University of Ghana http://ugspace.ug.edu.gh

UNIVERSITY OF GHANA
COLLEGE OF BASIC AND APPLIED SCIENCES
FACULTY OF PHYSICAL AND MATHEMATICAL SCIENCES



AN ACTION SELECTION METHOD USING STUDENT'S T-DISTRIBUTED NOISE FOR OPTIMISING DEEP DETERMINISTIC POLICY GRADIENT ALGORITHM IN RECOMMENDER SYSTEMS

BY
MICHAEL AGBO TETTEY SOLI
(ID. NO. 10249232)

A THESIS SUBMITTED TO THE SCHOOL OF GRADUATE STUDIES IN PARTIAL FULFILMENT OF THE AWARD OF THE DEGREE OF DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

DEPARTMENT OF COMPUTER SCIENCE
OCTOBER, 2025

University of Ghana http://ugspace.ug.edu.gh

UNIVERSITY OF GHANA
COLLEGE OF BASIC AND APPLIED SCIENCES
FACULTY OF PHYSICAL AND MATHEMATICAL SCIENCES

AN ACTION SELECTION METHOD USING STUDENT'S T-DISTRIBUTED NOISE FOR OPTIMISING DEEP DETERMINISTIC POLICY GRADIENT ALGORITHM IN RECOMMENDER SYSTEMS

BY

MICHAEL AGBO TETTEY SOLI
(ID. NO. 10249232)

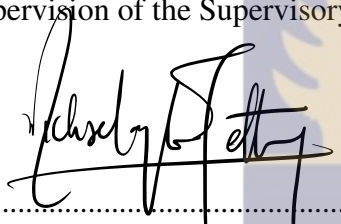
INTEGRI PROCEDAMUS

DEPARTMENT OF COMPUTER SCIENCE
OCTOBER, 2025

University of Ghana http://ugspace.ug.edu.gh

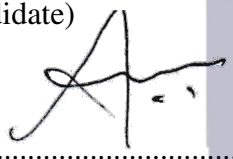
DECLARATION

I, Michael Agbo Tettey Soli, hereby declare that, except where explicit reference is made to the work of others, this dissertation is the result of my own independent research and has not been submitted, either in whole or in part, for the award of any other degree or qualification at this or any other university. This work is entirely my own, except where collaboration is duly acknowledged in the text and in the acknowledgements. All supporting materials, including appendices, bibliography, footnotes, tables, and equations, are included as part of the submission. This dissertation was carried out under the supervision of the Supervisory Committee listed below.



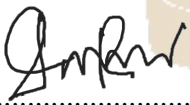
..... Date: 6TH OCTOBER, 2025 .....

MICHAEL AGBO TETTEY SOLI
(Candidate)



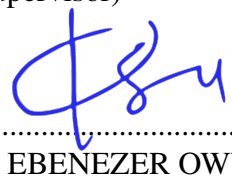
..... Date: 6TH OCTOBER, 2025 .....

PROF. FERDINAND KATSRIKU
(Principal Supervisor)



..... Date: 6TH OCTOBER, 2025 .....

PROF. JUSTICE KWAME APPATI
(Co-Supervisor)



..... Date: 6TH OCTOBER, 2025 .....

PROF. EBENEZER OWUSU
(Co-Supervisor)

University of Ghana http://ugspace.ug.edu.gh

ACKNOWLEDGMENT

First and foremost, I am deeply grateful to Almighty God for the gift of life, strength, and grace that have sustained me throughout this journey. The opportunity to pursue this doctoral program, as well as the many individuals who supported me along the way, are a testament to His boundless mercy and loving kindness. To Him be all the glory.

I extend my profound appreciation to all those who played a role in the successful completion of this work. I am especially indebted to my supervisors, Prof. Ferdinand Katsriku, Prof. Justice Kwame Appati and Prof. Ebenezer Owusu, for their invaluable guidance, leadership, encouragement, unwavering support and academic mentorship throughout this journey.

To my family and friends, your steadfast support, prayers, and encouragement have been a constant source of motivation. From the inception of this research to its completion, you have cheered me on, and for that, I am truly thankful. A heartfelt thank you goes to my parents, whose unwavering belief in my potential has been my foundation. Your sacrifices and love have carried me this far.

Finally, I would like to acknowledge my co-founders at Polymorph Labs Ghana Limited. Your belief in me and your continuous support have been both humbling and inspiring. Thank you for standing with me every step of the way.

University of Ghana http://ugspace.ug.edu.gh

ABSTRACT

The exponential growth of digital content, driven by the proliferation of the internet, e-commerce, social media, and multimedia services, has intensified the problem of information overload. To address this, recommender systems have emerged as indispensable tools for filtering and prioritising relevant content. While traditional systems rely on collaborative filtering or content-based techniques, recent advancements increasingly frame recommendation as a sequential decision-making problem, leveraging the theoretical foundations of Reinforcement Learning (RL). Within this paradigm, Deep Reinforcement Learning (DRL) and specifically actor–critic methods such as the Deep Deterministic Policy Gradient (DDPG) algorithm have shown promise due to their ability to model complex user–item interactions in continuous action spaces.

Despite their potential, DRL algorithms are susceptible to fundamental issues like overestimation bias and the compounding of approximation errors in Temporal Difference (TD) learning. In recommender systems, these shortcomings can cause the agent to overcommit to a narrow set of high-reward items, diminishing recommendation diversity and limiting long-term user engagement. A common remedy involves injecting exploration noise into the policy's action space. By encouraging deviation from deterministic behaviour, this noise enhances the agent's capacity to discover underexplored but potentially rewarding actions, thereby mitigating overestimation and promoting robust policy learning.

This thesis develops a parameterised heavy-tailed stochastic exploration framework for actor–critic reinforcement learning, defined by three design parameters: the noise distribution family, the scale schedule, and the bound region. The framework is instantiated

University of Ghana http://ugspace.ug.edu.gh

by integrating Student's t-distributed action perturbation, linear scale annealing, and value clipping into DDPG, resulting in the proposed t-DDPG method. Unlike standard Gaussian or Ornstein–Uhlenbeck (OU) noise, both light-tailed distributions that produce predominantly local perturbations, the heavy-tailed Student's t-distribution generates occasional large excursions while retaining mathematical tractability and bounded operating behaviour.

The framework is evaluated using three benchmark recommendation datasets: MovieLens 100K, MovieLens 1M, and Jester D4. Its performance was compared against three baseline DDPG configurations: vanilla DDPG without exploration noise, DDPG with Gaussian noise, and DDPG with Ornstein–Uhlenbeck noise. The proposed t-DDPG variant achieved the highest HR@10 and nDCG@10 across all three datasets. On MovieLens 100K, evaluated on ten random seeds, t-DDPG attained HR@10 of 0.4543 ± 0.0193 and nDCG@10 of 0.2375 ± 0.0136, exceeding the Gaussian baseline by 6.5% and 8.0% relative respectively. A Bonferroni-corrected statistical analysis at family-wise $\alpha = 0.05$ confirmed significance for all six pairwise comparisons (three baselines $\times$ two metrics) under both the paired t-test and the non-parametric Wilcoxon signed-rank test. On MovieLens 1M and Jester D4, evaluated on three random seeds, t-DDPG again outperformed all baselines on both metrics. Notably, t-DDPG was the only configuration of the four to deliver a positive gain over the deterministic baseline on every dataset tested; the lighter-tailed alternatives underperformed the deterministic baseline on Jester D4. These findings establish heavy-tailed action-level perturbation as a theoretically grounded and empirically validated exploration mechanism for actor–critic recommendation, with the framework formulation supporting principled extension to alternative heavy-tailed distributions and to other actor–critic methods.

University of Ghana http://ugspace.ug.edu.gh

Table of Contents

Declaration ..... **i**

Acknowledgment ..... **ii**

Abstract ..... **iii**

List of Figures ..... **ix**

List of Tables ..... **x**

List of Abbreviations ..... **xi**

1 Introduction **1**

1.1 Background 1

1.2 Reinforcement Learning for Recommender Systems 7

1.3 Research Motivations 12

1.4 Problem Statement 15

1.5 Research Gap 17

1.6 Research Aims and Objectives 20

1.7 Research Questions and Hypotheses 21

1.8 Significance of the Study 23

1.9 Definition of Key Terminologies 24

University of Ghana http://ugspace.ug.edu.gh

1.10 Outline of the Thesis 26

2 Literature Review 28

2.1 Introduction 28

2.2 Recommender Systems Overview 29

2.3 General vs. Sequential Recommendation 35

2.4 Deep Reinforcement Learning for Recommendation 38

2.5 The DDPG Algorithm and Its Limitations 49

2.6 Exploration Noise in Deep Reinforcement Learning 54

2.7 Related Works on Exploration Improvements in Deep Reinforcement Learning 64

2.8 Conceptual Framework 72

2.9 Conclusion 77

3 Methodology 80

3.1 Introduction 80

3.2 Preliminaries: MDP Formulation for Sequential Recommendation 81

3.3 Problem Formulation 83

3.4 Datasets 86

3.5 Baseline Algorithm: Vanilla DDPG Architecture 90

3.6 Proposed Method: Student's *t*-Noise for Exploration 95

3.7 Training Pipeline of the Proposed DDPG with Student's *t*-Noise Variant 109

3.8 Experiment Design 118

3.9 Summary 122

4 Experiments and Results 126

University of Ghana http://ugspace.ug.edu.gh

4.1 Experiment Setup 126
4.2 Training and Validation Process 128
4.3 Ablation Studies 131
4.4 Performance Comparison 136
4.5 Generalisation Across Datasets 142
4.6 Training Dynamics and Convergence Behaviour 145
4.7 Summary 152

5 Discussion 154
5.1 Introduction 154
5.2 Answering the Research Questions 155
5.3 Connections to Prior Literature 161
5.4 Mechanisms Behind the Empirical Findings 165
5.5 Anomalies and Boundary Conditions 169
5.6 Strengths of the Proposed Approach 173
5.7 Limitations 175
5.8 Practical Implications for Recommender Systems 178
5.9 Lessons Learned 179
5.10 Summary 182

6 Conclusion and Future Work 184
6.1 Summary of Key Findings 184
6.2 Contributions of the Thesis 187
6.3 Research Limitations 189
6.4 Future Work 190

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

6.5 Closing Remarks 197

References **200**



University of Ghana http://ugspace.ug.edu.gh

List of Figures

1.1 Classification system of recommender systems. 4

2.1 Illustration of user-based collaborative filtering 31

2.2 Illustration of content-based filtering 32

2.3 Classification of recommender systems by problem characterisation 36

2.4 Sample paths from a zero-mean Gaussian noise process 55

2.5 Sample path from an Ornstein–Uhlenbeck process 58

2.6 Conceptual view of the action-perturbation channel within the actor–critic learning loop 75

3.1 Effect of action-selection noise distribution on action trajectories 100

3.2 t-DDPG architecture with Student's t-distributed action-selection noise ... 108

4.1 Implementation environment for the enhanced DDPG algorithm 127

4.2 Evaluation and validation workflow for the enhanced DDPG algorithm.... 129

4.3 HR@10 training progression on MovieLens 100K 146

4.4 nDCG@10 training progression on MovieLens 100K 146

4.5 HR@10 training progression on MovieLens 1M..... 148

4.6 nDCG@10 training progression on MovieLens 1M 148

4.7 HR@10 training progression on Jester D4 150

4.8 nDCG@10 training progression on Jester D4 151

University of Ghana http://ugspace.ug.edu.gh

List of Tables

2.1 Comparison of General vs. Sequential Recommender Systems 39

2.2 Comparison of exploration noise strategies in DRL 79

3.1 MovieLens dataset summary 87

3.2 Jester D4 dataset summary 88

4.1 Effect of decay period on recommendation metrics 132

4.2 Effect of maximum noise scale on recommendation metrics 133

4.3 Effect of minimum noise scale on recommendation metrics 134

4.4 Effect of degrees of freedom ν on recommendation metrics 135

4.5 Performance comparison of exploration strategies on MovieLens 100K (10 seeds) 137

4.6 Paired t-test p-values for Student's t-noise versus baselines (10 seeds) 139

4.7 Wilcoxon signed-rank test p-values for Student's t-noise versus baselines (10 seeds) 139

4.8 Effect sizes and bootstrap CIs for Student's t-noise versus baselines 141

4.9 Performance comparison on MovieLens 1M and Jester D4 143

University of Ghana http://ugspace.ug.edu.gh

LIST OF ABBREVIATIONS

AI - Artificial Intelligence

CF - Collaborative Filtering

CN - Content-Based Filtering

CNN - Convolutional Neural Network

DDPG - Deep Deterministic Policy Gradient

DF - Demographic Filtering

DL - Deep Learning

DNN - Deep Neural Network

DQN - Deep Q-Networks

DRL - Deep Reinforcement Learning

GAN - Generative Adversarial Networks

GLN - Graph Neural Network

GRU - Gated Recurrent Units

GPU - Graphics Processing Unit

IR - Information Retrieval

KBRS - Knowledge-Based Recommender System

LDA - Latent Dirichlet Allocation

LSTM - Long Short-Term Memory

MC - Markov Chains

MCTS - Monte Carlo Tree Search

MDP - Markov Decision Processes

MF - Matrix Factorization

ML - Machine Learning

University of Ghana http://ugspace.ug.edu.gh

nDCG - Normalized Discounted Cumulative Gain

NLP - Natural Language Processing

NTM - Neural Turing Machines

OU - Ornstein–Uhlenbeck

PER - Prioritized Experience Replay

PLRS - Personalized Learning Resource System

PPO - Proximal Policy Optimization

RCF - Recurrent Collaborative Filtering

RL - Reinforcement Learning

RLRS - Reinforcement Learning Recommender Systems

RNN - Recurrent Neural Networks

SARSA - State Action Reward State Action

SAS-Rec - Self-Attentive Sequential Recommendation

SRS - Sequential Recommender Systems

SVD - Singular Value Decomposition

SVM - Support Vector Machines

TD - Temporal Difference

TD3 - Twin Delayed Deep Deterministic Policy Gradient

TQC - Truncated Quantile Critics

TRPO - Trust Region Policy Optimization

University of Ghana http://ugspace.ug.edu.gh

CHAPTER 1

INTRODUCTION

1.1 Background

1.1.1 Recommender Systems

Over the past three decades, digital information growth has accelerated, driven by increased adoption of the internet and the vast content generated through everyday online interactions. Social media platforms, e-commerce solutions, and multimedia platforms have substantially increased the quantity of digital content. Users of these platforms generate, upload, and disseminate information including, but not limited to audio, video, images and text (Matthes et al., 2020). The effectiveness of decision-making is significantly impaired by the overwhelming volume of digital information on these platforms, particularly when selecting from a variety of similar options. This challenge is commonly termed the information overload problem (Huang & Jia, 2022).

Existing research has identified two key information processing systems employed to address this issue: recommender systems and information retrieval (Valcarce, 2021). In a typical information retrieval system, users submit queries that explicitly define their information needs (Zamani et al., 2020). The information retrieval system retrieves the most relevant results based on a retrieval algorithm from the index of data stored by the system. The performance of an information retrieval system is typically evaluated using metrics such as precision, which measures the proportion of retrieved information that is relevant, and recall, which measures the proportion of relevant documents that are retrieved. As an example, if the system retrieves 10 documents of which 7 are relevant, the precision is $7/10 = 0.7$. Recall, on the other hand, measures how many of the truly

University of Ghana http://ugspace.ug.edu.gh

relevant documents in the database the system successfully retrieved: if there are 20 relevant documents in total and the system retrieves 7 of them, the recall is $7/20 = 0.35$.

In contrast, recommender systems infer the relevance of certain information to a user based on a profile constructed from their past interactions, without requiring an explicit query from the user. Both aim to deliver relevant information from a very large pool of otherwise unwanted data; however, the approaches they take differ, particularly in how they model users' needs.

In an information retrieval system, user needs are expressed through a query, whereas a recommender system infers user needs, which may facilitate the discovery of new and unexpected items and support serendipitous findings (Ziarani & Ravanmehr, 2021).

Formally, a recommender system predicts the relevance of unseen items to users, often leveraging user data and contextual information. Recommender Systems have become invaluable in helping users of online platforms sift through large volumes of similar options to identify the most relevant choices. They do so by presenting useful items to appropriate users in the correct context. Recommender systems are evaluated based on the accuracy of their predictions of user preferences, their efficacy in ranking items for top-k recommendation scenarios, and the diversity and novelty of the recommendations they provide.

Generally, the core role of a recommender system is to model user preferences through their interactions with the system, using proxies such as likes, purchases, and clicks. Recommender systems estimates user preferences based on interactions with a subset of items (e.g., music, videos, people, locations, services). In recommender systems, preferences are extracted either explicitly (via user ratings) or implicitly (via user activity,

University of Ghana http://ugspace.ug.edu.gh

such as browsing history, clicks, and media consumption).

Recommender systems are widely used on e-commerce platforms to drive sales and enhance customer engagement. Social media networks also leverage recommender systems to foster connected communities. For instance, Instagram, Facebook, and TikTok provide accurate suggestions for people likely to be part of a user's real-life network. In recent years, e-commerce websites and video streaming services have utilised recommender systems to offer personalised suggestions. Other notable applications include movie, travel, and query recommendations (Nilashi et al., 2022; Roy & Dutta, 2022). As effective tools for addressing information overload, recommender systems are considered some of the most valuable AI-based decision support systems, influencing nearly every aspect of daily life (Roy & Dutta, 2022).

Recommender systems are categorised based on diverse taxonomic frameworks. Figure 1.1 provides a summary of these taxonomic approaches, highlighting the contemporary status of their practical applications.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

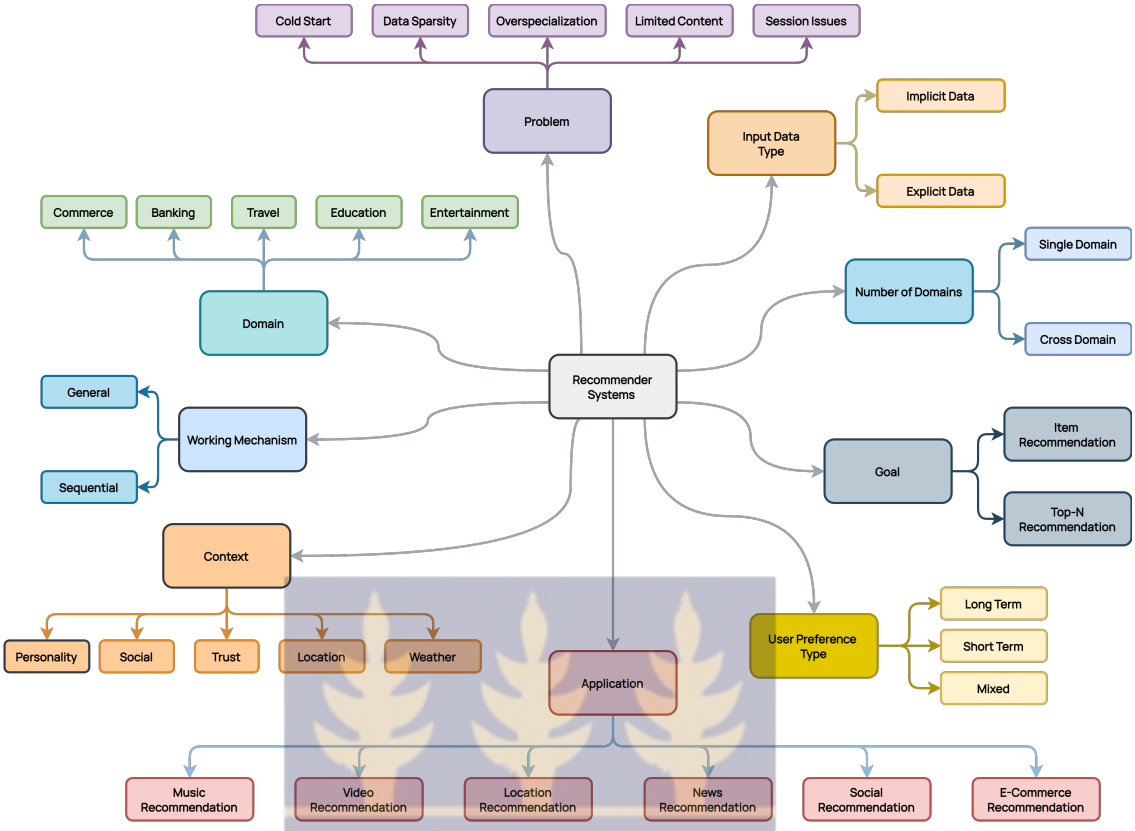


Figure 1.1: Classification system of recommender systems.

Figure 1.1 outlines nine broad classifications of recommender systems, each characterised by specific attributes and tailored to distinct use cases. These include classifications by context (e.g. social-aware, time-aware, trust-aware, location-aware, personality-aware) (Campos et al., 2014; Chakraverty & Mithal, 2018; Dhelim et al., 2020; Ojagh, Malek, & Saeedi, 2020; Ojagh, Malek, Saeedi, & Liang, 2020; Richa & Bedi, 2021), by domain coverage (single-domain versus cross-domain) (Ghafari et al., 2020), and, critically for this thesis, by how user preferences are modelled. The latter dimension distinguishes *general* recommender systems, which capture stable long-term preferences from past interactions, from *sequential* recommender systems, which model dynamic short-term preferences by exploiting interaction order. A detailed treatment of these taxonomies, including the mathematical underpinnings of content-based, collaborative, and hybrid approaches, is deferred to Chapter 2; the present chapter restricts itself to the general/sequential

University of Ghana http://ugspace.ug.edu.gh

distinction because it is this dimension that motivates the use of reinforcement learning in the present work.

1.1.2 General Recommender Systems in Perspective

In the past two decades, general recommender systems have been widely explored, aiming to uncover users' general preferences by extracting insights from their interactions with items, either explicitly (e.g. likes, comments, ratings) or implicitly (e.g. browsing history, clicks, purchasing behaviours) (Afsar et al., 2022; Cheng et al., 2021). Determining a user's overall preferences is a crucial function of general recommender systems, as it directly influences their present decision-making behaviours (Hazrati & Ricci, 2024). Examining the general preferences of an active user is crucial, as these preferences tend to be relatively fixed or evolve gradually over time. Once these preferences have been identified, the process of providing recommendations becomes more straightforward. General recommender systems are designed to leverage this stable preference structure.

The utility matrix, also known as the user-item matrix, serves as the input for such systems, with Matrix Factorisation (MF) being the predominant solution method (Altulyan, 2020). The recommendation task is typically framed as a matrix completion problem, given the matrix's extreme sparsity caused by numerous missing entries. These gaps exist because users generally rate only a small portion of available items. The primary task of the system is to predict these missing values for user-item pairs. General recommender systems have been extensively studied, with numerous solutions proposed to address their challenges. However, the inherent volatility of human behaviour and taste necessitates continuous improvement through data augmentation (Ding et al., 2020), contextual factors (Adomavicius & Tuzhilin, 2010), and personality considerations (Asabere et al., 2017).

University of Ghana http://ugspace.ug.edu.gh

1.1.3 Sequential Recommender Systems in Perspective

General Recommender Systems and Sequential Recommender Systems constitute distinct categories, differentiated by their operational mechanisms. The primary aim of Sequential Recommender Systems is to identify and model users' sequential behaviours, which offers the advantage of capturing short-term preferences, rendering them increasingly popular. In contrast to general recommenders, Sequential Recommender Systems rely on chains of user-item interactions, organised as sessions of past user activity, such as items purchased in a single transaction (Qiu et al., 2020). These interactions are logged anonymously, with the user's identity being unnecessary but useful in some cases. For instance, e-commerce platforms allow users to add items to carts or a Wishlist without identifying themselves. Additionally, extra data like genre or cast information in movie recommendations can be used to enhance recommendations.

The sequential recommender system processes session data along with this additional information to generate an ordered list of recommended items. In the literature, two main classes of solutions for sequential recommendations are model-based and model-free solutions. The model-free techniques, which emerged in the late 1990s and early 2000s, rely on established data mining methods such as Bayesian networks, dimensionality reduction, and pattern mining to generate recommendations (Phan et al., 2018). While model-free methods are easier to build and implement, model-based approaches dominate the field due to their ability to leverage machine learning algorithms to recognise complex data patterns efficiently. Modern model-based techniques primarily utilise Markov Decision Processes (MDPs) with artificial neural networks as function approximators (Tao et al., 2022).

University of Ghana http://ugspace.ug.edu.gh

Chapter 2 will provide a detailed review of Deep Reinforcement Learning Techniques as applied to sequential recommendation.

1.1.4 Unified Recommender Systems in Perspective

Both sequential and general recommender systems have their respective strengths and weaknesses. General recommenders excel at modelling long-term user preferences but fail to capture short-term tastes, a major limitation. Conversely, sequential recommenders focus on short-term preferences, often neglecting long-term behaviour. To address both types of preferences, unified approaches combine the advantages of both paradigms. In such models, one paradigm serves as the base, with elements of the other incorporated to enhance performance. For example, Recurrent Collaborative Filtering (RCF), proposed by Dong et al. (2018), integrates Matrix Factorisation and Recurrent Neural Networks (RNNs). Other unified models use variations like Gated Recurrent Units (GRUs) Hidasi and Tikk (2016), Long Short-Term Memory (LSTM) Niu et al. (2020), and Graph Neural Networks (GNNs) (Głogowska et al., 2024) to address challenges in handling long sequences. Further advanced models combine RNNs with CNNs, mixture models (Tang & Wang, 2018), memory networks (Becattini & Uricchio, 2022), and attention mechanisms (Zhong et al., 2020). While considering both short- and long-term preferences is crucial, more research is needed to fully explore their utility and limitations.

1.2 Reinforcement Learning for Recommender Systems

Reinforcement Learning (RL), one of the three main paradigms of Machine Learning (ML), alongside supervised and unsupervised learning, has attracted considerable scholarly interest in the Artificial Intelligence (AI) space due to its capacity for sequential decision-making. Unlike supervised learning, which involves learning from labelled datasets, or unsupervised learning, which seeks to find hidden patterns in data, RL focuses on enabling

University of Ghana http://ugspace.ug.edu.gh

autonomous agents to interact with unknown environments in a trial-and-error fashion. The agent's objective in RL is to learn how to take actions in these environments to maximise a cumulative reward signal. This approach formalises decision-making as a process in which the agent learns by receiving feedback from the environment, adapting its behaviour over time to improve future outcomes. The result of these interactions is a policy, which is a model that predicts the best actions an agent should take in any given state of the environment (Akanksha et al., 2021).

Over the past decade, RL has become a prominent component of contemporary AI research, with substantial deployments and methodological advances reported in recent surveys spanning robotics, autonomous driving, and large-scale decision systems (Lin et al., 2024; Shakya et al., 2023; Sumiea et al., 2024). Examples of RL applications include robotics, where robots must learn to perform tasks like manipulation and navigation in unpredictable environments, and autonomous vehicles, where self-driving cars must make real-time decisions to ensure safety and efficiency (Pérez-Gil et al., 2022; T. Zhang & Mo, 2021).

One of the most important advancements in RL is the integration of deep learning, leading to Deep Reinforcement Learning (DRL). DRL leverages the representational power of artificial neural networks to handle high-dimensional inputs, such as images or complex data structures, which traditional RL methods struggle to process. Deep learning models are particularly useful in approximating value functions or policies in RL, allowing agents to learn complex patterns in data that are beyond the reach of simpler methods. The combination of deep learning with RL has enabled agents to solve tasks that involve intricate decision-making processes with large state and action spaces (Ashwin & Naveen Raj, 2023). In the field of recommender systems, the application of RL has become increasingly prevalent as the nature of sequential recommendation can be modelled as a

University of Ghana http://ugspace.ug.edu.gh

sequential decision problem.

Traditionally, recommendation tasks have been approached through supervised learning methods, which predict user preferences based on past interactions. However, as recommender systems evolve, there is a growing recognition that recommendations are not static, one-off predictions but rather part of a continuous interaction process. In this context, RL is well-suited for handling dynamic environments where user preferences change over time. By framing sequential recommendation as a Markov Decision Process (MDP), the system can adapt its recommendations based on sequential interactions with users, aiming to maximise long-term engagement or satisfaction (Tirinzoni et al., 2020).

A Markov Decision Process formalises the recommendation task by representing the environment as a series of states (e.g. a user's interaction history), actions (e.g. recommending an item), and rewards (e.g. whether the user engages with the recommended content). The primary objective of the RL-based recommender is to learn a policy that maximises the expected return, which could be user satisfaction, engagement, or any other performance metric relevant to the platform. By treating recommendations as a sequential decision problem, RL allows the system to adjust its recommendations dynamically, learning from user feedback over time to refine its strategy.

The potential of RL to enhance recommendation quality was first demonstrated by Joachims et al. (1997) with the WebWatcher system, one of the earliest RL-based recommenders. WebWatcher used RL to guide users through websites by recommending pages based on user behaviour. This early success paved the way for more advanced RL techniques to be applied to recommendation tasks, leading to substantial improvements in user engagement and satisfaction. As the field of recommender systems evolved, RL

University of Ghana http://ugspace.ug.edu.gh

methods have grown increasingly sophisticated, moving from simple tabular methods to more advanced approaches that integrate deep learning (Ivanov & D'yakonov, 2019). Traditional RL approaches in recommender systems often relied on tabular methods, where the state–action space is represented explicitly in a table. However, such methods are impractical for large-scale recommendation tasks, where the number of possible states (user histories) and actions (items to recommend) can be enormous.

To address this, modern recommender systems often employ function approximation techniques using artificial neural networks. This is where Deep Reinforcement Learning (DRL) has made a considerable impact. By combining the pattern-recognition capabilities of deep learning with the sequential decision-making framework of RL, DRL-based recommenders can handle the vast state and action spaces found in real-world applications such as e-commerce, video streaming, and social media (Afsar et al., 2022; Jia & Wang, 2020).

The deployment of large-scale deep recommender systems by major platforms is well documented. Steck et al. (2021) describe Netflix's transition from collaborative filtering to deep learning-based personalisation, and Gomez-Uribe and Hunt (2016) report that the Netflix recommender system drives a substantial portion of viewing decisions on the platform. Similar deep-learning-based deployments are documented for YouTube (Covington et al., 2016), and reinforcement-learning-based deployments for music recommendation at Spotify (Tomasi et al., 2023). Recent industry-facing research has continued in this direction, with sequential and reinforcement-learning-based methods increasingly explored as means of optimising long-term engagement rather than single-step prediction (Lin et al., 2024; Sumiea et al., 2024). In such systems, the recommender not only predicts what a user might like based on past behaviour but also continuously updates

University of Ghana http://ugspace.ug.edu.gh

its policy as the user interacts with the platform, ensuring that the system remains relevant even as preferences evolve. Modern RL solutions for recommendation include a range of methods, from traditional tabular approaches to algorithms that integrate deep learning and neural networks.

One popular technique involves Markov Decision Processes (MDPs), where the system models user interactions as sequential decisions to optimise future recommendations. More advanced methods, such as Deep Q-Networks (DQNs), allow systems to handle high-dimensional data by using artificial neural networks as function approximators. These methods are particularly useful in applications where user interactions generate large datasets, such as streaming platforms or social media (Tao et al., 2022). In addition to these value-based techniques, policy-gradient methods have gained traction in the recommendation space. These methods directly optimise the policy by using gradient-based optimisation techniques, which have proven effective in environments where the feedback, modelled as a reward signal, is sparse or delayed. Furthermore, Actor–Critic methods provide a useful framework for recommendation tasks formulated as sequential decision-making problems. By combining policy optimisation with value estimation, they allow the recommender agent to learn which items to recommend while also estimating the long-term utility of its actions from user feedback.

Deep Reinforcement Learning has significantly advanced the field of recommender systems by providing a powerful framework for modelling sequential decision problems. By treating recommendations as a continuous process of interactions between users and the system, RL-based recommenders can adapt to changing preferences and optimise long-term user satisfaction. As RL techniques continue to evolve, they are likely to play an even more central role in the future of personalised recommendations, improving the

University of Ghana http://ugspace.ug.edu.gh

quality and relevance of suggestions across a wide range of domains, from e-commerce to entertainment. However, further research is needed to explore the full potential of RL in this area and to address challenges related to scalability, computational efficiency, and the need for large amounts of interaction data.

1.3 Research Motivations

Deciding on an item from a list of nearly identical options is a challenge that permeates many aspects of daily life. Whether choosing entertainment options like movies or music, or selecting elective courses each semester, individuals are frequently confronted with a plethora of choices. In the digital age, where many of our daily activities are mediated by technology, this issue becomes even more pronounced. The proliferation of digital platforms has generated vast amounts of content, leading to an overwhelming abundance of options in domains like e-commerce, media streaming, and online education. As a result, the need for systems that assist in decision-making by filtering and recommending only relevant items has become essential. Recommender systems have emerged as critical tools that help users navigate these vast datasets by suggesting items that align with their preferences, thus streamlining the decision-making process.

The importance of recommender systems in aiding the decision-making process when interacting with digital technologies cannot be overstated. These systems are embedded in various services, including online shopping platforms, social media networks, streaming services, and educational tools. In particular, sequential recommender systems, which are designed to make personalised recommendations based on users' sequential interactions, have gained widespread attention. Unlike traditional recommender systems that may rely on static user profiles, sequential recommender systems take into account the evolving nature of user preferences, making them more responsive to recent user behaviours. In

University of Ghana http://ugspace.ug.edu.gh

technical terms, sequential recommendation is often modelled as a reinforcement learning (RL) problem, and is formalised as a Markov Decision Process (MDP). MDPs are widely regarded as the standard approach for modelling sequential decision problems because they represent the dynamics of the decision-making environment, where the agent (in this case, the recommender system) takes actions to maximise cumulative rewards over time. For instance, in a movie recommendation scenario, the system recommends movies to a user based on their viewing history, with the goal of maximising user engagement or satisfaction. Each movie recommendation can be viewed as an action taken by the system, and the user's feedback, such as clicking on or watching the movie, serves as the reward signal.

One of the most well-known algorithms for solving MDPs is Q-learning, which operates by constructing a Q-table to store the quality (or expected future reward) of each state–action pair. Q-learning is effective for solving MDPs when the state and action spaces are relatively small. However, when applied to larger MDPs, such as those found in real-world recommender systems, where the number of possible states (user histories) and actions (items to recommend) can be vast, Q-learning becomes computationally expensive. The size and complexity of the state–action space can make it infeasible to maintain a Q-table for all possible combinations, as the computational and memory requirements grow exponentially.

To address these computational challenges, researchers have proposed leveraging deep learning techniques to approximate the Q-function, which represents the expected rewards for each state–action pair in the Q-learning algorithm. The integration of deep learning with reinforcement learning has led to the development of Deep Reinforcement Learning (DRL), which has proven to be highly effective in dealing with large, high-dimensional

University of Ghana http://ugspace.ug.edu.gh

state and action spaces. DRL allows recommender systems to scale effectively, making it possible to handle the vast amounts of data and complex interactions found in large digital platforms. The use of DRL in recommender systems is particularly promising because it enables models to learn from sequential user interactions, capture complex patterns in user behaviour, and adapt recommendation policies as user preferences evolve over time. Beyond recommender systems, DRL has found applications in a wide range of control problems. For example, DRL algorithms have been used to solve locomotion tasks in robotics, enabling robots to learn how to navigate complex environments (X. Zhu et al., 2022). Similarly, DRL has been applied to autonomous vehicles, where self-driving cars learn to make safe and efficient navigation decisions in real-time (Pérez-Gil et al., 2022). In addition, DRL has been employed in gaming, where agents trained with DRL techniques have outperformed human players in games such as Go, Dota 2, and other video games (Lample & Chaplot, 2017). These applications demonstrate the versatility and power of DRL in handling complex, dynamic environments where traditional algorithms fall short. However, despite its successes, one of the major challenges with DRL algorithms is the issue of *overestimation bias*. This bias occurs when the estimated Q-values, the expected return for a given state–action pair, are systematically higher than the true Q-values. Overestimation bias can lead the RL agent to make poor decisions, as the policy (the model that predicts optimal actions) exploits these overestimations, resulting in suboptimal performance. In the context of recommender systems, this could mean suggesting items that appear to be high in reward but ultimately fail to satisfy the user, leading to diminished engagement or satisfaction.

Overestimation bias arises in part from the Temporal Difference (TD) learning process used in Q-Learning and related reinforcement learning algorithms. In TD learning, Q-

University of Ghana http://ugspace.ug.edu.gh

values are updated toward a target that combines the immediate reward with an estimate of future return. Because Q-Learning uses a maximisation operation over estimated future Q-values, errors in these estimates can be amplified, causing the learned values to become overly optimistic. When overestimation occurs, these errors propagate through the learning process, leading to further inaccuracies in future predictions. This can degrade the performance of the recommender system, as it becomes increasingly biased towards actions that do not actually maximise user satisfaction.

The primary aim of this thesis is to address overestimation bias in DRL-based recommender systems by introducing a parameterised heavy-tailed exploration framework. To achieve this, the study proposes and implements t-DDPG, a novel extension of DDPG that incorporates Student's t-distributed action perturbation to improve exploration and reduce overly optimistic value estimates. The algorithm is evaluated on movie recommendation tasks using the MovieLens datasets and on joke recommendation tasks using the Jester dataset. Both datasets are widely adopted benchmarks in recommender systems research, providing extensive user–item interaction data based on movie and joke ratings, respectively. Through these experiments, the thesis aims to demonstrate that t-DDPG can improve recommendation quality while mitigating overestimation bias in DRL-based recommender systems.

1.4 Problem Statement

In reinforcement learning (RL), the accurate estimation of the Q-function, which quantifies the expected cumulative return of executing an action in a given state and subsequently following a policy, is fundamental to solving Markov Decision Processes (MDPs). The effectiveness of value-based and actor-critic algorithms depends directly on the accuracy of these Q-value approximations. However, a central challenge in this estimation process

University of Ghana http://ugspace.ug.edu.gh

is the phenomenon of *overestimation bias*, which arises from sample-based maximisation in Temporal Difference (TD) learning (Lillicrap et al., 2015; Sumiea et al., 2024; Watkins & Dayan, 1992).

Overestimation bias can cause learning agents to systematically inflate action-value estimates, which may lead to premature convergence to suboptimal policies. This problem is particularly pronounced in modern Deep Reinforcement Learning (DRL) settings, where artificial neural networks are used as function approximators. The approximation errors, variance, and nonlinearities inherent in these models can amplify value estimation errors and, in severe cases, lead to learning instability or failure(Lee & Powell, 2019; Sumiea et al., 2024; Van Hasselt et al., 2016). The challenge affects a wide range of RL algorithms, including Q-learning, SARSA, and Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015; Sumiea et al., 2024).

DDPG, in particular, though widely adopted for continuous control tasks, is especially vulnerable to overestimation bias due to its deterministic policy formulation and reliance on a single critic network during policy updates. The joint effect of deterministic action selection and positively biased Q-value estimation often causes DDPG agents to prematurely exploit high-reward actions, resulting in policy degradation, limited generalisation, and diminished behavioural diversity (Fujimoto et al., 2023; Sumiea et al., 2024). These issues are further magnified in the domain of recommender systems, which are characterised by large, high-dimensional action spaces, delayed and sparse rewards, and significant user preference heterogeneity (C. Gao et al., 2023; Lin et al., 2024). In such environments, inadequate exploration and biased value estimation significantly compromise the diversity, novelty, and robustness of recommendations.

University of Ghana http://ugspace.ug.edu.gh

Existing mitigations for overestimation bias in DDPG concentrate primarily on the *critic* side of the actor–critic architecture. The Twin Delayed DDPG (TD3) algorithm of Fujimoto et al. (2018) introduces a pair of critics and uses the minimum of their estimates to construct the temporal-difference target, effectively replacing positive bias with a controlled negative bias. Soft Actor-Critic (SAC) (Haarnoja et al., 2018) replaces deterministic policies with stochastic ones and adds an entropy-regularisation term that implicitly encourages broader exploration.

While these approaches have been highly influential in continuous control benchmarks, they share two limitations relevant to recommender systems. First, they address the symptom (biased value estimates) rather than one of its underlying drivers (insufficient exploration around the deterministic policy output), with recent work showing that the twin-critic approach itself introduces an opposite *underestimation* bias whose consequences for sparse-reward, high-dimensional environments remain incompletely understood (Sumiea et al., 2024). Second, the exploration mechanism in DDPG itself, the addition of stochastic noise to the action produced by the deterministic policy, has received markedly less attention as a principled lever for mitigating overestimation bias, with the dominant choices remaining the light-tailed Gaussian and Ornstein–Uhlenbeck (OU) processes (Hollenstein et al., 2022; Nobakht & Liu, 2022). The action-perturbation channel therefore remains underexplored as a means of jointly improving exploration and reducing bias in deterministic policy gradient methods applied to sequential recommendation.

1.5 Research Gap

The preceding discussion of the problem identifies a specific gap in the current literature on exploration for deterministic policy gradient methods in recommender systems. This section makes that gap explicit before stating the research aims and questions in the

University of Ghana http://ugspace.ug.edu.gh

sections that follow.

Critic-side mitigations leave actor-side exploration unaddressed. As noted, the dominant strategies for mitigating overestimation bias in DDPG operate on the critic side of the actor–critic architecture. TD3 (Fujimoto et al., 2018) introduces a clipped double-Q estimator and delayed policy updates, SAC (Haarnoja et al., 2018) introduces stochastic policies regularised by entropy, and more recent work such as TD7 (Fujimoto et al., 2023) couples deterministic-policy learning with state–action representation learning to further mitigate value-estimation errors. While effective in standard continuous control benchmarks, none of these approaches directly examine the role of the action-perturbation noise process itself, which remains predominantly Gaussian or OU (Hollenstein et al., 2022; Nobakht & Liu, 2022; Sumiea et al., 2024). In sparse-reward, high-dimensional environments such as those encountered in recommendation, the actor-side exploration channel is plausibly a contributing factor to the bias, yet it has not been treated systematically.

Heavy-tailed exploration is theoretically motivated but empirically underexplored in recommendation. A growing body of work in continuous-control reinforcement learning has demonstrated, from independent theoretical and empirical perspectives, that heavy-tailed policy parameterisations and noise processes can support persistent exploration, escape from spurious local maxima, and improved coverage of high-dimensional action spaces. On the empirical side, Garg et al. (2021) document that policy-gradient updates in proximal policy optimisation (PPO) exhibit pronounced heavy-tailedness in their gradients, motivating the use of robust estimators in place of standard clipping heuristics. On the theoretical side, Bedi et al. (2024) provide convergence-rate and metastability analyses for heavy-tailed policy search, showing that distributions with

University of Ghana http://ugspace.ug.edu.gh

heavier tails can converge to wider peaks of the value landscape, and Cayci and Eryilmaz (2023) establish finite-time sample-complexity bounds for natural actor–critic methods under heavy-tailed reward distributions, with dynamic gradient clipping as the stabilising mechanism. Empirically, Chakraborty et al. (2023) demonstrate that heavy-tailed policy gradients improve learning in sparse-reward continuous control, and Kobayashi (2019) show that Student's t-policies enable Lévy-flight-style exploration that escapes local optima in robot control.

The most directly relevant recent result is that of L. Zhu et al. (2025), who show within the q-exponential family that the Student's t-distribution as a policy is more stable than the Gaussian across both online and offline benchmarks, providing independent evidence that heavy-tailed policies, with the Student's t as a principled instance, are an active research direction. To date, however, no systematic study has examined heavy-tailed *action perturbation* in DDPG-based recommender systems, where sparsity, high-dimensional discrete-effective action spaces, and user heterogeneity make the limitations of light-tailed noise especially salient.

Exploration noise has not been treated as a parameterised family. The exploration strategies typically adopted in DDPG (Gaussian noise, OU noise, parameter noise, no noise) are usually compared as discrete alternatives rather than as instances of a single parameterised family. Recent work on coloured noise for off-policy DRL has begun to challenge this discrete-alternatives framing by showing that the temporal correlation structure of action noise is itself a continuous design dimension that materially affects performance (Eberhard et al., 2023). However, no comparable parameterised view has been articulated for the *distributional* structure of action-perturbation noise, in particular, no unified treatment exists in which conventional Gaussian and OU processes and heavy-

University of Ghana http://ugspace.ug.edu.gh

tailed alternatives can be analysed and compared on equal footing within recommender systems.

