Technical Appendix for "Deep Reinforcement Learning for Equilibrium Computation in Multi-Stage Auctions and Contests"

Fabian R. Pieroth, [1,*] Nils Kohring,[1] Martin Bichler, [1]

October 29, 2025

[1]School of Computation, Information and Technology, Technical University of Munich
(*) corresponding author fabian.pieroth@tum.de

A Verification Procedure

Let us outline our detailed methodology for verifying approximate equilibria in continuous multi-stage games. We start with some technical preliminaries and proceed with presenting the verification procedure formally.

A.1 Preliminaries

In Definition 1, we defined a multi-stage game with continuous signals and actions. In what follows, we summarize some additional assumptions required for bounding the error of our verifier.

Throughout, we assume the game to have *perfect recall*. This means that players remember all the information they received, and in particular, their actions. This assumption greatly simplifies theory and is usually made in literature (Myerson and Reny, 2020).

Definition A.1 (Perfect recall)**.** *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game. It is said to have* perfect recall *if for all $it \in L$ and $r > t$, there are measurable functions $\Psi : S_{ir} \to S_{it}$ and $\psi : S_{ir} \to A_{it}$ such that $\Psi(\sigma_{ir}(a_{<r})) = \sigma_{it}(a_{<t})$ and $\psi(\sigma_{ir}(a_{<r})) = a_{it}$, for all $a \in \mathcal{A}$.*

Under perfect recall, one can extract all of i's actions and received signals up to that point from a signal $s_{ir} \in S_{ir}$. That is, there exist functions $\psi_{irt} : S_{ir} \to \mathcal{A}_{it}$ and $\Psi_{irt} : S_{ir} \to S_{it}$ such that $\psi_{irt}(\sigma_{ir}(a_{<r})) = a_{it}$ and $\Psi_{irt}(\sigma_{ir}(a_{<r})) = \sigma_{it}(a_{<t})$ for all $a_{<r} = (a_{\cdot 1}, \cdots, a_{\cdot r-1}) \in \mathcal{A}_{<r}$. Denote with $\psi_{ir} = (\psi_{ir1}, \psi_{ir2}, \ldots, \psi_{irr-1})$ and $\Psi_{ir} = (\Psi_{ir1}, \Psi_{ir2}, \ldots, \Psi_{irr-1})$ the corresponding mappings from S_{ir} into $\mathcal{A}_{i<r}$ and from S_{ir} into $S_{i<r}$ respectively.

Later on, we are interested in two restrictions of the strategy spaces and their intersection. The first restriction allows only pure strategies, i.e., strategies that map onto Dirac measures for a given signal, which we denote by

$$\Sigma_{it}^{\mathrm{p}} := \{\beta \in \Sigma_{it} \mid \beta(s_{it}) = \delta_a \text{ for } s_{it} \in S_{it} \text{ and } a \in \mathcal{A}_{it}\}. \tag{1}$$

We can identify this with the set of measurable functions from signals to actions. With slight abuse of notation, we denote both sets by Σ_{it}^{p} and note where it does not become clear from context. We set $\Sigma_i^{\mathrm{p}} = \bigtimes_{t \in T} \Sigma_{it}^{\mathrm{p}}$.

For verification, we are interested in whether one can achieve the same best-response utility by restricting the space of a single agent to pure strategies. This is often satisfied under mild assumptions. For example, it is satisfied in most Bayesian games (Milgrom and Weber, 1985; Hosoya and Yu, 2022). Furthermore, it can often be guaranteed to be viable whenever a pure strategy equilibrium exists (Horst, 2005; Reny, 2011).

The second restriction is to consider the set of Lipschitz continuous functions, which we denote by

$$\Sigma_{it}^{\mathrm{Lip}} = \{\beta_{it} \in \Sigma_{it} \mid \beta_{it} : S_{it} \to \Delta(\mathcal{A}_{it}) \text{ is Lipschitz continuous in } d_W\}, \tag{2}$$

where d_W denotes the Wasserstein distance (Definition E.1). We set $\Sigma_i^{\mathrm{Lip}} = \bigtimes_{t \in T} \Sigma_{it}^{\mathrm{Lip}}$. Note that common function approximators for distributional strategies, such as neural networks, fall into the space of Lipschitz continuous strategies.

Finally, we consider the intersection of pure and Lipschitz continuous strategies, which we denote by $\Sigma_{it}^{\mathrm{Lip,\,p}} := \Sigma_{it}^{\mathrm{Lip}} \cap \Sigma_{it}^{\mathrm{p}}$, and $\Sigma_i^{\mathrm{Lip,\,p}} := \bigtimes_{t \in T} \Sigma_{it}^{\mathrm{Lip,\,p}}$.

For a given $\beta_{\cdot t}$, one can define a probability distribution from stage t to $t+1$. Let $B \subset \mathcal{A}_{\cdot t}$ be measurable and $a_{<t} \in \mathcal{A}_{<t}$, then the mapping P_t defines a transition probability from $\mathcal{A}_{<t}$ into the set of measurable subsets of $\mathcal{A}_{\cdot t}$ by

$$P_t(B \mid a_{<t}, \beta_{\cdot t}) = p_t(B_{0t} \mid a_{<t}) \prod_{i \in \mathcal{N}} \beta_{it}(B_{it} \mid \sigma_{it}(a_{<t})), \tag{3}$$

where $\beta_{it}(B_{it} \mid \sigma_{it}(a_{<t}))$ is the probability that player i takes actions from B_{it} when receiving the signal $\sigma_{it}(a_{<t})$. The players need to reason about several stages. Therefore, we inductively define probabilities that describe events from the beginning up to a certain stage.

Let $P_{<1}(\{\emptyset\}|\beta) = 1$, and for all $t \in T$ and measurable $B \subset \mathcal{A}_{<t+1}$, define the rollout measure up to stage t under strategy profile β as

$$P_{<t+1}(B|\beta) = \int_{\mathcal{A}_{<t}} P_t\left(\{a_{\cdot t} : (a_{<t}, a_{\cdot t}) \in B\} \mid a_{<t}, \beta_{\cdot t}\right) dP_{<t}\left(a_{<t} \mid \beta\right). \tag{4}$$

Intuitively speaking, this defines the probability of an intermediate history B up to stage t to occur when all players act according to β. The probability measure $P(\cdot \mid \beta) := P_{<T+1}(\cdot \mid \beta)$ denotes the probability measure over outcomes induced by the strategy profile β.

Finally, player i's ex-ante utility is defined as the expected utility over all possible outcomes of the game and is given by

$$\tilde{u}_i\left(\beta\right) = \int_{\mathcal{A}} u_i(a) dP(a \mid \beta). \tag{5}$$

In the game's interim stages, the individual agent reasons about what action to take after receiving a signal. To describe this optimization problem, we introduce the conditional probabilities and utilities to describe the expected utility given a certain signal and strategies. Let $\beta \in \Sigma$ be a strategy profile, $it \in L$, and $Z \subset S_{it}$ measurable, then define

$$P_{it}(Z \mid \beta) = P_{<t}(\sigma_{it}^{-1}(Z) \mid b) = P_{<t}(\{a_{<t} : \sigma_{it}(a_{<t}) \in Z\} \mid \beta), \tag{6}$$

to be the probability that player i's stage t signal is in Z under strategy profile β. Consequently, for any $it \in L$ and measurable $Z \subset S_{it}$ such that $P_{it}(Z|\beta) > 0$, we define, with a slight abuse of notation, the conditional probabilities as

$$P_{<t}(B \mid Z, \beta) = P_{<t}(B \cap \sigma_{it}^{-1}(Z) \mid \beta)/P_{it}(Z|\beta) \quad \forall B \subset \mathcal{A}_{<t} \text{ measurable}, \tag{7}$$

and

$$P(B \mid Z, \beta) = P(\{a \in B : \sigma_{it}(a_{<t}) \in Z\} \mid \beta)/P_{it}(Z|\beta) \quad \forall B \subset \mathcal{A} \text{ measurable}. \tag{8}$$

With all of this, the conditional expected utilities for a measurable $Z \subset S_{it}$ are defined as

$$\tilde{u}_i(\beta \mid Z) = \int_{\mathcal{A}} u_i(a) dP(a \mid Z, \beta). \tag{9}$$

A.2 Construction

This section gives a formal construction of the verification algorithm. For the reader's convenience, we repeat some explanations from the main body here. The verification procedure of an agent's learned strategy consists of two main parts. First, the strategy space must be discretized such that the search space is reduced to a finite size, and the number of simulations must be set such that the expected utilities (across nature and the opponents' action probabilities) can be approximated via sampling. Second, the deployment of all possible strategies is simulated, and the best-performing strategy with the highest utility is compared to the utility of the actual strategy.

A.2.1 Finite precision step-functions

For every $it \in L$ and $\mathrm{D} \in \mathbb{N}$ there exists a finite number of disjoint grid cells that consist of a product of half-open intervals, partitioning S_{it}. We denote these grid cells by C_{it}^k, where $1 \leq k \leq G_{S_{it}}^{\mathrm{D}}$ and $G_{S_{it}}^{\mathrm{D}} \in \mathbb{N}$ is the number of grid cells. Finally, let the tuple $\mathcal{G}_{it} = \left(S_{it}^{\mathrm{D}}, \mathcal{A}_{it}^{\mathrm{D}}, C_{it}, \mathrm{D}\right)$ denote the signal and action space discretization for $it \in L$.

We define the set of step functions with precision D for $it \in L$ by

$$\Sigma_{it}^{\mathrm{D}}\left(\mathcal{G}_{it}\right) := \left\{ s_{it} \mapsto \sum_{k=1}^{G_{S_{it}}^{\mathrm{D}}} \chi_{C_{it}^k}(s_{it}) a_{it}^k \;\middle|\; a_{it}^k \in \mathcal{A}_{it}^{\mathrm{D}} \right\}, \tag{10}$$

and $\Sigma_i^{\mathrm{D}}\left(\mathcal{G}_i\right) = \bigtimes_{t \in T} \Sigma_{it}^{\mathrm{D}}\left(\mathcal{G}_{it}\right)$. Any grid so that Σ_{it}^{D} can approximate any pure Lipschitz continuous function well for sufficiently high D (Lemma E.4) can be used. In the following, we restrict ourselves to the regular grid and show that it satisfies this property. Therefore, we drop the discretization and write Σ_i^{D} instead of $\Sigma_i^{\mathrm{D}}\left(\mathcal{G}_{it}\right)$. For any given finite precision $\mathrm{D} \in \mathbb{N}$, we make the *discretization error* (Equation 7).

A.2.2 Backward induction over finite precision step functions

For every finite precision D, there are finitely many elements in Σ_i^{D}, which can be translated into finitely many decision points for player i. That is, we can build a finite game or decision tree representing all possible step functions from Σ_i^{D}. We perform a backward induction scheme on this finite decision tree to get the maximal ex-ante utility from any step function.

To achieve this, we define player i's *counterfactual conditional utility* as the conditional utility for taking a specific action given a certain signal, excluding player i's influence of reaching this signal. This is similar to formulations of counterfactual reach probabilities and utilities from literature in finite games (Zinkevich et al., 2007). Before the counterfactual conditional utilities can be formally defined, we need to introduce some other objects first.

