ABSTRACT

CLOUD, KIRKWOOD ALEXANDER. Topics in Games and Reinforcement Learning. (Under the direction of Eric B. Laber and Ryan Martin).

The disciplines of statistics, game theory, and reinforcement learning provide tools to support decision making. This thesis explores intersections of these fields. It is composed of three investigations, covering (i) how a common statistical tool can be adapted to enable better understanding of games people play, (ii) how a classic game-theoretic algorithm can be improved to enable more efficient multiagent reinforcement learning, and (iii) how quantification of uncertainty can enable safe applications of reinforcement learning in high-stakes settings.

Topics in Games and Reinforcement Learning

by
Kirkwood Alexander Cloud

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2023

APPROVED BY:

Marie Davidian	Brian Reich
Michael Kosorok	Ralph Smith
Eric B. Laber	Ryan Martin Chair of Advisory Committee

DEDICATION

To my brother.

BIOGRAPHY

Alex Cloud was born and raised in Pleasanton, California. He grew from infant to teenager before leaving home to attend Pomona College, where he studied mathematics, played ultimate frisbee, and met many wonderful people. Upon graduating in 2015, he returned to the Bay Area to begin his adult life, working as a data analyst and slowly dying inside each day. During this time, while his Stata scripts and SQL jobs ran, he read about probability and epistemology and dreamt of a return to the intellectually fulfilling life that he used to take for granted.

Alex woke from this dream on a plane to North Carolina. What he found when he landed was a welcoming community of friends, colleagues, teammates, and mentors who ushered in a new chapter of his life. He learned a lot, started too many projects (finished too few), and served as a mentor to other students. In December 2017, he founded Doran's Lab, an undergraduate research group at NC State, where he supported cohorts of thoughtful and inquisitive students as they embarked on their first data science explorations. In 2019, he interned on the Artificial Intelligence team at Riot Games, which led to exciting research ideas, friendships, and employment which funded the rest of his time in graduate school. Then he wrote this.

ACKNOWLEDGEMENTS

I thank...

...Eric Laber for being the ultimate advocate, supporter, and enabler; for holding me to the highest standard, but doing so with compassion. Eric provided mentorship and resources (e.g. lab space, community, funding) beyond the call of duty. He showed me how to see through the noise and get to the heart of a statistical problem. I couldn't reasonably ask for a better advisor.

...Ryan Martin for sincerely engaging with me as I came into his office time and time again as a first-year to pester him with questions and concerns about the philosophy of statistics. Over the years, Ryan mentored and supported me despite having no real responsibility to do so. I am grateful for the many candid conversations which improved my understanding of statistics and the inner workings of academia.

...my committee, for taking time from their busy lives to read this and assist me in my return to the real world.

...Jesse Clifton for the long walks on campus, contributions to my intellectual development, and for modeling a compassionate and rational lifestyle that makes no compromise in its humanity ... Parker Trostle and Vaidehi Dixit for making the many hours spent on homework and qualifying exam preparation fly by, and Conor Artman and Kyle Duke for being good labmates and friends. You all made school fun for me.

...Wes Kerr, Albert Wang, and the AI team (past and present) at Riot Games for cultivating a friendly and intellectually vibrant environment that made work enjoyable, for teaching me to be a better programmer, and for giving me the support and freedom to contribute to AI at Riot in the ways that I was most excited about.

...Adam Venis for, among other things, an unparalleled willingness to hop on a video call to talk through whatever technical problem was currently confounding me.

...Lisa Wong, Nami Sumida, and Chris Waddell for going to incredible lengths to make Doran's Lab great, out of sheer conscientiousness and care for the students.

...the students of Doran's Lab for the curiosity, hard work, and positivity they brought to every meeting. I'm so proud to see what you've accomplished and excited for the next steps of your journeys.

...beloved friends who bolstered me throughout the desolate late-stage-PhD years during a pandemic: Jinny Riedel, James Pollard, Rachel Taube, Nami Sumida, Jane

Peabody, Lisa Wong, Will Curatolo, Ben Burkhardt (and the rest of Book Club), Alex Gruver, Mara Bandt-Law, Michael McLaren, Chris Brown, Kyle Gouchoe-Hanas, Emilia Morgan, Alon Brown, and Sydney Mayes.

... the Triangle Ultimate community and the NC State Statistics department faculty and staff, both of which welcomed me to North Carolina in their own ways and facilitated so many of my positive experiences here.

... my parents for supporting me in all things— most of all, my education.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Making decisions is hard! Thankfully, humans have developed many conceptual frameworks, such as decision theory, operations research, ethical theory, statistics, game theory, and reinforcement learning, to help with the process. Each of these frameworks can be characterized by which aspects of decision making that they place front and center and which aspects they neglect.

This thesis concerns topics at the intersection of statistics, game theory, and reinforcement learning. Broadly speaking, these areas can be characterized as follows.

- Statistics is about *quantification of uncertainty*: given some process that produces data, statistics provides formal tools for converting that data into statements about the process which are calibrated to the amount of evidence conveyed by the data. In a canonical statistics problem, scientists apply interventions to a system and collect numerical measurements of the system after each intervention. The statistician is tasked with using this data to provide (i) estimates of the effects of the interventions, and (ii) some principled quantification of the uncertainty in these estimates. The matter of what to *do* with the estimates is left squarely in the hands of the scientists (or other domain experts).

- Game theory is about *strategy*, i.e., decision making considerations in the presence of other decision makers. It places front and center questions about conflict, cooperation, and incentives between interacting decision makers, or *agents*. A canonical game theory question is: given the options available to a set of agents and the agents' preferences over outcomes, what assignment of agent behaviors will be stable, such that no agent will have incentive to change their behavior? Game theory takes for

granted that agents know their preferences exactly (represented numerically) and typically assumes that the consequences of actions are known with certainty once uncertainty about other agents' actions is accounted for.

- Reinforcement learning is about *learning from experience*. In reinforcement learning, a decision maker interacts with a system by taking actions and observing a numerical "reward" signal that is correlated with the quality of actions taken. The learning problem is to update a decision rule to obtain greater rewards. These updates may influence the distribution of data collected, introducing statistical challenges. A decision maker in this setting faces fundamental tradeoffs, such as whether to act to gather more information, or whether to act so as to earn greater rewards now (the so-called exploration-exploitation tradeoff). A canonical problem in reinforcement learning is the iterative playing of a finite number of slot machines, maintaining estimates of their payoff distributions, with the goal of earning the greatest cumulative payoff. Given its emphasis on behavior and the reward induced by behavior, reinforcement learning (narrowly construed) is not concerned with uncertainty quantification, nor is it concerned with the presence of other decision makers.

Each of the frameworks above offers a unique lens through which to analyze decision problems, and powerful tools for doing so. Each also has its limitations. For example, traditional statistics can be thought of as an entirely descriptive endeavor applied to a system with which the analyst does not interact: it does not explicitly incorporate decision making and often assumes that data was sampled in a convenient fashion, for example, as independent and identical draws from a population. This precludes that the analysis itself might change the data generated. On the other end of the spectrum, classical game theory makes little attempt to be descriptive and is instead quite normative, detailing the ways that 'rational' agents must act, and prescribing theoretically-stable configurations of such agents. Game theory is abstract, assuming away many sources of uncertainty, such as uncertainty in the environment or in one's own preferences over outcomes.

Loosely speaking, reinforcement learning compromises between being purely descriptive or normative: its tools can be used to analyze data generated by real-world agents, or to recommend or prescribe actions for an agent. It is often studied with a focus on manufacturing desirable behavior that makes it more suitable as an engineering discipline (c.f. control theory) than for data-scarce settings where careful quantification of uncertainty

is essential, or real-world settings where consideration of other agents is critical.

Given their strengths and weaknesses, each of these frameworks has something powerful to offer the others. It is this observation that motivates the work in this thesis, which seeks to meld the frameworks together in various ways in order to provide better tools for making decisions in a complicated world.[1] Below, we outline the motivation for, and result of, three different ways of melding the frameworks, each of which makes up a chapter of this thesis.

Combining statistics and game theory

In light of its normative orientation, game theory does not provide tools for data-driven analysis of games as they are played by people or algorithms. This type of analysis is important. Video games are a massively popular form of entertainment worldwide, and in order to create or balance them, game designers must understand how they are played. For example: do certain strategies have an undesirable impact on the outcome of the game? Does randomness play too large a role? Beyond game design, in many jurisdictions in the United States and worldwide, the legality of gambling on a game depends on whether skill or chance "predominates" in determining outcomes of that game. There is not an established formalization of this notion. So, in both game design and in gambling law, there is a need for statistical tools to characterize the influence of various aspects on the outcome of the game.

In **Chapter 2 (Variance decompositions for extensive-form games)**, we take a common statistical tool and develop a version of it for extensive-form games, which are a general model for games with discrete steps and partial observability that includes chess and poker as special cases. The tool is the variance decomposition, which is used in statistics to quantify sources of variation in an outcome of interest. The variance decomposition for games allows the user to attribute variation in a game's outcome to different players and to chance, enabling new ways of analyzing games that may have implications for gambling law or game design. Specifically, we derive a closed-form expression and estimators for the variance in the outcome of an extensive-form game that can be attributed to a single player, or to chance. We analyze poker hands, finding

[1] Of course, combinations of these frameworks have been explored extensively. Examples include the fields of multiagent reinforcement learning, statistical decision theory, and statistical approaches to reinforcement learning, a.k.a. dynamic treatment regimes.

that randomness in the cards dealt has a surprisingly small influence on the outcomes of each hand. We comment briefly on extensions of this idea that could be used to measure other interesting properties of games. *This chapter is a reproduction of Cloud and Laber (2021) with slight modifications.*

Combining game theory and reinforcement learning

In the past five years there has been rapid process in machine learning methods for decision making, especially in complicated, large-scale, multiplayer games like Chess, Go, Dota II, Starcraft, and Stratego. A clear difficulty in learning to play these games is figuring out how to act so as to effect certain changes in a complicated environment. This is the kind of problem that reinforcement learning is meant to solve. However, there is another significant difficulty, well understood in the literature but perhaps not at large: learning in the presence of multiple agents. In all of the games above, a machine learning algorithm must learn to act in a way that respects an implicit dynamism of the game: through the game environment, the agent may face many different kinds of opponents. These opponents might exploit behaviors which had produced favorable outcomes against other agents. For example, a naive reinforcement learning algorithm applied to the game Rock Paper Scissors, trained against a "Paper" opponent, will produce a "Scissors" agent, earning high reward against its training opponent, but being punished severely by a "Rock" agent. Considerations of these multi-agent dynamics exist in the realm of game theory. So, in order to define suitable notions of convergence and design algorithms that attain them, multi-agent reinforcement learning problems must bring game theory to bear.

In **Chapter 3 (Anticipatory fictitious play)**, we study a classic game-theoretic algorithm that has previously been extended to multi-agent reinforcement learning with success. The algorithm, called fictitious play, is used for finding equilibria in two-player competitive games. However, it converges slowly in some games of interest, as we show both theoretically and empirically. To address this deficiency, we propose a novel variant of fictitious play, called anticipatory fictitious play. Anticipatory fictitious play is proved to converge, demonstrated to have superior empirical performance, and extended to the setting of multi-agent reinforcement learning. In doing so, we provide an easy-to-implement multi-agent reinforcement learning algorithm with better performance than one powered by fictitious play. *This chapter is a reproduction of Cloud et al. (2022) with*

slight modifications and less extensive simulation results.

Combining statistics and reinforcement learning

There are many data-driven decision making problems where consideration of other agents is not essential, and hence, game theory is not needed. For example, in clinical trials, recommender systems, or other settings where many independent individuals interact with a system, it is common to treat the problem as though patients or website users as interchangable, independent, and stationary. This is reasonable because, for example, giving one patient a treatment is unlikely to cause another patient later to respond differently to that same treatment. Using data to tailor treatments to individuals in a clinical trial, or to recommend media to web users, is a reinforcement learning problem. However, in settings where a notion of "safety" is desired, a naive application of reinforcement learning is unlikely to be suitable. If a decision making system is to reliably respect constraints on the effects of its actions, it is important to be able to quantify uncertainty in the effects of its actions. Performing this quantification of uncertainty is a problem for statistical inference.

In **Chapter 4 (Safety-constrained online learning in contextual bandits)**, we formulate and study a constrained reinforcement learning problem, where in addition to reward maximization, a decision maker must also select actions according to a constraint on their "safety." Constraint satisfaction, like the underlying reward signal, is estimated from noisy data and thus requires careful handling of uncertainty. We propose a novel algorithmic framework which employs sample splitting in order to make more efficient use of data than existing safe algorithms. The generality of our framework means that it could potentially be applied to all sorts of real-world safety-critical decision problems, including ones that use hard-to-analyze function approximators like artificial neural networks. However, we study our framework in the more restricted linear contextual bandit setting in order to derive theoretical results that are suggestive of the practical safety and utiliy of the method. We prove that, under suitable conditions, our algorithm is guaranteed to produce optimal safe behavior in the limit, and is approximately safe even in small sample settings. In a variety of simulations, we validate the theory and demonstrate superior empirical performance. In this way, we provide a reliable algorithm that can be used for real-world safety-critical data-driven decision making problems. *This chapter is a preprint of a paper by Cloud, Laber, and Kosorok (forthcoming).*

Chapter 2

Variance decompositions for extensive-form games

The extent to which an individual or chance can influence the outcome of a game is a central question in the analysis of games. Consequently, the ability to characterize sources of variation in game outcomes may have significant implications in areas such as game design, law, and multi-agent reinforcement learning. We derive a closed-form expression and estimators for the variance in the outcome of a general multi-agent game that is attributable to a player or chance. We analyze poker hands to show that randomness in the cards dealt has surprisingly little influence on the outcomes of each hand. A simple example is given that demonstrates how variance decompositions can be used to measure other interesting properties of games.

2.1 Introduction

From game design studios to courtrooms, randomness in games has been the subject of extensive discussion. Game designers use random game elements to protect players' egos, increase gameplay variety, and limit the efficacy of mental calculation (Elias et al. 2012). In U.S. state law, the question of whether Poker is predominantly a game of chance or skill is considered to be central to the legality of online Poker (Kelly et al. 2007; Levitt and Miles 2014).

The question of how to measure the role of luck versus skill has proved difficult and produced many answers (DeDonno and Detterman 2008; Croson et al. 2008; Elias et al.

2012; Levitt and Miles 2014; Heubeck 2008a,b; Getty et al. 2018). For example, in *USA v. Lawrence Dicristina*, economic consultant and high-level amateur poker player Randal Heeb testified that "statistical analysis of poker hands confirms that skill predominates over chance." His conclusion was based on a series of heuristic data analyses combined with intuitive judgments (Heeb 2012). Others have argued that the strong association between player skill rating and future earnings constitute strong evidence that poker should be considered a game of skill (Levitt and Miles 2014; van Loon et al. 2015).

A first step in assessing the role of chance in a game is to quantify sources of uncertainty. We examine how variation in the outcomes of a game can be attributed to players or chance events by expressing variation in game outcomes as the sum of variance components associated with (i) the actions taken by a player of interest, and (ii) all remaining sources of variation. By applying this decomposition to a conceptual "chance player," we measure the degree to which randomness inherent in a game biases the results in favor of a given player.

We derive an analytical expression for these variance components and use it to obtain estimators which are model-free in the sense that they do not require access to an entire game model or other players' behavior. Our results apply to finite extensive-form games in general. As an illustrative example, we analyze poker hands played by the DeepStack poker agent against professional players (Moravcik et al. 2017) and find that chance events have very little influence on the expected per-hand profit for a player relative to the total variation in per-hand profit. A roadmap of the paper is as follows:

- Section 2.2 defines extensive-form games.
- Section 2.3.1 casts a general extensive-form game in terms of random variables. Section 2.3.2 describes how to decompose the variance of a game outcome into the sum of two nonnegative terms, gives a formula for the terms, and provides two ways of estimating them.
- Section 2.4 gives estimates for the variance component for chance in collections of Poker games.
- Section 2.5 offers an idea for a further variance decomposition for measuring skill, chance, and non-transitivity in games and applies it to a conceptual game.
- Section 2.6 discusses the interpretation and relevance of the results.

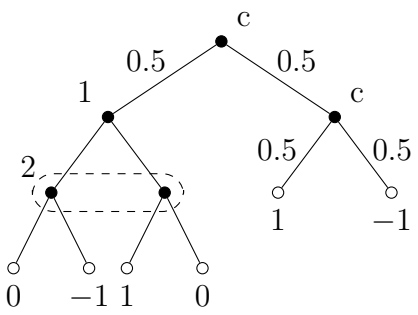

Figure 2.1: An example of an extensive-form game. Each node in the tree is a state $s \in \mathcal{S}$ and is annotated with the corresponding player, $\mathrm{P}(s)$. The dashed line represents Player 2's information state; in this example, they cannot tell what move Player 1 played. Rewards for Player 1 are shown below the terminal nodes.

2.2 Extensive-form games

An *extensive-form game* is a tree-based representation of a multi-agent system; Figure 2.1 displays a simple example. In this representation, the game is played by traversing the tree from the root to a leaf node, with a player's action at each node determining the next node visited. Our notation is based on (Lanctot et al. 2009) and (Heinrich et al. 2015), with some modifications.

Let $\mathcal{S}$ denote the set of possible game states which we assume is finite; each state is associated with a node in the game tree. Define $\mathcal{N} = \{1, \ldots, n\}$ to be the set of (non-chance) players and let c denote the chance player. The *player function* $\mathrm{P} : \mathcal{S} \to \mathcal{N} \cup \{\mathrm{c}\}$ associates each state with a player. At each state $s \in \mathcal{S}$, there are a finite number of available actions $\mathcal{A}(s)$, such that each $a \in \mathcal{A}(s)$ uniquely determines the next state visited in the tree (Shoham and Leyton-Brown 2008).

A sequence of actions $z = (a_1, \ldots, a_m)$ is a *terminal history* if it leads from the root to a leaf of the game tree; let $\mathcal{Z}$ denote the set of all terminal histories. For each player $i \in \mathcal{N}$ and terminal history $z \in \mathcal{Z}$, a reward $r^i(z) \in \mathbb{R}$ is obtained by player i upon reaching z. Each player $i \in \mathcal{N}$ has a set of *information states* $\mathcal{U}^i$ which represent collections of nodes which are indistinguishable to the player. In particular, $\mathcal{U}^i$ is a partition of $\{s \in \mathcal{S} : \mathrm{P}(s) = i\}$ with the additional condition that $\mathcal{A}(s) = \mathcal{A}(s')$ if s and s' are in the same information state. So, we can write $\mathcal{A}(u)$ for $u \in \mathcal{U}^i$ unambiguously. Define $\mathcal{U}^{\mathrm{c}} = \{\{s\} : \mathrm{P}(s) = \mathrm{c}\}$. We consider games of *perfect recall*, so that for every player i, each $u \in \mathcal{U}^i$ can be uniquely identified with the sequence of information states and

actions required to arrive there.

Finally, the behavior of each player $i \in \mathcal{N} \cup \{\mathrm{c}\}$ is described by a *policy* π^i (also known as a behavioral strategy), which is a function that maps each information state $u \in \mathcal{U}^i$ to a distribution over the allowable actions $\mathcal{A}(u)$. A *policy profile* is a tuple of player policies, $\pi = (\pi^1, \ldots, \pi^n)$. By convention, the policy of the chance player π^{c} is considered to be a fixed part of the extensive-form game itself and not a part of any policy profile.

2.3 Variance decompositions for game outcomes

To formalize our results, we represent an extensive-form game in terms of random variables. Having done this, the outcome of the game will be a random variable Y indicating the score obtained by a particular player.

2.3.1 Extensive-form games with random variables

First, we introduce random variables that represent the actions selected by players in a single play of the game. For each $i \in \mathcal{N} \cup \{\mathrm{c}\}$, and for each $u \in \mathcal{U}^i$, let $A(u)$ be a random variable taking values in $\mathcal{A}(u)$ which represents the action player i would take given information state u. This variable always realizes a value, even if u is not reached in a particular play of the game. Note that for $u \neq u' \in \mathcal{U}^i$, it need not be the case that $A(u)$ is independent of $A(u')$. This manner of specifying player behavior is quite general and can account for different models of player action selection. For example, a player may randomly precommit to a deterministic policy (this is known as a mixed strategy in the game theory literature), or select actions independently at random at each time step (a behavioral strategy) (Koller and Megiddo 1996).

For each terminal history $z \in \mathcal{Z}$ and player $i \in \mathcal{N} \cup \{\mathrm{c}\}$, let $m^i(z)$ be the number of actions selected by player i along z, so that for each $j \in \{1, \ldots, m^i(z)\}$, we can write $u^i_{z,j}$ and $a^i_{z,j}$ to denote the jth information state observed and action selected by player i along terminal history z. Define $I^i_{z,j} = \mathbb{1}[A(u^i_{z,j}) = a^i_{z,j}]$ to be the Bernoulli random variable that indicates whether player i selects $a^i_{z,j}$ at $u^i_{z,j}$. Finally, define $I^i_z = \prod_{j=1}^{m^i(z)} I^i_{z,j}$ to be the Bernoulli random variable that indicates whether player i selects all actions along z. (If $m^i(z) = 0$, set $I^i_z \equiv 1$.)

A terminal history occurs if and only if every action along it is selected. Therefore,

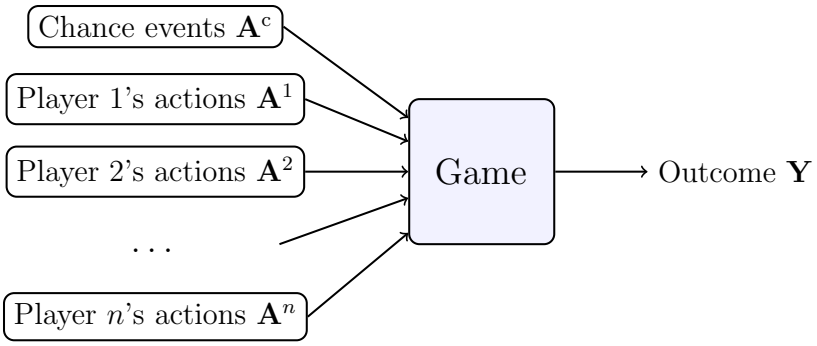

Figure 2.2: A game as a function of player actions.

for each $z \in \mathcal{Z}$, $I_z = \prod_{i \in \mathcal{N} \cup \{\mathrm{c}\}} I_z^i$ defines a Bernoulli random variable such that the success probability $P(I_z = 1)$ is the probability that terminal history z is realized. Let Z be a random terminal history variable such that $P(Z = z) = P(I_z = 1)$ for all z that represents a random play-through of the game. This allows us to cast the outcome of an extensive-form game as

$$\mathbf{Y} = [r^1(Z), \ldots, r^n(Z)],$$

as visualized in Figure 2.2. Write $Y = r(Z) = r^h(Z)$, the random variable representing the reward for a player $h \in \mathcal{N}$ upon one play of the game. Our goal is to express its variance, $V(Y) = E\{[Y - E(Y)]^2\}$, as a sum of nonnegative terms corresponding to meaningful properties of the game.

2.3.2 Variance decomposition

Let $i \in \mathcal{N} \cup \{\mathrm{c}\}$ be a player of interest, and let $\mathbf{A}^i = [A(u)]_{u \in \mathcal{U}^i}$ be the concatenation of all actions for player i. By the law of total variance we can decompose the variance in game outcomes as

$$V(Y) = V[E(Y|\mathbf{A}^i)] + E[V(Y|\mathbf{A}^i)]. \tag{2.1}$$

The term $E(Y|\mathbf{A}^i)$ is the average game outcome upon many traversals of the game tree when player i commits ahead of time to playing the actions in $\mathbf{A}^i$. For example, $E(Y|\mathbf{A}^{\mathrm{c}})$ represents the average outcome for a group of poker players who play the same hand from a deck with a particular card order many times, or the average outcome for a pair of chess

players who start with the same colors every game. Then $V[E(Y|\mathbf{A}^{\mathrm{c}})]$ is the variation in this mean as the chance actions $\mathbf{A}^{\mathrm{c}}$ vary, and represents the variation in game outcomes "explained by" chance events. The latter term of (2.1) has a similar interpretation as the variation in game outcomes not explained by actions selected by player i.

Let $i \in \mathcal{N} \cup \{\mathrm{c}\}$ be a player of interest. Suppose that player i plays according to a behavioral strategy π^i, meaning that $A(u)$ is independent of $A(u')$ for all $u \neq u' \in \mathcal{U}^i$ and action probabilities are given by a policy such that $P(A^i_{z,k} = 1) = \pi^i(a^i_{z,k}|u^i_{z,k})$ for all $z \in \mathcal{Z}$ and $k \in \{1, \ldots, m^i(z)\}$. No such assumption is required for the remaining players; we only require that other players' (precommitments to) actions are independent of the actions of player i.

Let $\eta^i(z) = P(I^i_z = 1)$ be the probability that player i assigns to actions along terminal history z; similarly, $\eta^{-i}(z) = P(\prod_{i' \in \mathcal{N} \cup \{\mathrm{c}\} \setminus \{i\}} I^{i'}_z = 1)$ is the probability assigned by other players, and $\eta(z) = \eta^i(z)\,\eta^{-i}(z) = P(I_z = 1)$. For each information state $u \in \mathcal{U}^i$, define $\mathcal{Z}(u) = \{z \in \mathcal{Z} : u \text{ is visited in } z\}$, $\eta(u) = \sum_{z \in \mathcal{Z}(u)} \eta(z)$, $\eta^i(u) = \sum_{z \in \mathcal{Z}(u)} \eta^i(z)$, and $\eta^{-i}(u)$ such that $\eta(u) = \eta^i(u)\,\eta^{-i}(u)$. For each $a \in \mathcal{A}(u)$, define $\mathcal{Z}(ua) = \{z \in \mathcal{Z} : u \text{ is visited in } z \text{ and action } a \text{ is selected at } u.\}$. Define $q(u, a) = E[r(Z)|Z \in \mathcal{Z}(u, a)]$ to be the expected outcome given that player i is at u and takes action a; similarly, define $v(u) = E[r(Z)|Z \in \mathcal{Z}(u)]$.

Our main result is an expression of the variance in game outcomes explained by player i's actions as a sum of weighted, squared action-value and value functions over all of player i's information states:

$$V[E(Y|\mathbf{A}^i)] = \sum_{u \in \mathcal{U}^i} \left(\sum_{a \in \mathcal{A}(u)} [q(u,a)]^2\, \pi^i(a|u) - [v(u)]^2 \right) \eta^{-i}(u)\, \eta(u). \tag{2.2}$$

A proof is provided in Appendix A.1. Computing this requires traversing the game tree a fixed number of times and hence is $O(|\mathcal{S}|)$. From this we obtain a formula for the other variance component by observing that $E[V(Y|\mathbf{A}^i)] = V(Y) - V[E(Y|\mathbf{A}^i)]$, where $V(Y)$ can be evaluated as $\sum_{z \in \mathcal{Z}} [r(z) - \sum_{z' \in \mathcal{Z}} r(z')\eta(z')]^2\, \eta(z)$.

Assuming η^{-i}, q, and v are known, given an i.i.d. sequence of ν playthroughs of the game, each generating a sequence $\overline{U}_k = (U_{k,1}, \ldots, U_{k,l_k})$ of observed information states in $\mathcal{U}^i$, then the following is a consistent estimator for $V[E(Y|\mathbf{A}^i)]$ as proved in Appendix

A.2:

$$\nu^{-1}\sum_{k=1}^{\nu}\sum_{l=1}^{l_k}\Big(\sum_{a\in\mathcal{A}(U_{k,l})}[q(U_{k,l},a)]^2\pi^i(a|U_{k,l})-[v(U_{k,l})]^2\Big)\eta^{-i}(U_{k,l}). \tag{2.3}$$

In practice, q and v can be estimated by supervised learning and $\eta^i = \eta/\eta^{-i}$ can be estimated with $\hat{\eta}(u) = \nu^{-1}\sum_{k=1}^{\nu}\sum_{l=1}^{l_k}\mathbb{1}(U_{k,l} = u)$ and $\eta^i(u) = \pi^i(u)$ (assuming the analyst does not have access to opponent policies and observations). However, if there are many possible information states, i.e., $|\mathcal{U}^i|$ is large, $\hat{\eta}(u)$ will greatly overestimate the visit probability. An alternative is a more straightforward regression-based estimator. Our regression-based estimator works by fitting a model for the conditional mean of the game outcome given a player's actions, then computing the empirical variance of the conditional mean estimator. The procedure is:

1. Specify a model f_θ that maps the collection of all actions for the player of interest to a real number, $f_\theta : \times_{u\in\mathcal{U}^i}\mathcal{A}(u) \to \mathbb{R}$.

2. For a each observed game $k \in \{1,\ldots,\nu\}$, record action-outcome pairs $(\mathbf{A}^i_k, Y_k)$. For each k, if an information state for the player of interest, $u \in \mathcal{U}^i$ was not visited in game k, sample $A(u) \sim \pi^i(\cdot|u)$ and include the sampled action in $\mathbf{A}^i_k$. Fit the model on the action-outcome pair data to find a $\hat{\theta}$ that minimizes the mean square error, $\nu^{-1}\sum_{k=1}^{\nu}[Y_k - f_{\hat{\theta}}(\mathbf{A}^i_k)]^2$, so $f_{\hat{\theta}}(\cdot)$ estimates $E(Y|\mathbf{A}^i = \cdot\,)$.

3. Estimate $V[E(Y|\mathbf{A}^i)]$ using the sample analog with $f_{\hat{\theta}}(\mathbf{A}^i)$ plugged in for $E(Y|\mathbf{A}^i)$:

$$\nu^{-1}\sum_{k=1}^{\nu}\left[f_{\hat{\theta}}(\mathbf{A}^i_k) - \nu^{-1}\sum_{h=1}^{\nu}f_{\hat{\theta}}(\mathbf{A}^i_h)\right]^2. \tag{2.4}$$

The regression-based estimator is consistent if $f_{\hat{\theta}}$ is consistent for $E(Y|\mathbf{A}^i = \cdot\,)$; a proof is provided in Appendix A.2.

2.4 Analysis of pro poker players versus DeepStack

We analyze 150 thousand hands of heads-up no-limit poker played by different players against the DeepStack poker agent, including 45 thousand hands played by self-identified professional players. For details on how the data were generated, see the supplemental

materials of the DeepStack paper (Moravcik et al. 2017). Our goal is to understand the role chance has in influencing the per-hand profits of a human playing against DeepStack, so we will estimate the variance component for chance for games played by each human player indexed by $j \in \{1, \ldots, 33\}$. We also include an algorithm used for poker agent evaluation called Local Best Response (index $j = 0$), which we include as a form of transfer learning in order to improve estimates of expected outcomes for the human players. Assume that player j plays according to a policy π_j and write $E_{\pi_j}(Y|\mathbf{A}^c)$ to denote the expected per-hand profit for player j against DeepStack given all chance events $\mathbf{A}^c$. Then we would like to know $V[E_{\pi_j}(Y|\mathbf{A}^c)]$ for each j.

We use a neural network to estimate $E_{\pi_j}(Y|\mathbf{A}^c)$ given a player and the realization of all chance events:

- The player's pocket cards (2 cards)
- DeepStack's pocket cards (2 cards)
- The flop (3 cards)
- The turn (1 card)
- The river (1 card)

The neural network shares a representation of cards across all inputs: each card rank (e.g. Ace) and suit (e.g. hearts) is associated with a learned vector embedding; a card is represented by the concatenation of these embeddings. To capture the unordered nature of players' pocket cards and the flop, the card representations for each of those groups is summed. The architecture is depicted in Figure 2.3 and hyperparameters are given in Appendix A.3.

Our model was trained by stochastic gradient descent with the Adam optimizer (Kingma and Ba 2014) with early stopping based on cross-validation loss using a 90%/10% train-test split. If a hand ended before all chance events were observed (for example, if a player folded before the river), the cards associated with that chance event were randomly sampled from the remaining cards in the deck at that point in the game. These cards were resampled in each epoch of training in order to decrease variance. We present our results in Table 2.1. For each player, the empirical variance of the regression estimator was computed over both the training and test data and is recorded in column "Chance var." Due to randomness in the training procedure (neural network initialization, train-

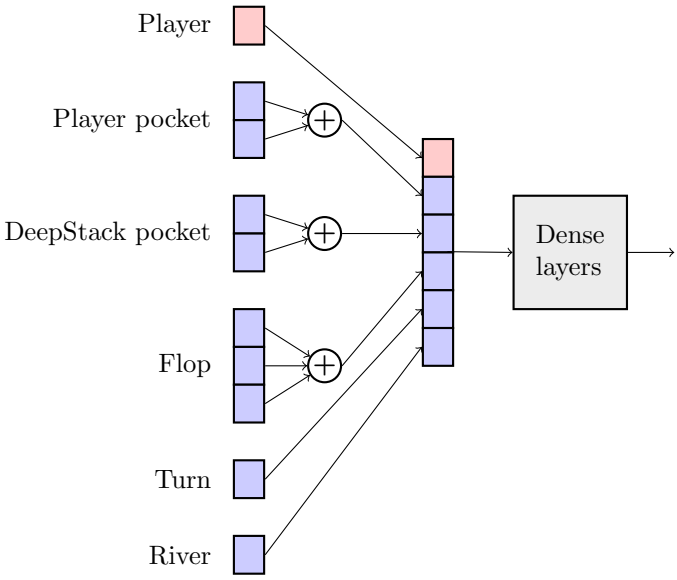

Figure 2.3: The neural network architecture used for analysis of DeepStack hands. The input for each card (shown in blue) is a concatenation of the rank and suit of the card. The rank and suit are each assigned a vector embedding, with the same weights shared for all card inputs.

test split, and sampled actions), we repeat the procedure 100 times and report average results and standard deviations.

The results are somewhat surprising: typical values for the percent of total variance "explained" by chance events fall between 0% and 5%, with low standard deviations. We conclude that the influence of chance events alone on per-hand outcomes is quite limited. Rather, the large amount of variation in per-hand profits is mostly explained by player randomization and the interaction between those actions and chance. We elaborate in the Discussion section.

2.5 A three-way decomposition for assessing skillfulness of a game

As another example of using variance decompositions to analyze games, we present a concept for measuring skill, chance, and non-transitivity that is inspired by prior work on decompositions of games (Candogan et al. 2011) and recent developments regarding

Table 2.1: Analysis of the variance in per-hand player profit (in $1,000's) for human professionals against the DeepStack poker agent. Standard deviations are given in parentheses and are based on 100 replications of the training procedure.

