Ph.D. DISSERTATION

A Distributional Perspective on Human-Aligned Decision Making under Uncertainty

불확실성 속 인간과 정렬된 의사결정에 관한 분포적 접근

February 2026

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

Taehyun Cho

A Distributional Perspective on Human-Aligned Decision Making under Uncertainty

불확실성 속 인간과 정렬된 의사결정에 관한 분포적 접근

지도교수 이 정 우

이 논문을 공학박사학위논문으로 제출함

2025 년 11 월

서울대학교 대학원

전기정보공학부

조 태 현

조 태 현의 공학박사 학위논문을 인준함

2025 년 12 월

위 원 장 _____ (인)
부위원장 _____ (인)
위 원 _____ (인)
위 원 _____ (인)
위 원 _____ (인)

Abstract

Sequential decision-making under uncertainty is a fundamental problem in artificial intelligence. Real-world environments rarely provide well-defined rewards or complete information, and feedback is often qualitative, subjective, or inconsistent. As AI systems are increasingly deployed in high-stakes domains such as finance, autonomous driving, and human–robot interaction, it becomes crucial to develop principled algorithms that can act reliably under uncertainty and align with human intentions. However, existing reinforcement learning (RL) paradigms, which lack explicit modeling of uncertainty, rely primarily on expectation-based objectives and handcrafted rewards, leaving a substantial gap between theoretical optimality and human-aligned behavior.

This dissertation addresses these challenges through two complementary perspectives—distributional reinforcement learning (DistRL) and reinforcement learning from human feedback (RLHF)—and unifies them under a common theoretical lens of regret minimization. The central goal is to establish a reliable foundation for learning human-aligned decision-making by interpreting the probabilistic nature inherent in human feedback.

The first part revisits the exploration problem in DistRL. Existing approaches based on "optimism under uncertainty" rely on estimates of return variance but conflate epistemic and aleatoric uncertainties, which induces persistent risk-seeking bias and distorted data collection. To address this, we propose the Perturbed Quantile Regression (`PQR`) algorithm, which introduces randomized perturbations of distorted risk measures to guide action selection. We theoretically establish that `PQR` avoids biased exploration and converges to the true

optimum, and empirically show that it outperforms variance-based exploration methods across diverse benchmarks, including 55 Atari games.

The second part tackles the fundamental challenge of infinite dimensionality in DistRL. Prior work introduced the notion of Bellman closedness, but this fails to guarantee unbiased updates from finite samples in online learning. We propose the concept of Bellman Unbiasedness, which characterizes functionals that are not only preserved under Bellman updates but also estimable without bias from finite samples. Our analysis shows that only moment functionals satisfy both conditions. Building on this result, we design the first provably efficient DistRL algorithm under general value function approximation—Statistical Functional Least-Squares Value Iteration (`SF-LSVI`)—which achieves a tight regret bound of $\tilde{O}(d_E H^{3/2}\sqrt{K})$, improving upon prior results.

The third part turns to RLHF, where agents learn from preference feedback instead of handcrafted rewards. Recent frameworks such as Direct Preference Optimization (DPO) optimize policies directly without an explicit reward model but implicitly assume that all preference data are generated by the optimal policy, leading to a likelihood mismatch. To overcome this, we reinterpret preferences through the lens of regret and propose Policy-labeled Preference Learning (`PPL`), which explicitly integrates policy labels into the learning process. Our method introduces contrastive KL regularization that aligns policies with preferred data while contrasting against less-preferred data. We theoretically show that `PPL` characterizes an equivalence class of reward models consistent with a given optimal policy and establishes statistical robustness via uniquely defined regret. Empirically, `PPL` substantially improves RLHF performance in offline robotic manipulation tasks and demonstrates robustness in online learning.

Collectively, these contributions establish regret minimization as a unifying theoretical principle that bridges distributional modeling and human feedback,

linking the mathematical efficiency of RL with the behavioral realism of human decision-making. This work contributes to the foundation of trustworthy and human-aligned artificial intelligence, providing theoretical and algorithmic insights for robust decision-making under uncertainty.

Keywords: Reinforcement Learning, Distributional Reinforcement Learning, Reinforcement Learning from Human Feedback, Regret Minimization

Student Number: 2020-24770

Contents

Abstract **i**

Chapter 1 Introduction **1**

1.1 The Shift from Perception to Action 1

1.2 The Cognitive Gap: Uncertainty and Regret in Human Decision-Making 3

1.3 Research Scope and Unified Hypothesis on Uncertainty 6

1.4 Core Research Areas and Contributions 8

1.4.1 Distributional Reinforcement Learning: Uncertainty, Efficiency, and Bias 8

1.4.2 Reinforcement Learning from Human Feedback: Robust Alignment via Regret 10

1.5 Organization of the Dissertation 11

1.6 Publications 12

Chapter 2 Background **13**

2.1 Reinforcement Learning and Markov Decision Processes 14

2.1.1 Markov Decision Processes 14

2.1.2 Limitations of Classical RL 15

2.2 Distributional Reinforcement Learning 17
2.2.1 Distributional Bellman Equation and Convergence Properties 17
2.2.2 Distributional Bellman Optimality and Instabilities 18
2.2.3 Approximation schemes in distributional RL 19
2.3 Reinforcement Learning from Human Feedback 23
2.3.1 Motivation and Origins 23
2.3.2 Canonical RLHF pipeline 23
2.3.3 Direct preference optimization (DPO) and extensions . . 25
2.3.4 Challenges and open directions 26
2.4 Regret Minimization Framework 28
2.4.1 From Expected Utility to Prospect Theory 28
2.4.2 Regret Theory: Anticipating Counterfactual Emotion . . 29
2.4.3 Algorithmic Regret in Reinforcement Learning 31
2.4.4 Bridging Behavioral and Algorithmic Perspectives 31
2.4.5 Toward a Unified View 32
2.5 Summary 33

Chapter 3 Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion 34

3.1 Backgrounds & Related works 37
3.1.1 Distributional RL 37
3.1.2 Exploration on Distributional RL 38
3.1.3 Risk in Distributional RL 39
3.2 Perturbation in Distributional RL 40
3.2.1 Perturbed Distributional Bellman Optimality Operator . 40

3.2.2 Convergence of the perturbed distributional Bellman optimality operator . . . 43
3.2.3 Practical Algorithm with Distributional Perturbation . . 44
3.3 Experiments on Stochastic Enviornments with High Intrinsic Uncertainty . . . 46
3.3.1 N-Chain Enviornment . . . 47
3.3.2 LunarLander-v2 . . . 52
3.3.3 55 Atari Games . . . 53
3.4 Related Works & Discussion . . . 59
3.4.1 Comparison with QUOTA . . . 59
3.4.2 Reproducibility issues on DLTV . . . 61
3.5 Summary . . . 62

Chapter 4 Bellman Unbiasedness: Toward Provably Efficient Distributional Reinforcement Learning with General Value Function Approximation 64

4.1 Related Work . . . 66
4.2 Preliminaries . . . 68
4.3 Statistical Functionals in Distributional RL . . . 70
4.3.1 Bellman Closedness . . . 71
4.3.2 Bellman Unbiasedness . . . 73
4.3.3 Statistical Functional Bellman Completeness . . . 76
4.4 `SF-LSVI`: Statistical Functional Least Squares Value Iteration . . 78
4.5 Theoretical Analysis . . . 79
4.6 Summary . . . 82

Chapter 5 Policy-labeled Preference Learning: Is Preference Enough for RLHF? 84

5.1 Preliminaries . . . 85
5.1.1 Preference-based Reinforcement Learning . . . 86
5.2 Policy-labeled Preference Learning . . . 89
5.2.1 Is Preference Enough for RLHF? . . . 90
5.2.2 Theoretical Analysis . . . 93
5.2.3 Practical Algorithm and Implementation Details . . . 99
5.3 Experiments . . . 102
5.3.1 Experimental Setup . . . 102
5.3.2 Can PPL be effectively trained on both homogeneous and heterogeneous offline dataset? . . . 103
5.3.3 Does incorporating policy labels improve learning performance? . . . 105
5.3.4 Can PPL be effectively applied to an online RLHF algorithm? 106
5.4 Summary . . . 107

Chapter 6 Conclusion **108**
6.1 Future Work . . . 109

Appendix A Appendix of Chapter 3 **111**
A.1 Main Proof . . . 111
A.1.1 Technical Lemma . . . 111
A.1.2 Proof of Theorem A.1.3 . . . 112
A.1.3 Proof of Theorem 3.2.3 . . . 113
A.1.4 Proof of Theorem 3.2.4 . . . 115
A.2 Implementation details . . . 117
A.2.1 Hyperparameter Setting . . . 117
A.3 Raw scores across 55 Atari games . . . 118

Appendix B Appendix of Chapter 4 **121**

B.1 Notation 121

B.2 Pseudocode of `SF-LSVI` and Technical Remarks 124

B.3 Related Work and Discussion 125

B.3.1 Technical Clarifications on Linearity Assumption in Existing Results 125

B.3.2 Existence of Nonlinear Bellman Closed Sketch. 126

B.3.3 Non-existence of sketch Bellman operator for quantile functional 127

B.4 Proof 131

Appendix C Appendix of Chapter 5 **144**

C.1 Main Proof 144

C.2 Further Theoretical Analysis & Discussion 147

C.2.1 Mathematical derivation of PPL framework 147

C.3 Variants of `PPL` and Baselines 149

C.4 Implementation Details 151

C.4.1 Hyperparameter Setting 151

C.4.2 MetaWorld Benchmark 152

C.4.3 Reproducibility Check 153

C.4.4 Offline dataset generation and its distribution 156

C.4.5 Online Implementation 158

C.5 Experimental Results on Homogeneous/ Heterogeneous Datasets (Section 5.3.2) 160

C.5.1 `Homogeneous Dense` Offline Dataset 160

C.5.2 `Homogeneous Sparse` Offline Dataset 161

C.5.3 `Heterogeneous Dense` Offline Dataset 162

C.5.4 `Heterogeneous Sparse` Offline Dataset 163
C.6 Comparison with Deterministic Pseudo-labels (Section 5.3.3) . . 164
C.6.1 `Homogeneous Dense` Offline Dataset 164
C.6.2 `Heterogeneous Dense` Offline Dataset 165
C.7 Experimental Results on Online Implementation (Section 5.3.4) . 166
C.7.1 Online Learning Curves 166
C.7.2 Ablation on Preference Query Count 167
C.7.3 Ablation on Rollout Length 168

초록 **185**

감사의 글 **187**

List of Figures

Figure 1.1 Four stages of human decision making 3

Figure 1.2 Conceptual structure of this dissertation. 11

Figure 2.1 This modified N-Chain MDP illustrates how outcomes are governed by full probability distributions rather than single scalar values. The agent at $S2$ faces a choice between a low-variance, safe path (Left) and a high-variance, risky path (Right). This structure highlights the representational limitations of expectation-based RL 15

Figure 2.2 Visualization of the Distributional Bellman Operator $\mathcal{T}$. The figure illustrates the three key transformation steps required to construct the target distribution in distRL. **(Left) Components of the Next-State Distribution:** Shows the individual probability components that collectively define the value distribution of the subsequent state. **(Center) Scaling by the Discount Factor:** The full distribution is compounded and scaled by the discount factor γ. This results in the discounted distribution (solid orange fill), relative to the original distribution (dashed line). **(Right) Shifting by the Immediate Reward:** The final step shifts the discounted distribution (dashed line) horizontally by the immediate reward, yielding the final target distribution (solid orange fill). 18

Figure 3.1 Illustrative example of why a biased risk criterion (naïve optimism) can degrade performance. Suppose two actions have similar expected returns, but different variances (intrinsic uncertainty). **(Left)** If an agent does not specify the risk criterion at the moment, the probability of selecting each action should be similar. **(Right)** As OFU principle encourages to decide uncertain behaviors, the empirical variance from quantiles was used as an estimate of uncertainty [54, 64, 66]. However, optimistic decision based on empirical variance inevitably leads a risk-seeking behavior, which causes biased action selection. 35

Figure 3.2 An illustrative example of proposed algorithm (PQR). Each distribution represents the empirical PDF of return. PQR benefits from excluding inferior actions and promoting unbiased selection with regards to high intrinsic uncertainty through randomized risk criterion. 36

Figure 3.3 Pipeline of PDBOO. 46

Figure 3.4 Illustration of the N-Chain environment [73] with high uncertainty starting from state s_2. To emphasize the intrinsic uncertainty, the reward of state s_4 was set as a mixture model composed of two Gaussian distributions. Blue arrows indicate the risk-neutral optimal policy in this MDPs. 47

Figure 3.5 Empirical return distribution plot in N-Chain environment. The ground truth of each distribution is $\gamma^2\mathcal{N}(10, 0.1^2)$ and $\gamma^2[\frac{1}{2}\mathcal{N}(5, 0.1^2) + \frac{1}{2}\mathcal{N}(13, 0.1^2)]$. Each dot represents an indicator for choosing action. Since QR-DQN does not depend on other criterion, the dots are omitted. 49

Figure 3.6 Total count of performing true optimal action. The oracle (dashed line) is to perform the optimal action from start to end. 50

Figure 3.7 2-Wasserstein distance between the empirical return distribution and the ground truth $\mathcal{N}(8.1, 0.081^2)$. We use QR-DQN with a fixed setting of ϵ-greedy as a reference baseline, because the hyperparameter of ϵ-greedy is not related to the scale of Q-values. 51

Figure 3.8 (**Left**) Three main environmental factors causing high intrinsic uncertainty on LunarLander-v2. (**Right**) Performance on LunarLander-v2 52

Figure 3.9 Evaluation curves on 8 Atari games with 3 random seeds for 50 million frames following *sticky actions* protocol [62]. Reference values are from Castro et al. [19]. 55

Figure 3.10 Evaluation curves on Atari games. All curves are smoothed over 10 consecutive steps with three random seeds. In case of Pong-v4, we resize the x-axis, since it can easily obtain the optimal policy with few interactions due to its environmental simplicity. 56

Figure 3.11 Evaluation curves on Pong-v4 environments. 58

Figure 4.1 **Venn-Diagram of Statistical Functional Classes.** The diagram illustrates categories of statistical functional. **(Yellow ∩ Blue)** Within the linear statistical functional class, Rowland et al. [81] showed that the only functionals satisfying Bellman closedness are moment functionals. **(Red ∩ Blue)** We extend this concept by introducing the notion of *Bellman unbiasedness*, which not only covers moment functionals but also includes central moment functionals from the broader class including nonlinear statistical functionals. **(Yellow ∩ Blue^c)** According to Lemmas 3.2 and 4.4 of Rowland et al. [81], categorical functionals are linear but not Bellman closed. **(A)** Maximum and minimum functionals are Bellman closed, while they are not unbiasedly estimatable. **(B)** Median and quantile functionals are neither Bellman closed nor unbiased, highlighting that they are not proper to encode the distribution in terms of exactness. The proofs corresponding to each region are provided in Appendix B.3. 65

Figure 4.2 Illustrative representation of sketch-based Bellman updates for a mixture distribution. Instead of updating the distributions directly, each sampled distribution is embedded through a sketch ψ (e.g., mean μ, quantile q_i). The transformation ϕ_ψ aims to compress the mixture distribution into the same number of parameters, ensuring unbiasedness to prevent information loss. 71

Figure 4.3 Bellman Closedness 72

Figure 4.4 Bellman Unbiasedness 72

Figure 4.5 Illustration of Bellman Closedness and Bellman Unbiasedness. The above path represents an ideal distributional Bellman update. Due to the infinite-dimensionality, the update process should be represented by using a finite-dimensional embedding (sketch) ψ. Since the transition kernel $\mathbb{P}$ is unknown, the below path describes that the implementation should sample the next state and update by using $\hat{\mathcal{T}}_\psi$ with the empirical transition kernel $\hat{\mathbb{P}}$. A sketch ψ is Bellman unbiased if $\hat{\mathcal{T}}_\psi \circ \psi$ can unbiasedly estimate $\psi \circ \mathcal{T}$ through some transformation ϕ_ψ, *i.e.*, $\psi(\mathcal{T}\eta) = \mathbb{E}_{\mathbb{P}}[\phi_\psi \circ \hat{\mathcal{T}}\psi(\eta)]$. 72

Figure 5.1 Visualization of 5000 samples in `Bin-Picking-v2` environment. While the ground-truth reward **(left)** is sparse and mainly provided upon task completion, regret **(right)** is more evenly distributed across all timesteps, making it a more informative score function for partial trajectory evaluation. 88

Figure 5.2 Unlike existing DPO algorithms, `PPL` aligns segment likelihoods by incorporating behavior policies. It reweights gradients based on closeness to the optimal policy, forming a contrastive learning framework. 90

Figure 5.3 Illustration of the likelihood mismatch problem. Although the behavior policy π differs from the optimal policy π^*, the learning process incorrectly assumes all data is generated by π^*. As a result, while π^* prefers s_1, this misinterpretation leads to the incorrect conclusion that s_2 is preferred, causing suboptimal learning outcomes. . 91

Figure 5.4 Distribution of returns in homogeneous vs heterogeneous offline dataset in `Button-Press-v2`. 104

Figure 5.5 Ablation on deterministic pseudo-labeling. We compare the average performance of `PPL` and `PPL`-deterministic across six environments in MetaWorld. The dashed line indicates the point where BC pretraining stops. 105

Figure 5.6 Online learning curves across five MetaWorld tasks, comparing `PPL` and `PEBBLE`. 106

Figure C.1 Visualization of the MetaWorld Benchmark Tasks. . . . 153

Figure C.2 Reproducibility check on `State Dense` dataset 154

Figure C.3 Reproducibility check on `State Sparse` dataset 155

Figure C.4 Comparison of return distributions across environments for different dataset configurations. The histograms illustrate the distribution of the partial returns for segments with 20% and 50% success rates generated using our method **(red and blue)** and the 50% success rate dataset from Hejna et al. [44] **(gray)**. 157

Figure C.5 Performance comparison of different methods on the `Homogeneous Dense` dataset across six MetaWorld tasks. The top row shows the success rate over training iterations, while the bottom row presents the corresponding return values. 160
Figure C.6 Performance comparison of different methods on the `Homogeneous Sparse` dataset across six MetaWorld tasks. 161
Figure C.7 Performance comparison of different methods on the `Heterogeneous Dense` dataset across six MetaWorld tasks. 162
Figure C.8 Performance comparison of different methods on the `Heterogeneous Sparse` dataset across six MetaWorld tasks. 163
Figure C.9 Comparison of `PPL` and `PPL`-deterministic on the `Homogeneous Dense` Offline Dataset. 164
Figure C.10 Comparison of `PPL` and `PPL`-deterministic on the `Heterogeneous Dense` Offline Dataset. 165
Figure C.11 `PPL` and `PEBBLE` learning curves in online learning. . . . 166
Figure C.12 Effect of preference query count in online learning. . . . 167
Figure C.13 Effect of rollout length in online learning. 168

List of Tables

Table 2.1 Structural Correspondence between Behavioral and Algorithmic Regret 32

Table 3.1 Total counts of performing true optimal action with 4 different seeds. 52

Table 3.2 Mean and median of best scores across 55 Atari games, measured as percentages of human baseline. Reference values are from Quan and Ostrovski [76] and Castro et al. [19]. 53

Table 3.3 Performance comparison among `QUOTA`, `DLTV`, and `PQR` on 55 Atari games. Values in the first block indicate the number of games (out of 55) where the row method outperforms the column method. 62

Table 4.1 Comparison for different methods under distributional RL framework. $\mathcal{H}$ represents a subspace of infinite-dimensional space $\mathcal{F}^{\infty}$. To bound the eluder dimesion d_E, Wang et al. [100] and Chen et al. [21] assumed the discretized reward MDP. 67

Table 5.1 Comparison for different preference models under PbRL framework. 87

Table 5.2 Success rates of all methods across six tasks on the Meta-World benchmark on different datasets. Each score is reported with the maximum average performance across four seeds over 200 episode evaluation window. 101

Table A.1 Table of hyperparameter setting 117

Table A.2 Raw scores across all 55 games, starting with 30 no-op actions. We report the best scores for `DQN`, `QR-DQN`, `IQN` and `Rainbow` on 50M frames, averaged by 5 seeds. Reference values were provided by DQN_Zoo framework [76]. **Bold** are wins against `DQN`, `QR-DQN` and `IQN`, and *asterisk are wins over `Rainbow`. 118

Table A.3 Raw scores across 55 games. We report the best scores for `DQN`, `QR-DQN`, `IQN*`, and `Rainbow` on 50M frames, averaged by 5 seeds. Reference values were provided by Dopamine framework [19]. **Bolds** are wins against `DQN`, `QR-DQN`, and *asterisk are wins over `IQN*` and `Rainbow`. Note that `IQN*` and `Rainbow` implemented in Dopamine framework applied n-step updates with $n = 3$ which improves performance. 119

Table A.4 Raw scores across all 49 games, starting with 30 no-op actions. We report the best scores for QR-DQN_zoo [76], QR-DQN_Zhang [111](implemented by QUOTA to evaluate the relative improvement) for a fair comparison and QUOTA [111], DLTV [64] on 40M frames, averaged by 3 seeds. **Bold** are wins against QUOTA and DLTV. 120

Table B.1 Table of notation (Part 1: core and statistical notation) . 122

Table B.2 Table of notation (Part 2: functionals and dataset-related quantities) 123

Table C.1 Hyperparameter settings for offline implementation. . . . 151

Table C.2 Hyperparameters for online implementation 151

Table C.3 Hyperparameters for PPL, CPL, SFT, and P-IQL 152

Table C.4 Success rates of all methods on six tasks from the Meta-World across different datasets from Hejna et al. [44]. Each score is reported as the highest average performance across four seeds over a 200-episode evaluation window. 154

Chapter 1

Introduction

1.1 The Shift from Perception to Action

Modern artificial intelligence has made extraordinary progress in *perception*—vision, language, and speech—domains concerned with how machines see, interpret, and generate information. Yet perception alone does not constitute intelligence. An autonomous system must not only understand its environment but also decide how to *act* within it, often under uncertainty and in alignment with human goals and values. In this sense, reinforcement learning (RL) represents the natural next step for AI: a shift from recognizing the world to interacting with it.

A familiar philosophical remark suggests that *"Life is Choice between Birth and Death"*, underscoring that existence—biological or artificial—is shaped by a continual sequence of decisions. Every agent must navigate uncertainty, evaluate consequences, and refine its behavior over time to survive and succeed. This perspective captures the essence of RL, which formalizes sequential decision-making under uncertainty.

The breakthroughs of the Alpha family of agents illustrate the remarkable potential of RL, deomonstrating how RL methodologies have scaled as game complexity progressively increased. Starting with AlphaGo's mastery of the sequential, high-search-depth game of Go [90], the methodology evolved into AlphaZero, which generalized zero-knowledge learning across perfect-information games like Chess and Shogi [91]. The culmination of this progress came with AlphaStar's success in the real-time, imperfect-information, and decentralized strategy game, StarCraft II [98]. This accomplishment not only demonstrated RL's capability to discover strategies that exceed human intuition in increasingly challenging environments but also provided crucial insights into the complexities of real-time, sequential decision-making under uncertainty. Notably, these successes were achieved in settings where objectives were clearly defined and explicit reward signals were directly available from the environment. Such conditions stand in sharp contrast to many real-world human-facing tasks, where objectives are ambiguous and rewards are not directly observable.

Under such reward-based formulations, RL can be interpreted as a computational analogue of human decision-making. This abstraction, however, relies not only on the availability of well-defined scalar rewards, but also on expectation-based value representations that collapse uncertainty into a single summary statistic. As a consequence, classical RL frameworks struggle to represent the uncertainty, variability, and asymmetry that are intrinsic to human evaluation and decision-making. In many real-world human-facing settings, where objectives are ambiguous and feedback is qualitative, both the specification of rewards and the modeling of uncertainty become fundamentally challenging. Bridging this gap requires principled frameworks that go beyond reward maximization, incorporating both uncertainty-aware evaluation and alternative feedback signals aligned with human judgment.

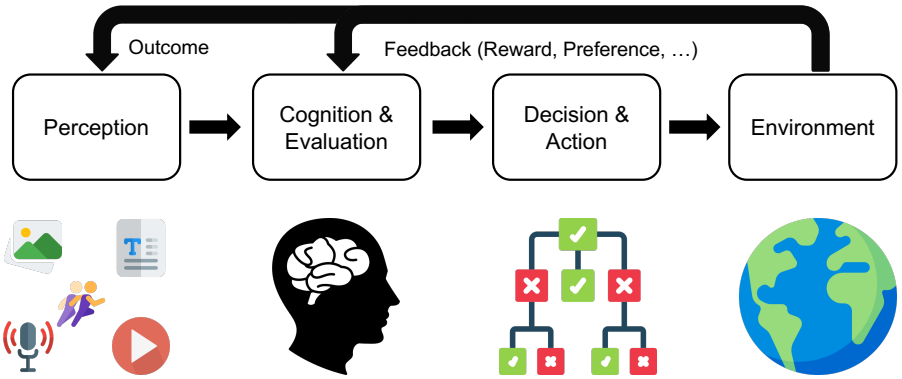


Figure 1.1: Four stages of human decision making

1.2 The Cognitive Gap: Uncertainty and Regret in Human Decision-Making

To understand the limitations of classical reinforcement learning in human-facing settings, it is essential to examine how humans actually make decisions. Human decision-making can be decomposed into four essential stages, forming a continuous cognitive loop:

1. **Information/State Perception:** This initial stage involves the agent (human or AI) collecting and interpreting external stimuli to form an internal representation of the current environment state. In RL terms, this is the observation phase, essential for subsequent prediction. The rapid advancement of generative AI (e.g., using models trained on vast datasets of images, audio, and text) has dramatically improved the quality of this stage. This foundational capability—accurate and rich internal modeling of the environment—is the prerequisite upon which all subsequent decision-making is built.

2. **Outcome Prediction and Value/Preference Evaluation:** Based on

the perceived state, the agent predicts the potential outcomes of available actions. Crucially, this stage includes evaluation, where subjective values, risk sensitivities, and personal preferences are applied to assign worth to those predicted outcomes. A central characteristic of this stage is the *inherent uncertainty* in prediction, which requires reasoning over the distribution of possible outcomes rather than simple averages. Evidence from neuroscience and cognitive psychology suggests that biological agents employ distributional representations of reward uncertainty, while behavioral economics has shown that risk-sensitive and non-expected utility behaviors cannot be captured by mean-based evaluation alone.

3. **Policy Formulation and Action Selection:** This stage concerns real-time execution, where evaluated predictions guide action selection. In classical reinforcement learning, this process is formalized through policies that maximize expected return given the current state. As a result, decision-making is largely driven by expectation-based optimization, abstracting away uncertainty beyond the mean.

4. **Interaction and Feedback Circulation:** The selected action is executed in the environment, leading to a new state and yielding an outcome (reward/cost) and feedback. This feedback then circulates back, refining the initial perception and future evaluation models. This stage generates two types of crucial feedback: explicit outcomes (next state, reward signal) from the environment and internal feedback (prediction error, human preference). In real-world human-facing systems, the lack of a clear, scalar reward signal necessitates moving beyond traditional extrinsic rewards to learning directly from qualitative human feedback (preferences, rankings). Crucially, unlike real-time action selection, the interpretation of preference

feedback is inherently *retrospective*, involving comparisons between realized outcomes and counterfactual alternatives. This retrospective comparison naturally aligns with the notion of *regret*.

