Rapport de stage
Algorithmes distribués pour la théorie des jeux

Jérôme Taupin

Préambule

Ce stage a été effectué durant ma césure de Février à Juillet 2023 à Thales SIX Genevilliers et sous la supervision de Christophe Le Martret et Xavier Leturc.

L'équipe dans laquelle j'ai travaillé étant spécialisée dans les télécoms, mon sujet était motivé par ce contexte mais mon travail concret se rapporte essentiellement a des maths appliquées indépendantes de l'application aux télécoms.

La motivation est la suivante: On se donne plusieurs "clusters" (petits réseaux) qui choisissent chacun un canal de fréquence pour communiquer. L'objectif est que les clusters choisissent des canaux différents de leurs voisins pour éviter de créer des interférences. Ces choix doivent être fait de manière distribuée, c'est-à-dire que chaque cluster ne sait pas ce que les autres clusters font.

Ce problème se modélise à l'aide de la théorie des jeux, où un cluster égal un joueur, qui reçoit une utilité dépendant des canaux choisis par chaque joueur. J'ai étudié des algorithmes qui agissent sur des jeux quelconques pour maximiser l'utilité moyenne, sans se restreindre à cette application précise. En revanche, j'ai implémenté ces algorithmes et fait des simulations numériques sur ce problème précis, mais les courbes ne sont pas incluses dans ce rapport, l'essentiel de mon travail étant théorique.

Un tel algorithme (le TEL, pour trial and error learning) existe déjà dans la littérature. Ils ont repéré quelques comportements sous-optimaux et ont élaborés des heuristiques pour y remédier. Mon travail dans un premier temps a été de lire et comprendre la preuve du TEL et de vérifier que l'ajout des heuristiques préserve le résultat de convergence. C'est effectivement le cas, et on appelle ITEL (improved TEL) l'algorithme résultant.

Le TEL utilise un paramètre $\varepsilon > 0$ qui contrôle le taux de perturbation de l'algorithme. Intuitivement, prendre ε trop grand rend l'algorithme instable, tandis que le prendre trop petit rend l'algorithme très lent. Dans un second temps j'ai brièvement étudié la possibilité de faire décroître epsilon durant l'algorithme. On peut étendre le résultat de convergence à ce cas mais les conditions théoriques sont difficiles à calculer en pratique.

Enfin, le plus gros du travail a été d'adapter l'ITEL aux jeux stochastiques en un nouvel algorithme RITEL (robust ITEL), où les utilités reçues par les joueurs sont aléatoires. Cet ajout complexifie significativement la discussion. Mon équipe avait déjà obtenu des résultats mais sous des hypothèses fortes sur le jeu. J'ai réussi à obtenir une convergence plus générale en utilisant des résultats de grande déviation.

Le rapport est structuré comme suit:

- Les sections 1 et 2 présente le problème et la modélisation mathématique, ainsi qu'un résultat existant dans la littérature qui permet de montrer la convergence de certains types de processus de Markov, dits perturbés.
- La section 3 rappelle le TEL et introduit l'ITEL, ainsi que les résultats théoriques correspondant.
- La section 4 traite de l'adaptation aux jeux stochastiques. Plus précisément, le RITEL est présenté en section 4.1. Les sections 4.2 et 4.3 permettent d'adapter les outils donnés en section 2 et ne sont pas essentiels à la compréhension de l'algorithme. La section 4.4 énonce les résultats théoriques.

- La section 5 est plus anecdotique et discute de la possibilité de faire décroître ε durant l'algorithme en adaptant des résultats de recuit simulé.
- Les appendices fournissent les preuves des résultats de convergence.

1 Introduction

In a general game theory setup, multiple agents (or players) can play different strategies (or actions). We are interested in the study of algorithms to try and find an optimal choice of strategies for multiple agents in a distributed manner. A distributed algorithm only allows each agent to have knowledge of its own action and observed utility when choosing a strategy. Despite this strong constraint, perturbation-based algorithms such as Trial and Error Learning (TEL) [1] and Online Distributed Learning (ODL) [2] were shown to converge to some sense towards optimal action profiles. These algorithm rely on a perturbation factor $\varepsilon > 0$. It is proven [1] that as the perturbation factor goes to zero, the asymptotic behavior of TEL is to almost only visit optimal states, favoring Nash equilibria. This result is proven using a key theorem [3, Theorem 4] which establishes a link between the convergence of a category of perturbed Markov processes and a notion of potential over a resistance graph.

Our work is separated into three main discussions. First, we review the reasoning of these algorithms and introduce heuristics to improve them. In particular, we show that the resulting algorithm Improved Trial and Error Learning (ITEL) retains the convergence properties of TEL while performing better in practical simulations. Secondly, we extend the scope of ITEL by adapting it into another algorithm Robust Improved Trial and Error Learning (RITEL) that is robust to stochastic games – where players receive random payoffs – and we derive similar convergence result as before. Thirdly, we draw a parallel between our approach and that of general simulated annealing. In particular, considering ITEL with a slowly decreasing noise schedule yields a stronger result of convergence when the perturbation vanishes along the algorithm.

2 Model and Notations

We consider a typical game theory setup where a set I of n players choose between a finite amount of actions. Given an action profile $\boldsymbol{a} = (a_i)_{i \in I}$ of chosen actions for all players, player i observes a utility $U_i(\boldsymbol{a})$. Without loss of generality, utilities are assumed to be bounded within $[0, 1]$. The quality of an action profile is measured via its average global welfare $W(\boldsymbol{a}) \triangleq \frac{1}{n} \sum_i U_i(\boldsymbol{a})$.

Remark. In this paper, a bold letter $\boldsymbol{a}$ will always denotes a vector of actions whose ith component will be denoted a_i. A regular letter a always denotes a single action. The same notational convention applies to other quantities.

In a general game, it is possible that a player i maximizes its utility by playing an action that is detrimental to the other players. If the other players have no way to influence i out of playing this action, then the algorithm will get stuck in a sub-optimal state. In order for a distributed algorithm to be capable of maximizing global welfare, an assumption on the nature of the game is introduces, stating that the games considered do not feature such situations.

Assumption A1 (Interdependence). We consider games that are *interdependent*, i.e., given any action profile $\boldsymbol{a}$, for any proper subset $\emptyset \subsetneq J \subsetneq I$ of players, there exists a player $i \in J$, an action $a_i' \neq a_i$, and a player $j \notin J$, such that $U_j(a_i', a_{-i}) \neq U_j(\boldsymbol{a})$.

We now introduce the formalism necessary to study TEL and the derived algorithms we shall introduce later. Consider a homogeneous Markov chain of transition matrix P^0 over a finite space $\mathcal{X}$ which we call the unperturbed process. We denote $P^0_{x,y}$ the probability of transitioning from x to y. We want to consider a perturbed version of this process controlled by a noise parameter $\varepsilon > 0$, such that the perturbation vanishes as ε converges to 0.

Definition 2.1 (Perturbed Markov Process)**.** A family of homogeneous Markov chains $(P^\varepsilon)_{0 \leq \varepsilon < \varepsilon_0}$ – referred to as P^ε for simplicity – forms a *perturbed Markov process (PMP)* if it satisfies the following properties:

$$\begin{cases} \forall \varepsilon \in (0, \varepsilon_0), \quad P^\varepsilon \text{ is aperiodic and irreducible,} & \text{(1a)} \\ \forall x, y \in \mathcal{X}, \quad P^\varepsilon_{x,y} \underset{\varepsilon \to 0}{\to} P^0_{x,y}. & \text{(1b)} \end{cases}$$

In order to analyze PMPs, an additional property is required.

Definition 2.2 (Regularity)**.** A family $(X^\varepsilon)_{0\leq\varepsilon<\varepsilon_0}$ of non-negative real numbers is *regular* if $X^\varepsilon = 0$ or there exists $r \geq 0$ such that

$$0 < \lim_{\varepsilon\to 0} \varepsilon^{-r} X^\varepsilon < +\infty. \tag{2}$$

r is unique and called the *resistance* of X^ε. When $X^\varepsilon = 0$ we say by convention that the resistance is $+\infty$. A PMP P^ε is a *regular perturbed Markov process (RPMP)* if all the transition probabilities $P^\varepsilon_{x,y}$ are regular.

Given a RPMP, one can define the underlying resistance graph over the state space, replacing transition probabilities with their respective resistance. The study of the resistance graph allows to describe the behavior of the perturbed process when $\varepsilon > 0$ is small.

Definition 2.3 (Resistance Graph)**.**

- The resistance of a transition $P^\varepsilon_{x,y}$ is denoted $r(x \to y)$. Intuitively, the regularity condition implies that $P^\varepsilon_{x,y}$ behaves like ε^r for small ε.
- Define $\mathcal{G}$ the directed graph over the vertex space $\mathcal{X}$ where there is a directed edge of weight $r(x \to y)$ from x to y when the resistance is finite.
- An edge $x \to y$ in $\mathcal{G}$ minimizing the resistance among all outwards edges from x is said to be an *easy edge*. The associated minimal resistance is called *outward resistance* and denoted $r^\star(x) = r(x \to y)$.
- The resistance of a path $x \rightsquigarrow y$ in $\mathcal{G}$ is the sum of the resistances of its edges.

Notice that as $\varepsilon \to 0$, $P^\varepsilon_{x,y}$ converges to a positive value if and only if $r(x \to y) = 0$. Hence a transition exists in the unperturbed process if and only if its resistance is equal to 0. Moreover, since all non-zero transition probabilities have finite resistance, the irreducibility of P^ε implies that their is a finite-resistance path from any state to any other in $\mathcal{G}$.

The tendency of the process to stay in a state x can be measured through a notion of potential. The higher its potential, the more difficult it is to reach the state, or the easier it is to leave. Let us detail the construction of the potential.

Definition 2.4 (Potential)**.**

- A *x-tree* is a spanning tree of the graph $\mathcal{G}$ rooted at x, i.e., an acyclic sub-graph of $\mathcal{G}$ such that every vertex $y \neq x$ has a unique outgoing edge and x has none.
- The resistance of a x-tree is the sum of the resistances of its edges.
- The *potential*[1] $\gamma(x)$ of a vertex x is the minimal resistance of a x-tree.
- The set of vertex minimizing the potential is denoted $\mathcal{X}^\star$.

We are now able to state the key result from [3], which essentially shows that a RPMP converges in some sense to states of minimal potential. The original formulation of this result [3, Theorem 4] is stated in terms of recurrence classes of the unperturbed process. It is proven in two steps, first showing the convergence in terms of individual states of the process, then showing that potential can be computed over recurrence classes using graph theory, which is easier to do in practice as it allows to consider less states. We choose to state both results separately, as the second part does not influence the convergence statement itself but only the practical computation of potentials.

Theorem 2.1 ([3], Lemma 1)**.** Let P^ε be a RPMP over a finite space $\mathcal{X}$. Denote π^ε its stationary distribution for every small positive ε. Then:

1. as $\varepsilon \to 0$, π^ε converges to a stationary distribution π^0 of P^0.

[1] γ is denoted ρ and called *stochastic potential* in [1].

2. $\pi_x^0 > 0$ if and only if $x \in \mathcal{X}^\star$.

States x verifying $\pi_x^0 > 0$ are said to be *stochastically stable*.

Let us recall a few notions of Markov chain theory. Two states of a Markov chain are said to communicate if a path exists from one to the other and back with positive probability. The communication class of a state is the set of all states that communicate with x, along with x itself. The recurrence states of the Markov chain are the communication classes such that no path exists from them to another communication class.

Going back to the study of a RPMP P^ε, two states communicate in the unperturbed process if and only if their is a path of zero resistance from one to another and back. A communication class is recurrent in P^0 if and only if all edges leaving the class have positive resistance. Consider two states x and y such that $r(x \to y) = 0$. A x-tree can be modified into a y-tree by removing the edge leaving y and adding the edge $x \to y$. Since $r(x \to y) = 0$, the obtained y-tree has smaller resistance than the original x-tree. It can be deduced that potential is constant over communication classes. This is proven in detail in [3, Lemma 2] Moreover, non-recurrent communication classes cannot minimize the potential as they have a higher potential then classes that can be reached from them.

For this reason, an alternative way to compute potentials is to consider an aggregated version of $\mathcal{G}$ over the communication classes of P^0. The resistance of an edge between two classes X and Y is defined as the minimum resistance of a path from any state of X to any state of Y. Note that a class of P^0 is recurrent if and only if its minimal outward resistance $r^\star$ is nonzero in the aggregated graph. Potential is then defined the same way as in $\mathcal{G}$. It is not necessary to compute the potential of non-recurrent classes since they cannot minimize it.

3 Perturbed Distributed Learning

In this section we introduce heuristics for TEL, resulting in an improved algorithm ITEL. ITEL is shown to perform better in practice, while keeping the same convergence properties as TEL, or even improving them in some cases. Recall that TEL [1] aims at identifying the best Nash equilibrium in the sense of maximizing global welfare.

On the other hand, ODL was later proposed [2], which acts as a closely related although simplified version of TEL. ODL maximizes global welfare without discriminating Nash equilibria. As for TEL, ODL can be improved into Improved Online Distributed Learning (IODL) in the same fashion. This section focuses on ITEL as it is more complex than IODL. IODL uses the same reasoning and is described briefly for completeness at the end of the section.

In the case of ITEL, the proposed heuristics act as extensions of the original TEL formulation, so that it can be modeled as a special case of TEL. The theoretical proof of convergence, which is deferred to Appendix A, follows the same overall reasoning, although some adaptations are required to fit the heuristics, and the way arguments are presented may differ slightly.

We present here improved versions of both algorithms which get rid of unnecessary restrictions on the choice of parameters and change some sub-optimal behaviors.

3.1 General Formalism

Let us introduce the notion of Perturbed Distributed Learning (PDL), which encompasses the general reasoning of both TEL and ODL and allows us to introduce heuristics efficiently. The main idea of a PDL algorithm is to attribute moods to each player along with a reference action and utility, called benchmark. Players behave differently based on their state, i.e., their mood m, benchmark action $\overline{a}$ and benchmark utility $\overline{u}$. In a *discontent* state, a player will systemically explore a new strategy and may accept it depending on the quality of the outcome. In a *content* state, a player will generally keep playing its benchmark action but may occasionally explore another strategy and accept it. Two intermediate moods, *hopeful* and *watchful*, are also used in TEL, allowing to identify Nash equilibria. A step of any PDL algorithm plays out as follows: for each player,

1. Choose an action a to play based on an *action policy* taking into account its state $(\overline{m}, \overline{a}, \overline{u})$.
2. Observe utility u resulting from the choice of all actions $\boldsymbol{a}$.
3. Update its state according to an *update policy* taking into account its state $(\overline{m}, \overline{a}, \overline{u})$ along with a and u.

Regarding the update policy, we list here the different behaviors a player can adopt to update its state $(\overline{m}, \overline{a}, \overline{u})$:

- Accept : Go to state $(\mathtt{C}, a, u)$,
- Revert : Go to state $(\mathtt{C}, \overline{a}, \overline{u})$,
- Reject : Go to state $(\mathtt{D}, a, u)$,
- Adopt intermediate mood m: Go to state $(m, \overline{a}, \overline{u})$.

The main structure of any PDL algorithm described above is summarized in Figure 1. A specific PDL algorithm is determined by its action and update policies.

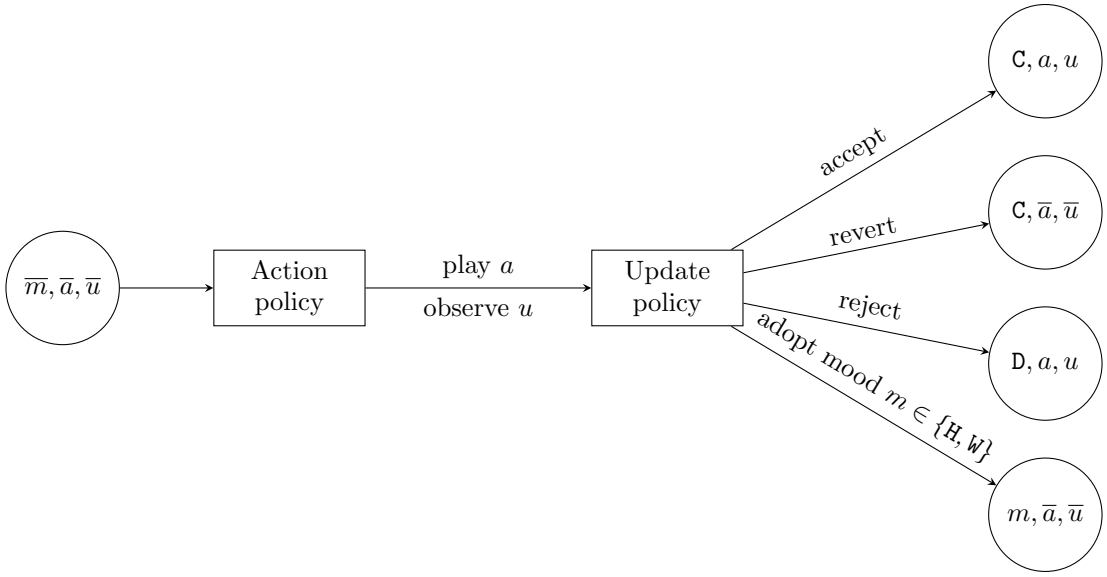

Figure 1: PDL Structure.

In the following, policies will be described by tables where each entry describes the decision of a player with the convention that the player is in state $(\overline{m}, \overline{a}, \overline{u})$, plays a, and observes u. Given $\overline{m}, \overline{a}, \overline{u}, a, u$, the decision of the player is always either deterministic or random involving constant probabilities or probabilities of the form ε^r for some $r \geq 0$. Such policies are therefore functions of ε and designed in a way such that the algorithm behaves as a RPMP as in Definition 2.1. Indeed, transition probabilities can be expressed as linear combinations of regular quantities, and are therefore regular, satisfying (2). Such transition probabilities also converge to a Markov chain we define to be the unperturbed process P^0, hence satisfying (1b). To justify the use of the RPMP framework, one would also need to show irreducibility and aperiodicity as required by (1a). In fact, we will see later that in ITEL irreducibility is not always satisfied, however the case where it is not is easy to analyze. This discussion will be detailed with Proposition 3.2. From now on we use the RPMP theoretical framework of resistances and potentials.

3.2 TEL

Before introducing our heuristics, we quickly recall the original form of Trial and Error Learning. Consider two resistance functions F and G. F is a non-increasing function of the observed utility u and controls the probability of acceptation from a discontent state. G is a non-increasing function of $u - \overline{u}$ and controls the probability of acceptation from a content state. TEL is defined under the following condition:

Condition C1 (Bounds on resistance functions). For all utilities $\overline{u}$ and u,

$$\begin{cases} 0 < F(u) < \frac{1}{2n} \\ 0 < G(u - \overline{u}) < \frac{1}{2} \end{cases} \tag{C1}$$

Action and update policies are detailed in Table 1. Explanations and examples can be found in [1].

Table 1: TEL Policies.

(a) Action Policy.

Mood	Decision
D	explore any a
C	explore $a \neq \overline{a}$ with probability ε, else play $\overline{a}$
H	play $\overline{a}$
W	play $\overline{a}$

(b) Update Policy.

Mood	Action	Utility	Decision
D	/	/	accept with probability $\varepsilon^{F(u)}$, else reject
C	$a \neq \overline{a}$	$u > \overline{u}$	accept with probability $\varepsilon^{G(u-\overline{u})}$, else revert
		$u \leq \overline{u}$	revert
	$a = \overline{a}$	$u > \overline{u}$	become H
		$u = \overline{u}$	revert
		$u < \overline{u}$	become W
H	/	$u > \overline{u}$	accept
		$u = \overline{u}$	revert
		$u < \overline{u}$	become W
W	/	$u > \overline{u}$	become H
		$u = \overline{u}$	revert
		$u < \overline{u}$	reject

The interpretation of Table 1 is the following: Given its mood, whether it explored or not, and how it performed compared to its benchmark, the player will behave according to the corresponding line. The behavior can either be deterministic or a random choice between two behaviors. Recall that the signification of behaviors "accept", "revert", etc. are given in Figure 1. Also note that when exploring, the action is chosen uniformly among all possible actions.

Under A1 and C1, TEL as described in Table 1 is proven [1] to minimize γ at Nash equilibria maximizing welfare, or at states maximizing a trade-off between welfare and stability if no Nash equilibrium exists. Stability will be defined in Section 3.4. Using Theorem 2.1, one concludes that these states are stochastically stable.

3.3 ITEL

In this section we extend TEL to integrate heuristics. The resulting algorithm is called ITEL.

The main idea behind our heuristics is based on the observation that a utility of 1 is optimal, hence a player should always accept it and keep it if possible. We introduce several heuristics to allow such behavior in the most general way. In particular C1 will be replaced by looser bounds.

Heuristic H1 (Acceptation from discontent state). A discontent player observing a sufficiently high utility will almost always accept it. Formally, we allow F to take values of 0, in contradiction with C1. For technical reasons, when $F(u) = 0$, we replace the accepting probability $\varepsilon^{F(u)} = 1$ by some fixed constant $c_F \in (0, 1)$.

The addition of c_F is due to the fact that when $\varepsilon^{F(u)} = 1$, the acceptation in the unperturbed process is not only possible but also systematic, hence we consider a constant c_F to ensure that there is still a positive probability $1 - c_F$ of rejecting the outcome, which will be needed later in the proof of convergence. To summarize, under H1, the resistance of accepting u from a discontent state is always equal to $F(u)$ and the resistance of rejecting it is always equal to 0. Note that TEL is still modeled under H1 as the c_F only appears when $F(u) = 0$.

Heuristic H2 (Exploration from content state). A content player with sufficiently high utility will never explore. Formally, let $A \subset [0, 1]$ be an *admissibility range*. Content players which benchmark utility falls within A are said to be *admissible* and shall never explore. A is assumed to be of the form $[u_0, 1]$ or $(u_0, 1]$.

Note that the original algorithm can still be modeled under H2 by taking $A = \emptyset$. A typical choice of A is $A = \{1\}$: if a player already maximizes its utility it has no reason to try other actions.

Heuristic H3 (Acceptation from content state). A content player observing a sufficiently high utility after exploring will always accept it. Formally, G is now a function of $(\overline{u}, u)$ such that G is non-decreasing in $\overline{u}$ and non-increasing in u, and can take values of 0.