Player name	Hands played	Mean profit	Variance	Chance var.	Chance var. %
Local best response	106,221	-0.07	4.71	0.11 (0.04)	2.4 (0.9)
Ivan Shabalin	3,122	-0.03	3.42	0.09 (0.03)	2.7 (1.0)
Pol Dmit	3,026	-0.09	4.99	0.11 (0.05)	2.2 (0.9)
Muskan Sethi	3,010	-0.21	8.07	0.12 (0.05)	1.5 (0.7)
Dmitry Lesnoy	3,007	0.01	4.42	0.09 (0.03)	1.9 (0.7)
Stanislav Voloshin	3,006	0.01	3.27	0.11 (0.05)	3.3 (1.4)
Lucas Schaumann	3,004	-0.02	2.59	0.11 (0.04)	4.1 (1.5)
Phil Laak	3,003	-0.08	3.58	0.10 (0.04)	2.8 (1.1)
Antonio Parlavecchio	3,003	-0.11	7.22	0.13 (0.05)	1.8 (0.7)
Kaishi Sun	3,002	-0.00	4.14	0.11 (0.04)	2.6 (1.0)
Martin Sturc	3,001	0.05	2.58	0.09 (0.03)	3.5 (1.3)
Prakshat Shrimankar	3,001	-0.02	3.47	0.10 (0.04)	2.9 (1.2)
Tsuneaki Takeda	1,901	0.03	7.46	0.09 (0.04)	1.3 (0.6)
Youwei Qin	1,759	-0.20	14.80	0.11 (0.04)	0.7 (0.3)
Fintan Gavin	1,555	0.00	10.97	0.11 (0.05)	1.0 (0.4)
Giedrius Talacka	1,514	-0.05	11.46	0.12 (0.05)	1.0 (0.5)
Juergen Bachmann	1,088	-0.18	7.80	0.17 (0.09)	2.2 (1.2)
Sergey Indenok	852	-0.03	13.90	0.11 (0.05)	0.8 (0.3)
Sebastian Schwab	516	-0.18	6.25	0.10 (0.05)	1.7 (0.8)
Dara Okearney	456	-0.02	3.37	0.15 (0.06)	4.5 (1.9)
Roman Shaposhnikov	330	0.09	3.95	0.09 (0.04)	2.3 (1.0)
Shai Zurr	330	-0.12	4.15	0.09 (0.04)	2.2 (0.8)
Luca Moschitta	328	-0.14	4.83	0.11 (0.07)	2.4 (1.4)
Stas Tishekvich	295	0.03	3.90	0.11 (0.05)	2.9 (1.2)
Eyal Eshkar	191	-0.07	8.77	0.13 (0.05)	1.5 (0.6)
Jefri Islam	176	-0.38	10.56	0.10 (0.05)	0.9 (0.4)
Fan Sun	122	0.13	9.27	0.12 (0.08)	1.2 (0.9)
Igor Naumenko	102	-0.09	0.61	0.08 (0.04)	13.1 (6.4)
Silvio Pizzarello	90	-0.51	10.44	0.17 (0.19)	1.7 (1.9)
Gaia Freire	76	-0.01	0.09	0.10 (0.06)	111.8 (63.9)
Alexander Bös	74	-0.00	1.29	0.05 (0.03)	3.9 (2.1)
Victor Santos	58	0.18	0.96	0.12 (0.07)	12.6 (7.5)
Mike Phan	32	1.12	25.58	0.07 (0.05)	0.3 (0.2)
Juan-Manuel Pastor	7	-0.73	1.14	0.06 (0.08)	5.7 (7.3)

learning in the context of complex games with nontransitive elements (Balduzzi et al. 2019; Omidshafiei et al. 2020). For simplicity, assume we are given a symmetric two-player zero-sum game and a population of players represented by a finite set of policies Π, each with a skill rating ρ_π for $\pi \in \Pi$. One notion of the skillfulness of the game is the variance in outcomes explained by players' skill ratings alone, assuming two policies (π_1, π_2) are sampled uniformly from Π:

$$V(Y) = V[E(Y|\rho_{\pi_1}, \rho_{\pi_2})] + E[V(Y|\rho_{\pi_1}, \rho_{\pi_2})]. \tag{2.5}$$

Applying the law of total variance to the conditional variance $V(Y|\rho_{\pi_1}, \rho_{\pi_2})$, we condition on chance actions $\mathbf{A}^c$ as in (2.1) to obtain $V(Y|\rho_{\pi_1}, \rho_{\pi_2}) = V[E(Y|\mathbf{A}^c, \rho_{\pi_1}, \rho_{\pi_2})|\rho_{\pi_1}, \rho_{\pi_2}] + E[V(Y|\mathbf{A}^c, \rho_{\pi_1}, \rho_{\pi_2})|\rho_{\pi_1}, \rho_{\pi_2}]$. Using linearity of expectation and the tower rule, this allows us to extend (2.5) to

$$\begin{aligned} V(Y) = &V[E(Y|\rho_{\pi_1}, \rho_{\pi_2})] && \text{(skill)} \\ &+E\{V[E(Y|\mathbf{A}^c, \rho_{\pi_1}, \rho_{\pi_2})|\rho_{\pi_1}, \rho_{\pi_2}]\} && \text{(chance)} \\ &+E[V(Y|\mathbf{A}^c, \rho_{\pi_1}, \rho_{\pi_2})]. && \text{(remaining)} \end{aligned} \tag{2.6}$$

We apply this formula to analyze a simple game parametrized by constants $n \in \mathbb{N}$, $c \in \mathbb{N} \cup \{0\}$, and $\alpha \in [0, 1]$ that is an abstract model of a game with a skill component (some strategies are strictly better than others), a nontransitive component (there exist cycles of pure strategies), and chance (some games are decided by events entirely out of the players' hands). Skillful Rock Paper Scissors, or SkillRPS(n, c, α), is defined as follows: each player $i \in \{1, 2\}$ simultaneously selects a number $N_i \in \{1, \ldots, n\}$ and a move $A_i \in \{\text{Rock, Paper, Scissors}\}$. Player 1's score is $S = N_1 - N_2 + c \cdot \text{RPS}(A_1, A_2)$, where RPS is the payoff function for Rock Paper Scissors depicted in Table 2.2.

Table 2.2: The payoff function RPS(a_1,a_2).

$a_1 \backslash a_2$	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0

The outcome of the game for player 1 is $Y = (1 - W)[\mathbb{1}(S > 0) - \mathbb{1}(S < 0)] + W(2Z - 1)$, where $W \sim$ Bernoulli(α) and $Z \sim$ Bernoulli(1/2) are chance events such that W determines whether the game is decided by a fair coin flip Z. Note that when $n = 1$, $c > 0$, $\alpha = 0$, the game is classic Rock Paper Scissors, when $\alpha = 1$ it is a coin flip, and when $c = 0$ it is a transitive game. The game can be represented in extensive form as shown in Figure 2.1, which depicts SkillRPS(2, 0, 0.5).

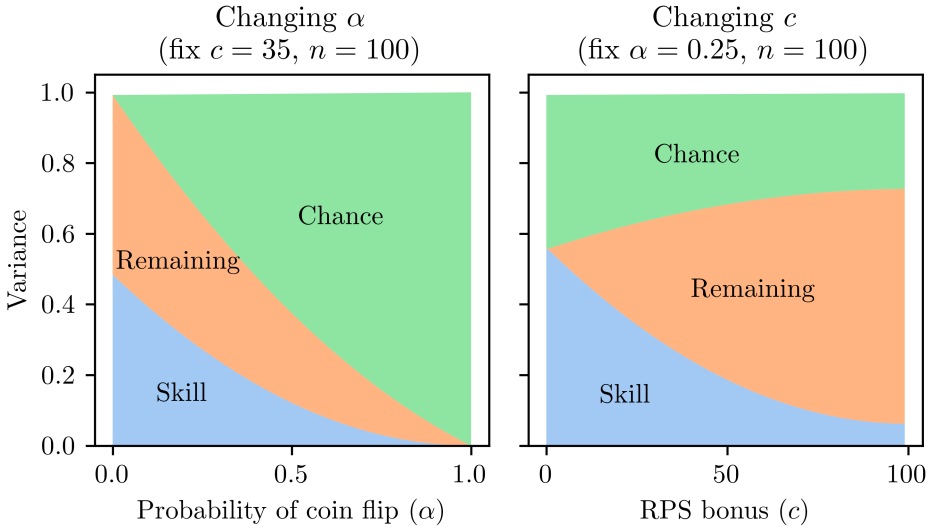

Figure 2.4: Three-way variance decompositions for SkillRPS with different game parameters under the assumption that players selects moves independently and uniformly at random, i.e., for $i \in \{1, 2\}$, $N_i \sim$ Uniform($\{1, \ldots, n\}$) and $A_i \sim$ Uniform({Rock, Paper, Scissors}) and are independent. Details on the variance components for SkillRPS are included in Appendix A.4.

In Figure 2.4, the three-way decomposition is given across many values of the SkillRPS game parameters, showing that the components correspond to meaningful properties of games: increasing the probability that the game outcome is determined by a coin flip increases the chance variance component to 1 as the other variance components decrease smoothly; increasing the bonus for winning at Rock Paper Scissors decreases the skill component. In this case, the "remaining" variation corresponds directly to the non-transitivity introduced by the RPS component of the game.

2.6 Discussion

One might hope that the variance component for chance $V[E(Y|\mathbf{A}^c)]$ measures how lucky a game is in the context of the players playing the game. We argue that this is not the case, and conclude with thoughts on the applicability of variance component estimation for the analysis of games.

First, the variance component for chance does not measure how lucky a game is because by design it avoids measuring variation introduced by random player actions. Consider the classic version of Rock Paper Scissors (RPS) depicted in Table 2.2. A cautious player can guarantee an expected payoff of 0 by assigning uniform probability to each action, causing the outcome of the game to be uniformly random over $\{-1, 0, 1\}$. For this reason, it is natural to view RPS as a game of luck— however, RPS as typically modeled does not have a chance player. All variation in RPS comes from randomness in player action selection. So, if we are to call RPS a game of luck, then a notion of luck that only considers chance events is inadequate.

Second, the variance component for chance is conservative in that it only measures the marginal (average) effect of chance actions on game outcomes. It does not capture the interaction between chance events and player actions. For example, consider a variant of RPS in which one of the players is replaced with a chance player. If the non-chance player employs a uniform random policy, then the expected outcome is 0 regardless of action is selected by chance. Thus $E(Y|\mathbf{A}^c = a) = 0$ for each $a \in \{\text{Rock}, \text{Paper}, \text{Scissors}\}$. This means that for any chance policy, the variance component for chance is 0, yet from the player's perspective, against a uniform chance policy, it is as though the game outcome is entirely determined by chance!

What the variance component for chance actually measures is the per-game amount that chance biases the outcome in favor of a player. In both the examples given above, luck plays a significant role in the game outcomes, but the realization of chance events alone does not tend to significantly tilt the game in the favor of either player— so our measure evaluates to 0. Returning to the analysis of DeepStack poker hands, we see that despite the large amount of variation in per-hand profits (of which any one realization could be called "lucky") the game (as played at a high level) is in some sense fair: on a hand-by-hand basis, the average amount that the random deck order advantages or disadvantages a particular player is small.

Video game designers may find the variance component for chance helpful in assessing

the per-play advantage gleaned by a player due to chance events. We speculate that for a rewarding game experience, the variance component should be kept low, or else players will feel a sense of limited agency. Returning to the question of the legality of poker, our measure could represent a sufficient (but not necessary) criterion for determining that a game is "predominantly due to chance:" if the ratio of the variance component for the chance player to the total variation is greater than 50%, then clearly the game outcomes could be said to be predominantly due to chance. The three-way variance decomposition in (2.6) offers a way to characterize meaningful properties of games that arise in the context of multiagent reinforcement learning and presents new research challenges such as (i) accounting for estimation error in the skill rating (however it is defined), and (ii) accounting for the actual distribution from which policies are sampled to play each other, which is often not uniform but rather skill-based, such that players with nearby skill ratings are likely to be placed together.

2.7 Acknowledgements

Members of the AI Accelerator at Riot Games and Mara Bandt-Law provided input on early drafts and helpful discussions. Three anonymous reviewers provided valuable feedback.

Chapter 3

Anticipatory fictitious play

3.1 Introduction

Matrix games (also known as *normal-form games*) are an abstract model for interactions between multiple decision makers. Fictitious play (FP) (Brown (1951)) is a simple algorithm for two-player matrix games. In FP, each player starts by playing an arbitrary strategy, then proceeds iteratively by playing the best strategy against the empirical average of what the other has played so far. In some cases, such as two-player, zero-sum games, the empirical average strategies will converge to a Nash equilibrium.

Although there are more efficient algorithms for computing Nash equilibria in matrix games (Adler 2013; Shoham and Leyton-Brown 2008), there are a few reasons why fictitious play remains a topic of interest. First, it serves as a model for how humans might arrive at Nash equilibria in real-world interactions (Luce and Raiffa 1989; Brown 1951; Conlisk 1993a). Second, FP is extensible to real-world games which are large and complicated. Our work is primarily motivated by the secondary application.

The initial step towards extending FP to real-world games was by Kuhn (1953), which established the equivalence of normal-form games (represented by matrices) and extensive-form games (represented by trees with additional structure). Loosely speaking, this means that results which apply for matrix games may also apply to much more complicated decision making problems, such as ones that that incorporate temporal elements or varying amounts of hidden information. Leveraging this equivalence, Heinrich et al. (2015) proposed an extension of FP to the extensive-form setting, full-width extensive-form fictitious play (XFP), and proved that it converges to a Nash equilibrium in two-

player, zero-sum games. Heinrich et al. (2015) also proposed Fictitious Self Play (FSP), a machine learning approximation to XFP. In contrast to XFP, which is intractable for real-world games whose states cannot be enumerated in practice, FSP relies only on basic operations which can be approximated in a machine learning setting, like averaging (via supervised learning) and computing best responses (via reinforcement learning). In this way, FSP provides a version of fictitious play suitable for arbitrarily complex two-player, zero-sum games. Not long after the introduction of FSP, Lanctot et al. (2017) presented Policy Space Response Oracles (PSRO), a general framework for fictitious-play-like reinforcement learning algorithms in two-player, zero-sum games. These ideas were employed as part of the groundbreaking AlphaStar system that defeated professional players at StarCraft II (Vinyals et al. 2019).

We introduce anticipatory fictitious play (AFP), a simple variant of fictitious play which is also reinforcement-learning-friendly. In contrast to FP, where players iteratively update to exploit an estimate of the opponent's strategy, players in AFP update proactively to respond to the strategy that the opponent would use to exploit them.

We prove that AFP is guaranteed to converge to a Nash equilibrium in two-player zero-sum games and establish an optimal convergence rate for two classes of games that are of particular interest in learning for real world games (Balduzzi et al. 2019), a class of "cyclic" games and a class of "transitive" games. Numerical comparisons suggest that in AFP eventually outperforms FP on virtually any game, and that its improvement over FP improves as games get larger. Finally, we propose a reinforcement learning version of AFP that is implementable as a one-line modification of an RL implementation of FP, such as FSP. These algorithms are applied to a stochastic, competitive multiagent environment with cyclic dynamics.

3.1.1 Related work

Aside from the literature on fictitious play and its extension to reinforcement learning, there has been substantial work on "opponent-aware" learning algorithms. These algorithms incorporate information about opponent updates and are quite similar to anticipatory fictitious play.

In the context of evolutionary game theory, Conlisk (1993a) proposed an "extrapolation process," whereby two players in a repeated game each forecast their opponents' strategies and respond to those forecasts. Unlike AFP, where opponent responses are

explicitly calculated, the forecasts are made by linear extrapolation based on the change in the opponent's strategy over the last two timesteps. Conlisk (1993b) proposed two types of "defensive adaptation," which are quite similar in spirit to AFP but differ in some important details; most importantly, while they consider the opponent's empirical payoffs at each step, they do not respond directly to what the opponent is likely to play given those payoffs.

Shamma and Arslan (2005) proposed derivative action fictitious play, a variant of fictitious play in the continous time setting in which a best response to a forecasted strategy is played, like in (Conlisk 1993a). The algorithm uses a derivative-based forecast that is analogous to the discrete-time anticipated response of AFP. However, their convergence results rely on a fixed, positive entropy bonus that incentivizes players to play more randomly, and they do not consider the discrete-time case.

Zhang and Lesser (2010) proposed Infinitesimal Gradient Ascent with Policy Prediction, in which two policy gradient learning algorithms continuously train against a forecast of the other's policy. Their algorithm represents the core idea of AFP, albeit implemented in a different setting. However, their proof of convergence is limited to $2x2$ games. Foerster et al. (2018) and Letcher et al. (2018) take this idea further, modifying the objective of a reinforcement learning agent so that it accounts for how changes in the agent will change the anticipated *learning* of the other agents. This line of research is oriented more towards equilibrium finding in general-sum games (e.g. social dilemmas), and less on efficient estimation of equilibria in strictly competitive two-player environments.

3.2 Preliminaries

A (finite) *two-player zero-sum game* (2p0s game) is represented by a matrix $A \in \mathbb{R}^{m \times n}$, so that when player 1 plays i and player 2 plays j, the players observe payoffs $(A_{i,j}, -A_{i,j})$ respectively. Let $\Delta^k \subset \mathbb{R}^k$ be the set of probability vectors representing distributions over $\{1, \ldots, k\}$ elements. Then a *strategy* for player 1 is an element $x \in \Delta^m$ and similarly, a strategy for player 2 is an element $y \in \Delta^n$.

A *Nash equilibrium* in a 2p0s game A is a pair of strategies (x^*, y^*) such that each strategy is optimal against the other, i.e.,

$$x^* \in \arg\max_{x \in \Delta^m} x^\intercal A y^* \quad \text{and} \quad y^* \in \arg\min_{y \in \Delta^n} (x^*)^\intercal A y.$$

The Nash equilibrium represents a pair of strategies that are "stable" in the sense that no player can earn a higher payoff by changing their strategy. At least one Nash equilibrium is guaranteed to exist in any finite game (Nash Jr 1950).

Nash equilibria in 2p0s games enjoy a nice property not shared by Nash equilibria in general: in 2p0s games, if (x_1, y_1) and (x_2, y_2) are Nash equilibria, then (x_2, y_1) is a Nash equilibrium. In a 2p0s game, we define a Nash *strategy* to be be one that occurs as part of some Nash equilibrium. Note that the aforementioned property does not hold in general, so normally it is only valid to describe collections of strategies (one per player) as equilibria.

A *solution* to a 2p0s game A is a pair of strategies (x^*, y^*) such that

$$\min_{y \in \Delta^n} (x^*)^\intercal A y \leq (x^*)^\intercal A y^* \leq \max_{x \in \Delta^m} x^\intercal A y^*.$$

We say $v^* = (x^*)^\intercal A y^*$, which is unique, the *value* of the game. Nash equilibria are equivalent to solutions of 2p0s games (Shoham and Leyton-Brown 2008), which is why we use the same notation. Finally, the *exploitability* of a strategy is the difference between the value of the game and the worst-case payoff of that strategy. So the exploitability of $x \in \Delta^m$ is $v^* - \min x^\intercal A$, and the exploitability of $y \in \Delta^n$ is $\max Ay - v^*$.

3.2.1 Fictitious play

Let $e_1, e_2, \ldots$ denote the standard basis vectors in $\mathbb{R}^m$ or $\mathbb{R}^n$. Let $\mathtt{BR}_A^k$ be the best response operator for player k, so that

$$\begin{aligned} (\forall y \in \Delta^n) \quad & \mathtt{BR}_A^1(y) = \{e_i \in \mathbb{R}^m : i \in \arg\max Ay\}; \\ (\forall x \in \Delta^m) \quad & \mathtt{BR}_A^2(x) = \{e_j \in \mathbb{R}^n : j \in \arg\min x^\intercal A\}. \end{aligned}$$

Fictitious play is given by the following process. Let $x_1 = \overline{x}_1 = e_i$ and $y_1 = \overline{y}_1 = e_j$ be initial strategies for some i, j. For each $t \in \mathbb{N}$, let

$$\begin{aligned} x_{t+1} &\in \mathtt{BR}_A^1(\overline{y}_t); & y_{t+1} &\in \mathtt{BR}_A^2(\overline{x}_t); \\ \overline{x}_{t+1} &= \frac{1}{t+1} \sum_{k=1}^{t+1} x_t; & \overline{y}_{t+1} &= \frac{1}{t+1} \sum_{k=1}^{t+1} y_t. \end{aligned}$$

In other words, at each timestep t, each player calculates the strategy that is the best response to their opponent's average strategy so far. Robinson (1951) proved that the pair of average strategies $(\overline{x}_t, \overline{y}_t)$ converges to a solution of the game by showing that the exploitability of both strategies converge to zero.

Theorem 1. (Robinson, 1951) If $\{(x_t, y_t)\}_{t\in\mathbb{N}}$ is a FP process for a 2p0s game with payoff matrix $A \in \mathbb{R}^{m\times n}$, then

$$\lim_{t\to\infty} \min \overline{x}_t^\intercal A = \lim_{t\to\infty} \max A\overline{y}_t = v^*,$$

where v^* is the value of the game. Furthermore, a bound on the rate of convergence is given by

$$\max A\overline{y}_t - \min \overline{x}_t^\intercal A = O(t^{-1/(m+n-2)}) \text{ for all } t \in \mathbb{N},$$

where $a = \max_{i,j} A_{i,j}$.

(Robinson did not explicitly state the rate, but it follows directly from her proof, as noted in Daskalakis and Pan (2014) and explicated in our Appendix B.2.1.)

3.3 Anticipatory Fictitious Play

Although FP converges to a Nash equilibrium in 2p0s games, it may take an indirect path. For example, in Rock Paper Scissors with tiebreaking towards the minimum strategy index, the sequence of average strategies $\{\overline{x}_1, \overline{x}_2, \dots\}$ orbits the Nash equilibrium, slowly spiraling in with decreasing radius, as shown on the left in Figure 3.1. This tiebreaking scheme is not special; under random tiebreaking, the path traced by FP is qualitatively the same, resulting in slow convergence with high probability, as shown in Figure 3.2.

Given that FP appears to do an especially poor job of decreasing exploitability in the case above, we consider alternatives. Inspired by gradient descent, we ask if there is there a way to follow the gradient of exploitability towards the Nash equilibrium (without explicitly computing it, as is done in Lockhart et al. (2019)). By definition, the best response to the average strategy is a strategy that maximally exploits the average strategy. So, a natural choice of update to reduce exploitability is to move the average strategy in a direction that counters this best response.

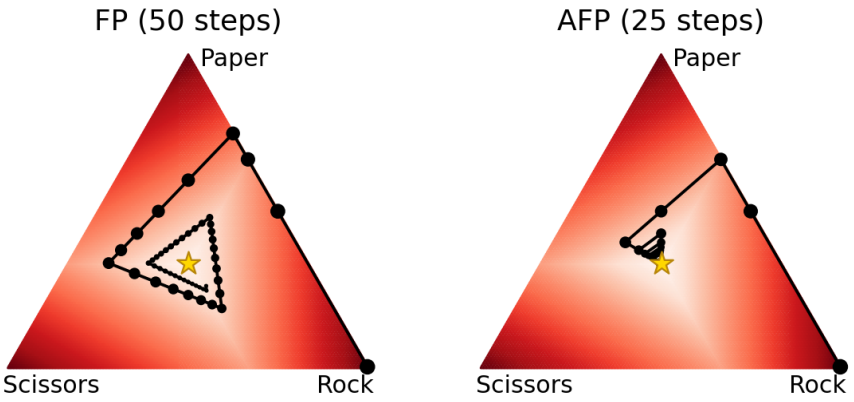

Figure 3.1: A visualization of the first 50 steps of FP $(\overline{x}_1^{\text{FP}}, \overline{x}_2^{\text{FP}}, \ldots, \overline{x}_{50}^{\text{FP}})$ and first 25 steps of AFP $(\overline{x}_1^{\text{AFP}}, \overline{x}_2^{\text{AFP}}, \ldots, \overline{x}_{25}^{\text{AFP}})$ on Rock Paper Scissors. This corresponds to an equal amount of computation per algorithm (50 best responses). Ties between best response strategies were broken according to the ordering 'Rock,' 'Paper,' 'Scissors.' The Nash equilibrium is marked by a star. The shading indicates the exploitability of the strategy at that point, with darker colors representing greater exploitability.

To this end, we propose *anticipatory fictitious play* (AFP), a version of fictitious play that "anticipates" the best response an adversary might play against the current strategy, and then plays the best response to an average of that and the adversary's current average strategy. (Simply responding directly to the opponent's response does not work; see Appendix B.1.) Alternatively, one can think of AFP as a version of FP that "forgets" every other best response it calculates. This latter interpretation enables a convenient implementation of AFP as a modification of FP as demonstrated in Algorithm 2.

Remark. The AFP update is seemingly consistent with human psychology: it is quite intuitive to imagine how an adversary might try to exploit oneself and to respond in order to best counter that strategy. Given that fictitious play provides a model for how humans or other non-algorithmic decision makers might arrive at an equilibrium in practice (Luce and Raiffa 1989; Brown 1951) anticipatory fictitious play offers a new model for how this may occur. We leave further consideration of this topic to future work.

AFP is given by the following process. For some $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$, let $x_1 = \overline{x}_1 = e_i$ and $y_1 = \overline{y}_1 = e_j$ be initial strategies for each player. For each $t \in \mathbb{N}$,

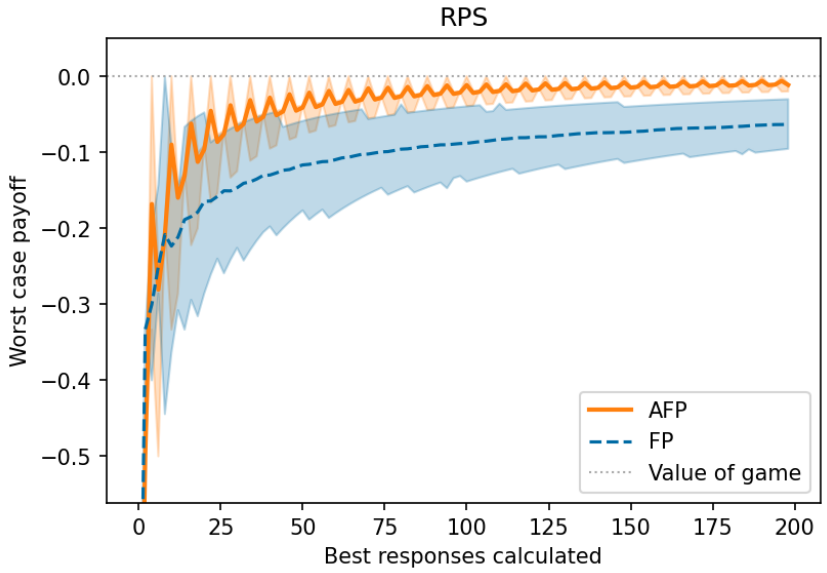

Figure 3.2: Comparison of FP and AFP performance ($\min \overline{x}_t^\intercal A$) on RPS with random tiebreaking. The highlighted region depicts the 10th and 90th percentiles across 10,000 runs. All variation is due to randomly sampled tiebreaking. The value of the game is $v^* = 0$.

define

$$
\begin{aligned}
x'_{t+1} &\in \mathtt{BR}^1_A(\overline{y}_t); & y'_{t+1} &\in \mathtt{BR}^2_A(\overline{x}_t);\\
\overline{x}'_{t+1} &= \tfrac{t}{t+1}\overline{x}_t + \tfrac{1}{t+1}x'_{t+1}; & \overline{y}'_{t+1} &= \tfrac{t}{t+1}\overline{y}_t + \tfrac{1}{t+1}y'_{t+1};\\
x_{t+1} &\in \mathtt{BR}^1_A(\overline{y}'_{t+1}); & y_{t+1} &\in \mathtt{BR}^2_A(\overline{x}'_{t+1});\\
\overline{x}_{t+1} &= \frac{1}{t+1}\sum_{k=1}^{t+1} x_t; & \overline{y}_{t+1} &= \frac{1}{t+1}\sum_{k=1}^{t+1} y_t.
\end{aligned}
\tag{3.1}
$$

Here, x'_{t+1} and y'_{t+1} are the best response to the opponent's average strategy. They are the strategies that FP would have played at the current timestep. In AFP, each player "anticipates" this attack and defends against it by calculating the opponent's average strategy that include this attack ($\overline{x}'_t$ and $\overline{y}'_t$), and then playing the best response to the anticipated average strategy of the opponent.

In Figure 3.1, we see the effect of anticipation geometrically: AFP "cuts corners," limiting the extent to which it overshoots its target. In contrast, FP aggressively overshoots, spending increasingly many steps playing strategies that take it further from its goal.

The effect on algorithm performance is pronounced, with AFP hovering near equilibrium while FP slowly winds its way there.

Of course, RPS is a very specific example. It is natural to wonder: is AFP good in general? The rest of the chapter seeks to answer that question. We begin by proving that AFP converges to a Nash equilibrium.

Proposition 1. If $\{(x_t, y_t)\}$ is an AFP process for a 2p0s game with payoff matrix $A \in \mathbb{R}^{m \times n}$, the conclusion of Theorem 1 holds for this process. Namely, AFP converges to a Nash equilibrium, and it converges no slower than the rate that bounds FP.

Proof. (Idea) Generalize the original proof of Theorem 1. We work with accumulating payoff vectors $U(t) = tA^\intercal \overline{x}_t$ and $V(t) = tA\overline{y}_t$. In the original proof, a player 1 strategy index $i \in \{1, \ldots, m\}$ is called *eligible* at time t if $i \in \arg\max V(t)$ (similarly for player 2). We replace eligibility with the notion of *E-eligibility*, satisfied by an index $i \in \arg\max[V(t) + E]$, for any $E \in \mathbb{R}^m$ with $\|E\|_\infty < \max_{i,j} |A_{i,j}|$. Essentially, an index is E-eligible if it corresponds to a best response to a perturbation of the opponent's history $\overline{y}_t$ or a perturbation of the game itself. The original proof structure can be preserved in light of this replacement, requiring only minor modifications to some arguments and new constants. Treating the in-between strategies in AFP, $\overline{x}'_t$ and $\overline{y}'_t$, as perturbations of $\overline{x}_t$ and $\overline{y}_t$, it follows that AFP satisfies the conditions for the generalized result. A complete proof is given in Appendix B.2. □

3.4 Application to normal form games

Proposition 1 establishes that AFP converges and that AFP's worst-case convergence rate satisfies the same bound as FP's, where the worst-case is with respect to games and tiebreaking rules. The next proposition shows that for two classes of games of interest, AFP not only outperforms FP, but attains an optimal rate. In both classes, our proofs will reveal that AFP succeeds where FP fails because AFP avoids playing repeated strategies. The results hold for general applications of FP and AFP rather than relying on specific tiebreaking rules.

The classes of games that we analyze are intended to serve as abstract models of two fundamental aspects of real-world games: transitivity (akin to "skillfullness;" some ways of acting are strictly better than others) and nontransitivity (most notably, in the form of strategy cycles like Rock < Paper < Scissors < Rock). Learning algorithms for

real-world games must reliably improve along the transitive dimension while accounting for the existence of strategy cycles; see Balduzzi et al. (2019) for further discussion.

For each integer $n \geq 3$, define payoff matrices C^n and T^n by

$$C^n_{i,j} = \begin{cases} 1 & \text{if } i = j+1 \mod n; \\ -1 & \text{if } i = j-1 \mod n; \\ 0 & \text{otherwise;} \end{cases} \quad \text{and} \quad T^n_{i,j} = \begin{cases} (n-i+2)/n & \text{if } i = j+1; \\ -(n-i+2)/n & \text{if } i = j-1; \\ 0 & \text{otherwise,} \end{cases}$$

for $i, j \in \{1, \ldots, n\}$. The game given by C^n is a purely cyclic game: each strategy beats the one before it and loses to the one after it; C^3 is the game Rock Paper Scissors. For each C^n, a Nash equilibrium strategy is $[n^{-1}, \ldots, n^{-1}]^\intercal$. The game given by T^n could be considered "transitive:" each strategy is in some sense better than the last, and $[0, \ldots, 0, 1]^\intercal$ is a Nash equilibrium strategy. The payoffs are chosen so that each strategy i is the unique best response to $i-1$, so that an algorithm that learns by playing best responses will progress one strategy at a time rather than skipping to directly to strategy n (as would happen if FP or AFP were applied to a game that is transitive in a stronger sense, such as one with a single dominant strategy; c.f. the definition of transitivity in (Balduzzi et al. 2019)). Note: T^n could be defined equivalently without the n^{-1} factor, but this would create a spurious dependence on n for the rates we derive.

The following proposition establishes a convergence rate of $O(t^{-1})$ for AFP applied to C^n and T^n. This rate is optimal within the class of time-averaging algorithms, because the rate at which an average changes is t^{-1}. Note: we say a random variable $Y_t = \Omega_p(g(t))$ if, for any $\epsilon > 0$, there exists $c > 0$ such that $P[Y_t < cg(t)] < \epsilon$ for all t.

Proposition 2. FP and AFP applied symmetrically to C^n and T^n obtain the rates given in Table 3.1. In particular, if $\{x_t, x_t\}_{t \in \mathbb{N}}$ is an FP or AFP process for a 2p0s game with payoff matrix $G \in \{C^n, T^n\}$ with tiebreaking as indicated, then $\max G\overline{x}_t = R(t)$. Tiebreaking refers to the choice of $x_{t+1} \in \arg\max \mathtt{BR}^1_G(\overline{x}_t)$ when there are multiple maximizers. The "random" tiebreaking chooses between tied strategies independently and uniformly at random. For entries marked with "arbitrary" tiebreaking, the convergence rate holds no matter how tiebreaks are chosen.

Proof. (Sketch, C^n) Full proofs of all cases are provided in Appendix B.3. Define $\Delta_0 = [0, \ldots, 0]^\intercal \in \mathbb{Z}^n$ and $\Delta_t = tC^n\,\overline{x}_t$ for each $t \in \mathbb{N}$. The desired results are equivalent to

Table 3.1: Convergence rates for FP and AFP on C^n and T^n.

Algorithm	Game G	Tiebreaking	Rate $R(t)$	Caveats
FP	C^n	random	$\Omega_p(t^{-1/2})$	
AFP	C^n	arbitrary	$O(t^{-1})$	$n = 3, 4$
FP	T^n	arbitrary	$\Omega(t^{-1/2})$	$t < t^*(n)$
AFP	T^n	arbitrary	$O(t^{-1})$	

$\max \Delta_t = O_p(\sqrt{t})$ under FP for the given tiebreaking rule, and $\max \Delta_t$ is bounded under AFP for $n = 3, 4$. Let i_t be the index played by FP (AFP) at time t (so $x_t = e_{i_t}$). It follows that

$$\Delta_{t+1,j} = \begin{cases} \Delta_{t,j} - 1 & \text{if } j = i_t - 1 \mod n; \\ \Delta_{t,j} + 1 & \text{if } j = i_t + 1 \mod n; \\ \Delta_{t,j} & \text{otherwise;} \end{cases} \tag{3.2}$$

for each $t \in \mathbb{N}_0$ and $j \in \{1, \ldots, n\}$. Note that the entries of Δ_t always sum to zero.

In the case of FP, it is easy to verify that $\max \Delta_t$ is nondecreasing for any choice of tiebreaking. For each $m \in \mathbb{N}_0$, define $t_m = \inf\{t \in \mathbb{N}_0 : \max \Delta_t = m\}$. Then by Markov's inequality,

$$P(\max \Delta_t < m) = P(t_m > t) \leq E(t_m)/t = \frac{1}{t} \sum_{k=0}^{m-1} E(t_{k+1} - t_k).$$

Examining the timesteps at which $i_{t+1} \neq i_t$ and relating them to $\{t_k\}$, we show in the appendix that the time to increment the max from k to $k+1$ satisfies $E(t_{k+1} - t_k) = O(k)$. Thus the bound above becomes $P(\max \Delta_t < m) \leq O(m^2)/t$. Now let $c \in \mathbb{R}^{\geq 0}$ be arbitrary and plug in $c\lceil\sqrt{t}\rceil$ for m, so we have $P(\max \Delta_t < c\sqrt{t}) \leq c^2 O(1) \to 0$ as $c \to 0$. So $\max \Delta_t = \Omega_p(\sqrt{t})$.

For the AFP case, consider the first timestep at which $\max \Delta_t = m + 1$. Working backwards and checking cases, it can be shown that in order for the maximum value to increment from m to $m + 1$, there must first be a timestep where there are two non-adjacent entries of m with an entry of $m - 1$ between them. This cannot happen in the $n = 3, m = 2$ case because three positive entries (2,1,2) don't sum to zero. Similarly,

in the $n = 4, m = 2$ case, it turns out by (3.2) that $\Delta_t = [a, b, -a, -b]$ for some a, b. So there cannot be three positive entries in this case either. Therefore $\max_t \Delta_t \leq 2$ for $n = 3, 4$. □

The proofs of Proposition 2 establish a theme: FP can be slow because it spends increasingly large amounts of time progressing between strategies (playing $x_t = x_{t+1} = \cdots = x_{t+k}$ with k increasing as t increases), whereas AFP avoids this. (Return to Figure 3.1 for a visual example.)

Some further comments on the results: we only obtain the $O(t^{-1})$ rate for AFP applied to C^n in the $n = 3, 4$ case. We conjecture that: (i) for a specific tiebreaking rule, AFP has the same worst-case rate as FP but with a better constant, (ii) under random tiebreaking, AFP is $O_p(t^{-1})$ for all n.

Our results are noteworthy for their lack of dependence on tiebreaking: worst-case analyses of FP typically rely on specific tiebreaking rules; see Daskalakis and Pan (2014), for example. As for the "$t < t^*(n)$" caveat for FP applied to T^n, this is an unremarkable consequence of analyzing a game with a pure strategy equilibrium (all probability assigned to a single strategy). We write $t^*(n)$ to indicate the first index at which FP plays e_n. Both FP and AFP will play e_n forever some finite number of steps after they play it for the first time, thus attaining a t^{-1} rate as the average strategy "catches up" to e_n. Our result shows that until this point, FP is slow, whereas AFP is always fast.

