Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Polynomial-Time Linear-Swap Regret Minimization in Imperfect-Information Sequential Games

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΑΡΙΛΑΟΣ ΠΙΠΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Αθήνα, Ιούλιος 2023

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών
και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Polynomial-Time Linear-Swap Regret Minimization in Imperfect-Information Sequential Games

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΧΑΡΙΛΑΟΣ ΠΙΠΗΣ

Επιβλέπων : Δημήτριος Φωτάκης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 18η Ιουλίου 2023.

Δημήτριος Φωτάκης Καθηγητής Ε.Μ.Π.	Αριστείδης Παγουρτζής Καθηγητής Ε.Μ.Π.	Χρήστος Τζάμος Αν. Καθηγητής Ε.Κ.Π.Α.

Αθήνα, Ιούλιος 2023

Χαρίλαος Πίπης

Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Οι no-regret learners επιδιώκουν να ελαχιστοποιήσουν τη διαφορά μεταξύ της απώλειας που συσσώρευσαν μέσω των ενεργειών που παίζουν και της απώλειας που θα είχαν συσσωρεύσει υποθετικά εάν είχαν τροποποιήσει συνεπώς τη συμπεριφορά τους σύμφωνα με μια συνάρτηση μετασχηματισμού στρατηγικής. Το μέγεθος του συνόλου των μετασχηματισμών που λαμβάνονται υπόψη από τον learner καθορίζει μια φυσική έννοια ορθολογισμού (rationality). Καθώς το σύνολο των μετασχηματισμών που κάθε learner λαμβάνει υπόψη μεγαλώνει, οι στρατηγικές που παίζονται από τους learners ανακτούν όλο και πιο περίπλοκες παιγνιοθεωρητικές ισορροπίες, συμπεριλαμβανομένων των correlated equilibria σε παιχνίδια κανονικής μορφής (normal-form games) και extensive-form correlated equilibria σε παιχνίδια εκτεταμένης μορφής (extensive-form games). Στην ακραία περίπτωση, ένας no-swap-regret αλγόριθμος είναι αυτός που ελαχιστοποιεί την μετάνοια έναντι του συνόλου όλων των συναρτήσεων από το σύνολο των στρατηγικών προς το ίδιο το σύνολο των στρατηγικών. Ενώ είναι γνωστό ότι η συνθήκη για το no-swap-regret μπορεί να επιτευχθεί αποδοτικά σε μη ακολουθιακά (normal-form) παιχνίδια, η κατανόηση του ποια είναι η ισχυρότερη έννοια του rationality που μπορεί να επιτευχθεί αποδοτικά στη χειρότερη περίπτωση σε ακολουθιακά (extensive-form) παιχνίδια αποτελεί ένα ανεπίλυτο πρόβλημα. Σε αυτήν την εργασία παρέχουμε ένα θετικό αποτέλεσμα, δείχνοντας ότι είναι δυνατό, σε οποιοδήποτε ακολουθιακό παιχνίδι, χρησιμοποιώντας επαναλήψεις επαναλήψεις πολυωνυμικού χρόνου (σε σχέση με το μέγεθος του δένδρου του παιχνιδιού) να επιτύχουμε υπογραμμική μετάνοια ως προς όλους τους γραμμικούς μετασχηματισμούς του χώρου μικτών στρατηγικών, μια έννοια που ονομάζεται no-linear-swap regret. Αυτή η έννοια του εκ των υστέρων ορθολογισμού είναι τόσο ισχυρή όσο το no-swap-regret σε μη ακολουθιακά παιχνίδια και ισχυρότερη από την έννοια no-trigger-regret σε ακολουθιακά παιχνίδια – αποδεικνύοντας έτσι την ύπαρξη ενός υποσυνόλου από εκτετατμένα correlated equilibria ανθεκτικά σε γραμμικές αποκλίσεις, τις οποίες ονομάζουμε linear-deviation correlated equilibria, που μπορούν να προσεγγιστούν αποδοτικά.

Λέξεις κλειδιά

άμεση μάθηση, αλγοριθμική θεωρία παιγνίων, παίγνια εκτεταμένης μορφής, correlated equilibria, swap regret, γραμμικές αποκλίσεις

Abstract

No-regret learners seek to minimize the difference between the loss they cumulated through the actions they played, and the loss they would have cumulated in hindsight had they consistently modified their behavior according to some strategy transformation function. The size of the set of transformations considered by the learner determines a natural notion of rationality. As the set of transformations each learner considers grows, the strategies played by the learners recover more complex game-theoretic equilibria, including correlated equilibria in normal-form games and extensive-form correlated equilibria in extensive-form games. At the extreme, a *no-swap-regret* agent is one that minimizes regret against the set of *all* functions from the set of strategies to itself. While it is known that the no-swap-regret condition can be attained efficiently in nonsequential (normal-form) games, understanding what is the strongest notion of rationality that can be attained *efficiently* in the worst case in *sequential* (extensive-form) games is a longstanding open problem. In this paper we provide a positive result, by showing that it is possible, in any sequential game, to retain polynomial-time (in the game tree size) iterations while achieving sublinear regret with respect to *all linear transformations* of the mixed strategy space, a notion called *no-linear-swap regret*. This notion of hindsight rationality is as strong as no-swap-regret in nonsequential games, and stronger than no-trigger-regret in sequential games—thereby proving the existence of a subset of extensive-form correlated equilibria robust to linear deviations, which we call *linear-deviation correlated equilibria*, that can be approached efficiently.

Key words

online learning, algorithmic game theory, extensive form games, correlated equilibrium, swap regret, linear swap regret

Ευχαριστίες

First and foremost, I would like to thank Gabriele Farina for proposing all the concrete directions for my thesis project and for being an exceptional supervisor and collaborator. I have learned a lot from him and truly appreciate him for always being available when needed and for carefully listening to my ideas and concerns at all times. It has been a pleasure working with him, and I look forward to continuing our collaboration during my PhD.

I would like to express my gratitude to the members of my diploma thesis committee, for much more than serving in the committee. I sincerely thank Dimitris Fotakis for inspiring my interest in algorithms during my participation in informatics competitions in high school, for then mentoring me and supporting me throughout the whole duration of my studies, and for giving me the opportunity to embark on my journey into TCS research. I feel truly grateful and fortunate that I had the chance to be mentored by him and I am certain that my academic path would not have been the same otherwise. I would also like to thank Christos Tzamos for playing an important role in shaping my thinking style through his active participation in my first research project, and for all his help during my PhD application period. Finally, I would like to thank Aris Pagourtzis for helping foster a welcoming environment in CoReLab, which made it feel like a second home to me.

Furthermore, I am truly indebted to my collaborators and friends, Evangelia Gergatsouli and Miltiadis Stouras. We have spent countless hours discussing and thinking about graph connectivity in my first research project, and their support has been invaluable to me. Additionally, I would like to extend special thanks to my friends Alkis Kalavasis, Vardis Kandiros, and Argyris Mouzakis. Each one of them has offered and taught me a lot in their own ways, and I am incredibly grateful for their unwavering support during my PhD applications and beyond.

Another significant aspect of my academic journey has been my involvement in the Greek National Competition in Informatics. In this regard, I am immensely grateful to Nikos Papaspyrou for his invaluable contribution to shaping the Greek competitive programming community and inspiring us to love algorithmic problem solving. I would also like to express my gratitude to all the other participants and friends I have made through these competitions, as the constant exchange of ideas and shared aspirations has been incredibly enriching. In particular, I extend my special thanks to Panagiotis Kostopanagiotis, Dionysis Zindros, and Evangelos Kipouridis for believing in me and offering their mentorship during my first steps into the beautiful world of algorithms and programming.

I would also like to thank my friends from all these years that have enriched my experiences and memories in diverse ways. Finally, I would like to express my gratitude to my family. I feel privileged to have been raised in an interesting environment that has taught me, among other things, the importance of independent thinking. I am deeply grateful that my parents and my brother have always been there for me, both in easy and difficult times.

Looking ahead, I am genuinely excited about the academic environment I am about to enter for my PhD, and I eagerly anticipate the opportunities that lie ahead.

Χαρίλαος Πίπης,

Αθήνα, 18η Ιουλίου 2023

Η εργασία αυτή είναι επίσης διαθέσιμη ως Τεχνική Αναφορά CSD-SW-TR-42-17, Εθνικό Μετσόβιο Πολυτεχνείο, Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών, Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών, Εργαστήριο Τεχνολογίας Λογισμικού, Ιούλιος 2023.

URL: `http://www.softlab.ntua.gr/techrep/`
FTP: `ftp://ftp.softlab.ntua.gr/pub/techrep/`

Περιεχόμενα

Εκτεταμένη Ελληνική Περίληψη

Η μέθοδος της ελαχιστοποίησης του regret παρέχει αλγορίθμους που οι παίκτες μπορούν να χρησιμοποιήσουν για να βελτιώσουν σταδιακά τις στρατηγικές τους σε ένα επαναλαμβανόμενο παίγνιο, επιτρέποντας την εκμάθηση ισχυρών στρατηγικών ακόμη και όταν αντιμετωπίζουν άγνωστους αντίπαλους. Μία από τις ελκυστικές ιδιότητες των αλγορίθμων μάθησης no-regret είναι ότι είναι αποσυζευγμένοι, πράγμα που σημαίνει ότι κάθε παίκτης βελτιώνει τη στρατηγική του με βάση τη δική του συνάρτηση αποπληρωμής, και τις στρατηγικές των άλλων παικτών, αλλά όχι τις συναρτήσεις αποπληρωμής των άλλων παικτών. Παρ' όλα αυτά, παρά την ασύζευκτη φύση τους και την εστίασή τους στην τοπική βελτιστοποίηση της ωφέλειας κάθε παίκτη, είναι ένα από τα πιο διάσημα αποτελέσματα στη θεωρία της μάθησης σε παίγνια ότι σε πολλές περιπτώσεις, όταν όλοι οι παίκτες μαθαίνουν χρησιμοποιώντας αυτούς τους αλγορίθμους, η εμπειρική πορεία του παιγνίου ανακτά τις κατάλληλες έννοιες της ισορροπίας – μια καθολική έννοια παιγνιοθεωρητικής βελτιστότητας. Οι στρατηγικές που κατασκευάζονται μέσω αλγορίθμων μάθησης no-regret (ή προσεγγίσεων αυτών) έχουν αποτελέσει βασικά στοιχεία για την κατασκευή ανθρώπινου επιπέδου και ακόμη και υπεράνθρωπων πρακτόρων TN σε μια ποικιλία ανταγωνιστικών παιγνίων, συμπεριλαμβανομένων του πόκερ [Moravčík et al., 2017, Brown and Sandholm, 2018, 2019], του Stratego [Perolat et al., 2022] και του Diplomacy [Bakhtin et al., 2023].

Στην ελαχιστοποίηση του regret, κάθε εκπαιδευόμενος πράκτορας προσπαθεί να ελαχιστοποιήσει τη διαφορά μεταξύ της απώλειας (αντίθετο της ανταμοιβής) που συσσώρευσε μέσω των ενεργειών που έπαιξε, και της απώλειας που θα είχε συσσωρεύσει εκ των υστέρων αν τροποποιούσε με συνέπεια τη συμπεριφορά του σύμφωνα με κάποια συνάρτηση μετασχηματισμού στρατηγικής. Το μέγεθος του συνόλου των συναρτήσεων μετασχηματισμού που εξετάζει ο εκπαιδευόμενος πράκτορας καθορίζει μια φυσική έννοια του ορθολογισμού του πράκτορα. Ήδη όταν οι πράκτορες επιδιώκουν να μάθουν στρατηγικές που συσσωρεύουν χαμηλή μετάνοια μόνο έναντι σταθερών μετασχηματισμών στρατηγικής–μια έννοια της μετάνοιας που ονομάζεται external regret–το μέσο παίξιμο των πρακτόρων συγκλίνει σε μια ισορροπία Nash σε παίγνια σταθερού αθροίσματος δύο παικτών και σε ένα coarse correlated equilibrium σε παίγνια πολλαπλών παικτών γενικού αθροίσματος. Καθώς τα σύνολα των μετασχηματισμών που εξετάζει κάθε πράκτορας αυξάνονται, μπορούν να επιτευχθούν πιο σύνθετες ισορροπίες, συμπεριλαμβανομένων συσχετισμένων ισορροπιών σε παίγνια κανονικής μορφής (Foster and Vohra [1997], Fudenberg and Levine [1995], Fudenberg and Levine [1995, 1999], Hart and Mas-Colell [2000, 2001], βλέπε επίσης τη μονογραφία των Fudenberg and Levine [1998]) και συσχετισμένων ισορροπιών εκτεταμένης μορφής σε παίγνια εκτεταμένης μορφής [Farina et al., 2022b]. Στο άκρο, ένας μέγιστα εκ των υστέρων ορθολογικός πράκτορας είναι εκείνος που ελαχιστοποιεί την μετάνοια έναντι του συνόλου όλων των συναρτήσεων από τον χώρο στρατηγικών προς τον εαυτό του (γνωστός και ως swap regret). Ενώ είναι γνωστό ότι ο μέγιστος εκ των υστέρων ορθολογισμός μπορεί να επιτευχθεί αποτελεσματικά σε μη διαδοχικά (κανονικής μορφής) παίγνια [Stoltz and Lugosi, 2007, Blum and Mansour, 2007], αποτελεί μεγάλο ανοικτό πρόβλημα να προσδιοριστεί αν το ίδιο ισχύει και για τα διαδοχικά (δηλ. εκτεταμένης μορφής) παίγνια, και γενικότερα ποια είναι η ισχυρότερη έννοια της ορθολογικότητας που μπορεί να επιτευχθεί αποτελεσματικά στη χειρότερη περίπτωση στο τελευταίο περιβάλλον.

Στην παρούσα εργασία, παρέχουμε ένα θετικό αποτέλεσμα προς αυτή την κατεύθυνση, δείχνοντας ότι ο εκ των υστέρων ορθολογισμός μπορεί να επιτευχθεί αποτελεσματικά σε γενικά παίγνια εκτεταμένης μορφής με ατελή πληροφόρηση, όταν κάποιος περιορίζεται στο σύνολο *όλων των γραμμικών μετασχηματισμών* του χώρου των μικτών στρατηγικών–μια έννοια που ονομάζεται *linear-swap*

regret, και που συμπίπτει με το swap regret στα παίγνια κανονικής μορφής. Προκειμένου να τεκμηριώσουμε το αποτέλεσμα, εισάγουμε διάφορα ενδιάμεσα αποτελέσματα που σχετίζονται με τη γεωμετρία των στρατηγικών ακολουθιακής μορφής σε παίγνια εκτεταμένης μορφής. Συγκεκριμένα, ένα κρίσιμο αποτέλεσμα δίνεται στο Θεώρημα 4.1.1, το οποίο δείχνει ότι το σύνολο των γραμμικών συναρτήσεων $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ από το σύνολο στρατηγικών ακολουθιακής μορφής $\mathcal{Q}$ ενός παίκτη σε ένα παίγνιο εκτεταμένης μορφής σε ένα γενικό κυρτό πολύτοπο $\mathcal{P}$ μπορεί να αποδοθεί χρησιμοποιώντας μόνο πολυωνυμικά πολλούς γραμμικούς περιορισμούς στο μέγεθος του δέντρου του παιγνίου και στον αριθμό των γραμμικών περιορισμών που ορίζουν το $\mathcal{P}$. Εφαρμόζοντας το αποτέλεσμα στην ειδική περίπτωση $\mathcal{P} = \mathcal{Q}$, είμαστε σε θέση να συμπεράνουμε ότι το πολύτοπο των γραμμικών μετασχηματισμών $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ από το σύνολο στρατηγικών ακολουθιακής μορφής στον εαυτό του μπορεί να καταγραφεί με πολυωνυμικά πολλούς γραμμικούς περιορισμούς στο μέγεθος του δέντρου παιγνίων και η νόρμα κάθε στοιχείου είναι πολυωνυμικά φραγμένη. Ο πολυωνυμικός χαρακτηρισμός και το φράγμα για το $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ χρησιμοποιείται σε συνδυασμό με μια ιδέα του Gordon et al. [2008] για την κατασκευή ενός linear-swap regret ελαχιστοποιητή για το σύνολο στρατηγικών $\mathcal{Q}$ ξεκινώντας από δύο αρχές: i) έναν no-external-regret αλγόριθμο για το σύνολο των μετασχηματισμών $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, και ii) έναν αλγόριθμο για τον υπολογισμό μιας στρατηγικής σταθερού σημείου για κάθε μετασχηματισμό στο $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$. Και στις δύο περιπτώσεις, η πολυωνυμική αναπαράσταση του $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ που καθιερώθηκε μέσω του Θεωρήματος 4.1.1 διαδραματίζει θεμελιώδη ρόλο. Επιτρέπει, αφενός, την ικανοποίηση της απαίτησης ii) χρησιμοποιώντας γραμμικό προγραμματισμό. Αφετέρου, μας επιτρέπει να κατασκευάσουμε έναν no-external regret αλγόριθμο που παράγει μετασχηματισμούς στο $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ με επαναλήψεις πολυωνυμικού χρόνου, αξιοποιώντας τις γνωστές ιδιότητες του online projected gradient descent, αξιοποιώντας την αποδοτικότητα της προβολής σε πολυωνυμικά αναπαριστώμενα πολύτοπα.

Τέλος, στο τελευταίο τμήμα της διατριβής στρέφουμε την προσοχή μας μακριά από την εκ των υστέρων ορθολογικότητα και επικεντρωνόμαστε στις ιδιότητες των ισορροπιών που ανακτούν οι no-linear-swap regret δυναμικές μας σε παίγνια εκτεταμένης μορφής. Το μέσο παίξιμο των no-linear-swap regret παικτών συγκλίνει σε ένα σύνολο ισορροπιών που ονομάζουμε *linear-deviation correlated equilibria (LCEs)*. Οι LCEs αποτελούν ένα υπερσύνολο των συσχετισμένων ισορροπιών και ένα υποσύνολο των συσχετισμένων ισορροπιών εκτεταμένης μορφής σε παίγνια εκτεταμένης μορφής. Στο Κεφάλαιο 6 δείχνουμε ότι αυτοί οι εγκλεισμοί είναι γενικά αυστηροί και παρέχουμε πρόσθετα αποτελέσματα σχετικά με την πολυπλοκότητα του υπολογισμού ενός LCE που μεγιστοποιεί την κοινωνική ευημερία.

Σχετικές εργασίες Όπως αναφέρθηκε στην εισαγωγή, η ύπαρξη ασύζευκτων no-regret δυναμικών που οδηγούν σε συσχετισμένη ισορροπία (CE) σε παίγνια κανονικής μορφής για πολλούς παίκτες είναι ένα περίφημο αποτέλεσμα που χρονολογείται τουλάχιστον από την εργασία των Foster and Vohra [1997]. Αυτή η εργασία ενέπνευσε τους ερευνητές να αναζητήσουν διαδικασίες ασύζευκτης μάθησης και σε άλλα περιβάλλοντα. Για παράδειγμα, οι Stoltz and Lugosi [2007] μελετούν τις δυναμικές μάθησης που οδηγούν σε CE σε παίγνια με άπειρο (αλλά συμπαγές) σύνολο δράσεων, ενώ οι Kakade et al. [2003] εστιάζουν σε γραφικά παίγνια. Τα πιο πρόσφατα χρόνια, μια αυξανόμενη προσπάθεια έχει καταβληθεί προς την κατεύθυνση της κατανόησης των σχέσεων μεταξύ της no-regret μάθησης και των ισορροπιών σε παίγνια εκτεταμένης μορφής με ατελή πληροφόρηση, το περιβάλλον στο οποίο εστιάζουμε. Τα παίγνια εκτεταμένης μορφής παρουσιάζουν πρόσθετες προκλήσεις σε σύγκριση με τα παίγνια κανονικής μορφής, λόγω της ακολουθιακής τους φύσης και της παρουσίας ατελούς πληροφόρησης. Ενώ είναι γνωστά αποτελεσματικά δυναμικά μάθησης no-regret για παίγνια εκτεταμένης μορφής (συμπεριλαμβανομένου του δημοφιλούς αλγορίθμου CFR [Zinkevich et al., 2008]), μέχρι σήμερα δεν γνωρίζουμε πολλά για το no-swap-regret και την πολυπλοκότητα της μάθησης CE σε παίγνια εκτεταμένης μορφής.

Η πλησιέστερη έννοια στην CE που είναι γνωστό ότι είναι αποτελεσματικά υπολογίσιμη σε παίγνια εκτεταμένης μορφής είναι η *extensive-form correlated equilibrium* (EFCE), που εισήχθη από τους von Stengel and Forges [2008]. Το ερώτημα κατά πόσο το σύνολο των EFCE θα μπορούσε να προσεγγιστεί μέσω μη συζευγμένων δυναμικών no-regret με επαναλήψεις πολυωνυμικού χρόνου στο

μέγεθος των παιγνίων εκτεταμένης μορφής απαντήθηκε θετικά πρόσφατα [Farina et al., 2022b, Celli et al., 2020]. Συγκεκριμένα, οι Farina et al. [2022b] δείχνουν ότι η EFCE προκύπτει από το μέσο παίξιμο των αλγορίθμων no-trigger-regret, όπου οι αποκλίσεις trigger είναι ένα συγκεκριμένο υποσύνολο γραμμικών μετασχηματισμών του πολυτόπου των στρατηγικών ακολουθιακής μορφής $\mathcal{Q}$ κάθε παίκτη. Δεδομένου ότι η παρούσα εργασία επικεντρώνεται σε δυναμικές μάθησης που εγγυώνται υπογραμμικό regret σε σχέση με *οποιονδήποτε* γραμμικό μετασχηματισμό του $\mathcal{Q}$, προκύπτει αμέσως ότι οι δυναμικές που παρουσιάζονται σε αυτή την εργασία ανακτούν την EFCE ως ειδική περίπτωση.

Η έννοια της ελαχιστοποίησης του linear-swap-regret έχει εξεταστεί στο παρελθόν στο πλαίσιο των *Μπεϋζιανών* παιγνίων. Οι Mansour et al. [2022] μελετούν ένα περιβάλλον όπου ένας no-regret *learner* ανταγωνίζεται σε ένα Μπεϋζιανό παίγνιο δύο παικτών με έναν ορθολογικό *maximizer* ωφέλειας, δηλαδή έναν αυστηρά ισχυρότερο αντίπαλο από τον learnner. Υπό αυτό το πλαίσιο, μπορεί να αποδειχθεί ότι σε κάθε γύρο ο βελτιστοποιητής είναι εγγυημένο ότι θα επιτύχει τουλάχιστον την Μπεϋζιανή αξία Stackelberg του παιγνίου. Στη συνέχεια, προχωρούν στην απόδειξη ότι η ελαχιστοποίηση του *linear-swap regret* είναι απαραίτητη αν θέλουμε να περιορίσουμε την απόδοση του βελτιστοποιητή στην τιμή Stackelberg, ενώ η ελαχιστοποίηση του polytope-swap regret (μια γενίκευση του swap regret για Bayesian παιχνίδια, και αυστηρά ισχυρότερη από το linear-swap) αρκεί για να περιορίσουμε την απόδοση του βελτιστοποιητή. Ως εκ τούτου, τα αποτελέσματα αυτά αναδεικνύουν τη σημασία της ανάπτυξης αλγορίθμων μάθησης υπό ισχυρότερες έννοιες *ορθολογισμού*, όπως είναι ο στόχος μας στην παρούσα διπλωματική εργασία. Επιπλέον, αυτά τα αποτελέσματα παρέχουν αποδείξεις ότι η κατασκευή ενός no-linear-swap regret learner, όπως είναι ο στόχος μας εδώ, μπορεί να επιφέρει οφέλη σε σύγκριση με άλλους λιγότερο ορθολογικούς learners. Σε μια παράλληλη εργασία, ο Fujii [2023] ορίζει την έννοια του *untruthful swap regret* για Μπεϋζιανά παίγνια και αποδεικνύει ότι, για τα Μπεϋζιανά παίγνια, είναι ισοδύναμη με το linear-swap regret που μας απασχολεί στην εργασία μας.

Τα Μπεϋζιανά παίγνια μπορούν να θεωρηθούν ως ειδική περίπτωση των παιγνίων εκτεταμένης μορφής, όπου ένας κόμβος τύχης επιλέγει αρχικά έναν από τους πιθανούς τύπους Θ για κάθε παίκτη. Έτσι, ο αλγόριθμός μας που ελαχιστοποιεί το linear-swap regret στα παίγνια εκτεταμένης μορφής ελαχιστοποιεί επίσης το linear-swap regret στα Μπεϋζιανά παίγνια. Ωστόσο, παρατηρούμε ότι η έκφρασή μας για το regret εξαρτάται πολυωνυμικά από τον αριθμό των τύπων $|\Theta|$ των παικτών καθώς αποτελούν μέρος της αναπαράστασης του δέντρου του παιγνίου, ενώ ο Fujii [2023] έχει επινοήσει έναν αλγόριθμο για Μπεϋζιανά παίγνια, του οποίου το regret εξαρτάται μόνο από το $\log |\Theta|$.

Τέλος, αναφέρουμε επίσης ότι μη γραμμικές αποκλίσεις έχουν διερευνηθεί σε παίγνια εκτεταμένης μορφής, αν και δεν γνωρίζουμε αξιοσημείωτα μεγάλα σύνολα για τα οποία μπορεί να επινοηθούν no-regret δυναμικές σε πολυωνυμικό χρόνο. Συγκεκριμένα, επισημαίνουμε την εργασία των Morrill et al. [2021], η οποία ορίζει την έννοια των "συμπεριφορικών αποκλίσεων". Αυτές οι αποκλίσεις είναι μη γραμμικές σε σχέση με την αναπαράσταση των στρατηγικών σε παίγνια εκτεταμένης μορφής. Οι συγγραφείς κατηγοριοποιούν διάφορους γνωστούς ή νέους τύπους περιορισμένων συμπεριφορικών αποκλίσεων σε ένα Τοπίο Αποκλίσεων που αναδεικνύει τις μεταξύ τους σχέσεις. Παρόλο που τόσο οι γραμμικές αποκλίσεις, που εξετάζουμε σε αυτή την εργασία, όσο και οι συμπεριφορικές αποκλίσεις φαίνεται να αποτελούν πλούσια μέτρα ορθολογισμού, καμία από αυτές δεν περιέχει την άλλη και έτσι, οι γραμμικές αποκλίσεις δεν ταιριάζουν στο Τοπίο Αποκλίσεων των Morrill et al. [2021] (βλ. επίσης την Παρατήρηση 6.1.2).

Εισαγωγικά

Υπενθυμίζουμε το τυπικό μοντέλο των παιγνίων εκτεταμένης μορφής, καθώς και το πλαίσιο της μάθησης στα παίγνια.

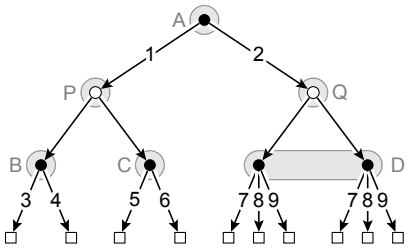
Παίγνια εκτεταμένης μορφής

Ενώ τα παίγνια κανονικής μορφής (NFGs) αντιστοιχούν σε μη διαδοχικές αλληλεπιδράσεις, όπως το πέτρα-ψαλίδι-χαρτί, όπου οι παίκτες επιλέγουν ταυτόχρονα μια ενέργεια και στη συνέχεια λαμβάνουν μια πληρωμή με βάση το τι επέλεξαν οι άλλοι, τα παίγνια εκτεταμένης μορφής (EFG) μοντελοποιούν παίγνια που παίζονται σε ένα δέντρο παιγνίων. Περιλαμβάνουν τόσο τις διαδοχικές όσο και τις ταυτόχρονες κινήσεις, καθώς και την ιδιωτική πληροφορία και επομένως αποτελούν ένα πολύ γενικό και εκφραστικό μοντέλο παιγνίων, που αποτυπώνει το σκάκι, το Go, το πόκερ, τις διαδοχικές δημοπρασίες και πολλές άλλες περιπτώσεις. Τώρα υπενθυμίζουμε βασικές ιδιότητες και συμβολισμούς για τα EFGs.