The more general form of G as a dual-input function is necessary for some cases, for example if one wants to have $G(\overline{u}, 1) = 0$ regardless of $\overline{u}$. It can still be a function of $u - \overline{u}$ hence this heuristic is compatible with the original formulation of TEL.

The policies are only slightly modified to include c_F, A, and the different form of G. They are given in Table 2. The ITEL algorithm is then described by Algorithm 1.

Table 2: ITEL Policies.

(a) Action Policy.

Mood	Utility	Decision
D	/	explore any a
C	$\overline{u} \notin A$	explore $a \neq \overline{a}$ with probability ε, else play $\overline{a}$
	$\overline{u} \in A$	play $\overline{a}$
H	/	play $\overline{a}$
W	/	play $\overline{a}$

(b) Update Policy.

Mood	Action	Utility	Decision
D	/	$F(u) > 0$	accept with probability $\varepsilon^{F(u)}$, else reject
		$F(u) = 0$	accept with probability c_F, else reject
C	$a \neq \overline{a}$	$u > \overline{u}$	accept with probability $\varepsilon^{G(\overline{u},u)}$, else revert
		$u \leq \overline{u}$	revert
	$a = \overline{a}$	$u > \overline{u}$	become H
		$u = \overline{u}$	revert
		$u < \overline{u}$	become W
H	/	$u > \overline{u}$	accept
		$u = \overline{u}$	revert
		$u < \overline{u}$	become W
W	/	$u > \overline{u}$	become H
		$u = \overline{u}$	revert
		$u < \overline{u}$	reject

Algorithm 1 ITEL

Initialize at any state $(\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}})$
for iterations $t = 1, 2, \dots$ **do**
  $a_i \leftarrow \text{ACTION}(\overline{m}_i, \overline{v}_i)$ according to Table 2a **for** $i = 1, \dots, n$
  $u_i \leftarrow \text{SAMPLE}(\boldsymbol{a}, i)$ **for** $i = 1, \dots, n$
  $(\overline{m}_i, \overline{a}_i, \overline{v}_i) \leftarrow \text{UPDATE}(\overline{m}_i, \overline{a}_i, \overline{u}_i, a_i, u_i)$ according to Table 2b **for** $i = 1, \dots, n$
end for

3.4 Theoretical Analysis of ITEL

The following discussion aims at extending the convergence result of TEL to ITEL. We call state a triplet $(\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) = (\overline{m}_i, \overline{a}_i, \overline{u}_i)_{i \in I}$ describing the states of all players at a given time in the algorithm, and $\mathcal{X}$ the set of all possible states. Remark that a discontent player's behavior does not depend on its benchmark $\overline{a}$ and $\overline{u}$, hence we reduce a discontent player's state to its discontent mood. In particular, the set of states where every player is discontent is identified with a single state we denote[2] D. Recall that our goal is to identify $\mathcal{X}^\star \subset \mathcal{X}$ the subset of states that minimize the potential γ.

Due to intermediate moods and discontent players playing at random, the benchmark utilities of a state are not always the result of the benchmark action profile. We identify the states where benchmark utilities correspond to the benchmark actions using the following definition.

[2] D is denoted $\overline{D}$ in [1].

Definition 3.1 (Aligned States)**.** A player of benchmark utility $\overline{u}$ is *aligned* with an action profile $\boldsymbol{a}$ if $\overline{u} = U(\boldsymbol{a})$. An all-content state is aligned if all players are aligned with the benchmark actions.

Denote[3] $\mathcal{C} \subset \mathcal{X}$ the subset of aligned all-content states. ITEL is designed so that content players tend to align themselves with the benchmark action profile when the noise ε is close to 0. Formally, we have the following proposition.

Proposition 3.1. The recurrence classes of the unperturbed process P^0 are the singletons $\{x\}$ for each aligned all-content states $x \in \mathcal{C}$, and possibly the communication class of the all-discontent state D.

Recall that TEL is shown to converge to Nash equilibria. For ITEL a similar result holds, although the class of favored states becomes slightly larger than Nash equilibria due to the introduction of admissibility. Hence, a different notion of equilibrium needs to be defined:

Definition 3.2 (Stable Equilibrium)**.** An action profile $\boldsymbol{a}$ with resulting utilities $\boldsymbol{u}$ is a stable equilibrium (SE) if for all player i one of the following two holds:

- i is admissible, i.e., $u_i \in A$.
- i is at an equilibrium position in the sense of Nash, i.e., for any action $a_i' \neq a_i$, $U_i(a_i', a_{-i}) \leq u_i$.

Note that Nash equilibria are SE as all players satisfy the second property. When $A = \emptyset$ or $A = \{1\}$, both notions are equal: in the first case there are no admissible players, in the second case admissible players maximize their utility hence are at an equilibrium position.

An aligned all-content state is said to be a SE if its benchmark action profile $\overline{\boldsymbol{a}}$ is itself a SE and we denote[4] $\mathcal{E} \subset \mathcal{C}$ the subset of aligned all-content SE states. Furthermore, a SE state is said to be admissible if all players are admissible and we denote $\mathcal{A} \subset \mathcal{E}$ the subset of aligned all-content admissible states.

Intuitively, in TEL Nash equilibria are particularly stable states, as no player can explore by itself and accept the outcome, which would necessarily be a deterioration in utility by definition of Nash equilibrium. In ITEL, the same reasoning applies when all players are at an equilibrium position. However it also applies when some players are admissible, as they cannot explore at all. Therefore one should expect ITEL to favor SE and not just Nash equilibria. This phenomenon will indeed appear in our main result. Moreover, admissible states are expected to be even more stable, as no player can explore. In fact, such states are absorbing states even in the unperturbed process. When no such state exists however, the perturbed process becomes irreducible, hence is a RPMP.

Proposition 3.2.

- If $\mathcal{A} \neq \emptyset$, states in $\mathcal{A}$ are absorbing for the perturbed process P^ε and the recurrence classes of P^ε are exactly the singletons $\{x\} \subset \mathcal{A}$.
- If $\mathcal{A} = \emptyset$, P^ε is aperiodic and irreducible.

It is common knowledge that an homogeneous Markov process converges almost surely (a.s.) to one of its recurrence classes. In the case where $\mathcal{A} \neq \emptyset$, Proposition 3.2 implies that ITEL converges a.s. to one of the states in $\mathcal{A}$. On the other hand, when $\mathcal{A} = \emptyset$, Proposition 3.2 implies that P^ε satisfies (1a), which was the remaining property to verify before concluding that P^ε is indeed a RPMP. Therefore, we will be able to apply Theorem 2.1 to the ITEL process.

Propositions 3.1 and 3.2 are highlighted key steps from the proof of our convergence result Theorem 3.3. They are proven along the proof of Theorem 3.3 in Appendix A.

Recall that, according to Theorem 2.1, the process will converge in some sense to states of minimal potential γ. We will see that in the case of ITEL – as it was for TEL [1] – the computation of a state's potential involves welfare and stability of the state, of which we recall the definition [1, (6)-(8)].[5]

[3] $\mathcal{C}$ is denoted C^0 in [1].
[4] $\mathcal{E}$ is denoted E^0 in [1].
[5] Here we consider average welfare instead of welfare, hence the normalization by n.

Definition 3.3 (Welfare and Stability)**.** Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{C}$. We define:

- $W(x) \triangleq \frac{1}{n} \sum_i \overline{u}_i$,
- $S(x) \triangleq \max\{U_i(a_i, \overline{a}_{-i}) - \overline{u}_i \quad | \quad i \text{ non-admissible}, a_i \neq \overline{a}_i : U_i(a_i, \overline{a}_{-i}) > \overline{u}_i\}$.

$S(x)$ is defined only if x is not a SE.

If F and G are chosen as affine function, the potential of states would feature W and S directly. However, for more general forms of F and G, we introduce notions of virtual welfare and virtual stability of a state which will be the ones appearing when computing potentials.

Definition 3.4 (Virtual Welfare and Stability)**.** Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{C}$. We define:

- $\tilde{W}(x) \triangleq 1 - \sum_i F(\overline{u}_i)$,
- $\tilde{S}(x) \triangleq \min\{G(\overline{u}_i, U_i(a_i, \overline{a}_{-i})) \quad | \quad i \text{ non-admissible}, a_i \neq \overline{a}_i : U_i(a_i, \overline{a}_{-i}) > \overline{u}_i\}$.

$\tilde{S}(x)$ is defined only if x is not a SE.

As F is non-increasing, $\tilde{W}(x)$ is non-decreasing in each individual utility $\overline{u}_i$. When F is chosen as a decreasing affine function of utility, $\tilde{W}(x)$ is an increasing affine function of $W(x)$.

The greater $\tilde{S}(x)$ is, the harder it is to leave state x, i.e., the more stable it is. On the other hand, stability corresponds to lower values of $S(x)$: $\tilde{S}(x)$ indicates the resistance needed to leave the state x, whereas $S(x)$ indicates how far it is from an equilibrium. Notice that $\tilde{S}(x) = G(S(x))$ when G is a non-increasing function of $u - \overline{u}$.

Reference [1] also features definitions $\tilde{W}$ and $\tilde{S}$ for any functions F which are similar to ours. We prefer our definitions as they are designed so that $\tilde{W}$ is non-negative, and $\tilde{S}$ supports the dual-input form of G.

In order to establish our result, some bounds on $\tilde{W}$ and $\tilde{S}$ are required. In the original TEL formulation C1 is introduced for this reason. We use a looser version of this condition under H1–H3.

Condition C2 (Minimal bounds on resistance functions in ITEL). There exist constants F_0 and G_0 such that for all utilities u and $\overline{u} < u$,

$$\begin{cases} 0 \leq F(u) \leq \frac{F_0}{n} \\ 0 \leq G(\overline{u}, u) < G_0 \\ F_0 + G_0 \leq 1 \end{cases} \tag{C2}$$

Note that if $F_0 = G_0 = \frac{1}{2}$, C2 is the same as C1 except some inequalities are no longer strict. We can finally state our convergence results:

Theorem 3.3. Assume that A1 and C2 hold and that $\mathcal{A} = \emptyset$. Then the ITEL process is a RPMP and the states of minimum potential are:

- If $\mathcal{E} \neq \emptyset$, $\mathcal{X}^\star = \arg\max_{x \in \mathcal{E}} \tilde{W}(x)$.
- Else, $\mathcal{X}^\star = \arg\max_{x \in \mathcal{C}} \tilde{W}(x) + \tilde{S}(x)$.

Theorem 3.3 is proven in Appendix A. Now, combining Proposition 3.2 and Theorems 2.1 and 3.3:

Theorem 3.4 (ITEL convergence)**.** Assume that A1 and C2 hold:

- The ITEL process converges a.s. to an admissible aligned all-content state.
- If there is no admissible state, the stochastically stable states are SE aligned all-content states maximizing $\tilde{W}$.
- If there is no SE state, the stochastically stable states are aligned all-content states maximizing $\tilde{W} + \tilde{S}$.

In the case where F and G are chosen of the form $F : u \mapsto \phi_F - \psi_F \cdot u$ and $G : (\overline{u}, u) \mapsto \phi_G - \psi_G \cdot (u - \overline{u})$, maximizing $\tilde{W}$ can be replaced with maximizing W in the second case and maximizing $\tilde{W} + \tilde{S}$ can be replaced with maximizing $\psi_F W - \psi_G S$ in the third case. The reasoning to obtain this result is detailed at the end of Appendix A. Note that the stability part of the trade-off in this last case goes from maximizing $\tilde{S}$ to maximizing $-S$. Indeed we have already discussed that the notion of high stability is associated with high values of $\tilde{S}$ and with low values of S.

3.5 IODL

In this section we briefly explain the differences in how ODL [2] is improved with regards to TEL. Recall that ODL does not feature intermediate moods and maximizes global welfare regardless of being at a Nash equilibria or not. Although the convergence is theoretically proven, ODL presents several behaviors that seem unintuitive, e.g. content players can become discontent even after having observed an improvement in utility. We will not describe in details the original ODL formulation in this document, but instead give a different formulation which is closer to TEL removing intermediate steps. Using the same heuristics as before, the policies of IODL are given in Table 3. The IODL algorithm is obtained by replacing the policies in Algorithm 1.

Table 3: IODL Policies.

(a) Action Policy.

Mood	Utility	Decision
D	/	explore any a
C	$\overline{u} \notin A$	explore $a \neq \overline{a}$ with probability ε, else play $\overline{a}$
	$\overline{u} \in A$	play $\overline{a}$

(b) Update Policy.

Mood	Action	Utility	Decision
D	/	$F(u) > 0$	accept with probability $\varepsilon^{F(u)}$, else reject
		$F(u) = 0$	accept with probability c_F, else reject
C	$a \neq \overline{a}$	$u > \overline{u}$	accept with probability $\varepsilon^{G(\overline{u},u)}$, else revert
		$u \leq \overline{u}$	revert
	$a = \overline{a}$	$u > \overline{u}$	accept
		$u = \overline{u}$	revert
		$u < \overline{u}$	reject

Note that contrarily to TEL and ITEL, ODL cannot be described as a particular choice of parameters in IODL. The removal of intermediate moods makes IODL unable to discriminate equilibria. Hence the definitions of $\mathcal{E}$ and $\tilde{S}$ are no longer needed, but we still consider $\mathcal{C}$, $\mathcal{A}$ and $\tilde{W}$. Contrarily to ITEL, G has a superficial role theoretically-wise and no particular bound is needed for it. The bound on F is the following:

Condition C3 (Minimal bounds on resistance functions in IODL). For all utilities u,

$$0 \leq F(u) < \frac{1}{n} \tag{C3}$$

Propositions 3.1 and 3.2 hold for IODL. Furthermore, a similar, simpler result to Theorem 3.4 holds:

Theorem 3.5 (IODL convergence)**.** Assume that A1 and C3 hold:

- The IODL process converges a.s. to an admissible aligned all-content state.
- If there is no admissible state, the stochastically stable states are aligned all-content states maximizing $\tilde{W}$.

As we previously said, G holds no theoretical role in IODL hence does not appear in the convergence statement. In the case where F is chosen of the form $F : u \mapsto \phi_F - \psi_F \cdot u$, $\tilde{W}$ can be replaced with W in Theorem 3.5.

Theorem 3.5 is proven by identifying $\mathcal{X}^\star$ in a similar way as in ITEL. The proof is mainly the same, and the few differences are highlighted in Appendix B.

4 Stochastic Games

In this section we consider stochastic games. In a given action profile $\boldsymbol{a}$, player i observes a utility $U_i(\boldsymbol{a})$ that is a random variable bounded a.s. in $[0, 1]$. Apart from this assumption, utilities may follow any probability distribution: continuous, discrete, and some may even be deterministic. The introduction of noisy utilities yields the following issues, both from a theoretical or practical standpoint.

Instability When no player explores, different payoffs may still be observed due to the random nature of utilities. Under ITEL, those differences would be wrongly interpreted as changes of behaviors from other players.

Infinite space state In ITEL, the proof of convergence is based on the study of the RPMP induced by the algorithm. This process acts over the space of all states $(\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}})$ of possible moods, actions, and utilities for each players. If the support of the random utilities is infinite, which would be the case for continuous distributions, then the state space becomes infinite and the theory of RPMPs no longer applies.

Non-Regularity An RPMP must have transition probabilities that behave as ε^r for some $r \geq 0$. We will see that this property is not verified in general when utilities are random.

The goal of this section is to adapt ITEL in order to tackle these issues and identify optimal states. As for ITEL, we aim at converging to the equilibrium of maximal welfare, or to the action profile of maximal trade-off between welfare and stability. Welfare and stability of action profiles, will be defined by considering the expected utilities.

4.1 RITEL

We introduce RITEL, an adaptation of ITEL in the stochastic setup. Let us describe the differences between ITEL and RITEL.

First, iterations are divided in periods of length τ. Within a period, each player commits to a fixed action so that the action profile played is constant. This way, the variations in utility a player may observe along the period are solely due to noise, so that it can estimate more reliably the average reward it should receive in this action profile. When the period ends, each player compares the estimate of their average reward to their benchmark utility and update their state in a similar manner as in ITEL.

Secondly, in order to consider only a finite amount of possible states – which is needed for theoretical analysis – benchmark utilities will be restricted to a finite amount of bins, slicing the utility range $[0, 1]$ into bins $[0, \delta), [\delta, 2\delta), \dots, [1-\delta, 1), [1, 1+\delta)$ for some $\delta > 0$ chosen as the inverse of an integer. The highest bin could be written as $\{1\}$ since utilities are bounded a.s. by 1, however we choose this notation for coherence. We denote B the function mapping a utility u to the bin v it belongs to. This implies a loss of accuracy that will eventually appear in our main result: RITEL will not be able to discriminate two action profiles when their welfare are too close. Choosing smaller δ will allow finer results, however we will see that longer periods are then needed to ensure convergence. Precisely, we will see that choosing a period length of the form $\tau = \lceil \tau_0 \log(\frac{1}{\varepsilon}) \rceil$ for some constant τ_0 ensures that the effect of offset observations is negligible. The particular choice of τ_0 will be guided by a condition which essentially boils down to the fact that better accuracy, i.e., smaller δ, requires higher reliability on the samples, i.e., higher τ_0.

Thirdly, when comparing utilities to choose a behavior, a player will actually compare the utilities' bins. Furthermore, to avoid that a player whose average utility is close to a border between two bins constantly

switches between both bins, we only allow a player to consider a change of utility as significant enough to act on it when the new utility is neither in the benchmark bin, nor in an adjacent bin, implying a gap in utilities of at least δ. This behavior will affect the accuracy of the convergence, and requires some changes in prior definition to take δ into account.

4.1.1 Notations

In the following, actual utilities will be denoted by the letter u whereas utility bins will be denoted by the letter v. v^- and v^+ will denote the two edges of a bin v, i.e., $v = [v^-, v^+)$ with $v^+ = v^- + \delta$. Given some bin $v = [v^-, v^+)$, we denote $v + k\delta = [(v + k\delta)^-, (v + k\delta)^+) = [v^- + k\delta, v^+ + k\delta)$ and use comparison operators $=$, $\leq$, $\geq$, $<$, and $>$ using the obvious order of the bins. We write $v' = v \pm \delta$ (resp. $v' \neq v \pm \delta$) if $v' \in \{v - \delta, v, v + \delta\}$ (resp. $v' \notin \{v - \delta, v, v + \delta\}$).

Given any action profile $\boldsymbol{a}$, we denote $M_i(\boldsymbol{a}) = \mathbb{E}[U_i(\boldsymbol{a})]$ the expected utility of player i in the action profile, and $N_i(\boldsymbol{a}) = B(M_i(\boldsymbol{a}))$ the corresponding bin. The convention for one step of the RITEL algorithm is as follows: a player is in state $(\overline{m}, \overline{a}, \overline{v})$ and plays a for τ steps, observing i.i.d. samples $U_1, U_2, \ldots, U_\tau$ of distribution $U(\boldsymbol{a})$. The player computes its average utility $u = U^{(\tau)} \triangleq \frac{1}{\tau}\sum_{k=1}^{\tau} U_k$ and the corresponding bin $v = B(u)$. Expected utilities will be denoted $\overline{\mu} = \mathbb{E}[U(\overline{\boldsymbol{a}})]$ and $\mu = \mathbb{E}[U(\boldsymbol{a})]$. Their respective bins will be denoted $\overline{\nu} = B(\overline{\mu})$ and $\nu = B(\mu)$, and referred to as expected bins.[6] From now on, the term "offset observation" will always refer to a player observing an utility bin $v \neq \nu \pm \delta$ where ν is the expected bin of the player in the played action profile.

Finally, F and G are introduced with the same purpose as in ITEL, although they are used as functions of bins instead of true utilities. We define them over bins by applying them to the lower bounds of the bins: $F(v) \triangleq F(v^-)$ and $G(\overline{v}, v) \triangleq G(\overline{v}^-, v^-)$. Similarly, $A \subset [0, 1]$ is introduced with the purpose as in ITEL. A utility u is admissible if $u \in A$ whereas a bin v is admissible – denoted $v \in A$ – if $v \cap A \neq \emptyset$.

4.1.2 Algorithm

The policies of RITEL and the RITEL algorithm itself are detailed in Table 4 and Algorithm 2. Notice that the policies are essentially the same as in ITEL, except that players make decision based on bins instead of utilities. They also have a small tolerance to change: observing bins $\overline{v} - \delta, \overline{v}, \overline{v} + \delta$ is interpreted as if no change had happened, whereas observing bins $\overline{v} + 2\delta, \overline{v} + 3\delta, \ldots$ is interpreted as an improvement and observing bins $\overline{v} - 2\delta, \overline{v} - 3\delta, \ldots$ is interpreted as a deterioration.

[6] Beware that here "expected bin" refers to the bin of the expected utility, not to the expectation of the observed bin, which may differ.

Table 4: RITEL Policies.

(a) Action Policy.

Mood	Utility	Decision
D	/	explore any a
C	$\overline{v} \notin A$	explore $a \neq \overline{a}$ with probability ε, else play $\overline{a}$
	$\overline{v} \in A$	play $\overline{a}$
H	/	play $\overline{a}$
W	/	play $\overline{a}$

(b) Update Policy.