We exclude player i's influence of reaching a certain signal grid cell $C_{it}^k \subset S_{it}$ by considering a strategy that deterministically plays to reach C_{it}^k. That is possible because, without loss of generality, one can assume that there exists a unique sequence of grid actions $a_{i<t}^{C_{it}^k} \in \mathcal{A}_{i<t}^{\mathrm{D}}$ that need to be taken for every grid cell C_{it}^k. To see this, note that due to perfect recall, agent i's actions taken prior to stage t can be extracted from every signal s_{it}. Due to our construction, this is a unique sequence for every grid cell C_{it}^k if, for example, each action space $\mathcal{A}_{ir}$ gets appended to the signaling space in the following stage S_{ir+1}. Therefore, for every C_{it}^k, there exist $s_{it} \in C_{it}^k$ and $a_{i<t} \in \mathcal{A}_{i<t}^{\mathrm{D}}$ such that $\psi_{it}(s_{it}) = a_{i<t}$. At the same time, there exists no $s'_{it} \in C_{it}^k$ such that $\psi_{it}(s'_{it}) \neq a_{i<t}$ and $\psi_{it}(s'_{it}) \in \mathcal{A}_{i<t}^{\mathrm{D}}$. Therefore, we define functions $\psi_{it}^{\mathrm{D}}(C_{it}^k) = a_{i<t}^{C_{it}^k}$ that map a grid cell to its unique history of grid actions.

Given a finite precision step function strategy $\beta_i \in \Sigma_i^{\mathrm{D}}$, we can now construct a strategy for player i that plays to reach C_{it}^k (adapting stages $1, \ldots, t-1$), takes a certain action in stage t, and remains the same for stages $t+1, \ldots, T$. More specifically, let $it \in L$, $\beta = (\beta_i, \beta_{-i})$ be a strategy profile with $\beta_i \in \Sigma_i^{\mathrm{D}}$ and $\beta_{-i} \in \Sigma_{-i}$, and $a_{it}^k \in \mathcal{A}_{it}^{\mathrm{D}}$. Then we define a function $(\beta_i)^{C_{it}^k, a_{it}^k} = \left((\beta_{i1})^{C_{it}^k, a_{it}^k}, \ldots, (\beta_{iT})^{C_{it}^k, a_{it}^k}\right)$ that is playing to reach C_{it}^k in the following way:

$$\begin{aligned}
(\beta_{ir})^{C_{it}^k, a_{it}^k} &= \beta_{ir} \quad \text{for } r > t, \\
(\beta_{it})^{C_{it}^k, a_{it}^k}(s_{it}) &= \begin{cases} a_{it}^k & \text{for } s_{it} \in C_{it}^k, \\ \beta_{it}(s_{it}) & \text{for } s_{it} \in S_{it} \setminus C_{it}^k, \end{cases} \\
(\beta_{i<t})^{C_{it}^k, a_{it}^k}(\Psi_{it}(s_{it})) &= \begin{cases} \psi_{it}^{\mathrm{D}}(C_{it}^k) & \text{for } s_{it} \in C_{it}^k, \\ \beta_{i<t}\left(\Psi_{it}(s_{it})\right) & \text{for } s_{it} \in S_{it} \setminus C_{it}^k, \end{cases}
\end{aligned}$$

The counterfactual conditional utilities for precision D are then defined as

$$\tilde{u}_i^{\mathrm{c},\,\mathrm{D}}\left(\beta \,|\, C_{it}^k, a_{it}^k\right) = \tilde{u}_i\left((\beta_i)^{C_{it}^k, a_{it}^k}, \beta_{-i} \,|\, C_{it}^k\right), \tag{11}$$

where $\tilde{u}_i(\cdot \mid \cdot)$ is the conditional utility defined in Equation 9. Note that $\tilde{u}_i^{\mathrm{c},\,\mathrm{D}}$ is independent of $\beta_{i<t} \in \Sigma_{i<t}^{\mathrm{D}}$, as only histories conditioned on observing signals from C_{it}^k are considered. All actions that may be taken off paths that lead to this set of signals do not matter. Therefore, we write $\tilde{u}_i^{\mathrm{c},\,\mathrm{D}}\left(\beta_{i>t}, \beta_{-i} \,|\, C_{it}^k, a_{it}^k\right)$ instead of $\tilde{u}_i^{\mathrm{c},\,\mathrm{D}}\left(\beta \,|\, C_{it}^k, a_{it}^k\right)$, where $\beta_{i>t} = (\beta_{it+1}, \ldots, \beta_{iT})$, to emphasize this independence.

We are now ready to define a best response over the finite strategy set Σ_i^{D} via backward induction. For a given opponent strategy profile β_{-i}, we inductively define a step function $\beta_i^{\mathrm{D},*} \in \Sigma_i^{\mathrm{D}}$. For the last stage $s_{iT} \in S_{iT}$, define

$$\beta_{iT}^{\mathrm{D},*}(s_{iT}) = \operatorname*{arg\,max}_{a_{iT} \in \mathcal{A}_{iT}^{\mathrm{D}}} \tilde{u}_i^{\mathrm{c},\,\mathrm{D}}\left(\beta_{-i} \,|\, C_{iT}^k, a_{iT}\right), \tag{12}$$

where C_{iT}^k is the unique set such that $s_{iT} \in C_{iT}^k$. For preceding stages $t < T$, we define

$$\beta_{it}^{\mathrm{D},*}(s_{it}) = \operatorname*{arg\,max}_{a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}} \tilde{u}_i^{\mathrm{c,\,D}} \left(\beta_{i>t}^{\mathrm{D},*}, \beta_{-i} \,|\, C_{it}^k, a_{it} \right), \tag{13}$$

again with $s_{it} \in C_{it}^k$. Note that the $\arg\max_{a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}}$ is non-empty, as the utility functions are bounded, and there are only finitely many values to consider. If there is more than one element in the $\arg\max_{a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}}$, then we simply choose one discrete action for a whole grid cell C_{it}^k. This backward induction procedure gives us a best-response over the set Σ_i^{D} (Lemma E.2).

A.2.3 Monte-Carlo integration for conditional utilities

The backward induction procedure above assumes that the conditional expected utilities from Equations 12 and 13 can be evaluated, which is, in general, impossible. It would require having access to the expectations of the conditional utilities (Equation 9) for which there may be no closed-form solutions. Therefore, we employ Monte-Carlo approximation to estimate $\beta_i^{\mathrm{D},*}$ and its ex-ante utility. We separate the approximation into a simulation and an aggregation phase.

We start with the simulation phase by sampling a single initial game state, and the players receive their respective signals $s_{\cdot\,1}$. We collect the opponent actions a_{-i1} according to β_{-i1}. For player i, we register into which grid cell C_{i1}^k the signal s_{i1} belongs and increase the cell's counter (aka. visitation count), which we denote by $M(C_{i1}^k)$. Then, we simulate the transition to the next stage for every possible action $a_{i1} \in \mathcal{A}_{i1}^{\mathrm{D}}$, multiplying the number of simulated games by a factor of $|\mathcal{A}_{i1}^{\mathrm{D}}|$. We proceed in this pattern; for each simulation, collect the opponent actions according to β_{-it}, register the corresponding grid cell C_{it}^k for player i's signal s_{it}, increase a respective counter $M(C_{it}^k)$, and simulate the state transition for every possible action $a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}$ multiplying the number of simulated games by a factor of $|\mathcal{A}_{it}^{\mathrm{D}}|$.[1] After T stages, there are $\prod_{t \in T} |\mathcal{A}_{it}^{\mathrm{D}}|$ complete histories a, for which the utility $u_i(a)$ is evaluated. This procedure is performed for $M_{\mathrm{IS}} \in \mathbb{N}$ initial states, resulting in a total of $M_{\mathrm{Tot}} = M_{\mathrm{IS}} \cdot \prod_{t \in T} |\mathcal{A}_{it}^{\mathrm{D}}|$ simulated histories and evaluated utilities, concluding the simulation phase. We denote the set of all simulated histories by $A_{M_{\mathrm{IS}}}$.

After performing M_{Tot} simulations, the aggregation phase starts. Depending on which subsets of $A_{M_{\mathrm{IS}}}$ are chosen, we get samples from different distributions. For example, let $\beta_i \in \Sigma_i^{\mathrm{D}}$ arbitrary. Consider the rollout procedure above with a single initial state. Then, as we explore every discrete action, there exists a simulated history a^l that is consistent with β_i. That is, $a_{it}^l = \beta_i(\sigma_{it}(a_{<t}^l))$ for $1 \leq t \leq T$. Due to construction, we have that $a^l \sim P(\,\cdot \,|\, \beta_i, \beta_{-i})$. Therefore, for every initial state, we get at least one sample for every possible $\beta_i \in \Sigma_i^{\mathrm{D}}$.

To perform the backward induction procedure as described above, we want to sample from conditional measures as well. So, for a grid cell $C_{it}^k \subset S_{it}$ and discretized action $a_{it}^k \in \mathcal{A}_{it}^{\mathrm{D}}$, let $\beta_i \in \Sigma_i^{\mathrm{D}}$ such that $\beta_i = (\beta_i)^{C_{it}^k, a_{it}^k}$. That is, β_i is playing to reach C_{it}^k and then plays a_{it}^k. The above procedure allows us to sample $a^l \sim P(\,\cdot \,|\, \beta_i, \beta_{-i})$. Suppose $P_{it}(C_{it}^k \,|\, \beta_i, \beta_{-i}) > 0$ and we only consider those $\tilde{a}^l$ with $\sigma_{it}(\tilde{a}_{<t}^l) \in C_{it}^k$, then $\tilde{a}^l \sim P(\,\cdot \,|\, C_{it}^k, (\beta_i, \beta_{-i}))$. The set of simulated histories $\tilde{a}^l$ for grid cell C_{it}^k, discrete action $a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}$, and step functions $\beta_{i>t}^{\mathrm{D}} \in \Sigma_{>t}^{\mathrm{D}}$ is given by

$$\begin{aligned} A\left(\beta_{i>t}^{\mathrm{D}}, C_{it}^k, a_{it}; M_{\mathrm{IS}}\right) := \big\{ a^l \in A_{M_{\mathrm{IS}}} \,\big|\, & \sigma_{it}(a_{<t}^l) \in C_{it}^k, a_{it}^l = a_{it}, \\ & \beta_{it+m}^{\mathrm{D}}\left(\sigma_{it+m-1}\left(a_{<t+m}^l\right)\right) = a_{it+m}^l \text{ for } 1 \leq m \leq T-t \big\}. \end{aligned}$$

It holds that $\left|A\left(\beta_{i>t}^{\mathrm{D}}, C_{it}^k, a_{it}; M_{\mathrm{IS}}\right)\right| = M(C_{it}^k)$ for any $a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}$ and $\beta_{i>t}^{\mathrm{D}} \in \Sigma_{>t}^{\mathrm{D}}$. That is, we get a valid sample from $P(\,\cdot \,|\, C_{it}^k, ((\beta_{i>t})^{C_{it}^k, a_{it}^k}, \beta_{-i}))$ whenever a simulation falls into C_{it}^k for every discrete action a_{it}^k. We define the estimated counterfactual conditional utility by

$$\hat{u}_i^{\mathrm{c,\,D}}\left(\beta_{i>t}^{\mathrm{D}}, \beta_{-i} \,\middle|\, A_t\left(\beta_{i>t}^{\mathrm{D}}, C_{it}^k, a_{it}; M_{\mathrm{IS}}\right)\right) := \frac{1}{M(C_{it}^k)} \sum_{l=1}^{M(C_{it}^k)} u_i\left(a^l\right), \tag{14}$$

for $\beta_{i>t}^{\mathrm{D}} \in \Sigma_{i>t}^{\mathrm{D}}$, $a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}$, grid cell C_{it}^k, and $a^l \in A\left(\beta_{i>t}^{\mathrm{D}}, C_{it}^k, a_{it}; M_{\mathrm{IS}}\right)$. If $M(C_{it}^k) = 0$, we set the value to zero.