For a visualization of the $n = 20$ case, demonstrating AFP's superiority despite these caveats, see Appendix B.4.

3.4.1 Numerical results

In order to compare FP and AFP more generally, we sample large numbers of random payoff matrices and compute aggregate statistics across them. Matrix entries are sampled as independent, identically distributed, standard Gaussian variables (note that the shift- and scale-invariance of matrix game equilibria implies that the choice of mean and variance is inconsequential). Since FP and AFP are so similar, and AFP computes two best responses per timestep, it's natural to wonder: is AFP's superior performance just an artifact of using more computation per timestep? So, in order to make a fair comparison, we compare the algorithms by *the number of best responses calculated* instead of the number of timesteps (algorithm iterations). Using the worst-case payoff as the measure

of performance, we compare FP and AFP based on the number of responses computed and based on matrix size in Figures 3.3 and 3.4.

The result is that AFP is clearly better on both counts. Although FP is better for a substantial proportion of 30×30 games at very early timesteps t, AFP quickly outpaces FP, eventually across each of 1,000 matrices sampled. In terms of matrix size, FP and AFP appear equivalent on average for small matrices, but quickly grow separated as matrix size grows, with AFP likely to be much better.

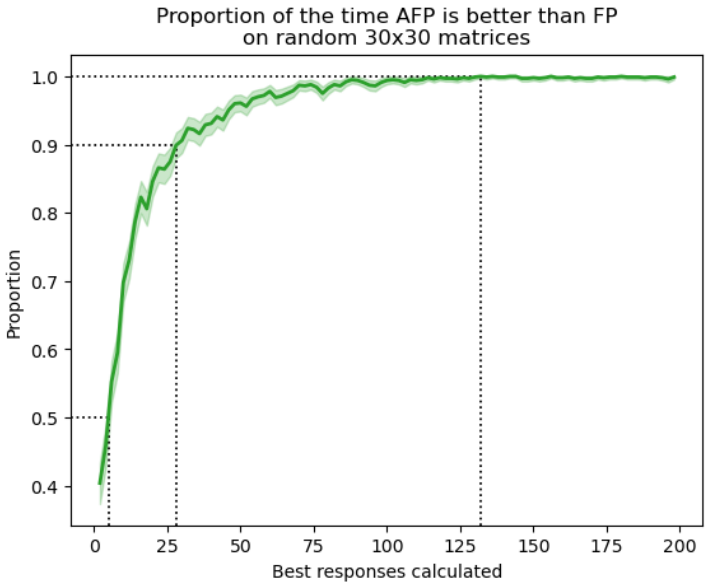

Figure 3.3: For 1,000 randomly sampled (30,30) matrices A, the proportion of the time that $\min(\overline{x}_{r/2}^{\text{AFP}})^\intercal A \geq \min(\overline{x}_r^{\text{FP}})^\intercal A$ for $r = 2, 4 \ldots, 200$. A 95% Agresti-Coull confidence interval (Agresti and Coull 1998) for the true proportion is highlighted. Note that after only about six best responses, AFP is better half the time, and by 130, AFP is better than FP essentially 100% of the time.

3.5 Application to reinforcement learning

We apply reinforcement learning (RL) (Sutton and Barto 2018) versions of FP and AFP in the context of a (two-player, zero-sum, symmetric) *stochastic game* (Shapley (1953)), defined by the tuple $(\mathcal{S}, \mathcal{O}, \mathcal{X}, \mathcal{A}, \mathcal{P}, \mathcal{R}, p_0)$, where $\mathcal{S}$ is the set of possible states of the

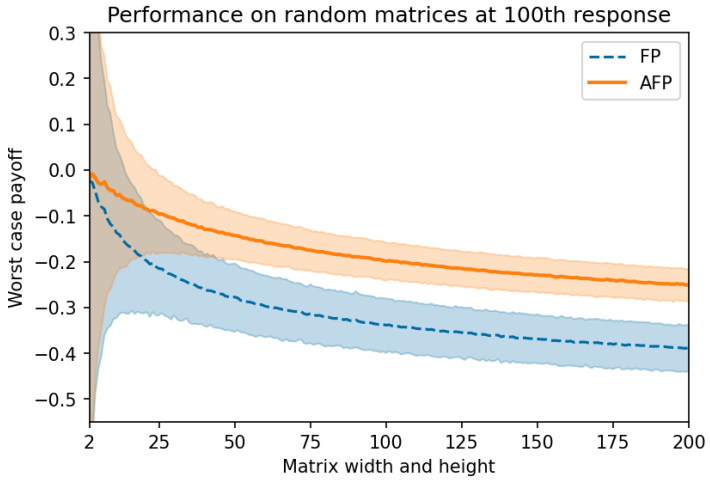

Figure 3.4: Average performance of FP vs. AFP at the 100th best response (timestep 100 for FP, timestep 50 for AFP) as matrix size is varied. All matrices are square. Highlighted regions show the 10th and 90th percentiles.

environment, $\mathcal{O}$ is the set of possible observations received by an agent, $\mathcal{X} : \mathcal{S} \to \mathcal{O} \times \mathcal{O}$ gives the observations for each player based on the current state, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{A} \to \Delta(\mathcal{S})$ defines the transition dynamics for the environment given each player's action, $\mathcal{R} : \mathcal{S} \to \mathbb{R} \times \mathbb{R}$ defines the reward for both players such that $\mathcal{R}(s_t) = (r_t, -r_t)$ are the rewards observed by each player at time t, and $p_0 \in \Delta(\mathcal{S})$ is the initial distribution of states, such that $s_0 \sim p_0$. Let H be the set of possible sequences of observations. Then a *policy* is a map $\pi : H \to \Delta(A)$. An *episode* is played by iteratively transitioning by the environment according to the actions sampled from each players' policies at each state. Players 1 and 2 earn *returns* $(\sum_t r_t, -\sum_t r_t)$. The reinforcement learning algorithms we consider take sequences of observations, actions, and rewards from both players and use them to incrementally update policies toward earning greater expected returns. For background on reinforcement learning, see Sutton and Barto (2018). For details on machine learning approximations to FP, see Heinrich et al. (2015). Table 3.2 gives a high-level overview of the relationship.

The stochastic game we choose is **TinyFighter**, a minimal version of an arcade-style fighting game shown in Figure 3.5. It features two players with four possible actions: `Move Left`, `Move Right`, `Kick`, and `Do Nothing`. Players are represented by a rectangular body and when kicking, extend a rectangular leg towards the opponent.

Table 3.2: The normal-form game analogies used to extend FP and AFP to reinforcement learning.

Normal-form game	Stochastic/extensive-form game
Strategy	Policy
Payoff $A_{i,j}$	Expected return $E_{\pi_i,\pi_j}(\sum_t r_t)$
Best response	Approximate best response by RL
Strategy mixture $\sum \alpha_i x_i$, $\sum \alpha_i = 1$, $\alpha_i \geq 0$	At start of episode, sample policy π_i with probability α_i. Play entire episode with π_i.

Kicking consists of three phases: Startup, Active, and Recovery. Each phase of a kick lasts for a certain number of frames, and if the Active phase of the kick intersects with any part of the opponent (body or leg), a hit is registered. When a hit occurs, the players are pushed back, the opponent takes damage, and the opponent is stunned (unable to take actions) for a period of time. In the Startup and Recovery phases, the leg is extended, and like the body, can be hit by the opponent if the opponent has a kick in the active phase that intersects the player. The game is over when a player's health is reduced to zero or when time runs out.

Player observations are vectors in $\mathbb{R}^{13}$ and contain information about player and opponent state: position, health, an 'attacking' indicator, a 'stunned' indicator, and how many frames a player has been in the current action. The observation also includes the distance between players, time remaining, and the direction of the opponent (left or right of self). The game is partially observable, so information about the opponent's state is hidden from the player for some number of frames (we use four, and the game runs at 15 frames per second). This means a strong player must reason about the distribution of actions the opponent may have taken recently and to respond to that distribution; playing deterministically will allow the opponent to exploit the player and so a stochastic strategy is required to play well.

3.5.1 Adapting FP and AFP to reinforcement learning

Neural Population Learning (NeuPL) (Liu et al. 2022) is a framework for multiagent reinforcement learning wherein a collection of policies is learned and represented by a single neural network and all policies train continuously. For our experiments, we implement FP and AFP within NeuPL, as shown in Algorithm 1. For reference, we also include a

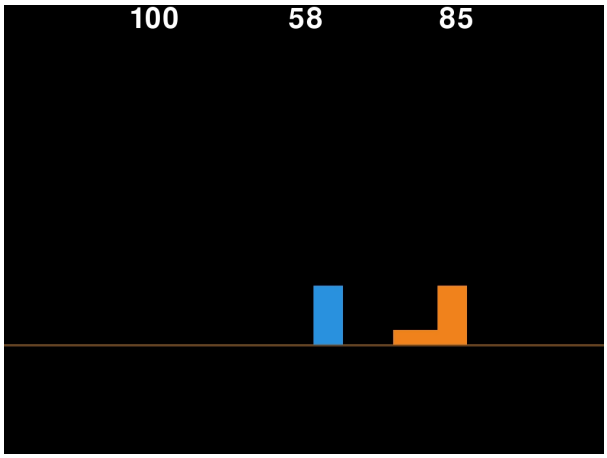

Figure 3.5: A screenshot of the TinyFighter environment, where players attempt to land kicks in order to reduce the other's health from 100 to zero.

simple RL version of FP and AFP in the style of PSRO in Appendix B.6.

Algorithm 1 NeuPL-FP/AFP

$\mathfrak{O} \in \{\mathfrak{O}^{\text{FP}}, \mathfrak{O}^{\text{AFP}}\}$ ▷ Input: FP or AFP opponent sampler.
$\{\Pi_\theta(t) : H \to \Delta(\mathcal{A})\}_{t=1}^{n}$ ▷ Input: neural population net.
for Batch $b = 1, 2, 3, \ldots,$ **do**
    $B \leftarrow \{\}$
    while per-batch compute budget remains **do**
        $T_{\text{learner}} \sim \text{Uniform}(\{1, \ldots, n\})$
        $T_{\text{opponent}} \sim \mathfrak{O}(T_{\text{learner}})$
        $D_{\text{learner}} \leftarrow \textsc{PlayEpisode}(\Pi_\theta(T_{\text{learner}}), \Pi_\theta(T_{\text{opponent}}))$
        $B \leftarrow B \cup D_{\text{learner}}$
    end while
    $\Pi_\theta \leftarrow \textsc{ReinforcementLearningUpdate}(B)$
end for

The terms in the algorithm are as follows: the opponent sampler $\mathfrak{O}$ determines the distributions of opponents that each agent faces and is the only difference between the

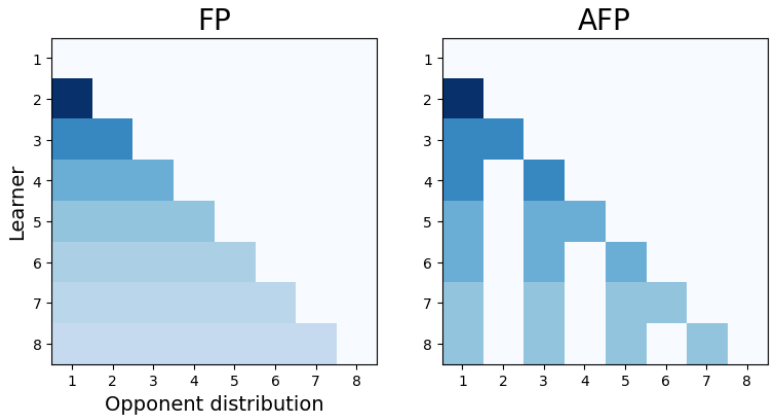

Figure 3.6: A visual depiction of the distributions of opponents ("meta-strategies" in PSRO or NeuPL) each learner faces in a population learning implementation of FP or AFP. The (i, j)-entry is the probability that, given that agent i has been sampled to train in a given episode, it will face agent j in that episode. Dark blue indicates probability 1, white indicates probability 0.

FP and AFP implementations. We have, for each $t > 1$,

$$\mathfrak{O}^{\mathrm{FP}}(t) = \mathrm{Uniform}(\{1, 2, 3, \ldots, t-1\}), \text{ and}$$
$$\mathfrak{O}^{\mathrm{AFP}}(t) = \mathrm{Uniform}(\{k < t : k \text{ odd}\} \cup \{t-1\}).$$

These distributions are depicted in Figure 3.6. Just as each step of FP involves computing a best response to an average against all prior strategies, sampling from $\mathfrak{O}^{\mathrm{FP}}(t)$ corresponds to training agent t uniformly against the prior policies; just as AFP can be thought of "forgetting" every other index, $\mathfrak{O}^{\mathrm{AFP}}(t)$ trains learner index t uniformly against every odd indexed policy plus the most recent policy. The neural population net $\Pi_\theta(t) : H \to \Delta(A)$ defines a different policy for each agent index t, and can equivalently be represented as $\Pi_\theta(a|s, t)$.

3.5.2 Experimental setup

For the population neural net, we used a simple actor-critic (Sutton and Barto 2018) architecture with a separate actor and critic network. The actor is a 3-layer dense MLP with ReLU activations and layer widths (512, 256, 256, 4) with action masking applied prior to a Softmax layer. Its inputs are a concatenation of the player observation and a vector

representing the distribution of opponents faced by the currently training agent index t (e.g., $[0.5, 0.5, 0, \ldots, 0]$ for agent $t = 3$ in FP or AFP). The critic is the same, except layer widths are (256, 256, 128, 1) and an additional input is used: the index of the opponent sampled at the beginning of this episode, in line with the original implementation Liu et al. (2022).

Note that the actor (policy network) does not observe which opponent it faces, only the *distribution* over agents it faces; this is important because otherwise our agent would not learning a best response to an average policy as intended in FP and AFP. The reason for including this information for the critic (value network) is that it may reduce the variance of the value function estimator (although, of course, it could increase the variance in estimated state values by conditioning on information that would normally be averaged over).

We implemented NeuPL within a basic self-play reinforcement learning loop by wrapping the base environment (TinyFighter) within a lightweight environment that handles NeuPL logic, such as opponent sampling. For reinforcement learning, we use the Asynchronous Proximal Policy Optimization (APPO) algorithm (Schulman et al. 2017), a distributed actor-critic RL algorithm, as implemented in RLLib (Moritz et al. 2018) with a single GPU learner and 80 workers. Hyperparameter settings are given in Appendix B.5. We use a population size of $n = 8$. We train the entire neural population net (agents 1-8) for 8,000 steps, where a step is roughly 450 minibatch updates of stochastic gradient descent. This corresponds to about 80 hours of training, during which time 31 million episodes are played. We repeat this procedure independently five times for FP and five times for AFP.

3.5.3 Results

To evaluate exploitability, we made use of the fact that each FP and AFP neural population are made up of agents trained to "exploit" the ones that came before them. Specifically, each agent is trained to approximate a best response to the average policy returned by the algorithm at the previous timestep. So, to estimate the exploitability of NeuPL-FP or NeuPL-AFP at step $t \in \{1, \ldots, n-1\}$, we simply use the average return earned by agent $t+1$ against agents $\{1, \ldots, t\}$ to obtain the *within-population exploitability* of agent t. This metric is convenient, but insufficient on its own. In order for it to be useful, the agents in the population must have learned approximate best responses that

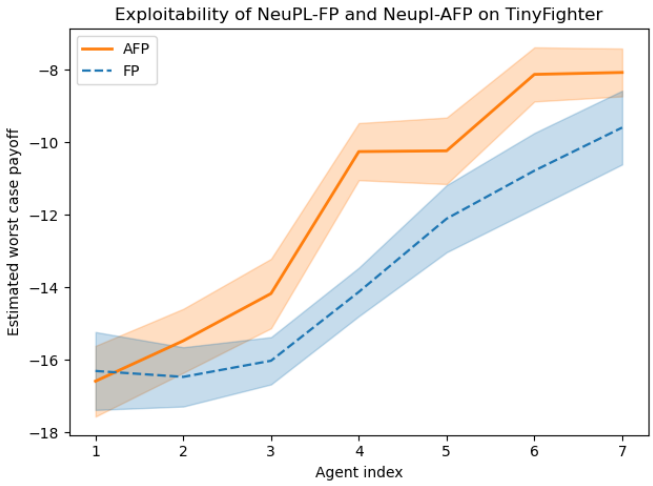

Figure 3.7: Comparison of NeuPL-FP and NeuPL-AFP on TinyFighter. Highlighting indicates a pointwise 90% confidence region.

are close to actual best responses; if they have not, it could be that within-population exploitability is low, not because the average policy approximates a Nash policy but because nothing had been learned at all. To account for this, we also evaluate the populations learned using *relative population performance* (Balduzzi et al. 2019), which measures the strength of one population of agents against the other. The purpose of using relative population performance is simply to verify that one algorithm did not produce generally more competent agents than the other.

We paired each of the five replicates of FP and AFP and computed the relative population performance for each, obtaining values of 1.85, 3.35, -4.60, -0.45, and -0.81, for an average of -0.13 and a Z-test based 90% confidence interval width of 1.99. This suggests that the relative strength of agents produced by FP and AFP is not significantly different from 0. So, it is reasonable to use within-population exploitability to compare FP and AFP, as shown in Figure 3.7.

We find that AFP has lower estimated exploitability, with the average policy at $t = 7$ earning -8.07 (standard error 0.61) average total reward against agent 8, whereas the same statistic for FP was -9.57 (standard error 0.40) average total reward. While these results are favorable for AFP, they are on the borderline of statistical significance and it is our view that further study is needed, including on other stochastic games. *For further experiments and experimental details, see Cloud et al. (2022), the preprint corresponding*

to this chapter.

3.6 Conclusion

We proposed a variant of fictitious play for faster estimation of Nash equilibria in two-player, zero-sum games. Anticipatory fictitious play is intuitive, easy to implement, and supported by theory and numerical simulations which suggest that it is virtually always preferable to fictitious play. Consequently, we shed new light on two motivating problems for fictitious play: primarily, large-scale multiagent reinforcement learning for complicated real-world games; also, modeling strategic decision making in humans.

3.7 Acknowledgements

Ryan Martin provided mentorship and guidance with the creation of this chapter, Philip Wardlaw helped orchestrate preliminary reinforcement learning experiments, Adam Venis and Angelo Olcese contributed ideas for the proofs of FP and AFP applied to C^n, Jesse Clifton and Eric Laber provided helpful comments on a draft, and Jesse Clifton and Marc Lanctot suggested related works that had been overlooked.

Chapter 4

Safety-constrained online learning in contextual bandits

4.1 Introduction

Bandit algorithms (Lai and Robbins 1985) allow for ongoing experimentation and optimization in sequential decision making problems in health care (Villar et al. 2015), ad targeting (Schwartz et al. 2017; Sawant et al. 2018), education (Rafferty et al. 2018), and other personalized services (Balakrishnan et al. 2018; Li et al. 2010). In the basic model of a multi-armed bandit, a decision maker faces a sequence of identical choices between k actions, or *arms*. Each arm has an unknown distribution of real-valued rewards that is sampled from independently and identically each time the arm is pulled. The goal of the learner is to maximize the cumulative reward earned. To achieve this, the learner must balance acting to gather information with acting to earn reward, facing the so-called exploration-exploitation tradeof. Contextual bandits (Wang et al. 2005; Goldenshluger and Zeevi 2011) extend the multi-armed bandit model by including an independent and identically random distributed variable on which the reward distribution depends. Bandits have been extensively studied and are commonly used for real-world decision making (Slivkins 2019).

Although multi-arm and contextual bandit models see wide use, they may not be suitable for safety-critical settings, where the cost of certain actions can be very high, or where there are other constraints on algorithm behavior that cannot be enforced by modifying the reward distributions of a bandit model. Problems in these settings

may be better modeled as safety-constrained learning problems, in which a learning algorithm seeks to maximize reward subject to a constraint on the effect of its actions which it respects with high probability. For example, an algorithm deployed within a mobile health application might tailor messages to diabetic patients to encourage them to exercise, subject to the constraint that the patient's glucose level stays in a safe range. In robotics, an autonomous driving system might be expected to navigate to its destination subject to the constraint that no humans are hurt or property is damaged. An ad-targeting system for a website might have a goal to maximize a user's chance of clicking an ad, subject to the constraint that the user does not terminate use of the website.

In this chapter we present an algorithmic framework for safety-constrained data-driven decision making that departs from previous approaches. We present our framework in the setting of a linear contextual bandit with two outcomes, the usual "reward" outcome and an additional real-valued "safety" outcome. In this setting, an algorithm is expected to maximize expected reward subject to a constraint on the expected value of the safety outcome. We prove that in the limit, an algorithm in our framework respects the safety constraint and achieves optimal reward subject to that constraint. We also derive the limiting distribution of scaled versions of the estimators used, allowing for approximate statistical inference. An outline of the chapter is as follows: in Section 4.2 we introduce contextual bandits and basic reinforcement learning terminology before formalizing a notion of a safe and consistent bandit learning algorithm. In Section 4.3 we outline different approaches to the problem of safety-constrained learning and give a detailed account of our proposed method, Split-Propose-Test. In Section 4.4 we provide theoretical results on the performance of our method. In Section 4.5 we compare our method with a baseline across a variety of simulated bandit environments. We close with a discussion of the contribution and the broad areas of the problem that are ripe for further study.

4.1.1 Related work

Past work on data-driven decision making under safety constraints has operationalized the problem in various ways. For example, some have assumed that the safety constraint is a functional of the same outcome that is being optimized (Laroche et al. 2019; Thomas et al. 2015; Wu et al. 2016; Kazerouni et al. 2017), which is a special case of the multi-outcome setting we consider here. An example of this would be "maximize profit in

expectation, subject to the constraint that profit must be nonnegative with high probability." A common criterion in this line of work is that an update to an algorithm's behavior must earn at least as much reward in expectation as the previous iteration with high probability.

Others consider the constrained multi-outcome learning problem in more general settings than bandits, like Markov Decision Processes (MDPs), but make strong assumptions about knowledge of environmental dynamics, such as Berkenkamp et al. (2021), which considers the case of MDPs but requires specification of a Gaussian process prior distribution over environment dynamics. Others address the problem in complicated settings but do not provide theoretical guarantees; for example, Lütjens et al. (2019) uses ensembles of neural networks to obtain heuristic uncertainty estimates to enable robots to navigate environments while avoiding novel obstacles; similarly, Balata et al. (2021) considers various methods for continuous time MDPs, none of which provably satisfy a safety constraint with high probability. In the context of medical decision making, Huang and Xu (2020) and Laber et al. (2018) deal with the offline setting, which allows for unsafe data collection and only requires safety of a newly estimated policy after data are collected.

Problem settings very similar to ours have been studied in Amani et al. (2019, 2020), Moradipari et al. (2021), and Pacchiano et al. (2021), which consider safety-constrained learning in different versions of a linear contextual bandit setting. Unlike the other works and ours, Amani et al. (2019) defines the set of safe actions in terms of a known linear transformation of the unknown *reward* parameter, with no observation of a safety outcome. Moradipari et al. (2021) and Pacchiano et al. (2021) use a formulation that is nearly equivalent to ours, wherein stochastic observations at each timestep provide data with which to estimate the set of safe actions. All of these works propose algorithms which work in the following way: at each timestep, (1) estimate a set of actions that, with high probability, contains no unsafe actions; (2) apply a standard reinforcement learning criterion for reward-maximization, i.e., Thompson Sampling (Thompson 1933; Russo et al. 2018) or UCB (Auer 2002), limited to the estimated action set. Our proposed algorithm uses a different strategy to guarantee safe action selection, enabling better performance in problems with large action spaces. In our simulation experiments, we use the Safe-LTS algorithm of Moradipari et al. (2021) as a competitive baseline for comparison.

4.2 Problem formulation

We consider a contextual bandit, which is defined by an independent and identically distributed (i.i.d.) collection of random variables (defined on a sample space Ω),

$$\{X_t, Y_t^*(a) : a \in \mathcal{A}\}_{t \in \mathbb{N}},$$

where $X_t : \Omega \to \mathcal{X}$ is a random context observed at time t, and $Y_t^*(a) : \Omega \to \mathcal{Y}$ is the potential outcome that would be observed at time t if action $a \in \mathcal{A}$ were taken at time t (Rubin 1978; Splawa-Neyman et al. 1990). We will sometimes drop the subscript to refer to a generic sample X. Let A_t be the action selected at time t, and assume that $Y_t = Y_t^*(A_t)$ is observed at time t. The history at time T is composed of the observed data up to that point, $H_T = \{(X_t, A_t, Y_t)\}_{t=1}^T$. Let $\Delta(\mathcal{A})$ be the set of probability distributions over $\mathcal{A}$. A *policy* $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ determines a distribution over actions based on the observed context. A *learning algorithm* is a collection of data-dependent policies $\{\pi_t\}_{t \in \mathbb{N}}$ such that each π_T is determined by H_{T-1}. Throughout this chapter, unless stated otherwise, we assume that actions are sampled independently from some learning algorithm, $A_T \sim \pi_T(X_T)|H_{T-1}$ for all $T \in \mathbb{N}$, so $\pi_T(a|x) = P(A_T = a|X_T = x, H_{T-1})$ for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $T \in \mathbb{N}$. We will occasionally abuse notation by using $\pi(x)$ to refer the random variable sampled from a policy evaluated at $x \in \mathcal{X}$.

In the safety-constrained contextual bandit setting, depicted in Figure 4.1, we are given functions $r : \mathcal{Y} \to \mathbb{R}$ and $s : \mathcal{Y} \to \mathbb{R}$ that define scalar outcomes of interest, *reward* $R_t = r(Y_t)$ and *safety* $S_t = s(Y_t)$. The goal is to maximize expected reward while satisfying a constraint on the expected safety. This constraint is defined in terms of a known, fixed *baseline policy* $\pi_0 : \mathcal{X} \to \mathcal{A}$ which represents a decision rule that is considered to be acceptable *a priori*. For example, in the clinical setting, π_0 might be "the standard of care:" a rule for assigning treatments that is well-established and considered proper in the healthcare community.

Define action-value functions for reward and safety,

$$Q^r(x, a) = E(R_t|X_t = x, A_t = a), \text{ and}$$
$$Q^s(x, a) = E(S_t|X_t = x, A_t = a).$$

Note that we could replace Q^s with another functional of the conditional distribution of

safety, such as a quantile (Leqi and Kennedy 2021). Throughout, we assume that Q^r and Q^s are linear with respect to a known feature vector $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, so that for some $\beta^r \in \mathbb{R}^d$ and $\beta^s \in \mathbb{R}^d$,

$$Q^r(x,a) = Q^r(x,a;\beta^r) = \phi(x,a)^\intercal \beta^r, \text{ and}$$
$$Q^s(x,a) = Q^s(x,a;\beta^r) = \phi(x,a)^\intercal \beta^s,$$

for all $x \in \mathcal{X}$, $a \in A$. All results presented here would also hold if the two outcomes had different feature vectors, ϕ^r and ϕ^s, but we state it this way for simplicity.

For each context $x \in \mathcal{X}$, and for some fixed, known safety tolerance $\tau \geq 0$, define the set of τ-safe actions

$$\mathcal{A}_x^\tau = \big\{a \in \mathcal{A} : Q^s(x,a) \geq Q^s[x, \pi_0(x)] - \tau\big\}.$$

Define a learning algorithm $\{\pi_t\}_{t\in\mathbb{N}}$ as (α,τ)-safe if

$$P[\pi_T(X) \notin \mathcal{A}_X^\tau] \leq \alpha + o_p(1). \tag{4.1}$$

Although (4.1) is stated in terms of limiting behavior, we are interested in algorithms for which the rate of unsafe action selection is controlled a rate close to α in finite samples, and for which safety with probability 1 is eventually attained. Finally, define a learning algorithm $\{\pi_t\}_{t\in\mathbb{N}}$ as (α,τ)-consistent if it is (α,τ)-safe and

$$P\left\{Q^r[X,\pi_T(X)] \geq \sup_{a\in\mathcal{A}_X^\tau} Q^r(X,a)\right\} \to 1 \text{ as } T \to \infty. \tag{4.2}$$

In other words, an algorithm is (α,τ)-consistent if it is safe and it earns at least as much reward as optimal under the safety constraint. Our goal is to construct an (α,τ)-consistent learning algorithm.

4.3 Split-Propose-Test

Given a context $X_t = x$, a natural way to decide whether action a is τ-safe is to perform a statistical test of the hypothesis $\mathcal{H}_1 : a \in \mathcal{A}_x^\tau$ against the null $\mathcal{H}_0 : a \notin \mathcal{A}_x^\tau$. In other words, we start with the default assumption that the action is not safe, and only conclude

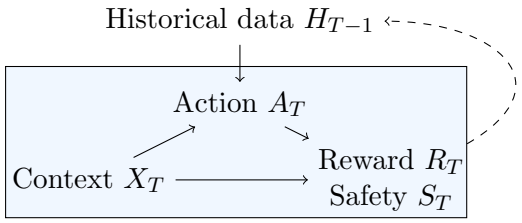

Figure 4.1: A contextual bandit with two outcomes of interest. At each timestep $T = 1, 2, \ldots$, an i.i.d. draw of X_T is observed, an action is A_T is selected based on X_T and historical data H_{T-1}, and outcomes R_T and S_T are observed.

it is safe it there is strong evidence that suggests that it is safe. If the null is rejected, we say that the action *passes the safety test*. If an action passes the safety test, then we can be confident that the action is safe. To help motivate our proposed procedure, consider three test-based options for selecting an action at time t, given context X_t and history H_{t-1}.

1. (Test first) For all actions $a \in \mathcal{A}$, use H_{t-1} to perform a hypothesis test of $\mathcal{H}_0 : a \notin \mathcal{A}^{\tau}_{X_t}$. Then, maximize estimated expected reward over the actions that pass the test.

2. (Optimize first, ignoring safety) Pick an action that is estimated to maximize reward, then test it for safety. Choose this action if it passes the safety test.

3. (Optimize first, accounting for safety) Pick an action that is estimated to be optimal according to a criterion defined in terms of both reward and safety, then test it for safety. Choose this action if it passes the safety test.

In each case, if no safe action is identified, the default $\pi_0(X_t)$ is chosen.

Each of these approaches has shortcomings. Option 1 (Test first) is the strategy used by Moradipari et al. (2021) and Pacchiano et al. (2021). It faces the challenges of multiple testing. If we do not correct for multiple testing, then the safety condition (4.1) will not hold; if we correct for multiple testing, we may suffer from a lack of power to detect safe actions which results in the algorithm following standard of care and failing to improve. Option 2 (Optimize first, ignoring safety) is a non-starter because in some problems reward and safety may trade off against each other. As a simple example, suppose that $Q^r(x, a) = -Q^s(x, a)$ for all x, a. In this case, optimizing for reward while ignoring safety

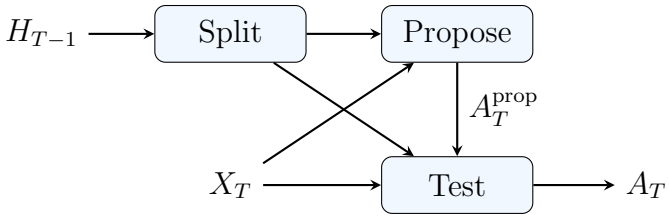

Figure 4.2: A schematic of Split-Propose-Test.

means picking unsafe actions. These actions will not pass the safety test and hence the algorithm will not learn. Finally, Option 3 (Optimize first, accounting for safety) avoids multiple testing and performs optimization in a way that is aware of safety. However, if we use the entirety of the historical data both to decide which action likely has high expected reward and to test whether the action is safe, the safety test may not be valid due to the dependence between the action selection and test.

We propose a variant of Option 3 called Split-Propose-Test (SPT), where the problem of data reuse is mitigated by sample splitting. For any $T \in \mathbb{N}$ and index set $I \subseteq \{1, \ldots, T\}$, let $H(I) = \{(X_t, A_t, Y_t) : t \in I\}$. A high-level overview of SPT is as follows: at each timestep $T \in \mathbb{N}$,

1. (Split) Partition H_{T-1} into a "proposal" dataset $H(I_{T-1}^{\text{prop}})$ and a "test" dataset $H(I_{T-1}^{\text{test}})$.
2. (Propose) Use the proposal dataset to propose an action A_T^{prop}.
3. (Test) Use the test dataset to test if A_T^{prop} is safe. If the test passes, choose A_T^{prop}. Otherwise, choose the default $\pi_0(X_T)$.

The SPT procedure is depicted in Figure 4.2. We specify the details of each step in subsequent subsections, but before this, we introduce some notation.

Define the estimators

$$\begin{aligned} \widehat{\beta}^{r,I} &= \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{t \in I} [R_t - \phi(X_t, A_t)^\intercal \beta]^2, \text{ and} \\ \widehat{\beta}^{s,I} &= \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{t \in I} [S_t - \phi(X_t, A_t)^\intercal \beta]^2. \end{aligned} \tag{4.3}$$

In Section 4.3.1 we will describe choices for I_T^{prop} and I_T^{test}. For now, write $\widehat{\beta}_T^{r,\text{prop}} = \widehat{\beta}^{r,I_T^{\text{prop}}}$, $\widehat{\beta}_T^{s,\text{prop}} = \widehat{\beta}^{s,I_T^{\text{prop}}}$, and $\widehat{\beta}_T^{s,\text{test}} = \widehat{\beta}^{s,I_T^{\text{test}}}$.

For $x \in \mathcal{X}$, $a \in \mathcal{A}$, and index set I, let $\Psi_x^{\tau,\eta}(a, I)$ be a nominally level-η hypothesis test for $\mathcal{H}_0 : a \notin \mathcal{A}_x^\tau$ computed as a function of data $H(I)$, where $\Psi_x^{\tau,\eta}(a, I) = 1$ indicates rejection of $\mathcal{H}_0$ (action a passes the safety test) and $\Psi_x^{\tau,\eta}(a, I) = 0$ indicates failure to reject $\mathcal{H}_0$ (action a does not pass the safety test). Let A_t^{prop} be an action proposed based on $H(I_T^{\text{prop}})$. SPT relies on the following parameters: a sequence of exploration parameters $\{\epsilon_t\}_{t\in\mathbb{N}}$ with $\epsilon_t \in [0, 1]$ for each t; a sequence of safety test levels $\{\eta_t\}_{t\in\mathbb{N}}$ with $\eta_t \in [0, 1]$ for each t.

Let $\{U_t\}_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \text{Uniform}([0, 1])$ be independent of the contextual bandit. The action returned by SPT at time T is

$$A_T = \pi_T^{\text{SPT}}(X_T) = \begin{cases} A_T^{\text{random}} & \text{if } U_T \leq \epsilon_T; \\ A_T^{\text{prop}} & \text{if } U_T > \epsilon_T \text{ and } \Psi_{X_T}^{\tau,\eta_T}(A_T^{\text{prop}}; I_{T-1}^{\text{test}}) = 1; \\ \pi_0(X_T) & \text{otherwise;} \end{cases} \tag{4.4}$$

where $A_T^{\text{random}} \overset{\text{iid}}{\sim} \text{Uniform}(\mathcal{A})$. Each of the three cases serves a purpose: choosing A_T^{random} a nonzero proportion of the time guarantees that we generate data that is diverse enough to reliably estimate β^r and β^s; we would like to choose A_T^{prop}, our best guess at a safe and high-reward action, but we only choose it if it passes the safety test; $\pi_0(X_T)$ is known to be safe so we fall back to it if A_T^{prop} does not pass the safety test.

4.3.1 Split step details

Sample splitting enables the use of an approximately valid hypothesis test of a single action when that action was chosen based on the data, thereby avoiding the need to test all actions for safety prior to selecting one based on its estimated reward.

