Delayed Rewards and Dynamic States in the Multi-Armed Bandit Problem

By: David Newman

Advisor: Giri Parameswaran

May 2, 2019

Abstract

This paper explores a variation of the multi-armed bandit problem and proposes a new strategy, the Pure Unknown strategy, to optimally maximize payoffs. In this game, the player chooses between two arms– one with known probability distributions and the other with unknown probability distributions–and does not realize the payoff of the arm she chooses until the next time period, where the probabilities of each time period are state dependent and those states are determined by a stochastic process. Additionally, elements of ambiguity aversion are incorporated into the model to reflect individuals' preferences for choices with known probabilities over those with unknown probabilities. Four strategies, including the Pure Unknown strategy, play this game to see which strategy produces the highest average payoff, and the other three strategies are inspired by previous multi-armed bandit literature. Results find that the Pure Unknown strategy is the most optimal strategy when the Normality assumption, which ultimately represents ambiguity averse preferences, is not present.

Contents

Acknowledgments

I would like to thank my parents, Jim and Teri, for their unconditional love and support throughout my life. Without them, I would not be able to attend a prestigious school such as Haverford, and I am beyond grateful for–among many things–their investment in my education and desire for me to succeed in life.

I would like to thank my thesis advisor, Giri, for his guidance and help with my thesis. Having taken several classes with him prior to writing my thesis, including my first economics class at Haverford, Giri has played an important role in my development as an economics student.

While I have made many friends in college, I would like to thank Tomas, Laura, and Phillip in particular. Ever since we were assigned to the same freshman hall, we have stayed close friends, and I could not have asked for better and more supportive friends during my time in college.

1 Introduction

People face decisions under uncertainty almost every day, and within a set of available choices, some information that can be crucial to choosing the correct action may not be immediately available, especially when the rewards associated with specific choices are delayed and not realized when the choice is made. For example, suppose a young investor sees that the market is in a bull phase and wishes to invest some of her savings in one of the two following options: a safe asset such as Treasury Bills where the expected payoff over six months is essentially known, or a risky asset such as stock in a newly listed start-up Biotech company that had record growth in the past few quarters but may not be able to sustain this growth through the next six months. Regardless of which asset the investor chooses, her payoff will not be realized until six months from now when the market could either be bullish or bearish. Both states of the market affect the investor's potential realized gains or losses for both assets. If the market continues to be bullish, then she will realize gains for both choices, but the gains in this state of the world will likely be higher if she chooses the Biotech stock because it has more upside potential. On the other hand, if the market enters a downturn, then her losses will be less severe–she could even realize some small gains–if she invests in Treasury Bills because there is a distinct chance that the Biotech company could go under. Given that the investor wants to earn the highest return off of her investment, which asset does she invest in today?

The only way that the investor can become more certain about which asset is the best investment is to choose the Biotech stock and learn about its expected payoff over time and compare this payoff to that of the Treasury Bills. Learning in this investment scenario, or any similar circumstance, can be costly because if the choice with the known probability distribution of outcomes (Treasury Bills) yields a higher reward than the choice with the unknown probability distribution

of outcomes (Biotech stock), then learning induces regret. Conversely, learning would be made worthwhile if the Biotech stock ultimately has higher realized gains than the Treasury Bills. This dilemma is an example of the multi-armed bandit[1] problem, which focuses on the tradeoff between exploitation (choosing the current best option) and exploration (choosing options with unknown probabilities to see if they are better). Thus, the player seeks to exploit the option with the greatest reward. To study this problem, this paper creates a model to characterize optimal strategies in a two-armed bandit problem when rewards are delayed by one time period and where players employ Bayesian learning to find optimal solutions to the bandit problem. In this model, there are two possible states for each time period: good and bad. This means that when a player is, for example, in the good state, they choose the arm based on their beliefs about which state the next period will likely be in, and which arm has the highest probability of generating a reward in the future state.

This paper is comprised of 6 sections. Section 2 reviews utility theory literature that lays the foundation for the multi-armed bandit problem, the concept of ambiguity aversion and its relevance to this topic, and the literature that specifically examines the multi-armed bandit problem. Section 3 presents this paper's variant of the multi-armed bandit problem and the structure of the game. Section 4 compares and constrasts the four different strategies that play the game, and briefly explains how the data to study these strategies is created. Section 5 presents the results of these strategies within the game under two different conditions that affect the degree of uncertainty in the game, as well as robustness checks, to determine which strategy is most optimal. Finally, Section 6 concludes and discusses the implications, takeaways, and possible extensions

[1]The term bandit refers to slot machines, which are often called one-armed bandits. It is easy to see that slot machines are the perfect real-world example of this problem because gamblers likely do not know the probability distribution of the slot machine they play. Out of all the machines they can play, they learn and subsequently choose which machine gives them the highest reward.

from the results.

2 Literature Review

There is extensive research of the bandit problem, and the common framework for solving the bandit problem comes from the seminal paper by Robbins (1952), which explores the concept of sequential sampling from multiple states with unknown probability distributions to maximize expected payoffs. An important lesson from Robbins (1952) is that players create a dynamically adapting rule that dictates their actions in each iteration of the game, which includes the ability to stop the game due to sequential sampling. Given this foundation, additional research examines optimal play and different strategies within the bandit problem.

Gittins (1979) provides a methodology to solve the bandit problem with dynamic allocation indices. Essentially, for every period of the game, each arm is assigned a Gittins Index, which is a simple integer; the arm with the highest Gittins Index in each time period is the current best choice, and the player chooses that arm so long as it has the highest Gittins Index. If another arm has the higher Gittins Index in a later time period, then that other arm is chosen. The current best choice locally, however, is not necessarily the best choice globally. The Gittins Index could prove to be useful to find optimal strategies in this paper since the index works with Bernoulli probability distributions and finitely repeated bandit problems. Additionally, the Gittens Index is commonly used as a method to solve various multi-armed bandit problems, such as in Anderson (2012).

Another approach to solving the bandit problem involves subjective and nonadditive probabilities since the probability distributions of the options are unknown, which requires the use of Choquet integration (Choquet 1955). While

the Gittins Index produces a scalar value from an analysis of stochastic processes, the Choquet Integral is a method of integrating specialized monotonic functions. Schmeidler (1989) takes classic von Neumann-Morgenstern expected utility theory—as defined in Morgenstern & Von Neumann (1947)—and amends the independence axiom to allow for comonotonic independence so that expected utility with subjective probabilities can be calculated with the Choquet integral. Since Schmeidler (1989) does not directly tackle the bandit problem, the important concepts to draw from his paper are how he builds his model around subjective probabilities and how they differ from the objective probabilities that are seen in standard expected utility models.

The literature above provides multiple angles to approach and design an analysis of the bandit problem, while supplementary research identifies specific strategies for general and specific iterations of the bandit problem. Since the bandit problem represents the exploitation/exploration tradeoff, there are several important strategies to highlight. One strategy is the ϵ-greedy strategy defined in Tokic (2010), in which the current best option is chosen with probability $1 - \epsilon$, while a random arm is chosen with probability ϵ. This strategy represents the player's belief that their current best choice may not be the global best choice, which is the main drawback of the Gittins Index strategy. So, a player believes the current best choice is the global best choice with probability $1 - \epsilon$. Tokic (2010) also defines the softmax strategy, where players choose a value of ϵ that is not significantly small to reflect greater uncertainty about which arm is the global best choice. For example, one version of the softmax strategy is to let $\epsilon = 0.5$, which, in the case of the two-armed bandit problem, makes the arm selection process a simple coin flip.