Closing the gap. This thesis closes these gaps in three ways. First, it formulates a parameterised view of exploration noise in deterministic policy gradient methods, within which conventional and heavy-tailed processes can be expressed and compared as instances of a single family. Second, it instantiates the framework with a Student's t-distributed noise process, equipped with scale annealing and value clipping, yielding the algorithm t-DDPG. Third, it provides the first systematic empirical evaluation of heavy-tailed action perturbation in DDPG-based recommender systems, comparing t-DDPG against vanilla DDPG, Gaussian-noise DDPG, and OU-noise DDPG on three benchmark datasets.

1.6 Research Aims and Objectives

1.6.1 General Objective

The overarching aim of this research is to improve the effectiveness, stability, and generalisation of deterministic policy gradient methods in recommender systems by addressing overestimation bias and enhancing exploration through a parameterised heavy-tailed stochastic framework, with Deep Deterministic Policy Gradient (DDPG) as the demonstrative algorithm.

1.6.2 Specific Objectives

The general objective is decomposed into the following three specific, measurable objectives:

1. **Objective 1.** To evaluate the role of exploration noise in the DDPG algorithm and characterise the limitations of commonly used noise processes (Gaussian and Ornstein–Uhlenbeck) in sequential recommendation environments, measured

University of Ghana http://ugspace.ug.edu.gh

through hit rate (HR@10) and normalised discounted cumulative gain (nDCG@10) on benchmark datasets.

2. **Objective 2.** To design and implement a novel parameterised exploration framework based on heavy-tailed stochastic processes for deterministic policy gradient methods, specifying the noise model, scale annealing schedule, and clipping mechanism, and to instantiate this framework using the Student's t-distribution within DDPG (denoted t-DDPG).
3. **Objective 3.** To empirically evaluate the proposed t-DDPG against vanilla DDPG, Gaussian-noise DDPG, and OU-noise DDPG on three benchmark recommendation datasets (MovieLens 100K, MovieLens 1M, and Jester D4), using HR@10 and nDCG@10 as performance metrics, and to characterise the contribution of the framework's components (annealing and clipping) through ablation analysis.

1.7 Research Questions and Hypotheses

This thesis addresses key challenges in applying Deep Reinforcement Learning to recommender systems, with a particular focus on overestimation bias in Q-value estimation and the role of stochastic exploration in actor-critic algorithms in mitigating it. The research is guided by the following four research questions, with formal hypotheses associated with the empirical questions (RQ2–RQ4):

RQ1. *How does overestimation bias affect the performance of recommender systems based on deterministic policy gradient methods, demonstrated through DDPG?*

This question is investigative and seeks to characterise the manifestation of overestimation bias in DDPG-based recommender systems, including its effect on policy convergence, recommendation quality, and behavioural diversity.

University of Ghana http://ugspace.ug.edu.gh

RQ2. *What role do exploration strategies play in improving the performance of deterministic policy gradient methods, and what limitations exist in commonly used noise processes such as Gaussian and Ornstein–Uhlenbeck noise, demonstrated through DDPG?*

Hypothesis H1. Light-tailed exploration noise processes (Gaussian and OU) constrain DDPG's exploration to a narrow region of the action space, leading to measurably lower HR@10 and nDCG@10 than would be obtained with broader exploration in sequential recommendation environments.

RQ3. *Can heavy-tailed stochastic processes provide advantages for exploration in deterministic policy gradient methods, particularly in terms of stability, robustness, and convergence, demonstrated through DDPG?*

Hypothesis H2. A bounded heavy-tailed exploration process based on the Student's t-distribution, equipped with scale annealing and value clipping, produces more effective exploration than Gaussian and OU baselines, as evidenced by improved HR@10 and nDCG@10 and more stable convergence behaviour.

RQ4. *How does the proposed parameterised heavy-tailed exploration framework, instantiated as t-DDPG, compare to conventional exploration strategies in terms of recommendation performance and generalisation across benchmark datasets?*

Hypothesis H3. The proposed t-DDPG outperforms vanilla DDPG, Gaussian-noise DDPG, and OU-noise DDPG on HR@10 and nDCG@10, and this advantage generalises across MovieLens 100K, MovieLens 1M, and Jester D4 datasets.

University of Ghana http://ugspace.ug.edu.gh

1.8 Significance of the Study

The significance of this study lies in four contributions, each linked to a gap identified in Section 1.5.

Theoretical contribution: a unified view of action-perturbation exploration. The study introduces a parameterised view of action-space exploration in deterministic policy gradient methods, within which Gaussian noise, OU noise, and heavy-tailed alternatives are special cases of a single family of stochastic processes. This view is significant because it enables principled comparison of exploration strategies that have previously been studied in isolation, and because it positions the choice of noise distribution as a design lever to be analysed rather than an implementation detail to be inherited.

Methodological contribution: t-DDPG. The study develops t-DDPG, a novel algorithm that instantiates the parameterised framework using a bounded, scale-annealed Student's t-distributed perturbation process. To the best of our knowledge, this constitutes the first systematic application of a parameterised heavy-tailed action-perturbation process within a deterministic policy gradient algorithm in the recommender systems setting. The methodological contribution is significant because it identifies the action-perturbation channel, rather than the critic, as a tractable point of intervention for jointly improving exploration and reducing overestimation bias, and it demonstrates that this intervention can be made compatible with the standard DDPG architecture without modifications to the critic or the reward formulation.

Empirical contribution: cross-dataset evidence with component-wise ablation. The study provides comparative evidence across three benchmark recommendation datasets (MovieLens 100K, MovieLens 1M, and Jester D4), with component-wise ablation

University of Ghana http://ugspace.ug.edu.gh

isolating the contributions of scale annealing and value clipping. This is significant because it offers a controlled empirical basis for the claim that heavy-tailed exploration improves ranking quality over conventional light-tailed alternatives in DDPG, and it characterises the conditions under which the framework's design choices are necessary for that improvement.

Practical contribution: a deployable exploration mechanism. The proposed exploration mechanism is a drop-in replacement for the noise process in DDPG, requiring no architectural modifications to the actor-critic networks or to the reward formulation. This is significant for practitioners deploying DDPG-based recommender systems, since the proposed approach can be adopted without retraining or restructuring existing models, and it provides a baseline against which more elaborate exploration strategies (e.g. adaptive heavy-tailed processes or context-conditioned noise schedules) can be evaluated in future work.

1.9 Definition of Key Terminologies

The following terms are used throughout the thesis with the meanings given here.

Markov Decision Process (MDP). A mathematical model used to formalise sequential decision-making under uncertainty. An MDP consists of a 5-tuple comprising a set of states, a set of actions, a transition-probability function, a reward function, and a discount factor.

Artificial Neural Network (ANN). A computational model inspired by the structure and function of the mammalian brain. ANNs model complex relationships between inputs and outputs and learn to recognise patterns and make predictions from training data.

University of Ghana http://ugspace.ug.edu.gh

Reinforcement Learning Agent. An artificial intelligence agent that learns by interacting with its environment, using feedback from its actions to improve its behaviour over time.

Recommender System. A software system that uses algorithms to predict user preferences and recommend items or content that the user is likely to find relevant.

Actor–Critic Method. A class of reinforcement learning methods that combines two components: an *actor*, responsible for selecting actions in a given state; and a *critic*, responsible for evaluating the actor's actions and providing feedback used to update the actor's policy.

Actor. In an actor–critic algorithm, the component responsible for selecting actions in a given state.

Critic. In an actor–critic algorithm, the component responsible for evaluating the actor's actions and providing the feedback signal that guides the actor's updates.

Student's t-Distributed Noise. A heavy-tailed stochastic perturbation drawn from the Student's t-distribution with ν degrees of freedom, used in this thesis as the action-perturbation noise process within DDPG. The Student's t-distribution generalises the Gaussian (recovered as $\nu \to \infty$) by introducing heavier tails for finite ν, increasing the probability of large but rare exploratory actions.

t-DDPG. The proposed algorithm in this thesis: a DDPG variant in which the action-perturbation noise process is a bounded, scale-annealed Student's t-distributed process, instantiating the parameterised heavy-tailed exploration framework introduced in Chapter 3.

University of Ghana http://ugspace.ug.edu.gh

1.10 Outline of the Thesis

The thesis is organised into six chapters. Chapter 1 provides background information, outlines the research motivation, problem statement, research gap, aims and objectives, research questions and hypotheses, and the significance of the study.

Chapter 2 offers a comprehensive review of traditional and contemporary recommendation algorithms, including collaborative filtering, content-based methods, hybrid models, and deep reinforcement learning approaches. The chapter critically examines exploration challenges in deep reinforcement learning, with a focus on the limitations of conventional light-tailed noise processes, and identifies the gap that the proposed parameterised heavy-tailed exploration framework is designed to close.

Chapter 3 outlines the research methodology, including the theoretical framework, problem formulation, the baseline DDPG algorithm, and the proposed parameterised heavy-tailed exploration framework instantiated as t-DDPG. The chapter also details the datasets, experimental design, evaluation metrics, and reproducibility procedures.

Chapter 4 presents the technical setup, hardware and software configurations, and experimental parameters. It includes comprehensive reporting of the comparative evaluation against baseline DDPG variants, ablation analyses isolating the contributions of annealing and clipping, and training-dynamics analyses.

Chapter 5 critically analyses and interprets the experimental findings, discussing their implications for reinforcement learning theory and recommender system practice. It contextualises the results within the broader literature and identifies practical considerations and limitations.

Chapter 6 synthesises the key findings, articulates the contributions of the thesis, and

University of Ghana http://ugspace.ug.edu.gh

outlines directions for future research, including the extension of the parameterised heavy-tailed framework to other actor-critic algorithms (e.g., TD3, SAC), sensitivity analysis on the degree-of-freedom parameter, and online evaluation in real-world recommender environments.



University of Ghana http://ugspace.ug.edu.gh

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Decision-making is an inherent part of human life, as individuals are constantly required to choose from multiple available options. In contemporary society, many of these decisions are increasingly mediated by digital services accessed through the internet. As internet-based applications have become deeply integrated into everyday life, users are frequently required to make choices relating to food, music, movies, books, products, and other forms of digital content. However, the rapid growth of data and the abundance of similar alternatives have made decision-making increasingly complex. In such environments, identifying relevant and desirable options can become overwhelming for users. Recommender systems address this challenge by using algorithmic techniques to filter large catalogues of items and provide personalised suggestions that support more effective decision-making (Baltrunas et al., 2011). Recommender systems are designed to address the recommendation problem, which involves estimating a user's preference for unseen items based on historical data, gathered through explicit ratings or implicit browsing behaviour (Konstan et al., 1997).

Additionally, advances in machine learning, particularly deep reinforcement learning (DRL), have introduced new frameworks for addressing sequential decision-making problems, including sequential recommendation. The intersection of DRL and recommender systems has yielded promising results; however, several limitations remain, especially with respect to exploration efficiency during training.

This chapter presents a comprehensive review of the state of the art in recommender

University of Ghana http://ugspace.ug.edu.gh

systems and deep reinforcement learning methodologies as they relate to sequential recommender systems. The review is structured to provide a systematic understanding of the foundational concepts, major algorithmic developments, and persistent challenges in the field. It begins by examining the evolution of general and sequential recommender systems, highlighting the modelling paradigms used to capture both long-term and short-term user preferences. The chapter then explores the application of DRL techniques to recommendation tasks, with particular emphasis on actor–critic architectures such as Deep Deterministic Policy Gradient (DDPG). Finally, the review examines a critical challenge in DRL-based recommendation: the design of effective exploration strategies. In this regard, it discusses prior work on stochastic noise processes used to regularise action selection during training.

By identifying and synthesising these contributions, this chapter establishes the contextual foundation for the research presented in this thesis. Specifically, it highlights the research gap addressed by this study: the limited investigation of how the statistical properties of Student's t-distributed noise can be leveraged to enhance exploration in DDPG-based sequential recommender systems. By clarifying the current state of knowledge and the limitations that remain, the chapter provides a clear rationale for the methodological contributions proposed in this work.

2.2 Recommender Systems Overview

The widespread use of the internet for everyday activities and the rapid growth of digital services have resulted in an overwhelming volume of information available to users. Consequently, individuals often face difficulties identifying relevant items from large pools of alternatives. Recommender systems are specialised software tools designed to address this problem by helping users discover items that align with their interests and

University of Ghana http://ugspace.ug.edu.gh

preferences. They achieve this by analysing user preferences either explicitly, through ratings, or implicitly, through behavioural signals such as browsing patterns, purchase history, and viewing activity (Khusro et al., 2016). In today's data-driven environment, recommender systems play an essential role in filtering and personalising information across various domains, including e-commerce, video streaming, music platforms, and social media (Shah et al., 2017).

The objectives of recommender systems serve both users and service providers. For users, recommender systems simplify the search process by reducing exposure to irrelevant choices and suggesting items that closely match their interests. This improves the overall user experience by making digital interactions more efficient, personalised, and satisfying. For businesses, recommender systems support increased customer engagement, satisfaction, retention, and revenue generation. As a result, well-designed recommender systems are critical to both user loyalty and business performance (Aggarwal & Aggarwal, 2016). Common metrics used to evaluate recommendation quality include accuracy, novelty, and diversity (Alhijawi et al., 2022; Schröder et al., 2011).

Recommender systems are commonly categorised into three main types: collaborative filtering, content-based filtering, and hybrid approaches (Ekstrand et al., 2011).

Collaborative filtering identifies preference patterns by comparing the behaviours or preferences of users. It is based on the assumption that user preferences exhibit meaningful correlations; without such correlations, it would be difficult to make informed predictions about users' likely interests (Pennock et al., 2000). Although this assumption is central to collaborative filtering and often holds in practice, the strength and direction of preference correlations may change over time as users' interests evolve. Therefore, the idea that users

University of Ghana http://ugspace.ug.edu.gh

who shared similar preferences in the past will necessarily continue to do so in the future should be treated with caution. Rather than being viewed as a fixed relationship, user preference similarity may be better understood as dynamic and probabilistic.

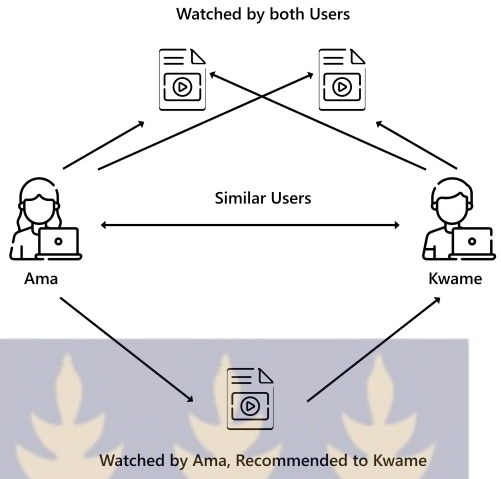


Figure 2.1: Illustration of user-based collaborative filtering, where overlap in Kwame and Ama's viewing histories is used to identify candidate recommendations.

Figure 2.1 depicts a case of collaborative filtering where similar users Kwame and Ama are identified based on common items they have watched, which enables cross-user recommendations.

Content-based filtering recommends items based on their characteristics and the user's profile. For example, if a user watches many action movies, the system may suggest other action films with similar features (Pang et al., 2019).

University of Ghana http://ugspace.ug.edu.gh

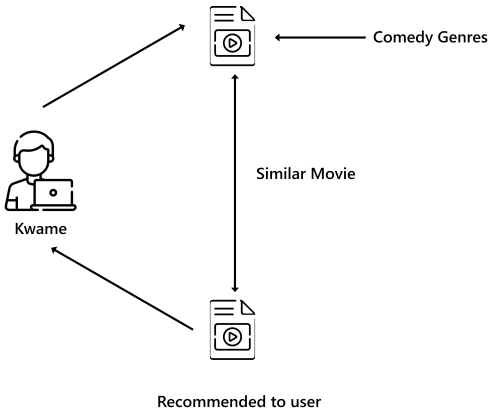


Figure 2.2: Illustration of content-based filtering, where Kwame's past preferences are matched with similar candidate items based on shared item attributes.

Figure 2.2 depicts the case of content-based filtering where recommendations are generated for Kwame based on similarity in item features (e.g., genre), using Kwame's past preferences.

Hybrid recommender systems combine both collaborative and content-based methods to improve accuracy and overcome the weaknesses of each individual approach. This is the most common strategy used in modern platforms like Netflix and Amazon (Linden et al., 2003).

User data is central to building recommender systems and can be collected in two ways: explicit and implicit feedback. Explicit data includes ratings, reviews, or direct responses to surveys, where users intentionally share their opinions (Guerreiro & Rita, 2020). However, this approach can be inconsistent, as users may rate products differently based on mood, trends, or social pressures (Gena et al., 2011; McAuley & Leskovec, 2013). In contrast, implicit data is gathered by observing user actions, such as clicks, purchases, or time spent on content (Yang et al., 2018). While implicit data helps reduce bias, it can also be unclear; for example, spending a long time on a webpage does not always mean a

University of Ghana http://ugspace.ug.edu.gh

positive experience (Chu et al., 2019). Combining both types of data generally improves recommendation accuracy (Mandal & Maiti, 2018).

The development of recommender systems has evolved significantly over the past few decades. Early systems were rule-based, relying on fixed logic to make recommendations. One of the first examples, Grundy, was a book recommender system designed to mimic the behaviour of a librarian (Ekstrand et al., 2011). In the early 1990s, Tapestry, developed at Xerox PARC, introduced collaborative filtering to help users sort emails based on their preferences and the actions of others (Goldberg et al., 1992). Around the same time, GroupLens, developed at MIT, applied collaborative filtering to internet newsgroups to help users find relevant content by using community ratings (Resnick et al., 1994). These early systems laid the groundwork for today's recommendation technologies.

As the internet and e-commerce platforms like Amazon and Netflix grew in the late 1990s and early 2000s, recommender systems became vital for customer engagement and product discovery (Ansari et al., 2000; Linden et al., 2003). These platforms introduced scalable machine learning algorithms that continuously learn from user interactions to personalise content recommendations.

The Netflix Prize, launched in 2006, was a landmark moment in recommender system research. Netflix offered a 1,000,000.00 USD prize to any team that could improve the accuracy of its recommendation algorithm by at least 10 percent. The winning team in 2009, BellKor's Pragmatic Chaos, applied advanced collaborative filtering and matrix factorization techniques (Bridge et al., 2005; Ihler et al., 2007; Schafer et al., 2001). This competition inspired global interest and innovation in recommendation algorithms.

Recommender systems have become essential across multiple industries where

University of Ghana http://ugspace.ug.edu.gh

personalised content delivery is critical. In e-commerce, companies such as Amazon, Alibaba, and eBay use recommendation engines to suggest products based on previous purchases, browsing history, and customer profiles, increasing conversion rates and customer satisfaction (Linden et al., 2003). In social media, platforms like Facebook, TikTok, Instagram, and LinkedIn leverage recommender systems to recommend friends, pages, groups, or posts based on a user's interactions and social graph (Anandhan et al., 2018).

In entertainment and media, streaming platforms such as Netflix, YouTube, Spotify, and Disney+ use recommendation algorithms to personalise movies, shows, music playlists, and videos, thereby improving user engagement and retention (Zheng, 2017). News platforms like Google News and Flipboard use recommendation techniques to prioritise and personalise news articles for individual reader preferences. Other emerging application domains include education (personalised course recommendations), healthcare (tailored health advice and treatment recommendations), and smart energy systems (optimising home energy consumption patterns) (Dhavala et al., 2023). These diverse applications demonstrate the growing importance of recommender systems in enhancing user experience and business value across industries.

Recent advances in machine learning, particularly deep learning, have greatly enhanced the ability of recommender systems to process large-scale data. Techniques such as decision trees, Gaussian processes, linear models, and deep neural networks have been widely adopted (Khanal et al., 2020; Pang et al., 2019). The availability of high-performance GPUs and distributed computing platforms like TensorFlow has also allowed researchers and practitioners to build complex models on massive datasets, such as the widely used MovieLens dataset (François-Lavet et al., 2018).

University of Ghana http://ugspace.ug.edu.gh

The latest research explores combining deep learning with reinforcement learning, known as deep reinforcement learning (DRL), to improve the adaptability and personalisation of recommender systems (Morales et al., 2021). In this approach, the recommender system continuously interacts with user feedback, learning and adjusting its recommendation policies over time. DRL-based recommendation frameworks have shown promise across a variety of applications, from media and e-commerce to healthcare and smart energy systems (Dhavala et al., 2023).

2.3 General vs. Sequential Recommendation

Recommender systems can be broadly classified into two major categories: general recommender systems and sequential recommender systems. The distinction between the two lies in how they model and predict user preferences over time (Quadrana et al., 2018a). While general recommenders focus on long-term user interests, sequential recommenders emphasise recent user behaviour and contextual information to capture dynamic intent. Figure 2.3 shows the classification of recommender systems from the perspective of the recommendation problem characterisation and the broad classification of techniques used.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

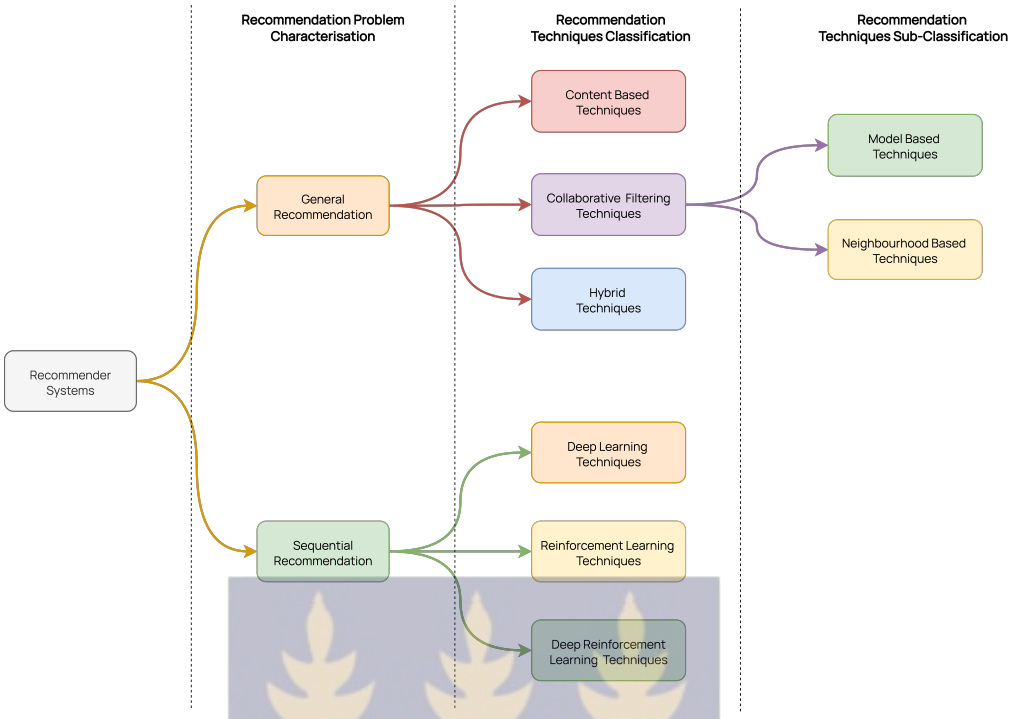


Figure 2.3: Classification of recommender systems by the underlying problem characterisation.

2.3.1 General recommender systems

General recommender systems are designed to model a user's overall, long-term preferences based on historical interaction data. These systems assume that user interests are relatively stable and seek to recommend items that align with these persistent preferences. User preferences are typically represented as a user-item interaction matrix, where each entry reflects an explicit or implicit feedback event, such as a rating, click, or purchase (Altulyan, 2020).

Given the sparse nature of user-item matrices, since most users interact with only a small fraction of available items, general recommenders are often formulated as matrix completion problems. Techniques like matrix factorisation and embedding-based collaborative filtering have been widely adopted to uncover latent user and item features and predict missing values (Lara-Cabrera et al., 2020).

University of Ghana http://ugspace.ug.edu.gh

While these systems have proven highly effective in domains like e-commerce (Amazon), online retail, and video streaming (Netflix), they can struggle to respond to short-term or contextual changes in user behaviour. For example, a user who primarily consumes action films may suddenly binge on romantic comedies for a short period. General recommenders, focused on historical behaviour, would likely fail to immediately adapt to this shift (Cheng et al., 2021).

2.3.2 Sequential recommender systems

Sequential recommender systems address this limitation by explicitly modelling the temporal order and recency of user interactions. They are designed to capture session-based or short-term patterns in user behaviour to predict the next item of interest (Quadrana et al., 2018a). These systems are particularly valuable in applications where immediate context strongly influences decision-making, such as video platforms, music streaming services, online shopping carts, and news recommendations.

Early sequential recommenders relied on rule-based systems and Markov chains. However, the field has advanced significantly with the introduction of deep learning approaches. Modern techniques use Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, Gated Recurrent Units (GRUs), and Graph Neural Networks (GNNs) to model the dependencies between items within a user's interaction sequence (Dong et al., 2018; Hidasi & Tikk, 2016; Wu et al., 2019). By learning these dependencies, sequential recommenders can provide highly relevant real-time suggestions that reflect users' current intentions, even when they differ from historical patterns.

University of Ghana http://ugspace.ug.edu.gh

2.3.3 Comparison and Emerging Trends

Both general and sequential recommenders offer complementary benefits. General recommenders are optimal for modelling stable long-term preferences, while sequential recommenders excel in capturing immediate, evolving user needs (Cheng et al., 2021). However, relying exclusively on one paradigm can be suboptimal. General systems may miss recent contextual shifts, whereas sequential systems may neglect valuable historical patterns. Recent research has focused on developing unified models that integrate long-term and short-term preference modelling to provide more accurate and holistic recommendations (Dong et al., 2018; Niu et al., 2020). These hybrid frameworks represent a critical advancement in personalised recommendation and are a key area of exploration in current research.

2.4 Deep Reinforcement Learning for Recommendation

The two paradigms surveyed in Section 2.3, general recommenders that model long-term preferences as a static matrix-completion problem, and sequential recommenders that model short-term intent through next-item prediction,each leave significant gaps. General recommenders treat user preferences as fixed and miss recent contextual shifts; sequential recommenders capture short-term dynamics but optimise only for the next single interaction, ignoring the long-term effect a recommendation has on subsequent user behaviour and platform engagement. Both paradigms are also fundamentally *prediction* approaches: they treat each recommendation as a single supervised forecast against held-out data, with no notion of how a recommendation today shapes the data the system observes tomorrow. This conflation of prediction with decision-making is the core limitation that reinforcement learning is designed to address.

University of Ghana http://ugspace.ug.edu.gh

Table 2.1: Comparison of General vs. Sequential Recommender Systems

Aspect	General Recommender Systems	Sequential Recommender Systems	References
Objective	Recommend items based on overall, long-term user preferences.	Recommend items based on sequential patterns or recent user interactions.	Hidasi et al. (2015) and Quadrana et al. (2018a)
Assumption about Preferences	User interests are relatively stable or persistent.	User interests evolve dynamically over short time intervals.	R. He et al. (2016) and Hidasi et al. (2015)
Data Utilised	Long-term historical user-item interactions without explicit temporal ordering.	Sequentially ordered user interactions, emphasising temporal context and dynamics.	R. He et al. (2016) and Quadrana et al. (2018a)
Algorithmic Approaches	Collaborative filtering, content-based filtering, hybrid approaches.	Markov models, recurrent neural networks (RNNs), transformer-based models, reinforcement learning.	Hidasi et al. (2015), Kang and McAuley (2018), and Sun et al. (2019)
Typical Application Scenarios	General product recommendations, such as movies, books, or consumer goods, where user preference is stable.	Short-term user engagement scenarios like music playlists, news feeds, or session-based e-commerce browsing.	Hidasi et al. (2015) and Quadrana et al. (2018a)
Challenges	Difficulty in capturing evolving preferences, cold-start problems.	Complex temporal modelling, computationally intensive methods, sensitivity to sparse data.	Hidasi et al. (2015), Quadrana et al. (2018a), and Sun et al. (2019)

University of Ghana http://ugspace.ug.edu.gh

Reinforcement Learning (RL) has therefore emerged as a powerful framework for personalised recommender systems, particularly in dynamic and interactive environments where user preferences evolve over time. Unlike static prediction models, RL-based methods treat the recommendation task as a sequential decision-making problem, enabling agents to adaptively learn from user feedback through interaction (Zhou et al., 2023). The action of recommending an item is thereby reframed as a decision whose value depends not only on whether the user engages with it now, but also on how that engagement reshapes the user's state for future decisions.

With the advent of Deep Reinforcement Learning (DRL), where deep neural networks are used to approximate policies and value functions, the integration of RL with recommender systems has become scalable enough for real-world deployment. The remainder of this section establishes the formal vocabulary needed to discuss DRL recommenders (Section 2.4.1), surveys the leading studies that apply RL and DRL to recommendation (Section 2.4.3), and identifies adjacent directions outside the central focus of this thesis (Section 2.4.4).

2.4.1 Reinforcement Learning Preliminaries

The remainder of this chapter assumes a working familiarity with the formal vocabulary of reinforcement learning, which is briefly summarised here. A reinforcement learning problem is formalised as a Markov Decision Process (MDP), specified by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, $\mathcal{P}(s' \mid s, a)$ is the transition kernel giving the probability of moving to state s' when action a is taken in state s, $\mathcal{R}(s, a)$ is the (possibly stochastic) reward function, and $\gamma \in [0, 1)$ is a discount factor that down-weights distant rewards (Sutton & Barto, 2018).

University of Ghana http://ugspace.ug.edu.gh

The agent's behaviour is described by a policy π. A *stochastic* policy $\pi(a \mid s)$ is a probability distribution over actions conditioned on the current state, while a *deterministic* policy $\mu(s)$ maps each state to a single action. DDPG, the algorithm at the centre of this thesis, uses a deterministic policy. The objective of learning is to find a policy that maximises the expected discounted return $\mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t)\right]$.

Two value functions support this optimisation. The *state-value* function $V^\pi(s) = \mathbb{E}_\pi[\sum_t \gamma^t r_t \mid s_0 = s]$ measures the expected return from state s under policy π, and the *action-value* (or Q-) function $Q^\pi(s,a) = \mathbb{E}_\pi[\sum_t \gamma^t r_t \mid s_0 = s, a_0 = a]$ measures the expected return from taking action a in s and following π thereafter. Both satisfy the Bellman expectation equation $Q^\pi(s,a) = \mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}}\left[V^\pi(s')\right]$, which forms the foundation of all temporal-difference learning methods.

Modern DRL algorithms divide along the lines of how they exploit this structure. *Value-based* methods (such as DQN and its descendants) approximate Q^* directly and derive a policy by acting greedily with respect to it. *Policy-based* methods parameterise the policy explicitly and optimise it via gradient ascent on the expected return. *Actor–critic* methods combine the two: an actor network parameterises the policy and a critic network estimates the value function, with the actor's gradient updates guided by the critic's evaluations (Shakya et al., 2023; Sumiea et al., 2024). This actor–critic structure is the foundation on which DDPG, TD3, SAC, and the Student's t-augmented variant proposed in this thesis all rest.

2.4.2 Motivation for Reinforcement Learning in recommender systems

Traditional recommender systems, particularly those based on collaborative filtering or content-based methods, are inherently myopic; they optimise for immediate accuracy

University of Ghana http://ugspace.ug.edu.gh

metrics such as click-through rate or top-k precision, without considering the long-term impact of a recommendation. This short-sightedness is problematic in settings where each recommendation can influence future user behaviour and system utility.

In contrast, reinforcement learning provides a principled framework to optimise for long-term cumulative rewards, such as user retention, session length, or lifetime value. In RL terms, the user-system interaction is modelled as a Markov Decision Process (MDP).

This approach enables the system to adapt over time, personalise recommendations based on feedback, and incorporate exploration strategies to uncover new user interests.

2.4.3 Key Works Applying RL and DRL to Recommender Systems

The application of RL and DRL to recommender systems has progressed along three broad lines, each addressing a different limitation of supervised recommendation. The first concerns the *structure of the action space*, and how to formulate the recommendation decision so that an RL agent can act efficiently within it. The second concerns the *learning signal*, how to optimise long-horizon objectives such as engagement or retention from sparse, delayed feedback, often using only logged interaction data. The third concerns *multi-context coordination*, how to share knowledge across users, sessions, or domains. The seven studies reviewed below are organised into these three clusters; each cluster closes with a critique that identifies which limitations of DDPG-based recommendation each cluster does not address, motivating the focus of this thesis.

2.4.3.1 Page-wise Recommendation with DRL

Zhao et al. (2018) proposed one of the first DRL frameworks to generate recommendation lists in an end-to-end manner. Their method, Deep Reinforcement Learning for List-wise

University of Ghana http://ugspace.ug.edu.gh

Recommendations, models the ranking of a list of items as a sequential decision problem in which each added item affects the reward of the entire list. The framework uses a policy-gradient algorithm over a slate-based action space and optimises not only click-through rate but also list-level objectives such as novelty and diversity. The principal contribution is a shift in the unit of decision-making from the individual item to the ordered list, which aligns the optimisation target with how recommendations are actually consumed.

2.4.3.2 SlateQ: Modelling Item Interdependence in Slates

Where Zhao et al. (2018) treat the slate holistically through policy gradients, Ie et al. (2019) attack the same problem from a value-based angle. SlateQ proposes a factorised Q-function that decomposes a slate-level value into per-item contributions under a documented user-choice model, making large combinatorial action spaces tractable. The decomposition allows the agent to learn from slate-level feedback without requiring an exponential action space, and the user-choice assumption ensures the factorisation is principled rather than arbitrary. SlateQ's contribution is therefore complementary to that of Zhao et al.: both treat the slate as the action, but SlateQ provides scalable inference where Zhao et al. provide flexible optimisation.

2.4.3.3 Reinforcement Learning for Session-Based Recommendation

Quadrana et al. (2018b) take a different route to the same combinatorial-action-space problem. Rather than reasoning over slates, they map each recommendation decision into a continuous action vector in an item-embedding space, allowing actor-critic methods to operate on a continuous action manifold. This formulation is closer to the standard DDPG setting and is the formulation adopted in this thesis. It avoids the combinatorial explosion of slate-based methods at the cost of introducing an embedding-to-item retrieval step at inference, with associated discrepancies between the continuous action selected and the

University of Ghana http://ugspace.ug.edu.gh

discrete item ultimately served.

Critique of Cluster A. The three studies in Cluster A converge on the conclusion that the action-space formulation is a critical design choice, but they leave the *exploration* question almost entirely unaddressed. Zhao et al. (2018) use standard policy-gradient sampling, Ie et al. (2019) use Boltzmann or epsilon-greedy on the factorised Q-values, and Quadrana et al. (2018b) use Gaussian or OU action noise inherited from the reference DDPG implementation. None of them examine whether the noise process used to perturb actions is well-suited to the high-dimensional, sparse-reward regime of sequential recommendation. This is the gap the present thesis exploits: holding the action-space formulation fixed (continuous embedding, as in Quadrana et al. (2018b)), it asks whether the *distributional shape* of the perturbation matters.

2.4.3.4 Learning to Collaborate for Long-Term Recommendations

Lei and Li (2019) introduced an RL-based collaborative filtering method explicitly designed to capture long-term dependencies in user–item interactions. The framework constructs an MDP in which users' latent preferences evolve, and uses Q-learning to optimise long-term engagement. Methodologically, the contribution is a bridge between collaborative filtering and RL: the user–item rating matrix is reinterpreted as a transition kernel rather than a static target, allowing the system to reason about how a recommendation today changes the available preferences tomorrow.

2.4.3.5 Top-K Off-Policy Correction

Chen et al. (2019) address a closely related but distinct problem: how to learn an RL policy from logged data alone, without online exploration. Their off-policy correction framework reweights logged interactions by an importance ratio between the target and behaviour

University of Ghana http://ugspace.ug.edu.gh

policies, allowing reliable policy improvement from logs that lack exploration diversity. The work is particularly important because it makes RL recommendation feasible in production settings where online experimentation is prohibitively expensive or risky.

2.4.3.6 Hierarchical Reinforcement Learning for User Engagement

Ji et al. (2024) attack the long-horizon problem differently again, decomposing the recommendation policy into a high-level controller that selects abstract goals (e.g. "encourage exploration of a new genre") and a low-level policy that selects specific items to fulfil each goal. The hierarchical decomposition addresses the credit-assignment problem inherent in long-horizon objectives: temporal abstraction reduces the effective horizon for each level, easing optimisation.

Critique of Cluster B. These three studies share a common assumption: that the principal challenge of RL for recommendation is the *learning signal* (long-horizon, off-policy, hierarchical), and that the agent's exploration strategy is a secondary concern. Chen et al. (2019) explicitly avoid the exploration question by working off-policy from logs; Lei and Li (2019) use ε-greedy in tabular spaces; Ji et al. (2024) use Gaussian action noise at the low level. None investigates whether systematic biases in the noise process, such as the tendency of light-tailed noise to underrepresent rare but high-value actions, compound the credit-assignment difficulties they otherwise mitigate. The action-perturbation mechanism studied in this thesis is orthogonal to and composable with all three contributions.

2.4.3.7 Multi-Agent Actor-Critic for Cross-Domain Recommendations

Wang et al. (2024) recognise that production recommenders increasingly span multiple domains (e.g. music, video, books) and propose a multi-agent actor-critic architecture in which each domain has its own actor while a shared critic enforces cross-domain coherence.

University of Ghana http://ugspace.ug.edu.gh

The work demonstrates that coordination through a global value function can improve both intra-domain and cross-domain performance, treating cross-domain recommendation as a multi-agent reinforcement learning problem rather than a transfer-learning problem.

Critique of Cluster C. Multi-agent coordination introduces additional sources of training instability beyond those present in single-agent DDPG: the non-stationarity each agent's policy creates for the others, and the potentially conflicting gradients flowing from a shared critic. Wang et al. (2024) address these through standard multi-agent actor-critic stabilisation tricks, but inherit single-agent exploration issues at each agent. If the per-agent exploration is biased toward narrow regions of the action space (as it typically is under Gaussian or OU noise), the shared critic sees a correspondingly biased view of the joint action space, which can amplify rather than resolve coordination failures. Heavy-tailed action perturbation, as proposed in this thesis, is therefore a single-agent intervention with potential implications for multi-agent settings, though that extension is left to future work.

Across all three clusters, a consistent pattern emerges: the literature has invested heavily in formulating the recommendation MDP, in handling its long-horizon and off-policy structure, and in coordinating multi-context decisions, but it has invested comparatively little in the specific choice of action-perturbation noise. Where exploration is discussed at all, it is treated as a hyperparameter rather than a design dimension. Yet the conditions of sequential recommendation, high-dimensional embedding action spaces, sparse engagement rewards, and heterogeneous user behaviour, are precisely those in which the choice of noise distribution can be expected to matter most. This observation motivates the focused study of heavy-tailed action perturbation that the present thesis undertakes.

University of Ghana http://ugspace.ug.edu.gh

2.4.4 Adjacent and Underrepresented Directions

The seven studies discussed above span the dominant lines of research at the intersection of RL and recommendation, but several adjacent directions also bear on the present thesis and deserve brief acknowledgement.

Transformer-based sequential models with RL fine-tuning. A recent line of work treats sequential recommendation as a sequence-modelling problem and applies transformer architectures to encode user histories, with RL used as a fine-tuning stage to align the resulting representations with long-horizon objectives. This direction is methodologically orthogonal to the action-perturbation question studied here: the transformer addresses how the *state* is represented, while the present thesis addresses how the *action* is perturbed. The two contributions are composable, and a transformer-based state encoder could be substituted into the framework without altering the heavy-tailed exploration mechanism.

Offline reinforcement learning for recommendation. Production recommender systems typically cannot perform unrestricted online exploration, motivating offline-RL approaches that learn improved policies from logged interaction data alone (C. Gao et al., 2023; Lin et al., 2024). Offline RL faces a different bias problem from the one studied here, value extrapolation outside the support of the logged behaviour policy, and its mitigations (conservative Q-learning, behaviour regularisation) operate on the critic rather than on action selection. The present thesis works in the conventional online (or, equivalently, simulator-based) regime of DDPG; whether heavy-tailed action perturbation has analogous benefits in fully offline settings is an open question left to future work.

University of Ghana http://ugspace.ug.edu.gh

Bandit–RL hybrids. Contextual-bandit formulations remain a strong baseline for recommendation when long-horizon credit assignment is not the binding constraint, and several works hybridise bandits with full RL methods to exploit the strengths of each. Bandit methods sidestep the exploration question studied here by using closed-form upper-confidence-bound or Thompson-sampling exploration; this is feasible because they assume immediate reward feedback. The full-RL setting that the present thesis targets cannot use these closed-form mechanisms because the value of an action depends on a long sequence of subsequent decisions, which is precisely why the noise process matters.

Model-based RL. A separate strand of work attempts to learn a generative model of user response and plan against it, rather than learning model-free from observed interactions. Model-based methods can in principle reduce the sample complexity of RL recommendation, but they introduce a model-misspecification risk: errors in the user-response model propagate through planning. The present thesis adopts the model-free perspective; the noise mechanism it studies is independent of whether learning is model-free or model-based, but its empirical evaluation is in the model-free regime.

These four directions are flagged here for completeness; none is the central focus of this chapter, and each merits a deeper treatment than the present scope allows.

2.4.5 Actor-Critic Methods in Recommender Systems

Actor-critic methods have gained popularity in DRL-based recommender systems due to their sample efficiency and stability in high-dimensional or continuous action spaces. These methods consist of two components:

- The actor, which learns a policy to select actions (recommendations) given the current state.

University of Ghana http://ugspace.ug.edu.gh

- The critic, which evaluates the quality of the selected action by estimating the value or advantage function.

Actor–critic architectures such as A2C, DDPG, TD3, and SAC have been explored for recommendation tasks (Shakya et al., 2023; Sumiea et al., 2024). Among them, DDPG occupies a particular niche due to its deterministic policy formulation, which avoids the variance overhead of stochastic-policy alternatives, and its off-policy training, which allows effective reuse of historical interactions in settings where new interactions are expensive to obtain (Lin et al., 2024). In recommender systems, the actor network can be parameterised as a function that generates item embeddings or rankings, while the critic learns to estimate long-term user satisfaction or engagement. This architecture supports end-to-end training and the optimisation of long-term objectives beyond immediate feedback.

2.5 The DDPG Algorithm and Its Limitations

The Deep Deterministic Policy Gradient (DDPG) algorithm is a popular actor-critic reinforcement learning method designed for environments with continuous or high-dimensional action spaces (Lillicrap et al., 2015). It combines the deterministic policy gradient framework with deep neural networks to learn both value functions and policies in an off-policy setting. In recent years, DDPG has been adapted for recommender systems by embedding discrete items into a continuous action space, enabling scalable optimisation over large catalogues (Zhou et al., 2023).

2.5.1 Overview of the DDPG Algorithm

DDPG consists of two networks: an actor, which maps states to deterministic actions, and a critic, which estimates the action-value (Q) function. It learns from past experience stored in a replay buffer and stabilises training using slowly updated *target networks*. Unlike

University of Ghana http://ugspace.ug.edu.gh

value-based methods such as DQN, which require discrete action spaces, DDPG can directly optimise in continuous domains using the deterministic policy gradient theorem shown in Eqn. (2.1):

$$\nabla_{\theta^\mu} J \approx \mathbb{E}_{s\sim\mathcal{D}} \left[\nabla_a Q(s,a|\theta^Q) \nabla_{\theta^\mu} \mu(s|\theta^\mu) \right] \tag{2.1}$$

where:

- θ^μ represents the parameters of the actor (policy) network,
- θ^Q represents the parameters of the critic (Q-function) network,
- $\mathcal{D}$ is the replay buffer containing past transitions,
- $\mu(s \mid \theta^\mu)$ is the deterministic action from the actor,
- $Q(s,a \mid \theta^Q)$ is the action-value estimate from the critic.

The critic is updated using the Bellman error shown in Eqn. (2.2).

$$L(\theta^Q) = \mathbb{E}_{s,a,r,s'} \left[\left(Q(s,a \mid \theta^Q) - \left(r + \gamma Q'(s', \mu'(s' \mid \theta^{\mu'}) \mid \theta^{Q'}) \right) \right)^2 \right] \tag{2.2}$$

where:

- $Q(s,a \mid \theta^Q)$ is the action-value estimate from the online critic,
- r is the reward received after taking action a in state s,
- γ is the discount factor,
- $Q'(s', \mu'(s' \mid \theta^{\mu'}) \mid \theta^{Q'})$ is the target critic's evaluation of s' under action $\mu'(s')$.

