การประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงิน

น.ส.ภัควัลย์ จันทรศิริภาส

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาสถิติ ภาควิชาสถิติ
คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

An Application of Reinforcement Learning to Financial Planning

Miss Pakawan Chansiripas

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 2019

หัวข้อวิทยานิพนธ์	การประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงิน
โดย	น.ส.ภัควัลย์ จันทรศิริภาส
สาขาวิชา	สถิติ
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์

คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะพาณิชยศาสตร์และการบัญชี

(รองศาสตราจารย์ ดร.วิเลิศ ภูริวัชร)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ

(อาจารย์ ดร.อัครินทร์ ไพบูลย์พานิช)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก

(รองศาสตราจารย์ ดร.เสกสรร เกียรติสุไพบูลย์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.วิฐรา พึ่งพาพงศ์)

..... กรรมการภายนอกมหาวิทยาลัย

(อาจารย์ ดร.ดลชัย ละออนวล)

ภัควัลย์ จันทรศิริภาส : การประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงิน. (An Application of Reinforcement Learning to Financial Planning) อ.ที่ปรึกษาหลัก : รศ. ดร.เสกสรร เกียรติสุไพบูลย์

งานวิจัยนี้มีวัตถุประสงค์ที่จะนำการเรียนรู้แบบเสริมกำลังมาประยุกต์กับการวางแผนทางการเงินเพื่อตัดสินใจเลือกอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงที่ดีที่สุดในแต่ละช่วงเวลาตลอดช่วงอายุของครัวเรือน ผลลัพธ์ที่ได้จากการเรียนรู้แบบเสริมกำลังซึ่งเป็นค่าประมาณ จะถูกนำมาเปรียบเทียบกับคำตอบที่ถูกต้องจากวิธี MDP สำหรับการเรียนรู้แบบเสริมกำลังในงานวิจัยนี้เป็นอัลกอริธึม SARSA โดยการเลือกการกระทำใช้วิธี ε-greedy ส่วนการประมาณค่าใช้ตัวแบบถดถอยที่มีตัวแปรต้นเป็นฟีเจอร์จากเคอร์เนล Radial Basis Function (RBF) จากการศึกษาพบว่าความผิดพลาดระหว่างค่าประมาณผลลัพธ์ที่ดีที่สุดเทียบกับคำตอบจาก MDP มีแนวโน้มลู่เข้าสู่ศูนย์ แสดงว่าการเรียนรู้แบบเสริมกำลังสามารถประยุกต์กับการวางแผนทางการเงินได้ อย่างไรก็ตาม SARSA แบบดั้งเดิมใช้เวลานานในการเรียนรู้ เมื่อปรับปรุงให้การเลือกการกระทำในช่วงแรกเน้นสำรวจมากขึ้น พบว่า ความผิดพลาดลดลง แสดงให้เห็นว่า SARSA ที่ปรับปรุงให้เน้นการสำรวจในช่วงแรกมีประสิทธิภาพดีขึ้นกว่าแบบดั้งเดิม

นอกจากนี้เมื่อพิจารณาผลของการปรับเปลี่ยนปัจจัยต่างๆ สำหรับ SARSA แบบเน้นการสำรวจในช่วงแรก พบว่า ความผิดพลาดระหว่างค่าประมาณผลลัพธ์ที่ดีที่สุดเทียบกับ MDP มีค่าน้อยสุดเมื่อใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 200 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับ ในขณะที่การนำคำตอบที่ดีที่สุดไปจำลองใช้จริง ผลของการวางแผนทางการเงินที่ได้มีความแตกต่างกับคำตอบจาก MDP มาก โดยการใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 300 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับให้ผลลัพธ์ที่ใกล้เคียงกับ MDP มากสุด แสดงว่าถึงแม้ความผิดพลาดของผลลัพธ์ที่ดีที่สุดจะมีค่าต่ำสุด คำตอบจากวิธีการเรียนรู้แบบเสริมกำลังยังมีความผิดพลาดสูงเมื่อเทียบกับคำตอบจาก MDP

สาขาวิชา สถิติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2562 ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6081586426 : MAJOR STATISTICS
KEYWORD:

Pakawan Chansiripas : An Application of Reinforcement Learning to Financial Planning. Advisor: Assoc. Prof. SEKSAN KIATSUPAIBUL, Ph.D.

In this study a reinforcement learning is applied to a financial planning problem to find an optimal consumption proportion and an optimal investment proportion in risky assets. The solutions from the reinforcement approach are compared with the exact solutions from an MDP approach. The algorithm used in this study is SARSA with $\boldsymbol{\varepsilon}$-greedy action selection where the value approximation employs a regression method with Radial Basis Function (RBF) features. From the experiments, the errors between the optimal value estimated from the reinforcement learning and the exact solution from the MDP have a tendency to converge, indicating the effectiveness of the reinforcement learning in solving a financial planning problem. The algorithm is then adjusted to emphasize more on exploration. The errors from the adjusted algorithm are lower than those from the original algorithm, showing that the adjusted algorithm is more efficient than the original algorithm.

In addition, considering the effects of factor adjustment of the SARSA algorithm focused on exploration in the first stage, it is found that the error between the optimal value of the reinforcement learning and the MDP is lowest when the initial weights from the linear regression model are used with 200 features and the initial decreased learning rate and epsilon are 0.1 and 0.9, respectively. When the optimal actions are used in the simulation, the obtained results of financial planning are very different compared to those from the MDP. The simulation in which 300 features are used instead gives the most similar result to the MDP. This shows that even though the error of the optimal value is lowest, the difference of the result from the reinforcement learning is still high compared to the result from the MDP.

Field of Study:	Statistics	Student's Signature .....
Academic Year:	2019	Advisor's Signature .....

กิตติกรรมประกาศ

ผู้วิจัยขอกราบขอบพระคุณรองศาสตราจารย์ ดร. เสกสรร เกียรติสุไพบูลย์ เป็นอย่างยิ่งที่ได้ให้โอกาสผู้วิจัยได้เป็นลูกศิษย์ในที่ปรึกษา สละเวลาให้คำแนะนำสั่งสอนและคำปรึกษาที่มีประโยชน์ อีกทั้งยังช่วยแก้ไขข้อบกพร่องต่างๆ จนกระทั่งวิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปได้ด้วยดี

ผู้วิจัยขอกราบขอบพระคุณอาจารย์ ดร. อัครินทร์ ไพบูลย์พานิช ประธานกรรมการสอบวิทยานิพนธ์ ผู้ช่วยศาสตราจารย์ ดร. วิฐรา พึ่งพาพงศ์ และอาจารย์ ดร. ดลชัย ละออนวล กรรมการสอบวิทยานิพนธ์ ที่ได้กรุณาสละเวลามาตรวจทานแก้ไขข้อบกพร่องในวิทยานิพนธ์ฉบับนี้ ตลอดจนให้คำแนะนำที่เป็นประโยชน์แก่ผู้วิจัยที่ช่วยให้วิทยานิพนธ์ฉบับนี้สมบูรณ์ยิ่งขึ้น

นอกจากนี้ผู้วิจัยขอกราบขอบพระคุณคณาจารย์ทุกท่านที่ให้ความรู้ทางวิชาการ รวมไปถึงเจ้าหน้าที่ของภาควิชาสถิติที่ช่วยจัดทำเอกสาร และอำนวยความสะดวกในด้านต่างๆ

สุดท้ายนี้ผู้วิจัยขอกราบขอบพระคุณบิดา มารดา และครอบครัวที่คอยสนับสนุนและให้กำลังใจเสมอมา

ภัควัลย์ จันทรศิริภาส

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

อันเนื่องมาจากสภาพสังคมและสภาพเศรษฐกิจ อาทิเช่น ปัญหาสังคมไทยกำลังก้าวเข้าสู่ยุคผู้สูงอายุอย่างรวดเร็ว การวางแผนทางการเงินจึงเป็นกิจกรรมที่ได้รับความนิยมและได้รับการตระหนักถึงความสำคัญมากขึ้น บางครั้งสำหรับนักวางแผนทางการเงินมีคำแนะนำให้ออมเงินหรือลงทุนในผลิตภัณฑ์ทางการเงินในอัตราส่วนต่างๆ ที่คงที่ ซึ่งขัดแย้งกับแนวความคิดของนักเศรษฐศาสตร์ว่าการทำเช่นนั้นอาจเป็นการหลีกเลี่ยงความเสี่ยง และทำให้เกิดการบริโภคมากหรือน้อยเกินไปจากความเป็นจริง (นราพงศ์ ศรีวิศาล และคณะ, 2561)

ภายหลังได้มีการเสนอแนะในการนำ Markov Decision Process (MDP) มาประยุกต์ใช้กับศาสตร์ทางการเงินรวมทั้งการวางแผนทางการเงิน ซึ่ง MDP เป็นการวางกรอบปัญหาในการเรียนรู้จากปฏิกิริยา (Interaction) เพื่อให้บรรลุเป้าหมาย โดยผู้ตัดสินใจ (Agent) จะมีปฏิสัมพันธ์กับสิ่งแวดล้อมภายนอก (Environment) เพื่อที่ผู้ตัดสินใจจะได้เลือกกระทำ (Action) แล้วสิ่งแวดล้อมจะตอบสนองการกระทำนั้นๆ ด้วยการแสดงสถานะใหม่ (State) และให้รางวัล (Reward) ให้แก่ผู้ตัดสินใจซึ่งรางวัลนี้จะเป็นค่าที่ผู้ตัดสินใจต้องการทำให้มีค่ามากที่สุดตามเป้าหมาย จากการประยุกต์ใช้ MDP จะทำให้สามารถตัดสินใจในการกำหนดอัตราส่วนเงินที่ใช้บริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง ส่งผลให้การวางแผนทางการเงินสามารถกำหนดแนวทางในการบริโภคและลงทุนรายปี เนื่องจากสามารถปรับเปลี่ยนอัตราส่วนเงินที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงตามสถานการณ์ในแต่ละปี (Bauerle & Rieder, 2011)

อย่างไรก็ตามการนำ MDP ไปประยุกต์ใช้ก็ยังมีข้อจำกัดกล่าวคือการแก้ปัญหาด้วย MDP จะต้องสามารถร่างปัญหานั้นให้เป็นไปตามรูปแบบของ MDP ได้ ซึ่งมีองค์ประกอบหนึ่งที่สำคัญคือต้องทราบพลวัตของสิ่งแวดล้อม (Dynamics) ซึ่งถูกแสดงผ่านความน่าจะเป็นในการเปลี่ยนไปสู่สถานะหนึ่งเมื่อถูกกำหนดว่าอยู่ในสถานะหนึ่งและเลือกกระทำหนึ่งๆ (State-transition probabilities) และอีกข้อจำกัดคือกรณีที่มีสถานะจำนวนมาก จะทำให้การคำนวณเพื่อแก้ปัญหาหาคำตอบของ MDP ต้องใช้ทรัพยากรจำนวนมากนั่นคือใช้เวลานานหรือประสิทธิภาพของคอมพิวเตอร์สูง หรือแม้กระทั่งอาจไม่สามารถหาคำตอบได้ในบางครั้ง

จากข้อจำกัดข้างต้นจึงมีการนำการเรียนรู้แบบเสริมกำลัง (Reinforcement Learning) เข้ามาช่วยโดยอาศัยกรอบโครงร่างของ MDP ในการแสดงสิ่งแวดล้อม แต่ไม่จำเป็นต้องทราบพลวัตของสิ่งแวดล้อม นอกจากนี้สำหรับปัญหาเมื่อมีสถานะจำนวนมากจะถูกหาคำตอบด้วยการประมาณจากการวางนัยทั่วไป (Generalization) คือการประมาณสถานะปัจจุบันจากสถานะที่เคยพบก่อนหน้า หรืออาจเรียกว่าเป็นการประมาณฟังก์ชัน (Function Approximation) แทนการหาคำตอบที่แท้จริงดังเช่นใน MDP ทำให้สามารถประหยัดทรัพยากรที่ใช้ในการคำนวณลงได้ (Sutton & Barto, 2018)

งานวิจัยนี้จึงจะศึกษาการประยุกต์ใช้การเรียนรู้แบบเสริมกำลังในการวางแผนทางการเงินสำหรับครัวเรือนที่มีมูลค่าปัจจุบันของทรัพย์สิน (Wealth) และสุขภาพของหัวหน้าครัวเรือน (Health) ในระดับต่างๆ เพื่อกำหนดอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภค (Consumption) และในการลงทุนในสินทรัพย์ที่มีความเสี่ยง (Risky Investment) แล้วจึงนำผลที่ได้ไปเปรียบเทียบกับผลจากการแก้ปัญหาด้วย MDP

1.2 วัตถุประสงค์ของการวิจัย

1.2.1 เพื่อศึกษาการประยุกต์ใช้การเรียนรู้แบบเสริมกำลังในการวางแผนทางการเงิน เช่น อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภค และการลงทุนในสินทรัพย์ที่มีความเสี่ยง

1.2.2 เพื่อเปรียบเทียบผลที่ได้จากการเรียนรู้แบบเสริมกำลังและ MDP โดยหาความผิดพลาดระหว่าง Optimal value ที่ได้จากทั้งสองวิธี

1.3 คำจำกัดความที่ใช้ในการวิจัย

$E[X]$	คือ ค่าคาดหวังของตัวแปรสุ่ม X โดย $E[X] = \sum_x p(x)x$
$\text{argmax}_a f(a)$	คือ การหาค่า a ที่ทำให้ $f(a)$ มีค่ามากที่สุด
$1_{predicate}$	คือ ฟังก์ชันอินดิเคเตอร์ โดยมีค่าเท่ากับ 1 เมื่อ predicate เป็นจริงและมีค่าเท่ากับ 0 เมื่อ predicate เป็นเท็จ
$\in$	คือ การเป็นสมาชิก
R	คือ เซตของจำนวนจริง

1.4 ขอบเขตของการวิจัย

1.4.1 ข้อมูลสำหรับงานวิจัยนี้ถูกเลือกมาจากงานวิจัยที่ทำการสำรวจพฤติกรรมของครัวเรือนที่อยู่นอกชุมชนเมือง จังหวัดตราด โดยกลุ่มเป้าหมายที่ทำการสำรวจครอบคลุมประชากรในวัยทำงาน

และวัยเกษียณ ซึ่งทำการเก็บข้อมูลจากกลุ่มตัวอย่างในจังหวัดตราด โดยข้อมูลที่เก็บรวบรวมเกี่ยวข้องกับการจัดสรรรายได้ รายจ่าย การออม การเป็นสมาชิกสถาบันหรือองค์กรชุมชน รวมถึงประโยชน์ที่ได้จากการเป็นสมาชิก (ณัตติฤดี เจริญรักษ์ และ สุภารัตน์ ตันทะนงศักดิ์กุล, 2559)

1.4.2 กรอบปัญหาเป็นปัญหาแบบที่เวลาไม่เป็นอนันต์ (Episodic)

1.4.3 เซตของสถานะเป็นเซตของข้อมูลแบบต่อเนื่อง (Continuous)

1.4.4 เซตของการกระทำเป็นเซตของข้อมูลแบบไม่ต่อเนื่อง (Discrete)

1.5 ประโยชน์ที่คาดว่าจะได้รับ

จากการศึกษาจะทำให้ทราบอัตราส่วนเงินที่ใช้ในการบริโภค และอัตราส่วนในการลงทุนในสินทรัพย์ที่มีความเสี่ยง รวมทั้ง Learning Curve ที่ได้จากการแก้ปัญหาด้วยการเรียนรู้แบบเสริมกำลังเทียบกับ MDP ซึ่งสามารถนำไปใช้เป็นแนวทางในการวางแผนและบริหารทางการเงิน และยังสามารถนำไปใช้เป็นแนวทางในการเลือกวิธีการในการแก้ปัญหาได้เมื่อปัจจัยในการวางแผนทางการเงินมีจำนวนมากขึ้นได้

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

2.1 องค์ประกอบของ MDP

2.1.1 นโยบาย (Policy)

คือการจับคู่ (Map) สถานะ (State) ของสิ่งแวดล้อม (Environment) กับความน่าจะเป็นในการเลือกแต่ละการกระทำ โดย Policy อาจเป็นฟังก์ชันอย่างง่าย หรืออาจเป็นกระบวนการที่ซับซ้อนต้องใช้การคำนวณเข้ามาช่วย โดยทั่วไปแล้ว Policy จะมีความไม่แน่นอน (Stochastic Policy)

หาก Agent ดำเนินการตาม Policy π ณ เวลา t แล้ว $\pi(a|s)$ เป็นความน่าจะเป็นที่เลือกการกระทำ $a_t = a$ ที่สถานะ $s_t = s$

2.1.2 รางวัล (Reward)

รางวัลเป็นการกำหนดเป้าหมายกล่าวคือที่ ณ เวลาหนึ่งๆ สิ่งแวดล้อมจะส่งค่ารางวัลให้แก่ Agent โดย Agent นั้นมีจุดประสงค์ต้องการให้รางวัลที่ได้มีค่ามากสุดในระยะยาว รางวัลจึงเป็นสิ่งที่บอกว่าเหตุการณ์นั้นดีหรือแย่สำหรับ Agent ดังนั้นรางวัลจึงเป็นสิ่งที่ทำให้ Policy เปลี่ยน เนื่องมาจากเมื่อ Policy กำหนดการกระทำที่ทำให้ได้รางวัลที่มีค่าน้อย Policy ก็จะถูกเปลี่ยนให้ไปเลือกการกระทำอื่นๆ ในครั้งต่อไป โดยทั่วไปแล้วรางวัลจะเป็นฟังก์ชันของสถานะและการกระทำที่ถูกเลือก

2.1.3 Value Function

เป็นสิ่งที่แสดงให้เห็นว่าอะไรดีในระยะยาว กล่าวคือ Value ของสถานะหนึ่งๆ มีค่าเท่ากับผลรวมของรางวัลที่คาดว่าจะได้รับในอนาคตนับตั้งแต่สถานะนั้น ซึ่งบางครั้งถูกเรียกว่า ผลตอบแทน (Return) ถูกแทนด้วยสัญลักษณ์ G_t ยกตัวอย่างเช่นที่สถานะหนึ่งมักได้รางวัลที่มีค่าน้อยแต่กลับได้ Value ที่มีค่าสูงเนื่องมาจากที่สถานะอื่นๆ หลังจากสถานะนั้นกลับได้รางวัลที่มีค่าสูง

2.1.4 โมเดล (Model)

โมเดลเป็นสิ่งที่เสมือนจำลองพฤติกรรมของสิ่งแวดล้อมไว้ เช่น ที่สถานะและการกระทำหนึ่งๆ โมเดลจะพยากรณ์สถานะและรางวัลที่จะเกิดขึ้นถัดไป

2.2 Markov Decision Process (MDP)

MDP เป็นการวางกรอบปัญหาในการเลือกตัดสินใจโดยที่การกระทำที่เลือกจะส่งผลต่อทั้งรางวัลในขณะนั้นรวมทั้งสถานะและรางวัลในเวลาถัดมา โดย MDP นั้นต้องการหา Policy ที่ให้รางวัลในระยะยาวมากที่สุดจากการประมาณ Value ของแต่ละการกระทำในแต่ละสถานะหรือ Value ของแต่ละสถานะ

ที่แต่ละเวลา t จะต้องตัดสินใจเลือกการกระทำ a_t จากเซตของการกระทำที่เป็นไปได้ทั้งหมด $A_t(s_t)$ ซึ่งแต่ละการกระทำ a_t จะถูกเลือกโดยพิจารณาจากสถานะในตอนนั้น s_t ดังนั้นอาจถือได้ว่าการตัดสินใจเลือกการกระทำเป็นการทำตาม Policy ซึ่งเป็นฟังก์ชัน A_t^π ที่จับคู่ระหว่างแต่ละสถานะกับการกระทำ

$$a_t = A_t^\pi(s_t) \in A_t(s_t)$$

ซึ่งในแต่ละสิ่งแวดล้อมจะมี Policy ที่สามารถเลือกได้หลากหลาย ซึ่ง Policy ทั้งหมดที่สามารถเลือกได้จะถูกแสดงด้วยสัญลักษณ์ Π

โดย Value ของสถานะ s ภายใต้ Policy π จะถูกแสดงโดยสัญลักษณ์ $v_\pi(s)$ ซึ่งหมายถึงค่าคาดหวังของผลตอบแทนเมื่อเริ่มต้นจากสถานะ s และดำเนินการตาม Policy π โดยฟังก์ชัน v_π ถูกเรียกว่า ฟังก์ชัน State-value สำหรับ Policy π ซึ่งสามารถเขียนในรูปสมการได้ดังนี้ สำหรับทุกสถานะ $s \in S$

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi[G_t|s_t = s] \\ &= \mathbb{E}_\pi[r_{t+1} + \alpha G_{t+1}|s_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\left[r + \alpha \mathbb{E}_\pi[G_{t+1}|s_{t+1} = s']\right] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r + \alpha v_\pi(s')] \end{aligned}$$

เมื่อ G_t คือผลตอบแทน, α คืออัตราคิดลด, การกระทำ a มาจากเซต $A(s)$, สถานะถัดไป s' มาจากเซต S และรางวัล r มาจากเซต R ซึ่งสมการข้างต้นนี้ถูกเรียกว่า Bellman equation for v_π

การหาทางเลือกที่ดีที่สุด (Optimal Policy) ซึ่งถูกแทนด้วย π_* ที่ให้ค่า Optimal State-value function ที่แทนด้วย v_* และถูกนิยามด้วย

$$v_*(s) = \max_\pi v_\pi(s) \quad \text{สำหรับทุกสถานะ } s \in S$$

ซึ่ง Value function สามารถแสดงได้ด้วย Bellman equation ส่งผลให้การหาทางเลือกที่ดีที่สุดสามารถแสดงในรูปของ Bellman equation ได้ดังนี้

$$
\begin{aligned}
v_*(s) &= \max_a \mathbb{E}_\pi[G_t|s_t = s, a_t = a] \\
&= \max_a \mathbb{E}_\pi[r_{t+1} + \alpha v_*(s_{t+1})|s_t = s, a_t = a] \\
&= \max_a \sum_{s',r} p(s', r|s, a)[r + \alpha v_*(s')]
\end{aligned}
$$

ในทางเดียวกัน Value สำหรับเลือกการกระทำ a ที่สถานะ s ภายใต้ Policy π จะถูกแสดงด้วยสัญลักษณ์ $q_\pi(s, a)$ ซึ่งหมายถึงค่าคาดหวังของผลตอบแทนเมื่อเริ่มต้นจากสถานะ s เลือกการกระทำ a และดำเนินการตาม Policy π โดยฟังก์ชัน q_π ถูกเรียกว่า ฟังก์ชัน Action-value สำหรับ Policy π ซึ่งทางเลือกที่ดีที่สุดโดยการพิจารณาจาก Action-value ก็สามารถแสดงในรูปของ Bellman equation ได้เช่นกันดังนี้

$$
\begin{aligned}
q_*(s, a) &= \mathbb{E}\left[r_{t+1} + \alpha \max_{a'} q_*(s_{t+1}, a')|s_t = s, a_t = a\right] \\
&= \sum_{s',r} p(s', r|s, a)\left[r + \alpha \max_{a'} q_*(s', a')\right]
\end{aligned}
$$

จากสมการการหาทางเลือกที่ดีที่สุดทั้งจากการพิจารณาจาก State-value และ Action-value จะพบว่าการใช้ MDP ในการหาทางเลือกที่เหมาะสมจะมีข้อจำกัด นั่นคือ การหาค่า Value นั้นขึ้นกับสถานะ ดังนั้นถ้าหากจำนวนสถานะมีจำนวนมากขึ้น การคำนวณสำหรับทุกสถานะทำให้ใช้ทรัพยากรมากขึ้นเป็นอย่างมากหรืออาจไม่สามารถหาคำตอบได้ นอกจากนี้ในสมการข้างต้น การคำนวณจะประกอบด้วยส่วนของความน่าจะเป็นที่จะเปลี่ยนไปในสถานะถัดไปจากแต่ละสถานะในขณะนั้น ทำให้การทราบความน่าจะเป็นนี้มีความจำเป็นในการหาทางเลือกที่ดีที่สุด

2.3 การเรียนรู้แบบเสริมกำลัง

จากข้อจำกัดของ MDP คือต้องทราบความน่าจะเป็นในการเปลี่ยนแปลงระหว่างสถานะ (State-transition probabilities) หรือพลวัตของสิ่งแวดล้อม (Dynamics) และอาจไม่สามารถหาคำตอบได้เมื่อมีสถานะจำนวนมาก จึงมีการใช้การเรียนรู้แบบเสริมกำลังในการแก้ปัญหาเพื่อหา Policy ที่ดีที่สุด

สำหรับการเรียนรู้แบบเสริมกำลัง การหา Policy ที่ดีที่สุดทำได้จากการประมาณ Value function ทั้ง State-value ซึ่งจะประมาณ v_π ด้วย $\hat{v}$ และ Action-value ที่จะประมาณ q_π ด้วย $\hat{q}$ โดยในที่นี้จะพิจารณาเฉพาะ Action-value

2.3.1 การประมาณ Action-value

2.3.1.1 การประมาณค่าฟังก์ชัน Action-value

การประมาณค่าฟังก์ชัน (Function Approximation) สามารถประมาณได้หลากหลายวิธีตามหลักของการเรียนรู้แบบมีผู้สอน (Supervised Learning) เช่น ฟังก์ชันเส้นตรง (Linear Methods), โครงข่ายประสาทเทียม (Artificial Neural Network), ต้นไม้การตัดสินใจ (Decision Tree) เป็นต้น ซึ่งในที่นี้จะกล่าวถึงเฉพาะฟังก์ชันเส้นตรง

การประมาณค่าฟังก์ชันด้วยฟังก์ชันที่เป็นเส้นตรงของเวกเตอร์น้ำหนัก $\mathbf{b}$ ก็คือผลคูณภายในระหว่าง $\mathbf{b}$ และ $\mathbf{F}(s,a)$ ซึ่งสามารถแสดงได้ตามสมการ

$$\hat{q}(s,a,\mathbf{b}) = \mathbf{b}^{\mathrm{T}}\mathbf{F}(s,a) = \sum_{i=1}^{d} b_i F_i(s,a)$$

เมื่อเวกเตอร์ $\mathbf{F}(s,a) = \big(F_1(s,a), F_2(s,a), \dots, F_d(s,a)\big)^{\mathrm{T}}$ คือเวกเตอร์ฟีเจอร์ที่แสดงถึงสถานะและการกระทำ (Feature Vector) และแต่ละองค์ประกอบ $F_i(s,a)$ ของ $\mathbf{F}$