Auer et. al (1995) create a version of the bandit problem where the probability distributions that determine rewards are controlled by an adversary who can

change these distributions in each time period, which means the player cannot optimize their strategy as they, for example, learn the stochastic process that determines the distributions. Their paper relaxes the statistical assumptions made in Lai & Robbins (1985), which finds an allocation rule that minimizes regret in the long-run of the bandit problem since players become more certain that they are not choosing the inferior option over time. So, instead of a normal rule in previous bandit problems, Auer et. al (1995) propose a randomized algorithm for the player to implement that best responds to the presence of an adversary. This algorithm draws parallels with the different strategies proposed by Tokic (2010) in that the player must to some extent randomize their choices between exploitation and exploration, but it does not take greed into account.

Another important concept to consider is ambiguity aversion, which is a preference for choices with known probabilities of outcomes over choices with unknown probabilities of outcomes. In the context of the bandit problem, ambiguity averse individuals prefer exploitation ahead of exploration; if the young investor from the introduction is ambiguity averse, she will choose to invest her money in Treasury Bills. The classic example of ambiguity aversion is the Ellsberg paradox, which finds that individuals have preferences that are inconsistent with standard expected utility theory because they are ambiguity averse when presented with bets that include partial unknown probability distributions (Ellsberg 1961). More explicitly, Ellsberg (1961) supposes that there are two urns: one with 50 red balls and 50 black balls, and another with 100 balls, but the number of red and black balls is unknown. Not only does Ellsberg (1961) find that a majority of people prefer to draw from the former urn, but also that their preferences violate the expected utility axioms created by Savage (1954). It is also important to highlight that ambiguity aversion is not the same as risk aversion. When an individual has risk averse preferences, they can compare

choices and their possible outcomes with probabilities to see which choice is the least risky. On the other hand, an individual with ambiguity averse preferences cannot compare the probabilities of both urns in the Ellsberg paradox example because the 50/50 chance of drawing a red ball from the known urn cannot be judged next to an unknown chance of drawing a red ball from the unknown urn.

While a majority of research that focuses on the bandit problem does not consider ambiguity aversion, Anderson (2012) directly studies the effects of ambiguity aversion on optimal decision making in the multi-armed bandit problem. Since ambiguity aversion intuitively suggests that individuals have some bias towards exploitation rather than exploration, the results of Anderson (2012) support this logic through an experiment that finds lower Gittins indexes than what the theory predicts, and Anderson (2012) argues this difference is due to ambiguity aversion. Thus, the effects of ambiguity aversion on how individuals approach the bandit problem are important to consider; ambiguity aversion motivates both the main strategy this paper adds to the literature and aspects of this paper's model.

This paper differs from the previous research in two main ways. First, the existence of dynamic states that are independent of choices made by the player is an application of the multi-armed bandit problem that has not been studied. While there exists a plethora of algorithms that characterize optimal solutions to the multi-armed bandit problem—some of which, such as the Gittins Index, may be useful for finding optimal solutions in this paper—these solutions may not be general enough to solve the multi-armed bandit problem in this context. Consequently, this paper explores different strategies within the context of the model to determine which strategy produces the highest reward.

The other way in which this paper differs from the previous research is a new strategy for the multi-armed bandit problem: the Pure Unknown strategy.

As the name suggests, a player that implements the Pure Unknown strategy commits to exploration of the unknown arms until theses arms are no longer unknown. The Pure Unknown strategy in this game also uses a unique learning process that incorporates the spirit of ambiguity aversion to determine the relationship between the known and unknown arms. After exploration is complete, the player can confidently choose the arm with the highest expected payoff because she now knows the probability distributions for all of the possible choices. In addition to exploring the nuances of this Pure Unknown strategy, Section 3 also creates the model.

3 Model

This model has a finite time horizon $t \in \{1, \ldots, n\}$ where each time period τ takes on a state $s_t \in [g, b]$, which represent a good and bad state, respectively. Each arm is characterized by a state dependent payoff probability that describes the likelihood of success. Regardless of the arm, players receive a payoff of 1 with a successful pull of the arm, and 0 otherwise., More formally:

$$r_t(a_t; s_t) = \begin{cases} 1, & p_{a,s_{t+1}} \\ 0, & 1 - p_{a,s_{t+1}} \end{cases}$$

Rewards in each time period τ are dependent on the arm chosen in the previous period $\tau - 1$ to reflect the delayed realization of rewards, and the states may or may not be the same between time periods. Each arm, denoted by $a_t \in [k, u]$, where k represents the known arm and u represents the unknown arm, has the

following rewards probabilities p_{as} for each state:

$$p_{k,s} = \begin{cases} 0.75 & s = g \\ 0.25 & s = b \end{cases}$$

$$p_{u,s} = \begin{cases} p_{u,g} & s = g \\ p_{u,b} & s = b \end{cases}$$

where $p_{u,g} > p_{u,b}$ and the true value of these probabilities are unknown. Thus, the player does not know, for example, if $p_{u,g} > 0.75$ and $p_{u,b} < 0.25$. However, she does have prior beliefs about the distributions of these probabilities, which implies that she can use Bayesian updating to learn an approximation of the true values and to determine which arm has a higher expected payoff, given the state in which the arm is actually pulled. Thus, the tradeoff between exploitation of the arm with known probabilities and the exploration of the arm with unknown probabilities motivates the player's objective function:

$$U = \max_{a_t \in \{k,u\}} \sum_t r_t(a_t; s_t) \tag{1}$$

The player wants to maximize her payoff, which means she must at least try to learn about the unknown arm for each state of the world.

While a player knows the existence of the two states and that each sequential time period is one of these states, she does not know for certain which state will characterize the next time period. Thus, states evolve according to a Markov

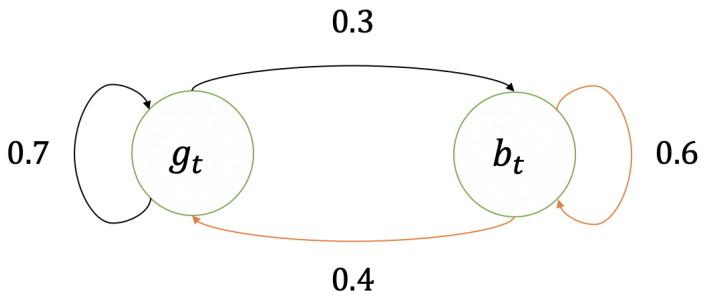


Figure 1: Markov chain diagram.

process with known probabilities[2] that are represented in the following matrix:

$$\begin{bmatrix} p_{gg} & p_{gb} \\ p_{bg} & p_{bb} \end{bmatrix} e.g. \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

Figure 1 shows a diagram for this Markov chain with the example transition probabilities. The choice of these probabilities is not entirely arbitrary, but this paper uses these probabilities merely as an example. For this specific case of p_{gg} and p_{bb}, the current state of the world should more likely than not reamin the same tomorrow. This Markov chain does, however, have a long run bias towards the good state; the steady state of this Markov chain dictates that for any future time period, the probability that that time period will be the good state is 0.57 and, consequently, 0.43 for the bad state. The higher frequency of the good state does not necessarily affect the rest of the game in any meaningful way so long as the game is played long enough to sufficiently experience both states.