For $t \leq T \in \mathbb{N}$, let $Z_{T,t}$ be a Bernoulli random variable that indicates membership of data point t in the propose set. For $T \in \mathbb{N}$, define $I_T^{\text{prop}} = \{t \in \{1, \ldots, T\} : Z_{T,t} = 1\}$ and $I_T^{\text{test}} = \{t \in \{1, \ldots, T\} : Z_{T,t} = 0\}$. To simplify the analysis, our proofs will assume that sample splitting indicators are independent and identically distributed, i.e., that $Z_{T,t} = Z_t$ for all t, T, and $\{Z_t\}_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} \text{Bernoulli}(p_{\text{prop}})$, with $p_{\text{test}} = 1 - p_{\text{prop}}$.

However, more general schemes should work as well. A more general condition on $\{Z_{T,t}\}_{t\leq T\in\mathbb{N}}$ (which we do not prove) would be: for each $T \in \mathbb{N}$, $Z_{T,1}, \ldots, Z_{T,T}$ are Bernoulli random variables expressible as a function only of $\{(X_t, A_t)\}_{t=1}^T$ and some external source of randomness, with $P(Z_{T,t} = 1|H_T)$ bounded away from 0 and 1 with

probability 1 for all $t \leq T \in \mathbb{N}$. In our simulations, we use stratified sampling based on actions. For each $a \in \mathcal{A}$ and $T \in \mathbb{N}$, define

$$\mathcal{T}_{T,a} = \big\{t \in \{1, \ldots, T\} : A_t = a\big\}.$$

At each T, for each $a \in \mathcal{A}$, form an independent random partition $\mathcal{T}_{T,a} = \widehat{\mathcal{T}}_{T,a}^{\text{prop}} \cup \widehat{\mathcal{T}}_{T,a}^{\text{test}}$. Then $Z_{T,t} = \mathbb{1}(t \in \cup_{a \in \mathcal{A}} \widehat{\mathcal{T}}_{T,a}^{\text{prop}})$, and $Z_{T,1}, \ldots, Z_{T,T}$ are (dependent) Bernoulli variables that satisfy the condition stated above.

4.3.2 Propose step details

In this section, we derive a data-dependent objective function over $\mathcal{A}$ that will be used to determine the action proposed at each timestep. The derivation is based on the goal of maximizing the reward earned on the current timestep. We begin with the insight that, as long as the safety test is valid at level η, *any* manner of choosing proposal actions is guaranteed to be safe, because unsafe proposed actions will not pass the safety test at a rate greater than η. Given this, it is acceptable to simply choose A_T^{prop} to maximize the expected reward of A_T.

In SPT, the *improvement* over the default policy of the selected action A_T is $Q^r[X_T, A_T] - Q^r[X_t, \pi_0(X_T)]$. During the Propose step, our goal is to choose A_T^{prop} so as to make this term large. However, we must do so without peeking at the test set. So, we consider maximizing the improvement term in expectation. We assume here that $\epsilon_T = 0$, as the chance of selecting an action at random is irrelevant to the analysis. Define the *expected improvement of* $a \in \mathcal{A}$ at context $X_T = x$ to be

$$J_T(x, a) = E\left\{Q^r[X_T, A_T] - Q^r[X_T, \pi_0(X_T)] \,|\, X_T = x, A_T^{\text{prop}} = a\right\},$$

where the conditioning in this expression is to be interpreted as an intervention to set A_T^{prop} equal to a. Note that the expectation here is over the distribution of $H(I_{T-1}^{\text{test}})$. The function $J_T(x, a)$ does not depend on the test set, so it is a good candidate for an action proposal criterion. However, it depends on unknown quantities which must be estimated.

In order to find an estimator, note that, by iterated expectation and (4.4),

$$\begin{aligned} J_T(x,a) &= E\left[E\left\{Q^r[x,A_T] - Q^r[x,\pi_0(x)]\,\middle|\,\Psi_x^{\tau,\eta_T}(A_T^{\text{prop}}; I_{T-1}^{\text{test}}), A_T^{\text{prop}} = a\right\}\,\middle|\, A_T^{\text{prop}} = a\right] \\ &= \left\{Q^r[x,\pi_0(x)] - Q^r[x,\pi_0(x)]\right\} P[\Psi_x^{\tau,\eta_T}(a; I_{T-1}^{\text{test}}) = 0] \\ &\quad + \left\{Q^r(x,a) - Q^r[x,\pi_0(x)]\right\} P[\Psi_x^{\tau,\eta_T}(a; I_{T-1}^{\text{test}}) = 1] \\ &= \left\{Q^r(x,a) - Q^r[x,\pi_0(x)]\right\} P[\Psi_x^{\tau,\eta_T}(a; I_{T-1}^{\text{test}}) = 1]. \end{aligned} \tag{4.5}$$

Although Q^r and $P[\Psi_x^{\tau,\eta_T}(a; I_{T-1}^{\text{test}}) = 1]$ are unknown quantities, we can estimate them using data $H(I_{T-1}^{\text{prop}})$. Define

$$\widehat{J}_T(x,a) = \left\{Q^r(x,a;\widehat{\beta}_{T-1}^{r,\text{prop}}) - Q^r[x,\pi_0(x);\widehat{\beta}_{T-1}^{r,\text{prop}}]\right\} P\left[\Psi_x^{\tau,\eta_T}(a; \widetilde{I}_{T-1}^{\text{prop}}) = 1 \,\middle|\, H(I_{T-1}^{\text{prop}})\right], \tag{4.6}$$

where $\widetilde{I}_{T-1}^{\text{prop}}$ is a multiset formed by resampling with replacement from I_{T-1}^{prop}. Specifically, $\widetilde{I}_{T-1}^{\text{prop}} = \{\widetilde{t}_1, \ldots \widetilde{t}_{|I_{T-1}^{\text{prop}}|}\} \overset{\text{iid}}{\sim} \text{Uniform}(I_{T-1}^{\text{prop}})$ is a bootstrap resampling of the proposal dataset. As such, the probability measure in (4.6) is with respect to these resamplings, and whatever source of variation is used to generate these samples (conditional on I_{T-1}^{prop}) is independent of all past and future data.

Finally, the proposed action is

$$A_T^{\text{prop}} \in \arg\max_{a \in \mathcal{A}} \widehat{J}_T(X_T, a),$$

which, based on the proposal data only, we expect to maximize the reward earned at this timestep.

4.3.3 Test step details

Let $x \in \mathcal{X}$, $a \in \mathcal{A}$ be fixed. In this section we develop the testing procedure for

$$\mathcal{H}_1 : a \in A_x^\tau \text{ against } \mathcal{H}_0 : a \notin A_x^\tau,$$

using data $H(I_T)$ for some index set $I_T \subseteq \{1, \ldots, T\}$. To do this, we use a normal approximation to the sampling distribution of the difference of action-value estimates.

The approximation is

$$Q^s(x,a;\widehat{\beta}^{s,I_T}) - Q^s[x,\pi_0(x);\widehat{\beta}^{s,I_T}] \overset{\text{approx.}}{\sim} N\left\{Q^s(x,a) - Q^s[x,\pi_0(x)]\,,\ \frac{(\widehat{\sigma}^{s,I_T})^2(x,a)}{T}\right\},$$

where $(\widehat{\sigma}^{s,I_T})^2(x,a)$ is defined by the following terms:

$$\begin{aligned}
\widehat{\boldsymbol{\phi}}^{I_T} &= \frac{1}{|I_T|}\sum_{t\in I_T}\phi(X_t,A_t)\phi(X_t,A_t)^\intercal,\\
\widehat{\Sigma}^{s,I_T} &= \frac{1}{|I_T|}\sum_{t\in I_T}\phi(X_t,A_t)\phi(X_t,A_t)^\intercal[S_t - \phi(X_t,A_t)^\intercal\widehat{\beta}^{s,I_T}]^2, \text{ and finally}\\
(\widehat{\sigma}^{s,I_T})^2(x,a) &= \{\phi(x,a)-\phi[x,\pi_0(x)]\}^\intercal(\widehat{\boldsymbol{\phi}}^{I_T})^{-1}\widehat{\Sigma}^{s,I_T}(\widehat{\boldsymbol{\phi}}^{I_T})^{-1}\{\phi(x,a)-\phi[x,\pi_0(x)]\}.
\end{aligned}$$

For generic $\tau\in\mathbb{R}$, $\eta\in[0,1]$, and $x\in\mathcal{X}$, the test for safety of action $a\in\mathcal{A}$ using data I_T is given by

$$\Psi_x^{\tau,\eta}(a;I_T) = \mathbb{1}\left\{\sqrt{T}\frac{\{\phi(x,a)-\phi[x,\pi_0(x)]\}^\intercal\widehat{\beta}^{s,I_T}+\tau}{\widehat{\sigma}^{s,I_T}(x,a)} > z_{1-\eta}\right\}, \tag{4.7}$$

where $z_{1-\eta}$ is the $1-\eta$ quantile of a standard normal distribution. Consistency of the safety test under various choices of I_T is given by Lemma 2.

4.4 Theory

In this section we state results about the behavior of SPT. Proofs are given in Appendix C.2. We begin with a general result adapted from Chen et al. (2021), that the OLS estimators used by SPT are consistent, then give a result that the safety tests used by SPT converge as desired. With appropriately chosen parameters, these results then imply our main theorem, which is that SPT is (α,τ)-consitent (as defined Section 4.2). We end with a result on the asymptotic normality of the sample-splitting OLS estimators in SPT and a corollary that relates the theorem to the structure of SPT. To obtain these results, we make the following assumptions.

(A1) The action space $\mathcal{A}$ is finite.

(A2) There is a constant $L_\phi<\infty$ such that $P\left[\|\phi(X,a)\|_\infty < L_\phi\right]=1$ for all $a\in\mathcal{A}$.

(A3) Sample splitting indicators are indendent and identically distributed, $Z_{T,t} = Z_t \overset{\text{iid}}{\sim} \text{Bernoulli}(p_{\text{prop}})$, with $0 < p_{\text{prop}} < 1$.

(A4) $\lambda_{\min}(\Sigma) > 0$, where $\Sigma = E_{A \sim \text{Uniform}(\mathcal{A})}[\phi(X, A)\phi(X, A)^\intercal]$.

(A5) Let $e^r = R - E(R|X, A)$ and $e^s = S - E(S|X, A)$. These errors are uniformly subgaussian in the sense that there exists $\sigma^2 > 0$ such that

$$E[\exp(ce^r)|X = x, A = a] \leq \exp(\sigma^2 c^2/2) \text{ for all } c \in \mathbb{R},$$
$$E[\exp(ce^s)|X = x, A = a] \leq \exp(\sigma^2 c^2/2) \text{ for all } c \in \mathbb{R},$$

for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

(A6) $P\{Q^s(X, a) - Q^s[X, \pi_0(X)] + \tau = 0\} = 0$ for all $a \in \mathcal{A}$.

Assumptions (A1), (A2), and (A3) are straightforward conditions which, in a real-world application, can be ensured by choosing an appropriate representation for the problem. Assumption (A1) could be weakened by introducing smoothness conditions on ϕ, and (A3) could be weakened to history-dependent sample splitting such as described in Section 4.3.1; we leave such extensions to future work. Assumptions (A4) and (A5) are similar to those in Chen et al. (2021) and allow for consistent estimation of β^r and β^s. Assumption (A6) is a kind of margin condition to ensure that the safety test converges to a non-stochastic limit. Without this condition, an action a might exist exactly on the safety threshold. Its safety test would have a nonzero probability of both passing and not passing a, even in the limit, which would complicate our analysis. However, the main results should be unchanged so long as the reward-maximizing action does not exist on the safety margin. Furthermore, SPT can still be guaranteed to perform reasonably without this condition, although it may not enjoy (α, τ)-consistency. See Appendix C.1 for a further discussion of assumptions.

Lemma 1. If (A1), (A2), (A3), (A4), and (A5) hold and if at each timestep t there is at least an ϵ_t-chance of selecting an action uniformly at random from $\mathcal{A}$, with $\{\epsilon_t\}$ non-increasing and $T\epsilon_T^2 \to \infty$, then for any choice of $I_T = I_T^{\text{prop}}, I_T^{\text{test}}, \{1, \ldots, T\}$, or $\widetilde{I}_T^{\text{prop}}$,

$$\widehat{\beta}^{r, I_T} \xrightarrow{\text{p}} \beta^r, \text{ and}$$
$$\widehat{\beta}^{s, I_T} \xrightarrow{\text{p}} \beta^s$$

as $T \to \infty$.

Lemma 2. If the conditions for Lemma 1 hold, (A6) holds, safety test levels $\{\eta_t\}_{t\in\mathbb{N}}$ are such that $0 < \liminf \eta_t \leq \limsup \eta_t < 1$, then for any choice of $I_T = I_T^{\text{prop}}, I_T^{\text{test}}$, $\{1, \ldots, T\}$, or $\widetilde{I}_T^{\text{prop}}$, the safety test as defined in (4.7) is consistent in the sense that

$$\Psi_{X_{T+1}}^{\eta_T,\tau}(a, I_T) \xrightarrow{\text{p}} \begin{cases} 1 & \text{if } Q^s(X_{T+1}, a) > Q^s[X_{T+1}, \pi_0(X_{T+1})] - \tau; \\ 0 & \text{otherwise;} \end{cases}$$

as $T \to \infty$.

Theorem 2. If (A1), (A2), (A3), (A4), (A5), and (A6) hold and SPT's parameters satisfy $\epsilon_1 \leq \eta$, $T\epsilon_T^2 \to \infty$, $\epsilon_T \to 0$, and ϵ_T is non-increasing, and $\eta_T = \alpha - \epsilon_T$, then Split-Propose-Test satisfies (4.1) and (4.2), i.e., it is (α, τ)-consistent.

An example choice of $\{\epsilon_t\}$ satisfying the above conditions is $\epsilon_t = \alpha\, t^{-1/3}/2$. Note that we choose the safety test parameters $\{\eta_t\}$ so that $\epsilon_t + \eta_t = \alpha$, so that the rate of exploratory actions plus the rate of unsafe actions that pass the safety test is no greater than the nominal safety level.

Theorem 2 establishes that SPT converges to optimal safe behavior in the limit. In fact, it is slightly stronger than stated, because as the proof in Appendix C.2 reveals, the probability of SPT selecting an unsafe action tends towards 0.

Two additional assumptions yield asymptotic normality of the OLS estimators, which is useful for performing statistical inference but also valuable for understanding why SPT uses sample splitting, as explained by a subsequent corollary. For $x \in \mathcal{X}$ and $a \in \mathcal{A}$, define

$$J(x, a) = \{Q^r(x, a) - Q^r[x, \pi_0(x)]\}\, \mathbb{1}(a \in \mathcal{A}_x^\tau). \tag{4.8}$$

The first assumption is that with probability 1 there is a unique best safe action.

(B1) $P\{J(X, a)$ has a unique maximum over $\mathcal{A}\} = 1$.

As argued at the end of the proof of Theorem 2, (B1) implies existence of a limiting policy $\pi_\infty : \mathcal{X} \to \Delta(\mathcal{A})$ such that

$$\max_{a\in\mathcal{A}} \left|\pi_T^{\text{SPT}}(a|X) - \pi_\infty(a|X)\right| \xrightarrow{\text{p}} 0 \text{ as } T \to \infty. \tag{4.9}$$

This is important because our proof of the asymptotic normality of the OLS estimators relies on existence of a limiting policy. The second assumption is that a covariance matrix defined with respect to the limiting policy is invertible.

(B2) The matrix $\boldsymbol{\phi}_{\pi_\infty} = E_{\pi_\infty}[\phi(X,A)\phi(X,A)^\intercal]$ is invertible.

Theorem 3. If the conditions for Theorem 2 hold and (B1) and (B2) hold, then the OLS estimators for reward and safety in the two-outcome bandit are asymptotically normal, and if estimated on different datasets, asymptotically independent. Specifically,

$$\sqrt{T}\begin{bmatrix}\widehat{\beta}_T^{r,\text{prop}} - \beta^r \\ \widehat{\beta}_T^{s,\text{prop}} - \beta^s \\ \widehat{\beta}_T^{s,\text{test}} - \beta^s\end{bmatrix} \rightsquigarrow N\left(\begin{bmatrix}0\\0\\0\end{bmatrix}, \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B} & \mathbf{0} \\ \boldsymbol{B} & \boldsymbol{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{D}\end{bmatrix}\right)$$

as $T \to \infty$, where

$$\begin{aligned}
\boldsymbol{A} &= (p_{\text{prop}})^{-1}(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^r_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1}, \\
\boldsymbol{B} &= (p_{\text{prop}})^{-1}(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^{r,s}_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1}, \\
\boldsymbol{C} &= (p_{\text{prop}})^{-1}(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^s_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1}, \\
\boldsymbol{D} &= (p_{\text{test}})^{-1}(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^s_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1},
\end{aligned}$$

and

$$\begin{aligned}
\Sigma^r_{\pi_\infty} &= E_{\pi_\infty}\{[R - Q^r(X,A)]^2\phi(X,A)\phi(X,A)^\intercal\} \\
\Sigma^{r,s}_{\pi_\infty} &= E_{\pi_\infty}\{[R - Q^r(X,A)][S - Q^s(X,A)]\phi(X,A)\phi(X,A)^\intercal\} \\
\Sigma^s_{\pi_\infty} &= E_{\pi_\infty}\{[S - Q^s(X,A)]^2\phi(X,A)\phi(X,A)^\intercal\}.
\end{aligned}$$

Based on the limiting distribution of the OLS estimators, Theorem 3 allows us to perform approximately valid statistical inference for β^r and β^s. It is a special case of Theorem 4, a more general result about estimating an arbitrary (finite) number of parameters based on different (possibly overlapping) splits of data in the contextual bandit setting. See Appendix C.3.2 for details.

As formalized in the following corollary, Theorem 3 also implies that the proposal action and safety test are asymptotically independent, validating the choice of sample

splitting and lack of multiple-testing correction in SPT. Let

$$Z_T(x,a) = \sqrt{T}\frac{\{\phi(x,a) - \phi[x,\pi_0(x)]\}^\intercal(\widehat{\beta}^{s,I_T^{\text{test}}} - \beta^s)}{\widehat{\sigma}^{s,I_T^{\text{test}}}(x,a)},$$

so the safety test can be written as

$$\Psi_x^{\tau,\eta}(a; I_T^{\text{test}}) = \mathbb{1}\left(Z_T(x,a) + \frac{Q^s(x,a) - Q^s[x,\pi_0(x)] + \tau}{\widehat{\sigma}^{s,I_T^{\text{test}}}(x,a)} > z_{1-\eta}\right).$$

Define

$$(\sigma^s)^2(x,a) = \{\phi(x,a) - \phi[x,\pi_0(x)]\}^\intercal(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^s_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1}\{\phi(x,a) - \phi[x,\pi_0(x)]\}.$$

Corollary 1. For any $x \in \mathcal{X}$ and $a \in \mathcal{A}$, $Z_T(x,a) \rightsquigarrow N(0,1)$ and $\widehat{\sigma}^{s,I_T^{\text{test}}}(x,a) \xrightarrow{\text{p}} \sigma^s(x,a)$ as $T \to \infty$ and are asymptotically independent of $(\widehat{\beta}_T^{r,\text{prop}}, \widehat{\beta}_T^{s,\text{prop}})$.

Corollary 1 elucidates the benefit of the Split step in Split-Propose-Test: splitting the data maintains the integrity of the safety test via approximate independence from the quantities used to compute the proposed action. Corollary 1 is an important addendum because, technically, Theorem 2 is provable without sample splitting, i.e. using $I_T^{\text{prop}} = I_T^{\text{test}} = \{1,\ldots,T\}$ for each $T \in \mathbb{N}$. This is because, broadly speaking, the only fundamental requirement for SPT to work in the limit is that the estimators for β^r and β^s are consistent. However, a version of SPT without sample splitting could have terrible performance in finite samples due to data reuse between the action proposal and the safety test. Such reuse would corrupt the safety test, removing then nominal error rate guaranteed by the testing procedure. This would cause problems in bandits when errors e^s and e^r are highly correlated, or, perhaps more strikingly, in cases where $R = S$. However, these conditions are not necessary for this version of SPT to exhibit unsafe behavior, which can happen even when $R \neq S$ and e^r and e^s are independent. We give such an example in Appendix C.4.2.

4.5 Simulation experiments

In this section, we evaluate SPT on a variety of simulated contextual bandit problems designed to highlight different aspects of safety-constrained learning. We compare SPT

with a simple competitor and a hybrid of SPT and the competitor, each of which is (α, τ)-consistent. We evaluate these three algorithms based on their finite sample performance in terms of mean reward and satisfaction of the safety constraint. Each of the algorithms is run independently many times on a variety of simulated bandit problems, with the problems chosen to emphasize different challenges of safe reinforcement learning. The algorithms are as follows:

- **Pretest All** (with Thompson Sampling): test all actions for safety and pick the one with the highest estimated reward among the ones that passed the safety test; this is essentially the algorithm proposed by Moradipari et al. (2021), extended to a contextual bandit setting. In particular, use the safety test from Section 4.3.3, plugging in the whole dataset H_{T-1} in for $H(I_{T-1}^{\text{test}})$ and applying a Bonferroni correction factor to preserve the level of the safety test. Select among the actions that passed this conservative safety test according to estimated reward. To improve exploration, we use the frequentist analogue to Thompson Sampling as follows. Let

$$\widetilde{\beta}_T^r = \underset{\beta \in \mathbb{R}^d}{\arg\min} \sum_{t=1}^{T} W_{T,t} \left[R_t - \phi(X_t, A_t)^\intercal \beta \right]^2,$$

  where $W_T \overset{\text{iid}}{\sim} \text{Multinomial}(T, T^{-1}, \ldots, T^{-1})$ are multinomial bootstrap weights (Efron and Tibshirani 1994). At timestep T, based on history H_{T-1}, with $\{U_t\}_{t \in \mathbb{N}} \overset{\text{iid}}{\sim} \text{Uniform}([0,1])$ as in the setup to SPT, Pretest All selects

$$A_T = \begin{cases} A_T^{\text{random}} & \text{if } U_T < \epsilon_T \\ \underset{\left\{a \in \mathcal{A} \,:\, \Psi_{X_T}^{\tau,\alpha/|\mathcal{A}|}(a, H_{T-1}) = 1\right\}}{\arg\max} \phi(X_T, a)^\intercal \widetilde{\beta}_{T-1}^r & \text{otherwise.} \end{cases}$$

  Like SPT, Pretest All is (τ, α)-consistent because the safety test is consistent and $\widetilde{\beta}_T^r \xrightarrow{\text{p}} \beta^r$ as $T \to \infty$, as proved in Appendix C.4.3.

- **SPT**: as described in Section 4.3, with action selection rule (4.4). We use $\eta_t = \alpha$ for all t. The sample splitting performed at every step uses stratified sampling by action, as described in Section 4.3.1.

- **SPT (fallback)**: SPT with $\eta_t = \alpha/2$, but if the proposed action does not pass the safety test, select the action returned by Pretest All with safety level $\alpha/2$.

Table 4.1: High-level summaries of each bandit setting.

Setting	Bandit type	Parameters	Summary
All safe	Standard		All actions are safe but this isn't known a priori.
Dosage bandit	Linear	$\rho_{r,s} = -0.5, 0, 0.5$	Reward and safety trade off against each other and information is pooled between nearby actions, allowing for gradual expansion of the action space.
Single-arm detection	Standard	$\lvert\mathcal{A}\rvert = 5, 10, 15$	There is a single safe, high-reward arm amongst many poor options.
Uniform-armed bandit	Standard		Reward and safety are functionals of the same outcome.
Orthogonal actions	Contextual	dot $= -0.5, 0, 0.5$	Each arm has its own parameters for reward and safety, which may point in opposing, orthogonal, or similar directions.
Noisy bandit	Contextual	$d_{\text{noise}} = 5, 10, 15$	A standard bandit augmented with features that have zero effect on reward and safety.

The following parameters are shared between all algorithms and all runs: $\pi_0(x) = 0$ for all x, $\epsilon_t = 0.1t^{-0.1}$, and $\alpha = 0.1$. For each algorithm-environment pair, we ran over 2,000 independent runs of the algorithm for $T = 350$ or 300 timesteps. For each experiment, at least four samples per arm ("burn-in samples") were gathered before initializing the algorithm.

4.5.1 Bandit setting descriptions

In this section we give detailed descriptions of each simulated bandit problem. High-level summaries of each setting are given in Table 4.1.

Let $\boldsymbol{e}_i \in \mathbb{R}^d$ denote the ith standard Euclidean basis vector. We say bandit is a multi-armed or "standard" bandit if $\phi(a) = \boldsymbol{e}_{a+1}$, we say it is a linear bandit if $\phi(x, a) = \phi(a)$ but it is not a standard bandit, and we say it is a contextual bandit otherwise. Except when otherwise stated, errors $e^r = R - E(R|X, A)$ and $e^s = S - E(S|X, A)$ are normally distributed and independent of (X, A). Also, unless otherwise stated, these errors have variance 1 and covariance 0. In bandits without contexts, we drop x as an input to Q^r, Q^s, and ϕ.

All safe

X	$\mathcal{A}$	τ	d
-	$\{0, 1, \ldots, 9\}$	0	10

This is a simple multi-arm bandit where all arms are safe and have varying expected rewards associated with them. In this setting, $\phi(a) = \boldsymbol{e}_{a+1} \in \mathbb{R}^d$, action 0 has mean safety 0, $Q^s(0) = 0$ and all other actions have safety 1, $Q^s(a) = 1$ for $a > 0$. At the start of each run, reward means are drawn from a multivariate standard normal distribution: $\beta_1^r, \ldots, \beta_d^r \overset{\text{iid}}{\sim} \text{Normal}(0, 1)$.

Dosage bandit

X	$\mathcal{A}$	τ	d
-	$[0, 1]$	0.1	10

The dosage bandit is a linear bandit that serves as a simple model of a medical decision making problem where the efficacy of a treatment trades off against its safety (e.g., tumor reduction vs. risk of toxicity (Thall and Cook 2004)). An action $a \in [0, 1]$ represents a dosage level of some treatment and $Q^r(a)$ is increasing in a while $Q^s(a)$ is decreasing in a. Consequently, the optimal action strikes a balance between efficacy and safety, achieving the greatest efficacy while having the lowest expected safety allowable based on the baseline policy π_0 and the safety tolerance τ.

The mean reward and safety functions are represented as linear combinations of d radial basis functions of a in order to approximate $Q^r(a) \approx 1 - \exp(-5a)$, $Q^s(a) \approx 1 - \exp[5(a - 1)]$. The centers of radial basis functions are evenly spaced in the unit interval $0 = c_1, \ldots, c_d = 1$, with $\varphi_c(a) = \exp(-d|a - c|)$. Errors are bivariate normal with variance 0.01 and covariance $0.01\rho_{r,s}$, where $\rho_{r,s} \in \{-0.5, 0, 0.5\}$.

Single-arm detection

X	$\mathcal{A}$	τ	d
-	$\{0, 1, \ldots, k\}$, $k = 5, 10, 15$	0	$\lvert\mathcal{A}\rvert$

This is a multi-armed bandit with many actions, one with much higher reward than the others. In particular, $Q^r(a) = 100 \cdot \mathbb{1}(a = 1)$, and $Q^s(a) = \mathbb{1}(a = 1)/2 - \mathbb{1}(a > 1)/2$. Because the reward-optimality of action 1 is so easy to detect, the problem essentially reduces to being able to detect that $a = 1$ is safe. The Single-arm detection bandit is a canonical motivating example for SPT, as SPT's ability to detect the safety of $a = 1$ crucially does not depend on the number of actions so long as the amount of data per action is fixed. On the other hand, typical multiple-testing procedures, including the one used by Pretest All, have power that tends to 0 as $|\mathcal{A}| \to 0$. This is illustrated in Appendix C.4.1.

Uniform-armed bandit

X	$\mathcal{A}$	τ	d
-	$\{0, 1, 2\}$	0.3	3

The uniform-armed bandit is a multi-armed bandit where the safety variable is defined in terms of reward. This corresponds to the special case studied in much of the safe reinforcement learning literature, where the safety constraint is a bound on a functional of the reward distribution.

Reward is distributed $R_t|A_t = a \sim \text{Uniform}([Q^r(a) - w_a, Q^r(a) + w_a])$, for some width $w_a > 0$, with $Q^r(a) > 0$ for each a. The safety outcome is defined as $S_t = \mathbb{1}(R_t \geq 0)$, so $Q^s(a) = P(R_t \geq 0|A_t = a)$. Consequently, the mean rewards $Q^r(a)$ determine the unconstrained quality of each action, and the widths w_a determine the associated risk that a safety-constrained algorithm must respect.

For the simulations here, $Q^r(0) = 0.5$, $Q^r(1) = 1$, and $Q^r(2) = 1.5$, with widths chosen so that $Q^s(0) = 1$, $Q^s(1) = 0.85$, and $Q^s(2) = 0.4$. With $\tau = 0.3$, this means that $a = 1$ is the optimal safe action, even though $a = 2$ earns greater reward in expectation and safety is defined in terms of reward.

Orthogonal actions

X	$\mathcal{A}$	τ	d
Normal$(\mathbf{0}_2, I_2)$	$\{0, 1, 2, 3, 4\}$	0	10

The orthogonal actions bandit is a contextual bandit where action a is represented by a unique parameter vector for reward and safety, so there is no information pooled across actions; equivalently, if $a \neq a'$ then $\phi(x, a)^\intercal \phi(x, a') = 0$ for all $x \in \mathcal{X}$. Each action $a \in \mathcal{A}$ is represented by a unique parameter vector $\beta_a^r \in \mathbb{R}^2$ and β_a^s. Then for $x \in \mathbb{R}^2$, $Q^r(x, a) = x^\intercal \beta_a^r$ and $Q^s(x, a) = x^\intercal \beta_a^s$. This means the full parameter vector $\beta^r = [(\beta_0^r)^\intercal, \ldots, (\beta_4^r)^\intercal]^\intercal \in \mathbb{R}^{10}$, and similarly for β^s, and $\phi(x, a) = [0, 0, x^\intercal, 0, \ldots, 0]^\intercal$ with the index of x occurring based on a. At the beginning of each run, parameter vector pairs (β_a^r, β_a^s) are sampled i.i.d. with marginal multivariate normal distributions such that

$$\frac{(\beta_a^r)^\intercal \beta_a^s}{\|\beta_a^r\|_2 \|\beta_a^s\|_2} = \text{dot}$$

for some value of dot $\in \{0.5, 0, -0.5\}$. Reward and safety errors are independently distributed, $e^r \sim N(0, 4^2)$, $e^s \sim N(0, 0.001)$.

Noisy bandit

X	$\mathcal{A}$	τ	d
Normal$(\mathbf{0}_{d_{\text{noise}}}, I_{d_{\text{noise}}})$	$\{0, 1, \ldots, 4\}$	0	$\lvert\mathcal{A}\rvert + d_{\text{noise}}$

The noisy bandit is a standard bandit augmented with a context vector in $\mathbb{R}^{d_{\text{noise}}}$. A learning algorithm must identify that the corresponding parameters are zero and instead learn appropriate estimates for single arms. The reward and safety parameters are defined as

$$\beta^r = [0, 10, 20, 30, 40, \mathbf{0}_{d_{\text{noise}}}^\intercal]^\intercal, \quad \beta^s = [0, -10, 33, -10, 20, \mathbf{0}_{d_{\text{noise}}}^\intercal]^\intercal,$$

and $\phi(x, a) = [\boldsymbol{e}_a^\intercal, \mathbf{0}_{d_{\text{noise}}}^\intercal]^\intercal$. Reward and safety errors are independently distributed, $e^r \sim N(0, 60^2)$, $e^s \sim N(0, 20^2)$.

4.5.2 Results

The average performance of each algorithm across timesteps $t = 1, \ldots, T$, with $T = 350$ or $T = 300$, is given in Table 4.2. Average performance at timestep $t = T$ is given in Table 4.3. Algorithm trajectories are plotted in Figure 4.3 for bandits without contexts and in Figure 4.4 for bandits with contexts.

The relative performance of SPT and Pretest All indicates how the algorithms handle different aspects of safety-constrained learning in the bandit setting. In general, we find that SPT has superior performance in all the context-free bandits studied, and that SPT, Pretest All, and SPT (fallback) all have very similar performance on the contextual bandits studied. This means that SPT has stronger performance with regards to the properties tested by those bandit settings, and taken together, suggest that SPT is a reliable choice in practice.

However, note that this does not suggest that SPT and Pretest All perform equivalently for bandits with contexts: it is straightforward to construct contextual bandit examples that inherit the same properties of the multi-arm bandit settings we consider; for example, one could modify the Orthogonal actions bandit to be (non-trivially) isomorphic to the All safe or Single-arm detection bandit by creating feature vectors that mean that all actions are safe, or such that one action has especially high reward and is uniquely safe. What our results in fact suggest is that the mere addition of contexts does not differentiate the performance of SPT and Pretest All: absent other distinguishing features, the two algorithms appear to perform roughly equivalently in general contextual bandits.

The standard bandits tested certain properties, as summarized in Table 4.1 and re-

capitulated here. The All safe bandit tests learning efficiency when everything is safe; the Dosage bandit tests learning that requires a gradual expansion of the estimated-to-be-safe action set when reward and safety trade off against each other; the single-arm detection bandit tests the power of safety testing procedures under large action spaces, and the Uniform-armed bandit tests learning when safety is defined as a functional of the reward.

As expected, the difference between SPT and Pretest All is most pronounced in the Single-arm detection setting, where SPT is much better able to identify a single best arm in the presence of many low quality options by virtue of the fact that it performs targeted hypothesis tests and has power to detect safe actions that is invariant to the size of the action space, in the sense discussed in Appendix C.4.1. We also see superior performance of SPT with respect to the other properties tested. Unsurprisingly, SPT (fallback) acts as an interpolation of SPT and Pretest All.

Of particular interest is the Dosage bandit example, where SPT's initial performance is poorer than that that of Pretest All. Despite this, SPT learns quickly, surpassing Pretest All after about 100 timesteps, without any cost in terms of satisfaction of the safety constraint. This result is encouraging as it demonstrates SPT's ability to explore plausibly good actions in a more targeted way.

In terms of safety, all algorithms obtain nominal safety on average and at the final timestep, although it takes a bit longer for SPT in some cases. This is not surprising, as the Z-test of SPT relies on an asymptotic approximation (see Section 4.3.3) which may break down in small samples. Although the same is true for Pretest All, the $1/|\mathcal{A}|$-multiple-testing correction applied to the safety test level introduces conservatism.