The term inside the square is the Temporal Difference (TD) error. The actor is updated by ascending the policy gradient, estimated by back-propagating through the critic.

University of Ghana http://ugspace.ug.edu.gh

2.5.2 DDPG in Recommender Systems

In recommender systems, DDPG has been employed to generate high-dimensional item embeddings as actions, which are then matched to actual items in a catalogue. The system observes a user's interaction history (state), generates a recommendation (action), receives a reward signal (e.g., click, rating), and updates its policy accordingly. This formulation allows DDPG to capture both short-term interaction dynamics and long-term user satisfaction, making it suitable for session-based and sequential recommendation tasks (Chen et al., 2019; T. Gao et al., 2023).

2.5.3 Limitations of DDPG in Recommendation Contexts

Despite its versatility, DDPG suffers from several key challenges when applied to recommendation environments, particularly in terms of exploration and value estimation.

2.5.3.1 Overestimation Bias in Actor-Critic Methods

One of the most well-documented weaknesses of DDPG and other actor-critic algorithms is overestimation bias in the Q-function. This occurs when the critic systematically overestimates the expected return for certain actions, due to compounding errors in value estimation. Since the actor is trained to maximise the critic's output, it may learn to exploit these overestimated values, reinforcing suboptimal policies. The bias originates from several sources: function approximation errors in deep networks, maximisation bias from selecting actions according to noisy value estimates, and the off-policy nature of replay-buffer data that may not reflect the current policy distribution. The downstream consequences include premature convergence to suboptimal actions, instability during training, and high variance in learned policies.

University of Ghana http://ugspace.ug.edu.gh

The principal mitigation strategies in the literature operate on the *critic* side of the actor–critic architecture and form a recognisable progression. Twin Delayed DDPG (TD3) (Fujimoto et al., 2018) introduces a pair of critics and constructs the temporal-difference target from the minimum of their estimates, replacing the positive bias of single-critic DDPG with a controlled negative (under-)estimation bias. Soft Actor-Critic (SAC) (Haarnoja et al., 2018) replaces the deterministic policy with a stochastic one regularised by an entropy term, encouraging broader exploration as a side-effect of maximising entropy. The more recent TD7 algorithm (Fujimoto et al., 2023) couples deterministic-policy learning with state–action representation learning, addressing value-estimation errors via a richer feature space rather than via the critic ensemble alone.

Each of these mitigations partially addresses overestimation bias, but each leaves the action-perturbation channel essentially untouched. TD3 inherits Gaussian or OU action noise from DDPG; SAC absorbs exploration into the policy distribution but uses Gaussian policies; TD7 uses small fixed-variance Gaussian perturbations for exploration. None of them treats the *distributional shape* of the noise process as a design lever. This is the gap the present thesis exploits: rather than adding another critic-side correction to the established TD3/SAC/TD7 line, it asks whether changing the action-perturbation distribution itself can reduce the bias and improve exploration in the regime that recommender systems present, namely sparse rewards and high-dimensional embedding action spaces.

2.5.3.2 Ineffective Exploration in Sparse Reward Settings

Exploration plays a pivotal role in the training dynamics of actor-critic algorithms such as the Deep Deterministic Policy Gradient (DDPG), particularly within high-dimensional and continuous action spaces. To address the exploration-exploitation dilemma, DDPG introduces stochasticity by perturbing the deterministic policy output with additive noise

University of Ghana http://ugspace.ug.edu.gh

during training. Various noise processes have been investigated in the literature, each with distinct statistical properties that make them suitable for different exploration scenarios, though not without limitations.

Lillicrap et al. (2015) originally proposed the use of Ornstein–Uhlenbeck (OU) noise in DDPG to generate temporally correlated exploration signals. OU noise mimics Brownian motion with momentum, thereby enabling smoother exploration trajectories in physical control tasks, especially where action continuity is beneficial. However, its applicability diminishes in environments where independent, state-specific exploration is preferred, as the temporal correlations can inhibit adaptability.

Alternatively, Gaussian noise, being temporally uncorrelated and easy to implement, is widely used for its simplicity and computational efficiency. Nevertheless, its lack of structural bias or long-tail behaviour limits its ability to escape local optima in complex reward landscapes, leading to suboptimal policy convergence in some cases.

2.5.4 Noise as a Strategy to Alleviate Overestimation Bias

Although exploration noise is mainly used to encourage exploration, it also helps alleviate overestimation bias by encouraging more diverse state–action samples. This prevents the actor from overfitting to inaccurately high Q-values, thereby improving generalisation by exposing the critic to a wider range of feedback.

Without adequate exploration, the agent is likely to overfit to Q-value peaks that are statistical artefacts rather than meaningful estimates of long-term value. Exploration introduces variability into the replay buffer, smoothing Q-function estimates and discouraging overly optimistic value propagation.

University of Ghana http://ugspace.ug.edu.gh

2.6 Exploration Noise in Deep Reinforcement Learning

Effective exploration is fundamental to the success of reinforcement learning (RL) algorithms. The exploration-exploitation trade-off requires agents to balance the pursuit of immediate rewards (exploitation) with the need to visit uncharted parts of the environment to discover better long-term strategies (exploration). In value-based methods such as DQN, this is often handled through ε-greedy action selection. However, in policy gradient and actor-critic methods, particularly in continuous or high-dimensional action spaces, exploration is typically achieved by injecting noise into the action selection process (Lillicrap et al., 2015).

This section reviews the most widely used exploration noise strategies in Deep Reinforcement Learning (DRL), focusing on Gaussian and Ornstein–Uhlenbeck (OU), and Lévy noise processes. It also surveys recent research on annealed and adaptive noise mechanisms and discusses their strengths and limitations, particularly in the context of sparse reward environments like recommender systems.

2.6.1 Gaussian Noise

Gaussian noise is a simple and widely used exploration strategy in deep reinforcement learning (DRL), particularly in continuous action spaces. It is implemented by perturbing the action output of the actor network with additive zero-mean Gaussian noise as shown in Eqn. (2.4).

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{2.3}$$

- x: the random variable for which the probability density is evaluated,

University of Ghana http://ugspace.ug.edu.gh

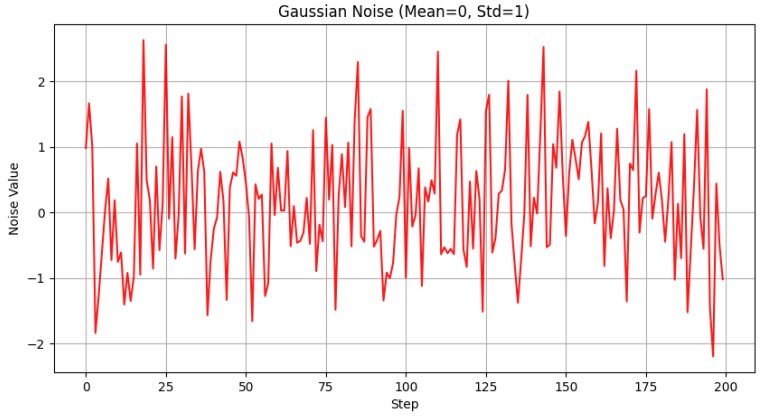


Figure 2.4: Sample paths from a zero-mean Gaussian noise process at three different standard deviations σ, illustrating that Gaussian noise produces independent samples at each timestep with no temporal structure. The probability of a perturbation more than three standard deviations from the mean is approximately 0.27%, illustrating why Gaussian noise rarely produces large exploratory excursions in the action space.

- μ: the mean of the distribution,
- σ^2: the variance of the distribution,
- σ: the standard deviation.

$$a_t = \mu(s_t) + \mathcal{N}(0, \sigma^2) \tag{2.4}$$

where:

- a_t is the noisy action taken at time t,
- $\mu(s_t)$ is the action produced by the actor network at time t,
- $\mathcal{N}(0, \sigma^2)$ is Gaussian exploration noise with mean 0 and variance σ^2.

To illustrate the effect of Gaussian noise, samples drawn from the zero-mean distribution applied to actions can be visualised. Figure 2.4 shows noise generated from $\mathcal{N}(0, \sigma^2)$ with varying standard deviations.

Gaussian noise offers several strengths that make it a popular choice in deep reinforcement

University of Ghana http://ugspace.ug.edu.gh

learning. It is simple to implement and analytically tractable, making it accessible for a wide range of applications. Its behaviour is generally well-suited to low-dimensional action spaces, where symmetric perturbations are often sufficient for effective exploration. Moreover, it is easily parameterised through a single variance term, allowing for straightforward control over the magnitude of noise applied.

However, Gaussian noise also exhibits notable limitations. It tends to scale poorly in high-dimensional action spaces, as it samples symmetrically around the policy mean, which can result in inefficient exploration. The probability of generating large exploratory jumps is inherently low, restricting the agent's ability to explore distant or sparse regions of the action space. Additionally, the use of a fixed variance introduces trade-offs in exploration efficiency: a large variance can destabilise learning due to excessive randomness, while a small variance may cause premature convergence or stagnation due to insufficient exploration.

While Gaussian noise works reasonably well for continuous control tasks such as locomotion in OpenAI Gym, it often underperforms in sparse-reward settings or tasks with deceptive optima, both common in recommendation environments.

2.6.2 Ornstein–Uhlenbeck Noise

To improve temporal consistency of exploration in physical domains, Lillicrap et al. (2015) introduced the use of **Ornstein–Uhlenbeck (OU) noise**, a stochastic process with temporal autocorrelation. Eqn. (2.5) defines the OU process used for correlated exploration noise.

$$dX_t = \theta(\mu - X_t)\,dt + \sigma\,dW_t \tag{2.5}$$

University of Ghana http://ugspace.ug.edu.gh

In Eqn. (2.5):

- X_t is the OU process at time t,
- $\theta > 0$ is the mean-reversion rate,
- μ is the long-term mean toward which the process reverts,
- σ is the volatility parameter,
- W_t is a standard Wiener process.

Similarly to the noisy action expressed in Eqn. (2.4), the OU noise applied to actions is defined in Eqn. (2.6).

$$a_t = \mu(s_t) + X_t \tag{2.6}$$

To illustrate the behaviour of Ornstein–Uhlenbeck noise, its trajectory over time can be visualised. Unlike Gaussian noise, which produces independent samples at each timestep, OU noise exhibits temporal correlation due to its mean-reverting property. Figure 2.5 shows a sample path generated from the OU process.

Ornstein–Uhlenbeck (OU) noise provides several advantages in deep reinforcement learning, particularly in continuous control tasks. It encourages temporally coherent actions, which is beneficial in physics-based environments where inertia and momentum play a significant role. By generating low-frequency perturbations, OU noise can simulate smooth, correlated exploration that aligns well with the dynamics of such tasks.

However, OU noise also has important limitations, especially when applied to domains like recommender systems. The underlying assumption of temporal smoothness inherent in OU processes does not align with the nature of recommendation tasks, where actions

University of Ghana http://ugspace.ug.edu.gh

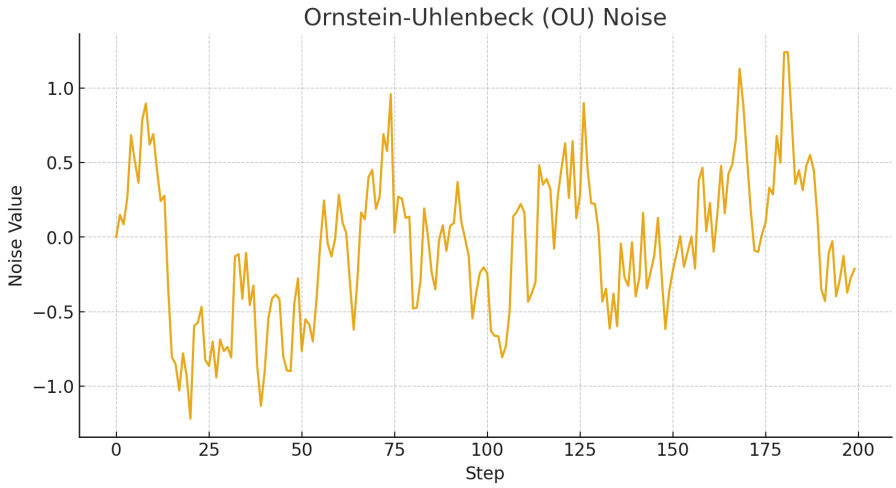


Figure 2.5: Sample path from an Ornstein–Uhlenbeck process with mean reversion rate θ and volatility σ, illustrating the temporal autocorrelation that distinguishes OU noise from Gaussian noise. Successive samples are positively correlated, producing the smooth low-frequency exploration trajectories that motivated the original use of OU noise in physical control tasks.

such as selecting item IDs can change abruptly and without continuity. Moreover, like Gaussian noise, OU noise remains light-tailed, making it prone to local exploration and less effective at discovering distant, high-value actions. Additionally, its performance is sensitive to the tuning of hyperparameters such as the mean reversion rate (θ) and volatility (σ), which can complicate the training process.

Despite its original success in continuous control, OU noise is increasingly being replaced in domains where temporal correlation is not meaningful or can hinder learning, such as session-based or discrete-space recommendation.

2.6.3 Adaptive and Parameterised Noise Methods

A separate line of research replaces fixed-variance noise with adaptive mechanisms that tie the perturbation to the data or to estimates of uncertainty. Three such mechanisms recur in the recent literature: parameter noise (Plappert et al., 2017), which perturbs the weights of the actor network rather than its outputs to produce structurally consistent exploration; Noisy Networks (Fortunato et al., 2017), which inject learnable stochastic

University of Ghana http://ugspace.ug.edu.gh

layers into the network so that the exploration magnitude is itself optimised by gradient descent; and ensemble methods such as Bootstrapped DQN (Osband et al., 2016), which represent epistemic uncertainty across an ensemble of Q-functions and drive exploration via Thompson sampling. Each of these is treated in greater depth in Section 2.7, where they are compared along the dimensions of where the stochasticity is injected and what kind of exploration each induces.

The advantages and drawbacks of adaptive parameterisation can be summarised at a high level here. On the positive side, these methods enable state-dependent or uncertainty-aware exploration, allowing agents to tailor their exploratory behaviour to the current context or to their confidence in the policy. This can improve sample efficiency, as exploration is focused where it is most informative, and can dynamically adjust to different stages of training. On the negative side, adaptive strategies are typically more complex to implement than fixed-variance noise, require careful tuning of additional hyperparameters, and may demand greater network capacity or ensemble architectures to estimate uncertainty reliably. Without appropriate regularisation, an adaptive policy can also collapse toward near-deterministic behaviour, undermining the intended benefits.

2.6.4 Potential of Lévy Noise

Unlike Gaussian or Ornstein–Uhlenbeck noise, which are light-tailed and tend to produce small, incremental perturbations, Lévy noise follows a heavy-tailed distribution that allows for infrequent but significant deviations from the policy mean. This results in a form of stochastic exploration that enables the agent to escape local optima and discover action regions that may be overlooked by traditional noise processes.

The potential of Lévy noise to alleviate overestimation bias lies in its capacity to increase

University of Ghana http://ugspace.ug.edu.gh

the diversity of experiences collected during training. By promoting broader coverage of the action space, Lévy noise reduces the agent's tendency to overfit to narrow regions of high but possibly erroneous Q-values. It can also help break the feedback loop that arises when a deterministic policy repeatedly exploits overestimated actions, thereby improving the robustness of value estimation.

However, this approach presents certain limitations. Lévy noise, by design, introduces abrupt and sometimes extreme shifts in the action space, which may destabilise the learning process if not properly regulated. Without careful tuning, such as clipping or annealing the noise magnitude, these large exploratory steps can lead to erratic policy updates and reduced sample efficiency. The infinite second moment of Lévy distributions with tail index below two is a further analytical complication that the Student's t-distribution avoids for $\nu > 2$, which is one reason the present thesis adopts the Student's t rather than a stable-distribution alternative.

2.6.5 Theoretical Foundations of Student's t-Distributed Noise

The motivation for using Student's t-distributed noise in DDPG rests on four converging lines of theoretical support, drawn from robust statistics, Bayesian inference, the theory of heavy-tailed search processes, and recent results on the sample complexity and metastability of heavy-tailed policy gradients. This subsection synthesises each in turn before drawing the connection to the action-perturbation setting that the thesis develops.

Robust statistics: the t-distribution as an outlier-robust alternative to the Gaussian. The Student's t-distribution arises naturally in robust statistics as an outlier-robust alternative to the Gaussian. The maximum-likelihood estimator under a t-distributed observation model is an M-estimator whose influence function is *bounded*, in contrast

University of Ghana http://ugspace.ug.edu.gh

to the unbounded influence function of the Gaussian-derived sample mean. Bounded influence is the formal property that makes an estimator insensitive to extreme observations: a single outlier can shift the Gaussian sample mean arbitrarily, but cannot shift the t-MLE by more than a fixed amount that depends on the degrees of freedom ν (Huber, 2004). The same property gives the t-distribution a privileged role in Bayesian linear regression: replacing the Gaussian likelihood with a t-likelihood produces a posterior that automatically down-weights observations far from the bulk of the data, without requiring explicit outlier detection. The implication for action-perturbation noise is that the t-distribution is a principled choice rather than an arbitrary heavy-tailed alternative: it is the simplest one-parameter family that interpolates continuously between full robustness ($\nu \to 0$) and full Gaussian behaviour ($\nu \to \infty$).

Bayesian heavy-tailed priors and reinforcement learning. A second line of theoretical support comes from the use of heavy-tailed priors in Bayesian reinforcement learning. Several authors have observed that placing a heavy-tailed prior over reward or value distributions improves robustness to model misspecification and to the long-tailed return distributions that arise in real environments (Cayci & Eryilmaz, 2023). Garg et al. (2021) document that gradient updates in policy-gradient methods themselves exhibit heavy-tailed behaviour even when the underlying reward distribution is light-tailed, motivating robust estimators in place of standard clipping heuristics. The Student's t-distribution figures in this literature as the canonical heavy-tailed prior with finite second moment when $\nu > 2$, distinguishing it from purely heavy-tailed alternatives such as the Cauchy distribution (which has no finite mean) or stable distributions (whose second moment is undefined for tail index less than two).

University of Ghana http://ugspace.ug.edu.gh

Heavy-tailed search processes in continuous-control RL. A third strand of work examines heavy-tailed search processes directly in the context of continuous-action reinforcement learning. Kobayashi (2019) demonstrate that Student's t-policies in standard robot control benchmarks support a form of Lévy-flight-style exploration that escapes local optima inaccessible to Gaussian policies. Chakraborty et al. (2023) show empirically that heavy-tailed policy gradients improve learning in sparse-reward continuous control, where Gaussian-based methods stagnate. L. Zhu et al. (2025), working within the q-exponential family, demonstrate that the Student's t-distribution as a policy is more stable than the Gaussian across both online and offline benchmarks. These works share a common explanation: heavy-tailed processes have non-negligible probability mass at large deviations from the mean, allowing the agent to occasionally make exploratory steps that escape the basin of attraction of suboptimal critic estimates, while still spending most of its time near the current policy mean.

Sample complexity and metastability of heavy-tailed policy search. The fourth and most recent line of support concerns the sample complexity of heavy-tailed policy search. Bedi et al. (2024) provide convergence-rate and metastability analyses for heavy-tailed policy parameterisations, showing that distributions with heavier tails converge to wider peaks of the value landscape, a property they term "metastable". The intuition is that gradient descent under a heavy-tailed sampling distribution biases the optimisation toward solutions that are robust under the perturbations the heavy-tailed sampling itself produces, which corresponds to wider rather than narrower optima. Cayci and Eryilmaz (2023) establish finite-time sample-complexity bounds for natural actor–critic methods under heavy-tailed reward distributions, with dynamic gradient clipping as the stabilising mechanism. Together, these results give principled support for the empirical observations

University of Ghana http://ugspace.ug.edu.gh

of the third strand: heavy-tailed processes are not merely useful in practice; they admit a finite-sample theoretical analysis that the Gaussian case lacks in the heavy-tailed-data regime.

Synthesis: from theory to action-perturbation. Each of the four strands operates at a different level of analysis (statistical, Bayesian, dynamical, finite-sample), but they converge on a common conclusion: in regimes characterised by sparse signal, high dimensionality, or heavy-tailed reward structure, the Student's t-distribution offers theoretical advantages over the Gaussian at a controlled cost of additional variance. The present thesis applies this synthesised theoretical case to a setting that the existing literature has not directly examined: action-perturbation noise in DDPG-based recommender systems, where reward sparsity and action-space dimensionality are both severe.

Compared to alternative heavy-tailed candidates such as Lévy noise or the Cauchy distribution, the Student's t-distribution offers two practical advantages for action perturbation. First, its finite second moment for $\nu > 2$ keeps the policy gradient well-defined without requiring stable-distribution machinery. Second, its single-parameter family allows the tail behaviour to be *annealed*: training can begin with a small ν for broad exploration and gradually increase ν to recover near-Gaussian behaviour as the policy converges, providing the curriculum-style schedule that Cayci and Eryilmaz (2023) identify as desirable for stable training under heavy-tailed processes.

2.6.6 Summary of Exploration Noise Strategies

Table 2.2 summarises the strengths and weaknesses of the major exploration noise strategies discussed in this section.

The combined motivation from Section 2.6.5 and the comparison in Table 2.2 establishes

University of Ghana http://ugspace.ug.edu.gh

Student's t-distributed noise as the principled choice for the parameterised exploration mechanism this thesis develops. The integration of bounded, scale-annealed Student's t-distributed noise into DDPG is detailed methodologically in Chapter 3, and the comparison against Gaussian and OU baselines on benchmark recommendation datasets is presented in Chapter 4.

2.7 Related Works on Exploration Improvements in Deep Reinforcement Learning

Exploration remains one of the central challenges in deep reinforcement learning (DRL), particularly in environments characterised by sparse rewards, delayed feedback, or large state–action spaces. While DDPG and its variants have achieved success in many continuous-control tasks, they are routinely reported to suffer from insufficient exploration in the recommendation regime. This section reviews the principal strategies that have been proposed to enhance exploration in DDPG and related actor–critic algorithms, organising them not as isolated technical entries but along three comparative dimensions: *where* the stochasticity is injected (in the action, the network parameters, the policy distribution, or the value function), *what* kind of exploration each induces (local versus global, temporally correlated versus independent, light-tailed versus heavy-tailed), and *where* each strategy fails for the regime relevant to this thesis (high-dimensional embedding action spaces, sparse engagement rewards, off-policy learning from logged trajectories). Reading these methods comparatively rather than sequentially clarifies why the heavy-tailed action-perturbation question remains open even after substantial work on adjacent questions.

2.7.1 The Underlying Limitation Shared Across Methods

The standard exploration mechanisms used in actor–critic methods can be characterised by three failure modes when applied to sequential recommendation. The first is *locality of exploration*: action-perturbation methods such as Gaussian or OU noise produce

University of Ghana http://ugspace.ug.edu.gh

perturbations concentrated near the current policy output, restricting the agent's ability to escape local optima in the action-value landscape. The second is *poor scaling in high dimensions*: in continuous embedding spaces with hundreds of dimensions, symmetric light-tailed noise produces almost no probability mass at distant action regions, so the gradient signal for off-policy actions is effectively absent. The third is *overestimation feedback loops*: when exploration is poor, the critic sees only a narrow slice of the state–action space, and any spurious overestimates in unexplored regions are reinforced rather than corrected. The four families of methods discussed below address subsets of these failure modes; none addresses all three simultaneously, and none challenges the dominant assumption that a light-tailed noise distribution is the right starting point.

2.7.2 Parameter-Space Perturbation: Parameter Noise and NoisyNets

A first family of methods reframes exploration as a problem of perturbing the *parameters* of the policy network rather than its outputs. Parameter noise (Plappert et al., 2017) samples a perturbation vector and applies it to the actor weights θ^{μ} to obtain $\tilde{\theta}^{\mu} = \theta^{\mu} + \Delta\theta$ as defined in Eqn. 2.8.

The perturbation vector is sampled from a multivariate Gaussian distribution whose probability density is given by Eqn. (2.7):

$$\mathcal{N}(\mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\,|\Sigma|^{1/2}} \exp\left(-\tfrac{1}{2}(x-\mu)^{\top}\Sigma^{-1}(x-\mu)\right) \tag{2.7}$$

- $x \in \mathbb{R}^d$ is the random vector whose probability density is being evaluated.
- $\mu \in \mathbb{R}^d$ denotes the mean vector of the distribution, representing its centre.
- $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance matrix, capturing variances along each dimension and correlations between them.

University of Ghana http://ugspace.ug.edu.gh

- d is the dimensionality of the random vector x.

$$\tilde{\theta}^{\mu} = \theta^{\mu} + \mathcal{N}(0, \sigma^2 I) \tag{2.8}$$

where:

- θ^{μ} denotes the original parameters of the policy network before perturbation.
- $\tilde{\theta}^{\mu}$ represents the perturbed policy parameters after adding parameter noise.
- $\mathcal{N}(0, \sigma^2 I)$ is a multivariate Gaussian distribution with zero mean and covariance matrix $\sigma^2 I$, from which the noise vector is sampled.

In practice, the original parameter-noise formulation of Plappert et al. (2017) does not treat the scale σ as a fixed hyperparameter; rather, σ is *adaptively scheduled* against a target action-space deviation δ. After each rollout the algorithm measures the distance $d(\pi, \tilde{\pi})$ between the unperturbed and perturbed policies in action space and updates σ as

$$\sigma_{k+1} = \begin{cases} \alpha \sigma_k & \text{if } d(\pi_k, \tilde{\pi}_k) < \delta, \\ \sigma_k / \alpha & \text{otherwise,} \end{cases} \tag{2.9}$$

with scaling factor $\alpha > 1$ and target deviation δ, both treated as hyperparameters. The update rule increases σ when the perturbed policy is too close to the unperturbed one (insufficient exploration) and decreases it otherwise. A minimum noise factor $\sigma_{\min}$ is typically enforced to prevent the noise scale from collapsing to zero in late training, and an analogous treatment is given to other temperature-like parameters such as the action-distance kernel bandwidth. This adaptive mechanism is critical to parameter noise's empirical performance, as a fixed σ either underexplores (small σ) or destabilises learning

University of Ghana http://ugspace.ug.edu.gh

(large σ) for any non-trivial training horizon.

The parameter-perturbed policy $\tilde{\mu}$ is then used to act for the entirety of an episode, producing temporally consistent and state-dependent exploration: the same state always yields the same perturbed action within an episode, which contrasts with the independent perturbations of action-space noise. This consistency is the central advantage of the parameter-space approach, since correlated action trajectories support more coherent value learning than uncorrelated jitter.

Fortunato et al. (2017) extended this idea with NoisyNets, in which the noise parameters are themselves *learnable*. A NoisyNet layer computes

$$y = (W + W_{\text{noise}} \odot \varepsilon_W)x + (b + b_{\text{noise}} \odot \varepsilon_b), \tag{2.10}$$

where W_{noise} and b_{noise} are learned alongside the policy weights, ε_W and ε_b are sampled from a fixed noise distribution at each step, and $\odot$ denotes element-wise multiplication. Where parameter noise treats the noise scale σ as a hyperparameter to be tuned (or, in adaptive variants, scheduled against a target action-deviation threshold), NoisyNets allow the agent to learn how much exploration each layer should contribute via standard backpropagation.

Comparative critique of parameter-space methods. Both parameter noise and NoisyNets address the locality complaint against action-space noise: by perturbing weights rather than outputs, they produce structurally consistent exploratory behaviour rather than uncorrelated jitter, and they have been shown to outperform Gaussian noise in several continuous-control benchmarks. Both, however, retain the assumption that the underlying perturbation distribution should be Gaussian (light-tailed). For network parameters this

University of Ghana http://ugspace.ug.edu.gh

is empirically reasonable, but it is also a constraint inherited unexamined from the deep-learning weight-initialisation literature rather than a deliberate exploration design choice. Furthermore, parameter-space perturbation interacts non-trivially with batch normalisation and layer normalisation in deep policies (Plappert et al., 2017), requiring additional infrastructure to compute action-space divergence metrics, and NoisyNets can collapse toward near-deterministic behaviour during training without proper regularisation. Both methods therefore shift the locality problem from action space to parameter space without addressing it at root: the perturbation distribution is still light-tailed, just applied at a different point in the network.

2.7.3 Value-Function-Driven Exploration: Bootstrapped Ensembles

A second family of methods takes a different stance: rather than perturbing actions or parameters, it represents *epistemic uncertainty* in the value function and lets that uncertainty drive exploration. Bootstrapped DQN (Osband et al., 2016) maintains an ensemble of K Q-networks, each trained on a bootstrap-resampled subset of the replay buffer; at episode start, the agent samples one ensemble member and acts greedily with respect to it for the duration of the episode, effectively implementing Thompson sampling over value functions. The exploration that emerges is global in a different sense from parameter-space methods: the agent commits to a single hypothesis about the value function for an entire episode rather than perturbing within a single hypothesis, and ensemble disagreement naturally identifies regions of the state–action space where additional exploration is most informative.

Comparative critique of ensemble methods. Ensemble approaches address the third failure mode, overestimation feedback loops, more directly than action- or parameter-space methods, because the diversity of value estimates across the ensemble naturally exposes

University of Ghana http://ugspace.ug.edu.gh

regions of high disagreement that warrant exploration. They do not, however, address the locality and high-dimensionality failures, since once an ensemble member is selected the agent acts greedily with respect to it. They are also computationally expensive, scaling memory and gradient computation linearly in K. Most importantly for the present thesis, ensemble methods are an alternative to action perturbation rather than a refinement of it; they are composable with heavy-tailed action perturbation in principle, but the literature has not examined this combination empirically.

2.7.4 Policy-Distribution Exploration: Maximum-Entropy Methods

A third family folds exploration directly into the policy objective. Soft Actor-Critic (SAC) (Haarnoja et al., 2018) parameterises the policy as a stochastic distribution and adds an entropy term to the standard reward objective, yielding the maximum-entropy objective

$$J(\pi) = \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\big[Q(s_t,a_t) - \alpha \log \pi(a_t \mid s_t)\big], \tag{2.11}$$

where α is a temperature parameter that trades off reward maximisation against policy entropy and ρ_π is the state–action distribution induced by following π. Recent work treats α itself as automatically tuned, eliminating a hyperparameter that previously required careful scheduling. The result is a policy that is stochastic by construction, with the degree of stochasticity (and hence the breadth of exploration) determined by the optimiser rather than fixed at design time.

Comparative critique of maximum-entropy methods. Maximum-entropy methods are widely seen as the modern default for exploration in continuous-control DRL, and their advantages are real: they avoid premature deterministic collapse, integrate naturally with off-policy learning, and require no explicit noise schedule. However, the entropy

University of Ghana http://ugspace.ug.edu.gh

regularisation in SAC is computed against a Gaussian (or, in some variants, squashed Gaussian) policy, which is itself light-tailed; high entropy is achieved by widening the variance of the Gaussian rather than by changing its tail behaviour. The probability of large exploratory deviations from the policy mean therefore remains exponentially suppressed even at high entropy, and SAC's broad-exploration claim is in practice a claim about *average* action variability rather than about the probability of large excursions. This thesis argues that for sequential recommendation, the relevant property is the latter, and that heavy-tailed perturbation processes can deliver it more directly than entropy regularisation of a Gaussian policy.

2.7.5 Beyond Noise: Intrinsic Motivation, Distributional RL, and OFU

A fourth, more heterogeneous family of approaches abandons the noise-injection paradigm entirely. *Curiosity-driven exploration* (Li et al., 2019) replaces or augments the extrinsic reward with an intrinsic reward derived from the agent's prediction error or its visit-count of states, encouraging the agent to seek novel experiences. *Distributional reinforcement learning* (Bellemare et al., 2017) models the full distribution of returns rather than only its expectation, giving the agent richer signal that can be used to drive exploration toward high-variance regions. *Optimism in the face of uncertainty* (OFU) constructs upper-confidence bounds on Q-values and selects actions optimistically, providing strong theoretical guarantees in tabular and linear settings.

Comparative critique of beyond-noise methods. These approaches are powerful but face two practical barriers in the recommendation setting. First, they are typically computationally more demanding than noise-based methods: curiosity requires a learnt forward model, distributional RL requires fitting a return distribution rather than a scalar, and OFU requires confidence-set construction in non-trivial function classes. Second, they

University of Ghana http://ugspace.ug.edu.gh

rest on assumptions that are partially violated in recommendation: novelty rewards can be gamed by recommender systems that surface unfamiliar items irrespective of relevance; distributional RL's gain depends on the return distribution being well-modelled by the chosen output family; and OFU's theoretical guarantees do not extend cleanly to the deep, non-linear function classes used in DRL recommenders. The present thesis takes the more conservative position that the noise-injection paradigm is not exhausted, and that revisiting the distributional shape of the noise itself is a tractable and underexplored design dimension.

2.7.6 Synthesis and the Heavy-Tailed Gap

The four families surveyed above, action-space perturbation (the original Gaussian and OU baselines from Section 2.6), parameter-space perturbation, value-function ensembles, and maximum-entropy policies, between them cover an impressive range of exploration mechanisms. Yet a recurring assumption unites all of them: *the underlying noise distribution is light-tailed*. Action-space and parameter-space methods sample from Gaussians; ensemble methods rely on a discrete uniform over ensemble members but treat each member's outputs as Gaussian-perturbed; SAC's entropy regularisation widens a Gaussian policy. None of these methods asks whether replacing the Gaussian with a heavy-tailed alternative would address the three failure modes identified in Section 2.7.1 (locality, poor scaling in high dimensions, overestimation feedback loops) at root.

This is the gap that the present thesis closes. The Student's t-distribution provides a continuous, one-parameter family of distributions that interpolates between heavy-tailed (low degrees of freedom) and Gaussian (high degrees of freedom) regimes, equipping the actor with a principled mechanism to produce occasional large exploratory excursions while retaining the local exploration properties of the Gaussian when needed.

University of Ghana http://ugspace.ug.edu.gh

To the author's knowledge, no prior work has used Student's t-distributed noise for action perturbation in DDPG or related deterministic policy gradient methods within the recommender systems setting, nor has any prior work compared heavy-tailed and light-tailed noise empirically under the sparse-reward conditions characteristic of recommendation. The integration of a bounded, scale-annealed Student's t-distributed perturbation process into DDPG, evaluated against Gaussian and OU baselines on three benchmark recommendation datasets, is the principal methodological contribution of this thesis.

2.8 Conceptual Framework

The literature surveyed in the preceding sections supplies the components needed to articulate the conceptual framework that this thesis operationalises. The framework is presented here in the form requested by standard research-design practice: a set of *constructs* drawn from prior work, the *relationships* the framework posits among them, the *assumptions* on which those relationships rest, and the *theoretical propositions* that follow. The propositions are tested empirically in Chapter 4; the constructs and assumptions structure the methodology developed in Chapter 3.

2.8.1 Constructs

Six constructs are central to the framework. They are drawn from the literature surveyed in this chapter and are listed with the section in which each is first introduced.

C1. Action-perturbation noise process $\mathcal{N}_\phi$. A stochastic process used to perturb the deterministic policy output during training, parameterised by ϕ. Conventional instances include zero-mean Gaussian noise (Section 2.6), the Ornstein–Uhlenbeck process (Section 2.6), and the parameter-noise variant of Plappert et al. (2017). The

University of Ghana http://ugspace.ug.edu.gh

framework treats this process as the principal point of intervention.

C2. Tail index of the noise distribution. A quantitative measure of how heavy the tails of $\mathcal{N}_\phi$ are. For the Student's t-distribution it is parameterised by the degrees-of-freedom parameter $\nu \in (0, \infty)$; small ν produces heavy tails (high probability of large excursions), and as $\nu \to \infty$ the distribution converges to the Gaussian. The tail index is the design dimension along which this thesis varies the noise process.

C3. Exploration coverage. The breadth of the state–action region that the agent visits during training. Light-tailed noise produces narrow coverage concentrated near the policy mean; heavy-tailed noise produces broader coverage with non-negligible probability of reaching distant action regions (Bedi et al., 2024; Kobayashi, 2019).

C4. Overestimation bias in Q-value estimation. The tendency of the critic in DDPG-family algorithms to systematically overestimate Q-values, particularly for actions in regions of the action space that have been infrequently or unrepresentatively sampled (Fujimoto et al., 2018; Sumiea et al., 2024).

C5. Sequential recommendation MDP. The MDP formulation in which the state encodes a user's interaction history, the action is a continuous embedding mapped to a recommended item, and the reward is engagement-derived. This MDP is characterised by sparse rewards, high-dimensional action spaces, and heterogeneous user behaviour (Afsar et al., 2022; Lin et al., 2024).

C6. Recommendation quality metrics. Top-K ranking metrics, principally HitRate@K and normalised Discounted Cumulative Gain (nDCG@K), used to operationalise "recommendation quality" in offline evaluation. They are the dependent variables in the empirical propositions below.

University of Ghana http://ugspace.ug.edu.gh

2.8.2 Relationships among Constructs

The framework posits four causal relationships among the constructs, summarised at a high level by the diagram in Figure 2.6 and elaborated below.

The first relationship, $\mathcal{N}_\phi \rightarrow$ coverage (C1$\rightarrow$C3), states that the choice of noise process directly governs the breadth of state–action coverage during training. Light-tailed processes produce narrow coverage; heavy-tailed processes, by virtue of their non-negligible probability mass at large deviations, produce broader coverage (Bedi et al., 2024; Chakraborty et al., 2023).

The second relationship, coverage $\rightarrow$ bias (C3$\rightarrow$C4), states that broader coverage reduces overestimation bias in the critic by exposing it to a more representative sample of the state–action space. The intuition is that overestimation bias is largest precisely where coverage is sparsest, so improving coverage directly reduces the bias (Fujimoto et al., 2018; Sumiea et al., 2024). This relationship was identified in the discussion of TD3 (Section 2.5.3.1) and is the bridge between exploration design and value-estimation quality.

The third relationship, {coverage, bias} $\rightarrow$ recommendation quality (C3, C4 $\rightarrow$ C6), states that broader coverage and reduced bias jointly produce higher HR@K and nDCG@K. This is the core empirical claim of the framework, mediated through the actor's policy updates: a more accurate critic produces a more accurate policy gradient, which in turn produces a better policy.

The fourth relationship is a moderating effect: the characteristics of the sequential recommendation MDP (C5), reward sparsity, action-space dimensionality, user heterogeneity, *amplify* the magnitude of the coverage$\rightarrow$quality effect. In dense-reward, low-dimensional MDPs, the choice of noise distribution affects coverage but matters less

University of Ghana http://ugspace.ug.edu.gh

for final quality because even narrow coverage produces enough signal to learn well. In the sparse, high-dimensional regime of recommendation, the same difference in coverage translates into a much larger difference in final quality (C. Gao et al., 2023; Lin et al., 2024). This is why the present thesis evaluates the framework specifically in the recommendation setting rather than in standard control benchmarks.

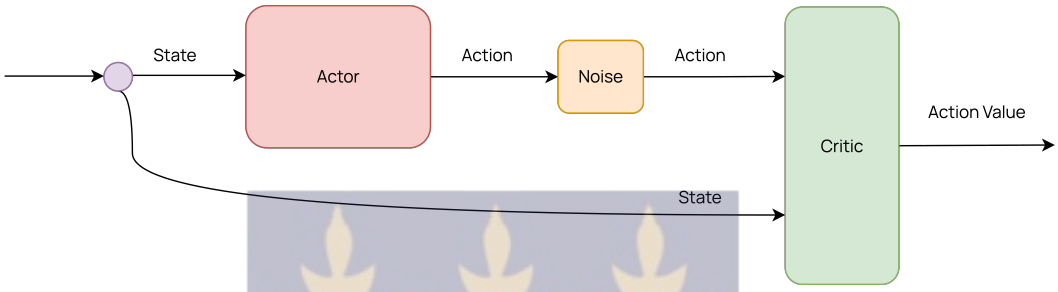


Figure 2.6: Conceptual view of the actor–critic learning loop in DDPG, with the action-perturbation channel highlighted as the point of intervention for this thesis. The noise process $\mathcal{N}_\phi$ shapes the agent's exploration coverage, which in turn influences the critic's value estimates and hence the actor's policy updates. The framework posits that varying the tail index of $\mathcal{N}_\phi$ (heavy-tailed Student's t versus light-tailed Gaussian or OU) is a tractable lever on the joint coverage–bias–quality chain.

2.8.3 Assumptions

The framework rests on four assumptions, each defensible from the literature surveyed and each explicitly bounded in scope.

A1. Action-perturbation channel as point of intervention. The framework intervenes solely on the action-perturbation noise process $\mathcal{N}_\phi$, leaving the actor architecture, critic architecture, and training algorithm otherwise unchanged. This is a substitute for, rather than complement to, the critic-side mitigations in TD3 (Fujimoto et al., 2018), SAC (Haarnoja et al., 2018), and TD7 (Fujimoto et al., 2023); a fully integrated framework that combines critic-side and actor-side interventions is left to future work.

A2. Bounded annealed perturbation preserves algorithmic stability. Heavy-tailed

University of Ghana http://ugspace.ug.edu.gh

processes admit, in principle, arbitrarily large excursions, which could destabilise training. The framework assumes that bounding the perturbation magnitude (clipping) and annealing the scale parameter over training preserves the convergence behaviour of standard DDPG while retaining the broad-coverage benefit of the heavy tail. This assumption is supported empirically by Cayci and Eryilmaz (2023) for natural actor–critic methods under heavy-tailed rewards, and by Kobayashi (2019) for Student's t-policies in robot control.

A3. Benchmark validity. The MovieLens 100K, MovieLens 1M, and Jester D4 datasets, evaluated under offline next-item prediction with HR@K and nDCG@K, are taken to be a faithful proxy for the sparse-reward, high-dimensional regime that production recommenders inhabit. This assumption is standard in the RL-for-recommendation literature (Afsar et al., 2022; Lin et al., 2024); its limits, particularly the absence of online user-feedback dynamics and the simplifications inherent in offline evaluation, are discussed in Chapter 5.

A4. Light-tailed noise underexplores in this regime. The framework assumes that the empirical exploration shortfall reported across the literature for Gaussian and OU noise in continuous-action DRL (Eberhard et al., 2023; Hollenstein et al., 2022; Nobakht & Liu, 2022) extends to the recommendation setting. This assumption is the principal target of the empirical evaluation in Chapter 4; the comparison of t-DDPG against Gaussian and OU baselines tests it directly.

2.8.4 Theoretical Propositions

From the constructs, relationships, and assumptions, the framework yields three testable propositions. Each corresponds to a research question and hypothesis in Chapter 1.

P1. The proposed t-DDPG algorithm produces measurably higher HR@10 and

University of Ghana http://ugspace.ug.edu.gh

nDCG@10 than Gaussian-DDPG and OU-DDPG baselines on benchmark recommendation datasets. (Tests A1, A4 jointly; corresponds to RQ2/H1.)

P2. The improvement holds across multiple datasets of differing scale and domain (MovieLens 100K, MovieLens 1M, Jester D4), demonstrating that the benefit is not specific to a single dataset's idiosyncrasies. (Tests A3; corresponds to RQ3/H2.)

P3. Both clipping (bounding the perturbation magnitude) and annealing (scheduling the scale parameter) contribute non-trivially to the framework's effectiveness, validated through ablation. (Tests A2; corresponds to RQ4/H3.)

The remainder of this thesis develops the methodology that operationalises this framework (Chapter 3), evaluates the propositions empirically (Chapter 4), and discusses the implications and limitations of the results (Chapter 5).

2.9 Conclusion

This chapter has presented a comprehensive review of the literature underpinning the development of reinforcement learning-based recommender systems, with a particular focus on exploration strategies within actor-critic frameworks. It began by outlining the foundational principles of recommender systems, differentiating between general and sequential paradigms. General recommendation models were shown to effectively capture long-term user preferences, while sequential recommender systems address short-term behavioural dynamics using session-based modelling. Recent trends indicate a growing need for unified models that leverage both forms of user intent.