Given the structure of the game and the player's objective function, the next step is to understand how the player uses Bayesian updating to learn the prob-

[2] A possible extension either within or to this paper would be to assume that either the transition probabilities are unknown. This case requires an additional learning process on top of the exploration of the unknown arm's true payoff probabilities, which further complicates the problem.

abilities of the unknown arm. In order to learn the true values of $p_{u,s}$ the player must have a prior belief that becomes a posterior belief—which takes the form of a prior probability distribution function multiplied by a likelihood function—after repeatedly pulling the unknown arm n times. Let $\pi(p_{u,s})$ represent a prior probability distribution, and $f(x|p_{u,s})$ be a likelihood function where x is a variable that represents the number of successful arm pulls. Given functional forms for the prior distribution and likelihood function, Bayes' Theorem states that the posterior probability distribution is:

$$\pi(p_{u,s}|x) = \frac{f(x|p_{u,s})\pi(p_{u,s})}{\int_0^1 f(x|p_{u,s})\pi(p_{u,s})dp_{u,s}} = f(x|p_{u,s})\pi(p_{u,s}).$$

As x increases, the law of large numbers posits that the estimation $\hat{p}_{us}$ by $\pi(p_{u,s}|x)$ will equal the true value of $p_{u,s}$ with probability 1.

While it is possible for the player to have any likelihood function or prior probability distribution, the most intuitive pairing for this game is the Bernoulli likelihood function and the Beta distribution. The Bernoulli likelihood function and the Beta prior distribution work well in tandem since the Beta distribution is the conjugate prior distribution for the Bernoulli likelihood function. This means that the Beta and Bernoulli distributions are part of the same probability distribution family, which guarantees that the Bernoulli likelihood function ensures that the posterior probability distribution remains a Beta distribution. The Bernoulli likelihood function has binary outcomes with some probability of a successful outcome, which perfectly matches the reward structure of the game. The Beta distribution nicely captures a player's prior beliefs with two hyperparameters, α and β, and these hyperparameters can represent successful and unsuccessful unknown arm pulls, respectively; let $\pi(p_{us}) \sim Beta(\alpha_s, \beta_s)$, where $\alpha_s, \beta_s > 0$. So, for n pulls of the unknown arm and x successful pulls of the arm, the posterior distribution is $\pi(p_{us}|x) \sim Beta(\alpha_s + x, \beta_s + n - x)$. The

Beta distribution's estimation of the expected probability of the unknown arm for a given state, $\hat{p}_{us}$, is characterized by the distribution's mean, $\frac{\alpha_s+x}{\beta_s+n-x}$, and thus reflects the learning process.

One important feature of the Bayesian updating process in this game is that α_s and β_s are state dependent such that there are values of α and β for each state. These state dependent hyperparameters represent the player's prior beliefs about the true probabilities of the unknown arm for both states. The player's prior beliefs are not entirely arbitrary in this game. This paper assumes that the player anchors her prior beliefs about the probabilities of the unknown arm on the probabilities of the known arm, such that $\pi(p_{ug}) = 0.75$ and $\pi(p_{ub}) = 0.25$. Ambiguity aversion justifies this anchoring assumption; if a player is ambiguity averse, then she has a preference to choose the known arm over the unknown arm. Whenever the player chooses the unknown arm over the known arm, especially in the early stages of the game where information is low, the player's beliefs about the true probaibilities of the unknown arm for each state depend on the probabilities of the known arm for each state since the player's only point of reference for the unknown arm is the known arm. For an ambiguity averse player, therefore, there must be ample evidence that the unknown arm is better than the known arm in order for them to choose it. So, this player begins with the prior belief that the two arms are equal to reflect her preference for the known arm until there is proof–through Bayesian learning–that the unknown arm is better. Overall, the anchoring assumption creates highly conservative prior beliefs, relative to the alternative that the player has uniform beliefs (no information whatsoever) about the probabilities of the unknown arm for each state, such that $\alpha = \beta = 1$. While pure uniform prior beliefs are a reasonable assumption, they do not capture the nature of ambiguity

aversion. Given the above logic, the player's prior beliefs are expressed as:

$$\overline{\pi(p_{u,s})} = \frac{\alpha_s}{\alpha_s + \beta_s} = p_{k,s} \tag{2}$$

where, given the player's anchored beliefs about the unknown arm, $\alpha_g = 3\beta_g$ and $\alpha_b = \frac{1}{3}\beta_b$. For example, in the good state, the prior belief $\overline{\pi(p_{u,g})} = \frac{\alpha_g}{\alpha_g + \beta_g} = 0.75$. Through simplification, $\alpha_g = 0.75\alpha_g + 0.75\beta_g \Rightarrow 0.25\alpha_g = 0.75\beta_g \Rightarrow \alpha_g = 3\beta_g$. The same process applies for the bad state when $p_{k,b} = 0.25$.

At the beginning of the game, the player has no information about the probabilities of the unknown arm, and so her uncertainty is "maximized." To best model the player's initial prior beliefs and uncertainty simultaneously, the hyperparameters α_s and β_s must maximize the Beta distribution's entropy η given Equation 2 as a constraint:

$$\max(\eta_s) = \ln(B(k\beta_s, \beta_s) - (k\beta_s - 1)\psi(\beta_s) - (\beta_s - 1)\psi(\beta_s) + (k\beta_s + \beta_s - 2)\psi(k\beta_s + \beta_s) \tag{3}$$

where $B(\cdot)$ is the Beta function, $\psi(\cdot)$ is the Digamma function, and $\alpha_s = k\beta_s$ for a given state s. The entropy function represents the amount of information in the Beta distribution, and when entropy is maximized, there is essentially no information, and so uncertainty is at its highest. Without any constraints, the entropy function is maximized when $\alpha, \beta = 1$, which equates the Beta distribution to the Uniform distribution, where all events are equiprobable (the probability density function of the Uniform distribution is flat). Thus, given the constraint of the player's prior beliefs, the entropy function is maximized to find values of α_s and β_s that brings the Beta distribution as close as possible to the Uniform distribution. Since the values of p_{ks} and the relationships between

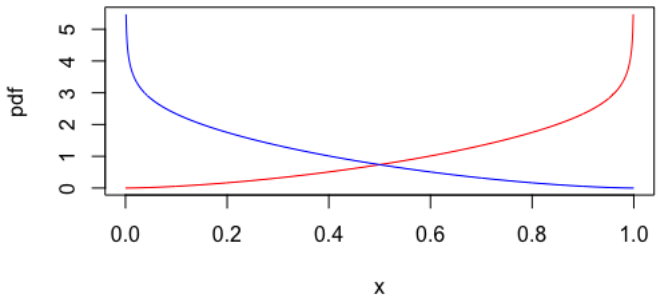


Figure 2: Probability density functions for the good (red) and bad (blue) states.