4.6 Discussion

We proposed the SPT framework for converting a hypothesis testing procedure into a safety-constrained online learning algorithm for contextual bandits, gave a theoretical account of its performance, and compared it to the baseline Pretest All algorithm, demonstrating comparable or superior performance across a variety of problem settings. In this work, our focus was on the general setting of multiple outcomes and on efficient exploration, especially in the case of multi-arm bandits with large action spaces which are challenging for past approaches. Our theory is fairly general, as the only margin condition

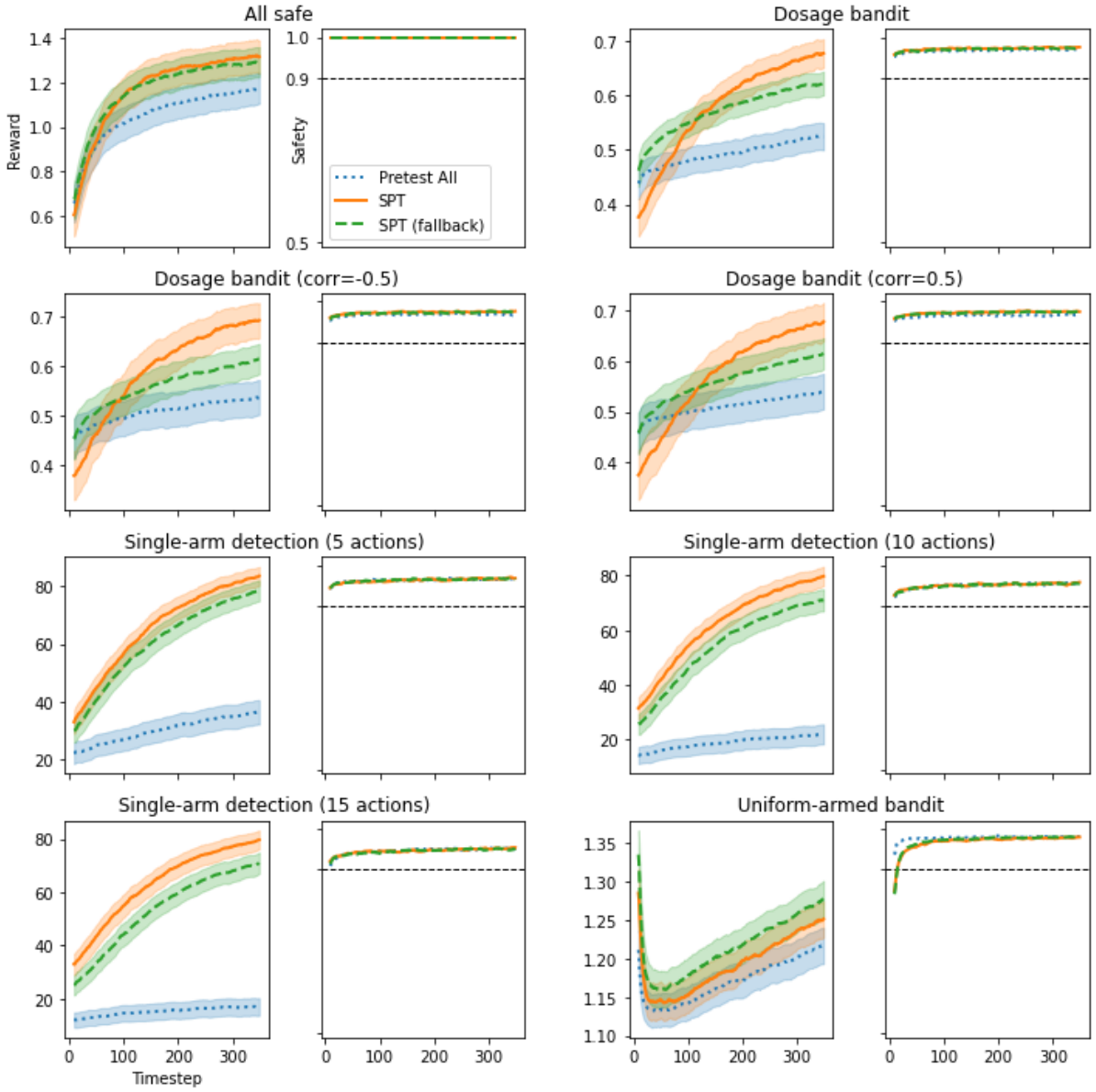

Figure 4.3: Algorithm performance over time for bandits without contexts. For each bandit setting, two plots are given: (left) the average (over runs) mean reward obtained per timestep, $E\{Q^r(X_t, \pi_t(X_t)\}$ for each t; (right) the proportion of the time that each algorithm selected a safe action, $E[\mathbb{1}[\pi_t(X_t) \in \mathcal{A}^{\tau}_{X_t}]$ for each t. These quantities are estimated with 2,000 independent runs of each algorithm. Highlighting indicates 95% pointwise confidence intervals. (The intervals are so narrow for safety estimates that we omit them for visual clarity.)

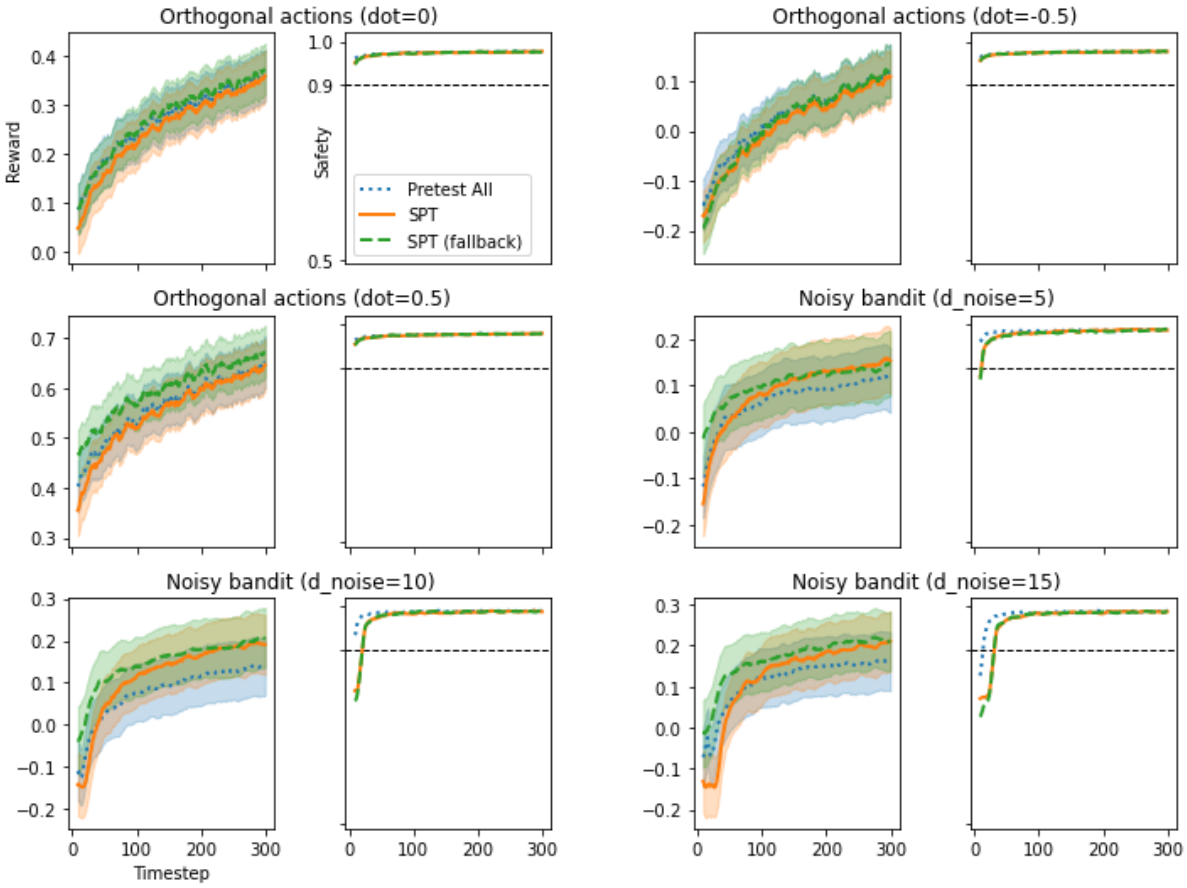

Figure 4.4: Algorithm performance over time for contextual bandits. For each bandit setting, two plots are given: (left) the average (over runs) mean reward obtained per timestep, $E\{Q^r(X_t, \pi_t(X_t)\}$ for each t; (right) the proportion of the time that each algorithm selected a safe action, $E[\mathbb{1}[\pi_t(X_t) \in \mathcal{A}^\tau_{X_t}]$ for each t. These quantities are estimated with 6,000 independent runs of each algorithm. Highlighting indicates indicate 95% pointwise confidence intervals. (The intervals are so narrow for safety estimates that we omit them for visual clarity.)

Table 4.2: The average performance (across all timesteps) of algorithms applied to different problem settings. Safety values above the $1-\alpha$ ($\alpha = 0.1$) threshold are highlighted. For each setting, the algorithm that achieved the highest reward among safe algorithms is marked in **bold**.

				Average reward			Average safety	
Setting	Burn-in	Algorithm Timesteps	Pretest All	SPT	SPT (fallback)	Pretest All	SPT	SPT (fallback)
All safe	40	350	1.052 (0.030)	**1.170** (0.031)	1.167 (0.028)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Dosage bandit	80	350	0.493 (0.011)	**0.580** (0.006)	0.572 (0.008)	0.968 (0.002)	0.972 (0.000)	0.971 (0.001)
Dosage bandit (corr=-0.5)	80	350	0.508 (0.015)	**0.592** (0.009)	0.559 (0.012)	0.967 (0.004)	0.972 (0.001)	0.971 (0.001)
Dosage bandit (corr=0.5)	80	350	0.510 (0.015)	**0.575** (0.009)	0.561 (0.012)	0.965 (0.004)	0.972 (0.001)	0.971 (0.001)
Single-arm detection (5 actions)	20	350	30.291 (1.785)	**65.913** (1.197)	60.777 (1.504)	0.963 (0.000)	0.962 (0.000)	0.963 (0.000)
Single-arm detection (10 actions)	40	350	18.932 (1.558)	**62.484** (1.293)	54.648 (1.581)	0.950 (0.001)	0.950 (0.001)	0.949 (0.001)
Single-arm detection (15 actions)	60	350	15.274 (1.448)	**63.300** (1.302)	53.585 (1.561)	0.946 (0.001)	0.946 (0.001)	0.945 (0.001)
Uniform-armed bandit	12	350	1.172 (0.008)	1.194 (0.006)	**1.218** (0.009)	0.977 (0.001)	0.967 (0.001)	0.969 (0.001)
Orthogonal actions (dot=0)	20	300	0.266 (0.005)	0.251 (0.005)	**0.274** (0.005)	0.974 (0.000)	0.973 (0.000)	0.973 (0.000)
Orthogonal actions (dot=-0.5)	20	300	**0.027** (0.003)	0.011 (0.003)	0.019 (0.003)	0.975 (0.000)	0.974 (0.000)	0.974 (0.000)
Orthogonal actions (dot=0.5)	20	300	0.563 (0.008)	0.552 (0.008)	**0.596** (0.008)	0.975 (0.000)	0.974 (0.000)	0.974 (0.000)
Noisy bandit (d=5)	12	300	0.069 (0.032)	0.088 (0.031)	**0.105** (0.032)	0.983 (0.002)	0.978 (0.001)	0.976 (0.003)
Noisy bandit (d=10)	12	300	0.079 (0.032)	0.115 (0.030)	**0.147** (0.032)	0.985 (0.001)	0.975 (0.002)	0.975 (0.001)
Noisy bandit (d=15)	12	300	0.113 (0.033)	0.124 (0.030)	**0.167** (0.032)	0.980 (0.001)	0.966 (0.002)	0.965 (0.002)

required for SPT to be (α, τ)-consistent can be weakened, and with appropriate choice of feature representation ϕ (e.g., using radial basis functions), linear models may be very expressive. That being said, the theory is meaningfully limited, as the assumption of linearity or of the bandit property may not be relevant for many reinforcement learning applications, for example, those in deep learning. However, we emphasize that SPT is a framework that can be applied on top of *any* method for statistical inference (hypothesis testing) on $Q^s(x, a)$, leaving the door open for future developments in inference in reinforcement learning to extend SPT's applicability.

In terms of future research directions, the problem of learning safely by reinforcement is multifaceted. Our approach highlighted and addressed the challenge of constrained optimization in a statistical settings, where datasets may be of small or moderate size and observations are noisy. In doing so, we avoided dealing with other aspects that make reinforcement learning (RL), and in particular *safe* reinforcement learning, hard:

- **Safe exploration** (as opposed to ϵ-random exploration): as is common in the reinforcement learning literature, SPT and Pretest All allow for a nonzero probability of selecting *any* action at each timestep in order to guarantee that the estimators for safety and reward will be consistent. SPT and Pretest All maintain nominal safety levels in finite samples by limiting the rate at which this random exploration occurs. However, in many settings where safe RL is required, this manner of action selection may be unacceptable. For example, a self-driving car system cannot rea-

Table 4.3: The final-timestep performance of algorithms applied to different problem settings. Safety values above the $1 - \alpha$ ($\alpha = 0.1$) threshold are highlighted. For each setting, the algorithm that achieved the highest reward among safe algorithms is marked in **bold**.

			Final reward			Final safety		
Setting	Burn-in	Algorithm Timesteps	Pretest All	SPT	SPT (fallback)	Pretest All	SPT	SPT (fallback)
All safe	40	350	1.157 (0.037)	**1.315** (0.038)	1.295 (0.034)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Dosage bandit	80	350	0.524 (0.013)	**0.677** (0.013)	0.626 (0.011)	0.972 (0.007)	0.975 (0.007)	0.975 (0.007)
Dosage bandit (corr=-0.5)	80	350	0.541 (0.018)	**0.697** (0.017)	0.612 (0.016)	0.962 (0.012)	0.982 (0.009)	0.973 (0.010)
Dosage bandit (corr=0.5)	80	350	0.539 (0.018)	**0.686** (0.019)	0.615 (0.016)	0.971 (0.011)	0.972 (0.011)	0.962 (0.012)
Single-arm detection (5 actions)	20	350	35.850 (2.145)	**83.550** (1.658)	77.850 (1.857)	0.972 (0.007)	0.968 (0.008)	0.960 (0.009)
Single-arm detection (10 actions)	40	350	21.900 (1.850)	**80.000** (1.789)	71.450 (2.020)	0.953 (0.009)	0.960 (0.009)	0.959 (0.009)
Single-arm detection (15 actions)	60	350	17.200 (1.688)	**79.700** (1.799)	71.850 (2.011)	0.960 (0.009)	0.950 (0.010)	0.959 (0.009)
Uniform-armed bandit	12	350	1.224 (0.012)	1.256 (0.012)	**1.279** (0.012)	0.972 (0.007)	0.980 (0.006)	0.977 (0.007)
Orthogonal actions (dot=0)	20	300	**0.377** (0.028)	0.362 (0.028)	0.374 (0.028)	0.980 (0.004)	0.977 (0.004)	0.975 (0.004)
Orthogonal actions (dot=-0.5)	20	300	**0.098** (0.027)	0.087 (0.027)	0.095 (0.028)	0.974 (0.004)	0.976 (0.004)	0.978 (0.004)
Orthogonal actions (dot=0.5)	20	300	0.667 (0.027)	0.673 (0.027)	**0.689** (0.027)	0.979 (0.004)	0.981 (0.004)	0.974 (0.004)
Noisy bandit (d=5)	12	300	0.104 (0.035)	**0.151** (0.036)	0.150 (0.036)	0.983 (0.006)	0.987 (0.005)	0.991 (0.004)
Noisy bandit (d=10)	12	300	0.142 (0.036)	0.192 (0.036)	**0.222** (0.036)	0.992 (0.004)	0.979 (0.006)	0.987 (0.005)
Noisy bandit (d=15)	12	300	0.158 (0.036)	**0.224** (0.037)	0.217 (0.037)	0.985 (0.005)	0.984 (0.006)	0.983 (0.006)

sonably be permitted to take entirely random actions any proportion of the time. On the other hand, there are settings where ϵ-random exploration make sense. In a clinical context, whatever action space a learning algorithm acts on typically comes pre-screened by clinicians to remove any interventions that are known to be unacceptably poor, so SPT could be applied in that setting.

- **Forward-looking exploration** (towards Bayes-optimality): the derivation of the SPT proposal given in Section 4.3.2 is explicitly based on the goal of maximizing the expected reward of the *next* arm pulled, not in terms of the expected reward of an arm pulled in the far future, nor in terms of future regret or another metric that better captures the desired limiting behavior of a safe learning algorithm. In this sense, SPT is a greedy algorithm whose good performance is merely due to the choice of heuristic for action selection (in this case, the form of the proposal objective). Of course, the same can be said of other successful and useful algorithms like Thompson Sampling and UCB. Although it is beyond the scope of this chapter, we expect that there are great gains to be made by applying ideas for more principled exploration to the problem of safe reinforcement learning (see Russo and Van Roy (2014) and Lu et al. (2021), for example). These could be incorporated into SPT via modifications to the proposal objective.

- **General problem settings** (beyond linearity, beyond bandits): SPT's structure

is fairly abstract and could plausibly be extended to nonlinear or non-stationary contextual bandit problems or even Markov Decision Processes (MDPs). A key challenge for such an extension would be obtaining a valid safety test. An example of a promising line of inquiry is Zhang et al. (2021), which provides inference for a more general class of contextual bandits, with weaker assumptions than ours. However, the form of their confidence regions does not allow for efficient testing of single hypothesis tests, which means it is unlikely to be competitive for SPT. For further discussion on theoretical limitations and Zhang et al. (2021), see Appendix C.1.

By focusing on the aspect of constrained optimization, SPT provides a general framework for reinforcement learning with safety constraints. The three areas above all provide broad directions for future work that could be used to generalize and improve SPT.

4.7 Acknowledgements

Adam Venis, Kyle Duke, and Joseph Lawson provided helpful comments on early drafts of this chapter.

REFERENCES

Adler, I. (2013). The equivalence of linear programs and zero-sum games. *International Journal of Game Theory*, 42(1):165–177.

Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32.

Amani, S., Alizadeh, M., and Thrampoulidis, C. (2020). Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3562–3566. IEEE.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Balakrishnan, A., Bouneffouf, D., Mattei, N., and Rossi, F. (2018). Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, pages 5802–5804.

Balata, A., Ludkovski, M., Maheshwari, A., and Palczewski, J. (2021). Statistical learning for probability-constrained stochastic optimal control. *European Journal of Operational Research*, 290(2):640–656.

Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W., Perolat, J., Jaderberg, M., and Graepel, T. (2019). Open-ended learning in symmetric zero-sum games. In *International Conference on Machine Learning*, pages 434–443. PMLR.

Berkenkamp, F., Krause, A., and Schoellig, A. P. (2021). Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pages 1–35.

Brown, G. W. (1951). Iterative solution of games by fictitious play. *Activity Analysis of Production and Allocation*, 13(1):374–376.

Candogan, O., Menache, I., Ozdaglar, A., and Parrilo, P. A. (2011). Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503.

Chen, H., Lu, W., and Song, R. (2021). Statistical inference for online decision making: In a contextual bandit setting. *Journal of the American Statistical Association*, 116(533):240–255.

Cloud, A. and Laber, E. B. (2021). Variance decompositions for extensive-form games. In *2021 IEEE Conference on Games (CoG)*, pages 1–8. IEEE.

Cloud, A., Wang, A., and Kerr, W. (2022). Anticipatory fictitious play.

Conlisk, J. (1993a). Adaptation in games: Two solutions to the crawford puzzle. *Journal of Economic Behavior & Organization*, 22(1):25–50.

Conlisk, J. (1993b). Adaptive tactics in games: Further solutions to the crawford puzzle. *Journal of Economic Behavior & Organization*, 22(1):51–68.

Croson, R., Fishman, P., and Pope, D. G. (2008). Poker superstars: Skill or luck? similarities between golf—thought to be a game of skill—and poker. *Chance*, 21(4):25–28.

Daskalakis, C. and Pan, Q. (2014). A counter-example to karlin's strong conjecture for fictitious play. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 11–20, Philadelphia, PA, USA. IEEE, IEEE.

DeDonno, M. A. and Detterman, D. K. (2008). Poker is a skill. *Gaming Law Review*, 12(1):31–36.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press.

Elias, G. S., Garfield, R., and Gutschera, K. R. (2012). *Characteristics of Games*. MIT Press.

Foerster, J., Chen, R. Y., Al-Shedivat, M., Whiteson, S., Abbeel, P., and Mordatch, I. (2018). Learning with opponent-learning awareness. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pages 122–130, '. AAMAS.

Getty, D., Li, H., Yano, M., Gao, C., and Hosoi, A. (2018). Luck and the law: Quantifying chance in fantasy sports and other contests. *SIAM Review*, 60(4):869–887.

Goldenshluger, A. and Zeevi, A. (2011). A note on performance limitations in bandit problems with side information. *IEEE Transactions on Information Theory*, 57(3):1707–1713.

Goldenshluger, A. and Zeevi, A. (2013). A linear response bandit problem. *Stochastic Systems*, 3(1):230–261.

Hall, P. and Heyde, C. C. (2014). *Martingale Limit Theory and its Application*. Academic press.

Heeb, R. D. (2012). Report of randall d. heeb, phd (united states of american against lawrence discristina). Case1:11-cr-00414 Document 77-1.

Heinrich, J., Lanctot, M., and Silver, D. (2015). Fictitious self-play in extensive-form games. In *International Conference on Machine Learning*, pages 805–813.

Heubeck, S. (2008a). Measuring skill in games: A critical review of methodologies. *Gaming Law Review and Economics*, 12(3):231–238.

Heubeck, S. (2008b). Measuring skill in games with random payoffs: Evaluating legality. *Review of Law & Economics*, 4(1):25–34.

Huang, X. and Xu, J. (2020). Estimating individualized treatment rules with risk constraint. *Biometrics*, 76(4):1310–1318.

Huber, P. J. (1992). Robust estimation of a location parameter. In *Breakthroughs in statistics*, pages 492–518. Springer.

Kazerouni, A., Ghavamzadeh, M., Abbasi Yadkori, Y., and Van Roy, B. (2017). Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30.

Kelly, J. M., Dhar, Z., and Verbiest, T. (2007). Poker and the law: is it a game of skill or chance and legally does it matter? *Gaming Law Review*, 11(3):190–202.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koller, D. and Megiddo, N. (1996). Finding mixed strategies with small supports in extensive form games. *International Journal of Game Theory*, 25(1):73–92.

Kuhn, H. (1953). Extensive games and the problem of information. In *Contributions to the Theory of Games*, pages 193–216. Princeton University Press.

Laber, E. B., Wu, F., Munera, C., Lipkovich, I., Colucci, S., and Ripa, S. (2018). Identifying optimal dosage regimes under safety constraints: An application to long term opioid treatment of chronic pain. *Statistics in Medicine*, 37(9):1407–1418.

Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. (2009). Monte carlo sampling for regret minimization in extensive games. In *Advances in neural information processing systems*, pages 1078–1086.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. (2017). A unified game-theoretic approach to multiagent reinforcement learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, -. Curran Associates, Inc.

Laroche, R., Trichelair, P., and Des Combes, R. T. (2019). Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning*, pages 3652–3661. PMLR.

Leqi, L. and Kennedy, E. H. (2021). Median optimal treatment regimes. *arXiv preprint arXiv:2103.01802*.

Letcher, A., Foerster, J., Balduzzi, D., Rocktäschel, T., and Whiteson, S. (2018). Stable opponent shaping in differentiable games. In *International Conference on Learning Representations*, page ', Vancouver Convention Center, Vancouver, BC, Canada. ICLR.

Levitt, S. D. and Miles, T. J. (2014). The role of skill versus luck in poker evidence from the world series of poker. *Journal of Sports Economics*, 15(1):31–44.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Liu, S., Marris, L., Hennes, D., Merel, J., Heess, N., and Graepel, T. (2022). Neupl: Neural population learning. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lockhart, E., Lanctot, M., Pérolat, J., Lespiau, J.-B., Morrill, D., Timbers, F., and Tuyls, K. (2019). Computing approximate equilibria in sequential adversarial games by exploitability descent. *arXiv preprint arXiv:1903.05614*, '(').

Lu, X., Van Roy, B., Dwaracherla, V., Ibrahimi, M., Osband, I., and Wen, Z. (2021). Reinforcement learning, bit by bit. *arXiv preprint arXiv:2103.04047*.

Luce, R. D. and Raiffa, H. (1989). *Games and Decisions: Introduction and Critical Survey*. Courier Corporation, '.

Lütjens, B., Everett, M., and How, J. P. (2019). Safe reinforcement learning with model uncertainty estimates. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8662–8668. IEEE.

Moradipari, A., Amani, S., Alizadeh, M., and Thrampoulidis, C. (2021). Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69:3755–3767.

Moravcik, M., Schmid, M., Burch, N., Lisỳ, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. (2017). Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513.

Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I., and Stoica, I. (2018). Ray: A distributed framework for emerging {AI} applications. In *13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18)*, pages 561–577.

Nash Jr, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49.

Omidshafiei, S., Tuyls, K., Czarnecki, W. M., Santos, F. C., Rowland, M., Connor, J., Hennes, D., Muller, P., Perolat, J., Vylder, B. D., Gruslys, A., and Munos, R. (2020). Navigating the landscape of multiplayer games. *Nature Communications*, 11(1):1–17.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR.

Rafferty, A. N., Ying, H., and Williams, J. J. (2018). Bandit assignment for educational experiments: Benefits to students versus statistical power. In *International Conference on Artificial Intelligence in Education*, pages 286–290. Springer.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, 54(2):296–301.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58.

Russo, D. and Van Roy, B. (2014). Learning to optimize via information-directed sampling. *Advances in Neural Information Processing Systems*, 27.

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., and Wen, Z. (2018). A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Sawant, N., Namballa, C. B., Sadagopan, N., and Nassif, H. (2018). Contextual multi-armed bandits for causal marketing. *arXiv preprint arXiv:1810.01859*.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms.

Schwartz, E. M., Bradlow, E. T., and Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 36(4):500–522.

Shamma, J. S. and Arslan, G. (2005). Dynamic fictitious play, dynamic gradient play, and distributed convergence to nash equilibria. *IEEE Transactions on Automatic Control*, 50(3):312–327.

Shapley, L. S. (1953). Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100.

Shoham, Y. and Leyton-Brown, K. (2008). *Multiagent Systems: Algorithmic, Game-theoretic, and Logical Foundations*. Cambridge University Press.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286.

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press, Cambridge, Massachusetts.

Thall, P. F. and Cook, J. D. (2004). Dose-finding based on efficacy–toxicity trade-offs. *Biometrics*, 60(3):684–693.

Thomas, P., Theocharous, G., and Ghavamzadeh, M. (2015). High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294.

Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge university press.

van Loon, R. J. P., van den Assem, M. J., and van Dolder, D. (2015). Beyond chance? the persistence of performance in online poker. *PLoS One*, 10(3).

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 30(2):199.

Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., Oh, J., Horgan, D., Kroiss, M., Danihelka, I., Huang, A., Sifre, L., Cai, T., Agapiou, J. P., Jaderberg, M., Vezhnevets, A. S., Leblond, R., Pohlen, T., Dalibard, V., Budden, D., Sulsky, Y., Molloy, J., Paine, T. L., Gulcehre, C., Wang, Z., Pfaff, T., Wu, Y., Ring, R., Yogatama, D., McKinney, K., Smith, O., Schaul, T., Lillicrap, T., Kavukcuoglu, K., Hassabis, D., Apps, C., and Silver, D. (2019). Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354.

Wang, C.-C., Kulkarni, S. R., and Poor, H. V. (2005). Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355.

Wu, Y., Shariff, R., Lattimore, T., and Szepesvári, C. (2016). Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR.

Zhang, C. and Lesser, V. (2010). Multi-agent learning with policy prediction. In *Twenty-fourth AAAI conference on artificial intelligence*, pages 927–937, '. AAAI.

Zhang, K., Janson, L., and Murphy, S. (2021). Statistical inference with m-estimators on adaptively collected data. *Advances in Neural Information Processing Systems*, 34.

APPENDICES

Appendix A

Variance decompositions for extensive-form games

A.1 Variance component formula derivation

Here we derive (2.2), the formula for $V[E(Y|\mathbf{A}^i)]$. The basic strategy is to write the game outcome Y as a sum of random variables representing different paths through the game tree, then to use basic properties of probability and algebra to manipulate the expression, concluding with an inductive argument.

Recall that $I_z = \prod_{i \in \mathcal{N} \cup \{c\}} I_z^i = \prod_{i \in \mathcal{N} \cup \{c\}} \prod_{j=1}^{m^i(z)} I_{z,j}^i$ is the indicator that all actions along terminal history z are selected, $Y = \sum_{z \in \mathcal{Z}} r(z)\, I_z$, and $u_{z,j}^i$ is the jth information state observed by player i in terminal history z. Write $I_{z,k:}^i = \prod_{j=k}^{m^i(z)} I_{z,j}^i$, the indicator that player i selects all actions in z at and after $u_{z,k}^i$. Let $d(u)$ be the depth of u in its trajectory; for example, if u is the first observation of a player in their trajectory, $d(u) = 1$. Define $W_u = \sum_{z \in \mathcal{Z}(u)} r(z)\, \eta^{-i}(z)\, I_{z,d(u):}^i$ and $\mathcal{U}_d^i = \{u \in \mathcal{U}^i : d(u) = d\}$. Then

$$V[E(Y|\mathbf{A}^i)] = V\left[\sum_{z \in \mathcal{Z}} r(z)\, \eta^{-i}(z)\, I_z^i\right] = V\left[\sum_{u \in \mathcal{U}_1^i} \sum_{z \in \mathcal{Z}(u)} r(z)\, \eta^{-i}(z)\, I_z^i\right] = V\left(\sum_{u \in \mathcal{U}_1^i} W_u\right). \tag{A.1}$$

Note that histories $z \in \mathcal{Z}$ that contain no information states for player i have $I_z^i \equiv 1$, so they are constant inside the conditional expectation, which is why the second and third expressions are equal.

By the perfect recall assumption, each information state u can be uniquely identified with the sequence of information states and actions required to reach u. Furthermore, the behavioral strategy assumption gives that $A(u)$ is independent of $A(u')$ if $u \neq u' \in \mathcal{U}^i$. Therefore if $u^i_{z,j} \neq u^i_{z',j}$ for some j, then $I^i_{z,h}$ is independent of $I^i_{z',h'}$ for all $h, h' \in \{j, \ldots, \min[m^i(z), m^i(z')]\}$. We conclude that W_u is independent of $W_{u'}$ if $u \neq u'$. This allows us to split up (A.1) as follows:

$$\begin{aligned} V[E(Y|\mathbf{A}^i)] &= V\bigg(\sum_{u \in \mathcal{U}^i_1} W_u\bigg) = \sum_{u \in \mathcal{U}^i_1} V(W_u) \\ &= \sum_{u \in \mathcal{U}^i_1} \bigg(V\{E[W_u|A(u)]\} + E\{V[W_u|A(u)]\}\bigg). \end{aligned} \tag{A.2}$$

The last equality holds by the law of total variance. To evaluate the components of (A.2), write $W_{ua} = \sum_{z \in \mathcal{Z}(ua)} r(z)\, \eta^{-i}(z)\, I^i_{z,[d(u)+1]:}$ for each $a \in \mathcal{A}(u)$ so we have that

$$\begin{aligned} W_u &= \sum_{z \in \mathcal{Z}(u)} r(z)\, \eta^{-i}(z)\, I^i_{z,d(u):} \\ &= \sum_{a \in \mathcal{A}(u)} \sum_{z \in \mathcal{Z}(ua)} r(z)\, \eta^{-i}(z)\, I^i_{z,d(u)} I^i_{z,[d(u)+1]:} \\ &= \sum_{a \in \mathcal{A}(u)} W_{ua} \mathbb{1}(A(u) = a). \end{aligned}$$

With this we obtain the following expressions for the components of (A.2):

$$\begin{aligned} V\{E[W_u|A(u)]\} &= \sum_{a \in \mathcal{A}(u)} E^2(W_{ua})\, \pi^i(a|u) - \bigg[\sum_{a \in \mathcal{A}(u)} E(W_{ua})\, \pi^i(a|u)\bigg]^2; \\ E\{V[W_u|A(u)]\} &= \sum_{a \in \mathcal{A}(u)} V[W_u|A(u) = a]\; P[A(u) = a]. \end{aligned}$$

Write $r(u,a) = E\{r(Z)\mathbb{1}[Z \in \mathcal{Z}(u,a)]\}$ and $r(u) = E\{r(Z)\mathbb{1}[Z \in \mathcal{Z}(u)]\}$. It can be shown that

$$E(W_{ua}) = r(u,a)\, [\eta^i(u)\pi(a|u)]^{-1}.$$

Substituting these terms back into the expression for the variance component, we get

that

$$V\{E[W_u|A(u)]\} = [\eta^i(u)]^{-2}\Bigg(\sum_{a\in\mathcal{A}(u)}[r(u,a)]^2\,\pi^i(a;u) - [r(u)]^2\Bigg);$$
$$E\{V[W_u|A(u)]\} = \sum_{a\in\mathcal{A}(u)} V\Bigg(\sum_{z\in\mathcal{Z}(ua)} r(z)\,\eta^{-i}(z)\,I^i_{z,[d(u)+1]:}\Bigg)\,\pi^i(a|u).$$

Now take $Y_2 = \sum_{z\in\mathcal{Z}(ua)} r(z)\,\eta^{-i}(z)\,I^i_{z,[d(u)+1]:}$ and repeat the steps shown in (A.2) inductively to obtain that:

$$V[E(Y|\mathbf{A}^i)] = \sum_{u\in\mathcal{U}^i}\Bigg(\sum_{a\in\mathcal{A}(u)}[r(u,a)]^2/\pi^i(a|u) - [r(u)]^2\Bigg)/\eta^i(u).$$

Because $r(u,a) = q(u,a)\,\eta(u)\,\pi^i(a|u)$ and $r(u) = v(u)\,\eta(u)$, this yields (2.2).

A.2 Consistency proofs

First, we prove consistency of (2.3). Let $\mu(u) = \eta(u)/\sum_{u\in\mathcal{U}^i}\eta(u)$ be a normalized reach probability. Then

$$\begin{aligned}V[E(Y|\mathbf{A}^i)] &= \sum_{u\in\mathcal{U}^i}\Bigg(\sum_{a\in\mathcal{A}(u)}[q(u,a)]^2\,\pi^i(a|u) - [v(u)]^2\Bigg)\,\eta^{-i}(u)\,\eta(u)\\ &= \Bigg(\sum_{u\in\mathcal{U}^i}\eta(u)\Bigg)\,E_{U\sim\mu}\Bigg[\Bigg(\sum_{a\in\mathcal{A}(U)}[q(U,a)]^2\,\pi^i(a|U) - [v(U)]^2\Bigg)\,\eta^{-i}(U)\Bigg].\end{aligned}$$

Note that

$$\sum_{u\in\mathcal{U}^i}\eta(u) = \sum_{u\in\mathcal{U}^i}\sum_{z\in\mathcal{Z}(u)}\eta(z) = \sum_{u\in\mathcal{U}^i}\sum_{z\in\mathcal{Z}}\eta(z)\mathbb{1}(u\in z) = \sum_{z\in\mathcal{Z}}\eta(z)\sum_{u\in\mathcal{U}^i}\mathbb{1}(u\in z) = E[d^i(Z)],$$

where $d^i(Z)$ is the length of the trajectory for player i in terminal history Z. So, by the law of large numbers, $\nu^{-1}\sum_{k=1}^{\nu} d^i(Z_k) \overset{\text{a.s.}}{\to} \sum_{u\in\mathcal{U}^i}\eta(u)$ as $\nu\to\infty$. Consider the Markov Chain $\{U_t\}_{t\in\mathbb{N}}$ defined by the information states for player i observed upon repeated independent playthroughs of the game and let $\phi(u) = \{\sum_{a\in\mathcal{A}(u)}[q(u,a)]^2\,\pi^i(a|u) - [v(u)]^2\}\,\eta^{-i}(u)$. Then $T^{-1}\sum_{t=1}^{T}\phi(U_t) \overset{\text{a.s.}}{\to} E_{U\sim\mu}[\phi(U)]$ as $T\to\infty$ by a Law of Large Numbers for Markov Chains since $\{U_t\}$ is irreducible and positive recurrent.

Converting both these results to the notation of the original statement of the estimator, we have $\nu^{-1}\sum_{k=1}^{\nu} l_k \overset{\text{a.s.}}{\to} \sum_{u\in\mathcal{U}^i}\eta(u)$ and $\left(\sum_{k=1}^{\nu} l_k\right)^{-1}\sum_{k=1}^{\nu}\sum_{l=1}^{l_k}\phi(U_{k,l}) \overset{\text{a.s.}}{\to} E_{U\sim\mu}[\phi(U)]$ as $\nu\to\infty$. Therefore their product converges to the estimand, as desired.

To prove consistency of the regression-based estimator (2.4), we employ a basic style of argument from empirical process theory. For background on notation and concepts, see Chapter 19 of Van der Vaart (2000). Let operator P denote expectation and operator $\mathbb{P}_n$ denote empirical expecation. Assume without loss of generality that $E(Y)=0$, so we can write $V[E(Y|\mathbf{A}^i)] = Pg$, where $g:\times_{u\in\mathcal{U}^i}\mathcal{A}(u)\to[-b,b]$ is given by $g(\mathbf{a}) = [E(Y|\mathbf{A}^i=\mathbf{a})]^2$, where the bound $b=\max_{z\in\mathcal{Z}}[r^i(z)]^2$ exists by finiteness of $\mathcal{Z}$. Let $\widehat{g}_\nu = \min(f^2_{\widehat{\theta}}, b)$, noting the implicit dependence of $\widehat{\theta}$ on the sample size ν. Then $\mathbb{P}_n\widehat{g}_\nu$ is our estimator and we have that

$$\begin{aligned}
|\mathbb{P}_n\widehat{g}_\nu - Pg| &= |\mathbb{P}_n\widehat{g}_\nu - P\widehat{g}_\nu + P\widehat{g}_\nu - Pg| \\
&\le |\mathbb{P}_n\widehat{g}_\nu - P\widehat{g}_\nu| + |P\widehat{g}_\nu - Pg| \\
&\le \sup_{g'\in\mathcal{G}} |\mathbb{P}_n g' - Pg'| + P|\widehat{g}_\nu - g|,
\end{aligned}$$

where $\mathcal{G}$ refers to the set of bounded functions $g':\times_{u\in\mathcal{U}^i}\mathcal{A}(u)\to[-b,b]$. It is not hard to show that the finite domain and uniform boundedness of $\mathcal{G}$ implies that its bracketing numbers are always finite. Thus, by a Glivenko-Cantelli theorem Van der Vaart (2000), the left-hand side converges to 0 in probability. The right-hand side converges to 0 in probability if $\sup_{\mathbf{a}}|\widehat{g}_\nu(\mathbf{a}) - g(\mathbf{a})| \overset{p}{\to} 0$, which by the continuous mapping theorem is equivalent to $f_{\widehat{\theta}}$ being consistent (in a uniform sense) for $E(Y|\mathbf{A}^i=\cdot\,)$.

