logistics Quarterly, 19(4):709–714, 1972.

[80] Junichi Minagawa.
On the uniqueness of nash equilibrium in strategic-form games.
Journal of Dynamics & Games, 7(2):97, 2020.

References XXX

[81] Lin Yang, Mohammad Hajiesmaili, Mohammad Sadegh Talebi, John Lui, and Wing Shing Wong.
Adversarial bandits with corruptions: Regret lower bound and no-regret algorithm.
In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[82] Xiaomin Lin, Stephen C. Adams, and Peter A. Beling.
Multi-agent inverse reinforcement learning for certain general-sum stochastic games.
Journal of Artificial Intelligence Research, 66:473–502, oct 2019.

References XXXI

[83] Sharada Mohanty, Erik Nygren, Florian Laurent, Manuel Schneider, Christian Scheller, Nilabha Bhattacharya, Jeremy Watson, Adrian Egli, Christian Eichenberger, Christian Baumberger, et al.
Flatland-rl: Multi-agent reinforcement learning on trains.
arXiv preprint arXiv:2012.05893, 2020.

[84] Dov Monderer and Moshe Tennenholtz.
k-implementation.
In *Proceedings of the 4th ACM conference on Electronic Commerce*, pages 19–28, 2003.

[85] Dov Monderer and Moshe Tennenholtz.
k-implementation.
Journal of Artificial Intelligence Research, 21:37–62, 2004.

References XXXII

[86] John F Nash Jr.
Equilibrium points in n-person games.
Proceedings of the national academy of sciences, 36(1):48–49, 1950.

[87] Norihiko Ono and Kenji Fukumoto.
A modular approach to multi-agent reinforcement learning.
In *Workshop on Learning in Distributed Artificial Intelligence Systems, Workshop on Learning, Interaction, and Organization in Multiagent Environments*, pages 25–39. Springer, Berlin, Heidelberg, 1997.

[88] Martin J Osborne.
An introduction to game theory, volume 3.
Oxford university press New York, 2004.

References XXXIII

[89] Ling Pan, Longbo Huang, Tengyu Ma, and Huazhe Xu.
Plan better amid conservatism: Offline multi-agent reinforcement learning with actor rectification.
In *International Conference on Machine Learning*, pages 17221–17237. PMLR, 2022.

[90] Anay Pattanaik, Zhenyi Tang, Shuijing Liu, Gautham Bommannan, and Girish Chowdhary.
Robust deep reinforcement learning with adversarial attacks.
arXiv preprint arXiv:1712.03632, 2017.

[91] HL Prasad and Shalabh Bhatnagar.
A study of gradient descent schemes for general-sum stochastic games.
arXiv preprint arXiv:1507.00093, 2015.

References XXXIV

[92] HL Prasad, Prashanth LA, and Shalabh Bhatnagar.
Two-timescale algorithms for learning nash equilibria in general-sum stochastic games.
In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1371–1379, 2015.

[93] Luis G Quintas.
Uniqueness of Nash equilibrium points in bimatrix games.
Center for Mathematical Studies in Economics and Management Science., 1988.

[94] Arthur L. Robinson.
New ways to make microcircuits smaller.
Science, 208(4447):1019–1022, 1980.

References XXXV

[95] Arthur L. Robinson.
New Ways to Make Microcircuits Smaller—Duplicate Entry.
Science, 208:1019–1026, 1980.

[96] James Rice.
Poligon: A System for Parallel Problem Solving.
Technical Report KSL-86-19, Dept. of Computer Science, Stanford Univ., 1986.

[97] TES Raghavan.
Zero-sum two-person games.
Handbook of game theory with economic applications, 2:735–768, 1994.

References XXXVI

[98] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla.
Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning.
In *International Conference on Machine Learning*, pages 7974–7984. PMLR, 2020.

[99] Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla.
Policy teaching in reinforcement learning via environment poisoning attacks.
Journal of Machine Learning Research, 22(210):1–45, 2021.

References XXXVII

[100] Amin Rakhsha, Xuezhou Zhang, Xiaojin Zhu, and Adish Singla.
Reward poisoning in reinforcement learning: Attacks against unknown learners in unknown environments.
arXiv preprint arXiv:2102.08492, 2021.

