การเปรียบเทียบอัลกอริทึมระหว่างการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบน สำหรับการเรียนรู้แบบเสริมแรงในเกมเป่ายิ้งฉุบ

นายธันยวุฒิ อักขระสมชีพ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2565

A Comparison between Thompson Sampling and Upper Confidence Bound Algorithm for Reinforcement Learning in the Game of Rock-Paper-Scissor

Mr. Thanyavuth Akarasomcheep

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Statistics
Department of Statistics
FACULTY OF COMMERCE AND ACCOUNTANCY
Chulalongkorn University
Academic Year 2022

หัวข้อวิทยานิพนธ์	การเปรียบเทียบอัลกอริทึมระหว่างการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบน สำหรับการเรียนรู้แบบเสริมแรงในเกมเป่ายิ้งฉุบ
โดย	นายธันยวุฒิ อักขระสมชีพ
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(อาจารย์ ดร.สุรณพีร์ ภูมิวุฒิสาร)

..... กรรมการภายนอกมหาวิทยาลัย

(ผู้ช่วยศาสตราจารย์ ดร.ดลชัย ละออนวล)

ธันยวุฒิ อักขระสมชีพ : การเปรียบเทียบอัลกอริทึมระหว่างการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบน สำหรับการเรียนรู้แบบเสริมแรงในเกมเป่ายิ้งฉุบ. (A Comparison between Thompson Sampling and Upper Confidence Bound Algorithm for Reinforcement Learning in the Game of Rock-Paper-Scissor) อ.ที่ปรึกษาหลัก : รศ. ดร.เสกสรร เกียรติสุไพบูลย์

งานวิจัยนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบน ในตัวแบบการเรียนรู้แบบเสริมแรงกับการตัดสินใจเชิงพฤติกรรมของมนุษย์ ทั้งสองอัลกอริทึมเป็นอัลกอริทึมที่มีประสิทธิภาพในการแก้ไขปัญหาแบนดิทหลายแขน แต่ไม่ชัดเจนว่าทั้งสองอัลกอริทึมจะมีประสิทธิภาพอย่างไรกับปัญหาการตัดสินใจเชิงพฤติกรรมของมนุษย์ที่ความซับซ้อนทางด้านพฤติกรรม งานวิจัยนี้จำลองเกมเป่ายิ้งฉุบแทนปัญหาการตัดสินใจของมนุษย์ โดยมีองค์ประกอบเชิงพฤติกรรม 2 องค์ประกอบ คือ พฤติกรรมการใช้กลยุทธตามเข็มนาฬิกาแบบผสม และพฤติกรรมการใช้กลยุทธยุติการสูญเสีย โดยตัวแบบเกมเป่ายิ้งฉุบถูกจำลองขึ้นตามกระบวนการตัดสินใจแบบมาร์คอฟ ตัวแทนตัวแบบจากทั้งสองอัลกอริทึมจะแก้ไขปัญหาดังกล่าวและวัดประสิทธิภาพด้วยผลรางวัลสะสมภายใต้เงื่อนไขการจำลองในรูปแบบต่าง ๆ ผลการเปรียบเทียบประสิทธิภาพพบว่า ตัวแทนตัวแบบจากอัลกอริทึมความเชื่อมั่นขอบเขตบนมีประสิทธิภาพดีกว่าตัวแทนตัวแบบจากอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันในการจำลองส่วนใหญ่ ยกเว้นกรณีการจำลองที่รูปแบบพฤติกรรมของมนุษย์มีความชัดเจนเป็นระยะเวลายาว ตัวแทนตัวแบบจากอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันมีประสิทธิภาพดีกว่าตัวแทนตัวแบบจากอัลกอริทึมความเชื่อมั่นขอบเขตบน

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2565 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6280144626 : MAJOR STATISTICS

KEYWORD: Thomson Sampling, Upper Confidence Bound, Reinforcement Learning, Markov Decision Process, Rock-Paper-Scissors

Thanyavuth Akarasomcheep : A Comparison between Thompson Sampling and Upper Confidence Bound Algorithm for Reinforcement Learning in the Game of Rock-Paper-Scissor. Advisor: Assoc. Prof. SEKSAN KIATSUPAIBUL, Ph.D.

The purpose of this study is to compare the efficiency of the Thompson sampling algorithm and the upper confidence bound algorithm in reinforcement learning models for human behavioral decision making. Both algorithms are known of being efficient in solving multi-armed bandit problems. However, little is known how well those two algorithms perform when they encounter a behaviorally complex human decision problem. In this study, simulated rock-paper-scissors games represent human decision problems with two human behavioral traits, a mixed clockwise strategy and a stop loss strategy. The simulated rock-paper-scissors game is modeled as a Markov decision process. The two reinforcement learning agents are then applied to solve the decision process with their cumulative rewards as the performance measures. The performances of the two agents are measured under various simulation settings. The comparison results show that the upper confidence bound agent outperforms the Thompson sampling agent in most cases. The only exception is when there exists a strong behavioral pattern that persists over a long decision horizon where the Thompson sampling agent outperforms the upper confidence bound agent.

Field of Study: Statistics Student's Signature .....

Academic Year: 2022 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์ด้วยการสนับสนุนจากรองศาสตราจารย์ดร. เสกสรร เกียรติสุไพบูลย์ ในการสละเวลาให้คำปรึกษาและแนะนำสิ่งที่เป็นประโยชน์ต่อการวิจัย ช่วยแก้ไขและให้ความคิดเห็นเกี่ยวกับข้อบกพร่องในประเด็นต่าง ๆ ผู้วิจัยขอกราบขอบพระคุณเป็น อย่างสูงไว้ ณ ที่นี้

ผู้วิจัยขอกราบขอบพระคุณผู้ช่วยศาสตราจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ประธานกรรมการสอบวิทยานิพนธ์ อาจารย์ ดร. สุรณพีร์ ภูมิวุฒิสาร และผู้ช่วยศาสตราจารย์ ดร. ดลชัย ละออนวล กรรมการสอบวิทยานิพนธ์ ที่กรุณาสละเวลามาตรวจทานแก้ไขข้อบกพร่องในวิทยานิพนธ์ฉบับนี้ตลอดจนให้คำแนะนำที่เป็นประโยชน์แก่ผู้วิจัยที่ช่วยให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

ผู้วิจัยขอกราบขอบพระคุณคณะอาจารย์ภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัยทุกท่านที่กรุณาอบรมความรู้ทางคณิตศาสตร์ สถิติ และคอมพิวเตอร์ ทำให้ผู้วิจัยนำความรู้ที่ได้รับมาบูรณาการณ์และประยุกต์ใช้ในวิทยานิพนธ์ฉบับนี้ ตลอดทั้งเจ้าหน้าที่ภาควิชาสถิติทุกท่านที่ช่วยสนับสนุนในการดำเนินการด้านเอกสารและการประสานงาน

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา ครอบครัว และมิตรสหาย ที่คอยสนับสนุนผู้วิจัยทั้งด้านกำลังกายและกำลังใจในการทำวิจัยมาโดยตลอด จนวิทยานิพนธ์ฉบับนี้เสร็จสมบูรณ์

ธันยวุฒิ อักขระสมชีพ

สารบัญ

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ข้อมูลในยุคปัจจุบันมีปริมาณมหาศาล เติบโตเพิ่มขึ้นอย่างรวดเร็วและหลากหลาย การคำนวณด้วยวิธีทางสถิติแบบเดิมจะมีข้อจำกัดในการประมวลผลข้อมูลที่มีปริมาณมาก ทำให้ไม่สามารถใช้ข้อมูลที่มีให้เกิดประสิทธิภาพสูงสุดได้ จึงนำมาสู่กระบวนการการเรียนรู้ด้วยเครื่อง (Machine Learning) คือตัวแบบจะเรียนรู้จากข้อมูลที่ป้อนเข้าสู่ระบบและสร้างหลักการในการตัดสินใจขึ้นตามอัลกอริทึม ตัวแบบนั้นจะมีหลายประเภทแต่ทว่ายังมีข้อจำกัดในการตัดสินใจเกี่ยวกับข้อมูลที่มีองค์ประกอบเชิงพฤติกรรมของมนุษย์ เนื่องจากพฤติกรรมของมนุษย์เป็นสิ่งที่ซับซ้อน และผลลัพธ์จากการตัดสินใจกระทำมีความไม่แน่นอน คือแม้ว่าจะอยู่ในสภาพแวดล้อมเดียวกัน การกระทำแบบเดียวกัน ผลลัพธ์ที่ได้อาจแตกต่างกัน นอกจากนี้การตัดสินใจเลือกการกระทำถัดไปจะมีความเกี่ยวข้องกับการกระทำก่อนหน้าที่เคยทำตามลำดับเวลา ซึ่งสอดคล้องกับกระบวนการตัดสินใจแบบมาร์คอฟ (Markov Decision Processes) (Soo-Chang et al., 2005; Sutton & Barto, 1999, 2018)

การเรียนรู้ด้วยเครื่องประเภทการเรียนรู้แบบเสริมแรง (Reinforcement Learning) เป็นประเภทการเรียนรู้ด้วยเครื่องที่สามารถนำมาประยุกต์กับการแก้ปัญหากระบวนการตัดสินใจแบบมาร์คอฟข้างต้น คือตัวแบบจะคำนวณหานโยบายสำหรับใช้โต้ตอบกับการกระทำของมนุษย์เพื่อให้ได้ผลรางวัลสะสมสูงสุด โดยเรียนรู้จากสภาพแวดล้อม (Environment) การกระทำ (Action) ผลรางวัล (Reward) และสถานะ (State) การเรียนรู้แบบเสริมแรงจะสร้างนโยบายสำหรับกำหนดการตัดสินใจระหว่าง การเอาผลประโยชน์ (Exploit) และการสำรวจ (Explore) ซึ่งการกำหนดนโยบายสามารถทำได้จากหลายวิธีการ(Auer et al., 2002; Gordan & Krishanan, 2014; Sutton & Barto, 2018) เช่น วิธีการจับคู่ความน่าจะเป็น (Probability Matching) สามารถช่วยตัดสินใจเพื่อให้ได้ผลรางวัลสะสมสูงสุด โดยขึ้นอยู่กับการแจกแจงก่อนสังเกต (Prior Distribution) เช่น อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน (Thompson Sampling) (Chapelle & Li, 2011; Daniel et al., 2018) วิธีการคำนวณหาช่วงความเชื่อมั่นของผลรางวัลและเลือกช่วงความเชื่อมั่นที่ให้ผลรางวัลสูงสุดโดยไม่จำเป็นต้องทราบการแจกแจงก่อนการสังเกต เช่น อัลกอริทึมความเชื่อมั่นขอบเขตบน (Upper Confidence Bound) (Auer, 2002; Hao et al., 2019; Soo-Chang et al., 2005)

ตัวแบบจากอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน เป็นสองอัลกอริทึมที่มีประสิทธิภาพและได้รับความนิยมเมื่อนำมาแก้ไขปัญหาแบนดิทตามบริบท (Contextual bandits) โดยหากการตอบสนองที่เกิดจากการกระทำมีความล่าช้ามาก อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันจะให้ประสิทธิภาพที่ดีกว่าอัลกอริทึมความเชื่อมั่นขอบเขตบน (Chapelle & Li, 2011)

งานวิจัยนี้เป็นการศึกษาต่อยอดเพื่อเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบนในการแก้ไขปัญหาที่ซับซ้อนมากขึ้นกว่าปัญหาแบนดิทตามบริบท คือการแก้ไขปัญหาการเรียนรู้แบบเสริมแรงที่มีองค์ประกอบเชิงพฤติกรรมของมนุษย์ที่มีความซับซ้อนกว่า เพราะการตัดสินใจจะต้องนำสถานะถัดไปมาคำนึงถึงด้วย เพื่อให้ได้ผลรางวัลสะสมสูงสุดในระยะยาว โดยงานวิจัยนี้ทำการศึกษาเปรียบเทียบประสิทธิภาพของตัวแบบจากทั้งสองอัลกอริทึมในสภาพแวดล้อมต่าง ๆ โดยศึกษาภายใต้การจำลองสถานการณ์เกมเป่ายิ้งฉุบ (Rock Paper Scissors: RPS) ระหว่างตัวแทนตัวแบบกับผู้เล่น ซึ่งมีความแตกต่างจากตัวแทนตัวแบบอื่นในเกมเป่ายิ้งฉุบ คือตัวแทนตัวแบบโดยทั่วไปจะทำการเล่นกับผู้เล่นและจบเป็นเกม ในขณะที่ตัวแทนตัวแบบในงานวิจัยนี้จะอยู่ในสถานการณ์ที่ต้องเล่นกับผู้เล่นคนเดิมอย่างต่อเนื่อง และผู้เล่นมีกระบวนการตัดสินใจแบบมาร์คอฟ

สำหรับเกมเป่ายิ้งฉุบเป็นหนึ่งในวิธีจัดการปัญหาของมนุษย์เมื่อเกิดความขัดแย้งของการตัดสินใจที่แตกต่างกัน โดยที่ผู้เล่นทั้งสองฝ่ายจะเลือกตัวเลือกหนึ่งในสามอย่าง คือ ค้อน (Rock) หรือ กระดาษ (Paper) หรือ กรรไกร (Scissor) เพื่อหาผู้ชนะ และได้รับสิทธิในการกำหนดการตัดสินใจ (Hai-Jun, 2016; Mccannon, 2007) หากมีการเล่นเกมเป่ายิ้งฉุบอย่างต่อเนื่องและมีการมอบรางวัลให้ผู้ชนะในทุกครั้ง สามารถอธิบายพฤติกรรมนั้นได้ด้วยทฤษฎีการเสริมแรง คือเมื่อบุคคลกระทำในสิ่งที่ตรงกับเงื่อนไขความสำเร็จจะได้รับรางวัลตอบแทน การกระทำเช่นนี้ซ้ำ ๆ จะทำให้บุคคลดังกล่าวปรับพฤติกรรมไปในทิศทางที่จะได้รับรางวัลมากขึ้น โดยมีองค์ประกอบสำคัญ 3 ส่วน คือ สถานการณ์ (Environment) การกระทำ (Action) และผลลัพธ์ (Response) ซึ่งเป็นแรงจูงใจที่ทำให้เกิดพฤติกรรมนิยม (Behaviorism) คือพฤติกรรมที่ตอบสนองต่อสิ่งเร้าที่เคยได้เรียนรู้ไว้ ซึ่งในที่นี้คือถ้าชนะในการเป่ายิ้งฉุบจะได้รับรางวัลจากการเป็นผู้ชนะ (Gordan & Krishanan, 2014)

งานวิจัยนี้ทำการจำลองสถานการณ์การเล่นเกมเป่ายิ้งฉุบระหว่างตัวแทนตัวแบบกับมนุษย์ โดยจำลองผู้เล่น 1 คน เล่นเกมเป่ายิ้งฉุบแข่งขันกับตัวแทนตัวแบบ และกำหนดให้การตัดสินใจของผู้เล่นมีองค์ประกอบเชิงพฤติกรรม 2 องค์ประกอบดังนี้ กลยุทธตามเข็มนาฬิกาแบบผสม (Mixed

Clockwise Strategy) คือเมื่อผู้เล่นแพ้จะมีโอกาสที่ผู้เล่นจะใช้กลยุทธตามเข็มนาฬิกาในเกมถัดไป และกลยุทธยุติการสูญเสีย (Stop-Loss Strategy) คือเมื่อผู้เล่นแพ้ต่อเนื่องติดต่อกันจำนวนหนึ่งผู้เล่นจะเลิกเล่นเกมต่อไป เพื่อหลีกเลี่ยงไม่ให้มีการสูญเสียเพิ่มขึ้นอีก ทำให้ตัวแทนตัวแบบต้องทำนายพฤติกรรมการตัดสินใจของผู้เล่น เพื่อเลือกตัวเลือกในการเป่ายิ้งฉุบที่ชนะ และใช้เทคนิคจูงใจเพื่อให้ผู้เล่นยังเล่นเกมต่อไปเพื่อสะสมผลรางวัลจากการชนะให้ได้มากที่สุดจนกว่าผู้เล่นจะเลิกเล่นเกม ซึ่งเมื่อมีการเล่นอย่างต่อเนื่องหลายเกม พฤติกรรมของผู้เล่นจะมีลักษณะเป็นรูปแบบที่ชัดเจนมากขึ้นตามทฤษฎีการเสริมแรง

การศึกษาพฤติกรรมการเล่นเกมเป่ายิ้งฉุบของมนุษย์ ผู้ชนะส่วนใหญ่มีพฤติกรรมยึดติดกับการกระทำที่ทำให้ชนะ ดังนั้นหากเป่ายิ้งฉุบชนะแล้วในเกมถัดไปผู้ชนะมีแนวโน้มจะเลือกการกระทำแบบเดียวกับที่ชนะไปในเกมก่อนหน้า ในขณะที่ผู้แพ้ส่วนใหญ่มีพฤติกรรมอยากชนะในเกมที่ตนแพ้ ดังนั้นหากการเป่ายิ้งฉุบในเกมก่อนหน้าแพ้ ในเกมถัดไปผู้แพ้จะเลือกกการกระทำที่สามารถชนะฝ่ายตรงข้ามในเกมที่แล้ว เช่น ในเกมแรก นาย ก เลือกออกค้อน และนาย ข เลือกกรรไกร ในเกมที่สอง นาย ข จะเลือกกระดาษ เพราะสามารถชนะนาย ก ในเกมที่แล้ว ซึ่งการกระทำของฝั่งผู้แพ้ในรูปแบบดังกล่าวจะเป็นไปตามทิศทางของเข็มนาฬิกา จึงเรียกว่า กลยุทธตามเข็มนาฬิกา (Wang et al., 2014)

การศึกษาพฤติกรรมการใช้กลยุทธยุติการสูญเสียของนักลงทุน เมื่อนักลงทุนตัดสินใจผิดพลาดและสูญเสียอย่างต่อเนื่อง จะมีการตัดสินใจหยุดลงทุนเพื่อหลีกเลี่ยงไม่ให้เกิดความสูญเสียมากขึ้นไปอีก ซึ่งเรียกว่า กลยุทธยุติการสูญเสีย (Kaminski & Lo, 2014)