Δέντρο παιγνίων Σε ένα παίγνιο εκτεταμένης μορφής n παικτών, κάθε κόμβος στο δέντρο παιγνίων αντιστοιχίζεται με ακριβώς έναν παίκτη από το σύνολο $\{1, \ldots, n\} \cup \{c\}$, όπου ο ειδικός παίκτης c–που ονομάζεται παίκτης *τυχαιότητας*—χρησιμοποιείται για τη μοντελοποίηση τυχαίων στοχαστικών αποτελεσμάτων, όπως η ρίψη ενός ζαριού ή το τράβηγμα καρτών από μια τράπουλα. Οι ακμές που φεύγουν από έναν κόμβο αντιπροσωπεύουν τις ενέργειες που μπορεί να κάνει ένας παίκτης σε αυτόν τον κόμβο. Για τη μοντελοποίηση ιδιωτικών πληροφοριών, το δέντρο του παιγνίου συμπληρώνεται με μια κατάτμηση πληροφοριών, που ορίζεται ως μια κατάτμηση των κόμβων σε σύνολα που ονομάζονται σύνολα πληροφοριών. Κάθε κόμβος ανήκει ακριβώς σε ένα σύνολο πληροφοριών και κάθε σύνολο πληροφοριών είναι ένα μη κενό σύνολο κόμβων του δέντρου για τον ίδιο παίκτη i. Ένα σύνολο πληροφοριών για τον παίκτη i υποδηλώνει μια συλλογή κόμβων μεταξύ των οποίων ο παίκτης i δεν μπορεί να κάνει διάκριση, με βάση όσα έχει παρατηρήσει μέχρι στιγμής. (Σημειώνουμε ότι όλοι οι κόμβοι σε ένα ίδιο σύνολο πληροφοριών πρέπει να έχουν το ίδιο σύνολο διαθέσιμων ενεργειών, αλλιώς ο παίκτης θα διέκρινε τους κόμβους). Το σύνολο όλων των συνόλων πληροφοριών του παίκτη i συμβολίζεται ως $\mathcal{J}_i$. Στην παρούσα εργασία, θα εξετάσουμε μόνο τα παίγνια με *τέλεια ανάκληση*, δηλαδή τα παίγνια στα οποία τα σύνολα πληροφοριών είναι διατεταγμένα σύμφωνα με το γεγονός ότι κανένας παίκτης δεν ξεχνά τι γνώριζε νωρίτερα.

Στρατηγικές ακολουθιακής μορφής Δεδομένου ότι οι κόμβοι που ανήκουν στο ίδιο σύνολο πληροφοριών για έναν παίκτη είναι δυσδιάκριτοι για τον παίκτη αυτό, ο παίκτης πρέπει να παίζει την ίδια στρατηγική σε κάθε έναν από τους κόμβους. Συνεπώς, μια στρατηγική για έναν παίκτη είναι ακριβώς μια απεικόνιση από ένα *σύνολο πληροφοριών* σε μια κατανομή πάνω στις ενέργειες. Με άλλα λόγια, είναι τα σύνολα πληροφοριών και όχι οι κόμβοι του δέντρου παιγνίων που αποτυπώνουν τα σημεία απόφασης του παίκτη. Μπορούμε τότε να αναπαραστήσουμε μια στρατηγική για έναν γενικό παίκτη i ως ένα διάνυσμα που δεικτοδοτείται από κάθε έγκυρο ζεύγος συνόλου πληροφοριών-δράσης (j, a). Κάθε τέτοιο έγκυρο ζεύγος ονομάζεται *ακολουθία* του παίκτη- το σύνολο όλων των ακολουθιών συμβολίζεται ως $\Sigma_i := \{(j, a) : j \in \mathcal{J}_i, a \in \mathcal{A}_j\} \cup \{\varnothing\}$, όπου το ειδικό στοιχείο $\varnothing$ ονομάζεται *κενή ακολουθία*. Δεδομένου ενός συνόλου πληροφοριών $j \in \mathcal{J}_i$, συμβολίζουμε με p_j την ακολουθία γονέα του j, η οποία ορίζεται ως το τελευταίο ζεύγος $(j, a) \in \Sigma_i$ που συναντάμε στο μονοπάτι από τη ρίζα προς οποιονδήποτε κόμβο $v \in j$- αν δεν υπάρχει τέτοιο ζεύγος, υποθέτουμε $p_j = \varnothing$. Τέλος, συμβολίζουμε με $\mathcal{C}_\sigma$ τα παιδιά της ακολουθίας $\sigma \in \Sigma_i$, που ορίζονται ως τα σύνολα πληροφορίας $j \in \mathcal{J}_i$ για τα οποία $p_j = \sigma$. Οι ακολουθίες σ για τις οποίες $\mathcal{C}_\sigma$ είναι ένα κενό σύνολο ονομάζονται *τερματικές*- το σύνολο όλων των τερματικών ακολουθιών συμβολίζεται ως $\Sigma_i^\perp$.

Example 0.0.1. *Θεωρήστε τη δενδροειδή διαδικασία λήψης αποφάσεων που αντιμετωπίζει ο παίκτης 1 στο μικρό παίγνιο του Σχήματος 0.1 (αριστερά). Η διαδικασία απόφασης έχει τέσσερις κόμβους απόφασης $\mathcal{J}_1 = \{A, B, C, D\}$ και εννέα ακολουθίες, συμπεριλαμβανομένης της κενής ακολουθίας $\varnothing$. Για τον κόμβο απόφασης D, η γονική ακολουθία είναι $p_D = A2$, για B και C είναι $p_B = A1$, για A είναι η κενή ακολουθία $p_A = \varnothing$.*



Sequence-form constraints:

$$\begin{cases} \boldsymbol{x}[\varnothing] = 1, \\ \boldsymbol{x}[\mathsf{A1}] + \boldsymbol{x}[\mathsf{A2}] = \boldsymbol{x}[\varnothing], \\ \boldsymbol{x}[\mathsf{B3}] + \boldsymbol{x}[\mathsf{B4}] = \boldsymbol{x}[\mathsf{A1}], \\ \boldsymbol{x}[\mathsf{C5}] + \boldsymbol{x}[\mathsf{C6}] = \boldsymbol{x}[\mathsf{A1}], \\ \boldsymbol{x}[\mathsf{D7}] + \boldsymbol{x}[\mathsf{D8}] + \boldsymbol{x}[\mathsf{D9}] = \boldsymbol{x}[\mathsf{A2}]. \end{cases}$$

Σχήμα 0.1: (Αριστερά) Δενδροειδής διαδικασία λήψης αποφάσεων που εξετάζεται στο παράδειγμα. Οι μαύροι στρογγυλοί κόμβοι ανήκουν στον παίκτη 1, οι λευκοί στρογγυλοί κόμβοι στον παίκτη 2. Οι τετράγωνοι λευκοί κόμβοι είναι τερματικοί κόμβοι στο δέντρο του παιγνίου, οι πληρωμές παραλείπονται. Οι γκρίζες σακούλες υποδηλώνουν σύνολα πληροφοριών. (Δεξιά) Οι περιορισμοί που ορίζουν το ακολουθιακής μορφής πολυτόπιο $\mathcal{Q}_1$ για τον παίκτη 1 (εκτός από τη μη αρνητικότητα).

Ένα *μειωμένο πλάνο κανονικής μορφής* για τον παίκτη i αναπαριστά μια ντετερμινιστική στρατηγική για τον παίκτη ως ένα διάνυσμα $\boldsymbol{x} \in \{0,1\}^{\Sigma_i}$ όπου η τιμή που αντιστοιχεί στη γενική ακολουθία $\boldsymbol{x}[ja]$ είναι ίση με 1 εάν ο παίκτης παίζει την ενέργεια a στο σύνολο πληροφοριών $j \in \mathcal{J}_i$. Τα σύνολα πληροφοριών που δεν μπορούν να προσεγγιστούν με βάση τη στρατηγική δεν έχουν καμία επιλεγμένη ενέργεια. Μια κρίσιμη ιδιότητα της αναπαράστασης μειωμένων πλάνων κανονικής μορφής των ντετερμινιστικών στρατηγικών είναι το γεγονός ότι η ωφέλεια κάθε παίκτη είναι μια πολυγραμμική συνάρτηση στο προφίλ των μειωμένων πλάνων κανονικής μορφής που παίζουν οι παίκτες. Το σύνολο όλων των μειωμένων πλάνων κανονικής μορφής του παίκτη i συμβολίζεται με το σύμβολο Π_i. Συνήθως, η πληθικότητα του Π_i είναι εκθετική στο μέγεθος του δέντρου του παιγνίου.

Το κυρτό περίβλημα του συνόλου των μειωμένων πλάνων κανονικής μορφής του παίκτη i ονομάζεται *πολύτοπο ακολουθιακής μορφής* του παίκτη και συμβολίζεται με το σύμβολο $\mathcal{Q}_i := \text{conv}(\Pi_i)$. Αντιπροσωπεύει το σύνολο όλων των τυχαιοποιημένων στρατηγικών στο παίγνιο. Ένα σημαντικό αποτέλεσμα του Romanovskii [1962], Koller et al. [1996], von Stengel [1996] δείχνει ότι το $\mathcal{Q}_i$ μπορεί να καταγραφεί από πολυωνυμικά πολλούς περιορισμούς στο μέγεθος του δέντρου του παιγνίου, όπως υπενθυμίζουμε στη συνέχεια.

Definition 0.0.2. *Το* πολύτοπο των στρατηγικών με μορφή ακολουθίας *του παίκτη i είναι ίσο με το κυρτό πολύτοπο*

$$\mathcal{Q}_i := \left\{ \boldsymbol{x} \in \mathbb{R}^{\Sigma}_{\geq 0} : \begin{array}{ll} (0.1) & \boldsymbol{x}[\varnothing] = 1 \\ (0.2) & \sum_{a \in \mathcal{A}_j} \boldsymbol{x}[ja] = \boldsymbol{x}[p_j] \quad \forall\, j \in \mathcal{J} \end{array} \right\}.$$

Για παράδειγμα, οι περιορισμοί που ορίζουν το πολυτόπιο ακολουθιακής μορφής για τον παίκτη 1 στο παιχνίδι 0.1 (αριστερά) εμφανίζονται στο 0.1 (δεξιά). Το πολύτοπο των στρατηγικών με μορφή ακολουθίας διαθέτει μία ισχυρή συνδυαστική δομή που επιτρέπει την επιτάχυνση πολλών κοινών διαδικασιών βελτιστοποίησης και θα είναι καίριας σημασίας για την ανάπτυξη αποτελεσματικών αλγορίθμων σύγκλισης σε ισορροπία.

Ορθολογισμός εκ των υστέρων και μάθηση σε παίγνια

Τα παίγνια είναι μία από τις πολλές καταστάσεις στις οποίες ένας λήπτης αποφάσεων πρέπει να ενεργήσει με άμεσο τρόπο. Για αυτές τις καταστάσεις, το πιο ευρέως χρησιμοποιούμενο πρωτόκολλο είναι αυτό της Άμεσης Μάθησης (π.χ., βλέπε Orabona [2022]). Συγκεκριμένα, κάθε εκπαιδευόμενος έχει ένα σύνολο ενεργειών ή συμπεριφορών που μπορεί να χρησιμοποιήσει $\mathcal{X} \subseteq \mathbb{R}^d$ (σε παίγνια εκτεταμένης μορφής, αυτό θα ήταν το σύνολο των στρατηγικών μειωμένης κανονικής μορφής). Σε κάθε χρονικό βήμα t ο εκπαιδευόμενος επιλέγει πρώτα, ενδεχομένως τυχαία, ένα στοιχείο $\boldsymbol{x}$ $\mathcal{X}$ και στη συνέχεια λαμβάνει μια συνάρτηση απώλειας (αντίθετη της ωφέλειας) $\ell^{(t)} : \mathcal{X} \mapsto \mathbb{R}$. Δεδομένου ότι όπως παρατηρήσαμε οι παραπάνω ωφέλειες στα EFGs είναι γραμμικές στα πλάνα μειωμένης

κανονικής μορφής κάθε παίκτη, στο υπόλοιπο της εργασίας θα επικεντρωθούμε στην περίπτωση που η συνάρτηση απώλειας $\ell^{(t)}$ είναι γραμμική, δηλαδή της μορφής $\ell^{(t)} : \boldsymbol{x} \mapsto \langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle$.

Ένας ευρέως υιοθετημένος στόχος για τον εκπαιδευόμενο είναι αυτός της εξασφάλισης εξαφανιζόμενου μέσου *regret* με μεγάλη πιθανότητα. Η μετάνοια ορίζεται ως η διαφορά μεταξύ της απώλειας που συσσώρευσε ο μαθητής μέσω της επιλογής της συμπεριφοράς του και της απώλειας που θα είχε συσσωρεύσει εκ των υστέρων αν είχε τροποποιήσει με συνέπεια τη συμπεριφορά του σύμφωνα με κάποια συνάρτηση μετασχηματισμού στρατηγικής. Ειδικότερα, έστω Φ ένα επιθυμητό σύνολο μετασχηματισμών στρατηγικής $\phi : \mathcal{X} \to \mathcal{X}$ που ο μαθητής μπορεί να θέλει να μάθει να μην μετανιώνει. Τότε, το Φ-regret του μαθητή ορίζεται ως η ποσότητα

$$\Phi\text{-Reg}^{(T)} := \max_{\phi \in \Phi} \sum_{t=1}^{T} \left(\langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle - \langle \boldsymbol{\ell}^{(t)}, \phi(\boldsymbol{x}^{(t)}) \rangle \right)$$

Ένας αλγόριθμος *no-Φ-regret* (επίσης γνωστός ως ελαχιστοποιητής Φ-regret) είναι ένας αλγόριθμος που, σε κάθε χρονική στιγμή T, εγγυάται με μεγάλη πιθανότητα ότι $\Phi\text{-Reg}^{(T)} = o(T)$ ανεξάρτητα από την ακολουθία των απωλειών που αποκαλύπτονται από το περιβάλλον. Το μέγεθος του συνόλου Φ των μετασχηματισμών στρατηγικής ορίζει ένα φυσικό μέτρο ορθολογισμού (που μερικές φορές ονομάζεται *ορθολογισμός εκ των υστέρων*) για τους παίκτες, και διάφορες επιλογές έχουν συζητηθεί στη βιβλιογραφία. Προφανώς, όσο το Φ γίνεται μεγαλύτερο, ο μαθητής γίνεται πιο ορθολογικός. Από την άλλη πλευρά, η εξασφάλιση υπογραμμικού regret σε σχέση με όλους τους μετασχηματισμούς του επιλεγμένου συνόλου Φ μπορεί να είναι γενικά δύσκολη. Στο ένα άκρο του φάσματος, ίσως η μικρότερη ενδιαφέρουσα επιλογή του Φ είναι το σύνολο όλων των *σταθερών* μετασχηματισμών $\Phi^{\text{const}} = \{\phi_{\hat{\boldsymbol{x}}} : \boldsymbol{x} \mapsto \hat{\boldsymbol{x}}\}_{\hat{\boldsymbol{x}} \in \mathcal{X}}$. Σε αυτή την περίπτωση, το Φ^{const}-regret ονομάζεται επίσης και *external regret* και έχει μελετηθεί εκτενώς στο πεδίο της online κυρτής βελτιστοποίησης. Στο άλλο άκρο του φάσματος, το *swap regret* αντιστοιχεί στη περίπτωση στην οποία Φ είναι το σύνολο *όλων* των μετασχηματισμών $\mathcal{X} \to \mathcal{X}$. Κάπως ενδιάμεση, και κεντρικής σημασίας στην παρούσα εργασία, είναι η έννοια του *linear-swap regret*, η οποία αντιστοιχεί στην περίπτωση στην οποία

$$\Phi := \{\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x} : \mathbf{A} \in \mathbb{R}^{d \times d}, \text{ with } \mathbf{A}\boldsymbol{x} \in \mathcal{X} \quad \forall\, \boldsymbol{x} \in \mathcal{X}\} \qquad \text{(linear-swap deviations)}$$

είναι το σύνολο όλων των γραμμικών μετασχηματισμών από $\mathcal{X}$ στον εαυτό του.[1]

Μια σημαντική παρατήρηση είναι ότι όταν όλες οι εξεταζόμενες συναρτήσεις απόκλισης στο Φ είναι γραμμικές, ένας αλγόριθμος που εγγυάται υπογραμμικό Φ-regret για το σύνολο $\mathcal{X}$ μπορεί να κατασκευαστεί άμεσα από έναν ντετερμινιστικό no-Φ-regret αλγόριθμο για το $\mathcal{X}' = \Delta(\mathcal{X})$ με δειγματοληψία $\mathcal{X} \ni \boldsymbol{x}$ από οποιοδήποτε $\boldsymbol{x}' \in \mathcal{X}'$ έτσι ώστε να εγγυάται ότι $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{x}'$. Δεδομένου ότι αυτό ακριβώς είναι το πλαίσιο που μελετάμε στην παρούσα εργασία, αυτή η διαδεδομένη παρατήρηση (βλ. επίσης Farina et al. [2022b]) μας επιτρέπει να επικεντρωθούμε στο ακόλουθο πρόβλημα: υπάρχει ένας ντετερμινιστικός αλγόριθμος χωρίς Φ-regret για το σύνολο των στρατηγικών ακολουθιακής μορφής $\mathcal{X} = \mathcal{Q}_i$ οποιουδήποτε παίκτη σε ένα παίγνιο εκτεταμένης μορφής, με εγγυημένο υπογραμμικό Φ-regret στη χειρότερη περίπτωση; Στην παρούσα εργασία απαντάμε θετικά σε αυτό το ερώτημα.

Από το regret στην ισορροπία Το πλαίσιο της Μάθησης στα Παίγνια αναφέρεται στην κατάσταση στην οποία όλοι οι παίκτες χρησιμοποιούν έναν αλγόριθμο μάθησης, λαμβάνοντας ως απώλεια το αρνητικό της κλίσης της δικής τους ωφέλειας που αποτιμάται στις στρατηγικές που εξάγονται από όλους τους άλλους παίκτες. Μια συναρπαστική πτυχή της δυναμικής μάθησης χωρίς Φ-regret είναι ότι αν κάθε παίκτης ενός παιγνίου χρησιμοποιεί έναν αλγόριθμο χωρίς Φ-regret, τότε η εμπειρική συχνότητα παιξίματος συγκλίνει σχεδόν σίγουρα στο σύνολο των Φ-ισορροπιών, οι οποίες είναι έννοιες συσχετισμένων ισορροπιών, στις οποίες ο ορθολογισμός των παικτών περιορίζεται από το μέγεθος του συνόλου Φ. Τυπικά, για ένα σύνολο Φ αποκλίσεων στρατηγικής, μια Φ-ισορροπία ορίζεται ως εξής.

[1] Για τους σκοπούς αυτής της εργασίας, το επίθετο *linear* αναφέρεται στο γεγονός ότι κάθε μετασχηματισμός μπορεί να εκφραστεί με τη μορφή $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ για έναν κατάλληλο πίνακα $\mathbf{A}$.

Definition 0.0.3. *Για ένα παίγνιο n παικτών εκτεταμένης μορφής G και ένα σύνολο Φ_i αποκλίσεων για κάθε παίκτη, μια $\{\Phi_i\}$-ισορροπία είναι μια κοινή κατανομή $\mu \in \Delta(\Pi_1 \times \cdots \times \Pi_n)$ τέτοια ώστε για κάθε παίκτη i και κάθε απόκλιση $\phi \in \Phi_i$ να ισχύει ότι*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(\boldsymbol{x})] \geq \mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(\phi(\boldsymbol{x}_i), \boldsymbol{x}_{-i})]$$

Δηλαδή, κανένας παίκτης i δεν έχει κίνητρο να αποκλίνει μονομερώς από τη συνιστώμενη κοινή στρατηγική $\boldsymbol{x}$ χρησιμοποιώντας οποιονδήποτε μετασχηματισμό $\phi \in \Phi_i$.

Αυτό το γενικό πλαίσιο καλύπτει αρκετές σημαντικές έννοιες ισορροπίας σε μια ποικιλία παιγνιοθεωρητικών μοντέλων. Για παράδειγμα, τόσο στα NFGs όσο και στα EFGs, οι no-external regret δυναμικές συγκλίνουν στο σύνολο των coarse correlated equilibria. Στα NFGs, οι no-swap regret δυναμικές συγκλίνουν στο σύνολο των correlated equilibria [Blum and Mansour, 2007]. Στα EFGs, οι Farina et al. [2022b] απέδειξαν πρόσφατα ότι ένα συγκεκριμένο υποσύνολο Φ γραμμικών μετασχηματισμών που ονομάζεται *trigger deviations* οδηγεί στο σύνολο των EFCE.

Αναγωγή του Φ-regret σε external regret Μια κομψή κατασκευή των Gordon et al. [2008] επιτρέπει την κατασκευή αλγορίθμων no-Φ-regret για ένα γενικό σύνολο $\mathcal{X}$ ξεκινώντας από έναν αλγόριθμο no-external-regret για το Φ. Υπενθυμίζουμε εν συντομία το αποτέλεσμα.

Theorem 0.0.4 (Gordon et al. [2008])**.** *Έστω $\mathcal{R}$ ένας ελαχιστοποιητής external regret που έχει ως χώρο δράσης το σύνολο των μετασχηματισμών Φ και επιτυγχάνει υπογραμμικό external regret $\text{Reg}^{(T)}$. Επιπλέον, υποθέτουμε ότι για όλα τα $\phi \in \Phi$ υπάρχει ένα σταθερό σημείο $\phi(\boldsymbol{x}) = \boldsymbol{x} \in \mathcal{X}$. Τότε, ένας ελαχιστοποιητής Φ-regret $\mathcal{R}_\Phi$ μπορεί να κατασκευαστεί ως εξής:*

- *Για την έξοδο μιας στρατηγικής $\boldsymbol{x}^{(t)}$ στην επανάληψη t του $\mathcal{R}_\Phi$, λάβε μια έξοδο $\phi^{(t)} \in \Phi$ του external regret ελαχιστοποιητή $\mathcal{R}$, και επέστρεψε ένα από τα σταθερά σημεία του $\boldsymbol{x}^{(t)} = \phi^{(t)}(\boldsymbol{x}^{(t)})$.*
- *Για κάθε γραμμική συνάρτηση απώλειας $\ell^{(t)}$ που λαμβάνει ο $\mathcal{R}_\Phi$, κατασκεύασε τη γραμμική συνάρτηση $L^{(t)} : \phi \mapsto \ell^{(t)}(\phi(\boldsymbol{x}^{(t)}))$ και πέρασέ την ως απώλεια στο $\mathcal{R}$.*

Έστω $\Phi\text{-Reg}^{(T)}$ το Φ-regret του $\mathcal{R}_\Phi$. Σύμφωνα με την προηγούμενη κατασκευή, ισχύει ότι

$$\Phi\text{-Reg}^{(T)} = \text{Reg}^{(T)} \qquad \forall\, T = 1, 2, \ldots$$

Έτσι, αν $\mathcal{R}$ είναι ένας ελαχιστοποιητής external regret, τότε ο $\mathcal{R}_\Phi$ είναι ένας ελαχιστοποιητής Φ-regret.

Ένας αλγόριθμος No-Linear-Swap Regret με επαναλήψεις πολυωνυμικού χρόνου

Σε αυτή την ενότητα, περιγράφουμε τον no-linear-swap regret αλγόριθμό μας για το σύνολο των στρατηγικών ακολουθιακής μορφής $\mathcal{Q}$ ενός γενικού παίκτη σε οποιοδήποτε παίγνιο εκτεταμένης μορφής με τέλεια ανάκληση και ατελή πληροφόρηση. Ο αλγόριθμος ακολουθεί το γενικό πρότυπο για την κατασκευή ελαχιστοποιητών Φ-regret που δίνεται από την εργασία Gordon et al. [2008] και υπενθυμίζεται στο Θεώρημα 0.0.4. Για το σκοπό αυτό χρειαζόμαστε δύο στοιχεία:

i) έναν αποδοτικό ελαχιστοποιητή external regret για το σύνολο $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ όλων των πινάκων που προκαλούν γραμμικούς μετασχηματισμούς από $\mathcal{Q}$ σε $\mathcal{Q}$,

ii) ένα αποδοτικά υπολογίσιμο μαντείο σταθερού σημείου για πίνακες $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, που επιστρέφει $\boldsymbol{x} = \mathbf{A}\boldsymbol{x} \in \mathcal{Q}$.

Η ύπαρξη ενός σταθερού σημείου, που απαιτείται στο ii), είναι εύκολο να διαπιστωθεί με το θεώρημα σταθερού σημείου του Brouwer, δεδομένου ότι το πολύτοπο των στρατηγικών ακολουθιακής μορφής είναι συμπαγές και κυρτό, και η συνεχής συνάρτηση $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ απεικονίζει το $\mathcal{Q}$ στον εαυτό του εξ ορισμού. Επιπλέον, όπως θα φανεί στη συνέχεια της ενότητας, όλα τα στοιχεία $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ έχουν τιμές στο $[0, 1]^{\Sigma\times\Sigma}$. Επομένως, η απαίτηση ii) μπορεί να ικανοποιηθεί άμεσα με την επίλυση του γραμμικού προγράμματος εφικτότητας $\{\text{find } \boldsymbol{x} : \mathbf{A}\boldsymbol{x} = \boldsymbol{x}, \boldsymbol{x} \in \mathcal{Q}\}$ χρησιμοποιώντας οποιονδήποτε από

τους γνωστούς αλγορίθμους πολυωνυμικού χρόνου για γραμμικό προγραμματισμό. Η ικανοποίηση της απαίτησης i) είναι η ουσία του θέματος και αποτελεί το επίκεντρο μεγάλου μέρους της εργασίας. Εδώ, δίνουμε τη διαίσθηση για τις κύριες ιδέες που συμβάλλουν στον αλγόριθμο. Όλες οι αποδείξεις παραπέμπονται στο παράρτημα.

Η δομή των γραμμικών μετασχηματισμών των πολυτόπων στρατηγικών ακολουθιακής μορφής

Το κρίσιμο βήμα στην κατασκευή μας είναι να καθιερώσουμε μια σειρά αποτελεσμάτων που ρίχνουν φως στη θεμελιώδη γεωμετρία του συνόλου $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ *όλων* των γραμμικών μετασχηματισμών από ένα πολύτοπο ακολουθιακής μορφής $\mathcal{Q}$ στον εαυτό του. Στην πραγματικότητα, τα αποτελέσματά μας επεκτείνονται πέρα από τις συναρτήσεις από το $\mathcal{Q}$ στο $\mathcal{Q}$ σε πιο γενικές συναρτήσεις από το $\mathcal{Q}$ σε ένα γενικό συμπαγές πολύτοπο $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\}$ για αυθαίρετα $\mathbf{P}$ και $\boldsymbol{p}$. Θεμελιώνουμε το ακόλουθο θεώρημα χαρακτηρισμού, το οποίο δείχνει ότι όταν οι συναρτήσεις εκφράζονται σε μορφή πίνακα, το σύνολο $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ μπορεί να προσδιοριστεί από έναν πολυωνυμικό αριθμό περιορισμών. Η απόδειξη αναβάλλεται για το Κεφάλαιο 5.

Theorem 0.0.5. *Έστω $\mathcal{Q}$ ένας χώρος στρατηγικών ακολουθιακής μορφής και έστω $\mathcal{P}$ οποιοδήποτε φραγμένο πολύτοπο της μορφής $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\} \subseteq [0,\gamma]^d$, όπου $\mathbf{P} \in \mathbb{R}^{k\times d}$. Τότε, για κάθε γραμμική συνάρτηση $f : \mathcal{Q} \to \mathcal{P}$, υπάρχει ένας πίνακας $\mathbf{A}$ στο πολύτοπο*

$$\mathcal{M}_{\mathcal{Q}\to\mathcal{P}} \coloneqq \left\{ \mathbf{A} = \left[\cdots \mid \mathbf{A}_{(\sigma)} \mid \cdots \right] \in \mathbb{R}^{d\times\Sigma} : \begin{array}{lll} (0.3) & \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & \forall\, ja \in \Sigma^{\perp} \\ (0.4) & \mathbf{A}_{(\sigma)} = \mathbf{0} & \forall\, \sigma \in \Sigma \setminus \Sigma^{\perp} \\ (0.5) & \sum_{j'\in\mathcal{C}_{\varnothing}} \boldsymbol{b}_{j'} = \boldsymbol{p} & \\ (0.6) & \sum_{j'\in\mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j & \forall\, ja \in \Sigma \setminus \Sigma^{\perp} \\ (0.7) & \mathbf{A}_{(\sigma)} \in [0,\gamma]^d & \forall\, \sigma \in \Sigma \\ (0.8) & \boldsymbol{b}_j \in \mathbb{R}^k & \forall\, j \in \mathcal{J} \end{array} \right\}$$

τέτοιος ώστε $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ για όλα τα $\boldsymbol{x} \in \mathcal{Q}$. Αντίστροφα, κάθε $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ ορίζει μια γραμμική συνάρτηση $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ από $\mathcal{Q}$ σε $\mathcal{P}$, δηλαδή τέτοια ώστε $\mathbf{A}\boldsymbol{x} \in \mathcal{P}$ για όλα τα $\boldsymbol{x} \in \mathcal{Q}$.