This approximates the counterfactual conditional utility from Equation 12. We use these to construct a step function $\beta_i^{\mathrm{D},M_{\mathrm{IS}}} \in \Sigma_i^{\mathrm{D}}$ according to the backward induction procedure from Equations 12 and 13. From this, using the relation between the counterfactual conditional and ex-ante utilities (see Lemma E.1), we get an estimated best response utility over the simulations $A_{M_{\mathrm{IS}}}$ which we define by

$$\hat{u}_i^{\mathrm{ver,\,D}}(\beta_{-i} \,|\, A_{M_{\mathrm{IS}}}) := \sum_{k=1}^{G_{S_{i1}}^{\mathrm{D}}} \frac{M(C_{i1}^k)}{\sum_j M(C_{i1}^j)} \hat{u}_i^{\mathrm{c,\,D}}\left(\beta_{i>1}^{\mathrm{D},M_{\mathrm{IS}}}, \beta_{-i} \,\middle|\, A_1\left(\beta_{i>1}^{\mathrm{D},M_{\mathrm{IS}}}, C_{i1}^k, \beta_{i1}^{\mathrm{D},M_{\mathrm{IS}}}(C_{i1}^k); M_{\mathrm{IS}}\right)\right). \tag{15}$$

[1] One can decrease this branching factor sometimes using game-specific knowledge. For example, if an agent loses in the first stage of the signaling contest, he or she may no longer bid in the second stage.

In the limit, the approximation recovers the maximum utility and best response over the set of step function Σ_i^{D} (see Lemma E.3). The *simulation error* $\varepsilon_{M_{\mathrm{IS}}}$ (Equation 8) denotes the discrepancy between ex-ante utilities of the best finite-precision step function evaluated analytically and approximated over the dataset $A_{M_{\mathrm{IS}}}$.

Finally, let $\beta \in \Sigma$ be a strategy profile. Then, we simulate M_{IS} complete histories from $P(\cdot \mid \beta)$, and collect them into a data set $B_{M_{\mathrm{IS}}}$. Using Equation 4, we obtain an estimation of the expected utility, which we denote by $\hat{u}_i(\beta|B_{M_{\mathrm{IS}}})$. The final estimation of our verification procedure for the utility loss over pure Lipschitz continuous strategies is then given by

$$\ell^{\text{ver}}(\beta) := \hat{u}_i^{\text{ver, D}}(\beta_{-i} \mid A_{M_{\mathrm{IS}}}) - \hat{u}_i(\beta|B_{M_{\mathrm{IS}}}). \tag{16}$$

B Hyperparameters

We employ common hyperparameters for our experiments, utilizing fully connected neural networks with two hidden layers, each consisting of 64 nodes, and employing SeLU activations on the inner nodes (Klambauer et al., 2017). The weights and biases of these networks determine the parameters θ_i. All experiments were conducted on a single Nvidia GeForce 2080Ti GPU with 11 gigabytes of RAM, accommodating a parallel simulation of $20,000$ environments. We employed the ADAM optimizer. We tuned the learning rate and the initial log-standard deviation for the individual settings. We used a learning rate of 8×10^{-6} for all experiments in the sequential auction and a learning rate of 5×10^{-5} for all experiments of the Bertrand competition. We set the initial log-standard deviation for the sequential auction and Bertrand competition to -3.0 for all experiments. For the elimination contest, we used the same learning rate and initial log-standard deviation as in the sequential auction environment as default. However, we made the following adaptations for the individual settings. We set the initial log-standard deviation in the standard setting where the winning bids are published to -2.0 for REINFORCE. We set the learning rate with risk-averse contestants for PPO to 6×10^{-6}. Finally, we set the learning rate to 1×10^{-5} and the initial log-standard deviation to -2.0 for the REINFORCE algorithm in the setting with asymmetric contestants. All remaining parameters are set to the default values used in the framework by Raffin et al. (2021).

For our verification procedure, we employ a discretization parameter of $\mathrm{D} = 6$ and an initial number of simulations $M_{\mathrm{IS}} = 2^{23}$ as default. We increase the discretization to $\mathrm{D} = 8$ in the Stackelberg Bertrand competition and reduce it to $\mathrm{D} = 4$ for the four-stage Sequential Auction so that the information set tree fits onto a single GPU.

The code is available at Github.

C Analytical Equilibrium Strategies

We report some of the known equilibrium strategies from literature here.

C.1 Elimination Contest

Zhang (2008) specifies the equilibrium strategies for the considered two-stage elimination contest with risk-neutral bidders via the solution of an abstract integral. By choosing a special case with uniform prior distributions, we can solve the related integrals. The equilibrium strategies are provided in the following proposition.

Proposition C.1 (Zhang (2008))**.** *Consider a four-bidder two stage-contest, where bidders privately learn their valuations $v = (v_1, v_2, v_3, v_4)$ for the prize, which are independently and uniformly distributed on the interval $[1.0, 1.5]$. Let i be some bidder, and denote with j the first round's winner of the other group. Then there exists a separating equilibrium for both information cases, which is given by the following.*

1. *Assuming the true valuations are revealed after the first stage, i.e., $\sigma_{i2}(a_{\cdot 1}) = v_j$, we have the following symmetric equilibrium:*

$$\beta_{i1}(v_i) = \mathit{WE}(v_i) \text{ and } \beta_{i2}(v_i, v_j) = \frac{v_i^2 v_j}{(v_i + v_j)^2}.$$

2. *Assuming the winning bids of the other group are revealed after the first stage, i.e., $\sigma_{i2}(a_{\cdot 1}) = a_{1j}$, we have the following equilibrium:*

$$\begin{aligned} \beta_{i1}(v_i) &= \mathit{WE}(v_i) + \mathit{SE}(v_i) \\ \beta_{i2}(v_i, a_{j1}) &= \frac{v_i^2 \beta_{i1}^{-1}(a_{j1})}{\left(v_i + \beta_{i1}^{-1}(a_{j1})\right)^2} \end{aligned}$$

where the functions WE and SE are defined as follows:

$$
\begin{aligned}
WE(v_i) &= 27\,\log\left(v_i+\frac{3}{2}\right)-\frac{17\,v_i}{2}-\frac{43\,\log\left(\frac{5}{2}\right)}{4}+\frac{7\,v_i^2}{2}-2\,v_i^3-4\,\log\left(v_i+1\right)\,\left(v_i^4-1\right)\\
&\quad+4\,\log\left(v_i+\frac{3}{2}\right)\,\left(v_i^4-\frac{81}{16}\right)+7\\
SE(1) &= 0\\
SE(v_i) &= 17\,\log\left(5\right)-8\,\log\left(v_i+1\right)-9\,\log\left(v_i+\frac{3}{2}\right)-17\,\log\left(2\right)-16\,v_i+8\,v_i^2\,\log\left(v_i+1\right)\\
&\quad+16\,v_i^3\,\log\left(v_i+1\right)-16\,v_i^4\,\log\left(v_i+1\right)-8\,v_i^2\,\log\left(v_i+\frac{3}{2}\right)-16\,v_i^3\,\log\left(v_i+\frac{3}{2}\right)\\
&\quad+16\,v_i^4\,\log\left(v_i+\frac{3}{2}\right)-\frac{135}{2\,v_i+3}+18\,v_i^2-8\,v_i^3+33 \textit{ for } v_i\in(1,1.5].
\end{aligned}
$$

C.2 Stackelberg Bertrand Competition

The analytically derived equilibrium strategy for the Stackelberg Bertrand competition by Arozamena and Weinschelbaum (2009) is given in the following proposition.

Proposition C.2 (Arozamena and Weinschelbaum (2009))**.** *Consider a two-firm Stackelberg-Bertrand competition as described in Section 6.4.1. Then for every measurable function $f:\mathbb{R}\to\mathbb{R}$ such that $f(x)>x$, an equilibrium is given as follows:*

$$
\begin{aligned}
\beta_{11}^{-1}(p_1) &= p_1-\frac{Q(p_1)(1-F(p_1))}{Q(p_1)F'(p_1)-Q'(p_1)(1-F(p_1))}=\frac{4p_1^3-27p_1^2-24p_1+20}{3p_1^2-18p_1-12},\\
\beta_{22}(c_2,p_1) &= \begin{cases}\min\{p_1,p^M(c_2)\}, & \textit{for } p_1\geq c_2,\\ f(b_1), & \textit{else,}\end{cases}
\end{aligned}
$$

where $p^M(c_2)=\max_{p_2}Q(p_2)(p_2-c_2)=5+\frac{c_2}{2}$ denotes the monopoly price, and β_{11}^{-1} is the leader's inverse equilibrium strategy.

The leader's equilibrium strategy β_{11} is guaranteed to be invertible in the above setting, so we recover it by numerically inverting β_{11}^{-1} from above.

D Stackelberg Betrand Competition under Risk Aversion

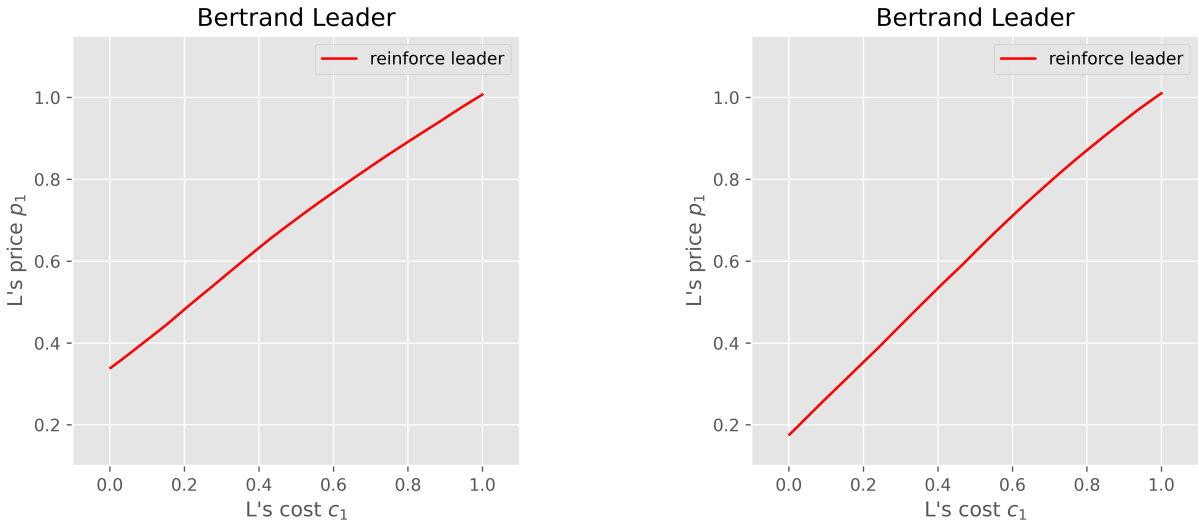
We also study the Stackelberg Bertrand model with risk-averse firms. Wambach (1999) examines how incomplete information and risk aversion affect price-setting behavior in a simultaneous Bertrand competition model. He demonstrates that contrary to the Bertrand paradox, risk-averse firms can sustain prices above the competitive level even as the number of firms increases due to the potential losses associated with cost uncertainty. Ma and Li (2014) examine a supply chain consisting of two manufacturers and a common retailer. The manufacturers first interact in a Bertrand competition, acting as leaders, whereas the retailer reacts to their prices as a follower. They study gradient dynamics of price setting under uncertain demand and risk aversion, finding that the level of risk aversion greatly influences the system's stability. None of the above settings incorporate a Stackelberg interaction and incomplete information under risk aversion. While risk-averse behavior has been extensively studied in the literature, we are unaware of a known equilibrium strategy for our setting.