Taken together, this cognitive loop reveals a fundamental mismatch with classical reinforcement learning. While RL has been highly successful in modeling prediction and action selection, it largely abstracts away two core aspects of human cognition: (i) uncertainty in evaluation, where values are subjective and distributional rather than scalar, and (ii) the retrospective, regret-driven nature of human feedback processing.

The philosophical drive of this dissertation is rooted in this gap between cognitive reality and mathematical idealization. Specifically, while human decision-making is shaped by uncertainty in evaluation and retrospective, regret-driven feedback, classical reinforcement learning—grounded in the Markov Decision Process (MDP) framework— abstracts decision-making through two strong assumptions: (i) the availability of a perfectly defined scalar reward, and (ii) the sufficiency of expectation-based dynamic programming for optimization. In many real-world human-facing settings, however, neither assumption reliably holds. Environments are inherently stochastic, and critical feedback is often qualitative, subjective, or inconsistent, rather than immediate and scalar. As a result, agents trained via standard expected RL may exhibit behaviors that are brittle, risk-indifferent, or misaligned with human values, which are deeply sensitive to variability and worst-case scenarios. Addressing this gap requires rethinking both how uncertainty is represented in value estimation and how feedback signals are modeled beyond scalar rewards. In particular, when feedback originates from human judgments rather than the environment, learning must account for uncertainty in evaluation and the regret-driven nature of feedback.

1.3 Research Scope and Unified Hypothesis on Uncertainty

This dissertation is motivated by precisely this challenge, which lies at the intersection of statistical modeling and behavioral alignment:

How can we design reinforcement learning systems that make decisions under uncertainty in ways that are theoretically sound, practically robust, and cognitively aligned with human judgment?

The complexity of this challenge arises from two fundamental limitations inherent in the classical expected utility paradigm of RL: its inability to rigorously model objective environmental stochasticity and its failure to capture nuanced subjective human values. These shortcomings mean that conventional mean-reward maximization is inadequate for real-world, high-stakes applications. Consequently, these dual limitations—the reliance on expected utility and the failure to capture human-like feedback—directly motivate the two distinct, yet interconnected, research trajectories explored in this dissertation:

- **Objective Uncertainty (Environmental Stochasticity):** The failure of expectation-based optimization to capture risk necessitates modeling the full distribution of returns (Distributional RL; DistRL) to rigorously account for environmental uncertainty and risk.
- **Subjective Uncertainty (Value Ambiguity and Heterogeneity):** This trajectory focuses on creating algorithms that interpret and align with subjective human feedback (Reinforcement Learning from Human Feedback; RLHF), specifically accounting for the uncertainty and heterogeneity inherent in human preferences.

While these two trajectories are conceptually distinct, they are not indepen-

dent lines of inquiry in this dissertation, but rather reflect a coherent progression of a single research agenda centered on uncertainty-aware decision making. My earlier work in DistRL focused on learning and exploiting environmental uncertainty to enable efficient exploration and risk-aware control in stochastic environments. By moving beyond expectation-based objectives and modeling full return distributions, this line of research addresses a foundational question: how should an agent act when outcomes are inherently stochastic, rare events are consequential, and uncertainty itself must inform decision making rather than be averaged away.

This perspective emphasizes that uncertainty is not merely a nuisance to be mitigated, but a structural property of the environment that must be explicitly represented and reasoned about. DistRL thus provides a principled framework for capturing variability, risk, and tail behavior of returns, enabling more robust policies in safety-critical or high-stakes domains.

At the same time, focusing on outcome distributions naturally raises a deeper question:

How are decisions evaluated when uncertainty is not only a property of the environment, but also a defining feature of human judgment?

In many settings, human evaluations are inherently sensitive to variability, risk, and counterfactual comparisons rather than point estimates of performance. That is, judgments depend not only on what happened, but on how a realized trajectory compares to plausible alternatives under uncertainty. This observation motivates a transition from modeling uncertainty over outcomes to modeling uncertainty over evaluation. While the former concerns the stochasticity of the environment, the latter concerns ambiguity and heterogeneity in human values. Importantly, these two forms of uncertainty are not orthogonal. Human judgments are often shaped by sensitivity to risk, missed opportunities, and

unfavorable comparisons, all of which depend on the underlying uncertainty of outcomes.

Here, regret provides a principled bridge between these two perspectives. In DistRL, distributional uncertainty quantifies the range and variability of possible outcomes, guiding exploration and risk-sensitive behavior. In RLHF, regret formalizes how humans implicitly assess actions by comparing realized behavior against unchosen alternatives under uncertainty. Crucially, the formulation of regret inherently incorporates both objective and subjective uncertainty: it integrates *distributional uncertainty*—the range of plausible counterfactual outcomes— with *subjective evaluation*—the human assessment of realized outcomes against the best unchosen alternative.

The central hypothesis of this dissertation is that uncertainty-aware decision making, grounded in *distributional modeling* and *regret-based evaluation*, provides a principled foundation for both statistically robust learning and cognitively aligned behavior.

1.4 Core Research Areas and Contributions

1.4.1 Distributional Reinforcement Learning: Uncertainty, Efficiency, and Bias

DistRL is the mathematical extension of classical RL, recognizing that the full distribution of returns contains information about risk and variability that is essential in high-stakes domains (finance, healthcare, robotics). Capturing this distributional asymmetry—a key component of human judgment formalized by prospect theory and regret theory—is paramount. While empirically successful, DistRL faces two central challenges addressed in this thesis:

Mitigating the Pitfall of Optimism in Exploration. DistRL exploration often employs optimism in the face of uncertainty (OFU), guiding action based on high variance estimates. However, this strategy suffers from a pitfall of optimism: it fundamentally conflates *epistemic uncertainty* (which should guide exploration) with *aleatoric uncertainty* (intrinsic environmental randomness). This leads to a persistent, systematic risk-seeking bias and sub-optimal data collection. We solve this bias by introducing Perturbed Quantile Regression (`PQR`), which replaces variance-based optimism with randomized perturbations applied to distortion risk measures. `PQR` ensures unbiased exploration while maintaining risk-neutral optimality, supported by theoretical convergence guarantees and state-of-the-art empirical performance across complex benchmarks like Atari.

Achieving Provable Statistical Efficiency with General Approximation DistRL algorithms must approximate the infinite-dimensional return distribution using finite statistical functionals (e.g., quantiles, moments). This introduces two fundamental issues: first, the functional must satisfy Bellman Closedness (it must be preserved under the Bellman update); second, it must ensure *unbiased estimability* from the finite samples collected online. Previous work focused only on the former, leaving algorithms vulnerable to accumulated approximation errors and failing to guarantee efficiency. We formally introduce Bellman Unbiasedness and prove that only moment functionals satisfy both this new property and Bellman Closedness. Based on this robust foundation, we propose the Statistical Functional Least-Squares Value Iteration (`SF-LSVI`) algorithm. SF-LSVI is the first distributional RL algorithm with provable efficiency under general value function approximation, achieving a tight regret bound of $\tilde{O}(d_E H^{3/2}\sqrt{K})$.

1.4.2 Reinforcement Learning from Human Feedback: Robust Alignment via Regret

RLHF addresses the inability of classical RL to handle subjective and qualitative human values, relying on pairwise preference comparisons to align agent behavior. The emergence of Direct Preference Optimization (DPO) has been transformative, simplifying the alignment process by directly updating policies from preferences without an explicit reward model. However, DPO's success, primarily demonstrated in LLM fine-tuning, rests on an assumption challenged by general RL environments: that preference data originates from policies near-optimal for the task.

Addressing Likelihood Mismatch in Stochastic Environments In standard and offline RL settings, preference data is generated by diverse, suboptimal policies under environmental stochasticity. Applying DPO's assumptions in this context creates a severe likelihood mismatch: the suboptimality of the behavior policy is incorrectly modeled as noise or inherent difficulty. This undermines stability and generalization, especially when data is heterogeneous. We propose Policy-labeled Preference Learning (`PPL`) to fundamentally resolve this mismatch. `PPL` reformulates human preference not through reward functions, but through the lens of regret, incorporating the behavior policy label directly into the learning objective. We show that regret, unlike reward, defines a unique, policy-aware equivalence class that is inherently robust to heterogeneity. This novel approach is further stabilized by a contrastive KL regularization. `PPL` provides a principled framework for robust RLHF, significantly improving offline alignment and extending its applicability beyond deterministic LLM settings to general sequential decision-making.

1.5 Organization of the Dissertation

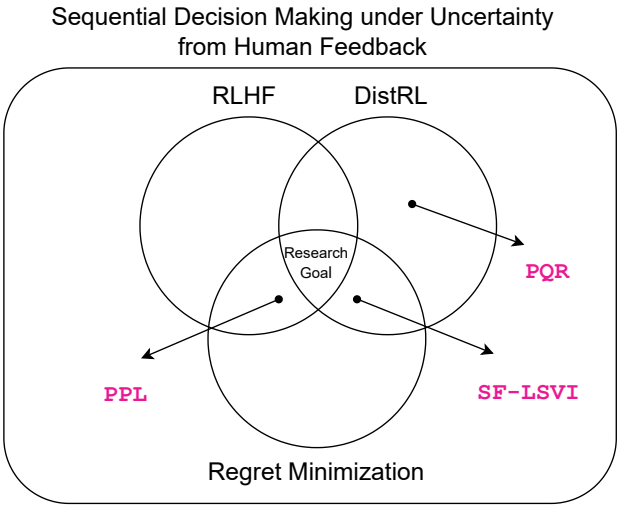


Figure 1.2: Conceptual structure of this dissertation.

This dissertation is organized as follows: Chapter 2 provides the necessary background on Markov Decision Processes, classical reinforcement learning, and the two central research areas of this dissertation: distributional reinforcement learning and reinforcement learning from human feedback. Chapter 3 revisits the exploration problem in DistRL and introduces the Perturbed Quantile Regression (`PQR`) algorithm, which addresses biased exploration by disentangling epistemic and aleatoric uncertainties. Chapter 4 develops the concept of Bellman Unbiasedness and presents Statistical Functional Least-Squares Value Iteration (`SF-LSVI`), the first provably efficient distributional algorithm under general value function approximation. Chapter 5 focuses on RLHF and introduces Policy-labeled Preference Learning (`PPL`), a regret-based framework that resolves the likelihood mismatch in preference optimization and achieves robust alignment

with human feedback. Finally, Chapter 6 concludes the dissertation, summarizing the key findings and outlining directions for future research. Supplementary results and additional implementation details are provided in the appendices.

1.6 Publications

The following publications have been selected as they closely align with the central themes of this dissertation. * indicates equal contribution.

- **Taehyun Cho**, Seungyub Han, Heesoo Lee, Kyungjae Lee, Jungwoo Lee. "Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion." *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

- **Taehyun Cho**, Seungyub Han, Seokhun Ju, Dohyeong Kim, Kyungjae Lee, Jungwoo Lee. "Bellman Unbiasedness: Toward Provably Efficient Distributional Reinforcement Learning with General Value Function Approximation." *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

- **Taehyun Cho***, Seokhun Ju*, Seungyub Han, Dohyeong Kim, Kyungjae Lee, Jungwoo Lee. "Policy-labeled Preference Learning: Is Preference Enough for RLHF?" *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025.

Chapter 2

Background

This chapter presents the theoretical and algorithmic foundations that underpin the main contributions of this dissertation. We begin by formulating the reinforcement learning (RL) problem through Markov decision processes (MDPs), outlining the standard framework for sequential decision-making. While classical RL provides a principled foundation for optimizing expected returns, it assumes that rewards are well-specified and that uncertainty can be adequately represented by expectations—assumptions that often fail in real-world settings. To address these limitations, we review two major extensions: *distributional reinforcement learning* (DistRL), which models the full distribution of returns to capture risk and uncertainty, and *reinforcement learning from human feedback* (RLHF), which replaces explicit reward signals with qualitative human judgments. Finally, we introduce the concept of *regret minimization*, a theoretical framework that unifies these perspectives by connecting statistical efficiency with behavioral alignment. Together, these topics provide the conceptual and mathematical background for the algorithms developed in Chapters 3–5.

2.1 Reinforcement Learning and Markov Decision Processes

Sequential decision-making is one of the central problems in artificial intelligence. The mathematical framework most widely used to formalize this problem is the notion of *Markov decision processes (MDPs)*. In this section, we introduce the preliminaries of reinforcement learning, including the definition of MDPs, value functions, and the Bellman equations. We also highlight the limitations of expectation-based RL, which motivate the development of distributional and preference-based frameworks.

2.1.1 Markov Decision Processes

An *episodic* Markov decision process is defined by the tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r),$$

where $\mathcal{S}$ is a (possibly infinite) state space, $\mathcal{A}$ is the action space, H is the horizon, $\mathbb{P} = \{\mathbb{P}_h\}_{h=1}^H$ are transition kernels, and $r = \{r_h\}_{h=1}^H$ are reward functions. At each step $h \in [H]$, the agent observes a state $s_h \in \mathcal{S}$, chooses an action $a_h \in \mathcal{A}$, receives reward $r_h(s_h, a_h)$, and transitions to a new state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h)$. A (stochastic) policy $\pi = \{\pi_h\}_{h=1}^H$ defines a distribution over actions given states: $\pi_h(a|s) = \mathbb{P}[a_h = a|s_h = s]$. Unless otherwise noted, we consider discounted returns with a discount factor $\gamma \in [0, 1)$.

The quality of a policy is measured by its value functions. The *state value function* is

$$V_h^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=h}^{H} r_t(s_t, a_t) \,\middle|\, s_h = s \right],$$

and the *state–action value function* is

$$Q_h^\pi(s, a) = r_h(s, a) + \mathbb{E}_{s' \sim \mathbb{P}_h(\cdot|s,a)}[V_{h+1}^\pi(s')].$$

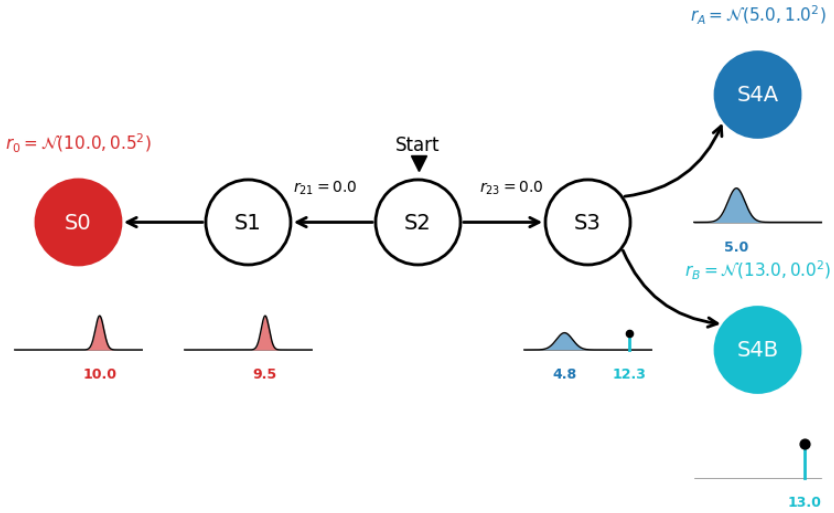


Figure 2.1: This modified N-Chain MDP illustrates how outcomes are governed by full probability distributions rather than single scalar values. The agent at $S2$ faces a choice between a low-variance, safe path (Left) and a high-variance, risky path (Right). This structure highlights the representational limitations of expectation-based RL

The Bellman operator $\mathcal{T}^\pi$ defines a recursive mapping of value functions, and it is a γ-contraction under the sup-norm. This property guarantees the existence and uniqueness of Q^π and underpins the convergence of classical RL algorithms such as value iteration and Q-learning.

2.1.2 Limitations of Classical RL

Despite its elegance, classical RL collapses all uncertainty into expectations. Two trajectories with identical expected returns but very different variances are treated as equivalent. This is problematic in domains where risk sensitivity matters, such as finance or healthcare, where variance and tail risks are critical.

To illustrate this representational deficiency, consider the N-Chain environment shown in Figure 2.1. A classical RL agent determines its policy solely by maximizing the expected return. In this example, the optimal decision based on the mean value may lead the agent to choose the Left Path, as its expected return is slightly higher than the average expectation of the Right Path. However, this expectation-based view cannot differentiate between the tightly concentrated, low-risk distribution of the Left Path and the bimodal distribution of the Right Path, which contains both high risk (low outcome potential) and high reward (certain high outcome potential). By summarizing the return only by its mean, classical RL is mathematically blind to these distinct risk profiles.

Moreover, this expectation-based view also conflicts with emerging findings in neuroscience. The canonical understanding posits that the firing of midbrain dopamine neurons encodes a scalar *Reward Prediction Error (RPE)*—the difference between the received reward and the expected mean return. However, this simple scalar model struggles to explain recent experimental evidence showing that individual dopamine neurons exhibit heterogeneous responses that reflect more than just the mean, effectively encoding a spectrum of optimism and pessimism related to the distribution of potential returns [32]. This suggests that biological value systems naturally represent *risk* and *uncertainty* by tracking the full probability distribution, not just its expectation.

Finally, classical RL presumes access to explicit scalar rewards, which are often unavailable in real-world settings such as dialogue, preference learning, or human–robot interaction. These limitations—the necessity of modeling risk or uncertainty for robustness and the need to align with human biological value systems and feedback—motivate the development of the distributional perspective and the study of RL from human feedback.

2.2 Distributional Reinforcement Learning

2.2.1 Distributional Bellman Equation and Convergence Properties

The seminal work of Bellemare et al. [12] formalized the concept of *return distributions*. Instead of learning the expected return, they proposed learning the entire probability distribution of the random return $Z^\pi(s,a)$ for a given policy π. The random return is the discounted sum of rewards from a state–action pair (s,a):

$$Z^\pi(s,a) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t),$$

and its law is denoted by $\eta^\pi(s,a)$. This framework acknowledges that the total return is a random variable, not a deterministic value, whose variability is often critical in real-world applications.

The core of this new perspective lies in the *distributional Bellman equation*, which defines a recursive relationship for the return distribution itself:

$$\mathcal{T}^\pi \eta(s,a) \overset{D}{=} r(s,a) + \gamma\, \eta(s',a'), \quad s' \sim \mathbb{P}(\cdot|s,a),\ a' \sim \pi(\cdot|s').$$

Here, $\overset{D}{=}$ denotes equality in distribution. This equation represents a shift from a functional mapping on scalars to an operator acting on distributions.

A crucial theoretical challenge was proving the convergence of this new Bellman operator. Bellemare et al. [12] showed that the distributional Bellman operator $\mathcal{T}^\pi$ is not a contraction mapping under conventional norms like the sup-norm or the total variation distance. However, they provided a novel convergence guarantee by proving that for any $p \geq 1$, $\mathcal{T}^\pi$ is a γ-contraction under the supremum p-Wasserstein distance:

$$\bar{W}_p(\mathcal{T}^\pi \eta, \mathcal{T}^\pi \eta') \leq \gamma\, \bar{W}_p(\eta, \eta'),$$

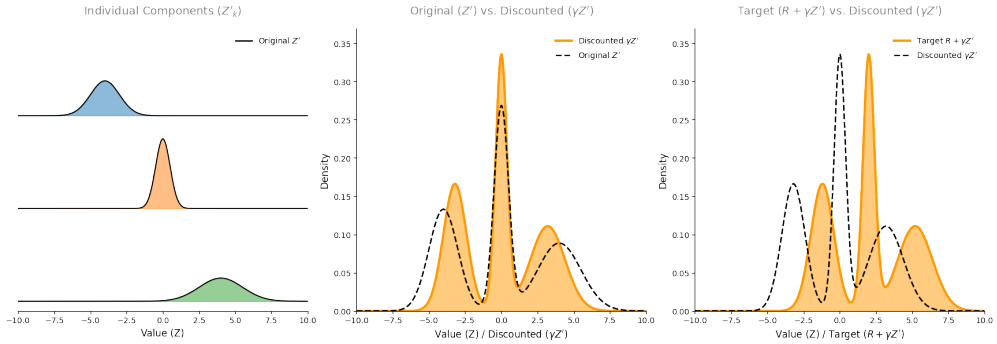


Figure 2.2: Visualization of the Distributional Bellman Operator $\mathcal{T}$. The figure illustrates the three key transformation steps required to construct the target distribution in distRL. **(Left) Components of the Next-State Distribution:** Shows the individual probability components that collectively define the value distribution of the subsequent state. **(Center) Scaling by the Discount Factor:** The full distribution is compounded and scaled by the discount factor γ. This results in the discounted distribution (solid orange fill), relative to the original distribution (dashed line). **(Right) Shifting by the Immediate Reward:** The final step shifts the discounted distribution (dashed line) horizontally by the immediate reward, yielding the final target distribution (solid orange fill).

where $\bar{W}_p(\eta, \eta') \coloneqq \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} W_p(\eta(s,a), \eta'(s,a))$. This result demonstrated that iterative application of the distributional Bellman operator converges to the unique true return distribution η^π, establishing DistRL as a mathematically sound generalization of classical RL.

2.2.2 Distributional Bellman Optimality and Instabilities

While the policy evaluation operator $\mathcal{T}^\pi$ enjoys the γ-contraction property in the supremum p-Wasserstein distance, the situation changes dramatically in the *control* setting. The *distributional Bellman optimality operator* $\mathcal{T}$ is defined by

applying a greedy policy with respect to the *expected returns*:

$$\mathcal{T}\eta(s,a) \overset{D}{=} r(s,a) + \gamma\, \eta(s', a^\star), \quad a^\star \in \arg\max_{a'} \mathbb{E}_{Z\sim\eta}[Z(s',a')],$$

where $Z(s',a') \sim \eta(s',a')$. This definition ensures that the means $\mathbb{E}[Z]$ evolve exactly as in the classical Bellman optimality operator.

However, Bellemare et al. [12] established that the distributional Bellman optimality operator $\mathcal{T}$ is fundamentally different from its policy-evaluation counterpart. First, $\mathcal{T}$ is not a contraction in any metric, i.e., there exist value distributions η, η' such that $\bar{W}_p(\mathcal{T}\eta, \mathcal{T}\eta') > \gamma\, \bar{W}_p(\eta, \eta')$.

Second, when multiple actions attain the same expected return, $\mathcal{T}$ may choose different greedy actions depending on the selection rule, and in this case no fixed point $\eta^\star = \mathcal{T}\eta^\star$ need exist. Third, even when a fixed point exists, the iterates $\eta_{k+1} := \mathcal{T}\eta_k$ are not guaranteed to converge; sequences can oscillate or converge only weakly to the broader set of nonstationary optimal value distributions.

By contrast, the expected values remain well behaved: for any η_1, η_2,

$$\left\| \mathbb{E}[\mathcal{T}\eta_1] - \mathbb{E}[\mathcal{T}\eta_2] \right\|_\infty \leq \gamma \left\| \mathbb{E}[\eta_1] - \mathbb{E}[\eta_2] \right\|_\infty,$$

which implies that $\mathbb{E}[\eta_k] \to Q^\star$ exponentially fast. This dichotomy highlights the central instability of distributional control: although the mean values converge reliably to the optimal Q-function, the underlying distributions can behave pathologically, exhibiting non-expansion, absence of fixed points, or non-convergence.

2.2.3 Approximation schemes in distributional RL

Categorical approximation A first practical algorithm in this line is `C51` [12], which represents return distributions on a fixed grid of N atoms $\{z_i\}_{i=1}^N$ equally spaced between $V_{\min}$ and $V_{\max}$. The agent learns probabilities $\{p_i(s,a)\}_{i=1}^N$ over

these atoms, and the distributional Bellman update $\mathcal{T}^\pi \eta$ is projected back onto this support using a projection operator Φ. Hence, the categorical update can be written as

$$\eta(s,a) \;\leftarrow\; \Phi_c\big(r + \gamma\, \eta(s',a')\big), \quad a' \sim \pi(\cdot|s'),$$

with Φ_c redistributing probability mass to the nearest atoms. Although this projection introduces bias (the support cannot adapt to the true distribution), Bellemare et al. [12] showed that the induced operator is non-expansive under the squared Cramér distance, which suffices to guarantee stability. Empirically, `C51` achieved state-of-the-art results on the Atari 2600 benchmark without other architectural modifications, suggesting that richer distributional targets can dramatically improve learning efficiency and representation capacity.

Quantile approximation To address the rigidity of fixed supports, Dabney et al. [31] proposed *Quantile Regression Deep Q-Network* (`QR-DQN`). Instead of fixed atoms, `QR-DQN` parameterizes the return distribution $\eta^\pi(s,a)$ by N learnable quantile values $\{\theta_i(s,a)\}_{i=1}^N$, corresponding to quantile levels $\tau_i = \frac{i}{N}$. The training objective minimizes the quantile regression loss

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{y\sim\eta}\Big[\rho_{\tau_i}\big(y - \theta_i(s,a)\big)\Big], \quad \rho_\tau(u) = u(\tau - \mathbf{1}\{u<0\}).$$

This is equivalent to minimizing the 1-Wasserstein distance between predicted and target distributions. In practice, `QR-DQN` employs the quantile Huber loss to improve robustness. Given quantile targets y sampled from the Bellman update and quantile predictions $\theta_i(s,a)$, the quantile Huber loss for quantile level τ_i is

$$\rho^\kappa_{\tau_i}(u) = \big|\tau_i - \mathbf{1}\{u<0\}\big|\, \mathcal{L}_\kappa(u), \quad u = y - \theta_i(s,a),$$

where $\mathcal{L}_\kappa(u)$ is the Huber loss defined by

$$\mathcal{L}_\kappa(u) = \begin{cases} \frac{1}{2}u^2, & |u| \le \kappa, \\ \kappa\big(|u| - \frac{1}{2}\kappa\big), & |u| > \kappa, \end{cases}$$

with $\kappa > 0$ a threshold parameter (typically set to 1). The full objective is then

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{y\sim\eta}\big[\rho_{\tau_i}^{\kappa}(y - \theta_i(s,a))\big].$$

This combines the robustness of the Huber loss with the asymmetry of quantile regression, yielding more stable optimization and reduced sensitivity to outliers. `QR-DQN` significantly outperformed `C51` on Atari benchmarks, and its theoretical grounding rests on the contraction of the projected Bellman operator in expectation.

Building on `QR-DQN`, Dabney et al. [30] introduced Implicit Quantile Networks (`IQN`), which approximate the entire quantile function $F_\eta^{-1} : [0,1] \to \mathbb{R}$ using a neural network. Rather than relying on a fixed set of quantile levels, `IQN` samples $\tau \sim \mathcal{U}[0,1]$ and outputs $\theta_\tau(s,a)$ as an estimate of $F_\eta^{-1}(\tau)$. This implicit formulation provides a flexible and fine-grained representation of return distributions, allowing the agent to capture distributional details beyond what fixed quantile schemes can offer.

Extending this idea, Yang et al. [105] proposed Fully Parameterized Quantile Functions (`FQF`), which jointly learn both the quantile fractions $\{\tau_i\}$ and the corresponding quantile values $\{\theta_i\}$. By adaptively allocating quantile fractions, FQF focuses model capacity on critical parts of the return distribution, leading to faster convergence and more accurate approximation. Empirically, both `IQN` and `FQF` improved sample efficiency and achieved state-of-the-art performance on Atari benchmarks, setting new records for distributional RL methods.