Mood	Action	Utility	Decision
D	/	$F(v) > 0$	accept with probability $\varepsilon^{F(v)}$, else reject
		$F(v) = 0$	accept with probability c_F, else reject
C	$a \neq \overline{a}$	$v \geq \overline{v} + 2\delta$	accept with probability $\varepsilon^{G(\overline{v},v)}$, else revert
		$v \leq \overline{v} + \delta$	revert
	$a = \overline{a}$	$v > \overline{v} + \delta$	become H
		$v = \overline{v} \pm \delta$	revert
		$v \leq \overline{v} - 2\delta$	become W
H	/	$v \geq \overline{v} + 2\delta$	accept
		$v = \overline{v} \pm \delta$	revert
		$v \leq \overline{v} - 2\delta$	become W
W	/	$v \geq \overline{v} + 2\delta$	become H
		$v = \overline{v} \pm \delta$	revert
		$v \leq \overline{v} - 2\delta$	reject

Algorithm 2 RITEL

Initialize at any state $(\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}})$
for periods $t = 1, 2, \ldots$ **do**
  $a_i \leftarrow \text{ACTION}(\overline{m}_i, \overline{v}_i)$ according to Table 4a **for** $i = 1, \ldots, n$
  for $k = 1, \ldots, \tau$ **do**
    $U_{k,i} \leftarrow \text{SAMPLE}(\boldsymbol{a}, i)$ **for** $i = 1, \ldots, n$
  end for
  $u_i \leftarrow \frac{1}{\tau} \sum_{k=1}^{\tau} U_{i,k}$ **for** $i = 1, \ldots, n$
  $v_i \leftarrow B(u_i)$ **for** $i = 1, \ldots, n$
  $(\overline{m}_i, \overline{a}_i, \overline{v}_i) \leftarrow \text{UPDATE}(\overline{m}_i, \overline{a}_i, \overline{v}_i, a_i, v_i)$ according to Table 4b **for** $i = 1, \ldots, n$
end for

4.1.3 Use of the RPMP Framework

Let us give insight as to why the RPMP framework may be used in the context of RITEL. The Markov process P^{ε} we will consider for some given ε is the state of the algorithm at the beginning of each period. To apply the RPMP framework, one would need to verify the conditions given by Definitions 2.1 and 2.2. Similarly to ITEL, the condition (1a) is not always verified however the cases where it is not are easily treated. Its validity is discussed in detail through Proposition 4.12. For now we focus on discussing the validity of (1b) and (2).

First of all, notice that since the process is updated only at the end of each periods, the only effect of the introduction of said periods is to reduce the variance of observations. Consider a process without periods but

where the distributions of utilities are replaced with the distributions of their respective empirical average over τ samples. This process has actually the same law as the RITEL process. This alternate view of the process allows to define the unperturbed process P^0. Indeed P^0 would correspond to infinite-length periods as τ is chosen as a function of ε such that $\lim_{\varepsilon\to 0}\tau(\varepsilon) = +\infty$. Such process cannot exist in reality, however we can model P^0 by defining it as the limit of P^ε when $\varepsilon \to 0$, with our alternate view where instead of periods the utilities themselves are replaced with averages. This is assuming, of course, that the limit exists, i.e., that (1b) holds. We can show that it is indeed the case.

In both ITEL and RITEL, a player's behavior depends on its observed utility. The difference being that in RITEL, the observed utility is the empirical mean of τ i.i.d. observations and is quantified to a bin of length δ. Hence, the probability that a player chooses some behavior is a combinations of probabilities of observing in given bins, and of probabilities of choosing a behavior given the observed bin. For example, the probability that a content player of benchmark bin $\overline{v}$ who explored and observes $U^{(\tau)}$ accepts its exploration with new benchmark bin v is $\mathbb{P}(U^{(\tau)} \in v)\varepsilon^{G(\overline{v},v)}$. The probability that a hopeful player of benchmark bin $\overline{v}$ who observes $U^{(\tau)}$ becomes watchful is $\sum_{v\leq\overline{v}-2\delta}\mathbb{P}(U^{(\tau)} \in v)$.

Since $\lim_{\varepsilon\to 0}\tau(\varepsilon) = +\infty$, the Law of Large Numbers implies that as $\varepsilon \to 0$, the quantities $\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ converge to 1 when $\mu = \mathbb{E}[U] \in v$, else converge to 0. There is one exception which is when U is not deterministic and $\mu = v^-$, that is the expectation of U is exactly between two bins. In this case both bins v and $v-\delta$ are observed with probability $\frac{1}{2}$. Indeed the distribution of the empirical mean becomes symmetric as $\tau \to +\infty$ due to the Central Limit Theorem. Regardless of the case, the key property to retain for now is that $\lim_{\varepsilon\to 0}\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ always exists. From there the same reasoning as in ITEL concludes that all transition probabilities $P^\varepsilon_{x,y}$ converge to a limit $P^0_{x,y}$ as $\varepsilon \to 0$.

In particular, in P^0 the observed utilities are deterministic (except for the special case of expected utilities at the edge of two bins). Therefore we should expect P^0 to be similar to the unperturbed process of ITEL.

Now that the case of (1b) is cleared, we need to discuss the validity of the regularity condition (2) for all $P^\varepsilon_{x,y}$. From the above discussion we can see that this condition would be satisfied if and only if $\mathbb{P}(U^{(\tau)} \in v)$ is regular for all random utilities U and all bins v.

This is unfortunately not true in general, and Section 4.2.2 gives some insight as to why. However, a slightly weaker result happens to be satisfied by any distribution. The following two sections are devoted to this discussion. Section 4.2 introduces large deviation results which eventually show that quantities $\mathbb{P}(U^{(\tau)} \in v)$ satisfy a weaker sense of regularity. Section 4.3 adapts the theory of RPMPs to fit this new regularity condition and yields a similar result to Theorem 2.1. From there we will be able to analyze the convergence of RITEL in Section 4.4.

4.2 Resistance and Noise

The goal of this section is to show that for any utility distribution U and bin v, the probability $\mathbb{P}(U^{(\tau)} \in v)$ satisfies the following definition.

Definition 4.1 (Almost Regularity)**.** A family $(X^\varepsilon)_{0\leq\varepsilon<\varepsilon_0}$ of non-negative real numbers is *almost regular* if there exists $r \geq 0$ such that

$$\begin{cases} \lim_{\varepsilon\to 0}\varepsilon^{-r'}X^\varepsilon = 0 & \text{when } r' < r, \\ \lim_{\varepsilon\to 0}\varepsilon^{-r'}X^\varepsilon = +\infty & \text{when } r' > r. \end{cases} \tag{3}$$

r is unique and called the *resistance* of X^ε. r can be equal to $+\infty$, which is the case in particular when $X^\varepsilon = 0$.

It is immediate that regularity defined in Definition 2.2 implies almost regularity, and that both notions of resistance coincide in this case. Almost regularity suggests that X^ε behaves somewhat like ε^r but with less accuracy than regularity. For example, the sequence $(\log(\frac{1}{\varepsilon})^\alpha\varepsilon^r)_{\varepsilon>0}$ is almost regular with resistance r regardless of $\alpha \in \mathbb{R}$, although $\log(\frac{1}{\varepsilon})^\alpha$ may converge to 0 or to $+\infty$.

The key result to show that almost regularity is verified in the context of RITEL is known as Cramér's theorem. For a real-valued random variable U, define its logarithmic moment generating function (LMGF)

$$\Lambda_U : t \mapsto \log(\mathbb{E}[\exp(tU)]) \tag{4}$$

along with its Legendre transform

$$\Lambda_U^* : x \mapsto \sup_{t \in \mathbb{R}} (tx - \Lambda_U(t)). \tag{5}$$

Theorem 4.1 (Cramér [4, Corollary 2.2.19])**.** Let U be any non-degenerate random variable – i.e., with non-zero variance – with finite LMGF and expectation μ. Then Λ_U^* is non-negative, convex, decreasing on $(-\infty, \mu]$ and increasing on $[\mu, +\infty)$, with $\Lambda_U^*(\mu) = 0$. Moreover, the empirical mean $U^{(\tau)}$ satisfies a large deviation principle with rate function Λ_U^*, i.e.,

- For all $x > \mu$,

$$\lim_{\tau \to +\infty} \frac{1}{\tau} \log\left(\mathbb{P}(U^{(\tau)} \geq x)\right) = -\Lambda_U^*(x). \tag{6}$$

- For all $x < \mu$,

$$\lim_{\tau \to +\infty} \frac{1}{\tau} \log\left(\mathbb{P}(U^{(\tau)} < x)\right) = -\Lambda_U^*(x). \tag{7}$$

In the above statement, $\mathbb{P}(U^{(\tau)} \geq x)$ could be replaced by $\mathbb{P}(U^{(\tau)} > x)$ with the same result. Similarly $\mathbb{P}(U^{(\tau)} < x)$ could be replaced by $\mathbb{P}(U^{(\tau)} \leq x)$. In fact, due to the fact that bins are of the form $[v^-, v^+)$, we only need the two results stated in Theorem 4.1. Note that the formulation of the theorem given in [4, Corollary 2.2.19] does not include the condition on $x > \mu$, however $\Lambda_U^*(x)$ is replaced with $\inf_{y \geq x} \Lambda_U^*(y)$. Adding the condition $x > \mu$ together with the fact that Λ_U^* is increasing over $[\mu, +\infty)$ allows us to write our statement. The same reasoning holds for the lower tail.

We have already discussed the need for τ to diverge to $+\infty$ when $\varepsilon \to 0$. We can now show that for a specific dependency in ε, Theorem 4.1 implies almost regularity.

Proposition 4.2. Let $\tau(\varepsilon) = \lceil \tau_0 \log(\frac{1}{\varepsilon}) \rceil$ for some constant τ_0. Under the same conditions as in Theorem 4.1, the following holds:

- For all $x > \mu$, $(\mathbb{P}(U^{(\tau(\varepsilon))} \geq x))_{\varepsilon > 0}$ is almost regular with resistance $\tau_0 \Lambda_U^*(x)$.
- For all $x < \mu$, $(\mathbb{P}(U^{(\tau(\varepsilon))} < x))_{\varepsilon > 0}$ is almost regular with resistance $\tau_0 \Lambda_U^*(x)$.

Proof. It suffices to show that if some quantity $X^{(\tau)}$ satisfies

$$\lim_{\tau \to +\infty} \frac{1}{\tau} \log\left(X^{(\tau)}\right) = -r \tag{8}$$

for some $r \geq 0$, then $X^{\varepsilon} = (X^{(\tau(\varepsilon))})_{\varepsilon > 0}$ is almost regular with resistance $\tau_0 r$. Indeed applying such result to $\mathbb{P}(U^{(\tau)} \geq x)$ and $\mathbb{P}(U^{(\tau)} < x)$ and using Theorem 4.1 yields Proposition 4.2. Assume (8) holds. Let $r' \geq 0$. We have

$$\frac{1}{\tau(\varepsilon)} \log\left(e^{\tau(\varepsilon) r'} X^{(\tau(\varepsilon))}\right) = r' + \frac{1}{\tau(\varepsilon)} \log\left(X^{(\tau(\varepsilon))}\right) \underset{\varepsilon \to 0}{\to} r' - r \tag{9}$$

as $\lim_{\varepsilon \to 0} \tau(\varepsilon) \to +\infty$. It follows that

- If $r' < r$, $\lim_{\varepsilon \to 0} \log\left(e^{\tau(\varepsilon) r'} X^{(\tau(\varepsilon))}\right) = -\infty$, hence $\lim_{\varepsilon \to 0} e^{\tau(\varepsilon) r'} X^{(\tau(\varepsilon))} = 0$.
- If $r' > r$, $\lim_{\varepsilon \to 0} \log\left(e^{\tau(\varepsilon) r'} X^{(\tau(\varepsilon))}\right) = +\infty$, hence $\lim_{\varepsilon \to 0} e^{\tau(\varepsilon) r'} X^{(\tau(\varepsilon))} = +\infty$.

Now, since $\tau(\varepsilon) = \lceil \tau_0 \log(\frac{1}{\varepsilon}) \rceil$, we have $\tau(\varepsilon) - 1 \leq \tau_0 \log(\frac{1}{\varepsilon}) \leq \tau(\varepsilon)$, hence $\frac{1}{e} e^{\tau(\varepsilon) r'} \leq \varepsilon^{-\tau_0 r'} \leq e^{\tau(\varepsilon) r'}$. Hence replacing $e^{\tau(\varepsilon) r'}$ with $\varepsilon^{-\tau_0 r'}$ in the above statements yields the same limits, and we conclude that $(X^{(\tau(\varepsilon))})_{\varepsilon > 0}$ is almost regular with resistance $\tau_0 r$. □

Recall that we are ultimately interested in $\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ for any bin v. This is deduced directly from the above using the fact that Λ_U^* is decreasing over $(-\infty, \mu]$ and increasing over $[\mu, +\infty)$.

Corollary 4.3. Let $\tau(\varepsilon) = \lceil \tau_0 \log(\frac{1}{\varepsilon}) \rceil$. Under the same conditions as in Theorem 4.1 and for all bin v, $\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ is almost regular. Moreover, denoting $r_U(v)$ the corresponding resistance, we have:

$$r_U(v) = \begin{cases} \tau_0 \Lambda_U^*(v^-) & \text{if } \mu < v^- \\ 0 & \text{if } v^- \leq \mu \leq v^+ \\ \tau_0 \Lambda_U^*(v^+) & \text{if } \mu > v^+ \end{cases} \tag{10}$$

Remark. If U is deterministic, i.e., $U = \mu$ a.s., then the same result holds, replacing Λ_U^* with $+\infty$ when $\mu \notin v$. When $\mu = v^+$, we also have $r_U(v) = +\infty$ and not 0, as $U = \mu = v^+ = (v + \delta)^-$ a.s., which is considered part of the bin $v + \delta$ and not v.

Proof. If $v^- \leq \mu \leq v^+$, we already know that this probability converges to a positive limit as $\varepsilon \to 0$ (precisely, to 1 if the inequalities are strict, else to $\frac{1}{2}$), which implies that $\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ is almost regular with resistance 0. If $\mu < v^-$, one can write $\mathbb{P}(U^{(\tau(\varepsilon))} \in v) = \mathbb{P}(U^{(\tau(\varepsilon))} \geq v^-) - \mathbb{P}(U^{(\tau(\varepsilon))} \geq v^+)$ and both terms are almost regular with resistance $\tau_0 \Lambda_U^*(v^-)$ and $\tau_0 \Lambda_U^*(v^+)$ respectively, according to Proposition 4.2. Since Λ_U^* is increasing over $[\mu, +\infty)$, $\varepsilon^{\tau_0 \Lambda_U^*(v^+)}$ vanishes exponentially faster than $\varepsilon^{\tau_0 \Lambda_U^*(v^-)}$, therefore the second term in $\mathbb{P}(U^{(\tau(\varepsilon))} \in v) = \mathbb{P}(U^{(\tau(\varepsilon))} \geq v^-) - \mathbb{P}(U^{(\tau(\varepsilon))} \geq v^+)$ becomes negligible as $\varepsilon \to 0$, hence $\mathbb{P}(U^{(\tau(\varepsilon))} \in v)$ is almost regular with resistance $\tau_0 \Lambda_U^*(v^-)$. If $\mu > v^+$, $\mathbb{P}(U^{(\tau(\varepsilon))} \in v) = \mathbb{P}(U^{(\tau(\varepsilon))} < v^+) - \mathbb{P}(U^{(\tau(\varepsilon))} < v^-)$ is almost regular with resistance $\tau_0 \Lambda_U^*(v^+)$ using the same reasoning. □

Regardless of the theoretical requirement of regularity, we also need to be able to control the probability of offset observations – in other words, to lower bound their resistance – in order to analyze the stability of aligned states. Two methods are possible to obtain such lower bounds. When utility distributions are known and the LMGF can be computed, a lower bound on Λ_U^* – or even a closed form formula – can be obtained. In general, several inequalities exist to upper bound $\mathbb{P}(U^{(\tau)} > x)$ with minimal information on the distribution of U, which in turn allow us to lower bound Λ_U^*. We will briefly discuss the case of gaussian distributions in Section 4.2.2 as it provides justification as to why the notion of almost regularity is required. The next section is devoted to general deviation bounds that hold for any bounded distributions.

4.2.1 General Deviation Bounds

In this section we state two common concentration inequalities which yield simple general bounds on Λ_U^*. Ultimately, such bounds allows us to lower bound the resistance of offset observations. Since we will eventually need such observations to have high enough resistance, the upcoming bounds will translate to a condition between τ_0, δ and σ^2 the variance of observed utilities.

Lemma 4.4 (Hoeffding's Inequality [5, 7.3])**.** Assume U is bounded a.s. in $[0, 1]$ and denote μ its expectation. Then,

$$\forall x \geq \mu, \quad \mathbb{P}(U^{(\tau)} \geq x) \leq \exp\big(-2\tau(x - \mu)^2\big), \tag{11}$$

$$\forall x \leq \mu, \quad \mathbb{P}(U^{(\tau)} \leq x) \leq \exp\big(-2\tau(x - \mu)^2\big). \tag{12}$$

Lemma 4.5 (Bernstein's Inequality [5, 7.5])**.** Assume U is bounded a.s. in $[0,1]$ and denote μ its expectation and σ^2 its variance. Then,[7]

$$\forall x \geq \mu, \quad \mathbb{P}(U^{(\tau)} \geq x) \leq \exp\left(-\tau \frac{(x-\mu)^2}{2\sigma^2 + 2|x-\mu|}\right), \tag{13}$$

$$\forall x \leq \mu, \quad \mathbb{P}(U^{(\tau)} \leq x) \leq \exp\left(-\tau \frac{(x-\mu)^2}{2\sigma^2 + 2|x-\mu|}\right). \tag{14}$$

Lemma 4.4 requires no information on the utility distributions besides the bounded assumption. On the other hand, Lemma 4.5 is able to yield a tighter bound when variance is known to be small. In the case where variance is unknown, being bounded in $[0,1]$ implies that it is at most $\frac{1}{4}$, however the bound given by Lemma 4.5 in this case is looser than that of Lemma 4.4, so that both inequalities can be useful. Combining Theorem 4.1 and Lemmas 4.4 and 4.5, we get for all $x \neq \mu$,

$$\Lambda_U^*(x) \geq \max\left(2(x-\mu)^2, \frac{(x-\mu)^2}{2\sigma^2 + 2|x-\mu|}\right). \tag{15}$$

Indeed, with Lemma 4.4, when $x > \mu$,

$$\frac{1}{\tau}\log\left(\mathbb{P}(U^{(\tau)} \geq x)\right) \leq \frac{1}{\tau}\log\left(\exp\left(-2\tau(x-\mu)^2\right)\right) = -2(x-\mu)^2,$$

which gives $\Lambda_U^*(x) \geq 2(x-\mu)^2$ when taking the limit $\tau \to +\infty$. The same reasoning gives $\Lambda_U^*(x) \geq \frac{(x-\mu)^2}{2\sigma^2+2|x-\mu|}$ using Lemma 4.5, and same goes for the lower tail.

Inequality (15) will be key to control the resistance associated with offset observations. Note that other inequalities such as subgaussian bounds or Bennett's inequality could be used to obtain tighter bounds depending on one's knowledge of the utility distributions. This discussion falls under the theory of large deviations and is independent from our work.

4.2.2 Case of a gaussian distribution

In the case of a gaussian distribution, we can analytically compute an equivalent of the deviation $\mathbb{P}(U^{(\tau)} > x)$. Gaussian distributions satisfy the following properties:

$$\begin{cases} \mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \\ \alpha \cdot \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(\alpha\mu, (\alpha\sigma)^2). \end{cases}$$

Then if $U \sim \mathcal{N}(\mu, \sigma^2)$, it follows that $U^{(\tau)} \sim \mathcal{N}(\mu, \frac{\sigma^2}{\tau})$, hence that $\frac{\sqrt{\tau}}{\sigma}(U^{(\tau)} - \mu) \sim \mathcal{N}(0,1)$. It can be shown ([6, (5) and (8)]) that when $Z \sim \mathcal{N}(0,1)$,

$$\mathbb{P}(Z > x) \underset{x \to +\infty}{\sim} \frac{1}{\sqrt{2\pi}x} e^{-\frac{x^2}{2}}, \tag{16}$$

hence, when $x > \mu$,

$$\mathbb{P}(U^{(\tau)} > x) = \mathbb{P}\left(\frac{\sqrt{\tau}}{\sigma}(U^{(\tau)} - \mu) > \frac{\sqrt{\tau}}{\sigma}(x-\mu)\right) \underset{\tau \to +\infty}{\sim} \frac{\sigma}{\sqrt{2\pi\tau}(x-\mu)} e^{-\tau\frac{(x-\mu)^2}{2\sigma^2}} \tag{17}$$

In particular,

$$\frac{1}{\tau}\log(\mathbb{P}(U^{(\tau)} > x)) \underset{\tau \to +\infty}{\sim} -\frac{1}{2\tau}\log\left(\frac{2\pi\tau(x-\mu)^2}{\sigma^2}\right) - \frac{(x-\mu)^2}{2\sigma^2} \underset{\tau \to +\infty}{\to} -\frac{(x-\mu)^2}{2\sigma^2} \tag{18}$$

[7] Bernstein's inequality is stated for centered variables bounded a.s. by some $a > 0$. Here we apply the result to $U - \mu$ which is bounded a.s. by 1. The result is deduced for the lower tail by symmetry.

The same reasoning applies for the lower tail. This proves Theorem 4.1 for the case of gaussian distributions, as computing explicitly the LMGF would also give $\Lambda_U^*(x) = \frac{(x-\mu)^2}{2\sigma^2}$.

The equivalent (17) shows that the introduction of almost regularity is necessary. Indeed, here $\mathbb{P}(U^{(\tau)} > x)$ is of order $\frac{1}{\sqrt{\tau}} e^{-\tau r}$, whereas regularity as defined in (2) would impose an order of $e^{-\tau r}$.

4.3 Almost Regular Perturbed Markov Processes

In this section we show that the theory of RPMPs can be extended by replacing the regularity condition with almost regularity. A theorem similar to Theorem 2.1 can then be derived to identify stochastically stable states, although at the cost of a small loss of accuracy.

Definition 4.2 (Almost Regular Perturbed Markov Process)**.** Let P^ε be a PMP over a state space $\mathcal{X}$. P^ε is said to be an *almost regular perturbed Markov process (ARPMP)* if all transition probabilities are almost regular, that is:

$$\forall x, y \in \mathcal{X},\ \exists r \geq 0 : \begin{cases} \lim_{\varepsilon \to 0} \varepsilon^{-r'} P^\varepsilon_{x,y} = 0 & \text{if } r' < r, \\ \lim_{\varepsilon \to 0} \varepsilon^{-r'} P^\varepsilon_{x,y} = +\infty & \text{if } r' > r. \end{cases} \tag{19}$$

As in Definition 2.2, r is called the resistance of the transition and can be equal to $+\infty$. This is the case in particular for non-existing transitions $P^\varepsilon_{x,y} = 0$. Note that the tansition probabilities of a RPMP satisfy (2) which implies (3). The resistance of a transition is unique and the same for both definitions. Hence we extend the notions of resistance $r(x \to y)$, resistance graph $\mathcal{G}$, rooted tree, and potential γ of the regular case to the almost regular one, along with all related definitions. In particular Corollary 4.3 implies that (19) is satisfied by the transition probabilities of RITEL, so that we can use the ARPMP framework in this context.