The results are summarized in Table 1 and show a very small estimated utility loss for all settings. Figure 1 shows the leader's strategy for two different levels of risk aversion, where the left plot displays the strategy for risk parameter $\rho=0.5$ and the right for $\rho=2.0$. Contrary to the prediction made by Wambach (1999), higher levels of risk aversion lead to lower prices and, therefore, higher competition. The follower's strategy aligns with expectations by slightly underbidding the leader's price whenever her cost is sufficiently low.

For a risk parameter of $\rho=2.0$, the exponential nature of CARA risk aversion results in utilities on the order of -10^{14} for the initialized bidding strategies. This scale of rewards is known to cause issues for learning algorithms. Therefore, we performed an additional normalization step of the rewards for the PPO algorithm. Our REINFORCE implementation normalizes rewards by default.

Table 1: Approximated utility losses of the Bertrand competition with risk-averse bidders. There is no analytical equilibrium available for comparison. We report the estimated utility loss ℓ^{ver} with its standard deviations over ten runs. The reward was normalized for the runs indicated with $(*)$.

risk ρ	agent	metric	REINFORCE	PPO
0.5	leader	ℓ^{est}	0.0001 (0.0001)	0.0003 (0.0002)
	follower	ℓ^{est}	-0.0058 (0.0007)	-0.0048 (0.0021)
1.0	leader	ℓ^{est}	0.0001 (0.0001)	0.0001 (0.0001)
	follower	ℓ^{est}	0.0007 (0.0013)	-0.0013 (0.0014)
2.0	leader	ℓ^{est}	0.0000 (0.0000)	0.0001 (0.0001)$(*)$
	follower	ℓ^{est}	0.0052 (0.0011)	-0.0014 (0.0008)$(*)$



(a) risk $\rho = 0.5$

(b) risk $\rho = 2.0$

Figure 1: REINFORCE-based learned strategies of the leader in the Stackelberg Bertrand competition with risk-averse firms.

E Proof of Verifier Convergence Theorem

Before stating the proof of the main theorem, we show several auxiliary results.

Lemma E.1. *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game under Assumption 1, $\beta_{-i} \in \Sigma_{-i}$, and $\beta_i \in \Sigma_i^D$ for $D \in \mathbb{N}$. For a grid cell $C_{it}^k \subset S_{it}$ that C_{it}^k is reachable under β_{-i} and $a_{it}^k \in \mathcal{A}_{it}^D$, consider $J \subset \{1, \dots, G_{S_{it+1}}\}$ such that C_{it+1}^j is reachable from C_{it}^k by taking a_{it}^k under β_{-i}, i.e., $P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) > 0$, then there is the following relationship between the conditional probabilities from stage $t+1$ to stage t:*

$$
\begin{aligned}
&P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) \cdot \tilde{u}_i^{c,D}\left(\beta_i, \beta_{-i} | C_{it}^k, a_{it}^k\right) \\
&= \sum_{j \in J} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{c,D}\left(\beta_i, \beta_{-i} | C_{it+1}^j, \beta_i(C_{it+1}^j)\right)
\end{aligned}
$$

In particular, it holds that

$$
\tilde{u}_i(\beta_i, \beta_{-i}) = \sum_{k=1}^{G_{S_{i1}}^D} P_{i1}\left(C_{i1}^k | \beta_i, \beta_{-i}\right) \cdot \tilde{u}_i^{c,D}\left(\beta_i, \beta_{-i} \,|\, C_{i1}^k, \beta_i(C_{i1}^k)\right).
$$

So, choosing $a_{it}^k = \beta_i(C_{it}^k)$ for every t, we can calculate player i's ex-ante utility $\tilde{u}_i(\beta_i, \beta_{-i})$ by iteratively summing up the conditional probabilities.

Proof. The second statement follows directly from the first by seeing that $G_{S_{i1}^{\mathrm{D}}} = 1$, as there is only a single signal that can be received in stage $t = 1$.

For the first statement, note that due to construction and perfect recall, it holds that $(\beta_i)^{C_{it}^k, a_{it}^k} = (\beta_i)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}$ for all $j \in J$. We then have

$$
\begin{aligned}
&\sum_{j \in J} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,D}}\left(\beta_i, \beta_{-i} | C_{it+1}^j, \beta_i(C_{it+1}^j)\right) && (17)\\
&= \sum_{j \in J} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \int_{\mathcal{A}} u_i(a) dP\left(a | C_{it+1}^j, \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) && (18)\\
&= \sum_{j \in J} \int_{\{a \in \mathcal{A} | \sigma_{it+1}(a_{<t+1} \in C_{it+1}^j)\}} u_i(a) dP\left(a | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) && (19)\\
&= \sum_{j \in J} \int_{\{a \in \mathcal{A} | \sigma_{it+1}(a_{<t+1} \in C_{it+1}^j)\}} u_i(a) dP\left(a | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) && (20)\\
&= \int_{\{a \in \mathcal{A} | \sigma_{it+1}(a_{<t+1} \in \bigcup_{j \in J} C_{it+1}^j)\}} u_i(a) dP\left(a | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) && (21)\\
&= P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) \cdot \tilde{u}_i\left(\left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i} | C_{it}^k\right) && (22)\\
&= P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,D}}\left(\beta_i, \beta_{-i} | C_{it}^k, a_{it}^k\right), && (23)
\end{aligned}
$$

where we used the definitions of the counterfactual conditional utilities and the conditional measures in Equations 18 and 19. Finally, we used that the C_{it+1}^j's are disjoint and $P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right) = \sum_{j \in J} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it}^k, a_{it}^k}, \beta_{-i}\right)$ in Equations 21 and 22 respectively. □

The following lemma establishes that the constructed step function $\beta_i^{\mathrm{D},*}$ defined in Equations 12 and 13 maximizes the ex-ante utility over the finite precision step functions from Σ_i^{D}.

Lemma E.2. *For a given multi-stage game $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$, under Assumption 1, opponent strategies $\beta_{-i} \in \Sigma_{-i}$, and precision parameter $D \in \mathbb{N}$, it holds that*

$$
\tilde{u}_i\left(\beta_i^{D,*}, \beta_{-i}\right) = \sup_{\beta_i' \in \Sigma_i^D} \tilde{u}_i\left(\beta_i', \beta_{-i}\right).
$$

Proof. First, note the following property for conditional probabilities. For $\beta_i, \beta_i' \in \Sigma_i^{\mathrm{D}}$ and any grid cell $C_{it}^k \subset S_{it}^k$, it holds that

$$P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, \beta_i(C_{it}^k)}, \beta_{-i}\right) = P_{it}\left(C_{it}^k | \left(\beta_i'\right)^{C_{it}^k, \beta_i'(C_{it}^k)}, \beta_{-i}\right). \tag{24}$$

That is due to two reasons. First, the conditional probabilities of stage t only depend on the strategies prior to stage t. Second, a discrete strategy $(\beta_i)^{C_{it}^k, a_{it}}$ conditioned on grid cell C_{it}^k is independent of $\beta_{i<t}$.

Next, we show that for all $it \in L$, $\beta_i \in \Sigma_i^{\mathrm{D}}$, and $1 \leq k \leq G_{S_{it}}$ the following holds

$$\tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{it}^k, \beta_i(C_{it}^k)\right) \leq \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i^{\mathrm{D},*}, \beta_{-i} | C_{it}^k, \beta_i^{\mathrm{D},*}(C_{it}^k)\right). \tag{25}$$

We perform a proof by induction. Let $k \in \{1, \dots G_{S_{iT}}\}$, then it holds that

$$\begin{aligned}\tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{iT}^k, \beta_i(C_{iT}^k)\right) &\leq \max_{a_{iT} \in \mathcal{A}_{iT}^{\mathrm{D}}} \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{iT}^k, a_{iT}\right) \\ &= \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i^{\mathrm{D},*}, \beta_{-i} | C_{iT}^k, \beta_i^{\mathrm{D},*}(C_{iT}^k)\right).\end{aligned}$$

This can be seen directly as for any C_{it}^k, the counterfactual conditional utility is independent of $\beta_{i<t}$. Suppose Equation 25 holds for $t+1, \dots, T$. Let $k \in \{1, \dots, G_{S_{it}}\}$ and $a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}$. Denote with $J^{a_{it}} \subset \left\{1, \dots, G_{S_{it+1}}\right\}$ the subset of reachable grid cells from cell C_{it}^k by taking action a_{it}. Then we get by Lemma E.1

$$\begin{aligned}
&P_{it}\left(C_{it}^k | \left(\beta_i\right)^{C_{it}^k, \beta_i(C_{it}^k)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{it}^k, \beta_i(C_{it}^k)\right) \\
&= \sum_{j \in J^{\beta_i(C_{it}^k)}} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{it+1}^j, \beta_i(C_{it+1}^j)\right) \\
&\leq \max_{a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}} \sum_{j \in J^{a_{it}}} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{it+1}^j, \beta_i(C_{it+1}^j)\right) \\
&\overset{\text{(IS)}}{\leq} \max_{a_{it} \in \mathcal{A}_{it}^{\mathrm{D}}} \sum_{j \in J^{a_{it}}} P_{it+1}\left(C_{it+1}^j | \left(\beta_i\right)^{C_{it+1}^j, \beta_i(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i^{\mathrm{D},*}, \beta_{-i} | C_{it+1}^j, \beta_i^{\mathrm{D},*}(C_{it+1}^j)\right) \\
&\overset{(*)}{=} \sum_{j \in J^{\beta_i^{\mathrm{D},*}(C_{it}^k)}} P_{it+1}\left(C_{it+1}^j | \left(\beta_i^{\mathrm{D},*}\right)^{C_{it+1}^j, \beta_i^{\mathrm{D},*}(C_{it+1}^j)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i^{\mathrm{D},*}, \beta_{-i} | C_{it+1}^j, \beta_i^{\mathrm{D},*}(C_{it+1}^j)\right) \\
&= P_{it}\left(C_{it}^k | \left(\beta_i^{\mathrm{D},*}\right)^{C_{it}^k, \beta_i^{\mathrm{D},*}(C_{it}^k)}, \beta_{-i}\right) \cdot \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i^{\mathrm{D},*}, \beta_{-i} | C_{it}^k, \beta_i^{\mathrm{D},*}(C_{it}^k)\right),
\end{aligned}$$

where (IS) denotes the induction step. Furthermore, we used Equation 24 and the definition of $\beta_{it}^{\mathrm{D},*}$ from Equation 13 in step (*). By applying Equation 24 again, we get the statement from Equation 25. Finally, by applying Lemma E.1, we get the statement. □

The next lemma shows that the simulation error for a dataset $A_{M_{\mathrm{IS}}}$ goes to zero as the number of simulations increases.