Moment-based approximation More recently, Nguyen-Tang et al. [70] proposed Moment Matching Distributional RL (MMDRL) and its deep variant *MMDQN*. Unlike `C51` and `QR-DQN`, which rely on predefined statistics (fixed atoms or quantiles), `MMDQN` represents each return distribution $\eta^\pi(s,a)$ using a set of learnable deterministic particles $\{Z_\theta(s,a)^i\}_{i=1}^N$. The update minimizes the Maximum Mean Discrepancy (MMD) between the current particle set and the Bellman target:

$$\mathcal{L}_{\mathrm{MMD}}(\theta) = \mathrm{MMD}^2\big(\{Z_\theta(s,a)^i\}_{i=1}^N, \{r + \gamma Z_{\theta^-}(s',a^\star)^i\}_{i=1}^N\big),$$

where $a^\star = \arg\max_{a'} \frac{1}{N}\sum_i Z_\theta(s',a')^i$ in the control setting. This formulation can be interpreted as implicitly matching all moments between the return distribution and its Bellman target. Theoretically, MMD provides sufficient conditions for contraction in certain kernel families and guarantees convergence at rate $O(1/\sqrt{n})$ regardless of the dimension. Empirically, `MMDQN` achieved superior performance on Atari-57, surpassing `C51` and `QR-DQN` while sharing the same network backbone, and achieving state-of-the-art mean human-normalized scores among non-distributed agents. By discarding the restriction of predefined statistics, `MMDQN` highlights a complementary perspective to quantile-based methods, with natural extensions toward `IQN`- and `FQF`-style architectures.

2.3 Reinforcement Learning from Human Feedback

2.3.1 Motivation and Origins

A central obstacle in reinforcement learning is the difficulty of designing reward functions that are both correct and aligned with human intentions. Even small misspecifications in a handcrafted reward can incentivize undesired behavior, a phenomenon broadly known as *reward hacking* or *reward exploitation*. Inverse reinforcement learning and imitation learning attempt to overcome this by inferring reward functions from expert demonstrations. However, expert data are often expensive to collect, and learned agents may be limited to imitating the demonstrated policy rather than surpassing it.

In contrast, human feedback provides a more flexible and scalable source of supervision. Instead of designing rewards directly, humans provide judgments about behaviors produced by the agent. This signal is easier to elicit: while a non-expert cannot always assign numerical scores, they can reliably indicate which of two outputs better reflects their preference. Early works in preference-based reinforcement learning (PbRL) demonstrated the feasibility of learning from such feedback [3]. The seminal study of Christiano et al. [27] scaled this idea, showing that collecting thousands of pairwise comparisons from non-expert annotators sufficed to train agents on Atari games and continuous-control robotics. Their work established RLHF as a practical methodology, and subsequent surveys [53] have consolidated RLHF as a central paradigm for aligning powerful AI systems with human values.

2.3.2 Canonical RLHF pipeline

The canonical RLHF framework can be described in three stages: (i) feedback collection, (ii) reward modeling, and (iii) policy optimization.

(i) Feedback collection Feedback can take various forms, including binary comparisons, rankings, scalar ratings, or textual critiques. Among these, pairwise trajectory or segment comparisons are the most common, as they balance cognitive simplicity with statistical efficiency [94, 118]. For language models, annotators often rank multiple responses for the same prompt, producing relative judgments that are robust to annotation noise. Granularity is another design choice: segment-level labels can improve sample efficiency by localizing signal, while trajectory-level labels provide global quality assessments. Active feedback strategies have also been explored, where the system selects queries that maximize expected information gain [3, 58].

(ii) Reward modeling Given preference data, the next step is to fit a parametric reward function R_ψ. A widely used formulation is the Bradley–Terry (BT) model [16]:

$$\mathbb{P}[\zeta^+ \succ \zeta^-] = \sigma\big(R_\psi(\zeta^+) - R_\psi(\zeta^-)\big), \quad R_\psi(\zeta) = \sum_t R_\psi(s_t, a_t),$$

where σ is the logistic sigmoid function. Training then reduces to maximum likelihood estimation over all annotated comparisons. This procedure ensures that R_ψ assigns higher scores to trajectories judged as better by humans, effectively transforming qualitative judgments into a quantitative reward landscape.

(iii) Policy optimization The final stage optimizes the agent's policy against R_ψ. To prevent divergence from the data distribution, the optimization is regularized relative to a reference policy π_{ref} (e.g., a supervised fine-tuned model). The canonical objective is

$$\max_\pi \ \mathbb{E}_{x,y\sim\pi}[R_\psi(x,y)] - \beta\,\mathrm{KL}\big(\pi(\cdot|x)\,\|\,\pi_{\text{ref}}(\cdot|x)\big),$$

which has a closed-form solution $\pi^\star(y|x) \propto \pi_{\text{ref}}(y|x)\exp(\beta R_\psi(x,y))$. In practice, this is approximated by PPO with KL penalties [8, 75, 94]. This pipeline enabled notable successes such as high-quality summarization models [94] and the alignment of large language models like InstructGPT [75].

2.3.3 Direct preference optimization (DPO) and extensions

To avoid the fragility of reward modeling, recent work proposes to optimize policies directly from preference data. Rafailov et al. [77] introduced *Direct Preference Optimization (DPO)*, which derives a surrogate objective by combining the Bradley–Terry likelihood with the KL-regularized optimal policy form. For a pair (ζ^+, ζ^-), the log-odds of preference simplify to a difference in log-policy ratios:

$$\log \frac{\mathbb{P}[\zeta^+ \succ \zeta^-]}{\mathbb{P}[\zeta^- \succ \zeta^+]} = \beta \left(\log \frac{\pi(\zeta^+)}{\pi_{\text{ref}}(\zeta^+)} - \log \frac{\pi(\zeta^-)}{\pi_{\text{ref}}(\zeta^-)} \right).$$

This yields the DPO loss

$$\mathcal{L}_{\text{DPO}}(\pi) = -\mathbb{E}_{(\zeta^+,\zeta^-)} \log \sigma \, \Big(\beta (\log \frac{\pi(\zeta^+)}{\pi_{\text{ref}}(\zeta^+)} - \log \frac{\pi(\zeta^-)}{\pi_{\text{ref}}(\zeta^-)}) \Big).$$

The loss resembles logistic regression on policy ratios, making optimization stable and bypassing the need for R_ψ. Empirically, DPO has proven more sample-efficient and less prone to reward miscalibration, and it is now widely used in LLM fine-tuning.

Variants and generalizations Following DPO, numerous extensions have been proposed. `SLiC-HF` [113] introduces sequence-likelihood calibration for greater robustness. Implicit Preference Optimization (`IPO`) [18] frames feedback as implicit gradient signals, while `ORPO` [46] modifies the functional form of the loss for stability and exploration. Weighted Preference Optimization (`WPO`) [116] emphasizes off-policy robustness by reweighting comparisons. General

frameworks such as ΨPO [7] unify these objectives under a single family. Active preference learning methods select informative queries to reduce annotation costs [10, 68]. Collectively, these advancements underscore a critical shift in preference learning research from simple functional alignment toward developing more robust, generalizable, and data-efficient algorithms that move beyond the explicit loss form of DPO.

2.3.4 Challenges and open directions

Although RLHF has advanced rapidly, several challenges remain fundamental.

Annotator heterogeneity Human preferences vary across individuals due to differences in knowledge, style, or bias. Modeling all annotators as sharing a single latent reward leads to noisy signals. Recent approaches address this by fitting mixture models, annotator-specific parameters, or hierarchical priors. Another direction is to collect feedback at varying granularity (token-, step-, or trajectory-level) to disambiguate local vs global preferences.

Off-policy data and likelihood mismatch Feedback datasets are often gathered from multiple behavior policies. Naively treating all data as optimal leads to *likelihood mismatch*, where suboptimality is conflated with stochasticity. This issue destabilizes offline RLHF and motivates corrections such as importance weighting, conservative sampling, or contrastive regularization. Policy-aware formulations that explicitly account for data-generation policies, such as regret-based models, provide a promising direction.

Data efficiency Collecting human preferences is costly. Active learning methods attempt to reduce annotation burden by querying comparisons that maximize expected information gain or model uncertainty [58]. Pairwise feedback can also

be supplemented with heuristics, synthetic labels, or preference propagation to reduce labeling requirements. Nonetheless, striking the balance between label cost and policy improvement remains an open challenge.

Theoretical guarantees While the Bradley–Terry likelihood enjoys asymptotic consistency, its finite-sample and robustness properties remain less understood. Recent theoretical analyses have begun to establish minimax rates and generalization bounds for preference learning [86, 104], as well as regret guarantees for preference-based policy optimization [22, 38]. Bridging the gap between these theoretical developments and empirical advances in large-scale RLHF remains an important research frontier.

2.4 Regret Minimization Framework

The study of decision-making under uncertainty has evolved through several theoretical frameworks, each attempting to explain how humans evaluate risk and choice. We begin with *Expected Utility Theory* and *Prospect Theory*, and then describe how *Regret Theory* extends these models by introducing counterfactual comparison as a core component of human decision-making. Finally, we discuss the formal notion of *regret* in reinforcement learning and the conceptual bridge that unites these perspectives.

2.4.1 From Expected Utility to Prospect Theory

Expected Utility Theory. Von Neumann and Morgenstern [99] assumes that a rational decision-maker assigns a scalar utility $u(x)$ to each outcome x and selects the action that maximizes expected utility:

$$A^* = \arg\max_i \sum_\omega p(\omega)\, u(x_i(\omega)).$$

While this framework provides a rigorous normative foundation for rational choice, it often fails to describe how humans actually make decisions. Empirical studies show that people systematically violate its axioms—exhibiting intransitive preferences, overweighting rare events, and shifting their choices depending on how equivalent outcomes are framed. These deviations highlight that human decision-making is not purely utility-maximizing but shaped by perception, context, and emotion, motivating the development of alternative descriptive theories such as prospect and regret theory.

Prospect Theory. Kahneman and Tversky [51] propose a psychologically grounded alternative to Expected Utility Theory in which outcomes are evaluated *relative to a reference point* r rather than in absolute terms. Let x denote an

outcome and write $v : \mathbb{R} \to \mathbb{R}$ for the *value function* applied to deviations from the reference point, with the following general properties: (i) **reference dependence** and normalization: $v(0) = 0$; (ii) **monotonicity** and continuity: v is continuous and strictly increasing; (iii) **diminishing sensitivity**: v is concave over gains ($x \geq r$) and convex over losses ($x < r$); (iv) **loss aversion**: the local slope at the reference is steeper for losses than for gains, e.g., $v'(0^-) > v'(0^+)$.

Given a prospect A_i that yields outcome $x_i(\omega)$ in state ω with probability $p(\omega)$, evaluation is

$$A^* = \arg\max_i \sum_\omega p(\omega)\, v\big(x_i(\omega) - r\big).$$

In the cumulative version (CPT) [97], objective probabilities are replaced by *decision weights* $w : [0,1] \to [0,1]$ (increasing, $w(0) = 0$, $w(1) = 1$) that typically overweight small probabilities and underweight large ones:

$$A^* = \arg\max_i \sum_\omega w\big(p(\omega)\big)\, v\big(x_i(\omega) - r\big).$$

Although Prospect Theory successfully explains various behavioral phenomena such as *framing effects* and *risk-aversion* in the loss domain, it still evaluates each option independently. It does not explicitly capture the emotional comparison between chosen and unchosen outcomes. The next refinement, *Regret Theory*, introduces this counterfactual component as a core element of decision-making.

2.4.2 Regret Theory: Anticipating Counterfactual Emotion

While Prospect Theory explains risk perception through reference-dependent valuation, it does not capture how individuals evaluate their realized outcomes relative to those that could have occurred. *Regret Theory*, first introduced by Loomes and Sugden [60] and later formalized by Sugden [95], extends the

analysis of risky choice by incorporating *counterfactual comparison*—a psychological mechanism through which people evaluate their choices against forgone alternatives.

The central idea is that decision-makers anticipate the emotional consequences of their choices, particularly the feeling of *regret* when an unchosen alternative would have led to a better outcome. Let x_a and x_b denote outcomes from two possible actions a and b, and $u(\cdot)$ be the utility function. The experienced utility of choosing a over b is expressed as

$$U(x_a; x_b) = u(x_a) - R\big(u(x_b) - u(x_a)\big),$$

where $R(\cdot)$ is an increasing function representing the psychological cost of regret (and rejoicing when the sign is reversed). Unlike Prospect Theory, which defines value relative to an *exogenous* reference point, Regret Theory defines the reference *endogenously*—it is determined by the outcome of the forgone option. Thus, satisfaction depends jointly on what is obtained and what is forgone, reflecting the inherently comparative and introspective nature of human decision-making.

This counterfactual structure enables Regret Theory to account for several empirically observed decision patterns in behavioral economics. First, the **anticipation of regret** can lead individuals to reverse preferences once feedback about alternatives becomes available, a phenomenon known as *preference reversal* [60, 61]. Second, experimental evidence shows a systematic asymmetry between *omissions* and *commissions*: people often prefer inaction when action carries a higher potential for self-blame or regret [39, 109]. Third, individuals tend to avoid high-variance options that may evoke intense regret, leading to patterns of *regret aversion* or cautious choice in repeated and feedback-driven environments [14, 15]. Together, these findings highlight how anticipated emotion

and self-evaluation influence real-world choice behavior.

In this sense, Regret Theory offers a perspective that complements Prospect Theory. Whereas Prospect Theory describes how people *perceive* risk and value through reference-dependent weighting, Regret Theory explains how they *evaluate* their own choices through counterfactual reasoning and emotional feedback, revealing the introspective dimension of human decision-making under uncertainty.

2.4.3 Algorithmic Regret in Reinforcement Learning

In contrast to behavioral regret, reinforcement learning (RL) interprets regret as a normative measure of learning efficiency. For a multi-armed bandit with optimal arm a^* and mean rewards μ_a, the cumulative regret after T rounds is defined as

$$\text{Regret}(T) = \sum_{t=1}^{T} (\mu_{a^*} - \mu_{a_t}),$$

quantifying the opportunity loss incurred by not always selecting the optimal arm. In a Markov decision process with horizon H and K episodes, the cumulative regret measures the discrepancy between the optimal value function $V_1^\star$ and the value realized by the learned policy π_k:

$$\text{Regret}(K) = \sum_{k=1}^{K} \left[V_1^\star(s_1^k) - V_1^{\pi_k}(s_1^k) \right].$$

A sublinear growth of this regret guarantees asymptotic optimality, establishing a theoretical foundation for efficient exploration and continual policy improvement.

2.4.4 Bridging Behavioral and Algorithmic Perspectives

Although behavioral and algorithmic notions of regret originate from distinct disciplines, their mathematical structures are remarkably aligned. Both quan-

Table 2.1: Structural Correspondence between Behavioral and Algorithmic Regret

Regret Theory	**Reinforcement Learning**
Utility of chosen action $u(x_i)$	Return under current policy $V^{\pi}(s_t)$
Utility of unchosen alternative $u(x_j)$	Return under optimal policy $V^{\star}(s_t)$
Difference $u(x_i) - u(x_j)$	Return gap $V^{\pi}(s_t) - V^{\star}(s_t)$

tify a counterfactual gap between realized and optimal outcomes, effectively measuring the cost of deviation from the best possible decision.

In both settings, learning and adaptation proceed through the reduction of this counterfactual difference: humans adjust their preferences to avoid future regret, while RL agents refine their policies to minimize performance loss. From this viewpoint, regret minimization emerges as a universal principle of adaptive decision-making under uncertainty—linking emotional reasoning with computational optimization.

2.4.5 Toward a Unified View

Seen from this unified perspective, regret serves as both a descriptive and normative construct. As a descriptive concept, it captures how humans psychologically evaluate their choices—through emotional and counterfactual comparisons with what might have been. As a normative concept, it defines how algorithms mathematically measure and minimize deviations from optimal behavior. In this way, regret bridges human introspection with computational rationality, revealing a shared structure between emotional learning and algorithmic optimization.

2.5 Summary

This chapter formalized Markov decision processes (MDPs) and emphasized why expectation-based objectives alone are inadequate for decision-making under uncertainty. We reviewed distributional reinforcement learning as a framework for modeling full return distributions and discussed RLHF as a practical approach to aligning policies with human judgment. Finally, we positioned regret as a unifying theoretical principle that connects behavioral realism with algorithmic efficiency. Chapters 3–5 build directly upon these foundations: mitigating exploration bias in distributional control, establishing unbiased functional updates with efficiency guarantees, and integrating regret into preference-based learning.

Chapter 3

Pitfall of Optimism: Distributional Reinforcement Learning by Randomizing Risk Criterion

Despite the richness of risk-sensitive information from return distribution, only a few DistRL methods [29, 64, 71, 96, 115] have tried to employ its benefits for exploration strategies which is essential in deep RL to find an optimal behavior within a few trials. The main reason is that the exploration strategies so far is based on *parametric (epistemic)* uncertainty which arise from insufficient or inaccurate data. In particular, *Optimism in the face of uncertainty* (OFU) is one of the fundamental exploration principles that employs parametric uncertainty to promote exploring less understood behaviors and to construct confidence set. In bandit or tabular MDP settings, OFU-based algorithms select an action with the highest upper-confidence bound (UCB) of parametric uncertainty which can be considered as the optimistic decision at the moment [20, 28].

However, in deep RL, it is hard to trivially estimate the parametric un-

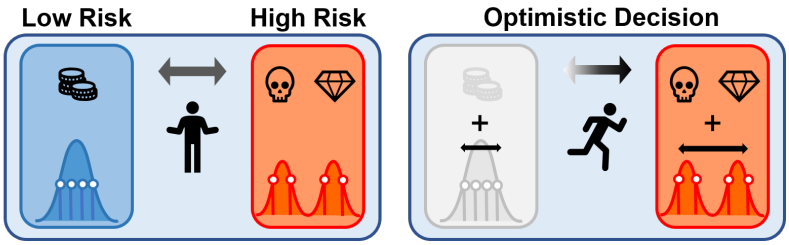


Figure 3.1: Illustrative example of why a biased risk criterion (naïve optimism) can degrade performance. Suppose two actions have similar expected returns, but different variances (intrinsic uncertainty). **(Left)** If an agent does not specify the risk criterion at the moment, the probability of selecting each action should be similar. **(Right)** As OFU principle encourages to decide uncertain behaviors, the empirical variance from quantiles was used as an estimate of uncertainty [54, 64, 66]. However, optimistic decision based on empirical variance inevitably leads a risk-seeking behavior, which causes biased action selection.

certainty accurately due to the black-box nature of neural networks and high-dimensionality of state-action space. Without further computational task, the estimated variance from distribution is extracted as a mixture of two types of uncertainty, making it difficult to decompose either component. For example, `DLTV` [64] was proposed as a distribution-based OFU exploration that decays bonus rate to suppress the effect of intrinsic uncertainty, which unintentionally induces a risk-seeking policy. Although `DLTV` is the first attempt to introduce OFU in distRL, we found that consistent optimism on the uncertainty of the estimated distribution still leads to biased exploration. We will refer to this side-effect as *one-sided tendency on risk*, where selecting an action based on a fixed risk criterion degrades learning performance. In Section 3.3, we will demonstrate the one-sided tendency on risk through a toy experiment and show that our proposed randomized approach is effective to avoid this side effect.

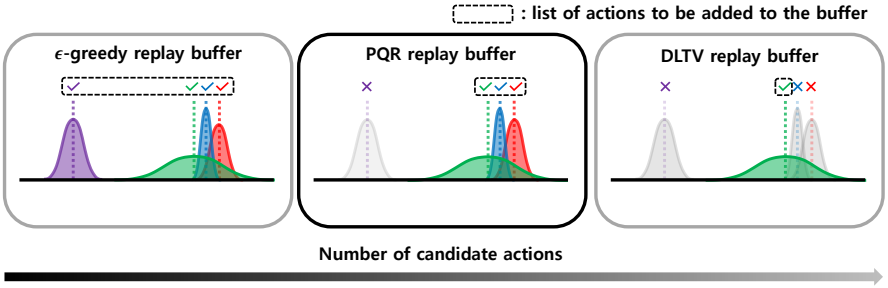


Figure 3.2: An illustrative example of proposed algorithm (PQR). Each distribution represents the empirical PDF of return. PQR benefits from excluding inferior actions and promoting unbiased selection with regards to high intrinsic uncertainty through randomized risk criterion.

In this paper, we introduce *Perturbed Distributional Bellman Optimality Operator (PDBOO)* to address the issue of biased exploration caused by a one-sided tendency on risk in action selection. We define the distributional perturbation on return distribution to re-evaluate the estimate of return by distorting the learned distribution with perturbation weight. To facilitate deep RL algortihm, we present *Perturbed Quantile Regression*(PQR) and test in Atari 55 games comparing with other distributional RL algorithms that have been verified for reproducibility by official platforms [19, 76].

In summary, our contributions are as follows.

- A randomized approach called perturbed quantile regression(PQR) is proposed without sacrificing the original (risk-neutral) optimality and improves over naïve optimistic strategies.

- A sufficient condition for convergence of the proposed Bellman operator is provided without satisfying the conventional contraction property.

3.1 Backgrounds & Related works

3.1.1 Distributional RL

We consider a Markov decision process (MDP) which is defined as a tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ where $\mathcal{S}$ is a finite state space, $\mathcal{A}$ is a finite action space, $P : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to [0, 1]$ is the transition probability, R is the random variable of rewards in $[-R_{\max}, R_{\max}]$, and $\gamma \in [0, 1)$ is the discount factor. We define a stochastic policy $\pi(\cdot|s)$ which is a conditional distribution over $\mathcal{A}$ given state s. For a fixed policy π, we denote $Z^\pi(s, a)$ as a random variable of return distribution of state-action pair (s, a) following the policy π. We attain $Z^\pi(s, a) = \sum_{t=0}^{\infty} \gamma^t R(S_t, A_t)$, where $S_{t+1} \sim P(\cdot|S_t, A_t),\ A_t \sim \pi(\cdot|S_t)$ and $S_0 = s,\ A_0 = a$. Then, we define an action-value function as $Q^\pi(s, a) = \mathbb{E}[Z^\pi(s, a)]$ in $[-V_{\max}, V_{\max}]$ where $V_{\max} = R_{\max}/(1 - \gamma)$. For regularity, we further notice that the space of action-value distributions $\mathcal{Z}$ has the first moment bounded by $V_{\max}$:

$$\mathcal{Z} = \left\{ Z : \mathcal{S} \times \mathcal{A} \to \mathscr{P}(\mathbb{R}) \middle|\ \mathbb{E}[|Z(s, a)|] \leq V_{\max}, \forall (s, a) \right\}.$$

In distributional RL, the return distribution for the fixed π can be computed via dynamic programming with the distributional Bellman operator defined as,

$$\mathcal{T}^\pi Z(s, a) \overset{D}{=} R(s, a) + \gamma Z(S', A'), \quad S' \sim P(\cdot|s, a),\ A' \sim \pi(\cdot|S')$$

where $\overset{D}{=}$ denotes that both random variables share the same probability distribution. We can compute the optimal return distribution by using the distributional Bellman optimality operator defined as,

$$\mathcal{T} Z(s, a) \overset{D}{=} R(s, a) + \gamma Z(S', a^*), \quad S' \sim P(\cdot|s, a),\ a^* = \operatorname*{argmax}_{a'}\ \mathbb{E}_Z[Z(S', a')].$$

Bellemare et al. [12] have shown that $\mathcal{T}^\pi$ is a contraction in a maximal form of the Wasserstein metric but $\mathcal{T}$ is not a contraction in any metric. Combining

with the expectation operator, $\mathbb{E} \circ \mathcal{T}$ is a contraction so that we can guarantee that the expectation of Z converges to the optimal state-action value. Another notable difference is that the convergence of a return distribution is not generally guaranteed to be unique, unless there is a total ordering $\prec$ on the set of greedy policies.

3.1.2 Exploration on Distributional RL

To combine with deep RL, a parametric distribution Z_θ is used to learn a return distribution. Dabney et al. [31] have employed a quantile regression to approximate the full distribution by letting $Z_\theta(s,a) = \frac{1}{N}\sum_{i=1}^{N} \delta_{\theta_i(s,a)}$ where θ represents the locations of a mixture of N Dirac delta functions. Each θ_i represents the value where the cumulative probability is $\tau_i = \frac{i}{N}$. By using the quantile representation with the distributional Bellman optimality operator, the problem can be formulated as a minimization problem as,

$$\begin{aligned}\theta &= \arg\min_{\theta'} D\left(Z_{\theta'}(s_t,a_t), \mathcal{T} Z_{\theta^-}(s_t,a_t)\right) \\ &= \arg\min_{\theta'} \sum_{i,j=1}^{N} \frac{\rho^{\kappa}_{\hat{\tau}_i}(r_t + \gamma\theta_j^-(s_{t+1},a') - \theta'_i(s_t,a_t))}{N}\end{aligned}$$

where (s_t, a_t, r_t, s_{t+1}) is a given transition pair, $a' := \text{argmax}_{a'}\ \mathbb{E}_Z[Z_\theta(s_{t+1},a')]$, $\hat{\tau}_i = \frac{\tau_{i-1}+\tau_i}{2}$, $\rho^{\kappa}_{\hat{\tau}_i}(x) := |\hat{\tau}_i - \delta_{\{x<0\}}|\mathcal{L}_\kappa(x)$, and $\mathcal{L}_\kappa(x) := x^2/2$ for $|x| \le \kappa$ and $\mathcal{L}_\kappa(x) := \kappa(|x| - \frac{1}{2}\kappa)$, otherwise.

Based on the quantile regression, Dabney et al. [31] have proposed a quantile regression deep Q network (`QR-DQN`) that shows better empirical performance than the categorical approach [12], since the quantile regression does not restrict the bounds for return.

As deep RL typically did, `QR-DQN` adjusts ϵ-greedy schedule, which selects the greedy action with probability $1-\epsilon$ and otherwise selects random available

actions uniformly. The majority of `QR-DQN` variants [30, 105] rely on the same exploration method. However, such approaches do not put aside inferior actions from the selection list and thus suffers from a loss [74]. Hence, designing a schedule to select a statistically plausible action is crucial for efficient exploration.

In recent studies, Mavrin et al. [64] modifies the criterion of action selection for efficient exploration based on optimism in the face of uncertainty. Using left truncated variance as a bonus term and decaying ratio c_t to suppress the intrinsic uncertainty, `DLTV` was proposed as an uncertainty-based exploration in DistRL without using ϵ-greedy schedule. The criterion of `DLTV` is described as:

$$a^* = \underset{a'}{\operatorname{argmax}} \left(\mathbb{E}_P[Z(s', a')] + c_t \sqrt{\sigma_+^2(s', a')} \right),$$
$$c_t = c\sqrt{\frac{\log t}{t}},\ \sigma_+^2 = \frac{1}{2N} \sum_{i=\frac{N}{2}}^{N} (\theta_{\frac{N}{2}} - \theta_i)^2,$$

where θ_i's are the values of quantile level τ_i.

3.1.3 Risk in Distributional RL

Instead of an expected value, risk-sensitive RL is to maximize a pre-defined risk measure such as Mean-Variance [112], Value-at-Risk (VaR) [26], or Conditional Value-at-Risk (CVaR) [78, 79] and results in different classes of optimal policy. Especially, Dabney et al. [30] interprets risk measures as the expected utility function of the return, i.e., $\mathbb{E}_Z[U(Z(s, a))]$. If the utility function U is linear, the policy obtained under such risk measure is called *risk-neutral*. If U is concave or convex, the resulting policy is termed as *risk-averse* or *risk-seeking*, respectively. In general, a *distortion risk measure* is a generalized expression of risk measure which is generated from the distortion function.