Η απόδειξη λειτουργεί με επαγωγή σε πολλαπλά βήματα. Στον πυρήνα της, εκμεταλλεύεται τη συνδυαστική δομή των πολυτόπων στρατηγικών ακολουθιακής μορφής, τα οποία μπορούν να αναλυθούν σε υποπροβλήματα χρησιμοποιώντας μια σειρά από καρτεσιανά γινόμενα και κυρτά περιβλήματα. Παρατηρούμε επίσης ότι ενώ το θεώρημα απαιτεί το πολύτοπο $\mathcal{P}$ να είναι της μορφής $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\}$, με λίγη δουλειά το αποτέλεσμα μπορεί επίσης να επεκταθεί για να χειρίζεται άλλες αναπαραστάσεις όπως $\{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} \leq \boldsymbol{p}\}$. Επιλέξαμε τη μορφή που ορίζεται στο θεώρημα, καθώς οδηγεί πιο άμεσα στην απόδειξη και καθώς οι περιορισμοί που ορίζουν το πολυτόπιο στρατηγικών ακολουθιακής μορφής (Ορισμός 0.0.2) είναι ήδη στη μορφή της έκφρασης.

Συγκεκριμένα, θέτοντας $\mathcal{P} = \mathcal{Q}$ στο Θεώρημα 0.0.5 (σε αυτή την περίπτωση, οι διαστάσεις του $\mathbf{P}$ θα είναι $k = |\mathcal{J}|+1$ και $d = |\Sigma|$), συμπεραίνουμε ότι το σύνολο των γραμμικών συναρτήσεων από το $\mathcal{Q}$ στον εαυτό του είναι ένα συμπαγές και κυρτό πολύτοπο $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}} \subseteq [0,1]^{\Sigma\times\Sigma}$, που ορίζεται από $O(|\Sigma|^2)$ γραμμικούς περιορισμούς. Όπως συζητήθηκε, αυτός ο πολυωνυμικός χαρακτηρισμός του $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ είναι η θεμελιώδης διαπίστωση που επιτρέπει την ελαχιστοποίηση σε πολυωνυμικό χρόνο του linear-swap regret σε γενικά παίγνια εκτεταμένης μορφής.

Ο αλγόριθμός μας για no-linear-swap regret

Από εδώ και πέρα, η κατασκευή ενός αλγορίθμου no-external-regret για το $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ είναι σχετικά απλή, χρησιμοποιώντας τυποποιημένα εργαλεία από την πλούσια βιβλιογραφία της online μάθησης. Για παράδειγμα, στον Αλγόριθμο 1, προτείνουμε μια λύση που χρησιμοποιεί online projected gradient descent [Gordon, 1999, Zinkevich, 2003].

Algorithm 1: Φ-Regret minimizer for the set $\Phi = \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$

Data: $\mathbf{A}^{(1)} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ and fixed point $\boldsymbol{x}^{(1)}$ of $\mathbf{A}^{(1)}$, learning rates $\eta^{(t)} > 0$

for $t = 1, 2, \ldots$ **do**

- Output $\boldsymbol{x}^{(t)}$
- Receive $\boldsymbol{\ell}^{(t)}$ and pay $\langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle$
- Set $\mathbf{L}^{(t)} = \boldsymbol{\ell}^{(t)} (\boldsymbol{x}^{(t)})^\top$
- $\mathbf{A}^{(t+1)} = \Pi_{\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}}(\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)}) = \arg\min_{\mathbf{Y}\in\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}} \|\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)} - \mathbf{Y}\|_F^2$
- Compute a fixed point $\boldsymbol{x}^{(t+1)} = \mathbf{A}^{(t+1)}\boldsymbol{x}^{(t+1)} \in \mathcal{Q}$ of matrix $\mathbf{A}^{(t+1)}$

Συνδυάζοντας αυτόν τον αλγόριθμο no-external-regret για το Φ με την κατασκευή του Gordon et al. [2008], μπορούμε στη συνέχεια να καθορίσουμε τα ακόλουθα όρια linear-swap regret και πολυπλοκότητας επανάληψης για τον Αλγόριθμο 1.

Theorem 0.0.6 (Informal)**.** *Έστω Σ το σύνολο των ακολουθιών του παίκτη που μαθαίνει στο παίγνιο εκτεταμένης μορφής, και έστω $\eta^{(t)} = 1/\sqrt{t}$ για όλα τα t. Τότε, για οποιαδήποτε ακολουθία διανυσμάτων απωλειών $\boldsymbol{\ell}^{(t)}$ $[0,1]^\Sigma$, ο Αλγόριθμος 1 εγγυάται linear-swap regret $O(|\Sigma|^2\sqrt{T})$ μετά από οποιονδήποτε αριθμό T επαναλήψεων και εκτελείται σε χρόνο $O(\text{poly}(|\Sigma|)\log^2 t)$ για κάθε επανάληψη t.*

Η επίσημη εκδοχή του θεωρήματος δίνεται στο Θεώρημα 4.2.1. Αξίζει να σημειωθεί ότι η περιγραφή του $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ σε πολυωνυμικό μέγεθος είναι κρίσιμη για την καθιέρωση του πολυωνυμικού χρόνου εκτέλεσης του αλγορίθμου, τόσο στο βήμα προβολής (5) όσο και στο βήμα υπολογισμού σταθερού σημείου (6). Παρατηρούμε επίσης ότι η επιλογή του online projected gradient descent σε συνδυασμό με τη μέθοδο ελλειψοειδούς για τις προβολές ήταν αυθαίρετη και οι χρήσιμες ιδιότητες του $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ διατηρούνται όταν χρησιμοποιείται με οποιονδήποτε αποδοτικό regret minimizer.

Linear-Deviation Correlated Equilibrium

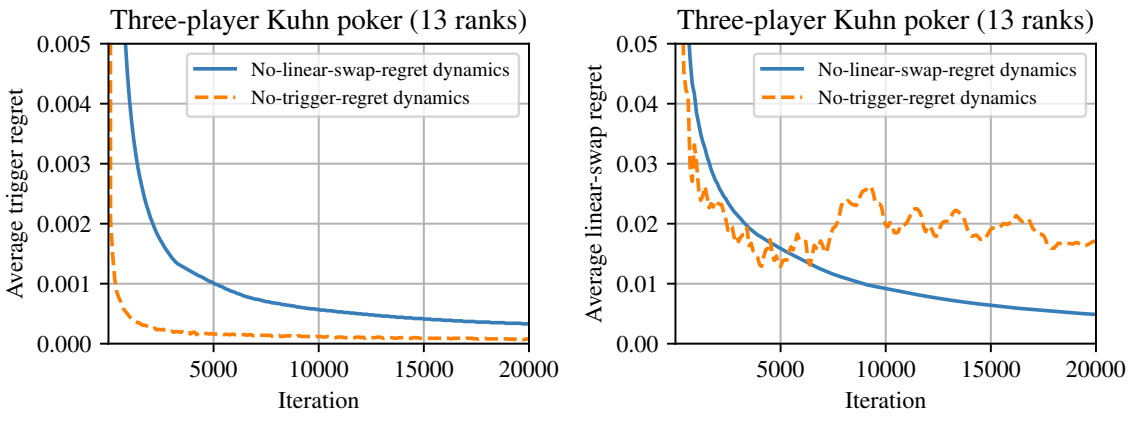
Όπως συζητήσαμε στα εισαγωγικά, όταν όλοι οι παίκτες σε ένα παίγνιο χρησιμοποιούν αλγορίθμους μάθησης no-Φ-regret, τότε η εμπειρική συχνότητα του παιξίματος συγκλίνει σχεδόν σίγουρα στο σύνολο των Φ-ισορροπιών. Παρομοίως, όταν $\Phi = \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ οι παίκτες ενεργούν με βάση δυναμικές "no-linear-swap regret" και συγκλίνουν σε μια έννοια της Φ-ισορροπίας που ονομάζουμε *linear-deviation correlated equilibrium* (LCE). Σε αυτή την ενότητα παρουσιάζουμε ορισμένες αξιοσημείωτες ιδιότητες της LCE. Συγκεκριμένα, συζητάμε τη σχέση της με άλλες ήδη καθιερωμένες ισορροπίες, καθώς και την υπολογιστική δυνατότητα επιλογής βέλτιστης ισορροπίας.

Σχέση με CE και EFCE

Το πλαίσιο ελαχιστοποίησης του Φ-regret, προσφέρει έναν φυσικό τρόπο για την οικοδόμηση μιας ιεραρχίας των αντίστοιχων Φ-ισορροπιών με βάση τη σχέση των Φ συνόλων αποκλίσεων. Ειδικότερα, εάν για τα σύνολα Φ_1, Φ_2 ισχύει ότι $\Phi_1 \subseteq \Phi_2$, τότε το σύνολο των Φ_2-ισορροπιών είναι υποσύνολο του συνόλου των Φ_1-ισορροπιών. Δεδομένου ότι η συσχετισμένη ισορροπία ορίζεται χρησιμοποιώντας το σύνολο όλων των αποκλίσεων ανταλλαγής, συμπεραίνουμε ότι κάθε Φ-ισορροπία, συμπεριλαμβανομένης της LCE, είναι υπερσύνολο της CE. Ποια είναι τότε η σχέση της LCE με τη συσχετισμένη ισορροπία εκτεταμένης μορφής (EFCE); Οι Farina et al. [2022b] έδειξαν ότι το σύνολο Φ^{EFCE} που οδηγεί στην EFCE είναι το σύνολο όλων των "trigger deviations", οι οποίες μπορούν να εκφραστούν ως γραμμικοί μετασχηματισμοί των στρατηγικών εκτεταμένης μορφής. Κατά συνέπεια, το σύνολο Φ^{EFCE} είναι ένα υποσύνολο όλων των γραμμικών μετασχηματισμών και συνεπώς, ισχύει ότι CE $\subseteq$ LCE $\subseteq$ EFCE. Στα παραδείγματα 6.1.1 και 6.1.3 του παραρτήματος δείχνουμε ότι υπάρχουν συγκεκριμένα παίγνια στα οποία είτε το CE $\neq$ LCE, ή LCE $\neq$ EFCE. Επομένως, συμπεραίνουμε ότι οι προηγούμενες εγκλείσεις είναι αυστηρές και ισχύει CE $\subset$ LCE $\subset$ EFCE.

Για το Παράδειγμα 6.1.1 χρησιμοποιούμε ένα παίγνιο σηματοδότησης από τους von Stengel and Forges [2008] με γνωστή EFCE και εντοπίζουμε έναν γραμμικό μετασχηματισμό που δεν αποδίδεται από τις trigger αποκλίσεις της EFCE. Συγκεκριμένα, είναι δυνατό να εκτελέσουμε γραμμικούς μετασχηματισμούς στις ακολουθίες ενός υποδέντρου με βάση τις στρατηγικές σε άλλα υποδέντρα του TFSDP. Για το Παράδειγμα 6.1.3 βρήκαμε ένα συγκεκριμένο παίγνιο μέσω υπολογιστικής αναζήτησης που έχει ένα LCE, το οποίο δεν είναι μια κανονικής μορφής συσχετισμένη ισορροπία. Για να το κάνουμε αυτό εντοπίζουμε μια συγκεκριμένη κανονικής μορφής ανταλλαγή που είναι μη γραμμική.

Εμπειρική αξιολόγηση Για να καταδείξουμε περαιτέρω τον διαχωρισμό μεταξύ των no-linear-swap regret δυναμικών και των no-trigger-regret δυναμικών, που χρησιμοποιούνται για την EFCE, παρέχουμε πειραματικές αποδείξεις ότι η ελαχιστοποίηση του linear-swap-regret ελαχιστοποιεί επίσης το trigger-regret (Σχήμα 0.2, αριστερά), ενώ η ελαχιστοποίηση του trigger-regret *δεν* ελαχιστοποιεί το linear-swap regret. Συγκεκριμένα, στο Σχήμα 0.2 συγκρίνουμε τη δική μας δυναμική μάθησης no-linear-swap-regret (που δίνεται στον Αλγόριθμο 1) με τον αλγόριθμο no-trigger-regret που εισήχθη από τους Farina et al. [2022b]. Περισσότερες λεπτομέρειες σχετικά με την υλοποίηση των αλγορίθμων είναι διαθέσιμες στο Κεφάλαιο B. Στο αριστερό διάγραμμα, μετράμε στον άξονα y το μέσο trigger regret που προκύπτει όταν όλοι οι παίκτες χρησιμοποιούν τη μία ή την άλλη δυναμική. Δεδομένου ότι οι trigger deviations είναι ειδικές περιπτώσεις των γραμμικών αποκλίσεων, όπως αναμενόταν, παρατηρούμε ότι και οι δύο δυναμικές είναι σε θέση να ελαχιστοποιήσουν το trigger regret. Αντίθετα, στο δεξιό διάγραμμα του Σχήματος 0.2, ο άξονας y μετράει το linear-swap-regret. Παρατηρούμε ότι ενώ οι δυναμικές μας επικυρώνουν τις επιδόσεις υπογραμμικού regret που αποδεικνύονται στο Θεώρημα 0.0.6, οι δυναμικές χωρίς trigger-regret των Farina et al. [2022b] παρουσιάζουν μια ακανόνιστη συμπεριφορά που δύσκολα είναι συμβατή με ένα εξαλειφόμενο μέσο regret. Αυτό υποδηλώνει ότι το no-linear-swap-regret είναι πράγματι μια αυστηρά ισχυρότερη έννοια της εκ των υστέρων ορθολογικότητας.



Σχήμα 0.2: (Αριστερά) Μέσο trigger regret ανά επανάληψη για έναν ελαχιστοποιητή linear-swap regret και έναν ελαχιστοποιητή trigger regret. (Δεξιά) Μέσο linear-swap regret ανά επανάληψη για τους ίδιους δύο ελαχιστοποιητές.

Δυσκολία της μεγιστοποίησης της κοινωνικής ευημερίας

Σε πολλές περιπτώσεις μας ενδιαφέρει να γνωρίζουμε αν είναι δυνατόν να επιλέξουμε μια ισορροπία με μέγιστη Κοινωνική Ευημερία. Έστω MAXPAY-LCE το πρόβλημα της εύρεσης μιας LCE σε EFGs που μεγιστοποιεί το άθροισμα (ή οποιονδήποτε γραμμικό συνδυασμό) των ωφελειών όλων των παικτών. Παρακάτω, αποδεικνύουμε ότι δεν μπορούμε να λύσουμε αποδοτικά το MAXPAY-LCE, εκτός αν P=NP, ακόμη και για 2 παίκτες αν επιτρέπονται τυχαίες κινήσεις, και ακόμη και για 3 παίκτες

διαφορετικά. Ακολουθούμε τη δομή της ίδιας απόδειξης δυσκολίας για το πρόβλημα MAXPAY-CE της εύρεσης ενός βέλτιστου CE σε EFGs. Συγκεκριμένα, οι von Stengel and Forges [2008] χρησιμοποιούν μια αναγωγή από το SAT για να αποδείξουν ότι η απόφαση για το αν το MAXPAY-CE μπορεί να επιτύχει τη μέγιστη τιμή είναι NP-δύσκολη ακόμη και για 2 παίκτες. Για να το κάνουν αυτό, επινοούν έναν τρόπο για να αντιστοιχίσουν οποιαδήποτε περίπτωση SAT σε ένα πολυωνυμικά μεγάλο δέντρο παιγνίων στο οποίο η ρίζα είναι ο παίκτης της τύχης, το δεύτερο επίπεδο αντιστοιχεί σε έναν παίκτη και το τρίτο επίπεδο αντιστοιχεί στον άλλο παίκτη. Οι ωφέλειες και για τους δύο παίκτες είναι ακριβώς οι ίδιες, επομένως οι παίκτες θα πρέπει να συντονιστούν για να μεγιστοποιήσουν την αμοιβή τους ανεξάρτητα από τον γραμμικό συνδυασμό των ωφελειών που στοχεύουμε να μεγιστοποιήσουμε.

Theorem 0.0.7. *Για παίγνια εκτεταμένης μορφής δύο παικτών, τέλειας ανάκλησης με τυχαίες κινήσεις, το πρόβλημα MAXPAY-LCE δεν είναι επιλύσιμο σε πολυωνυμικό χρόνο, εκτός αν P=NP.*

Remark 0.0.8. *Το πρόβλημα διατηρεί την δυσκολία του αν αφαιρέσουμε τον κόμβο τύχης και προσθέσουμε έναν τρίτο παίκτη. Όπως αποδείχθηκε από τους von Stengel and Forges [2008], σε αυτή την περίπτωση μπορούμε πάντα να κατασκευάσουμε ένα πολυωνυμικού μεγέθους δέντρο παιγνίων που αναγκάζει τον τρίτο παίκτη να ενεργεί ως κόμβος τύχης.*

Κείμενο στα αγγλικά

Chapter 1

Introduction

The framework of regret minimization provides algorithms that players can use to gradually improve their strategies in a repeated game, enabling learning strong strategies even when facing unknown and potentially adversarial opponents. One of the appealing properties of no-regret learning algorithms is that they are *uncoupled*, meaning that each player refines their strategy based on their own payoff function, and on other players' strategies, but not on the payoff functions of other players. Nonetheless, despite their uncoupled nature and focus on *local* optimization of each player's utility, it is one of the most celebrated results in the theory of learning in games that in many cases, when all players are learning using these algorithms, the empirical play recovers appropriate notions of *equilibrium*—a *global* notion of game-theoretic optimality. Strategies constructed via no-regret learning algorithms (or approximations thereof) have been key components in constructing human-level and even superhuman AI agents in a variety of adversarial games, including Poker [Moravčík et al., 2017, Brown and Sandholm, 2018, 2019], Stratego [Perolat et al., 2022], and Diplomacy [Bakhtin et al., 2023].

In regret minimization, each learning agent seeks to minimize the difference between the loss (opposite of reward) they accumulated through the actions they played, and the loss they would have accumulated in hindsight had they consistently modified their behavior according to some strategy transformation function. The size of the set of transformation functions considered by the learning agent determines a natural notion of rationality of the agent. Already when the agents seeks to learn strategies that cumulate low regret against *constant* strategy transformations only—a notion of regret called *external* regret—the average play of the agents converges to a Nash equilibrium in two-player constant-sum games, and to a coarse correlated equilibrium in general-sum multiplayer games. As the sets of transformations the each agent considers grows, more complex equilibria can be achieved, including correlated equilibria in normal-form games (Foster and Vohra [1997], Fudenberg and Levine [1995, 1999], Hart and Mas-Colell [2000, 2001]; see also the monograph by Fudenberg and Levine [1998]) and extensive-form correlated equilibria in extensive-form games [Farina et al., 2022b]. At the extreme, a maximally hindsight-rational agent is one that minimizes regret against the set of *all* functions from the strategy space to itself (aka. *swap* regret). While it is known that maximum hindsight rationality can be attained efficiently in nonsequential (normal-form) games [Stoltz and Lugosi, 2007, Blum and Mansour, 2007], it is a major open problem to determine whether the same applies to sequential (*i.e.*, extensive-form) games, and more generally what is the strongest notion of rationality that can be attained efficiently in the worst case in the latter setting.

In this work, we provide a positive result in that direction, by showing that hindsight rationality can be achieved efficiently in general imperfect-information extensive-form games when one restricts to the set of *all linear transformations* of the mixed strategy space—a notion called *linear-swap regret*, and that coincides with swap regret in normal-form games. In order to establish the result, we introduce several intermediate results related to the geometry of sequence-form strategies in extensive-form games. In particular, a crucial result is given in Theorem 4.1.1, which shows that the set of linear functions $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ from the sequence-form strategy set $\mathcal{Q}$ of a player in an extensive-form game to a generic convex polytope $\mathcal{P}$ can be captured using only polynomially many linear constraints in the size of the game tree and the number of linear constraints that define $\mathcal{P}$. Applying the result to the special case $\mathcal{P} = \mathcal{Q}$, we are then able to conclude that the the polytope of linear transformations $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ from the sequence-form strategy set to itself can be captured by polynomially many linear constraints

in the size of the game tree, and the norm of any element is polynomially bounded. The polynomial characterization and bound for $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ is used in conjunction with an idea of Gordon et al. [2008] to construct a no-linear-swap-regret minimizer for the set of strategies $\mathcal{Q}$ starting from two primitives: i) a no-external-regret algorithm for the set of transformations $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, and ii) an algorithm to compute a fixed point strategy for any transformation in $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$. In both cases, the polynomial representation of $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ established through Theorem 4.1.1 plays a fundamental role. It allows, on the one hand, to satisfy requirement ii) using linear programming. On the other hand, it enables us to construct a no-external-regret algorithm that outputs transformations in $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ with polynomial-time iterations, by leveraging the known properties of online projected gradient descent, exploiting the tractability of projecting onto polynomially-representable polytopes.

Finally, in the last section of the thesis we turn our attention away from hindsight rationality to focus instead on the properties of the equilibria that our no-linear-swap-regret dynamics recover in extensive-form games. The average play of no-linear-swap-regret players converges to a set of equilibria that we coin *linear-deviation correlated equilibria (LCEs)*. LCEs form a superset of correlated equilibria and a subset of extensive-form correlated equilibria in extensive-form games. In Chapter 6 we show that these inclusions are in general strict, and provide additional results about the complexity of computing a welfare-maximizing LCE.

Related work As mentioned in the introduction, the existence of uncoupled no-regret dynamics leading to correlated equilibrium (CE) in multiplayer normal-form games is a celebrated result dating back to at least the work by Foster and Vohra [1997]. That work inspired researchers to seek uncoupled learning procedures in other settings as well. For example, Stoltz and Lugosi [2007] studies learning dynamics leading to CE in games with an infinite (but compact) action set, while Kakade et al. [2003] focuses on graphical games. In more recent years, a growing effort has been spent towards understanding the relationships between no-regret learning and equilibria in imperfect-information extensive-form games, the settings on which we focus. Extensive-form games pose additional challenges when compared to normal-form games, due to their sequential nature and presence of imperfect information. While efficient no-external-regret learning dynamics for extensive-form games are known (including the popular CFR algorithm [Zinkevich et al., 2008]), as of today not much is known about no-swap-regret and the complexity of learning CE in extensive-form games.

The closest notion to CE that is known to be efficiently computable in extensive-form games is *extensive-form correlated equilibrium* (EFCE), introduced by von Stengel and Forges [2008]. The question of whether the set of EFCE could be approached via uncoupled no-regret dynamics with polynomial-time iterations in the size of the extensive-form games was recently settled in the positive [Farina et al., 2022b, Celli et al., 2020]. In particular, Farina et al. [2022b] show that EFCE arises from the average play of no-trigger-regret algorithms, where trigger deviations are a particular subset of linear transformations of the sequence-form strategy polytope $\mathcal{Q}$ of each player. Since this thesis focuses on learning dynamics that guarantee sublinear regret with respect to *any* linear transformation of $\mathcal{Q}$, it follows immediately that the dynamics presented in this thesis recover EFCE as a special case.

The concept of linear-swap-regret minimization has been considered before in the context of *Bayesian* games. Mansour et al. [2022] study a setting where a no-regret *learner* competes in a two-player Bayesian game with a rational utility *maximizer*, that is a strictly more powerful opponent than a learner. Under this setting, it can be shown that in every round the optimizer is guaranteed to obtain at least the Bayesian Stackelberg value of the game. Then they proceed to prove that minimizing *linear-swap regret* is necessary if we want to cap the optimizer's performance at the Stackelberg value, while minimizing polytope-swap regret (a generalization of swap regret for Bayesian games, and strictly stronger than linear-swap) is sufficient to cap the optimizer's performance. Hence, these results highlight the importance of developing learning algorithms under stronger notions of *rationality*, as is our aim in this thesis. Furthermore, these results provide evidence that constructing a no-linear-swap regret learner, as is our goal here, can present benefits when compared to other less rational learners. In a concurrent paper, Fujii [2023] defines the notion of *untruthful swap regret* for

Bayesian games and proves that, for Bayesian games, it is equivalent to the linear-swap regret which is of interest in our work.

Bayesian games can be considered as a special case of extensive-form games, where a chance node initially selects one of the possible types Θ for each player. Thus, our algorithm minimizing linear-swap regret in extensive-form games also minimizes linear-swap regret in Bayesian games. However, we remark that our regret bound depends polynomially on the number of player types $|\Theta|$ as they are part of the game tree representation, while Fujii [2023] has devised an algorithm for Bayesian games, whose regret only depends on $\log |\Theta|$.

Finally, we also mention that nonlinear deviations have been explored in extensive-form games, though we are not aware of notable large sets for which polynomial-time no-regret dynamics can be devised. Specifically, we point to the work by Morrill et al. [2021], which defines the notion of "behavioral deviations". These deviations are nonlinear with respect to the sequence-form representation of strategies in extensive-form games. The authors categorize several known or novel types of restricted behavioral deviations into a Deviation Landscape that highlights the relations between them. Even though both the linear deviations, we consider in this work, and the behavioral deviations seem to constitute rich measures of rationality, none of them contains the other and thus, linear deviations do not fit into the Deviation Landscape of Morrill et al. [2021] (see also Remark 6.1.2).

Chapter 2

Game Theory Basics

Game theory studies settings involving strategic interactions of multiple rational agents (or "players"). Game theory was initially established as a field of economics, and was mathematically developed mainly by von Neumann and Morgenstern [1944] and Nash [1951]. However, game-theoretic models have since been applied in a wide variety of disciplines such as political science, philosophy, evolution and biology more general (eg. see Lewis [1969], Bicchieri [1989], Smith and Price [1973], Smith and Harper [2003]). In this chapter we will present some of the basic game-theoretic concepts used in this thesis. The focus of this thesis is on extensive-form games, but we begin our discussion with the more widely studied normal-form games.

2.1 Normal-Form Games

One of the most widely studied descriptions of games is that of normal-form games (NFGs). Sometimes, NFGs are also called matrix games, because they can be represented using a payoff matrix. More specifically, in an n-player normal-form game each player i has a fixed set of *actions* (or pure strategies) S_i. Players choose their actions simultaneously and receive a payoff that represents their personal utility, or preference, which depends on the combination of chosen actions by all players. Consider for example the following rock-paper-scissors game. In this game there exist two players, one player choosing between rows and the other choosing columns. Depending on the outcome of the game, the *payoff matrix* shows the gained utilities for each of the two players.

Example 2.1.1 (Rock-Paper-Scissors)**.** *In the popular game of Rock-Paper-Scissors there exist 2 players, each having 3 actions: rock (R), paper (P), and scissors (S). Players pick their actions and reveal them simultaneously. The player with the stronger move wins the game, or if the moves are the same, the outcome is a draw. The rules are: rock beats scissors, scissors beat paper, and paper beats rock. The utility matrix for both players is shown below.*

1 \ 2	R	P	S
R	0, 0	-1, 1	1, -1
P	1, -1	0, 0	-1, 1
S	-1, 1	1, -1	0, 0

This is an example of a *zero-sum* game, because winning outcomes for one player correspond to losing outcomes for the other, and vice versa. In other words, the sum of utilities at each entry of the payoff matrix equals zero. This does not always have to be the case, as there might exist games with situations that are mutually beneficial for all players. For example, consider the following "traffic light" game.