α_s, β_s for each state are defined, the entropy-maximizing hyperparameters are:

$$(\alpha_s, \beta_s) = \begin{cases} (2.55, 0.85) & s = g \\ (0.85, 2.55) & s = b \end{cases}.$$

Figure 2 graphs the entropy-maximizing probability density functions for each state.

The main caveat to this Bayesian updating process, within the context of this model, is that the player does not know when $\hat{p}_{us}$ is not only close enough to the true value of $p_{u,s}$ for each state, but also significantly different from p_{ks}. In other words, when does the player confidently know which arm is the best to pull? The answer to this question lies in a comparison of $p_{k,s}$ and the estimated value of $p_{u,s}$. For the Pseudo-Gittins and ϵ-Greedy strategies, only the strict comparison of $\hat{p}_{u,s}$ and $p_{k,s}$ affects which arm the player pulls. The Pure Unknown strategy, however, compares the probabilities of the known arm and estimated probabilities of the unknown arm by constructing a difference in

means confidence interval with endpoints $(\gamma_{1s}, \gamma_{2s})$:

$$(\gamma_{1s}, \gamma_{2s}) = (p_{k,s} - \hat{p}_{u,s}) \pm t^* \sqrt{\frac{Var(\hat{p}_{u,s})}{n_{p_{u,s}}}} \tag{4}$$

where t^* represents the t-value from a Normal distribution for a player's given confidence level and their posterior probability distribution.[3] Since the t-value reflects a player's degree of confidence, different t-values reflect varying degrees of ambiguity aversion. For example, a player with a low confidence interval ($\sim 95\%$) is not as ambiguity averse as a player with a high confidence interval ($\sim 99\%$) because the latter player needs to be more certain that exploration of the unknown arm is worthwhile. Section 4 not only explains why only the Pure Unknown strategy uses the confidence interval as part of its learning process, but also outlines the different strategies and data collection process. Therefore, the remainder of the paper explores several different strategies under two variations of the game: one where the current state is known and another where the current state is unknown. These are herein referred to as the Current State Certainty (CSC) and Current State Uncertainty (CSU) conditions, respectively. The strategies are compared within and between the CSC and CSU conditions to determine which strategy produces the most optimal solution, or best maximizes the player's objective function. It is important to note that the most optimal solution is not necessarily the equilibrium strategy; this is because there is no true equilibrium strategy in this game since the stochastic movement between states and the learning process built into the game preclude any true equilibrium strategy from existing.

[3]The term $\frac{Var(p_{ks})}{n_{p_{ks}}}$ is not included in the standard error calculation because the true value of p_{ks} is known, so $Var(p_{ks}) = 0$.

4 Analysis

Given the structure of the model, there are four different strategies to explore within the context of this game: the Pure Unknown strategy, the Pseudo-Gittens strategy, the ϵ-Greedy strategy, and the Softmax strategy. The Pure Unknown strategy is this paper's novel strategy, and for this strategy, the player only chooses the unknown arm until she confidently learns the true probabilities of the unknown arm for each state. The player learns the probabilities of the unknown arm in each state with the difference in means confidence interval from Section 3. This confidence interval yields the following strategy for the player.

Proposition 1:

1. If $0 \notin [\gamma_{1s}, \gamma_{2s}]$ and $(\gamma_{1s}, \gamma_{2s})$ is positive, then the player exploits the known arm for the remainder of the game in state s.

2. If $0 \notin [\gamma_{1s}, \gamma_{2s}]$ and $(\gamma_{1s}, \gamma_{2s})$ is negative, then the player exploits the unknown arm for the remainder of the game in state s.

The proof of Proposition 1 is simple. Start with condition (1). Suppose $p_{k,g} > \hat{p}_{u,g}$ and the margin of error is sufficiently small such that $(\gamma_{1g}, \gamma_{2g})$ is positive. This means that $p_{k,g} > \hat{p}_{u,g}$ at a statistically significant level, so the known arm is the better option and condition (1) is true. The opposite is true for condition (2) when the inequality switches and the range of the confidence interval is negative. Therefore, the Pure Unknown strategy differentiates itself from the other strategies by using the confidence interval to complete exploration of the unknown arm before the global optimal arm is realized and exploited. Importantly, the Pure Unknown strategy is the only strategy that has a built-in flavor of ambiguity aversion since the player needs more evidence than a simple

comparison of probabilities that the unknown arm is better than the known arm.

The Pseudo-Gittins strategy is a variant of the methodology created by Gittins (1979) to derive the Gittins Index. This strategy does not produce a Gittins Index; rather, it follows the basic methodology that obtains the Gittins Index, such that the current best arm is chosen in each time period. The delayed reward structure and state-dependent payoffs in this game also complicate how to define the current best arm. Depending on the current state and the stochastic movement between states for each time period, the current best arm in time period τ may not be the current best arm in $\tau + 1$ when the arm is actually pulled. In other words, the current best arm relies on the expectation of which arm has the highest expected payoff in the next time period given the Markov chain and the relationship between $p_{k,s}$ and $\hat{p}_{u,s}$. This requires that the player calculates $p(s_{\tau+1}|s_\tau)$ to determine which state is most likely in period $\tau + 1$, and then choose the arm with the highest probability of a successful pull given the expected state. Proposition 2 outlines the pure strategy solution for the Pseudo-Gittins strategy in the CSC condition.

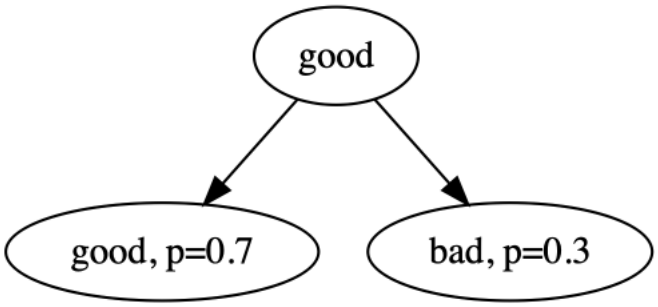
Proposition 2:

1. In state s, if $\hat{p}_{u,s} > p_{k,s}$, then the player chooses the unknown arm.

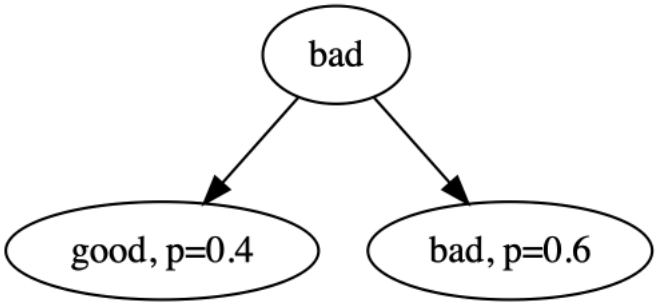
2. Otherwise, the player chooses the known arm.

3. $p(s_{\tau+1}|g_\tau) = \begin{cases} 0.7 & s = g \\ 0.3 & s = b \end{cases}$

4. $p(s_{\tau+1}|b_\tau) = \begin{cases} 0.4 & s = g \\ 0.6 & s = b \end{cases}$



(a) Starting in the good state.