Definition 3.1.1. Let $h : [0, 1] \to [0, 1]$ be a **distortion function** such that $h(0) = 0, h(1) = 1$ and non-decreasing. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and

a random variable $Z : \Omega \to \mathbb{R}$, a **distortion risk measure** ρ_h corresponding to a distortion function h is defined by:

$$\rho_h(Z) \coloneqq \mathbb{E}^{h(\mathbb{P})}[Z] = \int_{-\infty}^{\infty} z \frac{\partial}{\partial z}(h \circ F_Z)(z) dz,$$

where F_Z is the cumulative distribution function of Z.

In fact, non-decreasing property of h makes it possible to distort the distribution of Z while satisfying the fundamental property of CDF. Note that the concavity or the convexity of distortion function also implies risk-averse or seeking behavior, respectively. Dhaene et al. [34] showed that any distorted expectation can be expressed as weighted averages of quantiles. In other words, generating a distortion risk measure is equivalent to choosing a reweighting distribution.

Fortunately, DistRL has a suitable configuration for risk-sensitive decision making by using distortion risk measure. Chow et al. [25] and Stanko and Macek [93] considered risk-sensitive RL with a CVaR objective for robust decision making. Dabney et al. [30] expanded the class of policies on arbitrary distortion risk measures and investigated the effects of a distinct distortion risk measures by changing the sampling distribution for quantile targets τ. Zhang and Yao [111] have suggested `QUOTA` which derives different policies corresponding to different risk levels and considers them as options. Moskovitz et al. [67] have proposed `TOP-TD3`, an ensemble technique of distributional critics that balances between optimism and pessimism for continuous control.

3.2 Perturbation in Distributional RL

3.2.1 Perturbed Distributional Bellman Optimality Operator

To choose statistically plausible actions which may be maximal for certain risk criterion, we will generate a distortion risk measure involved in a pre-defined

constraint set, called an *ambiguity set*. The ambiguity set, originated from distributionally robust optimization (DRO) literature, is a family of distribution characterized by a certain statistical distance such as *ϕ-divergence* or *Wasserstein distance* [36, 88]. In this paper, we will examine the ambiguity set defined by the discrepancy between distortion risk measure and expectation. We say the sampled reweighting distribution ξ as *(distributional) perturbation* and define it as follows:

Definition 3.2.1. (Perturbation Gap, Ambiguity Set) Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, Define a random variable $X : \Omega \to \mathbb{R}$ and a set of probability density functions $\Xi = \left\{\xi : 0 \leq \xi(w) < \infty,\ \int_{w\in\Omega} \xi(w)\mathbb{P}(dw) = 1\right\}$. For a given constraint set $\mathcal{U} \subset \Xi$, we say $\xi \in \mathcal{U}$ as a **(distributional) perturbation** from $\mathcal{U}$ and denote the $\xi-$weighted expectation of X as follows:

$$\mathbb{E}_\xi[X] \coloneqq \int_{w\in\Omega} X(w)\xi(w)\mathbb{P}(dw),$$

which can be interpreted as the expectation of X under some probability measure $\mathbb{Q}$, where $\xi = d\mathbb{Q}/d\mathbb{P}$ is the Radon-Nikodym derivative of $\mathbb{Q}$ with respect to $\mathbb{P}$. We further define $d(X;\xi) = |\mathbb{E}[X] - \mathbb{E}_\xi[X]|$ as **perturbation gap** of X with respect to ξ. Then, for a given constant $\Delta \geq 0$, the **ambiguity set** with the bound Δ is defined as

$$\mathcal{U}_\Delta(X) = \Big\{\xi \in \Xi : |\mathbb{E}[X] - \mathbb{E}_\xi[X]| \leq \Delta\Big\}.$$

For brevity, we omit the input w from a random variable unless confusing. Since ξ is a probability density function, $\mathbb{E}_\xi[X]$ is an induced risk measure with respect to a reference measure $\mathbb{P}$. Intuitively, $\xi(w)$ can be viewed as a distortion to generate a different probability measure and vary the risk tendency. The aspect of using distortion risk measures looks similar to `IQN` [30]. However, instead of changing the sampling distribution of quantile level τ implicitly, we

reweight each quantile from the ambiguity set. This allows us to control the maximum allowable distortion with bound Δ, whereas the risk measure in `IQN` does not change throughout learning. In Section 3.2.3, we suggest a practical method to construct the ambiguity set.

Now, we characterize *perturbed distributional Bellman optimality operator* (PDBOO) $\mathcal{T}_\xi$ for a fixed perturbation $\xi \in \mathcal{U}_\Delta(Z)$ written as below:

$$\mathcal{T}_\xi Z(s,a) \overset{D}{=} R(s,a) + \gamma Z(S', a^*(\xi)),$$
$$S' \sim P(\cdot|s,a), \ a^*(\xi) = \underset{a'}{\operatorname{argmax}} \ \mathbb{E}_{\xi,P}[Z(s',a')].$$

Notice that $\xi \equiv 1$ corresponds to a base expectation, i.e., $\mathbb{E}_{\xi,P} = \mathbb{E}_P$, which recovers the standard distributional Bellman optimality operator $\mathcal{T}$. Specifically, PDBOO perturbs the estimated distribution only to select the optimal behavior, while the target is updated with the original (unperturbed) return distribution.

If we consider the time-varying bound of ambiguity set, scheduling Δ_t is a key ingredient to determine whether PDBOO will efficiently explore or converge. Intuitively, if an agent continues to sample the distortion risk measure from a fixed ambiguity set with a constant Δ, there is a possibility of selecting sub-optimal actions after sufficient exploration, which may not guarantee eventual convergence. Hence, scheduling a constraint of ambiguity set properly at each action selection is crucial to guarantee convergence.

Based on the quantile model Z_θ, our work can be summarized into two parts. First, we aim to minimize the expected discrepancy between Z_θ and $\mathcal{T}_\xi Z_{\theta^-}$ where ξ is sampled from ambiguity set $\mathcal{U}_\Delta$. To clarify notation, we write $\mathbb{E}_\xi[\cdot]$ as a $\xi-$weighted expectation and $\mathbb{E}_{\xi\sim\mathscr{P}(\mathcal{U}_\Delta)}[\cdot]$ as an expectation with respect to ξ which is sampled from $\mathcal{U}_\Delta$. Then, our goal is to minimize the perturbed distributional Bellman objective with sampling procedure $\mathscr{P}$:

$$\min_{\theta'} \mathbb{E}_{\xi_t\sim\mathscr{P}(\mathcal{U}_{\Delta_t})}[D(Z_{\theta'}(s,a), \mathcal{T}_{\xi_t} Z_{\theta^-}(s,a))] \tag{3.1}$$

where we use the Huber quantile loss as a discrepancy on $Z_{\theta'}$ and $\mathcal{T}_\xi Z_{\theta^-}$ at timestep t. In typical risk-sensitive RL or distributionally robust RL, the Bellman optimality equation is reformulated for a pre-defined risk measure [25, 92, 106]. In contrast, PDBOO has a significant distinction in that it performs dynamic programming that adheres to the risk-neutral optimal policy while randomizing the risk criterion at every step. By using min-expectation instead of min-max operator, we suggest unbiased exploration that can avoid leading to overly pessimistic policies. Furthermore, considering a sequence ξ_t which converges to 1 in probability, we derive a sufficient condition of Δ_t that the expectation of any composition of the operators $\mathbb{E} \circ \mathcal{T}_{\xi_{n:1}} \coloneqq \mathbb{E} \circ \mathcal{T}_{\xi_n} \circ \mathcal{T}_{\xi_{n-1}} \circ \cdots \circ \mathcal{T}_{\xi_1}$ has the same unique fixed point as the standard. These results are remarkable that we can apply the diverse variations of distributional Bellman operators for learning.

3.2.2 Convergence of the perturbed distributional Bellman optimality operator

Unlike conventional convergence proofs, PDBOO is time-varying and not a contraction, so it covers a wider class of Bellman operators than before. Since the infinite composition of time-varying Bellman operators does not necessarily converge or have the same unique fixed point, we provide the sufficient condition in this section. We denote the iteration as $Z^{(n+1)} \coloneqq \mathcal{T}_{\xi_{n+1}} Z^{(n)}$, $Z^{(0)} = Z$ for each timestep $n > 0$, and the intersection of ambiguity set as $\bar{\mathcal{U}}_{\Delta_n}(Z^{(n-1)}) \coloneqq \bigcap_{s,a} \mathcal{U}_{\Delta_n}\left(Z^{(n-1)}(s,a)\right)$.

Assumption 3.2.2. Suppose that $\sum_{n=1}^{\infty} \Delta_n < \infty$ and ξ_n is uniformly bounded.

Theorem 3.2.3. *(Weaker Contraction Property) Let ξ_n be sampled from an ambiguity set $\bar{\mathcal{U}}_{\Delta_n}(Z^{(n-1)})$ for every iteration. If Assumption 3.2.2 holds, then the expectation of any composition of operators $\mathbb{E}\mathcal{T}_{\xi_{n:1}}$ converges, i.e.,*

$\mathbb{E}\mathcal{T}_{\xi_{n:1}}[Z] \to \mathbb{E}[Z^*]$. *Moreover, the following bound holds,*

$$\sup_{s,a} \left\| \mathbb{E}[Z^{(n)}(s,a)] - \mathbb{E}[Z^*(s,a)] \right\|_{s,a} \leq \sum_{k=n}^{\infty} \left(2\gamma^{k-1} V_{max} + 2 \sum_{i=1}^{k} \gamma^i (\Delta_{k+2-i} + \Delta_{k+1-i}) \right).$$

Practically, satisfying Assumption 3.2.2 is not strict to characterize the landscape of scheduling. Theorem 3.2.3 states that even without satisfying γ-contraction property, we can show that $\mathbb{E}[Z^*]$ is the fixed point for the operator $\mathbb{E}\mathcal{T}_{\xi_{n:1}}$. However, $\mathbb{E}[Z^*]$ is not yet guaranteed to be "unique" fixed point for any $Z \in \mathcal{Z}$. Nevertheless, we can show that $\mathbb{E}[Z^*]$ is, in fact, the solution of the standard Bellman optimality equation, which is already known to have a unique solution.

Theorem 3.2.4. *If Assumption 3.2.2 holds,* $\mathbb{E}[Z^*]$ *is the unique fixed point of Bellman optimality equation for any* $Z \in \mathcal{Z}$.

As a result, PDBOO generally achieves the unique fixed point of the standard Bellman operator. Unlike previous distribution-based or risk-sensitive approaches, PDBOO has the theoretical compatibility to obtain a risk-neutral optimal policy even if the risk measure is randomly sampled during training procedure. For proof, see Appendix A.1.3.

3.2.3 Practical Algorithm with Distributional Perturbation

In this section, we propose a **perturbed quantile regression (`PQR`)** that is a practical algorithm for distributional reinforcement learning. Our quantile model is updated by minimizing the objective function (3.1) induced by PDBOO. Since we employ a quantile model, sampling a reweight function ξ can be reduced into sampling an N-dimensional weight vector $\boldsymbol{\xi} := [\xi_1, \cdots, \xi_N]$ where $\sum_{i=1}^{N} \xi_i = N$

Algorithm 1 Perturbed Quantile Regression (PQR)

Input: (s, a, r, s'), $\gamma \in [0, 1)$, timestep $t > 0$, $\epsilon > 0$, concentration $\boldsymbol{\beta}$

Initialize $\Delta_0 > 0$.

$\Delta_t \leftarrow \Delta_0 t^{-(1+\epsilon)}$. // Assumption 3.2.2

$\boldsymbol{\xi} \leftarrow \max\left(\mathbf{1}^N + \Delta_t(N\boldsymbol{x} - \mathbf{1}^N), 0\right)$ where $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\beta})$ // Sample

$\xi \sim \bar{\mathcal{U}}_{\Delta_t}(Z^{(t)})$

$\boldsymbol{\xi} \leftarrow N\boldsymbol{\xi}/\sum \xi_i$ // Refine as a weighting function

$a^* \leftarrow \mathrm{argmax}_{a'} \mathbb{E}_{\boldsymbol{\xi}}[Z(s', a')]$ // Select greedy action with perturbed return

$\mathcal{T}\theta_j \leftarrow r + \gamma\theta_j(s', a^*), \quad \forall j$ // Target update with unperturbed distribution

$t \leftarrow t + 1$

Output: $\sum_{i=1}^N \mathbb{E}_j[\rho^{\kappa}_{\hat{\tau}_i}(\mathcal{T}\theta_j - \theta_i(s, a))]$

and $\xi_i \geq 0$ for all $i \in \{1, \cdots, N\}$. Based on the QR-DQN setup, note that the condition $\int_{w\in\Omega} \xi(w)\mathbb{P}(dw) = 1$ turns into $\sum_{i=1}^N \frac{1}{N}\xi_i = 1$, since the quantile level is set as $\tau_i = \frac{i}{N}$.

A key issue is how to construct an ambiguity set with bound Δ_t and then sample $\boldsymbol{\xi}$. A natural class of distribution for practical use is the *symmetric Dirichlet distribution* with concentration $\boldsymbol{\beta}$, which represents distribution over distributions. (i.e. $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\beta})$.) We sample a random vector, $\boldsymbol{x} \sim \mathrm{Dir}(\boldsymbol{\beta})$, and define the reweight distribution as $\boldsymbol{\xi} := \mathbf{1}^N + \alpha(N\boldsymbol{x} - \mathbf{1}^N)$. From the construction of $\boldsymbol{\xi}$, we have $1 - \alpha \leq \xi_i \leq 1 + \alpha(N-1)$ for all i and it follows that $|1 - \xi_i| \leq \alpha(N-1)$. By controlling α, we can bound the deviation of ξ_i from 1 and bound the perturbation gap as

$$\sup_{s,a} |\mathbb{E}[Z(s,a)] - \mathbb{E}_{\xi}[Z(s,a)]| = \sup_{s,a} \left| \int_{w\in\Omega} Z(w; s, a)(1 - \xi(w))\mathbb{P}(dw) \right|$$

$$\leq \sup_{w\in\Omega} |1 - \xi(w)| \sup_{s,a} \mathbb{E}[|Z(s,a)|] \leq \sup_{w\in\Omega} |1 - \xi(w)| V_{\max} \leq \alpha(N-1)V_{\max}.$$

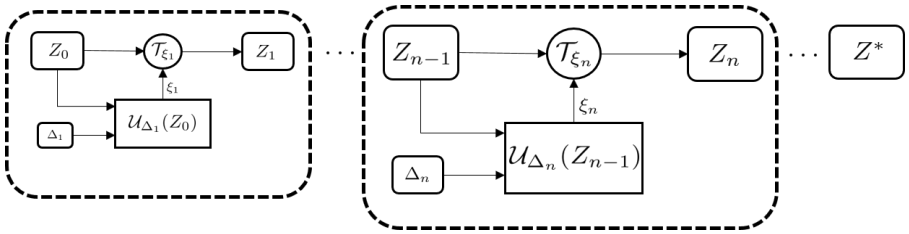


Figure 3.3: Pipeline of PDBOO.

Hence, letting $\alpha \leq \frac{\Delta}{(N-1)V_{\max}}$ is sufficient to obtain $d(Z;\xi) \leq \Delta$ in the quantile setting. We set $\boldsymbol{\beta} = 0.05 \cdot \mathbf{1}^N$ to generate a constructive perturbation ξ_n which gap is close to the bound Δ_n. For Assumption 3.2.2, our default schedule is set as $\Delta_t = \Delta_0 t^{-(1+\epsilon)}$ where $\epsilon = 0.001$.

Figure 3.3 shows the pipeline of our algorithm. With the schedule of perturbation bound $\{\Delta_n\}$, the ambiguity set $\mathcal{U}_{\Delta_n}(Z_{n-1})$ can be defined by previous Z_{n-1}. For each step, (distributional) perturbation ξ_n is sampled from $\mathcal{U}_{\Delta_n}(Z_{n-1})$ by the symmetric Dirichlet distribution and then PDBOO $\mathcal{T}_{\xi_n}$ can be performed.

3.3 Experiments on Stochastic Enviornments with High Intrinsic Uncertainty

Our experiments aim to investigate the following questions.

1. Does randomizing risk criterion successfully escape from the biased exploration in stochastic environments?

2. Can `PQR` accurately estimate a return distribution?

3. Can a perturbation-based exploration perform sucessfully as a behavior policy for the full Atari benckmark?

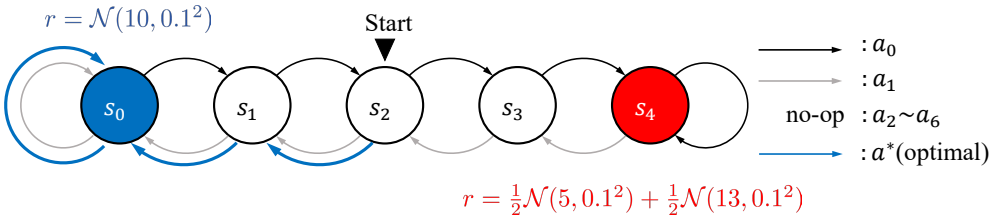


Figure 3.4: Illustration of the N-Chain environment [73] with high uncertainty starting from state s_2. To emphasize the intrinsic uncertainty, the reward of state s_4 was set as a mixture model composed of two Gaussian distributions. Blue arrows indicate the risk-neutral optimal policy in this MDPs.

Algorithm 2 Perturbed DLTV (`p-DLTV`)

Input: transition (s, a, r, s'), discount $\gamma \in [0, 1)$

$Q(s', a') = \frac{1}{N} \sum_j \theta_j(s', a')$

$c_t \sim c\, \mathcal{N}(0, \frac{\ln t}{t})$ // Randomize the coefficient

$a^* \leftarrow \text{argmax}_{a'}(Q(s', a') + c_t \sqrt{\sigma_+^2(s', a')})$

$\mathcal{T}\theta_j \leftarrow r + \gamma \theta_j(s', a^*), \quad \forall j$

Output: $\sum_{i=1}^{N} \mathbb{E}_j[\rho_{\hat{\tau}_i}^{\kappa}(\mathcal{T}\theta_j - \theta_i(s, a))]$

3.3.1 N-Chain Enviornment

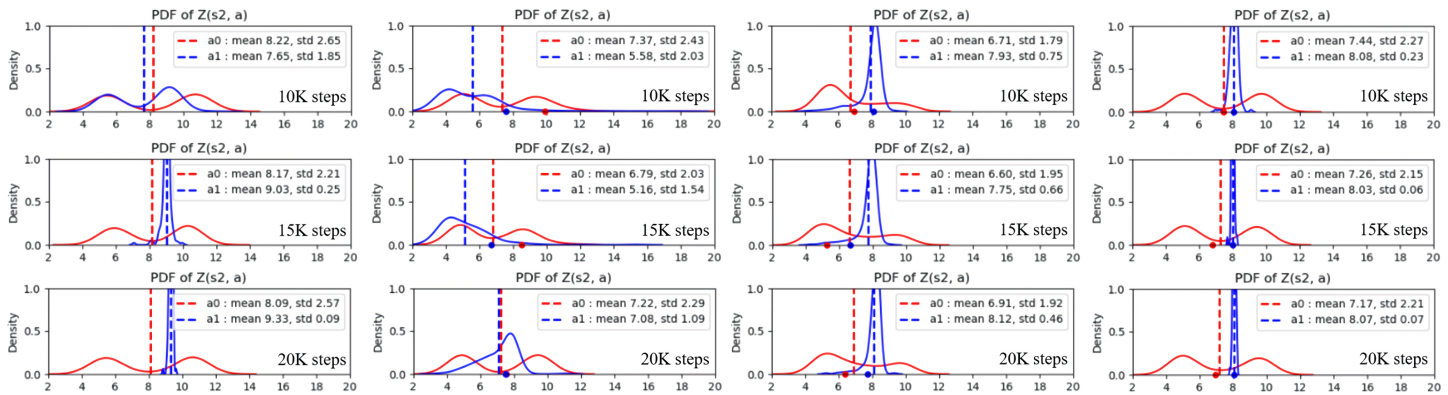
For intuitive comparison between optimism and randomized criterion, we design `p-DLTV`, a perturbed variant of `DLTV`, where coefficient c_t is multiplied by a normal distribution $\mathcal{N}(0, 1^2)$.

N-Chain with high intrinsic uncertainty. We extend N-Chain environment [73] with stochastic reward to evaluate action selection methods. A schematic diagram of the stochastic N-Chain environment is depicted in Figure 3.4. The reward is only given in the leftmost and rightmost states and the game termi-

nates when one of the reward states is reached. We set the leftmost reward as $\mathcal{N}(10, 0.1^2)$ and the rightmost reward as $\frac{1}{2}\mathcal{N}(5, 0.1^2) + \frac{1}{2}\mathcal{N}(13, 0.1^2)$ which has a lower mean as 9 but higher variance. The agent always starts from the middle state s_2 and should move toward the leftmost state s_0 to achieve the greatest expected return. For each state, the agent can take one of six available actions: left, right, and 4 no-op actions. The optimal policy with respect to mean is to move left twice from the start. We set the discount factor $\gamma = 0.9$ and the coefficient $c = 50$.

Despite the simple configuration, the possibility to obtain higher reward in suboptimal state than the optimal state makes it difficult for an agent to determine which policy is optimal until it experiences enough to discern the characteristics of each distribution. Thus, the goal of our toy experiment is to evaluate how rapidly each algorithm could find a risk-neutral optimal policy. The results of varying the size of variance are reported in Appendix 3.1.

Analysis of Experimental Results. As we design the mean of each return is intended to be similar, examining the learning behavior of the empirical return distribution for each algorithm can provide fruitful insights. Figure 3.5 shows the empirical PDF of return distribution by using Gaussian kernel density estimation. In Figure 3.5(b), `DLTV` fails to estimate the true optimal return distribution. While the return of $(s_2, \texttt{right})$ (red line) is correctly estimated toward the ground truth, $(s_2, \texttt{left})$ (blue line) does not capture the shape and mean due to the lack of experience. At 20K timestep, the agent begins to see other actions, but the monotonic scheduling already makes the decision like exploitation. Hence, decaying schedule of optimism is not a way to solve the underlying problem. Notably, `p-DLTV` made a much better estimate than `DLTV` only by changing from optimism to a randomized scheme. In comparison, `PQR`



(a) QR-DQN (b) DLTV (c) p-DLTV (d) PQR (Ours)

Figure 3.5: Empirical return distribution plot in N-Chain environment. The ground truth of each distribution is $\gamma^2\mathcal{N}(10, 0.1^2)$ and $\gamma^2[\frac{1}{2}\mathcal{N}(5, 0.1^2) + \frac{1}{2}\mathcal{N}(13, 0.1^2)]$. Each dot represents an indicator for choosing action. Since `QR-DQN` does not depend on other criterion, the dots are omitted.

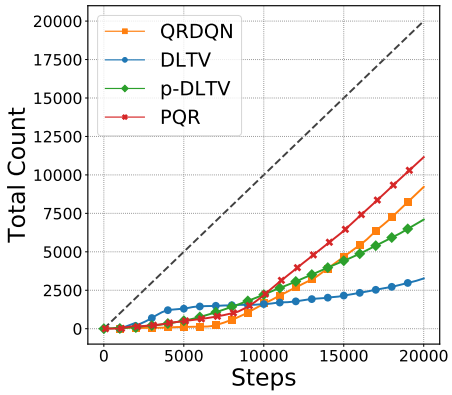


Figure 3.6: Total count of performing true optimal action. The oracle (dashed line) is to perform the optimal action from start to end.

estimates the ground truth much better than other baselines with much closer mean and standard-deviation.

Figure 3.6 shows the number of timesteps when the optimal policy was actually performed to see the interference of biased criterion. Since the optimal policy consists of the same index a_1, we plot the total count of performing the optimal action with 10 seeds. From the slope of each line, it is observed that `DLTV` selects the suboptimal action even if the optimal policy was initially performed. In contrast, `p-DLTV` avoids getting stuck by randomizing criterion and eventually finds the true optimal policy. The experimental results demonstrate that randomizing the criterion is a simple but effective way for exploration on training process.

Hyperparameter Sensitivity. In Figure 3.7, we compute the 2-Wasserstein distance from the ground truth return distribution $\mathcal{N}(10\gamma^2, (0.1\gamma^2)^2)$. Except for `QR-DQN` , each initial hyperparameter $\{c, \Delta_0\}$ was implemented with grid search on $[1, 5, 10, 50, 100, 500, 1000, 5000]$ in 5 different seeds. As the hyperparameter

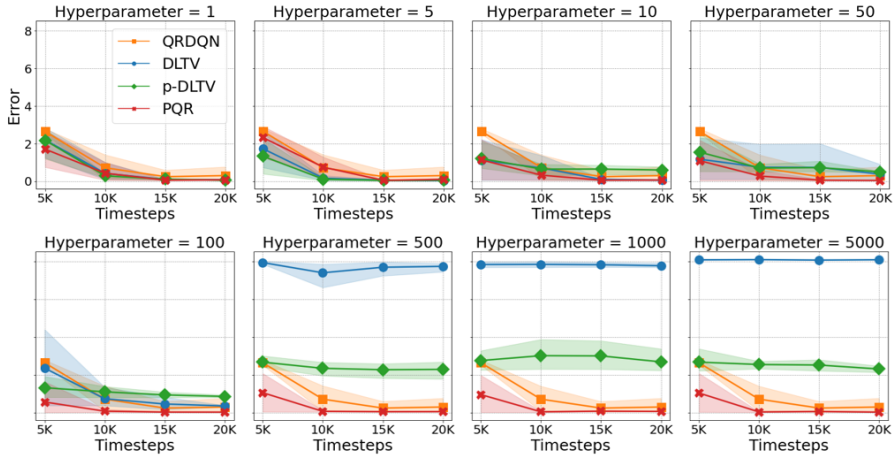


Figure 3.7: 2-Wasserstein distance between the empirical return distribution and the ground truth $\mathcal{N}(8.1, 0.081^2)$. We use `QR-DQN` with a fixed setting of ϵ-greedy as a reference baseline, because the hyperparameter of ϵ-greedy is not related to the scale of Q-values.

decreases, each agent is likely to behave as exploitation. One interesting aspect is that, while it may be difficult for `DLTV` and `p-DLTV` to balance the scale between the return and bonus term, `PQR` shows robust performance to the initial hyperparameter. This is because the distorted return is bounded by the support of return distribution, so that `PQR` implicitly tunes the scale of exploration. In practice, we set Δ_0 to be sufficiently large. See Table A.1 in Appendix A.2.1.

To explore the effect of intrinsic uncertainty, we run multiple experiments with various reward settings for the rightmost state as keeping their mean at 9. As the distance between two Gaussians was increased, the performance of `DLTV` decrease gradually, while other algorithms show consistent results. The result implies the interference of one-sided tendency on risk is proportional to the magnitude of the intrinsic uncertainty and the randomized criterion is effective in escaping from the issue.

Table 3.1: Total counts of performing true optimal action with 4 different seeds.