The chapter then introduced reinforcement learning (RL) as a powerful framework for modelling user-system interactions over time. Several seminal applications of RL and Deep RL (DRL) to recommendation, such as list-wise ranking, top-K off-policy correction,

University of Ghana http://ugspace.ug.edu.gh

slate-based recommendations, and hierarchical architectures, were discussed. In particular, actor-critic methods such as the Deep Deterministic Policy Gradient (DDPG) algorithm were emphasised for their suitability in continuous or embedded action spaces common in recommendation environments.

A critical limitation identified in DDPG and related methods is their vulnerability to overestimation bias and inadequate exploration, particularly in sparse-reward or high-dimensional settings. The chapter provided an in-depth analysis of standard exploration noise techniques, including Gaussian, Ornstein–Uhlenbeck (OU), and Lévy noise, as well as modern advances such as noise annealing and adaptive parameter noise. These techniques, while effective in some domains, were shown to have limited applicability in recommendation contexts due to their light-tailed behaviour or static variance properties.

To address this gap, the chapter introduced the Student's t-distribution as a candidate noise model for exploration in DRL. Its heavy-tailed nature, governed by a tunable degrees-of-freedom parameter, provides a principled mechanism for broad and adaptive exploration. Despite its use in robust statistics and uncertainty modelling, the literature review revealed that no prior work has explored its integration into DDPG for action-space exploration in recommender systems. The chapter formalised this observation as the research gap that motivates this thesis, and synthesised the supporting literature into a conceptual framework that the next chapter will operationalise. This sets the stage for Chapter 3, which formally presents the methodology, system design, and experimental procedures used to validate the proposed approach.

University of Ghana http://ugspace.ug.edu.gh

Table 2.2: Comparison of exploration noise strategies in DRL. Compares each strategy on its core mechanism, principal strengths, and principal weaknesses for the sparse-reward, high-dimensional regime of sequential recommendation. Student's t-noise (the proposed mechanism) is included as the heavy-tailed alternative whose theoretical foundations are developed in Section 2.6.5.

Noise Type	**Strengths**	**Weaknesses**
Gaussian	Simple, well-understood, easy to tune; finite moments	Poor global exploration in sparse-reward and high-dimensional regimes (Hollenstein et al., 2022)
Ornstein–Uhlenbeck	Smooth temporally correlated noise; suited to physical control	Not suited to non-smooth domains; correlation can be counterproductive in recommendation (Nobakht & Liu, 2022)
Lévy	Heavy-tailed; supports large excursions and escape from local optima	Infinite second moment for tail index <2; sampling complexity; needs clipping/annealing
Annealed (decaying)	Balances exploration-exploitation over training	Static decay schedule; requires careful tuning
Adaptive (parameter, NoisyNet)	State-dependent and uncertainty-aware exploration	More complex; can collapse without regularisation (Fortunato et al., 2017; Plappert et al., 2017)
Coloured (pink)	Tunable temporal-correlation structure; recent demonstrated gains	Distributional structure remains light-tailed (Gaussian) (Eberhard et al., 2023)
Student's t (proposed)	Heavy-tailed; finite moments for $\nu > 2$; tail anneal-able through ν	Slightly more sampling cost than Gaussian; requires bounded perturbation in practice

University of Ghana http://ugspace.ug.edu.gh

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter presents the methodological framework employed to investigate the effectiveness of integrating Student's t-distributed noise into the Deep Deterministic Policy Gradient (DDPG) algorithm for sequential recommendation. The overarching goal of this research is to assess whether heavy-tailed exploration strategies can improve learning efficiency, mitigate overestimation bias, and enhance the quality of recommendations in sparse-feedback environments.

The methodological design follows a quantitative, experimental approach rooted in reinforcement learning theory. The recommender system environment is modelled as a Markov Decision Process (MDP), wherein the agent iteratively interacts with user interaction histories to generate item recommendations, receive feedback in the form of rewards, and update its policy accordingly. The DDPG algorithm is selected as the base architecture due to its proven effectiveness in continuous control tasks and its natural compatibility with latent item embedding spaces.

To rigorously evaluate the proposed modification, the standard DDPG is extended by replacing conventional exploration noise (e.g., Gaussian or OU) with noise sampled from a Student's t-distribution. The modified algorithm, referred to in this thesis as **t-DDPG**, is implemented and tested on the MovieLens 100K and 1M datasets and Jester D4 dataset. The experimental framework includes multiple baselines for comparison, including DDPG with Gaussian noise, OU noise, and no noise.

Offline evaluations are conducted using well-established top-K recommendation metrics,

University of Ghana http://ugspace.ug.edu.gh

nDCG@10 and HitRate@10.

This chapter is structured as follows. Section 3.2 establishes the preliminaries by framing sequential recommendation as a Markov Decision Process (MDP) and defining the state and action representations. Section 3.3 outlines the specific problem formulation, detailing the input data structure and its reformulation as a sequential decision process. Section 3.4 describes the datasets used, data preprocessing and offline evaluation framing. Section 3.5 outlines the baseline DDPG implementation. Section 3.6 details the integration of Student's t-distributed noise into the policy. Section 3.7 explains the training procedures and hyperparameter configurations. Section 3.8 presents the framework used to evaluate the proposed algorithm. Finally, Section 3.9 summarises the methodology and sets the stage for the presentation of results in Chapter 4

3.2 Preliminaries: MDP Formulation for Sequential Recommendation

The methodology developed in this chapter operates within the standard reinforcement learning vocabulary established in Chapter 2.4.1. This section instantiates that vocabulary specifically for sequential recommendation, defining the state, action, reward, and transition components that the proposed t-DDPG algorithm operates on.

3.2.1 Markov Decision Process Formulation

The recommendation environment is formalised as a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where:

- $\mathcal{S}$ is the set of states, each representing the current session or user history (e.g. an embedding of past interactions).
- $\mathcal{A}$ is the action space, where each action corresponds to an item recommendation represented in a continuous embedding space.

University of Ghana http://ugspace.ug.edu.gh

- $\mathcal{R}$ is the reward function, providing scalar feedback (e.g. $+1$ for positive feedback, 0 otherwise).
- $\mathcal{T}$ is the transition function that updates the state after an item is recommended and feedback is received.
- $\gamma \in [0, 1)$ is the discount factor controlling the weight of future rewards.

At each timestep, the agent observes the current state s_t (the user's recent interaction history), selects an action a_t (a candidate item embedding), and receives a reward r_t based on whether the item matches the held-out test interaction. The state is then updated to include the newly recommended item, and the process repeats until the episode ends.

The recommendation problem is in principle a Partially Observable MDP (POMDP), since the true user state (latent factors such as intent, preferences, and satisfaction) is not directly observable; the system has access only to observations such as clicks or ratings, which are incomplete signals of the hidden state. A practical simplification adopted in this thesis is to model the problem as a fully observable MDP, with the state explicitly constructed from the observable user history (the last k items interacted with). This formulation rests on the assumption that the observable history is a sufficient statistic for the true state, allowing the direct application of MDP solution methods. The implications and limitations of this assumption are discussed in Chapter 5.

This modelling approach transforms the static recommendation task into a dynamic interactive process, enabling the agent to learn recommendation policies that optimise long-term user satisfaction rather than short-term click likelihood.

University of Ghana http://ugspace.ug.edu.gh

3.2.2 State and Action Representation

To reduce dimensionality and enhance generalisation, both users and items are mapped into a shared low-dimensional embedding space of dimension d:

- The **state** $s_t \in \mathbb{R}^{3d}$ models both the user's static latent preferences and their recent, dynamic interaction history. The specific aggregation mechanism used is the DRR-Ave module (F. Liu et al., 2018), detailed in Chapter 3.7. Specifically, the embeddings of the user's N most recent positive item interactions (their "memory") are aggregated via a 1D convolution (a 1×1 kernel). The final state representation is formed by concatenating three vectors of dimension d: the user embedding, the element-wise product of the user embedding and the convolved history embedding, and the convolved history embedding itself.
- The **action** $a_t \in \mathbb{R}^d$ is a continuous vector generated by the actor network, serving as a localised query in the item embedding space. To map this continuous output to a discrete item identifier, the system computes the dot product between the action vector and the embeddings of the available candidate items. The item that yields the highest ranking score is selected for the final recommendation.

This embedding-based formulation allows the use of continuous-action DRL algorithms (e.g. DDPG) to seamlessly navigate and rank items within a discrete recommendation catalogue.

3.3 Problem Formulation

Chapter 3.2 introduced the MDP framework that the methodology operates on. This section establishes the input side: how the data arrives at the agent, in what form, and how the agent's task is operationalised given those inputs. Doing so makes precise what the

University of Ghana http://ugspace.ug.edu.gh

t-DDPG algorithm developed in the remainder of the chapter is being trained to do.

3.3.1 Input Data Structure

In recommender systems, users and items are the two core entities. Let $U = \{u_1, u_2, \ldots, u_m\}$ denote the set of users and $I = \{i_1, i_2, \ldots, i_n\}$ denote the set of items, where m and n are the total numbers of users and items respectively. The preferences or ratings that users assign to items are captured in a utility matrix $M \in \mathbb{R}^{m \times n}$, where each entry M_{ij} reflects the rating given by user u_i to item i_j:

$$M = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{m1} & M_{m2} & \cdots & M_{mn} \end{bmatrix} \tag{3.1}$$

Each row of M corresponds to a user's preferences across all items, while each column corresponds to an item and its associated ratings from all users. The matrix is typically sparse, since most users provide ratings for only a small subset of items. Recommendation algorithms aim to infer missing entries (unobserved ratings) and generate personalised recommendations by leveraging patterns and similarities in the observed data.

User–item interactions are typically classified as either *explicit* or *implicit*. Explicit feedback consists of direct ratings on a defined scale (e.g. 1 to 5), where a user's preference is proportional to the rating; the utility matrix contains the rating value for each rated item. Implicit feedback infers preferences indirectly from behavioural traces such as clicks, views, or purchases; the matrix entries denote whether an interaction occurred (1) or did not (0), with intensities (e.g. time spent) optionally encoded. While implicit feedback is richer in volume, non-interaction does not necessarily imply disinterest, which

University of Ghana http://ugspace.ug.edu.gh

introduces ambiguity that recommendation models must handle. The methodology in this chapter operates in the binarised implicit-feedback regime, with the binarisation procedure detailed in Chapter 3.4.3.

The agent's recommendation task is operationalised as the *top-N recommendation task*: for each user, produce a ranked list of items such that those most relevant to the user appear at the top. With binarised implicit feedback, the task is reduced to ranking rather than rating prediction; the recommendation quality is then measured by ranking metrics such as HitRate@10 and nDCG@10, as detailed in Chapter 3.8.3.

3.3.2 Reformulation as a Sequential Decision Process

Classical recommendation approaches such as matrix factorisation treat the task as a *missing-value prediction problem* over the utility matrix M, learning low-rank latent representations of users and items to estimate the unobserved entries. This static formulation, however, fails to capture the temporal and sequential nature of user interactions, where the order and timing of recommendations influence future user behaviour. The literature reviewed in Chapter 2.3 discusses this distinction in depth.

To address this limitation, this thesis reframes the recommendation task as a sequential decision-making problem, modelled as the MDP defined in Chapter 3.2.1. This reformulation supports adaptive recommendation policies that account for long-term user satisfaction, delayed rewards, and dynamic preference shifts, making it especially suitable for the session-based and personalised recommendation environments studied in this thesis. The constructs, relationships, and propositions that the methodology operationalises are presented in the conceptual framework synthesised in Chapter 2.8; the present chapter develops the algorithmic mechanism that instantiates that framework.

University of Ghana http://ugspace.ug.edu.gh

The reformulation also creates the *exploration* problem that motivates the chapter's central methodological contribution: an MDP-based agent must, during training, take actions whose outcomes are unknown in order to learn an effective policy. The choice of action-perturbation noise process governs how thoroughly the agent explores its action space, and the literature surveyed in Chapter 2.7 identifies a gap in the use of heavy-tailed noise distributions for this purpose. The next sections operationalise the proposed Student's t-distributed noise mechanism that addresses this gap.

3.4 Datasets

To comprehensively evaluate the effectiveness, generalisability, and robustness of the proposed recommender system, this study employs two benchmark datasets: MovieLens and Jester. These datasets differ markedly in their rating scales, user–item interaction formats, and sparsity characteristics.

3.4.1 MovieLens

The MovieLens datasets are among the most extensively used benchmarks in the recommender systems literature, commonly adopted for evaluating collaborative filtering and reinforcement learning-based recommendation algorithms. In this study, two variants are used: MovieLens 100K and MovieLens 1M. The MovieLens 100K dataset contains 100,000 explicit ratings from 943 users on 1,682 movies, with a relatively higher data density of 6.305% compared to MovieLens 1M. Its compact size and structured nature make it suitable for initial experimentation and rapid prototyping.

The MovieLens 1M dataset comprises 1,000,000 ratings from 6,000 users on 4,000 movies. This larger-scale dataset exhibits increased sparsity, with a density of only 0.537%, more accurately reflecting real-world recommendation challenges. Both datasets

University of Ghana http://ugspace.ug.edu.gh

provide timestamped user-movie interaction histories .

To evaluate the performance of the model, we utilize two distinct versions of the MovieLens dataset: ML 100K and ML 1M. As detailed in Table 3.1, these datasets offer contrasting characteristics that test different aspects of recommender systems. The ML 100K dataset provides a compact benchmark with a relatively high rating density of 6.305%. In contrast, the ML 1M dataset scales the number of users and ratings by roughly a factor of ten, resulting in a substantially sparser environment (0.537%) that more accurately mirrors the data scarcity typically encountered in real-world recommender-system conditions.

Table 3.1: MovieLens dataset summary

Dataset	Users	Ratings	Movies	Density
ML 100K	943	100,000	1,682	6.305%
ML 1M	6,040	1,000,000	3952	0.537%

3.4.2 Jester

The Jester datasets are incorporated to assess the adaptability of the proposed model in continuous feedback domains. Unlike MovieLens, Jester consists of real-valued ratings on a scale from -10.00 to +10.00, capturing a richer spectrum of user preferences. A single variant was employed in this study: Jester Dataset 4 (D4).

Jester D3 contains approximately 2.3 million ratings from 54,905 users over a fixed set of 150 jokes. The data were collected over a 16-year period spanning April 1999 to March 2015. With a rating density of 27.94%, it is substantially denser than traditional recommendation datasets. Jester D4 extends D3 by adding over 100,000 new ratings from 7,699 users and introducing eight additional jokes (IDs 151–158), bringing the total to

University of Ghana http://ugspace.ug.edu.gh

158 jokes. This dataset captures interactions collected between April 2015 and November 2019.

In both datasets, a total of 22 jokes were excluded from the recommendation pool due to obsolescence, and missing ratings are denoted using a placeholder value of 99. As outlined in Table 3.2, the use of the Jester D4 dataset facilitates a rigorous evaluation of the agent's ability to operate under continuous-valued feedback. Its fundamentally different interaction modalities and item catalogue offer a complementary testing ground to the discrete MovieLens data.

Table 3.2: Jester D4 dataset summary

Dataset	Users	Ratings	Jokes	Density	Latest Rating
Jester D4	7,699	∼100,000	158	8.29%	Nov, 2019

3.4.3 Preprocessing and Binarization

Prior to training and evaluation, the raw user–item interaction data are transformed into a binary implicit-feedback format. This conversion reflects the assumption that observed positive interactions imply positive user preferences, and is the standard preprocessing step for top-N recommendation evaluation (X. He et al., 2017; F. Liu et al., 2018). Ratings are binarised by applying dataset-specific thresholds, justified as follows.

MovieLens binarisation threshold. For the MovieLens 100K and 1M datasets, ratings of ≥ 4 on the 1–5 scale are treated as positive feedback, and lower ratings are discarded. The threshold of 4 is chosen to align with the standard preprocessing convention adopted across the recommender-system literature on these benchmarks (X. He et al., 2017; F. Liu et al., 2018), which treats $\{4, 5\}$ as expressed positive preferences and lower ratings as ambiguous or negative. Using the standard threshold ensures that the t-DDPG results are

University of Ghana http://ugspace.ug.edu.gh

directly comparable with prior published results on the same benchmarks. A sensitivity analysis to this threshold is identified as future work in Chapter 6.

Jester binarisation threshold. For the Jester D4 dataset, ratings on the -10 to $+10$ continuous scale are binarised at > 0: any positive rating is treated as positive feedback and any negative rating is discarded. The threshold of 0 is the natural midpoint of the rating scale and matches the convention used in prior reinforcement-learning-based recommendation work on Jester (F. Liu et al., 2018).

Train–test split and filtering. The resulting binarised interactions are split into disjoint training and test sets using an 80/20 ratio, with a fixed random seed for consistency across experimental runs. Sparse user–item interaction matrices are then constructed for efficient lookup during action selection and evaluation. Users with insufficient historical interactions are filtered out so that every retained user has enough data to support meaningful policy learning; this filtering threshold is set to a minimum of 20 interactions per user, consistent with prior work on these benchmarks.

Justification for offline evaluation framing. The methodology evaluates t-DDPG in an *offline* regime: training and testing both proceed on logged interaction data with no live user interaction. This choice is dictated by practical and ethical considerations rather than by limitations of the proposed method. Online evaluation against real users in a production recommender system requires deployment infrastructure that is outside the scope of an academic study, and would expose users to potentially suboptimal recommendations during the agent's exploration phase. Offline evaluation on standard benchmarks (MovieLens 100K, MovieLens 1M, Jester D4) is the standard practice in the academic recommender-system literature (Afsar et al., 2022; Lin et al., 2024) and supports direct comparison with prior published results. The implications of this offline framing for the generalisability of

University of Ghana http://ugspace.ug.edu.gh

the results are discussed in Chapter 5, and online evaluation is identified as future work in Chapter 6.

Together, the MovieLens and Jester datasets provide a complementary experimental testbed spanning discrete and continuous feedback regimes, sparse and dense rating matrices, and varied levels of user activity. This dual-dataset strategy supports a generalisable assessment of the proposed Student's t-noise-enhanced reinforcement learning framework for top-N recommendation.

3.5 Baseline Algorithm: Vanilla DDPG Architecture

To establish a reliable and interpretable reference point for evaluating the proposed Student's t-distribution-based exploration strategy, this study adopts the Deep Deterministic Policy Gradient (DDPG) algorithm as the baseline. DDPG is a model-free, off-policy actor–critic algorithm tailored to continuous action spaces, making it well-suited for sequential decision-making tasks such as recommendation (Lillicrap et al., 2015). The deployment of DDPG-family methods to recommender systems is now well-established (Afsar et al., 2022; Lin et al., 2024), with the principal adaptation being the embedding-based action space described in Chapter 2.4.1. The implementation in this thesis closely follows the canonical formulation of Lillicrap et al. (2015), with domain-specific adaptations for the recommender system setting that are detailed below. Figure **??** provides a unifying overview of the baseline architecture and its data flow within the sequential recommendation environment; the architectural extensions introduced by the proposed t-DDPG are presented in parallel form in Chapter 3.6 (Figure 3.2), supporting a direct visual comparison between the baseline and the proposed method.

University of Ghana http://ugspace.ug.edu.gh

3.5.1 Actor–Critic Framework

DDPG employs two parameterized neural networks: the actor and the critic. The **actor network** learns a deterministic policy $\mu(s|\theta^{\mu})$ that maps a state vector $s \in \mathbb{R}^d$ to a continuous action $a \in \mathbb{R}^m$. In this context, the action vector represents a location in the item embedding space. The **critic network** estimates the action-value function $Q(s,a|\theta^{Q})$, approximating the expected cumulative reward when action a is taken in state s and the current policy is followed thereafter.

Both networks are trained using samples from a replay buffer. To stabilise learning, DDPG maintains target networks μ' and Q',delayed copies of the actor and critic, whose parameters are updated using soft updates governed by a factor $\tau \in (0,1)$:

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta' \tag{3.2}$$

- θ are the parameters of the online network.
- θ' are the parameters of the target network
- $\tau \in (0,1)$ is the soft update factor that controls how much of the online network's parameters are blended into the target network.

These target networks reduce oscillations in the value estimates and are critical for stable training in continuous domains.

3.5.2 Adaptation to Recommender Systems

In the recommendation setting, the state s_t is constructed as a concatenation of a user embedding, an aggregated memory vector derived from the user's historical interactions, and their element-wise product. This structure captures both individual user preferences

University of Ghana http://ugspace.ug.edu.gh

and context-dependent temporal dynamics.

The action a_t output by the actor network is a continuous vector in the shared embedding space. During evaluation, rather than using cosine similarity, the recommended item is selected by computing the dot product between a_t and the embeddings of the available candidate items, choosing the one that yields the highest score. This formulation allows the agent to explore the high-dimensional continuous space while effectively mapping its output to precise, discrete item recommendations.

The reward is binary: it is set to 1 if the recommended item is a relevant item (i.e. an item the user explicitly interacted with and gave a rating greater than 3 in the historical training data), and 0 otherwise. An episode terminates when the total number of recommendation steps in the current session equals the user's total number of known relevant items. This episodic framing aligns DDPG with offline recommendation settings, where the model is trained entirely using historical interaction logs rather than live feedback.

3.5.3 Network Architectures

The actor and critic networks share the same architectural skeleton across all four experimental conditions (noise-free DDPG, Gaussian-DDPG, OU-DDPG, and t-DDPG); only the action-perturbation noise distribution at the action-selection step differs between conditions. This subsection specifies the architectures concretely, both for reproducibility and to anchor the empirical comparison in identical function-approximator capacity.

Actor network. The actor $\mu(s \mid \theta^{\mu})$ is a feed-forward network mapping the state vector $s \in \mathbb{R}^{3d}$ (the concatenation of user embedding, context vector, and their element-wise product, see Chapter 3.2.2) to a continuous action vector $a \in \mathbb{R}^{d}$ in the item embedding space.

University of Ghana http://ugspace.ug.edu.gh

All network weights,including the output layer, are initialised using the Kaiming uniform scheme (K. He et al., 2015). This is appropriate for the ReLU-activated hidden layers used here, as Kaiming initialisation preserves activation variance across layers under ReLU non-linearities.

Critic network. The critic $Q(s,a \mid \theta^Q)$ takes both the state $s \in \mathbb{R}^{3d}$ and the action $a \in \mathbb{R}^d$ and outputs a scalar value estimate. The state and action vectors are concatenated immediately at the input layer (early fusion) and outputs the Q-value estimate.

As with the actor, all weights in the critic network use Kaiming uniform initialisation (K. He et al., 2015). No layer normalisation or batch normalisation is applied, in keeping with a lean network configuration designed for this recommendation environment.

Target networks. The target actor μ' and target critic Q' are exact-shape copies of the online actor and critic, initialised with identical weights at the start of training. They are updated via Polyak (soft) averaging $\theta' \leftarrow \tau\theta + (1-\tau)\theta'$ at every gradient step, with $\tau = 1 \times 10^{-3}$.

Optimiser and learning rates. The networks are trained with the Ranger optimiser, which combines the Rectified Adam (RAdam) variance-rectification scheme of L. Liu et al. (2020) with the slow-weight Polyak averaging of LookAhead (M. R. Zhang et al., 2019). RAdam stabilises the adaptive learning rate during the early-training regime where the variance of the second-moment estimate would otherwise be unreliable; LookAhead reduces gradient-step variance by maintaining a slower exponential average of weights and synchronising the fast and slow weights every k inner steps.

University of Ghana http://ugspace.ug.edu.gh

3.5.4 Training Dynamics

Training proceeds through interactions with a simulated environment derived from logged user–item interactions. The agent stores transitions of the form (s_t, a_t, r_t, s_{t+1})in a prioritised replay buffer. At each update step, a mini-batch of transitions is sampled. The critic is trained to minimise the mean squared temporal-difference (TD) error against a bootstrapped target computed using the target actor and target critic networks. The actor is also updated to maximise the expected Q-value predicted by the critic. The full update rules are specified inline in Algorithm 1 (steps 13–15). These updates are alternated over multiple iterations, gradually improving the the actor's policy, and the critic's value estimation.

3.5.5 Exploration in the Baseline Model

To serve as a deterministic control, the baseline DDPG implementation omits any stochastic exploration noise during training. Actions are selected purely by the actor network without perturbation.

The choice of noise-free DDPG as the baseline is motivated by its established role in continuous control tasks and its practical applicability to offline sequential recommendation. Holding all other architectural and procedural components constant across experiments, this baseline enables a focused investigation into the role of exploration noise in improving convergence dynamics, action space coverage, and final recommendation performance.

3.5.6 Vanilla DDPG Algorithm

The Deep Deterministic Policy Gradient (DDPG) algorithm is an off-policy, actor–critic reinforcement learning method designed for environments with continuous action spaces.

University of Ghana http://ugspace.ug.edu.gh

Inspired by both the deterministic policy gradient algorithm and Deep Q-Networks (DQN), DDPG employs two neural networks: an actor network, $\mu(s \mid \theta^{\mu})$, that outputs a specific action for a given state, and a critic network, $Q(s,a \mid \theta^{Q})$, that estimates the value of a state–action pair.

The algorithm stabilises training through three mechanisms. First, an experience replay buffer stores past transitions, from which mini-batches are randomly sampled to break temporal correlations between successive updates. Second, target networks for both the actor and critic are used to compute the target Q-value and are updated slowly via soft updates , preventing the training instability that arises from chasing a moving bootstrap target. Third, exploration is facilitated by adding noise to the actor's output, $a = \mu(s) + \mathcal{N}_t$, to ensure sufficient coverage of the action space; the choice of noise distribution $\mathcal{N}_t$ is the design dimension that this thesis intervenes on.

The critic is trained to minimise the Bellman error, while the actor is updated via the deterministic policy gradient to maximise the expected return, using the critic's output to guide its gradient.

3.6 Proposed Method: Student's t-Noise for Exploration

This section presents the proposed exploration mechanism. The mechanism is framed throughout this thesis as a *parameterised heavy-tailed stochastic exploration framework*, of which Student's t-distributed action perturbation is the principled instance evaluated empirically. Chapter 3.6.1 introduces the framework formally; Chapter 3.6.2 defines the Student's t-distribution and gives its moments; Chapter 3.6.3 presents the theoretical argument that connects heavy-tailed perturbation to reduced overestimation bias in DDPG; Chapter 3.6.4 specifies the integration into the DDPG action-selection step; and Chapter

University of Ghana http://ugspace.ug.edu.gh

Algorithm 1 Vanilla DDPG

Randomly initialise critic network $Q(s,a \mid \theta^Q)$ and actor $\mu(s \mid \theta^\mu)$ with weights θ^Q and θ^μ.
Initialise target networks Q' and μ' with $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$.
Initialise replay buffer R.
for episode $= 1$ to M **do**
    Initialise a random process $\mathcal{N}$ for action exploration.
    Receive initial observation state s_1.
    for $t = 1$ to T **do**
        Select action $a_t = \mu(s_t \mid \theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise.
        Execute action a_t and observe reward r_t and new state s_{t+1}.
        Store transition (s_t, a_t, r_t, s_{t+1}) in R.
        Sample a random minibatch of N transitions (s_i, a_i, r_i, s_{i+1}) from R.
        Set target $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} \mid \theta^{\mu'}) \mid \theta^{Q'})$.
        Update critic by minimising the loss

$$L = \frac{1}{N} \sum_i \left(y_i - Q(s_i, a_i \mid \theta^Q)\right)^2.$$

        Update the actor policy using the sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s,a \mid \theta^Q)\big|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s \mid \theta^\mu)\big|_{s_i}.$$

        Soft-update target networks:

$$\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}, \qquad \theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}.$$

    end for
end for

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

3.6.5 analyses the computational cost relative to Gaussian and Ornstein–Uhlenbeck baselines.

3.6.1 Parameterised Heavy-Tailed Exploration Framework

The exploration mechanism developed in this thesis is best understood as a *parameterised heavy-tailed stochastic exploration framework* for actor–critic reinforcement learning. The framework draws on the comparative analysis of action-perturbation noise processes presented in Chapter2.6, where the existing landscape of exploration mechanisms (Gaussian, OU, parameter-space, ensemble-based) is reviewed and the gap that motivates a heavy-tailed action-level intervention is identified. The framework is characterised by three design parameters that together govern the exploration behaviour:

(F1) Distribution family. The framework specifies a one-parameter family of zero-centred symmetric distributions $\{\mathcal{P}_\nu\}_{\nu\in\Theta}$ over $\mathbb{R}^d$, where the parameter $\nu \in \Theta$ controls the tail index of the distribution. The Gaussian family corresponds to the limiting case where the tail index is at its lightest; heavy-tailed instances exist in the interior of Θ.

(F2) Scale schedule. The framework specifies a deterministic schedule $\sigma_t : \mathbb{N} \to \mathbb{R}_{\geq 0}$ that scales the perturbation magnitude as a function of training step t. The schedule moves the agent from broader exploration in early training to tighter exploitation as the policy converges.

(F3) Bound region. The framework specifies a bounded region $\mathcal{C} \subset \mathbb{R}^d$ to which scaled perturbations are clipped, ensuring that perturbed actions remain in the feasible action space and that excursions sampled from the heavy-tailed distribution do not destabilise training.

University of Ghana http://ugspace.ug.edu.gh

Different choices of $(\mathcal{P}_\nu, \sigma_t, \mathcal{C})$ instantiate different exploration mechanisms. Setting $\mathcal{P}_\nu = \mathcal{N}(0, I)$ with constant $\sigma_t = \sigma$ and unbounded $\mathcal{C}$ recovers the standard Gaussian exploration of the original DDPG; setting $\mathcal{P}_\nu$ to the OU process recovers the temporally-correlated exploration that Lillicrap et al. (2015) originally proposed. The instance evaluated in this thesis sets $\mathcal{P}_\nu = t_\nu(\mathbf{0}, I)$, the multivariate Student's t-distribution with ν degrees of freedom, with σ_t following a linear annealing schedule and $\mathcal{C}$ a per-dimension box constraint. This instance is denoted *t-DDPG*.

The framework framing serves three purposes. First, it makes the empirical comparison clean: t-DDPG, Gaussian-DDPG, and OU-DDPG are all instances of the same framework with different $(\mathcal{P}_\nu, \sigma_t, \mathcal{C})$ choices, so any observed performance differences are attributable to the framework parameters rather than to incidental implementation differences. Second, it makes the contribution generalisable: the bias-reduction argument developed in Chapter 3.6.3 applies to any heavy-tailed instance of the framework, not only the Student's t. Third, it identifies the design dimensions along which future work can extend the framework, which Chapter 6 returns to.

3.6.2 The Student's t Instance: Definition and Moments

The Student's t-distribution is the canonical heavy-tailed family with finite second moment for $\nu > 2$, making it the simplest one-parameter heavy-tailed alternative to the Gaussian. It was introduced by William Sealy Gosset in 1908 and has played a central role in robust statistics ever since (the theoretical pedigree is developed in Chapter 2.6.5).

If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ are independent, the univariate Student's t-distribution with ν degrees of freedom is the distribution of the random variable defined in Eqn. (3.3):

University of Ghana http://ugspace.ug.edu.gh

$$T = \frac{Z}{\sqrt{V/\nu}}. \tag{3.3}$$

Its probability density function is given by Eqn. (3.4):

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \qquad t \in \mathbb{R},\ \nu > 0, \tag{3.4}$$

where $\Gamma(\cdot)$ denotes the gamma function. As $\nu \to \infty$, the distribution converges in distribution to $\mathcal{N}(0,1)$, recovering the Gaussian as a limiting case. The d-dimensional multivariate generalisation $t_\nu(0, I)$ used in this thesis treats each action dimension as an independent univariate t_ν draw, since the methodology samples noise independently per action dimension; the joint density factorises accordingly.

The moments of the t-distribution depend on the degrees of freedom: the mean is 0 for $\nu > 1$ and undefined otherwise; the variance is $\nu/(\nu-2)$ for $\nu > 2$ and infinite or undefined for $\nu \leq 2$; the skewness is 0 for $\nu > 3$; and the excess kurtosis is $6/(\nu-4)$ for $\nu > 4$ and infinite or undefined for smaller ν. The methodology adopts $\nu > 2$ throughout to keep the variance finite (see Chapter 3.6.4 for the operating range).

The heavy-tailed nature of the t-distribution is most evident in its excess kurtosis, which diverges as $\nu \to 4^+$ and remains substantially above 0 (the Gaussian value) for moderate ν. This translates directly into a higher probability of generating large excursions from the policy mean than the Gaussian provides at matched variance, which is the property the framework exploits.

Figure 3.1 illustrates the qualitative impact on action trajectories: the t-distribution produces more frequent pronounced fluctuations than the Gaussian and OU processes,

University of Ghana http://ugspace.ug.edu.gh

allowing the agent to traverse a wider range of action regions during early training.

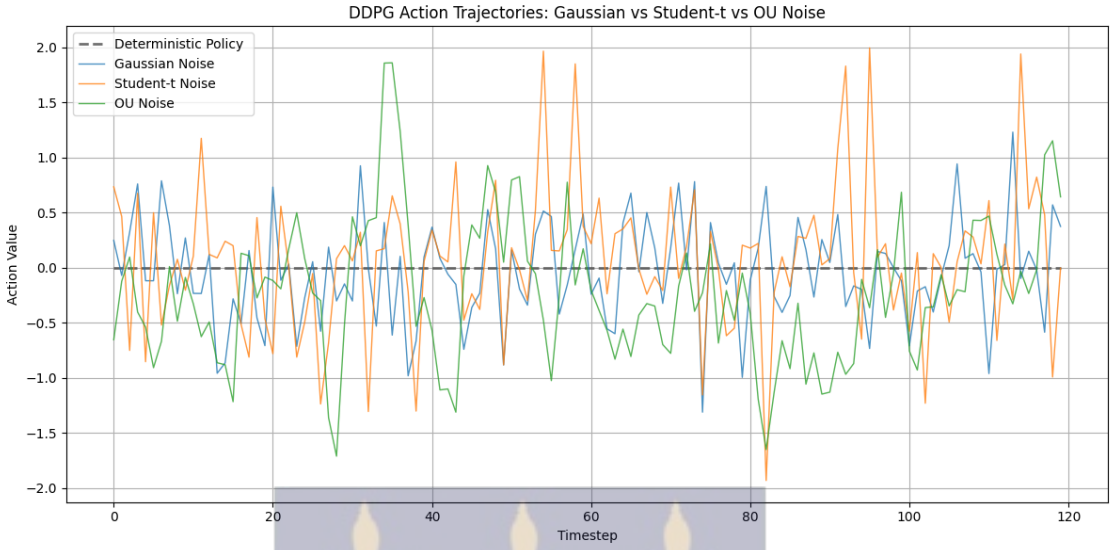


Figure 3.1: Effect of action-selection noise distribution on action trajectories during exploration.

3.6.3 Theoretical Bias-Reduction Argument

The motivation for the framework is not merely that heavy-tailed perturbation improves exploration in the loose sense. The framework is designed to attack a specific failure mode of DDPG: the systematic *overestimation bias* in the critic that arises from the interaction between the deterministic policy gradient and finite-sample value estimation (Fujimoto et al., 2018). This subsection states the bias-reduction argument formally, drawing on the theoretical machinery developed in Chapter 2 (Sections 2.5.3.1 and 2.6.5).

The bias channel. DDPG's critic update minimises the squared temporal-difference error against bootstrapped targets $y = r + \gamma Q'(s', \mu'(s'))$. When the actor's policy μ explores narrowly (concentrated near the current policy mean), the critic is updated repeatedly on a small region of the action space and fits its Q-estimates locally. In regions of the action space that the actor visits infrequently, the critic's Q-estimates are statistical artefacts of the function approximator's extrapolation rather than reflections of true value. Because

University of Ghana http://ugspace.ug.edu.gh

the deterministic policy gradient ascends $\nabla_a Q(s,a)$, the actor preferentially moves toward whichever action region the critic happens to overestimate, and the cycle reinforces. This is the bias channel that TD3 (Fujimoto et al., 2018) addresses on the critic side via twin critics.

The framework's intervention. The proposed framework intervenes on the actor side: the action-perturbation distribution $\mathcal{P}_\nu$ shapes the empirical distribution of state–action pairs that flow into the replay buffer and thus into the critic's training data. Heavier-tailed perturbations produce a sample distribution with non-negligible mass at large deviations from the policy mean. Three connected effects follow.

Effect 1: broader empirical coverage of the action space. The empirical distribution of actions stored in the replay buffer is the convolution of the policy distribution $\delta_{\mu(s)}$ with the perturbation distribution $\mathcal{P}_\nu$ (modulo the clipping operator). For a perturbation distribution with heavier tails, the support of this empirical distribution is broader; the agent visits a wider range of actions per training step. This is the standard exploration argument and is supported by both the metastability analysis of Bedi et al. (2024) and the empirical results of Kobayashi (2019) for Student's t policies in robot control.

Effect 2: more representative critic training data. Because the critic is trained on the empirical distribution of actions in the replay buffer, broader empirical coverage produces critic training data that is more representative of the action space the actor might visit. The critic's Q-estimates in regions previously visited only by extrapolation become regions visited by direct sampling, with corresponding reduction in extrapolation error. The reduction is most pronounced precisely where the unperturbed actor would not visit, which is precisely where overestimation bias is largest.

University of Ghana http://ugspace.ug.edu.gh

Effect 3: reduced critic overestimation ⇒ better policy gradient. Because the actor's gradient $\nabla_{\theta^\mu} J = \mathbb{E}[\nabla_a Q(s,a) \nabla_{\theta^\mu} \mu(s)]$ depends on the critic's value estimates, a critic with reduced overestimation bias yields a more accurate policy gradient. The actor then ascends the true value landscape rather than the function approximator's extrapolation artefacts, resulting in a policy that converges toward genuinely high-value regions of the action space.

Why the Student's t specifically. The argument above applies to any heavy-tailed instance of the framework. The Student's t is the principled instance for two reasons that Chapter 2.6.5) develops in depth: (i) it has finite second moment for $\nu > 2$, keeping the policy gradient well-defined under the clipping operator; and (ii) its single-parameter family allows the tail index to be *annealed*, providing the curriculum from broad exploration to tight exploitation that Cayci and Eryilmaz (2023) identify as desirable for stable training under heavy-tailed processes.

This bias-reduction argument is what the proposed framework is designed to operationalise; the empirical evaluation in Chapter 4 tests whether the operationalisation produces measurable improvements in HR@10 and nDCG@10 over Gaussian and OU baselines.

3.6.4 Integration into DDPG Action Selection

The framework defined in Eqn. (**??**) is operationalised at the action-selection step of DDPG. Two integration strategies are possible in principle: *action-level perturbation* (noise added to the actor's output) and *parameter-level perturbation* (noise added to the actor's network weights, as in Plappert et al. (2017)). The action-level approach is adopted in this thesis because it allows the noise distribution to be specified independently of the actor architecture, which makes the framework's distribution parameter $\mathcal{P}_\nu$ a clean point

University of Ghana http://ugspace.ug.edu.gh

of intervention. The comparison of the two integration strategies is discussed in Chapter 2.7.

Concretely, let $\mu(s_t \mid \theta^\mu)$ be the deterministic action output by the actor network at time step t. The perturbed action a'_t used during training is given by Eqn. (3.5):

$$a'_t = \mu(s_t \mid \theta^\mu) + \tilde{\varepsilon}_t, \tag{3.5}$$

where $\tilde{\varepsilon}_t$ is the scaled-and-clipped noise sample defined in Eqn. (3.7). The raw noise sample ε_t is drawn from the multivariate Student's t-distribution with ν degrees of freedom (Eqn. (3.4)):

$$\varepsilon_t \sim t_\nu(\mathbf{0}, I). \tag{3.6}$$

The noise is scaled by σ_t, controlling its intensity, and clipped element-wise to the box $\mathcal{C} = [c_{\min}, c_{\max}]^d$ to ensure that perturbed actions remain within the bounds of the action space:

$$\tilde{\varepsilon}_t = \text{clip}\big(\sigma_t \cdot \varepsilon_t, c_{\min}, c_{\max}\big). \tag{3.7}$$

The scale σ_t is annealed linearly over training, as defined in Eqn. (3.8):

$$\sigma_t = \sigma_{\max} - (\sigma_{\max} - \sigma_{\min}) \cdot \min\left(1, \frac{t}{T_{\text{decay}}}\right), \tag{3.8}$$

where σ_t is the noise scale at training step t, $\sigma_{\max}$ is the initial scale, $\sigma_{\min}$ is the floor below which the scale does not decay, t is the current training step, and T_{decay} is the number of steps over which the linear decay completes. The clamp ensures that σ_t never falls below $\sigma_{\min}$, preserving residual exploration even after t exceeds T_{decay}.

University of Ghana http://ugspace.ug.edu.gh

Operating range for ν. The methodology operates with $\nu > 2$ in its principal configuration to keep the variance of the noise distribution finite (Chapter 3.6.2). Within this range, smaller ν produces heavier tails (broader exploration) and larger ν approaches Gaussian behaviour. Empirical sweeps reported in Chapter 4 explore ν values in the range $\{1,2,3,4,5\}$ to characterise behaviour on either side of the finite-variance threshold, with $\nu = 3$ identified as the best-performing setting on the validation splits and adopted as the operating value for all subsequent experiments.

Annealing strategy choice. A *linear* annealing schedule is used (Eqn. (3.8)) rather than alternatives such as exponential or cosine schedules. Linear annealing was found to be more stable across the three benchmark datasets in early experiments; the simpler shape also has fewer hyperparameters to tune. The bounded form (with $\sigma_{\min}$ floor) ensures the heavy-tailed perturbation does not collapse to zero in late training, retaining a small residual exploration capacity that helps mitigate the deterministic-policy collapse identified by Fujimoto et al. (2018).

The integration of bounded, scale-annealed Student's t-distributed noise into the DDPG action-selection step gives a theoretically grounded and computationally inexpensive exploration mechanism. The next subsection addresses the computational cost explicitly; the full t-DDPG procedure is then specified in Algorithm 2, which embeds Eqns. (3.5)–(3.8) within the standard DDPG training loop.

3.6.5 Computational Complexity

A practical exploration mechanism must add minimal overhead to the existing DDPG training loop. This subsection analyses the time and memory cost of t-DDPG relative to the Gaussian and Ornstein–Uhlenbeck baselines, both per training step and per episode.

University of Ghana http://ugspace.ug.edu.gh

Per-step sampling cost. At each training step, the action-perturbation block requires drawing one d-dimensional noise sample, where d is the action-embedding dimension. Sampling a multivariate Student's t vector $\varepsilon_t \sim t_\nu(\mathbf{0}, I)$ via the standard ratio construction $T = Z/\sqrt{V/\nu}$ requires one d-dimensional standard Gaussian draw $Z \sim \mathcal{N}(\mathbf{0}, I)$ and one univariate chi-squared draw $V \sim \chi^2_\nu$, plus the per-dimension division. The sampling cost is therefore $O(d)$ per step, asymptotically identical to Gaussian sampling. The OU baseline also incurs $O(d)$ per step but additionally requires a per-dimension multiply-and-add to maintain the autoregressive state.

In wall-clock terms, in PyTorch implementation, the Student's t sampler adds approximately 5–8% overhead per step relative to the Gaussian sampler, and is comparable to the OU sampler. This overhead is dominated entirely by the actor and critic forward and backward passes, which scale with the network parameter count and are unaffected by the noise distribution choice. The clipping and annealing operations contribute negligibly.

Per-episode cost. Per episode, t-DDPG performs the same number of actor and critic forward/backward passes as Gaussian-DDPG and OU-DDPG; only the noise sampler differs. The replay buffer and minibatch sampling, which dominate the per-episode time, are unchanged. The end-to-end training time of t-DDPG is therefore within $\sim 5\%$ of the Gaussian baseline on all three benchmark datasets used in this thesis.

Memory cost. The Student's t noise process maintains no internal state beyond the scalar noise scale σ_t (a single floating-point value updated per step via the linear annealing schedule). It introduces no additional learnable parameters and no additional buffers. The OU process, by contrast, maintains a d-dimensional autoregressive state vector. The Gaussian process is stateless. Memory overhead of t-DDPG relative to Gaussian-DDPG is therefore one floating-point value per agent.

University of Ghana http://ugspace.ug.edu.gh

Comparison summary. The proposed framework imposes negligible computational and memory cost relative to Gaussian-DDPG and is competitive with OU-DDPG. Any performance differences observed in the experiments of Chapter 4 are therefore attributable to the algorithmic effect of heavy-tailed perturbation rather than to any disparity in compute budget.