Lemma E.3. *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game, $\beta_{-i} \in \Sigma_{-i}$, assuming Assumptions 1 and 2 hold, $D \in \mathbb{N}$, and $A_{M_{IS}}$ with a number of $M_{IS} \in \mathbb{N}$ initial simulations. It holds that*

$$\lim_{M_{IS} \to \infty} \varepsilon_{M_{IS}} = 0 \text{ almost surely.}$$

Proof. Let $C_{it}^k \subset S_{it}$ be reachable, i.e., $P_{it}(C_{it}^k \,|\, \beta_i', \beta_{-i}) > 0$ for some $\beta_i' \in \Sigma_i^{\mathrm{D}}$, and $a_{it}^k \in \mathcal{A}_{it}^{\mathrm{D}}$ and $\beta_i \in \Sigma_i^{\mathrm{D}}$ be arbitrary. Then it follows that $M(C_{it}^k) \to \infty$ for $M_{\mathrm{IS}} \to \infty$. That is, for any reachable C_{it}^k, we sample infinitely often from the conditional counterfactual measure $P(\,\cdot \,|\, C_{it}^k, (\beta_i)^{C_{it}^k, a_{it}^k}, \beta_{-i})$. Furthermore, it holds that

$$\tilde{u}_i^{\mathrm{c,\,D}}(\beta_i, \beta_{-i} \,|\, C_{it}^k, a_{it}^k) \leq \int_{\mathcal{A}} |u_i(a)| \, dP(a \,|\, C_{it}^k, (\beta_i)^{C_{it}^k, a_{it}^k}, \beta_{-i}) \leq ||u_i||_\infty < \infty.$$

Finally, as $a^l \in A\left(\beta_{i>t}, C_{it}^k, a_{it}^k; M_{\mathrm{IS}}\right)$ is distributed according to $P(\,\cdot \,|\, C_{it}^k, (\beta_i)^{C_{it}^k, a_{it}^k}, \beta_{-i})$, Kolmogorov's law of large numbers holds and

$$\lim_{M_{\mathrm{IS}} \to \infty} \hat{u}_i^{\mathrm{c,\,D}}\left(\beta_{i>t}, \beta_{-i} | A\left(\beta_{i>t}, C_{it}^k, a_{it}^k; M_{\mathrm{IS}}\right)\right) = \tilde{u}_i^{\mathrm{c,\,D}}\left(\beta_i, \beta_{-i} | C_{it}^k, a_{it}^k\right) \text{ almost surely.}$$

In particular, it holds that

$$\lim_{M_{\text{IS}}\to\infty} \hat{u}_i^{\text{c, D}}\left(\beta_{-i}|A\left(C_{iT}^k, \beta_i^{\text{D},M_{\text{IS}}}(C_{iT}^k); M_{\text{IS}}\right)\right) = \tilde{u}_i^{\text{c, D}}\left(\beta_i, \beta_{-i}|C_{iT}^k, \beta_i^{\text{D},*}(C_{iT}^k)\right) \text{ almost surely,}$$

as the utility is independent from the choice of the $\arg\max$ in Equation 12. Therefore, following the backward induction procedure from Equation 13, we get

$$\lim_{M_{\text{IS}}\to\infty} \hat{u}_i^{\text{c, D}}\left(\beta_{i>t}^{\text{D},M_{\text{IS}}}, \beta_{-i}|A\left(\beta_{i>t}^{\text{D},M_{\text{IS}}}, C_{it}^k, \beta_{i>t}^{\text{D},M_{\text{IS}}}(C_{it}^k); M_{\text{IS}}\right)\right) = \tilde{u}_i^{\text{c, D}}\left(\beta_i^{\text{D},*}, \beta_{-i}|C_{it}^k, \beta_i^{\text{D},*}(C_{it}^k)\right) \text{ a. s.}$$

Finally, we get

$$\lim_{M_{\text{IS}}\to\infty} \hat{u}_i^{\text{ver, D}}(\beta_{-i} \mid A_{M_{\text{IS}}}) = \tilde{u}_i\left(\beta_i^{\text{D},*}, \beta_{-i}\right) = \sup_{\beta_i'\in\Sigma_i^{\text{D}}} \tilde{u}_i\left(\beta_i', \beta_{-i}\right) \text{ almost surely,}$$

where we used Lemma E.1 for the first, and Lemma E.2 for the second equation. This gives us the desired statement. □

Next, we show that any pure Lipschitz continuous strategy can be approximated arbitrarily well in the infinity norm $\|\cdot\|_\infty$ for a sufficiently high discretization D.

Lemma E.4. *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game, where Assumption 1 holds. Let $\mathcal{G}_{it} = (S_{it}^D, \mathcal{A}_{it}^D, C_{it}, D)$ be the discretization that results from a regular grid of 2^D points along each dimension of S_{it} and $\mathcal{A}_{it}$. Then for every pure Lipschitz continuous strategy $\beta_{it} \in \Sigma_{it}^{Lip,\, p}$ there exists a sequence $\left\{\beta_{it}^D\right\}_{D\in\mathbb{N}}$ with $\beta_{it}^D \in \Sigma_{it}^D\left(\mathcal{G}_{it}\right)$ such that*

$$\lim_{D\to\infty} \left|\left|\beta_{it} - \beta_{it}^D\right|\right|_\infty = 0.$$

Proof. Let $\beta_{it} \in \Sigma_{it}^{\text{Lip, p}}$. Then there exists an $L > 0$ such that for $s, s' \in S_{it}$ it holds that

$$d_W\left(\beta_{it}(s), \beta_{it}(s')\right) = \left|\left|\beta_{it}(s) - \beta_{it}(s')\right|\right|_\infty \leq L \cdot \left|\left|s - s'\right|\right|.$$

As $\mathcal{A}_{it}$ and S_{it} are compact, there exists a $K > 0$ such that $\left|\left|a - a'\right|\right|_\infty \leq K$ and $\left|\left|s - s'\right|\right|_\infty \leq K$ for all $a, a' \in \mathcal{A}_{it}$ and $s, s' \in S_{it}$. Then, for every $a \in \mathcal{A}_{it}$ and $s \in S_{it}$, there exist $\tilde{a} \in \mathcal{A}_{it}^{\text{D}}$ and $\tilde{s} \in S_{it}^{\text{D}}$ such that $\left|\left|a - \tilde{a}\right|\right|_\infty \leq K2^{-\text{D}}$ and $\left|\left|s - \tilde{s}\right|\right|_\infty \leq K2^{-\text{D}}$.

Remember that $\left(C_{it}^k\right)_{1\leq k\leq G_{S_{it}^{\text{D}}}}$ partitions S_{it}. Let $s \in C_{it}^k$, then there is a unique $s^k \in S_{it}^{\text{D}} \cap C_{it}^k$. Define

$$\beta_{it}^{\text{D}}(s) = \underset{a_{it}^{Dk}\in\mathcal{A}_{it}^D}{\arg\min} \left|\left|a_{it}^k - \beta_{it}(s^k)\right|\right|_\infty \text{ for } s \in C_{it}^k, 1 \leq k \leq G_{S_{it}}^{\text{D}}.$$

Then $\beta_{it}^{\text{D}} \in \Sigma_{it}^{\text{D}}$ for every $\text{D} \in \mathbb{N}$. Finally, we get for all $s \in S_{it}$ that

$$\begin{aligned}\left|\left|\beta_{it}(s) - \beta_{it}^{\text{D}}(s^k)\right|\right|_\infty &\leq \left|\left|\beta_{it}(s) - \beta_{it}(s^k)\right|\right|_\infty + \left|\left|\beta_{it}(s^k) - \beta_{it}^{\text{D}}(s^k)\right|\right|_\infty \\ &\leq L\left|\left|s - s^k\right|\right| + K2^{-\text{D}} \leq K \cdot (1 + L)2^{-\text{D}}.\end{aligned}$$

As K and L are constants, and $2^{-\text{D}} \to 0$ for $\text{D} \to \infty$, we get the statement. □

We now turn to translate a close approximation of a strategy β_i by a step function β_i' into closeness of the outcome distribution in the Wasserstein distance. For completeness, we restate some well-known results about the Wasserstein distance.

Definition E.1 (Wasserstein distance)**.** *(Villani, 2009, p.93) Let (X, d) be a Polish metric space. For any probability measures μ, ν on X, the (1-)Wasserstein distance between μ and ν is defined by*

$$d_W(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_X d(x, y)d\pi(x, y),$$

where $\Pi(\mu, \nu)$ denotes the space of couplings between μ and ν. That is $\pi \in \Pi(\mu, \nu)$ is a probability measure on $X \times X$, such that $\int_X \pi(x, y)dy = \mu(x)$ and $\int_X \pi(x, y)dx = \nu(x)$.

In our applications, we always assume the metric d to be induced by some norm $||\cdot||$. As we consider finite-dimensional spaces by Assumption 1, the norm's choice is irrelevant.

The following lemma is a central result in optimal transport. It establishes a connection between the Wasserstein distance of two measures μ and ν and the difference of Lipschitz continuous functions measured by μ and ν. We use this relation frequently to establish upper bounds among outcome distributions under different strategies.

Lemma E.5 (Kantorovich–Rubinstein duality)**.** *(Villani, 2009, p.59) Let (X, d) be a Polish metric space. For any probability measures μ, ν on X, and $K > 0$, there holds the following equality*

$$d_W(\mu, \nu) = \frac{1}{K} \sup_{||f||_{Lip} \leq K} \left\{ \int f d\mu - \int f d\nu \right\},$$

where $|| \cdot ||_{Lip}$ denotes the Lipschitz norm.

The following theorem states that the Wasserstein distance metrizises weak convergence on the space of outcome distributions. This is central to establish our main result as it states that the outcome distribution's convergence in the Wasserstein distance means that the ex-ante utilities converge as well.

Theorem E.1 (Metrization of weak convergence)**.** *(Villani, 2009, p.96) Let (X, d) be a Polish metric space and $(\mu_k)_{k \in \mathcal{N}}$ is a sequence of measures in $P(X)$, and $\mu \in P(X)$, then the following two statements are equivalent*

1. $d_W(\mu_k, \mu) \to 0$
2. *For all bounded continuous functions $f : X \to \mathbb{R}$, one has*

$$\int f d\mu_k \to \int f d\mu.$$

Using Theorem E.1, it suffices to show that closeness to a Lipschitz continuous strategy translates to a small Wasserstein distance of the outcome distribution. More specifically, we show in Lemma E.10 that under Assumptions 1 and 2, for every $\beta_i \in \Sigma_i^{\text{Lip, p}}$ there exists a sequence $\left(\beta_i^{\text{D}}\right)_{\text{D} \in \mathcal{N}}$ with $\beta_i^{\text{D}} \in \Sigma_i^{\text{D}}$ such that $d_W \left(P\left(\cdot \mid \beta_i, \beta_{-i} \right), P(\cdot \mid \beta_i^{\text{D}}, \beta_{-i}) \right) \to 0$. We show several technical auxiliary lemmas first to establish this result. These intermediate results give relations among the Wasserstein distance with product measures (Lemma E.6), under different strategies in a specific stage (Lemma E.7), and under different strategies up to a specific stage (Lemmas E.8 and E.9).