When the process is not regular we loose the information on the behavior of $\varepsilon^{-r} P^\varepsilon_{x,y}$. In particular we are unable to know how two transition $P^\varepsilon_{x,y}$ and $P^\varepsilon_{x',y'}$ of equal resistance behave relatively, as their ratio can be any quantity that is sub-polynomial with regards to ε. This will lead to not being able to describe as accurately as before the behavior of π^ε_x for states $x \in \mathcal{X}^\star$ when $\varepsilon \to 0$. However we are able to adapt Theorem 2.1 to show that states outside of $\mathcal{X}^\star$ will be vanishing even under the looser assumption of almost regularity.

Theorem 4.6. Let P^ε be an ARPMP over a finite space $\mathcal{X}$. Denote π^ε its stationary distribution for every small positive ε. Then, the process only visits $\mathcal{X}^\star$ as $\varepsilon \to 0$, that is,

$$\lim_{\varepsilon \to 0} \pi^\varepsilon_{\mathcal{X}^\star} = 1. \tag{20}$$

Remark. The loss of accuracy here is that the visiting rate π^ε_x of a particular state $x \in \mathcal{X}^\star$ remains unknown, and may still vanish or even diverge. In comparison, for a regular process, π^ε_x was known to converge to a positive limit in Theorem 2.1

Proof. The proof of Theorem 2.1 [3, Lemma 1] states that π^ε can be expressed as follows: let

$$p^\varepsilon_x \triangleq \sum_{T : x\text{-tree}} \prod_{(y \to z) \in T} P^\varepsilon_{y,z} \tag{21}$$

for all $x \in \mathcal{X}$. Then

$$\pi^\varepsilon_x = \frac{p^\varepsilon_x}{\sum_{y \in \mathcal{X}} p^\varepsilon_y}. \tag{22}$$

Proving Theorem 2.1 consists in showing that for all $x \in \mathcal{X}$, $\lim_{\varepsilon \to 0} \varepsilon^{-\gamma(x)} p^\varepsilon_x \in (0, +\infty)$, i.e., p^ε_x is regular with resistance $\gamma(x)$. One can then easily conclude that the stochastically stable states are the states minimizing

γ. When the process is almost regular, we can only show that p^ε_x is almost regular with resistance $\gamma(x)$: for all $x \in \mathcal{X}$,

$$\begin{cases} \lim_{\varepsilon\to 0} \varepsilon^{-\gamma'} P^\varepsilon_{x,y} = 0 & \text{if } \gamma' < \gamma(x), \\ \lim_{\varepsilon\to 0} \varepsilon^{-\gamma'} P^\varepsilon_{x,y} = +\infty & \text{if } \gamma' > \gamma(x). \end{cases} \tag{23}$$

Indeed, let $x \in \mathcal{X}$ and $\gamma' < \gamma(x)$. Let T be any x-tree. Since $\gamma' < \gamma(x) \leq r(T) = \sum_{(y\to z)\in T} r(y \to z)$, one can find a constant $\alpha > 0$ such that $\gamma' = \sum_{(y\to z)\in T}(r(y \to z) - \alpha)$. Then,

$$\varepsilon^{-\gamma'} \prod_{(y\to z)\in T} P^\varepsilon_{y,z} = \prod_{(y\to z)\in T} \varepsilon^{-(r(y\to z)-\alpha)} P^\varepsilon_{y,z}.$$

By (3) with $r' = r(y \to z) - \alpha < r(y \to z)$, each term in the last product goes to 0 as $\varepsilon \to 0$. Summing over all x-tree, of which there is a finite amount, we conclude that $\lim_{\varepsilon\to 0} \varepsilon^{-\gamma'} p^\varepsilon_x = 0$.

Let $\gamma' > \gamma(x)$ and T be an optimal x-tree. Since $\gamma' > \gamma(x) = r(T) = \sum_{(y\to z)\in T} r(y \to z)$, one can find a constant $\alpha > 0$ such that $\gamma' = \sum_{(y\to z)\in T}(r(y \to z) + \alpha)$. Then,

$$\varepsilon^{-\gamma'} p^\varepsilon_x \geq \varepsilon^{-\gamma'} \prod_{(y\to z)\in T} P^\varepsilon_{y,z} = \prod_{(y\to z)\in T} \varepsilon^{-(r(y\to z)+\alpha)} P^\varepsilon_{y,z}.$$

By (3) with $r' = r(y \to z) + \alpha > r(y \to z)$, each term in the last product goes to $+\infty$ as $\varepsilon \to 0$. We conclude that $\lim_{\varepsilon\to 0} \varepsilon^{-\gamma'} p^\varepsilon_x = +\infty$.

We can now conclude by rewriting (22) as

$$\pi^\varepsilon_x = \frac{\varepsilon^{-\gamma'} p^\varepsilon_x}{\sum_{y\in\mathcal{X}^\star} \varepsilon^{-\gamma'} p^\varepsilon_y + \sum_{y\notin\mathcal{X}^\star} \varepsilon^{-\gamma'} p^\varepsilon_y}.$$

If $x \notin \mathcal{X}^\star$, choosing γ' so that $\gamma^\star < \gamma' < \gamma(x)$ makes the numerator go to 0 and the denominator go to $+\infty$. We conclude that $\lim_{\varepsilon\to 0} \pi^\varepsilon_x = 0$ for all $x \notin \mathcal{X}^\star$, and therefore that $\lim_{\varepsilon\to 0} \pi^\varepsilon_{\mathcal{X}^\star} = 1$. □

The same result as for Theorem 2.1 holds regarding the computation of γ over recurrence classes of P^0. Indeed, as we have previously discussed, this result only concerns the resistance graph $\mathcal{G}$ and relies on graph theory which is independent from our work to define the resistance graph.

4.4 Theoretical Analysis of RITEL

Now that we have introduced the tools to study ARPMPs, we can discuss the convergence properties of RITEL. This section follows the same reasoning as Section 3.4. We call state a triplet $(\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) = (\overline{m}_i, \overline{a}_i, \overline{v}_i)_{i\in I}$ describing the states of all players at a given time in the algorithm, and $\mathcal{X}$ the set of all possible states. As for ITEL, we remove the benchmark action and bin of discontent players, and identify the set of all-discontent states to a single state D. Recall that our goal is to identify $\mathcal{X}^\star \subset \mathcal{X}$ the subset of states that minimize the potential γ.

4.4.1 Definitions

The definitions given in Section 3.4 need to be extended in order to account for the loss of accuracy due to the introduction of noise and bins. The first notion to adapt is that of aligned states given in Definition 3.1. Since benchmark utilities are quantified into bins, one would intuitively define aligned states as states where for each players its expected utility is within its benchmark bin. In fact the definition needs to be a bit looser. Indeed, recall that by design of RITEL, observing an offset of one bin in utility is not sufficient to behave as if a change happened, so that a player which expected utility is offset from its benchmark bin by one bin will tend to keep this benchmark. For this reason we introduce two notions of alignment.

Definition 4.3 (Aligned States)**.**

- A player of benchmark bin $\overline{v}$ is *strongly aligned* with an action profile $\boldsymbol{a}$ if its average utility μ in $\boldsymbol{a}$ satisfies $\mu \in \overline{v}$.
- A player of benchmark bin $\overline{v}$ is *weakly aligned* with an action profile $\boldsymbol{a}$ if its average utility μ in $\boldsymbol{a}$ satisfies $\mu \in \overline{v} \pm \delta$. If $U(\boldsymbol{a})$ is not deterministic, we also impose that $\mu \neq (\overline{v} - \delta)^-$.

An all-content state is strongly aligned (resp. weakly aligned) if all players are strongly aligned (resp. weakly aligned) with the benchmark actions.

In general, a player is strongly aligned if its expected utility is contained in its benchmark bin, and weakly aligned if it is contained in its benchmark bin or in an adjacent bin. However, when the average utility μ lies exactly at the boundary between two bins ν and $\nu - \delta$ and the utility is not deterministic, the bin $\nu + \delta$ is not weakly aligned. This is due to the fact that as $\varepsilon \to 0$ – and thus $\tau \to +\infty$ – $U^{(\tau)}$ will tend to be observed in both ν and $\nu - \delta$ with equal probability. Note that in any case, given an action profile $\boldsymbol{a}$, there is exactly one strongly aligned all-content state with benchmark actions $\boldsymbol{a}$.

Recall that in ITEL, we defined subsets $\mathcal{A} \subset \mathcal{E} \subset \mathcal{C} \subset \mathcal{X}$ of the state space $\mathcal{X}$, namely $\mathcal{C}$ the subset of aligned all-content states, $\mathcal{E}$ the subset of SE states, and $\mathcal{A}$ the subset of admissible states. With each inclusion comes a stronger sense of stability: states in $\mathcal{C}$ are absorbing in the unperturbed process P^0, states in $\mathcal{E}$ have a strong outward resistance in the perturbed process P^ε, and states in $\mathcal{A}$ are absorbing in P^ε.

In RITEL, we can define similar sets with similar properties. However, and as for the previous definitions, each of them needs to come in both a "weak" and "strong" version. Intuitively, the "strong" set contains the states that are ensured to hold the associated property, whereas the weak set contains the states that may hold the property. For example, regarding alignment, we define $\mathcal{C}_\delta \subset \mathcal{X}$ the set of weakly aligned all-content states and $\mathcal{C} \subset \mathcal{C}_\delta$ the set of strongly aligned all-content states. Although states in $\mathcal{C}_\delta$ will be ensured to be absorbing in P^0, their outward resistance may be arbitrarily close to 0, whereas the outward resistance of states in $\mathcal{C}$ will be lower bounded by some quantity independent from the noise. Similar properties will be observed with the updated notions of SE states $\mathcal{E}_\delta$ and $\mathcal{E}$ and admissible states $\mathcal{A}_\delta$ and $\mathcal{A}$ that shall be defined later. The subscript δ identifies the weak version of the set. Eventually, we will be able to derive upper bounds on the potential of states that are part of the strong set, and lower bound on the potential of the states that are not part of the weak set. Finally, the weak set acts as a gray area that cannot be described as accurately.

Now, we must revise the interdependence assumption A1 into something stronger. Indeed, a player who observes a small change in utility ignores it. Due to the bin quantification, a "small" change can go up to a difference of 2δ if one utility is at the bottom of a bin v and the other at the top of $v + \delta$. Moreover, considering weakly aligned states, the benchmark can be offset by an additional δ. For this reason, we need to assume that interdependence can always happen with a change of at least 3δ in utility.

Assumption A2 (3δ-Interdependence). We consider games that are 3δ-*interdependent*, i.e., given any action profile $\boldsymbol{a}$, for any proper subset $\emptyset \subsetneq J \subsetneq I$ of players, there exists a player $i \in J$, an action $a_i' \neq a_i$, and a player $j \notin J$, such that $|M_j(a_i', a_{-i}) - M_j(a)| \geq 3\delta$.

Remark. Note that since there is a finite number of players and action profiles, a game that satisfies interdependence A1 always satisfies 3δ-interdependence A2 for small enough δ.

We can now extend the notion of SE given in Definition 3.2. Indeed, for a given action profile $\boldsymbol{a}$ there are now several weakly aligned states with benchmark $\overline{\boldsymbol{a}} = \boldsymbol{a}$ but with different benchmark bins. The stability of these states varies with their benchmark $\overline{\boldsymbol{v}}$, so the SE property must take both $\overline{\boldsymbol{a}}$ and $\overline{\boldsymbol{v}}$ into account. Whether of not a state is a SE shall reflects how stable it is in RITEL. Eventually, we are also interested in making a convergence statement in terms of action profiles and their actual expected utilities instead of benchmark bins. For this reason, the notion of SE is also defined for action profiles, and reflects the intrinsic equilibrium properties of $\boldsymbol{a}$ instead of its behavior in the RITEL algorithm. Moreover, the definition needs to account for δ to be able to quantify the inaccuracy that will arise due to bin quantification and weak alignment.

Definition 4.4 ((δ_1, δ_2)-Stable Equilibrium Action Profile)**.** An action profile $\boldsymbol{a}$ with resulting average utilities $\boldsymbol{\mu}$ is a (δ_1, δ_2)-SE if for all player i one of the following two holds:

- i is δ_1-admissible, i.e., $\mu_i + \delta_1 \in A$.
- i is at a δ_2-equilibrium position, i.e., for any action $a'_i \neq a_i$, $M_i(a'_i, a_{-i}) \leq \mu_i + \delta_2$.

Definition 4.5 ((δ_1, δ_2)-Stable Equilibrium State)**.** A weakly aligned all-content state of actions $\overline{\boldsymbol{a}}$ and utility bins $\overline{\boldsymbol{v}}$ is a (δ_1, δ_2)-SE if for all player i one of the following two holds:

- i is δ_1-admissible, i.e., $\overline{v}_i + \delta_1 \in A$.
- i is at a δ_2-equilibrium position, i.e., for any action $a_i \neq \overline{a}_i$, $N_i(a_i, \overline{a}_{-i}) \leq \overline{v}_i + \delta_2$.

An action profile is at a δ-equilibrium if no player can explore such that it would improve its expected utility by at least δ. A state is at a δ-equilibrium if no player can explore in a way so that it would observe an expected utility not aligned with its benchmark.

Notice that a $(0, 0)$-SE action profile is exactly the same as a SE action profile as defined in Definition 3.2. Likewise we say that a state is a SE if it is a $(0, 0)$-SE. Similarly, we say that an action profile or state is δ-admissible if all players are δ-admissible, and that it is admissible if all players are 0-admissible. We denote $\mathcal{E}_\delta \subset \mathcal{C}_\delta$ the set of weakly aligned $(0, \delta)$-SE states, $\mathcal{E} \subset \mathcal{E}_\delta \cap \mathcal{C}$ the set of strongly aligned SE states, and $\mathcal{A}_\delta \subset \mathcal{E}_\delta$ the set of weakly aligned admissible states. A "strong" set $\mathcal{A}$ could be defined as the set $\mathcal{A}_\delta \cap \mathcal{C}$ of strongly aligned admissible states, however we will see later that another definition of $\mathcal{A}$ is more suitable.

However closely related, the definitions of SE do not necessarily coincide in both contexts, that is, a (δ_1, δ_2)-SE state does not necessarily imply a (δ_1, δ_2)-SE action profile. In fact, Definitions 4.4 and 4.5 are designed so that a SE action profile translates well into a SE state, whereas a SE state translates into a weaker sense of SE action profile. The following lemma states the correspondences that will be required to convert our convergence result in terms of states and benchmark bins to a convergence result in terms of action profiles and expected utilities.

Lemma 4.7.

(i) If $\boldsymbol{a}$ is a SE action profile, then the corresponding strongly aligned all-content state x is in $\mathcal{E}$.

(ii) If $x \in \mathcal{A}_\delta$, then its benchmark action profile $\overline{\boldsymbol{a}}$ is 2δ-admissible.

(iii) If $x \in \mathcal{E}_\delta$, then its benchmark action profile $\overline{\boldsymbol{a}}$ is a $(2\delta, 3\delta)$-SE.

Proof.
(i). Let $\boldsymbol{a}$ be a SE action profile of expected utilities $\boldsymbol{\mu}$ with bins $\boldsymbol{\nu}$, and $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{C}$ with $\overline{\boldsymbol{a}} = \boldsymbol{a}$ and $\overline{\boldsymbol{v}} = \boldsymbol{\nu}$. Let i be some player. If i is admissible in $\boldsymbol{a}$, then since $\mu_i \in \nu_i = \overline{v}_i$, $\overline{v}_i \cap A \neq \emptyset$, hence $\overline{v}_i \in A$ and i is admissible in x. Else, if i is at an equilibrium position in $\boldsymbol{a}$, then for any action $a'_i \neq a_i = \overline{a}_i$, $M_i(a'_i, a_{-i}) \leq \mu_i$. It follows that $N_i(a'_i, \overline{a}_{-i}) = B(M_i(a'_i, a_{-i})) \leq B(\mu_i) = \overline{v}_i$, hence i is at an equilibrium in x. Therefore x is a SE state: $x \in \mathcal{E}$.

(ii). Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{A}_\delta$ and denote $\boldsymbol{\mu}$ the expected utilities of $\overline{\boldsymbol{a}}$, which then satisfies $\mu_i \geq \overline{v}_i^-$ for all i. Let i be some player. i is admissible in x, hence $\overline{v}_i \cap A \neq \emptyset$ which implies that utilities higher than $\overline{v}_i^+$ are within A, as A is of the form $[u_0, 1]$ or $(u_0, 1]$. In particular $\mu_i + 2\delta \geq \overline{v}_i^- + \delta = \overline{v}_i^+$ is in A, hence i is 2δ-admissible in $\overline{\boldsymbol{a}}$. Therefore $\overline{\boldsymbol{a}}$ is 2δ-admissible.

(iii). Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{E}_\delta$ and denote $\boldsymbol{\mu}$ the expected utilities of $\overline{\boldsymbol{a}}$, which then satisfies $\mu_i \geq \overline{v}_i^- - \delta$ for all i. Let i be some player. If i is admissible in x, we have shown in (ii) that it is 2δ-admissible in $\overline{\boldsymbol{a}}$. Else, if i is at a δ-equilibrium position in x, then for any action $a'_i \neq \overline{a}_i$, $M_i(a'_i, \overline{a}_{-i}) \leq N_i(a'_i, \overline{a}_{-i})^- + \delta \leq \overline{v}_i^- + 2\delta \leq \mu_i + 3\delta$, hence i is at a 3δ-equilibrium position in $\overline{\boldsymbol{a}}$. Therefore $\overline{\boldsymbol{a}}$ is a $(2\delta, 3\delta)$-SE. □

The last definitions to update are Definitions 3.3 and 3.4. As for Definitions 4.4 and 4.5, virtual welfare and stability shall be used both regarding states and regarding action profiles. Their definitions are thus given in both contexts. True welfare and stability will however be used only to state the final result in terms of action profiles, in the special case of affine functions F and G. Therefore we define them only in the context of action profiles.

Definition 4.6 (Welfare and Stability of an Action Profile)**.** Let $\boldsymbol{a}$ be an action profile with expected utilities $\boldsymbol{\mu}$. We define:

- $W(\boldsymbol{a}) \triangleq \frac{1}{n}\sum_i \mu_i$,
- $S(\boldsymbol{a}) \triangleq \max\{M_i(a'_i, a_{-i})) - \mu_i \quad | \quad i \text{ non-admissible}, a'_i \neq a_i : M_i(a'_i, a_{-i}) > \mu_i\}$.

$S(\boldsymbol{a})$ is defined only if $\boldsymbol{a}$ is not a SE.

Definition 4.7 (Virtual Welfare and Stability of an Action Profile)**.** Let $\boldsymbol{a}$ be an action profile with expected utilities $\boldsymbol{\mu}$. We define:

- $\tilde{W}(\boldsymbol{a}) \triangleq 1 - \sum_i F(\mu_i)$,
- $\tilde{S}(\boldsymbol{a}) \triangleq \min\{G(\mu_i, M_i(a'_i, a_{-i})) \quad | \quad i \text{ non-admissible}, a'_i \neq \overline{a}_i : M_i(a'_i, a_{-i}) > \mu_i\}$.

$\tilde{S}(\boldsymbol{a})$ is defined only if $\boldsymbol{a}$ is not a SE. We override notations by denoting

- $\tilde{W}(\boldsymbol{a} \pm \delta) \triangleq 1 - \sum_i F(\mu_i \pm \delta)$,
- $\tilde{S}(\boldsymbol{a} \pm \delta) \triangleq \min\{G(\mu_i \pm \delta, M_i(a'_i, a_{-i})) \quad | \quad i \text{ non-admissible}, a'_i \neq a_i : M_i(a'_i, a_{-i}) > \mu_i \pm \delta\}$.

Quantification implies an approximation error of order δ when estimating $\tilde{W}(\overline{\boldsymbol{a}})$ from the virtual welfare of state with benchmark actions $\overline{\boldsymbol{a}}$. This is the reason why we introduce $\tilde{W}(\boldsymbol{a} \pm \delta)$ and $\tilde{S}(\boldsymbol{a} \pm \delta)$. As for ITEL, when F and G are chosen as affine function of u and $u - \overline{u}$ respectively, $\tilde{W}(\boldsymbol{a})$ is an affine function of $W(\boldsymbol{a})$ and $\tilde{S}$ is an affine function of $S(\boldsymbol{a})$. This will eventually allow us to translate nicely a relative comparison between virtual welfare of states to a comparison between true welfare of their benchmark actions.

Before stating the definition of $\tilde{W}$ and $\tilde{S}$ in the context of states, note that the randomness of the observation makes the analysis of acceptations from a content state more complex. Indeed, a player of benchmark $\overline{v}$ can accept the expected bin ν of its exploration for a resistance equal to $G(\overline{v}, \nu)$. It can also accept an higher bin v for a resistance equal to $r_U(v) + G(\overline{v}, v)$ where U is the distribution of the observed utility. This resistance may be lower than that of accepting ν if the non-increasing nature of $G(\overline{v}, \cdot)$ compensate for the addition of $r_U(v)$. Eventually, the two quantities $\tilde{S}_+$ and $\tilde{S}_-$ we introduce will serve the respective roles of upper and lower bounds on the minimal resistance of accepting an exploration, assuming the state is not an equilibria.

Definition 4.8 (Virtual Welfare and Stability of a State)**.** Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{C}_\delta$. We define:

- $\tilde{W}(x) \triangleq 1 - \sum_i F(\overline{v}_i)$,
- $\tilde{S}_+(x) \triangleq \min\{G(\overline{v}_i, N_i(a_i, \overline{a}_{-i})) \quad | \quad i \text{ non-admissible}, a_i \neq \overline{a}_i : N_i(a_i, \overline{a}_{-i}) \geq \overline{v}_i + 2\delta\}$,
- $\tilde{S}_-(x) \triangleq \min\{G(\overline{v}_i, N_i(a_i, \overline{a}_{-i}) + \delta) \quad | \quad i \text{ non-admissible}, a_i \neq \overline{a}_i : N_i(a_i, \overline{a}_{-i}) \geq \overline{v}_i + \delta\}$.