(b) Starting in the bad state

Figure 3: Two period Markov chain movement for each state.

The proof for Proposition 2 depends on which state characterizes time period τ, and Figure 3 presents tree diagrams for each state to understand the stochastic movement of states between the time periods τ and $\tau + 1$. Since this strategy does not randomize between the selection of both arms, it is possible for the player to never learn the true probabilities of the unknown arm if she believes the current best arm is always the known arm. Consequently, if the player's local best choice is not the global best choice, then the player fails to maximize her payoff over the course of the game. The Pseudo-Gittins strategy also does

not require the confidence interval as part of its strategy because the player only cares about whether the posterior belief of the uknown arm in a given state is currently either greater or less than the corresponding probability of the known arm, and her expectations about which state will characterize the next period. In other words, there is no need for the posterior belief to be significantly different; rather, it must only be different.

The ϵ-Greedy strategy, which Tokic (2010) defines as a strategy where the current best arm is chosen with probability $1 - \epsilon$, is essentially the same as the Pseudo-Gittins strategy, except it allows for a small degree of randomization. Thus, ϵ-Greedy uses Proposition 2, but instead of choosing the current best arm with probability one, it is chosen with probability $1 - \epsilon$, where $\epsilon = 0.05$ in this game. On a more theoretical level, ϵ represents the probability that the player's belief about the current best arm are incorrect. Even if, for example, the player believes the current best arm is always the known arm for both states, she still learns about the probabilities of the unknown arm with $p = 0.05$. Consequently, suppose $p_{u,s} > p_{k,s}$ and the player correctly updates her beliefs about the probabilities of the unknown arm for n trials but incorrectly believes that $\hat{p}_{u,s} < p_{k,s}$ due to an unlucky sampling from the unknown arm. If the player uses the Pseudo-Gittins strategy, then she will always choose the known arm because she believes it is the current best arm, even though it is not. In this scenario, she always chooses the inferior arm and fails to explore the unknown arm; the ϵ-Greedy strategy lets the player explore the unknown arm–albeit infrequently–and over time learn that the unknown arm is actually the best current arm choice. Therefore, the ϵ-Greedy strategy addresses the main shortcoming of the Pseudo-Gittins strategy.

The final strategy is the Softmax strategy, which is the simplest of the four strategies. While the softmax strategy, according to Tokic (2010), allows $\epsilon \in$

$[0, 1]$, the Softmax strategy for this game can be thought of as the "softest" strategy because the player chooses each arm with probability $1/2$. The Softmax strategy is, consequently, not as extreme as the ϵ-Greedy strategy; the player does not treat either arms as the current best option since she does not yet know which arm is the best option.

The fact that there is no equilibrium solution to this game precludes any theoretical explanation for which strategy is optimal. At the same time, this two-armed bandit problem does not exist as one specific real-world scenario, so actual data cannot be collected. Consequently, this game is emulated by an alogrithm in R. The base algorithm for the game exactly follows the foundation in Section 3, and adjusts slightly for each strategy described earlier, as well as the CSC and CSU conditions. Other than the strategies and conditions, the main independent variables that change with each iteration of the algorithm are the two confidence interval significance levels for the confidence interval of the Pure Unknown strategy–$\alpha \in \{0.05, 0.01\}$–and the true values of $p_{u,g}$ and $p_{u,b}$, which are $p_{u,g} = \{0.9, 0.7, 0.5, 0.55\}$ and $p_{u,b} = \{0.5, 0.45, 0.3, 0.1\}$. To simulate all possible relationships between the probabilities of the unknown arm with those of the known arm, the different values of $p_{u,g}$ and $p_{u,b}$ are paired together as $(p_{u,g}, p_{u,b})$, such that there are 9 pairs. Since the model dictates that $p_{u,g} > p_{u,b}$, the pairing $(0.5, 0.5)$ is impossible and does not intuitively make sense–the good state should be strictly better than the bad state. Thus, to replace the $(0.5, 0.5)$ pairing, the $(0.55, 0.45)$ pairing stands in its place, and those probabilities are not used for any other pairing.

One iteration of this game is, for example, the Pure Unknown strategy with the $(0.9, 0.5)$ pairing at the 99% confidence level in the CSC condition. Figure 4 presents two tables as examples of the output. Given the different independent variables, there are a total of 63 iterations of the game to report.[4] Each iteration

[4]There are 18 iterations (9 unknown probability pairings and 2 confidence levels) for the

Pure Exploration 99% (0.9, 0.5)

Statistic	Mean	St. Dev.
n trials	500	0
payoff	392.556	11.017
p good	0.853	0.072
alpha good	321.222	40.187
beta good	36.196	6.923
p bad	0.415	0.108
alpha bad	70.17	6.705
beta bad	72.95	6.639

Pure Exploration 95% (0.9, 0.5)

Statistic	Mean	St. Dev.
n trials	500	0
payoff	393.892	9.322
p good	0.84	0.071
alpha good	324.352	25.116
beta good	36.548	6.207
p bad	0.394	0.126
alpha bad	70.952	6.903
beta bad	72.17	6.835

Figure 4: Output examples from Pure Unknown strategy.

of the game has a fixed trial length $n = 500$ and is run 500 times by the algorithm to satisfy the law of large numbers. All strategies also run over the same $n = 500$ trials, where the states are the same in each time period τ. After 500 trials, the Markov Chain does not reach its steady state, so in order to compare results for each strategy, the stochastic movement over 500 trials must be standardized. The randomly generated $n = 500$ standardized trial period for the results section has a "steady state" of $p(g) = 0.72$ and $p(b) = 0.28$. The dependent variables that each iteration outputs (a total of 8) are the length n of each trial, the final payoff, the final posterior probability estimates of the unknown arm $(\hat{p}_{u,g}, \hat{p}_{u,b})$, and the corresponding hyperparameters for each state (α_s, β_s).

5 Results

The results are presented in three main subsections: the CSC condition, the CSU condition, and then robustness checks. The CSC condition is the baseline

Pure Unknown strategy, and 9 iterations for the remaining 3 strategies. All 4 strategies are run in the CSC condition, and only 2 are run in the CSU condition (only the Pseudo-Gittins and ϵ-Greedy change in the CSU condition). This sums to 63 iterations.

condition, and the CSU condition is an extension of the CSC condition to consider how strategies and outcomes change when the player does not know what state she is currently in. The only strategies that actually change in the CSU condition are the Pseudo-Gittins and ϵ-Greedy strategy. How these strategies change is explained in detail in Section 5.2. Finally, the robustness checks are a brief investigation of the confidence interval for the Pure Unknown strategy and t^*. Thus far, the confidence interval assumes that t^* follows a Normal distribution instead of the Beta distribution. The justification for using a Normal distribution is twofold: it is more familiar and common than the Beta distribution, and the Central Limit Theorem suggests that large trials with random variables will resemble a Normal distribution. Section 5.3 explores the validity of this Normality assumption.