A.3 Neural network hyperparameters

- Embedding sizes: player - 3, card suit - 5, card rank - 14
- Card representation size: 10
- Dense layer sizes: 50, 20
- Activation functions: ReLU (hidden), identity (output)
- Batch size: 16
- Learning rate: 0.001

- Early stopping patience: 3 epochs

A.4 SkillRPS decomposition details

In this section, we show how to derive the exact variance components for SkillRPS. Recall that in SkillRPS, the outcome is $Y = (1-W)[\mathbb{1}(S>0) - \mathbb{1}(S<0)] + W(2Z-1)$, where $S = N_1 - N_2 + c \cdot \text{RPS}(A_1, A_2)$. In this case, a player's selection of N_i is considered to indicate their skill level, and $\mathbf{A}^c = (W, Z)$ is the collection of all chance actions. Adapting the three-way decomposition equation (2.6) to SkillRPS yields

$$\begin{aligned} V(Y) &= V[E(Y|N_1, N_2)] && \text{(skill)} \\ &\quad + E\{V[E(Y|W, Z, N_1, N_2)|N_1, N_2]\} && \text{(chance)} \\ &\quad + E[V(Y|W, Z, N_1, N_2)]. && \text{(remaining)} \end{aligned}$$

Assuming that $N_1, N_2 \overset{\text{iid}}{\sim} \text{Uniform}(\{1, \ldots, n\})$ and are independent of $A_1, A_2 \overset{\text{iid}}{\sim} \text{Uniform}(\{\text{Rock, Paper, Scissors}\})$, we can derive closed form expressions for each variance component.

Using routine probability manipulations, one can derive the following term for the variance in Y explained by the "skill" of the players in the case that the coin flip didn't happen ($W = 0$), for all $n \in \mathbb{N}$ and $c \in \mathbb{N} \cup \{0\}$. Begin by finding $E(Y|N_1 = n_1, N_2 = n_2, W = 0)$ for arbitrary $n_1, n_2 \in \{1, \ldots, n\}$, which is easy since the only remaining source of variation is $\text{RPS}(A_1, A_2) \sim \text{Uniform}(\{-1, 0, 1\})$. Next, treat this term as a discrete random variable depending on N_1 and N_2 and compute its variance. This yields:

$$V[E(Y|N_1, N_2, W=0)] = \begin{cases} 1 - \frac{1}{n} & \text{if } c = 0 \\ 1 - \frac{1}{3n} + \frac{8c^2+2c-16cn}{9n^2} & \text{if } 0 < c < n \\ (1 - \frac{1}{n})/9 & \text{if } c \geq n. \end{cases}$$

Call this term $\psi(n, c)$. From here it can be shown that:

$$\begin{aligned} V[E(Y|N_1, N_2)] &= (1-\alpha)^2 \, \psi(n, c) \\ E\{V[E(Y|W, Z, N_1, N_2)|N_1, N_2]\} &= \alpha + \alpha(1-\alpha) \, \psi(n, c). \end{aligned}$$

Finally, it follows that

$$E[V(Y|W,Z,N_1,N_2)] = \begin{cases} 0 & \text{if } c = 0 \\ (1-\alpha)\left[1 - \frac{1}{n} + \frac{2c}{3n^2} - \psi(n,c)\right] & \text{if } 0 < c < n \\ (1-\alpha)\left(\frac{8}{9} - \frac{2}{9n}\right) & \text{if } c \geq n. \end{cases}$$

Appendix B

Anticipatory fictitious play

B.1 Why naive AFP doesn't work

The original idea for the AFP algorithm, which we refer to as *naive AFP*, was: at each timestep, play the best response to the opponent's best response to your history. Formally, this given by the updates (c.f. the AFP update (3.1)):

$$\begin{aligned} x'_{t+1} &\in \mathtt{BR}^1_A(\overline{y}_t); & y'_{t+1} &\in \mathtt{BR}^2_A(\overline{x}_t); \\ x_{t+1} &\in \mathtt{BR}^1_A(y'_{t+1}); & y_{t+1} &\in \mathtt{BR}^2_A(x'_{t+1}); \\ \overline{x}_{t+1} &= \frac{1}{t+1}\sum_{k=1}^{t+1} x_k; & \overline{y}_{t+1} &= \frac{1}{t+1}\sum_{k=1}^{t+1} y_k. \end{aligned}$$

In preliminary simulations, naive AFP performed well in cyclic games and seemed to be competitive with FP in other games.

However, upon inspection it becomes clear that Naive AFP is not guaranteed to converge to a Nash equilibrium. This is because Naive AFP can only play best responses to strategies returned by the best response operator, which are pure strategies, but Nash equilibria of some games have nonzero probability assigned to strategies that are not best responses to any pure strategies. Thus there are some games where naive AFP is incapable of assigning any probability to actions which must be assigned nonzero probability in a Nash equilibrium, so naive AFP cannot converge to a Nash equilibrium. For example, see the game given in Table B.1.

Table B.1: Payoff matrix for Rock Paper Scissors SafeRock. In this game, SafeRock is not a best response to any of the opponent's pure strategies, so it won't be played by naive AFP. However, SafeRock is included in the Nash equilibrium which is given approximately by the following probabilities $([0.33, 0.47, 0.2], [0, 0.32, 0.18, 0.15])$, with value $v^* = 0.133$. If SafeRock were removed the game would be symmetric and have value 0, so we know SafeRock must be included in *any* Nash support.

	Rock	Paper	Scissors
Rock	0	-1	1
Paper	1	0	-1
Scissors	-1	1	0
SafeRock	0	0	0.99

B.2 Proof of Proposition 1

Our analysis closely follows the original proof of FP's convergence given in Robinson (1951), which consists of four key lemmas. We introduce the notion of a perturbed fictitious play system, a slight generalization of Robinson's "vector system." We modify Robinson's second lemma accordingly, which causes inconsequential changes in the third and fourth lemma, leaving the final result the same. In this way, we extend Robinson's proof to a broader class of algorithms which include fictitious play and anticipatory fictitious play as special cases.

Definition 1. An *iterative play system* (U, V) for $A \in \mathbb{R}^{m \times n}$ is a pair of sequences of vectors $U = \{U(0), U(1), \dots\}$ and $V = \{V(0), V(1), \dots\}$ with each $U(t) \in \mathbb{R}^n$ and $V(t) \in \mathbb{R}^m$ satisfying the following properties:

- $\min U(0) = \max V(0)$, and
- for each t, $U(t+1) = U(t) + A_{i(t),*}$ and $V(t+1) = V(t) + A_{*,j(t)}$,

where $A_{i,*}$ and $A_{*,j}$ are the ith row and jth column of A, and $i(t)$ and $j(t)$ are the row and column "played" by players 1 and 2 at time t.

The interpretation of an iterative play system is as follows. Suppose we choose $U(0) = 0$ and $V(0) = 0$. Write e_h to indicate a vector with a 1 at index h and 0's elsewhere. Then $\overline{x}_t = t^{-1} \sum_{k=1}^{t} e_{i(k)}$ and $\overline{y}_t = t^{-1} \sum_{k=1}^{t} e_{j(k)}$ are the empirical strategies played by players 1 and 2, and $t^{-1}V(t) = A\overline{y}_t$ and $t^{-1}U(t) = A^\intercal \overline{x}_t$ are the payoffs for player 1 and

2 when faced with those empirical strategies. In this way, $V(t)$ and $U(t)$ can be seen as "accumulating empirical payoffs" for players 1 and 2.

Definition 2. A *perturbed fictitious play* system (PFP-system) is an iterative play system with the additional property that $i(t)$ and $j(t)$ are best responses to a perturbed set of empirical payoffs. Precisely, an iterative play system is a PFP-system such that for any values $E_V(t) \in \mathbb{R}^m$ and $E_U(t) \in \mathbb{R}^n$ with $\|E_V(t)\|_\infty < a := \max_{i,j} |a_{i,j}|$ and $\|E_U(t)\|_\infty < a$ for each t,

$$i(t+1) \in \arg\max \big[V(t) + E_V(t)\big] \text{ and } j(t+1) \in \arg\min \big[U(t) + E_U(t)].$$

A special case of a PFP-system is what Robinson calls a "vector system," which describes fictitious play. This is obtained by setting all entries of $E_V(t)$ and $E_U(t)$ to zero at all timesteps.

Lemma 3. If (U, V) is an iterative play system for A, then

$$\liminf_{t\to\infty} t^{-1}\big\{\max V(t) - \min U(t)\big\} \geq 0.$$

This lemma follows from the minimax nature of two-player zero-sum games and holds regardless of what rows and columns of A are used to update elements of U and V.

Definition 3. Given an iterative play system (U, V) for matrix A, we say row $A_{i,*}$ is *E-eligible* in the interval $[t, t']$ if for some $t_1 \in [t, t']$, i could have been played as part of AFP. Precisely, the condition is that there exists $t_1 \in [t, t']$ such that

$$i \in \arg\max \big[V(t_1) + E\big]$$

for some $E \in \mathbb{R}^m$ with $\|E\|_\infty < a$. Or, equivalently,

$$v_i(t_1) \geq \max V(t_1) - 2a.$$

Similarly, we say column j is E-eligible if

$$u_j(t_1) \leq \min U(t_1) + 2a.$$

Lemma 4. If (U, V) is an iterative play system for A and all rows and columns are E-eligible in $[s, s+t]$, then we have

$$\max U(s+t) - \min U(s+t) \leq 2a(t+1), \text{ and}$$
$$\max V(s+t) - \min V(s+t) \leq 2a(t+1).$$

Proof. Let $j \in \arg\max U(s+t)$. By the E-eligibility of j, there must exist $t' \in [s, s+t]$ such that

$$u_j(t') - \min U(t') \leq 2a.$$

So, because the jth entry can't change by more than a per timestep,

$$\begin{aligned} \max U(s+t) &= u_j(s+t) \\ &\leq u_j(t) + at \\ &\leq \min U(t') + 2a + at \\ &\leq \min U(t+s) + at + 2a + at, \end{aligned}$$

where the last inequality holds because for $t' \in [s, s+t]$, the minimum of $U(t')$ versus $U(s+t)$ can't change by more than at in t timesteps. A similar argument shows the result for V. □

Lemma 5. If (U, V) is an iterative play system for A and all rows and columns are E-eligible in $[s, s+t]$, then

$$\max V(s+t) - \min U(s+t) \leq 4a(t+1).$$

Proof. As shown in Robinson (1951), this follows immediately from Lemma 4. The only difference here is that we replace $2at$ with $2a(t+1)$. □

Lemma 6. For every matrix A, $\epsilon > 0$, there exists a t_0 such that for any anticipatory fictitious play system,

$$\max V(t) - \min U(t) \leq \epsilon t \text{ for all } t \geq t_0.$$

Proof. We follow the proof of Robinson (1951), replacing the notion of eligibility with E-

eligibility. The strategy is induction. If $A \in \mathbb{R}^{1\times 1}$, the result is trivial because $V(t) = U(t)$ for all t. Now assume that the property holds for an arbitrary submatrix A' obtained by deleting any number of columns or rows from A. We wish to show that the property also holds for A.

Using the inductive hypothesis, pick t^* such that

$$\max V'(t) - \min U'(t) < \tfrac{1}{2}\epsilon t \text{ for all } t \geq t^*$$

for any A' a submatrix of A and (U', V') an anticipatory fictitious play system for A'.

As in Robinson (1951), we wish to show that if in (U, V), some row or column is not E-eligible in $[s, s + t^*]$, then

$$\max V(s + t^*) - \min U(s + t^*) < \max V(s) - \min U(s) + \tfrac{1}{2}\epsilon t^* \tag{B.1}$$

Suppose without loss of generality that row $A_{m,*}$ is not E-eligible $[s, s + t^*]$. Then we construct a new anticipatory fictitious play system (U', V') for matrix A', which is A with row m deleted. Define

$$\begin{aligned} U'(t) &= U(s + t) + c\,\mathbb{1}_n, \\ V'(t) &= \mathrm{Proj}_m V(s + t), \end{aligned}$$

for $t = 0, 1, \ldots, t^*$, where $\mathbb{1}_n$ is a vector of 1's with n entries, $c = \max V(s) - \min U(s)$, and $\mathrm{Proj}_k : \mathbb{R}^m \to \mathbb{R}^{m-1}$ is the operator that removes entry k. We now check the conditions for an anticipatory fictitious play system:

- We have $\min U'(0) = \min\{U(s) + [\max V(s) - \min U(s)]\mathbb{1}_n\} = \max V(s) = \max V'(0)$, where the last equality holds because m is not E-eligible, so it could not be a maximizer of $V(s)$, so deleting it to form $V'(0)$ does not change the maximum.

- We have that for each t,

$$\begin{aligned} U'(t+1) &= U(s+t+1) + c\mathbb{1}_n = U(s+t) + A_{i(s+t),*} + c\mathbb{1}_n = U'(t) + A'_{i(s+t),*}, \\ V'(t+1) &= \mathrm{Proj}_m V(s+t+1) = \mathrm{Proj}_m[V(s+t) + A_{*,j(s+t)}] = V'(t) + A'_{*,j(s+t)}, \end{aligned}$$

  where $A_{i(s+t),*} = A'_{i(s+t),*}$ because the AFP-ineligibility of m implies $i(s + t) \neq m$.

- Finally, we must show that the rows and columns selected still qualify as "anticipatory" responses within the context of (U', V') and A', i.e. that

$$v'_{i(s+t)}(t) \geq \max V'(t) - 2a \text{ and } u'_{j(s+t)}(t) \leq \min U'(t) + 2a$$

for each $t = 0, 1, \ldots, t^*$. By the definition of V' and fact that $i(s+t) \neq m$, we have

$$\begin{aligned} v'_{i(s+t)}(t) &= v_{i(s+t)}(s+t) \\ &\geq \max V(s+t) - 2a \qquad\qquad ((U,V) \text{ is a PFP-system}) \\ &= \max V'(s) - 2a, \end{aligned}$$

and $U'(t)$ is just a shifted version of $U(s+t)$, so the fact that $u_{j(s+t)}(s+t) \leq \min U(s+t) + 2a$ implies the result.

These points verify that (U', V') satisfy the conditions for an anticipatory fictitious play system for A' on $t = 0, \ldots, t^*$. We can choose remaining values for both sequences for $t = t^* + 1, t^* + 2, \ldots$ to satisfy the anticipatory fictitious play conditions. So, using the inductive hypothesis, we have

$$\begin{aligned} \max V(s+t^*) - \min U(s+t^*) &= \max V'(t^*) - \min\left\{U'(t^*) - [\max V(s) - \min U(s)]\mathbb{1}_n\right\} \\ &= \max V'(t^*) - \min U'(t^*) + \max V(s) - \min U(s) \\ &< \tfrac{1}{2}\epsilon t^* + \max V(s) - \min U(s). \end{aligned}$$

With (B.1) established, we are ready to finish the proof: under the inductive hypothesis, we will be able to deal with time intervals by splitting them into two cases: if a row or column is not E-eligible, we apply (B.1); if all rows or columns are E-eligible, we apply Lemma 5.

Specifically, we show that for any AFP system (U, V) for A and $t \geq 8at^*/\epsilon$,

$$\max V(t) - \min U(t) < \epsilon t.$$

Let $t > t^*$ and express it as $t = (\theta + q)t^*$, where $q \in \mathbb{N}$ and $\theta \in [0, 1)$. We consider the collection of length-t^* intervals $[(\theta + r - 1)t^*, (\theta + r)t^*)]$ for $r = 1, \ldots, q$.

- **Case 1.** There is at least one interval where all rows and columns are E-eligible.

Let $[(\theta+s-1)t^*, (\theta+s)t^*)]$ be the latest such interval. By Lemma 5,

$$\max V[(\theta+s)t^*] - \min V[(\theta+s)t^*] \leq 4a(t^*+1) \leq 8at^*,$$

where the last inequality holds because $t^* \geq 1$. By choice of the interval, all subsequent intervals with $r = s+1, \ldots, q$ have no E-eligible rows or columns, so (B.1) gives that

$$\max V(t) - \min U(t) \leq \max V[(\theta+s)t^*] - \min V[(\theta+s)t^*] + \tfrac{1}{2}\epsilon t^*(q-s),$$

noting that the result holds trivially if $q = s$. Combining the previous two results and loosening the bound, we have

$$\max V(t) - \min U(t) \leq 8at^* + \tfrac{1}{2}\epsilon t^*(q-s) \leq (8a + \tfrac{1}{2}\epsilon q)t^*. \tag{B.2}$$

- **Case 2.** In each interval $[(\theta+r-1)t^*, (\theta+r)t^*)]$ for $r = 1, \ldots, q$, some row or column is not E-eligible. Applying (B.1) repeatedly,

$$\begin{aligned}
\max V(t) - \min U(t) &= \max V[(\theta+q)t^*] - \min U[(\theta+q)t^*] \\
&< \max V[(\theta+q-1)t^*] - \min U[(\theta+q-1)t^*] + \tfrac{1}{2}\epsilon t^* \\
&< \max V[(\theta+q-2)t^*] - \min U[(\theta+q-2)t^*] + \tfrac{1}{2}\epsilon t^* + \tfrac{1}{2}\epsilon t^* \\
&< \ldots \\
&< \max V(\theta t^*) - \min U(\theta t^*) + \tfrac{1}{2}q\epsilon t^* \\
&\leq 2a\theta t^* + \tfrac{1}{2}q\epsilon t^* \qquad \text{(B.3)} \\
&= (2a\theta + \tfrac{1}{2}q\epsilon)t^*, \qquad \text{(B.4)}
\end{aligned}$$

where the last inequality holds because $\max V(\theta t^*) \leq a\theta t^*$ and $\min U(\theta t^*) \geq a\theta t^*$.

So, comparing (B.2) and (B.4) and noting that $\theta \in [0, 1)$, in either case we have that

$$\max V(t) - \min U(t) \leq 8at^* + \tfrac{1}{2}\epsilon(qt^*) \leq 8at^* + \tfrac{1}{2}\epsilon t \leq \epsilon t$$

for all $t \geq 16at^*/\epsilon$. □

Finally, we are ready for the proof of Proposition 1, which is essentially identical to

the final proof of Theorem 1 in Robinson (1951).

Proof. Let $V(0) = 0 \in \mathbb{R}^m$, $U(0) = 0 \in \mathbb{R}^n$, and $V(t) = tA\overline{y}_t$, $U(t) = tA^\intercal \overline{x}_t$ for $t \in \mathbb{N}$, where $\overline{x}_t$ and $\overline{y}_t$ are as given in (3.1). Clearly, (U, V) forms an iterative play system. It follows from (3.1) that (U, V) is also a PFP-system with $E_V(t) = A \cdot \mathtt{BR}_A^2(\overline{x}_t)$ and $E_U(t) = A^\intercal \cdot \mathtt{BR}_A^1(\overline{y}_t)$. This is because

$$\begin{aligned} \overline{y}'_t &= \tfrac{t-1}{t}\overline{y}_{t-1} + \tfrac{1}{t}\mathtt{BR}_A^2(\overline{x}_{t-1}) \text{ implies} \\ tA\overline{y}'_t &= (t-1)A\overline{y}_{t-1} + A \cdot \mathtt{BR}_A^2(\overline{x}_{t-1}) \\ &= V(t-1) + E_V(t-1), \end{aligned}$$

so $x_t = \mathtt{BR}_A^1(\overline{y}'_{t-1}) = e_{i(t)}$, where $i(t) \in \arg\max[V(t-1) + E_V(t-1)]$. A similar argument holds for y_t.

So, by Lemmas 3 and 6,

$$\lim_{t\to\infty} \left(\max A\overline{y}_t - \min \overline{x}_t^\intercal A\right) = \lim_{t\to\infty} \frac{\max V(t) - \min U(t)}{t} = 0,$$

where the first equality follows from the definition of $V(t)$ and $U(t)$. Combining this with the fact that, for all t,

$$\begin{aligned} \max A\overline{y}_t &\geq \inf_{y\in\Delta^n} (\max Ay) = v^*, \text{ and} \\ \min \overline{x}_t^\intercal A &\leq \sup_{x\in\Delta^m} (\min x^\intercal A) = v^*, \end{aligned}$$

we have that

$$\lim_{t\to\infty} \max A\overline{y}_t = \lim_{t\to\infty} \min \overline{x}_t^\intercal A = v^*,$$

concluding the proof of convergence of AFP. □

B.2.1 Convergence rate of perturbed fictitious play

Given a 2p0s matrix game with payoff matrix $A \in \mathbb{R}^{m\times n}$, write $t^*(\epsilon; m, n)$ to denote the value of t^* given by Lemma 6 such that

$$\max V(t) - \min U(t) < \tfrac{1}{2}\epsilon t \text{ for } t \geq t^*. \tag{B.5}$$

We have that $t^*(\epsilon; 1, 1) = 1$. So, by the inductive step of the proof of Lemma 6, $t^*(\epsilon; 2, 1) = t^*(\epsilon; 1, 2) = \frac{8a}{\epsilon}$, which then implies that $t^*(\epsilon; 3, 1) = t^*(\epsilon; 2, 2) = t^*(\epsilon; 1, 3) = (\frac{16a}{\epsilon})^2$. Continuing inductively, we see that $t^*(\epsilon; m, n) = (\frac{16a}{\epsilon})^{m+n-2}$. Substituting into (B.5) and rearranging terms gives that

$$\frac{\max V(t) - \min U(t)}{t} < \epsilon \text{ for } t \geq (\tfrac{8a}{\epsilon})^{m+n-2}.$$

Choosing $\epsilon_t = 8at^{-1/(m+n-2)}$ for each t gives the result.

B.3 Proof of Proposition 2

For first two subsections, we restate the definitions of Δ_t, $t \in \mathbb{N}_0$ from (3.2): $\Delta_0 = [0, \ldots, 0]^\intercal \in \mathbb{Z}^n$, $\Delta_t = tC^n\,\overline{x}_t$ for each $t \in \mathbb{N}$, and i_t is the index played by FP (AFP) at time t, so

$$\Delta_{t+1,j} = \begin{cases} \Delta_{t,j} - 1 & \text{if } j = i_t - 1 \mod n; \\ \Delta_{t,j} + 1 & \text{if } j = i_t + 1 \mod n; \\ \Delta_{t,j} & \text{otherwise;} \end{cases} \tag{3.2}$$

for each $t \in \mathbb{N}_0$ and $j \in \{1, \ldots, n\}$.

B.3.1 FP: C^n

We must show that $\max \Delta_t = \Omega_p(\sqrt{t})$ under random tiebreaking. Throughout, whenever performing arithmetic with indices, that arithetic is done modulo n. As in the body of , define $t_m = \inf\{t \in \mathbb{N}_0 : \max \Delta_t = m\}$ and note the Markov inequality bound:

$$P(\max \Delta_t < m) = P(t_m > t) \leq E(t_m)/t = \frac{1}{t} \sum_{k=0}^{m-1} E(t_{k+1} - t_k).$$

The bulk of the argument is in finding a bound for $E(t_{k+1} - t_k)$.

It follows by the definition of Δ_t and the FP update that $i_t \in \arg\max \Delta_t$. Index i_1

may be chosen arbitrarily, but for $t > 1$, it follows from (3.2) that

$$i_{(t+1)} \in \begin{cases} \{i_t\} & \text{if } \Delta_{t,i_t} > \Delta_{t,i_t+1}; \\ \{i_t, i_t+1\} & \text{if } \Delta_{t,i_t} = \Delta_{t,i_t+1}; \\ \{i_t+1\} & \text{if } \Delta_{t,i_t} < \Delta_{t,i_t+1}; \end{cases}$$

because the value that is incremented at time t is the value adjacent to index i_t, Δ_{t,i_t+1}. Let $\tau_1 = 1$ and inductively define, for $\ell \in \mathbb{N}$, $\tau_{\ell+1} = \inf\{t > \tau_\ell : \Delta_{t,i_t} = \Delta_{t,i_t+1}\}$ to be the next time at which there are two possible choices for i_{t+1}. This is depicted in Table B.2, writing $m = \max \Delta_{\tau_\ell}$ and $m' \in \{m, m+1\}$.

Table B.2: The process of incrementing the index played under FP on C^n.

$\Delta_{\tau_\ell} =$	$[\ldots$	≤ 0	m	m	≤ 0	≤ 0	$\ldots]$
		$\downarrow -1$	$\downarrow -a$	$\downarrow +1$	$\downarrow +a$		
$\Delta_{\tau_{(\ell+1)}} =$	$[\ldots$	≤ 0	$m-a$	m'	m'	≤ 0	$\ldots]$

As shown in the table, all entries of Δ_t other than the two maximum values must be nonpositive at each $t = \tau_\ell$. This follows by induction, since it holds for $\tau = 1$ (Δ_{τ_1} has one positive entry) and if it's true for some $\ell \in \mathbb{N}$, then in order to progress to $\tau_{(\ell+1)}$, we must add some number $a > m$ (over the course of a timesteps) to the next entry, which means by (3.2) we will subtract a from the previous entry m, with $m - a \leq 0$. Finally, note that between τ_ℓ and $\tau_{(\ell+1)}$, either we will have incremented the max from m to $m+1$ if $i_{\tau_\ell} = i_{(\tau_\ell - 1)}$, or we will not have not (if $i_{\tau_\ell} = i_{(\tau_\ell-1)} + 1$), and the max will remain at m until we repeat some further number of increments of the index played. It is only on these timesteps that the maximum value can increment.

Based on this reasoning, we know that for any k, there must exist $\ell(k)$ such that $t_k = \tau_{\ell(k)} + 1$. Consider the random variable $\ell(k+1) - \ell(k)$, which is the number of increments of the index played that occurred between the increment of the max from k to $k+1$. Under uniform random tiebreaking, we have that $\ell(k+1) - \ell(k) \sim \text{Geometric}(1/2)$, since at each τ_ℓ there is a $1/2$ chance of incrementing ("success") or not incrementing ("failure"). So, $E[\ell(k+1) - \ell(k)] = 2$. Now suppose that we had the bound $\tau_{\ell+1} - \tau_\ell \leq d \max \Delta_{\tau_\ell}$

for some $d > 0$. That would imply that

$$t_{k+1} - t_k = \tau_{\ell(k+1)} - \tau_{\ell(k)} = \sum_{r=\ell(k)}^{\ell(k+1)-1} \tau_{r+1} - \tau_r \leq \sum_{r=\ell(k)}^{\ell(k+1)-1} d \max \Delta_{t_{k+1}}$$
$$= d[\ell(k+1) - \ell(k)](k+1).$$

Taking the expectation of both sides, we get $E(t_{k+1} - t_k) \leq 2d(k+1)$. Plugging this into the Markov bound is sufficient to finish the argument, as explained in the proof sketch for Proposition 2 (in the chapter).

All that remains is to show that $\tau_{\ell+1} - \tau_\ell \leq d \max \Delta_{\tau_\ell} = dm$. From the argument depicted in Table B.2, we know $\tau_{\ell+1} - \tau_\ell \leq a + 1$, and that $a \leq m + 1 - \min \Delta_{\tau_\ell}$. Because Δ_{τ_ℓ} has only two positive entries and $\sum_{i=1}^n \Delta_{t,i} = 0$, we have $\min \Delta_{\tau_\ell} \geq -2m$, so $\tau_{\ell+1} - \tau_\ell \leq 3m + 2 \leq 5m$, concluding the proof.

B.3.2 AFP: C^n, $n = 3, 4$

For $n = 3, 4$ we show $\max_t \Delta_t < 3$, which proves the result.

Based on the AFP update, it is impossible to get to $\max_t \Delta_t = m+1$ unless there are at least two non-adjacent m's in Δ_{t-1} with an $m-1$ in between. Otherwise, the two-step nature of the AFP update will not allow an m to be incremented to $m + 1$. However, it is impossible to have two non-adjacent m's with an $m - 1$ in between for $n = 3, m = 2$ because the entries of Δ_{t-1} sum to 0. Furthermore, in the $n = 4$ case, for each t, it must be that $\Delta_t = [a, b, -a, -b]$ for some a and b, by (3.2). So there also cannot be three positive numbers in this case.

B.3.3 FP: T^n

Assume without loss of generality that $x_1 = e_1$. Let $\tau_k = \min\{t : x_t = e_k\}$ and note that the form of T^n implies that the strategies played by FP must be nondecreasing and increment by at most 1 at a time. We argue by strong induction that $\tau_{k+1} - \tau_k > \tau_k - \tau_{k-1}$ for each $k < n$. Checking the first few terms, we have

$$\tau_1 = 1, \text{ so } T^n \overline{x}_1 = n^{-1}[0, n, 0, \ldots, 0]^\intercal, \text{ so}$$
$$\tau_2 = 2, \text{ so } T^n \overline{x}_2 = 2^{-1} n^{-1}[-n, n, n-1, \ldots, 0]^\intercal, \text{ so}$$

$x_3 = e_2$, and therefore $\tau_3 > 3$, so $\tau_3 - \tau_2 > \tau_2 - \tau_1 = 1$. Now assume that for some fixed $k < n$ and all $k' \in \{1, \ldots, k\}$ that $\tau_{k'+1} - \tau_{k'} > \tau_{k'} - \tau_{k'-1}$. Note that

$$\begin{aligned}
\overline{x}_{\tau_{(k+1)}-1} &= [\tau_2 - \tau_1, \ldots, \tau_{k+1} - \tau_k, 0, 0, \ldots, 0]^\intercal, \text{ so,} \\
T^n \overline{x}_{\tau_{(k+1)}-1} &\propto [<, <, \ldots, <, (\tau_{k+1} - \tau_k)(n-k+1), 0, \ldots, 0]^\intercal, \text{ and} \\
T^n \overline{x}_\ell &\propto [<, <, \ldots, <, (\tau_{k+1} - \tau_k)(n-k+1), (\ell - \tau_{k+1} + 1)(n-k), \ldots, 0]^\intercal,
\end{aligned}$$

for $\ell \in \{\tau_{k+1}, \ldots, \tau_{k+2} - 1\}$, where the '$<$' signs indicate values that are no greater than their neighbors on the right; this holds by the inductive hypothesis and definition of T^n. We know that for steps $\tau_{k+1}, \ldots \tau_{k+2} - 1$, FP will play e_{k+1}, and we know that τ_{k+2} is the first timestep at which $(\tau_{k+2} - \tau_{k+1})(n-k) \geq (\tau_{k+1} - \tau_k)(n-k+1)$ (or else FP would have played $k+1$ at τ_{k+2}, a contradiction). It follows that $\tau_{k+2} - \tau_{k+1} > \tau_{k+1} - \tau_k$, as desired.

This result implies that $\tau_{k+1} - \tau_k \geq k$ for each $k < n$, so we have $\tau_k \geq \sum_{j=1}^{k} j = k(k+1)/2 \geq k^2/2$ for each k. Inverting this, we get that $t \mapsto \sqrt{2t}$ is an upper bound on $k(t)$, the index played by FP at time t. Combining the expression for $T^n \overline{x}_\ell$, $\ell \in \{\tau_{k+1}, \ldots, \tau_{k+2} - 1\}$, with this bound, we get that $\max T^n \overline{x}_t = n^{-1} t^{-1} (\tau_{k(t)+1} - \tau_{k(t)})(n - k(t) + 1) \geq n^{-1} t^{-1} (n - k(t) + 1) \geq n^{-1} t^{-1} (n - \sqrt{2t}) = \Omega(1/\sqrt{t})$.

B.3.4 AFP: T^n

We argue first by strong induction that $x_t = e_{\min(t,n)}$ for all t. Assume without loss of generality that $x_1 = e_1$. Now assume that, for some fixed τ, $x_t = e_{\min(t,n)}$ for $t \leq \tau$.

If $\tau < n$, then under the inductive hypothesis,

$$\begin{aligned}
T^n \overline{x}_\tau &= \tau^{-1} T^n [\overbrace{1, 1, \ldots, 1}^{\tau}, 0, \ldots, 0]^\intercal \\
&= \tau^{-1} n^{-1} [-n, 1, \ldots, (n - \tau + 2), (n - \tau + 1), \ldots, 0]^\intercal,
\end{aligned}$$

for which the largest value is at index τ, so $x' = e_\tau$, so then $T^n \overline{x}'_\tau$ will have largest value $2n^{-1}(n - \tau + 1)$ at index $\tau + 1$, so $x_{\tau+1} = e_{\tau+1}$. If $\tau \geq n$, then under the inductive hypothesis, $T^n \overline{x}_\tau = \tau^{-1} T^n [1, 1, \ldots, 1, \tau - n + 1]^\intercal = \tau^{-1} n^{-1} [-n, 1, 1, \ldots, 1, 1 - 2(\tau - n), 2]^\intercal$, at which point $x'_\tau = x_{\tau+1} = e_n$, as desired.

Finally, we are interested in $\max T^n \overline{x}_t$ for $t < n$, which we obtain from the calculation above, $\frac{n-t+2}{nt} = O(1/t)$.

B.4 Additional figures

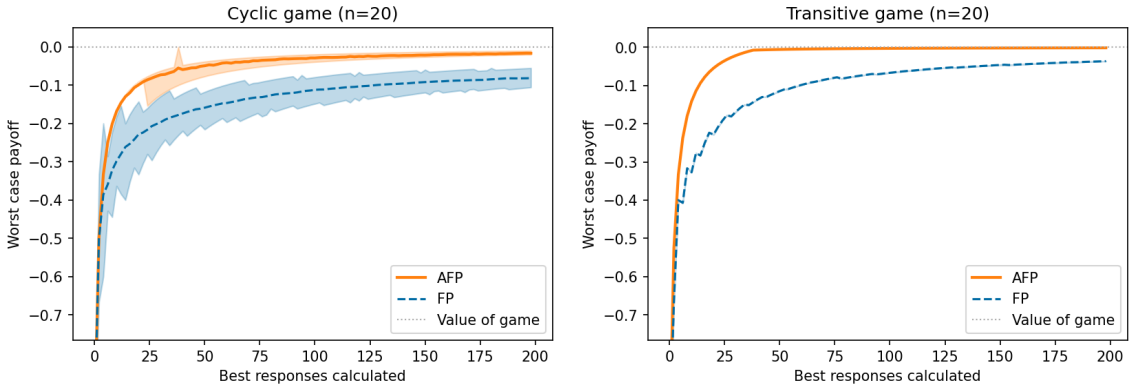

Figure B.1: Comparisons of FP and AFP on C^{20} and T^{20}. As before, highlighted regions indicate 10th and 90th percentiles across 10,000 runs.

B.5 RL experiment hyperparameters

APPO	TinyFighter
Discount factor γ	0.99
Value function coefficient	0.5
Entropy coefficient	0.01
Learning rate	3e-4
Gradient clipping	10
Batch size	5,120
Number of workers	80
Rollout fragment length	64
Other	Ray APPO defaults (as of November 2022)

B.6 Vanilla RL implementations of FP and AFP

Notes:

- $-i$ refers to the opponent of i.
- ReinforcementLearning(π, μ) plays π against μ for some number of episodes, gathers data, and updates π by reinforcement learning.
- Lines 8 and 9 are the approximate reinforcement learning analogue to computing a best response to the average of all previous strategies. For details, see Heinrich et al. (2015), which uses a more complicated setup in order to limit storage requirements (constantly learning β by supervised learning).