2.3.1.2 การอัพเดทค่าประมาณ Action-value

การประมาณ Action-value $q_\pi(s,a)$ ด้วย $\hat{q}(s,a)$ ด้วยการเรียนรู้แบบเสริมกำลังสามารถทำได้จากหลากหลายอัลกอริทึม เช่น Monte Carlo (MC), Temporal-Difference (TD), SARSA, Q-Learning เป็นต้น ซึ่งในที่นี้กล่าวถึงเฉพาะ SARSA

อัลกอริทึม SARSA จะพิจารณาการเปลี่ยนแปลงจากคู่ของสถานะ-การกระทำ (State-action pair) หนึ่งไปยังอีกคู่ของสถานะ-การกระทำ และเรียนรู้ Value จากคู่สถานะ-การกระทำ และการประมาณจะใช้แนวคิดของ Incremental Implementation คือการค่อยๆ อัพเดทค่าประมาณทีละน้อยอย่างต่อเนื่องในแต่ละลำดับเวลา ซึ่งสามารถเขียนในรูปแบบทั่วไปได้ดังสมการต่อไปนี้

$$NewEstimate \leftarrow OldEstimate + StepSize[Target - OldEstimate]$$

โดย $[Target - OldEstimate]$ คือความผิดพลาด (Error) ของการประมาณ ซึ่งสามารถลดได้โดยการประมาณให้ใกล้เคียงเป้าหมาย (Target) มากขึ้น ดังนั้นเป้าหมายจึงเป็นตัวบ่งบอกทิศทางที่ต้องการ

โดยการอัพเดทสามารถทำได้เมื่อเวลาผ่านไปหนึ่งลำดับเวลา ที่เวลา $t+1$ จะมีการสร้างเป้าหมายโดยอาศัยค่ารางวัลจากการสังเกต r_{t+1} และค่าประมาณ $\hat{q}(s_{t+1}, a_{t+1})$ เพื่ออัพเดทค่าประมาณที่เวลา t ซึ่งเป็นไปตามสมการต่อไปนี้

$$\hat{q}(s_t, a_t) \leftarrow \hat{q}(s_t, a_t) + \eta[r_{t+1} + \alpha\hat{q}(s_{t+1}, a_{t+1}) - \hat{q}(s_t, a_t)]$$

เมื่อ η คืออัตราการเรียนรู้ (Learning Rate)

การอัพเดทนั้นจะเกิดทุกครั้งภายหลังการเปลี่ยนแปลงจากสถานะ s_t ที่ไม่ใช่สถานะสุดท้าย ถ้าสถานะ s_{t+1} เป็นสถานะสุดท้ายแล้ว $q(s_{t+1}, a_{t+1})$ จะมีค่าเท่ากับศูนย์ การอัพเดทนี้จึงใช้ข้อมูลจาก $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$

และสำหรับการหาทางเลือกที่ดีที่สุดของ SARSA จะเป็นอัลกอริทึมแบบ On-policy คือวิธีที่จะประมาณและปรับปรุง Policy ซึ่งเป็น Policy เดียวกับที่ใช้ในกระบวนการตัดสินใจ ดังนั้นจะประมาณค่า q_π สำหรับ Policy π ในเวลาเดียวกันกับการเปลี่ยน π ให้เข้าสู่ทางเลือกทีดีที่สุดเมื่อพิจารณาจาก q_π

2.3.1.3 การอัพเดทค่าประมาณฟังก์ชัน Action-value

สำหรับการประมาณค่าฟังก์ชัน Action-value ด้วยฟังก์ชันเส้นตรง โดยทั่วไปแล้วไม่มีเวกเตอร์น้ำหนัก $\mathbf{b}$ ที่ทำให้ค่าประมาณ Action-value สำหรับทุกสถานะทุกต้องทั้งหมด จึงต้องการทำให้ค่าเฉลี่ยของความคลาดเคลื่อนระหว่างเป้าหมายและค่าประมาณยกกำลังสอง (Mean Squared Value Error, $\overline{VE}$) มีค่าต่ำที่สุด โดยการหาจุดที่เหมาะสมสมบูรณ์ (Global Optimum) ซึ่งมีเวกเตอร์น้ำหนัก $\mathbf{b}^*$ ที่ทำให้ $\overline{VE}(\mathbf{b}^*) \leq \overline{VE}(\mathbf{b})$ สำหรับทุกค่าที่เป็นไปได้ของ $\mathbf{b}$

การเรียนรู้ในการประมาณค่าฟังก์ชันจะทำโดยวิธี Semi-Gradient ซึ่งจะพยายามปรับค่าเวกเตอร์น้ำหนักทีละน้อยในทิศทางที่ทำให้ $\overline{VE}$ น้อยลง ตามสมการต่อไปนี้

$$\mathbf{b} \leftarrow \mathbf{b} + \eta[r + \alpha\hat{q}(s', a', \mathbf{b}) - \hat{q}(s, a, \mathbf{b})]\nabla\hat{q}(s, a, \mathbf{b})$$

เมื่อ $\nabla f(\mathbf{b})$ คือ เวกเตอร์ของอนุพันธ์ย่อยเทียบกับแต่ละค่าของเวกเตอร์น้ำหนัก $\mathbf{b}$ นั่นคือ

$$\nabla f(\mathbf{b}) = \left(\frac{\partial f(\mathbf{b})}{\partial b_1}, \frac{\partial f(\mathbf{b})}{\partial b_2}, \ldots, \frac{\partial f(\mathbf{b})}{\partial b_d}\right)^{\mathrm{T}}$$

สุดท้ายแล้วเวกเตอร์น้ำหนัก $\mathbf{b}$ สำหรับการประมาณด้วยฟังก์ชันเส้นตรงจะลู่เข้าสู่จุดที่อยู่ใกล้จุดที่เหมาะสมเฉพาะที่ (Local Optimum)

2.3.2 วิธีการในการเลือกการกระทำ (Action Selection Method)

เมื่อพิจารณาที่เวลาหนึ่งๆ จะมีอย่างน้อยหนึ่งการกระทำที่ทำให้ค่าประมาณของ Value มีค่าสูงสุด การกระทำนั้นจะถูกเรียกว่า Greedy Action และเมื่อเลือกกระทำการกระทำประเภทนี้จะเป็นการแสวงประโยชน์ (Exploitation) นั่นคือการเลือกการกระทำที่เคยกระทำในอดีต ในทางกลับกันถ้าเลือกการกระทำที่เป็น Nongreedy Action จะถือเป็นการสำรวจ (Exploration) นั่นคือ

การลองเลือกการกระทำใหม่ๆ เพื่อให้ได้ประโยชน์ในอนาคต เพราะเป็นการปรับปรุงการประมาณ Value ของ Nongreedy Action

ในการทำให้การแสวงประโยชน์และการสำรวจมีความสมดุลกัน สามารถใช้วิธีในการเลือกการกระทำได้หลากหลายวิธี โดยวิธีที่ซับซ้อนน้อยที่สุดคือเลือกหนึ่งในการกระทำที่ให้ค่าประมาณของ Value มากที่สุดนั่นคือการกระทำที่เป็น Greedy Action ดังนี้

$$a_t = \underset{a}{\operatorname{argmax}}\, \widehat{q_t}(a)$$

เมื่อ argmax_a แสดงถึงการกระทำ a ที่ทำให้สมการข้างหลังมีค่ามากที่สุด การเลือกการกระทำที่เป็น Greedy Action จะทำให้เกิดการแสวงประโยชน์เพื่อให้ได้รางวัลมากที่สุดเสมอ ส่งผลให้ไม่เกิดการลองสุ่มการกระทำที่ให้ค่า Value น้อยกว่าเพื่อเรียนรู้ว่าบางการกระทำอาจจะมีโอกาสให้รางวัลที่ดีกว่า

อีกวิธีในการเลือกการกระทำคือโดยส่วนใหญ่จะเลือก Greedy Action แต่บางครั้งสำหรับความน่าจะเป็น ε จะสุ่มเลือกการกระทำอย่างอิสระโดยไม่ขึ้นกับค่าประมาณของ Value ด้วยความน่าจะเป็นที่แต่ละการกระทำจะถูกเลือกเท่ากัน วิธีนี้ถูกเรียกว่า ε-greedy ข้อดีของวิธีนี้คือเมื่อลำดับเวลา (Time-step) เพิ่มขึ้นจะทำให้ $\widehat{q_t}(a)$ ลู่เข้าสู่ค่า $q_*(a)$

2.3.3 อัลกอริทึม SARSA (Poole & Mackworth, 2017; Szepesvári, 2010)

Input: a differentiable function with respect to $\mathbf{b}$, $\hat{q}: \mathcal{S} \times \mathcal{A} \times \mathbb{R}^{\mathrm{d}} \rightarrow \mathbb{R}$

Output: a weight vector $\mathbf{b} \in \mathbb{R}^{\mathrm{d}}$

Initialize value-function weights $\mathbf{b}$ arbitrarily

For each episode

Observe s, a (Select action by Exploration Techniques e.g. ε-greedy)

For each time-step $t \in T$

Take action a, Observe r, s'

If s' is terminal:

$$\mathbf{b} \leftarrow \mathbf{b} + \eta[r - \hat{q}(s, a, \mathbf{b})]\nabla\hat{q}(s, a, \mathbf{b})$$

Go to next episode

Choose a' as a function of $\hat{q}(s', \cdot, \mathbf{b})$ by Exploration Techniques

$$\mathbf{b} \leftarrow \mathbf{b} + \eta[r + \alpha\hat{q}(s', a', \mathbf{b}) - \hat{q}(s, a, \mathbf{b})]\nabla\hat{q}(s, a, \mathbf{b})$$

$$s \leftarrow s'$$

$$a \leftarrow a'$$

2.4 การกำหนดรูปแบบของปัญหา

เนื่องจากการแก้ปัญหาการวางแผนทางการเงินด้วยการหาทางเลือกที่ดีที่สุดหรือที่ทำให้เกิดประโยชน์สูงสุดด้วยวิธีที่สนใจศึกษา นั่นคือการเรียนรู้แบบเสริมกำลังและ MDP จะต้องมีการกำหนดรูปแบบของปัญหา (Problem Formulation) ดังต่อไปนี้

ปัญหาการวางแผนทางการเงินคือการหาค่าที่เหมาะสมในการบริโภคและการลงทุนสำหรับแต่ละเวลาตลอดช่วง T ปี ซึ่งจะกำหนดโครงร่างของปัญหาเพื่อใช้ในการหาทางเลือกที่ดีที่สุด โดยใช้แบบจำลองที่ใช้สำหรับอธิบายพฤติกรรมการจัดสรรรายได้ตลอดช่วงชีวิต เพื่อให้ได้รับอรรถประโยชน์จากการบริโภคที่เกิดขึ้นตลอดชีวิตมีค่าสูงสุด (Sripakdeevong, Stantcheva, & Townsend, 2011)

เมื่อพิจารณาครัวเรือนหนึ่งที่มีการเปลี่ยนแปลงในแต่ละเวลาแบบไม่ต่อเนื่องตลอดช่วงระยะเวลา T ปี ที่แต่ละเวลา $t = 0, \ldots, T$ สถานะของครัวเรือนจะถูกอธิบายด้วยตัวแปรของสถานะ s_t ซึ่งอยู่ในรูปดังต่อไปนี้

S = เวกเตอร์ที่แสดงลักษณะของครัวเรือน

$\mathcal{S}$ = เซตของลักษณะที่เป็นไปได้ทั้งหมดของครัวเรือน

โดยลักษณะของครัวเรือนจะถูกอธิบายด้วย 2 ลักษณะคือ

1. สุขภาพของหัวหน้าครัวเรือนที่ต้นปี t ซึ่งถูกแสดงด้วยสัญลักษณ์ h_t มีค่าที่เป็นไปได้ 3 แบบ ได้แก่ แข็งแรง, พิการ, และเสียชีวิต

2. มูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปี t ซึ่งถูกแสดงด้วยสัญลักษณ์ w_t มีค่าที่เป็นไปได้แบบต่อเนื่องในช่วงตั้งแต่ 1 ถึง $10w_0$ เมื่อ w_0 คือมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปีแรก

และที่แต่ละเวลา $t = 0, \dots, T-1$ ครัวเรือนจะต้องตัดสินใจเพื่อวางแผนทางการเงินหรือเลือกการกระทำ a_t โดยครัวเรือนสามารถเลือกกำหนดได้ 2 ค่าคือ

1. อัตราส่วนของมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ใช้ในการบริโภคที่สามารถเปลี่ยนแปลงได้ที่ต้นปี t ซึ่งถูกแสดงด้วยสัญลักษณ์ c_t และมีค่าที่เป็นไปได้ 21 ค่าแบบไม่ต่อเนื่องในช่วงตั้งแต่ 0.01 ถึง 0.99 โดยมีรายละเอียดดังต่อไปนี้
   - ค่าน้อยสุดที่เป็นไปได้คือ 0.01
   - ค่าในช่วงตั้งแต่ 0.05 ถึง 0.95 ถูกแบ่งเป็น 18 ช่วงเท่าๆ กัน ทำให้ได้ค่าที่เป็นไปได้ทั้งหมด 19 ค่า
   - ค่ามากสุดที่เป็นไปได้คือ 0.99

2. อัตราส่วนของมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่เหลือจากการบริโภคที่ใช้ในการลงทุนในทรัพย์สินที่มีความเสี่ยงที่ต้นปี t ซึ่งถูกแสดงด้วยสัญลักษณ์ z_t และมีค่าที่เป็นไปได้ 21 ค่าแบบไม่ต่อเนื่องในช่วงตั้งแต่ 0 ถึง 1 โดยมีรายละเอียดดังต่อไปนี้
   - ค่าน้อยสุดที่เป็นไปได้คือ 0
   - ค่าในช่วงตั้งแต่ 0.05 ถึง 0.95 ถูกแบ่งเป็น 18 ช่วงเท่าๆ กัน ทำให้ได้ค่าที่เป็นไปได้ทั้งหมด 19 ค่า
   - ค่ามากสุดที่เป็นไปได้คือ 1

นอกจากนี้ กรอบปัญหานี้ยังมีความไม่แน่นอนอันเป็นผลมาจากรายได้สุทธิของครัวเรือน และอัตราผลตอบแทนจากการลงทุนในสินทรัพย์ที่มีความเสี่ยง (Risky Return) สำหรับความไม่แน่นอนจากรายได้สุทธิของครัวเรือนที่เกิดขึ้นในแต่ละช่วงเวลาตลอดเวลา T ปีแสดงด้วย $y_1, \dots, y_T$ และมีการแจกแจงอย่างอิสระต่อกันและเหมือนกันแบบ $\log y_t \sim N\left(-\sigma_y^2/2, \sigma_y\right)$ และสำหรับความ

ไม่แน่นอนจากอัตราผลตอบแทนจากการลงทุนในสินทรัพย์ที่มีความเสี่ยง ซึ่งถูกแสดงด้วยสัญลักษณ์ r_t^r มีการกระจายเป็นไปตาม $\log r_t^r \sim N(\mu_r, \sigma_r)$

ความไม่แน่นอนและการตัดสินใจที่เกิดขึ้นในแต่ละช่วงเวลา t จะส่งผลต่อการเปลี่ยนแปลงมูลค่าปัจจุบันของทรัพย์สินของครัวเรือนที่ช่วงเวลาถัดไป $t+1$ ซึ่งเป็นไปตามสมการดังนี้

$$w_{t+1} = \left(r_t^r z_t + r^f(1-z_t)\right)(1-c_t)(w_t - C_t)$$
$$+\mu_t(h_t)(\gamma(y_t - 1) + 1) + \left(1 + r^i\right)^{-t-1} x 1(h_t \neq h_{t+1} = dead)$$

เมื่อ C_t คือ ค่าใช้จ่ายที่หลีกเลี่ยงไม่ได้

$\mu_t(.)$ คือ ค่าเฉลี่ยของรายได้สุทธิในปีที่ t ซึ่งขึ้นกับ h_t

γ คือ อัตราปันส่วนความเสี่ยง (Risk-sharing Parameter) ซึ่งค่า γ ที่สูงแสดงว่าครัวเรือนมีการช่วยเหลือกันกับคนนอกครัวเรือนมาก

x คือ จำนวนเงินสินไหมจากการประกันที่ครัวเรือนจะได้รับเมื่อหัวหน้าครัวเรือนเสียชีวิต

r^i คือ อัตราเงินเฟ้อที่ใช้ปรับมูลค่าให้เป็นมูลค่าจริง (Real Value)

r^f คือ อัตราผลตอบแทนในทรัพย์สินที่ไม่มีความเสี่ยง (Risk-free Return)

การเปลี่ยนแปลงของสุขภาพของหัวหน้าครอบครัวในปีถัดไป $t+1$ จะเป็นไปตามความน่าจะเป็นจากตารางมรณะโดยพิจารณาที่แต่ละช่วงอายุและสุขภาพของหัวหน้าครอบครัวในขณะนั้น

สำหรับรางวัลที่แต่ละช่วงเวลา t ที่จะนำไปคิดเป็นผลตอบแทนเพื่อคำนวณ Value นั้น ถูกแสดงด้วยสัญลักษณ์ r_t คือฟังก์ชันอรรถประโยชน์ (Utility) u ซึ่งเป็นฟังก์ชันของลักษณะสุขภาพของหัวหน้าครัวเรือนและอัตราส่วนของมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ใช้ในการบริโภคที่สามารถเปลี่ยนแปลงได้ มีค่าดังต่อไปนี้

$$r_t = u(h_t, c_t) = \begin{cases} \frac{((n-1)c_t)^{1-\kappa}}{1-\kappa} & \text{เมื่อ } h_t = dead \\ \frac{(nc_t)^{1-\kappa}}{1-\kappa} & \text{อื่นๆ} \end{cases}$$

เมื่อ κ คือ พารามิเตอร์ที่บ่งบอกถึง Relative Risk Aversion ซึ่ง κ ที่สูงจะแสดงว่าครัวเรือนมีความกลัวหรือความไม่ชอบความเสี่ยง (Risk Awareness) มาก

n คือ จำนวนสมาชิกในครัวเรือน

จากโครงร่างปัญหาและเป้าหมายเพื่อที่จะได้รับอรรถประโยชน์จากการบริโภคที่เกิดขึ้นตลอดชีวิตมีค่าสูงสุด สามารถแสดงในรูปของสมการเพื่อให้เป็นไปตาม Bellman Equation ได้ดังนี้

$$\max \mathbb{E}\left[\sum_{t=0}^{T-1} \alpha^t u(h_t, c_t) + \alpha^T \beta u(\text{dead}, w_T)\right]$$

เมื่อ β คือ อัตรามรดก (Bequest Factor) ซึ่ง β ที่สูงจะแสดงว่าครัวเรือนมีความห่วงใยต่อลูกหลานมาก

บทที่ 3
วิธีการดำเนินงานวิจัย

งานวิจัยนี้มีวัตถุประสงค์เพื่อนำวิธีการเรียนรู้แบบเสริมกำลังมาประยุกต์ใช้กับการวางแผนทางการเงินของแต่ละครัวเรือนโดยเฉพาะอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยพิจารณาจากลักษณะของแต่ละครัวเรือนที่ถูกกำหนดโดยสุขภาพของหัวหน้าครัวเรือนและมูลค่าปัจจุบันของทรัพย์สินของครัวเรือน และนำผลที่ได้ไปเปรียบเทียบกับผลที่ได้จากวิธี MDP

3.1 แหล่งข้อมูล

ข้อมูลของครัวเรือนสำหรับงานวิจัยนี้นำมาจากงานวิจัยที่ทำการสำรวจพฤติกรรมของครัวเรือนที่อยู่นอกชุมชนเมือง จังหวัดตราด (ณัตติฤดี เจริญรักษ์ และสุภารัตน์ ตันทะนงศักดิ์กุล, 2559) ซึ่งประกอบด้วยข้อมูลจากกลุ่มตัวอย่างจำนวนทั้งสิ้น 348 ครัวเรือน โดยอยู่ในวัยทำงานและวัยเกษียณ ใน 3 อำเภอในจังหวัดตราด ได้แก่ อำเภอเขาสมิง, อำเภอแหลมงอบ และอำเภอบ่อไร่ โดยข้อมูลที่เก็บรวบรวมเกี่ยวข้องกับการจัดสรรรายได้ รายจ่าย การออม การเป็นสมาชิกสถาบันหรือองค์กรชุมชน รวมถึงประโยชน์ที่ได้จากการเป็นสมาชิก

จากข้อมูลทั้งหมดได้ทำการตัดครัวเรือนที่มีจำนวนสมาชิกในครัวเรือนเพียงคนเดียวออก หรือ $n = 1$ เนื่องจากโปรแกรมสำหรับการวางแผนทางการเงินเพื่อหาทางเลือกที่เหมาะสมด้วยวิธี MDP ไม่เหมาะสมที่จะนำมาใช้คำนวณกับครัวเรือนที่มีสมาชิกเพียงคนเดียว

จากครัวเรือนที่เหลืออยู่ทำการตัดครัวเรือนที่หัวหน้าครัวเรือนมีอายุต่ำกว่า 30 ปี และมากกว่า 70 ปีไป เนื่องจากกรอบปัญหานี้มีสมมติฐานว่าบุคคลหนึ่งจะมีชีวิตอยู่ 100 ปี ดังนั้นสำหรับแต่ละบุคคล เวลาที่เหลือในการใช้ชีวิตจะมีค่าเท่ากับ 100 หักด้วยอายุของแต่ละบุคคล นั่นคือหัวหน้าครัวเรือนที่มีอายุต่ำกว่า 30 ปี และมากกว่า 70 ปี จะมีเวลาที่เหลือในการใช้ชีวิตหรือถูกแสดงด้วยสัญลักษณ์ T มีค่ามากกว่า 70 ปี และน้อยกว่า 30 ปีตามลำดับ ซึ่ง T นั้นส่งผลต่อเวลาที่ใช้ในการคำนวณของโปรแกรม กล่าวคือถ้า T มีค่ามากเกินไปจะใช้เวลาในการคำนวณมาก แต่ถ้า T มีค่าน้อยเกินไปจะไม่สามารถสะท้อนข้อมูลในความเป็นจริงได้

นอกจากนี้เพื่อให้ครัวเรือนที่เลือกสามารถสะท้อนความเป็นจริงได้ จึงคัดเลือกครัวเรือนที่อัตราการปันส่วนความเสี่ยง γ และอัตรามรดก β มีค่าไม่เท่ากับ 0 และค่าเฉลี่ยรายได้สุทธิอยู่ใกล้ค่ากลาง (Median) สุดท้ายจึงได้ครัวเรือนที่เหมาะสมที่มีข้อมูลของครัวเรือนดังตารางที่ 3.1

ตารางที่ 3.1 แสดงข้อมูลของครัวเรือนที่เลือก

ลักษณะ	**ค่าของข้อมูล**		
มูลค่าปัจจุบันของทรัพย์สินของครัวเรือน w_0	4,338,556 บาท		
จำนวนสมาชิกในครัวเรือน n	5 คน		
อัตราการปันส่วนความเสี่ยง γ	0.9650		
อัตรามรดก β	0.8744		
เพศของหัวหน้าครัวเรือน	เพศชาย		
อายุของหัวหน้าครัวเรือน	50 ปี		
จำนวนเงินสินไหมจากการประกันที่ครัวเรือนจะได้รับเมื่อหัวหน้าครัวเรือนเสียชีวิต	0 บาท		
ส่วนเบี่ยงเบนมาตรฐานของรายได้สุทธิ σ_y	0.5342		
ค่าเฉลี่ยของรายได้สุทธิที่แต่ละปี t	สุขภาพแข็งแรง	พิการ	เสียชีวิต
$t = 1$	500,000	250,000	250,000
$t = 2$	500,000	250,000	250,000
$t = 3$	500,000	250,000	250,000
$t = 4$	500,000	250,000	250,000
$t = 5$	500,000	250,000	250,000
$t = 6$	500,000	250,000	250,000
$t = 7$	500,000	250,000	250,000
$t = 8$	500,000	250,000	250,000
$t = 9$	500,000	250,000	250,000
$t = 10$	500,000	250,000	250,000
$t = 11$	500,000	250,000	250,000
$t = 12$	450,000	225,000	225,000

ตารางที่ 3.1 แสดงข้อมูลของครัวเรือนที่เลือก (ต่อ)

ลักษณะ	**ค่าของข้อมูล**		
ค่าเฉลี่ยของรายได้สุทธิที่แต่ละปี t	สุขภาพแข็งแรง	พิการ	เสียชีวิต
$t = 13$	405,000	202,500	202,500
$t = 14$	364,500	182,250	182,250
$t = 15$	328,050	164,025	164,025
$t = 16$	295,245	147,623	147,623
$t = 17$	265,721	132,860	132,860
$t = 18$	239,148	119,574	119,574
$t = 19$	215,234	107,617	107,617
$t = 20$	193,710	96,855	96,855
$t = 21$	174,339	87,170	87,170
$t = 22$ ถึง $t = 49$	0	0	0

3.2 เงื่อนไขที่ทำการศึกษา

อัลกอริทึมของการเรียนรู้แบบเสริมกำลังสำหรับงานวิจัยนี้จะศึกษาผลของปัจจัยต่างๆ ต่อไปนี้

3.2.1 ปรับปรุงอัลกอริทึมให้ช่วงแรกเน้นการสำรวจมากขึ้น

3.2.2 ค่าน้ำหนักเริ่มต้น คือ กำหนดค่าน้ำหนักเริ่มต้นเป็น 0 และกำหนดค่าน้ำหนักเริ่มต้นเมื่อตัวแปรตามเป็นรางวัลในขณะนั้น

3.2.3 จำนวนฟีเจอร์ที่ใช้เป็นตัวแปรต้น คือ 100, 200 และ 300

3.2.3 อัตราการเรียนรู้ (η) แบ่งเป็น 2 แบบคือ

1. อัตราการเรียนรู้แบบคงที่ เท่ากับ 0.1
2. อัตราการเรียนรู้แบบลดลงตามเวลา โดยมีค่าเริ่มต้นที่ 0.1 และ 0.01 แล้วมีค่าลดลงตามจำนวนรอบทุกๆ 1,000 รอบ

3.2.4 ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจ (ε) แบ่งเป็น 2 แบบคือ

1. ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่ เท่ากับ 0.7
2. ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลา โดยมีค่าเริ่มต้นที่ 0.9 และ 0.7 แล้วมีค่าลดลงตามจำนวนรอบทุกๆ 1,000 รอบ

3.3 ขั้นตอนในการดำเนินการวิจัย

3.3.1 การสุ่มเลือกสถานะเริ่มต้น

เพื่อที่จะประมาณ Action-value ที่ใช้ในการหาทางเลือกที่เหมาะสมที่สุด จะต้องจำลองสถานะของครัวเรือนที่แต่ละเวลา โดยที่เวลาเริ่มต้น $t = 0$ จะจำลองสถานะจากการสุ่ม

1. สุ่มสุขภาพของหัวหน้าครัวเรือนที่ต้นปี 0 หรือ h_0 จากเวกเตอร์สุขภาพที่เป็นไปได้ทั้งหมดซึ่งประกอบด้วย 3 ค่าคือ แข็งแรง, พิการ, และเสียชีวิตด้วยความน่าจะเป็นที่เท่ากัน

2. สุ่มมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปี 0 หรือ w_0 จากเวกเตอร์ของมูลค่าปัจจุบันของทรัพย์สินของครัวเรือนที่เป็นไปได้ทั้งหมดซึ่งประกอบด้วย 52 ค่าด้วยความน่าจะเป็นที่เท่ากัน โดยค่าที่เป็นไปได้มีรายละเอียดดังต่อไปนี้

- ค่าน้อยสุดที่เป็นไปได้คือ 1
- ค่าในช่วงตั้งแต่ $\frac{w_0}{10}$ ถึง w_0 ถูกแบ่งเป็น 5 ช่วงเท่าๆ กัน ทำให้ได้ค่าที่เป็นไปได้ทั้งหมด 6 ค่า
- ค่าในช่วงระหว่าง w_0 ถึง $9w_0$ ถูกแบ่งเป็น 44 ช่วงเท่าๆ กันทำให้ได้ค่าที่เป็นไปได้ 44 ค่า เนื่องจากไม่รวม w_0
- ค่ามากสุดที่เป็นไปได้คือ $10w_0$

3.3.2 การเลือกการกระทำ

1. สุ่มตัวเลขสุ่มที่มีค่าระหว่าง 0 ถึง 1

2. เปรียบเทียบตัวเลขสุ่มที่ได้จากข้อ 1. กับพารามิเตอร์ ε เพื่อพิจารณาว่าจะเลือกการกระทำที่เป็นแบบการสำรวจหรือการแสวงประโยชน์ โดยถ้าตัวเลขสุ่มมีค่าน้อยกว่า ε จะเลือกการกระทำที่เป็นการสำรวจโดยตามข้อ 3. แต่ถ้าตัวเลขสุ่มมีค่ามากกว่า ε จะเลือกการกระทำที่เป็นการแสวงประโยชน์ไปตามข้อ 4.