Reward Setting	(8,10)	(7,11)	(6,12)	(5,13)	(4,14)	(3,15)	(2,16)	(1,17)
QR-DQN	12293	11381	11827	12108	10041	11419	9696	11619
DLTV	9997	9172	9646	9251	7941	6964	7896	7257
p-DLTV	14344	14497	13769	**15507**	14469	**14034**	14068	13404
PQR	**14546**	**15018**	**14693**	15142	**15361**	13859	**14602**	**14354**

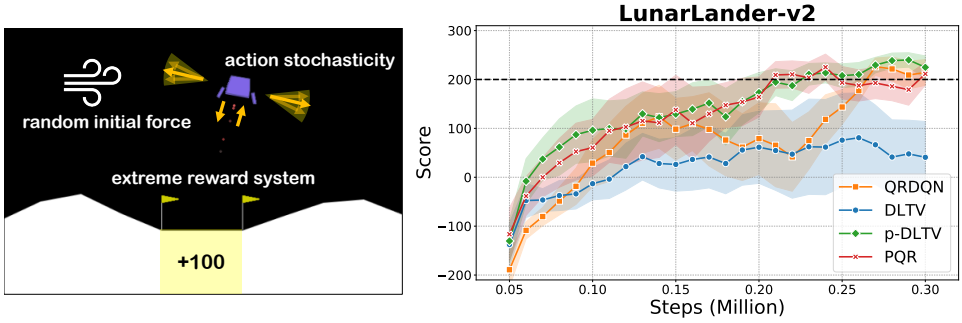


Figure 3.8: (**Left**) Three main environmental factors causing high intrinsic uncertainty on LunarLander-v2. (**Right**) Performance on LunarLander-v2

3.3.2 LunarLander-v2

To verify the effectiveness of the proposed algorithm in the complex environment with high intrinsic uncertainty, we conduct the experiment on LunarLander-v2. We have focused on three main factors that increase the intrinsic uncertainty from the structural design of LunarLander environment:

- **Random initial force:** The lander starts at the top center with an random initial force.
- **Action stochasticity:** The noise of engines causes different transitions with same action.

Table 3.2: Mean and median of best scores across 55 Atari games, measured as percentages of human baseline. Reference values are from Quan and Ostrovski [76] and Castro et al. [19].

50M Performance	Mean	Median	> human	> DQN
DQN-zoo (no-ops)	314%	55%	18	0
DQN-dopamine (sticky)	401%	51%	15	0
QR-DQN-zoo (no-ops)	559%	118%	29	47
QR-DQN-dopamine (sticky)	562%	93%	27	46
IQN-zoo (no-ops)	902%	131%	21	50
IQN-dopamine (sticky)	940%	124%	32	51
RAINBOW-zoo (no-ops)	1160%	154%	37	52
RAINBOW-dopamine (sticky)	965%	123%	35	53
PQR-zoo (no-ops)	1121%	124%	33	53
PQR-dopamine (sticky)	962%	123%	35	51

- **Extreme reward system:** If the lander crashes, it receives -100 points. If the lander comes to rest, it receives +100 points.

Therefore, several returns with a fixed policy have a high variance. As previously discussed about the fixedness from N-Chain environment, we can demonstrate that randomized approaches, `PQR` and `p-DLTV`, outperform other baselines in LunarLander-v2.

3.3.3 55 Atari Games

We compare our algorithm to various DistRL baselines, which have demonstrated good performance on RL benchmarks. In Table 3.2, we evaluated 55 Atari results,

averaging over 5 different seeds at 50M frames. We compared with the published score of QR-DQN [31], IQN [30], and Rainbow [45] via the report of DQN-Zoo [76] and Dopamine [19] benchmark for reliability. This comparison is noteworthy since our proposed method only applys perturbation-based exploration strategy and outperforms advanced variants of QR-DQN. [1]

No-ops Protocol. First, we follow the evaluation protocol of [11, 65] on full set of Atari games implemented in OpenAI's Gym [17]. Even if it is well known that the *no-ops* protocol does not provide enough stochasticity to avoid memorization, intrinsic uncertainty still exists due to the *random frame skipping* [62]. While PQR cannot enjoy the environmental stochasticity by the deterministic dynamics of Atari games, PQR achieved 562% performance gain in the mean of human-normalized score over QR-DQN, which is comparable results to Rainbow. From the raw scores of 55 games, PQR wins 39 games against QR-DQN and 34 games against IQN.

Sticky actions protocol. To prevent the deterministic dynamics of Atari games, Machado et al. [62] proposes injecting stochasticity scheme, called *sticky actions*, by forcing to repeat the previous action with probability $p = 0.25$. Sticky actions protocol prevents agents from relying on memorization and allows robust evaluation. In Figure 3.9, PQR shows steeper learning curves, even without any support of advanced schemes, such as n-step updates for Rainbow or IQN. In particular, PQR dramatically improves over IQN and Rainbow in Assault, Battlezone, Beamrider, Berzerk and Bowling. In Table 3.2, PQR shows robust median score against the injected stochasticity.

It should be noted that IQN benefits from the generalized form of distri-

[1] In Dopamine framework, IQN was implemented with $n-$step updates with $n = 3$, which improves performance.

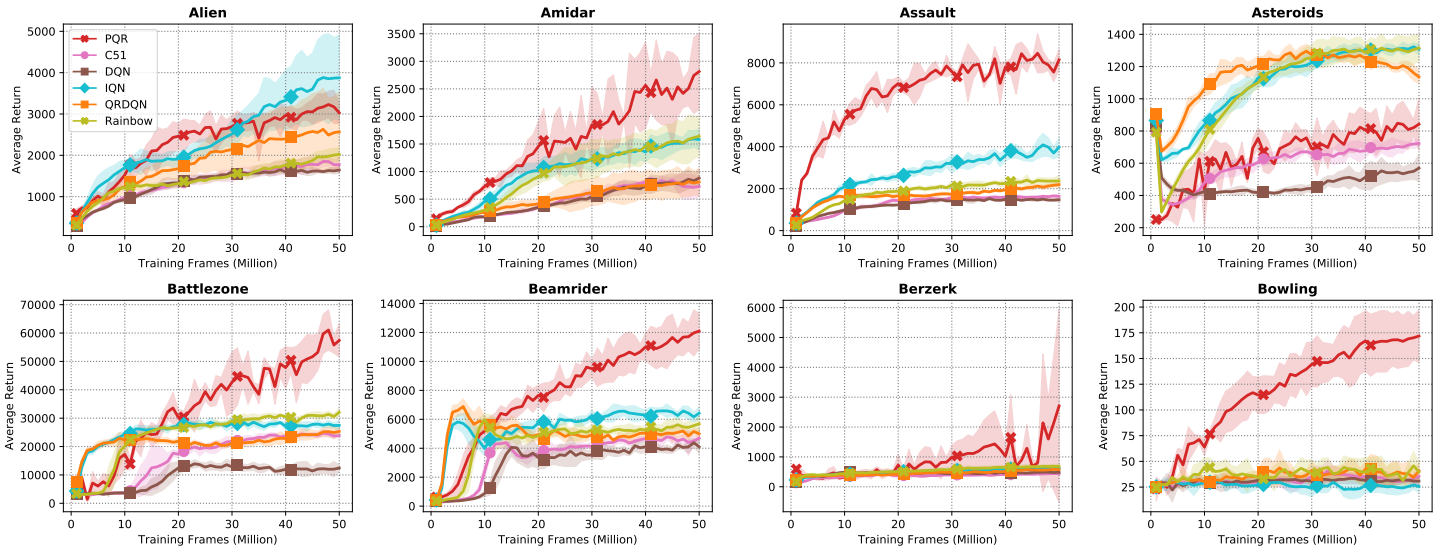


Figure 3.9: Evaluation curves on 8 Atari games with 3 random seeds for 50 million frames following *sticky actions* protocol [62]. Reference values are from Castro et al. [19].

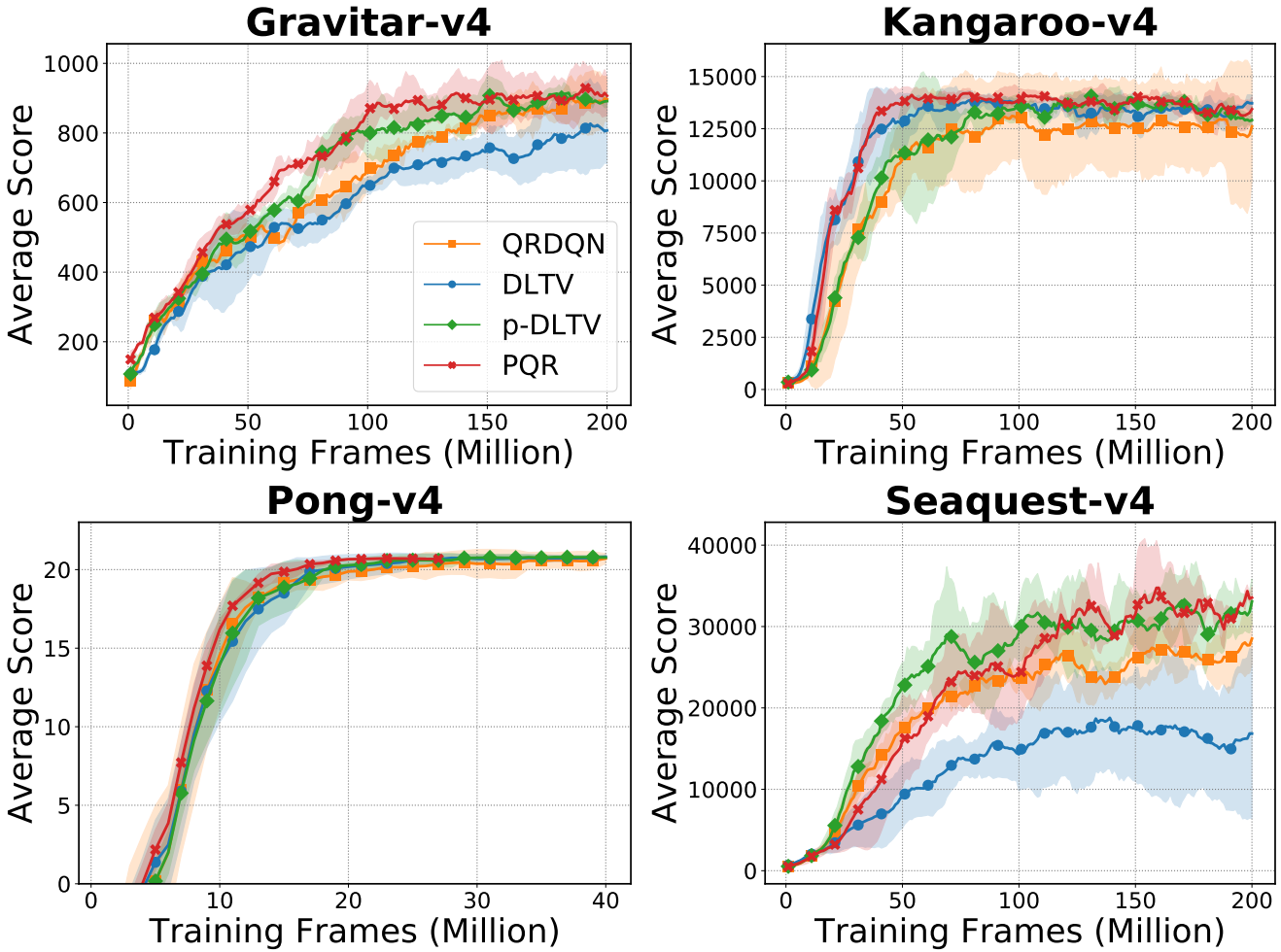


Figure 3.10: Evaluation curves on Atari games. All curves are smoothed over 10 consecutive steps with three random seeds. In case of Pong-v4, we resize the x-axis, since it can easily obtain the optimal policy with few interactions due to its environmental simplicity.

butional outputs, which reduces the approximation error from the number of quantiles output. Compare to `IQN`, `PQR` does not rely on prior distortion risk measure such as CVaR [24], Wang [102] or CPW [97], but instead randomly samples the risk measure and evaluates it with a risk-neutral criterion. Another notable difference is that `PQR` shows the better or competitive performance solely through its **exploration strategies**, compared to ϵ-greedy baselines, such as `QR-DQN`, `IQN`, and especially `Rainbow`. Note that `Rainbow` enjoys a combination of several orthogonal improvements such as double Q-learning, prioritized replay, dueling networks, and n-step updates.

We test our algorithm under 30 no-op settings to align with previous works. We compare our baseline results with results from the DQN-Zoo framework [76], which provides the full benchmark results on 55 Atari games at 50M and 200M frames. We report the average of the best scores over 5 seeds for each baseline algorithms up to 50M frames.

However, recent studies tried to follow the setting proposed by Machado et al. [62] for reproducibility, where they recommended using sticky actions. Hence, we provide all human normalized scores results across 55 Atari games for 50M frames including previous report of Dopamine and DQN-Zoo framework to help the follow-up researchers as a reference. We exclude `Defender` and `Surround` which is not reported on Yang et al. [105] because of relialbility issues in the Dopamine framework. In summary,

- DQN Zoo framework corresponds to 30 no-op settings (version **v4**).

- Dopamine framework corresponds to sticky actions protocol (version **v0**).

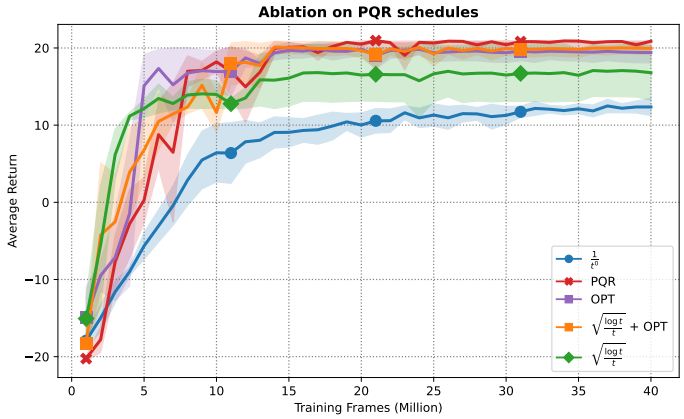


Figure 3.11: Evaluation curves on Pong-v4 environments.

Ablation on PQR schedules

To investigate the effect of the schedule of Δ_t, we run the experiment on Pong-v4 and set up several baselines as follows:

- $1/t^0$: A fixed size ambiguity set. $\Delta_t = O(1)$
- `PQR` : Our main algorithm. $\Delta_t = O(1/t^{1+\epsilon})$
- `OPT` : We fix the output vector, sampled from the Dirichlet distribution, as $[0, 0, \ldots, 1]$, forcing the agent to estimate only optimistically.
- $\sqrt{\log t/t}$: We imitate the schedule of p-DLTV (which does not satisfy the sufficient condition we presented). $\Delta_t = O(\sqrt{\log t/t})$
- $\sqrt{\log t/t}$ + `OPT` : We imitate the schedule of DLTV (which does not satisfy the sufficient condition we presented). We fixed the output vector, sampled from the Dirichlet distribution, as $[0, 0, \ldots, 1]$, forcing the agent to estimate only optimistically. $\Delta_t = |O(\sqrt{\log t/t})|$

In this experiment, our proposed PQR is the only method that stably achieves the maximum score with low variance. In the case of optimism (purple, orange curve), the agent learns quickly in the early stages, but converges without reaching the maximum score. In the case of fixed ambiguity set (blue curve), it converges to suboptimal and eventually shows low performance. This result implies the necessity of time-varying schedule of Δ_t. Finally, when imitating the schedule of p-DLTV (green curve), the performance also degrades implying that the proposed sufficient condition is quite tight.

3.4 Related Works & Discussion

Randomized or perturbation-based exploration has been focused due to its strong empirical performance and simplicity. In tabular RL, Osband et al. [74] proposed randomized least-squares value iteration (RLSVI) using random perturbations for statistically and computationally efficient exploration. Ishfaq et al. [47] leveraged the idea into optimistic reward sampling by perturbing rewards and regularizers. However, existing perturbation-based methods requires tuning of the hyperparameter for the variance of injected Gaussian noise and depend on well-crafted feature vectors in advance. On the other hand, PDBOO does not rely on the scale of rewards or uncertainties due to the built-in scaling mechanism of risk measures. Additionally, we successfully extend PQR to deep RL scenarios in distributional lens, where feature vectors are not provided, but learned during training.

3.4.1 Comparison with QUOTA

Zhang and Yao [111] have proposed Quantile Option Architecture(QUOTA) which derives different policies corresponding to different risk levels and consider them as options. By using an option-based framework, the agent learns a

high-level policy that adaptively selects a pessimistic or optimistic exploration strategy. While QUOTA has a similar approach in high-level idea, PQR gives a lot of improvements in both theoretical analysis and experimental results.

- **Theoretical guarantees of convergence toward risk-neutrality.** Since the agent selects via randomized risk criterion, the natural question is:

  "How should we control the injected randomness without sacrificing the original purpose of risk-neutrality?"

  In this work, we provide the sufficient condition for convergence without sacrificing risk-neutral perspective. Although QUOTA explores by using optimism or pessimism of a value distribution, there is no discussion whether the convergence is guaranteed toward a risk-neutral objective.

- **Explaining the effectiveness of randomized strategy.**

  QUOTA tested on two Markov chains to illustrate the inefficiency of expectation-based RL. It assumed that each task has an inherent, but unknown, preferred risk strategy, so agents should learn hidden preference. In contrast, we point out that the amount of inherent (intrinsic) uncertainty causes the inefficiency of fixed optimism or pessimism based exploration.

- **Significant performance difference in experimental results.**

  QUOTA is based on option-based learning which requires an additional option-value network. While QUOTA aims to control risk-sensitivity by transforming into an option O, the introduction of an option-value network requires the agent to explore an action space $|O| \times |A|$. This opposes the idea of efficient exploration as a factor that increases the complexity of learning. In contrast, PQR does not require a additional network and

explores over the original action space. In addition, PQR does not artificially discretize the ambiguity set of risk measurement. Another main reason is that PQR does not depend on an greedy schedule which is well-known for inefficient exploration strategies in tabular episodic MDP [48]. PQR solely explores its own strategies which is a simple yet effective approach. However, QUOTA depends on a greedy schedule in both quantile and option networks.

3.4.2 Reproducibility issues on DLTV

For the expected concerns about the comparison with DLTV, we address some technical issues to correct misconceptions of their performance. Before we reproduce the empirical results of DLTV, Mavrin et al. [64] did not report each raw scores of Atari games, but only the relative performance with cumulative rewards comparing with QR-DQN. While DLTV was reported to have a cumulative reward 4.8 times greater than QR-DQN, such gain mainly comes from VENTURE which is evaluated as 22,700% from their metric (i.e., 463% performance gain solely). However, the approximate raw score of VENTURE was 900 which is lower than our score of 993.3. Hence, the report with cumulative rewards causes a severe misconception that can be overestimated where the human-normalized score is commonly used for evaluation metrics. For a fair comparison, we computed based on mean and median of human-normalized scores and obtained results of 603.66% and 109.90%. Due to the absence of public results, however, DLTV was inevitably excluded from the comparison with human-normalized score in the main paper for reliability. In Table 3.3 and A.4, we report our raw scores and human-normalized score of DLTV based on QR-DQN_zoo performance.

Table 3.3: Performance comparison among QUOTA, DLTV, and PQR on 55 Atari games. Values in the first block indicate the number of games (out of 55) where the row method outperforms the column method.

Comparison	QUOTA > QR-DQN_Zhang	QR-DQN_zoo > QR-DQN_Zhang	PQR > QUOTA	PQR > QR-DQN_Zhang	PQR > DLTV
# wins (out of 55)	30	34	**42**	**42**	**39**

Human-normalized score summary

Metric	QR-DQN_zoo	QR-DQN_Zhang	QUOTA	DLTV	PQR
Average HN Score	505.02	463.47	383.70	603.66	**1078.00**
Median HN Score	120.74	78.07	91.08	109.90	**129.25**

3.5 Summary

This chapter introduces a general framework for perturbation in distributional Reinforcement Learning, leveraging the inherent characteristics of the return distribution. We identify a critical limitation in traditional Optimism Under Uncertainty exploration methods: they often conflate epistemic uncertainty with aleatoric uncertainty by relying on variance estimates of the return distribution. This confusion leads to a persistent risk-seeking bias and the collection of skewed data during exploration.

To resolve this issue, we propose the Perturbed Quantile Regression (PQR) algorithm. PQR facilitates robust action selection by introducing a randomly perturbed risk measure applied to the distorted risk scale. Theoretically, we demonstrate that PQR effectively avoids biased exploration while maintaining convergence to the true optimal policy. Empirically, PQR achieves superior performance over existing variance-based exploration methods across various

benchmarks, including 55 Atari games. The PQR algorithm thus provides a principled method to mitigate exploration bias in distributional RL, contributing significantly to the field of risk-sensitive exploration.

Chapter 4

Bellman Unbiasedness: Toward Provably Efficient Distributional Reinforcement Learning with General Value Function Approximation

While the distributional approach offers richer information about return uncertainty, it introduces two key theoretical challenges that distinguish it from expectation-based RL: *(i)* the *infinite-dimensionality* of the distribution and *(ii)* the complexity of *online distributional updates*. In practice, we must rely on approximations using a finite number of statistical functionals, such as categorical or quantile representations. However, previous work has shown that not all statistical functionals can be exactly learned through the Bellman operator, leading to the concept of *Bellman Closedness* [81], which characterizes preserved functionals. While Bellman closedness is a necessary structural property, it is insufficient for online learning; statistical functionals of the target distribution

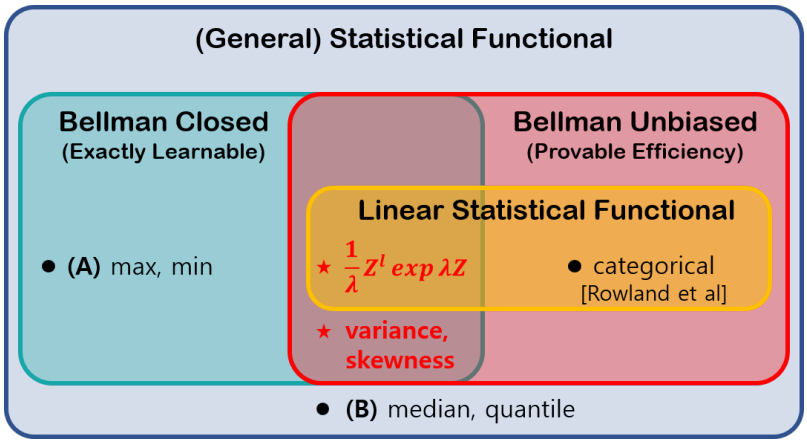


Figure 4.1: **Venn-Diagram of Statistical Functional Classes.** The diagram illustrates categories of statistical functional. **(Yellow ∩ Blue)** Within the linear statistical functional class, Rowland et al. [81] showed that the only functionals satisfying Bellman closedness are moment functionals. **(Red ∩ Blue)** We extend this concept by introducing the notion of *Bellman unbiasedness*, which not only covers moment functionals but also includes central moment functionals from the broader class including nonlinear statistical functionals. **(Yellow ∩ Blue^c)** According to Lemmas 3.2 and 4.4 of Rowland et al. [81], categorical functionals are linear but not Bellman closed. **(A)** Maximum and minimum functionals are Bellman closed, while they are not unbiasedly estimatable. **(B)** Median and quantile functionals are neither Bellman closed nor unbiased, highlighting that they are not proper to encode the distribution in terms of exactness. The proofs corresponding to each region are provided in Appendix B.3.

must also be unbiasedly estimated from the sampled distribution. This is critical in the context of developing an algorithm that efficiently explores from a regret minimization perspective while simultaneously performing distributional Bellman updates in an online manner.

To address this, we introduce the key concept of *Bellman Unbiasedness*, a property ensuring precise information learnability of a distribution from a finite number of samples in an online setting. We prove that the exponential-polynomial functional remains the unique solution in a class including nonlinear statistical functionals that satisfies both Bellman Closedness and Bellman Unbi-

asedness. Based on this, we propose Statistical Functional Least-Squares Value Iteration (`SF-LSVI`), an exactly learnable and provably efficient DistRL algorithm with general value function approximation. Our framework yields the tight regret upper bound $\tilde{O}(d_E H^{3/2}\sqrt{K})$, marking the first such result with a weaker structural assumption compared to prior work in distRL.

4.1 Related Work

Distributional RL. In classical RL, the Bellman equation, which is based on expected returns, has a closed-form expression. However, it remains unclear whether any statistical functionals of return distribution always have their corresponding closed-form expressions. Rowland et al. [81] introduced the notion of *Bellman closedness* for collections of statistical functionals that can be updated in a closed form via Bellman update. They showed that the only Bellman-closed statistical functionals in the discounted setting are the moments $\mathbb{E}_{Z\sim\eta}[Z^k]$. More recently, Marthe et al. [63] proposed a general framework for distRL, where the agent plans to maximize its own utility functionals instead of expected return, formalizing this property as *Bellman Optimizability*. They further demonstrated that in the undiscounted setting, the only W_1-continuous and linear Bellman optimizable statistical functionals are exponential utilities $\frac{1}{\lambda}\log\mathbb{E}_{Z\sim\eta}[\exp(\lambda Z)]$.

In practice, `C51` [12] and `QR-DQN` [31] are notable distributional RL algorithms where the convergence guarantees of sampled-based algorithms are proved [80, 82]. Dabney et al. [30] expanded the class of policies on arbitrary distortion risk measures by taking the based distribution non-uniformly and improve the sample efficiency from their implicit representation of the return distribution. Cho et al. [23] highlighted the drawbacks of optimistic exploration in distRL, introducing a randomized exploration that perturbs the distribution when the agent selects next action.

Table 4.1: Comparison for different methods under distributional RL framework. $\mathcal{H}$ represents a subspace of infinite-dimensional space $\mathcal{F}^\infty$. To bound the eluder dimesion d_E, Wang et al. [100] and Chen et al. [21] assumed the discretized reward MDP.

Algorithm	Regret	Eluder dimension d_E	Bellman Completeness	MDP assumption	Finite Representation	Exactly Learnable
O-DISCO [100]	$\tilde{\mathcal{O}}(\text{poly}(d_E H)\sqrt{K})$	$\dim_E(\mathcal{H}, \epsilon)$	distributional BC	discretized reward, small-loss bound	✗	✗
V-EST-LSR [21]	$\tilde{\mathcal{O}}(d_E H^2\sqrt{K})$ [1]	$\dim_E(\mathcal{H}, \epsilon)$	distributional BC	discretized reward, lipschitz continuity	✗	✗
SF-LSVI [Ours]	$\tilde{\mathcal{O}}(d_E H^{\frac{3}{2}}\sqrt{K})$	$\dim_E(\mathcal{F}^N, \epsilon)$	statistical functional BC	none	✓	✓

RL with General Value Function Approximation. Regret bounds have been studied for a long time in online RL, across various domains such as bandit [1, 57, 83], tabular RL [5, 48, 52, 72, 74], and linear function approximation [49, 103, 108]. In recent years, deep RL has shown significant performance using deep neural networks as function approximators, and attempts have been made to analyze whether it is efficient in terms of general function approximation [2, 50]. Wang et al. [101] established a provably efficient RL algorithm with general value function approximation based on the eluder dimension d_E [83] and achieves a regret upper bound of $\tilde{O}(\text{poly}(d_E H)\sqrt{K})$. To circumvent the intractability from computing the upper confidence bound, Ishfaq et al. [47] injected the stochasticity on the training data and get the optimistic value function instead of upper confidence bound, enhancing computationally efficiency. Beyond risk-neutral setting, several prior works have shown regret bounds under risk-sensitive objectives (e.g., entropic risk [37, 59], CVaR [9]), which align with our approach in that they are built on a distribution framework. Liang and Luo [59] achieved the regret upper bound of $\tilde{O}(\exp(H)\sqrt{|\mathcal{S}|^2|\mathcal{A}|H^2K})$ and the lower bound of $\Omega(\exp(H)\sqrt{|\mathcal{S}||\mathcal{A}|HK})$ in tabular setting.