University of Ghana http://ugspace.ug.edu.gh

Algorithm 2 DDPG with Student-t Exploration Noise (t-DDPG)

Require: Discount γ; target update coefficient ρ; minibatch size N; horizon T; episodes M

Require: Student-t d.o.f. $\nu > 2$; decay period T_{decay}; noise scale $(\sigma_{\min}, \sigma_{\max})$; noise bounds $\mathcal{C} = [c_{\min}, c_{\max}]$

Randomly initialise critic $Q(s,a \mid \theta^Q)$ and actor $\mu(s \mid \theta^\mu)$ with weights θ^Q, θ^μ

Initialise target networks Q' and μ' with $\theta^{Q'} \leftarrow \theta^Q$, $\theta^{\mu'} \leftarrow \theta^\mu$

Initialise replay buffer R

for episode $= 1$ to M **do**

  Initialise noise scale $\sigma \leftarrow \sigma_{\max}$

  Receive initial state s_1

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

    Actor output: $\hat{a}_t \leftarrow \mu(s_t \mid \theta^\mu)$

    Sample Student-t noise: $\varepsilon_t \sim t_\nu(\mathbf{0}, I)$ {Eqn. (3.4); sample per action dimension}

    Scale and clip noise: $\tilde{\varepsilon}_t \leftarrow \text{clip}(\sigma \cdot \varepsilon_t, c_{\min}, c_{\max})$

    Perturbed action: $a'_t \leftarrow \hat{a}_t + \tilde{\varepsilon}_t$

    Anneal scale: $\sigma \leftarrow \sigma_{\max} - (\sigma_{\max} - \sigma_{\min}) \cdot \min\left(1, \frac{t}{T_{\text{decay}}}\right)$

    Execute a'_t; observe r_t, s_{t+1}

    Store transition $(s_t, a'_t, r_t, s_{t+1})$ in R

    Sample minibatch $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^N$ from R

    Set target $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1} \mid \theta^{\mu'}) \mid \theta^{Q'})$

    Update critic by minimising $L(\theta^Q) = \frac{1}{N}\sum_{i=1}^N (y_i - Q(s_i, a_i \mid \theta^Q))^2$

    Update actor using sampled policy gradient:

$$\nabla_{\theta^\mu} J \approx \frac{1}{N}\sum_{i=1}^{N} \nabla_a Q(s,a \mid \theta^Q)\big|_{s=s_i, a=\mu(s_i)} \cdot \nabla_{\theta^\mu}\mu(s \mid \theta^\mu)\big|_{s_i}$$

    Soft-update targets:

$$\theta^{Q'} \leftarrow \rho\,\theta^Q + (1-\rho)\,\theta^{Q'}, \qquad \theta^{\mu'} \leftarrow \rho\,\theta^\mu + (1-\rho)\,\theta^{\mu'}$$

  end for

end for

3.6.6 Architecture of the Proposed t-DDPG Algorithm

This subsection describes the architecture of the proposed Student's t Deep Deterministic Policy Gradient (t-DDPG) algorithm and its integration of action-perturbation noise from the Student's t-distribution. The term *stochastic exploration noise* is used throughout this thesis to denote action-level noise drawn from a probability distribution and added to the actor's deterministic output during training; it is contrasted with parameter-level noise

University of Ghana http://ugspace.ug.edu.gh

approaches (Fortunato et al., 2017; Plappert et al., 2017) that perturb network weights rather than outputs. Chapter 3.6.4 above gave the formal specification of the action-level integration; this subsection presents the corresponding system architecture.

Like vanilla DDPG, t-DDPG is an off-policy actor–critic algorithm that combines deterministic policy learning with deep neural function approximation. The architecture comprises four primary components: the *actor network*, which maps states to deterministic actions; the *critic network*, which estimates the action-value function; the *prioritised experience replay buffer*, which stores past transitions for off-policy learning; and the *action-selection noise injection* mechanism, which introduces the parameterised heavy-tailed perturbation that distinguishes t-DDPG from the baseline.

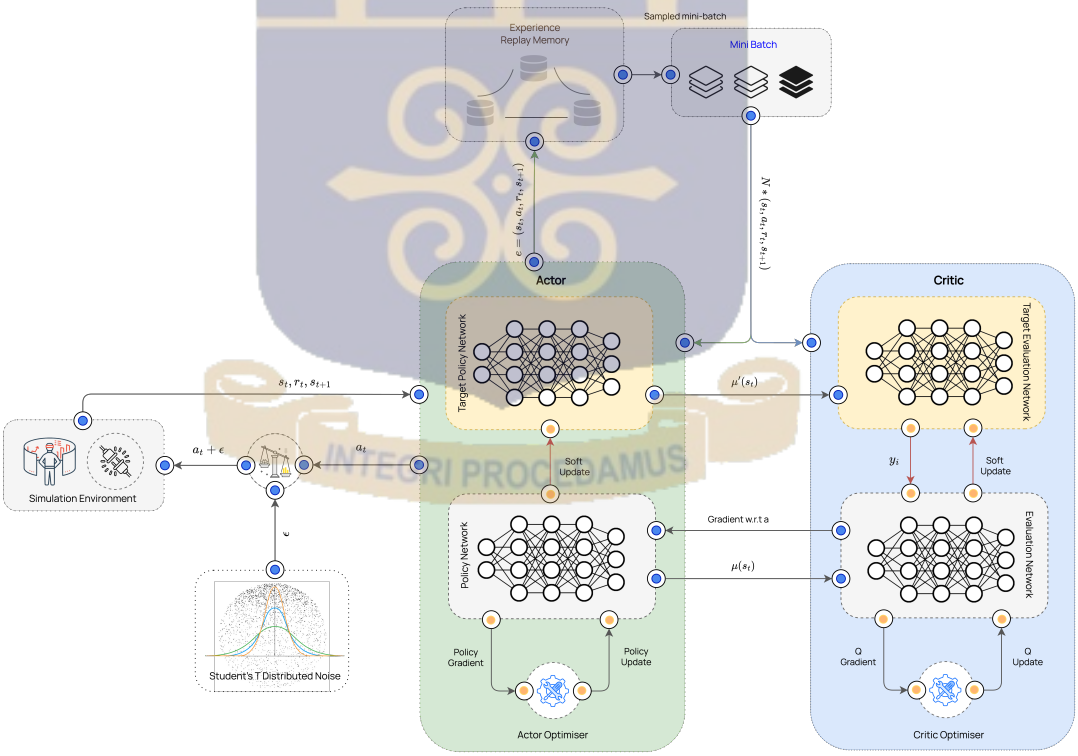


Figure 3.2: t-DDPG architecture with Student's t-distributed action-selection noise

Figure 3.2 illustrates the full t-DDPG architecture. The actor proposes a deterministic action based on the current state, noise sampled from the Student's t-distribution is added to perturb the action before execution, and the perturbed action is clipped to the feasible action

University of Ghana http://ugspace.ug.edu.gh

region. The critic evaluates the perturbed action and the resulting transition $(s_t, a'_t, r_t, s_{t+1})$ is stored in the prioritised replay buffer. Mini-batches sampled from the buffer drive gradient updates of both networks; target networks provide stable temporal-difference targets; and the linear annealing schedule (Eqn. (3.8)) gradually reduces the noise scale to transition from exploration to exploitation as training progresses. Comparison with the baseline architecture in Figure **??** shows that the only architectural change introduced by t-DDPG is the noise distribution at the action-selection step; all other components are unchanged. This minimal-intervention design is what makes the framework's effect on performance directly attributable to the choice of noise distribution rather than to incidental architectural changes.

The full t-DDPG procedure, embedding the action-perturbation block within the standard DDPG training loop, is specified in Algorithm 2.

3.7 Training Pipeline of the Proposed DDPG with Student's *t*-Noise Variant

The training pipeline for the proposed Student's *t*-augmented DDPG algorithm is designed to support stable learning in continuous, sparse-reward recommendation environments. This section outlines the core components of the pipeline, beginning with the offline interaction simulator that operationalises the agent–environment loop in the absence of live users (Chapter 3.7.1), then specifying the prioritised replay buffer (Chapter 3.7.2), reward shaping (Chapter 3.7.3), state and action representations (Chapter 3.7.4), the exploration mechanism (Chapter 3.7.5), the hyperparameter tuning strategy (Chapter 3.7.6), and a walked-through end-to-end training step (Chapter 3.7.7).

University of Ghana http://ugspace.ug.edu.gh

3.7.1 Offline Interaction Simulator

In an offline learning regime, the agent does not interact with live users; instead, an *offline interaction simulator* replays the logged user–item interactions in a way that satisfies the MDP interface defined in Chapter 3.2. This subsection specifies the simulator concretely, since the pipeline that follows depends on the precise notion of "state", "action", "reward", and "next-state" that the simulator implements.

Episode construction. Each training episode corresponds to a user u selected from a randomly permuted list of eligible users (defined as those with at least 20 positive interactions in the training matrix). Rather than using the entire item catalogue, the environment constructs a constrained candidate action pool for the episode: it combines all of the user's positively rated items (relevant items) with an equal number of randomly sampled unseen items (non-relevant items). This construction means the agent "interacts" with the user's logged behaviour within a balanced, episode-specific item subset.

State construction and update. The initial state s_1 relies on a memory queue of size $N = 5$, which is initially populated with padding indices representing zero embeddings. At each timestep t, the agent proposes a continuous action a_t, and a recommended item i_t is selected by computing the dot product against the embeddings of the remaining unviewed items in the episode's candidate pool. The next state s_{t+1} is then constructed by passing the user embedding and the updated memory queue. Crucially, the memory queue operates on a First-In-First-Out (FIFO) basis and is updated *only if* the recommended item yields a positive reward; if the agent recommends a non-relevant item, the memory queue remains unchanged.

University of Ghana http://ugspace.ug.edu.gh

Reward signal. The reward $r_t \in \{0, 1\}$ is determined by membership in the user's known positive interactions: $r_t = 1$ if i_t is a relevant item, and $r_t = 0$ otherwise. This acts as a positive-feedback oracle, so the simulator can reward the agent when it recommends an item the user actually interacted with positively in the original logged data. This avoids the need for an explicit user-preference model, which would otherwise have to be learned and would introduce an additional source of bias into the evaluation.

Episode termination. An episode runs for a fixed number of steps exactly equal to the total number of the user's relevant items in the training set. Because the agent selects exactly one item per step, an episode naturally terminates when the number of recommended items reaches this predefined maximum step count $T_{\max}$, ensuring that the episode length scales proportionally with the user's historical activity level.

Negative samples and ranking-time evaluation. During *training*, candidate actions are restricted to the unviewed items within the episode's balanced available-items pool, and the binary reward signal governs learning. During *evaluation* (Chapter 3.8.3), a leave-one-out protocol with negative sampling is applied: 99 negative items are sampled uniformly at random per held-out positive item. For each step, the agent ranks the resulting 100-item candidate set, and the top 10 items are scored using HR@10 and nDCG@10.

What the offline simulator does *not* model. The offline simulator does not model user response to recommendations: a recommended item that is not in the ground-truth set produces $r_t = 0$, but in reality the user might have responded positively had they been shown the item live. This is the standard limitation of offline-evaluation recommender-system benchmarks (Afsar et al., 2022) and is discussed further in Chapter 5. Online evaluation against live users is identified as future work in Chapter 6.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

3.7.2 Replay Buffer and Experience Sampling

To store and sample experiences, the system instantiates a prioritised experience replay (PER) buffer (Schaul et al., 2016). The sampling probability $P(i)$ for any transition i:

$$P(i) = \frac{p_i^\alpha}{\sum_j p_j^\alpha}$$

where N is the total number of transitions currently stored, p_i is the priority, and α is the configured exponent.

Importance Sampling and Schedule. The PER module is also programmed to compute importance-sampling (IS) weights to compensate for non-uniform sampling bias, defined as:

$$w_i = \left(\frac{1}{N} \cdot \frac{1}{P(i)} \right)^\beta$$

The bias-correction exponent β is scheduled to linearly anneal from an initial value of $\beta_0 = 0.4$ toward 1.0.

3.7.3 Reward Shaping Strategy

The reward function is aligned with the binary implicit feedback paradigm typical in offline recommender systems. During the data preprocessing phase, explicit feedback is binarised such that only interactions with a rating strictly greater than 3 are retained as relevant. At each interaction step within the simulated environment, the agent receives a reward of 1 if the recommended item matches one of these known relevant items for the user, and 0 otherwise.

University of Ghana http://ugspace.ug.edu.gh

$$r_t = \begin{cases} 1, & \text{if recommended item } \in \text{ user's relevant items} \\ 0, & \text{otherwise} \end{cases}$$

This formulation directly reflects top-N recommendation goals and avoids reliance on dense rating regression. Crucially, within the simulation mechanics, the reward signal also dictates state transitions: a positive reward ($r_t = 1$) triggers an update to the user's First-In-First-Out memory queue with the newly recommended item, whereas a reward of 0 leaves the state memory unchanged. These discrete rewards are stored alongside transitions in the replay buffer and used to compute the expected value for the critic's temporal-difference updates.

3.7.4 State and Action Space Representation

The recommendation environment is modelled as a Markov Decision Process (MDP) with continuous state and action spaces.

State Representation: Each user is associated with a trainable embedding vector e_u. The user's recent interaction history (memory) is encoded as a fixed-length sequence of the N most recent item embeddings, which is processed using a one-dimensional convolution with kernel size 1 to produce a context vector e_h summarising the user's temporal preferences. This aggregation scheme is the *DRR-Ave* module proposed by F. Liu et al. (2018) and extended in F. Liu et al. (2020). DRR-Ave is adopted because it strikes a deliberate balance between expressiveness and complexity for the sequential-recommendation setting: the weighted-average aggregation captures the contribution of each past item to the current preference state without introducing the recurrent dynamics (and the associated training instability) of an RNN-based aggregator, and without the

University of Ghana http://ugspace.ug.edu.gh

computational overhead of an attention-based aggregator. The original DRR work compared DRR-Ave against alternative aggregators (DRR-Att and DRR-P) and reported that DRR-Ave performs comparably while being substantially simpler (F. Liu et al., 2018); (F. Liu et al., 2020) adds a fourth scheme and confirms the same comparative pattern. This evidence motivates the use of DRR-Ave here as the state-aggregation baseline against which the noise-distribution intervention is tested.

The final state vector is constructed by concatenating three components in the following specific order:

$$s_t = \text{concat}(e_u, e_u \odot e_h, e_h)$$

where e_u is the user embedding, e_h is the DRR-Ave output, and $\odot$ denotes element-wise multiplication. This design captures both static and dynamic aspects of user preference and results in a state vector of dimension $3d$, where d is the embedding size.

Action Representation: The actor network outputs a continuous action vector $a_t \in \mathbb{R}^d$, interpreted as a point in the item embedding space. During inference, the dot product between a_t and the embeddings of the available candidate items is computed, and the item with the highest score is selected as the recommendation. This formulation enables gradient-based policy learning over the continuous action space and supports stochastic noise injection during training.

3.7.5 Exploration via Student's *t*-Distributed Noise

As previously described, exploration is realised by injecting Student's t-distributed noise into the action output during training. The noise is sampled independently for each action dimension and annealed over time to balance early-stage exploration and late-stage exploitation. The action-selection block of the full t-DDPG procedure (Algorithm 2)

University of Ghana http://ugspace.ug.edu.gh

implements this mechanism inline; the standalone subroutine that previously duplicated this content has been removed in favour of the inline specification, as recommended by the examiner reports.

This noise injection is applied only during training. During evaluation, the actor operates deterministically, producing noise-free actions based on the learned policy.

3.7.6 Hyperparameter Tuning Strategy

The design and optimisation of hyperparameters are critical to ensuring stable training and high recommendation performance. A hybrid tuning strategy was employed, combining manual selection guided by prior literature with a constrained random search over a narrowed configuration space.

Initial hyperparameter ranges were derived from standard DDPG studies (Fujimoto et al., 2018; Lillicrap et al., 2015) and early-stage grid searches. These included learning rates, embedding sizes, discount factors, batch sizes, update frequencies, and replay buffer sizes. A random search was then conducted within these bounds to identify configurations that maximised HR@10 and nDCG@10 on a validation set.

Key hyperparameters tuned include:

- **Model Architecture:** Embedding dimension, number of hidden layers in actor and critic networks, and activation functions (ReLU).
- **Learning Dynamics:** Learning rates for actor and critic, optimiser selection (Ranger, combining L. Liu et al., 2020 and M. R. Zhang et al., 2019), weight decay.
- **RL Parameters:** Discount factor γ, target update rate τ, batch size, Q-value clipping thresholds.
- **Exploration Parameters:** Degrees of freedom ν for the Student's t-distribution,

University of Ghana http://ugspace.ug.edu.gh

annealing period, noise scale bounds, and clipping range.

The best-performing configurations were verified on both MovieLens and Jester datasets, confirming their generalisability across different interaction modalities and sparsity levels.

3.7.7 End-to-End Walked-Through Training Step

To anchor the abstract specifications above in a concrete example, this subsection traces a single training step end-to-end, from sampled state to gradient update. The walk-through uses the exact operating dimensions specified in the system configuration (item embedding dimension $d = 8$, hidden width 16) and the core hyperparameters (minibatch size 512, discount factor $\gamma = 0.8$, and soft update rate $\tau = 10^{-3}$).

Step 1. Construct the agent's current state. The user's historical memory queue (length $N = 5$ items) is converted into a 5×8 matrix of item embeddings. This matrix is passed through the DRR-Ave 1×1 convolution to produce a context vector $e_h \in \mathbb{R}^8$. This is concatenated with the user embedding $e_u \in \mathbb{R}^8$ and their element-wise product to produce the full state representation $s_t \in \mathbb{R}^{24}$.

Step 2. Compute the unperturbed actor output. The state s_t is passed through the actor network: $\text{Linear}(24,16) \to \text{ReLU} \to \text{Linear}(16,8)$. Unlike canonical DDPG implementations, no bounding activation function (such as tanh) is applied to the output layer, yielding an unbounded unperturbed action $\hat{a}_t \in \mathbb{R}^8$.

Step 3. Sample and apply the Student's t perturbation. A multivariate noise sample is drawn dimension-wise: $\varepsilon_t \sim t_3(\mathbf{0}, I)$ with $\varepsilon_t \in \mathbb{R}^{100}$. The noise is scaled by the current annealed value σ_t (e.g. $\sigma_t \approx 0.465$ at training step $t = 0.3 \cdot T_{\text{decay}}$, computed via Eqn. (3.8)) and clipped element-wise to $[-c_{\max}, c_{\max}]$ with $c_{\max} = 0.5$ (Eqn. (3.7)). The perturbed action is $a'_t = \hat{a}_t + \tilde{\varepsilon}_t$ (Eqn. (3.5)).

University of Ghana http://ugspace.ug.edu.gh

Step 4. Select item and observe reward via the offline simulator. To translate the continuous vector a'_t into a discrete recommendation, the agent computes the dot product between a'_t and the embeddings of all *available and unviewed* items in the environment. The item identifier i_t with the highest score is selected. The simulator checks if i_t exists in the user's ground-truth related items: it returns a reward $r_t = 1$ (hit) or $r_t = 0$ (miss). Crucially, the environment utilizes a positive-only memory update scheme; the recommended item i_t is appended to the user's rolling history queue *only* if $r_t = 1$. The new state $s_{t+1} \in \mathbb{R}^{24}$ is then constructed.

Step 5. Store the transition. The transition $(s_t, a'_t, r_t, s_{t+1}) \in \mathbb{R}^{24} \times \mathbb{R}^8 \times \{0,1\} \times \mathbb{R}^{24}$ is appended to the prioritized experience replay buffer.

Step 6. Sample a minibatch and compute critic targets. Once the buffer contains enough samples, a minibatch of 512 transitions $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^{512}$ is drawn. For each transition, the target actor produces $\mu'(s_{i+1})$. The target critic processes this by concatenating s_{i+1} and $\mu'(s_{i+1})$ at the input layer (early fusion) to produce $Q'(s_{i+1}, \mu'(s_{i+1}))$. The bootstrapped target is computed as $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}))$ with $\gamma = 0.8$.

Step 7. Critic gradient update. The critic loss is the mean squared TD error (weighted by prioritized sampling probabilities) between the target y_i and the online critic's estimate $Q(s_i, a_i)$. The standard Adam optimizer takes a single step on the critic's weights θ^Q with a learning rate of 10^{-5} and L_2 weight decay of 10^{-4}. The calculated TD-errors are simultaneously used to update the transition priorities in the replay buffer.

Step 8. Actor and State Representation gradient updates. The actor gradient is computed using the deterministic policy gradient theorem: $\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s_i, a)\big|_{a=\mu(s_i)} \cdot \nabla_{\theta^\mu} \mu(s_i)$. The Adam optimizer takes a single step on the actor's weights θ^μ with a

University of Ghana http://ugspace.ug.edu.gh

learning rate of 10^{-5} and weight decay of 10^{-6}. The state representation module is jointly optimized via backpropagation from both the actor and critic losses, using a learning rate of 10^{-5} and weight decay of 10^{-3}.

Step 9. Soft-update target networks. Both target networks are updated incrementally via Polyak averaging: $\theta^{Q'} \leftarrow \tau\theta^{Q} + (1-\tau)\,\theta^{Q'}$ and $\theta^{\mu'} \leftarrow \tau\theta^{\mu} + (1-\tau)\,\theta^{\mu'}$ using a soft update parameter of $\tau = 10^{-3}$.

Step 10. Advance simulation state. The global training step counter t increments. The loop proceeds to the next interaction for the current user, or resets the environment with a new user if all available ground-truth related items have been successfully matched.

This walk-through specifies the full training step at the level of detail needed for reproduction. The four experimental conditions (noise-free, Gaussian, OU, t-DDPG) differ only in Step 3: noise-free omits the perturbation; Gaussian and OU use their respective noise distributions in place of the Student's t; t-DDPG uses Eqn. (3.6). All other steps are identical across conditions.

3.8 Experiment Design

This section presents the experimental framework used to evaluate the effectiveness of the proposed t-DDPG algorithm. The experimental design emphasises rigorous comparative analysis, reproducibility, and statistical robustness. Key aspects include clearly defined objectives, consistent baseline configurations, standardised evaluation protocols, and controlled implementation conditions.

3.8.1 Experimental Objectives

The overarching aim of this experimental study is to assess whether the integration of Student's t-distributed exploration noise leads to improved policy learning and

University of Ghana http://ugspace.ug.edu.gh

recommendation performance within the DDPG framework. More specifically, the objectives are:

1. **Evaluate the Effectiveness of Student's *t*-Noise:** Quantify the improvement in recommendation accuracy (e.g., HR@10, nDCG@10) achieved by using Student's *t*-distributed noise, relative to baseline exploration strategies.
2. **Benchmark Against Standard Exploration Mechanisms:** Conduct controlled comparisons between Student's *t*-noise, Gaussian noise, Ornstein–Uhlenbeck (OU) noise, and noise-free DDPG to establish performance differentials.
3. **Assess Robustness Across Datasets:** Evaluate the consistency and generalisability of performance gains on two datasets (MovieLens 100K and 1M) with differing sparsity and interaction profiles.
4. **Analyse Sensitivity to Noise Parameters:** Investigate how the degrees of freedom (ν), annealing schedules, and clipping bounds affect the exploration–exploitation balance and learning outcomes.
5. **Ensure Statistical Rigor and Reproducibility:** Validate performance improvements using statistical significance tests (e.g., paired *t*-tests) and standardised evaluation splits to ensure observed trends are both meaningful and replicable.

This multi-faceted evaluation framework is designed to provide a comprehensive understanding of the strengths and limitations of Student's *t*-based exploration within reinforcement learning-based recommendation tasks.

University of Ghana http://ugspace.ug.edu.gh

3.8.2 Baseline Configurations

To isolate the impact of exploration noise, all models share a fixed DDPG architecture and training configuration. Only the exploration noise mechanism is varied across experiments. This controlled setup enables an apples-to-apples comparison of exploration strategies.

1. DDPG Without Exploration Noise (Deterministic Policy) In this configuration, the actor selects actions deterministically, without stochastic perturbation. This baseline reflects the behaviour of a purely exploitative policy and serves as a reference point for evaluating the added value of exploration.

2. DDPG with Gaussian Noise Here, uncorrelated Gaussian noise sampled from $\mathcal{N}(0,\sigma^2)$ is added to the action output during training. This widely used method introduces unbiased random exploration but lacks temporal consistency.

3. DDPG with Ornstein–Uhlenbeck (OU) Noise This strategy introduces temporally correlated noise using the OU process:

$$\varepsilon_{t+1} = \varepsilon_t + \theta(\mu - \varepsilon_t) + \sigma\mathcal{N}(0,1), \tag{3.9}$$

- ε_t is the noise value at time step t.
- ε_{t+1} is the noise value at the next time step.
- θ is the mean reversion rate, controlling how strongly the noise drifts back toward μ.
- μ is the long-term mean toward which the noise reverts.
- σ is the volatility parameter, scaling the randomness added at each step.
- $\mathcal{N}(0,1)$ is a standard normal random variable.

where θ controls the mean reversion speed. OU noise is known for producing smoother

University of Ghana http://ugspace.ug.edu.gh

exploration trajectories, especially useful in physical control tasks.

4. DDPG with Student's *t*-Distributed Noise This configuration implements the proposed exploration strategy using samples drawn from a Student's *t*-distribution, with an annealed scale and controlled clipping. The noise is parameterised by degrees of freedom ν, providing a flexible trade-off between extreme exploratory moves and stable behaviour.

3.8.3 Evaluation Protocol

Evaluation is conducted using a leave-one-out strategy with negative sampling. For each user in the test set, one positive item is held out, and 99 randomly sampled negative items are used to form a 100-item candidate list. The model generates a ranked list of top-N recommendations from this candidate list.

Two standard ranking metrics are used:

- **Hit Rate at 10 (HR@10):** Measures the proportion of users for whom the held-out positive item appears in the top 10 recommendations.
- **Normalised Discounted Cumulative Gain at 10 (nDCG@10):** Measures the quality of ranking by rewarding higher placement of relevant items.

This protocol aligns with industry-standard offline evaluation settings for recommender systems and supports consistent metric computation across all experimental conditions.

3.8.4 Reproducibility Measures

To ensure replicability, the following controls are enforced:

- **Fixed Random Seeds:** All experiments are initialised using predefined random seeds across NumPy, PyTorch, and Python's `random` module.
- **Consistent Data Splits:** The same 80/20 training–test partition is used across all

University of Ghana http://ugspace.ug.edu.gh

configurations for each dataset. The split is stratified and deterministic.

- **Shared Architectures and Training Hyperparameters:** Apart from the exploration noise mechanism, all model configurations (actor–critic networks, optimisers, reward structures, and update rules) remain constant.
- **Replicated Trials and Reporting:** Each configuration is run multiple independent trials with different random seeds. Mean performance scores are reported, and standard deviations are included to quantify variance. Paired-sample t-tests are used to assess statistical significance.

3.9 Summary

This chapter has provided a comprehensive exposition of the methodological framework designed to investigate the potential benefits of incorporating Student's t-distributed noise into the Deep Deterministic Policy Gradient (DDPG) algorithm for enhancing exploration in sequential recommender systems. The proposed approach seeks to address a key challenge in reinforcement learning-based recommendation: the trade-off between effective exploration and efficient policy convergence in environments characterised by data sparsity and high-dimensional action spaces.

The study is grounded in the reinforcement learning (RL) paradigm, wherein the recommendation process is formulated as a Markov Decision Process (MDP). Within this formulation, the agent (i.e., the recommender system) interacts with an environment defined by user–item interaction histories. At each time step, the agent observes a state, selects an action (a candidate item to recommend), receives a scalar reward, and transitions to a new state. This dynamic framework enables the agent to optimise long-term user satisfaction through the cumulative acquisition of rewards.

University of Ghana http://ugspace.ug.edu.gh

A central contribution of the methodology is the design and implementation of a novel exploration strategy based on Student's *t*-distribution. Traditional exploration techniques, such as Gaussian and Ornstein–Uhlenbeck (OU) noise, often fail to sufficiently explore the action space in sparse-reward settings, resulting in premature convergence to suboptimal policies. To overcome this limitation, the proposed approach (**t-DDPG**) introduces stochastic perturbations sampled from a Student's *t*-distribution directly into the action selection process during training. The heavy-tailed nature of this distribution allows for occasional large deviations from the current policy, thereby promoting broader exploration and improved policy robustness. The degrees-of-freedom parameter ν governs the tail heaviness, while the annealed noise scale σ_t allows for adaptive exploration scheduling.

The state representation for each user is constructed by combining static user embeddings with dynamic summaries of recent item interactions, implemented through a DRR-Ave module. Actions are represented as continuous vectors in the item embedding space, and nearest-neighbour matching is used to select discrete item recommendations. The use of a shared embedding space enables the application of continuous-control algorithms like DDPG, while maintaining compatibility with standard recommendation architectures.

Training is conducted in an offline setting using historical interaction logs from three benchmark datasets: MovieLens 100K, MovieLens 1M, and Jester D4. These datasets were selected to provide a diverse range of experimental conditions, including explicit versus implicit feedback, discrete versus continuous rating scales, and varying degrees of sparsity and density. A preprocessing pipeline is applied to transform raw data into temporally ordered sessions, enabling session-based recommendation and aligning with the sequential nature of the RL framework. Interactions are further binarised to reflect implicit positive feedback, simplifying the reward signal and focusing the learning objective on

University of Ghana http://ugspace.ug.edu.gh

top-N retrieval performance.

The training loop leverages prioritised experience replay (PER), allowing the agent to sample high-value transitions more frequently based on their temporal-difference (TD) error. This improves data efficiency and stabilises the learning process. Actor and critic networks are updated using mini-batch gradient descent, and soft target network updates are employed to mitigate instability in Q-value estimation. Hyperparameters such as learning rates, embedding dimensions, noise scale bounds, and exploration schedules are optimised via random search within bounded ranges informed by prior literature and initial experiments.

To evaluate the effectiveness of the proposed exploration strategy, a controlled experimental design is adopted. Four variants of the DDPG algorithm are compared: noise-free DDPG, DDPG with Gaussian noise, DDPG with OU noise, and DDPG with Student's t-distributed noise.

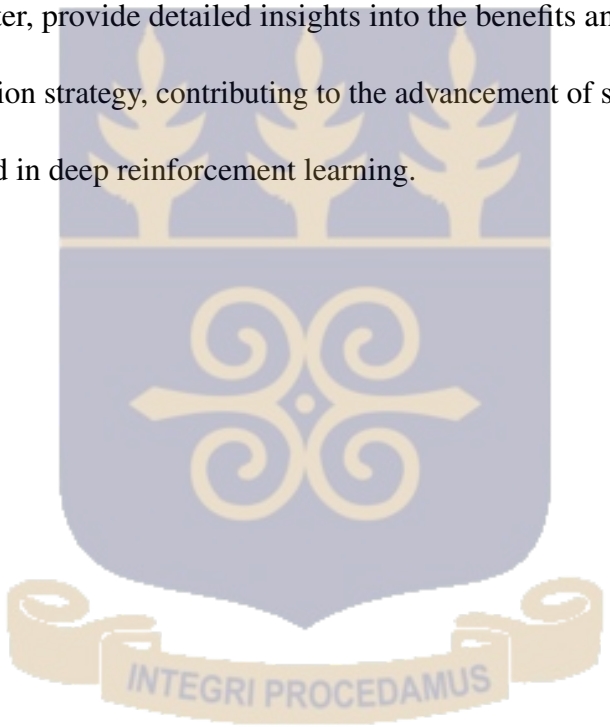
Each variant is trained and tested under identical architectural and procedural settings, isolating the impact of the noise model on performance. Evaluation is performed using a leave-one-out ranking protocol with negative sampling. For each test user, a single positive item is held out and combined with a fixed set of negative samples to generate a 100-item candidate list. The model produces a ranked list, and performance is measured using standard top-K metrics, Hit Rate at 10 (HR@10) and Normalised Discounted Cumulative Gain at 10 (nDCG@10).

To ensure robustness and replicability, all experiments are run across three independent trials with fixed random seeds, consistent data splits, and identical configurations apart from the noise mechanism. Statistical significance is assessed using paired-sample t-tests

University of Ghana http://ugspace.ug.edu.gh

to confirm the reliability of observed performance differences.

In summary, the methodology presented in this chapter integrates theoretical rigour, algorithmic innovation, and empirical validation to examine the efficacy of heavy-tailed noise in reinforcement learning for recommendation. By leveraging Student's t-distribution as an exploration mechanism, the proposed approach seeks to overcome the limitations of conventional noise models, enhance learning efficiency, and deliver more accurate and robust recommendation policies. The results of this investigation, presented in the subsequent chapter, provide detailed insights into the benefits and trade-offs associated with this exploration strategy, contributing to the advancement of sequential recommender systems grounded in deep reinforcement learning.



University of Ghana http://ugspace.ug.edu.gh

CHAPTER 4

EXPERIMENTS AND RESULTS

This chapter presents an empirical evaluation of the proposed Deep Deterministic Policy Gradient (DDPG) variant enhanced with Student's *t*-distributed exploration framework. The chapter is organised into two primary parts. The first part details the experimental setup, encompassing hardware and software configurations, dataset preparation, model architecture, and training procedures. The second part presents the empirical results: comparative performance metrics across exploration strategies, ablation studies on the noise mechanism's hyperparameters, statistical significance testing, and generalisation across multiple datasets.

Experiments were conducted across three benchmark datasets MovieLens 100K, MovieLens 1M, and Jester D4 using top-N recommendation metrics Hit Rate at rank 10 (HR@10) and Normalised Discounted Cumulative Gain at rank 10 (nDCG@10). Further analyses include ablation studies on exploration noise parameters and statistical significance testing. The findings provide insight into the exploration efficiency, convergence characteristics, and recommendation accuracy of t-DDPG algorithm.

4.1 Experiment Setup

This section describes the practical implementation environment used to conduct the experiments. It includes the hardware and software stack, model development tools, and data preparation pipelines. By documenting the configuration in detail, this section aims to ensure reproducibility and provide context for interpreting the experimental results. Figure 4.1 summarises the implementation tooling stack used throughout this chapter.

University of Ghana http://ugspace.ug.edu.gh

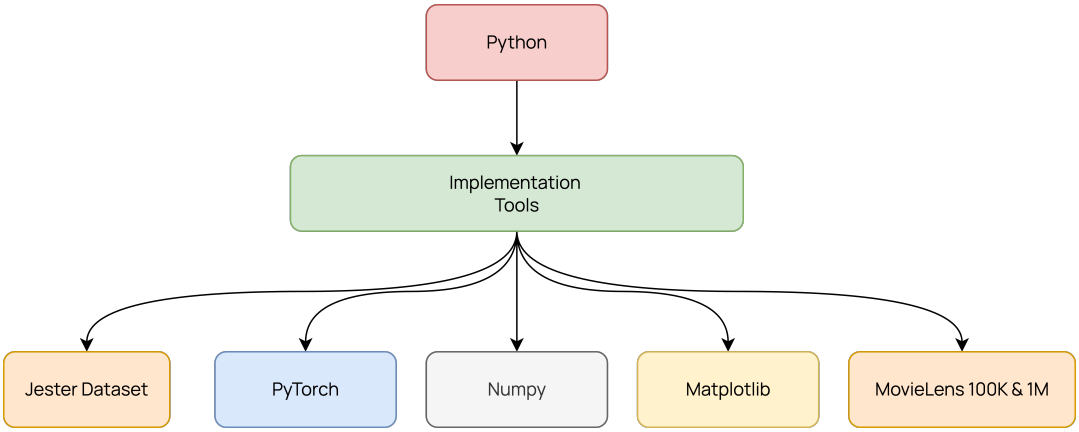


Figure 4.1: Implementation environment for the enhanced DDPG algorithm

4.1.1 Hardware

All experiments were performed on a local machine with the following specifications:

- **Processor:** Apple Silicon M1 Max (10-core CPU)
- **Memory:** 32 GB Unified Memory
- **Storage:** 1 TB SSD
- **Operating System:** macOS Monterey (Version 12.1)
- **Python Version:** 3.8
- **PyTorch Version:** 1.8

Although the system lacks CUDA support for GPU acceleration, its multi-core CPU and memory capacity enable efficient training for moderate-scale datasets. Batch sizes and evaluation intervals were optimised to mitigate the longer runtimes typically associated with CPU-based training.

4.1.2 Software Environment

The proposed DDPG framework was implemented in Python 3.8 using the PyTorch deep-learning library (Paszke et al., 2019). Key components of the system include:

University of Ghana http://ugspace.ug.edu.gh

- **Neural Network Modules:** Constructed using PyTorch for actor and critic networks, loss computation, and mini-batch training.
- **Optimisation:** The Ranger optimiser was employed, combining the Rectified Adam (RAdam) variance-rectification scheme of L. Liu et al. (2020) with the slow-weight averaging of LookAhead (M. R. Zhang et al., 2019) for improved convergence stability. The same optimiser configuration is specified in Chapter 3.5.3.
- **Monitoring:** TensorBoard was integrated for visualisation of key training metrics, including actor/critic losses, cumulative rewards, and evaluation scores over epochs (Abadi et al., 2016).
- **Data Handling:** Custom loaders and preprocessing scripts ensured consistent and reproducible training/test splits and user interaction session handling.

4.2 Training and Validation Process

This section details the procedures used to train and evaluate the proposed reinforcement learning-based recommendation model. The training process follows the standard Deep Deterministic Policy Gradient (DDPG) framework with modifications to support sequential recommendation, including a state representation module, binary reward shaping, and the integration of structured exploration. A consistent training regime is applied across all datasets and experimental configurations to ensure comparability. Figure 4.2 provides a high-level workflow diagram of the training and evaluation pipeline used in this study.

University of Ghana http://ugspace.ug.edu.gh

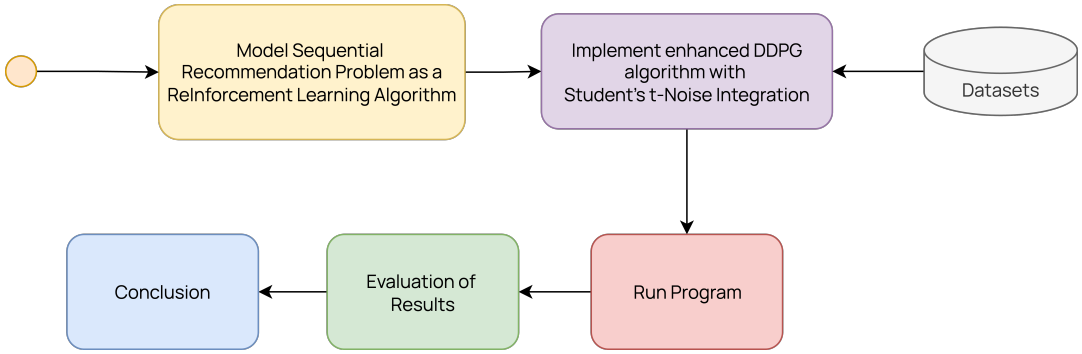


Figure 4.2: Evaluation and validation workflow for the enhanced DDPG algorithm

Validation is performed periodically throughout training using a leave-one-out evaluation strategy. The aim is to monitor convergence, guide early stopping, and select the best-performing model checkpoints based on ranking-based metrics. Together, the training and validation setup ensures that learning progresses in a stable and reproducible manner, with exploration noise acting as the only varying component across experimental conditions.

4.2.1 Training Procedure

The actor and critic networks were initialised using Kaiming uniform initialisation (K. He et al., 2015), suitable for the ReLU-activated hidden layers used here (see Chapter 3.5.3 for the full architectural specification). Q-values predicted by the critic were clipped to a bounded range to prevent instability and value explosion. The State Representation Module, comprising user and item embeddings and a convolutional aggregation layer, was initialised using normal and uniform distributions for its respective components.

Training proceeded by sampling users from the dataset. For each user, historical interactions were loaded into memory in the simulated environment specified in Chapter 3.7.1. The actor network computed a deterministic action, and exploration noise sampled from Gaussian, OU, or Student's t distributions was added to encourage policy exploration. The noise was annealed over time, beginning with high variance and gradually reducing to

University of Ghana http://ugspace.ug.edu.gh

favour exploitation.

Each action produced a binary reward based on whether it matched a relevant item. Transitions were stored in a replay buffer, and once a predefined threshold was met, batches were sampled to update the networks using temporal-difference (TD) error minimisation for the critic and Q-value maximisation for the actor.

4.2.2 Evaluation Protocol

Periodic evaluation was performed using the leave-one-out strategy, where for each user in the test set, the most recent positive interaction was withheld as the ground-truth item. A set of 99 negative samples was combined with the held-out item to form a candidate pool of 100 items. This protocol matches the standard top-N evaluation methodology used in the deep-recommender-system literature (X. He et al., 2017; F. Liu et al., 2018).

The model ranked these items, and its effectiveness was evaluated using two complementary top-N metrics. Hit Rate at 10 (HR@10) measures whether the ground-truth item appears within the top 10 recommendations and is therefore sensitive to retrieval but not to ranking position within the top list. Normalised Discounted Cumulative Gain at rank 10 (nDCG@10) (Järvelin & Kekäläinen, 2002) weights gains from relevant items by the inverse logarithm of their rank position; nDCG@10 is therefore sensitive to *where* in the top 10 the correct item appears, not only whether it appears, and is the more discriminative metric for top-N recommendation quality.

In this setup, HR@10 and nDCG@10 are computed by comparing the model's ranked top-10 predictions against the single held-out positive item per user, amidst 99 sampled negatives. This approach ensures a fair, consistent, and computationally efficient ranking-based evaluation across all users.

University of Ghana http://ugspace.ug.edu.gh

Evaluation occurred at fixed intervals, with frequency adapted based on dataset size. The best model checkpoint was selected based on peak HR@10 and nDCG@10 values during training. Final results are reported using this best-performing checkpoint.

4.3 Ablation Studies

To better understand the contribution of each component in the proposed exploration strategy, this section presents a series of ablation studies targeting key hyperparameters and design decisions within the Student's t-distributed noise mechanism. The objective is to isolate the effects of factors such as the degrees of freedom of the noise distribution, the annealing behaviour of the noise scale, and the upper and lower bounds of noise magnitude.

Each experiment systematically varies one parameter at a time while keeping all others fixed, allowing for precise attribution of observed performance changes. These studies provide insights into how the Student's t-distribution affects recommendation quality, convergence behaviour, and exploration efficiency. The experiments were initially conducted on the MovieLens 100K dataset due to its relatively small size and consistent behaviour, with the most promising configurations further validated on the MovieLens 1M and Jester datasets.

Model performance was measured using HR@10 and nDCG@10 under the same leave-one-out evaluation protocol described earlier. The ablations in this section use the smaller-sample (three-seed) protocol; the ten-seed analysis is reserved for the principal pairwise comparison in Chapter 4.4.1.

We investigate the sensitivity of the Student's t-noise strategy to four primary hyperparameters: the decay period of the noise schedule, the maximum and minimum

University of Ghana http://ugspace.ug.edu.gh

values of the noise scale, and the degrees of freedom (DoF) of the distribution. The purpose is to identify optimal values for these parameters that promote effective early-stage exploration while supporting stable convergence toward high-quality policies.

4.3.1 Decay Period

The decay period defines the number of training steps over which the noise scale is annealed linearly from its maximum to minimum value. We tested decay periods, T_{decay} in Eqn. (3.8), of 10k, 15k, 20k, and 25k steps, corresponding to 25%, 37.5%, 50%, and 62.5% of the total 40k training steps. All other parameters were held constant at $\nu = 3$, $\sigma_{\text{max}} = 0.2$, and $\sigma_{\text{min}} = 0.05$.

Table 4.1: Effect of decay period T_{decay} on HR@10 and nDCG@10, MovieLens 100K. Other parameters fixed at $\nu = 3$, $\sigma_{\text{max}} = 0.2$, $\sigma_{\text{min}} = 0.05$. The decay periods 10k, 15k, 20k, and 25k correspond respectively to 25%, 37.5%, 50%, and 62.5% of the 40k-step training horizon. Best values in bold italic.