5.1 Current State Certainty Condition

To determine the best strategy, the averaged payoffs are compared to not only see which strategy has the highest (the strategy that best maximizes Equation 1), but also compare these strategies to the theoretical maximum payoff for the game. This theoretical maximum payoff is characterized by the following equation:

$$p(r^*) = p(g)p^*_{a,g} + p(b)p^*_{a,b} \tag{5}$$

where $p(s)$ represents the steady state probability for each state and $p^*_{a,s}$ is the probability of success for the best arm in state s. $p(r^*)$ is multiplied by 500 to get the maximum payoff over 500 trials. The best arm in each state could be all possible combinations of the unknown and known arms. Recall that the steady state over 500 trials is $p(g) = 0.72$ and $p(b) = 0.28$ for these results, which is not the true steady state. Table 1 summarizes the theoretical and actual payoffs from each strategy for each unknown probability pairing. The

$(p_{u,g}, p_{u,b})$	Theoretical	PU 99%	PU 95%	Pseudo-Gittins	ϵ-Greedy	Softmax
$(0.9, 0.5)$	394	392.56 (11.02)	393.89 (9.32)	355.15 (32.44)	378.65 (17.16)	349.56 (9.62)
$(0.9, 0.3)$	366	365.23 (10.45)	365.15 (8.57)	339.74 (28.10)	353.27 (13.80)	336.04 (8.69)
$(0.9, 0.1)$	359	336.97 (10.11)	336.46 (7.65)	336.43 (28.51)	349.73 (15.01)	320.69 (8.68)
$(0.7, 0.5)$	340	324.41 (11.87)	323.22 (11.67)	320.55 (19.96)	330.57 (14.66)	313.81 (10.07)
$(0.7, 0.3)$	312	295.84 (10.78)	294.51 (10.55)	305.01 (9.77)	305.83 (10.25)	299.68 (9.78)
$(0.7, 0.1)$	305	269.90 (14.59)	267.64 (12.40)	303.22 (9.68)	301.35 (8.95)	284.88 (10.19)
$(0.5, 0.3)$	312	228.14 (18.63)	225.96 (16.46)	304.27 (9.99)	302.46 (10.71)	263.42 (10.57)
$(0.5, 0.1)$	305	198.33 (18.79)	198.44 (18.87)	304.21 (9.20)	297.93 (9.69)	249.59 (10.05)
$(0.55, 0.45)$	333	266.40 (19.38)	264.49 (16.27)	314.33 (17.24)	319.33 (13.80)	282.14 (10.70)
Average	336.22	297.53 (13.96)	296.64 (12.42)	320.32 (18.32)	326.57 (12.67)	269.98 (9.82)

Table 1: Payoff comparisons between the theoretical maximum payoff and all strategies.

standard deviation for each average payoff is included in parentheticals. To show the effects of the strategies on the player's anchored prior beliefs, Figure 5 presents the player's posterior beliefs for the Pure Unknown strategy at the 99% confidence level for the unknown probability pairing $(0.55, 0.45)$. As the player learns about the true probabilities of the unknown arm for each state, Figure 5 shows the effectiveness of this learning process. Relative to the prior beliefs displayed by Figure 2, not only are the posterior beliefs highly accurate, but also more normally distributed.

The average payoff over all unknown pairings indicates that the ϵ-Greedy strategy has the highest expected payoff. While the Pseudo-Gittins strategy has a payoff that is only six successful arm pulls shy of the ϵ-Greedy's payoff, the difference is statistically significant at $\alpha < 0.01$, so the ϵ-Greedy strategy is definitively the best strategy in this game. There are cases in which the Pure Unknown strategy is either more optimal than or on par with both the Pseudo-Gittins and ϵ-Greedy strategies. These cases are the first four unknown probability pairings, where the Pure Unknown strategy is essentially optimal for the top two pairings and roughly equivalent with the other strategies for the latter two pairings. For the first two pairings, $(0.9, 0.5)$ and $(0.9, 0.3)$, the Pure

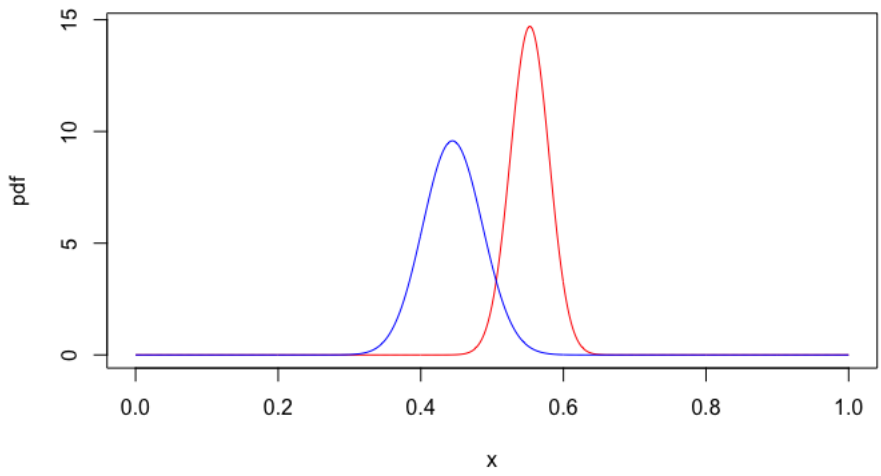


Figure 5: Posterior beliefs for the Pure Unknown strategy for the unknown probability pairing $(0.55, 0.45)$.

Unknown strategy is close to optimal because the unknown arm should always be chosen since it is the superior choice for both states; in other words, the Pure Unknown strategy strictly dominates these unknown probability pairings. The Pseudo-Gittins strategy does not match the Pure Unknown strategy in these cases because, through unlucky draws from the unknown arm, the player may believe that the known arm is the current best choice. Also, the ϵ-Greedy strategy does not match the Pure Unknown strategy because the known arm is chosen with probability ϵ. For the $(0.9, 0.1)$ and $(0.7, 0.5)$ unknown pairings, the dominance of the Pure Unknown breaks down because the unknown arm is no longer the best arm in both states, and the ϵ-Greedy strategy begins to stand out as the most optimal strategy.

As the known arm becomes the better option, the payoff for the Pure Unknown strategy lags behind the other strategies because of the extra trials needed by the player to determine that the unknown arm is inferior. These

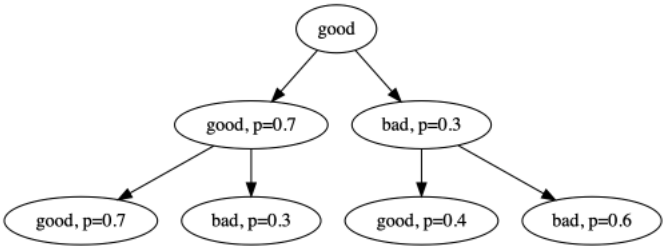
extra trials for the learning process reflect the ambiguity aversion built into the Pure Unknown strategy with the confidence interval, and there is no apparent difference in payoff between the 99% and 95% confidence levels. Given that the other strategies do not account for any degree of ambiguity aversion, it appears that ambiguity aversion does lead to a significant drop in payoffs. Additionally, payoffs do not change significantly between different levels of ambiguity aversion; rather, the existence of ambiguity aversion is the source of the payoff discrepancy for the Pure Unknown strategy and other strategies.