3. เลือกการกระทำที่แต่ละเวลา t ที่เป็นแบบการสำรวจ ดังนี้

3.1 สุ่มอัตราส่วนของมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ใช้ในการบริโภคที่สามารถเปลี่ยนแปลงได้ที่ต้นปี t หรือ c_t จากเวกเตอร์ของค่าที่เป็นไปได้ทั้งหมดซึ่งประกอบด้วย 21 ค่าด้วยความน่าจะเป็นที่เท่ากัน

3.2 สุ่มอัตราส่วนของมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่เหลือจากการบริโภคที่ใช้ในการลงทุนในทรัพย์สินที่มีความเสี่ยงที่ต้นปี t หรือ z_t จากเวกเตอร์ของค่าที่เป็นไปได้ทั้งหมดซึ่งประกอบด้วย 21 ค่าด้วยความน่าจะเป็นที่เท่ากัน

4. เลือกการกระทำที่แต่ละเวลา t ที่เป็นการแสวงประโยชน์จะแบ่งเป็น 2 กรณีคือ

กรณีที่ 1: การเลือกการกระทำที่เป็นการแสวงประโยชน์แบบปกติ ดังนี้

4.1 นำสถานะที่เวลา t ที่ได้จากขั้นตอนที่ 3.3.1 หรือ 3.3.3 มาคู่กับการกระทำที่เป็นไปได้ทุกแบบแล้วนำไปแปลงข้อมูล, ทำฟีเจอร์ และประมาณ Action-value ตามขั้นตอนที่ 3.3.4 - 3.3.6 ตามลำดับ

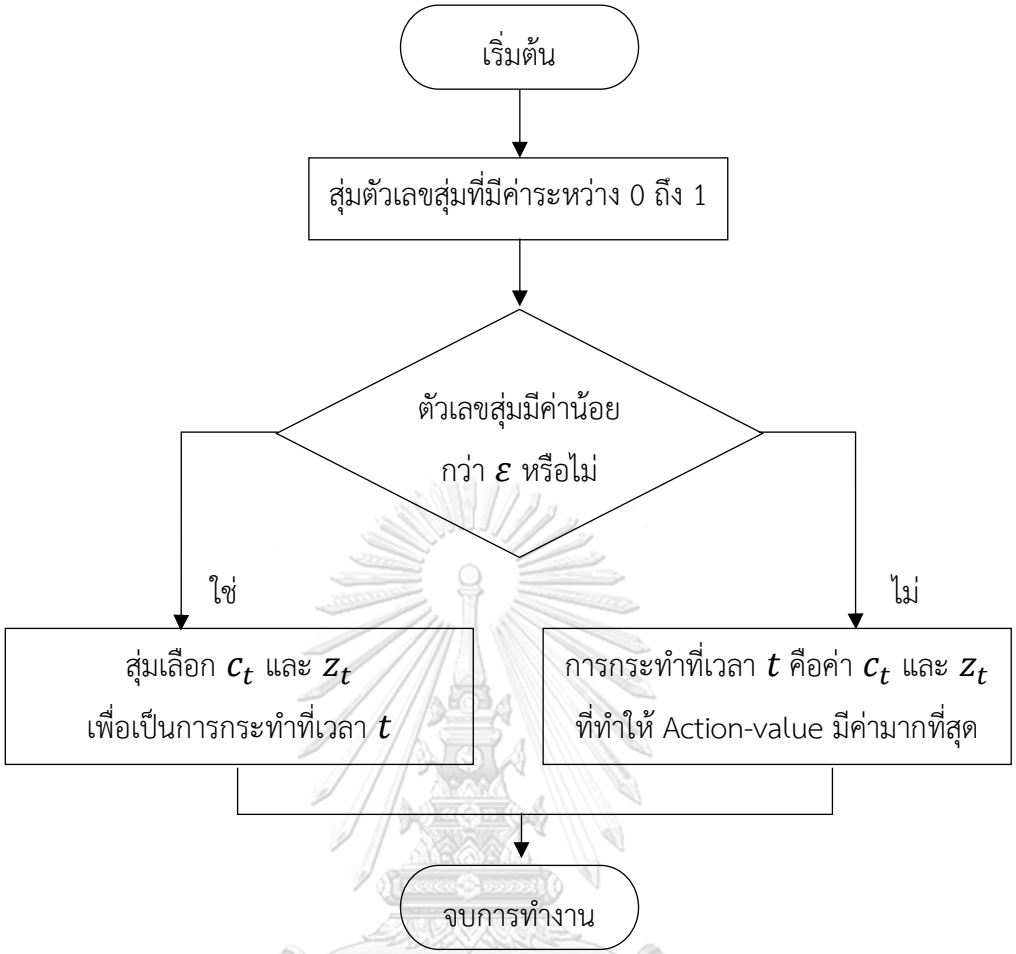
4.2 เปรียบเทียบค่า Action-value ที่ได้จากข้อ 4.1 โดยเลือกการกระทำที่ให้ค่า Action-value มากที่สุด

กรณีที่ 2: การเลือกการกระทำที่เป็นการแสวงประโยชน์ที่เน้นการสำรวจในช่วงแรก ดังนี้

4.1 ถ้าจำนวนรอบน้อยกว่า 35,000 รอบ สุ่มเลือกตัวอย่างการกระทำตามข้อ 3.1 – 3.2 ทั้งหมด 10 ค่าและเลือกการกระทำตามข้อ 4.1 – 4.2 ของกรณีที่ 1 โดยใช้ตัวอย่างการกระทำที่ถูกสุ่มแทนการกระทำที่เป็นไปได้ทั้งหมด

4.2 ถ้าจำนวนรอบมากกว่าหรือเท่ากับ 35,000 รอบ ทำตามกรณีที่ 1

จากขั้นตอนข้างต้นสามารถเขียนแผนผังแสดงขั้นตอนการเลือกการกระทำได้ดังต่อไปนี้



3.3.3 การหารางวัลและสถานะที่เวลาถัดไป

เมื่อทราบสถานะที่เวลา t นั่นคือสุขภาพของหัวหน้าครัวเรือนที่ต้นปี t หรือ h_t และมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปี t หรือ w_t และการกระทำที่เวลา t นั่นคืออัตราส่วนที่ใช้ในการบริโภคที่ต้นปี t หรือ c_t และอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงที่ต้นปี t หรือ z_t จะสามารถหารางวัลที่เวลา $t+1$ ได้จาก

$$r_{t+1} = \begin{cases} \dfrac{\left((n-1)c_t\right)^{1-\kappa}}{1-\kappa} & \text{เมื่อ } h_t = dead \\ \dfrac{(nc_t)^{1-\kappa}}{1-\kappa} & \text{อื่นๆ} \end{cases}$$

สำหรับสุขภาพของหัวหน้าครัวเรือนที่ต้นปีถัดไป $t+1$ หาได้จากการจำลองด้วยความน่าจะเป็นของการเปลี่ยนแปลงสุขภาพที่แต่ละเวลา t ซึ่งคำนวณได้จากตารางที่ 3.2

ตารางที่ 3.2 แสดงการคำนวณความน่าจะเป็นของการเปลี่ยนแปลงของสุขภาพของหัวหน้าครัวเรือน

h_t	h_{t+1}		
	แข็งแรง	พิการ	เสียชีวิต
แข็งแรง	$\left(1-p_t^{H,D}\right)\left(1-p_t^{H,Di}\right)$	$\left(1-p_t^{H,D}\right)p_t^{H,Di}$	$p_t^{H,D}$
พิการ	0	$1-p_t^{Di,D}$	$p_t^{Di,D}$
เสียชีวิต	0	0	1

เมื่อ $p_t^{H,Di}$ คือ ความน่าจะเป็นในการเปลี่ยนจากสุขภาพแข็งแรงเป็นพิการ

$p_t^{H,D}$ คือ ความน่าจะเป็นในการเปลี่ยนจากสุขภาพแข็งแรงเป็นเสียชีวิต

$p_t^{Di,D}$ คือ ความน่าจะเป็นในการเปลี่ยนจากพิการเป็นเสียชีวิต

โดยค่า $p_t^{H,Di}$, $p_t^{H,D}$ และ $p_t^{Di,D}$ นำมาจากตารางมรณะสำหรับเพศชายซึ่งมีค่าที่แต่ละอายุดังตารางที่ 3.3

ตารางที่ 3.3 แสดงความน่าจะเป็นของการเปลี่ยนแปลงของสุขภาพเพศชายที่ละแต่อายุ

อายุ (ปี)	$p_t^{H,Di}$	$p_t^{H,D}$	$p_t^{Di,D}$
50	0.002247	0.006808	0.009054
51	0.002429	0.007361	0.009790
52	0.002633	0.007979	0.010613
53	0.002862	0.008673	0.011535
54	0.003119	0.009450	0.012569
55	0.003406	0.010322	0.013728
56	0.002599	0.011300	0.013899
57	0.002852	0.012399	0.015251
58	0.003137	0.013637	0.016774
59	0.003458	0.015034	0.018492
60	0.003821	0.016612	0.020432
61	0.004229	0.018386	0.022615
62	0.004685	0.020371	0.025056
63	0.005191	0.022567	0.027758
64	0.005743	0.024970	0.030713

ตารางที่ 3.3 แสดงความน่าจะเป็นของการเปลี่ยนแปลงของสุขภาพเพศชายที่ละแต่อายุ (ต่อ)

อายุ (ปี)	$p_t^{H,Di}$	$p_t^{H,D}$	$p_t^{Di,D}$
65	0.006340	0.027564	0.033903
66	0.006976	0.030332	0.037308
67	0.007650	0.033262	0.040912
68	0.008362	0.036355	0.044717
69	0.009115	0.039631	0.048746
70	0.009920	0.043132	0.053052
71	0.010792	0.046922	0.057713
72	0.011748	0.051079	0.062827
73	0.012809	0.055690	0.068499
74	0.013992	0.060834	0.074826
75	0.015310	0.066564	0.081874
76	0.016764	0.072889	0.089653
77	0.018342	0.079749	0.098091
78	0.020011	0.087003	0.107013
79	0.021720	0.094436	0.116156
80	0.023415	0.101805	0.125220
81	0.025049	0.108910	0.133960
82	0.026607	0.115681	0.142287
83	0.028117	0.122248	0.150365
84	0.029665	0.128978	0.158643
85	0.031382	0.136445	0.167827
86	0.033425	0.145328	0.178753
87	0.035942	0.156272	0.192214
88	0.039040	0.169741	0.208781
89	0.042757	0.185902	0.228659
90	0.047051	0.204568	0.251618
91	0.051803	0.225232	0.277036

ตารางที่ 3.3 แสดงความน่าจะเป็นของการเปลี่ยนแปลงของสุขภาพเพศชายที่ละแต่อายุ (ต่อ)

อายุ (ปี)	$p_t^{H,Di}$	$p_t^{H,D}$	$p_t^{Di,D}$
92	0.056859	0.247213	0.304072
93	0.062054	0.269802	0.331856
94	0.067250	0.292392	0.359642
95	0.072351	0.314568	0.386918
96	0.077316	0.336158	0.413474
97	0.082161	0.357223	0.439384
98	0.086940	0.378000	0.464940
99	0.000000	1.000000	1.000000

และมูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปีถัดไป $t+1$ หาได้จากสมการแสดงการเปลี่ยนแปลงดังต่อไปนี้

$$w_{t+1} = \left(r_t^r z_t + r^f(1-z_t)\right)(1-c_t)(w_t - C_t)$$
$$+\mu_t(h_t)(\gamma(y_t-1)+1) + \left(1+r^i\right)^{-t-1} x1(h_t \neq h_{t+1} = dead)$$

3.3.4 การแปลงข้อมูลให้มีค่าเฉลี่ยเป็น 0 และส่วนเบี่ยงเบนมาตรฐานเป็น 1 (Standardization)

ก่อนที่จะนำแต่ละสถานะและการกระทำที่แต่ละเวลาไปทำเป็นฟีเจอร์จะแปลงข้อมูลก่อนเพื่อลดความเบี่ยงเบนจากลักษณะของข้อมูล โดยการนำค่าของแต่ละตัวแปรที่สนใจมาหักด้วยค่าเฉลี่ยและหารด้วยส่วนเบี่ยงเบนมาตรฐาน เพื่อให้ค่าของข้อมูลอยู่ในมาตราส่วนเดียวกัน

3.3.5 การสร้างฟีเจอร์ (Feature Extraction)

นำสถานะและการกระทำจากขั้นตอนที่ 3.3.4 มาทำเป็นฟีเจอร์ 100, 200 และ 300 ลักษณะด้วยเคอร์เนล Radial Basis Function (RBF) 4 แบบโดยมีค่าพารามิเตอร์สำหรับแต่ละจำนวนฟีเจอร์ดังตารางที่ 3.4

ตารางที่ 3.4 แสดงค่าพารามิเตอร์แกมมาสำหรับการทำฟีเจอร์จำนวน 100, 200 และ 300 ลักษณะ

	100 ลักษณะ	200 ลักษณะ	300 ลักษณะ
ค่าแกมมา (Gamma) เท่ากับ 5.0	25	50	75
ค่าแกมมา (Gamma) เท่ากับ 2.0	25	50	75
ค่าแกมมา (Gamma) เท่ากับ 1.0	25	50	75
ค่าแกมมา (Gamma) เท่ากับ 0.5	25	50	75

3.3.6 การกำหนดค่าน้ำหนักเริ่มต้น

การกำหนดค่าน้ำหนักเริ่มต้นแบ่งเป็น 2 กรณีคือ

กรณีที่ 1: กำหนดค่าน้ำหนักเริ่มต้นเป็นค่า 0

โดยสร้างเวกเตอร์ของ 0 ที่มีขนาดเท่ากับจำนวนฟีเจอร์ตามขั้นตอนที่ 3.3.5

กรณีที่ 2: กำหนดค่าน้ำหนักเริ่มต้นเมื่อตัวแปรตามเป็นรางวัลในขณะนั้น ดังนี้

1. สุ่มสุขภาพของหัวหน้าครัวเรือนที่ต้นปี t หรือ h_t , มูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนที่ต้นปี t หรือ w_t และอัตราส่วนที่ใช้ในการบริโภคที่ต้นปี t หรือ c_t จำนวน 1,000,000 ตัวอย่าง

2. แปลงข้อมูลในข้อ 1. ด้วยขั้นตอนตาม 3.3.4 – 3.3.5 เพื่อใช้เป็นตัวแปรต้นในการหาค่าน้ำหนักเริ่มต้น

3. หารางวัลในขณะนั้นด้วยขั้นตอนตาม 3.3.3 เพื่อใช้เป็นตัวแปรตามในการหาค่าน้ำหนัก

4. นำตัวแปรต้นและตัวแปรตามในข้อ 2. และ 3. ไปสร้างตัวแบบถดถอยเชิงเส้น (Linear regression) เพื่อนำค่าสัมประสิทธิ์ (Coefficient) มาใช้เป็นค่าน้ำหนักเริ่มต้น

3.3.7 การประมาณ Action-value

หา Action-value เพื่อใช้เปรียบเทียบว่าการกระทำใดเป็นการกระทำที่เหมาะสมที่สุด โดยที่แต่ละเวลา t สามารถประมาณ Action-value ได้จากฟังก์ชันเส้นตรงตามสมการ

$$\hat{q}(s,a,t,\mathbf{b}) = \mathbf{b}^{\mathrm{T}}\mathbf{F}(s,a,t) = \sum_{i=1}^{d} b_i F_i(s,a,t)$$

โดยการหาผลคูณภายใน (Dot product) ระหว่างฟีเจอร์ที่ได้จากข้อ 3.3.5 กับค่าน้ำหนักที่ได้จากข้อ 3.3.6 เมื่อ $t = 0$ หรือ 3.3.8 เมื่อ $t \neq 0$ โดยนำแต่ละองค์ประกอบของเวกเตอร์ฟีเจอร์มาคูณกับเวกเตอร์ค่าน้ำหนัก จากนั้นจึงหาผลรวมของทั้งหมด

อย่างไรก็ตาม เนื่องจากการอัพเดทค่าน้ำหนักต้องใช้ Action-value ที่ 2 ช่วงเวลาดังต่อไปนี้

1. Action-value ที่ได้จากฟีเจอร์ของสถานะและการกระทำที่เวลา t กับค่าน้ำหนัก
2. Action-value ที่ได้จากฟีเจอร์ของสถานะและการกระทำที่เวลาถัดไป $t + 1$ กับค่าน้ำหนัก

3.3.8 การอัพเดทค่าน้ำหนัก

1. กำหนด Target โดยพิจารณาจากเวลา t โดย
   - ถ้าหากสถานะถัดไปเป็นสถานะสุดท้าย หรือ $t = 50$ Action-value จากฟีเจอร์ของสถานะและการกระทำที่เวลาถัดไปจะมีค่าเป็น 0 ดังนั้น Target คือรางวัลในขณะนั้น r_{51}
   - ถ้าหากสถานะถัดไปไม่ใช่สถานะสุดท้าย หรือ $t < 50$ จะต้องนำ Action-value จากฟีเจอร์ของสถานะและการกระทำที่เวลาถัดไปที่ได้จากขั้นตอนที่ 3.3.7 มาพิจารณาด้วยอัตราคิดลด ดังนั้น Target คือ $r_{t+1} + \alpha\hat{q}(s_{t+1}, a_{t+1}, t + 1, \mathbf{b})$
2. หาความผิดพลาดโดยการหาผลต่างระหว่าง Target กับค่าประมาณในขณะนั้นหรือ Action-value จากฟีเจอร์ของสถานะและการกระทำที่เวลา t และปรับขนาดโดยการนำไปคูณกับพารามิเตอร์อัตราการเรียนรู้ η
3. หาทิศทางการอัพเดทของค่าน้ำหนักโดยการหาอนุพันธ์ของ Action-value เทียบกับเวกเตอร์ค่าน้ำหนัก ซึ่งก็คือฟีเจอร์ของสถานะและการกระทำที่เวลาขณะนั้น ดังสมการต่อไปนี้

$$\nabla\hat{q}(s_t, a_t, t, \mathbf{b}) = \mathbf{F}(s_t, a_t, t, \mathbf{b})$$

4. หาขนาดและทิศทางที่จะอัพเดทโดยการหาผลคูณภายในระหว่างความผิดพลาดที่ถูกปรับขนาดจากข้อ 2. และทิศทางการอัพเดทจากข้อ 3. ดังสมการ

$$\mathbf{b} \leftarrow \mathbf{b} + \eta[r_{t+1} + \alpha\hat{q}(s_{t+1}, a_{t+1}, t + 1, \mathbf{b}) - \hat{q}(s_t, a_t, t, \mathbf{b})]\nabla\hat{q}(s_t, a_t, t, \mathbf{b})$$

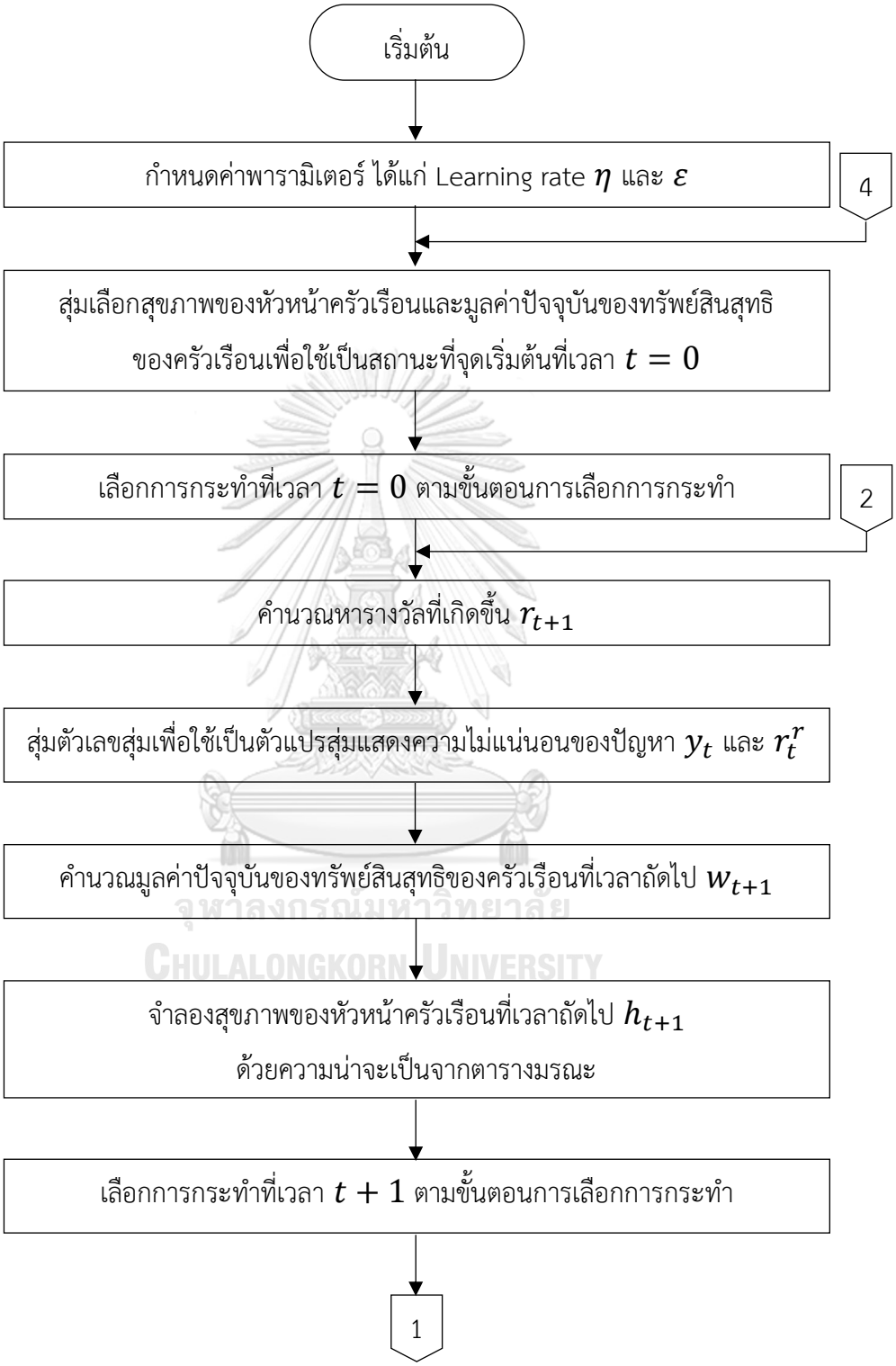
3.3.9 การเปรียบเทียบผลที่ได้จากการเรียนรู้แบบเสริมกำลังและ MDP จากความผิดพลาด