การเปรียบเทียบประสิทธิภาพระหว่างอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบนกับปัญหาดังกล่าวจะเปรียบเทียบด้วยผลรางวัลสะสมจากการเล่นเกมเป่ายิ้งฉุบ และเปรียบเทียบกับอัลกอริทึมบรรทัดฐานที่เป็นอัลกอริทึมการสุ่มแบบเอกรูปบนขอบเขตตัวเลือกที่ประกอบด้วย ค้อน กรรไกร กระดาษ ซึ่งโครงการวิจัยนี้นอกจากการที่แต่ละอัลกอริทึม (ยกเว้นอัลกอริทึมที่เป็นบรรทัดฐาน) จะต้องเรียนรู้พฤติกรรมของผู้เล่นในเกมเป่ายิ้งฉุบแล้ว จะต้องเรียนรู้การใช้เทคนิคในการหลอกล่อเพื่อจูงใจผู้เล่นให้เล่นอย่างต่อเนื่องและสะสมผลชนะรวมที่มากกว่า เพราะหากเรียนรู้พฤติกรรมผู้เล่นได้และตัดสินใจเลือกการกระทำที่ชนะผู้เล่นอย่างต่อเนื่อง อาจทำให้ผู้เล่นเลิกเล่นและทำให้เสียโอกาสในการสะสมผลรางวัลรวมในระยะยาว

1.2 วัตถุประสงค์การวิจัย

เพื่อศึกษาและเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้แบบเสริมแรง ระหว่างอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน สำหรับการตรวจจับและโต้ตอบกับมนุษย์ที่มีองค์ประกอบเชิงพฤติกรรมในการตัดสินใจในแต่ละสภาพแวดล้อม โดยจำลองด้วยเกมเป่ายิ้งฉุบ

1.3 ขอบเขตงานวิจัย

โครงการวิจัยนี้ใช้การจำลองข้อมูลภายใต้สถานการณ์การเล่นเกมเป่ายิ้งฉุบระหว่างผู้เล่นกับตัวแทนตัวแบบซึ่งเป็นคอมพิวเตอร์
กำหนดเงื่อนไขดังนี้

1. ผู้เข้าร่วมการแข่งขัน 1 คน
   1.1. เลือกการกระทำแบบสุ่มแบบเอกรูป (uniform random) บน $A = \{R, P, S\}$ เมื่อชนะในเกมก่อนหน้า
   1.2. เลือกการกระทำแบบกลยุทธตามเข็มนาฬิกา (clockwise strategy) เมื่อแพ้ในเกมก่อนหน้า ด้วยความน่าจะเป็น $c \in \{0.25, 0.75\}$
2. คอมพิวเตอร์ 1 เครื่อง
   2.1. เลือกการกระทำด้วยการสุ่มแบบเอกรูปบน $a = \{r, p, s\}$
   2.2. เลือกการกระทำด้วยการเรียนรู้แบบเสริมแรง อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน
   2.3. เลือกการกระทำด้วยการเรียนรู้แบบเสริมแรง อัลกอริทึมความเชื่อมั่นขอบเขตบน
3. เกมการแข่งขันจะจบลงเมื่อ
   3.1. ผู้เล่นแพ้ต่อเนื่องติดต่อกัน $l \in \{5,10\}$ เกม หรือ
   3.2. เล่นจนครบ 100 เกม
4. จำนวนรอบการจำลองสถานการณ์การเล่นเกมเป่ายิ้งฉุบในแต่ละสภาพแวดล้อม
   4.1. รอบการจำลองการเล่นเกมเป่ายิ้งฉุบ 2,000 รอบ (episode)

1.4 นิยามสัญลักษณ์

t คือจุดเหตุการณ์ t โดยที่ T คือจุดเหตุการณ์สุดท้าย และ $t < T$

A คือเซ็ตของการกระทำของผู้เล่น โดยที่ $A = \{R, P, S\}$

A_t คือการกระทำของผู้เล่น ณ จุดเหตุการณ์ t โดยที่ $A_t \in A$

a คือเซ็ตของการกระทำของคอมพิวเตอร์ โดยที่ $a = \{r, p, s\}$

$a_t = k$ คือการกระทำของตัวแทนตัวแบบ ณ จุดเหตุการณ์ t โดยที่ $a_t \in a$

l คือจำนวณครั้งที่ผู้เล่นแพ้ต่อเนื่องติดต่อกันแล้วจะใช้กลยุทธยุติการสูญเสีย $l \in \{5,10\}$

l_t คือจำนวณครั้งที่ผู้เล่นแพ้ต่อเนื่องติดต่อกัน ณ จุดเหตุการณ์ t

i คือปัจจัยแวดล้อม ณ จุดเหตุการณ์ t ประกอบไปด้วย (a_t, A_t, l_t)

j คือปัจจัยแวดล้อม ณ จุดเหตุการณ์ $t+1$ ประกอบไปด้วย $(a_{t+1}, A_{t+1}, l_{t+1})$

j' คือปัจจัยแวดล้อมทุกรูปแบบ ยกเว้นรูปแบบ j ณ จุดเหตุการณ์ $t+1$

$s_t = (t, i)$ คือสถานะ (state) ณ จุดเหตุการณ์ t ด้วยปัจจัยแวดล้อม i

w_t คือผลจากการกระทำของคอมพิวเตอร์ ณ จุดเหตุการณ์ t โดยที่ $w_t \in \{win, lose, draw\}$

$r_i^k(t)$ คือรางวัลที่ได้รับเมื่อกระทำ k ที่สถานะ (t, i)

c คือความน่าจะเป็นที่ผู้เล่นจะใช้กลยุทธตามเข็มนาฬิกา $c \in \{0.25, 0.75\}$

$P_{ij}^k(t)$ คือความน่าจะเป็นที่การกระทำ k นำ สถานะ (t, i) ไปสู่สถานะ $(t+1, j)$

π คือนโยบาย (policy) ประกอบด้วย (t, i)

$\pi(t, i)$ คือการตัดสินใจด้วยนโยบายที่สถานะ (t, i)

$v^{\pi}(t, i)$ คือฟังก์ชันคุณค่า (value function) ของนโยบาย π คือรางวัลทั้งหมดที่คาดว่าจะได้รับตั้งแต่สถานะ (t, i) เมื่อใช้นโยบาย π ที่จุดเหตุการณ์ t จนถึง T

$v^*(t, i)$ คือผลรางวัลทั้งหมดที่คาดว่าจะได้รับมากที่สุด (optimal value) ตั้งแต่สถานะ (t, i) ที่จุดเหตุการณ์ t จนถึง T

$Q_t^*(s,a)$ คือคุณค่าจากการกระทำ (Action-Values) a_t ที่สถานะ s_t

$\hat{Q}_t(s,a)$ คือค่าประมาณคุณค่าจากการกระทำ

θ_k คือความน่าจะเป็นในการเลือกการกระทำ k ของตัวแทนตัวแบบโดยที่ $\theta_k \in [0,1]$

n_k คือจำนวนครั้งที่กระทำ k ตั้งแต่ 0 ถึง t

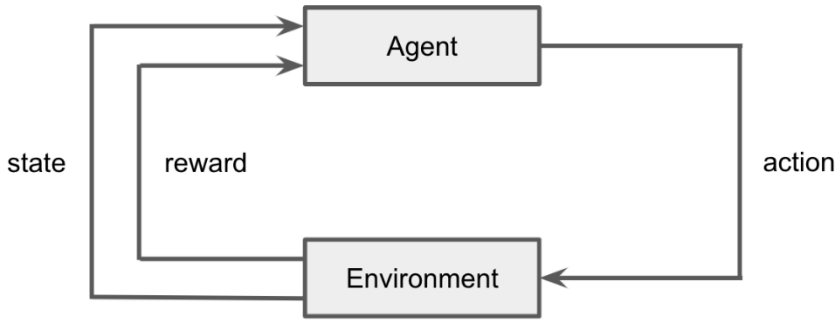
γ คือค่าถ่วงน้ำหนักในการเลือกสำรวจ (explore) งานวิจัยนี้กำหนดค่าเป็น 1

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้เป็นการศึกษาประสิทธิภาพระหว่างตัวแบบการเรียนรู้แบบเสริมแรงที่ได้รับความนิยม 2 ตัวแบบ ในการแก้ไขปัญหาที่มีองค์ประกอบพฤติกรรมของมนุษย์มาเกี่ยวข้อง โดยที่สถานะและผลจากการกระทำก่อนหน้าจะส่งผลต่อการตัดสินใจเลือกการกระทำถัดไป ซึ่งประกอบด้วยทฤษฎีดังต่อไปนี้

2.1 การเรียนรู้แบบเสริมแรง (Reinforcement Learning)

การเรียนรู้แบบเสริมแรงเป็นการเรียนรู้จากการกระทำ (action) และผลรางวัล (reward) ที่เกิดขึ้นของตัวแทนตัวแบบ (agent) ที่กระทำกับสภาพแวดล้อม (environment) โดยเมื่อตัวแทนตัวแบบกระทำไปยังสภาพแวดล้อมที่สถานะ (state) ปัจจุบัน จะส่งผลให้ได้รับรางวัลจากการกระทำนั้น และสภาพแวดล้อมจะมีสถานะเปลี่ยนไปสู่สถานะถัดไป จากนั้นตัวแทนตัวแบบจะเรียนรู้สถานะใหม่และรางวัลที่เกิดขึ้นจากการกระทำและกำหนดขึ้นเป็นนโยบาย (policy) สำหรับใช้ในการตัดสินใจเลือกการกระทำเพื่อใช้ในการตัดสินใจเลือกการกระทำถัดไป โดยที่การกระทำ ผลรางวัล และสถานะทั้งหมดจะเป็นไปตามลำดับเวลาของเหตุการณ์ที่เกิดขึ้น (Sutton & Barto, 2018)



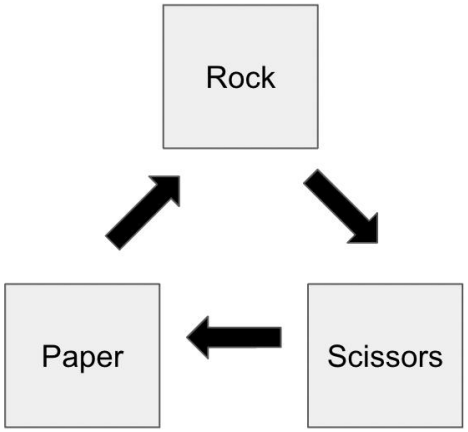
รูปที่ 2.1 รูปแบบของการเรียนรู้แบบเสริมแรง

2.2 กลยุทธตามเข็มนาฬิกา (Clockwise Strategy)

จากการศึกษาพฤติกรรมของผู้เล่นเกมเป่ายิ้งฉุบพบว่า ผู้ชนะมีพฤติกรรมยึดติดกับการชนะของตนเองและเลือกการกระทำแบบเดิมในเกมถัดไป ในขณะที่ผู้แพ้จะเลือกการกระทำที่สามารถชนะ

คู่แข่งในเกมที่แพ้ไปก่อนหน้า ซึ่งจะการกระทำจะเวียนในทิศทางตามเข็มนาฬิกาดังรูปที่ 2.2 จึงเรียกว่ากลยุทธตามเข็มนาฬิกา (Wang et al., 2014)

ตัวอย่าง นาย ก เลือกค้อน และนาย ข เลือกกรรไกร ในเกมนี้นาย ข แพ้ ดังนั้นด้วยกลยุทธตามเข็มนาฬิกาในเกมถัดไป นาย ข จะเลือก กระดาษ เพราะสามารถชนะนาย ก ในเกมที่แล้ว



รูปที่ 2.2 กลยุทธตามเข็มนาฬิกา

2.3 กลยุทธยุติการสูญเสีย (Stop-Loss Strategy)

กลยุทธยุติการสูญเสียคือหากมีการตัดสินใจเลือกการกระทำและการกระทำนั้นส่งผลให้ผลรางวัลลดลง หากมีเลือกการกระทำใหม่และผลรางวัลยังคงลดลงเรื่อย ๆ ติดต่อกันเป็นระยะเวลาหนึ่ง กลยุทธยุติการสูญเสียจะเป็นการหยุดตัดสินใจและหยุดการกระทำต่อไปเพื่อป้องกันไม่ให้ตัดสินใจผิดพลาดมากขึ้น

จากการศึกษาพฤติกรรมการใช้กลยุทธยุติการสูญเสียของนักลงทุน ผู้ที่มีการตัดสินใจผิดพลาดติดต่อกันจะใช้นโยบายยุติการสูญเสีย เพราะเป็นกลยุทธที่ช่วยในการป้องกันจากการตัดสินใจที่ผิดพลาด (Kaminski & Lo, 2014)

ในการเล่นเกมเป่ายิ้งฉุบผู้เล่นทั้งสองฝ่ายจะได้ผลรางวัลต่าง ๆ คือ ชนะ แพ้ หรือ เสมอ หากผู้เล่นคนหนึ่งแพ้อย่างต่อเนื่องก็จะมีการใช้กลยุทธยุติการสูญเสีย คือเลิกเล่นเกมถัดไป

2.4 กระบวนการตัดสินใจแบบมาร์คอฟ (Markov Decision Processes)

กระบวนการตัดสินใจแบบมาร์คอฟคือการสร้างนโยบายการตัดสินใจที่ให้ผลรางวัลได้ดีที่สุด (Optimal Policy) ในสถานการณ์ที่ผลลัพธ์เป็นไปแบบสุ่ม คือการกระทำแบบเดียวกันที่สถานะเดียวกันอาจได้ผลลัพธ์ที่แตกต่างกัน และเหตุการณ์ที่เกิดขึ้นเป็นไปตามลำดับเวลา เมื่อได้นโยบายแล้วตัวแทนตัวแบบจะตัดสินใจเลือกการกระทำตามนโยบายนั้น

ณ สถานะ $s_t = (t, i)$ เมื่อกระทำ k จะได้รับรางวัล $r_i^k(t)$ และนำไปอยู่ในสถานะ $s(t+1, j)$ ด้วยความน่าจะเป็น $P_{ij}^k(t)$ ดังแสดงในรูปที่ 2.3 และการตัดสินใจเลือกกระทำ k จะตัดสินใจด้วยนโยบาย $\pi(t, i)$ ซึ่งมาจากการคำนวณภายใต้เงื่อนไขของสมการ (2.1)

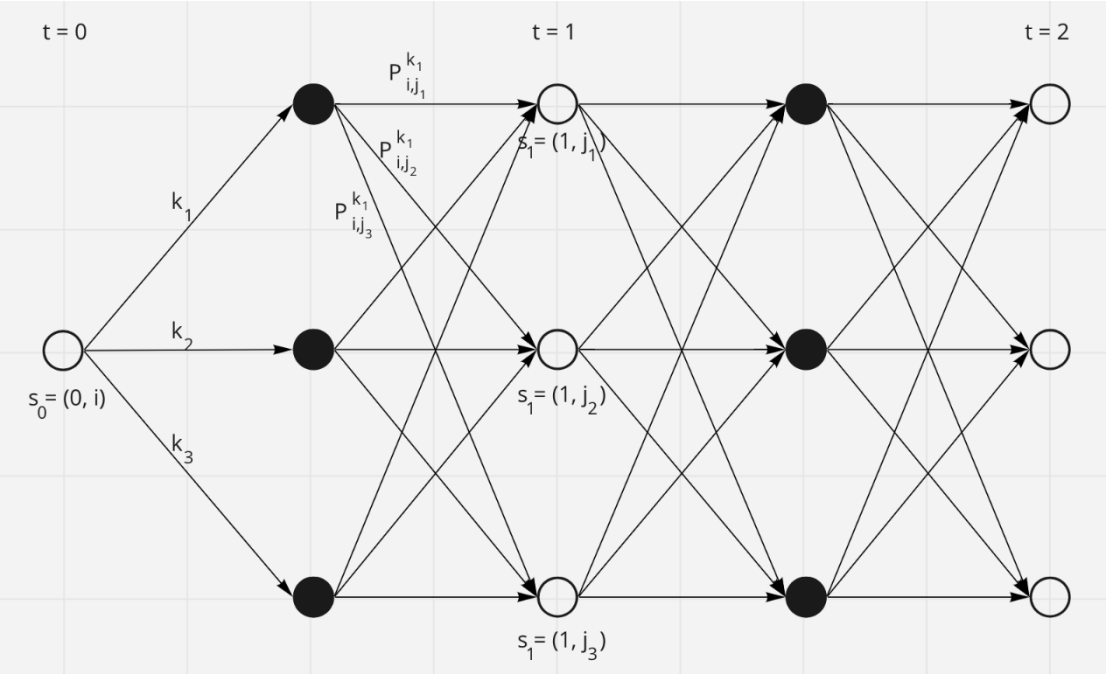
$$\sum_j P_{ij}^k(t) = 1 \text{ เมื่อรวมทุก } (t+1, j) \tag{2.1}$$

นโยบายการตัดสินใจ π จะได้ผลตอบแทนทั้งหมดที่คาดว่าจะได้รับตั้งแต่สถานะ (t, i) ไปจนถึงจุดเหตุการณ์สุดท้าย T แทนด้วย $v^\pi(t, i)$ และคำนวนได้ตามสมการเบลแมน (Bellman equation) ตามสมการ (2.2) (2.3) (2.4) (Sutton & Barto, 2018)

$$v^\pi(T, i) = r_i^k(T) \tag{2.2}$$

$$v^\pi(t, i) = r_i^k(t) + \sum_j P_{ij}^k(t)\, v^\pi(t+1, j) \text{ โดยที่ } k = \pi(t, i) \text{ ตั้งแต่ } t \text{ ถึง } T \tag{2.3}$$