Algorithm 2 FP/AFP with reinforcement learning

1: Choose setting: `FP` or `AFP`.
2: Initialize policies π_1^1 and π_1^2.
3: Initialize policy stores $\Pi^1 = \{\pi_1^1\}, \Pi^2 = \{\pi_1^2\}$.
4: **for** $t = 1, 2, \ldots$ **do**
5: Initialize π_{t+1}^1, π_{t+1}^2.
6: **for** $i = 1, 2$ **do**
7: **while** per-timestep compute budget remains **do**
8: Sample $\beta \sim \text{Uniform}(\Pi^{-i})$.
9: $\pi_{t+1}^i \leftarrow \text{ReinforcementLearning}(\pi_{t+1}^i, \beta)$.
10: **end while**
11: $\Pi^i \leftarrow \Pi^i \cup \{\pi_{t+1}^i\}$.
12: **end for**
13: **if** setting is `AFP` **and** t is odd **then**
14: $\Pi^1 \leftarrow \Pi^1 \setminus \{\pi_t^1\}$ and $\Pi^2 \leftarrow \Pi^2 \setminus \{\pi_t^2\}$.
15: **end if**
16: **end for**

Appendix C

Safety-constrained online learning in contextual bandits

C.1 Extended discussion of limitations and Zhang et al.

The theory presented in Section 4.4 is limited in two key ways: it relies on margin conditions: (A6) to establish consistency of SPT, and (B1) to establish asymptotic normality. The first assumption (A6) is somewhat inescapable, as identification of actions on the safety margin is intrinsically difficult; however it is likely replaceable by a weaker condition that only acts on actions which are candidates for being optimal in the reward sense. For example, (A6) could be weakened to

$$P\left\{Q^s(X,a) - Q^s[X,\pi_0(X)] + \tau = 0 \,\middle|\, Q^r(X,a) > Q^r[X,\pi_0(X)]\right\} = 0 \text{ for all } a \in \mathcal{A}.$$

On the other hand, (B1) was used only to ensure existence of a unique limiting policy in order to prove asymptotic normality. We also assume that expected rewards are linear in a feature vector. These are assumptions which are common, but not fundamentally necessary, to inference in bandit settings (as we discuss in the next paragraph).

A salient example of such a development is Zhang et al. (2021), which provides more general results than the theory of Chen et al. (2021), on which many of our results are based (see Appendix C.3). Zhang et al. (2021) considers the contexual bandit setting but

broadens the scope to M-estimation (Huber 1992). Crucially, it replaces the ordinary-least squares (OLS) estimators with adaptively-weighted least squares (AW-LS) estimators which use square root importance sampling weights. Under suitable conditions, these AW-LS estimators are asymptotically normal without any requirement that the action selection distribution $\widehat{\pi}_T$ converges to a fixed limiting policy π_∞. This is in contrast to our proof strategy, which required that we first establish the limiting behavior of SPT before proving asymptotic normality. In fact, using the results of Zhang et al. (2021), it is possible to strip away much of the theory. A proof based on their results would, using AW-LS estimators, begin by showing asymptotic normality, then give Theorem 2.

However, there is one critical drawback to the results of Zhang et al. (2021) which limits their applicability to the SPT framework: the form of the confidence regions given therein mean that, at the time of testing the safety of actions, it is only possible to get joint inference over all actions, not one at a time. This means that the tests formed by inverting their confidence regions are very conservative and already apply to all actions. As a consequence, the results of Zhang et al. (2021) would be suitable for Pretest All but overly conservative if applied in SPT.

C.2 Proofs of main results

This section contains proofs for each of the results given in Section 4.4. The proofs rely on auxiliary results which are stated and proved in Appendix C.3.

C.2.1 Lemma 1

Proof. The conditions for Lemma 1 imply Lemma 7. Lemma 7 gives that, for any $\zeta > 0$, $T\epsilon_T^2 \to \infty$ implies that $P(\|\widehat{\beta}^{r,I_T} - \beta^r\|_1 \leq \zeta) \to 1$ as $T \to \infty$ (similarly for $\widehat{\beta}^{s,I_T}$). □

C.2.2 Lemma 2

Proof. Let $a \in \mathcal{A}$ be fixed. Plugging in parameters η_T, and X_{T+1} to the safety test definition (4.7) and rearranging terms, we have

$$\begin{aligned}\Psi^{\tau,\eta_T}_{X_{T+1}}(a; I_T) &= \mathbb{1}\left\{\sqrt{T}\frac{\{\phi(X_{T+1}, a) - \phi[X_{T+1}, \pi_0(X_{T+1})]\}^\intercal \widehat{\beta}^{s,I_T} + \tau}{\widehat{\sigma}^{s,I_T}(X_{T+1}, a)} > z_{1-\eta_T}\right\} \\ &= \mathbb{1}\left\{\{\phi(X_{T+1}, a) - \phi[X_{T+1}, \pi_0(X_{T+1})]\}^\intercal \widehat{\beta}^{s,I_T} + \tau > T^{-1/2}\,\widehat{\sigma}^{s,I_T}(X_{T+1}, a)\, z_{1-\eta_T}\right\} \\ &= \mathbb{1}(A^{X_{T+1}} > B_T^{X_{T+1}}),\end{aligned}$$

where for any $x \in \mathcal{X}$,

$$\begin{aligned}A^x &= \{\phi(x, a) - \phi[x, \pi_0(x)]\}^\intercal \beta^s + \tau, \text{ and} \\ B_T^x &= \Delta_T^x + T^{-1/2}\,\widehat{\sigma}^{s,I_T}(x, a)\, z_{1-\eta_T},\end{aligned}$$

with $\Delta_T^x = \{\phi(x, a) - \phi[x, \pi_0(x)]\}^\intercal(\beta^s - \widehat{\beta}^{s,I_T})$. Because $A^x = Q^s(x, a) - Q^s[x, \pi_0(a)] + \tau$, proving Lemma 2 is equivalent to showing that

$$\Psi^{\tau,\eta_T}_{X_{T+1}}(a; I_T) \xrightarrow{\text{p}} \mathbb{1}(A^{X_{T+1}} > 0) \text{ as } T \to \infty. \tag{C.1}$$

Let X be a random context identically distributed to X_1 and independent from all data H_∞. Note that is sufficient for our claim to show that $\mathbb{1}(A^X > B_T^X) \xrightarrow{\text{p}} \mathbb{1}(A^X > 0)$ as $T \to \infty$, because for each T, the random tuple $\{\mathbb{1}(A^X > B_T^X), \mathbb{1}(A^X > 0)\}$ has the same distribution as $\{\mathbb{1}(A^{X_{T+1}} > B_T^{X_{T+1}}), \mathbb{1}(A^{X_{T+1}} > 0)\}$, by the independence of X_{T+1} and H_{T-1} and because $\{X_t\}_{t\in\mathbb{N}} \overset{\text{iid}}{\sim} X$.

First, we argue that $B_T^X \xrightarrow{\text{p}} 0$ because each of its terms converges to 0. By (A2) and Lemma 1, we have

$$|\Delta_T^X| \leq d \cdot L_\phi \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty \xrightarrow{\text{p}} 0 \text{ as } T \to \infty.$$

Clearly, $z_{1-\eta_T} = O(1)$ by the condition on $\{\eta_t\}$, and we prove in section C.2.2 that $\widehat{\sigma}^{s,I_T}(X_{T+1}, a) = O_p(1)$. Second, note that $P(A^X = 0) = 0$ is assumed in (A6).

To finish the proof, we note the general fact that if A is a random variable with $P(A = 0) = 0$ and $\{B_T\}_{T\in\mathbb{N}}$ is a sequence of random variables such that $B_T \xrightarrow{\text{p}} 0$ as $T \to \infty$, then $\mathbb{1}(A > B_T) \xrightarrow{\text{p}} \mathbb{1}(A > 0)$ as $T \to \infty$. A proof of this fact is given in section

C.2.2.

□

Stochastic boundedness of standard error

Here we prove that $\widehat{\sigma}^{s,I_T}(X_{T+1}, a) = O_p(1)$ by proving the stronger claim that $\sup_{x\in\mathcal{X}}(\widehat{\sigma}^{s,I_T})^2(x,a) = O_p(1)$. Let $u(x,a) = \{\phi(x,a) - \phi[x,\pi_0(x)]\}/(\sqrt{d}\, L_\phi)$. Starting with the definition of $\widehat{\sigma}^{s,I_T}(x,a)$ and using the fact that $\|v\|_2 \le \sqrt{d}\|v\|_\infty$ for $v \in \mathbb{R}^d$, we have that

$$
\begin{aligned}
&\sup_{x\in\mathcal{X}} (\widehat{\sigma}^{s,I_T})^2(x,a) \\
&\quad= \sup_{x\in\mathcal{X}} \left(\{\phi(x,a) - \phi[x,\pi_0(x)]\}^\intercal (\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \widehat{\Sigma}^{s,I_T} (\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \{\phi(x,a) - \phi[x,\pi_0(x)]\} \right) \\
&\quad= \sup_{x\in\mathcal{X}} \left[dL_\phi^2\, u(x,a)^\intercal (\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \widehat{\Sigma}^{s,I_T} (\widehat{\boldsymbol{\phi}}^{I_T})^{-1} u(x,a) \right] \\
&\quad\le dL_\phi^2\, \lambda_{\max}\left[(\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \widehat{\Sigma}^{s,I_T} (\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \right] \\
&\quad\le dL_\phi^2\, \lambda_{\max}\left[(\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \right] \lambda_{\max}\left(\widehat{\Sigma}^{s,I_T} \right) \lambda_{\max}\left[(\widehat{\boldsymbol{\phi}}^{I_T})^{-1} \right] \\
&\quad= dL_\phi^2\, \lambda_{\min}\left(\widehat{\boldsymbol{\phi}}^{I_T} \right)^{-1} \lambda_{\max}\left(\widehat{\Sigma}^{s,I_T} \right) \lambda_{\min}\left(\widehat{\boldsymbol{\phi}}^{I_T} \right)^{-1}.
\end{aligned}
$$

Next, we show that $\lambda_{\max}(\widehat{\Sigma}^{s,I_T}) = O_p(1)$.

$$
\begin{aligned}
\lambda_{\max}(\widehat{\Sigma}^{s,I_T}) &= \max_{w\in\mathbb{R}^d\,:\,\|w\|_2=1} w^\intercal \widehat{\Sigma}^{s,I_T} w \qquad \text{(C.2)} \\
&= \max_{w\in\mathbb{R}^d\,:\,\|w\|_2=1} \frac{1}{|I_T|} \sum_{t\in I_T} w^\intercal \phi(X_t,A_t)\phi(X_t,A_t)^\intercal w [S_t - \phi(X_t,A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 \\
&\le \frac{1}{|I_T|} \sum_{t\in I_T} d\, L_\phi^2\, [S_t - \phi(X_t,A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 \\
&= d\, L_\phi^2 \frac{1}{|I_T|} \sum_{t\in I_T} [S_t - \phi(X_t,A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 \\
&= d\, L_\phi^2 \frac{T}{|I_T|} \cdot \frac{1}{T} \sum_{t=1}^{T} W_{T,t} [S_t - \phi(X_t,A_t)^\intercal \widehat{\beta}^{s,I_T}]^2, \qquad \text{(C.3)}
\end{aligned}
$$

where for all $t \leq T \in \mathbb{N}$, $W_{T,t} = \sum_{\iota \in I_T} \mathbb{1}(t = \iota)$ so

$$
\begin{aligned}
&W_{T,t} = Z_{T,t} && \text{if } I_T = I_T^{\text{prop}};\\
&W_{T,t} = 1 - Z_{T,t} && \text{if } I_T = I_T^{\text{test}};\\
&W_{T,t} = 1 && \text{if } I_T = \{1, \ldots, T\};\\
&(W_{T,t} : t \in I_T^{\text{prop}}) \,|\, I_T^{\text{prop}} \sim \text{Multinomial}(|I_T^{\text{prop}}|, |I_T^{\text{prop}}|^{-1}, \ldots, |I_T^{\text{prop}}|^{-1}) && \text{if } I_T = \widetilde{I}_T^{\text{prop}}.
\end{aligned}
$$

Note that in the first three cases, $W_{T,t}$ is bounded above by 1, and in the final case, we have

$$
E[W_{T,t}|H(I_T)] = \begin{cases} 1 & \text{if } t \in I_T^{\text{prop}}; \\ 0 & \text{otherwise.} \end{cases}
$$

With the goal of bounding (C.3), first observe that $T/|I_T| = O_p(1)$ by (A3). Consider the last term:

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T} W_{T,t}[S_t - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 \\
&\qquad = \frac{1}{T}\sum_{t=1}^{T} W_{T,t}[S_t - \phi(X_t, A_t)^\intercal \beta^s + \phi(X_t, A_t)^\intercal \beta^s - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 \\
&\qquad = \frac{1}{T}\sum_{t=1}^{T} W_{T,t}[S_t - \phi(X_t, A_t)^\intercal \beta^s]^2 && \text{(C.4)} \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} W_{T,t}[S_t - \phi(X_t, A_t)^\intercal \beta^s][\phi(X_t, A_t)^\intercal \beta^s - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s,I_T}] && \text{(C.5)} \\
&\qquad + \frac{1}{T}\sum_{t=1}^{T} W_{T,t}[\phi(X_t, A_t)^\intercal \beta^s - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s,I_T}]^2 && \text{(C.6)}
\end{aligned}
$$

We argue that each of these components is $O_p(1)$. First, we have

$$\begin{aligned}
E[\text{(C.4)}] &= \frac{1}{T}\sum_{t=1}^{T} E[W_{T,t}|H(I_T)]E\{[S_t - \phi(X_t, A_t)^\intercal \beta^s]^2\} && \text{(iterated expectation)} \\
&\leq \frac{1}{T}\sum_{t=1}^{T} E\left\{\exp\left[|S_t - \phi(X_t, A_t)^\intercal \beta^s|\right]\right\} && (E[W_{T,t}|H(I_T)] \leq 1,\ x^2 < e^{|x|}) \\
&\leq \exp(\sigma^2/2). && (A5)
\end{aligned}$$

By Markov's inequality, boundedness in expectation implies stochastic boundedness, so (C.4) $= O_p(1)$. Next,

$$\begin{aligned}
\text{(C.6)} &= \frac{1}{T}\sum_{t=1}^{T} W_{T,t}[\phi(X_t, A_t)^\intercal(\beta^s - \widehat{\beta}^{s,I_T})]^2 \\
&\leq \frac{1}{T}\sum_{t=1}^{T} W_{T,t} L_\phi^2 \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty^2 \leq L_\phi^2 \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty^2 \xrightarrow{\text{p}} 0
\end{aligned}$$

by Lemma 1 and because $\frac{1}{T}\sum_{t=1}^{T} W_{T,t} \leq 1$. Finally,

$$\begin{aligned}
|\text{(C.5)}| &\leq \frac{1}{T}\sum_{t=1}^{T} W_{T,t}|S_t - \phi(X_t, A_t)^\intercal \beta^s||\phi(X_t, A_t)^\intercal \beta^s - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s,I_T}| \\
&\leq \frac{1}{T}\sum_{t=1}^{T} W_{T,t}|S_t - \phi(X_t, A_t)^\intercal \beta^s| L_\phi \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty \\
&= L_\phi \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty \cdot \frac{1}{T}\sum_{t=1}^{T} W_{T,t}|S_t - \phi(X_t, A_t)^\intercal \beta^s| \\
&\leq L_\phi \|\beta^s - \widehat{\beta}^{s,I_T}\|_\infty \cdot \frac{1}{T}\sum_{t=1}^{T} W_{T,t} \exp\{|S_t - \phi(X_t, A_t)^\intercal \beta^s|\} = o_p(1)\, O_p(1) \xrightarrow{\text{p}} 0,
\end{aligned}$$

combining the arguments from the previous two terms. Taking these results together, we have shown that (C.3), an upper bound for $\lambda_{\max}(\widehat{\Sigma}^{s,I_T})$, is $O_p(1)$.

Finally, we argue that $P[\lambda_{\min}(\widehat{\boldsymbol{\phi}}^{I_T}) > c] \to 1$ for some $c > 0$, which implies that $\lambda_{\min}(\widehat{\boldsymbol{\phi}}^{I_T})^{-1} = O_p(1)$. Choose $c = p_I \lambda_{\min}(\Sigma)/2$. Using the definitions of events $\mathcal{E}$ and $\mathcal{R}$

from the proof of Lemma 1, and the inequalities (C.10) and (C.11), we have

$$P\left[\lambda_{\min}(\widehat{\boldsymbol{\phi}}^{I_T}) > \frac{p_I}{2}\lambda_{\min}(\Sigma)\right] = P(\mathcal{E}) \geq P(\mathcal{E}|\mathcal{R})\,P(\mathcal{R})$$
$$\geq 1 - \exp\left\{-\frac{T\epsilon_T}{8}\right\} - d\exp\left\{-\frac{T\epsilon_T p_I \lambda^2_{\min}(\Sigma)}{16L^2_\phi}\right\} \quad \to 0$$

as $T \to \infty$ under the conditions for Lemma 1.

In conclusion,

$$(\widehat{\sigma}^{s,I_T})^2(x,a) \leq dL^2_\phi\,\lambda_{\min}\left(\widehat{\boldsymbol{\phi}}^{I_T}\right)^{-1}\lambda_{\max}\left(\widehat{\Sigma}^{s,I_T}\right)\lambda_{\min}\left(\widehat{\boldsymbol{\phi}}^{I_T}\right)^{-1}$$
$$= dL^2_\phi\,O_p(1)O_p(1)O_p(1) = O_p(1),$$

so $\widehat{\sigma}^{s,I_T}(x,a) = O_p(1)$.

Proof of general fact

Here we prove that if $P(A=0)=0$ and $B_T \xrightarrow{\text{p}} 0$ as $T\to\infty$, then $\mathbb{1}(A > B_T) \xrightarrow{\text{p}} \mathbb{1}(A>0)$ as $T\to\infty$. Let $\gamma > 0$ and note that $\{\mathbb{1}(A>B_T) = \mathbb{1}(A>0)\} = \{|A| > |B_T|\}$. Because $P(A=0)=0$, continuity of probability measures implies that there exists $\delta > 0$ such that $P(|A| \leq \delta) < \gamma$. On the other hand, $B_T \xrightarrow{\text{p}} 0$ implies $P(|B_T| < \delta) \to 1$. Thus, we have

$$\begin{aligned} P(|A| > |B_T|) &\geq P(|A| > \delta > |B_T|) \\ &\geq P(|A| > \delta) + P(|B_T| < \delta) - 1 \\ &\geq 1 - \gamma + P(|B_T| < \delta) - 1 \\ &= P(|B_T| < \delta) - \gamma \\ &\xrightarrow{\text{p}} 1 - \gamma. \end{aligned}$$

Since $\gamma > 0$ was chosen arbitrarily, we have $P(|A| > |B_T|) \xrightarrow{\text{p}} 1$, as desired.

C.2.3 Theorem 2

Proof. The conditions of the theorem imply Corollary 1 and Lemma 2, so we have consistency of the OLS estimators and safety test. Consistency of the OLS estimator gives

that

$$\begin{aligned}&\left|\widehat{Q}_T^r(X_T,a)-\widehat{Q}_T^r[X_T,\pi_0(X_T)]-\{Q^r(X_T,a)-Q^r[X_T,\pi_0(X_T)]\}\right|\\&=\left|\{\phi(X_T,a)-\phi[X_T,\pi_0(X_T)]\}^\intercal(\widehat{\beta}_T^{R,\text{prop}}-\beta^R)\right| \overset{(A2)}{\leq} L_\phi\|\widehat{\beta}_T^{R,\text{prop}}-\beta^R\|_2 \xrightarrow{\text{p}} 0\end{aligned}$$

as $T\to\infty$ for any $a\in\mathcal{A}$. By Lemma 2, $\Psi_{X_T}^{\tau,\eta_T}(a;\widetilde{I}_T^{\text{prop}}) \xrightarrow{\text{p}} \mathbb{1}(a\in A_{X_T}^\tau)$. Applying the conditional expectation operator $E[\,\cdot\,|\,X_T,H(I_T^{\text{prop}})]$, the Dominated Convergence Theorem gives that

$$P\left[\Psi_{X_T}^{\tau,\eta_T}(a;\widetilde{I}_T^{\text{prop}})=1\,\middle|\,X_T,H(I_T^{\text{prop}})\right]\xrightarrow{\text{p}}\mathbb{1}(a\in A_{X_T}^\tau)$$

for any $a\in\mathcal{A}$. Combining these via Slutsky's theorem implies that for each a,

$$\widehat{J}_T(X_T,a)-J(X_T,a)\xrightarrow{\text{p}}0 \text{ as } T\to\infty,$$

where $\widehat{J}_T(X_T,a)$ and $J(X_T,a)$ are as defined in (4.6) and (4.8). Since A_T^{prop} is a maximizer of $\widehat{J}_T(X_T,\cdot)$ over $\mathcal{A}$ and $\mathcal{A}$ is finite, this implies that

$$P[A_T^{\text{prop}}\in\operatorname*{Arg\,max}_{a\in\mathcal{A}}J(X_T,a)]\to 1 \tag{C.7}$$

as $T\to\infty$. We furthermore claim that

$$P[A_T\in\operatorname*{Arg\,max}_{a\in\mathcal{A}}J(X_T,a)]\to 1. \tag{C.8}$$

To prove this, define the following events: $E_T^1=\{U_T>\epsilon_T\}$, $E_T^2=\{A_T^{\text{prop}}\in \operatorname{Arg\,max}J(X_T,a)\}$, and $E_T^3=\{\Psi_{X_T}^{\tau,\eta_T}(A_T^{\text{prop}};I_T^{\text{test}})=\mathbb{1}(A_T^{\text{prop}}\in A_{X_T}^\tau)\}$. We have that

$$P[A_T\in\operatorname*{Arg\,max}_{a\in\mathcal{A}}J(X_T,a)]\geq P\left[A_T\in\operatorname*{Arg\,max}_{a\in\mathcal{A}}J(X_T,a)\,\middle|\,E_T^1,E_T^2,E_T^3\right]P(E_T^1,E_T^2,E_T^3).$$

We know $P(E_T^1)=1-\epsilon_T\to 1$ by choice of $\{\epsilon_t\}$; we showed $P(E_T^2)\to 1$ in (C.7); and $P(E_T^3)\to 1$ by consistency of the safety test. Thus $P(E_T^1,E_T^2,E_T^3)\to 1$ as $T\to\infty$. To

finish the proof, we claim that

$$P\left[A_T \in \operatorname*{Arg\,max}_{a\in\mathcal{A}} J(X_T, a) \,\middle|\, E_T^1, E_T^2, E_T^3\right] = 1.$$

To see why, consider the following cases, given the occurrence of events E_T^1, E_T^2, and E_T^3. By the definition of J:

1. If $J(X_T, A_T^{\text{prop}}) = 0$, then $A_T^{\text{prop}} = \pi_0(X_T)$ or $A_T^{\text{prop}} \notin A_{X_T}^{\tau}$. Given E_T^3 (correctness of the safety test), in the former case the proposal will pass the safety test, resulting in $A_T^{\text{prop}} = \pi_0(X_T)$ since E_T^1 means an exploratory random action was not taken. In the latter case, the proposal will fail the test and we will have $A_T^{\text{prop}} = \pi_0(X_T)$. In either case, we have $J(X_T, A_T) = J[X_T, \pi_0(X_T)] = J(X_T, A_T^{\text{prop}}) = 0$. Given E_T^2, this means that A_T is a maximizer of $J(X_T, \cdot)$.

2. If $J(X_T, A_T^{\text{prop}}) > 0$, then $A_T^{\text{prop}} \in A_{X_T}^{\tau}$, so by similar reasoning, we will have $A_T = A_T^{\text{prop}}$, a maximizer of $J(X_T, \cdot)$.

This establishes (C.8), that the action selected by Split-Propose-Test maximize $J(X_T, \cdot)$ in the limit. It follows immediately that (4.2) holds, since an action that maximizes $J(X_T, \cdot)$ attains the maximally-attainable reward subject to the safety constraint. Furthermore, the consistency of the safety test plus the definition of A_T ensures (4.1).

As a corollary, if $\operatorname{Arg\,max} J(X, a)$ set is finite with probability 1 (assumption (B1)), then (C.8) implies that the actions selected under SPT will converge to the policy π_∞ defined by $\pi_\infty(x) = \arg\max J(x, a)$ w.p. 1. □

C.2.4 Theorem 3

Theorem 3, about the asymptotic normality of $(\widehat{\beta}_T^{r,\text{prop}}, \widehat{\beta}_T^{s,\text{prop}}, \widehat{\beta}_T^{s,\text{test}})$, follows as a special case of Theorem 4, which concerns the estimation of multiple outcomes on multiple splits of data. The correspondence is given in the following table.

Table C.1: A comparison of Theorems 4 and 3.

Description	Theorem 4 (general)	Theorem 3 (special case)
Number of estimators	k	3
Outcomes	$f^1(Y_t), f^2(Y_t), f^3(Y_t)$	R_t, S_t, S_t
Split probabilities	p_1, p_2, p_3	$p_{\text{prop}}, p_{\text{prop}}, 1-p_{\text{prop}}$
Sample overlap probabilities	$p_{1\cap 2}$	p_{prop}
	$p_{1\cap 3}, p_{2\cap 3}$	0, 0
Assumptions	(C1), (C2), (C3), (C4), (C5)	(A2), (A5), (4.9), (B2), (A3)

C.2.5 Corollary 1

Proof. We have the following, and wish to prove the results written below. For any $x \in \mathcal{X}$, $a \in \mathcal{A}$,

$$Z_T(x,a) = \underbrace{\sqrt{T}\frac{\{\phi(x,a)-\phi[x,\pi_0(x)]\}^\intercal(\widehat{\beta}^{s,I_T^{\text{test}}}-\beta^s)}{p_{\text{test}}^{-1/2}\,\sigma^s(x,a)}}_{\rightsquigarrow N(0,1) \text{ as } T\to\infty} \cdot \underbrace{\frac{p_{\text{test}}^{-1/2}\,\sigma^s(x,a)}{\widehat{\sigma}^{s,I_T^{\text{test}}}(x,a)}}_{\xrightarrow{\text{p}} 1 \text{ as } T\to\infty}.$$

The first result holds directly by Theorem 3, the continuous mapping theorem, and the property of multivariate normal random variables that $x \sim N(\mu,\Sigma)$ implies $a^\intercal x \sim N(a^\intercal\mu, a^\intercal\Sigma a)$. It remains to prove that $\widehat{\sigma}^{s,I_T^{\text{test}}}(x,a) \xrightarrow{\text{p}} p_{\text{test}}^{-1/2}\,\sigma^s(x,a)$ as $T\to\infty$.

Recall that

$$(\widehat{\sigma}^{s,I_T})^2(x,a) = \{\phi(x,a)-\phi[x,\pi_0(x)]\}^\intercal(\widehat{\boldsymbol{\phi}}^{I_T})^{-1}\widehat{\Sigma}^{s,I_T}(\widehat{\boldsymbol{\phi}}^{I_T})^{-1}\{\phi(x,a)-\phi[x,\pi_0(x)]\}, \text{ and}$$
$$(\sigma^s)^2(x,a) = \{\phi(x,a)-\phi[x,\pi_0(x)]\}^\intercal(\boldsymbol{\phi}_{\pi_\infty})^{-1}\,\Sigma^s_{\pi_\infty}\,(\boldsymbol{\phi}_{\pi_\infty})^{-1}\{\phi(x,a)-\phi[x,\pi_0(x)]\}$$

Using the arguments in Appendices C.3.2 and C.3.2, we have that

$$(\widehat{\boldsymbol{\phi}}^{I_T^{\text{test}}})^{-1} = \left(\frac{1}{|I_T^{\text{test}}|}\sum_{t\in I_T^{\text{test}}}\phi_t\phi_t^\intercal\right)^{-1} = \frac{|I_T^{\text{test}}|}{T}\cdot\left(\frac{1}{T}\sum_{t=1}^{T}(1-Z_{T,t})\phi_t\phi_t^\intercal\right)^{-1}$$
$$\xrightarrow{\text{p}} p_{\text{test}}\cdot(p_{\text{test}}\boldsymbol{\phi}_{\pi_\infty})^{-1} = (\boldsymbol{\phi}_{\pi_\infty})^{-1} \text{ as } T\to\infty.$$

We reuse another martingale argument to prove that $\widehat{\Sigma}^{s,I_T} \xrightarrow{\text{p}} \Sigma^s_{\pi_\infty}$ as $T\to\infty$. Let $v\in\mathbb{R}^d$. Using the definition of $U_{T,t} = T^{-1/2}v^\intercal\phi_t e_t^{\boldsymbol{\gamma}}$ from Appendix C.3.2 with $\boldsymbol{\gamma} = [0,0,1]^\intercal$ per

Table C.1, we have that

$$
\begin{aligned}
&v^\intercal \widehat{\Sigma}^{s, I_T^{\text{test}}} v \\
&\quad = \frac{1}{|I_T^{\text{test}}|} \sum_{t \in I_T^{\text{test}}} \phi(X_t, A_t)\phi(X_t, A_t)^\intercal [S_t - \phi(X_t, A_t)^\intercal \widehat{\beta}^{s, I_T^{\text{test}}}]^2 \\
&\quad = \frac{T}{|I_T^{\text{test}}|} \cdot \frac{1}{T} \sum_{t=t}^{T} (1 - Z_{T,t}) \phi(X_t, A_t)\phi(X_t, A_t)^\intercal [S_t - \phi(X_t, A_t)^\intercal \beta^s]^2 + o_p(1) + o_p(1) \\
&\quad = \frac{T}{|I_T^{\text{test}}|} \cdot p_{\text{test}} \sum_{t=1}^{T} U_{T,t}^2 + o_p(1),
\end{aligned}
$$

where the second equality is a consequence of the analysis of (C.3) in Appendix C.2.2.

Consider each summand in the second line. We have $1 - Z_{T,t} \in \{0, 1\}$, $v^\intercal \phi(X_t, A_t)\phi(X_t, A_t) v \leq d L_\phi^2$, and finally, by the uniform gaussian errors assumption (A5), $[S_t - \phi(X_t, A_t)^\intercal \beta^s]^2$ is stochastically dominated by $\overline{Z}^2$, where $\overline{Z} \sim N(0, \sigma^2)$. So, all together, each summand is stochastically dominated by $d L_\phi^2 \overline{Z}^2$ satisfying $E(d L_\phi^2 \overline{Z}^2) = d L_\phi^2 \sigma^2 < \infty$. Thus Theorem 2.19 of Hall and Heyde (2014) applies. Recall that $\mathcal{F}_T = \sigma(H(I_T))$ is the sigma algebra generated by all events up to time T. Noting the factor T^{-1} contained within $U_{T,t}^2$, the theorem gives that

$$
\sum_{t=1}^{T} \left[U_{T,t}^2 - E(U_{T,t}^2 | \mathcal{F}_{t-1}) \right] \xrightarrow{\text{p}} 0 \text{ as } T \to \infty.
$$

We have from the argument in Appendix C.3.2 that $\sum_{t=1}^{T} E(U_{T,t}^2 | \mathcal{F}_{t-1}) = V_T^2 \xrightarrow{\text{p}} v^\intercal \boldsymbol{G}^{\boldsymbol{\gamma}} v$. In our case,

$$
\begin{aligned}
\boldsymbol{G}^{\boldsymbol{\gamma}} &= \sum_{j=1}^{k} \sum_{\ell=1}^{k} \gamma_j \gamma_\ell \frac{p_{j \cap \ell}}{p_j p_\ell} E_{\pi_\infty} \left[\phi(X, A)\phi(X, A)^\intercal E(e^j e^\ell | X, A) \right] \\
&= \frac{p_{\text{test}}}{p_{\text{test}}\, p_{\text{test}}} E_{\pi_\infty} \left\{ \phi(X, A)\phi(X, A)^\intercal E\left[(e^s)^2 | X, A \right] \right\} \\
&= p_{\text{test}}^{-1} \Sigma_{\pi_\infty}^s.
\end{aligned}
$$

Noting that $T / |I_T^{\text{test}}| \xrightarrow{\text{p}} p_{\text{test}}^{-1}$ and combining the above results by the continuous mapping theorem, we have that $\widehat{\Sigma}^{s, I_T^{\text{test}}} \xrightarrow{\text{p}} p_{\text{test}}^{-1} \Sigma_{\pi_\infty}^s$. The final desired result follows again by applying the continuous mapping theorem. Asymptotic independence holds based on the

covariance matrix given by Theorem 3. □

C.3 Auxiliary results

C.3.1 Tail inequality for OLS estimator

In this section we present slight generalizations of lemmas from Chen et al. (2021). The lemmas from the aforementioned paper correspond to the special case where $\phi(x, a) = [\mathbb{1}(a' = a)x]_{a' \in \mathcal{A}}$, i.e. that there is no generalization across actions. Our case allows for arbitrary $\phi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ satisfying conditions similar to the conditions imposed on $x \in \mathcal{X}$ in Chen et al. (2021).

Lemma 7. In the two-outcome online contextual bandit, if (A1), (A2), (A3), (A4), and (A5) hold and if at each timestep t, there is at least an ϵ_t-chance of selecting an action uniformly at random from $\mathcal{A}$, with $\{\epsilon_t\}$ non-increasing, then for data $I_T = I_T^{\text{prop}}$, I_T^{test}, $\{1, \ldots, T\}$ or $\widetilde{I}_T^{\text{prop}}$, and $* = r$ or s, then

$$P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta\right)$$
$$\geq 1 - \exp\left\{-\frac{T\epsilon_T}{8}\right\} - d\exp\left\{-\frac{T\epsilon_T p_I \lambda_{\min}(\Sigma)}{16 L_\phi^2}\right\} - 2d\exp\left\{-\frac{T\epsilon_T^2 (p_I)^2 \lambda_{\min}^2(\Sigma)\,\zeta^2}{32\sigma^2 d^2 L_\phi^2}\right\}$$

for any $\zeta > 0$, where $p_I = P(t \in I_T)$ for arbitrary t, so $p_I = p_{\text{prop}}, p_{\text{test}}$, 1, or $p_{\text{prop}} \cdot [1 - (1 - T^{-1})^T]$ respectively based on the choice of I_T.

Let $\Sigma = E_{A \sim \text{Uniform}(\mathcal{A})}[\phi(X, A)\phi(X, A)^\intercal]$ and let $\widehat{\Sigma}(I) = \frac{1}{|I|}\sum_{t \in I} \phi(X_t, A_t)\phi(X_t, A_t)^\intercal$ for some finite index set $I \subset \mathbb{N}$.