Example 2.1.2 (traffic light game)**.** *Consider a situation where two cars are driving towards a crossing. Each car has the option to either STOP or GO. If one car decides to GO and the other to STOP, then the passing car gets a utility of 1 and the waiting a utility of 0. If both cars STOP then they both*

get a utility of 0 because they are wasting time waiting. However, if both cars decide to GO then an accident will happen and the utility is extremely negative for both of them. The payoff matrix for this game is given below.

1 \ 2	STOP	GO
STOP	0, 0	0, 1
GO	1, 0	-100, -100

Now that we have become more familiar with some basic 2-player normal-form games, we are ready to formally define the general mathematical model.

Definition 2.1.3 (normal-form game)**.** *The normal-form representation of a finite n-player game contains for each player i:*

- *a set of actions (or pure strategies) S_i*
- *and a utility (or payoff) function $u_i : S_1 \times \cdots \times S_n \mapsto \mathbb{R}$*

A tuple $(s_1, \ldots, s_n) \in S_1 \times \cdots \times S_n$ containing actions for each of the n players is called an action profile.

Each player of a game acts by picking a specific pure strategy $s \in S_i$. A more general kind of behavior for a player is to play a *mixed strategy*. That is, the player selects a probability distribution over all pure strategies and acts by randomly picking the played action based on this distribution. In this case, players are rewarded with their expected payoff based on the selected mixed action profiles.

But what kinds of questions can one hope to answer in such settings? Game theorists are interested in predicting the outcomes of games by understanding what strategies the players will adopt in a given setting. The prototypical solution concept in games is an *equilibrium*, which describes an outcome in which players do not have an incentive to unilaterally deviate from their chosen strategies. The most popular such concept is the Nash Equilibrium introduced by John Forbes Nash Nash [1951], which prescribes players' behavior in a setting where no prior communication is allowed and players have to take into account the individual strategic behavior of others.

First consider the case where each player can only choose to play a pure strategy. Then an action profile is a *Pure* Nash Equilibrium if no player can increase their payoff by unilaterally deviating and choosing a different pure strategy. However, the applicability of the Pure Nash Equilibrium is limited, as there might exist games which do not have any such solution. Take for instance the Rock-Paper-Scissors game of example 2.1.1. It is not hard to see that in this case there does not exist any pure Nash equilibrium, since at least one of the two players will always be better off by deviating (eg. in the state S-P, the column player can always change to S-R and receive payoff 1 instead of -1). This motivates the definition of the *Mixed Nash Equilibrium*, in which players choose mixed strategies and no one has an incentive to unilaterally deviate from following their chosen mixed strategy. In a celebrated result Nash [1951], Nash proved that any finite n-player non-zero-sum game always has a mixed Nash equilibrium.

To see this more concretely, for the Rock-Paper-Scissors game of Example 2.1.1 there exists a unique mixed Nash equilibrium in which both players select between their actions with equal probability of $1/3$ each. As a second example, in the traffic light game (Example 2.1.2) there exist 3 Nash equilibria in total. These are shown in Figure 2.1. Notice that in every equilibrium, there exists at least one player that receives 0 utility.

1 \ 2	STOP	GO
STOP	0	1
GO	0	0

1 \ 2	STOP	GO
STOP	0	0
GO	1	0

1 \ 2	STOP	GO
STOP	98%	< 1%
GO	< 1%	0.01%

Figure 2.1: Nash equilibria for the traffic light game

In spite of its great success as a solution concept, the Nash Equilibrium is often not enough to capture realistic aspects of games such as coordination and furthermore, in a seminal result, Daskalakis et al. [2009] proved that it is computationally intractable to compute a Nash equilibrium unless P = PPAD. For these and other reasons, Game Theorists have turned to other generalized notions of equilibrium over the years. The Correlated Equilibrium (CE), first defined by Aumann [1974], allows players' strategies to be correlated with each other.

Take for instance the traffic light game from example 2.1.2. As we saw, the Nash equilibrium always forces at least one player to receive 0 utility in expectation. In reality, players are not restricted to act independently and can make use of the traffic light signals. Thus, in the traffic light game players correlate their strategies so that they do not perform the same action at the same time. The state reached in this case is called a Correlated Equilibrium, which extends the notion of Nash equilibrium beyond just product distributions of strategies. The traffic light acts as a "mediator", meaning that it is the coordination device that helps players achieve this Correlated Equilibrium.

Definition 2.1.4 (normal-form Correlated Equilibrium)**.** *In a n-player normal-form game, a Correlated Equilibrium is a joint distribution $\mu \in \Delta(S_1 \times \cdots \times S_n)$ such that for each player i, and every action $s_i^* \in S_i$ it holds that*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(\boldsymbol{x})] \geq \mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(s_i^*, \boldsymbol{x}_{-i}) \mid \boldsymbol{x}_i]$$

An intuitive way to think of the correlated equilibrium is to assume that there exists an external trusted mediator that draws an action profile $\boldsymbol{x} \sim \mu$ and then recommends each individual action $\boldsymbol{x}_i$ to player i. Furthermore, the joint distribution μ is common knowledge to all players of the game. Then, each player, after observing the recommended action, has the choice to either follow the recommendation or to deviate an act with a different action $s_i^* \in S_i$. In a correlated equilibrium, no player has an incentive to deviate from the action that was recommended by the mediator. This can describe behaviors where players infer the actions of others based only on their own private recommendation. This is, for example, what happens in the case of the traffic light game, where players can always infer what is the recommended action of the mediator (traffic light) to the other players.

Note that the set of correlated equilibria is a superset of the Nash equilibria, because Nash equilibria are simply product distributions of strategies. Consequently, by the existence of Nash equilibria it follows that every finite normal-form game always has a CE. Furthermore, it has been proven that, contrary to the Nash equilibrium, the CE can be computed in polynomial time in normal-form games [Foster and Vohra, 1997, Blum and Mansour, 2007, Papadimitriou and Roughgarden, 2008]. From these, we will present the algorithm of Blum and Mansour [2007] using uncoupled no-regret learning dynamics in section 3.3. The fact that learning agents repeatedly playing the game can converge to a correlated equilibrium is a fascinating aspect of this solution concept that further highlights its fundamental nature.

Another interesting equilibrium concept involving correlation is the Coarse Correlated Equilibrium (CCE) defined below.

Definition 2.1.5 (normal-form Coarse Correlated Equilibrium)**.** *In a n-player normal-form game, a Coarse Correlated Equilibrium is a joint distribution $\mu \in \Delta(S_1 \times \cdots \times S_n)$ such that for each player i, and every action $s_i^* \in S_i$ it holds that*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(\boldsymbol{x})] \geq \mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(s_i^*, \boldsymbol{x}_{-i})]$$

Coarse correlated equilibria (CCE) can be interpreted in a similar way to CEs, by assuming that a trusted mediator recommends actions to players based on a commonly known joint distribution on action profiles. The distinguishing difference between the two is that in the case of CCE it should not be beneficial for players to deviate even before observing the recommended action. Of course, this immediately implies that the set of CE is a subset of the set of all CCE which makes it even easier to efficiently compute them.

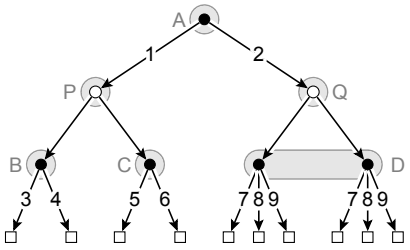
2.2 Extensive-Form Games

While normal-form games (NFGs) correspond to nonsequential interactions, such as Rock-Paper-Scissors, where players simultaneously pick one action and then receive a payoff based on what others picked, extensive-form games (EFGs) model games that are played on a game tree. They capture both sequential and simultaneous moves, as well as private information and are therefore a very general and expressive model of games, capturing chess, go, poker, sequential auctions, and many other settings as well. We now recall basic properties and notation for EFGs.

Game tree In an n-player extensive-form game, each node in the game tree is associated with exactly one player from the set $\{1, \ldots, n\} \cup \{c\}$, where the special player c—called the *chance* player—is used to model random stochastic outcomes, such as rolling a die or drawing cards from a deck. Edges leaving from a node represent actions that a player can take at that node. To model private information, the game tree is supplemented with an information partition, defined as a partition of nodes into sets called information sets. Each node belongs to exactly one information set, and each information set is a nonempty set of tree nodes for the same Player i. An information set for Player i denotes a collection of nodes that Player i cannot distinguish among, given what she has observed so far. (We remark that all nodes in a same information set must have the same set of available actions, or the player would distinguish the nodes). The set of all information sets of Player i is denoted $\mathcal{J}_i$. In this paper, we will only consider *perfect-recall* games, that is, games in which the information sets are arranged in accordance with the fact that no player forgets what the player knew earlier,.

Sequence-form strategies Since nodes belonging to the same information set for a player are indistinguishable to that player, the player must play the same strategy at each of the nodes. Hence, a strategy for a player is exactly a mapping from an *information set* to a distribution over actions. In other words, it is the information sets, and not the game tree nodes, that capture the decision points of the player. We can then represent a strategy for a generic player i as a vector indexed by each valid information set-action pair (j, a). Any such valid pair is called a *sequence* of the player; the set of all sequences is denoted as $\Sigma_i \coloneqq \{(j, a) : j \in \mathcal{J}_i, a \in \mathcal{A}_j\} \cup \{\varnothing\}$, where the special element $\varnothing$ is called *empty sequence*. Given an information set $j \in \mathcal{J}_i$, we denote by p_j the parent sequence of j, defined as the last pair $(j, a) \in \Sigma_i$ encountered on the path from the root to any node $v \in j$; if no such pair exists we let $p_j = \varnothing$. Finally, we denote by $\mathcal{C}_\sigma$ the children of sequence $\sigma \in \Sigma_i$, defined as the information sets $j \in \mathcal{J}_i$ for which $p_j = \sigma$. Sequences σ for which $\mathcal{C}_\sigma$ is an empty set are called *terminal*; the set of all terminal sequences is denoted $\Sigma_i^{\perp}$.

Example 2.2.1. *Consider the tree-form decision process faced by Player 1 in the small game of Figure 2.2 (Left). The decision process has four decision nodes $\mathcal{J}_1 = \{\mathsf{A}, \mathsf{B}, \mathsf{C}, \mathsf{D}\}$ and nine sequences including the empty sequence $\varnothing$. For decision node D, the parent sequence is $p_{\mathsf{D}} = \mathsf{A2}$; for B and C it is $p_{\mathsf{B}} = \mathsf{A1}$; for A it is the empty sequence $p_{\mathsf{A}} = \varnothing$.*



Sequence-form constraints:

$$\begin{cases} \boldsymbol{x}[\varnothing] = 1, \\ \boldsymbol{x}[\mathsf{A1}] + \boldsymbol{x}[\mathsf{A2}] = \boldsymbol{x}[\varnothing], \\ \boldsymbol{x}[\mathsf{B3}] + \boldsymbol{x}[\mathsf{B4}] = \boldsymbol{x}[\mathsf{A1}], \\ \boldsymbol{x}[\mathsf{C5}] + \boldsymbol{x}[\mathsf{C6}] = \boldsymbol{x}[\mathsf{A1}], \\ \boldsymbol{x}[\mathsf{D7}] + \boldsymbol{x}[\mathsf{D8}] + \boldsymbol{x}[\mathsf{D9}] = \boldsymbol{x}[\mathsf{A2}]. \end{cases}$$

Figure 2.2: (Left) Tree-form decision process considered in the example. Black round nodes belong to Player 1; white round nodes to Player 2. Square white nodes are terminal nodes in the game tree, payoffs are omitted. Gray bags denote information sets. (Right) The constraints that define the sequence-form polytope $\mathcal{Q}_1$ for Player 1 (besides nonnegativity).

A *reduced-normal-form plan* for Player i represents a deterministic strategy for the player as a vector $\boldsymbol{x} \in \{0,1\}^{\Sigma_i}$ where the entry corresponding to the generic sequence $\boldsymbol{x}[ja]$ is equal to 1 if the player plays action a at (the nodes of) information set $j \in \mathcal{J}_i$. Information sets that cannot be reached based on the strategy do not have any action select. A crucial property of the reduced-normal-form plan representation of deterministic strategies is the fact that the utility of any player is a multilinear function in the profile of reduced-normal-form plans played by the players. The set of all reduced-normal-form plans of Player i is denoted with the symbol Π_i. Typically, the cardinality of Π_i is exponential in the size of the game tree.

The convex hull of the set of reduced-normal-form plans of Player i is called the *sequence-form polytope* of the player, and denoted with the symbol $\mathcal{Q}_i := \text{conv}(\Pi_i)$. It represents the set of all randomized strategies in the game. An important result by Romanovskii [1962], Koller et al. [1996], von Stengel [1996] shows that $\mathcal{Q}_i$ can be captured by polynomially many constraints in the size of the game tree, as we recall next.

Definition 2.2.2. *The* polytope of sequence-form strategies *of Player i is equal to the convex polytope*

$$\mathcal{Q}_i := \left\{ \boldsymbol{x} \in \mathbb{R}^{\Sigma}_{\geq 0} : \begin{array}{ll} (2.1) & \boldsymbol{x}[\varnothing] = 1 \\ (2.2) & \sum_{a \in \mathcal{A}_j} \boldsymbol{x}[ja] = \boldsymbol{x}[p_j] \quad \forall\, j \in \mathcal{J} \end{array} \right\}.$$

As an example, the constraints that define the sequence-form polytope for Player 1 in the game of Figure 2.2 (Left) are shown in Figure 2.2 (Right). The polytope of sequence-form strategies possesses a strong combinatorial structure that enables speeding up several common optimization procedures and will be crucial in developing efficient algorithms to converge to equilibrium.

To further highlight the importance of the sequence-form representation we note that all EFGs have an equivalent representation as normal-form games. Namely, the actions of each player i are the reduced normal-form plans Π_i. This means that all equilibrium concepts from normal-form games transfer to extensive-form games as well. An EFG can have a normal-form Correlated Equilibrium, or Nash Equilibrium. However, as explained earlier, the size of the equivalent normal-form game for an EFG might be exponentially larger than the size of the game tree and, furthermore, the standard normal-form equilibrium concepts might fail to exploit the sequential nature of these games to achieve a goal such as maximizing the social welfare of the game.

The *extensive-form correlated equilibrium (EFCE)*, first introduced by von Stengel and Forges [2008], is a natural extension of the CE for extensive-form games. Like in the normal-form CE, the mediator in an EFCE selects an action profile from a joint probability distribution. Unlike in the CE, the recommendations are not fully revealed to the players since the beginning. That is, players are not recommended a full reduced normal-form plan with a specified action at each information set. Rather, the mediator sequentially reveals the recommended action at each individual information

set only when the player reaches that information set. Furthermore, if the player decides to deviate from the recommended action then they receive no further recommendations from the mediator. This allows for a richer set of equilibria (a superset of the normal-form correlated equilibria), which include solutions with higher social welfare than what is achievable by employing a normal-form view of the game. The signaling game in Example 6.1.1 is a notable example in which an EFCE achieves a far better outcome than a CE, similarly to how the CE achieved a better outcom than the Nash equilibrium in the traffic light game (Example 2.1.2).

Another important aspect of the EFCE is that it is efficiently computable, both in a centralized manner via a variation of the Ellipsoid Against Hope algorithm [Huang and von Stengel, 2008] and, as was recently proved, in a decentralized manner using uncoupled no-regret dynamics [Farina et al., 2022b]. This is currently an additional advantage of the EFCE over the normal-form CE, which is not yet known whether it can be efficiently computed or learned in extensive-form games. Similarly to normal-form games, there exist efficiently computable equilibrium concepts for extensive-form games that are even coarser, such as the extensive-form coarse-correlated equilibrium [Farina et al., 2020]. However, it is a major challenge to understand what is the strongest notion of equilibrium in extensive-form games that can be efficiently computed. Morrill et al. [2021] defined a Deviation Landscape with their corresponding equilibria that aim to generalize the behavioral deviations prescribing the EFCE. This thesis explores yet another way of generalizing the concept of EFCE by considering all linear deviations, as we describe starting from Chapter 4.

2.3 Additional Extensive-Form Game Notation

In the proofs, we will make use of the following symbols and notation.

Symbol	Description
$\mathcal{J}_i$	Set of all Player i's infosets.
$\mathcal{A}_j$	Set of actions available at any node in the information set j.
Σ_i	Set of sequences for Player i, defined as $\Sigma^{(i)} \coloneqq \{(j,a) : j \in \mathcal{J}, a \in A_j\} \cup \{\varnothing\}$,
$\varnothing$	where the special element $\varnothing$ is called the *empty sequence*.
$\Sigma_i^{\perp}$	Set of terminal sequences for Player i.
p_j	Parent sequence of j, defined as the last pair $(j,a) \in \Sigma_i$ encountered on the path from the root to any information set j.
$\mathcal{C}_\sigma$	Set of all "children" of sequence σ, defined as the information sets $j \in \mathcal{J}$ having as parent $p_j = \sigma$.
$j' \prec j$	Information set $j \in \mathcal{J}$ is an ancestor of $j' \in \mathcal{J}$, that is, there exists a path in the game tree connecting a node $h \in j$ to some node $h' \in j'$.
$\sigma \prec \sigma'$	Sequence σ precedes sequence σ', where σ, σ' belong to the same player.
$\sigma \succeq j$	Sequence $\sigma = (j', a')$ is such that $j' \succeq j$.
$\Sigma_{\succeq j}$	Sequences at $j \in \mathcal{J}$ and all of its descendants, $\Sigma_{\succeq j} \coloneqq \{\sigma \in \Sigma : \sigma \succeq j\}$.
$\mathcal{Q}_i$	Sequence-form strategies of Player i (Definition 2.2.2).
$\mathcal{Q}_{\succeq j}$	Sequence-form strategies for the subtree rooted at $j \in \mathcal{J}$ (Definition 2.3.1).
Π_i	Reduced-normal-form plans (a.k.a. deterministic sequence-form strategies) of Player i.
$\Pi_{\succeq j}$	Reduced-normal-form plans (a.k.a. deterministic sequence-form strategies) for the subtree rooted at $j \in \mathcal{J}$.

Table 2.1: Summary of game-theoretic notation used in this paper. Note that we might skip player-specific subscripts when they can be inferred.

Furthermore, when the subscript referring to players can be inferred or is irrelevant (that is, the quantities are referred to a generic player), then we might skip it.

As hinted by some of the rows in the above table, we will sometimes find it important to consider partial strategies that only specify behavior at a decision node j and all of its descendants $j' \succ j$. We make this formal through the following definition.

Definition 2.3.1. *The* set of sequence-form strategies for the subtree rooted at j*, denoted* $\mathcal{Q}_{\succeq j}$*, is the set of all vectors* $\boldsymbol{x} \in \mathbb{R}_{\geq 0}^{\Sigma_{\succeq j}}$ *such that probability-mass-conservation constraints hold at decision node* j *and all of its descendants* $j' \succ j$*, specifically*

$$\mathcal{Q}_{\succeq j} \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}_{\geq 0}^{\Sigma_{\succeq j}} : \begin{array}{ll} (2.3) & \sum_{a \in \mathcal{A}_j} \boldsymbol{x}[ja] = 1 \\ (2.4) & \sum_{a \in \mathcal{A}_{j'}} \boldsymbol{x}[j'a] = \boldsymbol{x}[p_{j'}] \quad \forall\, j' \succ j \end{array} \right\}.$$

Finally, we define the symbol $\Sigma_i^{\perp}$ to be the set of all terminal sequences for Player i. Thus, the set $\Sigma_i \setminus \Sigma_i^{\perp}$ would give us all the non-terminal sequences of that player.

Access to coordinates By definition, sequence-form strategies are vectors indexed by sequences. To access the coordinate corresponding to sequence σ, we will use the notation $\boldsymbol{x}[\sigma]$. Occasionally, we will need to extract a subvector corresponding to all sequences that are successor of an information set j, that is, all sequences $\sigma \succeq j$. For that, we use the notation $\boldsymbol{x}[\succeq j]$.

Remark on the structure of sequence-form strategies We further remark the following known fact about the structure of sequence-form strategies. Intuitively, it crystallizes the idea that sequence-form strategies encode product of probabilities of actions on the path from the root to any decision point. The proof follows directly from the definitions.

Lemma 2.3.2. *Let $j \in \mathcal{J}_i$ be an information set for a generic player. Then, given any sequence-form strategy $\boldsymbol{x} \in \mathcal{Q}_{\succeq j}$, action $a \in \mathcal{A}_j$, and child information set $j' \in \mathcal{C}_{ja}$, there exists a sequence-form strategy $\boldsymbol{x}_{\succeq j'} \in \mathcal{Q}_{\succeq j'}$ such that*

$$\boldsymbol{x}[\succeq j'] = \boldsymbol{x}[ja]\boldsymbol{x}_{\succeq j'}.$$

Chapter 3

Hindsight Rationality and Learning in Games

3.1 The Online Learning framework

Games are one of many situations in which a decision-maker has to act in an online manner. For these situations, the most widely used protocol is that of Online Learning (e.g., see Orabona [2022]). Specifically, the learner has a set of actions or behavior they can employ $\mathcal{X} \subseteq \mathbb{R}^d$ (in extensive-form games, this would typically be the set of reduced-normal-form strategies). At each timestep t the learner first selects, possibly at random, an element $\boldsymbol{x} \in \mathcal{X}$, and then receives a loss (opposite of utility) function $\ell^{(t)} : \mathcal{X} \mapsto \mathbb{R}$ from an adversary. Both the learner and the adversary observe the history of all prior actions and losses chosen in the past. A widely adopted objective for the learner is that of ensuring vanishing average *regret* with high probability. Regret is defined as the difference between the loss the learner cumulated through their choice of behavior, and the loss they would have cumulated had they chosen the best fixed action in hindsight. More formally, the *realized* regret is defined as

$$\text{Reg}^{(T)} := \max_{\boldsymbol{x} \in \mathcal{X}} \sum_{t=1}^{T} \left(\ell^{(t)}(\boldsymbol{x}^{(t)}) - \ell^{(t)}(\boldsymbol{x}) \right)$$

where T is the time horizon over which the learner acts. Note that, in general, this quantity might be a random variable. Often, we are interested in the *expected* regret, which is the expected value of the realized regret. Additionally, the time-averaged regret is just the regret $\text{Reg}^{(T)}$ divided by T. Thus, the objective of an online learner is to ensure that its regret is *sublinear* with respect to T or, equivalently, that its time-averaged regret is $o(1)$.

In this work we are interested in applying the online learning framework to players of extensive-form games. Since as we observed the utilities in EFGs are linear in each player's reduced-normal-form plans, for the rest of the paper we focus on the case in which the loss function $\ell^{(t)}$ is linear, that is, of the form $\ell^{(t)} : \boldsymbol{x} \mapsto \langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle$. Thus, we only care about the setting of so called *Online Linear Optimization*. To construct our main algorithm in chapter 4, we use the results from this section.

Online Projected Gradient Descent. In the rest of this section we will present the Online Projected Gradient Descent, one of the most widely used algorithms for online convex optimization, which is an even more general setting than online linear optimization.

Algorithm 2: Online Projected Gradient Descent

Data: Non-empty and closed convex set $V \subseteq \mathbb{R}^d$, $\boldsymbol{x}^{(1)} \in V$, learning rates $\eta^{(t)} > 0$

for $t = 1, 2, \dots T$ **do**

- Output $\boldsymbol{x}^{(t)} \in V$
- Receive $\ell^{(t)} : V \to \mathbb{R}$ and pay $\ell^{(t)}(\boldsymbol{x}^{(t)})$
- Set $\boldsymbol{g}^{(t)} = \nabla \ell^{(t)}(\boldsymbol{x}^{(t)})$
- $\boldsymbol{x}^{(t+1)} = \Pi_V(\boldsymbol{x}^{(t)} - \eta^{(t)} \boldsymbol{g}^{(t)}) = \arg\min_{\boldsymbol{y} \in V} \|\boldsymbol{x}^{(t)} - \eta^{(t)} \boldsymbol{g}^{(t)} - \boldsymbol{y}\|_2$

We will use the following important Projection Lemma directly without proof. It states that the Euclidean projections of a point $\boldsymbol{x} \in \mathbb{R}^d$ to a convex set $V \subseteq \mathbb{R}^d$ always decrease the distance from all points of the set.

Lemma 3.1.1 (Proposition 2.11 from Orabona [2022])**.** *Let $\boldsymbol{x} \in \mathbb{R}^d$ and $\boldsymbol{y} \in V$, where $V \subseteq \mathbb{R}^d$ is a non-empty closed convex set and define $\Pi_V(\boldsymbol{x}) := \arg\min_{\boldsymbol{z} \in V} \|\boldsymbol{x} - \boldsymbol{z}\|_2$. Then, $\|\Pi_V(\boldsymbol{x}) - \boldsymbol{y}\|_2 \leq \|\boldsymbol{x} - \boldsymbol{y}\|_2$.*

Lemma 3.1.2. *Let $V \subseteq \mathbb{R}^d$ be a non-empty convex set and $\ell^{(t)} : V \to \mathbb{R}$ be a convex function that is differentiable in an open set containing V. Set $\boldsymbol{g}^{(t)} = \nabla \ell^{(t)}(\boldsymbol{x}^{(t)})$. Then, for all $\boldsymbol{u} \in V$, the following inequality holds*

$$\eta^{(t)}(\ell^{(t)}(\boldsymbol{x}^{(t)}) - \ell^{(t)}(\boldsymbol{u})) \leq \eta^{(t)}\langle \boldsymbol{g}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{u}\rangle \leq \frac{1}{2}\|\boldsymbol{x}^{(t)} - \boldsymbol{u}\|_2^2 - \frac{1}{2}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{u}\|_2^2 + \frac{(\eta^{(t)})^2}{2}\|\boldsymbol{g}^{(t)}\|_2^2$$

Proof. Recall [Rockafellar, 1970] that since $\ell^{(t)}$ are convex differentiable functions, it holds that

$$\ell^{(t)}(\boldsymbol{y}) \geq \ell^{(t)}(\boldsymbol{x}) + \langle \nabla \ell^{(t)}(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle, \forall \boldsymbol{x}, \boldsymbol{y} \in V,$$

which immediately gives the first inequality. For the second inequality we have

$$\begin{aligned}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}^{(t)} - \boldsymbol{u}\|_2^2 &\leq \|\boldsymbol{x}^{(t)} - \eta^{(t)}\boldsymbol{g}^{(t)} - \boldsymbol{u}\|_2^2 - \|\boldsymbol{x}^{(t)} - \boldsymbol{u}\|_2^2 \\ &= -2\eta^{(t)}\langle \boldsymbol{g}^{(t)}, \boldsymbol{x}^{(t)} - \boldsymbol{u}\rangle + (\eta^{(t)})^2\|\boldsymbol{g}^{(t)}\|_2^2\end{aligned}$$

where we used Lemma 3.1.1 in the first line. Reordering, we get the second inequality and the proof is complete. □

Theorem 3.1.3 (Theorem 2.13 from Orabona [2022])**.** *Let $V \subseteq \mathbb{R}^d$ be a non-empty closed convex set with diameter D. Let $\ell^{(t)} : V \to \mathbb{R}$ for $t = 1, \ldots, T$ be an arbitrary sequence of convex functions that are differentiable in open sets containing V. Pick any $\boldsymbol{x}^{(1)} \in V$ and assume $\eta^{(t+1)} \leq \eta^{(t)}$ for all $1 \leq t < T$. Then for all $\boldsymbol{u} \in V$, it holds*

$$\sum_{t=1}^{T}(\ell^{(t)}(\boldsymbol{x}^{(t)}) - \ell^{(t)}(\boldsymbol{u})) \leq \frac{D^2}{2\eta^{(T)}} + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2.$$

Proof. To bound the regret, we begin by using the inequality of Lemma 3.1.2 for each iteration of the algorithm

$$\sum_{t=1}^{T}(\ell^{(t)}(\boldsymbol{x}^{(t)}) - \ell^{(t)}(\boldsymbol{u})) \leq \sum_{t=1}^{T}\left(\frac{1}{2\eta^{(t)}}\|\boldsymbol{x}^{(t)} - \boldsymbol{u}\|_2^2 - \frac{1}{2\eta^{(t)}}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{u}\|_2^2\right) + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2.$$

We now subtract and add $\frac{1}{2\eta^{(t+1)}}\|\boldsymbol{x}^{(t+1)} - \boldsymbol{u}\|_2^2$ for each t and compute the appearing telescoping sum to get

$$\frac{1}{2\eta^{(1)}}\|\boldsymbol{x}^{(1)} - \boldsymbol{u}\|_2^2 - \frac{1}{2\eta^{(T)}}\|\boldsymbol{x}^{(T)} - \boldsymbol{u}\|_2^2 + \sum_{t=1}^{T-1}\left(\frac{1}{2\eta^{(t+1)}} - \frac{1}{2\eta^{(t)}}\right)\|\boldsymbol{x}^{(t+1)} - \boldsymbol{u}\|_2^2 + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2.$$

By applying the assumption about the diameter D it follows that this expression is at most

$$\begin{aligned}&\frac{1}{2\eta^{(1)}}D^2 + D^2\sum_{t=1}^{T-1}\left(\frac{1}{2\eta^{(t+1)}} - \frac{1}{2\eta^{(t)}}\right) + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2 \\ =&\frac{1}{2\eta^{(1)}}D^2 + D^2\left(\frac{1}{2\eta^{(T)}} - \frac{1}{2\eta^{(1)}}\right) + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2 \\ =&\frac{D^2}{2\eta^{(T)}} + \sum_{t=1}^{T}\frac{\eta^{(t)}}{2}\|\boldsymbol{g}^{(t)}\|_2^2.\end{aligned}$$

□

If we additionally assume that we have an upper bound L on the L2 norm of all gradients, then by setting

$$\eta^{(1)} = \cdots = \eta^{(T)} = \frac{D}{L\sqrt{T}}$$

we achieve a regret bound of $DL\sqrt{T}$, which is sublinear with respect to T as was our goal.