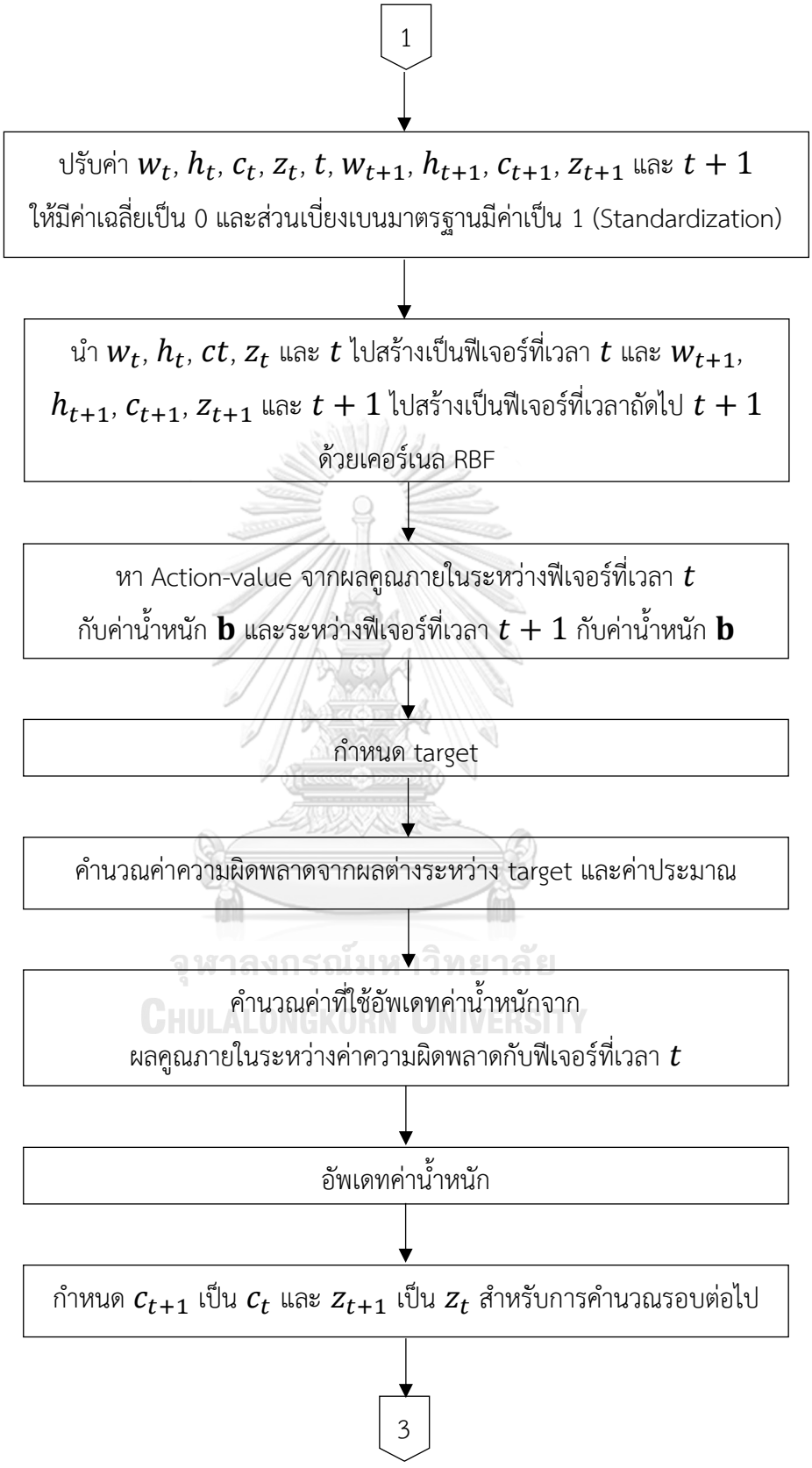
1. ภายหลังจาก Action-value ลู่เข้าสู่ค่าหนึ่ง หาค่า Action-value สำหรับสถานะที่มูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนมีค่าเท่ากับ w_0 และสุขภาพของหัวหน้าครัวเรือนแข็งแรงที่เวลาเริ่มต้นโดยทำตามข้อ 4. ของขั้นตอนที่ 3.3.2

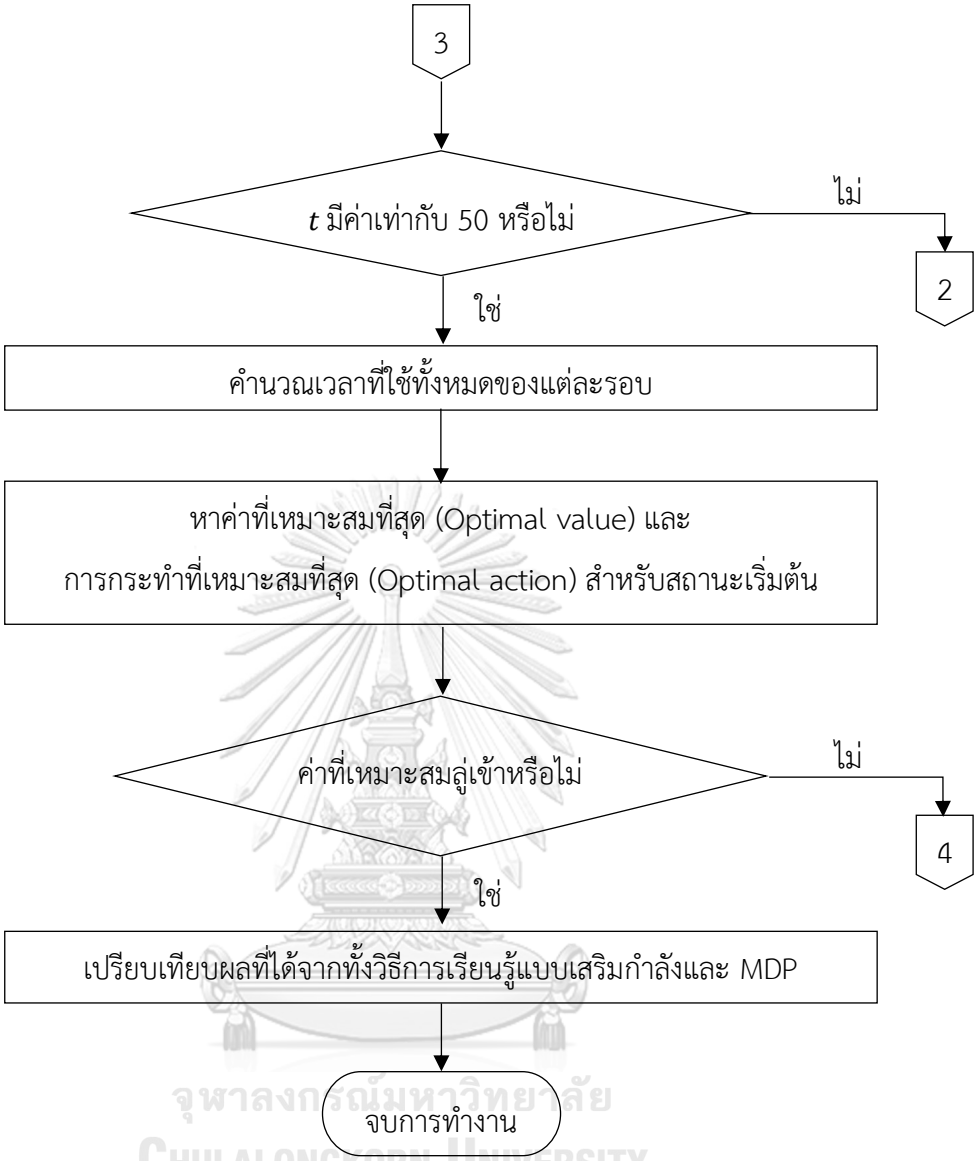
2. หาค่า Optimal value สำหรับสถานะที่มูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนมีค่าเท่ากับ w_0 และสุขภาพของหัวหน้าครัวเรือนแข็งแรงที่เวลาเริ่มต้นด้วยโปรแกรมแก้ปัญหาด้วย MDP

3. นำค่าที่ได้จากข้อ 1. และ 2. มาลบกันเพื่อหาความผิดพลาด

3.4 แผนผังแสดงขั้นตอนการทำงาน







บทที่ 4
ผลการวิจัย

งานวิจัยนี้มีจุดประสงค์เพื่อศึกษาการประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงินเทียบกับ MDP โดยจะศึกษาผลของการเปลี่ยนแปลงปัจจัยต่างๆ และการพิจารณาความสัมพันธ์ของ Optimal action ดังนั้นผลการวิจัยจึงจะถูกแบ่งออกเป็น 7 ส่วนดังต่อไปนี้

1) การเรียนรู้แบบเสริมกำลังแบบปกติกับการวางแผนทางการเงิน
2) ผลของการปรับปรุงอัลกอริทึมให้ช่วงแรกเน้นการสำรวจมากขึ้น
3) ผลของการกำหนดค่าน้ำหนักเริ่มต้น
4) ผลของจำนวนฟีเจอร์ที่ใช้เป็นตัวแปรต้น
5) ผลของพารามิเตอร์อัตราการเรียนรู้ (η)
6) ผลของพารามิเตอร์ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจ (ε)
7) การพิจารณาความสัมพันธ์ของ Optimal action

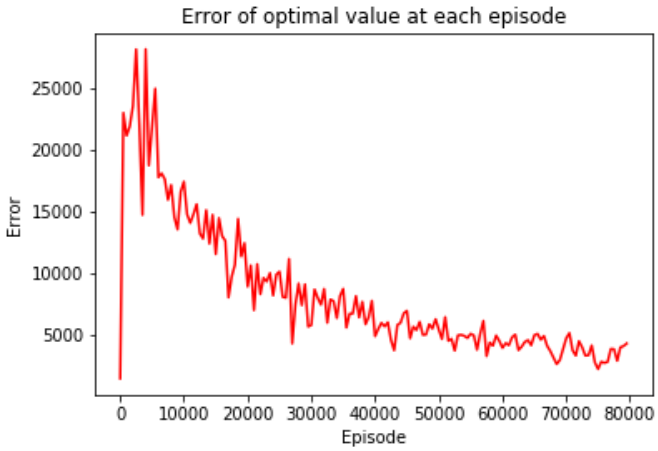
4.1 การเรียนรู้แบบเสริมกำลังแบบปกติกับการวางแผนทางการเงิน

เพื่อเปรียบเทียบการเรียนรู้แบบเสริมกำลังโดยใช้อัลกอริทึม SARSA แบบปกติเมื่อกำหนดค่าน้ำหนักเริ่มต้นจากตัวแบบความถดถอยเชิงเส้นเมื่อตัวแปรตามเป็นรางวัลในขณะนั้น, จำนวนฟีเจอร์ทั้งหมดเป็น 200 ลักษณะ, อัตราการเรียนรู้แบบลดลงตามเวลาโดยมีค่าเริ่มต้นที่ 0.1 และความน่าเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาโดยมีค่าเริ่มต้นเป็น 0.9 กับ MDP ความผิดพลาด (Error) ของ Optimal value ซึ่งคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบ (Episode) และของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้ง แล้วเลือกครั้งที่ให้ความผิดพลาดต่ำที่สุด

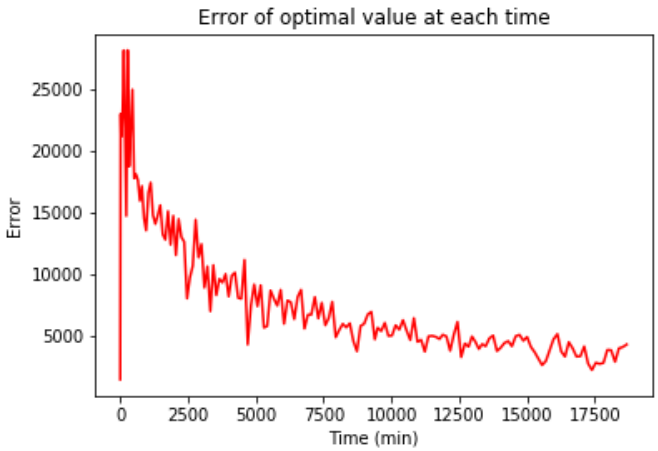
การเปรียบเทียบพิจารณาจากสถานะเริ่มต้นหรือที่มูลค่าปัจจุบันของทรัพย์สินสุทธิของครัวเรือนมีค่าเท่ากับ w_0 และสุขภาพของหัวหน้าครัวเรือนแข็งแรงที่เวลาเริ่มต้น โดยการเปรียบเทียบของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.1 และการเปรียบเทียบที่แต่ละเวลาที่ใช้ในการคำนวณถูกแสดงดังในรูปภาพที่ 4.2

จากรูปภาพที่ 4.1 และ 4.2 จะเห็นว่าในช่วงแรกจนถึงประมาณรอบที่ 30,000 หรือที่เวลาประมาณ 5,000 นาที เส้นกราฟแสดงความผิดพลาดของ Optimal Value ระหว่างวิธีการเรียนรู้แบบ

เสริมกำลังและ MDP มีการแกว่งตัวในช่วงกว้างกว่าความผิดพลาดภายหลังจากช่วงนี้ เนื่องจากในช่วงนี้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจยังมีค่ามากคือประมาณ 0.9 – 0.225 ทำให้เกิดการเลือกการกระทำแบบสำรวจหรือการสุ่มการกระทำ ส่งผลให้ค่าที่ได้มีความผันผวน และอีกสาเหตุหนึ่งคืออัตราการเรียนรู้ยังมีค่าสูงคือประมาณ 0.1 – 0.025 ทำให้การอัพเดทค่าน้ำหนักในแต่ละรอบส่งผลให้เกิดการเปลี่ยนแปลงมาก



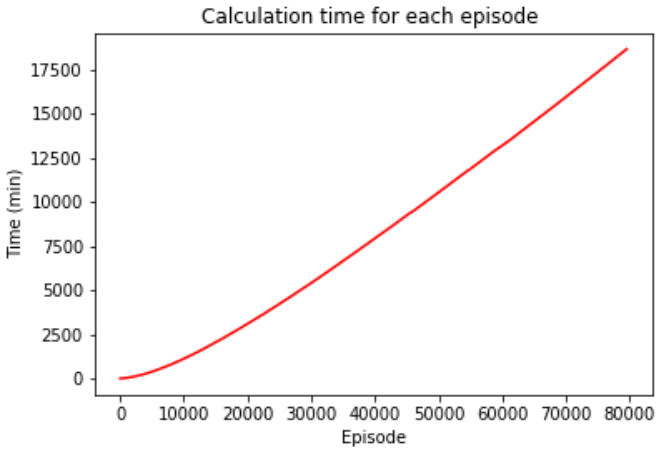
รูปภาพที่ 4.1 ความผิดพลาดของ Optimal Value ระหว่างวิธีการเรียนรู้แบบเสริมกำลังแบบปกติและ MDP ที่แต่ละรอบ



รูปภาพที่ 4.2 ความผิดพลาดของ Optimal Value ระหว่างวิธีการเรียนรู้แบบเสริมกำลังแบบปกติและ MDP ที่แต่ละเวลาในหน่วยนาที

สำหรับช่วงหลังภายหลังรอบที่ 30,000 หรือเวลา 5,000 นาที เส้นกราฟแสดงความผิดพลาดของ Optimal Value ระหว่างวิธีการเรียนรู้แบบเสริมกำลังและ MDP มีการแกว่งตัวในช่วงแคบกว่าและเริ่มมีการลู่เข้าสู่ค่าประมาณ 4,000 ที่ประมาณ 70,000 รอบหรือที่เวลาประมาณ 16,000 นาที

อันเนื่องมาจากความน่าจะเป็นในการเลือกการกระทำแบบสำรวจและอัตราการเรียนรู้ลดลงเข้าใกล้ 0 แสดงว่าคำตอบเริ่มลู่เข้าสู่ค่าหนึ่งๆ



รูปภาพที่ 4.3 เวลาที่ใช้ในการคำนวณของการเรียนรู้แบบเสริมกำลังแบบปกติที่แต่ละรอบ

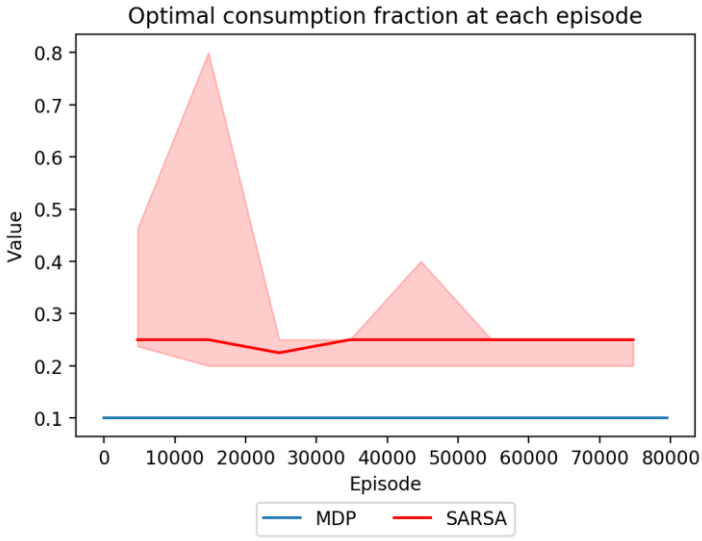
จากรูปภาพที่ 4.3 พบว่า เส้นกราฟในช่วงรอบที่ 0 – 10,000 มีความชันน้อยและค่อยๆ มีความชันเพิ่มขึ้น จนกระทั่งช่วงประมาณรอบที่ 50,000 – 80,000 เส้นกราฟมีความชันคงที่แต่มีค่ามากกว่าในช่วงรอบที่ 0 – 10,000 แสดงว่าในช่วงแรกที่ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าสูง ทำให้การกระทำส่วนมากเป็นแบบสำรวจใช้เวลาในการคำนวณน้อย แต่ในช่วงหลังที่ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าน้อย ทำให้การกระทำส่วนมากเป็นแบบการแสวงประโยชน์ใช้เวลาในการคำนวณมาก

เมื่อกำหนดให้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจเป็น 1 และ 0 เพื่อให้เลือกการกระทำแบบสำรวจและแสวงประโยชน์ตามลำดับ โดยทำการคำนวณซ้ำทั้งหมด 100 รอบและหาค่าเฉลี่ยของเวลาที่ใช้ในการคำนวณแต่ละรอบ จะได้ผลดังตารางที่ 4.1 พบว่า การเลือกการกระทำแบบแสวงประโยชน์ใช้เวลาในการคำนวณมากกว่าการเลือกการกระทำแบบสำรวจถึง 189 เท่า และเมื่อพิจารณาจากรูปภาพที่ 4.1 พบว่า รอบที่ 35,000 ค่า Optimal value ยังไม่ลู่เข้าสู่ค่าคงที่ค่าหนึ่ง แสดงว่าคำตอบที่ได้ยังไม่ถูกต้องและไม่เหมาะสมที่จะนำไปใช้สำหรับการเลือกการกระทำแบบแสวงประโยชน์ จึงมีแนวคิดในการปรับปรุงอัลกอริทึมให้ช่วงแรกก่อน 35,000 รอบเน้นการเลือกการกระทำแบบสำรวจดังผลการวิจัยในกรณีต่อไป

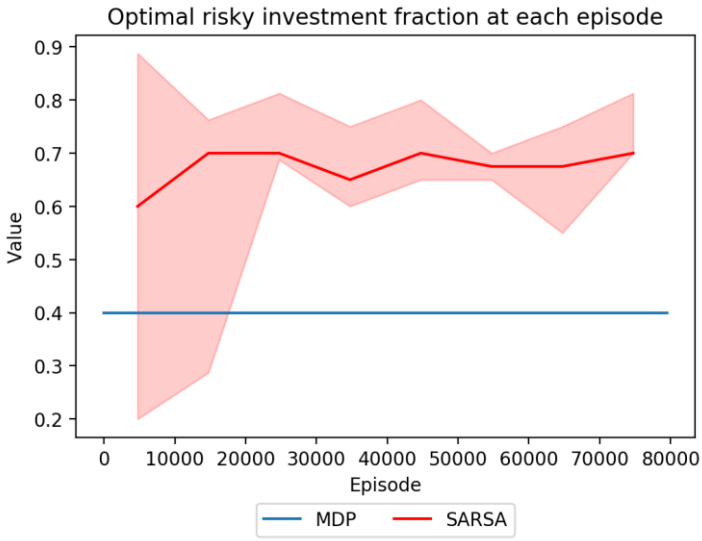
ตารางที่ 4.1 เวลาที่ใช้ในการคำนวณแต่ละรอบของขั้นตอนการเลือกการกระทำในหน่วยวินาที สำหรับอัลกอริทึมแบบปกติ

การเลือกการกระทำ	เวลาที่ใช้ในการคำนวณเฉลี่ย (วินาที/รอบ)
แบบสำรวจ	0.069
แบบแสวงประโยชน์	13.086

เมื่อพิจารณาการวางแผนทางการเงินโดยพิจารณาอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงตามวิธีการเรียนรู้แบบเสริมกำลังเทียบกับ MDP ที่สถานะเริ่มต้นซึ่งถูกแสดงด้วยค่ามัธยฐาน (Median) พร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.4 และ 4.5 ตามลำดับ จากรูปภาพที่ 4.4 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธีการเรียนรู้แบบเสริมกำลังมีการลู่เข้าสู่ค่าประมาณ 0.25 ในขณะที่ผลจากวิธี MDP มีค่า 0.10 และจากรูปภาพที่ 4.5 จะเห็นว่าอัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธีการเรียนรู้แบบเสริมกำลังมีการลู่เข้าสู่ค่าประมาณ 0.65 ในขณะที่ผลจากวิธี MDP มีค่า 0.40 แสดงให้เห็นว่าคำตอบที่ได้จากวิธีการเรียนรู้แบบเสริมกำลังยังมีความผิดพลาดสูงอยู่



รูปภาพที่ 4.4 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังแบบปกติ โดยมีขอบบนจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และขอบล่างจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 75 ของทุก 10,000 รอบ

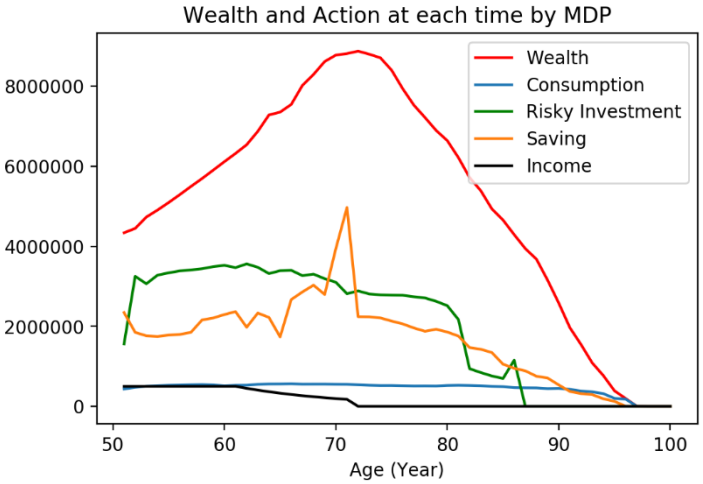


รูปภาพที่ 4.5 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังแบบปกติ โดยมีขอบบนจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และขอบล่างจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 75 ของทุก 10,000 รอบ

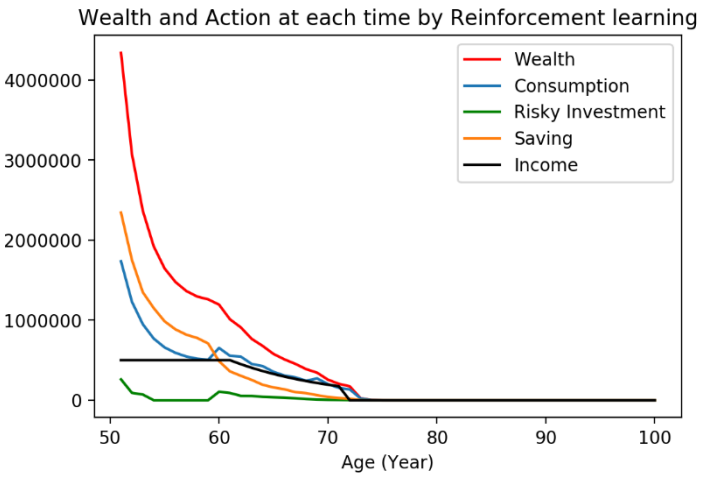
อย่างไรก็ตามเมื่อพิจารณาขอบบนและขอบล่างของทั้งอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงตามวิธีการเรียนรู้แบบเสริมกำลัง พบว่า ทั้งขอบบนและขอบล่างลู่เข้าเมื่อจำนวนรอบเพิ่มมากขึ้น ซึ่งสอดคล้องกับรูปภาพที่ 4.1 ที่แสดงถึงการลู่เข้าของคำตอบสู่ค่าหนึ่งเมื่อจำนวนรอบเพิ่มมากขึ้น

ผลจากการนำคำตอบที่ได้จากทั้งโปรแกรม MDP และโปรแกรมการเรียนรู้แบบเสริมกำลังแบบปกติไปจำลองประยุกต์ใช้จริงกับการวางแผนทางการเงินของครัวเรือนตลอดช่วงอายุของหัวหน้าครัวเรือน โดยทำซ้ำทั้งหมด 5,000 ครั้งแล้วหาค่ามัธยฐานของข้อมูลที่แต่ละเวลาถูกแสดงดังรูปภาพที่ 4.6 พบว่า สำหรับการตัดสินใจตามโปรแกรม MDP ตามรูปภาพที่ 4.6 (ก) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นจนถึงเวลาหนึ่งแล้วลดลงเข้าใกล้ 0 ตามเป้าหมายที่ต้องการให้อรรถประโยชน์ที่เกิดขึ้นตลอดชีวิตมีค่าสูงสุด โดยในช่วงอายุ 50 – 60 ปีเน้นการลงทุนในสินทรัพย์ที่มีความเสี่ยงมากกว่าการออมและการบริโภค เมื่อภายหลังอายุ 60 ปีรายได้ของครัวเรือนลดลง การลงทุนในสินทรัพย์ที่มีความเสี่ยงลดลงและเพิ่มการออม และเมื่ออายุประมาณ 70 ปีมูลค่าสินทรัพย์ของครัวเรือนลดลง ปริมาณเงินที่ใช้ในกิจกรรมทางการเงินต่างๆ ก็ลดลงเช่นกัน โดยในช่วงหลังอายุ 90 ปีจะเน้นการบริโภคมากกว่าการออมและการลงทุน ในขณะที่จากรูปภาพที่ 4.6 (ข) ผลจากการตัดสินใจด้วยโปรแกรมการเรียนรู้แบบเสริมกำลังได้ลักษณะที่ต่างออกไป โดยมูลค่าทรัพย์สินของครัวเรือนไม่มีการสะสมเพิ่มขึ้นตามเวลา เน้นการออมมากกว่าการบริโภคและการลงทุนในสินทรัพย์ที่

มีความเสี่ยง ที่ช่วงอายุ 60 ปีอัตราการลดลงของมูลค่าทรัพย์สินของครัวเรือนน้อยลง เมื่อรายได้ของครัวเรือนลดลง การลงทุนในสินทรัพย์ที่มีความเสี่ยงและการบริโภคเพิ่มขึ้น แต่การออมลดลง โดยในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน



(ก)



(ข)

รูปภาพที่ 4.6 มูลค่าสินทรัพย์ของครัวเรือน, รายได้ของครัวเรือน และการวางแผนทางการเงิน จาก