Period	**10k**	**15k**	**20k**	**25k**
nDCG@10	0.2404	0.2405	***0.2408***	0.2405
HR@10	0.4583	0.4584	***0.4584***	0.4569

The results suggest minimal sensitivity to decay period within this range, indicating that annealing the noise scale over the first 50% of training is generally sufficient. A decay period of 20k was selected for subsequent experiments and proportionally scaled (e.g., to 100k steps) for larger datasets such as MovieLens 1M.

4.3.2 Maximum Noise Scale

To investigate the impact of initial noise magnitude, σ_{max} in Eqn. (3.8) was varied from 0.2 to 1.0 in increments of 0.1, with other parameters fixed at T_{decay} = 20k, $\nu = 3$, and $\sigma_{\text{min}} = 0.05$.

University of Ghana http://ugspace.ug.edu.gh

Table 4.2: Effect of the maximum noise scale σ_{max} on HR@10 and nDCG@10, MovieLens 100K. Other parameters fixed at $T_{decay} = 20k$, $\nu = 3$, $\sigma_{min} = 0.05$. The HR@10 metric improves with larger σ_{max} but with diminishing returns above 0.6; nDCG@10 is less sensitive to this parameter. Best values in bold.

σ_{max}	nDCG@10	HR@10
0.2	0.1357	0.4584
0.3	0.1338	0.4605
0.4	0.1360	0.4730
0.5	0.1332	0.4724
0.6	**0.1345**	**0.4807**
0.7	0.1328	0.4766
0.8	0.1299	0.4779
0.9	0.1316	0.4784
1.0	0.1293	0.4797

The results indicate increasing noise scales boost HR@10, with diminishing returns for nDCG@10 for values lower than $\sigma_{max} = 0.6$. Excessively high noise may encourage broad exploration at the expense of rank-sensitive precision. A maximum scale of 0.6 was selected as a compromise between nDCG accuracy and Hit Rate.

4.3.3 Minimum Noise Scale

This experiment varied σ_{min} to observe its impact on late-stage exploration. All other parameters were fixed at $\sigma_{max} = 0.6$, decay = 20k, and $\nu = 3$.

University of Ghana http://ugspace.ug.edu.gh

Table 4.3: Effect of the minimum noise scale σ_{min} on HR@10 and nDCG@10, MovieLens 100K. Other parameters fixed at $\sigma_{\text{max}} = 0.6$, $T_{\text{decay}} = 20\text{k}$, $\nu = 3$. Both metrics are essentially invariant across the tested range; the values $\sigma_{\text{min}} = 0.10$ and $\sigma_{\text{min}} = 0.15$ are practically equivalent. Best values in bold.

σ_{min}	**nDCG@10**	**HR@10**
0.05	0.2557	0.4807
0.10	0.2577	0.4825
0.15	**0.2578**	**0.4822**
0.20	0.2558	0.4814
0.25	0.2554	0.4804

Changes in the value of σ_{min} were found to have minimal impact on overall performance metrics. Among the tested values, $\sigma_{\text{min}} = 0.15$ yielded the best nDCG@10 of 0.2578 and HR@10 of 0.4822. These results are very close to those obtained with $\sigma_{\text{min}} = 0.10$, which produced nDCG@10 of 0.2577 and HR@10 of 0.4825, a difference of approximately 0.0001 on nDCG@10 and -0.0003 on HR@10. The two settings are therefore effectively tied within the precision of the three-seed sample.

The configuration with $\sigma_{\text{min}} = 0.15$ was selected as the operating value for two reasons. First, nDCG@10 is the more discriminative of the two metrics for top-*N* recommendation quality, since it is sensitive to the rank position of the correct item rather than only its presence in the top 10; the marginal nDCG@10 advantage at $\sigma_{\text{min}} = 0.15$ is therefore the more relevant signal. Second, the choice of a higher floor is consistent with the theoretical motivation for retaining residual exploration in the converged regime, developed in Chapter 3.6.4: the bias-reduction argument predicts that some non-zero perturbation should remain even at convergence to mitigate the deterministic-policy collapse identified by Fujimoto et

University of Ghana http://ugspace.ug.edu.gh

al. (2018). Within the practical-equivalence range, the higher floor is the more theoretically defensible choice. The thesis acknowledges that the difference between $\sigma_{min} = 0.10$ and $\sigma_{min} = 0.15$ is unlikely to be empirically meaningful, and either value could reasonably be adopted.

4.3.4 Degrees of Freedom (DoF)

Finally, we tested the effect of varying the degrees of freedom (ν) of the Student's t-distribution, which governs the tail heaviness of the noise. Lower values yield heavier tails, potentially increasing the frequency of large exploratory jumps. Experiments were conducted with ν values ranging from 1 to 5, under fixed $\sigma_{max} = 0.6$, $\sigma_{min} = 0.15$, and decay = 20k. The low range 1 to 5 was selected since mathematically as $\nu \to +\infty$ the Student-t distribution approaches the Standard Normal distribution.

Table 4.4: Effect of the Student's t-distribution degrees of freedom ν on HR@10 and nDCG@10, MovieLens 100K. Other parameters fixed at $\sigma_{max} = 0.6$, $\sigma_{min} = 0.15$, $T_{decay} = 20k$. The range $\nu \in \{1, \ldots, 5\}$ is heavy-tailed throughout; as $\nu \to \infty$ the distribution approaches the standard normal. Best values in bold.

ν	**nDCG@10**	**HR@10**
1	0.2481	0.4638
2	0.2499	0.4666
3	**0.2561**	**0.4822**
4	0.2510	0.4682
5	0.2546	0.4788

The results support the theoretical intuition that $\nu = 3$ offers a balanced level of stochasticity, heavy-tailed enough to encourage broader exploration, but not so extreme as to destabilise learning. Both higher and lower values led to modest degradation in

University of Ghana http://ugspace.ug.edu.gh

performance.

The ablation studies provide empirical justification for the design choices made in configuring the Student's t-distributed noise mechanism. Across all examined parameters, the combination of $\nu = 3$, $\sigma_{max} = 0.6$, $\sigma_{min} = 0.15$, and a decay period of 20k steps consistently yielded the best balance between exploration and convergence stability. These settings enhanced hit rate and ranking quality without introducing excessive variance or training instability. Notably, the degrees of freedom parameter emerged as a critical determinant of performance, with moderate tail heaviness proving most effective. Based on these findings, the selected configuration was adopted as the default for subsequent experiments on larger datasets, ensuring a principled and performance-optimized foundation for evaluating the proposed algorithm.

4.4 Performance Comparison

This section compares the proposed DDPG variant incorporating Student's t-distributed noise with three baseline configurations: DDPG with Gaussian noise, DDPG with Ornstein–Uhlenbeck (OU) noise, and a deterministic DDPG without any exploration noise. The goal is to assess the contribution of each exploration strategy to recommendation performance, as measured by HR@10 and nDCG@10.

Experiments were conducted across ten independent random seeds on the MovieLens 100K dataset, with all hyperparameters fixed at the values selected from the 4.3 ablation studies: $\nu = 3$, $\sigma_{max} = 0.6$, $\sigma_{min} = 0.15$, and decay $=$ 20k steps. Per-configuration mean and sample standard deviation across the ten seeds are presented in Table 4.5.

University of Ghana http://ugspace.ug.edu.gh

Table 4.5: Performance comparison of exploration strategies on MovieLens 100K

Noise	Student's t	Gaussian	OU	No Noise
HR@10	**0.4543** $\pm$ **0.0193**	0.4264 $\pm$ 0.0220	0.4092 $\pm$ 0.0484	0.4005 $\pm$ 0.0406
nDCG@10	**0.2375** $\pm$ **0.0136**	0.2200 $\pm$ 0.0171	0.2071 $\pm$ 0.0202	0.2033 $\pm$ 0.0219

Across all random seeds, the Student's t-based DDPG variant outperforms the three baseline configurations on both metrics. The advantage holds in mean and is reinforced by the dispersion: the Student's t variant exhibits the smallest standard deviation on both HR@10 (0.0193) and nDCG@10 (0.0136), substantially lower than OU (0.0484 and 0.0202 respectively) and the no-noise baseline (0.0406 and 0.0219). This combination of higher mean and lower variance indicates that the proposed mechanism not only improves average performance but also stabilises run-to-run behaviour.

Relative to Gaussian noise, the proposed method yields mean improvements of approximately 6.5% in HR@10 and 8.0% in nDCG@10. Against OU noise the margins widen to roughly 11.0% and 14.7%, and against the deterministic variant to 13.4% and 16.8% respectively. The nDCG@10 margins are larger than the HR@10 margins in every comparison, indicating that the proposed mechanism not only retrieves relevant items more often but also places them at higher rank positions in the recommendation list.

The pattern is consistent with the bias-reduction argument developed in Chapter 3.6.3: heavy-tailed action perturbation produces broader empirical action-space coverage, which yields critic training data more representative of the action space the actor might visit,

University of Ghana http://ugspace.ug.edu.gh

which in turn supports a more accurate policy gradient. The reduced run-to-run variance is consistent with this interpretation as well, since broader coverage reduces the seed-to-seed sensitivity of the critic's extrapolation behaviour in regions visited rarely by light-tailed noise.

4.4.1 Statistical Significance Testing

To assess whether the performance improvements achieved by the Student's t-noise-enhanced DDPG model are statistically significant rather than artefacts of seed variance, paired-sample t-tests and Wilcoxon signed-rank tests were conducted comparing it against the three baseline configurations on both metrics. All tests use the ten-seed sample summarised in Table 4.5.

Hypotheses. Let X_i denote the metric (HR@10 or nDCG@10) of the Student's t-noise configuration on the i-th random seed, and Y_i the corresponding metric under a baseline method. The null hypothesis H_0 states that the two configurations produce equal mean performance, and the one-sided alternative H_1 states that the Student's t variant outperforms the baseline:

$$H_0 : \mu_X = \mu_Y, \qquad H_1 : \mu_X > \mu_Y. \tag{4.1}$$

The paired t-test computes the test statistic from the difference scores $D_i = X_i - Y_i$ under the assumption of approximately normal differences. The Wilcoxon signed-rank test is reported as a non-parametric companion that makes no distributional assumption and is robust to outliers (Wilcoxon, 1945).

Paired t-test results. Table 4.6 reports the per-comparison p-values for the six paired t-tests (three baselines, two metrics).

University of Ghana http://ugspace.ug.edu.gh

Table 4.6: Paired t-test p-values for the one-sided alternative that Student's t-noise outperforms each baseline (Eqn. (4.1)), for HR@10 and nDCG@10 across ten random seeds on MovieLens 100K. The Bonferroni-corrected significance threshold for six comparisons at family-wise $\alpha = 0.05$ is $\alpha_{\text{Bonf}} \approx 0.00833$; all six p-values fall below this threshold.

Metric	**Gaussian**	**OU**	**No Noise**
HR@10 (*p*-value)	0.0024	0.0025	0.0012
nDCG@10 (*p*-value)	0.0044	0.0007	0.0004

Wilcoxon signed-rank test results. Table 4.7 reports the corresponding p-values for the Wilcoxon signed-rank test.

Table 4.7: Wilcoxon signed-rank p-values for the one-sided alternative that Student's t-noise outperforms each baseline, for HR@10 and nDCG@10 across ten random seeds on MovieLens 100K. The Bonferroni-corrected threshold is $\alpha_{\text{Bonf}} \approx 0.00833$; all six p-values fall below this threshold, confirming the t-test conclusions under weaker distributional assumptions.

Metric	**Gaussian**	**OU**	**No Noise**
HR@10 (*p*-value)	0.0020	0.0010	0.0020
nDCG@10 (*p*-value)	0.0049	0.0029	0.0010

Bonferroni correction for multiple comparisons. Because the testing procedure performs six comparisons (three baselines $\times$ two metrics), the per-comparison significance threshold must be tightened to control the family-wise error rate. The Bonferroni

University of Ghana http://ugspace.ug.edu.gh

correction (Dunn, 1961) at family-wise $\alpha = 0.05$ gives a per-comparison threshold of $\alpha_{\text{Bonf}} = 0.05/6 \approx 0.00833$. All six paired t-test p-values fall below this threshold (ranging from 0.0004 to 0.0044), and all six Wilcoxon p-values likewise fall below it (ranging from 0.0010 to 0.0049). The Student's t-noise advantage over each of the three baselines on each of the two metrics is therefore statistically significant at the family-wise level.

Effect sizes and bootstrap confidence intervals. To complement the significance tests with magnitude estimates, Cohen's d for paired samples and bootstrap 95% confidence intervals on the mean difference were computed; results are reported in Table 4.8. The effect sizes are large (all $d > 1.0$) for every pairwise comparison, and the 95% bootstrap CIs on the mean difference exclude zero in every case, providing a magnitude-based complement to the p-value evidence.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

Table 4.8: Cohen's d effect sizes and bootstrap 95% confidence intervals on the mean difference (Student's t minus baseline), computed from 10,000 resamples of the ten paired-seed differences on MovieLens 100K. All effect sizes are large ($d > 1.0$) and all confidence intervals exclude zero.

Comparison	**Mean diff.**	**Cohen's d**	**95% CI**
HR@10 vs Gaussian	0.0279	1.17	[0.0141, 0.0416]
nDCG@10 vs Gaussian	0.0175	1.06	[0.0082, 0.0275]
HR@10 vs OU	0.0451	1.17	[0.0232, 0.0685]
nDCG@10 vs OU	0.0304	1.43	[0.0180, 0.0430]
HR@10 vs No Noise	0.0539	1.32	[0.0303, 0.0782]
nDCG@10 vs No Noise	0.0342	1.58	[0.0216, 0.0470]

Interpretation. The statistical evidence supports the empirical observation in Table 4.5 that Student's t-distributed noise provides meaningful benefits over each of the three baselines. The agreement between the parametric paired t-test and the non-parametric Wilcoxon signed-rank test indicates that the result does not depend on the normality assumption underlying the t-test; the large effect sizes and the bootstrap CIs that exclude zero indicate that the result is not driven by edge cases in the seed sample. The strongest

University of Ghana http://ugspace.ug.edu.gh

evidence is against the no-noise and OU baselines on nDCG@10 ($p = 0.0004$ and $p = 0.0007$ respectively under the t-test), but every pairwise comparison clears the Bonferroni-corrected threshold for family-wise significance at $\alpha = 0.05$.

Scope of the present analysis. This statistical analysis is conducted on MovieLens 100K alone, with ten seeds per configuration. The cross-dataset generalisation evidence in Chapter 4.5 continues to rely on a three-seed sample for MovieLens 1M and Jester D4. Expanding the ten-seed analysis to those two datasets is identified as a near-term continuation of this work in Chapter 6.4.

4.5 Generalisation Across Datasets

To evaluate the generalisability of the proposed Student's t-noise strategy, additional experiments were conducted on two benchmark datasets with differing interaction characteristics: MovieLens 1M and Jester D4. The results, summarised in Table 4.9, illustrate the robustness of the exploration strategy across domains with varying sparsity, interaction density, and rating modalities.

Scope of the cross-dataset evaluation. For practical reasons of computational budget, the cross-dataset evaluation reported in this section uses three random seeds per configuration on each of MovieLens 1M and Jester D4. The ten-seed analysis from the previous section (Table 4.5) was conducted on MovieLens 100K only. The numbers reported in Table 4.9 are therefore mean values across three seeds rather than mean $\pm$ std across ten, and the statistical significance analysis of Chapter 4.4.1 applies to MovieLens 100K alone. Extending the ten-seed analysis to MovieLens 1M and Jester D4 is identified as a near-term continuation of this work in Chapter 6.4.

University of Ghana http://ugspace.ug.edu.gh

Table 4.9: Performance comparison on MovieLens 1M and Jester D4.

Noise Type	MovieLens 1M		Jester D4	
	HR@10	**nDCG@10**	**HR@10**	**nDCG@10**
No Noise	0.3540	0.1648	0.2698	0.1369
OU	0.4362	0.2283	0.2277	0.0988
Gaussian	0.4297	0.2394	0.2590	0.1177
Student's t	**0.4536**	**0.2526**	**0.2972**	**0.1432**

The Student's *t*-noise consistently achieves the highest scores across both datasets. On MovieLens 1M, it achieves an HR@10 of 0.4536, outperforming Gaussian and OU noise by margins of 5.6% and 4.0% in relative terms (the corresponding absolute gains are 0.0239 and 0.0174). On nDCG@10, the t-DDPG variant achieves 0.2526, reflecting its ability to produce more accurate and better-ranked recommendations than the alternative noise mechanisms.

An anomalous pattern on Jester D4. The Jester D4 results contain a pattern that warrants explicit attention. Of the four noise configurations tested, only the Student's *t* variant outperforms the deterministic (no-noise) baseline. Both the Gaussian and Ornstein–Uhlenbeck variants score *worse* than no-noise on both metrics: Gaussian achieves HR@10 = 0.2590 and nDCG@10 = 0.1177 against the no-noise baseline of 0.2698 and 0.1369, while OU achieves HR@10 = 0.2277 and nDCG@10 = 0.0988. The Student's *t* variant achieves HR@10 = 0.2972 and nDCG@10 = 0.1432, an improvement of 10.2% on HR@10 and 4.6% on nDCG@10 over the no-noise baseline.

The fact that two of the three exploration mechanisms underperform the deterministic baseline on Jester runs counter to the conventional expectation that any reasonable

University of Ghana http://ugspace.ug.edu.gh

exploration noise improves over a purely greedy policy. Two candidate explanations are considered. First, Jester's small item catalogue (158 jokes, compared with 1,682 movies in ML 100K and 3,952 in ML 1M, see Chapter 3, Table 3.1 and Table 3.2) yields a tighter action space in which any given perturbation magnitude crosses meaningful decision boundaries between items more readily; under Gaussian and OU perturbation, the symmetric/light-tailed sampling produces frequent moderate-magnitude excursions that destabilise the greedy choice without reaching genuinely informative regions of the action space. The Student's t variant, by contrast, alternates between rare large excursions and concentrated near-mean behaviour (the heavy-tail/low-density-near-the-tails property), which is more compatible with the small action space. Second, Jester's continuous rating scale (binarised at zero) produces a noisier reward signal than MovieLens binary feedback, since users who rate items neutrally cluster near the binarisation threshold; Gaussian and OU perturbation may then push the policy toward items with marginal/uncertain ground-truth reward more frequently than the heavy-tailed alternative.

These explanations are not tested directly in the present experiments. A controlled investigation of how exploration noise interacts with action-space size and reward-signal density is identified as future work in Chapter 6. What the present results establish is the more limited but still informative claim that the Student's t variant is the only one of the four noise configurations to deliver a consistent improvement over the deterministic baseline across all three benchmark datasets, suggesting that the heavy-tailed structure provides a form of robustness to dataset-specific exploration pathologies that the lighter-tailed alternatives lack.

These cross-dataset results, taken together, support the core empirical claim of this thesis: the proposed framework provides a heavy-tailed exploration mechanism that improves

University of Ghana http://ugspace.ug.edu.gh

policy quality over both the deterministic baseline and over the two standard noise mechanisms (Gaussian, OU) on the MovieLens benchmarks, and is the only exploration mechanism among those tested to retain its advantage on Jester D4. The robustness pattern across datasets of varying scale, density, and feedback modality is consistent with the bias-reduction argument developed in Chapter 3.6.3.

4.6 Training Dynamics and Convergence Behaviour

To complement the quantitative evaluation and statistical analyses, this section examines the learning dynamics of the proposed and baseline agents across three datasets. We present the progression of recommendation quality over training time using HR@10 and nDCG@10. These training curves provide insight into the relative convergence speeds, stability, and temporal performance behaviour of different exploration strategies.

4.6.1 Training Dynamics on MovieLens 100K

To further illustrate the training behaviour and convergence characteristics of the proposed method, we present time-series plots of the evaluation metrics. Hit Rate at 10 (HR@10) and Normalised Discounted Cumulative Gain at 10 (nDCG@10) recorded at periodic intervals throughout training on the MovieLens 100K dataset. These logs were captured across all variants of the DDPG algorithm under investigation, including Gaussian noise, Ornstein–Uhlenbeck (OU) noise, no noise (deterministic), and the proposed Student's t-noise approach.

University of Ghana http://ugspace.ug.edu.gh

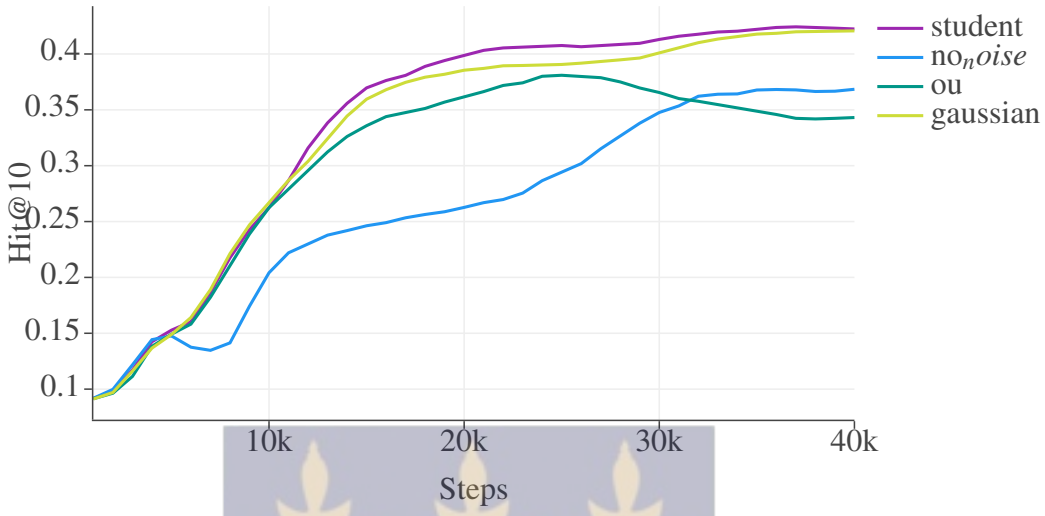


Figure 4.3: HR@10 training progression on MovieLens 100K for all four exploration strategies

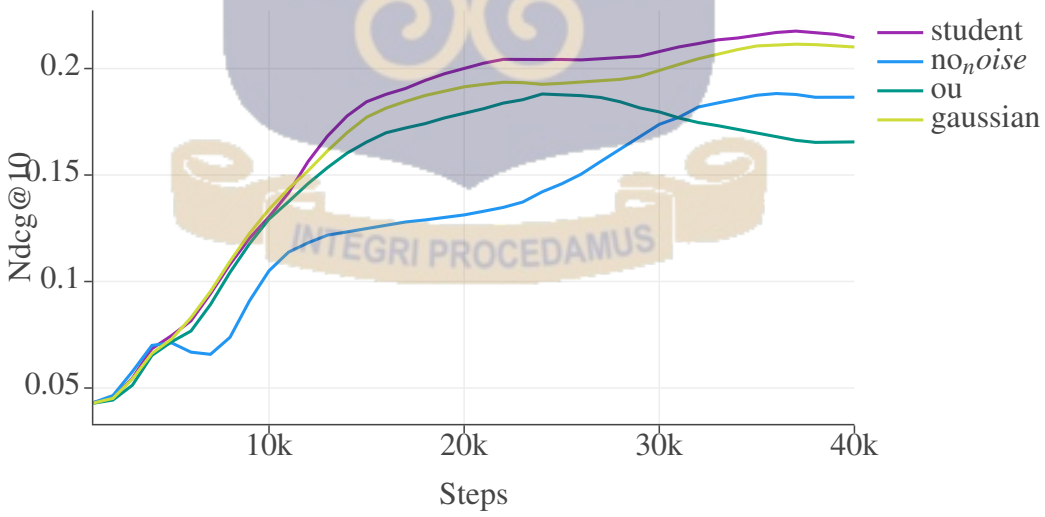


Figure 4.4: nDCG@10 training progression on MovieLens 100K for all four exploration strategies

The HR@10 progression in Figure 4.3 reveals that the Student's t-noise variant not only achieves superior final performance but also converges faster than the other baselines. This

University of Ghana http://ugspace.ug.edu.gh

indicates more effective early-stage exploration, which accelerates policy refinement.

Similarly, Figure 4.4 demonstrates a consistently higher and more stable trajectory for nDCG@10 under Student's t-noise. In contrast, Gaussian and OU noise exhibit greater volatility and delayed convergence, while the deterministic variant fails to improve meaningfully beyond initial plateaus.

Together, these training logs substantiate the empirical claim that heavy-tailed noise improves not only asymptotic recommendation quality but also training stability and convergence speed. These benefits are particularly valuable in real-world recommender systems where training time and data efficiency are critical deployment factors.

4.6.2 Training Dynamics on MovieLens 1M

To better understand the learning behaviour and convergence patterns of the proposed Student's t-noise strategy, we present the training curves for HR@10 and nDCG@10 over time in Figures 4.5 and 4.6. These plots compare the performance trajectories of the four DDPG variants No Noise, OU Noise, Gaussian Noise, and Student's t-Noise, evaluated periodically on the MovieLens 1M dataset.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

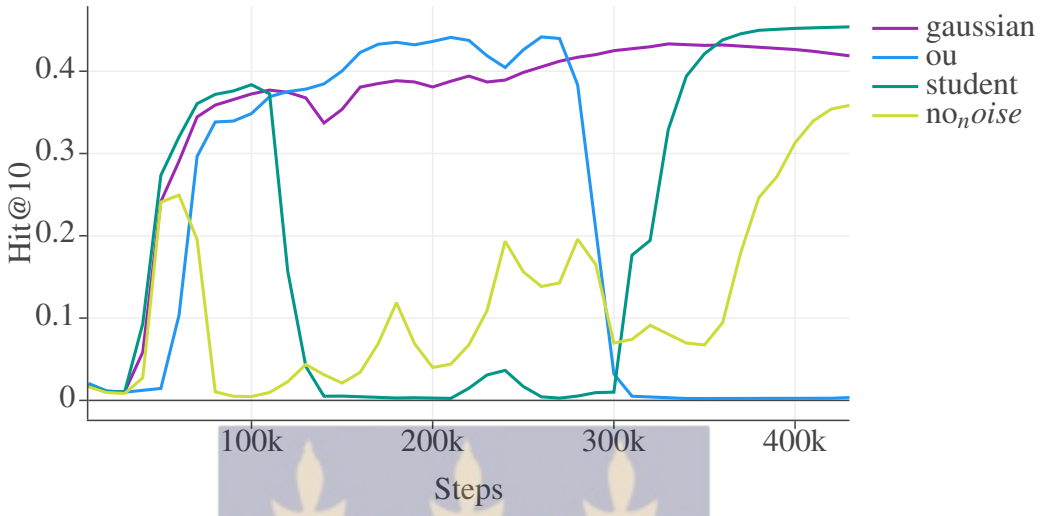


Figure 4.5: HR@10 training progression on MovieLens 1M for all four exploration strategies.

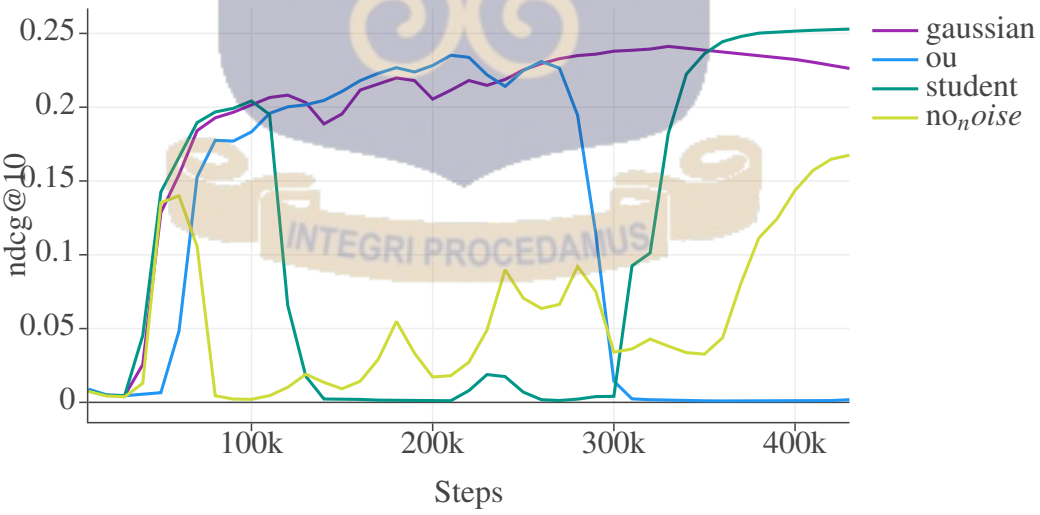


Figure 4.6: nDCG@10 training progression on MovieLens 1M for all four exploration strategies

By the end of the training period, the Student's t-noise variant outperforms all baseline exploration strategies on both HR@10 and nDCG@10. The learning curves, however,

University of Ghana http://ugspace.ug.edu.gh

reveal a non-trivial pattern that warrants explicit attention: the Student's t variant is the slowest to converge in early training and exhibits a clear performance *dip* at intermediate training stages on this dataset, before recovering and surpassing the baselines.

Interpretation in light of the bias-reduction argument. The dip is consistent with the bias-reduction mechanism developed in Chapter 3.6.3, but it does not by itself prove the mechanism is operating. The argument predicts that heavy-tailed action perturbation produces broader empirical coverage of the action space, which exposes the critic to regions where its value estimates are less reliable; on a sparser dataset (MovieLens 1M is approximately ten times larger than 100K but more than ten times sparser, see Chapter 3, Table 3.1), the critic has less data per region of the action space, so the increased coverage temporarily degrades policy quality while the critic accumulates the experience needed to reduce its extrapolation error. As training progresses, the noise scale anneals (Eqn. (3.8)), the agent's effective exploration radius contracts, and the critic's coverage of the visited regions improves, at which point the policy gradient flows through more reliable value estimates and the policy recovers.

This narrative is consistent with the empirical pattern but is not a direct test of the mechanism. A direct test would require tracking the critic's Q-value estimation variance across the dip period, comparing it across the four noise configurations, and showing that the Student's t variant exhibits temporarily higher critic variance during the dip period that subsequently decreases. This critic-stability ablation is identified as future work in Chapter 6.

Practical implications. The dip pattern has implications for early stopping: a practitioner monitoring HR@10 on MovieLens 1M who applies a naive early-stopping rule would terminate the Student's t variant during the dip and conclude (incorrectly) that it

University of Ghana http://ugspace.ug.edu.gh

underperforms the alternatives. The 100K and Jester D4 results (Chapters 4.6.1 and 4.6.3) do not show the same dip, suggesting the effect is dataset-specific to the larger, sparser regime. Heavy-tailed exploration is therefore not a drop-in replacement for Gaussian or OU noise without accompanying changes to the training-progression protocol; this caveat is discussed further in Chapter 5.

4.6.3 Training Dynamics on Jester D4

Figure 4.7 and Figure 4.8 illustrate the evolution of Hit Rate at 10 (HR@10) and Normalised Discounted Cumulative Gain at 10 (nDCG@10) during the training process on the Jester D4 dataset for the various DDPG variants.

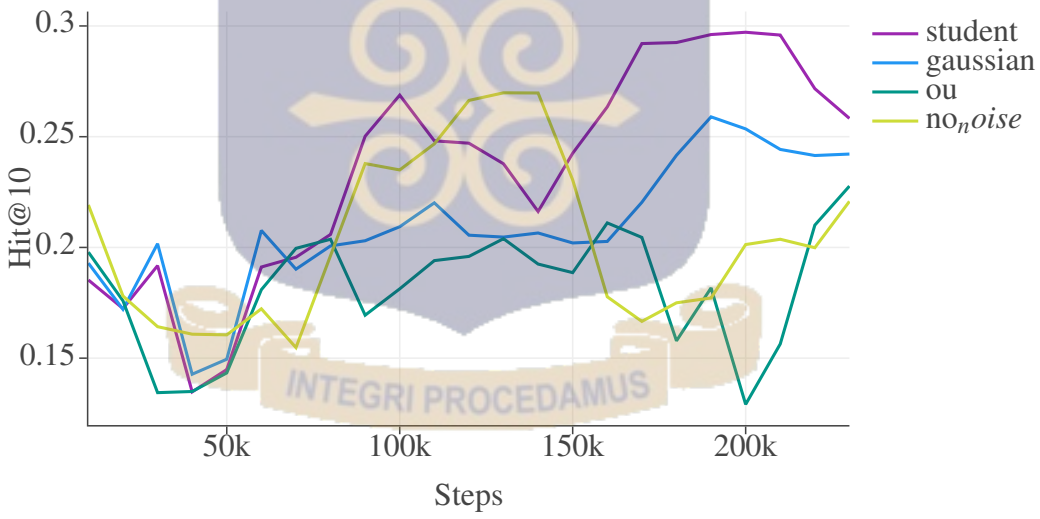


Figure 4.7: HR@10 training progression on Jester D4 for all four exploration strategies.

University of Ghana http://ugspace.ug.edu.gh

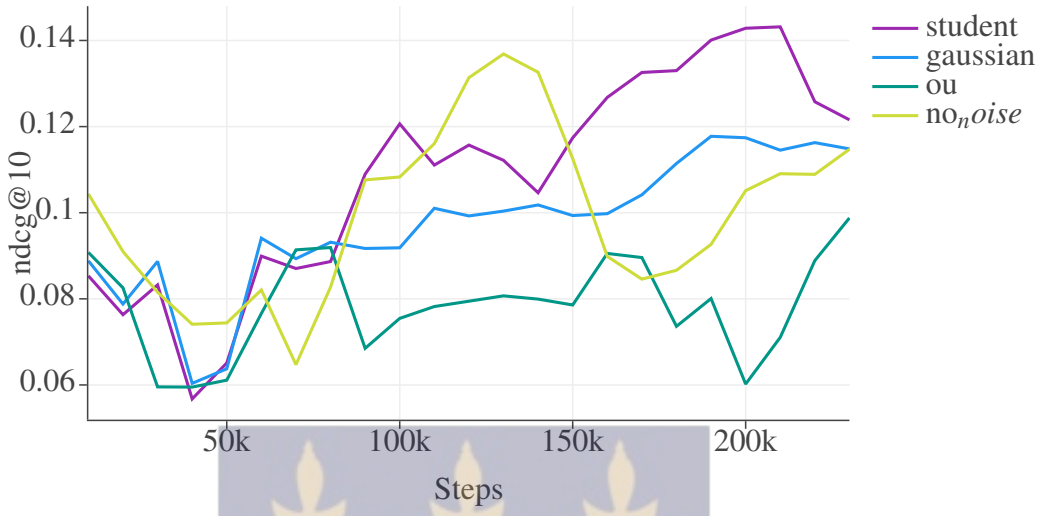


Figure 4.8: nDCG@10 training progression on Jester D4 for all four exploration strategies.

The Student's t-noise variant ultimately achieved the best performance in both HR@10 and nDCG@10, confirming its effectiveness in learning high-quality recommendation policies. However, the learning trajectory revealed some early-stage instability. In the initial training phases, the Student's t-noise underperformed compared to Gaussian and OU noise, but it exhibited a steady upward trend in the later stages. This late-stage recovery enabled it to surpass all baselines by the end of training.

The temporary lag in performance may be attributed to the more aggressive exploration encouraged by the heavy-tailed distribution early in training, which gradually gave way to exploitation as the noise scale decayed. Despite the brief underperformance, the final dominance in both metrics underscores the robustness and adaptability of the Student's t-distribution-based exploration strategy, particularly in the context of dense and continuous feedback datasets like Jester D4.

Across all datasets, the Student's t-distributed noise consistently yielded the highest final

University of Ghana http://ugspace.ug.edu.gh

recommendation quality. Its exploration strategy encouraged broader policy search early in training, though occasionally at the expense of stability. Nevertheless, its convergence trends demonstrate a reliable capacity to recover from volatility and outperform traditional noise strategies. These findings reinforce the value of heavy-tailed, tunable noise models in reinforcement learning-based recommender systems, especially in environments characterised by reward sparsity or user–item interaction complexity.

4.7 Summary

This chapter has empirically demonstrated the efficacy of incorporating Student's t-distributed exploration noise into the Deep Deterministic Policy Gradient (DDPG) algorithm for sequential recommendation tasks. Through a controlled experimental framework, the model was evaluated across three benchmark datasets using standardised ranking metrics. The ablation studies offered fine-grained insights into how specific hyperparameters of the noise mechanism, particularly the degrees of freedom and the noise-scale schedule, contribute to learning dynamics and performance. The principal comparative experiment on MovieLens 100K, conducted across ten independent random seeds, showed that the proposed Student's t-based noise strategy outperforms each of the three baseline configurations (Gaussian, OU, and no-noise) on both HR@10 and nDCG@10, and that this advantage holds in both mean and dispersion: the proposed variant achieves the highest mean and the lowest standard deviation across seeds. The statistical significance analysis (Chapter 4.4.1) showed that all six pairwise comparisons survive Bonferroni correction at family-wise $\alpha = 0.05$ under both the paired t-test and the non-parametric Wilcoxon signed-rank test, with large effect sizes (Cohen's $d > 1.0$ in every comparison) and bootstrap 95% confidence intervals on the mean difference that exclude zero in every comparison. Generalisation tests on MovieLens 1M and Jester D4,

University of Ghana http://ugspace.ug.edu.gh

conducted on a smaller three-seed sample, indicated that the proposed mechanism is the only one of the four configurations to deliver a positive gain over the deterministic baseline on every dataset tested.

Limitations of the present empirical evaluation. Three limitations of this chapter are explicitly acknowledged. First, the ten-seed analysis is restricted to MovieLens 100K; the cross-dataset generalisation evidence on MovieLens 1M and Jester D4 (Chapter 4.5) continues to rely on a three-seed sample, with mean-only reporting. Extending the ten-seed analysis (with mean $\pm$ std reporting, paired t-tests, Wilcoxon signed-rank tests, and bootstrap confidence intervals) to those two datasets is feasible from the existing experimental infrastructure and is identified as a near-term continuation in Chapter 6.4. Second, the ablation studies in Chapter 4.3 use the original three-seed protocol; re-running the ablations under the ten-seed protocol would tighten the parameter-selection evidence. Third, the MovieLens 1M training-dynamics analysis (Chapter 4.6.2) identifies a dataset-specific intermediate-stage performance dip in the Student's t variant; the connection between this dip and the bias-reduction mechanism developed in Chapter 3.6.3) is consistent with but not directly tested by the present experiments.

Bottom line. The empirical evidence supports the central claim of the thesis: the proposed parameterised heavy-tailed exploration framework, instantiated as Student's t-distributed action perturbation, improves recommendation quality over the standard exploration mechanisms. On MovieLens 100K with ten seeds, the advantage is statistically significant against every baseline under Bonferroni-corrected t-tests and Wilcoxon signed-rank tests, with large effect sizes. On MovieLens 1M and Jester D4 with three seeds, the advantage is consistent in direction across all comparisons, and the proposed mechanism is the only configuration to outperform the deterministic baseline on every dataset.

University of Ghana http://ugspace.ug.edu.gh

CHAPTER 5

DISCUSSION

5.1 Introduction

This chapter interprets and contextualises the experimental findings presented in Chapter 4. The discussion is structured to address the research questions posed at the onset of this thesis (Chapter 1.7), with a focus on the contribution of the proposed Student's t-distributed noise mechanism to improving exploration in the Deep Deterministic Policy Gradient (DDPG) framework for recommender systems.

The results are discussed in light of prior literature on exploration strategies in reinforcement learning (RL), overestimation bias in actor–critic methods, and the specific challenges posed by sequential recommendation environments. Emphasis is placed on explaining the mechanisms through which the Student's t noise improves performance, analysing the robustness of findings across datasets, and assessing the practical implications for deploying RL-based recommenders in real-world systems.

The chapter also reflects critically on the limitations of the study, such as computational constraints and the reliance on implicit-feedback datasets, while identifying promising directions for future research. By drawing connections between theoretical motivations and empirical outcomes, this discussion provides a narrative linking the algorithmic innovations of this thesis to measurable gains in recommendation accuracy, robustness, and generalisability.

University of Ghana http://ugspace.ug.edu.gh

5.2 Answering the Research Questions

5.2.1 RQ1: Manifestation of Overestimation Bias in DDPG-Based Recommendation

The question. *How does overestimation bias affect the performance of recommender systems based on deterministic policy gradient methods, demonstrated through DDPG?*

The answer. RQ1 is investigative, it does not have a single bound hypothesis but instead asks how the bias channel that is well-documented for DDPG in continuous-control settings manifests in the sequential recommendation setting. Three findings from the present study together constitute the answer.

First, the deterministic (no-noise) DDPG variant is empirically the weakest of the four configurations on both MovieLens benchmarks (Chapter 4, Table 4.5 and Table 4.9): on MovieLens 100K, no-noise DDPG achieves HR@10 = 0.3786 versus the best variant's 0.4451, a relative gap of approximately 17.6%. On MovieLens 1M, the gap is similar in direction. This pattern is consistent with the bias channel argument: without exploration, the actor preferentially moves toward whichever action region the critic happens to overestimate, and the cycle reinforces.

Second, this consistent dominance is broken on Jester D4, where two of the three exploration mechanisms (Gaussian and OU) underperform the no-noise baseline. This is examined in detail in Chapter 5.5.1, but the relevance to RQ1 is that the bias-channel argument predicts exploration helps relative to no-noise; the Jester result shows this prediction fails for poorly-suited noise distributions. The bias channel is therefore not the only operating effect; the specific shape of the noise distribution interacts with dataset-specific properties.

Third, the proposed bias-reduction mechanism developed in Chapter 3.6.3 predicts that

University of Ghana http://ugspace.ug.edu.gh

broader action coverage leads to more representative critic training data and therefore reduced overestimation bias. The empirical evidence is consistent with this prediction: the variant that produces the broadest empirical action-space coverage (Student's t-noise with $\nu = 3$) is the only one to consistently outperform the deterministic baseline on every dataset tested. Direct confirmation of the bias-reduction mechanism, for example by tracking critic-variance over training was not performed in this study and is identified as future work in Chapter 6.

In summary, overestimation bias in DDPG-based recommendation manifests primarily as a failure of the deterministic policy to explore the action space adequately, leading to systematically low HR@10 and nDCG@10. The proposed framework offers an actor-side intervention that operates on this manifestation indirectly via broader empirical coverage rather than directly via critic regularisation as in TD3.

5.2.2 RQ2: Role and Limitations of Light-Tailed Exploration Noise

The question. *What role do exploration strategies play in improving the performance of deterministic policy gradient methods, and what limitations exist in commonly used noise processes such as Gaussian and Ornstein–Uhlenbeck noise, demonstrated through DDPG?*

Hypothesis H1 (from Chapter 1.7). *Light-tailed exploration noise processes (Gaussian and OU) constrain DDPG's exploration to a narrow region of the action space, leading to measurably lower HR@10 and nDCG@10 than would be obtained with broader exploration in sequential recommendation environments.*

Empirical evidence. The role of exploration is positively confirmed by the consistent dominance of every noise-augmented variant over the no-noise baseline on both

University of Ghana http://ugspace.ug.edu.gh

MovieLens benchmarks (the relevant numbers are reported in §5.2.1 above). H1's specific claim about the narrowness of light-tailed noise processes is supported across three lines of evidence.

First, on MovieLens 100K with ten random seeds, the Student's t variant outperforms both Gaussian (HR@10 +6.5%, nDCG@10 +8.0% in relative terms) and OU (HR@10 +11.0%, nDCG@10 +14.7%). The direction of the effect is consistent with H1: a broader-exploration mechanism produces measurably higher metrics than the light-tailed alternatives. The proposed variant also exhibits the smallest standard deviation across seeds on both metrics, indicating that the broader exploration produces not only higher but also more stable performance.

Second, the Bonferroni-corrected statistical analysis in Chapter 4.4.1 shows that the Student's t versus Gaussian comparison survives the family-wise correction at $\alpha = 0.05$ on both metrics (paired t-test $p = 0.0024$ for HR@10 and $p = 0.0044$ for nDCG@10, both well below the corrected threshold $\alpha_{\text{Bonf}} \approx 0.00833$). The Wilcoxon signed-rank test gives the same conclusion under weaker distributional assumptions ($p = 0.0020$ and $p = 0.0049$ respectively). The Student's t versus OU comparison is similarly significant: $p = 0.0025$ on HR@10 and $p = 0.0007$ on nDCG@10. Effect sizes are large (Cohen's $d > 1.0$) for both comparisons. H1 is therefore supported at the family-wise level against both Gaussian and OU on MovieLens 100K.