Lemma E.6. *Let μ_1, ν_1 and μ_2, ν_2 be measures on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively. Define the product measures $\mu := \mu_1 \otimes \mu_2$, $\nu := \nu_1 \otimes \nu_2$. Then the following inequality holds*

$$d_W(\mu, \nu) \leq d_W(\mu_1, \nu_1) + d_W(\mu_2, \nu_2)$$

Proof. By Theorem 4.1 of (Villani, 2009, p.43), there exist optimal couplings π_1, π_2 for μ_1, ν_1 and μ_2, ν_2 respectively, such that

$$d_W(\mu_1, \nu_1) = \int_{\mathbb{R}^n \times \mathbb{R}^n} d_{\mathbb{R}^n}(x_1, x_2) d\pi_1(x_1, x_2) \, , \, d_W(\mu_2, \nu_2) = \int_{\mathbb{R}^m \times \mathbb{R}^m} d_{\mathbb{R}^m}(y_1, y_2) d\pi_2(y_1, y_2).$$

Then $\pi := \pi_1 \otimes \pi_2$ is the trivial coupling for μ and ν, which can be readily checked by

$$\int_{\mathbb{R}^n \times \mathbb{R}^m} \pi(x_1, x_2, y_1, y_2) d(x_1, y_1) = \int_{\mathbb{R}^n} \pi_1(x_1, x_2) dx_1 \int_{\mathbb{R}^m} \pi_2(y_1, y_2) dy_1 = \nu_1(x_2) \nu_2(y_2),$$
$$\int_{\mathbb{R}^n \times \mathbb{R}^m} \pi(x_1, x_2, y_1, y_2) d(x_2, y_2) = \int_{\mathbb{R}^n} \pi_1(x_1, x_2) dx_2 \int_{\mathbb{R}^m} \pi_2(y_1, y_2) dy_2 = \mu_1(x_1) \mu_2(y_1).$$

Therefore, we get

$$\begin{aligned}
d_W(\mu, \nu) &\leq \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} d_{\mathbb{R}^{n+m}}(x_1, y_1, x_2, y_2) d\pi(x_1, x_2, y_1, y_2) \\
&\leq \int_{\mathbb{R}^{n+m} \times \mathbb{R}^{n+m}} d_{\mathbb{R}^n}(x_1, y_1) + d_{\mathbb{R}^m}(x_2, y_2) d\pi(x_1, x_2, y_1, y_2) \\
&= \int_{\mathbb{R}^n \times \mathbb{R}^n} d_{\mathbb{R}^n}(x_1, y_1) d\pi_1(x_1, y_1) + \int_{\mathbb{R}^m \times \mathbb{R}^m} d_{\mathbb{R}^m}(x_2, y_2) d\pi_2(x_2, y_2) \\
&= d_W(\mu_1, \nu_1) + d_W(\mu_2, \nu_2)
\end{aligned}$$

□

Lemma E.7. *Let $\beta_{it}, \beta'_{it} \in \Sigma_{it}^p$, $\beta_{-it} \in \Sigma_{-it}$ and $\beta_{<t} \in \Sigma_{<t}$. Under Assumption 1, it holds that*

$$d_W\left(P_{<t+1}\left(\cdot \mid (\beta_{i<t}, \beta_{it}), (\beta_{-i<t}, \beta_{-it})\right), P_{<t+1}\left(\cdot \mid (\beta_{i<t}, \beta'_{it}), (\beta_{-i<t}, \beta_{-it})\right)\right) \leq ||\beta_{it} - \beta'_{it}||_\infty.$$

Proof. We start with showing the following step first

$$d_W\left(P_t\left(\cdot \mid a_{<t}, \beta_{it}, \beta_{-it}\right), P_t\left(\cdot \mid a_{<t}, \beta'_{it}, \beta_{-it}\right)\right) \leq ||\beta_{it} - \beta'_{it}||_\infty \text{ for all } a_{<t} \in \mathcal{A}_{<t}. \tag{26}$$

Let $a_{<t} \in \mathcal{A}_{<t}$ be arbitrary. Then note first that as β_{it}, β'_{it} are pure strategies, $\beta_{it}(\cdot \mid \sigma_{it}(a_{<t}))$ and $\beta'_{it}(\cdot \mid \sigma_{it}(a_{<t}))$ are Dirac-measures. Therefore, we can use the well-known fact that the Wasserstein distance between these is simply the distance between the points with positive measure (see Example 6.3 in (Villani, 2009, p.94)), i.e.,

$$d_W\left(\beta_{it}(\cdot \mid \sigma_{it}(a_{<t})), \beta'_{it}(\cdot \mid \sigma_{it}(a_{<t}))\right) = ||\beta_{it}(\sigma_{it}(a_{<t})) - \beta'_{it}(\sigma_{it}(a_{<t}))||_\infty \leq ||\beta_{it} - \beta'_{it}||_\infty,$$

where we abused notation and treated β_{it}, β'_{it} once as mapping to Dirac-measures and once as mapping to elements in $\mathcal{A}_{it}$. By Assumption 1, P_t is a product measure on $\mathbb{R}^m$ for some $m \in \mathbb{N}$. Therefore, one can use Lemma E.6 and get

$$\begin{aligned}
&d_W\left(P_t\left(\cdot \mid a_{<t}, \beta_{it}, \beta_{-it}\right), P_t\left(\cdot \mid a_{<t}, \beta'_{it}, \beta_{-it}\right)\right) \\
&\leq d_W\left(\beta_{it}(\cdot \mid \sigma_{it}(a_{<t})), \beta'_{it}(\cdot \mid \sigma_{it}(a_{<t}))\right) \leq ||\beta_{it} - \beta'_{it}||_\infty,
\end{aligned}$$

which shows Equation 26. Consequently, we get

$$\begin{aligned}
&d_W\left(P_{<t+1}\left(\cdot \mid (\beta_{i<t}, \beta_{it}), (\beta_{-i<t}, \beta_{-it})\right), P_{<t+1}\left(\cdot \mid (\beta_{i<t}, \beta'_{it}), (\beta_{-i<t}, \beta_{-it})\right)\right) \\
&= \sup_{||f||_{\text{Lip}} \leq 1} \int_{\mathcal{A}_{<t}} \int_{\mathcal{A}_{\cdot t}} f\left(a_{<t}, a_{\cdot t}\right) dP_t\left(a_{\cdot t} \mid a_{<t}, \beta_{it}, \beta_{-it}\right) \\
&- \int_{\mathcal{A}_{\cdot t}} f\left(a_{<t}, a_{\cdot t}\right) dP_t\left(a_{\cdot t} \mid a_{<t}, \beta'_{it}, \beta_{-it}\right) dP_{<t}\left(a_{<t} \mid \beta_{<t}\right) \\
&\overset{(E.5)}{\leq} \int_{\mathcal{A}_{<t}} d_W\left(P_t\left(\cdot \mid a_{<t}, \beta_{it}, \beta_{-it}\right), P_t\left(\cdot \mid a_{<t}, \beta'_{it}, \beta_{-it}\right)\right) dP_{<t}\left(a_{<t} \mid \beta_{<t}\right) \\
&\overset{Equ.(26)}{\leq} \int_{\mathcal{A}_{<t}} ||\beta_{it} - \beta'_{it}||_\infty dP_{<t}\left(a_{<t} \mid \beta_{<t}\right) = ||\beta_{it} - \beta'_{it}||_\infty.
\end{aligned}$$

□

Lemma E.8. *Let $(X, d_X), (Y, d_Y)$ be metric Polish spaces, $f : X \times Y \to \mathbb{R}$ be a K_f-Lipschitz continuous function and $\mu(\cdot \mid x)$ be a measure on Y for every $x \in X$. Furthermore, the mapping $x \mapsto \mu(\cdot \mid x)$ is K_μ-Lipschitz continuous with respect to the Wasserstein distance d_W. Then it holds that*

$$g_f : X \to \mathbb{R}, x \mapsto \int_Y f(x, y) d\mu(y|x) \text{ is } (K_f + K_f K_\mu) - \text{Lipschitz}.$$

Proof. Let $x, x' \in X$, then

$$\begin{aligned}
&|g_f(x) - g_f(x')| = \left| \int_Y f(x, y) d\mu(y|x) - \int_Y f(x', y) d\mu(y|x') \right| \\
&\leq \left| \int_Y f(x, y) - f(x', y) d\mu(y|x) \right| + \left| \int_Y f(x', y) d\mu(y|x) - \int_Y f(x', y) d\mu(y|x') \right| \\
&\leq \int_Y K_f \cdot d_{X \times Y}\left((x, y)^T, (x', y)^T\right) d\mu(y|x) + \sup_{||g||_{\text{Lip}} \leq K_f} \left| \int_Y g(y) d\mu(y|x) - \int_Y g(y) d\mu(y|x') \right| \\
&\overset{(E.5)}{=} K_f \int_Y d_X(x, x') d\mu(y|x) + K_f \cdot d_W\left(\mu\left(\cdot \mid x\right), \mu\left(\cdot \mid x'\right)\right) \\
&= K_f \left(d_X(x, x') + d_W\left(\mu\left(\cdot \mid x\right), \mu\left(\cdot \mid x'\right)\right)\right) \\
&\leq K_f \left(d_X(x, x') + L_\mu d_X(x, x')\right) = \left(K_f + K_f K_\mu\right) d_X(x, x').
\end{aligned}$$

□

Lemma E.9. *Let Assumptions 1 and 2 hold, and let $\beta_{<t}, \beta'_{<t} \in \Sigma_{<t}$ and $\beta_{\cdot t} \in \Sigma_{\cdot t}^{Lip}$. Then, there exists $K > 0$ such that*

$$d_W\left(P_{<t+1}\left(\cdot \mid \beta_{<t}, \beta_{\cdot t}\right), P_{<t+1}\left(\cdot \mid \beta'_{<t}, \beta_{\cdot t}\right)\right) \leq K \cdot d_W\left(P_{<t}\left(\cdot \mid \beta_{<t}\right), P_{<t}\left(\cdot \mid \beta'_{<t}\right)\right)$$

Proof. By Assumption 2, there exist constants $K_{\sigma_{it}} > 0$ for $it \in L$, such that σ_{it} is $K_{\sigma_{it}}$-Lipschitz continuous in $\mathcal{A}_{<t}$. Also, denote with K_{0t} nature's Lipschitz constant with respect to d_W in stage t. Similarly, as $\beta_{\cdot t} \in \Sigma_{\cdot t}^{\text{Lip}}$, there exist constants $K_{\beta_{it}} > 0$ such that $\beta_{it}(\cdot \mid s_{it})$ is $K_{\beta_{it}}$-Lipschitz with respect to the Wasserstein distance. Overall, we get for $it \in L$, $a_{<t}, a'_{<t} \in \mathcal{A}_{<t}$

$$d_W\left(\beta_{it}(\cdot \mid \sigma_{it}(a_{<t})), \beta_{it}(\cdot \mid \sigma_{it}(a'_{<t}))\right) \leq K_{\beta_{it}} K_{\sigma_{it}} \cdot ||a_{<t} - a'_{<t}||.$$