(ก) โปรแกรม MDP ด้วยวิธี Backward Recursive

(ข) โปรแกรมการเรียนรู้แบบเสริมกำลังด้วยอัลกอริทึม SARSA แบบปกติ

หากเปรียบเทียบคำตอบที่ได้จากทั้ง 2 โปรแกรมในแต่ละช่วง พบว่า ช่วงหลังมีลักษณะใกล้เคียงกันคือเน้นการบริโภคมากกว่าการออมและการลงทุนในสินทรัพย์ที่มีความเสี่ยง ในขณะที่ช่วงแรกคำตอบแตกต่างกันมาก เมื่อพิจารณาโปรแกรมการเรียนรู้แบบเสริมกำลัง พบว่า ที่ลำดับเวลาสุดท้ายจะอัพเดทค่าน้ำหนักโดยคำนวณจากรางวัลในขณะนั้นเทียบกับค่าประมาณที่สถานะนั้น

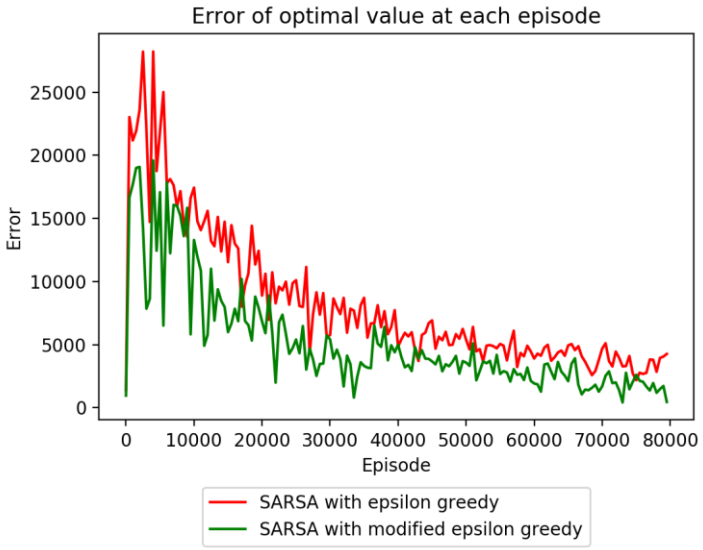
ในขณะที่ลำดับเวลาอื่นๆ จะอัพเดทค่าน้ำหนักโดยคำนวณจากผลรวมของรางวัลในขณะนั้นกับค่าประมาณของสถานะถัดไปเทียบกับค่าประมาณที่สถานะนั้น ทำให้การอัพเดทสำหรับลำดับเวลาช่วงหลังได้รับผลจากความเบี่ยงเบนของค่าประมาณน้อยกว่าที่ลำดับเวลาช่วงแรก

4.2 ผลของการปรับปรุงอัลกอริทึมให้ช่วงแรกเน้นการสำรวจมากขึ้น

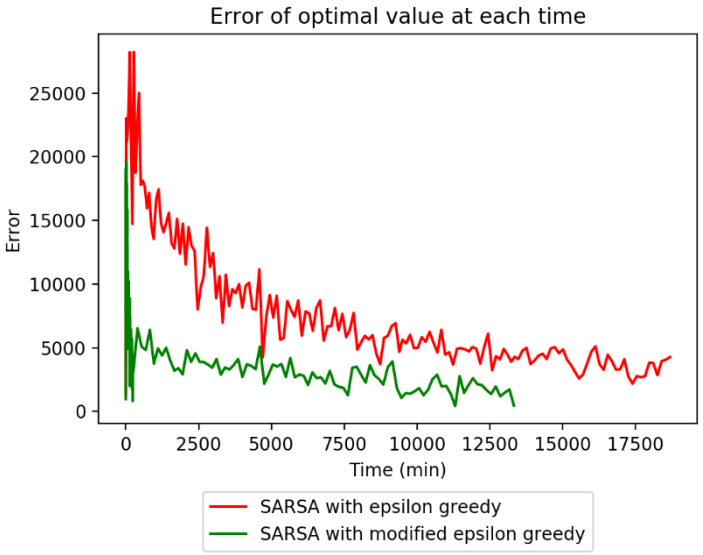
เนื่องจากการกระทำที่เป็นการสำรวจจะถูกสุ่มเลือกจากการกระทำที่เป็นไปได้ทั้งหมด ในขณะที่การกระทำที่เป็นการแสวงประโยชน์จะถูกเลือกจากการกระทำที่ทำให้ได้ประโยชน์สูงสุด ซึ่งมาจากการคำนวณ Action value สำหรับทุกการกระทำ แล้วจึงเลือกการกระทำที่ทำให้เกิด Action value สูงสุด การเลือกการกระทำที่เป็นการสำรวจจึงใช้เวลาคำนวณน้อยกว่าการเลือกการกระทำที่เป็นการแสวงประโยชน์ ดังนั้นในหัวข้อนี้จึงต้องการศึกษาว่าถ้าหากปรับปรุงอัลกอริทึมให้ช่วงแรกเน้นการสำรวจมากขึ้นจะมีผลต่อการวางแผนทางการเงินอย่างไรบ้าง

เมื่อเปรียบเทียบการเรียนรู้แบบเสริมกำลังโดยใช้อัลกอริทึม SARSA แบบปกติตามกรณี 4.1 และอัลกอริทึมแบบเน้นการสำรวจในช่วงแรก (SARSA with modified epsilon greedy) กับ MDP โดยคำนวณความผิดพลาดของ Optimal value จากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบ (Episode) และของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วเลือกครั้งที่ให้ความผิดพลาดต่ำสุด และการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น โดยการเปรียบเทียบของแต่ละรอบและแต่ละเวลาถูกแสดงดังในรูปภาพที่ 4.7 และ 4.8 ตามลำดับ พบว่า สำหรับวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจ ในช่วงแรกจนถึงรอบที่ 40,000 หรือที่เวลา 1,200 นาที เส้นกราฟแสดงความผิดพลาดของ Optimal Value มีการแกว่งตัวในช่วงกว้าง โดยในช่วงนี้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจยังมีค่ามากคือช่วงประมาณ 0.90 – 0.18 และอัตราการเรียนรู้ยังมีค่าสูงคือช่วงประมาณ 0.1 – 0.02 สำหรับช่วงภายหลังรอบที่ 40,000 หรือที่เวลา 1,200 นาที ความผิดพลาดมีการแกว่งตัวแคบลง ซึ่งการแกว่งตัวเช่นนี้เป็นลักษณะเช่นเดียวกับอัลกอริทึมแบบปกติ แต่ในช่วงรอบที่ 10,000 – 35,000 ความผิดพลาดจากอัลกอริทึมแบบเน้นการสำรวจในช่วงแรกมีการแกว่งตัวกว้างกว่าความผิดพลาดจากอัลกอริทึมแบบปกติ ซึ่งเป็นผลจากการปรับปรุงอัลกอริทึมให้เน้นการสำรวจในช่วง 35,000 รอบแรก

หากเปรียบเทียบความผิดพลาดระหว่างอัลกอริทึมแบบปกติและแบบเน้นการสำรวจในช่วงแรก พบว่า ความผิดพลาดจากอัลกอริทึมแบบเน้นการสำรวจในช่วงแรกสามารถลู่เข้าสู่ค่าหนึ่งได้เช่นเดียวกับอัลกอริทึมแบบปกติ โดยลู่เข้าสู่ค่าประมาณ 1,700 ที่รอบที่ 70,000 หรือเวลา 10,000 นาทีซึ่งต่ำกว่าผลจากอัลกอริทึมแบบปกติ



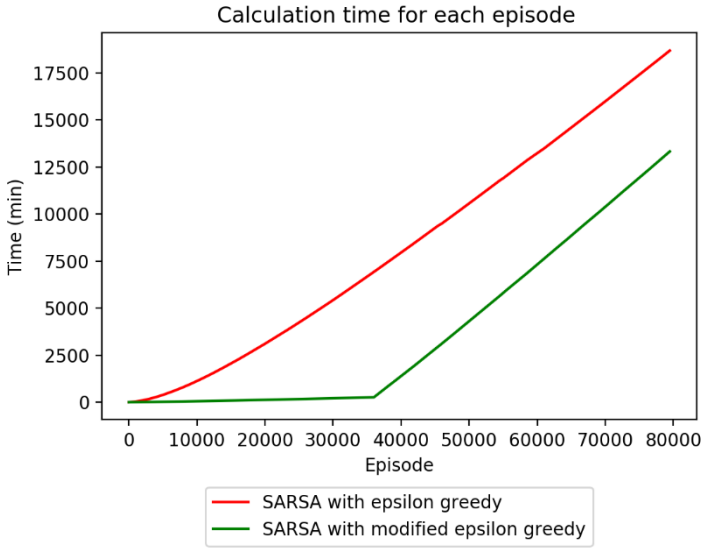
รูปภาพที่ 4.7 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบ



รูปภาพที่ 4.8 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่แต่ละเวลาในหน่วยนาที

จากรูปภาพที่ 4.9 พบว่า เส้นกราฟในช่วงรอบที่ 0 – 35,000 ของการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกมีความชันน้อยมากและน้อยกว่าเส้นกราฟของการเรียนรู้แบบเสริมกำลังแบบปกติอย่างชัดเจน ภายหลังรอบที่ 35,000 เส้นกราฟมีความชันเพิ่มขึ้นอย่างคงที่และมีความชันใกล้เคียงกับเส้นกราฟของการเรียนรู้แบบเสริมกำลังแบบปกติ แสดงว่าการปรับปรุงให้เน้นการสำรวจในช่วง 35,000 รอบแรกช่วยลดเวลาที่ใช้ในการคำนวณได้จริง โดยทำให้เวลาที่ใช้ในการ

คำนวณสำหรับ 80,000 รอบลดลงไป 5,360 นาทีจากทั้งหมด 18,687 นาที คิดเป็น 28.68% ของเวลาที่ใช้ของการเรียนรู้แบบเสริมกำลังแบบปกติ



รูปภาพที่ 4.9 เวลาที่ใช้ในการคำนวณของการเรียนรู้แบบเสริมกำลังแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบ

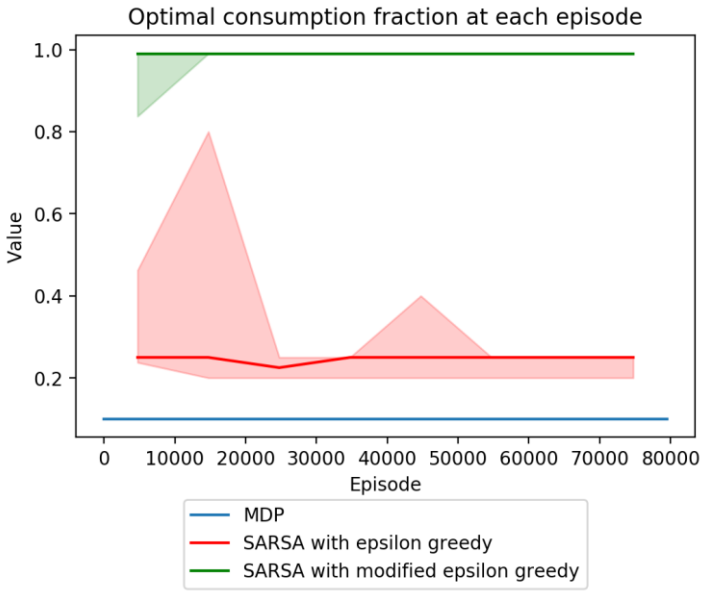
สำหรับการเรียนรู้ของอัลกอริทึมที่ปรับปรุงให้ช่วงแรกเน้นการสำรวจ โดยเลือกการกระทำแบบแสวงประโยชน์จากการสุ่มการกระทำจากการกระทำที่เป็นไปได้ทั้งหมด แล้วเลือกการกระทำที่ให้ประโยชน์สูงสุด เมื่อกำหนดให้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจเป็น 1 และ 0 เพื่อให้เลือกการกระทำแบบสำรวจและแสวงประโยชน์ตามลำดับ โดยสำหรับการกระทำแบบแสวงประโยชน์แบ่งเป็นการกระทำแบบปกติและแบบปรับปรุง และทำการคำนวณซ้ำทั้งหมด 100 รอบ และหาค่าเฉลี่ยของเวลาที่ใช้ในการคำนวณแต่ละรอบเป็นไปดังตารางที่ 4.2 พบว่า เวลาที่ใช้ในการคำนวณสำหรับการเลือกการกระทำแบบแสวงประโยชน์แบบปกติมากกว่าการเลือกการกระทำแบบสำรวจ 162 เท่า ในขณะที่การเลือกการกระทำแบบแสวงประโยชน์แบบปรับปรุงใช้เวลามากกว่าการเลือกการกระทำแบบสำรวจ 5 เท่า จะเห็นว่าการปรับปรุงช่วยทำให้เวลาที่ใช้ในการคำนวณสำหรับการกระทำแบบแสวงประโยชน์ลดลงไปถึง 97% ในแต่ละรอบ

ตารางที่ 4.2 เวลาที่ใช้ในการคำนวณแต่ละรอบของขั้นตอนการเลือกการกระทำในหน่วยวินาที สำหรับอัลกอริทึมแบบปรับปรุงให้ช่วงแรกเน้นการสำรวจ

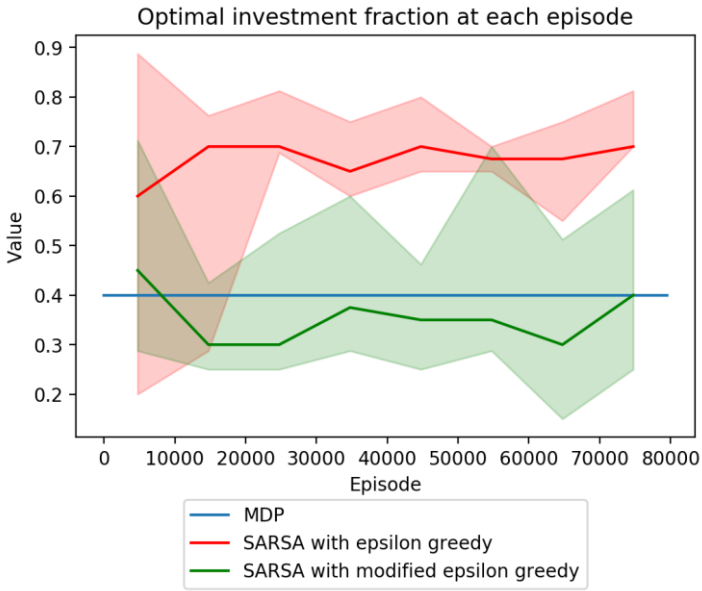
การเลือกการกระทำ	เวลาที่ใช้ในการคำนวณเฉลี่ย (วินาที/รอบ)
แบบสำรวจ	0.077
แบบแสวงประโยชน์แบบปกติ	12.508
แบบแสวงประโยชน์แบบปรับปรุง	0.354

เมื่อพิจารณาผลของวิธีการเรียนรู้แบบเสริมกำลังแบบปกติและแบบเน้นการสำรวจในช่วงแรกกับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้นซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.10 และ 4.11 ตามลำดับ จากรูปภาพที่ 4.10 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกลู่เข้าสู่ค่า 0.99 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.1 น้อยกว่าผลจากวิธีการเรียนรู้แบบเสริมกำลังแบบปกติที่ลู่เข้าสู่ค่า 0.25 และจากรูปภาพที่ 4.11 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกมีการลู่เข้าค่า 0.4 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.4 มากกว่าผลจากวิธีการเรียนรู้แบบเสริมกำลังแบบปกติที่ลู่เข้าสู่ค่า 0.7 แสดงให้เห็นว่าถึงแม้ความผิดพลาดของ Optimal value ของวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกมีค่าน้อยกว่าวิธีการเรียนรู้แบบเสริมกำลังแบบปกติ คำตอบก็ยังมีความผิดพลาดสูง

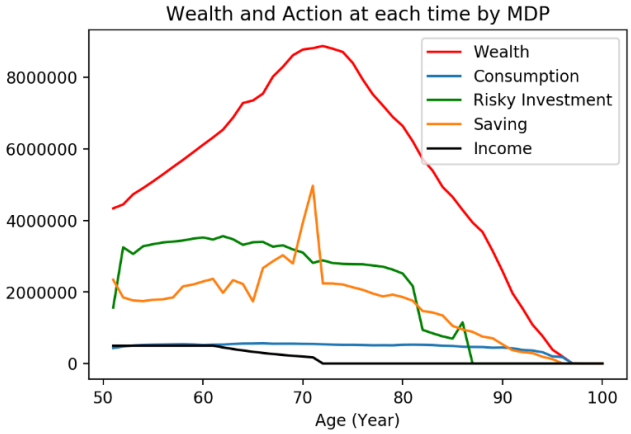
สำหรับอัลกอริทึมแบบเน้นการสำรวจในช่วงแรก ขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคมีการลดลงจากช่วงแรกอย่างชัดเจน โดยในช่วงหลังข้อมูลไม่มีความแปรปรวน ทำให้ขอบบนและขอบล่างตรงกับค่ามัธยฐาน แต่สำหรับความกว้างระหว่างขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงค่อนข้างคงที่ ไม่แสดงให้เห็นถึงการลู่เข้าอย่างชัดเจน



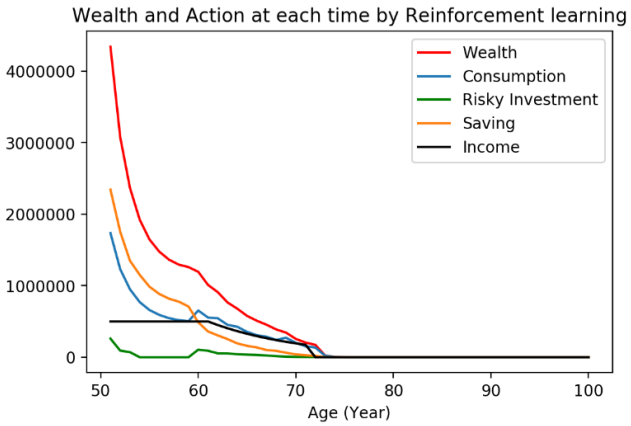
รูปภาพที่ 4.10 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบ



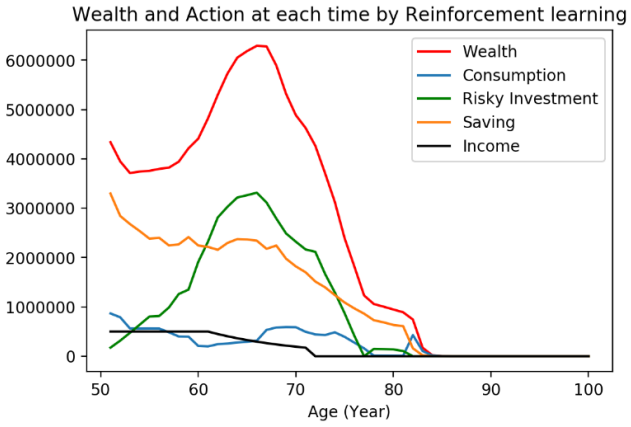
รูปภาพที่ 4.11 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบ



(ก)



(ข)



(ค)

รูปภาพที่ 4.12 มูลค่าสินทรัพย์ของครัวเรือน, รายได้ของครัวเรือน และการวางแผนทางการเงิน จาก

(ก) โปรแกรม MDP ด้วยวิธี Backward Recursive

(ข) โปรแกรมการเรียนรู้แบบเสริมกำลังด้วยอัลกอริทึม SARSA แบบปกติ

(ค) โปรแกรมการเรียนรู้แบบเสริมกำลังด้วยอัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก

ผลจากการใช้อัลกอริทึมแบบปกติและแบบเน้นการสำรวจในช่วงแรกต่อการนำคำตอบที่ได้จากทั้งโปรแกรม MDP และโปรแกรมการเรียนรู้แบบเสริมกำลังไปจำลองประยุกต์ใช้จริงกับการวางแผนทางการเงินของครัวเรือนตลอดช่วงอายุของหัวหน้าครัวเรือน โดยทำซ้ำทั้งหมด 5,000 ครั้ง แล้วหาค่ามัธยฐานของข้อมูลที่แต่ละเวลาถูกแสดงดังรูปภาพที่ 4.12 พบว่า สำหรับการตัดสินใจตามโปรแกรมการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกตามรูปภาพที่ 4.12 (ค) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นตามเวลาจนมีค่ามากสุดที่ 6,300,000 บาท ในช่วงแรกเน้นการออมมากกว่าการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง แต่ในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน

หากเปรียบเทียบคำตอบที่ได้จากการใช้อัลกอริทึมทั้ง 2 แบบกับผลจาก MDP พบว่า การใช้อัลกอริทึมแบบเน้นการสำรวจในช่วงแรก มูลค่าทรัพย์สินของครัวเรือนมีความใกล้เคียงกับผลจาก MDP มากกว่า เนื่องจากสามารถมูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นและถูกรักษาไว้ได้จนถึงอายุประมาณ 80 ปี นอกจากนี้หากพิจารณาการตัดสินใจทางการเงินของทั้ง 2 กรณีในแต่ละช่วง ช่วงหลังมีลักษณะใกล้เคียงกันคือเน้นการบริโภคมากกว่าการออมและการลงทุนในสินทรัพย์ที่มีความเสี่ยง ในขณะที่ช่วงแรกคำตอบแตกต่างกันมาก ซึ่งเป็นผลจากการอัพเดทค่าน้ำหนักที่ช่วงหลังจะได้รับผลกระทบจากการประมาณน้อยกว่า

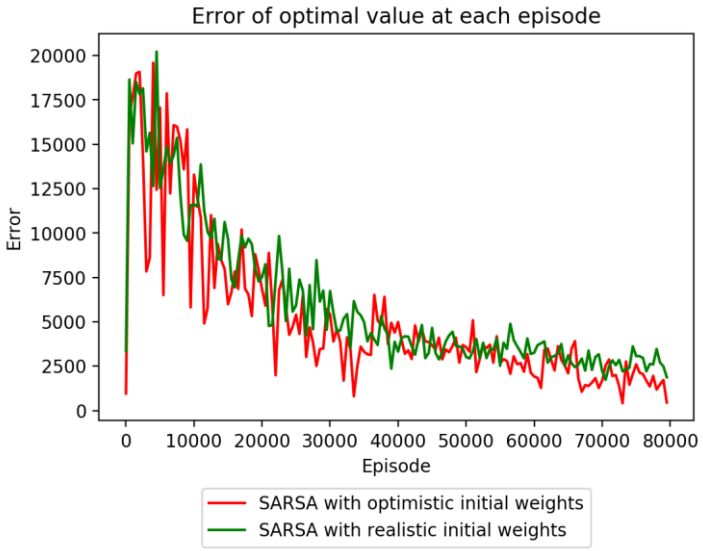
4.3 ผลของการกำหนดค่าน้ำหนักเริ่มต้น

การกำหนดค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมด (Realistic Initial Weights) อาจส่งผลให้เกิดความเบี่ยงเบนของ Action value ทำให้มีผลกระทบต่อการเลือกการกระทำโดยเฉพาะการกระทำแบบสำรวจ การกำหนดค่าน้ำหนักเริ่มต้นโดยอาศัยข้อมูลที่มีอยู่หรือที่เรียกว่าแบบ Optimistic ซึ่งใช้ค่าน้ำหนักเริ่มต้นจากจากตัวแบบการถดถอยแบบเชิงเส้นที่มีตัวแปรตามเป็นรางวัลในขณะนั้น (Optimistic Initial Weights) สามารถช่วยให้เกิดการสำรวจในแต่ละคู่ของสถานะ-การกระทำมากขึ้นสำหรับช่วงการเรียนรู้ค่า Action-value ดังนั้นในหัวข้อนี้จะศึกษาผลของการกำหนดค่าน้ำหนักเริ่มต้น โดยเปรียบเทียบค่าน้ำหนักเริ่มต้นเมื่อกำหนดให้เป็น 0 ทั้งหมดหรือแบบ Realistic กับแบบ Optimistic

เมื่อปรับค่าน้ำหนักเริ่มต้นและใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก เพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดย

ทำซ้ำทั้งหมด 3 ครั้งแล้วเลือกครั้งที่ให้ความผิดพลาดต่ำสุด โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.13 พบว่า สำหรับทั้งกรณีที่ใช้ค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดและใช้ค่าน้ำหนักจากตัวแบบการถดถอยเชิงเส้นในช่วงแรกจนถึงรอบที่ 40,000 เส้นกราฟแสดงความผิดพลาดของ Optimal Value มีการแกว่งตัวในช่วงกว้าง เนื่องจากความน่าจะเป็นในการเลือกการกระทำแบบสำรวจยังมีค่ามากคือประมาณ 0.90 – 0.18 และอัตราการเรียนรู้ยังมีค่าสูงคือประมาณ 0.1 – 0.02 สำหรับช่วงภายหลังรอบที่ 40,000 ความผิดพลาดมีการแกว่งตัวในช่วงแคบกว่า โดยในช่วง 40,000 รอบแรกความผิดพลาดจากอัลกอริทึมที่ใช้ค่าน้ำหนักเริ่มต้นเป็น 0 มีการแกว่งตัวแคบกว่าความผิดพลาดจากอัลกอริทึมที่ใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นที่มีตัวแปรตามเป็นรางวัลในขณะนั้น

หากเปรียบเทียบระหว่างการใช้ค่าน้ำหนักเริ่มต้นทั้ง 2 แบบ พบว่า ที่ 70,000 รอบ ความผิดพลาดจากกรณีที่ใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นลู่เข้าสู่ค่า 1,700 ซึ่งต่ำกว่าความผิดพลาดจากกรณีที่ใช้ค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดซึ่งลู่เข้าสู่ค่า 2,500

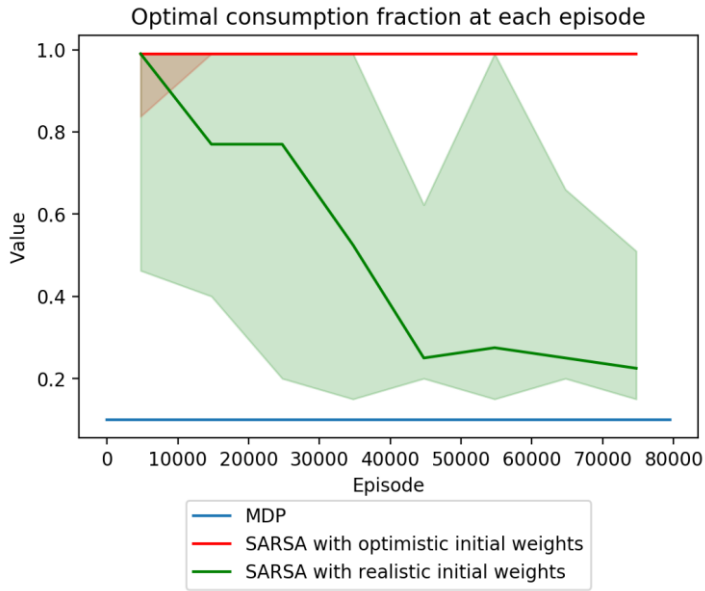


รูปภาพที่ 4.13 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อกำหนดค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นที่มีตัวแปรตามเป็นรางวัลในขณะนั้นและค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมด

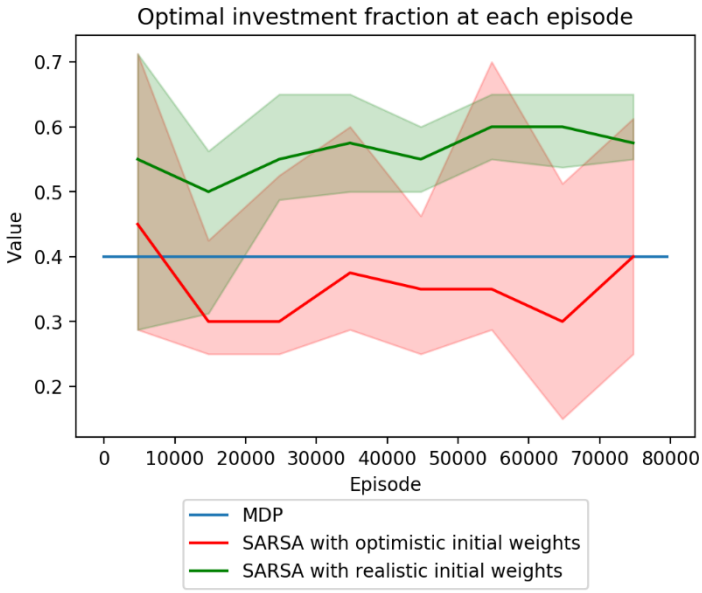
เมื่อพิจารณาผลของการกำหนดค่าน้ำหนักเริ่มต้นกับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้น ซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่

25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.14 และ 4.15 ตามลำดับ จากรูปภาพที่ 4.14 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นลู่เข้าสู่ค่า 0.99 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.1 น้อยกว่าผลจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดที่ลู่เข้าสู่ค่า 0.25 และจากรูปภาพที่ 4.15 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นลู่เข้าสู่ค่า 0.4 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.4 มากกว่าผลจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดที่ลู่เข้าสู่ค่า 0.6 แสดงให้เห็นว่าคำตอบที่ได้ยังมีความผิดพลาดสูงอยู่

อย่างไรก็ตามความกว้างระหว่างขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากทั้งสองกรณีเริ่มแคบลงเมื่อจำนวนรอบมากขึ้น ในขณะที่ขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นค่อนข้างคงที่ ไม่แสดงให้เห็นถึงการลู่เข้าอย่างชัดเจน แต่ผลจากอัลกอริทึมที่มีค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดมีการลู่เข้าอย่างชัดเจน



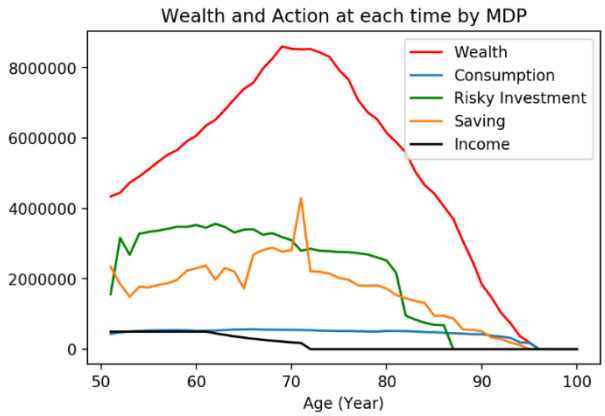
รูปภาพที่ 4.14 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อกำหนดค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นที่มีตัวแปรตามเป็นรางวัลในขณะนั้นและค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมด



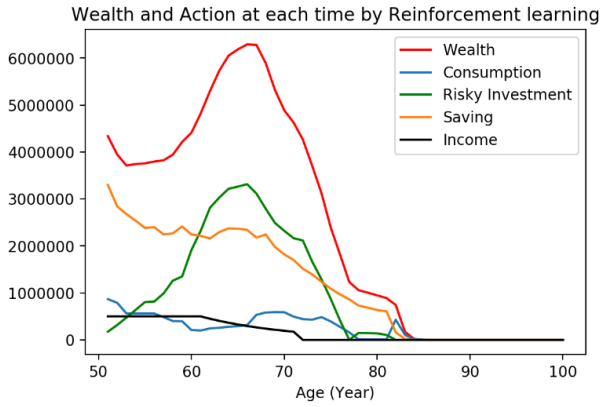
รูปภาพที่ 4.15 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อกำหนดค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นที่มีตัวแปรตามเป็นรางวัลในขณะนั้นและค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมด

ผลของการใช้ค่าน้ำหนักเริ่มต้นต่อการนำคำตอบที่ได้จากทั้งโปรแกรม MDP และโปรแกรมการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกไปจำลองใช้กับการวางแผนทางการเงินของครัวเรือนตลอดช่วงอายุของหัวหน้าครัวเรือน โดยทำซ้ำทั้งหมด 5,000 ครั้งแล้วหาค่ามัธยฐานของข้อมูลที่แต่ละเวลาถูกแสดงดังรูปภาพที่ 4.16 พบว่า เมื่อใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้นดังรูปภาพท่ 4.16 (ข) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นตามเวลาจนมีค่ามากสุดที่ 6,300,000 บาท ในช่วงแรกเน้นการออมมากกว่าการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง แต่ในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน สำหรับการกำหนดค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมดตามรูปภาพที่ 4.16 (ค) มูลค่าทรัพย์สินของครัวเรือนไม่มีการสะสมเพิ่มขึ้นตามเวลา ในช่วงแรกเน้นการบริโภคมากกว่าการออมและการลงทุนในสินทรัพย์ที่มีความเสี่ยง แต่ในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน

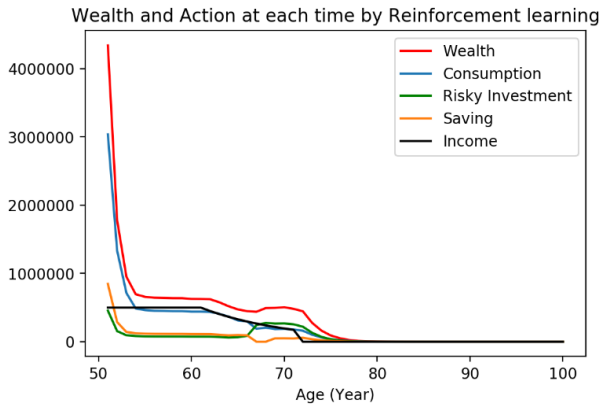
หากเปรียบเทียบคำตอบที่ได้จากการใช้ค่าน้ำหนักเริ่มต้นทั้ง 2 แบบกับผลจาก MDP พบว่าการใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น มูลค่าทรัพย์สินของครัวเรือนมีความใกล้เคียงกับผลจาก MDP มากกว่าเนื่องจากมีการสะสมมูลค่าทรัพย์สินของครัวเรือนและถูกรักษาไว้ได้ถึงอายุ 82 ปี นอกจากนี้หากพิจารณาการตัดสินใจทางการเงินของทั้ง 2 แบบในแต่ละช่วง ช่วงหลังมีลักษณะ



(ก)



(ข)



(ค)

รูปภาพที่ 4.16 มูลค่าสินทรัพย์ของครัวเรือน, รายได้ของครัวเรือน และการวางแผนทางการเงิน จาก

(ก) โปรแกรม MDP ด้วยวิธี Backward Recursive

(ข) การเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น

(ค) การเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อค่าน้ำหนักเริ่มต้นเป็น 0 ทั้งหมด

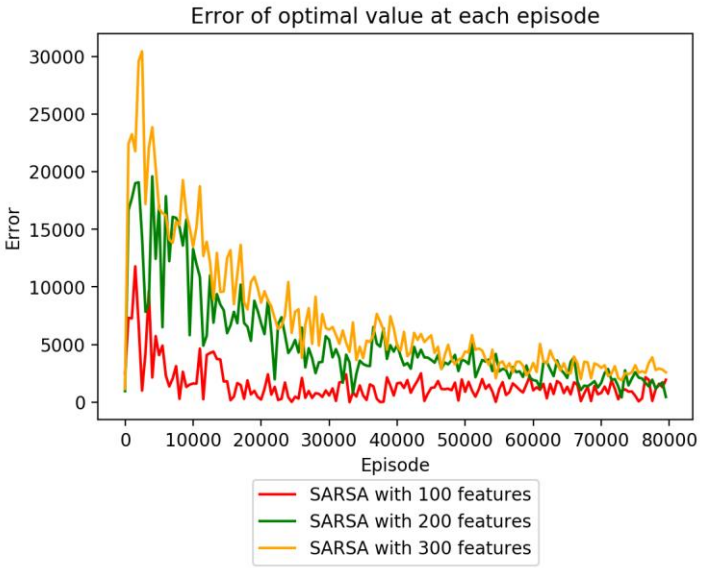
ใกล้เคียงกันคือเน้นการบริโภคมากกว่าการออมและการลงทุนในสินทรัพย์ที่มีความเสี่ยง ในขณะที่ช่วงแรกคำตอบแตกต่างกันมาก ซึ่งเป็นผลจากการอัพเดทค่าน้ำหนักที่ช่วงหลังจะได้รับผลกระทบการประมาณน้อยกว่า

4.4 ผลของจำนวนฟีเจอร์ที่ใช้เป็นตัวแปรต้น

จากกรณี 4.2 วิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกใช้จำนวนฟีเจอร์ทั้งหมด 200 ลักษณะเป็นตัวแปรต้นสำหรับการประมาณค่า Action-value ของกรอบปัญหา ในหัวข้อนี้จึงต้องการเปลี่ยนจำนวนฟีเจอร์ที่ใช้เป็นตัวแปรต้นดังตารางที่ 3.3 เพื่อศึกษาผลที่เกิดขึ้นกับความผิดพลาดของ Optimal value, ผลต่อการวางแผนทางการเงิน และการจำลองใช้จริง

เมื่อปรับจำนวนฟีเจอร์ทั้งหมดเป็น 100 และ 300 ลักษณะและใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก เพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วเลือกครั้งที่ให้ความผิดพลาดต่ำสุด โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.17 พบว่าสำหรับทั้งกรณีจำนวนฟีเจอร์ 100, 200 และ 300 ลักษณะในช่วงแรกจนถึงรอบที่ 25,000 เส้นกราฟแสดงความผิดพลาดของ Optimal Value ระหว่างวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจและ MDP มีการแกว่งตัวในช่วงกว้าง โดยในช่วงนี้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่ามากคือช่วงประมาณ 0.90 – 0.26 และอัตราการเรียนรู้ยังมีค่าสูงคือช่วงประมาณ 0.1 – 0.029 สำหรับช่วงภายหลังรอบที่ 25,000 ความผิดพลาดของ Optimal value มีการแกว่งตัวในช่วงแคบกว่า

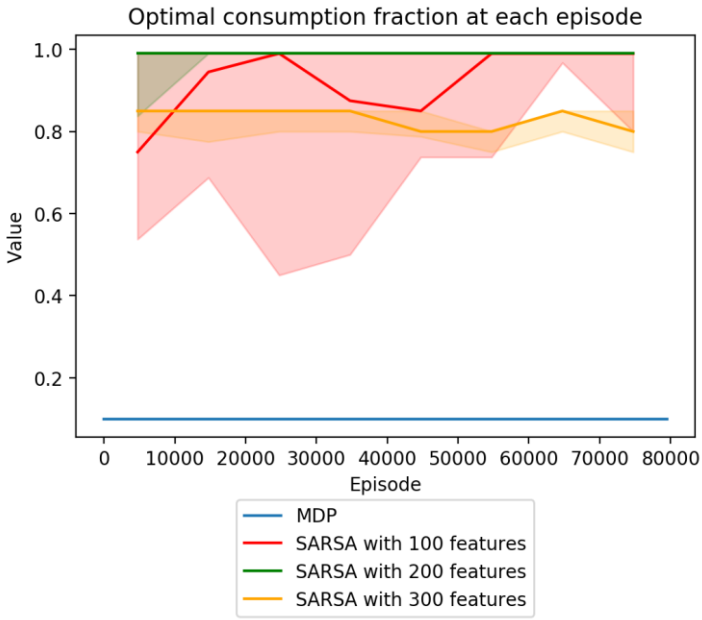
หากเปรียบเทียบระหว่างการใช้จำนวนฟีเจอร์ที่แตกต่างกัน พบว่า ในช่วง 20,000 รอบแรกการใช้จำนวนฟีเจอร์มากขึ้น เส้นกราฟความผิดพลาดจะมีการแกว่งตัวมากขึ้น แต่สำหรับการใช้จำนวนฟีเจอร์ 100 ลักษณะ ภายหลังจากเมื่อความผิดพลาดลดลงและเริ่มแสดงการลู่เข้าที่รอบที่ 20,000 ความผิดพลาดยังสามารถเพิ่มขึ้นและลู่เข้าสู่อีกค่าหนึ่งได้ แสดงให้เห็นว่าจำนวนฟีเจอร์อาจยังไม่เพียงพอในการแสดงลักษณะของครัวเรือน ในขณะที่ความผิดพลาดของการใช้จำนวนฟีเจอร์ 200 และ 300 ลักษณะสามารถลู่เข้าสู่ค่าหนึ่งได้ โดยเมื่อเพิ่มจำนวนฟีเจอร์ความผิดพลาดกลับเพิ่มขึ้นด้วย โดยที่ประมาณรอบที่ 65,000 ความผิดพลาดเริ่มมีการลู่เข้าสู่ค่าประมาณ 1,900, 1,700 และ 2,500 สำหรับการใช้จำนวนฟีเจอร์ 100, 200 และ 300 ลักษณะตามลำดับ



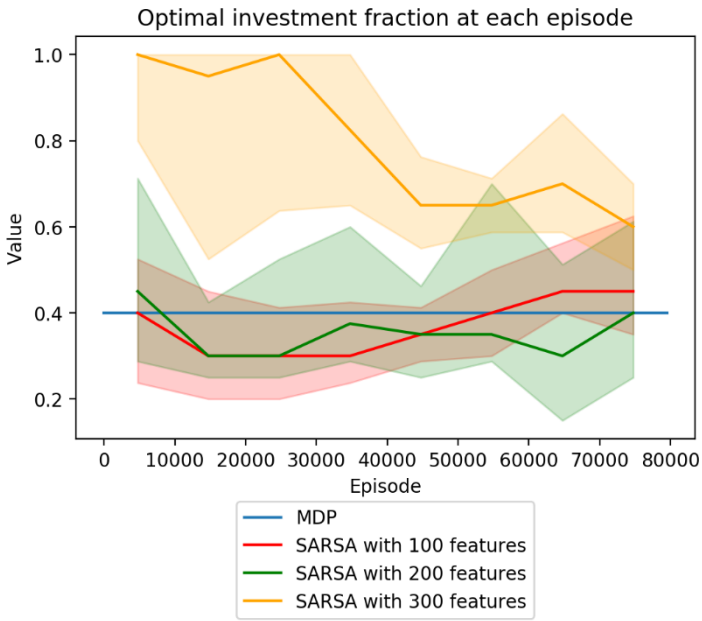
รูปภาพที่ 4.17 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้จำนวนฟีเจอร์ 100, 200 และ 300 ลักษณะ

เมื่อพิจารณาผลของจำนวนฟีเจอร์ที่ใช้กับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้นซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.18 และ 4.19 ตามลำดับ จากรูปภาพที่ 4.18 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้จำนวนฟีเจอร์ 300 ลักษณะลู่เข้าค่า 0.8 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.1 มากกว่าผลจากการใช้จำนวนฟีเจอร์ 100 และ 200 ลักษณะที่ลู่เข้าสู่ค่า 0.99 ทั้งสองกรณี และจากรูปภาพที่ 4.19 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้จำนวนฟีเจอร์ 300 ลักษณะลู่เข้าสู่ค่า 0.6 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.4 น้อยกว่าผลจากการใช้จำนวนฟีเจอร์ 100 และ 200 ลักษณะซึ่งลู่เข้าสู่ค่า 0.45 และ 0.4 ตามลำดับ แสดงให้เห็นว่าถึงแม้ความผิดพลาดของ Optimal value จะแสดงถึงการลู่เข้าสู่ค่าหนึ่ง คำตอบที่ได้ยังมีความผิดพลาดสูงอยู่

อย่างไรก็ตามขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคทั้ง 3 กรณีแคบลงจากช่วงแรกอย่างชัดเจน เช่นเดียวกับขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้ฟีเจอร์จำนวน 100 และ 300 ลักษณะที่แคบลงจากช่วงแรกอย่างชัดเจน



รูปภาพที่ 4.18 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้จำนวนฟีเจอร์ 100, 200 และ 300 ลักษณะ

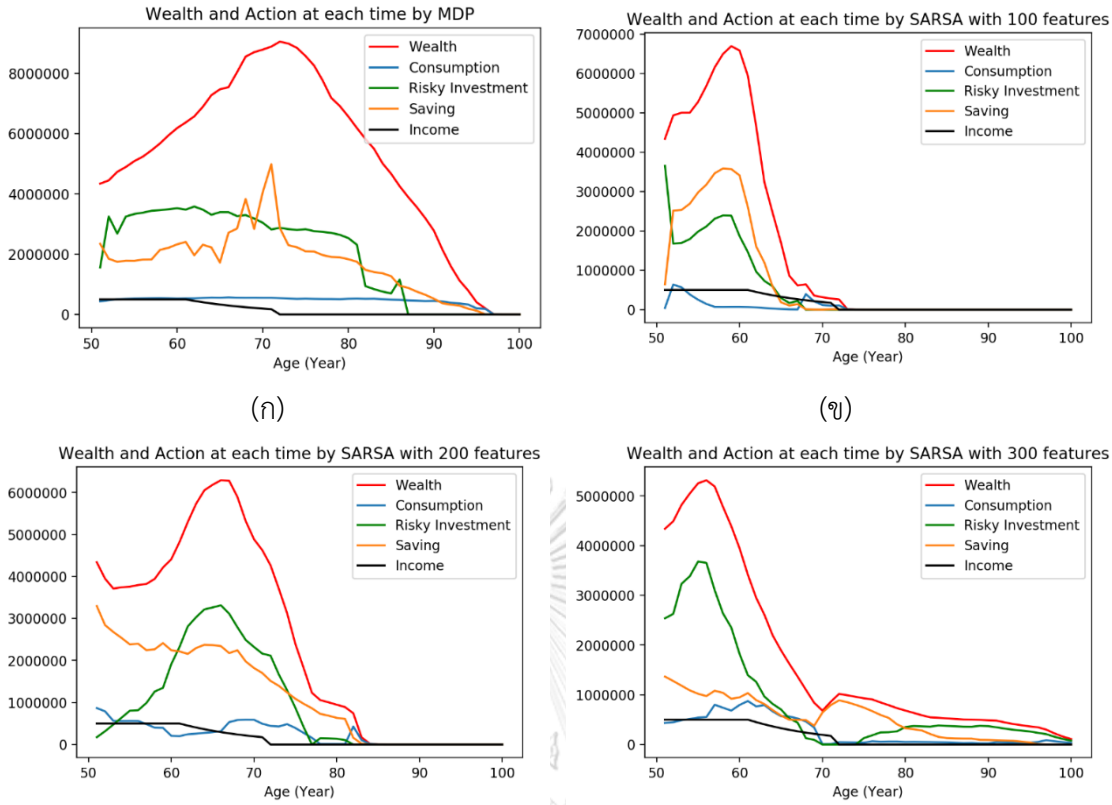


รูปภาพที่ 4.19 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้จำนวนฟีเจอร์ 100, 200 และ 300 ลักษณะ

ในขณะที่ความกว้างระหว่างขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้ฟีเจอร์ 200 ลักษณะค่อนข้างคงที่ ไม่แสดงให้เห็นถึงการลู่เข้าอย่างชัดเจน

ผลของการใช้จำนวนฟีเจอร์ที่แตกต่างกันต่อการนำคำตอบที่ได้จากทั้งโปรแกรม MDP และโปรแกรมการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกไปจำลองประยุกต์ใช้จริงกับการวางแผนทางการเงินของครัวเรือนตลอดช่วงอายุของหัวหน้าครัวเรือน โดยทำซ้ำทั้งหมด 5,000 ครั้ง แล้วหาค่ามัธยฐานของข้อมูลที่แต่ละเวลาถูกแสดงดังรูปภาพที่ 4.20 พบว่า เมื่อใช้จำนวนฟีเจอร์ 100 ลักษณะดังรูปภาพที่ 4.20 (ข) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นตามเวลาจนมีค่ามากสุดที่ 6,700,000 บาท ในช่วงแรกเน้นการลงทุนในสินทรัพย์ที่มีความเสี่ยงมากกว่าการออมและบริโภค แต่ในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน สำหรับการใช้จำนวนฟีเจอร์ 200 ลักษณะดังรูปภาพที่ 4.20 (ค) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นตามเวลาจนมีค่ามากสุดที่ 6,300,000 บาท ในช่วงแรกเน้นการออมมากกว่าการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง แต่ในช่วงหลังจะเน้นการบริโภคมากกว่าการออมและการลงทุน สำหรับการใช้จำนวนฟีเจอร์ 300 ลักษณะดังรูปภาพที่ 4.20 (ง) มูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นตามเวลาจนมีค่ามากสุดที่ 5,300,000 เน้นการลงทุนในสินทรัพย์ที่มีความเสี่ยงมากกว่าการออมและบริโภคตลอดช่วงอายุ

หากเปรียบเทียบคำตอบที่ได้จากการใช้จำนวนฟีเจอร์ทั้ง 3 กรณีกับผลจาก MDP จากรูปภาพที่ 4.20 (ค) พบว่า การใช้ฟีเจอร์จำนวน 300 ลักษณะ มูลค่าทรัพย์สินของครัวเรือนมีความใกล้เคียงกับผลจาก MDP มากที่สุดเนื่องจากสามารถรักษามูลค่าทรัพย์สินของครัวเรือนได้เกือบตลอดช่วงชีวิต และภายหลังจากรายได้ครัวเรือนลดลงที่อายุ 60 ปี การออมมีการเพิ่มขึ้นในขณะที่การลงทุนลดลงเช่นเดียวกับผลจาก MDP นอกจากนี้หากพิจารณาการตัดสินใจทางการเงินของทั้ง 3 กรณีในแต่ละช่วง ช่วงหลังมีลักษณะใกล้เคียงกันคือเน้นการบริโภคมากกว่าการออมและการลงทุนในสินทรัพย์ที่มีความเสี่ยง ในขณะที่ช่วงแรกคำตอบแตกต่างกันมาก ซึ่งเป็นผลจากการอัพเดทค่าน้ำหนักที่ช่วงหลังจะได้รับผลกระทบการประมาณน้อยกว่า



รูปภาพที่ 4.20 มูลค่าสินทรัพย์ของครัวเรือน, รายได้ของครัวเรือน และการวางแผนทางการเงินที่แต่ละเวลาจาก (ก) โปรแกรม MDP ด้วยวิธี Backward Recursive

(ข) การเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้ฟีเจอร์ 100 ลักษณะ

(ค) การเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้ฟีเจอร์ 200 ลักษณะ

(ง) การเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้ฟีเจอร์ 300 ลักษณะ

4.5 ผลของพารามิเตอร์อัตราการเรียนรู้ (η)

อัตราการเรียนรู้ส่งผลต่อการอัพเดทค่าน้ำหนักที่ใช้ในการประมาณ Action-value ซึ่งใช้ในการเลือกการกระทำหรือคำตอบของกรอบปัญหา จาก Sutton an Barto (2018) กล่าวว่าการประมาณ Action-value ควรใช้อัตราการเรียนรู้ที่มีค่าลดลงตามเวลาเพื่อให้คำตอบลู่เข้าสู่ค่าหนึ่ง เนื่องจากถ้าอัตราการเรียนรู้มีค่ามากเกินไป การอัพเดทค่าน้ำหนักแต่ละครั้งจะส่งผลให้เกิดการเปลี่ยนแปลงของคำตอบมากจนอาจทำให้ไม่เกิดการลู่เข้าค่าหนึ่งๆ ของคำตอบได้ แต่ถ้าอัตราการเรียนรู้มีค่าน้อยเกินไป การอัพเดทค่าน้ำหนักแต่ละครั้งจะทำให้เกิดการเปลี่ยนแปลงของคำตอบน้อยมากซึ่งส่งผลให้ต้องใช้เวลาในการคำนวณหลายรอบมากขึ้นเพื่อให้คำตอบลู่เข้าสู่ค่าหนึ่งๆ

ในหัวข้อนี้จึงต้องการศึกษาผลของอัตราการเรียนรู้ใน 2 กรณีคือ

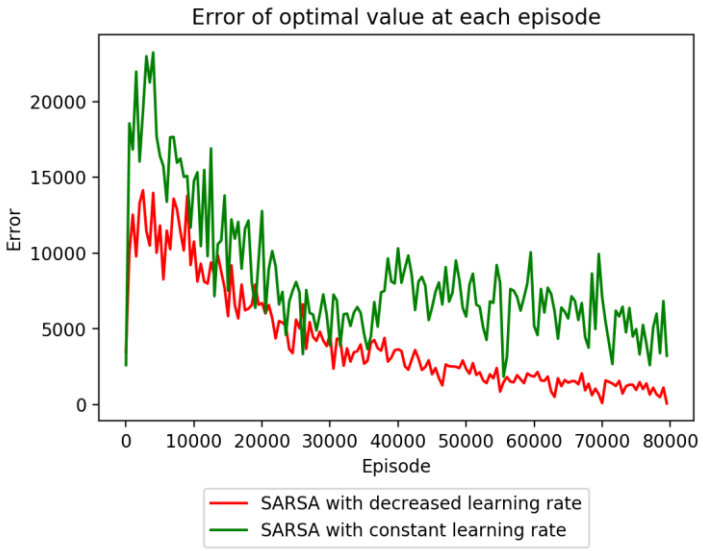
1.) เปรียบเทียบอัตราการเรียนรู้แบบลดลงตามเวลากับอัตราการเรียนรู้คงที่