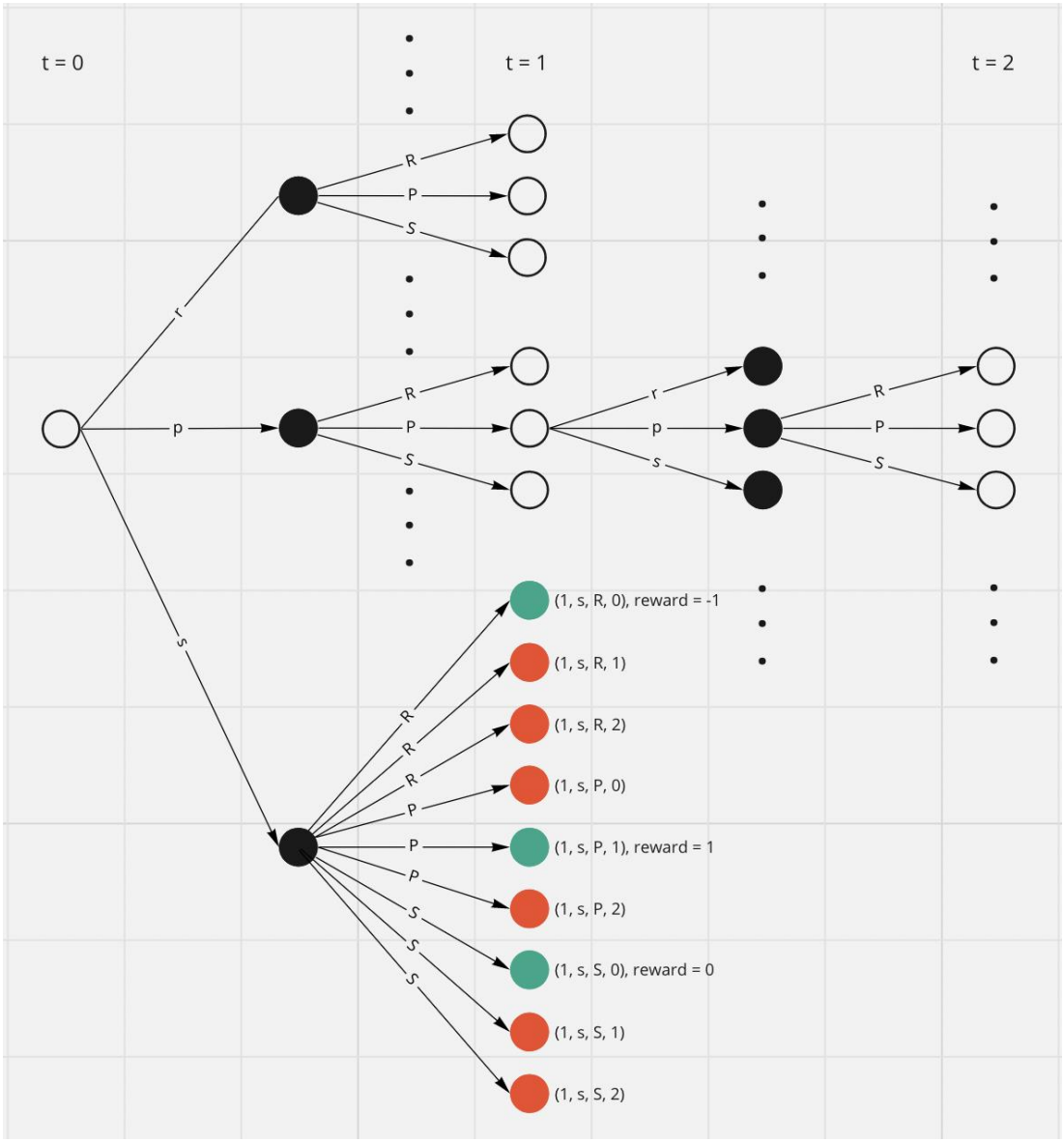
$$v^*(t, i) = max_k\{r_i^k(t) + \sum_j P_{ij}^k(t)\, v^*(t+1, j)\} \text{ ตั้งแต่ } t \text{ ถึง } T \tag{2.4}$$



รูปที่ 2.3 ตัวแบบกระบวนการตัดสินใจแบบมาร์คอฟ

ในงานวิจัยนี้กำหนดให้ สถานะ $s_t = (t, a_t, A_t, l_t)$ หากกำหนดเงื่อนไขการใช้กลยุทธยุติการสูญเสีย $l = 2$ จะได้ว่า $l_t \in \{0, 1, 2\}$ จำนวนสถานะของ s_{t+1} คือ 3 รูปแบบการกระทำของตัวแทนตัวแบบ × 3 รูปแบบการกระทำของผู้เล่น × 3 รูปแบบจำนวณครั้งที่ผู้เล่นแพ้ต่อเนื่องติดต่อกัน เป็นสถานะทั้งหมดคือ 27 สถานะ แต่สถานะถัดไป s_{t+1} ที่เป็นไปได้จะแตกต่างกันโดยขึ้นกับค่าของ l_t ว่าเป็นเท่าไร เช่น

สถานะ $\mathbf{s_t}$ มีค่าของ $l_t = 0$ ที่สถานะ s_{t+1} ค่าของ l_t สามารถเป็นได้เพียง 0 หรือ 1 เท่านั้น ขึ้นอยู่กับว่าที่สถานะ s_t ผู้เล่นแพ้ ชนะ หรือ เสมอ จะได้ว่า $l_{t+1} \in \{0, 1\}$ และไม่มีความเป็นไปได้ที่ $l_{t+1} = 2$ ดังแสดงในรูปที่ 2.4 สถานะที่เป็นไปได้คือจุดสีเขียว สถานะที่เป็นไปไม่ได้คือจุดสีแดง เมื่อตัวแทนตัวแบบเลือกการกระทำ $a_t = s$



รูปที่ 2.4 รายละเอียดของสถานะของตัวแบบ

2.5 อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน (Thompson Sampling Algorithm)

อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันเป็นอัลกอริทึมในการแก้ปัญหาแบนดิทหลายแขน โดยในเกมเป่ายิ้งฉุบจะให้แต่ละตัวเลือกคือแต่ละแขนของแบนดิท ตัวแทนตัวแบบจะทำการเลือกการกระทำ a_t ที่สถานะ s_t และได้รับผล w_t ด้วยความน่าจะเป็น θ_{a_t} และสังเกตผลรางวัล r_t ตามสมการ (2.5)

$$r_t = r(w_t) = \begin{cases} 1 & ; w_t = win \\ 0 & ; w_t = draw \\ -1 & ; w_t = lose \end{cases} \quad (2.5)$$

ถ้า θ เป็นเซ็ตจำกัด ค่าของ $\hat{\theta}$ จะสุ่มค่าจาก P โดยการแจกแจงของ P จะถูกอัพเดทด้วยเงื่อนไขจากการสังเกต $\widehat{w}_t$ ซึ่งเงื่อนไขการแจกแจงจะอธิบายได้ด้วยกฎของเบย์ (Bayes rule) ตามสมการที่ (2.6)

$$P(u = k|v = w_t) = \frac{P(u)P(v|u)}{P(v)} = \frac{\theta_k \, P_{ij}^k(t)}{\sum_a \theta_a P_{ij'}^a(t)} \tag{2.6}$$

อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน

1: $win \leftarrow 0$

2: $not\_win \leftarrow 0$

3: **for** t = 1 to T **do**

4: $\forall a: \theta_{a_t} \sim beta(win + 1, not\_win + 1)$

5: $a_t \leftarrow argmax_a \{\theta_{a_t}\}$

6: $w_t \leftarrow$ action a_t

7: **if** $w_t = win$: $win \leftarrow win + 1$

8: **else**: $not\_win \leftarrow not\_win + 1$

9: **end for**

(Daniel et al., 2018)

2.6 อัลกอริทึมความเชื่อมั่นขอบเขตบน (Upper Confidence Bound Algorithm)

อัลกอริทึมความเชื่อมั่นขอบเขตบนเป็นอัลกอริทึมในการแก้ปัญหาแบนดิทหลายแขน โดยในเกมเป่ายิ้งฉุบจะให้แต่ละตัวเลือกคือแต่ละแขนของแบนดิท ค่าความเชื่อมั่นของแต่ละตัวเลือกจะแตกต่างกันในแต่ละสถานะ และค่าความเชื่อมั่นของแต่ละตัวเลือกจะลู่เข้าเมื่อมีการเลือกตัวเลือกนั้นซ้ำ ๆ โดยการตัดสินใจเลือกตัวเลือกที่ให้ผลรางวัลที่คาดหวังสูงที่สุด สามารถดูจากตัวเลือกที่มีค่าความเชื่อมั่นขอบเขตบนสูงที่สุด

การคำนวณหาคุณค่าจากการกระทำ (Action-Values) $Q_t^*\ (s,a)$ ทำได้โดยดูจากรางวัลที่ได้รับเมื่อกระทำ a_t ที่สถานะ s_t โดยคำนวณตามฟังก์ชันคุณค่าของการกระทำ (Action-Value Function) ตามสมการที่ (2.7)

$$Q_t^*(s,a)\ = \mathrm{E}[r_t \mid a_t] \tag{2.7}$$

เนื่องจากไม่ทราบค่า $Q_t^*(s,a)$ จึงใช้การค่าประมาณ $\hat{Q}_t(s,a)$ โดยคำนวณจากค่าเฉลี่ยตัวอย่าง (sample mean) ตามสมการ (2.9)

$$\hat{Q}_t(s,a) = \frac{\sum_t (r_t|a_t)}{n_k} \tag{2.9}$$

การเลือกการกระทำที่มีค่าความเชื่อมั่นขอบเขตบนสูงที่สุด คำนวณตามสมการ (2.10)

$$a_t = argmax_a \left\{ \hat{Q}_t(s,a)\ + \gamma \sqrt{\frac{ln(t)}{n_k}} \right\} \tag{2.10}$$

อัลกอริทึมความเชื่อมั่นขอบเขตบน

1: $\forall a$: $\hat{Q}_t(s,a) \leftarrow 0$

2: **for** t = 1 to T **do**

3: $a_t \leftarrow argmax_a \left\{ \hat{Q}_t(s,a) + \sqrt{\frac{ln(t)}{n_k}} \right\}$

4: $w_t \leftarrow$ action a_t

5: $n_k \leftarrow n_k + 1$

6: $\hat{Q}_t(s,a) \leftarrow \frac{\sum_t (r_t(w_t)|a_t)}{n_k}$

7: **end for**

(Auer, 2002; Hao et al., 2019)

2.7 อัลกอริทึมการสุ่มแบบเอกรูป (Uniform Random Algorithm)

อัลกอริทึมการสุ่มแบบเอกรูปจะถูกใช้เป็นตัวแบบบรรทัดฐาน (baseline model) สำหรับการเปรียบเทียบประสิทธิภาพของอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน

อัลกอริทึมการสุ่มแบบเอกรูปมีคุณสมบัติคือตัวแทนตัวแบบจะตัดสินใจเลือกการกระทำจากการะทำที่เป็นไปได้ทั้งหมดโดยการสุ่มแบบเอกรูป ซึ่งการกระทำแต่ละอย่างที่เป็นไปได้จะมีความน่าจะเป็นเท่ากัน และเป็นการสุ่มแบบอิสระไม่ขึ้นกับผลแพ้ชนะของเกมก่อนหน้า หรือสถานะใดๆ

อัลกอริทึมการสุ่มแบบเอกรูป

1: **for** t = 1 to T **do**

2: $a_t \sim uniform(a)$

4: **action** a_t

3: **end for**

บทที่ 3
การดำเนินงานวิจัย

3.1 วิธีการจำลองข้อมูล

การจำลองข้อมูลจะจำลองการตัดสินใจเลือกการกระทำของผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบ 1 คน ในสภาพแวดล้อมต่าง ๆ คือ

ผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบไม่มีรูปแบบพฤติกรรมในการตัดสินใจ $c = 0$

ผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบมีรูปแบบพฤติกรรมในการตัดสินใจเพียงเล็กน้อย $c = 0.25$

ผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบมีรูปแบบพฤติกรรมในการตัดสินใจอย่างชัดเจน $c = 0.75$

ผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบมีรูปแบบพฤติกรรมระยะเวลาสั้น (เลิกเล่นเร็ว) $l = 5$

ผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบมีรูปแบบพฤติกรรมระยะเวลายาว (เลิกเล่นช้า) $l = 10$

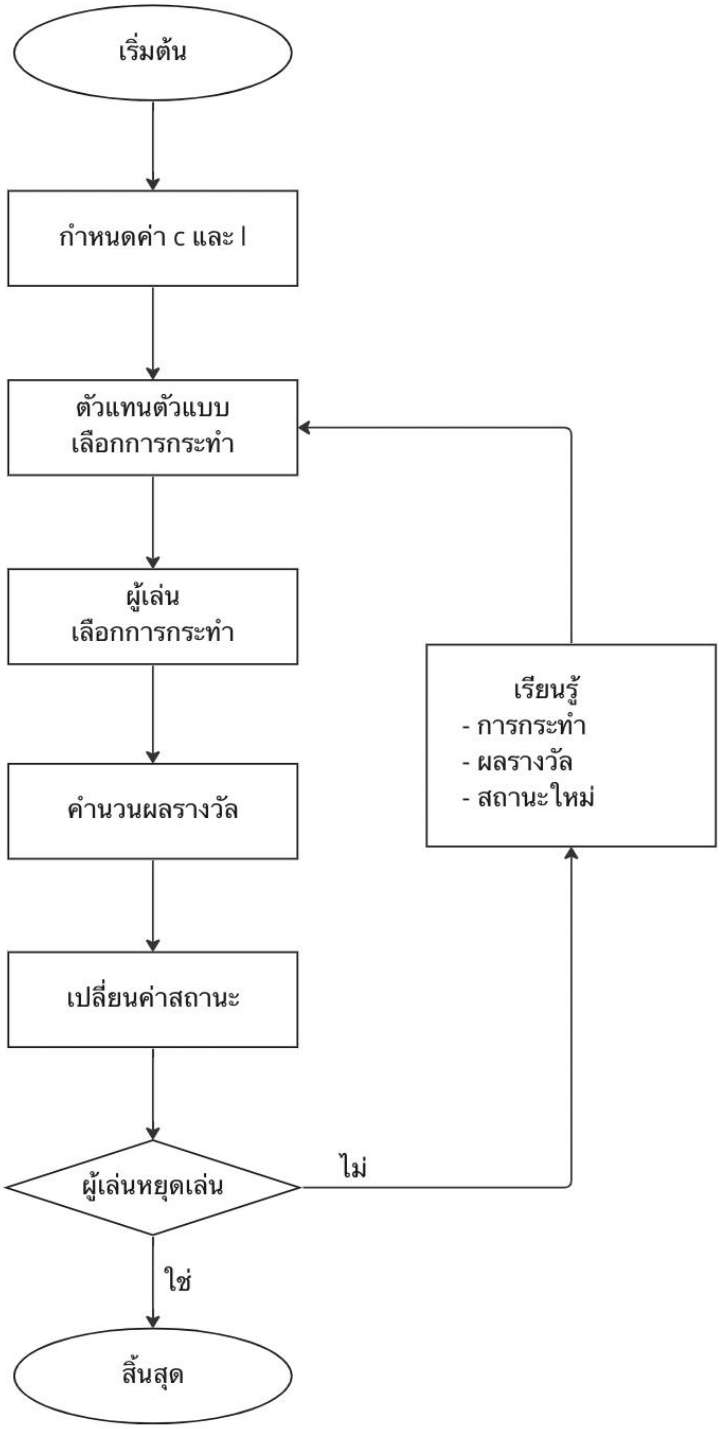
การจำลองการตัดสินใจเลือกการกระทำของผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบ 1 คน ในรูปแบบการแจกแจงเอกรูป (uniform distribution) เป็นอิสระต่อกันในแต่ละเกม โดยไม่ขึ้นกับผลแพ้ชนะของเกมก่อนหน้า และหยุดเล่นเมื่อผู้เล่นแพ้ต่อเนื่องติดต่อกัน l เกม โดย $l = \{5, 10\}$ หรือเล่นครบ 100 เกม

การจำลองการตัดสินใจเลือกการกระทำของผู้เข้าร่วมเล่นเกมเป่ายิ้งฉุบ 1 คน โดยขึ้นกับผลแพ้ชนะในเกมก่อนหน้า 1 เกม ตามเงื่อนไขดังนี้

1. หากผู้เล่นชนะในเกมรอบที่แล้ว ผู้เล่นจะตัดสินใจเลือกการกระทำในเกมถัดไปโดยการสุ่มแบบเอกรูป
2. หากผู้เล่นแพ้ในเกมรอบที่แล้ว ผู้เล่นจะตัดสินใจเลือกการกระทำในเกมถัดไปโดยกลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น $c = \{0.25, 0.75\}$
3. หากผู้เล่นแพ้ต่อเนื่องติดต่อกัน l เกม โดย $l = \{5, 10\}$ หรือเล่นครบ 100 เกม ผู้เล่นจะหยุดเล่น

เริ่มการจำลองโดยการกำหนดค่า c และ l ให้กับสภาพแวดล้อม จากนั้นตัวแทนตัวแบบจะเลือกการกระทำตามอัลกอริทึมของตัวแบบ ผู้เล่นจะเลือกการกระทำตามเงื่อนไขข้างต้น และนำมาคำนวณผลรางวัล อัพเดทสถานะใหม่ตามผลการเล่น ตรวจสอบว่าผู้เล่นหยุดเล่นหรือไม่ กรณีผู้เล่นไม่

หยุดเล่นตัวแทนตัวแบบจะเรียนรู้สิ่งที่เกิดขึ้น และเล่นต่อเกมถัดไป กรณีผู้เล่นหยุดเล่นให้สิ้นสุดรอบการจำลอง ดังแสดงในรูปที่ 3.1