[1] In Chen et al. [21], the regret bound is written as $\tilde{O}(d_E L_\infty(\rho)H\sqrt{K})$, where $L_\infty(\rho)$ represents the lipschitz constant of the risk measure ρ, i.e., $|\rho(Z)-\rho(Z')| \leq L_\infty(\rho)\|F_Z - F_{Z'}\|_\infty$. Since $L_\infty(\rho) \geq H$ in risk-neutral setting, we translate the regret bound into $\tilde{O}(d_E H^2\sqrt{K})$.

DistRL with General Value Function Approximation. Recently, only few efforts have aimed to bridge the gap between two fields. Wang et al. [100] proposed a distributional RL algorithm, `O-DISCO`, which enjoys small-loss bound by using a log-likelihood objective. Similarly, Chen et al. [21] provided a risk-sensitive RL framework with static lipschitz risk measure. While these studies analyze within a distributional framework, they do not address the intractability of implementation in infinite-dimensional space of distributions. In contrast, our approach focuses on a statistical functional framework, providing a detailed comparison with other distRL methods as shown in Table 5.1.

4.2 Preliminaries

Episodic MDP. We consider a episodic Markov decision process which is defined as a $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r)$ characterized by state space $\mathcal{S}$, action space $\mathcal{A}$, horizon length H, transition kernels $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$, and reward $r = \{r_h\}_{h\in[H]}$ at step $h \in [H]$. The agent interacts with the environment across K episodes. For each $k \in [K]$ and $h \in [H]$, $\mathbb{H}_h^k = (s_1^1, a_1^1, \ldots, s_H^1, a_H^1, \ldots, s_h^k, a_h^k)$ represents the history up to step h at episode k. We assume the reward is bounded by $[0, 1]$ and the agent always transit to terminal state s_{end} at step $H + 1$ with $r_{H+1} = 0$.

Policy and Value Functions. A (deterministic) policy π is a collection of H functions $\{\pi_h : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$. Given a policy π, a step $h \in [H]$, and a state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the Q and V-function are defined as $Q_h^\pi(s, a)(: \mathcal{S} \times \mathcal{A} \to \mathbb{R}) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s, a_h = a\right]$ and $V_h^\pi(s)(: \mathcal{S} \to \mathbb{R}) \coloneqq \mathbb{E}_\pi \left[\sum_{h'=h}^H r_{h'}(s_{h'}, a_{h'}) \mid s_h = s\right]$.

Random Variables and Distributions. For a sample space Ω, we extend the definition of the Q-function into a random variable and its distribution,

$$Z_h^\pi(s,a)(:\mathcal{S}\times\mathcal{A}\times\Omega\to\mathbb{R}) \coloneqq \sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) \mid s_h=s, a_h=a, a_{h'}=\pi_{h'}(s_{h'}),$$

$$\eta_h^\pi(s,a)(:\mathcal{S}\times\mathcal{A}\to\mathscr{P}(\mathbb{R})) \coloneqq \mathrm{law}(Z_h^\pi(s,a)).$$

Analogously, we extend the definition of V-function by introducing a bar notation.

$$\bar{Z}_h^\pi(s)(:\mathcal{S}\times\Omega\to\mathbb{R}) \coloneqq \sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) \mid s_h=s, a_{h'}=\pi_{h'}(s_{h'}),$$

$$\bar{\eta}_h^\pi(s)(:\mathcal{S}\to\mathscr{P}(\mathbb{R})) \coloneqq \mathrm{law}(\bar{Z}_h^\pi(s)).$$

Note that $\bar{Z}_h^\pi(s) = Z_h^\pi(s,\pi(s))$ and $\bar{\eta}_h^\pi(s) = \eta_h^\pi(s,\pi(s))$. We use $\pi^\star$ to denote an optimal policy (*i.e.,* $\pi_h^\star(\cdot|s) = \arg\max_\pi V_h^\pi(s)$) and denote $V_h^\star(s) = V_h^{\pi^\star}(s), Q_h^\star(s,a) = Q_h^{\pi^\star}(s,a)$, $\eta_h^\star(s,a) = \eta_h^{\pi^\star}(s,a)$, and $\bar{\eta}_h^\star(s) = \bar{\eta}_h^{\pi^\star}(s)$. For notational simplicity, we denote the expectation over transition, $[\mathbb{P}_h V_{h+1}^\pi](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} V_{h+1}^\pi(s')$, $[\mathbb{P}_h \bar{Z}_{h+1}^\pi](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} \bar{Z}_{h+1}^\pi(s')$, and $[\mathbb{P}_h \bar{\eta}_{h+1}^\pi](s,a) = \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} \bar{\eta}_{h+1}^\pi(s')$. [2] For brevity, we refer to $\bar{\eta}^\pi$ simply as $\bar{\eta}$.

In the episodic MDP, the agent aims to learn the optimal policy through a fixed number of interactions with the environment across a number of episodes. At the beginning of each episode $k(\in[K])$, the agent starts at the initial state s_1^k and choose a policy π^k. In step $h(\in[H])$, the agent observes $s_h^k(\in\mathcal{S})$, takes an action $a_h^k(\in\mathcal{A}) \sim \pi_h^k(\cdot|s_h^k)$, receives a reward $r_h(s_h^k,a_h^k)$, and the environment transits to the next state $s_{h+1}^k \sim \mathbb{P}_h(\cdot|s_h^k,a_h^k)$. Finally, we measure the suboptimality of an agent by its regret, which is the accumulated difference between the ground truth optimal and the return received from the interaction. The regret after K episodes is defined as $\mathrm{Reg}(K) = \sum_{k=1}^K V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)$.

[2] Note that $\mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a)} \bar{\eta}_{h+1}^\pi(s')$ is a mixture distribution.

Distributional Bellman Optimality Equation. Recall that $\eta_h^\star$ satisfies the following optimality equation:

$$\begin{aligned}\eta_h^\star(s,a) &= (\mathcal{T}_h \eta_{h+1}^\star)(s,a) \\ &\coloneqq \mathbb{E}_{s'\sim\mathbb{P}_h(\cdot|s,a),a'\sim\pi_h^\star(\cdot|s')}[(\mathcal{B}_{r_h})_\# \eta_{h+1}^\star(s',a')] \\ &= (\mathcal{B}_{r_h})_\#[\mathbb{P}_h \eta_{h+1}^\star](s,a)\end{aligned}$$

where $\mathcal{B}_r : \mathbb{R} \to \mathbb{R}$ is defined by $\mathcal{B}_r(x) = r + x$, and $g_\# \eta \in \mathscr{P}(\mathbb{R})$ is the pushforward of the distribution η through g (*i.e.,* $g_\#\eta(A) = \eta(g^{-1}(A))$ for any Borel set $A \subseteq \mathbb{R}$).

Additional Notations. For a given N, we denote an $N-$dimensional function class $\mathcal{F}^N \coloneqq \mathcal{F}^{(1)} \times \cdots \times \mathcal{F}^{(N)} \subseteq \left\{ f = [f^{(1)}, \cdots, f^{(N)}] : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N \right\}$. Given a dataset $\mathcal{D} = \{(s_t, a_t, [z_t^{(1)}, \ldots, z_t^{(N)}])\}_{t=1}^{|\mathcal{D}|} \subseteq \mathcal{S} \times \mathcal{A} \times \mathbb{R}^N$, a set of state-action pairs $\mathcal{Z} = \{(s_t, a_t)\}_{t=1}^{|\mathcal{Z}|} \subseteq \mathcal{S} \times \mathcal{A}$ and for a function $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$, we define the norm $\|f^{(n)}\|_\infty, \|f\|_{\infty,1}, \|f\|_\mathcal{D}, \|f\|_\mathcal{Z}$ as written in Appendix B.1. For a set of (vector-valued) functions $\mathcal{F}^N \subseteq \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N\}$, the width function of (s,a) is defined as $w^{(n)}(\mathcal{F}^N, s, a) \coloneqq \max_{f,g\in\mathcal{F}^N} |f^{(n)}(s,a) - g^{(n)}(s,a)|$.

4.3 Statistical Functionals in Distributional RL

In this section, we define two key concepts in the distRL framework: the *statistical functional* and the *sketch*. We also illustrate *Bellman closedness*, a crucial property from Bellemare et al. [13]. Next, we introduce *Bellman unbiasedness*, a novel concept that complements the previous property and is essential for provable efficiency. As shown in Figure 4.2, quantile functionals cannot be updated in an unbiased manner (as proved in Theorem 4.3.3), demonstrating that only certain sketches can be updated exactly. We then show that the only sketch satisfying both properties is the moment functional, which is unique

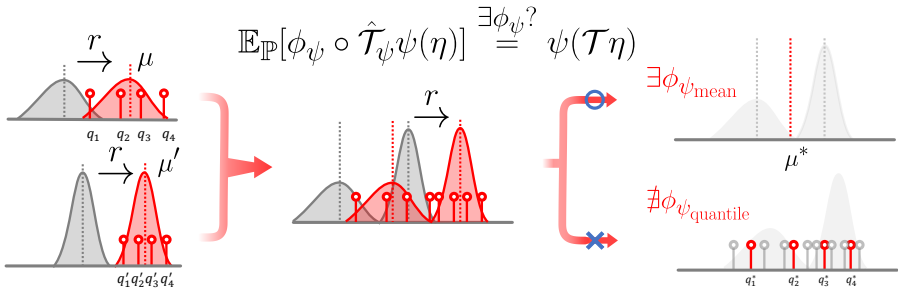


Figure 4.2: Illustrative representation of sketch-based Bellman updates for a mixture distribution. Instead of updating the distributions directly, each sampled distribution is embedded through a sketch ψ (e.g., mean μ, quantile q_i). The transformation ϕ_ψ aims to compress the mixture distribution into the same number of parameters, ensuring unbiasedness to prevent information loss.

among statistical functionals. Finally, we discuss the intractability of the previous structural assumption, distributional Bellman Completeness, and its tendency to cause linear regret. To address this, we introduce *statistical functional Bellman Completeness*, a relaxed assumption, and explain why it satisfies both properties.

4.3.1 Bellman Closedness

Definition 4.3.1 (Statistical functionals, Sketch; [13])**.** A **statistical functional** is a mapping from a probability distribution to a real value $\psi : \mathscr{P}(\mathbb{R}) \to \mathbb{R}$. A **sketch** is a vector-valued function $\psi_{1:N} : \mathscr{P}(\mathbb{R}) \to \mathbb{R}^N$ specified by an N-tuple where each component is a statistical functional,

$$\psi_{1:N}(\cdot) = (\psi_1(\cdot), \cdots, \psi_N(\cdot)).$$

We denote the domain of sketch as $\mathscr{P}_{\psi_{1:N}}(\mathbb{R})$ and its image as $I_{\psi_{1:N}} = \{\psi_{1:N}(\bar{\eta}) : \bar{\eta} \in \mathscr{P}_{\psi_{1:N}}(\mathbb{R})\}$. We further extend to state return distribution functions $\psi_{1:N}(\bar{\eta}) = \left(\psi_{1:N}(\bar{\eta}(s)) : s \in \mathcal{S}\right)$.

Definition 4.3.2 (Bellman closedness; [81])**.** A sketch $\psi_{1:N}$ is **Bellman closed** if there exists an operator $\mathcal{T}_{\psi_{1:N}} : I^{\mathcal{S}}_{\psi_{1:N}} \to I^{\mathcal{S}}_{\psi_{1:N}}$ such that

$$\psi_{1:N}(\mathcal{T}\bar{\eta}) = \mathcal{T}_{\psi_{1:N}} \psi_{1:N}(\bar{\eta}) \quad \text{for all } \bar{\eta} \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$$

which is closed under a distributional Bellman operator $\mathcal{T} : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}}$.

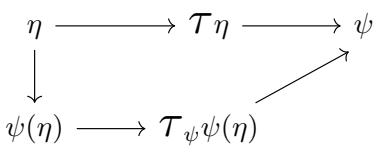


Figure 4.3: Bellman Closedness

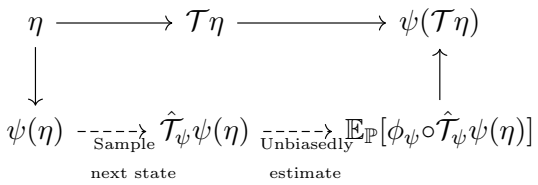


Figure 4.4: Bellman Unbiasedness

Figure 4.5: Illustration of Bellman Closedness and Bellman Unbiasedness. The above path represents an ideal distributional Bellman update. Due to the infinite-dimensionality, the update process should be represented by using a finite-dimensional embedding (sketch) ψ. Since the transition kernel $\mathbb{P}$ is unknown, the below path describes that the implementation should sample the next state and update by using $\hat{\mathcal{T}}_\psi$ with the empirical transition kernel $\hat{\mathbb{P}}$. A sketch ψ is Bellman unbiased if $\hat{\mathcal{T}}_\psi \circ \psi$ can unbiasedly estimate $\psi \circ \mathcal{T}$ through some transformation ϕ_ψ, *i.e.,* $\psi(\mathcal{T}\eta) = \mathbb{E}_{\mathbb{P}}[\phi_\psi \circ \hat{\mathcal{T}}\psi(\eta)]$.

Bellman closedness is the property that a sketch are exactly learnable when updates are performed from the infinite-dimensional distribution space to the finite-dimensional embedding space. While classical Bellman equation implies the existence of Bellman operator for expected values, not all statistical functional has such corresponding Bellman operator. Precisely, Rowland et al. [81] showed that the only finite linear statistical functionals that are Bellman closed are given by the collections of statistical functionals where its linear span is equal to the set of exponential-polynomial functionals. [3]

Theorem 4.3.3. *Quantile functional cannot be Bellman closed under any additional sketch.*

While Rowland et al. [81] focused on "linear" statistical functionals in defining a sketch (i.e., $\psi(\bar{\eta}) = \mathbb{E}_{Z\sim\bar{\eta}}[h(Z)]$ for some h), leaving questions about nonlinear functionals, we extend this by showing that "nonlinear" statistical functionals, such as maximum or minimum, can also be Bellman closed. Additionally, while

[3] In discounted setting, a unique solution becomes moments. We've overwritten it for convinience.

their proof implicitly treated quantiles as linear functionals, we provide a technical clarification in Appendix B.3.1 where we formally demonstrate that no sketch Bellman operator exists for quantiles.

4.3.2 Bellman Unbiasedness

While the intractability caused by infinite-dimensionality was addressed in Bellman closedness, another intractable element which has not yet fully tackled is the *sampling of the next state.* During the implementation, note that the agent does not have access to the transition kernel $\mathbb{P}$. Instead, the agent can only access the empirical transition kernel $\hat{\mathbb{P}}(\cdot|s,a) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\{s'_k = \cdot\ |s,a\}$ which is derived from K sampled next states. This limitation implies that the operator should be treated as an empirical operator $\hat{\mathcal{T}}_\psi$, rather than $\mathcal{T}_\psi$ (*i.e.,* $\hat{\mathcal{T}}_\psi\psi(\bar{\eta}) \coloneqq \psi((\mathcal{B}_r)_\#[\hat{\mathbb{P}}\bar{\eta}])$). Therefore, we naturally introduce a new notion of *Bellman unbiasedness* to unbiasedly estimate the expected distribution $(\mathcal{B}_r)_\#\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')]$, which is a mixture by transitions, from the sample distribution $(\mathcal{B}_r)_\#\bar{\eta}(s')$.

Definition 4.3.4 (Bellman unbiasedness)**.** A sketch $\psi(=\psi_{1:N})$ is **Bellman unbiased** if a vector-valued estimator $\phi_\psi = \phi_\psi(\psi(\cdot),\cdots,\psi(\cdot)) : (I_\psi^{\mathcal{S}})^k \to I_\psi^{\mathcal{S}}$ exists where the sketch of expected distribution $(\mathcal{B}_r)_\#\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')]$ can be unbiasedly estimated by ϕ_ψ using the k sampled sketches from the sample distribution $(\mathcal{B}_r)_\#\bar{\eta}(s')$, i.e.,

$$\mathbb{E}_{s'_i\sim\mathbb{P}}\Bigg[\phi_\psi\Bigg(\underbrace{\psi\Big((\mathcal{B}_r)_\#\bar{\eta}(s'_1)\Big),\cdots,\psi\Big((\mathcal{B}_r)_\#\bar{\eta}(s'_k)\Big)}_{k \text{ sampled sketches from sample distribution } \hat{\mathcal{T}}_\psi\psi(\bar{\eta}(s))}\Bigg)\Bigg]$$
$$= \psi\Big((\mathcal{B}_r)_\#\mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')]\Big).$$

Bellman unbiasedness is another natural definition, similar to Bellman closedness, which takes into account a finite number of samples for the transition.

For example, mean-variance sketch is Bellman unbiased as the following unbiased estimator $\phi_{(\mu,\sigma^2)}$ exists for k sample estimates:

$$\begin{aligned}(\mu, \sigma^2) &= \phi_{(\mu,\sigma^2)}\Big((\hat{\mu}_1, \hat{\sigma}_1^2), \cdots, (\hat{\mu}_k, \hat{\sigma}_k^2)\Big) \\ &= \Big(\frac{1}{k}\sum_{i=1}^{k}\hat{\mu}_i,\ \frac{1}{k}\sum_{i=1}^{k}(\hat{\mu}_i - \frac{1}{k}\sum_{i=1}^{k}\hat{\mu}_i)^2 + \hat{\sigma}_i^2\Big)\end{aligned}$$

On the other hand, median functional is not Bellman unbiased since there is no unbiased estimator for median. Then, the following question naturally arises;

"Which sketches are unbiasedly estimatable under the sketch-based Bellman update?"

The following lemma answers this question.

Lemma 4.3.5. *Let $F_{\bar{\eta}}$ be a CDF of the probability distribution $\bar{\eta} \in \mathscr{P}_\psi(\mathbb{R})^{\mathcal{S}}$. Then a sketch is Bellman unbiased if and only if the sketch is homogeneous over $\mathscr{P}_\psi(\mathbb{R})^{\mathcal{S}}$ of degree k, i.e., there exists some vector-valued function $h = h(x_1, \cdots, x_k) : \mathcal{X}^k \to \mathbb{R}^N$ such that*

$$\psi(\bar{\eta}) = \int \cdots \int h(x_1, \cdots, x_k) dF_{\bar{\eta}}(x_1) \cdots dF_{\bar{\eta}}(x_k).$$

Lemma 4.3.5 states that in statistical functional dynamic programming, the unbiasedly estimatable embedding of a distribution can only be structured in the form of functions that are *homogeneous* of finite degree [42]. To illustrate that homogenity defines a broader class than linear functionals, consider the variance as a simple example. Variance is clearly not a linear functional, as it is non-additive. However, it can be written as

$$\begin{aligned}&\mathrm{Var}(\bar{\eta}) = \mathbb{E}_{Z_1\sim\bar{\eta}}[(Z_1 - \mathbb{E}_{Z_2\sim\bar{\eta}}[Z_2])^2] \\ &= \mathbb{E}_{Z_1,Z_2\sim\bar{\eta}}[Z_1^2 - 2Z_1Z_2 + Z_2^2] = \mathbb{E}_{Z_1,Z_2\sim\bar{\eta}}[h(Z_1, Z_2)]\end{aligned}$$

which implies the homogenity of degree 2. Taking this concept further and combining it with the results on Bellman closedness, we prove that even when

including a nonlinear statistical functional, the only sketch that can be exactly learned and unbiasedly estimated in a finite-dimensional embedding space is the moment sketch.

Theorem 4.3.6. *The only finite sketches that are both Bellman unbiased and Bellman closed are given by collections of* $\psi_1, \ldots, \psi_N$ *where its linear span* $\{\sum_{n=0}^{N} \alpha_n \psi_n |\ \alpha_n \in \mathbb{R}\ , \forall N\}$ *is equal to the linear span of the set of exponential polynomial functionals* $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l \exp(\lambda Z)] |\ l = 0, 1, \ldots, L, \lambda \in \mathbb{R}\}$*, where* ψ_0 *is the constant functional equal to* 1*.*

Compared to Rowland et al. [81], we extend beyond linear statistical functionals to include nonlinear statistical functionals, showing the uniqueness of the moment functional. As shown in Figure 4.1, our theoretical results not only show that high-order central moments such as variance or skewness are exactly learnable and unbiasedly estimatable, but also reveal that other nonlinear statistical functionals like median or quantiles inevitably involve approximation errors due to biased estimations.

Necessity of Bellman unbiasedness. Bellman unbiasedness ensures that updates can be unbiasedly performed when only a finite number of sampled sketches are available. In other words, it guarantees that the sequence of sampled sketches forms a martingale, enabling the construction of confidence regions through concentration inequalities. This property is crucial for establishing provable efficiency in terms of regret minimization.

Complementary roles of unbiasedness and closedness. At first glance, Bellman Unbiasedness (BU) may appear to be a stricter subset of Bellman Closedness (BC). However, as illustrated in Figure 4.1, the relationship is more subtle: for example, the categorical sketch is BU but not BC, whereas

functionals like the maximum or minimum are BC but not BU. More precisely, BU guarantees the existence of an unbiased estimator of the ground-truth sketch given a finite number of sampled sketches. In contrast, BC plays a complementary role by ensuring that the update process consistently provide such sketches. If a sketch is BU but not BC–as in the case of the categorical sketch–then the update process cannot continue providing new sampled sketches, making dynamic programming infeasible.

4.3.3 Statistical Functional Bellman Completeness

We consider distributional reinforcement learning with general value function approximation (GVFA). For successful TD learning, GVFA framework for classical RL commonly requires the assumption, *Bellman Completeness*, that after applying Bellman operator, the output lies in the function class $\mathcal{F}$ [6, 47, 101]. As a natural extension, our approach receives a tuple of function class $\mathcal{F}^N \subseteq \{f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N\}$ as input to represent N moments of distribution. Building on this, we assume that for any $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$, the sketch of target function lies in the function class $\mathcal{F}^N$.

Assumption 4.3.7 (Statistical Functional Bellman Completeness)**.** For any distribution $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$ and $h \in [H]$, there exists $f_{\bar{\eta}} \in \mathcal{F}^N$ which satisfies

$$f_{\bar{\eta}}(s, a) = \psi_{1:N}\big((\mathcal{B}_{r_h})_{\#}[\mathbb{P}_h \bar{\eta}](s, a)\big) \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}$$

DistBC inevitably leads to linear regret. In the seminal works, Wang et al. [100] and Chen et al. [21] assumed that the function class $\mathcal{H} \subseteq \{\eta : \mathcal{S} \times \mathcal{A} \to \mathscr{P}([0, H])\}$ follows the *distributional Bellman Completeness* (distBC) assumption (*i.e.,* if $\eta \in \mathcal{H}$ for all $\pi, h \in [H]$, $\mathcal{T}_h^\pi \eta \in \mathcal{H}$). This seems natural, but constructing a finite-dimensional subspace $\mathcal{H}$ that satisfies distBC is quite challenging. Since the distributional Bellman operator is a composition of translation and mixing distributions for the next state, it implies that a function class $\mathcal{H}$ must be closed

under translation and mixture. However, when considering the representation of infinite-dimensional distributions using a finite number of representations, it is not trivial that the mixture of distributions can also be represented with the same number of representations. For example, while a Gaussian distribution can be represented using two parameters (μ, σ^2), a mixture of K Gaussians generally requires $2K$ representations.

To avoid the issue of closedness under mixture, both previous studies assumed a discretized reward MDP where all outcomes of the return distribution are able to discretized into an uniform grid of finite points. Unfortunately, the approximation error introduced by the discretization is not negligible when it comes to regret. This is because *model misspecification*, which is the error when the model fails to represent the target, typically leads to linear regret.

Definition 4.3.8 (Model Misspecification in distBC)**.** For a given distribution class $\mathcal{H}$ which is the finite-dimensional subspace of the space of all distribution $\mathcal{F}^\infty$, we call ζ the **misspecification error**

$$\zeta \coloneqq \inf_{f_{\bar{\eta}} \in \mathcal{H}} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \| f_{\bar{\eta}}(s,a) - (\mathcal{B}_{r_h})_{\#}[\mathbb{P}_h \bar{\eta}](s,a) \|$$

for any $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0,H])$ and $h \in [H]$.

Note that ζ is strictly positive unless the function approximator $f_{\bar{\eta}}$ can represent any distribution in the finite-dimensional subspace $\mathcal{H}$ generated by translation and mixture. In a classical linear bandit setting [108], a lower bound with misspecification error ζ is known to yield linear regret $\Omega(\zeta K)$. Therefore, redefining Bellman Completeness within the infinite-dimensional distribution space is not appropriate, as it either imposes strong constraints on the MDP structure or leads to linear regret. To circumvent model misspecification, we revisit the distributional BC through the statistical functional lens. We propose a novel framework that matches a finite number of statistical functionals to the target, rather than the entire distribution itself.

4.4 SF-LSVI: Statistical Functional Least Squares Value Iteration

In this section, we propose SF-LSVI for distRL framework with general value function approximation. Leveraging the result from Theorem 4.3.6, we introduce a *moment least square regression*. This allows us to capture a finite set of moment information from the distribution, which can be unbiasedly estimated, thereby leading to the *truncated moment problem* [84, 89]. Unlike previous work [21, 100] that estimates in infinite-dimensional distribution spaces, our method enables to estimate distribution unbiasedly in finite-dimensional embedding spaces without misspecification error.

Algorithm 3 Statistical Functional Least Squares Value Iteration (SF-LSVI)

Input: failure probability $\delta \in (0,1)$ and the number of episodes K

for episode $k = 1, 2, \ldots, K$ **do**
  Receive initial state s_1^k
  Initialize $\psi_{1:N}(\bar{\eta}_{H+1}^k(\cdot)) \leftarrow \mathbf{0}^N$
  for step $h = H, H-1, \ldots, 1$ **do**
    $\mathcal{D}_h^k \leftarrow \left\{ s_{h'}^\tau, a_{h'}^\tau, \psi_{1:N}\Big((\mathcal{B}_{r_{h'}^\tau})_\# \bar{\eta}_{h+1}^k(s_{h'+1}^\tau)\Big) \right\}_{(\tau,h') \in [k-1]\times[H]}$
    // Data collection
    $\tilde{f}_{h,\bar{\eta}}^k \leftarrow \arg\min_{f \in \mathcal{F}^N} \|f\|_{\mathcal{D}_h^k}$
    // Distribution Estimation
    $b_h^k(\cdot,\cdot) \leftarrow w^{(1)}((\mathcal{F}^N)_h^k, \cdot, \cdot)$
    $Q_h^k(\cdot,\cdot) \leftarrow \min\{(\tilde{f}_{h,\bar{\eta}}^k)^{(1)}(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\}$
    $\pi_h^k(\cdot) = \arg\max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$, $V_h^k(\cdot) = Q_h^k(\cdot, \pi_h^k(\cdot))$
    // Optimistic planning
    $\psi_1\Big(\eta_h^k(\cdot,\cdot)\Big) \leftarrow Q_h^k(\cdot,\cdot)$
    $\psi_{2:N}\Big(\eta_h^k(\cdot,\cdot)\Big) \leftarrow \Big(\min\{(\tilde{f}_{h,\bar{\eta}}^k)^{(n)}(\cdot,\cdot), H\}\Big)_{n \in [2:N]}$
    $\psi_1\Big(\bar{\eta}_h^k(\cdot)\Big) \leftarrow V_h^k(\cdot)$, $\psi_{2:N}\Big(\bar{\eta}_h^k(\cdot)\Big) \leftarrow \psi_{1:N}\Big(\eta_h^k(\cdot, \pi_h^k(\cdot))\Big)_{n \in [2:N]}$
  for $h = 1, 2, \ldots, H$ **do**
    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$
    Observe reward $r_h^k(s_h^k, a_h^k)$ and get next state s_{h+1}^k.