Denote $K_t := K_{0t} + \sum_{i \in \mathcal{N}} K_{\beta_{it}} K_{\sigma_{it}}$. Then we get that the mapping $a_{<t} \mapsto P_t(\cdot \mid a_{<t}, \beta_{\cdot t})$ is K_t-Lipschitz continuous with respect to d_W, which can be seen by

$$\begin{aligned}
&d_W\left(P_t(\cdot \mid a_{<t}, \beta_{\cdot t}), P_t(\cdot \mid a'_{<t}, \beta_{\cdot t})\right) \\
&\overset{(E.6)}{\leq} d_W\left(p_{0t}(\cdot \mid a_{<t}), p_{0t}(\cdot \mid a'_{<t})\right) + \sum_{i \in \mathcal{N}} d_W\left(\beta_{it}(\cdot \mid \sigma_{it}(a_{<t})), \beta_{it}(\cdot \mid \sigma_{it}(a'_{<t}))\right) \\
&\leq \left(K_{0t} + \sum_{i \in \mathcal{N}} K_{\beta_{it}} K_{\sigma_{it}}\right) \cdot ||a_{<t} - a'_{<t}||_\infty,
\end{aligned}$$

for all $a_{<t}, a'_{<t} \in \mathcal{A}_{<t}$. Let $f : \mathcal{A}_{<t+1} \to \mathbb{R}$ be 1-Lipschitz continuous. Then, we get by Lemma E.8, that the function $g_f(a_{<t}) := \int_{\mathcal{A}_{<t}} f(a_{<t}, a_{\cdot t}) dP_t(a_{\cdot t} \mid a_{<t}, \beta_{\cdot t})$ is $(1 + K_t)$-Lipschitz continuous. With this, we get

$$\begin{aligned}
&d_W\left(P_{<t+1}(\cdot \mid \beta_{<t}, \beta_{\cdot t}), P_{<t+1}(\cdot \mid \beta'_{<t}, \beta_{\cdot t})\right) \\
&\overset{(E.5)}{=} \sup_{||f||_{\text{Lip}} \leq 1} \int_{\mathcal{A}_{<t+1}} f(a_{<t+1}) P_{<t+1}(a_{<t+1} \mid \beta_{<t}, \beta_{\cdot t}) - \int_{\mathcal{A}_{<t+1}} f(a_{<t+1}) P_{<t+1}(a_{<t+1} \mid \beta'_{<t}, \beta_{\cdot t}) \\
&= \sup_{||f||_{\text{Lip}} \leq 1} \int_{\mathcal{A}_{<t}} g_f(a_{<t}) P_{<t}(a_{<t} \mid \beta_{<t}) - \int_{\mathcal{A}_{<t}} g_f(a_{<t}) P_{<t}(a_{<t} \mid \beta'_{<t}) \\
&\overset{(E.8)}{\leq} \sup_{||g||_{\text{Lip}} \leq 1 + K_t} \int_{\mathcal{A}_{<t}} g(a_{<t}) P_{<t}(a_{<t} \mid \beta_{<t}) - \int_{\mathcal{A}_{<t}} g(a_{<t}) P_{<t}(a_{<t} \mid \beta'_{<t}) \\
&\overset{(E.5)}{=} (1 + K_t) \cdot d_W\left(P_{<t}(\cdot \mid \beta_{<t}), P_{<t}(\cdot \mid \beta'_{<t})\right),
\end{aligned}$$

which shows the statement. □

Lemma E.10. *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game, where Assumptions 1 and 2 hold. For strategies $\beta_{-i} \in \Sigma_{-i}^{Lip}$, $\beta_i \in \Sigma_i^{Lip,\, p}$ and $\epsilon > 0$, there exists a $\delta > 0$ such that for all $\beta'_i \in \Sigma_i^p$ with $||\beta_i - \beta'_i||_\infty < \delta$, it holds that*

$$d_W\left(P(\cdot \mid \beta_i, \beta_{-i}), P(\cdot \mid \beta'_i, \beta_{-i})\right) < \epsilon.$$

Proof. Let $\epsilon > 0$, $\beta_{-i} \in \Sigma_{-i}^{\text{Lip}}$, $\beta_i \in \Sigma_i^{\text{Lip, p}}$, and $\beta'_i \in \Sigma_i^{\text{p}}$. Then we get

$$\begin{aligned}
&d_W\left(P(\cdot \mid \beta_i, \beta_{-i}), P(\cdot \mid \beta'_i, \beta_{-i})\right) \\
&\overset{(\triangle\text{-inequ.})}{\leq} \sum_{t=1}^{T} d_W\left(P(\cdot \mid (\beta'_{i<t}, \beta_{it}, \beta_{i>t}), \beta_{-i}), P(\cdot \mid (\beta'_{i<t}, \beta'_{it}, \beta_{i>t}), \beta_{-i})\right) \\
&\overset{(E.9)}{\leq} \sum_{t=1}^{T} \left(\prod_{v>t} K_v\right) \cdot d_W\left(P_{<t+1}(\cdot \mid (\beta'_{i<t}, \beta_{it}), \beta_{-i}), P_{<t+1}(\cdot \mid (\beta'_{i<t}, \beta'_{it}), \beta_{-i})\right) \\
&\overset{(E.7)}{\leq} \sum_{t=1}^{T} K_{>t} ||\beta_{it} - \beta'_{it}||_\infty,
\end{aligned}$$

where $K_{>t} = \prod_{v=t+1}^{T} K_v$ with $K_t := K_{0t} + \sum_{i \in \mathcal{N}} K_{\beta_{it}} K_{\sigma_{it}}$ (see proof of Lemma E.9) for all $1 \leq t \leq T$. By choosing $\delta < \max_{t \in T} \frac{\epsilon}{K_{>t} \cdot T}$, we get the statement. □

With the established results, we can finally give the proof of our main result.

E.1 Proof of Theorem 1

We repeat the theorem for the reader's convenience.

Theorem E.2. *Let $\Gamma = (\mathcal{N}, T, S, \mathcal{A}, p, \sigma, u)$ be a multi-stage game, where Assumptions 1 and 2 hold, and that the utility function u_i is continuous. Further, let $\beta_{-i} \in \Sigma_{-i}^{Lip}$, $\beta_i \in \Sigma_i$, and $A_{M_{IS}}$ and $B_{M_{IS}}$ be simulated data sets with initial simulation size $M_{IS} \in \mathbb{N}$ as described in Section A.2.3. Then, we have*

$$\lim_{D\to\infty} \varepsilon_D \leq 0 \text{ and } \lim_{M_{IS}\to\infty} \varepsilon_{M_{IS}} = 0 \text{ almost surely.}$$

Furthermore, we receive an upper bound on the utility loss over pure Lipschitz continuous strategies for the strategy profile $\beta = (\beta_i, \beta_{-i})$ by

$$\begin{aligned}\lim_{D\to\infty}\lim_{M_{IS}\to\infty} \ell_i^{ver}(\beta) &= \lim_{D\to\infty}\lim_{M_{IS}\to\infty} \hat{u}_i^{ver,\,D}(\beta_{-i} \mid A_{M_{IS}}) - \hat{u}_i(\beta \mid B_{M_{IS}}) \\ &\geq \sup_{\beta_i' \in \Sigma_i^{Lip,\,p}} \tilde{u}_i(\beta_i', \beta_{-i}) - \tilde{u}_i(\beta) = \tilde{\ell}_i^{Lip,\,p}(\beta) \text{ a. s.}\end{aligned}$$

Proof. By Lemma E.3, we have almost sure convergence of $\lim_{M_{\text{IS}}\to\infty} \varepsilon_{M_{\text{IS}}} = 0$. To finish the first statement, it remains to show $\lim_{\text{D}\to\infty} \varepsilon_{\text{D}} \leq 0$.

Let $\epsilon > 0$ and $\bar{\beta}_i \in \Sigma_i^{\text{Lip, p}}$ such that $\sup_{\beta_i' \in \Sigma_i^{\text{Lip, p}}} \tilde{u}_i(\beta_i', \beta_{-i}) - \tilde{u}_i\left(\bar{\beta}_i, \beta_{-i}\right) \leq \epsilon$. Then, by Lemma E.4, there exists a sequence $\left\{\beta_i^{\text{D}}\right\}_{\text{D}\in\mathbb{N}}$ with $\beta_i^{\text{D}} \in \Sigma_i^{\text{D}}$ such that

$$\lim_{\text{D}\to\infty} \left|\left|\bar{\beta}_i - \beta_i^{\text{D}}\right|\right|_\infty = 0.$$

By Lemma E.10, we further get

$$\lim_{\text{D}\to\infty} d_W\left(P\left(\cdot \mid \bar{\beta}_i, \beta_{-i}\right), P\left(\cdot \mid \beta_i^{\text{D}}, \beta_{-i}\right)\right) = 0.$$

The utility functions u_i are bounded and continuous by assumption. Therefore, we can use Theorem E.1 and get

$$\lim_{\text{D}\to\infty} \tilde{u}_i\left(\beta_i^{\text{D}}, \beta_{-i}\right) = \lim_{\text{D}\to\infty} \int_{\mathcal{A}} u_i(a) dP\left(a \mid \beta_i^{\text{D}}, \beta_{-i}\right) = \tilde{u}_i\left(\bar{\beta}_i, \beta_{-i}\right).$$

As this holds for every $\epsilon > 0$, we get for $\Sigma_i^{\text{SF}} := \bigcup_{\text{D}\in\mathbb{N}} \Sigma_i^{\text{D}}$

$$\sup_{\beta_i' \in \Sigma_i^{\text{Lip, p}}} \tilde{u}_i(\beta_i', \beta_{-i}) \leq \sup_{\beta_i^{\text{SF}} \in \Sigma_i^{\text{SF}}} \tilde{u}_i\left(\beta_i^{\text{SF}}, \beta_{-i}\right),$$

finishing the first statement. For the second statement, note that due to the boundedness of u_i, we can use Kolmogorov's law of large numbers and get

$$\lim_{M_{\text{IS}}\to\infty} \hat{u}_i(\beta \mid B_{M_{\text{IS}}}) = \tilde{u}_i(\beta). \tag{27}$$

Furthermore, we get that

$$\begin{aligned}\left|\ell_i^{\text{ver}}(\beta) - \tilde{\ell}_i^{\text{Lip, p}}(\beta)\right| &= \left|\hat{u}_i^{\text{ver, D}}(\beta_{-i} \mid A_{M_{\text{IS}}}) - \hat{u}_i(\beta \mid B_{M_{\text{IS}}}) - \sup_{\beta_i' \in \Sigma_i^{\text{Lip, p}}} \tilde{u}_i(\beta_i', \beta_{-i}) + \tilde{u}_i(\beta)\right| \\ &\leq \left|\hat{u}_i^{\text{ver, D}}(\beta_{-i} \mid A_{M_{\text{IS}}}) - \sup_{\beta_i' \in \Sigma_i^{\text{D}}} \tilde{u}_i(\beta_i', \beta_{-i})\right| + \left|\sup_{\beta_i' \in \Sigma_i^{\text{D}}} \tilde{u}_i(\beta_i', \beta_{-i}) - \sup_{\beta_i' \in \Sigma_i^{\text{Lip, p}}} \tilde{u}_i(\beta_i', \beta_{-i})\right| \\ &\quad + \left|\hat{u}_i(\beta \mid B_{M_{\text{IS}}}) - \tilde{u}_i(\beta)\right| \\ &= \varepsilon_{\text{D}} + \varepsilon_{M_{\text{IS}}} + \left|\hat{u}_i(\beta \mid B_{M_{\text{IS}}}) - \tilde{u}_i(\beta)\right|.\end{aligned}$$

From the first statement and using the relation of Equation 27, we get

$$\lim_{D\to\infty}\lim_{M_{\text{IS}}\to\infty} \left|\ell_i^{\text{ver}}(\beta) - \tilde{\ell}_i^{\text{Lip, p}}(\beta)\right| \geq 0,$$

finishing the statement. □

F Additional Related Work: Convergence Results to Equilibrium

The literature on convergence in games is vast and rapidly growing. Here, we state some central results from this field, with a focus on policy gradient algorithms and continuous multi-stage games. For a broader discussion, refer to the survey articles by Yang and Wang (2020) and Zhang et al. (2021).