[101] Anshuka Rangi, Long Tran-Thanh, Haifeng Xu, and Massimo Franceschetti.
Saving stochastic bandits from poisoning attacks via limited data verification.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8054–8061, 2022.

References XXXVIII

[102] Anshuka Rangi, Haifeng Xu, Long Tran-Thanh, and Massimo Franceschetti.
Understanding the limits of poisoning attacks in episodic reinforcement learning.
In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3394–3400. International Joint Conferences on Artificial Intelligence Organization, 7 2022.
Main Track.

[103] Tummalapalli Sudhamsh Reddy, Vamsikrishna Gopikrishna, Gergely Zaruba, and Manfred Huber.
Inverse reinforcement learning for decentralized non-cooperative multiagent systems.
In *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1930–1935. IEEE, 2012.

References XXXIX

[104] Martin Riedmiller, Thomas Gabel, Roland Hafner, and Sascha Lange.
Reinforcement learning for robot soccer.
Autonomous Robots, 27:55–73, 2009.

[105] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel.
Adversarial attacks on neural network policies, 2017.

[106] Jernej Kos and Dawn Song.
Delving into adversarial attacks on deep policies, 2017.

[107] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua.
Safe, multi-agent, reinforcement learning for autonomous driving.
arXiv preprint arXiv:1610.03295, 2016.

References XL

[108] Lloyd S Shapley.
Stochastic games.
Proceedings of the national academy of sciences, 39(10):1095–1100, 1953.

[109] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al.
Mastering the game of go with deep neural networks and tree search.
nature, 529(7587):484–489, 2016.

References XLI

[110] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, and Adrian Bolton.
Mastering the game of go without human knowledge.
nature, 550(7676):354–359, 2017.

[111] Aleksandrs Slivkins.
Introduction to multi-armed bandits.
arXiv preprint arXiv:1904.07272, 2019.

[112] Nobuo Suematsu and Akira Hayashi.
A multiagent reinforcement learning algorithm using extended optimal response.
In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: Part 1*, pages 370–377, 2002.

References XLII

[113] Jianwen Sun, Tianwei Zhang, Xiaofei Xie, Lei Ma, Yan Zheng, Kangjie Chen, and Yang Liu.
Stealthy and efficient adversarial attacks against deep reinforcement learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5883–5891, 2020.

[114] Yanchao Sun, Da Huo, and Furong Huang.
Vulnerability-aware poisoning mechanism for online rl with unknown dynamics.
arXiv preprint arXiv:2009.00774, 2020.

References XLIII

[115] Yanchao Sun, Ruijie Zheng, Yongyuan Liang, and Furong Huang.
Who is the strongest enemy? towards optimal and efficient evasion attacks in deep rl.
arXiv preprint arXiv:2106.05087, 2021.

[116] Peter Szilágyi.
On the uniqueness of the optimal solution in linear programming.
Revue d'analyse numérique et de théorie de l'approximation, 35(2):225–244, 2006.

References XLIV

[117] Rustam Tagiew.
Hypotheses about typical general human strategic behavior in a concrete case.
In *Congress of the Italian Association for Artificial Intelligence*, pages 476–485. Springer, 2009.

[118] Ming Tan.
Multi-agent reinforcement learning: Independent vs. cooperative agents.
In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

References XLV

[119] J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al.
Pettingzoo: Gym for multi-agent reinforcement learning.
Advances in Neural Information Processing Systems, 34:15032–15043, 2021.

[120] Emanuel Tewolde.
Game transformations that preserve Nash equilibria or best response sets.
arxiv preprint arXiv:2111.00076, 2023.

References XLVI

[121] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, and Petko Georgiev.
Grandmaster level in starcraft ii using multi-agent reinforcement learning.
Nature, 575(7782):350–354, 2019.

[122] Sergey Vorotnikov, Konstantin Ermishin, Anaid Nazarova, and Arkady Yuschenko.
Multi-agent robotic systems in collaborative robotics.
In *International Conference on Interactive Collaborative Robotics*, pages 270–279. Springer, 2018.