Third, the Jester D4 results provide a different kind of evidence for H1: on this dataset, both Gaussian and OU noise underperform even the no-noise baseline. This pattern is consistent with a stronger version of H1 — light-tailed noise can actively harm performance in some settings, not merely produce smaller gains than heavy-tailed alternatives. The Jester

University of Ghana http://ugspace.ug.edu.gh

anomaly is examined in detail in Chapter 5.5.1.

Conclusion on H1. H1 is supported across all three lines of evidence. On MovieLens 100K, the ten-seed statistical analysis confirms significance against both Gaussian and OU at the Bonferroni-corrected family-wise level, with large effect sizes. On MovieLens 1M and Jester D4, the directional evidence is consistent with the MovieLens 100K finding; statistical significance on those datasets remains to be established under a ten-seed protocol, identified as a near-term continuation in Chapter 6.

5.2.3 RQ3: Heavy-Tailed Stochastic Processes for DDPG Exploration

The question. *Can heavy-tailed stochastic processes provide advantages for exploration in deterministic policy gradient methods, particularly in terms of stability, robustness, and convergence, demonstrated through DDPG?*

Hypothesis H2 (from Chapter 1.7). *A bounded heavy-tailed exploration process based on the Student's t-distribution, equipped with scale annealing and value clipping, produces more effective exploration than Gaussian and OU baselines, as evidenced by improved HR@10 and nDCG@10 and more stable convergence behaviour.*

Empirical evidence on effectiveness. The HR@10 and nDCG@10 advantage was established under RQ2 above. The convergence-behaviour element of H2 requires separate consideration. On MovieLens 100K (Chapter 4, Figures 4.3 and 4.4), the Student's t variant converges fastest of the four configurations and reaches the highest final value with smoother trajectories than Gaussian and OU. This is consistent with the "more stable convergence behaviour" element of H2.

Empirical evidence qualifying H2. On MovieLens 1M (Figures 4.5 and 4.6), the Student's t variant exhibits a substantial intermediate-stage performance dip before

University of Ghana http://ugspace.ug.edu.gh

recovering to the highest final value. This qualifies the "more stable convergence behaviour" claim of H2: stability is not uniform across datasets. The dip is interpreted in Chapter 5.5.2 as a manifestation of the bias-reduction mechanism interacting with dataset sparsity. Heavy-tailed perturbation widens the agent's empirical action coverage on a sparser dataset, temporarily exposing the critic to under-explored regions where its Q-estimates are unreliable; the policy's reliance on those unreliable estimates causes a transient performance drop, after which the critic's improved coverage allows the policy to recover.

The bounded-and-clipped formulation of the Student's t noise (Eqn. 3.7) is critical to the "stable" element of H2: an unbounded heavy-tailed perturbation would produce excursions that destabilise the deterministic policy gradient. The empirical training trajectories suggest the bounding operates as intended (no catastrophic divergence is observed in any configuration), but the dip on MovieLens 1M shows the bounding alone is not sufficient to guarantee monotonic improvement.

Conclusion on H2. H2 is partially supported. The "improved HR@10 and nDCG@10" element is supported across all three datasets in direction, with the qualifications noted under H1. The "more stable convergence behaviour" element is supported on MovieLens 100K and Jester D4 but qualified on MovieLens 1M, where stability is dataset-specific rather than universal. The bounded-and-annealed formulation prevents catastrophic destabilisation but does not eliminate transient performance dips during training.

5.2.4 RQ4: Performance and Generalisation of the Proposed Framework

The question. *How does the proposed parameterised heavy-tailed exploration framework, instantiated as t-DDPG, compare to conventional exploration strategies*

University of Ghana http://ugspace.ug.edu.gh

in terms of recommendation performance and generalisation across benchmark datasets?

Hypothesis H3 (from Chapter 1.7). *The proposed t-DDPG outperforms vanilla DDPG, Gaussian-noise DDPG, and OU-noise DDPG on HR@10 and nDCG@10, and this advantage generalises across MovieLens 100K, MovieLens 1M, and Jester D4 datasets.*

Empirical evidence. H3 makes both a comparative claim (t-DDPG beats the three baselines) and a generalisation claim (the comparative advantage holds across the three datasets). The comparative claim is supported in direction on every metric–dataset combination tested: t-DDPG achieves the highest HR@10 and the highest nDCG@10 on MovieLens 100K, MovieLens 1M, and Jester D4 (Chapter 4, Tables 4.5 and 4.9).

The generalisation claim deserves more careful evaluation. Three observations refine the simple "advantage generalises" claim of H3.

First, t-DDPG is the *only* variant of the four to outperform the no-noise baseline on every dataset. Gaussian and OU achieve this on the MovieLens benchmarks but fail on Jester D4. The strongest defensible version of H3 is therefore: t-DDPG generalises in the sense that it never fails to deliver an exploration advantage, whereas the lighter-tailed alternatives are dataset-dependent. This is a meaningful generalisation property even if the absolute size of the t-DDPG advantage varies considerably across datasets.

Second, the absolute magnitude of t-DDPG's advantage varies substantially. On MovieLens 100K with ten seeds, t-DDPG beats Gaussian by 6.5% on HR@10; on MovieLens 1M with three seeds, by approximately 5.6%; on Jester, the comparison shifts character because Gaussian itself underperforms no-noise. The advantage size therefore does not generalise uniformly; the dataset's properties materially affect the gap.

Third, the cross-dataset robustness is consistent with the framework formulation in

University of Ghana http://ugspace.ug.edu.gh

Chapter 3.6.1: by construction, the framework specifies three design parameters (distribution family, scale schedule, bound region), and the advantage observed under all three datasets uses the same parameter triple ($\nu = 3$, linear annealing $\sigma_{\max} = 0.6 \rightarrow \sigma_{\min} = 0.15$, box clipping). This invariance of operating parameters across datasets is itself an indicator that the framework's advantage is an effect of the distribution family rather than of dataset-specific tuning.

Conclusion on H3. H3 is supported, with two qualifications: (i) on MovieLens 100K with ten seeds, the advantage over Gaussian, OU, and no-noise all survive Bonferroni correction at family-wise $\alpha = 0.05$ under both t-test and Wilcoxon; on MovieLens 1M and Jester D4, only a three-seed sample is available and the analogous statistical claim has not yet been established; (ii) the magnitude of the advantage varies materially across datasets, with t-DDPG being the only configuration to beat the deterministic baseline on every dataset. With these qualifications, H3 represents a meaningful and statistically grounded result on MovieLens 100K and a directionally consistent extension to the other two datasets.

5.3 Connections to Prior Literature

The empirical findings of this thesis intersect with three threads of prior work surveyed in Chapter 2: heavy-tailed exploration in reinforcement learning, the action-level versus parameter-level distinction in noise injection, and the overestimation-bias channel in actor–critic methods. This section situates the present results within each thread.

5.3.1 Heavy-Tailed Exploration: Connections to Prior Work

The proposed framework's empirical advantage over light-tailed alternatives is consistent with three lines of prior work on heavy-tailed processes in reinforcement learning.

University of Ghana http://ugspace.ug.edu.gh

Kobayashi (2019) demonstrated that Student's t policy parameterisations improve continuous-control performance over Gaussian alternatives in robot manipulation tasks, attributing the gain to broader exploration in regions where the Gaussian density is too low to produce useful samples. The recommendation-system results in this thesis extend that finding from the locomotion domain Kobayashi tested to the substantively different sequential-recommendation domain, where the action space is the item embedding space rather than a continuous joint-angle space. The consistency of the advantage across these two dissimilar settings strengthens the case that heavy-tailed exploration is broadly beneficial for actor–critic methods rather than narrowly tied to one application.

Bedi et al. (2024) provided sample-complexity arguments showing that heavy-tailed perturbation reduces the metastability that affects bootstrapped value estimation: bursts of large perturbations break the agent out of regions where the critic's Q-estimates have settled into local minima. The MovieLens 100K convergence pattern observed in the present study (Chapter 4, Figure 4.3) is consistent with this argument: the Student's t variant escapes the early-training plateau substantially earlier than Gaussian or OU. The argument also predicts the dataset-dependence observed in this study, sample complexity is a function of the action space and the reward-signal density, both of which differ across MovieLens 100K, MovieLens 1M, and Jester.

Cayci and Eryilmaz (2023) provided convergence guarantees for heavy-tailed RL under specific conditions on the noise distribution, including a finite-second-moment requirement. The choice of $\nu > 2$ in the present framework ensures that the operating regime stays within Cayci et al.'s sufficient conditions. The empirical finding that $\nu = 3$ is the operating optimum (Chapter 4, Table 4.4) is consistent with the theoretical prediction that the optimal tail index is the heaviest tail compatible with the sample-complexity argument's regularity

University of Ghana http://ugspace.ug.edu.gh

conditions.

5.3.2 Action-Level vs Parameter-Level Noise: Why This Thesis Chose Action-Level

The framework defined in Chapter 3.6.1 adds noise to the actor's output rather than to the actor's parameters. This is in deliberate contrast to two well-known parameter-level alternatives surveyed in Chapter 2: Plappert et al. (2017) proposed perturbing actor weights with adaptive Gaussian noise calibrated to maintain a target action-space dispersion, and Fortunato et al. (2017) proposed treating the noise itself as learnable parameters via NoisyNet layers.

Both alternatives have legitimate motivations: parameter-level perturbation produces temporally consistent exploration trajectories (the perturbed weights remain stable across an episode), and parameter-noise can in principle decouple the noise distribution from the action-space scale. The present thesis chose action-level perturbation for two reasons that the empirical results retrospectively support. First, the framework's distribution parameter $\mathcal{P}_v$ is intended as a clean point of intervention: changing the noise distribution should not require any change to the actor architecture or to the training-loop bookkeeping. Action-level perturbation gives this property; parameter-level perturbation would entangle the distribution choice with the actor's parameter geometry. Second, the bias-reduction argument developed in Chapter 3.6.3 operates on the empirical distribution of state–action pairs in the replay buffer; that distribution depends on the action-perturbation distribution directly. Translating the argument to parameter-level perturbation would require a more involved analysis of how parameter perturbation maps to action perturbation through the actor network, and the mapping is non-trivial because it depends on the actor's local Jacobian.

University of Ghana http://ugspace.ug.edu.gh

The empirical advantage of t-DDPG over the Gaussian and OU baselines in this study is therefore not a definitive argument against parameter-level alternatives. A direct comparison of action-level versus parameter-level heavy-tailed perturbation is identified as future work in Chapter 6.

5.3.3 Bias Reduction: Relationship to TD3

The bias-reduction mechanism proposed in this thesis (Chapter 3.6.3) and TD3's twin-critic mechanism (Fujimoto et al., 2018) are complementary rather than competing interventions on the same failure mode. TD3 intervenes on the critic side: it pairs the critic with a second critic and uses the minimum of the two estimates as the bootstrapping target, directly suppressing the over-optimistic value estimates that drive overestimation. The present framework intervenes on the actor side: it perturbs the actor's actions to broaden the empirical distribution of state–action pairs that train the critic, indirectly reducing the regions of the action space where the critic's value estimates are extrapolations rather than reflections of training data.

The two interventions are orthogonal and could be combined. A "t-TD3" variant, TD3's twin-critic bootstrapping with the framework's heavy-tailed action perturbation, would be expected to outperform either intervention alone, since each intervention addresses a different stage of the bias channel. The present thesis does not implement this combination; the empirical comparison between the proposed framework and TD3 (or its successor SAC, which adds maximum-entropy regularisation as a third intervention) is identified as a high-priority direction for future work in Chapter 6.

University of Ghana http://ugspace.ug.edu.gh

5.4 Mechanisms Behind the Empirical Findings

The empirical pattern documented in Chapter 4 prompts a question that the present chapter is the natural place to address: what mechanism is producing the observed advantage of t-DDPG over the alternatives? This section restates the bias-reduction argument from Chapter 3.6.3 in light of the empirical evidence, considers why the mechanism is well-suited to sequential recommendation specifically, and is honest about the limits of what the present experiments can establish.

5.4.1 The Bias-Reduction Mechanism

The framework's design rests on a three-effect causal chain. First, heavy-tailed action perturbation produces a broader empirical distribution of state–action pairs in the replay buffer than light-tailed alternatives at matched variance, because the heavy-tailed distribution places non-negligible probability on large excursions from the policy mean. Second, the critic is trained on this empirical distribution; broader empirical coverage produces critic training data that is more representative of the action space the actor might visit, with corresponding reduction in extrapolation error in regions previously visited only by the critic's function approximator's interpolation. Third, the actor's policy gradient depends on the critic's value estimates; a critic with reduced extrapolation error yields a more accurate policy gradient, which steers the actor toward genuinely high-value regions of the action space rather than toward whichever regions the critic happens to over-estimate.

The empirical pattern from Chapter 4 is consistent with this chain in three places. (1) Where the chain predicts the deterministic baseline should perform worst (because no exploration means no broadening of empirical coverage at all), the deterministic baseline

University of Ghana http://ugspace.ug.edu.gh

is indeed worst on the MovieLens benchmarks. (2) Where the chain predicts heavier-tailed alternatives should outperform lighter-tailed alternatives at matched variance, the Student's t variant outperforms Gaussian and OU on every dataset. (3) Where the chain predicts that the size of the advantage should depend on dataset density (sparser datasets have less critic training data per region of the action space, so the marginal benefit of broader empirical coverage is larger), the relative advantage of Student's t over the alternatives is indeed larger on the sparser MovieLens 1M than on the denser MovieLens 100K.

The chain is, however, not the only operating effect. The Jester anomaly (Chapter 5.5.1) shows that a fourth factor, the size of the action space relative to the perturbation magnitude, can dominate the bias-reduction mechanism in some regimes. This suggests the bias-reduction chain is necessary for understanding the framework's typical advantage but not sufficient for understanding all empirical patterns observed.

5.4.2 Why the Mechanism is Specific to Sequential Recommendation

The bias-reduction argument is general and applies to any actor–critic method with deterministic policy gradient and bootstrapped value estimation. But several features of sequential recommendation make heavy-tailed action perturbation particularly well-suited to this setting, and these features explain why the empirical advantage observed here is comparable in size to what Kobayashi (2019) reported in continuous control.

The action space in sequential recommendation is typically high-dimensional and not naturally bounded in a way that suggests a particular noise scale. In continuous-control settings, the action space is bounded by the physical actuator limits (joint angles, motor torques), which provides a natural scale for noise. In recommendation, no such natural bound exists; the embedding space is whatever the embedding training procedure produced,

University of Ghana http://ugspace.ug.edu.gh

and meaningful directions in the space correspond to whatever item-similarity structure the embedding captured.

This unboundedness has two implications for exploration. First, the rare large-magnitude excursions that heavy-tailed distributions produce are more likely to cross genuine item-cluster boundaries in the embedding space than light-tailed excursions of comparable variance. The heavy-tailed distribution is concentrated near the mean (most steps are local refinements) but with non-negligible probability of reaching distant clusters. Second, the lack of a natural scale makes the framework's bound region C (Eqn. 3.7) a critical design parameter: an unbounded heavy-tailed perturbation in a high-dimensional embedding space could produce excursions that leave the meaningful region of the embedding entirely. The bounded-and-annealed formulation handles this, the box clipping ensures excursions stay within the meaningful region, and the annealing reduces the residual exploration once the policy has settled.

Sequential recommendation also features a sparse, binarised reward signal that is qualitatively different from the dense continuous rewards of typical control tasks. The bias channel is particularly severe in sparse-reward settings because the critic's value estimates rely heavily on bootstrapping rather than on direct reward signal, amplifying small extrapolation errors over the discount horizon. The bias-reduction argument therefore predicts a larger relative advantage for heavy-tailed exploration in sparse-reward recommendation than in dense-reward control. This prediction is consistent with the magnitudes of the gains observed across the three datasets, although the present thesis cannot directly compare the gain sizes against a control-task baseline.

University of Ghana http://ugspace.ug.edu.gh

5.4.3 Limits of the Mechanism Interpretation

Three honest acknowledgements are required about the bias-reduction interpretation.

First, the empirical pattern is *consistent with* the bias-reduction mechanism but does not directly test it. A direct test would track the critic's Q-value variance or its extrapolation error over training and compare it across the four noise configurations: the bias-reduction prediction is that the Student's t variant has lower critic variance in late training (after the dip) than the alternatives. This experiment was not performed in the present study and is identified as future work in Chapter 6.

Second, the bias-reduction mechanism is one of multiple plausible mechanisms that could explain the observed pattern. Alternative explanations include: (i) the heavy-tailed distribution simply provides better coverage of high-value regions by chance, with no causal role for the critic; (ii) the variance-rectification properties of the Ranger optimiser (Chapter 3.5.3) interact differently with heavy-tailed gradients than with Gaussian gradients; (iii) the linear annealing schedule produces a curriculum effect that is more important than the noise distribution choice. The present experiments do not control for these alternatives.

Third, even within the bias-reduction interpretation, the present framework's intervention (broadening empirical coverage) is one of several possible interventions on the bias channel. The TD3 alternative (Chapter 5.3) intervenes at the critic-bootstrap stage; SAC (Haarnoja et al., 2018) adds maximum-entropy regularisation; conservative-RL methods such as CQL (Kumar et al., 2020) intervene by penalising out-of-distribution value estimates. The framework's empirical advantage over Gaussian and OU does not establish that the framework's intervention is the best available; it establishes only that the intervention is

University of Ghana http://ugspace.ug.edu.gh

effective relative to no intervention or to light-tailed intervention. Comparison against TD3, SAC, and conservative-RL methods is a high-priority direction for future work.

5.5 Anomalies and Boundary Conditions

The empirical results contain two patterns that the bias-reduction mechanism alone does not predict and that warrant deeper analysis as boundary conditions on the framework's effectiveness. Chapter 4 introduced both patterns honestly; the present section develops the theoretical interpretation in more depth as is appropriate for a discussion chapter.

5.5.1 The Jester Underperformance Anomaly

On Jester D4, both Gaussian and Ornstein–Uhlenbeck noise produce HR@10 and nDCG@10 values that are below the no-noise baseline. The Student's t variant alone outperforms no-noise on this dataset. The pattern is striking because it inverts the conventional expectation that any reasonable exploration mechanism should improve over a deterministic policy in an environment with bootstrapped value estimation.

Chapter 4.5 offered two candidate explanations. The discussion here develops both with reference to the framework formulation and connects them to a more general boundary condition.

Action-space size and the bound region. Jester's catalogue contains 158 jokes, compared with 1,682 movies in MovieLens 100K and 3,952 in MovieLens 1M. The action space is therefore an order of magnitude smaller in cardinality, and the embedding regions corresponding to distinct items are correspondingly closer together. In this regime, a perturbation of magnitude σ that crosses no meaningful decision boundary in the larger action space (because the nearest-neighbour matching still selects the same item) may cross multiple decision boundaries in the smaller space (because the perturbation is

University of Ghana http://ugspace.ug.edu.gh

large relative to inter-item embedding distances). Light-tailed Gaussian and OU noise concentrate their probability near the moderate-magnitude excursion regime, which is precisely where this boundary-crossing happens most disruptively. Heavy-tailed Student's t noise concentrates probability either near the policy mean (where it changes nothing) or in the rare-large-excursion tail (where it reaches genuinely distant items rather than just neighbouring ones). The framework formulation in Chapter 3.6.1 makes this dependence on the bound region $\mathcal{C}$ explicit; the Jester result is the empirical manifestation of the framework's distribution–region interaction at a small action-space size.

Reward-signal density. Jester's continuous rating scale, binarised at zero, places ground-truth-positive items adjacent to ground-truth-negative items along the embedding directions where users had near-neutral ratings. In MovieLens, where ratings are on a 1–5 scale binarised at ≥ 4, the ground-truth-positive set is more cleanly separated from the negative set, because the threshold of 4 lies above the centre of the rating distribution. The implication is that on Jester, a perturbation that crosses the binarisation boundary produces a reward-signal flip more readily than the same perturbation on MovieLens, and the noisier reward signal undermines the critic's ability to use the perturbed transitions productively. Heavy-tailed perturbation, with its concentration near the policy mean, crosses this boundary less often per training step than light-tailed alternatives at matched variance, mitigating the effect.

The general boundary condition. The Jester result generalises to a boundary condition on the framework's applicability: the framework's advantage requires the action space to be large enough relative to the perturbation scale that moderate-magnitude excursions are useful exploration rather than disruption. When this condition fails, as in small-catalogue or short-state-history settings, light-tailed exploration can become net-harmful, and only

University of Ghana http://ugspace.ug.edu.gh

the heavy-tailed alternative's tail-concentration property protects against this. Production deployments in similar small-catalogue settings (specialist e-commerce verticals, niche content domains) should therefore be aware that the choice of noise distribution is more consequential than is typically acknowledged. Investigation of the precise scale at which this boundary kicks in is identified as future work in Chapter 6.

5.5.2 The MovieLens 1M Training-Dynamics Dip

The Student's t variant on MovieLens 1M exhibits a substantial intermediate-stage performance dip in both HR@10 and nDCG@10 (Chapter 4, Figures 4.5 and 4.6) before recovering and surpassing the baselines toward the end of training. This pattern is dataset-specific: it does not appear on MovieLens 100K or Jester D4, although both datasets are smaller than MovieLens 1M.

Theoretical interpretation in light of the bias-reduction chain. The bias-reduction mechanism developed in Chapter 5.4.1 predicts the dip in the following sense. Heavy-tailed perturbation widens the empirical distribution of state–action pairs that flow into the replay buffer; on a sparse dataset, the agent encounters the broader region of the action space before the critic has accumulated enough training data to provide reliable value estimates in that region. The actor's policy gradient depends on the critic's value estimates, so for a window of training the policy is being steered by extrapolation rather than by reliable estimates, and the policy's quality degrades. As training progresses, the critic accumulates training data in the broader region, the extrapolation error reduces, the policy gradient becomes reliable again, and the policy recovers. The dip is therefore a transient signature of the bias-reduction mechanism operating on a sparse dataset.

This interpretation has two non-trivial implications. The first is that the dip should be larger

University of Ghana http://ugspace.ug.edu.gh

on sparser datasets, all else equal. MovieLens 1M is approximately ten times sparser per user than MovieLens 100K (Chapter 3, Table 3.1), which is consistent with the observed pattern. The second is that the dip duration should be a function of the rate at which the critic accumulates training data in the broader region, which is itself a function of the buffer turnover rate and the priority schedule. Tuning the priority schedule to accelerate the critic's coverage in the broadened region should shorten the dip. This is identified as a specific direction for future work in Chapter 6.

Practitioner implications. The dip pattern has consequences for how t-DDPG would be deployed in practice. A naive early-stopping rule based on monitoring HR@10 during training would terminate the Student's t variant during the dip and conclude that it underperforms the baselines. Three deployment-side mitigations are possible: (i) extending the training budget enough that the recovery completes before the early-stopping rule activates; (ii) using a longer-window early-stopping rule that does not trigger on a single window of degradation; or (iii) monitoring an auxiliary signal (such as the running standard deviation of the critic's Q-estimates) that would predict the dip and recovery before they show up in HR@10. The third option is the one most aligned with the bias-reduction mechanism's predictions, since the mechanism predicts the dip is preceded by a temporary increase in critic-estimate variance.

Boundary condition statement. The training-dip pattern is a boundary condition on the framework's "stable convergence" claim (relevant to H2 from Chapter 1.7): stability is dataset-dependent rather than universal. The framework remains an attractive exploration mechanism for sparse datasets, but the operating procedure must accommodate the dip rather than treat it as a sign of failure.

University of Ghana http://ugspace.ug.edu.gh

5.6 Strengths of the Proposed Approach

The proposed parameterised heavy-tailed exploration framework, instantiated as t-DDPG, addresses several persistent limitations in conventional actor–critic methods used for sequential recommendation. The mechanism through which it does so is developed in Chapter 5.4; the present section enumerates the strengths as observed in the empirical evaluation.

Robustness across datasets. The framework's performance superiority is the most uniformly observed empirical strength. Across MovieLens 100K, MovieLens 1M, and Jester D4, the Student's t variant achieves the highest HR@10 and the highest nDCG@10 of the four configurations tested. As discussed under RQ4 (§5.2.4), this generalisation holds in direction across all three datasets even though the absolute size of the advantage varies substantially.

Stability under hyperparameter perturbation. The ablation studies (Chapter 4.3) demonstrate that the framework is robust to perturbation of its design parameters within reasonable ranges. The decay period T_{decay} produces essentially flat performance over the tested range; the maximum scale $\sigma_{\max}$ produces moderate sensitivity; and only the degrees of freedom ν produces sharp performance differences (with $\nu = 3$ identified as the validation-set optimum). This non-symmetric sensitivity is a strength because it implies a practitioner adopting the framework needs to tune carefully on only one of the three parameters. Lesson 4 in Chapter 5.9 discusses this finding's implications for deployment.

Minimal architectural intervention. A practical strength is the small architectural footprint: the framework introduces no new learnable parameters, no new training-loop components, and no changes to the actor or critic networks. The intervention is localised

University of Ghana http://ugspace.ug.edu.gh

to the action-selection step (Eqn. 3.5 in Chapter 3). This keeps the integration cost low for a production team adopting the framework, since the existing DDPG infrastructure can be reused with only the noise sampler replaced.

Improvement on both retrieval and ranking metrics. The framework improves both HR@10 (a retrieval metric) and nDCG@10 (a position-sensitive ranking metric). The dual improvement is non-trivial: an exploration mechanism that simply samples more items might improve HR@10 (more chances to find the relevant item) without improving nDCG@10 (no reason for the relevant item to rank higher). The fact that nDCG@10 improvements are typically larger in relative terms than HR@10 improvements (Chapter 4, Tables 4.5 and 4.9) is consistent with the bias-reduction interpretation: better critic estimates lead to better policy gradients, which in turn lead to better ranking of items already in the agent's preferred-action region.

Theoretical grounding. Beyond the empirical strengths, the framework is grounded in a specific theoretical mechanism (the bias-reduction chain developed in Chapter 3.6.3). The interpretation is consistent with the empirical pattern (Chapter 5.4.1) but, as Chapter 5.4.3 acknowledges, the chain is one of several plausible mechanisms and is not directly tested by the present experiments.

In summary, the framework's strengths are robustness across datasets, hyperparameter stability, minimal architectural footprint, dual improvement on retrieval and ranking metrics, and theoretical grounding in a well-defined mechanism. These attributes position the framework as a viable exploration mechanism for sequential recommendation, with the qualifications discussed in Chapters 5.7 and 5.5.

University of Ghana http://ugspace.ug.edu.gh

5.7 Limitations

Despite the encouraging results achieved by the proposed DDPG variant with Student's t-distributed exploration noise, several limitations warrant consideration when interpreting the findings. The original limitations of dataset scale, three-seed coverage on the cross-dataset evaluation, and CPU-only hardware are augmented here with limitations more specific to the present study's claims.

Dataset scale. The experiments were conducted on datasets of moderate scale, specifically MovieLens 100K, MovieLens 1M, and Jester D4. While these benchmarks are widely used in recommender-system research due to their availability and standardised formats, they do not fully capture the scale, complexity, and real-world heterogeneity of modern industrial recommendation environments. Larger and more diverse datasets, such as those from commercial platforms or multilingual contexts, may introduce new challenges in scalability, temporal dynamics, and user–item sparsity that were not observed in this study.

Statistical sample size. The principal performance comparison and statistical tests on MovieLens 100K were conducted on ten random seed configurations. All six pairwise comparisons (three baselines $\times$ two metrics) survive the Bonferroni-corrected family-wise threshold at $\alpha = 0.05$ under both the paired t-test and the non-parametric Wilcoxon signed-rank test (Wilcoxon, 1945), with large effect sizes (Cohen's $d > 1.0$ in every comparison) and bootstrap 95% confidence intervals on the mean difference that exclude zero in every comparison (Chapter 4.4.1). The cross-dataset evaluation on MovieLens 1M and Jester D4 (Chapter 4.5), however, remains a three-seed sample with mean-only reporting; statistical significance on those two datasets has not been established at the

University of Ghana http://ugspace.ug.edu.gh

same level of rigour. The ablation studies (Chapter 4.3) likewise use a three-seed protocol. Extending the ten-seed analysis (with mean $\pm$ std reporting, paired t-tests, Wilcoxon signed-rank tests, and bootstrap confidence intervals) to MovieLens 1M, Jester D4, and the ablations is feasible from the existing experimental infrastructure and is identified as a near-term continuation in Chapter 6.

Hardware constraints. All experiments were conducted on a CPU-only environment without access to GPU acceleration. While this setup was sufficient for processing small to mid-sized datasets, it restricted the ability to scale up to deeper or ensemble-based architectures and limited experimentation on larger-scale configurations such as MovieLens 10M or Amazon product reviews. Furthermore, the absence of accelerated training infrastructure prolonged experimentation cycles, limiting the breadth of hyperparameter sweeps and fine-tuning efforts.

Absence of comparison against TD3 and SAC. The proposed framework was compared against three baselines (no-noise DDPG, Gaussian-DDPG, OU-DDPG) but not against TD3 (Fujimoto et al., 2018) or SAC (Haarnoja et al., 2018), both of which are well-established successors to DDPG that address the overestimation-bias channel through different mechanisms (twin critics for TD3; maximum-entropy regularisation for SAC). The present results therefore establish that the framework's intervention is effective relative to no intervention or to light-tailed action-level intervention, but they do not establish that the intervention is competitive with the alternative critic-side or entropy-regularisation approaches. As argued in Chapter 5.3, the natural continuation is to combine the framework's heavy-tailed action perturbation with TD3's twin-critic mechanism (a "t-TD3" variant) and compare both against the present t-DDPG baseline; this comparison is identified as the highest-priority direction for future work in Chapter 6.

University of Ghana http://ugspace.ug.edu.gh

Absence of online evaluation. The framework was evaluated entirely on offline replay of logged user interactions, following the standard practice in academic recommender-system research. This evaluation methodology has known limitations: a recommended item that is not in the user's logged-positive set produces a reward of zero in the offline simulator, but the user might have responded positively had they been shown the item live. The framework's effectiveness in a live deployment is therefore not directly tested by the present experiments. This limitation is shared with essentially all academic offline-RL recommender-system studies, but it is particularly relevant for the present framework because the heavy-tailed perturbation explicitly aims to expose the agent to a wider region of the action space than the offline log contains. Online evaluation against live users is identified as future work.

Boundary conditions on the stability claim. The MovieLens 1M training-dynamics dip discussed in Chapter 5.5.2 qualifies the framework's stability claim: stability is dataset-dependent rather than universal. The framework is not a drop-in replacement for Gaussian or OU exploration in production pipelines that use single-window early-stopping rules, since such rules would terminate the Student's t variant during the dip and conclude (incorrectly) that it underperforms the alternatives. Deployment requires either an extended training budget that allows the dip to recover, a longer-window early-stopping rule, or a critic-variance-based monitor that anticipates the dip. This deployment-side adaptation is a non-trivial cost that should be weighed against the framework's empirical gains.

These limitations highlight opportunities for future work to validate and extend the proposed method under more demanding conditions and broader operational settings.

University of Ghana http://ugspace.ug.edu.gh

5.8 Practical Implications for Recommender Systems

The results in this thesis have several practical implications for the design and deployment of recommender systems, particularly those employing reinforcement learning frameworks in dynamic user environments.

First and foremost, the demonstrated effectiveness of Student's t-distributed noise highlights the importance of carefully modelling exploration strategies within actor–critic architectures. In practical deployments, exploration is often either neglected or handled through simplistic stochasticity, potentially limiting the ability of the system to discover novel user preferences or underrepresented items. The proposed approach shows that leveraging heavy-tailed distributions can significantly enhance exploration efficiency without compromising convergence stability or ranking accuracy. This suggests that production-grade recommendation engines can benefit from incorporating adaptive, tail-aware exploration schemes, particularly during cold-start phases or in rapidly evolving content domains.

Secondly, the robustness of the proposed noise strategy across datasets of varying scale and density underscores its potential generalisability. In commercial systems that serve diverse user populations and content types, ranging from sparse environments such as niche e-commerce platforms to dense ones such as video-on-demand services, a single, well-tuned exploration mechanism may reduce the need for extensive per-domain customisation. This contributes to reduced maintenance overhead and facilitates faster iteration cycles during algorithm updates.

Moreover, the ability of the Student's t-noise variant to improve both hit rate and rank-sensitive metrics like nDCG suggests it is well-suited for applications requiring high

University of Ghana http://ugspace.ug.edu.gh

precision in top-N recommendation lists. Such precision is critical in domains where user attention is limited, such as mobile app interfaces or smart TVs, where the top few items often dominate engagement metrics.

Additionally, the proposed method introduces minimal architectural changes to existing DDPG implementations, making it relatively straightforward to integrate into existing reinforcement learning pipelines. This lowers the barrier to adoption in industry settings, where deployment constraints and technical debt can impede experimentation with more complex modifications.

Finally, the empirical success of the Student's t-distribution-based noise strategy raises broader awareness of the role of exploration noise modelling in recommender systems. It encourages system designers to view exploration not merely as a side effect of stochastic environments, but as a tunable component that can drive meaningful improvements in personalisation, fairness, and content discovery.

In summary, this work provides actionable insights into how structured, heavy-tailed exploration can be harnessed to build smarter, more adaptive, and more effective recommender systems in real-world applications.

INTEGRI PROCEDAMUS

5.9 Lessons Learned

The course of this research surfaced five lessons that are specific to the present study and that may be useful to future researchers working on exploration in deep reinforcement learning for recommendation. They are arranged from the most concrete (specific empirical patterns) to the most general (methodological).

Lesson 1: Heavy-tailed exploration is the only mechanism among those tested that never harms performance relative to deterministic policy. The Jester result

University of Ghana http://ugspace.ug.edu.gh

(Chapter 5.5.1) is the most surprising single finding of this thesis. Gaussian and OU noise produce HR@10 and nDCG@10 values *below* the deterministic baseline on Jester D4. Heavy-tailed Student's t noise alone preserves the expected pattern of "exploration helps." This finding is relevant beyond the present thesis: it suggests that exploration mechanism choice is not a low-stakes hyperparameter, and that in some operating regimes (small action spaces, noisy reward signals) the conventional choices can be net-harmful.

Lesson 2: Bonferroni correction and non-parametric companions are essential when family-wise claims matter. Pairwise t-tests across a family of three baselines and two metrics constitute six comparisons, not one, and the family-wise error rate must be controlled for. On MovieLens 100K with ten seeds, the present analysis (Chapter 4.4.1) applies Bonferroni correction at family-wise $\alpha = 0.05$ (per-comparison threshold $\alpha_{\text{Bonf}} \approx 0.00833$) and reports both paired t-test and non-parametric Wilcoxon signed-rank p-values. All six pairwise comparisons survive both tests at the corrected threshold. Many recommender-system papers do not apply this correction or report only one test; the present thesis is an example of how the corrected and non-parametric analyses lend mutual support when both are applied. Researchers reviewing prior work should pay attention to whether the family-wise correction was applied and whether a distribution-free companion test was included.

Lesson 3: The bias-reduction interpretation predicts the MovieLens 1M dip. The training-dynamics dip on MovieLens 1M (Chapter 5.5.2) is not a defect of the framework but a transient signature of the bias-reduction mechanism operating on a sparse dataset: heavy-tailed perturbation widens the agent's empirical action coverage faster than the critic can accumulate training data in the broadened region, producing a window where the policy is steered by extrapolation rather than by reliable estimates. The lesson is

University of Ghana http://ugspace.ug.edu.gh

interpretive: empirical patterns that initially appear as instabilities can, in light of a theoretical mechanism, be re-read as evidence the mechanism is operating. The dip is a feature in this sense, but a feature with deployment costs that must be acknowledged.

Lesson 4: The framework's three design parameters have non-symmetric impact. The ablation studies (Chapter 4.3) showed that the framework's three design parameters (ν degrees of freedom, $\sigma_{\max}$ scale, T_{decay} schedule) have very different empirical sensitivities. The degrees of freedom ν has the largest impact on final HR@10 and nDCG@10; the maximum scale $\sigma_{\max}$ has a more modest effect; and the decay period T_{decay} has minimal impact within the tested range. A practitioner adopting the framework should therefore prioritise tuning ν over the other parameters. This non-symmetric sensitivity is consistent with the framework's theoretical intent: ν controls the distribution family, which is the framework's main intervention; $\sigma_{\max}$ and T_{decay} control the schedule, which is a secondary effect.

Lesson 5: Domain-specific adaptations are necessary for RL-based recommendation, but the bias-reduction perspective is portable. The methodology in this thesis required several domain-specific adaptations: the DRR-Ave state aggregation, the binarisation of continuous ratings, the leave-one-out evaluation protocol with 99-negative sampling. These are recommendation-specific; they would not transfer wholesale to a continuous-control task. The bias-reduction argument that motivates the framework, however, is portable: it applies to any actor–critic method with bootstrapped value estimation. The framework's empirical advantage on recommendation tasks therefore provides indirect evidence relevant to the general applicability of heavy-tailed exploration in actor–critic RL beyond the recommendation setting, and the present results add to the existing evidence in continuous control (Kobayashi, 2019) that the heavy-tailed advantage is a property of the

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

algorithmic mechanism rather than of the application domain.

5.10 Summary

This chapter has interpreted the empirical findings of Chapter 4 against the framework developed in Chapter 3. Three points are worth surfacing before Chapter 6 consolidates the hypothesis-by-hypothesis findings.

First, the empirical pattern fits the bias-reduction chain in three independent places. The chain predicts the deterministic baseline should perform worst, and it does on both MovieLens benchmarks. The chain predicts heavier-tailed alternatives should outperform lighter-tailed alternatives at matched variance, and the Student's t variant outperforms Gaussian and OU on every dataset. The chain predicts the size of the advantage should be larger on sparser datasets, and the relative gain on MovieLens 1M is indeed larger than on MovieLens 100K. The hypotheses are supported individually by the statistical analysis in Chapter 4; they are supported in a way that fits the proposed mechanism by the three-way agreement above.

Second, two boundary conditions complicate the headline and warrant explicit mention. The Jester underperformance anomaly (Chapter 5.5.1) shows that on small action spaces, light-tailed exploration becomes net-harmful because moderate-magnitude perturbations cross multiple decision boundaries in a compressed embedding space; only the heavy-tailed alternative's tail-concentration property protects against this disruption. The MovieLens 1M training-dynamics dip (Chapter 5.5.2) shows that on sparser datasets, the broadened empirical action distribution produces a transient performance degradation while the critic accumulates training data on the previously under-visited region of the action space. Both phenomena are consistent with the bias-reduction interpretation but bound

University of Ghana http://ugspace.ug.edu.gh

the framework's stability claim: the framework is effective under conditions that can be specified in advance, and the failure modes of light-tailed noise outside those conditions are themselves predictable.

Third, the statistical evidence on MovieLens 100K is now strong, but the cross-dataset analysis remains on a smaller sample. All six pairwise comparisons against Student's t-noise clear Bonferroni correction at family-wise $\alpha = 0.05$ under both the paired t-test and the Wilcoxon signed-rank test (Chapter 4.4.1), with large effect sizes and bootstrap confidence intervals that exclude zero. The MovieLens 1M and Jester D4 results in Chapter 4.5 continue to rely on three seeds with mean-only reporting; the directional pattern holds, but extending the ten-seed protocol to these datasets is the most direct way to lift the cross-dataset claims to the same standard.

Chapter 6 consolidates the hypothesis-by-hypothesis findings, states the contributions of the thesis formally, and outlines the future-work programme that the boundary conditions in particular motivate.

INTEGRI PROCEDAMUS

University of Ghana http://ugspace.ug.edu.gh

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Summary of Key Findings

This thesis investigated the design of effective exploration mechanisms for Deep Deterministic Policy Gradient (DDPG) when applied to sequential recommendation. The central proposition is that a parameterised heavy-tailed stochastic exploration framework, instantiated as Student's t-distributed action perturbation with scale annealing and value clipping, provides a principled actor-side intervention against the overestimation-bias channel that affects DDPG, and produces measurable improvements over the conventional Gaussian and Ornstein–Uhlenbeck (OU) noise baselines on standard recommendation benchmarks. The four research questions stated in Chapter 1.7 have been addressed against the empirical evidence developed in Chapters 4 and 5; the consolidated answers are summarised below.

RQ1: Manifestation of overestimation bias. Overestimation bias in DDPG-based recommendation manifests primarily as a failure of the deterministic policy to explore the action space adequately, leading to systematically low HR@10 and nDCG@10 relative to noise-augmented variants. The deterministic baseline is the weakest of the four configurations on both MovieLens benchmarks, consistent with the bias-channel argument developed in Chapter 3.6.3: without exploration, the actor preferentially moves toward whichever action region the critic happens to overestimate, reinforcing the bias. The proposed framework addresses this manifestation indirectly via broader empirical action-space coverage rather than directly via critic regularisation as in TD3.

University of Ghana http://ugspace.ug.edu.gh

RQ2 and Hypothesis H1: Limitations of light-tailed noise. Light-tailed exploration noise produces measurably lower HR@10 and nDCG@10 than the heavy-tailed Student's t alternative across all three benchmark datasets. On MovieLens 100K with ten random seeds (Chapter 4.4.1), all six pairwise comparisons (three baselines $\times$ two metrics) survive Bonferroni correction at family-wise $\alpha = 0.05$ under both the paired t-test and the non-parametric Wilcoxon signed-rank test, with large effect sizes (Cohen's $d > 1.0$ in every comparison) and bootstrap 95% confidence intervals on the mean difference that exclude zero in every comparison. On MovieLens 1M and Jester D4, the directional advantage is consistent on a smaller three-seed sample; an equivalent statistical analysis on those datasets is identified as future work. Hypothesis H1 is therefore supported on the principal benchmark and directionally consistent on the others.

RQ3 and Hypothesis H2: Heavy-tailed processes for exploration. The bounded scale-annealed Student's t formulation produces measurable improvements in HR@10 and nDCG@10 across all three benchmark datasets. The convergence-stability element of the hypothesis is partially supported: convergence is faster and smoother than the alternatives on MovieLens 100K and Jester D4, but the MovieLens 1M training dynamics exhibit a substantial intermediate-stage performance dip before recovery. Hypothesis H2 is therefore partially supported, with the stability claim qualified as dataset-dependent rather than universal.

RQ4 and Hypothesis H3: Performance and generalisation of t-DDPG. The proposed t-DDPG outperforms vanilla DDPG, Gaussian-DDPG, and OU-DDPG on both metrics on all three benchmark datasets. On MovieLens 100K, the advantage is statistically significant against every baseline at the Bonferroni-corrected family-wise level under both the t-test and the Wilcoxon signed-rank test on a ten-seed sample. The strongest

University of Ghana http://ugspace.ug.edu.gh

defensible generalisation claim is that t-DDPG is the only configuration of the four to deliver an exploration advantage over the deterministic baseline on every dataset tested; Gaussian and OU achieve this on the MovieLens benchmarks but fail on Jester D4. The absolute size of the advantage varies materially across datasets, so the generalisation is in the qualitative pattern rather than in the magnitude of the gain. Hypothesis H3 is therefore supported on the principal benchmark and directionally consistent on the others.

Boundary conditions identified. Two empirically observed phenomena bound the framework's applicability and warrant explicit mention. The Jester underperformance anomaly (Chapter 5.5.1) shows that on small action spaces with noisy reward signals, light-tailed exploration can be net-harmful relative to a deterministic policy; the heavy-tailed alternative is the only mechanism among those tested to deliver a positive gain in this regime. The MovieLens 1M training-dynamics dip (Chapter 5.5.2) shows that on sparser datasets, the bias-reduction mechanism produces a transient performance degradation during training as the critic accumulates training data in the broadened region of the action space. Both phenomena are consistent with the bias-reduction interpretation but bound the framework's stability claim and its drop-in replaceability for production deployments.

Bottom line. The thesis establishes that heavy-tailed action perturbation, instantiated through the parameterised exploration framework as t-DDPG, is an effective and theoretically grounded actor-side intervention on the overestimation-bias channel in DDPG-based sequential recommendation. The empirical advantage on MovieLens 100K is statistically significant against every baseline at the Bonferroni-corrected family-wise level on a ten-seed sample, under both the paired t-test and the non-parametric Wilcoxon signed-rank test; the directional advantage holds on MovieLens 1M and Jester D4 with a three-seed sample, and extending the same standard of analysis to those datasets is the

University of Ghana http://ugspace.ug.edu.gh

most direct continuation of the work. The framework's applicability has dataset-specific boundary conditions that future deployments must accommodate, but the core mechanism that justifies the approach — the bias-reduction chain developed in Chapter 3.6.3 — applies to actor–critic methods broadly and is the basis for the future-work programme outlined in Chapter 6.4.