In games with finitely many states and actions, some algorithms are guaranteed to converge to equilibrium in specific game classes. The time-average policies of no-regret algorithms converge to the set of coarse correlated equilibria (Blum and Mansour, 2007). In two-player zero-sum games, this implies that the time-average policies converge to the set of Nash equilibria. Hennes et al. (2020) show that a popular policy gradient algorithm, where the policy is tabular and parametrized by a softmax function, satisfies the no-regret property. Other MARL variants with this property include neural fictitious play (Heinrich and Silver, 2016) and neural replicator dynamics (Hennes et al., 2020). Additionally, some approaches use a meta-game solver, where RL agents learn best-responses in an inner loop to compute best-responses against an average of previous policies (Lanctot et al., 2017; McAleer et al., 2022). These works consider convergence in average policy. Some algorithms also achieve last-iterate convergence in two-player zero-sum games with finitely many states and actions (Lockhart et al., 2019; Farina et al., 2022).

Another game class learnable by MARL algorithms is Markov potential games (Marden, 2012; Macua et al., 2018, april; Leonardos et al., 2022, april). Leonardos et al. (2022, april) provide convergence guarantees to Nash equilibrium policies for independent policy gradient algorithms. Ding et al. (2022, july) extend these results to independent policy gradient methods with function approximation, offering sharper convergence rates.

For general-sum stochastic games, convergence remains elusive. Giannou et al. (2022) describe properties of Nash equilibrium policies in finite Markov games, such as strictness and second-order stationarity that ensure policy gradient algorithms in self-play converge to NE with high probability.

Convergence guarantees in multi-stage games with continuous signals and actions are even rarer. Zhang et al. (2019, december) show that policy gradient algorithms in self-play converge to NE in two-player zero-sum quadratic games. However, finding the NE of zero-sum Markov games generally becomes a nonconvex-nonconcave saddle-point problem (Mazumdar et al., 2019; Bu et al., 2019; Chasnov et al., 2020). This inherent difficulty persists even in the simplest linear quadratic setting with linear function approximation (Bu et al., 2019; Chasnov et al., 2020). Consequently, most convergence results are local, addressing behavior around local NE points (Mescheder et al., 2017, december; Daskalakis and Panageas, 2018, december; Mertikopoulos et al., 2019, may). Moreover, policy gradient updates in MARL often fail to converge to local NEs due to non-convergent behaviors such as limit cycling (Mescheder et al., 2017, december; Daskalakis and Panageas, 2018, december; Mertikopoulos et al., 2019, may) or the existence of non-Nash stable limit points (Mazumdar et al., 2019).

While some results about Markov potential games apply to games with continuous states (Leonardos et al., 2022, april; Ding et al., 2022, july), we know of no formal extension of potential games to games with multiple stages and continuous actions or imperfect information.

In summary, the above results are either restricted to games with finitely many states and actions, two-player zero-sum, or Markov potential games. The class of games we consider, namely, multi-stage games with continuous signals and actions, does not fall into any of these categories. Therefore, to the best of our knowledge, no convergence guarantees directly apply to our setting.

References

Arozamena, Leandro, Federico Weinschelbaum. 2009. Simultaneous vs. sequential price competition with incomplete information. *Economics Letters* **104**(1) 23–26. doi:10.1016/j.econlet.2009.03.017.

Blum, Avrim, Yishay Mansour. 2007. Learning, Regret Minimization, and Equilibria. Noam Nisan, Tim Roughgarden, Eva Tardos, Vijay V. Vazirani, eds., *Algorithmic Game Theory*. Cambridge University Press, Cambridge, 79–102. doi:10.1017/CBO9780511800481.006. URL https://www.cambridge.org/core/product/identifier/CBO9780511800481A051/type/book_part.

Bu, Jingjing, Lillian J. Ratliff, Mehran Mesbahi. 2019. Global convergence of policy gradient for sequential zero-sum linear quadratic dynamic games. *ArXiv e-prints* **abs/1911.04672**.

Chasnov, Benjamin, Lillian Ratliff, Eric Mazumdar, Samuel Burden. 2020. Convergence Analysis of Gradient-Based Learning in Continuous Games. *Conference on Uncertainty in Artificial Intelligence (UAI)*. PMLR, 935–944.

Daskalakis, Constantinos, Ioannis Panageas. 2018, december. The limit points of (optimistic) gradient descent in min-max optimization. Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi,

Roman Garnett, eds., *Conference on Neural Information Processing Systems (NeurIPS)*. Montréal, Canada, 9256–9266.

Ding, Dongsheng, Chen-Yu Wei, Kaiqing Zhang, Mihailo R. Jovanovic. 2022, july. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, Sivan Sabato, eds., *International Conference on Machine Learning (ICML)*. PMLR, Baltimore, Maryland, USA.

Farina, Gabriele, Chung-Wei Lee, Haipeng Luo, Christian Kroer. 2022. Kernelized multiplicative weights for 0/1-Polyhedral games: Bridging the gap between learning in extensive-form and normal-form games. Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, Sivan Sabato, eds., *International Conference on Machine Learning (ICML)*, *Proceedings of Machine Learning Research*, vol. 162. PMLR, Baltimore, Maryland, USA, 6337–6357.

Giannou, Angeliki, Kyriakos Lotidis, Panayotis Mertikopoulos, Emmanouil-Vasileios Vlatakis-Gkaragkounis. 2022. On the convergence of policy gradient methods to Nash equilibria in general stochastic games.

Heinrich, Johannes, David Silver. 2016. Deep Reinforcement Learning from Self-Play in Imperfect-Information Games. *arXiv:1603.01121 [cs]* URL http://arxiv.org/abs/1603.01121.

Hennes, Daniel, Dustin Morrill, Shayegan Omidshafiei, Rémi Munos, Julien Perolat, Marc Lanctot, Audrunas Gruslys, Jean-Baptiste Lespiau, Paavo Parmas, Edgar Duèñez-Guzmán, Karl Tuyls. 2020. Neural Replicator Dynamics: Multiagent Learning via Hedging Policy Gradients. *Proceedings of International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*. AAMAS '20, International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 492–501.

Horst, Ulrich. 2005. Stationary equilibria in discounted stochastic games with weakly interacting players. *Games and Economic Behavior* **51**(1) 83–108. doi:10.1016/j.geb.2004.03.003.

Hosoya, Yuhki, Chaowen Yu. 2022. On the approximate purification of mixed strategies in games with infinite action sets. *Economic Theory Bulletin* **10**(1) 69–93. doi:10.1007/s40505-022-00219-1.

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, Sepp Hochreiter. 2017. Self-normalizing neural networks. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, Roman Garnett, eds., *30th Annual Conference on Neural Information Processing Systems*. Long Beach, California, USA, 971–980.

Lanctot, Marc, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, karl Tuyls, Julien Perolat, David Silver, Thore Graepel. 2017. A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning. I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 4190–4203. URL http://papers.nips.cc/paper/7007-a-unified-game-theoretic-approach-to-multiagent-reinforcement-learning.pdf.

Leonardos, Stefanos, Will Overman, Ioannis Panageas, Georgios Piliouras. 2022, april. Global convergence of multi-agent policy gradient in markov potential games. *International Conference on Learning Representations (ICLR)*. OpenReview.net, virtual event.

Lockhart, Edward, Marc Lanctot, Julien Pérolat, Jean-Baptiste Lespiau, Dustin Morrill, Finbarr Timbers, Karl Tuyls. 2019. Computing approximate equilibria in sequential adversarial games by exploitability descent. Sarit Kraus, ed., *Proceedings of the Joint Conference on Artificial Intelligence (IJCAI)*. ijcai.org, Macao, SAR, China, 464–470.

Ma, Junhai, Qiuxiang Li. 2014. The Complex Dynamics of Bertrand-Stackelberg Pricing Models in a Risk-Averse Supply Chain. *Discrete Dynamics in Nature and Society* **2014**(1) 749769.

Macua, Sergio Valcarcel, Javier Zazo, Santiago Zazo. 2018, april. Learning parametric closed-loop policies for markov potential games. *International Conference on Learning Representations (ICLR)*. OpenReview.net, Vancouver, BC, Canada.

Marden, Jason R. 2012. State based potential games. *Automatica (Journal of IFAC)* **48**(12) 3075–3088.

Mazumdar, Eric, Michael I. Jordan, S. Shankar Sastry. 2019. On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games. *ArXiv e-prints* **abs/1901.00838**.

McAleer, Stephen, Kevin Wang, John Lanier, Marc Lanctot, Pierre Baldi, Tuomas Sandholm, Roy Fox. 2022. Anytime PSRO for Two-Player Zero-Sum Games.

Mertikopoulos, Panayotis, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, Georgios Piliouras. 2019, may. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *International Conference on Learning Representations (ICLR)*. OpenReview.net, New Orleans, LA, USA.

Mescheder, Lars M., Sebastian Nowozin, Andreas Geiger. 2017, december. The numerics of gans. Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, Roman Garnett, eds., *Conference on Neural Information Processing Systems (NeurIPS)*. Long Beach, California, USA, 1825–1835.

Milgrom, Paul R., Robert J. Weber. 1985. Distributional strategies for games with incomplete information. *Mathematics of Operations Research* **10**(4) 619–632.

Myerson, Roger B., Philip J. Reny. 2020. Perfect conditional ϵ-equilibria of multi-stage games with infinite sets of signals and actions. *Econometrica* **88**(2) 495–531.

Raffin, Antonin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* **22**(268) 1–8. URL http://jmlr.org/papers/v22/20-1364.html.

Reny, Philip J. 2011. On the existence of monotone pure-strategy equilibria in bayesian games. *Econometrica* **79**(2) 499–553.

Villani, Cédric. 2009. *Optimal Transport*, *Grundlehren Der Mathematischen Wissenschaften*, vol. 338. Springer, Berlin, Heidelberg. doi:10.1007/978-3-540-71050-9.

Wambach, Achim. 1999. Bertrand competition under cost uncertainty. *International Journal of Industrial Organization* **17**(7) 941–951.

Yang, Yaodong, Jun Wang. 2020. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583* .

Zhang, Jun. 2008. Simultaneous signaling in elimination contests. Tech. rep., Queen's Economics Department Working Paper.

Zhang, Kaiqing, Zhuoran Yang, Tamer Basar. 2019, december. Policy optimization provably converges to Nash equilibria in zero-sum linear quadratic games. Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, Roman Garnett, eds., *Conference on Neural Information Processing Systems (NeurIPS)*. Vancouver, BC, Canada, 11598–11610.

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

Zinkevich, Martin, Michael Johanson, Michael Bowling, Carmelo Piccione. 2007. Regret minimization in games with incomplete information. *Advances in neural information processing systems* **20**.