5.2 Current State Uncertainty Condition

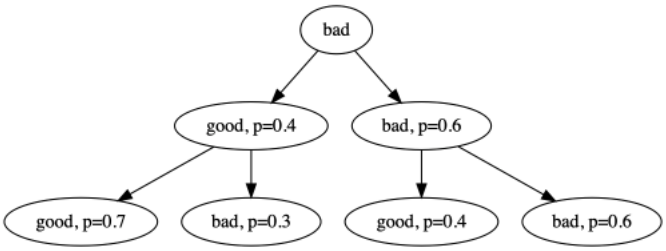
The CSU condition amends the game such that the player does not know which state she is currently in for time period τ, but knows $s_{\tau-1}$. This additional layer of uncertainty in the model only affects the Pseudo-Gittins and ϵ-Greedy strategies because these are the only strategies that select one of the arms by maximizing the expected payoff. Since the player still chooses the arm that has the highest expected payoff in time period $\tau + 1$, she must form expectations around $s_{\tau-1}$; the player now determines which state is most likely by calculating $p(s_{\tau+1}|s_{\tau-1})$. Proposition 3 outlines the new arm selection strategy in the CSU, and Figure 6 shows the three-stage Markov chain for each state.

Proposition 3:

1. If $\hat{p_{ug}} > p_{kg}$ and $\hat{p_{ub}} > p_{kb}$, then the player chooses the unknown arm.
2. If $\hat{p_{ug}} > p_{kg}$ and $\hat{p_{ub}} < p_{kb}$, then the player chooses the unknown arm.
3. If $\hat{p_{ug}} < p_{kg}$ and $\hat{p_{ub}} > p_{kb}$, then the player chooses the known arm.
4. If $\hat{p_{ug}} < p_{kg}$ and $\hat{p_{ub}} < p_{kb}$, then the player chooses the known arm.



(a) Starting in the good state.



(b) Starting in the bad state.

Figure 6: Three period Markov chain movement for each state.

$(p_{u,g}, p_{u,b})$	Theoretical	Pseudo-Gittins CSC	Pseudo-Gittins CSU	ϵ-Greedy CSC	ϵ-Greedy CSU
$(0.9, 0.5)$	394	355.15 (32.44)	358.45 (44.83)	378.65 (17.16)	380.84 (20.00)
$(0.9, 0.3)$	366	339.74 (28.10)	342.49 (32.41)	353.27 (13.80)	356.12 (15.24)
$(0.9, 0.1)$	359	336.43 (28.51)	324.52 (19.23)	349.73 (15.01)	333.62 (9.43)
$(0.7, 0.5)$	340	320.55 (19.96)	305.24 (9.58)	330.57 (14.66)	307.51 (10.82)
$(0.7, 0.3)$	312	305.01 (9.77)	305.31 (9.77)	305.83 (10.25)	303.34 (10.17)
$(0.7, 0.1)$	305	303.22 (9.68)	304.21 (9.50)	301.35 (8.95)	301.92 (9.55)
$(0.5, 0.3)$	312	304.27 (9.99)	304.54 (9.65)	302.46 (10.71)	300.31 (10.02)
$(0.5, 0.1)$	305	304.21 (9.20)	304.45 (9.48)	297.93 (9.69)	298.85 (10.06)
$(0.55, 0.45)$	333	314.33 (17.24)	304.49 (9.55)	319.33 (13.80)	302.44 (9.40)
Average	336.22	320.32 (18.32)	317.08 (17.11)	326.57 (12.67)	320.55 (11.63)

Table 2: Payoff comparisons between the CSC and CSU conditions

5. $p(s_{\tau+1}|g_{\tau-1}) = \begin{cases} 0.61 & s = g \\ 0.39 & s = b \end{cases}$

6. $p(s_{\tau+1}|b_{\tau-1}) = \begin{cases} 0.52 & s = g \\ 0.48 & s = b \end{cases}$

The proof for Proposition 3 follows the same logic as the proof for Proposition 2. The change in arm selection for Proposition 3 is due to the fact that the good state is always more likely to occur, regardless of $s_{\tau-1}$. Now that the arm selection is slightly different in the CSU condition, Table 2 compares the payoffs of the Pseudo-Gittins and ϵ-Greedy strategies in both the CSC and CSU conditions to the same theoretical maximum payoff. At first glance, the strategies have slightly higher payoffs in the CSC condition than in the CSU condition. In fact, these differences in payoffs are statistically significant for both strategies at $\alpha < 0.01$. Thus, the CSU condition produces significantly lower payoffs than the CSC condition for the Pseudo-Gittins and ϵ-Greedy strategies. This result matches basic intuition; the CSU condition forces the player to essentially predict the state of a time period that is two periods in the future instead of one. Given that this adds additional uncertainty to the model, it is no surprise that the player receives a lower payoff. Overall, the ϵ-Greedy strategy

remains the superior strategy, and at $\alpha < 0.01$, the payoff for the ϵ-Greedy strategy is significantly greater than that of the Pseudo-Gittins strategy in the CSU condition.

5.3 Robustness Checks

Thus far, the Pure Unknown strategy assumes a Normal distribution in its learning process because the confidence interval uses critical t^* values from the Normal distribution. While the justification for this Normality assumption is substantiated in Section 4, this subsection removes this assumption and maintains the same learning process for the Pure Unknown strategy. Theoretically, the Pure Unknown strategy does not change if a hypothesis test is used instead of a confidence interval because they run the same significance tests with slightly different methodologies. The hypothesis test for the Pure Unknown strategy, however, does not borrow anything from the Normal distribution; rather, this test maintains the Beta distribution and does not invoke the Central Limit Theorem. More formally:

$$H_\emptyset : \hat{p}_{u,s} = p_{k,s}$$

$$H_a : \hat{p}_{u,s} \neq p_{k,s}$$

The player chooses the unknown arm until she can accept the alternative hypothesis for a given state; at this point, the player knows which arm is better, and subsequently chooses that arm for the remainder of the game. This is still the Pure Unknown strategy, but now the hypothesis test replaces the confidence interval. The hypothesis test is also two-sided instead of one-sided because when the player pulls the unknown arm, she does not know whether–for a given state– the unknown arm is better or worse than the known arm. In other words, the player only needs to determine that the unknown arm is different than the known

$(p_{u,g}, p_{u,b})$	Theoretical	PU 99%	PU Robust 99%	PU 95%	PU Robust 95%
$(0.9, 0.5)$	394	392.56 (11.02)	393.56 (8.35)	393.89 (9.32)	393.53 (11.84)
$(0.9, 0.3)$	366	365.23 (10.45)	363.90 (12.69)	365.15 (8.57)	363.87 (13.26)
$(0.9, 0.1)$	359	336.97 (10.11)	349.75 (14.96)	336.46 (7.65)	350.81 (16.29)
$(0.7, 0.5)$	340	324.41 (11.87)	336.36 (9.97)	323.22 (11.67)	338.04 (9.84)
$(0.7, 0.3)$	312	295.84 (10.78)	310.74 (9.82)	294.51 (10.55)	310.38 (10.83)
$(0.7, 0.1)$	305	269.90 (14.59)	296.95 (9.49)	267.64 (12.40)	300.12 (9.28)
$(0.5, 0.3)$	312	228.14 (18.63)	305.88 (9.88)	225.96 (16.46)	307.16 (9.76)
$(0.5, 0.1)$	305	198.33 (18.79)	294.61 (9.94)	198.44 (18.87)	297.20 (9.91)
$(0.55, 0.45)$	333	266.40 (19.38)	322.69 (11.00)	264.49 (16.27)	326.67 (10.23)
Average	336.22	297.53 (13.96)	330.49 (10.68)	296.64 (12.42)	331.98 (11.25)

Table 3: Payoff comparisons between the Pure Unknown and PU Robust strategies.

arm, and since there is no information at the beginning of the game about the unknown arm relative to the known arm (ignoring the player's anchored beliefs), the hypothesis test must be two-sided.