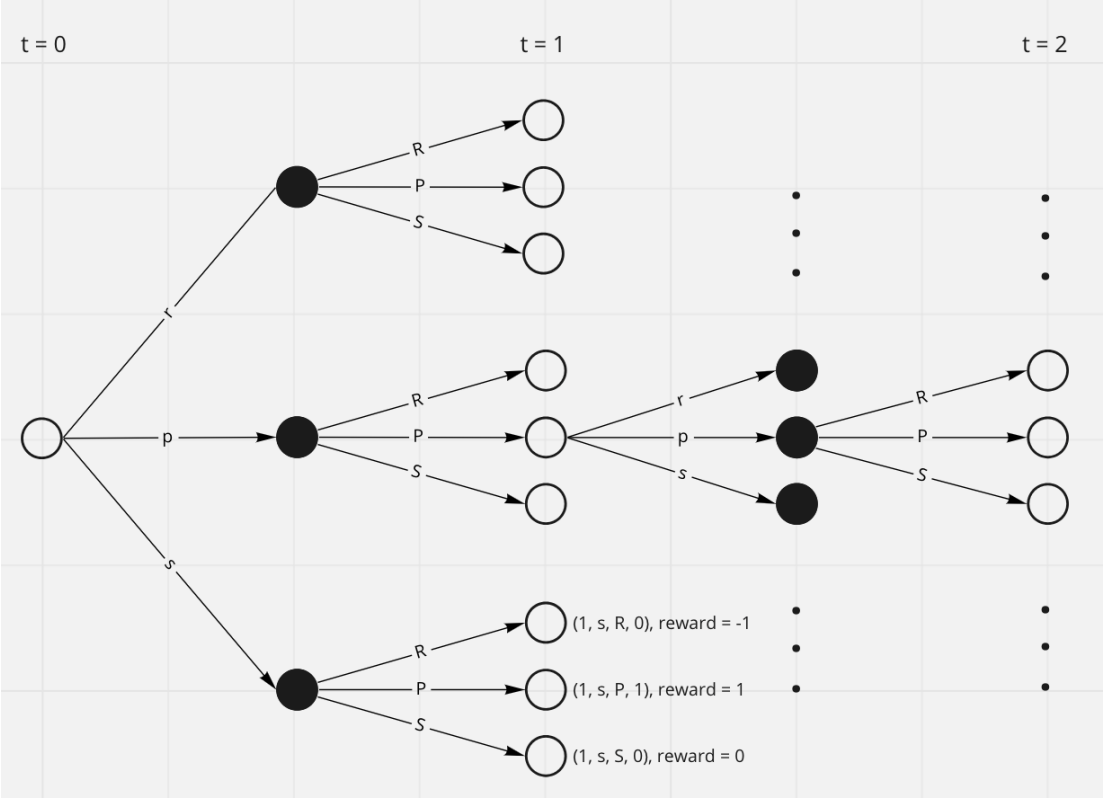
รูปที่ 3.1 ขั้นตอนการจำลองข้อมูล

3.2 วิธีการสร้างตัวแบบ

กำหนดให้ สถานะ $s_t = (t, a_t, A_t, l_t)$

สถานะเริ่มต้น $s_0 = (t = 0, a_0, A_0, l_0 = 0)$ โดย $a_0 \in a$ โดยการสุ่มแบบเอกรูป และ $A_0 \in A$ โดยการสุ่มแบบเอกรูป

คำนวณหานโยบายสำหรับให้ตัวแทนตัวแบบเลือกการกระทำโดย อัลกอริทึมการสุ่มแบบเอกรูป อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน โดยมีรายละเอียดของตัวแบบ ตามรูปที่ 3.2



รูปที่ 3.2 ตัวแบบสถานะและการกระทำ

จากรูปที่ 3.2 เมื่อระบบเริ่มต้น ที่ $t = 0$ สภาพแวดล้อมของระบบจะมี สถานะ $s_0 = (t = 0, a_0, A_0, l_0 = 0)$ ตัวแบบมีความน่าจะเป็นในการเลือกการกระทำ $a_0 = \{r, p, s\}$ และผู้เล่นมีความน่าจะเป็นในการเลือกการกระทำ $A_0 = \{R, P, S\}$ จากนั้นสภาพแวดล้อมของระบบจะเปลี่ยนสถานะไปเป็น สถานะ $s_1 = (t = 1, a_1, A_1, l_1)$ โดยที่ l_1มีค่าที่เป็นไปได้คือ 0 หรือ 1 เท่านั้น เพราะที่ $t = 1$ ผู้เล่นสามารถแพ้ต่อเนื่องติดกันได้แค่ 0 หรือ 1 รอบ และตัวแบบจะคำนวนนโยบายการตัดสินใจโดยดูเฉพาะสถานะที่เป็นไปได้ในขณะนั้น ค่าสถานะที่ t ในลำกับถัด ๆ

ไปนั้น สถานะที่เป็นไปได้ในแต่ละจุดก็จะปรับเปลี่ยนไป โดยขึ้นอยู่กับสถานะก่อนหน้าและผลจากการกระทำของตัวแบบและผู้เล่นในสถานะก่อนหน้า

3.2.1 การเลือกการกระทำถัดไปโดยอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน

การเลือกการกระทำถัดไปโดยอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน ตามอัลกอริทึมบทที่ 2.5 กำหนดให้แต่ละตัวเลือกการกระทำแทนด้วยแต่ละแขนของแบนดิทหลายแขน โดยที่แขนทั้งหมด คือ รูปแบบตัวเลือกการกระทำ × รูปแบบจำนวณครั้งที่ผู้เล่นแพ้ต่อเนื่องติดต่อกัน หรือ $a \times l$ และการแจกแจงของแต่ละแขนคือ $beta(win, not\_win)$ โดยที่ $win = 0,\ not\_win = 0$

การตัดสินใจเลือกการกระทำถัดไป โดยการสุ่มค่าให้แต่ละแขนที่เป็นไปได้ในสถานะนั้น $\theta_{a_t} \sim beta(win + 1, not\_win + 1)$ และเลือกกระทำตามค่าของแขนที่ θ_{a_t} มากที่สุด หากการกระทำที่เลือกชนะ ให้อัพเดทค่า win เพิ่มขึ้น 1 หน่วย หากการกระทำที่เลือกไม่ชนะ ให้อัพเดทค่า not_win เพิ่มขึ้น 1 หน่วย เมื่อแขนนั้นถูกเลือกซ้ำ ๆ การแจกแจงของแขนจะลู่เข้าสู่ค่าแท้จริง

3.2.2 การเลือกการกระทำถัดไปโดยอัลกอริทึมความเชื่อมั่นขอบเขตบน

การเลือกการกระทำถัดไปโดยอัลกอริทึมความเชื่อมั่นขอบเขตบน ตามอัลกอริทึมบทที่ 2.6 กำหนดให้แต่ละตัวเลือกการกระทำแทนด้วยแต่ละแขนของแบนดิทหลายแขน โดยที่แขนทั้งหมด คือ รูปแบบตัวเลือกการกระทำ × รูปแบบจำนวณครั้งที่ผู้เล่นแพ้ต่อเนื่องติดต่อกัน หรือ $a \times l$ และคุณค่าจากการกระทำของทุกแขนคือ $\hat{Q}_t(s, a) = 0$

การตัดสินใจเลือกการกระทำถัดไป ให้เลือกแขนที่เป็นไปได้ในสถานะนั้น และเป็นแขนมีค่าความเชื่อมั่นสูงสุดเมื่อรวมปัจจัยการสำรวจแล้ว ดังนี้ $argmax_a \left\{ \hat{Q}_t(s, a) + \sqrt{\frac{ln(t)}{n_k}} \right\}$ จากนั้นอัพเดทค่าคุณค่าจากการกระทำของแขนที่เลือกไป เมื่อแขนนั้นถูกเลือกซ้ำ ๆ ค่าความเชื่อมั่นของแขนจะลู่เข้าสู่ค่าแท้จริง

3.2.3 การเลือกการกระทำถัดไปโดยอัลกอริทึมการสุ่มแบบเอกรูป

การเลือกการกระทำถัดไปโดยอัลกอริทึมการสุ่มแบบเอกรูป ตามอัลกอริทึมบทที่ 2.7 กำหนดให้แต่ละตัวเลือกการกระทำแทนด้วยแต่ละแขนของแบนดิทหลายแขน โดยที่แขนทั้งหมด คือ รูปแบบตัวเลือกการกระทำ หรือ a

การตัดสินใจเลือกการกระทำถัดไป ให้เลือกโดยการสุ่ม $a_t \sim uniform(a)$ โดยไม่ขึ้นกับผลแพ้ชนะและการกระทำในรอบก่อนหน้า

3.3 วิธีการเปรียบเทียบผล

เปรียบเทียบผลโดยแสดงในรูปแบบกราฟ ประกอบด้วยข้อมูลรางวัลสะสมบนแกนเวลา ซึ่งตัวแทนตัวแบบ (agent) เรียนรู้รูปแบบการตัดสินใจ โดยทำการเล่นกับผู้เล่นเพื่อเรียนรู้การตัดสินใจที่จะทำให้ได้ผลรางวัลสะสมมากที่สุด

กำหนดให้การเรียนรู้ 1 รอบ (episode) คือการที่ตัวแทนตัวแบบเล่นเป่ายิ้งฉุบกับผู้เล่นจนกระทั่งผู้เล่นหยุดเล่นเกมเมื่อ ผู้เล่นแพ้ต่อเนื่องติดต่อกัน l เกม หรือเล่นจนครบ 100 เกม จากนั้นจะทำการเปรียบเทียบว่าตัวแบบอัลกอริทึมใด ใช้จำนวนรอบการเรียนรู้น้อยกว่าเพื่อให้ได้ผลรางวัลสะสมที่มากกว่า โดยทำการเปรียบเทียบกับอัลกอริทึมการสุ่มแบบเอกรูป ที่เป็นตัวแบบบรรทัดฐาน (baseline model) โดยรอบการเรียนรู้ในการเปรียบเทียบทั้งหมด 2,000 รอบ

1. เปรียบเทียบผลรางวัลสะสมระหว่างตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน คู่กับตัวแบบบรรทัดฐานอัลกอริทึมการสุ่มแบบเอกรูป เมื่อการตัดสินใจของผู้เข้าร่วมเล่นเป็นการสุ่มแบบเอกรูปโดยไม่ขึ้นกับผลแพ้ชนะของเกมก่อนหน้า
2. เปรียบเทียบผลรางวัลสะสมระหว่างอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน คู่กับตัวแบบบรรทัดฐานอัลกอริทึมการสุ่มแบบเอกรูปบน a เมื่อการตัดสินใจของผู้เข้าร่วมเล่นขึ้นกับผลแพ้ชนะของเกมก่อนหน้า 1 เกม ว่าจะกลยุทธตามเข็มนาฬิกาที่ความน่าจะเป็น 0.25 และ 0.75

บทที่ 4
ผลการวิจัย

งานวิจัยนี้ศึกษาตัวแบบการเรียนรู้แบบเสริมแรงด้วยอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบน โดยที่การเปรียบเทียบประสิทธิภาพของตัวแบบทั้งสองจะเปรียบเทียบจากผลรางวัลสะสมของตัวแบบกับตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน ภายใต้สถานการณ์การจำลองเกมเป่ายิ้งฉุบว่า ตัวแบบใดสามารถคาดการณ์พฤติกรรมของมนุษย์ และหาวิธีโต้ตอบเพื่อให้ได้ผลรางวัลสะสมที่มากกว่าในระยะยาว ทั้งนี้ผลการวิจัยจะแบ่งเป็น 4 ส่วน ดังนี้

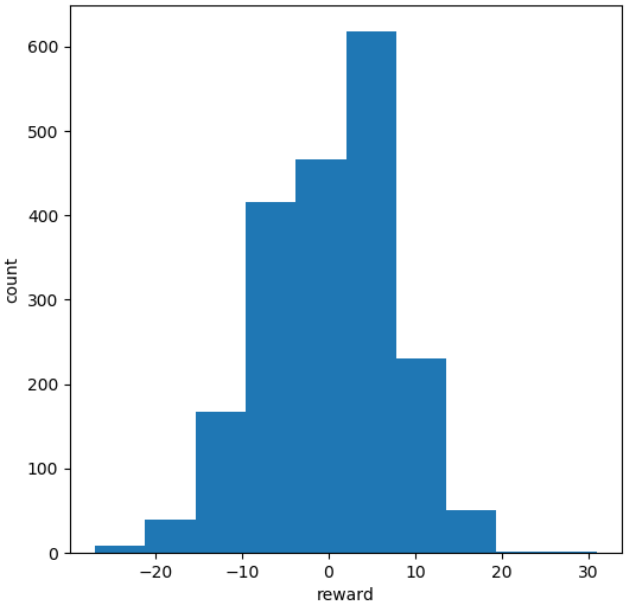
1. ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป (ตัวแบบบรรทัดฐาน)
2. ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน
3. ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน
4. เปรียบเทียบผลรางวัลสะสมของตัวแบบ

4.1 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป

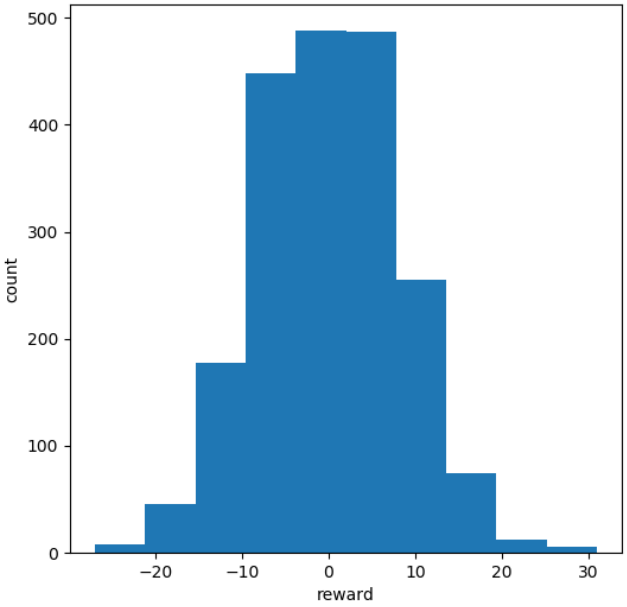
การตอบสนองของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปเมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่นเมื่อผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะและแพ้ใกล้เคียงกันภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.1 และ รูปที่ 4.2 ตามลำดับ

เมื่อดูผลรางวัลสะสมของตัวแบบ พบว่าคาดเดารูปแบบของผลรางวัลสะสมไม่ได้ คือในบางช่วงที่ตัวแบบชนะมากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะสูงขึ้น แต่ในบางช่วงที่ตัวแบบแพ้มากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะต่ำลง โดยที่ไม่สามารถคาดการณ์แนวโน้มว่าจะสูงขึ้นหรือต่ำลงได้เลย ดังรูปที่ 4.3 และรูปที่ 4.4

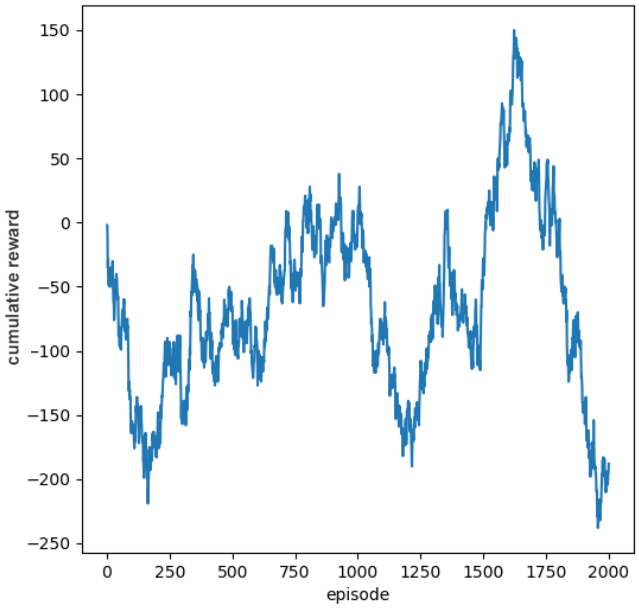
ทั้งนี้หากทำการเพิ่มรอบการจำลองหรือลดรอบการจำลองการเล่นเกมเป่ายิ้งฉุบของตัวแบบนี้กับผู้เล่น ก็ไม่สามารถคาดเดาผลรางวัลสะสมได้เช่นกันว่า ตัวแบบจะได้ผลรางวัลสะสมออกมาในรูปแบบใด



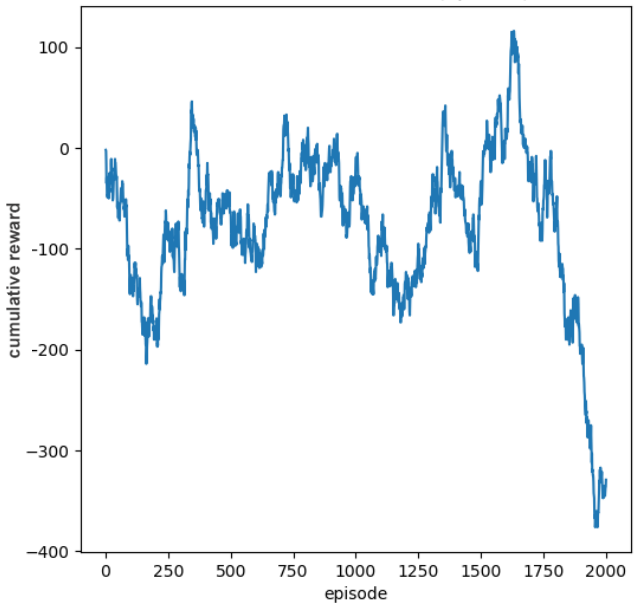
รูปที่ 4.1 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.2 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.3 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



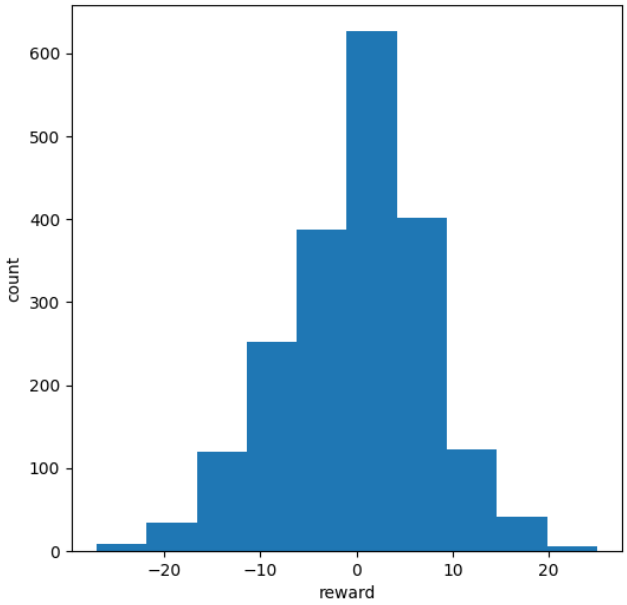
รูปที่ 4.4 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม

การตอบสนองของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปเมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่นเมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม และเมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.75 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะและแพ้ใกล้เคียงกันภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.5 รูปที่ 4.6 รูปที่ 4.7 และรูปที่ 4.8 ตามลำดับ