3.2 Learning in Games and Φ-regret minimization

In this section we explore what happens if all agents in a repeated normal-form game independently employ an online no-regret learning algorithm. Does this natural acting process converge to some specific solution for the game? It turns out that, indeed, the empirical frequency of play arising from these uncoupled regret dynamics converges almost surely to a Coarse Correlated Equilibrium. Actually, a more fine-grained result holds. If after T steps, the time-averaged regret of all agents is at most ϵ, then the empirical frequency of play converges to an ϵ-approximate CCE, as shown below.

Theorem 3.2.1. *Consider a normal-form game of n players with pure strategies $S_1, \ldots, S_n$. Assume that players repeatedly play the game and each one is acting based on a no-regret algorithm. If after T steps, the time-averaged regret is at most ϵ, then the empirical frequency of play is an ϵ-approximate Coarse Correlated Equilibrium $\mu \in \Delta(S_1 \times \cdots \times S_n)$. Specifically, for each player i, and every pure strategy $s^* \in S_i$ it holds*

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(\boldsymbol{x})] \geq \mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(s^*, \boldsymbol{x}_{-i})] - \epsilon.$$

That is, no player can gain more than ϵ expected utility by unilaterally deviating.

Proof. The condition of the CCE for each player i and action $s^* \in S_i$ can be equivalently written as

$$\mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(s^*, \boldsymbol{x}_{-i}) - u_i(\boldsymbol{x})] \leq \epsilon. \tag{3.1}$$

Now, in the interaction of the n no-regret learners, assume that $\boldsymbol{a}^{(t)} \in S_1 \times \cdots \times S_n$ are the action profiles played at each repetition of the game $t = 1, \ldots, T$. Then the distribution μ will be the empirical frequency of play after T steps

$$\mu = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{a}^{(t)}. \tag{3.2}$$

If after T steps the time-averaged regret for each player i is at most ϵ then for all $s^* \in S_i$ we get

$$\mathrm{Reg}_i^{(T)} = \frac{1}{T}\sum_{t=1}^{T}\left(\ell_i^{(t)}(\boldsymbol{a}_i^{(t)}) - \ell_i^{(t)}(s^*)\right) \leq \epsilon$$

where $\ell_i^{(t)}(z) = -u_i(z, \boldsymbol{a}_{-i}^{(t)})$. Consequently, we can write this regret expression as follows

$$\begin{aligned}\mathrm{Reg}_i^{(T)} &= \frac{1}{T}\sum_{t=1}^{T}\left(u_i(s^*, \boldsymbol{a}_{-i}^{(t)}) - u_i(\boldsymbol{a}^{(t)})\right) \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mu}[u_i(s^*, \boldsymbol{x}_{-i}) - u_i(\boldsymbol{x})]\end{aligned}$$

where the last equation follows from (3.1) and (3.2). Thus, we conclude that when $\mathrm{Reg}_i^{(T)} \leq \epsilon$ the empirical frequency of play μ is an ϵ-approximate Coarse Correlated Equilibrium. □

However, as we explained in the previous chapter, the Coarse Correlated Equilibrium is a rather weak solution concept. Can we hope to construct online learning algorithms that converge to stronger solution concepts, such as the Correlated Equilibrium in normal-form games? It seems unlikely that the regret is a suitable objective to aim for stronger solution concepts. Instead, we will generalize the notion of regret to that of Φ-regret, which defines a spectrum of stronger objectives for the online learner with respect to transformations of the previously played strategies. In particular, let Φ be a desired set of strategy transformations $\phi : \mathcal{X} \to \mathcal{X}$ that the learner might want to learn not to regret. Then, the learner's Φ-regret is defined as the quantity

$$\Phi\text{-Reg}^{(T)} := \max_{\phi \in \Phi} \sum_{t=1}^{T} \left(\langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle - \langle \boldsymbol{\ell}^{(t)}, \phi(\boldsymbol{x}^{(t)}) \rangle \right)$$

A *no-Φ-regret algorithm* (also known as a Φ-regret minimizer) is one that, at all times T, guarantees with high probability that $\Phi\text{-Reg}^{(T)} = o(T)$ no matter what is the sequence of losses revealed by the environment. The size of the set Φ of strategy transformations defines a natural measure of rationality (sometimes called *hindsight rationality*) for players, and several choices have been discussed in the literature. Clearly, as Φ gets larger, the learner becomes more rational. On the flip side, guaranteeing sublinear regret with respect to all transformations in the chosen set Φ might be intractable in general. On one end of the spectrum, perhaps the smallest meaningful choice of Φ is the set of all *constant* transformations $\Phi^{\text{const}} = \{\phi_{\hat{\boldsymbol{x}}} : \boldsymbol{x} \mapsto \hat{\boldsymbol{x}}\}_{\hat{\boldsymbol{x}} \in \mathcal{X}}$. In this case, Φ^{const}-regret is also called *external regret* and has been extensively studied in the field of online convex optimization. On the other end of the spectrum, *swap regret* corresponds to the setting in which Φ is the set of *all* transformations $\mathcal{X} \to \mathcal{X}$. Somewhat intermediate, and of central importance in this paper, is the notion of *linear-swap regret*, which corresponds to the case in which

$$\Phi := \{\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x} : \mathbf{A} \in \mathbb{R}^{d \times d}, \text{ with } \mathbf{A}\boldsymbol{x} \in \mathcal{X} \quad \forall\, \boldsymbol{x} \in \mathcal{X}\} \qquad \text{(linear-swap deviations)}$$

is the set of all linear transformations from $\mathcal{X}$ to itself.[1]

An important observation is that when all considered deviation functions in Φ are linear, an algorithm guaranteeing sublinear no-Φ regret for the set $\mathcal{X}$ can be constructed immediately from a deterministic no-Φ-regret algorithm for $\mathcal{X}' = \Delta(\mathcal{X})$ by sampling $\mathcal{X} \ni \boldsymbol{x}$ from any $\boldsymbol{x}' \in \mathcal{X}'$ so as to guarantee that $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{x}'$. Since this is exactly the setting we study in this paper, this folklore observation (see also Farina et al. [2022b]) enables us to focus on the following problem: does a deterministic no-Φ-regret algorithm for the set of sequence-form strategies $\mathcal{X} = \mathcal{Q}_i$ of any player in an extensive-form game, with guaranteed sublinear Φ-regret in the worst case, exist? In this paper we answer the question for the positive.

From regret to equilibrium The setting of Learning in Games refers to the situation in which all players employ a learning algorithm, receiving as loss the negative of the gradient of their own utility evaluated in the strategies output by all the other players. A fascinating aspect of no-Φ-regret learning dynamics is that if each player of a game employs a no-Φ-regret algorithm, then the empirical frequency of play converges almost surely to the set of Φ-equilibria, which are notions of correlated equilibria, in which the rationality of players is bounded by the size of the set Φ. Formally, for a set Φ of strategy deviations, a Φ-equilibrium is defined as follows.

Definition 3.2.2. *For a n-player extensive-form game G and a set Φ_i of deviations for each player, a $\{\Phi_i\}$-equilibrium is a joint distribution $\mu \in \Delta(\Pi_1 \times \cdots \times \Pi_n)$ such that for each player i, and every deviation $\phi \in \Phi_i$ it holds that*

$$\mathbb{E}_{\boldsymbol{x} \sim \mu}[u_i(\boldsymbol{x})] \geq \mathbb{E}_{\boldsymbol{x} \sim \mu}[u_i(\phi(\boldsymbol{x}_i), \boldsymbol{x}_{-i})]$$

That is, no player i has an incentive to unilaterally deviate from the recommended joint strategy $\boldsymbol{x}$ using any transformation $\phi_i \in \Phi_i$.

[1] For the purposes of this paper, the adjective *linear* refers to the fact that each transformation can be expressed in the form $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ for an appropriate matrix $\mathbf{A}$.

This general framework captures several important notions of equilibrium across a variety of game theoretic models. For example, in both NFGs and EFGs, no-external regret dynamics converge to the set of Coarse Correlated Equilibria. In NFGs, no-swap regret dynamics converge to the set of Correlated Equilibria [Blum and Mansour, 2007]. In EFGs, Farina et al. [2022b] recently proved that a specific subset Φ of linear transformations called *trigger deviations* lead to the set of EFCE.

Reducing Φ-regret to external regret An elegant construction by Gordon et al. [2008] enables constructing no-Φ-regret algorithms for a generic set $\mathcal{X}$ starting from a no-external-regret algorithm for Φ. We briefly recall the result.

Theorem 3.2.3 (Gordon et al. [2008])**.** *Let $\mathcal{R}$ be an external regret minimizer having the set of transformations Φ as its action space, and achieving sublinear external regret $\text{Reg}^{(T)}$. Additionally, assume that for all $\phi \in \Phi$ there exists a fixed point $\phi(\boldsymbol{x}) = \boldsymbol{x} \in \mathcal{X}$. Then, a Φ-regret minimizer $\mathcal{R}_\Phi$ can be constructed as follows:*

- *To output a strategy $\boldsymbol{x}^{(t)}$ at iteration t of $\mathcal{R}_\Phi$, obtain an output $\phi^{(t)} \in \Phi$ of the external regret minimizer $\mathcal{R}$, and return one of its fixed points $\boldsymbol{x}^{(t)} = \phi^{(t)}(\boldsymbol{x}^{(t)})$.*
- *For every linear loss function $\ell^{(t)}$ received by $\mathcal{R}_\Phi$, construct the linear function $L^{(t)} : \phi \mapsto \ell^{(t)}(\phi(\boldsymbol{x}^{(t)}))$ and pass it as loss to $\mathcal{R}$.*

Let $\Phi\text{-Reg}^{(T)}$ be the Φ-regret of $\mathcal{R}_\Phi$. Under the previous construction, it holds that

$$\Phi\text{-Reg}^{(T)} = \text{Reg}^{(T)} \qquad \forall\, T = 1, 2, \dots$$

Thus, if $\mathcal{R}$ is an external regret minimizer then $\mathcal{R}_\Phi$ is a Φ-regret minimizer.

3.3 A No-Swap Regret Algorithm for Normal-Form Games

Theorem 3.3.1 (Theorem 2 from Blum and Mansour [2007])**.** *In the online learning from K experts problem, for any algorithm with sublinear external regret $R(T)$, there exists an algorithm with sublinear swap regret at most $KR(T)$.*

Proof. To construct the no-swap regret algorithm, we create K copies $\mathcal{R}_1, \dots, \mathcal{R}_K$ of the no-external regret algorithm. Conceptually, the i-th copy $\mathcal{R}_i$ will be responsible for minimizing regret with respect to all swaps of the form $i \to j$. Then, the algorithm is as follows.

At each timestep $t = 1, 2, \dots, T$:

(a) Receive the output distributions $\boldsymbol{q}_1^{(t)}, \dots, \boldsymbol{q}_K^{(t)}$ from the K algorithms $\mathcal{R}_1, \dots, \mathcal{R}_K$.

(b) Then combine these into a single final distribution $\boldsymbol{p}^{(t)}$ for this step.

(c) Receive a loss vector $\boldsymbol{\ell}^{(t)}$.

(d) Give to each algorithm $\mathcal{R}_i$ the loss vector $\boldsymbol{p}_i^{(t)} \boldsymbol{\ell}^{(t)}$.

We want to select a suitable construction of the distribution $\boldsymbol{p}^{(t)}$ and guarantee that the final algorithm has small swap regret. In other words, we want to bound the expected loss of the central algorithm

$$\sum_{t=1}^{T} \sum_{j=1}^{K} \boldsymbol{p}_j^{(t)} \cdot \boldsymbol{\ell}_j^{(t)} \tag{3.3}$$

with the expected loss under a deviation function $\phi : [K] \to [K]$, which is

$$\sum_{t=1}^{T} \sum_{j=1}^{K} \boldsymbol{p}_j^{(t)} \cdot \boldsymbol{\ell}_{\phi(j)}^{(t)} \tag{3.4}$$

If we focus our attention to a single algorithm $\mathcal{R}_i$, then by its regret guarantee and the losses it receives, we get that for all actions j^*

$$\sum_{t=1}^{T}\sum_{j=1}^{K} \boldsymbol{q}_{i,j}^{(t)} \cdot (\boldsymbol{p}_i^{(t)} \boldsymbol{\ell}_j^{(t)}) \leq \sum_{t=1}^{T} \boldsymbol{p}_i^{(t)} \boldsymbol{\ell}_{j^*}^{(t)} + R(T).$$

Now fix a specific deviation function ϕ. Our aim is to construct the expression of (3.4). To this end, we add up the previous inequalities for all algorithms $\mathcal{R}_1, \dots, \mathcal{R}_K$ and for each $\mathcal{R}_i$ we set $j^* = \phi(i)$. This gives

$$\sum_{t=1}^{T}\sum_{i=1}^{K}\sum_{j=1}^{K} \boldsymbol{q}_{i,j}^{(t)} \cdot (\boldsymbol{p}_i^{(t)} \boldsymbol{\ell}_j^{(t)}) \leq \sum_{t=1}^{T}\sum_{i=1}^{K} \boldsymbol{p}_i^{(t)} \boldsymbol{\ell}_{\phi(i)}^{(t)} + KR(T).$$

Finally, it remains to select suitable distributions $\boldsymbol{p}^{(t)}$. To make the last expression match the desired expression of swap regret, we would like its left-hand side to equal the expected loss of our algorithm from (3.3). Thus, combining these for each $t = 1, \dots, T$ and each $j = 1, \dots, K$ we get

$$\boldsymbol{p}_j^{(t)} = \sum_{i=1}^{K} \boldsymbol{q}_{i,j}^{(t)} \cdot \boldsymbol{p}_i^{(t)}.$$

We can observe that in this last equation, the matrix having $\boldsymbol{q}_{i,j}^{(t)}$ as entries is a right-stochastic matrix or, in other words, the transition matrix of a Markov chain. Then, the equation requires that the distribution $\boldsymbol{p}^{(t)}$ is a stationary distribution of this Markov chain. It is well-known that at least one such distribution always exists, and can be computed efficiently as an eigenvector of the transition matrix. This completes the proof and the construction of the no-swap regret algorithm. □

At this point it is worth noting the resemblance of this algorithm to the construction of Theorem 3.2.3. This is not a coincidence, since the Φ-regret minimization framework can be seen as a generalization of the described algorithm. The set Φ in this case is the set of all right-stochastic matrices and the construction with the K regret minimizers is an algorithm that minimizes regret with respect to these matrices.

We conclude this chapter by briefly mentioning a line of work that seeks to improve the rate of convergence of no-regret learning algorithms. All algorithms that we presented in this chapter achieve a rate of convergence of $O(1/\sqrt{T})$ and guarantee robustness even against fully adversarial environments. The same holds for the no-swap regret learning algorithm presented previously, as it makes use of a no-external-regret algorithm. Furthermore, this convergence rate of $O(1/\sqrt{T})$ is known to be tight in fully adversarial environments.

However, in games consisting entirely of competing no-regret learners, the utilities acquired by players exhibit a more structured and predictable behavior as the environment is not fully adversarial. Can we then take advantage of this structure to achieve faster convergence rates in environments involving other online learners? This question was addressed by Daskalakis et al. [2011], Rakhlin and Sridharan [2013a,b], who devised algorithms that achieve an $O(\log T/T)$ rate of convergence in two-player zero-sum games. Later, Syrgkanis et al. [2015] and Daskalakis et al. [2021] described ways to construct predictive learning algorithms that, if followed by all players of a general-sum normal-form game, guarantee convergence rates of $O(T^{-3/4})$ and $\tilde{O}(T^{-1})$ respectively. Subsequently, Farina et al. [2022a] extended these results for general convex games. All these results improve the rates of external-regret minimizers and, thus, converge to coarse correlated equilibria. However, as we discussed earlier, this is a rather weak notion of equilibrium. Recently, Anagnostides et al. [2022a,c] settled the problem of constructing no-swap regret minimizers achieving a rate of $\tilde{O}(T^{-1})$. Finally, Anagnostides et al. [2022b] established the first no-regret learning dynamics that converge to the extensive-form correlated equilibrium and the extensive-form coarse correlated equilibrium at an improved rate of $O(T^{-3/4})$ compared to the previous $O(T^{-1/2})$.

Chapter 4

A No-Linear-Swap Regret Algorithm with Polynomial-Time Iterations

In this section, we describe our no-linear-swap-regret algorithm for the set of sequence-form strategies $\mathcal{Q}$ of a generic player in any perfect-recall imperfect-information extensive-form game. The algorithm follows the general template for constructing Φ-regret minimizers given by Gordon et al. [2008] and recalled in Theorem 3.2.3. For this we need two components:

i) an efficient no-external regret minimizer for the set $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ of all matrices inducing linear transformations from $\mathcal{Q}$ to $\mathcal{Q}$,

ii) an efficiently computable fixed point oracle for matrices $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, returning $\boldsymbol{x} = \mathbf{A}\boldsymbol{x} \in \mathcal{Q}$.

The existence of a fixed point, required in ii), is easy to establish by Brouwer's fixed point theorem, since the polytope of sequence-form strategies is compact and convex, and the continuous function $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ maps $\mathcal{Q}$ to itself by definition. Furthermore, as it will become apparent later in the section, all elements $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ have entries in $[0,1]^{\Sigma\times\Sigma}$. Hence, requirement ii) can be satisfied directly by solving the linear feasibility program $\{\text{find } \boldsymbol{x} : \mathbf{A}\boldsymbol{x} = \boldsymbol{x}, \boldsymbol{x} \in \mathcal{Q}\}$. using any of the known polynomial-time algorithms for linear programming. Establishing requirement i) is where the heart of the matter is, and it is the focus of much of the paper. Here, we give intuition for the main insights that contribute to the algorithm.

4.1 The Structure of Linear Transformations of Sequence-Form Strategy Polytopes

The crucial step in our construction is to establish a series of results shedding light on the fundamental geometry of the set $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ of *all* linear transformations from a sequence-form polytope $\mathcal{Q}$ to itself. In fact, our results extend beyond functions from $\mathcal{Q}$ to $\mathcal{Q}$ to more general functions from $\mathcal{Q}$ to a generic compact polytope $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\}$ for arbitrary $\mathbf{P}$ and $\boldsymbol{p}$. We establish the following characterization theorem, which shows that when the functions are expressed in matrix form, the set $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ can be captured by a polynomial number of constraints. The proof is deferred to Chapter 5.

Theorem 4.1.1. *Let $\mathcal{Q}$ be a sequence-form strategy space and let $\mathcal{P}$ be any bounded polytope of the form $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\} \subseteq [0,\gamma]^d$, where $\mathbf{P} \in \mathbb{R}^{k\times d}$. Then, for any linear function $f : \mathcal{Q} \to \mathcal{P}$, there exists a matrix $\mathbf{A}$ in the polytope*

$$\mathcal{M}_{\mathcal{Q}\to\mathcal{P}} \coloneqq \left\{ \mathbf{A} = \left[\, \cdots \mid \mathbf{A}_{(\sigma)} \mid \cdots \,\right] \in \mathbb{R}^{d\times\Sigma} : \begin{array}{lll} (4.1) & \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & \forall\, ja \in \Sigma^{\perp} \\ (4.2) & \mathbf{A}_{(\sigma)} = \mathbf{0} & \forall\, \sigma \in \Sigma \setminus \Sigma^{\perp} \\ (4.3) & \sum_{j'\in\mathcal{C}_\varnothing} \boldsymbol{b}_{j'} = \boldsymbol{p} & \\ (4.4) & \sum_{j'\in\mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j & \forall\, ja \in \Sigma \setminus \Sigma^{\perp} \\ (4.5) & \mathbf{A}_{(\sigma)} \in [0,\gamma]^d & \forall\, \sigma \in \Sigma \\ (4.6) & \boldsymbol{b}_j \in \mathbb{R}^k & \forall\, j \in \mathcal{J} \end{array} \right\}$$

such that $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Conversely, any $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ defines a linear function $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ from $\mathcal{Q}$ to $\mathcal{P}$, that is, such that $\mathbf{A}\boldsymbol{x} \in \mathcal{P}$ for all $\boldsymbol{x} \in \mathcal{Q}$.

The proof operates by induction in several steps. At its core, it exploits the combinatorial structure of sequence-form strategy polytopes, which can be decomposed into sub-problems using a series of Cartesian products and convex hulls. We also remark that while the theorem calls for the polytope $\mathcal{P}$ to be in the form $\mathcal{P} = \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\}$, with little work the result can also be extended to handle other representations such as $\{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} \leq \boldsymbol{p}\}$. We opted for the form specified in the theorem since it most directly leads to the proof, and since the constraints that define the sequence-form strategy polytope (Definition 2.2.2) are already in the form of the statement.

In particular, by setting $\mathcal{P} = \mathcal{Q}$ in Theorem 4.1.1 (in this case, the dimensions of $\mathbf{P}$ will be $k = |\mathcal{J}| + 1$, and $d = |\Sigma|$), we conclude that the set of linear functions from $\mathcal{Q}$ to itself is a compact and convex polytope $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}} \subseteq [0,1]^{\Sigma\times\Sigma}$, defined by $O(|\Sigma|^2)$ linear constraints. As discussed, this polynomial characterization of $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ is the fundamental insight that enables polynomial-time minimization of linear-swap regret in general extensive-form games.

4.2 Our No-Linear-Swap Regret Algorithm

From here, constructing a no-external-regret algorithm for $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ is relatively straightforward, using standard tools from the rich literature of online learning. For example, in Algorithm 3, we propose a solution employing online projected gradient descent [Gordon, 1999, Zinkevich, 2003].

Algorithm 3: Φ-Regret minimizer for the set $\Phi = \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$

Data: $\mathbf{A}^{(1)} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ and fixed point $\boldsymbol{x}^{(1)}$ of $\mathbf{A}^{(1)}$, learning rates $\eta^{(t)} > 0$

for $t = 1, 2, \ldots$ **do**

- Output $\boldsymbol{x}^{(t)}$
- Receive $\boldsymbol{\ell}^{(t)}$ and pay $\langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \rangle$
- Set $\mathbf{L}^{(t)} = \boldsymbol{\ell}^{(t)}(\boldsymbol{x}^{(t)})^\top$
- $\mathbf{A}^{(t+1)} = \Pi_{\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}}(\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)}) = \arg\min_{\mathbf{Y}\in\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}} \|\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)} - \mathbf{Y}\|_F^2$
- Compute a fixed point $\boldsymbol{x}^{(t+1)} = \mathbf{A}^{(t+1)}\boldsymbol{x}^{(t+1)} \in \mathcal{Q}$ of matrix $\mathbf{A}^{(t+1)}$

Combining that no-external-regret algorithm for Φ with the construction by Gordon et al. [2008], we can then establish the following linear-swap regret and iteration complexity bounds for Algorithm 3.

Theorem 4.2.1. *Let Σ denote the set of sequences of the learning player in the extensive-form game, and let $\eta^{(t)} = 1/\sqrt{t}$ for all t. Then, for any sequence of loss vectors $\boldsymbol{\ell}^{(t)} \in [0,1]^\Sigma$, Algorithm 3 guarantees linear-swap regret $O(|\Sigma|^2\sqrt{T})$ after any number T of iterations, and runs in $O(|\Sigma|^{10}\log(|\Sigma|)\log^2 t)$ time for each iteration t.*

Proof of Theorem 4.2.1. First we focus on the linear-swap regret bound. Based on Gordon et al. [2008] the Φ-regret equals external regret over the set Φ of transformations. In our case Φ is the set $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ of all valid linear transformations and the losses for the external regret minimizer are functions $\mathbf{A} \mapsto \langle \boldsymbol{\ell}^{(1)}, \mathbf{A}\boldsymbol{x}^{(1)} \rangle, \mathbf{A} \mapsto \langle \boldsymbol{\ell}^{(2)}, \mathbf{A}\boldsymbol{x}^{(2)} \rangle, \ldots$ Equivalently, we can write these as $\mathbf{A} \mapsto \langle \mathbf{L}^{(t)}, \mathbf{A} \rangle_F$, where $\langle \cdot, \cdot \rangle_F$ is the component-wise inner product for matrices and $\mathbf{L}^{(t)} = \boldsymbol{\ell}^{(t)}(\boldsymbol{x}^{(t)})^\top$, which is a rank-one matrix. Let D be an upper bound on the diameter of $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, and L be such that $\|\mathbf{L}^{(t)}\|_F \leq L$ for all t. Then, based on Orabona [2022] we can bound the external regret for this instance of Online Linear Optimization by picking $\eta^{(t)} = \frac{D}{L\sqrt{t}}$ which gives a regret of $O(DL\sqrt{T})$. Since $\mathbf{A} \in [0,1]^{\Sigma\times\Sigma}$ we get $D = |\Sigma|$, and since $\boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} \in [0,1]^\Sigma$ we get $L = |\Sigma|$. This results in the desired linear-swap regret of $O(|\Sigma|^2\sqrt{T})$.

However, in our previous analysis we assumed that it is possible to compute an exact solution $\mathbf{A}^{(t+1)} = \Pi_{\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}}(\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)})$ of the projection step, which is an instance of convex quadratic programming, meaning that we can only get an approximate solution [Vishnoi, 2021]. In the following Lemma we prove that an ϵ-approximate projection does not affect the regret for a single iteration of the algorithm, if ϵ is sufficiently small. The inequality we prove is similar to the one in Lemma 2.12 from Orabona [2022].