The same data collection method from Section 4 applies to the new variant of the Pure Unknown strategy (herein referred to as the PU Robust strategy). Table 3 reports the average payoffs of the Pure Unknown and PU Robust strategies at the 99% and 95% levels. There is a striking and noticeable difference between the two versions of the Pure Unknown strategy. The PU Robust strategy is close to the optimal theoretical maximum of the game, and is, as a result, easily the best strategy in the entire game. If these two strategies are essentially the same–given that the confidence interval and hypothesis test are two sides of the same coin–then how is the PU Robust strategy so much better than the Pure Unknown strategy? The answer is the Normality assumption.

Since the PU Robust strategy is so close to the maximum expected payoff possible, the PU Robust strategy has a very quick and efficient learning process. Clearly, the confidence interval takes longer to learn the true probabilities of the unknown arm for each state. The main reason why the learning process for the confidence interval is longer than that of the hypothesis test is because the Normality assumption relies on the Central Limit Theorem, which

holds true only for large random variable samples. Consequently, the confidence interval requires that the sample size of these trials, which have a Beta distribution, is sufficiently large in order to become a Normal distribution to determine any statistical significance. On the other hand, the hypothesis test determines signficance with the cumulative distribution function of the Beta distribution, where there are no assumptions regarding either the data having a Normal distribution or a sufficiently large sample size. Therefore, though theoretically equivalent, the hypothesis test and confidence interval methodologies for the Pure Unknown strategy differ because the confidence interval makes a strong Normality assumption; the hypothesis test shows that the player learns faster and has the most optimal strategy when the learning process stays with the Beta distribution.

6 Conclusion

The results from Section 5 reveal that the Pure Unknown strategy is the most optimal strategy in this version of the multi-armed bandit problem, which suggests that this paper's original strategy is better at maximizing Equation 1 (a relatively simple utility maximization equation) than the strategies introduced by Gittins (1979) and Tokic (2010). At the same time, however, the two possible learning processes–that are at a theoretical level the same–for the Pure Unknown strategy produce different results. This, naturally, calls into question whether the Pure Unknown strategy is actually the most optimal strategy, and compels further discussion. The strength of the Normality assumption, which does account for the difference in payoffs between the confidence interval and hypothesis test, has already been discussed. What has not yet been addressed is the implications of these differences within the context of ambiguity aversion. Recall that ambiguity aversion reveals a preference for the known arm over the

unknown arm, and that in order to reverse that preference, the player must sufficiently learn that the unknown arm is superior to the known arm. Given this, Section 5.1 argues that the Pure Unknown strategy is not as optimal as the other strategies because the confidence interval characterizes ambiguity aversion and subsequently makes the learning process longer for the player. Section 5.3 to some extent nullifies the argument that the Pure Unknown strategy is more ambiguity averse than other strategies–assuming that ambiguity aversion leads to a longer learning process and, therefore, lower average payoffs–because the PU Robust strategy is overall the most optimal strategy. These two results, however, are not contradictory. Rather, the Normality assumption (the critical value in the confidence interval) is the source of ambiguity aversion within the Pure Unknown strategy because it elongates the learning process. Thus, if a player has ambiguity averse preferences, then she plays the Pure Unknown strategy; otherwise, she plays the PU Robust strategy.

While the previous literature often finds that the Gittins Index is consistently the most optimal across many variants of the multi-armed bandit problem, this paper finds that both the ϵ-Greedy and PU Robust strategies are significantly better in the context of this game. More importantly, the PU Robust strategy is the superior choice in this game; this suggests that a player should prioritize learning the unknown arm instead of choosing what she believes to be the current best arm, with the caveat that the learning process is fast enough. Additionally, the fact that the Pseudo-Gittins strategy is the third best strategy implies that a player suffers from situations in which she should have explored the other alternative options more instead of focusing on the local best choice. Intuitively, strategies that encourage more exploration of unknown alternatives–in a game that requires some degree of learning–should reward the player.

There are also several extensions to consider. First, as a type of stress test,

the Pure Unknown strategy should be applied to other variants of the multi-armed bandit problem to see whether it remains the most optimal strategy. One variant in which it seems likely that the Pure Unknown strategy begins to lag behind other strategies is if there are mutliple unknown arms available. The Pure Unknown strategy requires the player to learn the true probabilities of all unknown alternatives before the global best arm is chosen. As the number of uknown arms increases, so does the learning process for the Pure Unknown strategy. The differences in payoffs between the confidence interval and hypothesis test iterations of the Pure Unknown strategy are evidence of how a longer learning process decreases the viability of the Pure Unknown strategy. Another extension, more specific to this paper's game, is to remove the assumption that the player knows the state transition probabilities from the Markov chain. This creates an additional learning process for the player and overall increases uncertainty in the model. Given the lower payoffs from the CSU condition, the average payoffs with the transition probability uncertainty should be lower.

References

[1] Anderson, C. M. (2012). Ambiguity aversion in multi-armed bandit problems. Theory and decision, 72(1), 15-33.

[2] Auer, P., Cesa-Bianchi, N., Freund, Y., & Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. 36th Annual IEEE Symposium on Foundations of Computer Science (FOCS), 1-24.

[3] Choquet, G. (1955). Theory of Capacities. Ann. Inst. Fourier (Grenoble), 5, 131-295.

[4] Ellsberg, D. (1961). Risk, ambiguity, and the Savage axioms. The quarterly journal of economics, 643-669.

[5] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. Journal of the Royal Statistical Society. Series B (Methodological), 148-177.

[6] Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

[7] Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1), 4-22.

[8] Morgenstern, O., & Von Neumann, J. (1947). Theory of games and economic behavior, 2nd ed. Princeton: Princeton university press.

[9] Robbins, H. (1952). Some aspects of the sequential design of experiments. Bulletin of the American Mathematical Society, 58, 527-535.

[10] Savage, L.J. (1954). Foundations of statistics. Oxford, England: Wiley.

[11] Schmeidler, D. (1989). Subjective probability and expected utility without additivity. Econometrica: Journal of the Econometric Society, 571-587.

[12] Tokic, M. (2010). Adaptive ε-greedy exploration in reinforcement learning based on value differences. In Annual Conference on Artificial Intelligence. Springer, Berlin, Heidelberg. 203-210.