2.) เปรียบเทียบอัตราการเรียนรู้แบบลดลงตามเวลาที่มีค่าเริ่มต้นแตกต่างกัน

4.5.1 การเปรียบเทียบอัตราการเรียนรู้แบบลดลงตามเวลากับอัตราการเรียนรู้คงที่

เมื่อกำหนดให้อัตราการเรียนรู้มีค่าเริ่มต้นที่ 0.1 แล้วลดลงตามเวลาทุกๆ 1,000 รอบและกำหนดให้อัตราการเรียนรู้มีค่าคงที่ที่ 0.1 ตลอดทุกจำนวนรอบ จากนั้นใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก เพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วคำนวณหาค่าเฉลี่ย โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.21 พบว่า สำหรับการใช้อัตราการเรียนรู้แบบคงที่ ความกว้างของการแกว่งตัวของเส้นกราฟแสดงความผิดพลาดของ Optimal value ค่อนข้างสม่ำเสมอตลอดช่วง ในขณะที่การใช้อัตราการเรียนรู้แบบลดลงตามเวลาในช่วงแรกจนถึงรอบที่ 25,000 เส้นกราฟมีการแกว่งตัวในช่วงกว้าง เนื่องจากความน่าจะเป็นในการเลือกการกระทำแบบสำรวจและอัตราการเรียนรู้ยังมีค่าสูงคือช่วงประมาณ 0.1 – 0.029 จากนั้นจึงแกว่งตัวในช่วงแคบกว่า

นอกจากนี้หากเปรียบค่าความผิดพลาดของการใช้อัตราการเรียนรู้ทั้ง 2 แบบ พบว่า ความผิดพลาดของการใช้อัตราการเรียนรู้แบบคงที่มีมากกว่าการใช้อัตราการเรียนรู้แบบลดลงตามเวลาอย่างชัดเจน โดยมีการลู่เข้าที่ค่าประมาณ 5,000 และ 1,700 ตามลำดับ

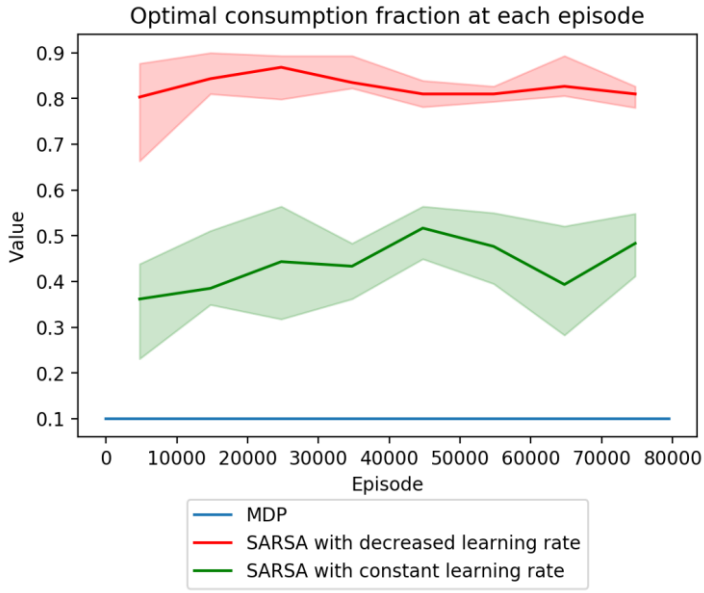


รูปภาพที่ 4.21 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาและแบบคงที่

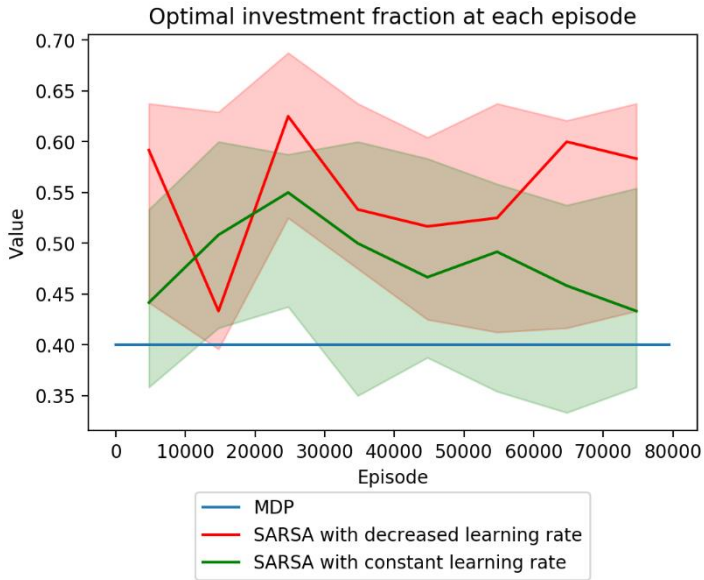
เมื่อพิจารณาผลของการใช้อัตราการเรียนรู้แบบลดลงตามเวลาและแบบคงที่กับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงโดยเปรียบเทียบที่สถานะเริ่มต้น ซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.22 และ 4.23 ตามลำดับ จากรูปภาพที่ 4.22 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาลู่เข้าค่า 0.8 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.1 ในขณะที่ผลจากการใช้อัตราการเรียนรู้แบบคงที่ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจน และจากรูปภาพที่ 4.23 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาลู่เข้าสู่ค่า 0.6 ซึ่งมีค่าใกล้เคียงกับผลจากวิธี MDP ที่มีค่า 0.4 ในขณะที่ผลจากการใช้อัตราการเรียนรู้แบบคงที่ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจนเช่นเดียวกัน และถึงแม้ความผิดพลาดของ Optimal value จะแสดงถึงการลู่เข้าสู่ค่าหนึ่ง คำตอบที่ได้ยังมีความผิดพลาดสูงอยู่

อย่างไรก็ตามขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคของการใช้อัตราการเรียนรู้แบบลดลงตามเวลาแคบลงเมื่อจำนวนรอบมากขึ้นอย่างชัดเจน ในขณะที่ความกว้างระหว่างขอบบนและขอบล่างของการใช้อัตราการเรียนรู้แบบคงที่ค่อนข้างคงที่ สำหรับขอบบนและ

ขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้อัตราการเรียนรู้ทั้งแบบลดลงตามเวลาและแบบคงที่ค่อนข้างสม่ำเสมอ ไม่แสดงการแคบลงอย่างชัดเจน



รูปภาพที่ 4.22 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาและแบบคงที่

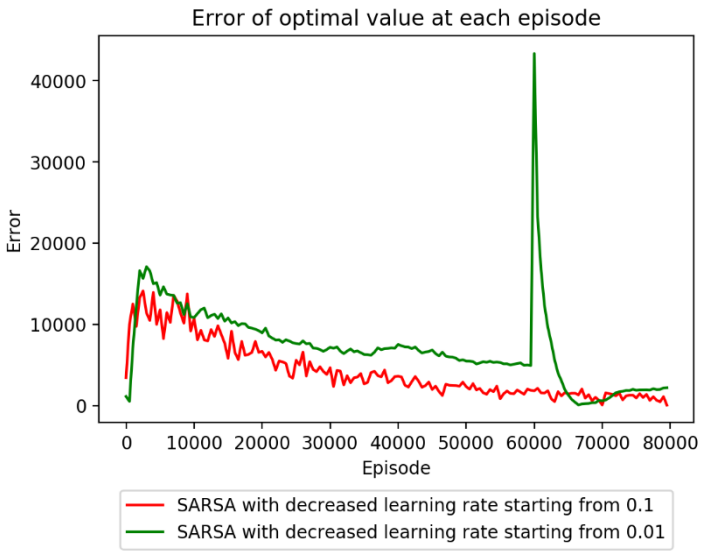


รูปภาพที่ 4.23 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาและแบบคงที่

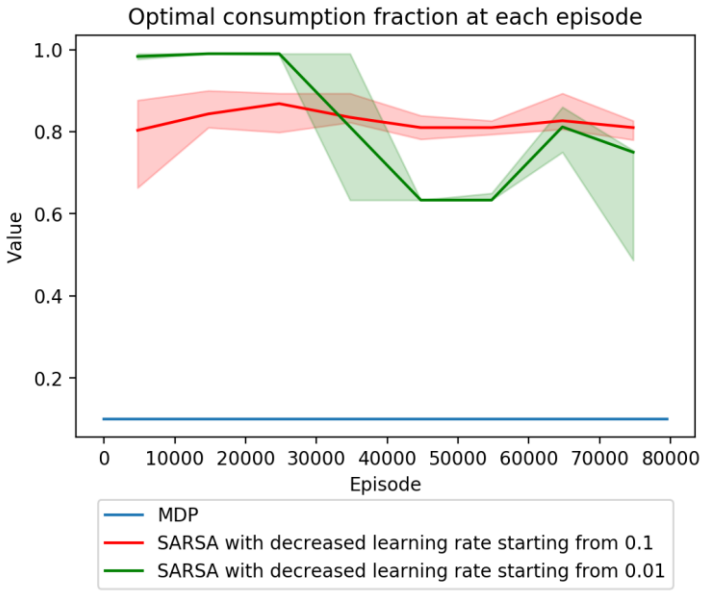
4.5.2 เปรียบเทียบอัตราการเรียนรู้แบบลดลงตามเวลาที่มีค่าเริ่มต้นแตกต่างกัน

เมื่อกำหนดให้อัตราการเรียนรู้มีค่าเริ่มต้นที่ 0.1 และ 0.01 แล้วลดลงตามเวลาทุกๆ 1,000 รอบ จากนั้นใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก เพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วคำนวณหาค่าเฉลี่ย โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.24 พบว่า สำหรับการใช้อัตราการเรียนรู้แบบลดลงที่มีค่าเริ่มต้น 0.01 เส้นกราฟแสดงความผิดพลาดของ Optimal value มีการแกว่งตัวแคบกว่าการใช้อัตราการเรียนรู้แบบลดลงที่มีค่าเริ่มต้น 0.1 อย่างชัดเจน ยกเว้นช่วงประมาณรอบที่ 60,000 ที่ความผิดพลาดมีการแกว่งตัวกว้าง คาดว่าเกิดจากการเลือกการกระทำแบบสำรวจที่ทำให้เกิดการอัพเดทค่าน้ำหนักซึ่งส่งผลต่อการเปลี่ยนแปลงของ Optimal value

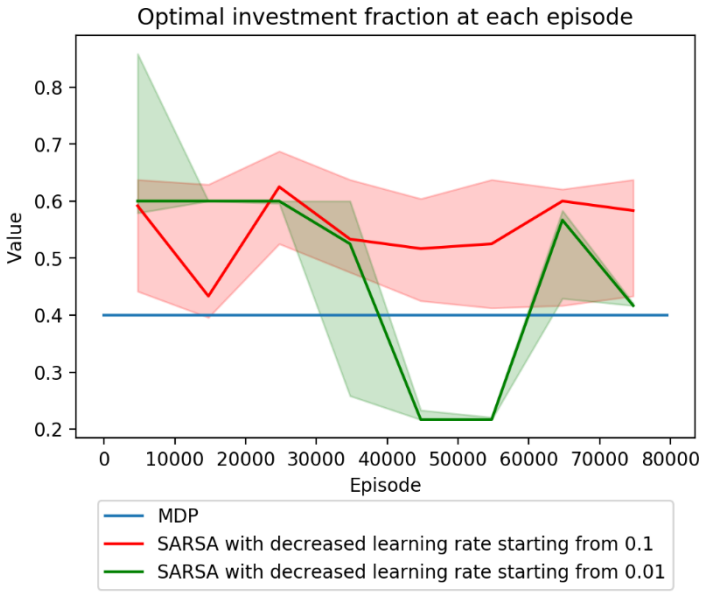
นอกจากนี้หากเปรียบค่าความผิดพลาดของการใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้นทั้ง 2 แบบ พบว่า ความผิดพลาดของการใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.1 สามารถปรับค่าเพื่อลู่เข้าสู่ค่าหนึ่งได้เร็วกว่า โดยลู่เข้าสู่ค่าประมาณ 1,700 ที่ 65,000 รอบ ในขณะที่การใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.01 ยังไม่สามารถสรุปได้แน่ชัดว่ามีการลู่เข้าหรือไม่ที่ 80,000 รอบ



รูปภาพที่ 4.24 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลง



รูปภาพที่ 4.25 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลา



รูปภาพที่ 4.26 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลา

เมื่อพิจารณาผลของการใช้อัตราการเรียนรู้แบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.01 กับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้น ซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.25 และ 4.26 ตามลำดับ จากรูปภาพที่ 4.25 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้อัตราการเรียนรู้แบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 ลู่เข้าค่า 0.8 ในขณะที่ผลจากการใช้อัตราเรียนรู้ที่มีค่าเริ่มต้น 0.01 ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจน และจากรูปภาพที่ 4.26 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธีการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกเมื่อใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.1 ลู่เข้าสู่ค่า 0.6 ในขณะที่ผลจากการใช้อัตราเรียนรู้ที่มีค่าเริ่มต้น 0.01 ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจนเช่นเดียวกัน

อย่างไรก็ตามขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคของการใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.1 แคบลงตามเมื่อจำนวนรอบมากขึ้นอย่างชัดเจน ในขณะที่ความกว้างระหว่างขอบบนและขอบล่างของการใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.01 ไม่มีรูปแบบที่ชัดเจน สำหรับขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.1 ค่อนข้างสม่ำเสมอ ไม่แสดงการแคบลงอย่างชัดเจน ในขณะที่ความกว้างระหว่างขอบบนและขอบล่างของการใช้อัตราการเรียนรู้ที่มีค่าเริ่มต้น 0.01 ไม่มีรูปแบบที่ชัดเจน

4.6 ผลของพารามิเตอร์ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจ (ε)

ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจส่งผลต่อการเลือกการกระทำหรือคำตอบของกรอบปัญหา จาก Sutton and Barto (2018) กล่าวว่าการประมาณ Action-value ควรใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าลดลงตามเวลาเพื่อให้คำตอบลู่เข้าสู่ค่าหนึ่ง และถ้าความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่ามากเกินไป ทำให้อาจเลือกการกระทำที่เป็นการสำรวจมากเกินไปจนการประมาณไม่เกิดการลู่เข้าค่าหนึ่งๆ ของคำตอบได้ แต่ถ้าความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าน้อยเกินไป ทำให้เลือกการกระทำที่เป็นการแสวงประโยชน์มากเกินไปจนไม่ได้เลือกการกระทำอื่นๆ เพียงพอ ส่งผลให้การประมาณ Action-value เกิดความเบี่ยงเบน

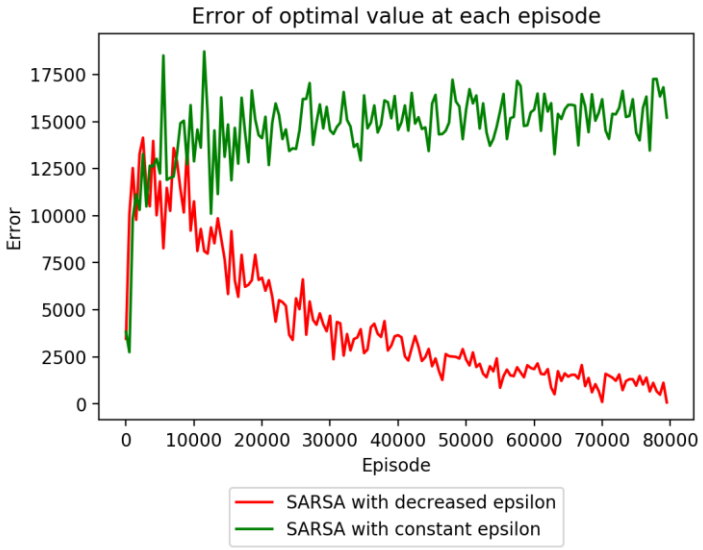
ในหัวข้อนี้จึงต้องการศึกษาผลของความน่าจะเป็นในการเลือกการกระทำแบบสำรวจใน 2 กรณีคือ

1.) เปรียบเทียบความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลากับแบบคงที่
2.) เปรียบเทียบความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้นแตกต่างกัน

4.6.1 การเปรียบเทียบความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลากับแบบคงที่

เมื่อกำหนดให้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าเริ่มต้นที่ 0.9 แล้วลดลงตามเวลาทุกๆ 1,000 รอบและกำหนดให้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าคงที่ที่ 0.7 ตลอดทุกจำนวนรอบ จากนั้นใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรก เพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วคำนวณหาค่าเฉลี่ย โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.27 พบว่า สำหรับการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่ ความกว้างของการแกว่งตัวของเส้นกราฟแสดงความผิดพลาดของ Optimal value ค่อนข้างสม่ำเสมอตลอดช่วง ในขณะที่การใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลา ในช่วงแรกจนถึงรอบที่ 25,000 เส้นกราฟมีการแกว่งตัวในช่วงกว้าง เนื่องจากความน่าจะเป็นในการเลือกการกระทำแบบสำรวจยังมีค่าสูงคือช่วงประมาณ 0.9 – 0.26 จากนั้นจึงแกว่งตัวในช่วงแคบกว่าเมื่อความน่าจะเป็นในการเลือกการกระทำแบบสำรวจเข้าใกล้ 0

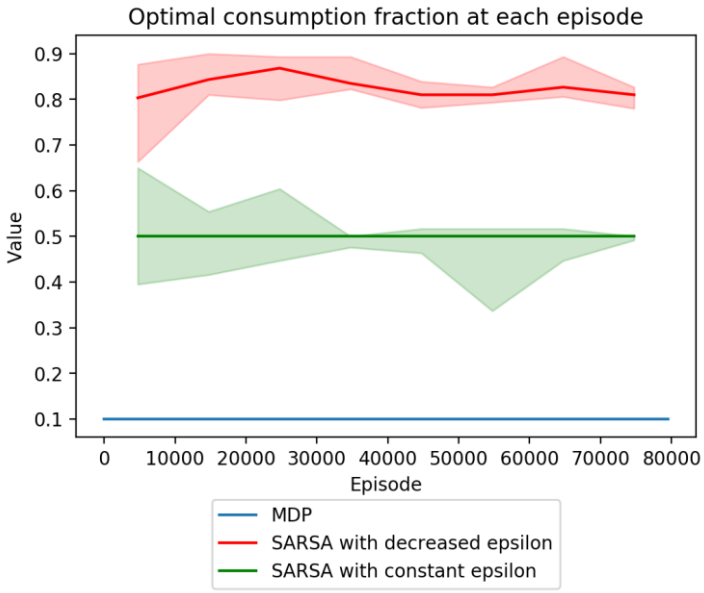
นอกจากนี้หากเปรียบค่าความผิดพลาดของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้ง 2 แบบ พบว่า ความผิดพลาดของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่ไม่สามารถสรุปได้ว่าลู่เข้าหรือไม่ที่ 80,000 รอบ แต่ก็มีค่ามากกว่าความผิดพลาดเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาอย่างชัดเจน เนื่องจากการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่มีโอกาสน้อยที่จะเลือกการกระทำแบบแสวงประโยชน์ ทำให้ Optimal value ที่ได้มีความผิดพลาดสูง



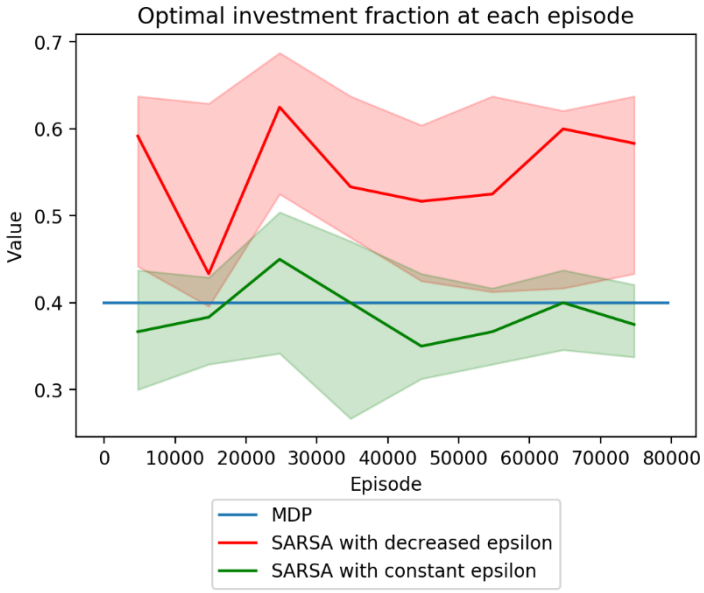
รูปภาพที่ 4.27 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาและแบบคงที่

เมื่อพิจารณาผลของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาและแบบคงที่กับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้น ซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.28 และ 4.29 ตามลำดับ จากรูปภาพที่ 4.28 พบว่า ขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้ง 2 กรณีแคบลงเมื่อจำนวนรอบมากขึ้น ซึ่งในกรณีของความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่เป็นไปได้ว่าการเลือกการกระทำแบบสำรวจที่เกิดขึ้นไม่กระทบต่อ Optimal value ทำให้การกระทำที่เหมาะสมมีการเปลี่ยนแปลงไม่มาก และจากรูปภาพที่ 4.29 พบว่า ขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้งแบบลดลงตามเวลาและแบบคงที่ค่อนข้างสม่ำเสมอ ไม่แสดงการแคบลงอย่างชัดเจน

เนื่องจากคำตอบจากวิธีการเรียนรู้แบบเสริมกำลังเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบคงที่ไม่สามารถสรุปได้ว่าลู่เข้าสู่ค่าหนึ่ง การนำคำตอบจากการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้ง 2 แบบมาเปรียบเทียบจึงไม่เหมาะสมนัก แต่ก็สามารถเห็นได้ว่าคำตอบที่ได้ยังมีความผิดพลาดสูงอยู่



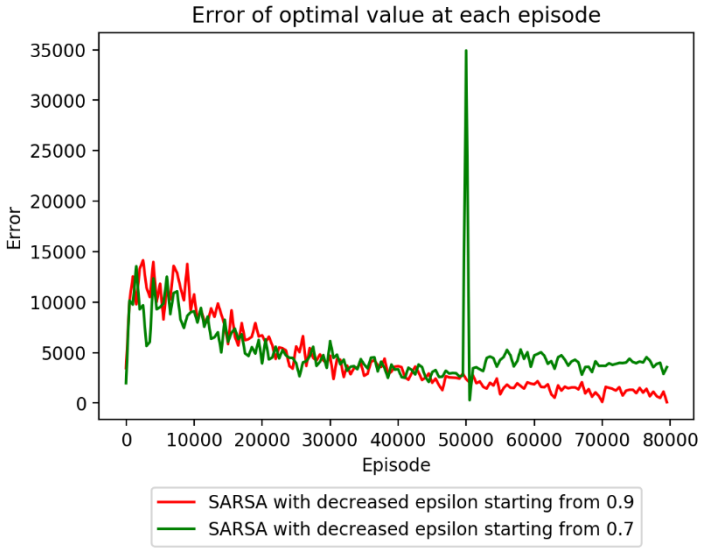
รูปภาพที่ 4.28 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาและแบบคงที่



รูปภาพที่ 4.29 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาและแบบคงที่

4.6.2 เปรียบเทียบความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้นแตกต่างกัน

เมื่อกำหนดให้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจมีค่าเริ่มต้นที่ 0.9 และ 0.7 แล้วลดลงตามเวลาทุกๆ 1,000 รอบ จากนั้นใช้อัลกอริทึม SARSA แบบเน้นการสำรวจในช่วงแรกเพื่อเปรียบเทียบความผิดพลาดของ Optimal value ซึ่งถูกคำนวณจากความแตกต่างระหว่าง Optimal value ของการเรียนรู้แบบเสริมกำลังจำนวน 80,000 รอบและของวิธี Backward recursive โดยทำซ้ำทั้งหมด 3 ครั้งแล้วคำนวณหาค่าเฉลี่ย โดยการเปรียบเทียบพิจารณาจากสถานะเริ่มต้น ซึ่งผลที่ได้ของแต่ละรอบถูกแสดงดังในรูปภาพที่ 4.30 พบว่า สำหรับการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงที่มีค่าเริ่มต้น 0.9 และ 0.7 เส้นกราฟแสดงความผิดพลาดของ Optimal value มีการแกว่งตัวในลักษณะใกล้เคียงกัน โดยในช่วงแรกจะแกว่งกว้างกว่าช่วงหลัง ยกเว้นช่วงประมาณรอบที่ 50,000 ที่ความผิดพลาดมีการแกว่งตัวกว้าง คาดว่าเกิดจากการเลือกการกระทำแบบสำรวจที่ทำให้เกิดการอัพเดทค่าน้ำหนักซึ่งส่งผลต่อการเปลี่ยนแปลงของ Optimal value

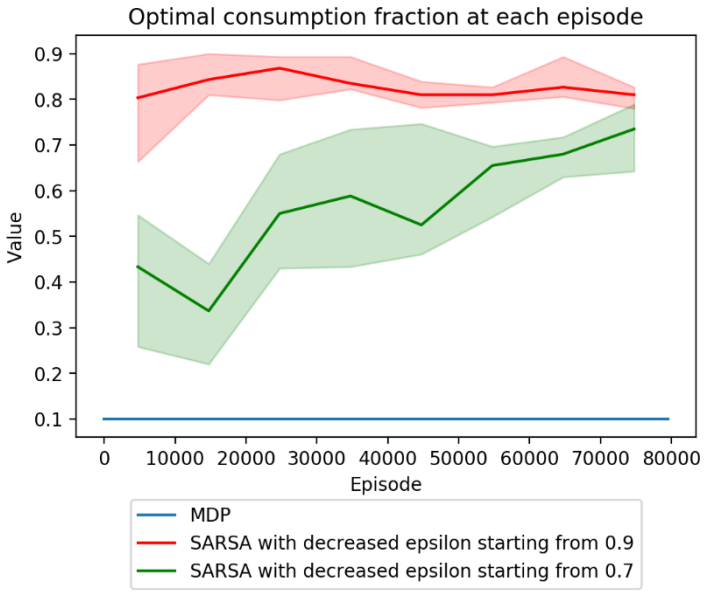


รูปภาพที่ 4.30 ความผิดพลาดของ Optimal Value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลา

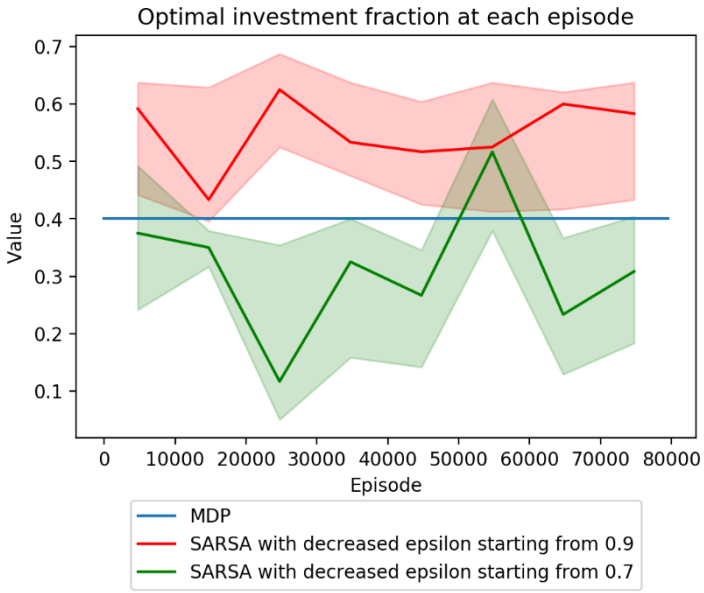
นอกจากนี้ค่าความผิดพลาดของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้นทั้ง 2 แบบสามารถลู่เข้าสู่ค่าหนึ่งได้ที่จำนวนรอบใกล้เคียงกันคือที่ประมาณ 65,000 รอบ อย่างไรก็ตามความผิดพลาดของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.9 สามารถลู่เข้าสู่ค่าประมาณ 1,700 ซึ่งน้อยกว่าความผิดพลาดกรณีที่มีค่าเริ่มต้น 0.7 ซึ่งลู่เข้าสู่ค่าประมาณ 5,000

เมื่อพิจารณาผลของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.9 และ 0.7 กับการวางแผนทางการเงินจากอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยง โดยเปรียบเทียบที่สถานะเริ่มต้น ซึ่งถูกแสดงด้วยค่ามัธยฐานพร้อมทั้งขอบบนและขอบล่างซึ่งมาจากข้อมูลที่เปอร์เซนต์ไทล์ที่ 25 และ 75 ตามลำดับสำหรับข้อมูลทุกๆ 10,000 รอบดังรูปภาพที่ 4.31 และ 4.32 ตามลำดับ จากรูปภาพที่ 4.31 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธีการเรียนรู้แบบเสริมกำลังเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.9 ลู่เข้าค่า 0.8 ในขณะที่ผลจากการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.7 ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจน และจากรูปภาพที่ 4.32 พบว่า อัตราส่วนของสินทรัพย์ที่ใช้ในลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธีการเรียนรู้แบบเสริมกำลังเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.9 ลู่เข้าสู่ค่า 0.6 ในขณะที่ผลจากการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจที่มีค่าเริ่มต้น 0.7 ยังไม่แสดงถึงการลู่เข้าสู่ค่าหนึ่งอย่างชัดเจนเช่นเดียวกัน

ขอบบนและขอบล่างของอัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคของการใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้งสองกรณีแคบลงเมื่อจำนวนรอบมากขึ้น ในขณะที่ขอบบนและขอบล่างของอัตราส่วนที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจทั้งสองกรณีค่อนข้างสม่ำเสมอ ไม่แสดงการแคบลงอย่างชัดเจน



รูปภาพที่ 4.31 อัตราส่วนของสินทรัพย์ที่ใช้ในการบริโภคจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลา



รูปภาพที่ 4.32 อัตราส่วนของสินทรัพย์ที่ใช้ในการลงทุนในสินทรัพย์ที่มีความเสี่ยงจากวิธี MDP และวิธีการเรียนรู้แบบเสริมกำลังจากโปรแกรมแบบเน้นการสำรวจในช่วงแรกที่แต่ละรอบเมื่อใช้ความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลา

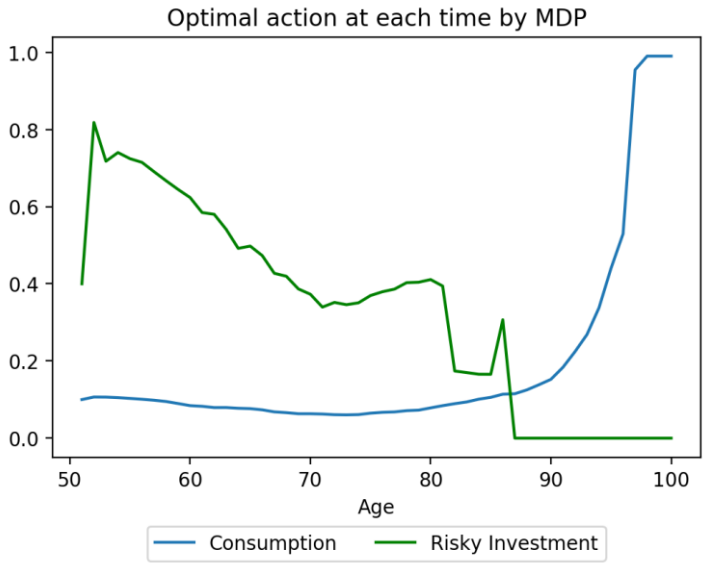
4.7 การพิจารณาความสัมพันธ์ของ Optimal action

เมื่อนำผลลัพธ์ที่ได้จากทั้งโปรแกรม MDP และการเรียนรู้แบบเสริมกำลังไปจำลองกับข้อมูลจำเพาะของครัวเรือน จะได้นโยบายที่เหมาะสมที่สุด (Optimal policy) ซึ่งเป็นการเลือกการกระทำที่เหมาะสมที่สุด (Optimal action) นั่นคือกิจกรรมทางการเงิน ได้แก่ การบริโภค, การลงทุนในสินทรัพย์ที่มีความเสี่ยง และการออมที่อายุต่างๆ ตลอดช่วงชีวิตของหัวหน้าครัวเรือน ที่แต่ละอายุการบริโภค, การลงทุนในสินทรัพย์ที่มีความเสี่ยง และการออมจะที่ค่าแตกต่างกันไป ดังนั้นในหัวข้อนี้จึงต้องการพิจารณาความสัมพันธ์ระหว่างอัตราส่วนที่ใช้ในการบริโภคและอัตราส่วนที่ใช้ในการลงทุนสินทรัพย์ที่มีความเสี่ยงทั้งจากโปรแกรม MDP และการเรียนรู้แบบเสริมกำลัง

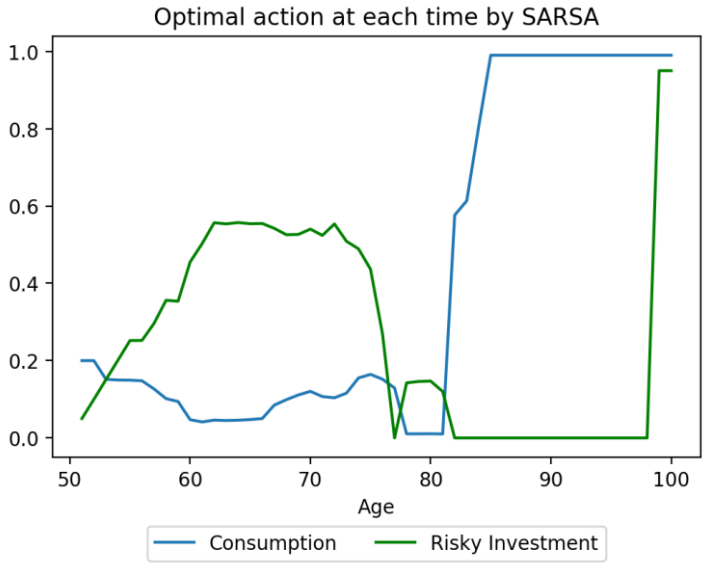
ซึ่งจะพิจารณาเฉพาะกรณีที่ความผิดพลาด Optimal value ระหว่างวิธี MDP และการเรียนรู้แบบเสริมกำลังมีค่าต่ำสุด นั่นคือเมื่อใช้โปรแกรมการเรียนรู้แบบเสริมกำลังที่เน้นการสำรวจในช่วงแรก, ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 200 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับ โดยนโยบายที่เหมาะสมที่สุดที่ได้จากการจำลองทั้งหมด 5,000 ครั้งด้วยโปรแกรม MDP และการเรียนรู้แบบเสริมกำลัง แล้วนำไปหาค่ามัธยฐานที่แต่ละอายุถูกแสดงดังรูปภาพที่ 4.33

จากรูปภาพที่ 4.33 (ก) ในช่วงอายุ 50 – 80 ปีอัตราส่วนที่ใช้ในการบริโภคจากโปรแกรม MDP มีค่าค่อนข้างคงที่ ในขณะที่อัตราส่วนที่ใช้ในการลงทุนมีความผันผวนค่อนข้างมาก กล่าวคือมีการเพิ่มขึ้นแล้วลดลงในช่วงแรก และเพิ่มขึ้นแล้วลดลงอีกครั้งในช่วงอายุ 70 – 80 ปี และที่ช่วงอายุหลัง 80 ปี อัตราส่วนที่ใช้ในการบริโภคเพิ่มขึ้น ในขณะที่อัตราส่วนที่ใช้ในการลงทุนลดลงและคงที่ แสดงให้เห็นว่าอัตราส่วนทั้งสองไม่แสดงถึงความสัมพันธ์ระหว่างกัน

จากรูปภาพที่ 4.33 (ก) ในช่วงอายุ 50 – 60 ปีอัตราส่วนที่ใช้ในการบริโภคจากโปรแกรมการเรียนรู้แบบเสริมกำลังลดลง ในขณะที่อัตราส่วนที่ใช้ในการลงทุนเพิ่มขึ้น ที่ช่วงอายุ 60 – 75 ปี อัตราส่วนที่ใช้ในการบริโภคลดลง ในขณะที่อัตราส่วนที่ใช้ในการลงทุนค่อนข้างคงที่ และที่ช่วงอายุหลัง 80 ปี อัตราส่วนที่ใช้ในการบริโภคเพิ่มขึ้นแล้วคงที่ ในขณะที่อัตราส่วนที่ใช้ในการลงทุนค่อนข้างคงที่แล้วจึงเพิ่มขึ้นในช่วงหลัง แสดงให้เห็นว่าอัตราส่วนทั้งสองไม่แสดงถึงความสัมพันธ์ระหว่างกัน



(ก)



(ข)

รูปภาพที่ 4.33 อัตราส่วนที่ใช้ในการบริโภคและการลงทุนในสินทรัพย์ที่มีความเสี่ยงที่เหมาะสมที่สุดที่แต่ละอายุตลอดช่วงชีวิตของหัวหน้าครัวเรือน ซึ่งเป็นค่ามัธยฐานของการจำลองทั้งหมด 5,000 รอบ จาก

(ก) โปรแกรม MDP ด้วยวิธี Backward recursive

(ข) โปรแกรมการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรก

บทที่ 5
สรุปผลการวิจัยและข้อเสนอแนะ

5.1 สรุปผลการวิจัย

ตามจุดประสงค์ของงานวิจัยนี้ที่ต้องการศึกษาการประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงิน และเปรียบเทียบผลที่ได้กับผลจาก MDP โดยศึกษาจากความผิดพลาดระหว่างทั้งสองวิธี ดังนั้นการสรุปผลจะแบ่งออกเป็น 4 ส่วนดังต่อไปนี้

5.1.1 การประยุกต์ใช้การเรียนรู้แบบเสริมกำลังกับการวางแผนทางการเงิน

MDP เป็นการหาคำตอบที่แท้จริงด้วยวิธี Backward recursive ในขณะที่การเรียนรู้แบบเสริมกำลังเป็นการหาคำตอบด้วยการประมาณ Action-value ซึ่งในงานวิจัยนี้ Action-value ถูกประมาณด้วยตัวแบบการถดถอยเชิงเส้นที่ใช้ฟีเจอร์แบบ RBF เป็นตัวแปรต้น การอัพเดทค่าน้ำหนักสำหรับตัวแบบใช้อัลกอริทึม SARSA โดยแบ่งเป็นอัลกอริทึมแบบปกติและแบบเน้นการสำรวจในช่วงแรกที่จะเลือกการกระทำที่ให้ค่า Action-value สูงสุดจากการกระทำที่ถูกสุ่มเลือกมาถึงแม้ว่าต้องการเลือกการกระทำแบบแสวงประโยชน์

จากการศึกษาคำตอบที่ได้จากวิธีการเรียนรู้แบบเสริมกำลังด้วยอัลกอริทึมทั้งแบบปกติและแบบเน้นการสำรวจในช่วงแรกเมื่อกำหนดพารามิเตอร์ที่เหมาะสมมีแนวโน้มในการลู่เข้าสู่ค่าหนึ่ง นั่นแสดงว่าวิธีการเรียนรู้แบบเสริมกำลังสามารถหาคำตอบจากการประมาณโดยใช้ตัวแบบการถดถอยเชิงเส้นได้

ความแตกต่างระหว่างอัลกอริทึมทั้งสองคือเวลาที่ใช้ในการคำนวณสำหรับการเลือกการกระทำแบบแสวงประโยชน์ในช่วงแรก เนื่องจากอัลกอริทึมแบบเน้นการสำรวจในช่วงแรกถูกปรับปรุงให้การเลือกการกระทำแบบแสวงประโยชน์ในช่วง 35,000 รอบแรกจะทำงานคล้ายกับการเลือกการกระทำแบบสำรวจ ทำให้ช่วยลดเวลาในการคำนวณลงได้ถึง 97% สำหรับการเลือกการกระทำแบบแสวงประโยชน์

เมื่อพิจารณาเทียบวิธีการเรียนรู้แบบเสริมกำลังกับ MDP สำหรับกรอบปัญหาการวางแผนทางการเงินที่มีจำนวนสถานะทั้งหมด 156 สถานะและการกระทำทั้งหมด 441 โดยในกรณีศึกษาหัวหน้าครัวเรือนมีอายุเหลืออีก 50 ปี โปรแกรม MDP ใช้เวลาในการคำนวณประมาณ 80 วินาที ในขณะที่โปรแกรมการเรียนรู้แบบเสริมกำลังแบบเน้นการสำรวจในช่วงแรกต้องใช้เวลาในการเรียนรู้

เพื่ออัพเดทค่าน้ำหนักจนกระทั่งเกิดการลู่เข้าสู่ค่าหนึ่งประมาณ 10,000 นาที ดังนั้นสำหรับกรณีศึกษาวิธีการเรียนรู้แบบเสริมกำลังไม่เหมาะสมที่จะใช้ในการแก้ปัญหาเนื่องจากต้องใช้เวลาในการเรียนรู้นาน แต่ถ้ากรอบปัญหามีความซับซ้อนมากขึ้นทั้งจากจำนวนสถานะหรือจำนวนการกระทำจนทำให้โปรแกรม MDP ใช้เวลาคำนวณมากกว่า 10,000 นาที วิธีการเรียนรู้แบบเสริมกำลังอาจมีความเหมาะสมที่จะใช้ในการหาคำตอบ

5.1.2 ความผิดพลาดระหว่าง MDP และการเรียนรู้แบบเสริมกำลัง

เนื่องจากเวลาที่ใช้ในการคำนวณเป็นสิ่งที่ให้ความสำคัญในการใช้งานจริง จึงพิจารณาเปรียบเทียบผลโดยอาศัยอัลกอริทึมแบบเน้นการสำรวจในช่วงแรกเป็นหลัก โดยจากการศึกษาผลของปัจจัยในกรณีต่างๆ สามารถสรุปผลเกี่ยวกับความผิดพลาดระหว่างสองวิธีได้ใน 2 ส่วนดังนี้

ประการแรก ความผิดพลาดของ Optimal value ที่สถานะเริ่มต้นระหว่างสองวิธีมีลักษณะคล้ายกันในทุกกรณีคือช่วงแรกมีการแกว่งตัวกว้างแล้วจึงแกว่งตัวแคบลง เมื่ออัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจลดลง แสดงให้เห็นถึงการลู่เข้าสู่ค่าหนึ่งๆ ของคำตอบที่จำนวนรอบต่างกันไปในแต่ละกรณี ยกเว้นเมื่อกำหนดให้อัตราการเรียนรู้หรือความน่าจะเป็นในการเลือกการกระทำแบบสำรวจเป็นค่าคงที่ ไม่แสดงการลู่เข้าของคำตอบอย่างชัดเจน หากพิจารณาถึงความกว้างของการแกว่งตัวของค่าความผิดพลาด บางปัจจัยส่งผลให้มีการแกว่งตัวกว้างกว่า ได้แก่ จำนวนฟีเจอร์ที่ใช้มากขึ้น และการกำหนดค่าน้ำหนักเริ่มต้นแบบ Optimistic ซึ่งในที่นี้เป็นค่าน้ำหนักจากตัวแบบการถดถอยเชิงเส้นที่มีรางวัลในขณะนั้นเป็นตัวแปรตาม สำหรับขนาดของความผิดพลาดของ Optimal value มีค่าน้อยสุดที่ 1,700 เมื่อใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 200 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับ

ประการที่สอง Optimal action สำหรับการวางแผนทางการเงินที่สถานะเริ่มต้นจากวิธีการเรียนรู้แบบเสริมกำลังลู่เข้าสู่ค่าหนึ่งๆ อย่างไรก็ตามถึงแม้ความผิดพลาดของ Optimal value จะมีค่าต่ำและแสดงว่าหาคำตอบได้ Optimal action จากวิธีการเรียนรู้แบบเสริมกำลังยังมีความแตกต่างสูงเมื่อเทียบกับ Optimal action จาก MDP นอกจากนี้จากการศึกษาผลของปัจจัยต่างๆ ไม่สามารถสรุปได้ว่ากรณีใดที่ให้ Optimal action ใกล้เคียงกับวิธี MDP มากที่สุดได้ แต่สามารถสรุปถึงความสัมพันธ์ระหว่างอัตราส่วนที่ใช้ในการบริโภคและการลงทุนจากวิธีการเรียนรู้แบบเสริมกำลังได้ว่า

มีความผกผันกัน นั่นคือเมื่ออัตราส่วนที่ใช้ในการบริโภคมีค่ามาก อัตราส่วนที่ใช้ในการลงทุนจะมีค่าน้อยสำหรับทุกกรณีที่คำตอบที่ดีที่สุดมีการลู่เข้า

5.1.3 การนำคำตอบที่ดีที่สุดไปใช้กับการวางแผนทางการเงิน

การนำคำตอบที่ได้ไปจำลองกับการวางแผนทางการเงินตลอดช่วงอายุของหัวหน้าครัวเรือน เมื่อเทียบผลที่ได้จากวิธีการเรียนรู้แบบเสริมกำลังและ MDP ในทุกกรณี พบว่า ผลที่ได้มีความแตกต่างกันมากซึ่งสอดคล้องกับข้อสรุปในประการที่สอง โดยคำตอบในช่วงลำดับเวลาหลังมีความใกล้เคียงกับ MDP มากกว่าคำตอบในช่วงลำดับเวลาแรก เนื่องจากการอัพเดทค่าน้ำหนักที่ใช้ในการประมาณ Action-value ได้รับผลกระทบจากความเบี่ยงเบนของค่าประมาณน้อยกว่า อย่างไรก็ตาม อัลกอริทึมที่เน้นการสำรวจในช่วงแรกเมื่อใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 300 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับให้ผลลัพธ์ที่ใกล้เคียงกับ MDP มากสุดคือมูลค่าทรัพย์สินของครัวเรือนมีการสะสมเพิ่มขึ้นและถูกรักษาไว้ได้นานที่สุดตามวัตถุประสงค์ที่ต้องการให้เกิดอรรถประโยชน์สูงสุดตลอดชีวิต ในช่วงแรกที่รายได้ครัวเรือนสูงเน้นการลงทุนในสินทรัพย์ที่มีความเสี่ยงมากกว่าการออมและการบริโภคตามลำดับ เมื่อรายได้ครัวเรือนลดลงมีการออมเพิ่มขึ้นจนมากกว่าการลงทุนช่วงหนึ่ง จากนั้นจึงกลับมาเน้นการลงทุน และในช่วงท้ายของลำดับเวลาเลือกบริโภคมากกว่าการออม

5.1.4 ความสัมพันธ์ระหว่างอัตราส่วนที่ใช้ในการบริโภคและการลงทุน

สำหรับโปรแกรม MDP และการเรียนรู้แบบเสริมกำลังที่เน้นการสำรวจในช่วงแรกเมื่อใช้ค่าน้ำหนักเริ่มต้นจากตัวแบบการถดถอยเชิงเส้น, จำนวนฟีเจอร์ 200 ลักษณะ, อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจแบบลดลงตามเวลาที่มีค่าเริ่มต้น 0.1 และ 0.9 ตามลำดับ เมื่อนำ Optimal action สำหรับแต่ละสถานะที่แต่ละเวลาไปจำลองกับการวางแผนทางการเงิน พบว่า อัตราส่วนที่ใช้ในการบริโภคไม่แสดงความสัมพันธ์กับอัตราส่วนที่ใช้ในการลงทุนทั้งในกรณีของโปรแกรม MDP และการเรียนรู้แบบเสริมกำลัง

5.2 ข้อเสนอแนะ

5.2.1 การใช้อัลกอรึทึม SARSA มีความเหมาะสมกับกรอบปัญหาที่มีเซตของการกระทำเป็นเซตของข้อมูลแบบไม่ต่อเนื่อง แต่มีข้อเสียคือต้องใช้เวลาในการคำนวณเพิ่มตามขนาดของเซตของการกระทำ ดังนั้นหากเซตของการกระทำมีขนาดใหญ่มาก อัลกอริทึม SARSA อาจไม่เหมาะสม แต่อาจประยุกต์ทำให้เซตของการกระทำเป็นเซตของข้อมูลแบบต่อเนื่อง แล้วจึงเลือกใช้อัลกอริทึมที่เรียนรู้จากการกระทำโดยตรง เพื่อลดเวลาที่ใช้ในการคำนวณ

5.2.2 ในงานวิจัยนี้ใช้ตัวแบบการถดถอยเชิงเส้นที่ใช้ฟีเจอร์จากเคอร์เนล RBF ประมาณ Action-value เนื่องจากมีความซับซ้อนน้อย อย่างไรก็ตามตัวแบบนี้อาจไม่เหมาะสมกับกรอบปัญหาที่มีเซตของสถานะหรือการกระทำมีขนาดใหญ่ขึ้น จึงอาจใช้ตัวแบบอื่นที่มีความซับซ้อนมากขึ้นเพื่อประมาณ Action-value นอกจากนี้การใช้ตัวแบบการถดถอยเชิงเส้นไม่จำเป็นต้องใช้ตัวแบบเดียวกันตลอดทุกช่วงเวลา จึงอาจใช้กำหนดตัวแบบเพื่อประมาณ Action-value สำหรับที่แต่ละเวลา

5.2.3 การศึกษาผลของพารามิเตอร์ในที่นี้ทำการปรับเปลี่ยนเพียงบางพารามิเตอร์ และเพียงบางค่าของพารามิเตอร์ จึงยังมีรายละเอียดเพิ่มเติมในประเด็นอื่นๆ เช่น

1. การกำหนดค่าน้ำหนักเริ่มต้นแบบ Optimistic ที่มีการใช้ข้อมูลเบื้องต้น ในที่นี้ใช้ค่าน้ำหนักจากตัวแบบการถดถอยเชิงเส้นเมื่อใช้รางวัลในขณะนั้นเป็นตัวแปรตาม สามารถใช้ค่าน้ำหนักจากตัวแบบอื่นๆ ที่ลักษณะคล้ายกัน หรืออาจใช้ค่าอื่นเป็นตัวแปรตามสำหรับตัวแบบการถดถอยเชิงเส้น เช่น การจำลองค่าตัวแปรตามเป็นผลรวมค่ารางวัลที่จะเกิดขึ้น

2. การทำฟีเจอร์ในที่นี้ใช้เคอร์เนล RBF ที่กำหนดพารามิเตอร์ต่างๆ และศึกษาผลของการปรับเปลี่ยนจำนวนฟีเจอร์ทั้งหมด แต่อย่างไรก็ตามยังสามารถศึกษาผลของการใช้เคอร์เนลที่แตกต่างกัน หรืออาจใช้เคอร์เนล RBF แล้วศึกษาผลของการปรับเปลี่ยนพารามิเตอร์สำหรับเคอร์เนล RBF

3. การศึกษาพารามิเตอร์อัตราการเรียนรู้และความน่าจะเป็นในการเลือกการกระทำแบบสำรวจนั้นศึกษาเพียงแค่บางค่า ซึ่งพารามิเตอร์เหล่านี้มีค่าที่เหมาะสมแตกต่างกันในแต่ละกรอบปัญหา และไม่มีแนวทางที่แน่ชัดในการกำหนดค่าเหล่านี้ จึงอาจทำการศึกษาพารามิเตอร์ในช่วงที่กว้างขึ้น เพื่อให้ได้อัลกอริทึมที่มีประสิทธิภาพมากขึ้น

บรรณานุกรม

ณัตติฤดี เจริญรักษ์ และ สุภารัตน์ ตันทะนงศักดิ์กุล. (2559). โครงการศึกษาระดับการออมและความเพียงพอของผู้สูงอายุในชุมชน. รายงานวิจัยเสนอต่อสำนักงานคณะกรรมการวิจัยแห่งชาติ.

นราพงศ์ ศรีวิศาล และคณะ. (2561). โครงการวิเคราะห์ความเพียงพอของการออมเพื่อการดำรงชีพของผู้สูงอายุในชุมชน: ระยะที่ 2. สำนักงานคณะกรรมการวิจัยแห่งชาติ. กรุงเทพมหานคร.

Bauerle, N., & Rieder, U. (2011). Markov Decision Processes with Applications to Finance. *Springer*.

Geramifard, A., Walsh, T. J., Tellex, S., Chowdhary, G., Roy, N., & How, J. P. (2013). A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning. *Foundations and Trends in Machine Learning,* 6(4), 375-454.

Poole, D. L., & Mackworth, A. K. (2017). *Artificial Intelligence: Foundations of Computational Agents*. Cambridge: Cambridge University Press.

Puterman, M. (1994). *Markov Decision Processes – Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons, Inc.

Sripakdeevong, P., Stantcheva, S., & Townsend, R. (2011). *Wealth Management for the Poor*. Presentation at Massachusetts Institute of Technology (MIT).

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*.

Szepesvári, C. (2010). *Algorithms for Reinforcement Learning*: Morgan & Claypool Publishers.

ประวัติผู้เขียน

ชื่อ-สกุล	นางสาว ภัควัลย์ จันทรศิริภาส
วัน เดือน ปี เกิด	22 กรกฎาคม 2536
สถานที่เกิด	กรุงเทพฯ
วุฒิการศึกษา	วิศวกรรมศาสตร์บัณฑิต
ที่อยู่ปัจจุบัน	591/23 ซอยลาดพร้าว 87 ถนนลาดพร้าว แขวงคลองเจ้าคุณสิงห์ เขตวังทองหลาง กรุงเทพฯ 10310