Lemma 4.2.2. *Let $\mathbf{Y}^* = \Pi_{\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}}(\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)})$ and suppose that $\mathbf{A}^{(t+1)} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ is such that $\|\mathbf{A}^{(t+1)} - \mathbf{Y}^*\|_F^2 \leq \epsilon^{(t)}$, then for any $\mathbf{X} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ it holds*

$$\langle \mathbf{L}^{(t)}, \mathbf{X} - \mathbf{A}^{(t)} \rangle_F \leq \frac{1}{2\eta^{(t)}}\|\mathbf{A}^{(t)} - \mathbf{X}\|_F^2 - \frac{1}{2\eta^{(t)}}\|\mathbf{A}^{(t+1)} - \mathbf{X}\|_F^2 + \frac{\eta^{(t)}}{2}\|\mathbf{L}^{(t)}\|_F^2 + \frac{D}{\eta^{(t)}}\frac{1}{\epsilon^{(t)}}$$

Proof. From Lemma 2.12 of Orabona [2022] we know that

$$\langle \mathbf{L}^{(t)}, \mathbf{X} - \mathbf{A}^{(t)} \rangle_F \leq \frac{1}{2\eta^{(t)}}\|\mathbf{A}^{(t)} - \mathbf{X}\|_F^2 - \frac{1}{2\eta^{(t)}}\|\mathbf{Y}^* - \mathbf{X}\|_F^2 + \frac{\eta^{(t)}}{2}\|\mathbf{L}^{(t)}\|_F^2.$$

Additionally, it holds that

$$\begin{aligned}
\|\mathbf{A}^{(t+1)} - \mathbf{X}\|_F^2 &= \|\mathbf{A}^{(t+1)} - \mathbf{Y}^* + \mathbf{Y}^* - \mathbf{X}\|_F^2 \\
&= \|\mathbf{A}^{(t+1)} - \mathbf{Y}^*\|_F^2 + \|\mathbf{Y}^* - \mathbf{X}\|_F^2 + 2\langle \mathbf{A}^{(t+1)} - \mathbf{Y}^*, \mathbf{Y}^* - \mathbf{X} \rangle_F \\
&\leq \|\mathbf{Y}^* - \mathbf{X}\|_F^2 + 2\langle \mathbf{A}^{(t+1)} - \mathbf{Y}^*, \mathbf{A}^{(t+1)} - \mathbf{Y}^* + \mathbf{Y}^* - \mathbf{X} \rangle_F \\
&= \|\mathbf{Y}^* - \mathbf{X}\|_F^2 + 2\langle \mathbf{A}^{(t+1)} - \mathbf{Y}^*, \mathbf{A}^{(t+1)} - \mathbf{X} \rangle_F \\
&\leq \|\mathbf{Y}^* - \mathbf{X}\|_F^2 + 2\|\mathbf{A}^{(t+1)} - \mathbf{Y}^*\|_F \|\mathbf{A}^{(t+1)} - \mathbf{X}\|_F \\
&\leq \|\mathbf{Y}^* - \mathbf{X}\|_F^2 + 2D\epsilon^{(t)}.
\end{aligned}$$

The Lemma follows by combining the previous two inequalities. □

If we set $\epsilon^{(t)} = 1/t^{5/2}$ then for our choice of $\eta^{(t)} = \frac{D}{L\sqrt{t}}$, the error term becomes L/t^2. Summing over T timesteps we thus get an additive error of $O(L) = O(|\Sigma|)$ in the regret, and our total linear-swap regret bound remains $O(|\Sigma|^2\sqrt{T})$.

We now move to analyzing the per-iteration time complexity of Algorithm 3. The most computationally heavy steps are (5) and (6). To compute the fixed point at step (6) we can use any polynomial-time LP algorithm. We can also perform the projection step (5) in polynomial time using the ellipsoid method [Vishnoi, 2021]. For this we reduce it to a suitable Semidefinite Program with $O(|\Sigma|^2)$ variables and use the Cholesky factorization [Gärtner and Matousek, 2014] as the separation oracle, thus responding to separation queries in $O(|\Sigma|^6)$ time. Note that it is possible to guarantee $\|\mathbf{A}^{(t+1)} - \mathbf{Y}^*\|_F^2 \leq \epsilon^{(t)}$ as the ellipsoid method can output a point $\mathbf{A}^{(t+1)} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ such that $\|\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)} - \mathbf{A}^{(t+1)}\|_F \leq \|\mathbf{A}^{(t)} - \eta^{(t)}\mathbf{L}^{(t)} - \mathbf{Y}^*\|_F + \epsilon$ and furthermore, the Frobenius norm is a 2-strongly convex function.

To apply the ellipsoid method we further need bounds on the Frobenius norm of $\mathbf{Y} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ and the maximum projection distance. For the norm of $\mathbf{Y}$, we pick an upper bound of $R = D = |\Sigma|$. For the lower bound r we note that $\mathbf{Y}\boldsymbol{x} \in \mathcal{Q}$ for all $\boldsymbol{x} \in \mathcal{Q}$, which implies $\mathbf{Y}[\varnothing]\boldsymbol{x} = 1 \implies \|\mathbf{Y}[\varnothing]\|_1 \geq 1$. Thus we get $r \geq \frac{1}{|\Sigma|}$. Similarly, we bound the projection distance between 0 and D. Based on these bounds and on Theorem 13.1 from Vishnoi [2021] we conclude that the total per-iteration time complexity of Algorithm 3 is $O\left(|\Sigma|^{10}\log(|\Sigma|)\log^2(t)\right)$. □

It is worth noting that the polynomial-sized description of $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ is crucial in establishing the polynomial running time of the algorithm, both in the projection step (5) and in the fixed point computation step (6). We also remark that the choice of online projected gradient descent combined with the ellipsoid method for projections were arbitrary and the useful properties of $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ are retained when using it with any efficient regret minimizer.

Chapter 5

Proof of the Characterization Theorem (Theorem 4.1.1)

In this section we prove the central result of this paper, the characterization given in Theorem 4.1.1 of linear functions from the sequence-form strategy polytope $\mathcal{Q}$ to the generic polytope $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\}$, where $\mathbf{P} \in \mathbb{R}^{k \times d}$ and $\boldsymbol{p} \in \mathbb{R}^k$.

We will prove the characterization theorem by induction on the structure of the extensive-form strategy polytope. To do so, it will be useful to introduce a few additional objects and notations. We do so in the next subsection.

5.1 Additional Objects and Notation Used in the Proof

First, we introduce a parametric version of the polytope $\mathcal{P}$, where the right-hand side vector is made variable.

Definition 5.1.1. *Given any $\boldsymbol{b} \in \mathbb{R}^k$, we will denote with $\mathcal{P}(\boldsymbol{b})$ the polytope*

$$\mathcal{P}(\boldsymbol{b}) \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{b}, \boldsymbol{x} \geq \mathbf{0}\}.$$

In particular, $\mathcal{P} = \mathcal{P}(\boldsymbol{p})$.

Furthermore, we introduce the equivalence relation $\cong_{\mathcal{D}}$ to indicate that two matrices induce the same linear function when restricted to domain $\mathcal{D}$.

Definition 5.1.2. *Given two matrices $\mathbf{A}, \mathbf{B}$ of the same dimension, we write $\mathbf{A} \cong_{\mathcal{D}} \mathbf{B}$ if $\mathbf{A}\boldsymbol{x} = \mathbf{B}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{D}$. Similarly, given two sets $\mathcal{U}, \mathcal{V}$ of matrices we write $\mathcal{U} \cong_{\mathcal{D}} \mathcal{V}$ to mean that for any $\mathbf{A} \in \mathcal{U}$ there exists $\mathbf{B} \in \mathcal{V}$ with $\mathbf{A} \cong_{\mathcal{D}} \mathbf{B}$, and vice versa.*

Additionally, we introduce a symbol to denote the set of all matrices that induce linear functions from a set $\mathcal{U}$ to a set $\mathcal{V}$.

Definition 5.1.3. *Given any sets $\mathcal{U}, \mathcal{V}$ we denote with $\mathcal{L}_{\mathcal{U} \to \mathcal{V}}$ the set of all matrices that induce linear transformations from $\mathcal{U}$ to $\mathcal{V}$, that is,*

$$\mathcal{L}_{\mathcal{U} \to \mathcal{V}} \coloneqq \{\mathbf{A} : \mathbf{A}\boldsymbol{x} \in \mathcal{V} \text{ for all } \boldsymbol{x} \in \mathcal{U}\}.$$

Finally, we remark that for a matrix $\mathbf{A}$ whose columns are indexed using sequences $\sigma \in \Sigma$, we represent its columns as $\mathbf{A}_{(\sigma)}$. Furthermore, for sequence-form strategies $\boldsymbol{x} \in \mathcal{Q}$, we use $\boldsymbol{x}[\sigma]$ to represent their entries, and $\boldsymbol{x}[\Sigma_{\succeq j}]$ to represent a vector consisting only of the entries corresponding to sequences $\sigma \in \Sigma_{\succeq j}$.

5.2 A Key Tool: Linear Transformations of Cartesian Products

We are now ready to introduce the following Proposition, which will play an important role in the proof of Theorem 4.1.1.

Proposition 5.2.1. *Let $\mathcal{U}_1, \dots, \mathcal{U}_m$ be sets, with $\mathbf{0} \notin \operatorname{aff}\mathcal{U}_i$ [1] for all $i = 1, \dots, m$. Furthermore, for any $i = 1, \dots, m$ and any $\boldsymbol{b}_i \in \mathbb{R}^k$, let $\mathcal{M}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)}$ be such that $\mathcal{M}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)} \cong_{\mathcal{U}_i} \mathcal{L}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)}$. Then, for all $\boldsymbol{b} \in \mathbb{R}^k$,*

$$\mathcal{L}_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m) \to \mathcal{P}(\boldsymbol{b})} \cong_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m)} \left\{ \begin{bmatrix} \mathbf{A}_1 \mid \cdots \mid \mathbf{A}_m \end{bmatrix} : \begin{array}{lll} (5.1)\ \mathbf{A}_i \in \mathcal{M}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)} & \forall\, i \in \{1, \dots, m\} \\ (5.2)\ \boldsymbol{b}_1 + \cdots + \boldsymbol{b}_m = \boldsymbol{b} & \\ (5.3)\ \boldsymbol{b}_i \in \mathbb{R}^k & \forall\, i \in \{1, \dots, m\} \end{array} \right\}. \tag{5.4}$$

Proof. We prove the result by showing the two directions of the inclusion separately.

($\supseteq$) First, we show that for any $\boldsymbol{b} \in \mathbb{R}^k$, any matrix $\mathbf{A} = [\mathbf{A}_1 \mid \cdots \mid \mathbf{A}_m]$ that belongs to the set on the right-hand side of (5.4) induces a linear transformation from $\mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ to $\mathcal{P}(\boldsymbol{b})$ and thus belongs to $\mathcal{L}_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m) \to \mathcal{P}(\boldsymbol{b})}$. To that end, we note that for any $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_m) \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$,

$$\mathbf{A}\boldsymbol{x} = \sum_{i=1}^{m} \mathbf{A}_i \boldsymbol{x}_i \overset{(5.1)}{\geq} \mathbf{0}, \qquad \text{and} \quad \mathbf{P}(\mathbf{A}\boldsymbol{x}) = \sum_{i=1}^{m} \mathbf{P}\mathbf{A}_i \boldsymbol{x}_i \overset{(5.1)}{=} \sum_{i=1}^{m} \boldsymbol{b}_i \overset{(5.2)}{=} \boldsymbol{b},$$

where in both cases we used the fact that $\mathbf{A}_i$ maps any point in $\mathcal{U}_i$ to a point in $\mathcal{P}(\boldsymbol{b}_i) = \{\boldsymbol{y} : \mathbf{P}\boldsymbol{y} = \boldsymbol{b}_i, \boldsymbol{y} \geq 0\}$ by (5.1). Hence $\mathbf{A}\boldsymbol{x} \in \mathcal{P}(\boldsymbol{b})$ for all $\boldsymbol{x} \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$, as we wanted to show.

($\subseteq$) We now look at the converse, showing that for any $\boldsymbol{b} \in \mathbb{R}^k$ and matrix $\mathbf{B} = [\mathbf{B}_1 \mid \cdots \mid \mathbf{B}_m] \in \mathcal{L}_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m) \to \mathcal{P}(\boldsymbol{b})}$, there exists a matrix $\mathbf{A} = [\mathbf{A}_1 \mid \cdots \mid \mathbf{A}_m]$ that satisfies constraints (5.1)-(5.3) and such that $\mathbf{B}\boldsymbol{x} = \mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$. As a first step, in the next lemma we show that $\mathbf{B}$ is always equivalent to another matrix $\mathbf{B}' = [\mathbf{B}'_1 \mid \cdots \mid \mathbf{B}'_m] \cong_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m)} \mathbf{B}$ that satisfies $\mathbf{B}'_i \boldsymbol{x}_i \geq \mathbf{0}$ for all $i = 1, \dots, m$ and $\boldsymbol{x}_i \in \mathcal{U}_i$.

Lemma 5.2.2. *There exist $\mathbf{B}'_1, \dots, \mathbf{B}'_n$, such that $\mathbf{B} \cong_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m)} \mathbf{B}' \coloneqq \left[\mathbf{B}'_1 \;\middle|\; \dots \;\middle|\; \mathbf{B}'_n\right]$, and furthermore, for all $i = 1, \dots, m$, $\boldsymbol{B}'_i \boldsymbol{x}_i \geq \mathbf{0}$ for all $\boldsymbol{x}_i \in \mathcal{U}_i$.*

Proof of Lemma 5.2.2. Since $\mathbf{0} \notin \operatorname{aff}\mathcal{U}_i$ for all $i = 1, \dots, m$ by hypothesis, then there exist vectors $\boldsymbol{\tau}_i$ such that $\boldsymbol{\tau}_i^\top \boldsymbol{x}_i = 1$ for all $\boldsymbol{x}_i \in \mathcal{U}_i$. For any $k \in \{1, \dots, d\}$ and $i \in \{1, \dots, m-1\}$, let

$$\boldsymbol{\beta}_i[k] \coloneqq \min_{\boldsymbol{x} \in \mathcal{U}_i} (\mathbf{B}_i \boldsymbol{x})[k] \quad \forall\, k \in \{1, \dots, d\}, \qquad \mathbf{B}'_i \coloneqq \mathbf{B}_i - \boldsymbol{\beta}_i \boldsymbol{\tau}_i^\top$$

Furthermore, let $\boldsymbol{\beta}_m \coloneqq \sum_{i=1}^{m-1} \boldsymbol{\beta}_i$ and $\mathbf{B}'_m \coloneqq \mathbf{B}_i + \boldsymbol{\beta}_m \boldsymbol{\tau}_m^\top$. It is immediate to check that the matrix $\mathbf{B}' \coloneqq [\mathbf{B}'_1 \mid \cdots \mid \mathbf{B}'_m]$ is such that $\mathbf{B}'\boldsymbol{x} = \mathbf{B}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$, that is, $\mathbf{B}' \cong_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m)} \mathbf{B}$. We now show that $\mathbf{B}'_i \boldsymbol{x}_i \geq \mathbf{0}$ for all $i = 1, \dots, m$ and $\boldsymbol{x}_i \in \mathcal{U}_i$. Expanding the definition of $\mathbf{B}'_i$ and $\boldsymbol{\beta}_i$, for all $i \in \{1, \dots, m-1\}$, $\boldsymbol{x}_i \in \mathcal{U}_i$ and $k \in \{1, \dots, d\}$,

$$(\mathbf{B}'_i \boldsymbol{x}_i)[k] = (\mathbf{B}_i \boldsymbol{x}_i)[k] - (\boldsymbol{\beta}_i)[k] \cdot (\boldsymbol{\tau}_i^\top \boldsymbol{x}_i) = (\mathbf{B}_i \boldsymbol{x}_i)[k] - \min_{\hat{\boldsymbol{x}}_i \in \mathcal{U}_i} (\mathbf{B}_i \hat{\boldsymbol{x}}_i)[k] \geq 0.$$

Hence, it only remains to prove that the same holds for $i = m$. To that end, fix any $k \in \{1, \dots, d\}$ and $\boldsymbol{x}_m \in \mathcal{U}_m$, and let $\boldsymbol{x}_i^* \in \arg\min_{\boldsymbol{x}_i \in \mathcal{U}_i} (\mathbf{B}_i \boldsymbol{x}_i)[k]$ for all $i \in \{1, \dots, m-1\}$. Using the fact that all vectors in $\mathcal{P}(\boldsymbol{b})$ are nonnegative, $\boldsymbol{x}^* \coloneqq (\boldsymbol{x}_1^*, \dots, \boldsymbol{x}_{m-1}^*, \boldsymbol{x}_m) \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ must satisfy $\mathbf{B}\boldsymbol{x}^* \geq \mathbf{0}$. Hence,

$$0 \leq (\mathbf{B}\boldsymbol{x}^*)[k] = (\mathbf{B}_m \boldsymbol{x}_m)[k] + \sum_{i=1}^{m-1} (\mathbf{B}_i \boldsymbol{x}_i^*)[k] = (\mathbf{B}_m \boldsymbol{x}_m)[k] + \sum_{i=1}^{m-1} \boldsymbol{\beta}_i[k] = (\mathbf{B}'_m \boldsymbol{x}_m)[k],$$

thus concluding the proof of the lemma. □

[1] Instead of the condition $\mathbf{0} \notin \operatorname{aff}\mathcal{U}_i$, we could equivalently state that there exists $\boldsymbol{\tau}_i$ such that $\boldsymbol{\tau}_i^\top \boldsymbol{x}_i = 1$ for all $\boldsymbol{x}_i \in \mathcal{U}_i$ using the properties of affine sets.

Since $\mathbf{B}' \cong_{(\mathcal{U}_1 \times \cdots \times \mathcal{U}_m)} \mathbf{B}$, and $\mathbf{B}$ maps to $\mathcal{P}(\boldsymbol{b})$, for all $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \mathcal{U}_1 \times \cdots \times \mathcal{U}_m$ we must have

$$\boldsymbol{b} = \mathbf{P}(\mathbf{B}'\boldsymbol{x}) = \sum_{i=1}^{m} \mathbf{P}\mathbf{B}'_i \boldsymbol{x}_i.$$

Since we can pick the $\boldsymbol{x}_i$ for different indices i independently, it follows that $\mathbf{P}\mathbf{B}_i\boldsymbol{x}_i$ must be a constant function of $\boldsymbol{x}_i \in \mathcal{U}_i$, that is, there must exist vectors $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_m \in \mathbb{R}^k$ such that

$$\boldsymbol{b}_1 + \cdots + \boldsymbol{b}_m = \boldsymbol{b}, \qquad \text{and} \quad \mathbf{P}\mathbf{B}'_i\boldsymbol{x}_i = \boldsymbol{b}_i \qquad \forall\, \boldsymbol{x}_i \in \mathcal{U}_i.$$

Since in addition $\mathbf{B}'_i\boldsymbol{x}_i \geq \mathbf{0}$ (by construction of the $\mathbf{B}'_i$), this means that $\mathbf{B}'_i \in \mathcal{L}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)}$. Finally, using the hypothesis that $\mathcal{L}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)} \cong_{\mathcal{U}_i} \mathcal{M}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)}$, there must exist $\mathbf{A}_i \in \mathcal{M}_{\mathcal{U}_i \to \mathcal{P}(\boldsymbol{b}_i)}$, with $\mathbf{A}_i \cong_{\mathcal{U}_i} \mathbf{B}'_i$, for all $i = 1, \ldots, m$. This concludes the proof. □

5.3 Characterization of Linear Functions of Subtrees

The following result can be understood as a version of Theorem 4.1.1 stated for each subtree, rooted at some decision node, of the decision space.

Theorem 5.3.1. *For any decision node $j \in \mathcal{J}$ and vector $\boldsymbol{b}_j \in \mathbb{R}^k$, let*

$$\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)} \coloneqq \left\{ \underbrace{\begin{bmatrix} \cdots \mid \mathbf{A}_{(ja)} \mid \cdots \end{bmatrix}}_{\in \mathbb{R}^{d \times \Sigma_{\succeq j}}} : \begin{array}{lll} (5.5) & \mathbf{P}\mathbf{A}_{(j'a')} = \boldsymbol{b}_{j'} & \forall\, j'a' \in \Sigma_{\succeq j} \cap \Sigma^{\perp} \\ (5.6) & \mathbf{A}_{(j'a')} = \mathbf{0} & \forall\, j'a' \in \Sigma_{\succeq j} \setminus \Sigma^{\perp} \\ (5.7) & \sum_{j'' \in \mathcal{C}_{j'a'}} \boldsymbol{b}_{j''} = \boldsymbol{b}_{j'} & \forall\, j'a' \in \Sigma_{\succeq j} \setminus \Sigma^{\perp} \\ (5.8) & \mathbf{A}_{(j'a')} \geq \mathbf{0} & \forall\, j'a' \in \Sigma_{\succeq j} \\ (5.9) & \boldsymbol{b}_{j'} \in \mathbb{R}^k & \forall\, j' \succ j \end{array} \right\}. \tag{5.10}$$

Then, $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)} \cong_{\mathcal{Q}_{\succeq j}} \mathcal{L}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$.

Before continuing with the proof, we remark a subtle point: unlike (4.5), which constraints each column to have entries in $[0, \gamma]$, (5.8) only specifies the lower bound at zero, but no upper bound. Hence the tilde above the symbol of this Theorem. Consequently, the matrices in the set $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ need not have bounded entries. In that sense, Theorem 5.3.1 is slightly different from Theorem 4.1.1. We will strengthen (5.8) to enforce a bound on each column when completing the proof of Theorem 4.1.1 in the next subsection.

Proof. To aid us with the proof, we first express the definition of $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ in a way that better captures the inductive structure we need. By direct inspection of the constraints, the set $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ satisfies the inductive definition

$$\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)} = \left\{ \begin{array}{lll} \mathbf{A} = \left[\, \cdots \mid \mathbf{A}_{(ja)} \mid \cdots \right] \in \mathbb{R}^{d\times\Sigma_{\succeq j}}, \text{ such that:} & & \\ (5.11) \quad \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & & \forall\, a\in\mathcal{A}_j : ja \in \Sigma_{\succeq j}\cap\Sigma^{\perp} \\ (5.12) \quad \mathbf{A}_{(ja)} = \mathbf{0} & & \forall\, a\in\mathcal{A}_j : ja \notin \Sigma^{\perp} \\ (5.13) \quad \sum_{j'\in\mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j & & \forall\, a\in\mathcal{A}_j : ja \notin \Sigma^{\perp} \\ (5.14) \quad [\mathbf{A}_{(\sigma)}]_{\sigma\succeq j'} \in \tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j'}\to\mathcal{P}(\boldsymbol{b}_{j'})} & & \forall\, a\in\mathcal{A}_j, j'\in\mathcal{C}_{ja} \\ (5.15) \quad \mathbf{A}_{(ja)} \geq \mathbf{0} & & \forall\, a\in\mathcal{A}_j \\ (5.16) \quad \boldsymbol{b}_{j'} \in \mathbb{R}^k & & \forall\, a\in\mathcal{A}_j, j'\in\mathcal{C}_{ja} \end{array} \right\}. \tag{5.17}$$

We prove the result by structural induction on the tree-form decision process.

- **Base case.** We start by establishing the result for any terminal decision node $j\in\mathcal{J}$, that is, one for which all sequences $\{ja : a\in\mathcal{A}_j\}$ are terminal. In this case, the set $\mathcal{Q}_{\succeq j}$ is the probability simplex $\Delta(\{ja : a\in\mathcal{A}_j\})$. Thus, for a matrix $\mathbf{A}$ to map all $\boldsymbol{x}\in\mathcal{Q}_{\succeq j}$ to elements in the convex polytope $\mathcal{P}(\boldsymbol{b}_j)$ it is both necessary and sufficient that all columns of $\mathbf{A}$ be elements of $\mathcal{P}(\boldsymbol{b}_j)$. It is necessary because if $\mathbf{A}\boldsymbol{x}\in\mathcal{P}(\boldsymbol{b}_j)$ for all $\boldsymbol{x}\in\mathcal{Q}_{\succeq j}$, then for the indicator vector $\boldsymbol{x}$ with $\boldsymbol{x}[ja]=1$ we get $\mathbf{A}\boldsymbol{x}=\mathbf{A}_{(ja)}\in\mathcal{P}(\boldsymbol{b}_j)$. And, it is sufficient because any $\boldsymbol{x}\in\mathcal{Q}_{\succeq j}$ represents a convex combination of the columns $\mathbf{A}_{(ja)}$.

  The set defined by these constraints matches exactly the set $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$ defined in the statement: since all sequences ja are terminal, in this case it reduces to

$$\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)} = \left\{ \left[\, \cdots \mid \mathbf{A}_{(ja)} \mid \cdots \right] \in \mathbb{R}^{d\times\Sigma_{\succeq j}} : \begin{array}{lll} (5.18) & \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & \forall\, a\in\mathcal{A}_j \\ (5.19) & \mathbf{A}_{(ja)} \geq \mathbf{0} & \forall\, a\in\mathcal{A}_j \end{array} \right\},$$

  that is, the set of matrices whose columns are elements of $\mathcal{P}(\boldsymbol{b}_j)$. So, we have $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)} = \mathcal{L}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$ with equality, which immediately implies the claim $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)} \cong_{\Sigma_{\succeq j}} \mathcal{L}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$.

- **Inductive step.** We now look at a general decision node $j\in\mathcal{J}$, assuming as inductive hypothesis that the claim holds for any $j'\succ j$. Below we prove that $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)} \cong_{\mathcal{Q}_{\succeq j}} \mathcal{L}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$ as well.

  ($\subseteq$) We start by showing that for any $\boldsymbol{b}_j\in\mathbb{R}^k$, $\boldsymbol{x}\in\Pi_{\succeq j}$ and $\mathbf{A}\in\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$, we have $\mathbf{A}\boldsymbol{x}\in\mathcal{P}(\boldsymbol{b}_j)$. From (5.8) it is immediate that $\mathbf{A}$ has nonnegative entries, and since any vector $\boldsymbol{x}\in\Pi_{\succeq j}$ also has nonnegative entries, it follows that $\mathbf{A}\boldsymbol{x}\geq\mathbf{0}$. Hence, it only remains to show that $\mathbf{P}(\mathbf{A}\boldsymbol{x})=\boldsymbol{b}_j$. Using Lemma 2.3.2, for any $j'\in\sqcup_{a\in\mathcal{A}_j}\mathcal{C}_{ja}$ there exists $\boldsymbol{x}_{\succeq j'}\in\mathcal{Q}_{\succeq j'}$ such that $\boldsymbol{x}[\Sigma_{\succeq j'}] = \boldsymbol{x}[ja]\cdot\boldsymbol{x}_{\succeq j'}$. Hence, we have

$$\begin{aligned} \mathbf{P}(\mathbf{A}\boldsymbol{x}) &= \sum_{a\in\mathcal{A}_j} \mathbf{P}\mathbf{A}_{(ja)}\boldsymbol{x}[ja] + \sum_{\substack{a\in\mathcal{A}_j \\ ja\notin\Sigma^{\perp}}} \sum_{j'\in\mathcal{C}_{ja}} \mathbf{P}\mathbf{A}_{\succeq j'}(\boldsymbol{x}[ja]\cdot\boldsymbol{x}_{\succeq j'}) \\ &= \sum_{\substack{a\in\mathcal{A}_j \\ ja\in\Sigma^{\perp}}} \boldsymbol{x}[ja]\cdot\mathbf{P}\mathbf{A}_{(ja)} + \sum_{\substack{a\in\mathcal{A}_j \\ ja\notin\Sigma^{\perp}}} \left(\boldsymbol{x}[ja] \sum_{j'\in\mathcal{C}_{ja}} \mathbf{P}\mathbf{A}_{\succeq j'}\boldsymbol{x}_{\succeq j'} \right) \qquad \text{(from (5.12))} \end{aligned}$$

$$= \sum_{\substack{a \in \mathcal{A}_j \\ ja \in \Sigma^\perp}} \boldsymbol{x}[ja] \cdot \boldsymbol{b}_j + \sum_{\substack{a \in \mathcal{A}_j \\ ja \notin \Sigma^\perp}} \left(\boldsymbol{x}[ja] \sum_{j' \in \mathcal{C}_{ja}} \boldsymbol{b}_{j'} \right) \qquad \text{(from (5.11) and (5.14))}$$

$$= \sum_{\substack{a \in \mathcal{A}_j \\ ja \in \Sigma^\perp}} \boldsymbol{x}[ja] \cdot \boldsymbol{b}_j + \sum_{\substack{a \in \mathcal{A}_j \\ ja \notin \Sigma^\perp}} \boldsymbol{x}[ja] \cdot \boldsymbol{b}_j \qquad \text{(from (5.13))}$$

$$= \boldsymbol{b}_j \cdot \sum_{a \in \mathcal{A}_j} \boldsymbol{x}[ja] = \boldsymbol{b}_j. \qquad \text{(from Definition 2.2.2)}$$

($\supseteq$) Conversely, consider any $\mathbf{B} \in \mathcal{L}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$. We will show that there exists a matrix $\mathbf{A} \in \tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ such that $\mathbf{B} \cong_{\mathcal{Q}_{\succeq j}} \mathbf{A}$. First, we argue that there exists a matrix $\mathbf{B}' \cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}$ with the property that the column $\mathbf{B}'_{(ja)}$ corresponding to any *nonterminal* sequence is identically zero.