Proof. (Lemma 7) Our proof is a modified version of the proof of Proposition 3.1 of Chen et al. (2021). Let $\mathcal{R}_T = \{t \in \{1, \ldots, T\} : U_t \leq \epsilon_t\}$ be the indices at which uniformly random decisions were selected, per (4.4). Let $\mathcal{R}_T^I = \mathcal{R}_T \cap \text{supp}\, I_T$, where supp denotes the support, or set of unique elements, of a multiset; this detail is relevant only for the $I_T = \widetilde{I}_T^{\text{prop}}$ case. Define events

$$\mathcal{E} = \left\{\lambda_{\min}\left[\widehat{\Sigma}(\mathcal{R}_T^I)\right] > \frac{p_I}{2}\lambda_{\min}(\Sigma)\right\} \text{ and } \mathcal{R} = \left\{|\mathcal{R}_T| > \frac{T\epsilon_T}{2}\right\}.$$

To obtain the desired bound, first observe that

$$P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta\right) \geq P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta \,\middle|\, \mathcal{E}, \mathcal{R}\right) P\left(\mathcal{E} \,|\, \mathcal{R}\right) P\left(\mathcal{R}\right). \tag{C.9}$$

We find lower bounds for each of these terms. To bound the first term, begin by verifying that the condition for Lemma 8 holds under $\mathcal{E}$ and $\mathcal{R}$. We have

$$\widehat{\Sigma}(I_T) = \frac{|\mathcal{R}_T^I|}{|I_T|}\widehat{\Sigma}(\mathcal{R}_T^I) + \frac{|I_T| - |\mathcal{R}_T^I|}{|I_T|}\widehat{\Sigma}(I_T \setminus \mathcal{R}_T^I).$$

Noting in general that for square matrices A and B, $\lambda_{\min}(A+B) \geq \lambda_{\min}(A) + \lambda_{\min}(B)$, we have that

$$\begin{aligned}
\lambda_{\min}[\widehat{\Sigma}(I_T)] &\geq \lambda_{\min}\left[\frac{|\mathcal{R}_T^I|}{|I_T|}\widehat{\Sigma}(\mathcal{R}_T^I)\right] + \lambda_{\min}\left[\frac{|I_T| - |\mathcal{R}_T^I|}{|I_T|}\widehat{\Sigma}(I_T \setminus \mathcal{R}_T^I)\right] \\
&\geq \frac{|\mathcal{R}_T^I|}{|I_T|}\lambda_{\min}\left[\widehat{\Sigma}(\mathcal{R}_T^I)\right] + 0 \\
&\overset{(\mathcal{E})}{\geq} \frac{|\mathcal{R}_T^I|}{T}\frac{p_I}{2}\lambda_{\min}(\Sigma) \overset{(\mathcal{R})}{\geq} \frac{\epsilon_T\, p_I}{4}\lambda_{\min}(\Sigma).
\end{aligned}$$

So, applying Lemma 8 with $c = \frac{\epsilon_T\, p_I}{4}\lambda_{\min}(\Sigma)$, we have that

$$P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta \,\middle|\, \mathcal{E}, \mathcal{R}\right) \geq 1 - 2d\exp\left\{-\frac{T\epsilon_T^2 (p_I)^2 \lambda_{\min}^2(\Sigma)\,\zeta^2}{32\sigma^2 d^2 L_\phi^2}\right\}.$$

Moving on to the second term of (C.9), Lemma 9 implies that

$$P(\mathcal{E}) \geq 1 - d\exp\left\{-\frac{|\mathcal{R}_T| p_I \lambda_{\min}^2(\Sigma)}{8L_\phi^2}\right\}.$$

Conditioning on $\mathcal{R}$ does not change the conclusion of the lemma; it only introduces the bound on $|\mathcal{R}_T|$,

$$P\left(\mathcal{E} \,|\, \mathcal{R}\right) \geq 1 - d\exp\left\{-\frac{T\epsilon_T p_I \lambda_{\min}^2(\Sigma)}{16L_\phi^2}\right\}. \tag{C.10}$$

To bound the third term of (C.9), note that $|\mathcal{R}_T| = \sum_{t=1}^T \mathbb{1}(U_t \leq \epsilon_t)$ is the sum of

independent Bernoulli variables. We apply a Chernoff lower tail bound. For any $c \in (0,1)$,

$$P\left(|\mathcal{R}_T| \leq (1-c)\sum_{t=1}^{T}\epsilon_t\right) \leq \exp\left\{-\frac{c^2}{2}\sum_{t=1}^{T}\epsilon_t\right\}.$$

Choose $c = 1 - \frac{T\epsilon_T}{2}(\sum_{t=1}^{T}\epsilon_t)^{-1} = 1 - (2T^{-1}\sum_{t=1}^{T}\epsilon_t/\epsilon_T)^{-1}$. Clearly $c \leq 1$. By the non-increasing property of ϵ_T, $\sum_{t=1}^{T}\epsilon_t/\epsilon_T \geq T$, so $c \geq 1/2$. So, we can apply the above bound to get

$$P\left(|\mathcal{R}_T| \leq \frac{T\epsilon_T}{2}\right) \leq \exp\left\{-\frac{c^2}{2}\frac{\frac{T\epsilon_T}{2}}{1-c}\right\} = \exp\left\{-\frac{T\epsilon_T}{4}\frac{c^2}{1-c}\right\} \leq \exp\left\{-\frac{T\epsilon_T}{8}\right\},$$

where the last inequality holds because $c \geq 1/2$ implies $c^2/(1-c) \geq 1/2$. We conclude that

$$P(\mathcal{R}) \geq 1 - \exp\left\{-\frac{T\epsilon_T}{8}\right\}. \tag{C.11}$$

Multiplying these bounds and dropping positive exponential terms yields the result. □

Lemma 8. (Lemma 2 of Chen et al. (2021)) Writing $\widehat{\Sigma} = \widehat{\Sigma}(I_T)$, then if (A2) holds and $\lambda_{\min}(\widehat{\Sigma}) > c$ for some $c > 0$ almost surely, then for any $\zeta > 0$ and $* = r, s$,

$$P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta\right) \geq 1 - 2d\exp\left\{-\frac{Tc^2\zeta^2}{2\sigma^2 d^2 L_\phi^2}\right\}.$$

For the next two lemmas, let $Z_{T,t}^I = \mathbb{1}(t \in I_T)$.

Proof. (Adapted from Chen et al. (2021)) Using the fact that for a positive semi-definite matrix D, $\|D^{-1}\|_2 = \lambda_{\max}(D^{-1}) = [\lambda_{\min}(D)]^{-1}$, we have

$$\begin{aligned}
\|\widehat{\beta}^{*,I_T} - \beta^*\|_2 &= \left\|(\widehat{\Sigma})^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} Z_{T,t}^I \phi_t e_t^*\right)\right\|_2 \\
&\leq \frac{1}{T}\|(\widehat{\Sigma})^{-1}\|_2 \left\|\sum_{t=1}^{T} Z_{T,t}^I \phi_t e_t^*\right\|_2 \\
&\leq \frac{1}{Tc}\left\|\sum_{t=1}^{T} Z_{T,t}^I \phi_t e_t^*\right\|_2.
\end{aligned}$$

Denote the jth element of $\phi(X_t, A_t)$ as $\phi_{t,j}$. Then

$$\begin{aligned}
P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_2 \leq \zeta\right) &\geq P\left(\left\|\sum_{t=1}^{T} Z_{T,t}^I \phi_t e_t^*\right\|_2 \leq Tc\zeta\right) \\
&\geq P\left(\left|\sum_{t=1}^{T} Z_{T,t}^I \phi_{t,1}\, e_t^*\right| \leq \frac{Tc\zeta}{\sqrt{d}}, \ldots, \left|\sum_{t=1}^{T} Z_{T,t}^I \phi_{t,d}\, e_t^*\right| \leq \frac{Tc\zeta}{\sqrt{d}}\right) \\
&= 1 - P\left(\bigcup_{j=1}^{d}\left\{\left|\sum_{t=1}^{T} Z_{T,t}^I \phi_{j,1}\, e_t^*\right| \leq \frac{Tc\zeta}{\sqrt{d}}\right\}\right) \\
&\geq 1 - \sum_{j=1}^{d} P\left(\left|\sum_{t=1}^{T} Z_{T,t}^I \phi_{j,1}\, e_t^*\right| \leq \frac{Tc\zeta}{\sqrt{d}}\right) \\
&\geq 1 - \sum_{j=1}^{d} 2\exp\left(-\frac{Tc^2\zeta^2}{2d\sigma^2 L_\phi^2}\right) \\
&= 1 - 2d\exp\left(-\frac{Tc^2\zeta^2}{2d\sigma^2 L_\phi^2}\right).
\end{aligned}$$

The last inequality follows from a slight generalization of Lemma 2 of Goldenshluger and Zeevi (2013), simply replacing the i.i.d. normality of $\{e_t^*\}_{t=1}^T$ with (A5), the assumption that errors are uniformly subgaussian, plus a Chernoff bound. The Chernoff bound is given in the proof of Lemma 1 of Chen et al. (2021).

Finally, noting that for $a \in \mathbb{R}^d$, $\|a\|_1 \leq \|a\|_2\sqrt{d}$, we have

$$P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_1 \leq \zeta\right) \geq P\left(\|\widehat{\beta}^{*,I_T} - \beta^*\|_2 \leq \frac{\zeta}{\sqrt{d}}\right) \geq 1 - 2d\exp\left\{-\frac{Tc^2\zeta^2}{2\sigma^2 d^2 L_\phi^2}\right\}.$$

□

Lemma 9. (Lemma 3 from Chen et al. (2021)) Let $\{(X_t, A_t)\}_{t=1}^T$ be i.i.d. context-action pairs where $A_t \sim \text{Uniform}(\mathcal{A})$, and assume (A2), uniformly bounded features. If $\lambda_{\min}(\Sigma) > \lambda$ for some $\lambda > 0$, then

$$P\left\{\lambda_{\min}\left[T^{-1}\sum_{t=1}^{T} Z_{T,t}^I \phi(X_t, A_t)\phi(X_t, A_t)^\intercal\right] \leq \frac{p_I\lambda}{2}\right\} \leq d\exp\left\{-\frac{Tp_I\lambda}{8L_\phi^2}\right\}.$$

Proof. The proof mirrors the proof of Lemma 3 of Chen et al. (2021), except we replace

the action indicator, which they denote $a_t \sim \text{Bernoulli}(p)$, with the dataset indicator $Z^I_{T,t} \sim \text{Bernoulli}(p_I)$.

□

C.3.2 Asymptotic OLS theory with multiple outcomes and multiple splits

In this section we consider the problem of fitting an OLS model for a contextual bandit with arbitrary data splits and arbitrarily many real-valued outcomes. This result is a generalization of Theorem 3.1 of Chen et al. (2021), incorporating multiple outcomes, multiple sample splits with overlap, and, as in the rest of the paper, a more general feature representation.

We simultaneously estimate k vector-valued parameters based on k different outcomes associated with the same context-action pairs, using k possibly-overlapping subsets of the data. More precisely, let $f^1, f^2, \ldots, f^k : \mathcal{Y} \to \mathbb{R}$ be outcomes of interest, and for each $T \in \mathbb{N}$ let $I^1_T, I^2_T, \ldots, I^k_T$ be random subsets of $\{1, 2, \ldots, T\}$. Our parameter of interest is $\boldsymbol{\beta} = [\beta^1, \ldots, \beta^k]$, where for each $j \in \{1, \ldots, k\}$,

$$E[f^j(Y_t)|X_t = x, A_t = a] = \phi(X_t, A_t)^\intercal \beta^j, \tag{C.12}$$

which we estimate with, for each j,

$$\widehat{\beta}^j_T \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \sum_{t \in I^j_T} \left[f^j(Y_t) - \phi(X_t, A_t)^\intercal \beta\right]^2. \tag{C.13}$$

We construct the sample splits in the following manner. For each $j \in \{1, \ldots, k\}$, let B^j be a Borel-measurable subset of $[0, 1]$ and let $p_j = \lambda(B^j)$ and $p_{j \cap \ell} = \lambda(B^j \cap B^\ell)$. For $O_1, O_2, \ldots \overset{\text{iid}}{\sim} \text{Uniform}([0, 1])$ observable and mutually independent of all other variables, define

$$I^j_T = \Big\{t \in \{1, \ldots, T\} : O_t \in B^j\Big\}.$$

We now state some assumptions. Throughout, for each $T \in \mathbb{N}$, let $\mathcal{F}_T = \sigma(H(I_T))$ be the sigma algebra generated by all events up to time T and let $\widehat{\pi}_T$ be a data-dependent policy, i.e. an $\mathcal{F}_T$-measurable map from $\mathcal{X}$ to $\Delta(\mathcal{A})$.

(C1) There is a constant $L_\phi < \infty$ such that $P\left[\|\phi(X,a)\|_\infty < L_\phi\right] = 1$ for all $a \in \mathcal{A}$.

(C2) Let $e^j = f^j(Y) - E[f^j(Y)|X,A]$ for each j. These errors are uniformly subgaussian in the sense that there exists $\sigma^2 > 0$ such that for all j,

$$E[\exp(ce^j)|X = x, A = a] \leq \exp(\sigma^2 c^2/2) \text{ for all } c > 0,$$

for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$.

(C3) There is a limiting policy $\pi_\infty : \mathcal{X} \to \Delta(\mathcal{A})$ such that $\max_{a\in\mathcal{A}} |\widehat{\pi}_T(a|X) - \pi_\infty(a|X)| \xrightarrow{\text{p}} 0$ as $T \to \infty$.

(C4) The matrix $\boldsymbol{\phi}_{\pi_\infty} = E_{\pi_\infty}[\phi(X,A)\phi(X,A)^\intercal]$ has an inverse.

(C5) Sample splitting probabilities are non-zero; $p_j > 0$ for each $j \in \{1,\ldots,k\}$.

Theorem 4. (Asymptotic normality with multiple outcomes, sample splitting) Under (C1), (C2), (C3), (C4), and (C5),

$$\sqrt{T}\begin{bmatrix} \widehat{\beta}_T^1 - \beta^1 \\ \widehat{\beta}_T^2 - \beta^2 \\ \vdots \\ \widehat{\beta}_T^k - \beta^k \end{bmatrix} \rightsquigarrow N(\mathbf{0}, \boldsymbol{\Sigma}) \text{ as } T \to \infty,$$

where $\mathbf{0} \in \mathbb{R}^{kd}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{kd\times kd}$ is a $k \times k$ block matrix with (j,ℓ)-block

$$\boldsymbol{\Sigma}_{j,\ell} = \frac{p_{j\cap\ell}}{p_j p_\ell} \boldsymbol{\phi}_{\pi_\infty}^{-1} E_{\pi_\infty}\left[E(e^j e^\ell|X,A)\,\phi(X,A)\,\phi(X,A)^\intercal\right] \boldsymbol{\phi}_{\pi_\infty}^{-1} \in \mathbb{R}^{d\times d},$$

recalling that $\boldsymbol{\phi}_{\pi_\infty} = E_{\pi_\infty}[\phi(X,A)\phi(X,A)^\intercal]$.

Proof. We follow the proof structure of Chen et al. (2021), with an added Cramer-Wold-

style argument. Let $\phi_t = \phi(X_t, A_t)$. Let $\boldsymbol{\gamma} = [\gamma_1, \ldots, \gamma_k] \in \mathbb{R}^k$ and note that

$$\begin{aligned}\sqrt{T}\sum_{j=1}^{k}\gamma_j(\widehat{\beta}_T^j - \beta^j) &= \sqrt{T}\sum_{j=1}^{k}\gamma_j\left(\sum_{t\in I_T^j}\phi_t\,\phi_t^\intercal\right)^{-1}\sum_{1=t}^{T}\phi_t Z_t^j e_t^j \\ &= \sum_{j=1}^{k}\gamma_j\frac{T}{|I_t^j|}\left(\frac{1}{|I_T^j|}\sum_{t\in I_T^j}\phi_t\,\phi_t^\intercal\right)^{-1}\frac{1}{\sqrt{T}}\sum_{1=t}^{T}\phi_t Z_t^j e_t^j,\end{aligned}$$

where $e_t^j = f^j(Y_t) - E[f^j(Y_t)|X_t, A_t]$. By further rearranging terms, we obtain the following expression (results to be shown are stated underneath):

$$\sqrt{T}\sum_{j=1}^{k}\gamma_j(\widehat{\beta}_T^j - \beta^j) = \underbrace{\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t\,\phi_t^\intercal\right)^{-1}}_{\xrightarrow{\text{P}}\boldsymbol{\phi}_{\pi_\infty}}\underbrace{\frac{1}{\sqrt{T}}\sum_{1=t}^{T}\phi_t\left(\sum_{j=1}^{k}\gamma_j p_j^{-1}Z_t^j e_t^j\right)}_{\rightsquigarrow N(0,\boldsymbol{G}^{\boldsymbol{\gamma}})} + \underbrace{\delta_T^{\boldsymbol{\gamma}}}_{\xrightarrow{\text{P}}0}, \quad \text{(C.14)}$$

where

$$\delta_T^{\boldsymbol{\gamma}} = \sum_{j=1}^{k}\gamma_j\left\{\left[\frac{T}{|I_t^j|}\left(\frac{1}{|I_T^j|}\sum_{t\in I_T^j}\phi_t\,\phi_t^\intercal\right)^{-1} - p_j^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t\,\phi_t^\intercal\right)^{-1}\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_t Z_t^j e_t^j\right\}.$$

Each of the convergence results stated in (C.14) is proved in a subsection below.

Asymptotic normality

We begin with the normality result. We use the Cramer-Wold device here as well. Let $v \in \mathbb{R}^d$ and let

$$e_t^{\boldsymbol{\gamma}} = \sum_{j=1}^{k}\gamma_j p_j^{-1}Z_t^j e_t^j.$$

We want to show that for

$$\begin{aligned}M_{T,j} &:= \frac{1}{\sqrt{T}}\sum_{t=1}^{j}v^\intercal\phi_t e_t^{\boldsymbol{\gamma}}, \\ M_{T,T} &\rightsquigarrow N(0, v^\intercal\boldsymbol{G}^{\boldsymbol{\gamma}}v) \text{ as } T\to\infty,\end{aligned}$$

for some matrix $\boldsymbol{G}^\gamma \in \mathbb{R}^{d\times d}$ not depending on v. Note that $E(v^\intercal\phi_t e_t^\gamma|\mathcal{F}_{t-1}) = E[v^\intercal\phi_t E(e_t^\gamma|X_t, A_t)|\mathcal{F}_{t-1}] = 0$, so $\{M_{T,j}\}_{j=1}^T$ is a martingale for each $T \in \mathbb{N}$. By Theorem 3.2 in Hall and Heyde (2014), if we define $\mathcal{F}_{T,j} = \mathcal{F}_j$ and $U_{T,t} = T^{-1/2}v^\intercal\phi_t e_t^\gamma$ and consider the martingale array $\{M_{T,j}, \mathcal{F}_{T,j}, 1 \le j \le T\}_{T=1}^\infty$, then the following two conditions are sufficient to establish asymptotic normality:

1. For all $\epsilon > 0$, $\sum_{t=1}^T E\left[U_{T,t}^2 \mathbb{1}(|U_{T,t}| > \epsilon) \,\middle|\, \mathcal{F}_{t-1}\right] \xrightarrow{\text{p}} 0$ as $t \to \infty$.
2. $V_T^2 = \sum_{t=1}^T E(U_{T,t}^2|\mathcal{F}_{t-1})$ converges in probability to a finite number.

We begin with the first condition. First note that,

$$T^{-1}(v^\intercal\phi_t)^2 = T^{-1}\|v\|_2^2\left[(v/\|v\|_2)^\intercal\phi_t\right]^2 \le T^{-1}\|v\|_2^2 \sup_{\|w\|_2=1}(w^\intercal\phi_t)^2$$
$$= T^{-1}\|v\|_2^2\left(\sup_{\|w\|_2=1}|w^\intercal\phi_t|\right)^2 = T^{-1}\|v\|_2^2\|\phi_t\|_\infty^2 \le T^{-1}\|v\|_2^2 L_\phi^2,$$

where the last inequality follows by (A2). Now let $\epsilon > 0$. We have that

$$\begin{aligned}\sum_{t=1}^T E\left[U_{T,t}^2\mathbb{1}(|U_{T,t}| > \epsilon) \,\middle|\, \mathcal{F}_{t-1}\right] &= \sum_{t=1}^T E\left[T^{-1}(v^\intercal\phi_t)^2(e_t^\gamma)^2\mathbb{1}(T^{-1}(v^\intercal\phi_t)^2(e_t^\gamma)^2 > \epsilon^2) \,\middle|\, \mathcal{F}_{t-1}\right] \\ &\le \frac{\|v\|_2^2 L_\phi^2}{T}\sum_{t=1}^T E\left[(e_t^\gamma)^2\,\mathbb{1}\left((e_t^\gamma)^2 > \frac{T\epsilon^2}{\|v\|_2^2 L_\phi^2}\right)\,\middle|\,\mathcal{F}_{t-1}\right] \\ &\le \|v\|_2^2 L_\phi^2 E\left[\overline{e}^2\,\mathbb{1}\left(\overline{e}^2 > \frac{T\epsilon^2}{\|v\|_2^2 L_\phi^2}\right)\right],\end{aligned}$$

where the last inequality follows from (C2), choosing $\overline{e}^2 \sim N(0, \sigma^2)$. The term inside the expectation is bounded above by integrable $\overline{e}^2$ and converges to 0 as $T \to \infty$, so the desired result holds by the Dominated Convergence Theorem.

Now to the second condition. We have

$$V_T^2 = \sum_{t=1}^T E(U_{T,t}^2|\mathcal{F}_{t-1}) = \frac{1}{T}\sum_{t=1}^T v^\intercal E[\phi_t\phi_t^\intercal(e_t^\gamma)^2|\mathcal{F}_{t-1}]v$$
$$= v^\intercal\frac{1}{T}\sum_{t=1}^T E\left(E\left\{\phi_t\phi_t^\intercal E\left[(e_t^\gamma)^2|X_t, A_t)\right] \,\middle|\, \mathcal{F}_{t-1}, X_t\right\} \,\middle|\, \mathcal{F}_{t-1}\right)v. \tag{C.15}$$

By independence of the sample splitting indicators, the inner term simplifies:

$$\begin{aligned}
E\left[(e_t^\gamma)^2 | X_t, A_t\right] &= E\left[\left(\sum_{j=1}^k \gamma_j p_j^{-1} Z_t^j e_t^j\right)^2 \middle| X_t, A_t\right] \\
&= E\left[\sum_{j=1}^k \sum_{\ell=1}^k \gamma_j \gamma_\ell p_j^{-1} p_\ell^{-1} Z_t^j Z_t^\ell e_t^j e_t^\ell \middle| X_t, A_t\right] \\
&= \sum_{j=1}^k \sum_{\ell=1}^k \gamma_j \gamma_\ell (p_j p_\ell)^{-1} E(Z_t^j Z_t^\ell) E(e_t^j e_t^\ell | X_t, A_t) \\
&= \sum_{j=1}^k \sum_{\ell=1}^k \gamma_j \gamma_\ell \frac{p_{j\cap\ell}}{p_j p_\ell} E(e_t^j e_t^\ell | X_t, A_t).
\end{aligned}$$

For policy $\pi : \mathcal{X} \to \Delta(A)$ and context $x \in \mathcal{X}$, let

$$\begin{aligned}
g^\gamma(\pi, x) :&= E_\pi\left\{\phi(X,A)\phi(X,A)^\intercal E\left[(e^\gamma)^2 | X, A\right] \,\middle|\, X = x\right\} \\
&= \sum_{j=1}^k \sum_{\ell=1}^k \gamma_j \gamma_\ell \frac{p_{j\cap\ell}}{p_j p_\ell} E_\pi\left[\phi(X,A)\phi(X,A)^\intercal E(e^j e^\ell | X, A) \,\middle|\, X = x\right] \\
&\propto \sum_{a\in\mathcal{A}} \Big[\phi(x,a)\phi(x,a)^\intercal E(e^j e^\ell | X = x, A = a)\Big]\pi(a|x),
\end{aligned}$$

where $\pi(a|x) = P[\pi(x) = a]$. From this we see that $g^\gamma(\pi, x)$ is linear and hence continuous in π for any x. Therefore $g^\gamma(\pi) = E[g^\gamma(\pi, X)]$ is continuous in the following sense: for any $\xi > 0$, there exists $\delta > 0$ such that $\max_a P[\pi(a|X) \neq \pi'(a|X)] < \delta$ implies $\|g^\gamma(\pi) - g^\gamma(\pi')\|_\infty < \epsilon$.

Returning to (C.15), note that $V_T^2 = v^\intercal \frac{1}{T}\sum_{t=1}^T g^\gamma(\widehat{\pi}_t)$. So, the continuous mapping theorem (CMT) plus the existence of a limiting policy (C3) implies that $g^\gamma(\widehat{\pi}_T) \xrightarrow{\text{p}} g^\gamma(\pi_\infty)$ as $T \to \infty$. By i.i.d.-ness of $\{X_t\}_{t\in\mathbb{N}}$ and Lemma 4 of Chen et al. (2021), this means that

$$V_T^2 = v^\intercal \frac{1}{T}\sum_{t=1}^T g^\gamma(\widehat{\pi}_t) v \xrightarrow{\text{p}} v^\intercal g^\gamma(\pi_\infty) v.$$

Thus the conditions for the Martingale Central Limit Theorem are satisfied, and we

conclude that

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_t\, e_t^{\gamma} \rightsquigarrow N(0, \boldsymbol{G}^{\gamma}),$$

where

$$\boldsymbol{G}^{\gamma} = g^{\gamma}(\pi_\infty) = \sum_{j=1}^{k}\sum_{\ell=1}^{k}\gamma_j\gamma_\ell\frac{p_{j\cap\ell}}{p_j p_\ell}E_{\pi_\infty}\left[\phi(X,A)\phi(X,A)^{\intercal}E(e^j e^\ell|X,A)\right]. \tag{C.16}$$

Matrix limit

Now we turn to the leading term of (C.14). We wish to show

$$\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t\phi_t^{\intercal}\right)^{-1} \xrightarrow{\text{p}} \boldsymbol{\phi}_{\pi_\infty}^{-1}.$$

To do this, we begin by showing

$$\frac{1}{T}\sum_{t=1}^{T}U_t \xrightarrow{\text{p}} v^{\intercal}\boldsymbol{\phi}_{\pi_\infty}v \text{ as } T\to\infty, \text{ where } U_t := v^{\intercal}\phi_t\phi_t^{\intercal}v.$$

Since U_t is $\mathcal{F}_t$-measurable and $|U_t| = v^{\intercal}\phi_t\phi_t^{\intercal}v \le \|v\|_2^2 L_\phi^2$, Theorem 2.19 of Hall and Heyde (2014) gives that

$$\frac{1}{T}\sum_{t=1}^{T}U_t - \frac{1}{T}\sum_{t=1}^{T}E(U_t|\mathcal{F}_{t-1}) \xrightarrow{\text{p}} 0.$$

Therefore it is sufficient to find the limit for the term with conditional expectations. Let $h(\widehat{\pi}_t) = E[\phi_t\phi_t^{\intercal}|\mathcal{F}_{t-1}]$, so we have

$$\frac{1}{T}\sum_{t=1}^{T}E(U_t|F_{t-1}) = \frac{1}{T}\sum_{t=1}^{T}v^{\intercal}h(\widehat{\pi}_t)v \xrightarrow{\text{p}} v^{\intercal}h(\pi_\infty)v,$$

by a similar argument as before, using Lemma 4 of Chen et al. (2021), (C3), and the CMT. Finally, Lemma 6 of Chen et al. (2021) says that convergence of a symmetric

matrix is equivalent to convergence of all its quadratic forms, so we have

$$\frac{1}{t}\sum_{t=1}^{T}\phi_t\phi_t^\intercal \xrightarrow{\text{p}} h(\pi_\infty) = E_{\pi_\infty}[\phi(X,A)\phi(X,A)^\intercal] = \boldsymbol{\phi}_{\pi_\infty}.$$

The desired result holds under the invertibility assumption (C4) by noting the continuity of matrix inversion and applying the CMT.

Remainder term is negligible

To prove that $\delta_T^{\boldsymbol{\gamma}} \xrightarrow{\text{p}} 0$, let $j \in \{1,\ldots,k\}$ and pick $\boldsymbol{\gamma} = \boldsymbol{e}_j \in \mathbb{R}^k$, the jth standard Euclidean basis vector. Then

$$\delta_T^{\boldsymbol{\gamma}} = \left[\frac{T}{|I_t^j|}\left(\frac{1}{|I_T^j|}\sum_{t\in I_T^j}\phi_t\,\phi_t^\intercal\right)^{-1} - p_j^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\phi_t\,\phi_t^\intercal\right)^{-1}\right]\frac{1}{\sqrt{T}}\sum_{t=1}^{T}\phi_t Z_t^j e_t^{\boldsymbol{\gamma}}.$$

By the results in Section C.3.2, the term on the right hand side is asymptotically normal and hence $O_p(1)$. Rearranging terms, we have

$$\delta_T^{\boldsymbol{\gamma}} = \left[\left(\frac{1}{T}\sum_{t=1}^{T}Z_t^j\phi_t\,\phi_t^\intercal\right)^{-1} - \left(p_j\cdot\frac{1}{T}\sum_{t=1}^{T}\phi_t\,\phi_t^\intercal\right)^{-1}\right]O_p(1).$$

As shown in Section C.3.2, $\frac{1}{T}\sum_{t=1}^{T}\phi_t\,\phi_t^\intercal \xrightarrow{\text{p}} \boldsymbol{\phi}_{\pi_\infty}$. Noting that $\{Z_t^j\}_{t\in\mathbb{N}}$ are i.i.d., the same argument goes through to prove that $\frac{1}{T}\sum_{t=1}^{T}Z_t^j\phi_t\,\phi_t^\intercal \xrightarrow{\text{p}} p_j\,\boldsymbol{\phi}_{\pi_\infty}$. Thus by the CMT and continuity of the matrix inverse, $\boldsymbol{\delta}_T^{\boldsymbol{\gamma}} \xrightarrow{\text{p}} 0$. This result implies the result for general $\boldsymbol{\gamma} \in \mathbb{R}^k$, so we are done.

Putting it all together

Combining the results from the previous subsections, (C.14) implies that for any $\boldsymbol{\gamma} \in \mathbb{R}^k$,

$$\sqrt{T}\sum_{j=1}^{k}\gamma_j(\widehat{\beta}_T^j - \beta^j) \rightsquigarrow N\big(0, \boldsymbol{\phi}_{\pi_\infty}^{-1}\boldsymbol{G}^{\boldsymbol{\gamma}}\boldsymbol{\phi}_{\pi_\infty}^{-1}\big) \text{ as } T\to\infty.$$

Plugging in (C.16) for $\boldsymbol{G}^{\boldsymbol{\gamma}}$, we get that

$$\begin{aligned}
\boldsymbol{\phi}_{\pi_\infty}^{-1}\boldsymbol{G}^{\boldsymbol{\gamma}}\boldsymbol{\phi}_{\pi_\infty}^{-1} &= \boldsymbol{\phi}_{\pi_\infty}^{-1}\sum_{j=1}^{k}\sum_{\ell=1}^{k}\gamma_j\gamma_\ell\frac{p_{j\cap\ell}}{p_j p_\ell}E_{\pi_\infty}\left[\phi(X,A)\phi(X,A)^\intercal E(e^j e^\ell|X,A)\right]\boldsymbol{\phi}_{\pi_\infty}^{-1} \\
&= \sum_{j=1}^{k}\sum_{\ell=1}^{k}\gamma_j\gamma_\ell\left\{\boldsymbol{\phi}_{\pi_\infty}^{-1}\frac{p_{j\cap\ell}}{p_j p_\ell}E_{\pi_\infty}\left[\phi(X,A)\phi(X,A)^\intercal E(e^j e^\ell|X,A)\right]\boldsymbol{\phi}_{\pi_\infty}^{-1}\right\} \\
&= \sum_{j=1}^{k}\sum_{\ell=1}^{k}\gamma_j\gamma_\ell\boldsymbol{\Sigma}_{j,\ell}.
\end{aligned}$$

Because $\boldsymbol{\gamma}$ was chosen arbitrarily, the desired result holds by the Cramer-Wold device.

□

C.4 Miscellanea

C.4.1 Power of multiple testing vs. splitting in single arm detection

Consider a standard k-armed bandit defined as follows:

- Action 0 is the baseline, with known safety level 0.
- Action 1 has known infinite reward and strictly positive safety ("effect size").
- Actions $2, 3, \ldots, k$ have reward that is known to be finite, and safety $-\infty$.

A hypothesis-test-based safe learning algorithm (see introduction to Section 4.3) is given n samples per arm, then queried for a recommended arm.

By construction, actions 2 and higher will never pass the safety test. Action 1 is guaranteed to be proposed by SPT as long as nonzero probability is assigned to it passing the safety test, so for both SPT and Pretest All, it is guaranteed that action 1 will be selected as long as it passes the safety test. So, in this abstract example, the only thing that matters is each algorithm's ability to allow action 1 to pass the safety test. Figure C.1 shows each algorithm's probability of passing action 1. Notably, for SPT, the probability is invariant to the number of arms, but for Pretest All, the power decays to 0.

This contrived example illustrates the key benefit of SPT: by not testing every action, it maintains the possibility of high performance even in settings with many actions.

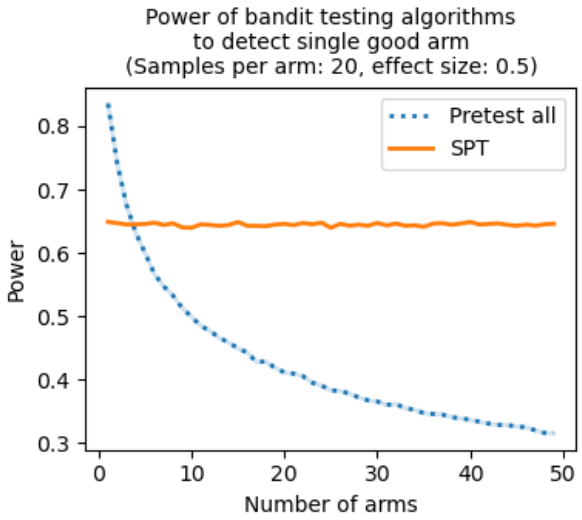

Figure C.1: The power (probability of detection) of Pretest All vs. SPT to select a single good action from many.

C.4.2 The peril of data reuse

At the end of Section 4.4 we commented that, technically, SPT does not require sample splitting to attain (α, τ)-consistency, but that sample splitting is a crucial component for maintaining safety constraint satisfaction in small samples. Here, we give simulation results that illustrate the problem with data reuse in SPT. We compare Pretest All and SPT as defined in Section 4.5, with an additional version of SPT that uses $I_T^{\text{prop}} = I_T^{\text{test}} = \{1, \ldots, T\}$ at all timesteps T, which we call 'SPT (no split)'. Except where otherwise stated, we use the same experimental parameters as in Section 4.5. Simulations are performed on the 'All unsafe' bandit.

All unsafe

X	$\mathcal{A}$	τ	d
-	$\{0, 1, \ldots, 9\}$	0	10

A simple multi-arm bandit where all arms except 0 are unsafe; $Q^s(0) = 0$ and $Q^s(a) = -1$ for $a > 0$. All unsafe actions have high reward; $Q^r(0) = 0$ and $Q^r(a) = 100$ for $a > 0$. Errors are independent and distributed as $e^r \sim N(0, 1)$, $e^s \sim N(0, 10^2)$.

Simulation results from this setting are shown in Figure C.2. As before, all runs were initialized with 4 i.i.d. samples from each arm and a safety level of $\alpha = 0.1$ were used. The results show that, while Pretest All and SPT do fall slightly below this at the earliest steps, it still selects safe actions at a rate higher than 0.8 from the beginning, and quickly transitions to above the $1-\alpha = 0.9$ level. SPT (no split), on the other hand, selects unsafe actions almost half the time.

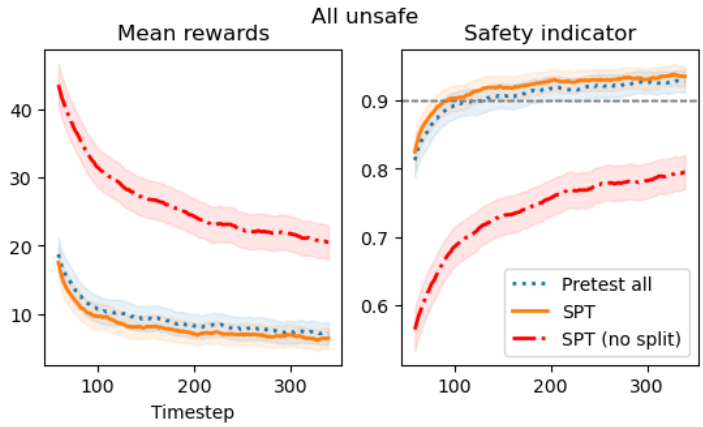

Figure C.2: Comparison of algorithms on the All Unsafe bandit. Highlighting indicates 95% pointwise confidence intervals based on 1,000 independent runs.

C.4.3 Safety of Pretest All

We sketch the argument that the Pretest All algorithm defined in Section 4.5 is also consistent in the sense of SPT. By Lemma 1 with $p_{\text{test}} = 1$, the multinomial bootstrap estimator of reward $\widehat{\beta}^r_T \xrightarrow{\text{p}} \beta^r_T$ as $T \to \infty$. By Lemma 2 with $I_T = \{1, \ldots, T\}$ and finiteness of $\mathcal{A}$, the set of estimated safe actions converges to the set of truly safe actions. Thus, under the same conditions as Theorem 2 and using reasoning found in its proof, given in Appendix C.2.3, Pretest All is also safe and consistent.