Overview. At the beginning of episode $k \in [K]$, we maintain all previous samples $\{(s^\tau_{h'}, a^\tau_{h'}, r^\tau_{h'})\}_{(\tau,h')\in[k-1]\times[H]}$ and initialize a sketch $\psi_{1:N}(\bar{\eta}^k_{H+1}(\cdot)) = \mathbf{0}^N$. For each step $h = H, \dots, 1$, we compute the normalized sample moments of target distribution $\{(\mathcal{B}_{r^\tau_{h'}})_\# \bar{\eta}^k_{h+1}(s^\tau_{h'+1})\}_{h'\in[H]}$ with the help of binomial theorem,

$$\begin{aligned}\psi_n\Big((\mathcal{B}_{r^\tau_{h'}})_\#\bar{\eta}_h(s^\tau_{h'+1})\Big) &:= \frac{\mathbb{E}[(\bar{Z}^k_{h+1}(s^\tau_{h'+1}) + r^\tau_{h'})^n]}{H^{n-1}} \\ &= \frac{\sum_{n'=0}^n H^{n'}\psi_{n'}\Big(\bar{\eta}_h(s^\tau_{h'+1})\Big)(r^\tau_{h'})^{n-n'}}{H^{n-1}}\end{aligned}$$

and iteratively solve the N-moment least squares regression

$$\tilde{f}^k_{h,\bar{\eta}} \leftarrow \arg\min_{f\in\mathcal{F}} \sum_{\tau=1}^{k-1}\sum_{h'=1}^{H}\Big(\sum_{n=1}^N f^{(n)}(s^\tau_{h'}, a^\tau_{h'}) - \psi_n\Big((\mathcal{B}_{r^\tau_{h'}})_\#\bar{\eta}^k_{h+1}(s^\tau_{h'+1})\Big)\Big)^2$$

based on the dataset $\mathcal{D}^k_h$ which contains the sketch of temporal target distribution $\psi_{1:N}\Big((\mathcal{B}_{r^\tau_{h'}})_\#\bar{\eta}^k_{h+1}(s^\tau_{h'+1})\Big)$. Then we define $Q^k_h(\cdot,\cdot) = \min\{(\tilde{f}^k_{h,\bar{\eta}})^{(1)}(\cdot,\cdot) + b^k_h(\cdot,\cdot), H\}$ and choose the greedy policy $\pi^k_h(\cdot)$ with respect to Q^k_h. Next, we update all N normalized moments of Q-distribution $\psi_{1:N}\Big(\eta^h_k(\cdot,\cdot)\Big)$ and V-distribution $\psi_{1:N}\Big(\bar{\eta}^h_k(\cdot)\Big)$. We repeat the procedure until all the K episodes are completed.

Remark 4.4.1. For an optimistic planning, we define the bonus function as the width function $b^k_h(s,a) := w^k_h((\mathcal{F}^N)^k_h, s, a)$ where $(\mathcal{F}^N)^k_h$ denotes a confidence region at step h, episode k. When $\mathcal{F}$ is a linear function class, the width function can be evaluated by simply computing the maximal distance of weight vector. For a general function class $\mathcal{F}$, computing the width function requires to solve a set-constrained optimization problem, which is known as NP-hard [33]. However, a width function is computed simply for optimistic exploration, and approximation errors are known to have a small effect on regret [1]. We leave the study of computationally efficient algorithms for the width function as a promising future work, and replace with one of the numerical approximations mentioned above.

4.5 Theoretical Analysis

In this section, we provide the theoretical guarantees for `SF-LSVI` under Assumption 4.3.7. Applying proof techniques from Wang et al. [101] and extending the

result to a statistical functional lens, we generalize *eluder dimension* [83] to the vector-valued function, which has been widely used in RL literatures [6, 49, 101] to measure the complexity of learning with the function approximators.

Definition 4.5.1 (ϵ-dependent, ϵ-independent, Eluder dimension for vector–valued function)**.** Let $\epsilon \geq 0$ and $\mathcal{Z} = \{(s_i, a_i)\}_{i=1}^n \subseteq \mathcal{S} \times \mathcal{A}$ be a sequence of state-action pairs.

- A state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ is **ϵ-dependent** on $\mathcal{Z}$ with respect to $\mathcal{F}^N$ if $\|f - g\|_{\mathcal{Z}} \leq \epsilon$ for any vector-valued function $f, g \in \mathcal{F}^N$, then $|f^{(1)}(s, a) - g^{(1)}(s, a)| \leq \epsilon$.
- An (s, a) is **ϵ-independent** on $\mathcal{Z}$ with respect to $\mathcal{F}^N$ if (s, a) is not ϵ-dependent on $\mathcal{Z}$.
- The **ϵ-eluder dimension** $\dim_E(\mathcal{F}^N, \epsilon)$ of a vector-valued function class $\mathcal{F}^N$ is the length of the longest sequence of elements in $\mathcal{S} \times \mathcal{A}$ such that, for some $\epsilon' \geq \epsilon$, every element is ϵ'-independent on its predecessors.

We assume that the function class $\mathcal{F}^N$ and state-action space $\mathcal{S} \times \mathcal{A}$ have bounded covering numbers.

Assumption 4.5.2 (Covering number)**.** For any $\epsilon > 0$, the following holds:

- there exists an ϵ-cover $\mathcal{C}(\mathcal{F}^N, \epsilon) \subseteq \mathcal{F}^N$ with size $|\mathcal{C}(\mathcal{F}^N, \epsilon)| \leq \mathcal{N}(\mathcal{F}^N, \epsilon)$, such that for any $g \in \mathcal{F}^N$, there exists $g' \in \mathcal{C}(\mathcal{F}^N, \epsilon)$ with $\|g - g'\|_{\infty,1} \leq \epsilon$.
- there exists an ϵ-cover $\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)$ with size $|\mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)| \leq \mathcal{N}(\mathcal{S} \times \mathcal{A}, \epsilon)$, such that for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, there exists $(s', a') \in \mathcal{C}(\mathcal{S} \times \mathcal{A}, \epsilon)$ with $\max_{f \in \mathcal{F}} |f(s, a) - f(s', a')| \leq \epsilon$

The following two lemmas give confidence bounds on the sum of the l_2 norms of all normalized moments.

Lemma 4.5.3 (Single Step Optimization Error)**.** *Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau \in [k-1]}$ be a state-action pairs and its dataset $\mathcal{D}_{h,\bar{\eta}}^k = \left\{\left(s_h^\tau, a_h^\tau, \psi_{1:N}\left((\mathcal{B}_{r_{h'}}^\tau)_{\#}\bar{\eta}(s_{h'+1}^\tau)\right)\right)\right\}_{\tau \in [k-1]}$ for any $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0, H])$. Define $\tilde{f}_{h,\bar{\eta}}^k = \arg\min_{f \in \mathcal{F}^N} \|f\|_{\mathcal{D}_{h,\bar{\eta}}^k}^2$. For any $\bar{\eta}$ and $\delta \in (0, 1)$, there is an event*

$\mathcal{E}(\bar{\eta}, \delta)$ such that conditioned on $\mathcal{E}(\bar{\eta}, \delta)$, with probability at least $1-\delta$, for any $\bar{\eta}' : \mathcal{S} \to \mathscr{P}([0, H])$ with $\|\psi_{1:N}(\bar{\eta}') - \psi_{1:N}(\bar{\eta})\|_{\infty,1} \leq 1/T$,we have

$$\begin{aligned}&\left\|\tilde{f}_{h,\bar{\eta}'}(\cdot,\cdot) - \psi_{1:N}\Big((\mathcal{B}_{r(\cdot,\cdot)})_{\#}[\mathbb{P}\bar{\eta}'](\cdot,\cdot)\Big)\right\|_{\mathcal{Z}_h^k} \\ &\qquad \leq c'\left(N^{\frac{1}{2}} H\sqrt{\log(1/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T)}\right)\end{aligned}$$

for some constant $c' > 0$.

Due to the definition of Bellman unbiasedness, we remark that moment sketch has a corresponding vector-valued estimator $\phi_{\psi_{1:N}}$ as an identity and leads to a concentration results as the sampled sketches forms a martingale with respect to the filtration $\mathbb{F}_h^\tau$ induced by the history of $\{(s_h^\tau, a_h^\tau)\}_{\tau\in[k-1]}$ (*i.e.*, $\mathbb{E}\Big[\psi_{1:N}\Big((\mathcal{B}_{r_h})_{\#}\bar{\eta}(s_h^\tau)\Big)\Big|\mathbb{F}_h^\tau\Big] = \psi_{1:N}\Big((\mathcal{B}_{r_h})_{\#}[\mathbb{P}_h\bar{\eta}](s_h^\tau, a_h^\tau)\Big)$).

Another notable aspect in Lemma 4.5.3 is using normalized moments $\mathbb{E}[Z^n]/H^{n-1}$ instead of moments $\mathbb{E}[Z^n]$, as it reduces the size of the confidence region from $O(H^N)$ to $O(\sqrt{N})$. This adjustment is akin to scaling the optimization function in multi-objective optimization to treat each objective equally, which effectively prevents the model from favoring objectives with larger scales.

Lemma 4.5.4 (Confidence Region)**.** *Let $(\mathcal{F}^N)_h^k = \{f \in \mathcal{F}^N | \|f - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N, \delta)\}$, where*

$$\beta(\mathcal{F}^N, \delta) \geq c' \cdot NH^2(\log(T/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T))$$

for some constant $c' > 0$. Then with probability at least $1-\delta/2$, for all $k, h \in [K] \times [H]$, we have

$$\psi_n\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h\bar{\eta}_{h+1}^k](\cdot,\cdot)\Big) \in (\mathcal{F}^N)_h^k$$

Lemma 4.5.4 guarantees that the sequence of moments from the target distribution $\psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h\bar{\eta}_{h+1}^k](\cdot,\cdot)\Big)$ lies in the confidence region $(\mathcal{F}^N)_h^k$ with high probability. Supported by the aforementioned lemma, we can further

guarantee that all Q-functions are optimistically estimated with high probability and derive our final result.

Theorem 4.5.5. *Under Assumption 4.3.7, with probability at least* $1-\delta$, `SF-LSVI` *achieves a regret bound of*

$$Reg(K) \leq 2H\, dim_E(\mathcal{F}^N, 1/T) + 4H\sqrt{KH\log(1/\delta)}$$

Under weaker structural assumptions, we show that `SF-LSVI` enjoys near-optimal regret bound of order $\tilde{O}(d_E H^{\frac{3}{2}}\sqrt{K})$, which is $\sqrt{H}$ better than the state-of-the-art distRL algorithm `V-EST-LSR` [21]. For the linear MDP setting, we have $d_E = \tilde{O}(d)$ and thus `SF-LSVI` achieves a tight regret bound as $\tilde{O}(\sqrt{d^2H^3K})$ which matches a lower bound $\Omega(\sqrt{d^2H^3K})$ [114]. In our analysis, we highlight two main technical novelties which significantly reduces the degree of regret in distRL framework;

1. We refine previous lemma of Osband et al. [74] and Wang et al. [101] to remove the dependency of $\beta(\mathcal{F}^N, 1/\delta)$ (See Appendix B.4.4), ensuring that regret bound depends only on the pre-defined function class, not on the number of moment extracted.

2. As shown in Table 5.1, we define the eluder dimension d_E in a finite-dimensional embedding space $\mathcal{F}^N$, while other methods rely on an infinite-dimensional distribution space $\mathcal{H} \subseteq \mathcal{F}^\infty$.

4.6 Summary

This chapter addresses the fundamental challenge of statistical efficiency in Distributional Reinforcement Learning when combined with general value function approximation. Prior work introduced the concept of Bellman closedness; however, it fails to guarantee that statistical functionals can be updated without bias from finite samples in online learning settings. To resolve this, we propose

the novel concept of Bellman Unbiasedness. This concept precisely characterizes which statistical functionals are both preserved under Bellman updates and unbiasedly estimable from a finite number of samples. Our analysis theoretically demonstrates that only the family of exponential-polynomial functional satisfies these two necessary properties. Based on this insight, we design the Statistical Functional Least-Squares Value Iteration (`SF-LSVI`) algorithm, which is the first theoretically efficient DistRL algorithm capable of handling general value function approximation. The `SF-LSVI` algorithm achieves a tight regret bound of $\tilde{O}(d_E H^{\frac{3}{2}}\sqrt{K})$, which represents a significant improvement over previous theoretical results.

Chapter 5

Policy-labeled Preference Learning: Is Preference Enough for RLHF?

Preference-based Reinforcement Learning (PbRL), a branch of RLHF, focuses on learning optimal policies directly from human preferences, mitigating the difficulty of crafting explicit, numerical rewards. Recent advancements, notably Direct Preference Optimization (DPO) [77], simplify this process by directly optimizing the policy based on preferences, bypassing the need for an explicit reward model. While DPO has shown strong performance in domains like LLM fine-tuning, its underlying assumptions are largely shaped by deterministic environments. However, in standard RL settings, state transitions involve environmental stochasticity, introducing inherent uncertainty that complicates policy optimization and inference.

This contrast highlights a key limitation when applying DPO to general RL problems. We found that DPO's framework implicityly assumes that the observed data was generated by the optimal policy, creating a *likelihood mismatch* that is exacerbated by environmental randomness. Furthermore, this challenge

is critical in offline RL, where the pre-collected datasets often originate from diverse, unlabelled policies, making it difficult to distinguish whether outcome quality stems from policy suboptimality or external stochasticity. This leads to the fundamental question:

Can preference data generated by diverse policies sufficiently guide sequential decision-making, or is additional information required?

To address this, we propose Policy-labeled Preference Learning (`PPL`), a novel RLHF framework that leverages regret-based preference modeling while explicitly labeling the behavior policy. `PPL` incorporates policy information directly into the learning process to disentangle the effects of environmental stochasticity and behavior policy suboptimality. We provide theoretical insights by defining a reward equivalence class and deriving a bijective mapping that allows regret to be expressed as a uniquely defined function of the optimal policy. We further introduce contrastive KL regularization for stable policy alignment. Empirically, `PPL` is evaluated on homogeneous and heterogeneous offline datasets in the MetaWorld environment, demonstrating superior performance in policy alignment compared to conventional preference-based methods.

5.1 Preliminaries

Maximum Entropy Framework. We define the MDP as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P}, r, \gamma)$ characterized by state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernels $\mathbb{P}$ which represents the probability of the next state s' given the current state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$, underlying reward $r \in [r_{\min}, r_{\max}]$, and discount factor γ. For notational simplicity, we denote the expectation over trajectories $\tau = (s_0, a_0, s_1, a_1, \cdots)$ generated by a policy π and the transition kernel $\mathbb{P}$ as $\mathbb{E}_{\tau \sim \mathbb{P}^\pi}[\cdot]$.

The MaxEnt framework provides an optimal policy which not only maximizes

the expected cumulative return, but also the entropy for each visited state:

$$\pi^*_{\text{MaxEnt}} = \arg\max_{\pi} \mathbb{E}_{\tau\sim\mathbb{P}^\pi}\Big[\sum_{t\geq 0}\gamma^t(r(s_t,a_t)+\alpha\mathcal{H}^\pi(\cdot|s_t))\Big],$$

where $\mathcal{H}^\pi(\cdot|s) = -\mathbb{E}_\pi[\log\pi(\cdot|s)]$ is the entropy of policy π at state s. Here, α is a temperature hyperparameter that determine the relative importance of entropy and reward. For clarity, we say π^*_{MaxEnt} as α-optimal. In addition, soft Q-function $Q^\pi(s,a)$ is defined as the expected cumulative return augmented by an entropy terms, expressed as;

$$Q^\pi(s,a) = r(s,a) + \mathbb{E}_{\tau\sim\mathbb{P}^\pi}\Big[\sum_{t>0}\gamma^t(r(s_t,a_t)+\alpha\mathcal{H}^\pi(\cdot|s_t))\Big].$$

Analogously, we can derive soft value function $V^\pi(s)$ and soft Bellman equation as follows:

$$V^\pi(s) = \mathbb{E}_{a\sim\pi}[Q^\pi(s,a) - \alpha\log\pi(a|s)],$$
$$Q^\pi(s,a) = r(s,a) + \gamma\mathbb{E}_{s'\sim\mathbb{P}}\Big[V^\pi(s')\Big]$$

for all state-action pairs $(s,a)\in\mathcal{S}\times\mathcal{A}$. Note that the interpretation of the value function is modified by involving the entropy term in the MaxEntRL, *i.e.,* $V^\pi(s)\neq\mathbb{E}_\pi[Q^\pi(s,a)]$. For an α-optimal policy π^*, Ziebart [117] derived the relationship between the optimal policy and optimal soft Q-function Q^{π^*}:

$$\pi^*(a|s) = \exp\Big(\alpha^{-1}(Q^{\pi^*}(s,a) - V^{\pi^*}(s))\Big),$$
$$V^{\pi^*}(s) = \alpha\log\int_{a\in\mathcal{A}}\exp\Big(\alpha^{-1}Q^{\pi^*}(s,a)da\Big).$$

5.1.1 Preference-based Reinforcement Learning

Designing a reward function that accurately aligns with human behaviors is inherently challenging. To address this, PbRL focuses on learning the optimal policy directly from human preferences rather than relying on predefined rewards.

Table 5.1: Comparison for different preference models under PbRL framework.

Algorithm	Score Function	Meaning
PEBBLE [58]	$r_\psi(s_t, a_t)$	reward
DPO [77]	$\log \pi_\psi(y\|s)/\pi_{\text{ref}}(y\|s)$	relative likelihood
DPPO [4]	$-\mathbb{E}_{a\sim\pi_\psi(\cdot\|s_t)}[\|\|a - a_t\|\|_2]$	action distance
CPL [44]	$Q^{\pi_\psi}(s_t, a_t) - V^{\pi_\psi}(s_t)$	optimal advantage
PPL [Ours]	$-(V^{\pi_\psi}(s_t) - Q^{\pi}(s_t, a_t))$	regret

In this context, we adopt a reward-free MDP $\mathcal{M}\backslash r$ within the MaxEnt framework. We define a segment $\zeta = (s_0, a_0, \ldots, s_k, a_k)$ as a sequence sampled from a dataset $\mathcal{D}$. Specifically, human annotators or AI systems are tasked with comparing pairs of trajectory segments (ζ^+, ζ^-), where ζ^+ is preferred over ζ^- (*i.e.*, $\zeta^+ \succ \zeta^-$).

Score-based Preference Model. Score-based preference model is a natural generalization of RLHF for modeling human preferences through score functions, instead of partial sum of rewards [58]. This approach extends the Bradley-Terry model [16], where pairwise comparisons are used to infer relative preferences, by introducing a score function that evaluates all observed state-action pairs within a segment. The preference model then assigns probabilities proportional to the sum of these scores, aligning with the Bradley-Terry framework. To implement the preference model using a neural network, the score function is parametrized as S_ψ, and the model is trained by minimizing the cross-entropy loss between its predictions and the preference labels derived from the dataset $\mathcal{D}$, as follows:

$$P_{S_\psi}[\zeta^+ \succ \zeta^-] = \sigma\Big(\sum_{t\geq 0} S_\psi(s_t^+, a_t^+) - S_\psi(s_t^-, a_t^-)\Big),$$

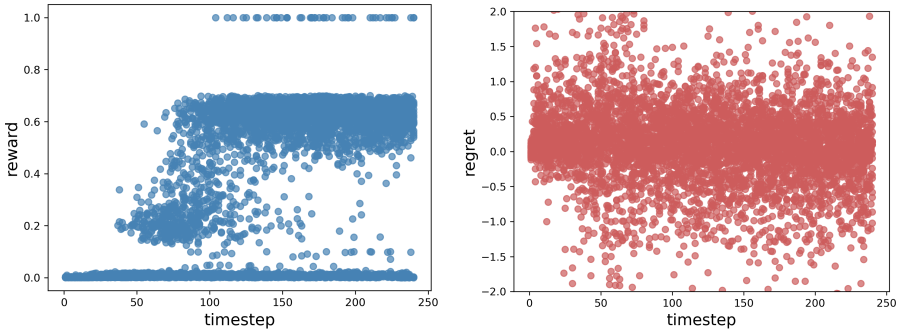


Figure 5.1: Visualization of 5000 samples in `Bin-Picking-v2` environment. While the ground-truth reward **(left)** is sparse and mainly provided upon task completion, regret **(right)** is more evenly distributed across all timesteps, making it a more informative score function for partial trajectory evaluation.

$$\mathcal{L}(S_\psi; \mathcal{D}) = -\mathbb{E}_{(\zeta^+,\zeta^-)\sim\mathcal{D}}\Big[\log P_{S_\psi}[\zeta^+ \succ \zeta^-]\Big],$$

where $\sigma(x) = 1/(1+e^{-x})$ and each (s_t^+, a_t^+) and (s_t^-, a_t^-) is the t-th state and action of preferred segment ζ^+ and less preferred segment ζ^-, respectively. For notational simplicity, we abbreviate $\mathbb{E}_{(\zeta^+,\zeta^-)\sim\mathcal{D}}$ as $\mathbb{E}_{\mathcal{D}}$.

Although it is unclear how humans evaluate their preferences, preference models can be improved to better align with human judgment by refining them based on intuitive examples. If the score function does not align with human preference evaluation, the model may produce counterintuitive outcomes. For example, Knox et al. [55] demonstrated that using the partial sum of rewards as a score function overlooks a critical issue in sparse reward MDPs: all segments that fail to reach the goal are treated as equally preferable, regardless of their contributions.

As shown in Figure 5.1, sparse reward MDPs provide little feedback for the states that do not reach the terminal goal, leading to meaningless comparison of the preferences in the early- and mid-stage segments based solely on return sums. In contrast, regret is more evenly distributed across timesteps, making

it a more effective score for comparing segment preferences regardless of their position in the trajectory. This highlights the importance of modeling preference with the score function that aligns with human intuition. Various approaches to designing such score functions have been proposed, as summarized in Table 5.1.

Optimal Advantage-based Preference Model. Hejna et al. [44] proposed *Contrastive Preference Learning* (`CPL`), which is based on an optimal advantage-based preference model [55], treating as a regret-based preference model. The `CPL` score is defined as the difference between the value of the action taken and the average value under the optimal policy, (*i.e.,* $A_{\pi^*}(s_t, a_t) \coloneqq Q^{\pi^*}(s_t, a_t) - V^{\pi^*}(s_t) = \alpha \log \pi^*(a_t|s_t)$.) Leveraging the relationship between the optimal advantage and the optimal policy within the MaxEnt framework, their objective can be reformulated into a policy-based expression, enabling the optimal policy to be learned directly without relying on reward:

$$\mathcal{L}_{\mathrm{CPL}(\lambda)}(\pi_\psi; \mathcal{D}) = -\alpha \mathbb{E}_{\mathcal{D}} \left[\log \sigma \Big(\sum_{t \geq 0} \log \pi_\psi(a_t^+|s_t^+) - \lambda \log \pi_\psi(a_t^-|s_t^-) \Big) \right]. \quad (5.1)$$

However, the standard score-based preference loss is convex but not strictly convex, leading to the existence of multiple optimal solutions. Hejna et al. [44] identified that the shift-invariance property of the loss function (*i.e.,* $P_{S(\pi_\psi)+C} = P_{S(\pi_\psi)}$) causes out-of-distribution actions to be overly weighted, deteriorating learning performance. To mitigate this issue, they introduced an asymmetric regularizer λ, which reduces the gradient weight on less preferred actions, breaking the inherent symmetry and stabilizing the learning process.

5.2 Policy-labeled Preference Learning

This section introduces the *regret-based preference model* and its distinctions from prior work, with a focus on the issue of *likelihood mismatch*, where sampled

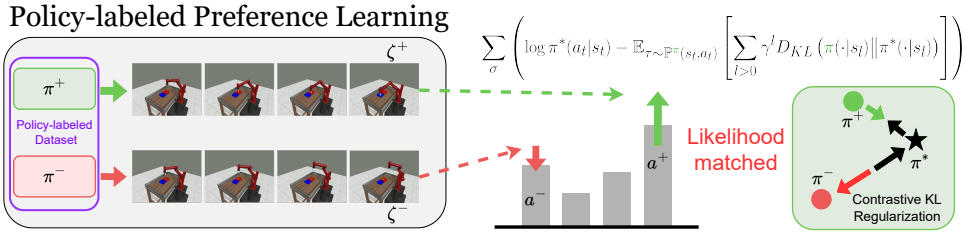


Figure 5.2: Unlike existing DPO algorithms, PPL aligns segment likelihoods by incorporating behavior policies. It reweights gradients based on closeness to the optimal policy, forming a contrastive learning framework.

segments are misinterpreted as optimal, leading to suboptimal learning. To address this, we propose **Policy-labeled Preference Learning (PPL)**, which employs a regret-based model to accurately estimate segment likelihoods. Finally, we present theoretical results derived from the PPL framework.

5.2.1 Is Preference Enough for RLHF?

Negative Regret vs Optimal Advantage. In prior work, Hejna et al. [44] utilized the *optimal advantage* as the score in CPL to introduce a *regret*-based preference model. While they presented these two concepts as equivalent, they differ significantly in their precise definitions and implications. Optimal advantage refers to the relative benefit of taking a specific action a under the optimal policy π^* $(i.e., Q^{\pi^*}(s,a) - V^{\pi^*}(s))$. In contrast, negative regret captures the performance difference between the behavior policy π and the optimal policy π^* $(i.e., Q^{\pi}(s,a) - V^{\pi^*}(s))$. The key difference between these concepts lies in whether the *behavior policy* is incorporated into the score.

From a perspective of regret, the optimal advantage disregards the source of the trajectories and evaluates the actions taken solely based on Q^{π^*}. Consequently, it implicitly treats all trajectories as if they were generated by the optimal policy. This raises an important question: what impact does this as-

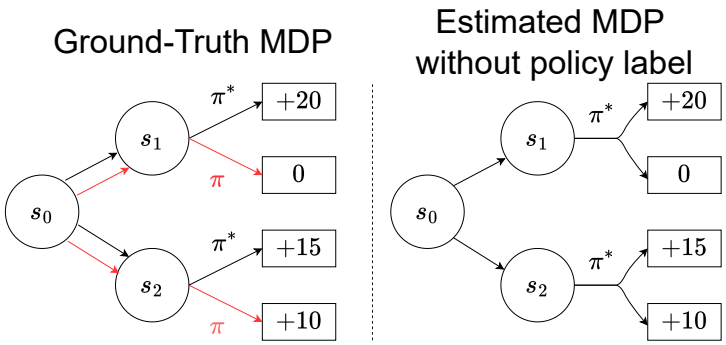


Figure 5.3: Illustration of the likelihood mismatch problem. Although the behavior policy π differs from the optimal policy π^*, the learning process incorrectly assumes all data is generated by π^*. As a result, while π^* prefers s_1, this misinterpretation leads to the incorrect conclusion that s_2 is preferred, causing suboptimal learning outcomes.

sumption – *treating all behavior polices as optimal* – have on the regret-based learning process?

Likelihood Mismatch. Likelihood mismatch occurs when outcome differences between two segments, which actually stem from behavior policy differences, are mistakenly attributed to environmental stochasticity. This misinterpretation leads to incorrect likelihood assignments. Figure 5.3 illustrates this issue in an offline setting where offline data from both a suboptimal policy π and an optimal policy π^* lacks explicit policy labels. In this scenario, all data is mistakenly assumed to be generated by the optimal policy π^*, leading to misinterpretations during learning.