Lemma 5.3.2. *There exists* $\mathbf{B}' \cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}$ *such that* $\mathbf{B}'_{(ja)} = \mathbf{0}$ *for all* $a \in \mathcal{A}_j$ *such that* $ja \in \Sigma_{\succeq j} \setminus \Sigma^\perp$.

Proof. Fix any $a \in \mathcal{A}_j$ such that ja is nonterminal. Then, by definition there exists at least one decision node j' whose parent sequence is ja. Consider now the matrix $\mathbf{B}''$ obtained from "spreading" column $\mathbf{B}_{(ja)}$ onto $\mathbf{B}_{(ja')}$ ($a' \in \mathcal{A}_{j'}$), that is, the matrix whose columns are defined according to the following rules: (i) $\mathbf{B}''_{(ja)} = \mathbf{0}$, (ii) $\mathbf{B}''_{(ja')} = \mathbf{B}_{(ja')} + \mathbf{B}_{(ja)}$ for all $a' \in \mathcal{A}_{j'}$, (iii) $\mathbf{B}''_{(\sigma)} = \mathbf{B}_{(\sigma)}$ everywhere else. The column $\mathbf{B}''_{(ja)}$ is identically zero by construction, and all other columns $\mathbf{B}''_{(ja')}$, $a' \in \mathcal{A}_j \setminus \{a\}$, are the same as $\mathbf{B}$. Most importantly, since from the sequence-form constraints Definition 2.2.2 any sequence-form strategy $\boldsymbol{x} \in \mathcal{Q}_{\succeq j}$ satisfies the equality $\boldsymbol{x}[ja] = \sum_{a' \in \mathcal{A}_{j'}} \boldsymbol{x}[ja']$, the matrix $\mathbf{B}''$ satisfies $\mathbf{B}''\boldsymbol{x} = \mathbf{B}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}_{\succeq j}$, *i.e.*, $\mathbf{B}'' \cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}$. Iterating the argument for all actions $a \in \mathcal{A}_j$ yields the statement. □

Consider now any $a \in \mathcal{A}_j$ that leads to a terminal sequence $ja \in \Sigma_{\succeq j} \cap \Sigma^\perp$. The vector $\mathbf{1}_{ja}$ defined as having a 1 in the position corresponding to ja and 0 everywhere else is a valid sequence-form strategy vector, that is, $\mathbf{1}_{ja} \in \mathcal{Q}_{\succeq j}$. Hence, since $\mathbf{B}'$ maps $\mathcal{Q}_{\succeq j}$ to $\mathcal{P}(\boldsymbol{b}_j)$, it is necessary that $\mathbf{B}'_{(ja)} \in \mathcal{P}(\boldsymbol{b}_j)$, that is, $\mathbf{B}'_{(ja)} \geq \mathbf{0}$ and $\mathbf{P}\mathbf{B}'_{(ja)} = \boldsymbol{b}_j$. In other words, we have just proved the following.

Lemma 5.3.3. *For any* $a \in \mathcal{A}_j$ *such that* $ja \in \Sigma_{\succeq j} \cap \Sigma^\perp$, $\mathbf{B}'_{(ja)} \geq \mathbf{0}$ *and* $\mathbf{P}\mathbf{B}'_{(ja)} = \boldsymbol{b}_j$.

Combined, Lemmas 5.3.2 and 5.3.3 show that $\mathbf{B}'$ satisfies constraints (5.11), (5.12), and (5.15). Consider now any action $a \in \mathcal{A}_j$ that defines a *nonterminal* sequence $ja \in \Sigma_{\succeq j} \setminus \Sigma^\perp$. For each child decision point $j' \in \mathcal{C}_{ja}$, let $\boldsymbol{x}_{\succeq j'} \in \mathcal{Q}_{\succeq j'}$ be a choice of strategy for that decision point, and denote $\mathbf{B}'_{\succeq j'}$ the submatrix of $\mathbf{B}'$ obtained by only considering the columns $\mathbf{B}'_{(\sigma)}$ corresponding to sequences $\sigma \succeq j'$. The vector $\boldsymbol{x}$ defined according to $\boldsymbol{x}[ja] = 1$, $\boldsymbol{x}[\Sigma_{\succeq j'}] = \boldsymbol{x}_{\succeq j'}$ for all $j' \in \mathcal{C}_{ja}$, and 0 everywhere else is a valid sequence-form strategy $\boldsymbol{x} \in \mathcal{Q}_{\succeq j}$, and therefore $\mathbf{B}'\boldsymbol{x} \in \mathcal{P}(\boldsymbol{b}_j)$ since $\mathbf{B}' \cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}$ and $\mathbf{B} \in \mathcal{L}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ by hypothesis. Therefore, using the fact that $\mathbf{B}'_{(ja)} = \mathbf{0}$ by Lemma 5.3.2, we conclude that

$$\mathcal{P}(\boldsymbol{b}_j) \ni \mathbf{B}'\boldsymbol{x} = \sum_{j' \in \mathcal{C}_{ja}} \mathbf{B}'_{\succeq j'} \boldsymbol{x}_{\succeq j'}.$$

Because the above holds for any choice of $\boldsymbol{x}_{\succeq j'} \in \mathcal{Q}_{\succeq j'}$, it follows that the matrix $[\cdots \mid \mathbf{B}'_{\succeq j'} \mid \cdots] \in \mathcal{L}_{\times_{j' \in \mathcal{C}_{ja}} \mathcal{Q}_{\succeq j'} \to \mathcal{P}(\boldsymbol{b}_j)}$. Hence, applying Proposition 5.2.1 (note that $\mathbf{0} \notin \operatorname{aff} \mathcal{Q}_{\succeq j'}$ since $\sum_{a \in \mathcal{A}_{j'}} \boldsymbol{x}[j'a] = 1$ for all $\boldsymbol{x} \in \mathcal{Q}_{\succeq j'}$ by Definition 2.3.1) together with the inductive

hypothesis, we conclude that for each $j' \in \mathcal{C}_{ja}$ there exist a vector $\boldsymbol{b}_{j'} \in \mathbb{R}^k$ and a matrix $\mathbf{A}_{\succeq j'} \in \mathcal{M}_{\mathcal{Q}_{\succeq j'} \to \mathcal{P}(\boldsymbol{b}_{j'})}$, such that $\sum_{j' \in \mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j$ and $[\cdots \mid \mathbf{B}'_{\succeq j'} \mid \cdots] \cong_{\times_{j' \in \mathcal{C}_{ja}} \mathcal{Q}_{\succeq j'}} [\cdots \mid \mathbf{A}_{\succeq j'} \mid \cdots]$. We can therefore replace all columns corresponding to $\mathbf{B}'_{\succeq j'}$ with those of $\mathbf{A}_{\succeq j'}$, obtaining a new matrix $\cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}'$. Repeating the argument for each $ja \in \Sigma_{\succeq j} \setminus \Sigma^{\perp}$ finally yields a new matrix that is $\cong_{\Sigma_{\succeq j}} \mathbf{B}$ and satisfies all constraints given in (5.17), as we wanted to show. □

5.4 Putting all the Pieces Together

Finally, we are ready to prove the main result of the paper.

Theorem 4.1.1. *Let $\mathcal{Q}$ be a sequence-form strategy space and let $\mathcal{P}$ be any bounded polytope of the form $\mathcal{P} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \mathbf{P}\boldsymbol{x} = \boldsymbol{p}, \boldsymbol{x} \geq \mathbf{0}\} \subseteq [0, \gamma]^d$, where $\mathbf{P} \in \mathbb{R}^{k \times d}$. Then, for any linear function $f : \mathcal{Q} \to \mathcal{P}$, there exists a matrix $\mathbf{A}$ in the polytope*

$$\mathcal{M}_{\mathcal{Q} \to \mathcal{P}} \coloneqq \left\{ \mathbf{A} = \left[\cdots \mid \mathbf{A}_{(\sigma)} \mid \cdots \right] \in \mathbb{R}^{d \times \Sigma} : \begin{array}{lll} (4.1) & \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & \forall\, ja \in \Sigma^{\perp} \\ (4.2) & \mathbf{A}_{(\sigma)} = \mathbf{0} & \forall\, \sigma \in \Sigma \setminus \Sigma^{\perp} \\ (4.3) & \sum_{j' \in \mathcal{C}_{\varnothing}} \boldsymbol{b}_{j'} = \boldsymbol{p} & \\ (4.4) & \sum_{j' \in \mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j & \forall\, ja \in \Sigma \setminus \Sigma^{\perp} \\ (4.5) & \mathbf{A}_{(\sigma)} \in [0, \gamma]^d & \forall\, \sigma \in \Sigma \\ (4.6) & \boldsymbol{b}_j \in \mathbb{R}^k & \forall\, j \in \mathcal{J} \end{array} \right\}$$

such that $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Conversely, any $\mathbf{A} \in \mathcal{M}_{\mathcal{Q} \to \mathcal{P}}$ defines a linear function $\boldsymbol{x} \mapsto \mathbf{A}\boldsymbol{x}$ from $\mathcal{Q}$ to $\mathcal{P}$, that is, such that $\mathbf{A}\boldsymbol{x} \in \mathcal{P}$ for all $\boldsymbol{x} \in \mathcal{Q}$.

Proof of Theorem 4.1.1. We prove the result in two steps. First, we show that

$$\mathcal{L}_{\mathcal{Q} \to \mathcal{P}} \cong_{\mathcal{Q}} \tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{P}} \coloneqq \left\{ \mathbf{A} = \left[\cdots \mid \mathbf{A}_{(\sigma)} \mid \cdots \right] \in \mathbb{R}^{d \times \Sigma} : \begin{array}{lll} (5.20) & \mathbf{P}\mathbf{A}_{(ja)} = \boldsymbol{b}_j & \forall\, ja \in \Sigma^{\perp} \\ (5.21) & \mathbf{A}_{(\sigma)} = \mathbf{0} & \forall\, \sigma \in \Sigma \setminus \Sigma^{\perp} \\ (5.22) & \sum_{j' \in \mathcal{C}_{\varnothing}} \boldsymbol{b}_{j'} = \boldsymbol{p} & \\ (5.23) & \sum_{j' \in \mathcal{C}_{ja}} \boldsymbol{b}_{j'} = \boldsymbol{b}_j & \forall\, ja \in \Sigma \setminus \Sigma^{\perp} \\ (5.24) & \mathbf{A}_{(\sigma)} \geq \mathbf{0} & \forall\, \sigma \in \Sigma \\ (5.25) & \boldsymbol{b}_j \in \mathbb{R}^k & \forall\, j \in \mathcal{J} \end{array} \right\},$$

where the difference between $\tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{P}}$ and $\mathcal{M}_{\mathcal{Q} \to \mathcal{P}}$ lies in constraint (5.24), which only sets a lower bound (at zero) for each entry of the matrix, as opposed to a bound $[0, \gamma]$ as in (4.5). Using the definition of $\tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)}$ given in (5.10) (Theorem 5.3.1), the set $\tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{P}}$ can be equivalently written as

$$\tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{P}} = \left\{ \mathbf{A} = \left[\cdots \mid \mathbf{A}_{(\sigma)} \mid \cdots \right] \in \mathbb{R}^{d \times \Sigma} : \begin{array}{lll} (5.26) & \mathbf{A}_{(\varnothing)} = \mathbf{0} & \\ (5.27) & \sum_{j \in \mathcal{C}_{\varnothing}} \boldsymbol{b}_j = \boldsymbol{p} & \\ (5.28) & [\mathbf{A}_{(\sigma)}]_{\sigma \succeq j} \in \tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j} \to \mathcal{P}(\boldsymbol{b}_j)} & \forall\, j \in \mathcal{C}_{\varnothing} \\ (5.29) & \boldsymbol{b}_j \in \mathbb{R}^k & \forall\, j \in \mathcal{C}_{\varnothing} \end{array} \right\}.$$

To show that $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}} \cong_{\mathcal{Q}} \mathcal{L}_{\mathcal{Q}\to\mathcal{P}}$, we proceed exactly like in the inductive step of the proof of Theorem 5.3.1. Specifically, let $\mathbf{A} \in \tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}}$ and $\boldsymbol{x} \in \mathcal{Q}$ be arbitrary. From Definition 2.2.2 it follows that $\boldsymbol{x}[\Sigma_{\succeq j}] \in \mathcal{Q}_{\succeq j}$ for any $j \in \mathcal{C}_\varnothing$ and therefore, denoting $\mathbf{A}_{\succeq j} \coloneqq [\mathbf{A}_{(\sigma)}]_{\sigma\succeq j}$,

$$\mathbf{P}(\mathbf{A}\boldsymbol{x}) \overset{(5.26)}{=} \sum_{j\in\mathcal{C}_\varnothing} \mathbf{P}\mathbf{A}_{(\sigma)}\boldsymbol{x}[\Sigma_{\succeq j}] \overset{(5.28)}{=} \sum_{j\in\mathcal{C}_\varnothing} \boldsymbol{b}_j \overset{(5.28)}{=} \boldsymbol{p}, \qquad \mathbf{A}\boldsymbol{x} \overset{(5.26)}{=} \sum_{j\in\mathcal{C}_\varnothing} \mathbf{A}_{(\sigma)}\boldsymbol{x}[\Sigma_{\succeq j}] \overset{(5.28)}{\geq} \mathbf{0}.$$

which shows that $\mathbf{A}\boldsymbol{x} \in \mathcal{P}$. Since $\mathbf{A}$ and $\boldsymbol{x}$ were arbitrary, it follows that $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}} \subseteq \mathcal{L}_{\mathcal{Q}\to\mathcal{P}}$. Conversely, let $\mathbf{B} \in \mathcal{L}_{\mathcal{Q}\to\mathcal{P}}$ be arbitrary, and fix a root decision node $j \in \mathcal{C}_\varnothing$. Then, we can "spread out" the column $\mathbf{B}_{(\varnothing)}$ by adding it to each $\mathbf{B}_{(ja)} : a \in \mathcal{A}_j$ by constructing the matrix $\mathcal{L}_{\mathcal{Q}\to\mathcal{P}} \ni \mathbf{B}' \cong_{\mathcal{Q}} \mathbf{B}$ defined by (i) $\mathbf{B}'_{(\varnothing)} = \mathbf{0}$, (ii) $\mathbf{B}'_{(ja)} = \mathbf{B}_{(ja)} + \mathbf{B}_{(\varnothing)}$ for any $a \in \mathcal{A}_j$, and (iii) $\mathbf{B}'_{(\sigma)} = \mathbf{B}_{(\sigma)}$ everywhere else. Pick now any vectors $\{\boldsymbol{x}_{\succeq j} \in \mathcal{Q}_{\succeq j} : j \in \mathcal{C}_\varnothing\}$, and consider the vector $\boldsymbol{x}$ defined as $\boldsymbol{x}[\varnothing] = 1$, and $\boldsymbol{x}[\Sigma_{\succeq j}] = \boldsymbol{x}_{\succeq j}$ for all $j \in \mathcal{C}_\varnothing$. The vector $\boldsymbol{x}$ is a valid sequence-form strategy, that is, $\boldsymbol{x} \in \mathcal{Q}$. Let now $\mathbf{B}'_{\succeq j} \coloneqq [\mathbf{B}'_{(\sigma)}]_{\sigma\succeq j}$. From the fact that $\mathbf{B}'\boldsymbol{x} \in \mathcal{P}$, together with the fact that by construction $\mathbf{B}'_{(\varnothing)} = \mathbf{0}$, we conclude that

$$\mathbf{B}'\boldsymbol{x} = \sum_{j\in\mathcal{C}_\varnothing} \mathbf{B}'_{\succeq j}\boldsymbol{x}_{\succeq j} \in \mathcal{P}.$$

Since the inclusion above holds for any choice of $\{\boldsymbol{x}_{\succeq j} \in \mathcal{Q}_{\succeq j} : j \in \mathcal{C}_\varnothing\}$, and since for all $j \in \mathcal{C}_\varnothing$ the vector $\mathbf{0} \notin \operatorname{aff}\mathcal{Q}_{\succeq j}$ (indeed, $\sum_{a\in\mathcal{A}_j} \boldsymbol{x}[ja] = 1$ for all $\boldsymbol{x} \in \Sigma_{\succeq j}$ by Definition 2.3.1), from Proposition 5.2.1 together with Theorem 5.3.1 we conclude that for each $j \in \mathcal{C}_\varnothing$ there exists a vector $\boldsymbol{b}_j \in \mathbb{R}^k$ and a matrix $\mathbf{A}_{\succeq j} \in \tilde{\mathcal{M}}_{\mathcal{Q}_{\succeq j}\to\mathcal{P}(\boldsymbol{b}_j)}$, such that $\sum_{j\in\mathcal{C}_\varnothing} \boldsymbol{b}_j = \boldsymbol{p}$ and $\mathbf{A}_{\succeq j} \cong_{\mathcal{Q}_{\succeq j}} \mathbf{B}'_{\succeq j}$. By replacing the submatrices $\mathbf{B}'_{\succeq j}$ with $\mathbf{A}_{\succeq j}$ in $\mathbf{B}'$ we then obtain an equivalent matrix that satisfies all constraints that define $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}}$. In summary, we have $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}} \cong \mathcal{L}_{\mathcal{Q}\to\mathcal{P}}$.

To conclude the proof, we now show that $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}} = \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$. First, we make the straightforward observation that any $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ also belongs to $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}}$, as the constraint that define the latter set are only looser. Hence, we only need to show that any $\mathbf{B} \in \tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}}$ also satisfies constraint (4.5). Since $\mathbf{B} \in \tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}}$, all columns of $\mathbf{B}$ are nonnegative (constraint (5.24)). Furthermore, since $\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{P}} \cong_{\mathcal{Q}} \mathcal{L}_{\mathcal{Q}\to\mathcal{P}}$, clearly $\mathbf{B}\boldsymbol{x} \in \mathcal{P}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Fix now any sequence $\sigma \in \Sigma$, and consider any strategy $\boldsymbol{x} \in \mathcal{Q}$ that puts proability mass 1 on all the actions on the path from the root to σ included, that is, any $\boldsymbol{x} \in \mathcal{Q}$ with $\boldsymbol{x}[\sigma] = 1$. Then, from the nonnegativity of the columns of $\mathbf{B}$, it follows that

$$\mathcal{P} \ni \mathbf{B}\boldsymbol{x} \geq \mathbf{B}_{(\sigma)}.$$

Since by definition of γ any point in $\mathcal{P}$ belongs to $[0,\gamma]^d$, we then conclude that $\mathbf{B}_{(\sigma)} \in [0,\gamma]^d$, implying that $\mathbf{B} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ as we wanted to show. □

Chapter 6

Linear-Deviation Correlated Equilibrium

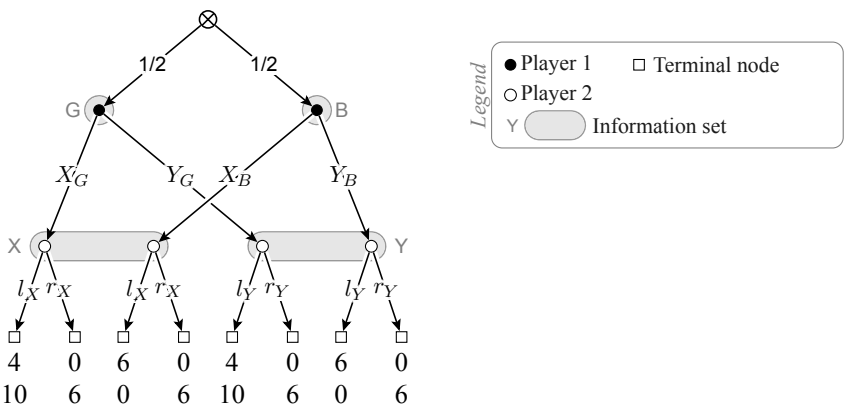
As we discussed in the preliminaries, when all players in a game employ no-Φ-regret learning algorithms, then the empirical frequency of play converges to the set of Φ-equilibria almost surely. Similarly, when $\Phi = \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ the players act based on "no-linear-swap regret" dynamics and converge to a notion of Φ-equilibrium we call *linear-deviation correlated equilibrium* (LCE). In this section we present some notable properties of the LCE. In particular, we discuss its relation to other already established equilibria, as well as the computational tractability of optimal equilibrium selection.

6.1 Relation to CE and EFCE

The Φ-regret minimization framework, offers a natural way to build a hierarchy of the corresponding Φ-equilibria based on the relationship of the Φ sets of deviations. In particular, if for the sets Φ_1, Φ_2 it holds that $\Phi_1 \subseteq \Phi_2$, then the set of Φ_2-equilibria is a subset of the set of Φ_1-equilibria. Since the Correlated Equilibrium is defined using the set of all swap deviations, we conclude that any Φ-equilibrium, including the LCE, is a superset of CE. What is the relationship then of LCE with the extensive-form correlated equilibrium (EFCE)? Farina et al. [2022b] showed that the set Φ^{EFCE} inducing EFCE is the set of all "trigger deviations", which can be expressed as linear transformations of extensive-form strategies. Consequently, the set Φ^{EFCE} is a subset of all linear transformations and thus, it holds that CE $\subseteq$ LCE $\subseteq$ EFCE. In examples 6.1.1 and 6.1.3 we show that there exist specific games in which either CE $\neq$ LCE, or LCE $\neq$ EFCE. Hence, we conclude that the previous inclusions are strict and it holds CE $\subset$ LCE $\subset$ EFCE.

For Example 6.1.1 we use a signaling game from von Stengel and Forges [2008] with a known EFCE and we identify a linear transformation that is not captured by the trigger deviations of EFCE. Specifically, it is possible to perform linear transformations on sequences of a subtree based on the strategies on other subtrees of the TFSDP. For Example 6.1.3 we have found a specific game through computational search that has a LCE, which is not a normal-form correlated equilibrium. To do that we identify a particular normal-form swap that is non-linear.

Example 6.1.1 (EFCE $\neq$ LCE)**.** *Consider the following 2-player signaling game presented by von Stengel and Forges [2008]*



Below are the normal-form payoff matrix of the signaling game (left) and the extensive-form correlated equilibrium (right) given by von Stengel and Forges [2008]:

1 \ 2	$l_X l_Y$	$l_X r_Y$	$r_X l_Y$	$r_X r_Y$
$X_G X_B$	5, 5	5, 5	0, 6	0, 6
$X_G Y_B$	5, 5	2, 8	3, 3	0, 6
$Y_G X_B$	5, 5	3, 3	2, 8	0, 6
$Y_G Y_B$	5, 5	0, 6	5, 5	0, 6

	$l_X l_Y$	$l_X r_Y$	$r_X l_Y$	$r_X r_Y$
$X_G X_B$	0	1/4	0	0
$X_G Y_B$	0	1/4	0	0
$Y_G X_B$	0	0	1/4	0
$Y_G Y_B$	0	0	1/4	0

However, we can observer that this EFCE is not a linear-deviation correlated equilibrium because, for example, Player 1 can increase their payoff by a value of $3/2$ *using the following transformation:*

$X_G X_B \mapsto X_G X_B$ *i.e., map the reduced-normal-form plan* $(1,0,1,0) \mapsto (1,0,1,0)$
$X_G Y_B \mapsto X_G X_B$ *i.e., map the reduced-normal-form plan* $(1,0,0,1) \mapsto (1,0,1,0)$
$Y_G X_B \mapsto Y_G Y_B$ *i.e., map the reduced-normal-form plan* $(0,1,1,0) \mapsto (0,1,0,1)$
$Y_G Y_B \mapsto Y_G Y_B$ *i.e., map the reduced-normal-form plan* $(0,1,0,1) \mapsto (0,1,0,1)$

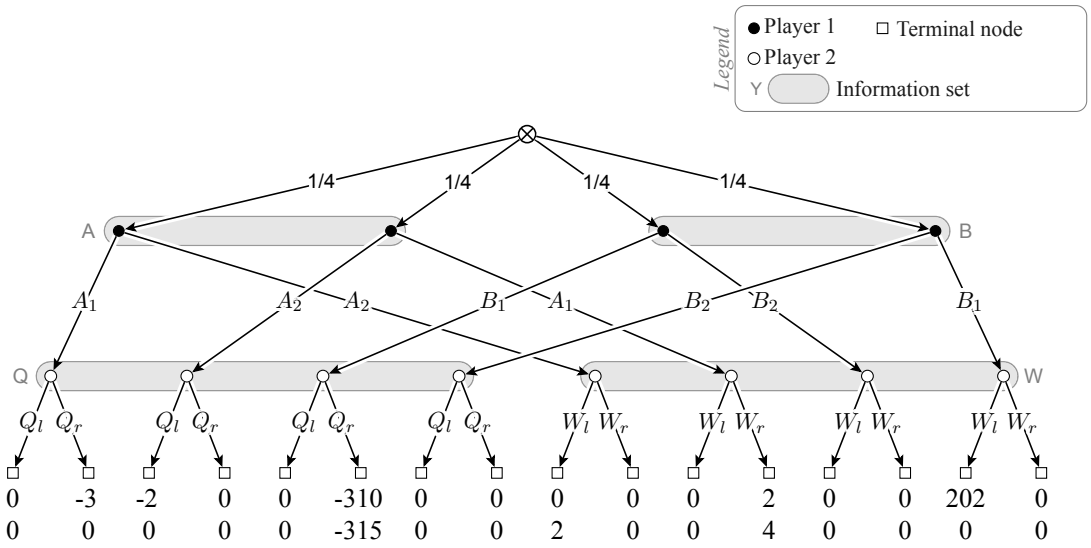
(Above, we have implicitly assumed that the strategy vectors encode probability of actions in the arbitrary order X_G, Y_G, X_B, Y_B*). The above transformation is linear, since it can be represented via the matrix*

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

This would swap the pure strategy $X_G Y_B$ *with* $X_G X_B$ *and strategy* $Y_G X_B$ *with* $Y_G Y_B$*. Crucially, the strategy at the subtree of information set* B *is determined by the strategy at the subtree of information set* G*. Thus, the transformed value for each sequence does not purely depend on the ancestors of that sequence, but can also depend on strategies belonging to "sibling" subtrees. Hence, this example proves that LCE* $\neq$ *EFCE.*

Remark 6.1.2. *The linear transformation given in Example 6.1.1 also serves to show that the Behavioral Deviations defined in Morrill et al. [2021] are not a superset of all linear transformations. Additionally, behavioral deviations are not a subset of linear transformations, as the latter act only on reduced strategies. Thus the two sets of deviations are incomparable.*

Example 6.1.3 (LCE $\neq$ CE)**.** *This example was found through computational search. Consider the 2-player game with the following game tree:*



Based on the previous game tree, the normal-form payoff matrix of the game is shown below.

1 \ 2	Q_lW_l	Q_lW_r	Q_rW_l	Q_rW_r
A_1B_1	*50.5, 0.0*	*50.5, 0.25*	*-27.75, -78.75*	*-27.75, -78.5*
A_1B_2	*0.0, 0.0*	*0.5, 1.0*	*-0.75, 0.0*	*-0.25, 1.0*
A_2B_1	*50.0, 0.5*	*49.5, -0.75*	*-27.0, -78.25*	*-27.5, -79.5*
A_2B_2	*-0.5, 0.5*	*-0.5, 0.0*	*0.0, 0.5*	*0.0, 0.0*

We can now verify that the following is a linear-deviation correlated equilibrium for this game. The verifcation can be done computationally by expressing all constraints of Theorem 4.1.1 as a Linear Program.

	Q_lW_l	Q_lW_r	Q_rW_l	Q_rW_r
A_1B_1	*1/5*	*0*	*0*	*0*
A_1B_2	*0*	*1/5*	*0*	*0*
A_2B_1	*1/5*	*0*	*0*	*0*
A_2B_2	*0*	*0*	*1/5*	*1/5*

However, the swap $\{A_1B_2 \mapsto A_1B_1, A_2B_1 \mapsto A_1B_1\}$ *can increase player 1's payoff by* 50.5.