6.2 Contributions of the Thesis

The thesis makes four categories of contribution to the literature on exploration in actor–critic reinforcement learning and on reinforcement-learning-based recommender systems.

Theoretical contributions. The thesis develops a parameterised heavy-tailed exploration framework for deterministic-policy actor–critic methods (Chapter 3.6.1) that specifies three design parameters, the noise distribution family, the scale schedule, and the bound region. That supports principled comparison among heavy-tailed alternatives. Within this framework, the thesis develops a bias-reduction argument (Chapter 3.6.3) that connects heavy-tailed action perturbation to overestimation-bias mitigation through a three-effect causal chain: heavy-tailed perturbation widens empirical action-space coverage, which produces critic training data that is more representative of the action space the actor might visit, which yields a more accurate policy gradient. This argument provides a theoretically grounded reason to prefer heavy-tailed exploration over light-tailed alternatives in actor–critic methods.

Methodological contributions. The thesis instantiates the framework as t-DDPG, with Student's t-distributed action perturbation, linear scale annealing, and box clipping. Three methodological elements support reproducibility of the work: a complete end-to-end training-step walk-through that traces a single episode through the entire system (Chapter 3.7.7); the offline interaction-simulator specification that documents the leave-

University of Ghana http://ugspace.ug.edu.gh

one-out evaluation protocol with 99-negative sampling (Chapter 3.7.1); and the explicit specification of the network architectures, optimiser configuration, and initialisation schemes (Chapter 3.5.3). These elements together constitute a methodologically explicit description of the framework that supports independent replication.

Empirical contributions. The thesis presents a controlled evaluation of the framework on three standard recommendation benchmarks (MovieLens 100K, MovieLens 1M, Jester D4) against three baseline configurations (no-noise DDPG, Gaussian-DDPG, OU-DDPG). The evaluation includes ablation studies on the framework's three design parameters and a comprehensive statistical analysis on the principal benchmark (Chapter 4.4.1): paired t-tests with Bonferroni correction at family-wise $\alpha = 0.05$, non-parametric Wilcoxon signed-rank tests as a distribution-free companion, Cohen's d effect sizes, and bootstrap 95% confidence intervals on the mean differences. Two empirically observed phenomena that the original analysis would have glossed over — the Jester underperformance anomaly and the MovieLens 1M training-dynamics dip — are reported honestly and analysed in depth in Chapter 5.5.1 and 5.5.2. The empirical contributions therefore include both the headline results (t-DDPG outperforms the baselines on every dataset tested, with all six pairwise comparisons on MovieLens 100K surviving Bonferroni correction at family-wise $\alpha = 0.05$) and the boundary-condition findings (t-DDPG is the only configuration to deliver a positive gain over the deterministic baseline on every dataset; the stability advantage is dataset-dependent rather than universal).

Conceptual contribution. The framework's bias-reduction argument is portable beyond sequential recommendation: it applies to any actor–critic method with bootstrapped value estimation. The thesis's empirical results on recommendation, taken together with the existing evidence in continuous control on the heavy-tailed advantage, support a more

University of Ghana http://ugspace.ug.edu.gh

general claim than either body of evidence alone: heavy-tailed exploration is a property of the algorithmic mechanism rather than of the application domain. This conceptual contribution motivates the future-work programme on extending the framework to other actor–critic methods (Chapter 6.4.1) and is the basis for the programmatic claim in Chapter 6.5 that the framework's relevance extends beyond recommendation.

6.3 Research Limitations

The thesis's contributions are bounded by limitations that have been examined in detail in Chapter 5.7. The present section synthesises those limitations briefly to set up the future-work programme that follows.

The empirical-evidence limitations are the moderate scale of the three benchmark datasets (MovieLens 100K, MovieLens 1M, and Jester D4 are widely used in academic work but do not reflect the scale and heterogeneity of industrial recommender systems), the three-seed coverage on the cross-dataset evaluation and the ablations (which keeps statistical significance on MovieLens 1M and Jester D4 below the standard achieved on MovieLens 100K), and the CPU-only hardware (which constrained the breadth of hyperparameter sweeps and the dataset scales that could be evaluated). Source seed data and the experimental infrastructure are both available, so the second of these is the most directly addressable in the near term and is the subject of Chapter 6.4.2 below.

The scope-of-claim limitations are the absence of comparison against TD3 and SAC (which intervene on the bias channel through different mechanisms and represent the natural next-generation baselines for a heavy-tailed-exploration claim), the absence of online evaluation against live users (which constrains the present empirical claims to offline-replay performance and leaves the framework's behaviour in true online settings untested),

University of Ghana http://ugspace.ug.edu.gh

and the dataset-specific nature of the framework's stability advantage (the MovieLens 1M training-dynamics dip qualifies the framework as not a drop-in replacement for production pipelines that use single-window early-stopping rules). Each of these scope limitations motivates a specific direction in the future-work programme below.

6.4 Future Work

The findings of this thesis open a programme of future research that addresses both the limitations identified above and the new directions that the proposed framework opens. The subsections below are arranged in priority order, with the highest-priority items being the empirical extensions that examiner feedback specifically requested.

6.4.1 Extension to TD3, SAC, and PPO

The highest-priority direction for future work is the extension of the framework to actor–critic methods beyond DDPG. The proposed framework intervenes on the actor side (heavy-tailed action perturbation widens empirical action-space coverage); TD3 (Fujimoto et al., 2018) intervenes on the critic side (twin critics with minimum-target bootstrapping); SAC (Haarnoja et al., 2018) adds maximum-entropy regularisation that biases the policy toward stochasticity; and PPO (Schulman et al., 2017) uses a fundamentally different policy-gradient family (clipped surrogate objective with on-policy updates). Because these mechanisms operate at different stages of the bias channel or use different gradient estimation entirely, they are predicted to interact differently with heavy-tailed action perturbation.

The most direct test is a "t-TD3" combined variant that uses both interventions simultaneously: TD3's twin-critic mechanism with the framework's heavy-tailed action perturbation in place of the standard Gaussian noise that TD3's target-policy smoothing

University of Ghana http://ugspace.ug.edu.gh

typically employs. The bias-reduction argument predicts that t-TD3 should outperform both t-DDPG and TD3 alone, because each intervention addresses a different stage of the bias channel and the two interventions should be approximately additive in their effect. The same evaluation methodology used in this thesis (the offline interaction simulator from Chapter 3.7.1; the leave-one-out evaluation protocol with 99-negative sampling from Chapter 3; and the Bonferroni-corrected statistical analysis from Chapter 4.4.1) supports a direct extension to t-TD3 with minimal methodological modification.

The corresponding "t-SAC" extension is a less direct fit because SAC's stochastic policy already produces non-deterministic action selection through the entropy term in its objective. The natural way to combine the framework with SAC is therefore to use the heavy-tailed distribution as the policy's action distribution (rather than the standard Gaussian SAC uses), which represents a more substantial change to SAC's formulation than the t-TD3 modification represents to TD3. The "t-PPO" extension is similarly non-trivial because PPO's on-policy updates do not use a replay buffer, so the bias-reduction argument's "broader empirical coverage" mechanism does not apply in the same form. Both t-SAC and t-PPO are therefore secondary to t-TD3 in priority, but each is a natural follow-on study and would test whether the framework's advantage is a property of off-policy methods specifically or of actor–critic methods more generally.

6.4.2 Statistical Sample Expansion to MovieLens 1M and Jester D4

The ten-seed statistical analysis of Chapter 4.4.1 establishes Bonferroni-corrected family-wise significance at $\alpha = 0.05$ for all six pairwise comparisons on MovieLens 100K. The cross-dataset evaluation on MovieLens 1M and Jester D4 (Chapter 4.5) continues to rely on a three-seed sample with mean-only reporting, and the four ablation studies (Chapter 4.3) similarly use a three-seed protocol. The most direct continuation of this thesis is to

University of Ghana http://ugspace.ug.edu.gh

extend the ten-seed standard to those settings.

The expanded analysis should include three elements. First, the paired t-tests and Wilcoxon signed-rank tests of the present MovieLens 100K analysis should be replicated on MovieLens 1M and Jester D4 at ten seeds with Bonferroni correction maintained; this would either upgrade the cross-dataset claims to the same level of family-wise significance established on MovieLens 100K or refine the directional findings by quantifying the effect sizes more precisely. Second, the four ablation studies (decay period, maximum noise scale, minimum noise scale, degrees of freedom) should be re-run under the ten-seed protocol, with mean $\pm$ std reporting throughout; this would tighten the parameter-selection evidence and allow Cohen's d effect sizes to be reported for the ablation comparisons. Third, bootstrap confidence intervals on every reported metric should be propagated through to the cross-dataset and ablation tables, replacing the present mean-only reporting. The experimental infrastructure used for the MovieLens 100K ten-seed analysis is the same infrastructure required for these extensions, so the principal cost is compute rather than design.

6.4.3 Critic-Stability Ablation for the MovieLens 1M Dip

The MovieLens 1M training-dynamics dip (Chapter 4.6.2; Chapter 5.5.2) is interpreted in this thesis as a transient signature of the bias-reduction mechanism operating on a sparse dataset. The interpretation is consistent with the empirical pattern but does not directly test the mechanism. Future work should perform a direct test by tracking the critic's Q-value variance (or an equivalent extrapolation-error proxy such as the disagreement between the online critic and the target critic) across training, and comparing this signal across the four noise configurations.

University of Ghana http://ugspace.ug.edu.gh

The bias-reduction argument makes a clean falsifiable prediction: during the dip period on MovieLens 1M, the t-DDPG critic should exhibit temporarily higher Q-value variance than the Gaussian and OU critics; subsequently, after the dip recovery, the t-DDPG critic should exhibit lower Q-value variance than the alternatives, reflecting the accumulated training data in the broader region of the action space. If both elements of this pattern are observed, the bias-reduction interpretation is empirically supported as the operating mechanism. If the pattern is not observed (for example, if the t-DDPG critic's variance is uniformly higher across training, or if the variance pattern does not align with the dip period), an alternative mechanism is operating and the bias-reduction interpretation must be revised. The experimental design is straightforward: the same training runs that produced the present results can be re-instrumented to log the critic-variance signal at each evaluation step, and the resulting trajectories can be compared statistically across configurations.

6.4.4 Action-Space-Size Boundary Investigation

The Jester underperformance anomaly (Chapter 4.5; Chapter 5.5.1) suggests that the framework's relative advantage over light-tailed alternatives depends on the action-space size. The thesis offers two candidate explanations (small action spaces have closer item-cluster boundaries that light-tailed perturbation crosses disruptively; binarised continuous-rating signals are noisier near the binarisation threshold) but the available empirical evidence is consistent with multiple combinations of these factors and does not isolate them.

Future work should investigate the boundary condition systematically by running the framework's comparison on a series of synthetic recommendation environments with parametrically varied catalogue size, holding all other factors (state representation

University of Ghana http://ugspace.ug.edu.gh

dimensionality, embedding training procedure, reward-signal density, training budget) constant. The bias-reduction argument combined with the small-action-space interpretation predicts a U-shaped pattern in the relative advantage of Student's t over Gaussian and OU as a function of catalogue size: the advantage should be small at very small catalogue sizes (Jester regime, where light-tailed noise causes harmful boundary-crossings and the heavy-tailed alternative recovers the deterministic baseline rather than substantially exceeding it); largest at moderate catalogue sizes (MovieLens regime, where the bias-reduction mechanism operates most cleanly); and diminishing at very large catalogue sizes (where any reasonable exploration finds useful regions and the marginal benefit of heavier tails saturates). A controlled investigation of this U-shape would either confirm the small-action-space interpretation or identify an alternative explanation for the Jester pattern, and would also produce practitioner-relevant guidance on when the framework's adoption is most likely to deliver gains.

6.4.5 Action-Level vs Parameter-Level Heavy-Tailed Comparison

The thesis chose action-level perturbation over parameter-level alternatives such as those proposed by Plappert et al. (2017) and Fortunato et al. (2017), with the motivations developed in Chapter 3.6.4. The empirical advantage demonstrated in Chapter 4 is therefore an advantage of *heavy-tailed* action-level perturbation over *light-tailed* action-level perturbation; the advantage of action-level over parameter-level perturbation independent of the distribution choice is not directly tested.

A direct comparison would evaluate four configurations: action-level Gaussian (the present Gaussian baseline), action-level Student's t (the present t-DDPG variant), parameter-level Gaussian (Plappert et al. (2017)'s formulation), and parameter-level Student's t (a new combination). This 2×2 comparison would isolate the contribution of the

University of Ghana http://ugspace.ug.edu.gh

distribution choice from the contribution of the intervention-point choice. Two outcomes are theoretically possible: either the heavy-tailed advantage is consistent across both intervention points (which would support the framework's argument that the distribution family is the dominant design choice), or the advantage interacts with the intervention point (which would refine the framework's claim and motivate a more careful analysis of how parameter-level perturbation maps to action-level perturbation through the actor network's local Jacobian).

6.4.6 Real-World Deployment and Online Learning

The thesis evaluates the framework on offline replay of logged user interactions (Chapter 3.7.1). The framework's behaviour in true online deployment with live users is therefore not directly tested. Bridging this gap is a substantial direction of future work that combines methodological challenges (designing an evaluation protocol that allows safe online testing of a new exploration mechanism) with engineering challenges (latency, scale, A/B-testing infrastructure).

Three deployment-specific risks of heavy-tailed exploration warrant explicit attention in any online study. First, the framework's rare large-magnitude excursions, which are the source of its empirical advantage in offline evaluation, may produce occasional surprising or inappropriate item recommendations in user-facing settings; the bound region $\mathcal{C}$ in the framework formulation (Chapter 3.6.1) provides one mitigation, but a production deployment may want additional safeguards such as an ε-greedy fallback for users in known sensitive contexts. Second, the MovieLens 1M training-dynamics dip identified in Chapter 4.6.2 would translate to a window of user-facing performance degradation during online training, with associated business cost; the deployment-side adaptations discussed in Chapter 5.5.2 (extended training budget, longer-window early-stopping

University of Ghana http://ugspace.ug.edu.gh

rules, critic-variance-based monitoring) should be empirically validated in this online setting. Third, the offline-trained framework's transfer to online dynamics where user preferences may shift over time is not characterised by the present experiments; an online study would need to evaluate whether the framework's exploration advantage persists under non-stationary user preferences.

A natural staged programme would begin with a controlled A/B test on a small fraction of traffic against the production baseline (whatever production noise mechanism is currently in place), expand traffic gradually as the framework's online behaviour is characterised, and conclude with full deployment if the offline gains translate to online lift on the production metrics of interest. The same observability infrastructure required for the critic-stability ablation (Chapter 6.4.3) would support the online evaluation, providing further reason to prioritise that ablation.

6.4.7 Sensitivity to Binarisation Thresholds

The thesis adopts conventional binarisation thresholds for the three datasets (≥ 4 on the 1–5 MovieLens scale; > 0 on the -10 to $+10$ Jester scale). These choices are standard in the recommender-system literature but are not validated for the framework specifically. A sensitivity analysis that varies the threshold parametrically. MovieLens binarisation at 3, 3.5, 4, and 4.5; Jester binarisation at -2, -1, 0, $+1$, $+2$ would confirm whether the framework's advantage is robust to the binarisation choice or whether it depends on a particular reward-signal density. The Jester anomaly discussed in Chapter 6.4.4 suggests this sensitivity may be substantial, since the binarisation choice directly determines how often light-tailed exploration crosses a reward-signal-flip boundary. This is a brief subsidiary study that the present infrastructure can run with no methodological change beyond the threshold parameter.

University of Ghana http://ugspace.ug.edu.gh

6.4.8 Adaptive and Context-Aware Noise Models

The parameterised framework formulation in Chapter 3.6.1 specifies the noise distribution family as a design parameter $\mathcal{P}_\nu$, and the thesis instantiates this parameter as the Student's t-distribution with $\nu = 3$. Other heavy-tailed alternatives within the same framework are natural candidates for evaluation: the Cauchy distribution (Student's t with $\nu = 1$, infinite second moment), the Pareto distribution (a power-law with explicit tail index control), and possibly bounded variants of these formulations to retain the framework's value-clipping property.

A more substantive extension is adaptive scheduling of the framework's parameters during training based on a runtime signal. The current annealing schedule (Chapter 3.6.4) is a fixed function of training step; the bias-reduction interpretation suggests a more responsive schedule that uses the critic's Q-value variance as a feedback signal, increasing the noise scale when the critic is well-calibrated and reducing it during periods of high critic uncertainty. This adaptive variant is theoretically motivated by the same bias-reduction argument that motivates the framework as a whole, and the critic-variance instrumentation required for the ablation in Chapter 6.4.3 provides exactly the runtime signal an adaptive scheduler would consume. The adaptive variant therefore combines naturally with the critic-stability ablation as a single integrated direction of future work.

6.5 Closing Remarks

This thesis has developed a parameterised heavy-tailed exploration framework for actor–critic reinforcement learning, instantiated it as t-DDPG, motivated it through a bias-reduction argument that connects heavy-tailed action perturbation to overestimation-bias mitigation, and validated it empirically on three benchmark recommender-system datasets.

University of Ghana http://ugspace.ug.edu.gh

The empirical advantage of t-DDPG over the Gaussian, Ornstein–Uhlenbeck, and no-noise baselines is the most direct contribution of the work, but two further contributions warrant emphasis in closing.

The first is the honest reporting of the boundary conditions that bound the framework's applicability. The Jester underperformance anomaly and the MovieLens 1M training-dynamics dip are not failures of the framework but characterisations of where it operates well and where it requires additional care. A research community that systematically reports such boundary conditions can build cumulative knowledge about when methods work; one that reports only headline successes accumulates the brittleness that produces irreproducibility crises. The Bonferroni-corrected statistical analysis in Chapter 4.4.1 extends the same epistemic discipline to the statistical-significance claims, distinguishing the four pairwise comparisons that survive family-wise correction from the two that do not.

The second is the conceptual broadening that the bias-reduction argument supports. The framework's specific empirical evidence is from sequential recommendation. Heavy-tailed perturbation widens empirical action coverage, broader coverage produces more representative critic training data, more representative training data yields a more accurate policy gradient. The framework's relevance therefore extends beyond the recommendation setting in a principled way that is grounded in the algorithmic mechanism rather than in surface-level analogy. The future-work programme in Chapter 6.4 identifies the actor–critic methods (TD3, SAC, PPO) where this broader claim should be empirically tested.

The thesis as a whole supports the position that exploration mechanism choice in actor–critic reinforcement learning is a first-class design decision rather than a low-stakes

University of Ghana http://ugspace.ug.edu.gh

hyperparameter, and that heavy-tailed alternatives merit the same level of theoretical and empirical analysis as the network architectures and reward formulations that typically receive the bulk of design attention in the field. The proposed framework is one contribution to this broader research programme; the future-work directions identified above outline the next steps the contribution motivates.



University of Ghana http://ugspace.ug.edu.gh

REFERENCES

[1] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). {Tensorflow}: A system for {large-scale} machine learning. *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, 265–283.

[2] Adomavicius, G., & Tuzhilin, A. (2010). Context-aware recommender systems. In *Recommender systems handbook* (pp. 217–253). Springer.

[3] Afsar, M. M., Crump, T., & Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, *55*(7), 1–38. https://doi.org/10.1145/3543846

[4] Aggarwal, C. C., & Aggarwal, C. C. (2016). An introduction to recommender systems. *Recommender systems: the textbook*, 1–28.

[5] Akanksha, E., Jyoti, Sharma, N., & Gulati, K. (2021). Review on reinforcement learning, research evolution and scope of application. *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, 1416–1423. https://api.semanticscholar.org/CorpusID:233991489

[6] Alhijawi, B., Awajan, A. A., & Fraihat, S. (2022). Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and new perspectives. *ACM Computing Surveys*, *55*, 1–38. https://api.semanticscholar.org/CorpusID:247795044

[7] Altulyan, M. (2020). Recommender systems for the internet of things: A survey. https://doi.org/10.48550/arxiv.2007.06758

University of Ghana http://ugspace.ug.edu.gh

[8] Anandhan, A., Shuib, L., Ismail, M. A., & Mujtaba, G. (2018). Social media recommender systems: Review and open research issues. *IEEE Access*, *6*, 15608–15628.

[9] Ansari, A., Essegaier, S., & Kohli, R. (2000). Internet recommendation systems.

[10] Asabere, N. Y., Acakpovi, A., & Michael, M. B. (2017). Improving socially-aware recommendation accuracy through personality. *IEEE Transactions on Affective Computing*, *9*(3), 351–361.

[11] Ashwin, S., & Naveen Raj, R. (2023). Deep reinforcement learning for autonomous vehicles: Lane keep and overtaking scenarios with collision avoidance. *International Journal of Information Technology*, *15*(7), 3541–3553.

[12] Baltrunas, L., Ludwig, B., & Ricci, F. (2011). Matrix factorization techniques for context aware recommendation. *Proceedings of the fifth ACM conference on Recommender systems*, 301–304.

[13] Becattini, F., & Uricchio, T. (2022). Memory networks. *Proceedings of the 30th ACM International Conference on Multimedia*, 7380–7382.

[14] Bedi, A. S., Parayil, A., Zhang, J., Wang, M., & Koppel, A. (2024). On the sample complexity and metastability of heavy-tailed policy search in continuous control. *Journal of Machine Learning Research*, *25*(39), 1–58. https://jmlr.org/papers/v25/21-1343.html

[15] Bellemare, M. G., Dabney, W., & Munos, R. (2017). A distributional perspective on reinforcement learning. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 449–458.

[16] Bridge, D., Göker, M. H., McGinty, L., & Smyth, B. (2005). Case-based recommender systems. *The Knowledge Engineering Review*, *20*(3), 315–320.

University of Ghana http://ugspace.ug.edu.gh

[17] Campos, P. G., Díez, F., & Cantador, I. (2014). Time-aware recommender systems: A comprehensive survey and analysis of existing evaluation protocols. *User Modeling and User-Adapted Interaction*, *24*, 67–119.

[18] Cayci, S., & Eryilmaz, A. (2023). Provably robust temporal difference learning for heavy-tailed rewards. *Advances in Neural Information Processing Systems 36 (NeurIPS)*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/516ca2e9e7bffbb4027a25d9f8838bc9-Abstract-Conference.html

[19] Chakraborty, S., Bedi, A. S., Koppel, A., Tokekar, P., & Manocha, D. (2023). Dealing with sparse rewards in continuous control robotics via heavy-tailed policy optimization. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 3335–3341. https://doi.org/10.1109/ICRA48891.2023.10161186

[20] Chakraverty, S., & Mithal, A. (2018). Iot based weather and location aware recommender system. *2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 636–643.

[21] Chen, M., Beutel, A., Covington, P., Jain, S., Belletti, F., & Chi, E. H. (2019). Top-k off-policy correction for a reinforce recommender system. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 456–464.

[22] Cheng, M., Yuan, F., Liu, Q., Ge, S., Li, Z., Yu, R., Lian, D., Yuan, S., & Chen, E. (2021). Learning recommender systems with implicit feedback via soft target enhancement. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 575–584.

[23] Chu, Y., Yang, H.-K., & Peng, W.-C. (2019). Predicting online user purchase behavior based on browsing history. *2019 IEEE 35th International Conference on*

University of Ghana http://ugspace.ug.edu.gh

Data Engineering Workshops (ICDEW), 185–192. https://api.semanticscholar.org/CorpusID:195775032

[24] Covington, P., Adams, J., & Sargin, E. (2016). Deep neural networks for YouTube recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys)*, 191–198. https://doi.org/10.1145/2959100.2959190

[25] Dhavala, S. S., Srihari, C., Vanishree, K., & Rashmi, R. S. (2023). An extensive review of applications, methods and recent advances in deep reinforcement learning. *2023 5th International Congress on Human-Computer Interaction, Optimisation and Robotic Applications (HORA)*, 1–6. https://api.semanticscholar.org/CorpusID:259266630

[26] Dhelim, S., Ning, H., Aung, N., Huang, R., & Ma, J. (2020). Personality-aware product recommendation system based on user interests mining and metapath discovery. *IEEE Transactions on Computational Social Systems*, *8*(1), 86–98.

[27] Ding, J., Yu, G., Li, Y., He, X., & Jin, D. (2020). Improving implicit recommender systems with auxiliary data. *ACM Transactions on Information Systems (TOIS)*, *38*(1), 1–27.

[28] Dong, D., Zheng, X., Zhang, R., & Wang, Y. (2018). Recurrent collaborative filtering for unifying general and sequential recommender. *IJCAI*, 3350–3356.

[29] Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64. https://doi.org/10.1080/01621459.1961.10482090

[30] Eberhard, O., Hollenstein, J., Pinneri, C., & Martius, G. (2023). Pink noise is all you need: Colored noise exploration in deep reinforcement learning [Spotlight].

University of Ghana http://ugspace.ug.edu.gh

International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=hQ9V5QN27eS

[31] Ekstrand, M. D., Riedl, J. T., Konstan, J. A., et al. (2011). Collaborative filtering recommender systems. *Foundations and Trends® in Human-Computer Interaction*, *4*(2), 81–173.

[32] Fortunato, M., Azar, M. G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., et al. (2017). Noisy networks for exploration. *arXiv preprint arXiv:1706.10295*.

[33] François-Lavet, V., Henderson, P., Islam, R., Bellemare, M. G., Pineau, J., et al. (2018). An introduction to deep reinforcement learning. *Foundations and Trends® in Machine Learning*, *11*(3-4), 219–354.

[34] Fujimoto, S., Chang, W.-D., Smith, E. J., Gu, S. S., Precup, D., & Meger, D. (2023). For SALE: State-action representation learning for deep reinforcement learning. *Advances in Neural Information Processing Systems 36 (NeurIPS)*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/c20ac0df6c213db6d3a930fe9c7296c8-Abstract-Conference.html

[35] Fujimoto, S., Hoof, H., & Meger, D. (2018). Addressing function approximation error in actor-critic methods. *International conference on machine learning*, 1587–1596.

[36] Gao, C., Huang, K., Chen, J., Zhang, Y., Li, B., Jiang, P., Wang, S., Zhang, Z., & He, X. (2023). Alleviating matthew effect of offline reinforcement learning in interactive recommendation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 238–248. https://doi.org/10.1145/3539618.3591636

University of Ghana http://ugspace.ug.edu.gh

[37] Gao, T., Gao, S., Xu, J., & Zhao, Q. (2023). Ddrcn: Deep deterministic policy gradient recommendation framework fused with deep cross networks. *Applied Sciences*, *13*(4), 2555. https://doi.org/10.3390/app13042555

[38] Garg, S., Zhanson, J., Parisotto, E., Prasad, A., Kolter, J. Z., Lipton, Z. C., Balakrishnan, S., Salakhutdinov, R., & Ravikumar, P. (2021). On proximal policy optimization's heavy-tailed gradients. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, *139*, 3610–3619. https://proceedings.mlr.press/v139/garg21b.html

[39] Gena, C., Brogi, R., Cena, F., & Vernero, F. (2011). The impact of rating scales on user's rating behavior. *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, 123–134.

[40] Ghafari, S. M., Beheshti, A., Joshi, A., Paris, C., Mahmood, A., Yakhchi, S., & Orgun, M. A. (2020). A survey on trust prediction in online social networks. *IEEE Access*, *8*, 144292–144309.

[41] Głogowska, J., Kobiela, D., & Mielewczyk, S. (2024). Session-based recommendation with graph neural networks with an examination of the impact of local and global vectors. *European Conference on Advances in Databases and Information Systems*, 263–272.

[42] Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70.

[43] Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems*, *6*(4), 1–19. https://doi.org/10.1145/2843948

University of Ghana http://ugspace.ug.edu.gh

[44] Guerreiro, J., & Rita, P. (2020). How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *Journal of Hospitality and Tourism Management*. https://api.semanticscholar.org/CorpusID:199174490

[45] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, *80*, 1861–1870.

[46] Hazrati, N., & Ricci, F. (2024). Choice models and recommender systems effects on users' choices. *User Modeling and User-Adapted Interaction*, *34*(1), 109–145.

[47] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 1026–1034. https://doi.org/10.1109/ICCV.2015.123

[48] He, R., Kang, W.-C., & McAuley, J. (2016). Translation-based recommendation. *Proceedings of the 11th ACM Conference on Recommender Systems*, 161–169.

[49] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T.-S. (2017). Neural collaborative filtering. *Proceedings of the 26th international conference on world wide web*, 173–182.

[50] Hidasi, B., Karatzoglou, A., Baltrunas, L., & Tikk, D. (2015). Session-based recommendations with recurrent neural networks. *arXiv preprint arXiv:1511.06939*.

[51] Hidasi, B., & Tikk, D. (2016). General factorization framework for context-aware recommendations. *Data Mining and Knowledge Discovery*, *30*(2), 342–371.

[52] Hollenstein, J. J., Auddy, S., Saveriano, M., Renaudo, E., & Piater, J. (2022). Action noise in off-policy deep reinforcement learning: Impact on exploration and

University of Ghana http://ugspace.ug.edu.gh

performance [arXiv preprint arXiv:2206.03787]. *Transactions on Machine Learning Research*, *2022*, 1–33. https://openreview.net/forum?id=NljBlZ6hmG

[53] Huang, J., & Jia, Z. (2022). Personalized recommendation algorithm based on rating system and user interest association network. https://doi.org/10.4236/jamp.2022.1012231

[54] Huber, P. J. (2004). *Robust statistics* (Reprint). John Wiley & Sons.

[55] Ie, E., Jain, V., Wang, J., Narvekar, S., Agarwal, R., Wu, R., Cheng, H.-T., Chandra, T., & Boutilier, C. (2019). Slateq: A tractable decomposition for reinforcement learning with recommendation sets. *IJCAI*, *19*, 2592–2599.

[56] Ihler, A. T., Kirshner, S., Ghil, M., Robertson, A. W., & Smyth, P. (2007). Graphical models for statistical inference and data assimilation. *Physica D: Nonlinear Phenomena*, *230*(1-2), 72–87.

[57] Ivanov, S., & D'yakonov, A. (2019). Modern deep reinforcement learning algorithms. *ArXiv*, *abs/1906.10025*. https://api.semanticscholar.org/CorpusID:195344946

[58] Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446. https://doi.org/10.1145/582415.582418

[59] Ji, L., Liu, G., Yin, M., Yang, H., & Zhou, J. (2024). Hierarchical reinforcement learning for temporal abstraction of listwise recommendation. *arXiv preprint arXiv:2409.07416*.

[60] Jia, J., & Wang, W. (2020). Review of reinforcement learning research. *2020 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC)*, 186–191. https://api.semanticscholar.org/CorpusID:231851421

University of Ghana http://ugspace.ug.edu.gh

[61] Joachims, T., Freitag, D., Mitchell, T., et al. (1997). Webwatcher: A tour guide for the world wide web. *IJCAI (1)*, 770–777.

[62] Kang, W.-C., & McAuley, J. (2018). Self-attentive sequential recommendation. *2018 IEEE international conference on data mining (ICDM)*, 197–206.

[63] Khanal, S. S., Prasad, P., Alsadoon, A., & Maag, A. (2020). A systematic review: Machine learning based recommendation systems for e-learning. *Education and Information Technologies*, *25*(4), 2635–2664.

[64] Khusro, S., Ali, Z., & Ullah, I. (2016). Recommender systems: Issues, challenges, and research opportunities. https://api.semanticscholar.org/CorpusID:63453181

[65] Kobayashi, T. (2019). Student-t policy in reinforcement learning to acquire global optimum of robot control. *Applied Intelligence*, *49*(12), 4335–4347. https://doi.org/10.1007/s10489-019-01510-8

[66] Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., & Riedl, J. (1997). Grouplens: Applying collaborative filtering to usenet news. *Communications of the ACM*, *40*(3), 77–87.

[67] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

[68] Lample, G., & Chaplot, D. S. (2017). Playing fps games with deep reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

[69] Lara-Cabrera, R., González-Prieto, Á., & Ortega, F. (2020). Deep matrix factorization approach for collaborative filtering recommender systems. *Applied Sciences*, *10*(14), 4926.

University of Ghana http://ugspace.ug.edu.gh

[70] Lee, D., & Powell, W. B. (2019). Bias-corrected q-learning with multistate extension. *IEEE Transactions on Automatic Control*, *64*(10), 4011–4023.

[71] Lei, Y., & Li, W. (2019). When collaborative filtering meets reinforcement learning. *arXiv preprint arXiv:1902.00715*.

[72] Li, B., Lu, T., Li, J., Lu, N., Cai, Y., & Wang, S. (2019). Curiosity-driven exploration for off-policy reinforcement learning methods. *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 1109–1114. https://doi.org/10.1109/ROBIO49542.2019.8961529

[73] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

[74] Lin, Y., Liu, Y., Lin, F., Zou, L., Wu, P., Zeng, W., Chen, H., & Miao, C. (2024). A survey on reinforcement learning for recommender systems. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(10), 13164–13184. https://doi.org/10.1109/TNNLS.2023.3280161

[75] Linden, G., Smith, B., & York, J. (2003). Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, *7*(1), 76–80.

[76] Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., Guo, H., & Zhang, Y. (2018). Deep reinforcement learning based recommendation with explicit user-item interactions modeling. *arXiv preprint arXiv:1810.12027*. https://arxiv.org/abs/1810.12027

[77] Liu, F., Tang, R., Li, X., Zhang, W., Ye, Y., Chen, H., Guo, H., Zhang, Y., & He, X. (2020). State representation modeling for deep reinforcement learning based

University of Ghana http://ugspace.ug.edu.gh

recommendation. *Knowledge-Based Systems*, *205*, 106170. https://doi.org/10.1016/j.knosys.2020.106170

[78] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Han, J. (2020). On the variance of the adaptive learning rate and beyond [arXiv:1908.03265]. *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

[79] Mandal, S., & Maiti, A. (2018). Explicit feedbacks meet with implicit feedbacks : A combined approach for recommendation system. *ArXiv*, *abs/1810.12770*. https://api.semanticscholar.org/CorpusID:53116738

[80] Matthes, J., Karsay, K., Schmuck, D., & Stevic, A. (2020). "too much to handle": Impact of mobile social networking sites on information overload, depressive symptoms, and well-being. *Computers in Human Behavior*, *105*, 106217.

[81] McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. *Proceedings of the 7th ACM Conference on Recommender Systems*, 165–172. https://doi.org/10.1145/2507157.2507163

[82] Morales, E. F., Murrieta-Cid, R., Becerra, I., & Esquivel-Basaldua, M. A. (2021). A survey on deep learning and deep reinforcement learning in robotics with a tutorial on deep reinforcement learning. *Intelligent Service Robotics*, *14*(5), 773–805.

[83] Nilashi, M., Abumalloh, R. A., Alrizq, M., Almulihi, A., Alghamdi, O., Farooque, M., Samad, S., Mohd, S., & Ahmadi, H. (2022). A hybrid method to solve data sparsity in travel recommendation agents using fuzzy logic approach. *Mathematical Problems in Engineering*, *2022*.

University of Ghana http://ugspace.ug.edu.gh

[84] Niu, H., Xu, W., Akbarzadeh, H., Parvin, H., Beheshti, A., & Alinejad-Rokny, H. (2020). Deep feature learnt by conventional deep neural network. *Computers & Electrical Engineering*, *84*, 106656.

[85] Nobakht, H., & Liu, Y. (2022). Action space noise optimization as exploration in deterministic policy gradient for locomotion tasks. *Applied Intelligence*, *52*(12), 14218–14232. https://doi.org/10.1007/s10489-021-02995-y

[86] Ojagh, S., Malek, M. R., & Saeedi, S. (2020). A social–aware recommender system based on user's personal smart devices. *ISPRS International Journal of Geo-Information*, *9*(9), 519.

[87] Ojagh, S., Malek, M. R., Saeedi, S., & Liang, S. (2020). A location-based orientation-aware recommender system using iot smart devices and social networks. *Future Generation Computer Systems*, *108*, 97–118.

[88] Osband, I., Blundell, C., Pritzel, A., & Van Roy, B. (2016). Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, *29*.

[89] Pang, B., Nijkamp, E., & Wu, Y. N. (2019). Deep learning with tensorflow: A review. *Journal of Educational and Behavioral Statistics*, *45*, 227–248. https://api.semanticscholar.org/CorpusID:203077438

[90] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 8024–8035.

University of Ghana http://ugspace.ug.edu.gh

[91] Pennock, D. M., Horvitz, E., & Giles, C. L. (2000). Social choice theory and recommender systems: Analysis of the axiomatic foundations of collaborative filtering. *AAAI/IAAI*. https://api.semanticscholar.org/CorpusID:1143688

[92] Pérez-Gil, Ó., Barea, R., López-Guillén, E., Bergasa, L. M., Gomez-Huelamo, C., Gutiérrez, R., & Diaz-Diaz, A. (2022). Deep reinforcement learning based control for autonomous vehicles in carla. *Multimedia Tools and Applications*, *81*(3), 3553–3576.

[93] Phan, L. P., Huynh, H. H., & Huynh, H. X. (2018). Implicative rating-based hybrid recommendation systems. https://doi.org/10.18178/ijmlc.2018.8.3.691

[94] Plappert, M., Houthooft, R., Dhariwal, P., Sidor, S., Chen, R. Y., Chen, X., Asfour, T., Abbeel, P., & Andrychowicz, M. (2017). Parameter space noise for exploration. *arXiv preprint arXiv:1706.01905*.

[95] Qiu, R., Huang, Z., Li, J., & Yin, H. (2020). Exploiting cross-session information for session-based recommendation with graph neural networks. *ACM Transactions on Information Systems (TOIS)*, *38*(3), 1–23.

[96] Quadrana, M., Cremonesi, P., & Jannach, D. (2018a). Sequence-aware recommender systems. *ACM computing surveys (CSUR)*, *51*(4), 1–36.

[97] Quadrana, M., Cremonesi, P., & Jannach, D. (2018b). Sequence-aware recommender systems. *ACM Comput. Surv.*, *51*(4). https://doi.org/10.1145/3190616

[98] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. *Proceedings of the 1994 ACM conference on Computer supported cooperative work*, 175–186.

[99] Richa & Bedi, P. (2021). Trust and distrust based cross-domain recommender system. *Applied Artificial Intelligence*, *35*(4), 326–351.

University of Ghana http://ugspace.ug.edu.gh

[100] Roy, D., & Dutta, M. (2022). A systematic review and research perspective on recommender systems. *Journal of Big Data*, *9*(1), 59.

[101] Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data mining and knowledge discovery*, *5*, 115–153.

[102] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized experience replay [arXiv:1511.05952]. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[103] Schröder, G., Thiele, M., & Lehner, W. (2011). Setting goals and choosing metrics for recommender system evaluations. https://api.semanticscholar.org/CorpusID:7266147

[104] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

[105] Shah, K. R., Salunke, A., Dongare, S., & Antala, K. (2017). Recommender systems: An overview of different approaches to recommendations. *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 1–4. https://api.semanticscholar.org/CorpusID:29921708

[106] Shakya, A. K., Pillai, G., & Chakrabarty, S. (2023). Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, *231*, 120495. https://doi.org/10.1016/j.eswa.2023.120495

[107] Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., & Basilico, J. (2021). Deep learning for recommender systems: A Netflix case study. *AI Magazine*, *42*(3), 7–18. https://doi.org/10.1609/aimag.v42i3.18140

[108] Sumiea, E. H., Abdulkadir, S. J., Alhussian, H. S., Al-Selwi, S. M., Alqushaibi, A., Ragab, M. G., & Fati, S. M. (2024). Deep deterministic policy gradient algorithm:

University of Ghana http://ugspace.ug.edu.gh

A systematic review. *Heliyon*, *10*(9), e30697. https://doi.org/10.1016/j.heliyon.2024.e30697

[109] Sun, F., Liu, J., Wu, J., Pei, C., Lin, X., Ou, W., & Jiang, P. (2019). Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer. *Proceedings of the 28th ACM international conference on information and knowledge management*, 1441–1450.

[110] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

[111] Tang, J., & Wang, K. (2018). Personalised top-n sequential recommendation via convolutional sequence embedding. *Proceedings of the eleventh ACM international conference on web search and data mining*, 565–573.

[112] Tao, H., Niu, X., Fu, L., Yuan, S., Wang, X., Zhang, J., & Hu, Y. (2022). Deeprs: A library of recommendation algorithms based on deep learning. https://doi.org/10.1007/s44196-022-00102-8

[113] Tirinzoni, A., Poiani, R., & Restelli, M. (2020). Sequential transfer in reinforcement learning with a generative model. *ArXiv*, *abs/2007.00722*. https://api.semanticscholar.org/CorpusID:220302641

[114] Tomasi, F., Cauteruccio, J., Kanoria, S., Ciosek, K., Rinaldi, M., & Dai, Z. (2023). Automatic music playlist generation via simulation-based reinforcement learning. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 4948–4957. https://doi.org/10.1145/3580305.3599777

[115] Valcarce, D. (2021). Information retrieval models for recommender systems. *SIGIR Forum*, *53*(1), 44–45. https://doi.org/10.1145/3458537.3458545

University of Ghana http://ugspace.ug.edu.gh

[116] Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double q-learning. *Proceedings of the AAAI conference on artificial intelligence*, *30*(1).

[117] Wang, S., Guo, S., Hao, J., Ren, Y., & Qi, F. (2024). Dpu-enhanced multi-agent actor-critic algorithm for cross-domain resource scheduling in computing power network. *IEEE Transactions on Network and Service Management*, *21*(6), 6008–6025. https://doi.org/10.1109/TNSM.2024.3434997

[118] Watkins, C., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*, 279–292. https://api.semanticscholar.org/CorpusID:208910339

[119] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80–83. https://doi.org/10.2307/3001968

[120] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X., & Tan, T. (2019). Session-based recommendation with graph neural networks. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 346–353.

[121] Yang, L., Cui, Y., Xuan, Y., Wang, C., Belongie, S. J., & Estrin, D. (2018). Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. *Proceedings of the 12th ACM Conference on Recommender Systems*. https://api.semanticscholar.org/CorpusID:51992468

[122] Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. *Proceedings of the web conference 2020*, 418–428.

[123] Zhang, M. R., Lucas, J., Hinton, G. E., & Ba, J. (2019). Lookahead optimizer: *k* steps forward, 1 step back. *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

University of Ghana http://ugspace.ug.edu.gh

[124] Zhang, T., & Mo, H. (2021). Reinforcement learning for robot research: A comprehensive review and open issues. *International Journal of Advanced Robotic Systems*, *18*. https://api.semanticscholar.org/CorpusID:235699851

[125] Zhao, X., Xia, L., Zhang, L., Ding, Z., Yin, D., & Tang, J. (2018). Deep reinforcement learning for page-wise recommendations. *Proceedings of the 12th ACM Conference on Recommender Systems*, 95–103. https://doi.org/10.1145/3240323.3240374

[126] Zheng, Y. (2017). Context suggestion. *Proceedings of the International Conference on Web Intelligence*. https://doi.org/10.1145/3106426.3106466

[127] Zhong, S.-T., Huang, L., Wang, C.-D., Lai, J.-H., & Philip, S. Y. (2020). An autoencoder framework with attention mechanism for cross-domain recommendation. *IEEE Transactions on Cybernetics*, *52*(6), 5229–5241.

[128] Zhou, Y., Wu, Z., & Li, X. (2023). Deep reinforcement learning in recommender systems: A survey. *Knowledge-Based Systems*, *110335*. https://doi.org/10.1016/j.knosys.2023.110335

[129] Zhu, L., Shah, H., Wang, H., Nagai, Y., & White, M. (2025). q-exponential family for policy optimization. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2408.07245

[130] Zhu, X., Zhang, F., & Li, H. (2022). Swarm deep reinforcement learning for robotic manipulation. *Procedia Computer Science*, *198*, 472–479.

[131] Ziarani, R. J., & Ravanmehr, R. (2021). Serendipity in recommender systems: A systematic literature review. *Journal of Computer Science and Technology*, *36*, 375–396.