$\tilde{S}_+(x)$ is defined only if x is not a $(0, \delta)$-SE, $\tilde{S}_-(x)$ is defined only if x is not a SE.

As for Definitions 4.4 and 4.5, we will eventually need to estimate the welfare and stability corresponding to the expected utilities of a state. The following lemma states the correspondences that will be required in order to translate our convergence result in terms of states to a result in terms of action profiles.

Lemma 4.8. Let x be a state with benchmark actions $\overline{\boldsymbol{a}}$.

(i) If $x \in \mathcal{C}_\delta$, $\tilde{W}(x) \leq \tilde{W}(\overline{\boldsymbol{a}} + \delta)$.

(ii) If $x \in \mathcal{C}$, $\tilde{W}(x) \geq \tilde{W}(\overline{\boldsymbol{a}} - \delta)$.

(iii) If $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$, $\tilde{S}_+(x) \leq \tilde{S}(\overline{\boldsymbol{a}} + \delta)$.

(iv) If $x \in \mathcal{C} \setminus \mathcal{E}$, $\tilde{S}_-(x) \geq \tilde{S}(\overline{\boldsymbol{a}} - \delta)$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{C}_\delta$ and denote $\boldsymbol{\mu}$ the expected utilities of $\overline{\boldsymbol{a}}$. Recall that F and G are defined over bins by replacing the bins with their lower edge: $F(v) = F(v^-)$ and $G(\overline{v}, v) = G(\overline{v}^-, v)$.

(i) If x is weakly aligned, we have $\mu_i \geq \overline{v}_i - \delta$ for all players i. F is non-increasing, hence

$$\tilde{W}(x) = 1 - \sum_i F(\overline{v}_i^-) \leq 1 - \sum_i F(\mu_i + \delta) = \tilde{W}(\overline{\boldsymbol{a}} + \delta).$$

(ii) If x is strongly aligned, we have $\mu_i \leq \overline{v}_i^- + \delta$ for all players i. F is non-increasing, hence

$$\tilde{W}(x) = 1 - \sum_i F(\overline{v}_i^-) \geq 1 - \sum_i F(\mu_i - \delta) = \tilde{W}(\overline{\boldsymbol{a}} - \delta).$$

(iii) Denote $\overline{v}_i$ and $v_i = N_i(a_i, \overline{a}_{-i})$ the bins such that $\tilde{S}_+(x) = G(\overline{v}_i, v_i)$. x is weakly aligned so we have $\mu_i \geq \overline{v}_i^- - \delta$. G is non-decreasing in the first argument and non-increasing in the second, hence

$$\tilde{S}_+(x) = G(\overline{v}_i^-, v_i^-) \leq G(\mu_i + \delta, v_i^-) \leq G(\mu_i + \delta, M_i(a_i, \overline{a}_{-i})) = \tilde{S}(\overline{\boldsymbol{a}} + \delta).$$

(iv) Denote $\overline{v}_i$ and $v_i = N_i(a_i, \overline{a}_{-i})$ the bins such that $\tilde{S}_-(x) = G(\overline{v}_i, v_i + \delta)$. x is strongly aligned so we have $\mu_i \leq \overline{v}_i^- + \delta$. G is non-decreasing in the first argument and non-increasing in the second, hence

$$\tilde{S}_-(x) = G(\overline{v}_i^-, v_i^- + \delta) \geq G(\mu_i - \delta, v_i^- + \delta) \geq G(\mu_i - \delta, M_i(a_i, \overline{a}_{-i})) = \tilde{S}(\overline{\boldsymbol{a}} - \delta).$$

□

4.4.2 Noise Stability

Before stating our convergence result, let us describe more precisely the resistance of a player's behavior. In ITEL such resistances were derived directly from the policies given in Table 2. Here, Table 4 describes how a player behaves based on its random observation. To be able to derive actual resistances, one would want to rewrite Table 4 replacing the condition on the observed utility bin by its expectation. In fact, the resistance of choosing some behavior becomes the sum of the resistance of observing a bin allowing this behavior and of the resistance of choosing the behavior given the bin.

The second resistance is essentially the same as in ITEL. The first can be estimated using Corollary 4.3, which introduces the resistance $r_U(v)$ associated with U falling in bin v. In particular, observing the expected bin happens with no resistance. The probability of observing other bins is lower bounded using (15). In particular, a lower bound on the resistance of offset observations will be needed in order to argue that they play a negligible role in the RITEL process. From now on, we consider σ^2 an upper bound on the variance of all utility distributions involved in the game. Since utilities are bounded within $[0, 1]$, one can always choose $\sigma^2 \leq \frac{1}{4}$.

Lemma 4.9. Let U be a random utility of expected bin ν. Then for all $v \neq \nu \pm \delta$,

$$r_U(v) \geq R_0 \triangleq \tau_0 \delta^2 \max\left(\frac{1}{2\sigma^2 + 2\delta}, 2\right). \tag{24}$$

Proof. Let U be a random utility, μ its expected utility and $\nu = B(\mu)$ its expected bin. According to Corollary 4.3, the resistance associated with observing bin $v \geq \overline{v} + 2\delta$ is

$$r_U(v) = \tau_0 \Lambda_U^*(v^-) \geq \tau_0 \Lambda_U^*(\mu + \delta),$$

as Λ_U^* is increasing over $[\mu, +\infty)$ and $\mu + \delta \leq \overline{v}^+ + \delta = \overline{v}^- + 2\delta \leq v^-$. Plugging (15), we get

$$r_U(v) \geq \tau_0 \max\left(2(\mu + \delta - \mu)^2, \frac{(\mu + \delta - \mu)^2}{2\sigma^2 + 2|\mu + \delta - \mu|}\right) = \tau_0 \max\left(2\delta^2, \frac{\delta^2}{2\sigma^2 + 2\delta}\right) = R_0.$$

The same reasoning holds when $v \leq \overline{v} - 2\delta$: $r_U(v) \geq \tau_0 \Lambda_U^*(\mu - \delta) \geq R_0$. □

We can now define the noise stability of a state, that is the resistance needed to make an observation that is not aligned with the benchmark without the influence of other players playing different actions. Intuitively, this quantity is expected to be high when the state is aligned and low when it is not. In particular the noise stability of strongly aligned states will be lower bounded using Lemma 4.9.

Definition 4.9 (Noise Stability)**.** Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{X}$. We define the *noise stability* of x as

$$\tilde{R}(x) \triangleq \min\{r_{U_i(\overline{\boldsymbol{a}})}(v_i) \quad | \quad i \text{ player}, v_i \neq \overline{v}_i \pm \delta\}. \tag{25}$$

Recall that the minimum is considered over bins $v_i \neq \overline{v}_i \pm \delta$ as observing these bins implies a change of behavior according to Table 4b. In order to describe easy edges and outward resistance of states, estimations of $\tilde{R}$ are needed. The following lemma states the bounds that will be crucial for our proof.

Lemma 4.10. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{X}$.

(i) If $x \notin \mathcal{C}_\delta$, $\tilde{R}(x) = 0$.

(ii) If $x \in \mathcal{C}_\delta$, $\tilde{R}(x) > 0$.

(iii) If $x \in \mathcal{C}$, $\tilde{R}(x) \geq R_0$.

Proof.
(i). If $x \notin \mathcal{C}_\delta$, let i be a non-aligned player in x. That is, the expected utility μ_i of bin ν_i satisfies $\nu_i \neq \overline{v}_i \pm \delta$. According to Corollary 4.3, $r_{U_i(\overline{\boldsymbol{a}})}(\nu_i) = 0$, hence $\tilde{R}(x) = 0$. If $\mu_i = \nu_i^-$ and $U_i(\overline{\boldsymbol{a}})$ is not deterministic, it is actually possible that $\nu_i = \overline{v}_i - \delta$ (see Definition 4.3). In this case, we have $\nu_i - \delta = \overline{v}_i - 2\delta \neq \overline{v}_i \pm \delta$ and Corollary 4.3 states that $r_{U_i(\overline{\boldsymbol{a}})}(\nu_i - \delta) = 0$ as $\mu_i = (\nu_i - \delta)^+$, hence $\tilde{R}(x) = 0$.

(ii). If $x \in \mathcal{C}_\delta$, then $\nu_i = \overline{v}_i \pm \delta$ for all players i. Then, $v_i \neq \overline{v}_i \pm \delta$ implies $v_i \neq \nu_i$, hence $r_{U_i(\overline{\boldsymbol{a}})}(v_i) > 0$ according to Corollary 4.3. Therefore $\tilde{R}(x) > 0$. Actually it would be possible that $r_{U_i(\overline{\boldsymbol{a}})}(v_i) = 0$ if $v_i = \nu_i - \delta$, $\mu_i = \nu_i^- = v_i^+$, and $U_i(\overline{\boldsymbol{a}})$ is not deterministic, however this corresponds to the excluded case in Definition 4.3.

(iii). If $x \in \mathcal{C}$, then $\nu_i = \overline{v}_i$ for all players i. The fact that $\tilde{R}(x) \geq R_0$ is then directly implied from Lemma 4.9. □

Lemma 4.10 tends to indicate that the recurrence classes of RITEL will be the weakly aligned states. Indeed, putting together this lemma with the reasoning that was done in ITEL, we are able to show the following result similar to Proposition 3.1. Proposition 4.11 is proven along the proof of Theorem 4.13 in Appendix C.

Proposition 4.11. The recurrence classes of the unperturbed process P^0 are the singletons $\{x\}$ for each weakly aligned all-content states $x \in \mathcal{C}$, and possibly the communication class of the all-discontent state D.

Notice that Lemma 4.10 allows to lower bound the outward resistance of strongly aligned states using R_0. R_0 can be seen as an indicator of how utilities sampled by RITEL are reliable. Assuming R_0 is large enough, offset observations will play a negligible role in leaving such states compared to the other paths that were already described in ITEL. In fact we will see that in order to obtain our convergence result, we need to ensure that $R_0 \geq 1$, hence we introduce a new condition C4. As we mentioned before, $\mathcal{C}_\delta$ acts as a gray area: the outward resistance of weakly aligned states may be arbitrarily close to 0 ; we only know that it is positive.

Condition C4 (Noise control).

$$R_0 = \tau_0 \delta^2 \max\left(\frac{1}{2\sigma^2 + 2\delta}, 2\right) \geq 1. \tag{C4}$$

Remark. In general, the key condition one must satisfy is that $\tau_0 \Lambda_U^*(\mu + \delta) > 1$ for all utility distributions U of expectation μ. This is ensured by C4 thanks to Lemmas 4.4 and 4.5. However, if one has more precise knowledge of the distributions involved, a better bound could be derived. Again, this discussion depends on the practical application and is totally independent from our work. The above condition is sufficient for our reasoning and provides an easy way to integrate knowledge of variance.

Let us now discuss the role of admissible states. As in ITEL, in an admissible state $x \in \mathcal{A}_\delta$ no player can explore. However, players may make offset observations, so it is not true that all states in $\mathcal{A}_\delta$ will be absorbing even in P^ε, regardless of whether or not they are strongly aligned. In fact, $x \in \mathcal{A}_\delta$ will be absorbing only if $\tilde{R}(x) = +\infty$. For this reason, we choose to define $\mathcal{A} \subset \mathcal{A}_\delta$ the set of admissible states with infinite noise stability $\tilde{R}$. Although strongly aligned states are more likely to verify this property as suggested by Lemma 4.10, weakly aligned states may verify it too, for example when players' observations are bounded a.s. in two consecutive bins, hence they cannot leave the weakly aligned one with an offset observation. As for Proposition 4.11, Proposition 4.12 is proven along the proof of Theorem 4.13 in Appendix C.

Proposition 4.12.

- If $\mathcal{A} \neq \emptyset$, states $x \in \mathcal{A}$ are absorbing for the perturbed process P^ε and the recurrence classes of P^ε are exactly the corresponding singletons $\{x\}$.
- If $\mathcal{A} = \emptyset$, P^ε is aperiodic and irreducible.

As for ITEL, the first case is then treated by common Markov chain theory, whereas the second can be studied using the ARPMP theory.

4.4.3 Convergence of RITEL

We are finally able to state our convergence result.

Theorem 4.13. Assume that A2, C2, and C4 hold and that $\mathcal{A} = \emptyset$. Then the RITEL process is a ARPMP and the states of minimum potential verify:

- If $\mathcal{E} \neq \emptyset$, $\mathcal{X}^\star \subset \mathcal{A}_\delta \cup \{x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta : \tilde{W}(x) \geq \tilde{W}^\star\}$ where $\tilde{W}^\star = \max_{x \in \mathcal{E}} \tilde{W}(x)$.
- If $\mathcal{E} = \emptyset$, $\mathcal{X}^\star \subset \mathcal{E}_\delta \cup \{x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta : \tilde{S}_+(x) + \tilde{W}(x) \geq \tilde{S}_-^\star + \tilde{W}^\star\}$ where $\tilde{S}_-^\star + \tilde{W}^\star = \max_{x \in \mathcal{C}} \tilde{S}_-(x) + \tilde{W}(x)$.

Theorem 4.13 is proven in Appendix C. Combining Proposition 4.12 and Theorems 4.6 and 4.13 and using Lemmas 4.7 and 4.8, we can translate the above result to a result in terms of action profiles.

Theorem 4.14 (RITEL convergence)**.** Assume that A2, C2, and C4 hold.

- If there exist SE action profiles, the RITEL process spends most of the time as $\varepsilon \to 0$ in action profiles $\boldsymbol{a}$ that are 2δ-admissible or $(2\delta, 3\delta)$-SE with $\tilde{W}(\boldsymbol{a} + \delta) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$, where $\boldsymbol{a}^\star$ maximizes $\tilde{W}$ over SE action profiles.

- Else, the RITEL process spends most of the time as $\varepsilon \to 0$ in action profiles $\boldsymbol{a}$ that are $(2\delta, 3\delta)$-SE or with $\tilde{S}(\overline{\boldsymbol{a}}+\delta)+\tilde{W}(\overline{\boldsymbol{a}}+\delta) \geq \tilde{S}(\boldsymbol{a}^\star-\delta)+\tilde{W}(\boldsymbol{a}^\star-\delta)$, where $\boldsymbol{a}^\star$ maximizes $\tilde{S}+\tilde{W}$ over all action profiles.

By "spending most of the time as $\varepsilon \to 0$", we mean that either the process has absorbing states which benchmark actions are among the ones described, or that it is aperiodic and irreducible and its stationary distribution vanishes at states with benchmark actions different from the ones described.

Notice that since the game is finite, the set of all expected utilities involved is discrete. In particular, assuming δ is chosen small enough, 3δ-interdependence implies interdependence, δ-admissibility implies admissibility, δ-equilibrium implies equilibrium, $\tilde{W}(\boldsymbol{a}+\delta) \geq \tilde{W}(\boldsymbol{a}^\star-\delta)$ implies $\tilde{W}(\boldsymbol{a}) \geq \tilde{W}(\boldsymbol{a}^\star)$, and finally $\tilde{S}(\boldsymbol{a}+\delta) \geq \tilde{S}(\boldsymbol{a}^\star-\delta)$ implies $\tilde{S}(\boldsymbol{a}) \geq \tilde{S}(\boldsymbol{a}^\star)$. In this case, one can therefore replace all occurrences of δ with 0 in Theorem 4.14. In other words, RITEL is capable of identifying optimal action profiles provided δ is small enough. This is a purely theoretical consideration, as choosing such δ would imply a choice of τ_0 that would be deemed unreasonable in most applications in order to satisfy C4.

Theorem 4.14 is hard to interpret in general. In the case where F and G are chosen of the form $F : u \mapsto \phi_F - \psi_F \cdot u$ and $G : (\overline{u}, u) \mapsto \phi_G - \psi_G \cdot (u - \overline{u})$, we get:

- If there exist SE action profiles, the RITEL process spends most of the time as $\varepsilon \to 0$ in action profiles $\boldsymbol{a}$ that are 2δ-admissible or $(2\delta, 3\delta)$-SE with $W(\boldsymbol{a}) \geq W^\star - 2\delta$ where $W^\star$ is the maximal welfare over SE action profiles.
- Else, the RITEL process spends most of the time as $\varepsilon \to 0$ in action profiles $\boldsymbol{a}$ that are $(2\delta, 3\delta)$-SE or with $\psi_F W(\boldsymbol{a}) - \psi_G S(\boldsymbol{a}) \geq \psi_F W^\star - \psi_G S^\star - 2\delta$ where $W^\star$ and $S^\star$ are defined as the welfare and stability of an action profile maximizing $\psi_F W - \psi_G S$.

The reasoning required to deduce Theorem 4.14 from Theorem 4.13 along with the special case of affine functions is detailed at the end of Appendix C.

5 Dynamic Noise Schedule

The convergence result stated by Theorem 2.1 can be summarized as follows: if $(X_k^\varepsilon)_{k\geq 0}$ is a Markov process controlled by a RPMP P^ε, then

$$\lim_{\varepsilon \to 0} \lim_{k \to +\infty} \mathbb{P}[X_k^\varepsilon \in \mathcal{X}^\star] = 1. \tag{26}$$

In practice this means that for a fixed $\varepsilon > 0$ sufficiently small, the process X^ε will spend most of the time in optimal states $x \in \mathcal{C}^\star$. One would be interested in decreasing the noise parameter ε as the algorithm is running. A stronger result would be

$$\lim_{k \to +\infty} \mathbb{P}[X_k \in \mathcal{X}^\star] = 1, \tag{27}$$

where $(X_k)_{k\geq 0}$ follows an inhomogeneous Markov chain $(P^{\varepsilon_k})_{k\geq 0}$ for some well chosen sequence $(\varepsilon_k)_{k\geq 0}$. In fact, this kind of process can be interpreted as a form of generalized simulated annealing (GSA), which have been studied thoroughly in the literature [7].

In this section we apply results taken from [7] to improve Theorem 2.1. We will see that (27) holds assuming $(\varepsilon_k)_{k\geq 0}$ is decreasing slowly enough. Moreover, given a finite horizon H, one can compute a "cooling" schedule $(\varepsilon_k)_{0\leq k\leq H}$ with optimal convergence speed.

First let us discuss how both notions of RPMP and GSA coincide. GSA as defined in [7, Definitions 1.1-1.3] can be reformulated with our notations as follows:

Definition 5.1 (Generalized Simulated Annealing)**.** Let q be an aperiodic irreducible[8] Markov kernel over the state space $\mathcal{X}$, $\kappa \in [1, +\infty)$, and $V : \mathcal{X} \times \mathcal{X} \to [0, +\infty]$. A family of homogeneous Markov chains

[8] [7] does not mention aperiodicity, however we suspect it is an oversight as irreducibility alone is not enough to justify the existence of a unique invariant distribution.

$(Q^T)_{0\leq T\leq T_0}$ over $\mathcal{X}$ is *admissible* for (q,κ,V) if it satisfies the following properties:

$$\forall x,y\in\mathcal{X},\quad \begin{cases} Q^T_{x,y} \underset{T\to 0}{\to} Q^0_{x,y}, & \text{[9]} \quad (28a)\\ V(x,y)<+\infty \Leftrightarrow q_{x,y}>0, & (28b)\\ \forall T\in[0,T_0],\ \frac{1}{\kappa}q_{x,y}e^{-V(x,y)/T}\leq Q^T_{x,y}\leq \kappa q_{x,y}e^{-V(x,y)/T}. & (28c)\end{cases}$$

When Q is admissible, a *generalized simulated annealing process (GSA)* is an inhomogeneous Markov process $(Q^{T_k})_{k\geq 0}$ for some non-negative cooling schedule $(T_k)_{k\geq 0}$.

In GSA, we see that the transition from state x to state y behaves like $e^{-V(x,y)/T}$ where $V\geq 0$ is a cost function and $T>0$ is the temperature of the system. The parallel with PMPs is done by setting $\varepsilon=e^{-1/T}$, $P^\varepsilon=Q^T$ and $r(x\to y)=V(x,y)$, in which case the previous quantity is exactly $\varepsilon^{r(x\to y)}$. Low temperature T is then equivalent to small noise ε. Using this correspondence, we see that a RPMP as in Definition 2.1 is also a GSA as in Definition 5.1. In fact, both definitions are equivalent except for the fact that the condition (28c) is weaker than the regularity condition (2). Indeed, if $\lim_{\varepsilon\to 0}\varepsilon^{-r(x\to y)}P^\varepsilon_{x,y}=p_{x,y}>0$, then $\frac{1}{\kappa}p_{x,y}\varepsilon^{r(x\to y)}\leq P^\varepsilon_{x,y}\leq \kappa p_{x,y}\varepsilon^{r(x\to y)}$ for some $\kappa\in[1,+\infty)$. However the converse is not true in general.

The next section presents results from [7] translated into the RPMP vocabulary. These results are stronger than Theorem 2.1, so that applying them in the context of ITEL and IODL yields a more powerful statement than Theorem 3.4. Note that while (28c) is weaker than regularity, it is still stronger than almost regularity (3), hence we may not apply the theory of GSA to RITEL. It is possible however that some adaptation could be done to accommodate both frameworks.

5.1 GSA Theoretical Results

In [7], our notion of resistance appears via the cost V. The notion of x-trees also appears as "$\{x\}$-graphs" [7, Definition 1.4], and potential appears as "virtual energy" [7, Definition 1.5].[10]

The results from [7] are proven following a classical reasoning based on a cycle decomposition. This decomposition is done by aggregating states from $\mathcal{X}$ into "cycles" and deriving a new cost function over the graph of cycles, then iterating the process until the graph is reduced to a single state. Without discussing too much detail, a key idea is that two states will be merged into a cycle if there is a path going from one to the other and back using only easy edges. The results are presented using quantities that depend on the cycle decomposition, which can be tedious to analyze in general. This is the case for ITEL, however in IODL we can explicitly describe the involved quantities.

For a given RPMP over a state space $\mathcal{X}$ with potential γ and for $\lambda>0$, define $\mathcal{X}_\lambda=\{x\in\mathcal{X}:\gamma(x)\geq\gamma^\star+\lambda\}$ where $\gamma^\star=\min_{x\in\mathcal{X}}\gamma(x)$. For all cycle Π, denote $\gamma(\Pi)=\min_{x\in\Pi}\gamma(x)$. The results we present below discuss the probability of the process to be in $\mathcal{X}_\lambda$, i.e., to be in a state more than λ-sub-optimal. Note that since there is a finite amount of state, when λ is small enough $\mathcal{X}_\lambda=\mathcal{X}\setminus\mathcal{X}^\star$. Similarly the quantities depending on λ that will be introduced are constant at the neighborhood of $\lambda=0$, so that we can eventually control $\mathbb{P}(X_k\in\mathcal{X}^\star)$ by applying the result to small enough λ.