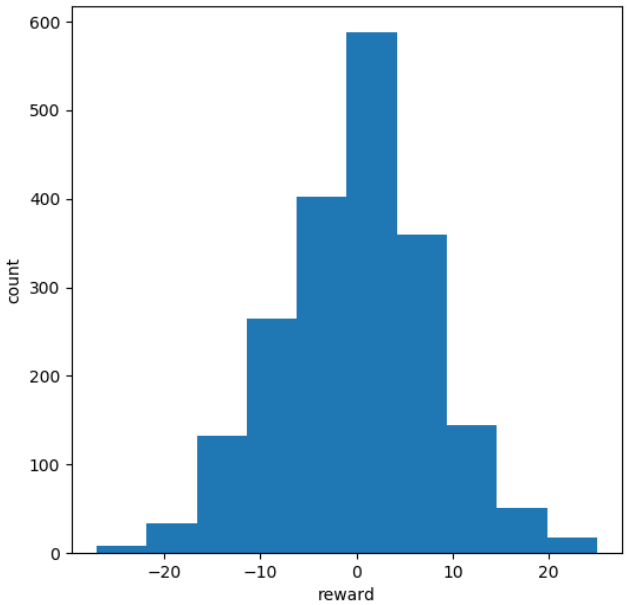
เมื่อดูผลรางวัลสะสมของตัวแบบ พบว่าคาดเดารูปแบบของผลรางวัลสะสมไม่ได้ คือในบางช่วงที่ตัวแบบชนะมากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะสูงขึ้น แต่ในบางช่วงที่ตัวแบบแพ้มากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะต่ำลง โดยที่ไม่สามารถคาดการณ์แนวโน้มว่าจะสูงขึ้นหรือต่ำลงได้เลย ดังรูปที่ 4.9 รูปที่ 4.10 รูปที่ 4.11 และรูปที่ 4.12

ทั้งนี้หากทำการเพิ่มรอบการจำลองหรือลดรอบการจำลองการเล่นเกมเป่ายิ้งฉุบของตัวแบบนี้กับผู้เล่น ก็ไม่สามารถคาดเดาผลรางวัลสะสมได้เช่นกันว่า ตัวแบบจะได้ผลรางวัลสะสมออกมาในรูปแบบใด

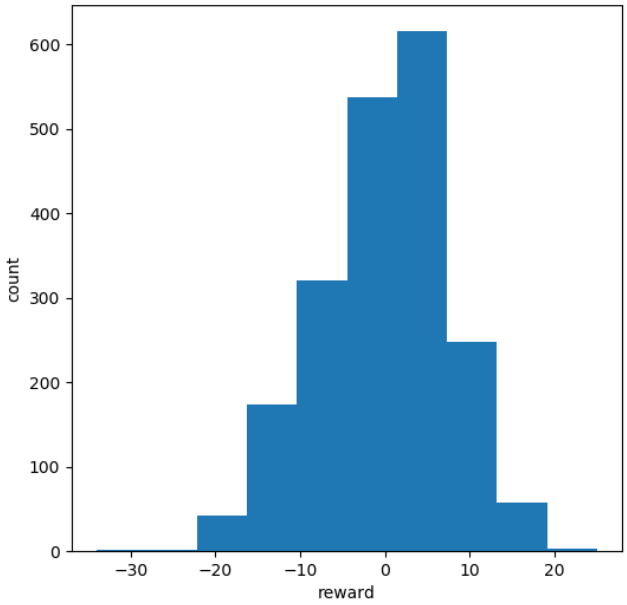
จากผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปทั้งในกรณีที่ผู้เล่นมีการใช้ และไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบนี้ไม่มีการคาดการณ์พฤติกรรมของผู้เล่น และตัวแบบตอบสนองผู้เล่นโดยการสุ่มแบบเอกรูป ซึ่งไม่มีแบบแผนที่ใช้ในการโต้ตอบการกระทำของผู้เล่น ส่งผลให้ไม่สามารถคาดการณ์ผลรางวัลสะสมของตัวแบบได้ ไม่ว่าจะจำลองรอบการเล่นเกมให้สั้นลงหรือยาวมากขึ้นก็ตาม ดังรูปที่ 4.13



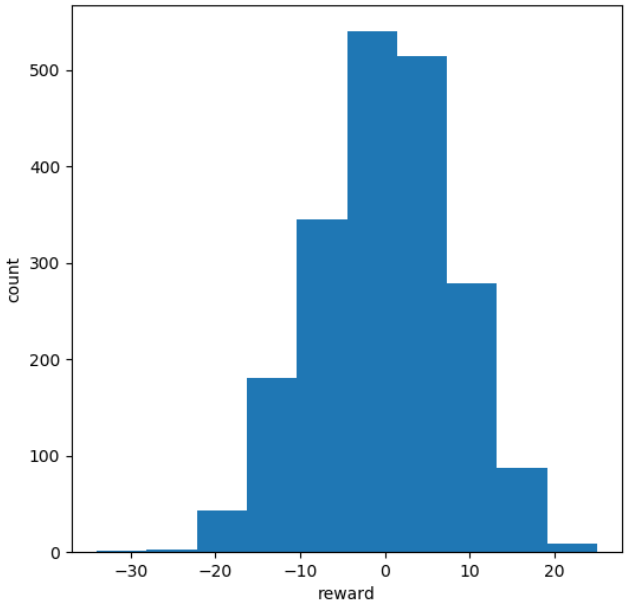
รูปที่ 4.5 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



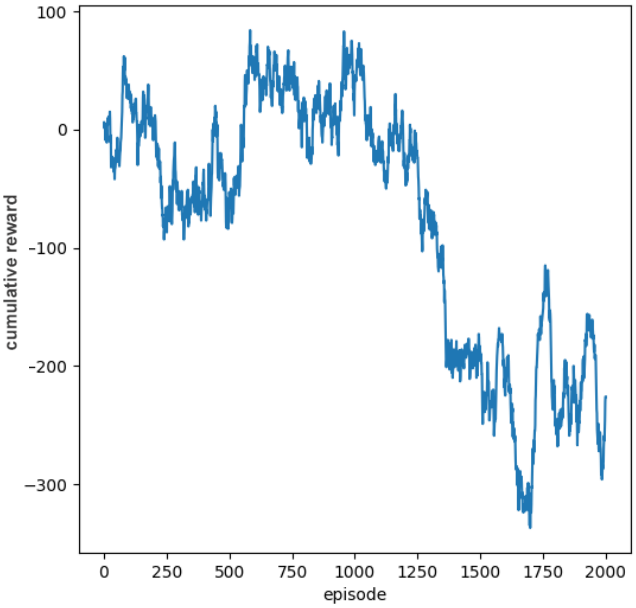
รูปที่ 4.6 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



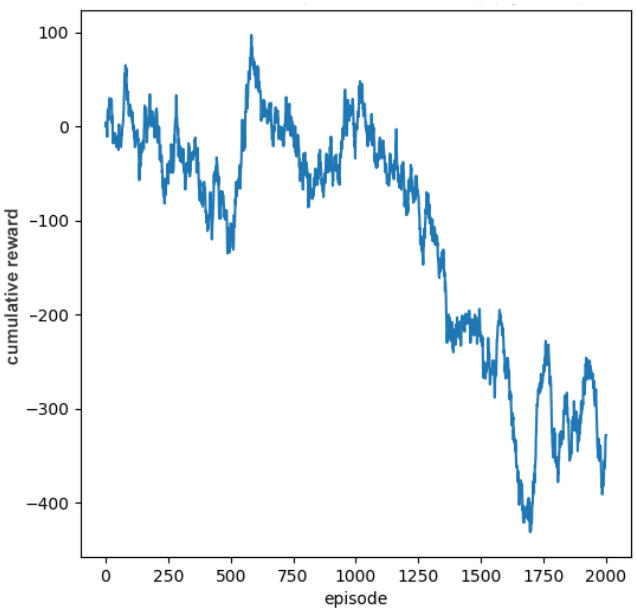
รูปที่ 4.7 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



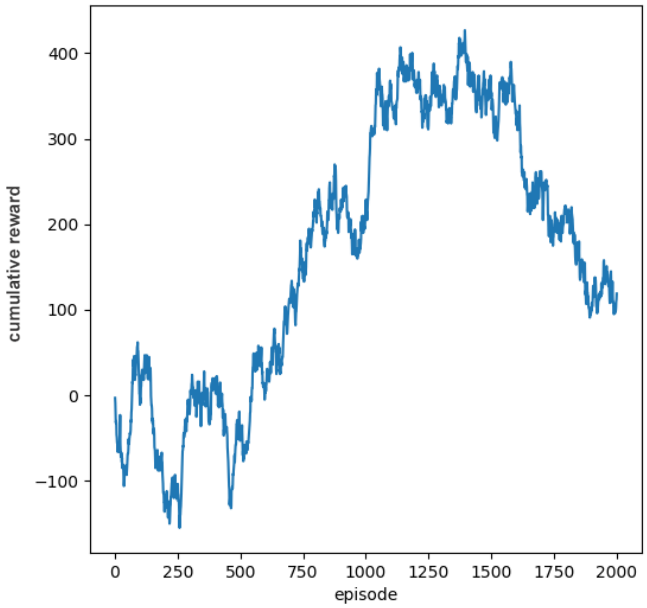
รูปที่ 4.8 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



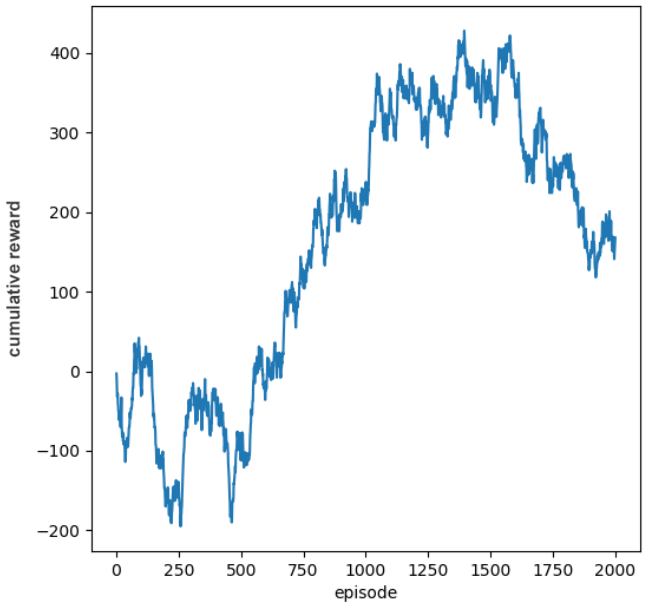
รูปที่ 4.9 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



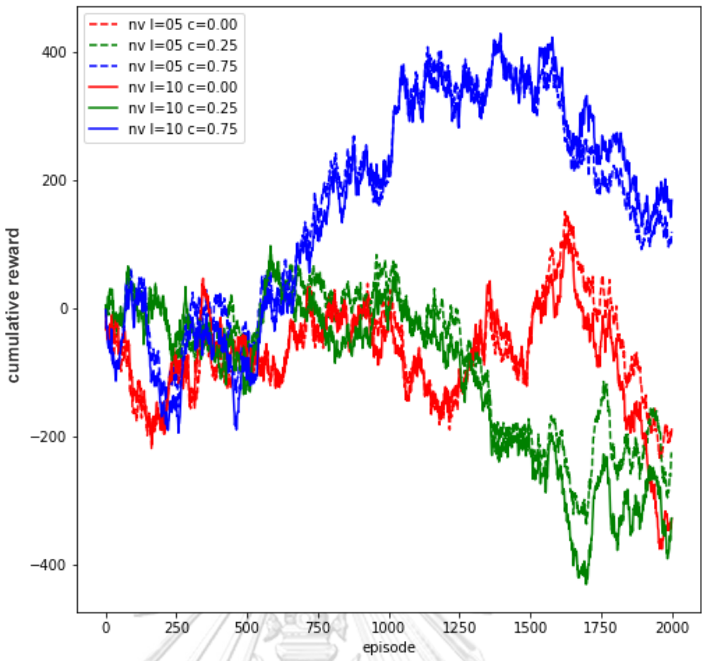
รูปที่ 4.10 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.11 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.12 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.13 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็นที่ต่างกัน และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

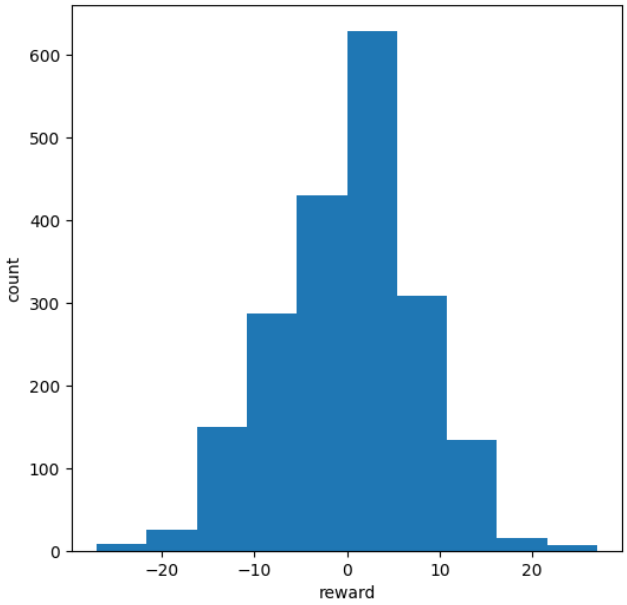
4.2 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน

การตอบสนองของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันเมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่น เมื่อผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะและแพ้ใกล้เคียงกันภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.14 และ รูปที่ 4.15 ตามลำดับ

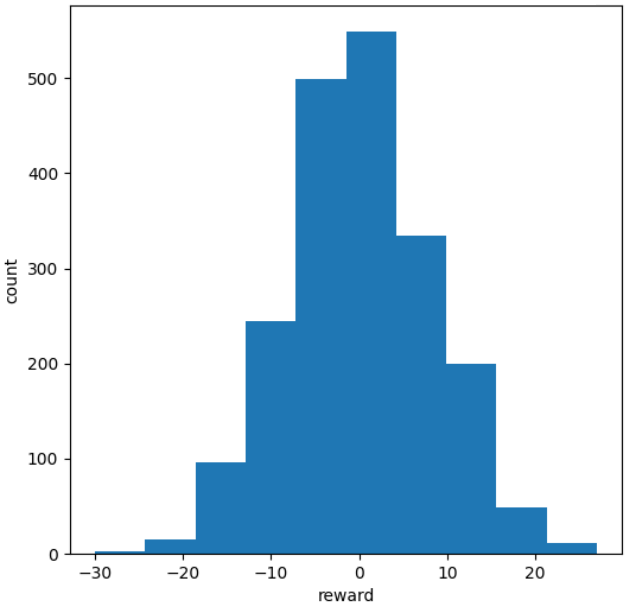
เมื่อดูผลรางวัลสะสมของตัวแบบ พบว่าคาดเดารูปแบบของผลรางวัลสะสมไม่ได้ คือในบางช่วงที่ตัวแบบชนะมากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะสูงขึ้น แต่ในบางช่วงที่ตัวแบบแพ้มากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะต่ำลง โดยที่ไม่สามารถคาดการณ์แนวโน้มว่าจะสูงขึ้นหรือต่ำลงได้เลย ดังรูปที่ 4.16 และรูปที่ 4.17 ซึ่งแม้ว่าผลรางวัลสะสมในรูปที่ 4.17 จะดูเหมือนผลรางวัลสะสมสูงขึ้นก็ตาม แต่เมื่อรอบการจำลองมาขึ้นเรื่อย ๆ ผลรางวัลสะสมกลับมากขึ้นและลดลงกลับไปกลับมา ไม่สามารถคาดการณ์พฤติกรรมได้เช่นเดียวกัน

ทั้งนี้หากทำการเพิ่มรอบการจำลองหรือลดรอบการจำลองการเล่นเกมเป่ายิ้งฉุบของตัวแบบนี้กับผู้เล่น ก็ไม่สามารถคาดเดาผลรางวัลสะสมได้เช่นกันว่า ตัวแบบจะได้ผลรางวัลสะสมออกมาในรูปแบบใด

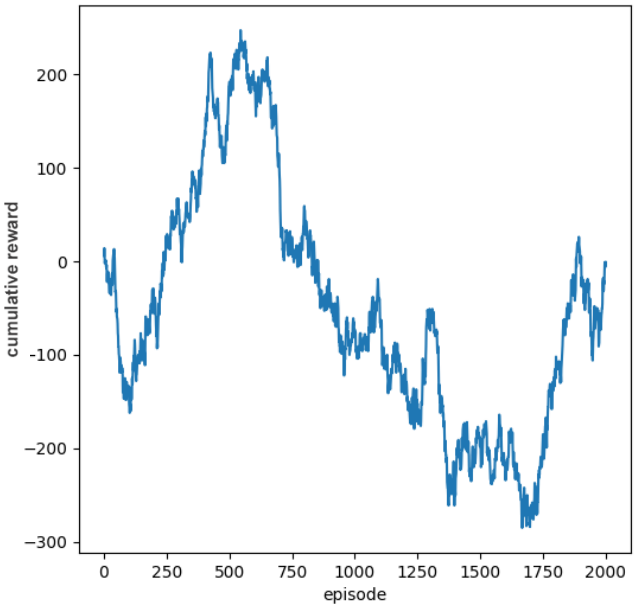
การที่ไม่สามารถคาดเดาผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันเช่นเดียวกับที่ไม่สามารถคาดเดาผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปนั้น เพราะผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา จึงไม่มีรูปแบบของพฤติกรรมของผู้เล่น ซึ่งส่งผลให้ตัวแบบไม่สามารถตรวจจับรูปแบบการกระทำของผู้เล่นได้ ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันจึงคาดเดาไม่ได้ ไม่ต่างจากผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน



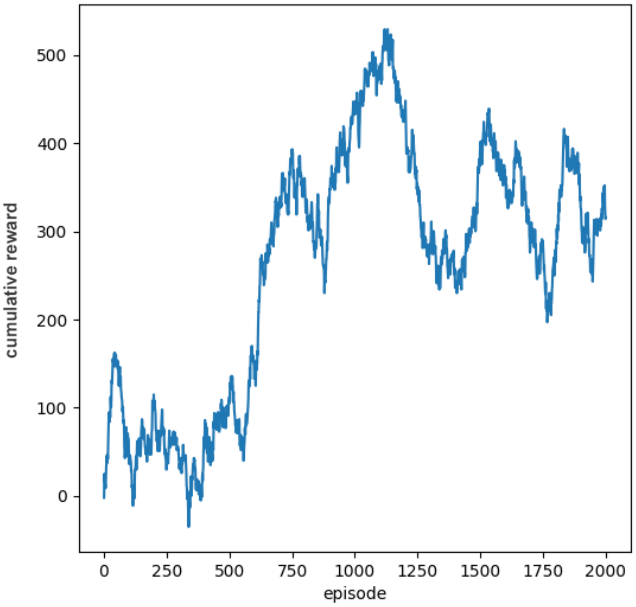
รูปที่ 4.14 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.15 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.16 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.17 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม

การตอบสนองของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน เมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่น เมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะมากกว่าแพ้ ดังรูปที่ 4.18 รูปที่ 4.19 ตามลำดับ

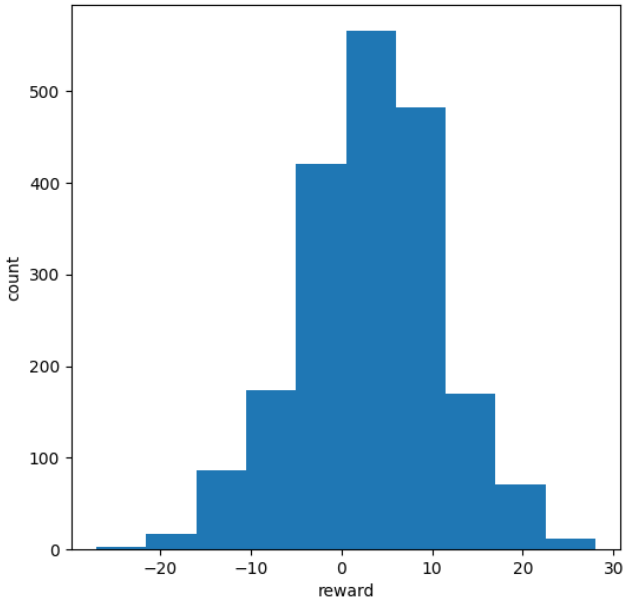
เมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.75 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะมากกว่าแพ้มากขึ้นกว่าตอนที่ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และยิ่งผู้เล่นใช้กลยุทธยุติการสูญเสียด้วยจำนวนเกมที่มากขึ้น ตัวแบบจะยิ่งชนะมากขึ้นตามไปด้วย ภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.20 รูปที่ 4.21 ตามลำดับ

เมื่อดูผลรางวัลสะสมของตัวแบบ สามารถคาดเดารูปแบบของผลรางวัลสะสมได้ คือผลรางวัลสะสมของตัวแบบจะเพิ่มขึ้นเรื่อย ๆ ตามจำนวนรอบการจำลอง ดังรูปที่ 4.22 รูปที่ 4.23 รูปที่ 4.24 และรูปที่ 4.25

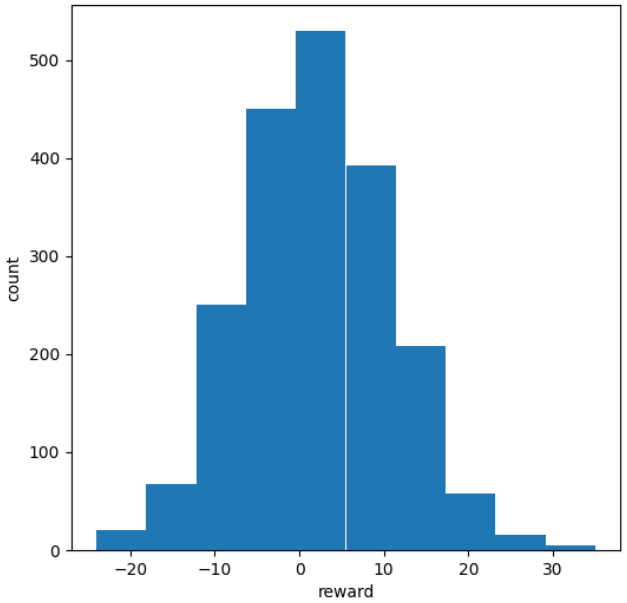
หากทำการเพิ่มรอบการจำลองมากขึ้นผลรางวัลสะสมก็จะมากขึ้นตาม หรือหากลดรอบการจำลองการลงผลรางวัลสะสมก็จะลดลงตาม ผลรางวัลสะสมจะขึ้นอยู่กับจำนวนรอบที่มี

จากผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน ในกรณีที่ผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบนี้ไม่สามารถตรวจจับและคาดการณ์พฤติกรรมของผู้เล่นได้ เนื่องจากผู้เล่นไม่มีรูปแบบพฤติกรรมให้ตัวแบบตรวจจับ ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันจะไม่สามารถคาดการณ์ได้ ไม่ต่างตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป ไม่ว่าจะจำลองรอบการเล่นเกมให้สั้นลงหรือยาวมากขึ้นก็ตาม

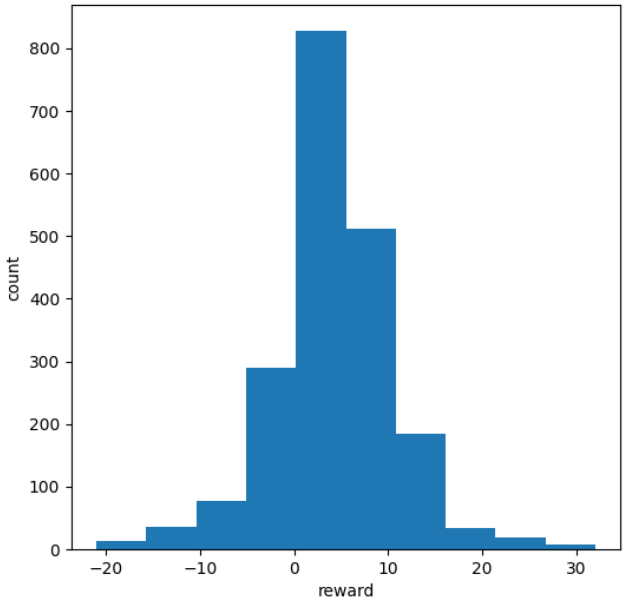
จากผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน ในกรณีที่ผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และ 0.75 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบนี้สามารถตรวจจับและคาดการณ์พฤติกรรมของผู้เล่นได้ และเลือกวิธีการโต้ตอบโดยคำนึงถึงผลรางวัลสะสมในระยะยาว ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันแปรผันตรงกับจำนวนรอบการจำลอง คือผลรางวัลสะสมมีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ตามจำนวนรอบการจำลอง ดังรูปที่ 4.26 ซึ่งให้ผลรางวัลสะสมที่ดีกว่าตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป



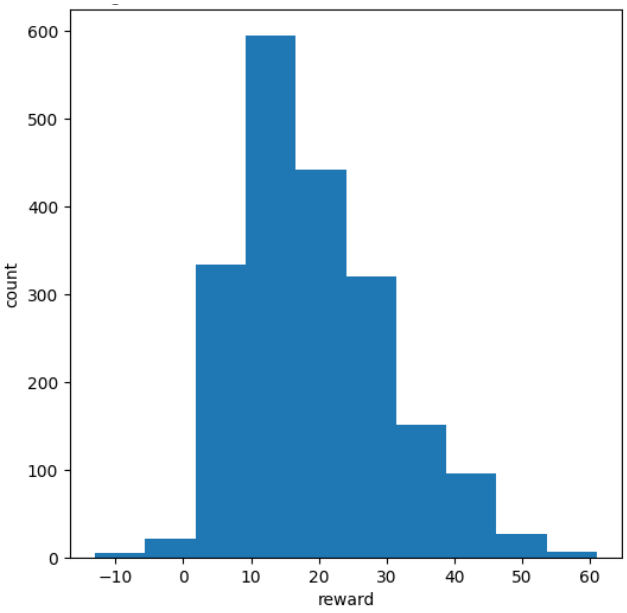
รูปที่ 4.18 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



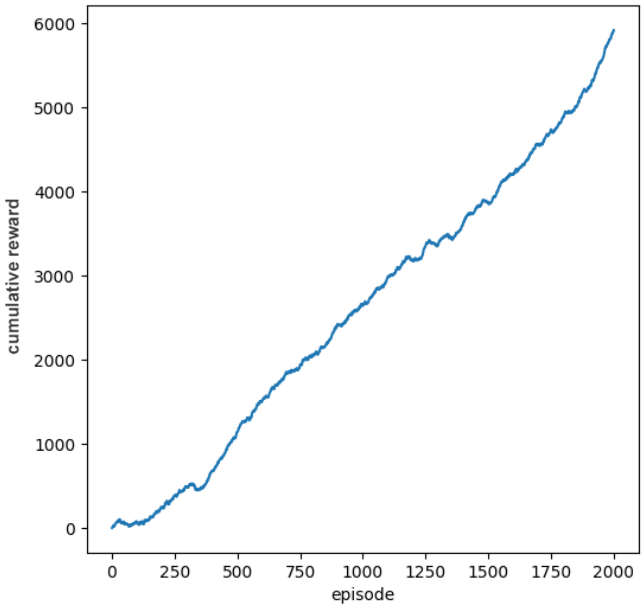
รูปที่ 4.19 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



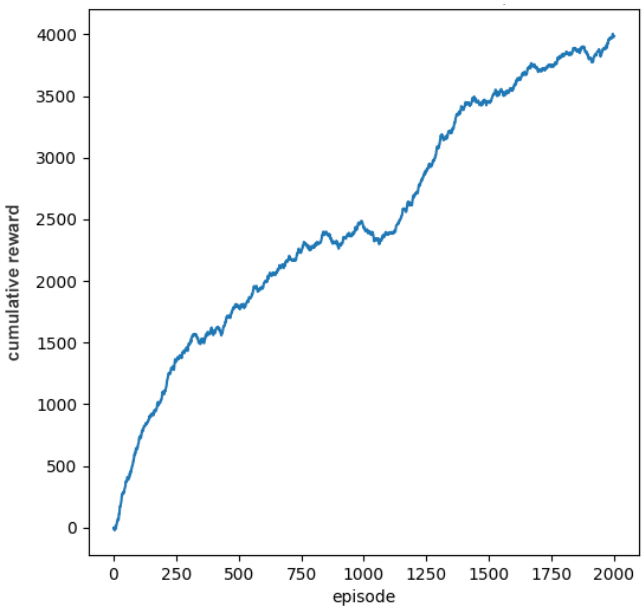
รูปที่ 4.20 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



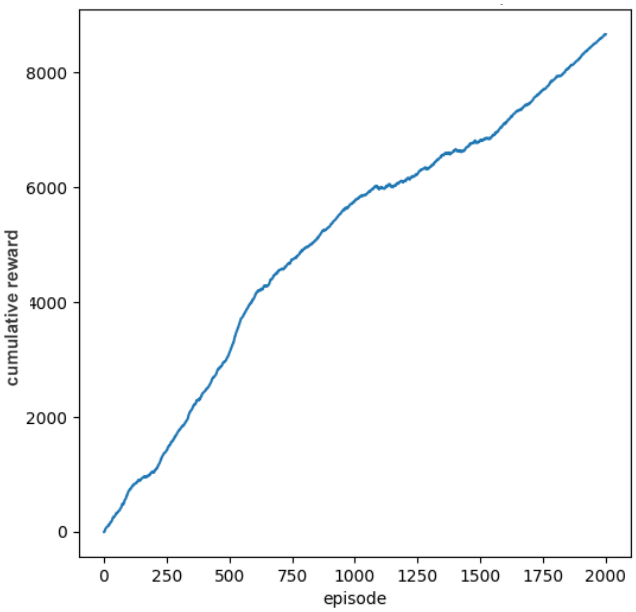
รูปที่ 4.21 ผลรางวัลของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



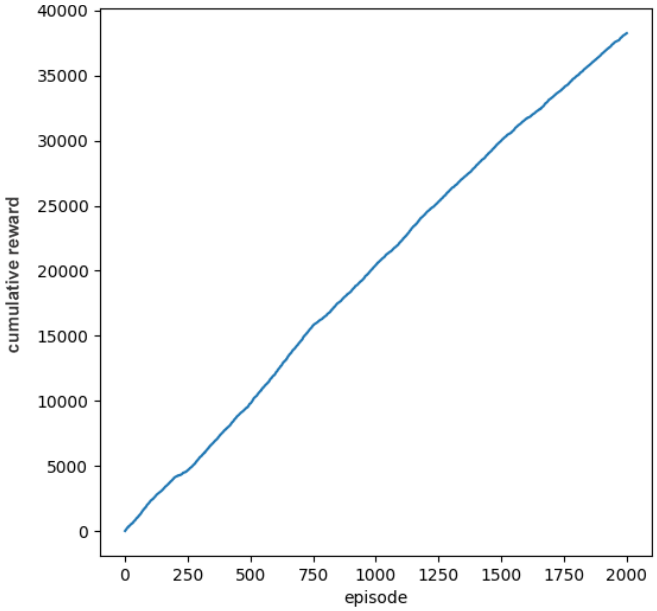
รูปที่ 4.22 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



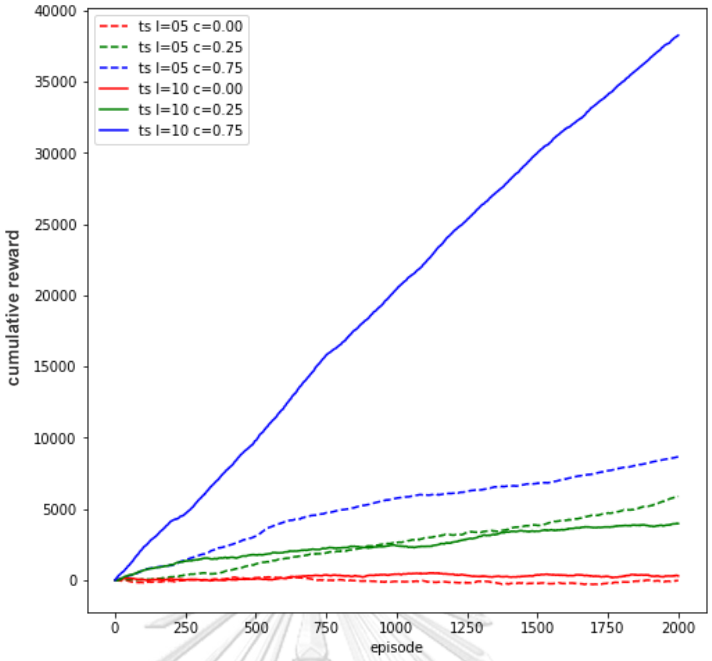
รูปที่ 4.23 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.24 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.25 ผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.26 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็นที่ต่างกัน และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

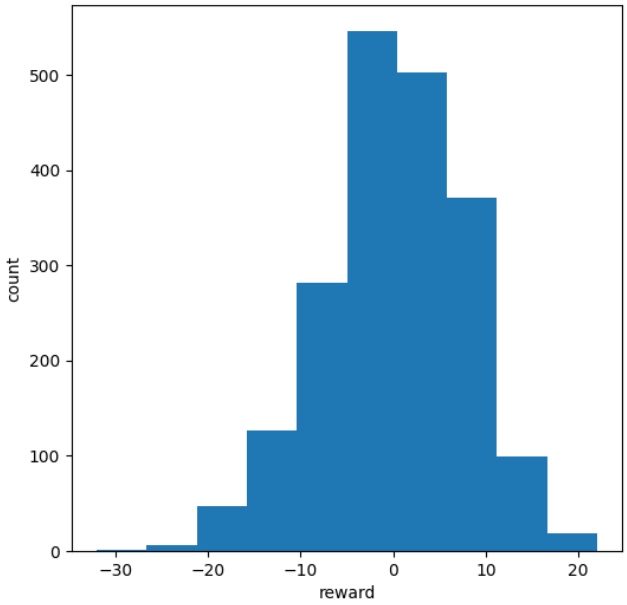
4.3 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน

การตอบสนองของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน เมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่น เมื่อผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะและแพ้ใกล้เคียงกันภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.27 และ รูปที่ 4.28 ตามลำดับ

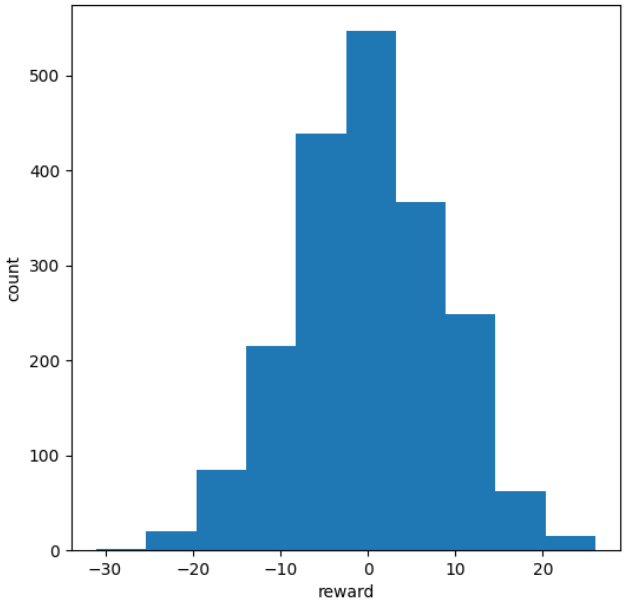
เมื่อดูผลรางวัลสะสมของตัวแบบ พบว่าคาดเดารูปแบบของผลรางวัลสะสมไม่ได้ คือในบางช่วงที่ตัวแบบชนะมากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะสูงขึ้น แต่ในบางช่วงที่ตัวแบบแพ้มากกว่าติดต่อกันในหลายรอบการจำลอง ผลรางวัลสะสมก็จะต่ำลง โดยที่ไม่สามารถคาดการณ์แนวโน้มว่าจะสูงขึ้นหรือต่ำลงได้เลย ดังรูปที่ 4.29 และรูปที่ 4.30 ซึ่งแม้ว่าผลรางวัลสะสมในรูปที่ 4.29 จะมากขึ้นเรื่อยๆในช่วงแรกแต่เมื่อระยะเวลาผ่านไปผลรางวัลสะสมกลับลดลง และแม้ว่าผลรางวัลสะสมในรูปที่ 4.30 จะน้อยลงเรื่อยๆในช่วงแรกแต่เมื่อเวลาผ่านไปผลรางวัลสะสมกลับมากขึ้น ทั้งสองกรณีผลรางวัลสะสมกลับมากขึ้นและลดลงกลับไปกลับมา โดยไม่มีรูปแบบและไม่สามารถคาดการณ์พฤติกรรมได้เช่นเดียวกัน