References XLVII

[123] Xingyu Wang and Diego Klabjan.
Competitive multi-agent inverse reinforcement learning with sub-optimal demonstrations.
In *International Conference on Machine Learning*, pages 5143–5151. PMLR, 2018.

[124] Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy.
Optimism in reinforcement learning with generalized linear function approximation.
arXiv preprint arXiv:1912.04136, 2019.

References XLVIII

[125] Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song.
Backdoorl: Backdoor attack against competitive reinforcement learning.
arXiv preprint arXiv:2105.00579, 2021.

[126] Chen-Yu Wei, Christoph Dann, and Julian Zimmert.
A model selection approach for corruption robust reinforcement learning.
In *International Conference on Algorithmic Learning Theory*, pages 1043–1096. PMLR, 2022.

[127] Stephen J Wright.
Numerical optimization.
New York, NY: Wiley, 2006.

References XLIX

[128] Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li.
Crop: Certifying robust policies for reinforcement learning through functional smoothing.
arXiv preprint arXiv:2106.09292, 2021.

[129] Fan Wu, Linyi Li, Chejian Xu, Huan Zhang, Bhavya Kailkhura, Krishnaram Kenthapadi, Ding Zhao, and Bo Li.
Copa: Certifying robust policies for offline reinforcement learning against poisoning attacks.
arXiv preprint arXiv:2203.08398, 2022.

References L

[130] Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie.
Reward poisoning attacks on offline multi-agent reinforcement learning.
In *The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*, 2023.

[131] Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie.
On faking a Nash equilibrium.
arXiv preprint arXiv:2306.08041, 2023.

References LI

[132] Young Wu, Jeremy McMahan, Xiaojin Zhu, and Qiaomin Xie.
Reward poisoning attacks on offline multi-agent reinforcement learning.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10426–10434, 2023.

[133] Qiaomin Xie, Yudong Chen, Zhaoran Wang, and Zhuoran Yang.
Learning zero-sum simultaneous-move markov games using function approximation and correlated equilibrium.
In *Conference on learning theory*, pages 3674–3682. PMLR, 2020.

References LII

[134] Yang Yang, Li Juntao, and Peng Lingling.
Multi-robot path planning based on a deep reinforcement learning dqn algorithm.
CAAI Transactions on Intelligence Technology, 5(3):177–183, 2020.

[135] Haoqi Zhang and David C Parkes.
Value-based policy teaching with active indirect elicitation.
In *AAAI*, volume 8, pages 208–214, 2008.

[136] Haoqi Zhang, David C Parkes, and Yiling Chen.
Policy teaching through reward function learning.
In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 295–304, 2009.

References LIII

[137] Xuezhou Zhang, Yuzhe Ma, Adish Singla, and Xiaojin Zhu.
Adaptive reward-poisoning attacks against reinforcement learning.
In *International Conference on Machine Learning*, pages 11225–11234. PMLR, 2020.

[138] Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Mingyan Liu, Duane Boning, and Cho-Jui Hsieh.
Robust deep reinforcement learning against adversarial perturbations on state observations.
Advances in Neural Information Processing Systems, 33:21024–21037, 2020.

[139] Xuezhou Zhang, Yiding Chen, Jerry Zhu, and Wen Sun.
Corruption-robust offline reinforcement learning.
arXiv preprint arXiv:2106.06630, 2021.

References LIV

[140] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar.
Multi-agent reinforcement learning: A selective overview of theories and algorithms.
Handbook of Reinforcement Learning and Control, pages 321–384, 2021.

[141] Xuezhou Zhang, Yiding Chen, Xiaojin Zhu, and Wen Sun.
Robust policy gradient against strong data corruption.
In *International Conference on Machine Learning*, pages 12391–12401. PMLR, 2021.

[142] Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C Parkes, and Richard Socher.
The ai economist: Improving equality and productivity with ai-driven tax policies.
arXiv preprint arXiv:2004.13332, 2020.

References LV

[143] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang.
Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets.
arXiv preprint arXiv:2202.07511, 2022.

[144] Shiliang Zuo.
Near optimal adversarial attack on ucb bandits.
arXiv preprint arXiv:2008.09312, 2020.

[145] Avrim Blum and Yishay Mansour.
Learning, regret minimization, and equilibria.
Algorithmic Game Theory, 2007.