Theorem 5.1 ([7], Theorem 5.2)**.** Let P^ε be a RPMP and $(\varepsilon_k)_{k\geq 0}$ be a non-increasing cooling schedule. Let $(X_k)_{k\geq 0}$ be an inhomogeneous Markov process following $(P^{\varepsilon_k})_{k\geq 0}$ with any starting distribution. Then for all $\lambda>0$,

$$\lim_{k\to 0}\mathbb{P}(X_k\in\mathcal{X}_\lambda)=0 \quad \text{if and only if} \quad \sum_{k\geq 0}\varepsilon_k^{\Gamma_\lambda}=+\infty, \tag{29}$$

[9] In [7] this property is replaced by $(Q^T)_{T\geq 0}$ being a continuous family, which is stronger. However only the limit as $T\to 0$ is ever needed. In the case of ITEL transition probabilities are continuous functions of ε anyway.

[10] Beware not to confuse virtual energy, denoted W in [7], with welfare in our setup. Stochastically stable states will be states minimizing potential/virtual energy, which implies maximizing welfare.

where

$$\Gamma_\lambda \triangleq \sup\{H_e(\Pi) \quad | \quad \Pi \text{ cycle} : \gamma(\Pi) \geq \gamma^\star + \lambda\} \tag{30}$$

and $H_e(\Pi)$ is called the *exit height* of the cycle Π and depends on the cycle decomposition.

The general definition of H_e is tedious, but can be interpreted as the minimal cost to leave a cycle. In the case of singletons, which are the starting cycles of the cycle construction, $H_e(\{x\}) = r^\star(x)$. Applying Theorem 5.1 with small enough λ yields the following result.

Corollary 5.2. Let P^ε be a RPMP and $(\varepsilon_k)_{k\geq 0}$ be a non-increasing cooling schedule. Let $(X_k)_{k\geq 0}$ be an inhomogeneous Markov process following $(P^{\varepsilon_k})_{k\geq 0}$ with any starting distribution. Then,

$$\lim_{k\to+\infty} \mathbb{P}(X_k \notin \mathcal{X}^\star) = 0 \quad \text{if and only if} \quad \sum_{k\geq 0} \varepsilon_k^{\Gamma_0} = +\infty, \tag{31}$$

where

$$\Gamma_0 \triangleq \sup\{H_e(\Pi) \quad | \quad \Pi \text{ cycle} : \gamma(\Pi) > \gamma^\star\}. \tag{32}$$

A more theoretical result states that there exists a cooling schedule with optimal convergence speed.

Theorem 5.3 ([7], Theorem 6.3)**.** Let P^ε be a RPMP and $\lambda > 0$. There exists a constant $c > 0$ such that for all finite horizon N, there exists a non-increasing cooling schedule $(\varepsilon_k^N)_{0\leq k\leq N}$ such that if $(X_k)_{k\geq 0}$ is an inhomogeneous Markov process following $(P^{\varepsilon_k^N})_{0\leq k\leq N}$ with any starting distribution,

$$\mathbb{P}(X_N \in \mathcal{X}_\lambda) \leq \frac{c}{N^{\hat{\alpha}_\lambda}}, \tag{33}$$

where

$$\hat{\alpha}_\lambda \triangleq \min\left\{\frac{\max(\gamma(\Pi) - \gamma^\star, \lambda)}{H_e(\Pi)} \quad | \quad \Pi \text{ cycle} : \gamma(\Pi) > \gamma^\star\right\}. \tag{34}$$

The above convergence speed becomes optimal when λ is close to 0. Applying Theorem 5.3 with small enough λ yields the following result.

Corollary 5.4. Let P^ε be a RPMP. There exists a constant $c > 0$ such that for all finite horizon N, there exists a non-increasing cooling schedule $(\varepsilon_k(N))_{0\leq k\leq N}$ such that if $(X_k)_{k\geq 0}$ is an inhomogeneous Markov process following $(P^{\varepsilon_k(N)})_{0\leq k\leq N}$ with any starting distribution,

$$\mathbb{P}(X_N \notin \mathcal{X}^\star) \leq \frac{c}{N^{\hat{\alpha}_0}} \tag{35}$$

where

$$\hat{\alpha}_0 \triangleq \min\left\{\frac{\gamma(\Pi) - \gamma^\star,}{H_e(\Pi)} \quad | \quad \Pi \text{ cycle} : \gamma(\Pi) > \gamma^\star\right\}. \tag{36}$$

5.2 Application to ITEL and IODL

The above results apply to all RPMPs, hence in particular to our algorithms ITEL and IODL. They mostly serve a theoretical purpose: Corollary 5.2 implies that the cooling schedule should be of order $\varepsilon_k = k^{-1/\Gamma_0}$ at the fastest. Moreover, the precise definition of $\varepsilon_k(N)$ is given in the proof of [8, Theorem 7.1] on which [7, Theorem 6.3] is based.[11] This definition depends on problem-dependent quantities that make it hard to compute in practice. Finally, the optimal convergence speed given by Corollary 5.4 is of order $\frac{c}{N^{\hat{\alpha}_0}}$. The

[11] Recall that $\varepsilon_k(N) = \exp(1/T_k(N))$.

cooling rate and the convergence speed are both of polynomial order, which is not very fast in practice. In particular, an exponentially decreasing cooling schedule may give better result, although it will theoretically not converge. These result, however, allow to visualize how the problem properties will affect the convergence of the RPMP.

In the case of ITEL and IODL, the precise values of constants Γ and $\hat{\alpha}$ are hard to compute despite our knowledge of these processes. In ITEL (resp. IODL) one may conjecture that the exit height of cycles should not go beyond 2 (resp. 1) as all outward resistances $r^\star$ are bounded this way in the resistance graph over recurrence classes. It would follow that $\Gamma_0 \leq 2$ (resp. $\Gamma_0 \leq 1$) and $\hat{\alpha}_0 \geq \frac{\Delta\gamma}{2}$ (resp. $\hat{\alpha}_0 \geq \Delta\gamma$), where $\Delta\gamma$ is equal to the sub-optimal gap between the lowest potential and the second lowest, which can be expressed in terms of sub-optimal gap of welfare and stability depending on the context.

A Proof of ITEL convergence

In this section we prove results stated in Section 3.4, that is Propositions 3.1 and 3.2 and Theorems 3.3 and 3.4. Let P^ε be the Markov process defined by the ITEL algorithm for any $\varepsilon \in [0, 1)$. As stated in Section 1, the computation of $\mathcal{X}^\star$ can be simplified by reasoning over the recurrence classes of P^0 instead of the whole Markov chain. Hence the first step of the proof is to identify these recurrence classes. The second step will be to compute resistances and then potentials in order to describe $\mathcal{X}^\star$.

Remark. The proof given in this section follows the same reasoning as in [1], although the organization of lemmas and some details may differ slightly.

A.1 Recurrence Classes of the Unperturbed Process

The goal of this first section[12] is to study paths of zero resistance, i.e., paths in the unperturbed process. In particular, we are going to show that in P^0 any state can reach a state in $\mathcal{C} \cup \{D\}$, and that each state $x \in \mathcal{C}$ is absorbing. From there we will deduce Proposition 3.1: the recurrence classes of P^0 are the singletons $\{x\} \subset \mathcal{C}$ and possibly the communication class of D.

First of all, let us show that intermediate moods `H` and `W` and non-aligned benchmarks are temporary events in some sense. Formally, the following lemma holds:

Lemma A.1. In the unperturbed process P^0 there is a path from any state to a state where all players are either content with aligned benchmark or discontent, and so that discontent players stay so along the path.

Remark. Recall that discontent players store no benchmark actions. When saying that a state $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}})$ featuring discontent players is aligned, we mean that $\overline{\boldsymbol{a}}$ can be extended to all players in a way that $\overline{\boldsymbol{u}}$ is aligned with $\overline{\boldsymbol{a}}$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{X}$ be any state and extend $\overline{\boldsymbol{a}}$ with arbitrary actions for discontent players. Denote $\boldsymbol{u}$ the utilities resulting from $\overline{\boldsymbol{a}}$, which may differ from $\overline{\boldsymbol{u}}$ when the latter is not aligned with $\overline{\boldsymbol{a}}$. Let every player play according to $\overline{\boldsymbol{a}}$ for two steps. This is possible according to Table 2a. Indeed, in the unperturbed process, a non-discontent player always play its benchmark action. Moreover, a discontent player plays at random uniformly among all actions, hence has a positive probability of playing according to $\boldsymbol{a}$ twice in a row. Players observe $\boldsymbol{u}$ for two steps and the following behaviors may happen:

- Discontent players stay discontent.
- Watchful players with $u < \overline{u}$ reject after one step and end up discontent.
- Content or hopeful players with $u < \overline{u}$ become watchful, then discontent.
- Content players with $u = \overline{u}$ stay in their content state.
- Hopeful players with $u \geq \overline{u}$ and watchful players with $u = \overline{u}$ accept the outcome after one step and end up content.
- Content or watchful players with $u > \overline{u}$ become hopeful, then content.

According to Table 2b, the behaviors described above happen in P^0 with probability 1 assuming $\overline{\boldsymbol{a}}$ is played twice in a row. The only exception is the case of a discontent player observing u such that $F(u) = 0$, in which case rejection happens only with probability $1 - c_F$, which is still positive. Therefore, this path has a positive probability to happen in P^0 and leads after two steps to a state $y = (\overline{\boldsymbol{m}}', \overline{\boldsymbol{a}}, \boldsymbol{u})$ where all players are content or discontent and where content players have aligned benchmarks by definition of $\boldsymbol{u}$. □

[12] Appendix A.1 is equivalent to [1, Lemma 1], although the structure of the proof differs.

We can now study the influence of a player's exploration over all players. In fact, using the interdependence assumption A1, we are able to show that a discontent player can explore in a way that makes other players become discontent. Iterating the process, all players can eventually become discontent and the following lemma holds:

Lemma A.2. In the unperturbed process P^0 there is a path from any state featuring at least one discontent player, to D.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{X}$ be a state with at least one discontent player. Using Lemma A.1, we can assume that x only features aligned content players and discontent players. Let i be a discontent player and $a_i \neq \overline{a}_i$ be any action. Denote $\boldsymbol{u}$ the utilities resulting from $(a_i, \overline{a}_{-i})$. Let every player play according to $(a_i, \overline{a}_{-i})$ for two steps, then back to $\overline{\boldsymbol{a}}$ for another two steps. With the same justification as in Lemma A.1, this has positive probability to happen in P^0 according to Table 2a. Players observe $\boldsymbol{u}$ for two steps then $\overline{\boldsymbol{u}}$ for the other two steps, and the following behaviors may happen:

- Discontent players – including i – stay discontent.
- Content players with $u < \overline{u}$ become watchful then discontent. They stay discontent for the following two steps.
- Content players with $u > \overline{u}$ become hopeful then accept u as their new benchmark. They then become watchful then discontent, as their new benchmark is u but they observe $\overline{u}$.
- Content players with $u = \overline{u}$ stay in their content state.

According to Table 2b and with the same arguments as in Lemma A.1, the behaviors described above happen in P^0 with positive probability under the aforementioned choices of actions.

Now, using A1 over the set of discontent players in x, one can choose i and a_i such that $u_j \neq \overline{u}_j$ for some content player j. In other words, at least one player fits in one of the last two categories, hence eventually becomes discontent along the described path. We have shown that there is a path in P^0 from x to another state with strictly more discontent players. Iterating the reasoning until all players are discontent, we deduce that there is a path in P^0 from the original state x to D. □

Now, combining both Lemmas A.1 and A.2, we conclude that in P^0, there is a path from any state to either D or a state in $\mathcal{C}$. Indeed, there is a first path to a state where all players are either discontent or content and aligned. If this state is not in $\mathcal{C}$ then there is at least one discontent player hence a path to D. This implies that the recurrence classes of P^0 are included in the communication classes of $\mathcal{C}$ and D. If all players are content and aligned then no one explores or observes a change in its utility, so the corresponding state is absorbing. Hence every $\{x\} \subset \mathcal{C}$ is a recurrence class of P^0. This proves Proposition 3.1.

Regarding whether or not D is recurrent, the introduction of H1 makes the discussion more complex than in TEL. Without H1, a discontent player would always stay so, hence D would be absorbing and $\{D\}$ would be a recurrence class of P^0. Under H1, it is possible that the communication class of D contains other states as a discontent player accepts an exploration when the observed utility u satisfies $F(u) = 0$. Furthermore, it is possible that D is not recurrent if all players can simultaneously observe a utility satisfying $F(u) = 0$, as this implies the existence of a path from D to an aligned all-content state, which is absorbing. Either way this does not influence our reasoning: the goal of this section was to get rid of other states when reasoning over the resistance graph.

From now on we consider the resistance graph $\mathcal{G}$ over the class of D and classes $\{x\} \subset \mathcal{C}$. We remove the brackets and refer to them as D and $x \in \mathcal{C}$ to ease notations. Note that if D is not recurrent then its outward resistance will be $r^\star(D) = 0$ and we know right away that it will not minimize γ.

A.2 Resistances and Potentials

Now that we have isolated the recurrence classes of P^0, we can compute the resistance between these classes in the perturbed process P^ε and eventually their potential. From now on we reason over the graph of the

classes $x \in \mathcal{C}$ and D, where the resistance of any edge $x \to y$ is the lowest resistance of a path $x \rightsquigarrow y$ in P^ε. Recall that $\mathcal{E} \subset \mathcal{C}$ is the set of SE aligned all-content state and $\mathcal{A} \subset \mathcal{E}$ the subset of admissible states, see Definition 3.2. Definitions of easy edges, rooted trees, and potentials were given in Definitions 2.3 and 2.4. Definitions of virtual welfare and stability were given in Definition 3.4.

It is not necessary to compute all possible resistances in $\mathcal{G}$ to compute potentials, as we are only interested in optimal rooted trees. In fact, it is sufficient to study the resistances of edges $D \to x$ along with the easy edges of states $x \in \mathcal{C}$. Regarding the latter, the following lemma holds:[13]

Lemma A.3. Let $x \in \mathcal{C}$.

(i) If $x \in \mathcal{A}$, $r^*(x) = +\infty$.

(ii) If $x \in \mathcal{E} \setminus \mathcal{A}$, $r^*(x) = 2 = r(x \to D)$.

(iii) If $x \in \mathcal{C} \setminus \mathcal{E}$, $r^*(x) = 1 + \tilde{S}(x)$, and if $x \to D$ is not easy then any easy edge $x \to y \in \mathcal{C}$ leads to a strictly better welfare: $W(y) > W(x)$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{C}$.

(i). When x is admissible, no player can explore, so that $\overline{\boldsymbol{a}}$ is played continuously and so is observed $\overline{\boldsymbol{u}}$. Players never change state, and the state is absorbing, i.e., $r^\star(x) = +\infty$.

(ii). When x is a SE, all players are either admissible, hence never explore, or at an equilibrium, hence cannot accept their exploration if they are the only player to explore as the observed utility would be lower than their benchmark. Hence, a single exploration cannot leave the recurrence class of x: other players may become hopeful or watchful but then would require another exploration to change state, else they revert to their original state. Therefore at least two explorations are needed to leave the recurrence class of x, so $r^*(x) \geq 2$.

Reciprocally, one can construct a path $x \rightsquigarrow D$ of resistance equal to 2 using a reasoning almost identical to that of Lemma A.2. Indeed, consider the same path as the one described, except that player i is chosen as a non-admissible content player instead of a discontent player. Such player exists since $x \notin \mathcal{A}$. Let i play some action $a_i \neq \overline{a}_i$ twice in a row then go back to playing $\overline{a}_i$ while all other players play according to $\overline{a}$. As x is a SE and i is not admissible, i must be at an equilibria, so that its observation u_i in the action profile $(a_i, \overline{a}_{-i})$ is lower than its benchmark $\overline{u}_i$. Therefore i always reverts after playing a_i. This choice of actions implies a resistance of 2, as two explorations are performed. Similarly to Lemma A.2, players observe $\boldsymbol{u}$ for two steps then $\overline{\boldsymbol{u}}$ for the other two steps, and the following behaviors may happen:

- i reverts twice then stays in its content state.
- Content players with $u < \overline{u}$ become watchful then discontent. They stay discontent for the following two steps.
- Content players with $u > \overline{u}$ become hopeful then accept u as their new benchmark. They then become watchful then discontent, as their new benchmark is u but they observe $\overline{u}$.
- Content players with $u = \overline{u}$ stay in their content state.

According to Table 2b, the behaviors described above happen with no additional resistance under the aforementioned choices of actions. Now, using the interdependence assumption A1 over the singleton $\{i\}$, one can choose a_i such that $u_j \neq \overline{u}_j$ for player $j \neq i$. In other words, at least one player fits in one of the last two categories, hence eventually becomes discontent along the described path. We have shown that there is a path of resistance equal to 2 from x to another state with at least one discontent player. We conclude with Lemma A.2 that with no additional resistance the path can be extended to D. Therefore, $r(x \to D) = 2 = r^\star(x)$.

[13] Lemma A.3 is equivalent to [1, Lemmas 2 and 3], with the addition of admissible states.

(iii). When x is not a SE, one exploration is still not enough to change state, but a player i neither admissible nor at an equilibrium may accept the outcome of its exploration with some resistance controlled by G. Let i be such player, and let a_i be an action such that the observed utility $u_i = U_i(a_i, \overline{a}_{-i})$ when i explores a_i satisfies $u_i > \overline{u}_i$. The resistance of accepting this exploration is $G(\overline{u}_i, u_i)$, which is exactly $\tilde{S}(x)$ according to Definition 3.4 when i and a_i are chosen as to minimize $G(\overline{u}_i, u_i)$. Adding the cost of exploration we get the lower bound $r^*(x) \geq 1 + \tilde{S}(x)$. Note that this bound is strictly lower than 2 as $\tilde{S}(x) < G_0 \leq 1$ under C2.

Let us now show that accepting this exploration is sufficient to change state, so that $r^\star(x) = 1 + \tilde{S}(x)$. Consider the path similar to the first two steps of described in (ii) except i accepts its exploration after the first step and plays it again instead of reverting and exploring again. This causes the others content players to either become discontent or to stay content with a utility $u_j \geq \overline{u}_j$. If all players stay content, the new state y is an aligned all-content state and all players have improved in utility (i strictly), hence $W(y) > W(x)$. Else there is a path of zero resistance to D according to Lemma A.2. Overall this path has an associated resistance of $1 + \tilde{S}(x)$ due to i exploring and accepting. We conclude that $r^*(x) = 1 + \tilde{S}(x)$. □

Now, let us compute the resistance of edges $D \to x \in \mathcal{C}$.[14]

Lemma A.4. For all $x \in \mathcal{C}$, $r(D \to x) = 1 - \tilde{W}(x)$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{C}$ and consider any path $D \rightsquigarrow x$. Each player i eventually ends up content with utility $\overline{u}_i$. Since a content player cannot decrease in benchmark utility without first getting discontent, there is a point in the path where i goes from `D` to `C` with a utility $u_i \leq \overline{u}_i$. This costs a resistance of $F(u_i) \geq F(\overline{u}_i)$ as F is non-increasing. This reasoning holds for each player, hence the total resistance of the path is at least $\sum_i F(\overline{u}_i) = 1 - \tilde{W}(x)$.

Reciprocally, there is a direct path $D \to x$ of resistance $\sum_i F(\overline{u}_i)$ when all players choose to play according to $\overline{\boldsymbol{a}}$ simultaneously and to accept the outcome, hence $r(D \to x) = 1 - \tilde{W}(x)$. □

Before moving on to computing potentials, notice that the previous lemmas allow us to conclude on the nature of the ITEL process. First of all, states $x \in \mathcal{A}$ are absorbing even in the perturbed process according to Lemma A.3. Moreover, it appears that there is a path in P^ε of finite resistance from D to any state $x \in \mathcal{C}$ due to Lemma A.4. We can show that there is also a path from x back to either D or a state in $\mathcal{A}$ when $x \notin \mathcal{A}$. Indeed, Lemma A.3 shows that an easy edge leaving x leads either to D or to another state in $\mathcal{C}$ with strictly higher welfare. Since there is a finite amount of states in $\mathcal{C}$, following easy edges eventually leads to D or to $\mathcal{A}$. Indeed, if it was not the case, the easy edges would create a cycle $x = x_1 \to x_2 \to \cdots \to x_m = x$ at one point, but this cycle would verify $W(x) = W(x_1) < W(x_2) < \cdots < W(x_m) = W(x)$, which is a contradiction.

Therefore, if $\mathcal{A} \neq \emptyset$, there is a path from any state to a state in $\mathcal{A}$ which is absorbing. Hence the perturbed process is not irreducible and its recurrence classes are exactly the singletons $\{x\} \subset \mathcal{A}$. On the other hand, when $\mathcal{A} = \emptyset$, D communicates with all states in $\mathcal{C}$, so that all these states belong to the same recurrence class. Hence the perturbed process is constituted of a unique recurrence class, i.e., is irreducible[15]. It is also aperiodic as the transition $D \to D$ has positive probability. This concludes Proposition 3.2. The case of Theorem 3.3 where $\mathcal{A} \neq \emptyset$ is treated easily by common Markov processes arguments: the process converges a.s. to one of its recurrence classes, which happens to be one of the absorbing states $x \in \mathcal{A}$.

The rest of this section is devoted to the case where $\mathcal{A} = \emptyset$, where Theorem 2.1 can be applied. We now study optimal rooted trees for each state, in order to compute their potential.[16]

Lemma A.5.

(i) $\gamma(D) = \sum_{x \in \mathcal{C}} r^*(x)$.

[14] Lemma A.4 is equivalent to [1, Lemma 4].

[15] Actually, there may be states that are not be part of the recurrence class of P^ε. In this case we restrict ourselves to the study of the single recurrence class, which is a RPMP. This is without loss of generality, as the algorithm ends up a.s. in this recurrence class regardless of the starting state.