Furthermore, we can verify that this swap is not linear as follows. First assume that it was linear and could be written as a matrix $\mathbf{A} \in [0,1]^{\Sigma\times\Sigma}$. *Then the matrix has to be consistent with the following*

transformations:

$$A_1B_1 \mapsto A_1B_1$$
$$A_1B_2 \mapsto A_1B_1$$
$$A_2B_1 \mapsto A_1B_1$$
$$A_2B_2 \mapsto A_2B_2$$

For convenience of referring to matrix rows and columns we number the four sequences as:

$$A_1 : 0, \quad A_2 : 1, \quad B_1 : 2, \quad B_2 : 3$$

If we focus on the second transformation, $A_1B_2 \mapsto A_1B_1$, we conclude that $\mathbf{A}[:,0] + \mathbf{A}[:,3] = (1,0,1,0)^\top$ which implies that $\mathbf{A}[1,0] = \mathbf{A}[3,0] = \mathbf{A}[1,3] = \mathbf{A}[3,3] = 0$. Now, if we subtract the respective equations of the last two swaps we get $\mathbf{A}[:,2] - \mathbf{A}[:,3] = (1,-1,1,-1)^\top$. Since $\mathbf{A} \in [0,1]^{\Sigma\times\Sigma}$, the last equation implies $\mathbf{A}[1,3] = 1$ which contradicts the previous constraint of $\mathbf{A}[1,3] = 0$. Thus, we have found a valid normal-form swap that cannot be expressed as a linear transformation of sequence-form strategies. Hence, this example proves that LCE $\neq$ CE.

Empirical evaluation To further illustrate the separation between no-linear-swap-regret dynamics and no-trigger-regret dynamics, used for EFCE, we provide experimental evidence that minimizing linear-swap-regret also minimizes trigger-regret (Figure 6.1, left), while minimizing trigger-regret does *not* minimize linear-swap regret. Specifically, in Figure 6.1 we compare our no-linear-swap-regret learning dynamics (given in Algorithm 3) to the no-trigger-regret algorithm introduced by Farina et al. [2022b]. More details about the implementation of the algorithms is available in Appendix B. In the left plot, we measure on the y-axis the average trigger regret incurred when all players use one or the other dynamics. Since trigger deviations are special cases of linear deviations, as expected, we observe that both dynamics are able to minimize trigger regret. Conversely, in the right plot of Figure 6.1, the y-axis measures linear-swap-regret. We observe that while our dynamics validate the sublinear regret performance proven in Theorem 4.2.1, the no-trigger-regret dynamics of Farina et al. [2022b] exhibit an erratic behavior that is hardly compatible with a vanishing average regret. This suggests that no-linear-swap-regret is indeed a strictly stronger notion of hindsight rationality.

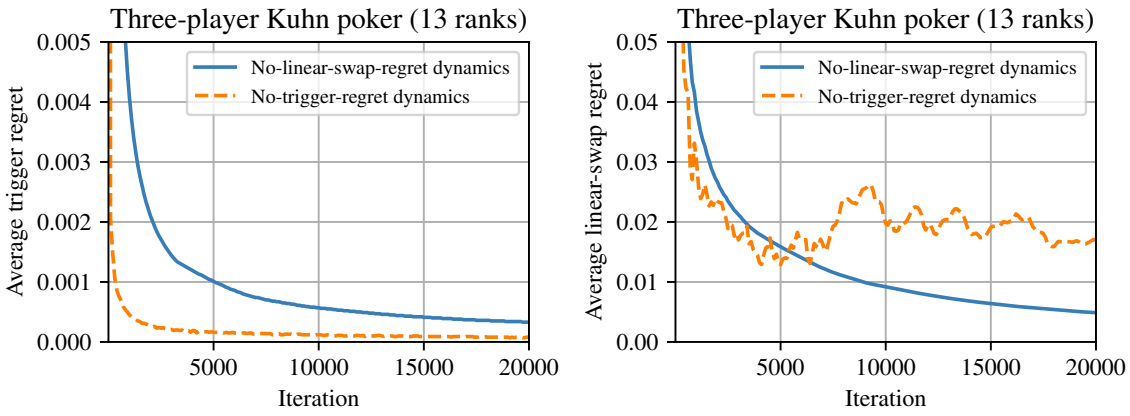


Figure 6.1: (Left) Average trigger regret per iteration for both a linear-swap-regret minimizer and a trigger-regret minimizer. (Right) Average linear-swap regret per iteration for the same two minimizers.

6.2 Hardness of Maximizing Social Welfare

In many cases we are interested in knowing whether it is possible to select an Equilibrium with maximum Social Welfare. Let MAXPAY-LCE be the problem of finding an LCE in EFGs that maximizes the sum (or any linear combination) of all player's utilities. Below, we prove that we cannot solve MAXPAY-LCE, unless P=NP, even for 2 players if chance moves are allowed, and even for 3 players otherwise. We follow the structure of the same hardness proof for the problem MAXPAY-CE of finding an optimal CE in EFGs. Specifically, von Stengel and Forges [2008] use a reduction from SAT to prove that deciding whether MAXPAY-CE can attain the maximum value is NP-hard even for 2 players. To do that, they devise a way to map any SAT instance into a polynomially large game tree in which the root is the chance player, the second level corresponds to one player, and the third level corresponds to the other player. The utilities for both players are exactly the same, thus the players will have to coordinate to maximize their payoff irrespective of the linear combination of utilities we aim to maximize.

Theorem 6.2.1. *For two-player, perfect-recall extensive-form games with chance moves, the problem MAXPAY-LCE is not solvable in polynomial time, unless P=NP.*

Proof. We use the exact same argument employed in the paper by von Stengel and Forges [2008] by reducing SAT to the MAXPAY-LCE problem. Specifically, for each instance of SAT having n clauses and m variables we construct a two-player extensive-form game of size polynomial in n and m. In the beginning, there is a chance move that picks one of n possible actions uniformly at random – one for each SAT clause. Then is the turn of Player 2 who has n distinct singleton information sets corresponding to the chance node actions, and respectively to the n clauses of the SAT instance. Let L_i be the set of literals (negated or non-negated variables) included in the i-th clause of the SAT instance. In information set i of Player 2 there exist $|L_i|$ actions, one for each literal in L_i. Finally, each literal leads to a different decision node for Player 1 who has as many decision nodes as the number of literals in the SAT instance, and in each node there exist exactly 2 possible actions: TRUE and FALSE. However, Player 1 only has m information sets corresponding to the m SAT variables with each information set x grouping together all nodes corresponding to literals of the variable x. This way, Player 1 only chooses the truth value of a variable without knowing from which literal Player 2 has picked this variable. The utilities for both players are equal to 1 if the truth value picked by Player 1 satisfies the literal picked by Player 2, and both utilities are 0 otherwise.

In this game, there exists a pure strategy attaining payoff 1 for each player if and only if the SAT instance is satisfiable – namely Player 1 always acts based on the satisfying assignment and Player 2 picks for every clause a literal that is known to be TRUE. Otherwise, the maximum payoff for each pure strategy is at most $1 - 1/n$. Given that a LCE describes a convex combination of pure strategies, it follows that the maximum total expected payoff of the the sum of the two players' utilities will either be 2 when the SAT instance is satisfiable, or it will be at most $2(1 - 1/n)$ when the instance is not satisfiable. Additionally, this holds not just for the sum but for any linear combination of player utilities. Thus, the problem of deciding whether MAXPAY-LCE can attain a value of at least k is NP-hard for any linear combination of utilities. □

Remark 6.2.2. *The problem retains its hardness if we remove the chance node and add a third player instead. As showed in von Stengel and Forges [2008], in that case we can always build a polynomially-sized game tree that forces the third player to act as a chance node.*

Chapter 7

Conclusions and Future Work

In this paper we have shown the existence of uncoupled no-linear-swap regret dynamics with polynomial-time iteration complexity in the game tree size in any extensive-form game. This significantly extends prior results related to extensive-form correlated equilibria, and begets learning agents that learn not to regret a significantly larger set of strategy transformations than what was known to be possible before. A crucial technical contribution we made to establish our result, and which might be of independent interest, is providing a polynomial characterization of the set of all linear transformations from a sequence-form strategy polytope to itself. Specifically, we showed that such a set of transformations can be expressed as a convex polytope with a polynomial number of linear constraints, by leveraging the rich combinatorial structure of the sequence-form strategies. Moreover, these no-linear-swap regret dynamics converge to linear-deviation correlated equilibria in extensive-form games, which are a novel type of equilibria that lies strictly between normal-form and extensive-form correlated equilibria.

These new results leave open a few interesting future research directions. Even though we know that there exist polynomial-time uncoupled dynamics converging to linear-deviation correlated equilibrium, we conjecture that it is also possible to obtain an efficient centralized algorithm similar to the Ellipsoid Against Hope for computing EFCE in extensive-form games by Huang and von Stengel [2008]. Furthermore, it would be interesting to further explore problems of equilibrium selection related to LCE, possibly by devising suitable Fixed-Parameter Algorithms in the spirit of Zhang et al. [2022]. Finally, the problem of understanding what is the most hindsight rational type of deviations based on which we can construct *efficient* regret minimizers in extensive-form games remains a major open question.

Bibliography

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm. Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games. In *ACM Symposium on Theory of Computing*, 2022a.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Andrea Celli, and Tuomas Sandholm. Faster no-regret learning dynamics for extensive-form correlated and coarse correlated equilibrium. In *Economics and Computation*, 2022b.

Ioannis Anagnostides, Gabriele Farina, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Tuomas Sandholm. Uncoupled learning dynamics with $o(\log t)$ swap regret in multiplayer games. In *Neural Information Processing Systems (NeurIPS)*, 2022c.

Robert J Aumann. Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96, 1974.

Anton Bakhtin, David J Wu, Adam Lerer, Jonathan Gray, Athul Paul Jacob, Gabriele Farina, Alexander H Miller, and Noam Brown. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning. In *International Conference on Learning Representations (ICLR)*, 2023.

Cristina Bicchieri. Self-refuting theories of strategic interaction: A paradox of common knowledge. *Erkenntnis (1975-)*, 30(1/2):69–85, 1989. ISSN 01650106, 15728420.

Avrim Blum and Yishay Mansour. From external to internal regret. *J. Mach. Learn. Res.*, 8, 2007.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 365(6456): 885–890, 2019. doi: 10.1126/science.aay2400.

Andrea Celli, Alberto Marchesi, Gabriele Farina, and Nicola Gatti. No-regret learning dynamics for extensive-form correlated equilibrium. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. *Near-Optimal No-Regret Algorithms for Zero-Sum Games*, pages 235–254. 2011.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich. Near-optimal no-regret learning in general games. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27604–27616. Curran Associates, Inc., 2021.

Gabriele Farina, Tommaso Bianchi, and Tuomas Sandholm. Coarse correlation in extensive-form games. In *AAAI Conference on Artificial Intelligence*, 2020.

Gabriele Farina, Ioannis Anagnostides, Haipeng Luo, Chung-Wei Lee, Christian Kroer, and Tuomas Sandholm. Near-optimal no-regret learning dynamics for general convex games. In *Neural Information Processing Systems (NeurIPS)*, 2022a.

Gabriele Farina, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. *Journal of the ACM*, 69(6), 2022b.

Dean P. Foster and Rakesh V. Vohra. Calibrated learning and correlated equilibrium. *Games and Economic Behavior*, 21(1):40–55, 1997.

Drew Fudenberg and David K Levine. Consistency and cautious fictitious play. *Journal of Economic Dynamics and Control*, 19(5-7):1065–1089, 1995.

Drew Fudenberg and David K Levine. *The theory of learning in games*, volume 2. MIT press, 1998.

Drew Fudenberg and David K Levine. Conditional universal consistency. *Games and Economic Behavior*, 29(1-2):104–130, 1999.

Kaito Fujii. Bayes correlated equilibria and no-regret dynamics, 2023.

Bernd Gärtner and Jirí Matousek. *Approximation Algorithms and Semidefinite Programming*. Springer Publishing Company, Incorporated, 2014.

Geoffrey Gordon. Regret bounds for prediction problems. In *Proceedings of 12th Annual Conference on Computational Learning Theory (COLT '99)*, pages 29 – 40, July 1999.

Geoffrey J Gordon, Amy Greenwald, and Casey Marks. No-regret learning in convex games. In *International Conference on Machine learning*, pages 360–367, 2008.

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual, 2023. URL `https://www.gurobi.com`.

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

Sergiu Hart and Andreu Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26–54, 2001.

Wan Huang and Bernhard von Stengel. Computing an extensive-form correlated equilibrium in polynomial time. In *International Workshop on Internet and Network Economics*, pages 506–513. Springer, 2008.

Sham Kakade, Michael Kearns, John Langford, and Luis Ortiz. Correlated equilibria in graphical games. In *Proceedings of the 4th ACM Conference on Electronic Commerce*, pages 42–47, 2003.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2):247–259, 1996.

H. W. Kuhn. A simplified two-person poker. In H. W. Kuhn and A. W. Tucker, editors, *Contributions to the Theory of Games*, volume 1 of *Annals of Mathematics Studies, 24*, pages 97–103. Princeton University Press, Princeton, New Jersey, 1950.

David Kellogg Lewis. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell, 1969.

Yishay Mansour, Mehryar Mohri, Jon Schneider, and Balasubramanian Sivan. Strategizing against learners in bayesian games. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5221–5252. PMLR, 02–05 Jul 2022.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):pp. 508–513, 2017.

Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R. Wright, Michael Bowling, and Amy R. Greenwald. Efficient deviation types and learning for hindsight rationality in extensive-form games. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7818–7828. PMLR, 2021.

John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

Francesco Orabona. A modern introduction to online learning, 2022.

Christos H. Papadimitriou and Tim Roughgarden. Computing correlated equilibria in multi-player games. *Journal of the ACM*, 55(3), 2008.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 993–1019, Princeton, NJ, USA, 12–14 Jun 2013a. PMLR.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3066–3074, Red Hook, NY, USA, 2013b. Curran Associates Inc.

Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970. ISBN 9781400873173.

I. Romanovskii. Reduction of a game with complete memory to a matrix game. *Soviet Mathematics*, 3, 1962.

J. Maynard Smith and George R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

J.M. Smith and D. Harper. *Animal Signals*. Animal Signals. OUP Oxford, 2003. ISBN 9780198526858.

Gilles Stoltz and Gábor Lugosi. Learning correlated equilibria in games with compact sets of strategies. *Games and Economic Behavior*, 59(1):187–208, 2007.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E. Schapire. Fast convergence of regularized learning in games. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2989–2997, Cambridge, MA, USA, 2015. MIT Press.

Nisheeth K. Vishnoi. *Algorithms for Convex Optimization*. Cambridge University Press, 2021.

John von Neumann and Oskar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, NJ, 1944.

B. von Stengel and F. Forges. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008.

Bernhard von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Brian Hu Zhang, Gabriele Farina, Andrea Celli, and Tuomas Sandholm. Optimal correlated equilibria in general-sum extensive-form games: Fixed-parameter algorithms, hardness, and two-sided column-generation. In David M. Pennock, Ilya Segal, and Sven Seuken, editors, *EC '22: The 23rd ACM Conference on Economics and Computation, Boulder, CO, USA, July 11 - 15, 2022*, pages 1119–1120. ACM, 2022.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, page 928–935. AAAI Press, 2003.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, pages 1729–1736, 2008.

Appendix A

Corollaries of the Characterization Theorem

We mention two direct corollaries of Theorem 4.1.1 that slightly extend the scope of the characterization. The first corollary asserts that the polytope $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ characterizes not only all *linear* functions from $\mathcal{Q}$ to $\mathcal{P}$, but also all *affine* functions.

Corollary A.0.1 (From linear to affine functions)**.** *Let $\mathcal{Q}$ be a sequence-form strategy space and let $\mathcal{P}$ be any polytope. Then, for any* affine *function $g : \mathcal{Q} \to \mathcal{P}$, there exists a matrix $\mathbf{A}$ in the polytope $\mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ defined in Theorem 4.1.1 such that $g(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Conversely, any $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ induces an affine function from $\mathcal{Q}$ to $\mathcal{P}$.*

Proof. The second part of the statement is trivial since any linear function is also affine, and any $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ induces a linear function from $\mathcal{Q}$ to $\mathcal{P}$. Let $g(\boldsymbol{x}) = f(\boldsymbol{x}) + \boldsymbol{b}$ be any affine function, where f is an appropriate linear function from $\mathcal{Q}$ to $\mathcal{P}$ and $\boldsymbol{b} \in \mathbb{R}^n$. Since $\boldsymbol{q}[\varnothing] = 1$ for all $\boldsymbol{q} \in \mathcal{Q}$, the function g coincides on $\mathcal{Q}$ with the function $\tilde{g} : \mathcal{Q} \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) + \boldsymbol{b}\cdot\boldsymbol{x}[\varnothing]$, which is a linear function of $\boldsymbol{x}$. Hence, from the first part of Theorem 4.1.1 there exists $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{P}}$ such that $\mathbf{A}\boldsymbol{x} = \tilde{g}(\boldsymbol{x}) = g(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{Q}$. □

The second corollary of Theorem 4.1.1 extends the characterization to the alternative definition of polytope as a bounded set of the form $\mathcal{C} \coloneqq \{\boldsymbol{y} \in \mathbb{R}^n : \mathbf{C}\boldsymbol{y} \le \boldsymbol{c}\}$, by first introducing slack variables and rewriting the polytope in the form handled by Theorem 4.1.1.

Corollary A.0.2 (Alternative polytope representations)**.** *Let Q be a sequence-form strategy polytope, and $\mathcal{C} \coloneqq \{\boldsymbol{y} \in \mathbb{R}^n : \mathbf{C}\boldsymbol{y} \le \boldsymbol{c}\} \subseteq [-\gamma, \gamma]^n$ be a bounded polytope, where $\mathbf{C} \in \mathbb{R}^{m\times n}$. Let $k \coloneqq \max\{\|\mathbf{C}\|_\infty, \|\boldsymbol{c}\|_\infty\}$ and introduce the polytope*

$$\tilde{\mathcal{P}} \coloneqq \left\{ (\tilde{\boldsymbol{y}}, \boldsymbol{s}) \in \mathbb{R}^n \times \mathbb{R}^m : \left[\begin{array}{c|c} \mathbf{C} & kn\mathbf{I} \end{array}\right] \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{s} \end{bmatrix} = \boldsymbol{c} + \gamma\mathbf{C}\mathbf{1}, \quad \begin{bmatrix} \tilde{\boldsymbol{y}} \\ \boldsymbol{s} \end{bmatrix} \ge \mathbf{0} \right\} \subseteq [0, 2\gamma]^{n+m},$$

which is of the form handled by Theorem 4.1.1. For any affine function $g : \mathcal{Q} \to \mathcal{C}$, there exists a matrix $\mathbf{A}$ in the polytope

$$\tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{C}} \coloneqq \left\{ \left[\tilde{\mathbf{M}}_{(\varnothing)} - \gamma\mathbf{1} \mid \cdots \mid \tilde{\mathbf{M}}_{(\sigma)} \mid \cdots \right] \in \mathbb{R}^{n\times\Sigma} : \begin{bmatrix} \tilde{\mathbf{M}} \\ \tilde{\mathbf{Z}} \end{bmatrix} \in \mathcal{M}_{\mathcal{Q}\to\tilde{\mathcal{P}}},\ \tilde{\mathbf{M}} \in \mathbb{R}^{n\times\Sigma},\ \tilde{\mathbf{Z}} \in \mathbb{R}^{m\times\Sigma} \right\}$$

such that $g(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Conversely, any $\mathbf{A} \in \tilde{\mathcal{M}}_{\mathcal{Q}\to\mathcal{C}}$ induces an affine function from $\mathcal{Q}$ to $\mathcal{C}$.

Proof. We begin by proving that $\mathcal{C} \subseteq [-\gamma, \gamma]^n$ implies $\tilde{\mathcal{P}} \subseteq [0, 2\gamma]^{n+m}$. Consider any $(\tilde{\boldsymbol{y}}, \boldsymbol{s}) \in \tilde{\mathcal{P}}$ and set $\boldsymbol{y} = \tilde{\boldsymbol{y}} - \gamma\mathbf{1}$. Then, this is a valid $\boldsymbol{y} \in \mathcal{C} \subseteq [-\gamma, \gamma]^n$ and, consequently, $\tilde{\boldsymbol{y}} \in [0, 2\gamma]^n$. For the slack variables $\boldsymbol{s}$ it holds that $kn\boldsymbol{s} = \boldsymbol{c} - \mathbf{C}\boldsymbol{y}$, where $\boldsymbol{y} = \tilde{\boldsymbol{y}} - \gamma\mathbf{1}$ from before. By definition of k and by $\boldsymbol{y} \ge -\gamma\mathbf{1}$, we conclude that $\boldsymbol{c} - \mathbf{C}\boldsymbol{y} \le (k + n\gamma k)\mathbf{1} \implies \boldsymbol{s} \in [0, 2\gamma]^m$.

Now let $g : \mathcal{Q} \to \mathcal{C}$ be any affine function from $\mathcal{Q}$ to $\mathcal{C}$. Then we can define an affine function $f : \mathcal{Q} \to \tilde{\mathcal{P}}$ such that

$$f(\boldsymbol{x}) = \begin{bmatrix} g(\boldsymbol{x}) + \gamma \mathbf{1} \\ \boldsymbol{s} \end{bmatrix}$$

for all $\boldsymbol{x} \in \mathcal{Q}$. By Corollary-A.0.1 we know that $\mathcal{M}_{\mathcal{Q} \to \tilde{\mathcal{P}}}$ characterizes all affine functions from $\mathcal{Q}$ to $\tilde{\mathcal{P}}$, including the previous function f. Thus, there exists an $\tilde{\mathbf{M}} \in \mathbb{R}^{n \times \Sigma}$ such that $g(\boldsymbol{x}) + \gamma \mathbf{1} = \tilde{\mathbf{M}} \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathcal{Q}$. Since $\boldsymbol{x}[\varnothing] = 1$ for all $\boldsymbol{x} \in \mathcal{Q}$, we conclude that there exists $\mathbf{A} \in \tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{C}}$ such that $g(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} = \tilde{\mathbf{M}}\boldsymbol{x} - \gamma \mathbf{1}$ for all $\boldsymbol{x} \in \mathcal{Q}$.

Conversely, consider any $\mathbf{A} \in \tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{C}}$ and define $g(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$. That is, there exist suitable $\tilde{\mathbf{M}} \in \mathbb{R}^{n \times \Sigma}, \tilde{\mathbf{Z}} \in \mathbb{R}^{m \times \Sigma}$ that satisfy the constraints of polytope $\tilde{\mathcal{M}}_{\mathcal{Q} \to \mathcal{C}}$. Then for all $\boldsymbol{x} \in \mathcal{Q}$ it holds $g(\boldsymbol{x}) = \tilde{\mathbf{M}}\boldsymbol{x} - \gamma \mathbf{1}$, and $(\tilde{\boldsymbol{y}}, \boldsymbol{s}) \in \tilde{\mathcal{P}}$ where $\tilde{\boldsymbol{y}} = \tilde{\mathbf{M}}\boldsymbol{x}$ and $\boldsymbol{s} = \tilde{\mathbf{Z}}\boldsymbol{x}$. Thus, by construction of $\tilde{\mathcal{P}}$, as we also argued in the beginning of the proof, we conclude that $g(\boldsymbol{x}) = \tilde{\boldsymbol{y}} - \gamma \mathbf{1} = \boldsymbol{y} \in \mathcal{C}$. □

Appendix B

Details on Empirical Evaluation

In this section we provide details about the implementation of our algorithm, as well as the compute resources and game instances used.

Implementation of our no-linear-swap-regret dynamics We implemented our no-linear-swap-regret algorithm (Algorithm 3) in the C++ programming language using the Gurobi commercial optimization solver [Gurobi Optimization, LLC, 2023], version 10. We use Gurobi for the following purposes.

- To compute the projection needed on Algorithm 5 of Algorithm 3. We remark that while Gurobi is typically recognized as a linear and integer linear programming solver, modern versions include tuned code for convex quadratic programming. In particular, we used the barrier algorithm to compute the Euclidean projections onto the polytope $\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$ required at every iteration of our algorithm.
- To compute the fixed points of the matrices $\mathbf{A} \in \mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}$, that is, finding $\mathcal{Q} \ni \boldsymbol{x} = \mathbf{A}\boldsymbol{x}$. As discussed in Chapter 4 this is a polynomially-sized linear program.
- To measure the linear-swap regret incurred after any T iterations, which is plotted on the y-axes of Figure 6.1. This corresponds to solving the linear optimization problem

$$\min_{\mathbf{A}\in\mathcal{M}_{\mathcal{Q}\to\mathcal{Q}}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} - \mathbf{A}\boldsymbol{x}^{(t)} \rangle \right\}.$$

We did very minimal tuning of the constant learning rate η used for online projected gradient descent, trying values $\eta \in \{0.05, 0.1, 0.5\}$ (we remark that a constant value of $\eta \approx 1/\sqrt{T}$ is theoretically sound). We found that $\eta = 0.1$, which is used in the plots of Figure 6.1, performed best.

Implementation of no-trigger-regret dynamics We implemented the no-trigger-regret algorithm of Farina et al. [2022b] in the C++ programming language. In this case, there is no need to use Gurobi, since, as the original authors show, the polytope of trigger deviation functions admits a convenient combinatorial characterization that enables us to sidestep linear programming. Rather, we implemented the algorithm and the computation of the trigger regret directly leveraging the combinatorial structure.

Computational resources used Minimal computational resources were used. All code ran on a personal laptop for roughly 12 hours.

Game instance used We ran our code on the standard benchmark game of Kuhn poker Kuhn [1950]. We used a three-player variant of the game. Compared to the original game, which only considers a simplified deck make of cards out of only three possible ranks (Jack, Queen, or Kind), we use a full deck of 13 possible card ranks. The game has 156 information sets, 315 sequences, and 22308 terminal states.

Appendix C

Further Remarks on the Reduction from No-Φ-Regret to External Regret

A no-Φ-regret algorithm is typically defined as outputting *deterministic* behavior from a finite set $\mathcal{X}$ (for example, deterministic reduced-normal-form plans in extensive-form games, or actions in normal-form games), which can be potentially sampled at random. This is the setting used by, for example, Hart and Mas-Colell [2000] and Farina et al. [2022b]. However, we remark that when the transformations $\phi \in \Phi$ are linear, any such device can be constructed starting from an algorithm that outputs points $\boldsymbol{x}' \in \mathcal{X}' \coloneqq \text{conv}(\mathcal{X})$, and then sampling $\boldsymbol{x}$ unbiasedly in accordance with $\boldsymbol{x}'$, that is, so that $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{x}'$. The reason why this is useful is that constructing the latter object is usually simpler, as $\mathcal{X}'$ is a closed and convex set, and is therefore amenable to the wide array of online optimization techniques that have been developed over the years.

More formally, let Φ be a set of linear transformations that map $\mathcal{X}$ to itself. A no-Φ-regret algorithm for $\mathcal{X}' = \text{conv}(\mathcal{X})$ guarantees that, no matter the sequence of loss vectors $\boldsymbol{\ell}^{(t)}$,

$$R'^{(T)} = \min_{\phi \in \Phi} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}'^{(t)} - \phi(\boldsymbol{x}'^{(t)}) \rangle$$

grows sublinearly. Consider now an algorithm that, after receiving $\boldsymbol{x}'^{(t)} \in \text{conv}(\mathcal{X})$, samples $\boldsymbol{x}^{(t)} \in \mathcal{X}$ unbiasedly, that is, so that $\mathbb{E}[\boldsymbol{x}^{(t)}] = \boldsymbol{x}'^{(t)}$. Then, by linearity of the transformations and using the Azuma-Hoeffind concentration inequality, we obtain that for any $\epsilon > 0$,

$$\mathbb{P}\left[\left| R'^{(T)} - \min_{\phi \in \Phi} \sum_{t=1}^{T} \langle \boldsymbol{\ell}^{(t)}, \boldsymbol{x}^{(t)} - \phi(\boldsymbol{x}^{(t)}) \rangle \right| \leq \Theta\left(\sqrt{T \log \frac{1}{\epsilon}} \right) \right] \geq 1 - \epsilon,$$

where the big-theta notation hides constants that depend on the payoff range of the game and the diameter of $\mathcal{X}$ (a polynomial quantity in the game tree size). This shows that as long as the regret of the no-Φ-algorithm that operates over $\mathcal{X}'$ is sublinear, then so is that of an algorithm that outputs points on $\mathcal{X}$ by sampling unbiasedly from $\mathcal{X}$. We refer the interested reader to Section 4.2 ("From Deterministic to Mixed Strategies") of Farina et al. [2022b].