ทั้งนี้หากทำการเพิ่มรอบการจำลองหรือลดรอบการจำลองการเล่นเกมเป่ายิ้งฉุบของตัวแบบนี้กับผู้เล่น ก็ไม่สามารถคาดเดาผลรางวัลสะสมได้เช่นกันว่า ตัวแบบจะได้ผลรางวัลสะสมออกมาในรูปแบบใด

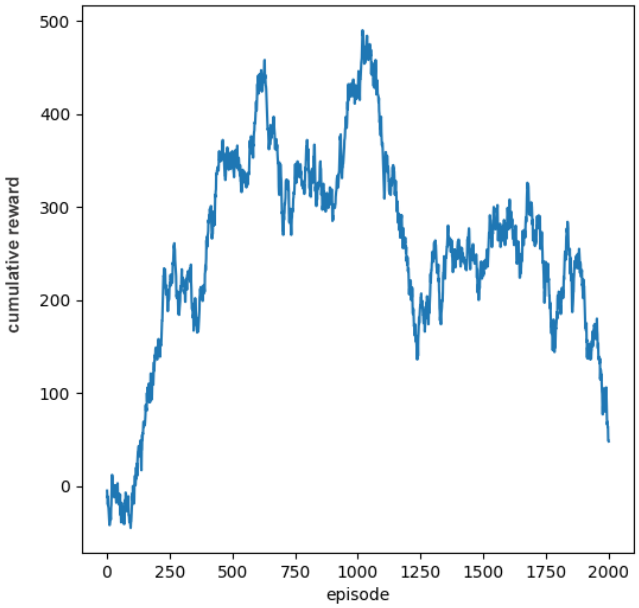
การที่ไม่สามารถคาดเดาผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนเช่นเดียวกับที่ไม่สามารถคาดเดาผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปนั้น เพราะผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา จึงไม่มีรูปแบบของพฤติกรรมของผู้เล่น ซึ่งส่งผลให้ตัวแบบไม่สามารถตรวจจับรูปแบบการกระทำของผู้เล่นได้ ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนจึงคาดเดาไม่ได้ ไม่ต่างจากผลรางวัลสะสมของตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน



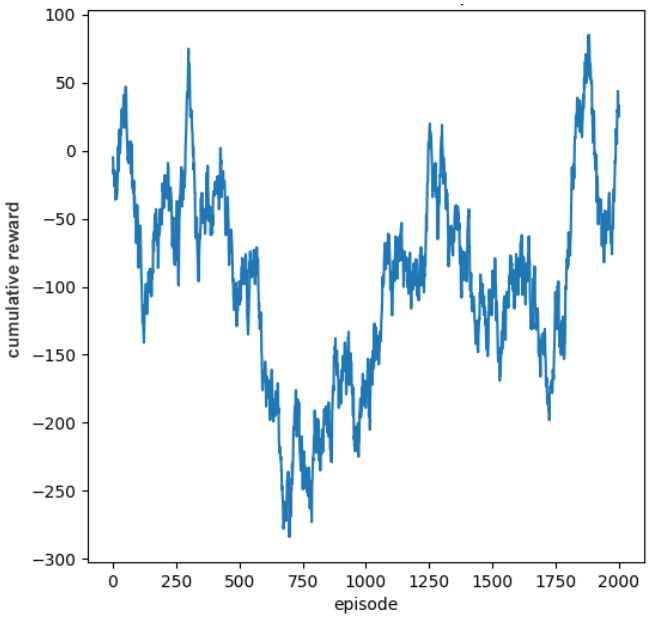
รูปที่ 4.27 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.28 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.29 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.30 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม

การตอบสนองของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน เมื่อต้องทำการเล่นเป่ายิ้งฉุบกับผู้เล่น เมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะมากกว่าแพ้ ดังรูปที่ 4.31 รูปที่ 4.32 ตามลำดับ

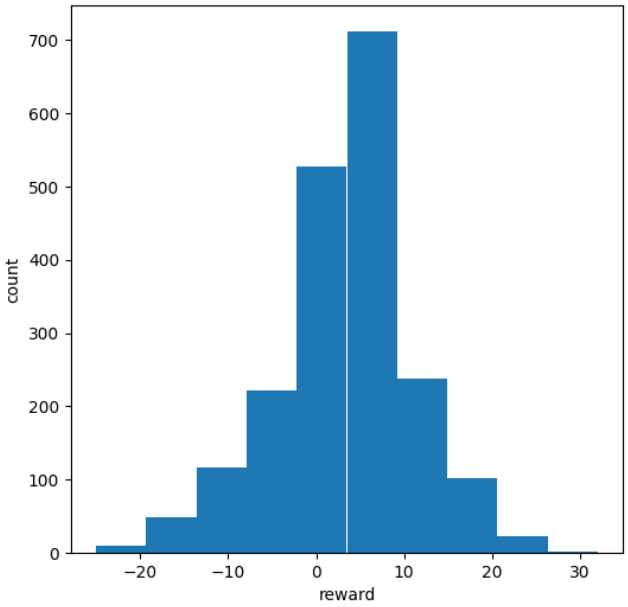
เมื่อผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.75 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบมีผลการชนะมากกว่าแพ้มากขึ้นกว่าตอนที่ผู้เล่นใช้กลยุทธตามเข็มนาฬิกา ด้วยความน่าจะเป็น 0.25 และยิ่งผู้เล่นใช้กลยุทธยุติการสูญเสียด้วยจำนวนเกมที่มากขึ้น ตัวแบบจะยิ่งชนะมากขึ้นตามไปด้วย ภายใต้การจำลองทั้งหมด 2,000 รอบ ดังรูปที่ 4.33 รูปที่ 4.34 ตามลำดับ

เมื่อดูผลรางวัลสะสมของตัวแบบ สามารถคาดเดารูปแบบของผลรางวัลสะสมได้ คือผลรางวัลสะสมของตัวแบบจะเพิ่มขึ้นเรื่อย ๆ ตามจำนวนรอบการจำลอง ดังรูปที่ 4.35 รูปที่ 4.36 รูปที่ 4.37 และรูปที่ 4.38

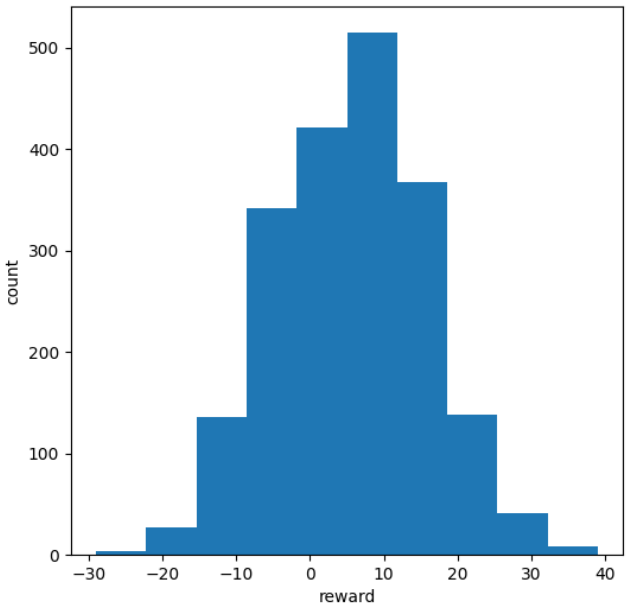
หากทำการเพิ่มรอบการจำลองมากขึ้นผลรางวัลสะสมก็จะมากขึ้นตาม หรือหากลดรอบการจำลองการลงผลรางวัลสะสมก็จะลดลงตาม ผลรางวัลสะสมจะขึ้นอยู่กับจำนวนรอบที่มี

จากผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน ในกรณีที่ผู้เล่นไม่มีการใช้กลยุทธตามเข็มนาฬิกา และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบนี้ไม่สามารถตรวจจับและคาดการณ์พฤติกรรมของผู้เล่นได้ เนื่องจากผู้เล่นไม่มีรูปแบบพฤติกรรมให้ตัวแบบตรวจจับ ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนจะไม่สามารถคาดการณ์ได้ ไม่ต่างตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป ไม่ว่าจะจำลองรอบการเล่นเกมให้สั้นลงหรือยาวมากขึ้นก็ตาม

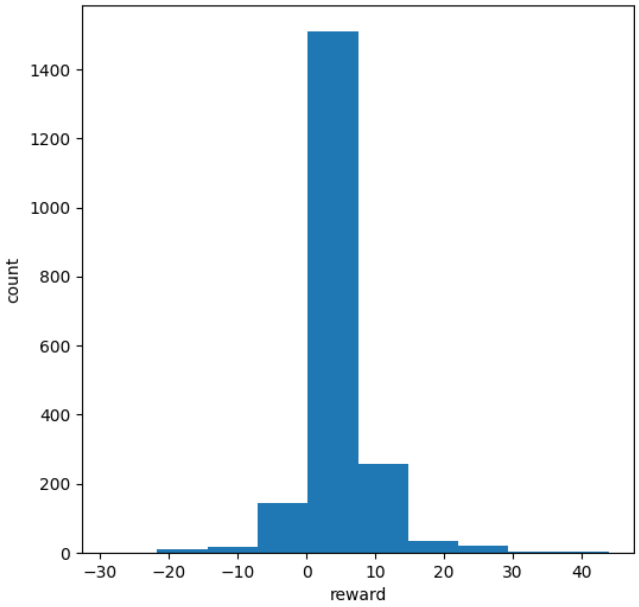
จากผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน ในกรณีที่ผู้เล่นมีการใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และ 0.75 และผู้เล่นมีการใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดต่อกัน 5 เกม และ 10 เกม ตัวแบบนี้สามารถตรวจจับและคาดการณ์พฤติกรรมของผู้เล่นได้ และเลือกวิธีการโต้ตอบโดยคำนึงถึงผลรางวัลสะสมในระยะยาว ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนแปรผันตรงกับจำนวนรอบการจำลอง คือผลรางวัลสะสมมีแนวโน้มเพิ่มขึ้นเรื่อย ๆ ตามจำนวนรอบการจำลอง ดังรูปที่ 4.39 ซึ่งให้ผลรางวัลสะสมที่ดีกว่าตัวแบบอัลกอริทึมการสุ่มแบบเอกรูป



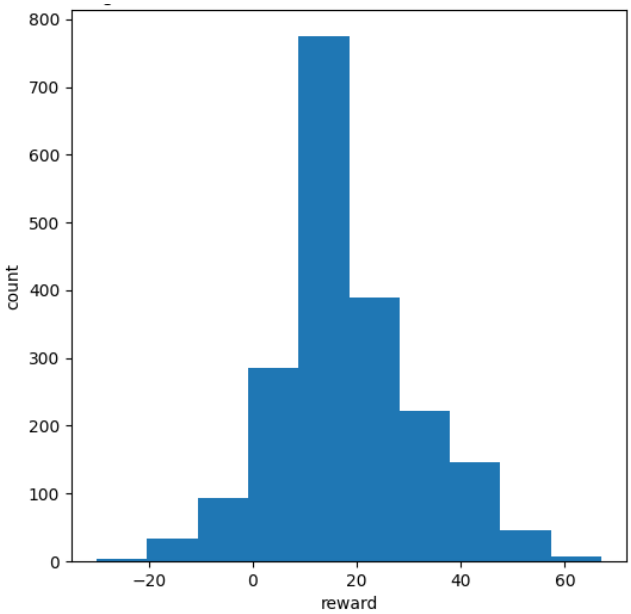
รูปที่ 4.31 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



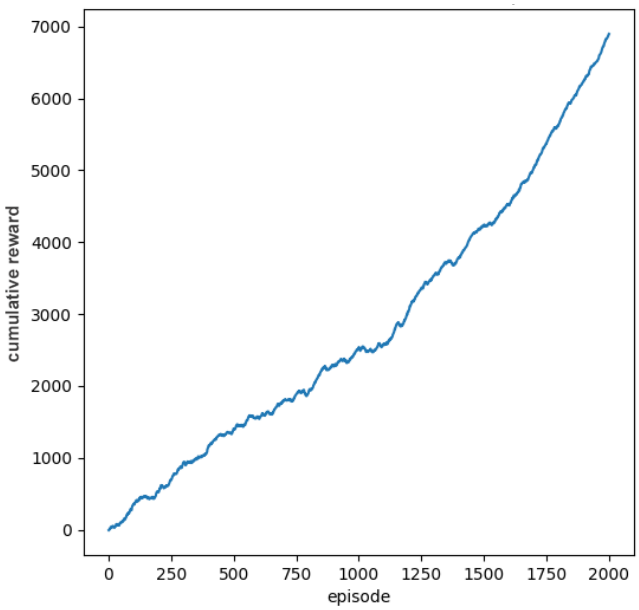
รูปที่ 4.32 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



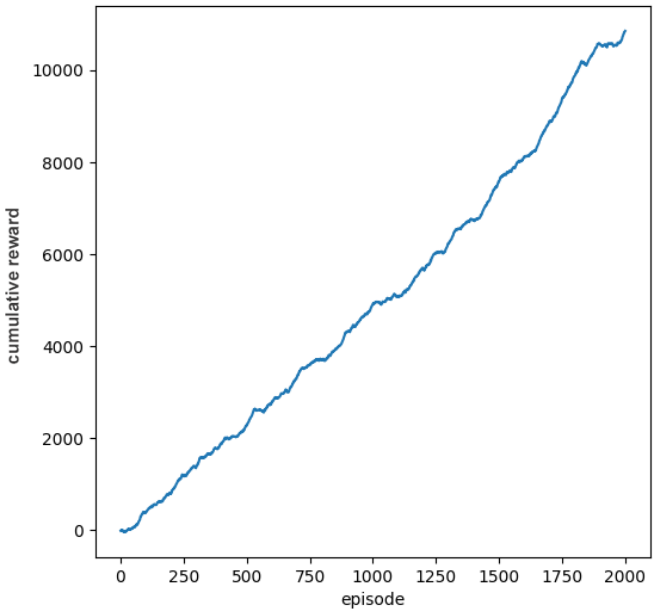
รูปที่ 4.33 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



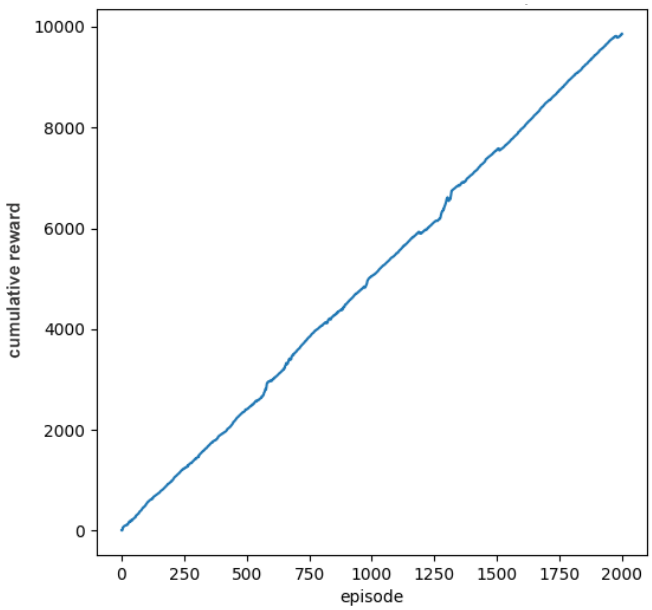
รูปที่ 4.34 ผลรางวัลของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



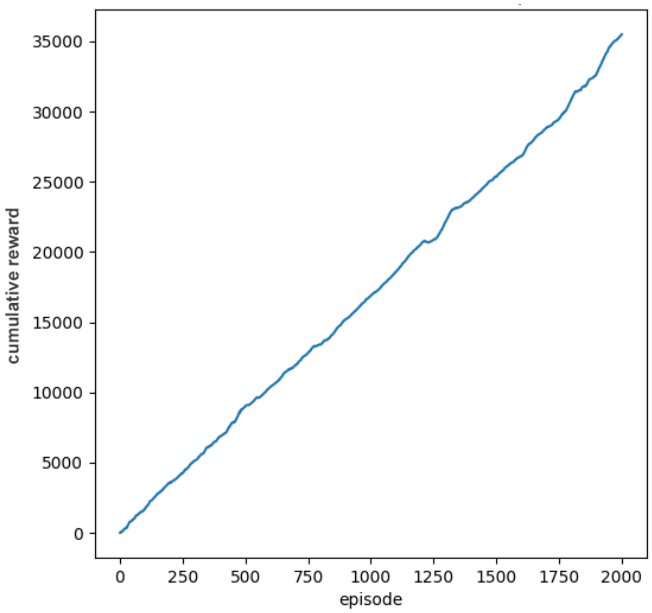
รูปที่ 4.35 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



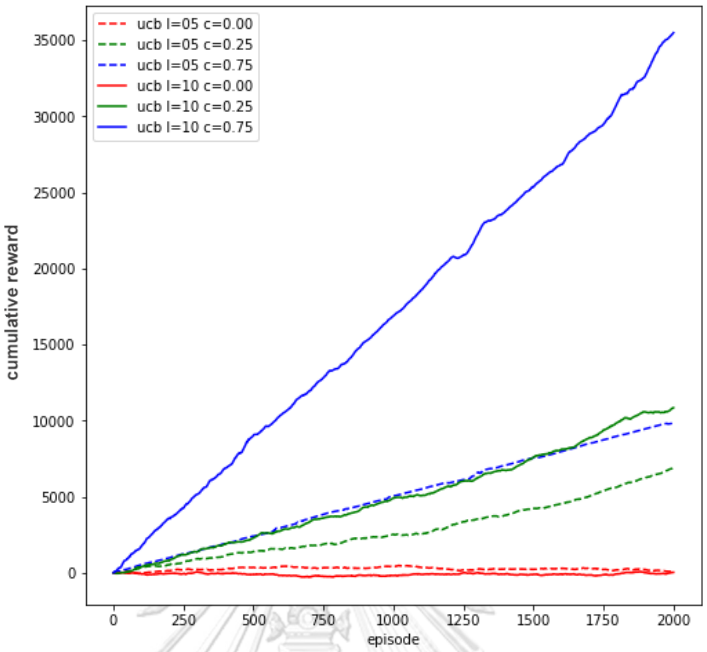
รูปที่ 4.36 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม



รูปที่ 4.37 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 5 เกม



รูปที่ 4.38 ผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกัน 10 เกม

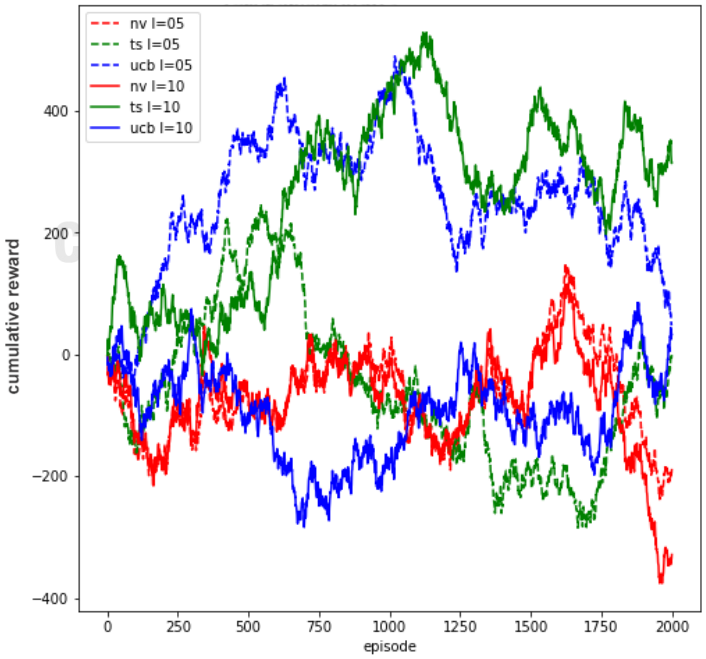


รูปที่ 4.39 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็นที่ต่างกัน และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

4.4 เปรียบเทียบผลรางวัลสะสมของตัวแบบ

การเปรียบเทียบผลจะเปรียบเทียบโดยแสดงในรูปแบบกราฟ ประกอบด้วยข้อมูลผลรางวัลสะสมบนแกนรอบการจำลองที่ตัวแบบทำการเล่นเกมเป่ายิ้งฉุบกับผู้เล่น กำหนดให้การจำลอง 1 รอบ คือการที่ตัวแบบเล่นเป่ายิ้งฉุบกับผู้เล่น จนกระทั่งผู้เล่นแพ้ต่อเนื่องติดต่อกัน l เกม และหยุดเล่นเกม หรือเล่นจนครบ 100 เกม จากนั้นทำการเปรียบเทียบว่าตัวแบบอัลกอริทึมใด ใช้จำนวนรอบการเรียนรู้น้อยกว่าเพื่อให้ได้ผลรางวัลสะสมที่มากกว่า และจะทำการเปรียบเทียบกับตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน โดยรอบของการเรียนรู้ในการเปรียบเทียบทั้งหมดที่ 2,000 รอบ

การเปรียบเทียบผลรางวัลสะสมระหว่างตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน เปรียบเทียบกับตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน เมื่อการตัดสินใจของผู้เล่นเป็นการสุ่มแบบเอกรูป โดยไม่ขึ้นกับผลแพ้ชนะของเกมก่อนหน้า ตัวแบบทั้งสามมีผลรางวัลสะสมกระจัดกระจาย ไม่ว่าผู้เล่นจะหยุดเล่นเมื่อแพ้ติดต่อกัน 5 เกม หรือ 10 เกม ตัวแบบทั้งสามอัลกอริทึมไม่สามารถคาดการณ์รูปแบบผลรางวัลสะสมได้ ดังรูปที่ 4.40

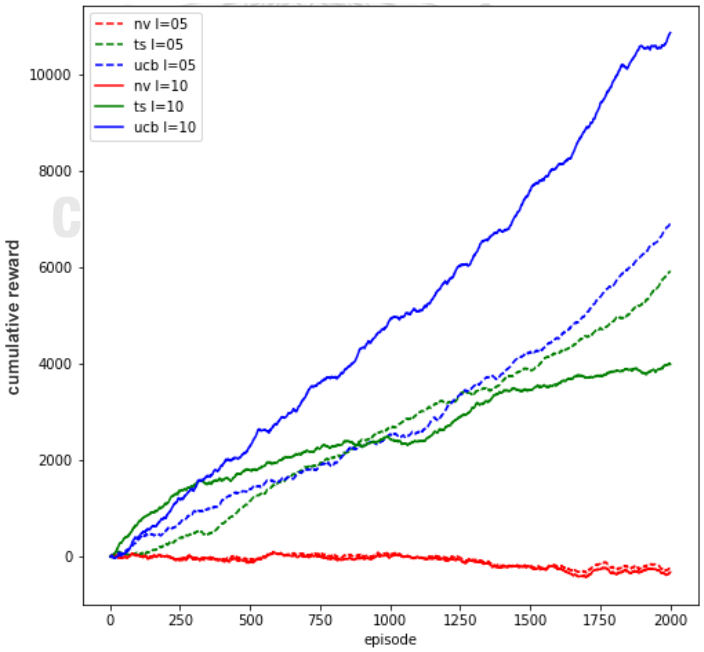


รูปที่ 4.40 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมต่าง ๆ กรณี ผู้เล่นไม่ใช้กลยุทธตามเข็มนาฬิกา และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

การเปรียบเทียบผลรางวัลสะสมระหว่างตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน เปรียบเทียบกับตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน เมื่อการตัดสินใจของผู้เล่นขึ้นกับผลแพ้ชนะของเกมก่อนหน้า 1 เกม ในการเลือกใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 ตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนมีผลรางวัลสะสมแปรผันตรงกับจำนวนรอบ คือผลรางวัลสะสมเพิ่มมากขึ้นเรื่อย ๆ ตามจำนวนรอบที่เพิ่มมากขึ้น ดังรูปที่ 4.41

ในระยะยาวตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนมีผลรางวัลสะสมที่มากกว่าตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน ทั้งในกรณีที่ผู้เล่นใช้กลยุทธยุติการสูญเสียที่ 5 เกม และ 10 เกม ดังรูปที่ 4.41

ข้อสังเกต แม้พฤติกรรมจะมีรูปแบบเพียงเล็กน้อย ตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนจะให้ผลดีกว่าตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน เนื่องจากอัลกอริทึมความเชื่อมั่นขอบเขตบนจะเลือกการกระทำที่มีค่าความเชื่อมั่นมากที่สุด ในขณะที่อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันจะเลือกการกระทำจากการสุ่ม ซึ่งค่าที่สุ่มออกโดยปกติจะใกล้เคียงกับค่าเฉลี่ย การเลือกการกระทำจากขอบเขตบนจึงให้ผลตอบแทนที่สูงกว่าการเลือกการกระทำจากค่าเฉลี่ย

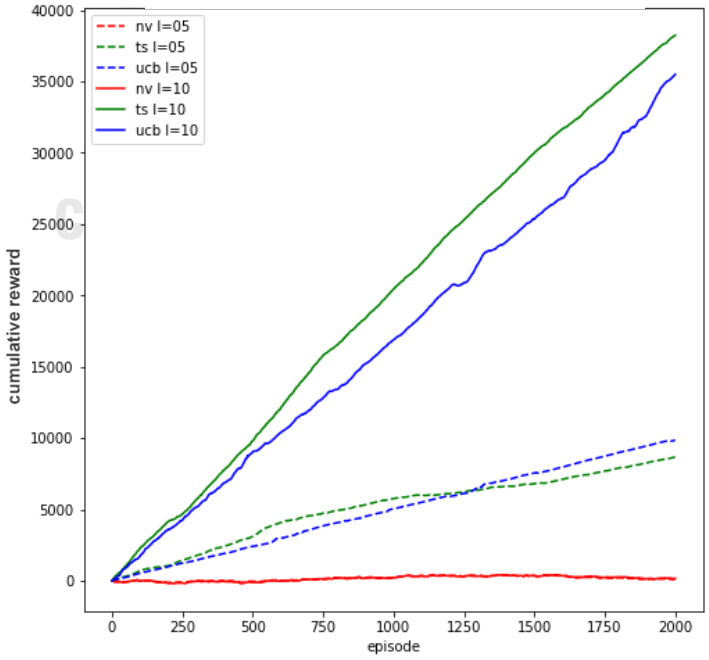


รูปที่ 4.41 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมต่าง ๆ กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.25 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

การเปรียบเทียบผลรางวัลสะสมระหว่างตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน เปรียบเทียบกับตัวแบบอัลกอริทึมการสุ่มแบบเอกรูปที่เป็นตัวแบบบรรทัดฐาน เมื่อการตัดสินใจของผู้เล่นขึ้นกับผลแพ้ชนะของเกมก่อนหน้า 1 เกม ในการเลือกใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 ตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนมีผลรางวัลสะสมแปรผันตรงกับจำนวนรอบ คือผลรางวัลสะสมเพิ่มมากขึ้นเรื่อย ๆ ตามจำนวนรอบที่เพิ่มมากขึ้น ดังรูปที่ 4.42

ในระยะยาวตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบนมีผลรางวัลสะสมที่มากกว่าตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน ทั้งในกรณีที่ผู้เล่นใช้กลยุทธยุติการสูญเสียที่ 5 เกม แต่เมื่อผู้เล่นใช้กลยุทธยุติการสูญเสียที่ 10 เกม ตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันมีผลรางวัลสะสมที่มากกว่าตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน ดังรูปที่ 4.42

ข้อสังเกต เมื่อต้องรอผลจากการกระทำเป็นระยะเวลานาน ตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันมีผลรางวัลสะสมที่มากกว่าตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน สอดคล้องกับงานวิจัยก่อนหน้าว่า หากการตอบสนองที่เกิดจากการกระทำมีความล่าช้า อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันจะให้ประสิทธิภาพที่ดีกว่าอัลกอริทึมความเชื่อมั่นขอบเขตบน (Chapelle & Li, 2011)



รูปที่ 4.42 เปรียบเทียบผลรางวัลสะสมของตัวแบบอัลกอริทึมต่าง ๆ กรณี ผู้เล่นใช้กลยุทธตามเข็มนาฬิกาด้วยความน่าจะเป็น 0.75 และใช้กลยุทธยุติการสูญเสียเมื่อแพ้ติดกันต่างกัน

บทที่ 5
สรุปผลการวิจัยและประโยชน์ที่ได้รับ

5.1 สรุปผลการวิจัย

ตามจุดประสงค์การวิจัยการเปรียบเทียบประสิทธิภาพของตัวแบบอัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบอัลกอริทึมความเชื่อมั่นขอบเขตบน กับปัญหาการตัดสินใจที่มีลำดับโดยมีองค์ประกอบของพฤติกรรมของมนุษย์ และทำการเปรียบเทียบประสิทธิภาพด้วยการเปรียบเทียบผลรางวัลสะสมของตัวแบบทั้งสองอัลกอริทึมในสถานกาณ์จำลองเกมเป่ายิ้งฉุบในสถานการณ์ต่าง ๆ พบว่า

กรณีที่พฤติกรรมของผู้เล่นมีรูปแบบเป็นแบบแผนแม้เพียงเล็กน้อย ตัวแบบที่ใช้อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน และตัวแบบที่ใช้อัลกอริทึมความเชื่อมั่นขอบเขตบน ตัวแบบทั้งสองอัลกอริทึมสามารถตรวจจับแบบแผนพฤติกรรมนั้นและตัดสินใจเลือกการกระทำเพื่อโต้ตอบ และทำการสะสมผลรางวัลในระยะยาวได้อย่างดี นอกจากนี้หากพฤติกรรมของผู้เล่นมีความเป็นแบบแผนที่ชัดเจนมากยิ่งขึ้น ผลรางวัลสะสมของตัวแบบทั้งสองอัลกอริทึมจะยิ่งมากขึ้น แปรผันตามกับความชัดเจนของแบบแผนพฤติกรรมผู้เล่น

กรณีที่พฤติกรรมของผู้เล่นไม่มีรูปแบบ หรือไม่มีแบบแผนใด ๆ ผลรางวัลสะสมของตัวแบบทั้งสองอัลกอริทึมจะมีความกระจัดกระจาย ไม่สามารถคาดการณ์แนวโน้มของผลรางวัลสะสม หรือการันตีผลรางวัลสะสมได้ว่าจะเป็นไปในรูปแบบใด อ้างอิงตามผลรางวัลสะสม รูปที่ 4.40

กรณีที่พฤติกรรมของผู้เล่นมีรูปแบบเป็นแบบแผนที่ไม่ค่อยชัดเจน หรือพฤติกรรมของผู้เล่นมีรูปแบบเป็นแบบแผนชัดเจนในระยะเวลาสั้น ผลรางวัลสะสมในระยะยาวของตัวแบบที่ใช้อัลกอริทึมความเชื่อมั่นขอบเขตบนทำผลรางวัลสะสมได้ดีกว่า ตัวแบบที่ใช้อัลกอริทึมการสุ่มตัวอย่างแบบทอมสัน อ้างอิงตามผลรางวัลสะสม รูปที่ 4.41 และรูปที่ 4.42

กรณีที่พฤติกรรมของผู้เล่นมีรูปแบบเป็นแบบแผนชัดเจนเป็นระยะเวลายาว ผลรางวัลสะสมของตัวแบบที่ใช้อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันทำผลรางวัลสะสมได้ดีกว่า ตัวแบบที่ใช้อัลกอริทึมความเชื่อมั่นขอบเขตบน อ้างอิงตามผลรางวัลสะสม รูปที่ 4.42

5.2 ประโยชน์ที่ได้รับ

งานวิจัยนี้เป็นแนวทางในการศึกษาการตัวแบบการเรียนรู้แบบเสริมแรงที่ใช้ตรวจจับรูปแบบพฤติกรรมของมนุษย์ และการตัดสินใจเลือกการกระทำในสถานะที่แตกต่างกันเพื่อจูงใจให้ได้ผลรางวัลสะสมสูงสุดในระยะยาว ซึ่งทำการเปรียบเทียบประสิทธิภาพของอัลกอริทึมการเรียนรู้แบบเสริมแรงที่มีองค์ประกอบเชิงพฤติกรรมของมนุษย์ ระหว่างอัลกอริทึมทอมสัน และอัลกอริทึมความเชื่อมั่นขอบเขตบน ซึ่งทั้งสองเป็นอัลกอริทึมที่ใช้สถิติคนละแขนง คือ อัลกอริทึมการสุ่มตัวอย่างแบบทอมสันใช้วิธีคิดของสถิติแบบเบย์ (Bayesian) และอัลกอริทึมความเชื่อมั่นขอบเขตบนใช้วิธีคิดของสถิติแบบความถี่ (Frequentist statistics) ซึ่งในการวิจัยนี้หากมนุษย์มีพฤติกรรมที่เป็นแบบแผนเพียงเล็กน้อย ตัวแบบจากทั้งสองอัลกอริทึมสามารถตรวจจับพฤติกรรมได้และตัดสินใจเลือกการกระทำโต้ตอบเพื่อสะสมผลตอบแทนในระยะยาวได้ทั้งคู่ และประสิทธิภาพของตัวแบบทั้งสองอัลกอริทึมจะแตกต่างกันไปในแต่ละสถานการณ์ตาม บทที่ 5.1

5.3 แนวทางการพัฒนาต่อยอด

งานวิจัยนี้ยังมีประเด็นให้ศึกษาต่อยอดได้โดยการนำอัลกอริทึมอัลกอริทึมการสุ่มตัวอย่างแบบทอมสันและอัลกอริทึมความเชื่อมั่นขอบเขตบนไปทดสอบกับผู้เล่นจริงซึ่งจะมีพฤติกรรมหลากหลายมากกว่า 2 พฤติกรรมในงานวิจัยนี้ เพื่อศึกษาว่าในสถานการณ์ที่พฤติกรรมมีความหลากหลายมากขึ้นทั้งสองอัลกอริทึมจะยังให้ประสิทธิภาพอย่างไร หรือสามารถเปลี่ยนจากสถานการณ์เกมเป่ายิ้งฉุบเป็นสถานการณ์อื่น เช่น ระบบการแนะนำข้อมูล ระบบการแนะนำสินค้า ระบบการแนะนำโฆษณา

บรรณานุกรม

Auer, P. (2002). Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, *3*, 397-422.

Auer, P., Cesa-Bianchi, N., et al. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, *47*, 235-256.

Chapelle, O., & Li, L. (2011). An Empirical Evaluation of Thompson Sampling. *Advances in Neural Information Processing Systems*, *24*, 2249–2257.

Daniel, R. J., Van-Roy, B., et al. (2018). A Tutorial on Thompson Sampling.

Gordan, M., & Krishanan, I. A. (2014). A Review of B. F. Skinner's 'Reinforcement Theory of Motivation'. *International Journal of Research in Education Methodology*, *5*, 680-688.

Hai-Jun, Z. (2016). The rock–paper–scissors game. *Contemporary Physics*, *57*(2), 151-163.

Hao, B., Abbasi-Yadkori, Y., et al. (2019). Bootstrapping Upper Confidence Bound. *NeurIPS*, *32*, 12123-12133.

Kaminski, K. M., & Lo, A. W. (2014). When do stop-loss rules stop losses? *Journal of Financial Markets*, *18*, 234–254.

Mccannon, B. C. (2007). Rock Paper Scissors. *Journal of Economics*, *92*(1), 67-88.

Soo-Chang, H., Fu, M. C., et al. (2005). An Adaptive Sampling Algorithm for Solving Markov Decision Processes. *Operation Research*, *53*(1), 126-139.

Sutton, R. S., & Barto, A. G. (1999). The Reinforement Learning Problem. *Journal of Cognitive Neuroscience*, *11*, 126-134.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction. 2nd ed*. MIT Press.

Wang, Z., Xu, B., et al. (2014). Social cycling and conditional responses in the Rock-Paper-Scissors game. *Scientific Reports*, *4*, 1-21.

ประวัติผู้เขียน

ชื่อ-สกุล	ธันยวุฒิ อักขระสมชีพ
วัน เดือน ปี เกิด	9 กุมภาพันธ์ 2536
สถานที่เกิด	กรุงเทพมหานคร
วุฒิการศึกษา	2016, Bachelor of Computer Engineering, Kasetsart University, Thailand
ผลงานตีพิมพ์	Thanyavuth Akarasomcheep, Anan Phonphoem, Chaiporn Jaikaeo, Aphirak Jansang, Nattika Penglee, Natrapee Polyai, and Wichan Mawinthorn. (2016). Data Collection and Analysis System for Measuring Physical Fitness. ECTI Conference on Application Research and Development (ECTI-CARD 2016). 8, 9-12.