[16] Lemma A.5 is equivalent to [1, Lemmas 5 and 6].

(ii) $\gamma(x) = \gamma(D) - r^*(x) + r(D \to x) = \begin{cases} \gamma(D) - 1 - \tilde{W}(x) & \text{if } x \in \mathcal{E}, \\ \gamma(D) - \tilde{S}(x) - \tilde{W}(x) & \text{if } x \in \mathcal{C} \setminus \mathcal{E}. \end{cases}$

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{u}}) \in \mathcal{C}$.

(i). We want to show that it is possible to create a D-tree using only easy edges. For each $x \in \mathcal{C}$, choose an easy edge from x, going to D if possible, and consider the resulting sub-graph with only the chosen edges. According to Lemma A.3, the chosen easy edge either leads directly to D or leads to a state of strictly better welfare. We already argued that this choice of edges cannot create a cycle. Therefore, the sub-graph is acyclic with exactly one outward edge for each $x \in \mathcal{C}$, hence a D-tree. We conclude that $\gamma(D) = \sum_{x \in \mathcal{C}} r^*(x)$.

(ii). It is immediate that in the previous tree, removing the edge leaving x and adding $D \to x$ instead gives a x-tree of resistance $\gamma(D) - r^*(x) + r(D \to x)$.

Let us now show that this x-tree is optimal. Consider any x-tree T and the path $D \to y_1 \to y_2 \to \cdots \to y_m \to x$ in T. Denote $\boldsymbol{u}$ the utilities of x. Reusing the proofs of Lemmas A.3 and A.4, we know that along this path:

- As shown in Lemma A.3, each state y_k has a minimal outward resistance $r^\star(y_k)$ due to explorations and acceptations from content players.
- As shown in Lemma A.4, each player i must accept at some point a utility $u_i \leq \overline{u}_i$ from a discontent state, for a resistance $F(u_i) \geq F(\overline{u}_i)$.

Both points reason over behaviors of different nature ; the resistances due to content players exploring or accepting do not intersect with the resistances due to discontent players accepting. Hence we deduce from these observations that the total resistance of the path is greater than the sum of all the resistance mentioned above, that is $\sum_{k=1}^m r^*(y_k) + \sum_i F(\overline{u}_i)$. Lower bounding the resistances of the other edges of T via $r^\star$, we have

$$r(T) \geq \sum_{y \in \mathcal{C} \setminus \{x\}} r^*(y) + \sum_i F(\overline{u}_i) = \gamma(D) - r^*(x) + r(D \to x).$$

This proves that $\gamma(x) = \gamma(D) - r^*(x) + r(D \to x)$. The specific formulas depending on whether $x \in \mathcal{E}$ or not are derived directly from Lemma A.3. □

We can now conclude using the bounds on F and G from C2, we recall:

$$\begin{cases} 0 \leq F \leq \frac{F_0}{n}, \\ 0 \leq G < G_0, \\ F_0 + G_0 \leq 1. \end{cases}$$

C2 was essentially chosen so that the following inequalities hold:

$$0 \leq \tilde{S} < \tilde{W} \leq 1 \tag{37}$$

for any possible value of $\tilde{S}$ and $\tilde{W}$. Indeed, denote $\overline{\boldsymbol{u}}$ the utilities of x, so that $\tilde{W}(x) = 1 - \sum_i F(\overline{u}_i)$. $\tilde{W}(x) \leq 1$ as $F \geq 0$. Denote $\overline{u}', u'$ utilities such that $\tilde{S}(x) = G(\overline{u}', u')$. $\tilde{S}(x) \geq 0$ as $G \geq 0$. Finally, C2 implies that $\tilde{S} - \tilde{W} = G(\overline{u}', u') + \sum_i F(\overline{u}_i) - 1 < G_0 + n\frac{F_0}{n} - 1 \leq 0$, hence $\tilde{S} < \tilde{W}$. It follows that if $x \in \mathcal{E}$ and $y \in \mathcal{C} \setminus \mathcal{E}$, and according to Lemma A.5,

$$\begin{aligned} \gamma(x) &= \gamma(D) - 1 - \tilde{W}(x) < \gamma(D) - 1 - \tilde{S}(y) \leq \gamma(D) - \tilde{W}(y) - \tilde{S}(y) = \gamma(y), \\ \gamma(y) &= \gamma(D) - \tilde{W}(y) - \tilde{S}(y) < \gamma(D). \end{aligned}$$

Moreover, γ is minimized over $\mathcal{E}$ at states maximizing $\tilde{W}$, and minimized over $\mathcal{C} \setminus \mathcal{E}$ at states maximizing $\tilde{W} + \tilde{S}$. We conclude that, if $\mathcal{E} \neq \emptyset$, $\mathcal{X}^\star$ is the set of aligned all-content states SE maximizing $\tilde{W}$. Else, if $\mathcal{E} = \emptyset$, $\mathcal{X}^\star$ is the set of aligned all-content states maximizing $\tilde{W} + \tilde{S}$. This concludes the proof of Theorem 3.3.

Let us discuss the special case where F and G are chosen of the form $F : u \mapsto \phi_F - \psi_F \cdot u$ and $G : (\overline{u}, u) \mapsto \phi_G - \psi_G \cdot (u - \overline{u})$. In this case, for any state $x \in \mathcal{C}$, denote $\overline{\boldsymbol{u}}$ the utilities of x and $\overline{u}', u'$ utilities such that $S(x) = u' - \overline{u}'$. We have

$$\tilde{W}(x) = 1 - \sum_i F(\overline{u}_i) = 1 - n\phi_F + \psi_F \sum_i \overline{u}_i = 1 - n\phi_F + n\psi_F W(x),$$
$$\tilde{S}(x) = G(S(x)) = \phi_G - \psi_G S(x).$$

It follows that maximizing $\tilde{W}$ is equivalent to maximizing W, and that maximizing $\tilde{W} + \tilde{S}$ is equivalent to maximizing $\psi_F W - \psi_G S$.

B Proof of IODL convergence

The proof of convergence for IODL is the same as for ITEL. In this section we highlight the few differences due to the removal of intermediate moods `H` and `W` which yield Theorem 3.5.

B.1 Recurrence Classes of the Unperturbed Process

All of the discussion done in Appendix A.1 also holds for IODL. The paths described must be adapted by removing intermediate step, but the reasoning is identical. Recall that the policies allowing these paths are given in Table 3.

In particular, Lemmas A.1 and A.2 are true for IODL, and so is Proposition 3.1. In the next section we study the resistance graph which states are the communication classes of all $x \in \mathcal{C}$ and D.

B.2 Resistances and Potentials

Lemmas A.3 and B.2 are adapted as follows:

Lemma B.1.

(i) If $x \in \mathcal{A}$, $r^\star(x) = +\infty$.

(ii) If $x \in \mathcal{C} \setminus \mathcal{A}$, $r^*(x) = 1$, and if $x \to D$ is not easy than any easy edge $x \to y \in \mathcal{C}$ leads to a strictly better welfare: $W(y) > W(x)$.

Proof. When $x \in \mathcal{A}$, the reasoning is the same as in Lemma A.3. Let $x \in \mathcal{C} \setminus \mathcal{A}$. Leaving x requires some exploration hence $r^\star(x) \geq 1$.

Now let us describe a path leaving x of resistance equal to 1. By interdependence assumption A1, one can find a non-admissible player i that is able to explore in a way that affects other players. First, let us assume that i is able to do so in a way so that $G(\overline{u}_i, u_i) > 0$, so that it can reverts with no resistance. Let i explore once than playing its benchmark action. According to Table 3 and with the same arguments that were used in Lemma A.3 (ii), the following path may happen with resistance equal to 1 due to a single observation from i:

- i reverts once then stays in its content state.
- Content players with $u < \overline{u}$ become discontent. They stay discontent after the second step.
- Content players with $u > \overline{u}$ accept u as their new benchmark. They then become discontent, as their new benchmark is u but they observe $\overline{u}$.
- Content players with $u = \overline{u}$ stay in their content state.

Along this path some players become discontent. We conclude with Lemma A.2 that with no additional resistance the path can be extended to D. Therefore, $r(x \to D) = 1 = r^\star(x)$.

Now, in the case where $G(\overline{u}_i, u_i) = 0$, i can accept its exploration with no resistance. Then we can conclude with the same reasoning as in Lemma A.3 (iii) that there is a path to D or to another state $y \in \mathcal{C}$ with $W(y) > W(x)$. □

Lemma B.2. For all $x \in \mathcal{C}$, $r(D \to x) = 1 - \tilde{W}(x)$.

Proof. The reasoning is exactly the same as in Lemma A.4. □

One can see that Proposition 3.2 holds in the case of IODL for the same reasons as for ITEL. The case where $\mathcal{A} \neq \emptyset$ in Theorem 3.5 is then treated the same way.

When $\mathcal{A} = \emptyset$, although the numerical values of resistances are not the same as in ITEL, they do share the same properties and applying Lemma A.5's reasoning yields the following:

Lemma B.3.

(i) $\gamma(D) = \sum_{x \in \mathcal{C}} r^*(x)$.

(ii) If $x \in \mathcal{C}$, $\gamma(x) = \gamma(D) - r^*(x) + r(D \to x) = \gamma(D) - \tilde{W}(x)$.

Now recall that C3 imposes $nF < 1$, hence $\tilde{W}(x) = 1 - \sum_i F(u_i) \in (0, 1]$ for any $x \in \mathcal{C}$ of utilities $\boldsymbol{u}$. From there, it is immediate with Lemma B.3 that γ is minimized at states $x \in \mathcal{C}$ maximizing $\tilde{W}(x)$, and applying Theorem 2.1 concludes Theorem 3.5.

C Proof of RITEL convergence

In this section we prove all results stated in Section 4.4, that is Propositions 4.11 and 4.12 and Theorems 4.13 and 4.14. Let P^ε be the Markov process defined by the RITEL algorithm for any $\varepsilon \in [0, 1)$. As for ITEL, the first step of the proof is to identify the recurrence classes of the unperturbed process P^0. The second step will be to compute resistances and then potentials in order to describe $\mathcal{X}^\star$. Contrarily to ITEL, resistances will here be estimated with some margin of error that we can control in order to identify states that may be in $\mathcal{X}^\star$.

The proof detailed in this section is very similar to Appendix A. In fact, most adaptations have already been discussed in Section 4.4, and it only remains to integrate them to the regular ITEL reasoning.

C.1 Recurrence Classes of the Unperturbed Process

The goal of this first section is to study paths of zero resistance, i.e., paths in the unperturbed process. In particular, we are going to show that in P^0 any state can reach a state in $\mathcal{C}_\delta \cup \{D\}$, and that each state $x \in \mathcal{C}_\delta$ is absorbing. From there we will deduce Proposition 4.11: the recurrence classes of P^0 are the singletons $\{x\} \subset \mathcal{C}_\delta$ and possibly the communication class of D.

Apart from $\mathcal{C}$ being replaced with $\mathcal{C}_\delta$, this section is proven exactly as in Appendix A.1 and Lemmas C.1 and C.2 hold the same statements as Lemmas A.1 and A.2. We prove them by referring to the ITEL case, highlighting the few required adaptations. Recall that the policies of RITEL are given in Table 4.

Lemma C.1. In the unperturbed process P^0 there is a path from any state to a state where all players are either content with weakly aligned benchmark or discontent, and so that discontent players stay so along the path.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{X}$ be any state and extend $\overline{\boldsymbol{a}}$ with arbitrary actions for discontent players. Denote $\overline{\boldsymbol{\nu}}$ the expected bins resulting from $\overline{\boldsymbol{a}}$. We have shown in Corollary 4.3 that the resistance of observing the expected bin is zero. Hence the event where all players play according to $\overline{\boldsymbol{a}}$ and observe $\boldsymbol{v} = \overline{\boldsymbol{\nu}}$ has positive probability to happen in P^0.[17] Then, the path described Lemma A.1 – replacing comparisons $u = \overline{u}$ (resp. $u < \overline{u}$ and $u > \overline{u}$) by $v = \overline{v} \pm \delta$ (resp. $v \leq \overline{v} - 2\delta$ and $v \geq \overline{v} + 2\delta$) – also happens here with positive probability. This path leads to a state where players are either discontent or content. Moreover, content players either kept their benchmark if it was already weakly aligned, or accepted $v = \overline{\nu}$. Either way content players are all weakly-aligned with $\overline{\boldsymbol{a}}$ in this new state, which concludes the proof. □

Lemma C.2. In the unperturbed process P^0 there is a path from any state featuring at least one discontent player, to D.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{X}$ be a state with at least one discontent player. Using Lemma C.1, we can assume that x only features aligned content players and discontent players. Let i be a discontent player and

[17] Regarding the special case where the expected utility μ lies exactly at the edge between bins $\overline{\nu} - \delta$ and $\overline{\nu}$ and the utility distribution is not deterministic, both bins are observed with probability $\frac{1}{2}$ in P^0. A benchmark $\overline{v} = \overline{\nu} + \delta$ is not considered weakly aligned in this case according to Definition 4.3. We assume that players in such situation observe $\overline{\nu} - \delta$, so that they act as if they observed a lower utility and eventually become discontent.

$a_i \neq \overline{a}_i$. Denote $\boldsymbol{\nu}$ the expected bins of $(a_i, \overline{a}_{-i})$. As for Lemma C.1, the path described in Lemma A.2 – replacing comparisons $u = \overline{u}$ (resp. $u < \overline{u}$ and $u > \overline{u}$) by $\nu = \overline{v} \pm \delta$ (resp. $\nu \leq \overline{v} - 2\delta$ and $\nu \geq \overline{v} + 2\delta$) – can be used again and leads to another state with more discontent players, assuming that a discontent player i can explore in a way such that some non discontent player j observes $\nu_j \neq \overline{v}_j \pm \delta$.

It remains to show that the stronger interdependence assumption A2 satisfies this condition. Indeed, according to A2, one can find a discontent player i and an action a_i such that $M_j(a_i, \overline{a}_{-i}) \leq M_j(\overline{\boldsymbol{a}}) - 3\delta$ or $M_j(a_i, \overline{a}_{-i}) \geq M_j(\overline{\boldsymbol{a}}) + 3\delta$ for some non discontent player j. In the first case, $\nu_j^- \leq M_j(a_i, \overline{a}_{-i}) \leq M_j(\overline{\boldsymbol{a}}) - 3\delta < \overline{v}_j^- - \delta$ as weak alignment implies $M_j(\overline{\boldsymbol{a}}) \in [\overline{v}_j^- - \delta, \overline{v}_j^- + 2\delta)$. Hence $\nu_j < \overline{v}_j - \delta$, thus $\nu_j \leq \overline{v}_j - 2\delta$, and similarly $\nu_j \geq \overline{v}_j + 2\delta$ in the second case.

We conclude that the path described above happens with positive probability in P^0. Iterating it until all players are discontent concludes the proof. □

As for ITEL, Lemmas C.1 and C.2 imply that in P^0, there is a path from any state to either D or a state in $\mathcal{C}_\delta$. In particular, the recurrence classes of P^0 are included in the communication classes of $\mathcal{C}_\delta$ and D. If all players are content then no one explores. Moreover, observed utilities always fall in the expected bin as P^0 has deterministic payoffs. If an all-content state is weakly aligned, then the observed bin is always within δ from the benchmark,[18] hence all players stay in their state, so that the state is absorbing. Hence every $\{x\} \subset \mathcal{C}_\delta$ is a recurrence class of P^0. This proves Proposition 3.1.

Regarding whether or not D is recurrent, the same discussion as in ITEL holds: the communication class of D is not necessarily reduced to D itself, and may not be recurrent. Either way from now on we consider the resistance graph $\mathcal{G}$ over the class of D and classes $\{x\} \subset \mathcal{C}_\delta$. We remove the brackets and refer to them as D and $x \in \mathcal{C}_\delta$ to ease notations. Note that if D is not recurrent then its outward resistance will be $r^\star(D) = 0$ and we know right away that it will not minimize γ.

C.2 Resistances and Potentials

Now that we have isolated the recurrence classes of P^0, we can compute the resistance between these classes in the perturbed process P^ε and eventually their potential. From now on we reason over the graph of the classes $x \in \mathcal{C}_\delta$ and D, where the resistance of any edge $x \to y$ is the lowest resistance of a path $x \rightsquigarrow y$ in P^ε. Recall that sets $\mathcal{C}_\delta$, $\mathcal{C}$, $\mathcal{E}_\delta$, $\mathcal{E}$, $\mathcal{A}_\delta$ and $\mathcal{A}$ were defined in Section 4.4. Definitions of easy edges, rooted trees, and potentials were given in Definitions 2.3 and 2.4. Definitions of virtual welfare and stability of a state were given in Definition 4.8.

As for ITEL, we only need to study the resistances of edges $D \to x$ along with the easy edges of states $x \in \mathcal{C}_\delta$. Before computing the outward resistances from states $x \in \mathcal{C}_\delta$, we discuss the nature of an easy edge leaving x. As in the case of ITEL, it is required in order to leave x that either a player observes a non-aligned utility twice, or that a player accepts its own exploration. However, to observe a non-aligned utility, a player can be influenced by another as in ITEL, or can make an offset observation due to noise. The following lemma discusses the resistance of each three kinds of paths depending on the nature of the starting state x. An easy edge leaving x will correspond to the path with minimal resistance among them.

Lemma C.3.

(i) If $x \in \mathcal{C}_\delta$, there exists a path from x to D where a player makes an offset observation twice, with resistance equal to $2\tilde{R}(x)$.

(ii) If $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$, there exists a path from x to D where a player explores twice and influences another player into changing state,with resistance equal to 2. If $x \in \mathcal{A}_\delta$, no such path exists.

(iii) If $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$, there exists a path from x to another state where a player explores and accepts, with resistance equal to $1 + \tilde{S}_+(x)$. If $x \in \mathcal{C} \setminus \mathcal{E}$, such path has a resistance greater than $1 + \tilde{S}_-(x)$. If

[18] Regarding the special case where the expected utility μ lies exactly at the edge between bins $v - \delta$ and v and the utility distribution is not deterministic, both bins are observed with probability $\frac{1}{2}$ in P^0. By Definition 4.3, $v + \delta$ is not considered weakly aligned in this case, so the reasoning is not hindered.

$x \in \mathcal{E} \setminus \mathcal{A}_\delta$, such path has a resistance greater than 2. If $x \in \mathcal{A}_\delta$, no such path exists. Moreover, this path leads either to D or to a state $y \in \mathcal{C}_\delta$ verifying $W(y) > W(x)$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{C}_\delta$.

(i). Is no one explores, the minimal resistance for a player to make a non-aligned observation is exactly $\tilde{R}(x)$ by Definition 4.9. Assume that such observation happens twice in a row while other players observe their expected bin, which is within δ of their benchmark bin by weak alignment. The associated resistance is $2\tilde{R}(x)$ and the player making offset observations becomes either watchful then discontent if the observation was lower than its benchmark, or becomes hopeful then accepts the offset observation as its new benchmark if it was higher. In the second case, the player is no longer aligned, hence with no resistance it observes its expected bin which is now perceived as a deterioration. It then becomes watchful, then discontent. Either way the player eventually becomes discontent and the path can be extended from there to D with no additional resistance. Note that if $\tilde{R}(x) = +\infty$, the path may actually not exist, however this is equivalent to saying that it exists with infinite resistance.

(ii). If $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$, the path described in Lemma A.3 (ii) can be used again and leads to D, with the same adaptations as in Lemmas C.1 and C.2. Its resistance is equal to 2, corresponding to both explorations. If $x \in \mathcal{A}_\delta$, all players are admissible so no exploration can happen to influence other players and the path does not exist.

(iii). If $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$, one can find a player that is neither admissible nor at a δ-equilibrium position. Such player i can explore with resistance 1 some action a_i and observe the bin $\nu_i = N_i(a_i, \overline{a}_{-i}) \geq \overline{v}_i + 2\delta$ with no additional resistance. It can then accept it with resistance $G(\overline{v}_i, \nu_i)$. Choosing i and a_i such that this last quantity is minimal, the total resistance of this event is exactly $1 + \tilde{S}_+(x)$ by Definition 4.8. i can then play a_i again with no resistance as it is its new benchmark action. Meanwhile, any other player j whose benchmark bin $\overline{v}_j$ is not aligned with its new expected bin $\nu_j = N_j(a_i, \overline{a}_{-i})$ can observe a bin $v_j \neq \overline{v}_j \pm \delta$ with no resistance.[19] If all players observe an improvement, they become hopeful then accept their observation and the new state is a state $y \in \mathcal{C}_\delta$ with strictly higher welfare. Else, some player becomes discontent and the path can be extended from there to D with no additional resistance. Either way the path has resistance $1 + \tilde{S}_+(x)$.

The path described above is no longer guaranteed to exist when $x \in \mathcal{E}_\delta$, as no player i can explore in a way so that its new expected bin ν_i is lower than $\overline{v}_i + \delta$. However, if the resistance of observing a bin $v_i > \nu_i$ is low, it is possible that i accepts v_i. The total resistance of the path is then $1 + r_{U_i(a_i, \overline{a}_{-i})}(v_i) + G(\overline{v}_i, v_i)$. Since $G(\overline{v}_i, \cdot)$ is non-increasing and $r_{U_i(a_i, \overline{a}_{-i})}(v_i)$ is minimal – and equal to 0 – for $v_i = \nu_i$, the total resistance is minimized for some $v_i \geq \nu_i$. Accepting the bin ν_i costs in total $1 + G(\overline{v}_i, \nu_i)$, whereas accepting the bin $\nu_i + \delta$ costs in total $1 + r_{U_i(a_i, \overline{a}_{-i})}(\nu_i + \delta) + G(\overline{v}_i, \nu_i + \delta)$. One cannot know whichever is lower in general, however both can be lower bounded by $1 + G(\overline{v}_i, \nu_i + \delta)$. This quantity also lower bounds the resistance of accepting a bin $v_i \geq \nu_i + 2\delta$, as we saw in Corollary 4.3 that $r_{U_i(a_i, \overline{a}_{-i})}(v_i) \geq R_0$ in this case. Hence accepting such bin would give a total resistance of $1 + R_0 + G(\overline{v}_i, v_i) \geq 2 > 1 + G(\overline{v}_i, \nu_i + \delta)$ under C2 and C4. We have shown that the resistance of the path corresponding to i exploring a_i and accepting is lower bounded by $1 + G(\overline{v}_i, \nu_i + \delta)$.

If $x \in \mathcal{C} \setminus \mathcal{E}$, taking the minimum of this lower bound over all possible i and a_i such that $\nu_i \geq \overline{v}_i + \delta$ gives $1 + G(\overline{v}_i, \nu_i + \delta) = 1 + \tilde{S}_-(x)$ according to Definition 4.8. If $x \in \mathcal{E} \setminus \mathcal{A}_\delta$, all players are either admissible or at a δ-equilibrium position, which implies that any choice of i and a_i would verify $\nu_i \leq \overline{v}_i$, hence accepting would require an offset observation of at least two bins which we have shown to imply a total resistance greater than 2. If $x \in \mathcal{A}_\delta$, all players are admissible so no exploration can happen, let alone be accepted, so the path considered here cannot happen. □

Combining the above results yields the following estimates of the outward resistances.

[19] In general this bin would be $v_j = \nu_j$. The only exception being when the utility is not deterministic and $M_j(a_i, \overline{a}_{-i}) = \nu_j^- = \overline{v}_j^- - \delta$, in which case one should choose $v_j = \nu_j - \delta = \overline{v}_j - 2\delta$. This kind of reasoning has been described with more details in Appendix C.1.

Lemma C.4.

(i) If $x \in \mathcal{A}_\delta$, $r^\star(x) = 2\tilde{R}(x)$.

(ii) If $x \in \mathcal{E} \setminus \mathcal{A}_\delta$, $r^\star(x) = 2$.

(iii) If $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$, $r^\star(x) \leq 2$.

(iv) If $x \in \mathcal{C} \setminus \mathcal{E}$, $r^\star(x) \geq 1 + \tilde{S}_-(x)$.

(v) If $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$, $r^\star(x) \leq 1 + \tilde{S}_+(x)$.

Moreover, in all cases an easy edge leads either to D or to a state $y \in \mathcal{C}_\delta$ verifying $W(y) > W(x)$.

Proof. We already justified that an easy edge will necessarily correspond to one of the three paths of Lemma C.3, which implies that easy edges will lead to D or to $y \in \mathcal{C}_\delta$ with $W(y) > W(x)$. The rest of the proof is essentially a matter of which of the three paths – which we refer to as paths (i), (ii), and (iii) as stated in Lemma C.3 – has minimal resistance depending on the nature of x, eventually leading to an estimate of $r^\star(x)$.

(i). If $x \in \mathcal{A}_\delta$, Lemma C.3 states that path (i) is the only way to leave x, hence it is the easy edge and $r^\star(x) = 2\tilde{R}(x)$.

(ii). If $x \in \mathcal{E} \setminus \mathcal{A}_\delta$, x is strongly aligned, hence path (i) has resistance equal to $2\tilde{R}(x) \geq 2R_0 \geq 2$ according to Lemma 4.10 and C4. Moreover, path (ii) has resistance equal to 2, whereas path (iii) has resistance greater than 2. Therefore, path (ii) implies an easy edge and $r^\star(x) = 2$.

(iii). If $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$, path (ii) has resistance equal to 2, hence $r^\star(x) \leq 2$.

(iv). If $x \in \mathcal{C} \setminus \mathcal{E}$, x is strongly aligned, hence path (i) has resistance equal to $2\tilde{R}(x) \geq 2$. Path (ii) has resistance 2 and path (iii) has resistance greater than $1 + \tilde{S}_-(x)$, which itself is lower than 2. Therefore, $r^\star(x) \geq 1 + \tilde{S}_-(x)$.

(v). If $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$, path (iii) has resistance lower than $1 + \tilde{S}_+(x)$, hence $r^\star(x) \leq 1 + \tilde{S}_+(x)$. □

Now, let us compute the resistance of edges $D \to x \in \mathcal{C}$.

Lemma C.5.

(i) If $x \in \mathcal{C}_\delta$, then $r(D \to x) \geq 1 - \tilde{W}(x)$.

(ii) If $x \in \mathcal{C}$, then $r(D \to x) = 1 - \tilde{W}(x)$.

Proof. Let $x = (\overline{\boldsymbol{m}}, \overline{\boldsymbol{a}}, \overline{\boldsymbol{v}}) \in \mathcal{C}_\delta$.

(i). In RITEL as in ITEL, a content player cannot decrease in benchmark utility without first becoming discontent. As in Lemma A.4, it follows that for a discontent player to accept the bin $\overline{v}$, it is needed at some point that it accepts a bin $v \leq \overline{v}$ with resistance $F(v) \geq F(\overline{v})$. Applying the reasoning to all players, it follows that $r(D \to x) \geq \sum_i F(\overline{v}_i) = 1 - \tilde{W}(x)$.

(ii). When x is strongly aligned, the direct path $D \to x$ corresponding to all players simultaneously playing according to $\overline{\boldsymbol{a}}$ and accepting $\overline{\boldsymbol{v}}$ has a resistance equal to $1 - \tilde{W}(x)$, so that the lower bound is actually an equality: $r(D \to x) = 1 - \tilde{W}(x)$. □

Before moving on to computing potentials, notice that the previous lemmas allow us to conclude on the nature of the RITEL process. First of all, states $x \in \mathcal{A}_\delta$ are absorbing in the perturbed process if and only if $\tilde{R}(x) = +\infty$ according to Lemma C.4. Moreover, Lemmas C.4 and C.5 show that there is a path in P^ε of finite resistance from D to any state $x \in \mathcal{C}_\delta$, and back when x is not absorbing. The key argument to derive the above being exactly the same as in Appendix A.2: easy edges between states in $\mathcal{C}_\delta$ cannot create cycles as they always increase benchmark welfare.

Therefore, if $\mathcal{A} \neq \emptyset$, there is a path from any state to an absorbing state. Hence the perturbed process is not irreducible and its recurrence classes are exactly the singletons $\{x\} \subset \mathcal{A}$. On the other hand, when $\mathcal{A} = \emptyset$, D communicates with all states in $\mathcal{C}_\delta$, so that all these states belong to the same recurrence class. Hence the perturbed process is constituted of a unique recurrence class, i.e., is irreducible[20]. It is also aperiodic as the transition $D \to D$ has positive probability. This concludes Proposition 4.12. The case of Theorem 4.13 where $\mathcal{A} \neq \emptyset$ is treated easily by common Markov processes arguments: the process converges a.s. to one of its recurrence classes, which happens to be one of the absorbing states $x \in \mathcal{A}$.

The rest of this section is devoted to the case where $\mathcal{A} = \emptyset$, where Theorem 4.6 can be applied. We now study optimal rooted trees for each state, in order to compute their potential.

Lemma C.6.

(i) $\gamma(D) = \sum_{x \in \mathcal{C}} r^*(x)$.

(ii) $\gamma(x) = \gamma(D) - r^*(x) + r(D \to x)$
$$\begin{cases} \leq \gamma(D) - 2\tilde{R}(x) + 1 - \tilde{W}(x) & \text{if } x \in \mathcal{A}_\delta, \\ = \gamma(D) - 1 - \tilde{W}(x) & \text{if } x \in \mathcal{E} \setminus \mathcal{A}_\delta, \\ \geq \gamma(D) - 1 - \tilde{W}(x) & \text{if } x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta, \\ \leq \gamma(D) - \tilde{S}_-(x) - \tilde{W}(x) & \text{if } x \in \mathcal{C} \setminus \mathcal{E}, \\ \geq \gamma(D) - \tilde{S}_+(x) - \tilde{W}(x) & \text{if } x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta. \end{cases}$$

Proof. The exact same reasoning as in Lemma A.5 shows that a D-tree can be constructed using only easy edges, proving (i). Using again the same reasoning as in Lemma A.5 proves that $\gamma(x) = \gamma(D) - r^*(x) + r(D \to x)$. The different cases are deduced by replacing $r^*(x)$ and $r(D \to x)$ by the values given by Lemmas C.4 and C.5, concluding (ii). □

We can now conclude using the bounds on F and G from C2 which are common to ITEL and RITEL. Inequalities similar to (37) hold:

$$0 \leq \tilde{S}_+ < \tilde{W} \leq 1 \tag{38}$$

for any possible value of $\tilde{S}_+$ and $\tilde{W}$. These bounds are proven the same way as (37), by writing $\tilde{W} = 1 - \sum_i F(\overline{v}_i)$ and $\tilde{S}_+ = G(\overline{v}', v')$ and using the bounds from C2.

When $\mathcal{E} \neq \emptyset$, we want to show that $\mathcal{X}^\star \subset \mathcal{A}_\delta \cup \{x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta : \tilde{W}(x) \geq \tilde{W}^\star\}$ where $\tilde{W}^\star = \max_{x \in \mathcal{E}} \tilde{W}(x)$. Denoting $x^\star \in \mathcal{E}$ such that $\tilde{W}(x) = \tilde{W}^\star$, Lemma C.6 states that

$$\begin{cases} \gamma(x^\star) = \gamma(D) - 1 - \tilde{W}(x) = \gamma(D) - 1 - \tilde{W}^\star & \text{if } x \notin \mathcal{A}_\delta, \\ \gamma(x^\star) \leq \gamma(D) - 2\tilde{R}(x) + 1 - \tilde{W}(x) \leq \gamma(D) - 2R_0 + 1 - \tilde{W}^\star \leq \gamma(D) - 1 - \tilde{W}^\star & \text{if } x \in \mathcal{A}_\delta, \end{cases}$$

as x is strongly aligned hence $\tilde{R}(x) \geq R_0 \geq 1$ due to Lemma 4.10 and C4. It follows that states in $\mathcal{X}^\star$ must have a potential lower than $\gamma(D) - 1 - \tilde{W}^\star$. To conclude, it is then sufficient to show that states outside of $\mathcal{A}_\delta \cup \{x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta : \tilde{W}(x) \geq \tilde{W}^\star\}$ have potential greater than $\gamma(D) - 1 - \tilde{W}^\star$. Consider all the possible cases: let $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$ with $\tilde{W}(x) < \tilde{W}^\star$ and $y \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$. Lemma C.6 states that

$$\begin{aligned} &\gamma(x) \geq \gamma(D) - 1 - \tilde{W}(x) > \gamma(D) - 1 - \tilde{W}^\star, \\ &\gamma(y) \geq \gamma(D) - \tilde{W}(y) - \tilde{S}_+(y) \geq \gamma(D) - 1 - \tilde{S}_+(y) > \gamma(D) - 1 - \tilde{W}^\star, \\ &\gamma(D) > \gamma(D) - 1 - \tilde{W}^\star. \end{aligned}$$

This concludes the first case of Theorem 4.13.

[20] Actually, there may be states that are not be part of the recurrence class of P^ε. In this case we restrict ourselves to the study of the single recurrence class, which is a RPMP. This is without loss of generality, as the algorithm ends up a.s. in this recurrence class regardless of the starting state.

When $\mathcal{E} = \emptyset$, we want to show that $\mathcal{X}^\star \subset \mathcal{E}_\delta \cup \{x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta : \tilde{S}_+(x) + \tilde{W}(x) \geq \tilde{S}_-^\star + \tilde{W}^\star\}$ where $\tilde{S}_-^\star + \tilde{W}^\star = \max_{x\in\mathcal{C}} \tilde{S}_-(x) + \tilde{W}(x)$. Denoting $x^\star \in \mathcal{C}$ such that $\tilde{S}_-(x) + \tilde{W}(x) = \tilde{S}_-^\star + \tilde{W}^\star$, Lemma C.6 states that

$$\gamma(x^\star) \leq \gamma(D) - \tilde{S}_-(x) - \tilde{W}(x) = \gamma(D) - \tilde{S}_-^\star - \tilde{W}^\star.$$

It follows that states in $\mathcal{X}^\star$ must have a potential lower than $\gamma(D) - \tilde{S}_-^\star - \tilde{W}^\star$. To conclude, it is then sufficient to show that states outside of $\mathcal{E}_\delta \cup \{x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta : \tilde{S}_+(x) + \tilde{W}(x) \geq \tilde{S}_-^\star + \tilde{W}^\star\}$ have potential greater than $\gamma(D) - 1 - \tilde{W}^\star$. Let $y \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$ with $\tilde{S}_+(y) + \tilde{W}(y) < \tilde{S}_-^\star + \tilde{W}^\star$. Lemma C.6 states that

$$\gamma(y) \geq \gamma(D) - \tilde{S}_+(y) - \tilde{W}(y) > \gamma(D) - \tilde{S}_-^\star - \tilde{W}^\star,$$
$$\gamma(D) > \gamma(D) - \tilde{S}_-^\star - \tilde{W}^\star.$$

This concludes the second case of Theorem 4.13.

It remains to translate Theorem 4.13 to a convergence result regarding action profiles as stated by Theorem 4.14. This is done thanks to Lemmas 4.7 and 4.8. The first step is to identify what are the action profiles of states that can be part of $\mathcal{X}^\star$ according to Theorem 4.13.

Assume that $\mathcal{A} \neq \emptyset$. We have shown that the process converges a.s. to states in $\mathcal{A}$. In particular $\mathcal{A} \subset \mathcal{A}_\delta$ so these states have benchmark actions that are 2δ-admissible due to Lemma 4.7 (ii).

Assume that $\mathcal{A} = \emptyset$ and $\mathcal{E} \neq \emptyset$. We have shown that $\mathcal{X}^\star \subset \mathcal{A}_\delta \cup \{x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta : \tilde{W}(x) \geq \tilde{W}(x^\star)\}$ where $\tilde{W}(x^\star) = \max_{x\in\mathcal{E}} \tilde{W}(x)$. Note that while $x^\star$ has a maximal $\tilde{W}$ among SE, it is not guaranteed that its benchmark action profile has maximal welfare, not even that it is a SE. Denoting $\boldsymbol{a}^\star$ an action profile maximizing welfare among SE and $y \in \mathcal{E}$ its associated state (y is in $\mathcal{E}$ due to Lemma 4.7 (i)), we can however notice that $\tilde{W}(x^\star) \geq \tilde{W}(y) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$ due to Lemma 4.8 (ii).

A state $x \in \mathcal{A}_\delta$ has a 2δ-admissible benchmark action profile $\overline{\boldsymbol{a}}$ due to Lemma 4.7 (ii). A state $x \in \mathcal{E}_\delta \setminus \mathcal{A}_\delta$ with $\tilde{W}(x) \geq \tilde{W}^\star$ has a $(2\delta, 3\delta)$-SE benchmark action profile $\overline{\boldsymbol{a}}$ due to Lemma 4.7 (iii). Moreover, Lemma 4.8 (i) implies that $\tilde{W}(\overline{\boldsymbol{a}} + \delta) \geq \tilde{W}(x) \geq \tilde{W}(x^\star) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$. We conclude that when $\mathcal{E} \neq \emptyset$, the process will asymptotically visit action profiles $\boldsymbol{a}$ that are either 2δ-admissible or $(2\delta, 3\delta)$-SE with $\tilde{W}(\overline{\boldsymbol{a}}+\delta) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$.

Assume that $\mathcal{E} = \emptyset$. We have shown that $\mathcal{X}^\star \subset \mathcal{E}_\delta \cup \{x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta : \tilde{S}_+(x) + \tilde{W}(x) \geq \tilde{S}_-(x^\star) + \tilde{W}(x^\star)\}$ where $\tilde{S}_-(x^\star) + \tilde{W}(x^\star) = \max_{x\in\mathcal{C}} \tilde{S}_-(x) + \tilde{W}(x)$. Denoting $\boldsymbol{a}^\star$ an action profile maximizing $\tilde{S} + \tilde{W}$ among SE and $y \in \mathcal{C}$ its associated state, we have $\tilde{S}_-(x^\star) + \tilde{W}(x^\star) \geq \tilde{S}_-(y) + \tilde{W}(y) \geq \tilde{S}(\boldsymbol{a}^\star - \delta) + \tilde{W}(\boldsymbol{a}^\star - \delta)$ due to Lemma 4.8 (ii) and (iv).

A state $x \in \mathcal{E}_\delta$ has a $(2\delta, 3\delta)$-SE benchmark action profile $\overline{\boldsymbol{a}}$ due to Lemma 4.7 (iii). A state $x \in \mathcal{C}_\delta \setminus \mathcal{E}_\delta$ with $\tilde{S}_+(x)+\tilde{W}(x) \geq \tilde{S}_-(x^\star)+\tilde{W}(x^\star)$ satisfies $\tilde{S}(\overline{\boldsymbol{a}}+\delta)+\tilde{W}(\overline{\boldsymbol{a}}+\delta) \geq \tilde{S}_+(x)+\tilde{W}(x) \geq \tilde{S}_-(x^\star)+\tilde{W}(x^\star) \geq \tilde{S}(\boldsymbol{a}^\star-\delta)+\tilde{W}(\boldsymbol{a}^\star-\delta)$ due to Lemma 4.8 (i) and (iii). We conclude that when $\mathcal{E} \neq \emptyset$, the process will asymptotically visit action profiles $\boldsymbol{a}$ that are either $(2\delta, 3\delta)$-SE or satisfy $\tilde{S}(\overline{\boldsymbol{a}} + \delta) + \tilde{W}(\overline{\boldsymbol{a}} + \delta) \geq \tilde{S}(\boldsymbol{a}^\star - \delta) + \tilde{W}(\boldsymbol{a}^\star - \delta)$.

Now, depending on our knowledge of the existence of SE action profiles, the following holds:

- If there exist SE action profiles, the RITEL process spends most of the time in action profiles $\boldsymbol{a}$ that are 2δ-admissible or $(2\delta, 3\delta)$-SE with $\tilde{W}(\boldsymbol{a} + \delta) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$, where $\boldsymbol{a}^\star$ maximizes $\tilde{W}$ over SE action profiles.
- Else, the RITEL process spends most of the time in action profiles $\boldsymbol{a}$ that are $(2\delta, 3\delta)$-SE or with $\tilde{S}(\overline{\boldsymbol{a}} + \delta) + \tilde{W}(\overline{\boldsymbol{a}} + \delta) \geq \tilde{S}(\boldsymbol{a}^\star - \delta) + \tilde{W}(\boldsymbol{a}^\star - \delta)$, where $\boldsymbol{a}^\star$ maximizes $\tilde{S} + \tilde{W}$ over all action profiles.

Notice that the non-existence of SE action profiles does not guarantee that $\mathcal{E} = \emptyset$. Our statement remains true, as the action profiles described in the case of $\mathcal{E} \neq \emptyset$ are included in the ones described in the case $\mathcal{E} = \emptyset$. Similarly, one cannot know if $\mathcal{A} = \emptyset$ or not knowing only the expected utilities of action profiles, however the action profiles described in the case $\mathcal{A} \neq \emptyset$ are also included in the ones described in the case $\mathcal{A} = \emptyset$ (whether $\mathcal{E} \neq \emptyset$ or not). This concludes Theorem 4.14.

Let us discuss the special case where F and G are chosen of the form $F : u \mapsto \phi_F - \psi_F \cdot u$ and $G : (\overline{u}, u) \mapsto \phi_G - \psi_G \cdot (u - \overline{u})$, we want to show that:

- $\tilde{W}(\boldsymbol{a}+\delta) \geq \tilde{W}(\boldsymbol{a}^\star - \delta)$ implies $W(\boldsymbol{a}) \geq W(\boldsymbol{a}^\star) - 2\delta$.
- $\tilde{S}(\overline{\boldsymbol{a}}+\delta) + \tilde{W}(\overline{\boldsymbol{a}}+\delta) \geq \tilde{S}(\boldsymbol{a}^\star - \delta) + \tilde{W}(\boldsymbol{a}^\star - \delta)$ implies $\psi_F W(\boldsymbol{a}) - \psi_G S(\boldsymbol{a}) \geq \psi_F W(\boldsymbol{a}^\star) - \psi_G S(\boldsymbol{a}^\star) - 2\delta$.

Indeed, denote $\boldsymbol{\mu}$ the expected utilities of $\boldsymbol{a}$. Then

$$\tilde{W}(\boldsymbol{a}+\delta) = 1 - \sum_i F(\mu_i) = 1 - n\phi_F + \psi_F \sum_i \mu_i + n\psi_F \delta = 1 - n\phi_F + n\psi_F(W(\boldsymbol{a}) + \delta).$$

Similarly, $\tilde{W}(\boldsymbol{a}^\star - \delta) = 1 - n\phi_F + n\psi_F(W(\boldsymbol{a}^\star) - \delta)$. Now, Denote μ and μ' the expected utilities such that $S(\boldsymbol{a}) = \mu' - \mu$. Then

$$\tilde{S}(\overline{\boldsymbol{a}} + \delta) = G(\mu' - (\mu + \delta)) = G(S(\boldsymbol{a}) - \delta) = \phi_G - \psi_G(S(\boldsymbol{a}) - \delta).$$

Similarly, $\tilde{S}(\overline{\boldsymbol{a}}^\star - \delta) = \phi_G - \psi_G(S(\boldsymbol{a}^\star) + \delta)$.

The implications needed are obtained by plugging the above bounds in the inequalities, and noticing that C2 implies $\psi_F + \psi_G \leq 2$ to derive the error of 2δ in the second case.

References

[1] Bary S.R. Pradelski and H. Peyton Young. Learning efficient nash equilibria in distributed systems. *Games and Economic Behavior*, 75(2):882–897, 2012.

[2] Jason R. Marden, H. Peyton Young, and Lucy Y. Pao. Achieving pareto optimality through distributed learning. In *2012 IEEE 51st IEEE Conference on Decision and Control (CDC)*, pages 7419–7424, 2012.

[3] H. Peyton Young. The evolution of conventions. *Econometrica*, 61(1):57–84, 1993.

[4] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg, 2009.

[5] Z. Lin and Z. Bai. *Probability Inequalities*. Springer Berlin Heidelberg, 2011.

[6] P. Borjesson and C.-E. Sundberg. Simple approximations of the error function q(x) for communications applications. *IEEE Transactions on Communications*, 27(3):639–643, 1979.

[7] Alain Trouvé. Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms. *Annales de l'I.H.P. Probabilités et statistiques*, 32(3):299–348, 1996.

[8] Olivier Catoni. Rough Large Deviation Estimates for Simulated Annealing: Application to Exponential Schedules. *The Annals of Probability*, 20(3):1109 – 1146, 1992.