To understand how preference labels are assigned in this setting, let us first consider the left-side figure. The red trajectory, generated by the suboptimal policy π, assigns a higher score (+10) to s_2, making it appear more preferable than s_1. In contrast, the black trajectory, generated by the optimal policy π^*,

assigns a higher score (+20) to s_1, leading to the opposite preference. These conflicting results can be properly distinguished when policy labels are available, allowing the model to infer the suboptimality of π by evaluating preferences separately for each policy.

Now consider the right-side Figure 5.3, where the same data is used but without policy labels. Since all data is incorrectly assumed to originate from π^*, the model observes contradictory outcomes—s_2 being preferred in one case and s_1 in another—despite assuming a single policy. Lacking policy labels, the model misinterprets this discrepancy as environmental stochasticity rather than differences in policies, distorting the learned MDP and leading to incorrect likelihood estimates for trajectories. To mitigate this issue, it is crucial to explicitly track and incorporate the behavior policy π for each segment, ensuring accurate interpretation and proper differentiation of feedback. Thus, replacing optimal advantage with regret, which reflects the suboptimality of the behavior policy, provides a principled solution.

Regret-based Model Requires the Behavior Policy. In essence, regret quantifies how much better we could have done if we had followed the optimal policy instead of the behavior policy. A larger regret indicates that the behavior policy is significantly less efficient compared to the optimal policy. We remark that the regret is the difference between the *expected return under optimal policy* and the *achieved return under behavior policy*. Based on the conventional definition of *regret*, we reformulate negative regret in a policy-based form within

the MaxEnt framework:

$$
\begin{aligned}
& -\operatorname{Reg}^{\pi}_{\pi^*}(s_t, a_t) \\
& := - \underbrace{V^{\pi^*}(s_t)}_{\text{expected return under } \pi^*} + \underbrace{Q^{\pi}(s_t, a_t)}_{\text{achieved return under } \pi} && (5.2) \\
& \overset{\text{(Thm 5.2.4)}}{=} \alpha\Big(\underbrace{\log \pi^*(a_t|s_t)}_{\text{increase likelihood}} - \underbrace{\bar{D}_{\mathrm{KL}}\big(\pi||\pi^*; s_t, a_t\big)}_{\text{decrease sequential forward KL}} \Big). && (5.3)
\end{aligned}
$$

In summary, the regret for the preferred segment can be decomposed into two components: First, it increases the likelihood of actions taken in preferred segments, aligning the behavior policy with human preferences. Second, it reduces the sequential forward KL divergence, correcting for likelihood mismatch by considering long-term differences between the behavior policy and the optimal policy. Analogously, for the less preferred segment, the regret exhibits the opposite tendencies. Based on Equation (5.3), our objective can be formulated as follows:

$$
\mathcal{L}_{\mathrm{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\mathcal{D}}\left[\log \sigma\Big(- \sum_{t \geq 0} \operatorname{Reg}^{\pi^+}_{\pi_\psi}(s_t^+, a_t^+) - \operatorname{Reg}^{\pi^-}_{\pi_\psi}(s_t^-, a_t^-) \Big) \right]
$$

where the policy label for the preferred and less preferred segments are denoted as π^+ and π^-, respectively. The detailed derivation of this formulation will be introduced in the next section and Appendix C.2.1.

5.2.2 Theoretical Analysis

Consider a triplet $(\pi^*, (\zeta^+, \pi^+), (\zeta^-, \pi^-))$, where the segments ζ^+ and ζ^- are generated by policies π^+ and π^-, respectively. The (unknown) optimal policy π^* serves as the basis for determining the underlying reward and ensuring consistent preferences. During the learning process, we assume that each segment is labeled by its behavior policy. Under this setup, the policy-labeled preference model is expressed as:

$$P_{\pi^*}^{(\pi^+,\pi^-)} = \sigma\Big(\sum_{t\geq 0} \underbrace{\Big[V^{\pi^*}(s_t^-) - V^{\pi^*}(s_t^+)\Big]}_{A_t(\pi^*)} + \underbrace{\Big[Q^{\pi^+}(s_t^+, a_t^+) - Q^{\pi^-}(s_t^-, a_t^-)\Big]}_{B_t(\pi^*,\pi^+,\pi^-)}\Big).$$

This expression is decomposed into two components: (i) $A_t(\pi^*)$, which depends solely on Q^{π^*} (note that $V^{\pi^*}(s) = \mathbb{E}_{a\sim\pi^*}[Q^{\pi^*}(s,a)] + \alpha\mathcal{H}^{\pi^*}(\cdot|s)$), and (ii) $B_t(\pi^*, \pi^+, \pi^-)$, which involves Q^{π^+} and Q^{π^-}.

The main theoretical challenge in performing a direct policy update is expressing the soft optimal Q-function and soft Q-function of a given policy π in closed-form with respect to the optimal policy π^*. Before proceeding, we introduce the concept of *equivalence classes* within the MaxEnt framework to analyze the reward structures that make a given policy optimal.

Definition 5.2.1. The set of reward functions where π^* is α-optimal is defined as (α, π^*)-*equivalence class of reward function*, denoted by $\mathcal{R}_{\alpha,\pi^*}$. For every policy π, the set of Q^π-function generated by any reward function $r_{\alpha,\pi^*} \in \mathcal{R}_{\alpha,\pi^*}$ is defined as the (α, π^*)-*equivalence class of* Q^π*-function*, denoted by $\mathcal{Q}^\pi_{\alpha,\pi^*}$.

Definition 5.2.1 indicates that a reward function class $\mathcal{R}$ or a Q^π-function class $\mathcal{Q}^\pi$ can be partitioned based on the α-optimal policy π^*. For notational simplicity, we denote the ground truth reward function corresponding to the α-optimal policy π^* as r_* and the Q^π-function induced by r_* as Q^π_*, simplifying the subscript to $*$.

Lemma 5.2.2 (Structural Condition for α-optimality)**.** *A reward function and a soft optimal Q-function where $\pi^*(\cdot|s)$ is α-optimal have a one-to-one correspondence with a state-dependent function $\beta : \mathcal{S} \to \mathbb{R}$, defined as follows:*

$$\mathcal{R}_{\alpha,\pi^*} = \{r_*(s,a) = \alpha \log \pi^*(a|s) + \beta(s) - \gamma \mathbb{E}_{\mathbb{P}}[\beta(s')]\}$$
$$\mathcal{Q}^{\pi^*}_{\alpha,\pi^*} = \{Q^{\pi^*}_*(s,a) = \alpha \log \pi^*(a|s) + \beta(s)\}$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Lemma 5.2.2 demonstrates that the (α, π^*)-equivalence class of soft optimal Q-functions can be uniquely expressed as the sum of a log-probability term,

$\alpha \log \pi^*(a|s)$, and a state-dependent function, $\beta(s)$. This result improves upon the prior lemma of Rafailov et al. [77], which established only a *surjection* from reward functions to optimal policies. By contrast, we ensure a *bijection*, rigorously defining the equivalence class of reward functions. Furthermore, Lemma 5.2.2 refines the concept of *policy invariance* introduced by Gleave et al. [40], Ng et al. [69] by specifying that the action-dependent term must be $\alpha \log \pi^*(a|s)$ to guarantee π^* is the α-optimal policy.

Lemma 5.2.3 (Unique Fixed Point of Soft Bellman π-operator)**.** *Let π^* be α-optimal. For a given policy π and Q-function $Q_A^\pi \in \mathcal{Q}^\pi$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, define the Bellman π-operator $\mathcal{T}_*^\pi : \mathcal{Q}^\pi \to \mathcal{Q}^\pi$ where*

$$\mathcal{T}_*^\pi Q_A^\pi(s,a) \coloneqq Q_*^{\pi^*}(s,a) - \gamma \mathbb{E}_{\mathbb{P}}\Big[\alpha\Big(\mathcal{H}^{\pi^*}(\cdot|s') - \mathcal{H}^\pi(\cdot|s')\Big) + \mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a')] - \mathbb{E}_\pi[Q_A^\pi(s',a')]\Big].$$

Then, $\mathcal{T}_*^\pi$ has a unique fixed point Q_*^π.

Lemma 5.2.3 describes an operator that links the soft Q-function of a given policy π to the optimal soft Q-function $Q_*^{\pi^*}$, identifying Q_*^π as its unique fixed point. Notably, this relationship is established without requiring explicit knowledge of the reward function r_*. From the novel design of the soft Bellman π-operator, we now derive the following important theorem.

Theorem 5.2.4 (Policy Deviation Theorem)**.** *If a policy π^* is α-optimal, then for any policy π,*

$$Q_*^{\pi^*}(s,a) - Q_*^\pi(s,a) = \alpha \bar{D}_{KL}(\pi||\pi^*; s, a)$$

*where the **sequential forward KL divergence** is defined as*

$$\bar{D}_{KL}(\pi||\pi'; s, a) \coloneqq \mathbb{E}_{\tau \sim \mathbb{P}_{s,a}^\pi}\left[\sum_{l>0} \gamma^l D_{KL}(\pi(\cdot|s_l)||\pi'(\cdot|s_l))\right].$$

Here, $\mathbb{P}_{s,a}^\pi$ is the distribution of trajectories $\tau = (s_0, a_0, \cdots, s_l, a_l, \cdots)$ generated by policy π and the transition $\mathbb{P}$, starting at $(s_0, a_0) = (s, a)$.

Theorem 3.4 establishes that the difference between the soft Q-function of any policy π and the optimal soft Q-function is constant and can be expressed as the sequential forward KL divergence. Intuitively, $\bar{D}_{KL}(\pi||\pi^*; s, a)$ represents the discounted sum of the forward KL divergence between π and π^* over the states visited during a rollout starting from (s, a). This property is particularly valuable, as it quantifies the performance gap using only π and π^*.

While related results were proposed by Shaikh et al. [87] and Zeng et al. [110], their proofs were restricted to contextual bandits and token-level MDPs with deterministic transitions, respectively. Moreover, their formulation depends on a KL-regularized objective that explicitly incorporates a reference policy. In contrast, Theorem 3.4, formulated within the MaxEnt framework, does not require a reference policy to be well-defined, making it more broadly applicable.

Corollary 5.2.5. *For a given (α, π^*) and a policy π, $Reg^{\pi}_{\pi^*}(\cdot, \cdot)$ is uniquely determined regardless of $\beta(s)$.*

Since regret is invariant to transformations of $\beta(s)$, it does not require additional variance reduction techniques [85] to ensure stable learning. As noted in Lemma 5.2.2, any policy-invariant transformation can be expressed as a combination of a state-dependent function $\beta(s)$ and a scaled log-likelihood term $\alpha \log \pi(a|s)$, where α represents the temperature parameter in the MaxEnt framework. Specifically, for any transformation of the reward function that preserves the optimal policy, we can rewrite the modified reward as:

$$r(s, a) = \alpha \log \pi(a|s) + \beta(s).$$

This formulation extends the classical reward shaping result of Ng et al. [69] by explicitly incorporating the policy-dependent term $\alpha \log \pi(a|s)$, which accounts for transformations in the likelihood space. This insight allows us to

generalize policy-invariant transformations and directly integrate them into preference-based learning objectives.

Using this representation, we can reformulate the sequential DPO objective with a policy-invariant transformation as follows:

$$\begin{aligned}
&\mathcal{L}_{\mathrm{DPO}(\beta)}(\pi_\psi; \mathcal{D}) \\
&= -\mathbb{E}_{\mathcal{D}}\Big[\log\sigma\Big(\sum_{t\geq 0}\Big\{\log\frac{\pi_\psi(a_t^+|s_t^+)}{\pi_{\mathrm{ref}}(a_t^+|s_t^+)} + \beta(s_t^+) - \gamma\mathbb{E}_{s_t'\sim\mathbb{P}(\cdot|s_t^+,a_t^+)}[\beta(s_t')]\Big\} \\
&\qquad - \Big\{\log\frac{\pi_\psi(a_t^-|s_t^-)}{\pi_{\mathrm{ref}}(a_t^-|s_t^-)} + \beta(s_t^+) - \gamma\mathbb{E}_{s_t'\sim\mathbb{P}(\cdot|s_t^-,a_t^-)}[\beta(s_t')]\Big\}\Big)\Big]. \qquad (5.4)
\end{aligned}$$

The existence of multiple objectives that preserve the optimal policy through reward shaping has been explored in previous work, particularly in the *variance reduction* schemes of policy gradient methods. Schulman et al. [85] introduced the *generalized advantage estimate* (GAE) as a method to reduce the variance of policy gradient estimates, effectively selecting an appropriate $\beta(s)$ to improve stability and efficiency in learning. Similarly, in Equation 5.4, the standard DPO framework assumes $\beta(s) = 0$, but optimizing $\beta(s)$ to minimize the variance of gradient estimates could lead to more stable training.

In contrast, as shown in Equation C.2.1 on Appendix C.2.1, regret-based formulations naturally eliminate $\beta(s)$ by definition, avoiding the challenges associated with policy-invariant transformations. This property ensures that regret serves as a unique and well-defined objective function, making it inherently robust without requiring explicit variance reduction techniques.

Corollary 5.2.6. *Maximizing the MaxEnt objective with negative regret as the reward is equivalent to minimizing the sequential forward KL divergence between the learned policy and the behavior policy for each preferred state-action pair in the dataset, i.e.,*

$$\begin{aligned}
&\arg\max_{\pi_\psi}\Big(\mathbb{E}_{\zeta^+\sim\mathcal{D}}[-Reg_{\pi_\psi}^{\pi^+}(s^+,a^+) - \alpha\log\pi_\psi(a^+|s^+)]\Big) \\
&\quad\equiv \arg\min_{\pi_\psi}\Big(\mathbb{E}_{\zeta^+\sim\mathcal{D}}[\bar{D}_{KL}(\pi^+||\pi_\psi; s^+, a^+)]\Big). \qquad (5.5)
\end{aligned}$$

Corollary 5.2.6 implies that regret-based RLHF operates by aggregating behavior policies from preferred segments, aligning the learned policy toward preferred actions. Notably, if all preferred segments are assumed to be generated by the optimal policy, the formulation reduces to the standard `CPL` objective, highlighting its connection to prior methods.

Proof. Consider a dataset $\mathcal{D}$ and a set of sampled preferred segments $\{\zeta_i^+\}_{i=1}^N$ which are generated by behavior policy π_i^+ respectively. To avoid notation ambiguity, we emphasize that the subscript i in this proof denotes the index of each individual samples. When defining the reward function as the negative regret, the optimal policy of Maxent objective π^*_{Reg} can be reformulated as:

$$
\begin{aligned}
&\pi^*_{Reg} \\
&:= \arg\max_{\pi_\psi} \Big(\frac{1}{N} \sum_{i=1}^{N} \Big[-\text{Reg}^{\pi_i^+}_{\pi_\psi}(s_i^+, a_i^+) - \alpha \log \pi_\psi(a_i^+|s_i^+) \Big] \Big) \\
&= \arg\max_{\pi_\psi} \Big(\frac{1}{N} \sum_{i=1}^{N} \Big[Q^{\pi_i^+}_{\pi_\psi}(s_i^+, a_i^+) - V^{\pi_\psi}_{\pi_\psi}(s_i^+) - \alpha \log \pi_\psi(a_i^+|s_i^+) \Big] \Big) \\
&= \arg\max_{\pi_\psi} \Big(\frac{1}{N} \sum_{i=1}^{N} \Big[\alpha \log \pi_\psi(a_i^+|s_i^+) - \alpha \bar{D}_{KL}(\pi||\pi_\psi; s_i^+, a_i^+) - \alpha \log \pi_\psi(a_i^+|s_i^+) \Big] \Big) \\
&= \arg\min_{\pi_\psi} \Big(\frac{1}{N} \sum_{i=1}^{N} \bar{D}_{KL}(\pi_i^+||\pi_\psi; s_i^+, a_i^+) \Big)
\end{aligned}
$$

■

Notably, the minimum is achieved if and only if $\pi_\psi(a_i^+|s_i^+) = \pi(a_i^+|s_i^+)$ for each $i \in [N]$. This formulation demonstrates that maximizing the MaxEnt objective with a regret-based reward is fundamentally equivalent to minimizing the sequential forward KL divergence for each segment.

Discussion. The regret-based DPO framework can be reinterpreted as a process that aggregates the behavior policies underlying the given dataset, aligning the learned policy to preferred actions by reducing the sequential

forward KL divergence. If, as assumed in CPL, the behavior policies of all preferred segments in dataset $\mathcal{D}$ correspond to the optimal policy π^* (or can be constructed as such), then PPL is guaranteed to converge to the optimal policy.

However, in practical RLHF settings, such an assumption rarely holds. Unlike standard reinforcement learning, where an agent maximizes a predefined reward function, RLHF optimizes for policy alignment rather than absolute optimality. In the DPO framework, the reward function is implicitly constructed to make the aligned policy the optimal one within the given preference dataset. As a result, the optimal policy under the learned reward function is already the policy obtained through alignment, making it unnecessary to perform an additional RL algorithm to reach the optimal policy.

To achieve further improvements, it is crucial to expand the dataset by rolling out new policies and incorporating additional preference data. This process enhances dataset coverage while enabling the learned reward function to extrapolate more effectively. Without such iterative expansion, RLHF remains constrained by the limitations of the static dataset, preventing meaningful policy improvements beyond the scope of the initially collected preferences.

5.2.3 Practical Algorithm and Implementation Details

In this section, we present PPL, a practical algorithm that leverages the policy label to solve the likelihood mismatch. Our setting follows the classical DPO, but with the difference that we manage preference queries by labeling the behavior policy for each trajectory in the dataset.

Pseudo-labels. In general RL settings, the behavior policy that generated a trajectory is typically known or accessible, making policy labeling relatively inexpensive. However, in offline datasets, the behavior policy is often unknown.

Algorithm 4 Policy-labeled Preference Learning (PPL)

Input: number of queries N, trajectory dataset $\mathcal{E}$, minibatch size D

Initialize policy parameters ψ
for $n = 1, \cdots, N$ **do**
  Sample $\zeta, \zeta' \sim \mathcal{E}$
  if policy label $\pi(a_t|s_t), \pi(a'_t|s'_t)$ unknown **then**
    $\pi(\cdot|s_t) \leftarrow \delta_{a_t}$ $\pi(\cdot|s'_t) \leftarrow \delta_{a'_t}$
  end if
  Label the behavior policy $p = (\pi, \pi')$
  Instruct the preference label $y = (y(0), y(1))$
  Store preference $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\zeta, \zeta', y, p)\}$
end for
// Create Policy-labeled Preference Queries
for $t = 1$ **to** T **do**
  Sample minibatch $\{(\zeta, \zeta', y, p)_d\}_{d=1}^{D} \sim \mathcal{D}$
  $\psi \leftarrow \arg\min_\psi \mathcal{L}_{\text{PPL}}(\pi_\psi; \mathcal{D})$
end for
// Policy Learning

To address this, we assign *pseudo-labels* as an alternative, assuming each segment was generated by a deterministic policy that executed the observed actions.

Contrastive KL Regularization. As previously discussed, the regret is decomposed into two components. In particular, the sequential KL divergence plays a pivotal role in aligning the learned policy with the preferred policy while diverging from the less preferred policy:

$$
\begin{aligned}
&- \sum_{t \geq 0} \text{Reg}^{\pi^+}_{\pi_\psi}(s_t^+, a_t^+) - \text{Reg}^{\pi^-}_{\pi_\psi}(s_t^-, a_t^-) \\
&= \alpha \sum_{t \geq 0} \Big(\log \frac{\pi_\psi(a_t^+|s_t^+)}{\pi_\psi(a_t^-|s_t^-)} \underbrace{-\bar{D}_{KL}(\pi^+||\pi_\psi; s_t^+, a_t^+) + \bar{D}_{KL}(\pi^-||\pi_\psi; s_t^-, a_t^-)\Big]}_{\text{contrastive KL regularization } \mathcal{R}(\pi_\psi;\ \pi^+, \pi^-)} \Big).
\end{aligned}
$$

We call this term as *contrastive KL regularization*, which requires performing rollouts for each (s_t, a_t) with respect to π^+ or π^-. This regularization term

Table 5.2: Success rates of all methods across six tasks on the MetaWorld benchmark on different datasets. Each score is reported with the maximum average performance across four seeds over 200 episode evaluation window.

		Bin Picking	Button Press	Door Open	Drawer Open	Plate Slide	Sweep Into
Homogeneous Dense	SFT	39.7 ± 19.2	71.5 ± 3.3	**48.0 ± 15.6**	56.2 ± 1.8	64.8 ± 0.8	70.0 ± 6.5
	P-IQL	62.0 ± 4.4	72.3 ± 1.0	47.7 ± 5.1	58.0 ± 5.7	**70.5 ± 6.1**	65.8 ± 1.3
	CPL	22.7 ± 5.5	64.3 ± 1.4	29.0 ± 4.3	54.0 ± 4.3	65.5 ± 3.1	69.8 ± 3.3
	PPL	**83.5 ± 4.4**	**79.8 ± 4.8**	39.3 ± 2.0	**69.2 ± 5.5**	64.7 ± 2.0	**72.8 ± 4.8**
Homogenous Sparse	SFT	33.5 ± 5.4	67.4 ± 1.5	31.3 ± 2.1	54.9 ± 2.7	67.1 ± 3.7	78.3 ± 2.5
	P-IQL	72.4 ± 6.6	74.5 ± 0.0	**58.5 ± 1.4**	51.4 ± 4.6	**76.3 ± 1.6**	**79.0 ± 2.6**
	CPL	26.5 ± 1.0	63.7 ± 1.3	28.5 ± 5.8	50.1 ± 4.5	65.1 ± 2.8	72.9 ± 6.1
	PPL	**87.2 ± 3.5**	**87.3 ± 2.8**	49.3 ± 6.5	**68.5 ± 5.3**	64.0 ± 6.4	73.9 ± 3.5
Heterogeneous Dense	SFT	18.5 ± 23.8	63.7 ± 12.2	26.0 ± 12.5	32.0 ± 5.7	62.8 ± 1.6	53.0 ± 9.1
	P-IQL	51.2 ± 5.3	62.5 ± 4.9	**32.0 ± 3.5**	41.8 ± 3.8	67.0 ± 3.0	**59.3 ± 3.7**
	CPL	1.2 ± 0.8	49.7 ± 3.0	17.3 ± 2.5	26.0 ± 2.2	59.2 ± 7.7	51.2 ± 3.0
	PPL	**59.7 ± 18.6**	**73.8 ± 3.3**	25.8 ± 2.0	**58.5 ± 3.8**	**69.8 ± 2.3**	57.3 ± 8.6
Heterogeneous Sparse	SFT	12.2 ± 1.0	63.7 ± 4.7	17.8 ± 0.8	38.7 ± 3.0	70.7 ± 3.8	60.7 ± 2.5
	P-IQL	48.0 ± 5.6	71.0 ± 6.6	**44.1 ± 3.2**	47.5 ± 3.0	**72.0 ± 4.0**	**64.3 ± 1.0**
	CPL	18.0 ± 6.1	50.8 ± 0.8	18.5 ± 3.0	32.1 ± 1.6	67.3 ± 5.5	55.5 ± 3.3
	PPL	**83.8 ± 3.8**	**83.5 ± 1.8**	34.3 ± 7.6	**60.8 ± 7.3**	71.2 ± 1.9	63.3 ± 4.2

ensures that the learned policy π_ψ aligns more closely with the preferred policy π^+ while pushing away from the less preferred policy π^-.

In practice, implementing contrastive KL regularization can result in a computational overhead, as it requires multiple rollouts with each state-action pair as the initial point at every timestep until the terminal is reached. This approach can also increase memory usage as it requires additional timesteps outside of the sampled segment. To address these technical challenges, we replace the discounted sum with an L-horizon undiscounted sum. We normalize the contrastive KL regularization to balance their scale, and the process is further simplified by reusing segments ζ^+, ζ^- as a single rollout of policy π^+, π^-, respectively.

$$\mathcal{R}(\pi_\psi;\ \pi^+, \pi^-) \approx \frac{1}{L}\sum_{l=1}^{L}\Big[-\log\frac{\pi^+(a^+_{t+l}|s^+_{t+l})}{\pi_\psi(a^+_{t+l}|s^+_{t+l})} + \log\frac{\pi^-(a^-_{t+l}|s^-_{t+l})}{\pi_\psi(a^-_{t+l}|s^-_{t+l})}\Big].$$

Here, L corresponds to the step of look-ahead during rollouts. When $L = 0$, the framework fully reduces to CPL, which does not account rollout for sequential

planning. Another interesting observation is that if we assume the segments in the offline dataset were generated by the reference policy (i.e., $\pi^+, \pi^- = \pi_{\text{ref}}$), the framework recovers the original DPO formulation, i.e., *forward KL-constrained RLHF implicitly minimizes regret.*

5.3 Experiments

In our experiments, we aim to answer the following questions: (1) Can `PPL` effectively learn in offline settings composed of heterogeneous data generated by diverse policies? (2) Does incorporating policy labels improve learning performance? (3) Can `PPL` be effectively applied to online RLHF algorithm? A full report for each question is provided in the Appendix C.5, C.6 and C.7.

5.3.1 Experimental Setup

For a fair comparison, we first evaluate the performance of `PPL` on six robotic manipulation tasks in MetaWorld [107], using the same rollout data provided by Hejna et al. [44]. Results from the reproducibility check are included in Appendix C.4.3. To evaluate performance on offline datasets generated from diverse policies, we aimed to follow `CPL`'s preference dataset generation procedure. However, there are two key differences in our implementation of the critic. First, we utilize raw rollout data without any trajectory truncation. Second, whereas `CPL` applies a specific technique to reduce TD-error by re-training the critic with all rollout data added to the replay buffer, we generated preference labels without such retraining. As a result, our labels may be noisier than those in `CPL`. Nevertheless, to ensure a fair comparison, all algorithms were trained using the same set of labels. For further details, please see Appendix E.4.

Baselines. We consider CPL as our primary baseline, where the key distinction between PPL and CPL lies in whether the label of the behavior policy is utilized. For additional baselines, we include supervised fine-tuning (SFT) and Preference-based Implicit Q-Learning (P-IQL). Specifically, SFT first trains a policy via behavior cloning on all preferred segments in the preference dataset. P-IQL [43] is a reward-based RLHF algorithm that first learns a reward function from preference data and then derives an optimal policy using the Implicit Q-Learning (IQL) algorithm [56]. Notably, P-IQL is expected to achieve higher performance, as it not only learns a policy but also simultaneously optimizes a reward function, Q-function, and value function.

Implementation Details. To generate preference queries without human supervision, we pretrain an SAC model as an oracle that achieves a 100% success rate. Using this pretrained model as a critic, we uniformly sampled segments of length 64 and assigned labels based on estimated regret. To evaluate performance in heterogeneous datasets, we further construct an additional offline dataset by rolling out suboptimal policies with 20% and 50% success rates and combining them. For preference datasets, we conduct experiments under two settings: Dense, where comparisons are made between all segment pairs, and Sparse, where only one comparison is made per segment.

5.3.2 Can PPL be effectively trained on both homogeneous and heterogeneous offline dataset?

In the previous works, the evaluation of offline datasets has been conducted under homogeneous conditions. However, in practice, offline datasets are more commonly generated by a multiple different policies. Thus, we investigate the following question:

How would PPL and the baselines perform if the offline dataset were hetero-

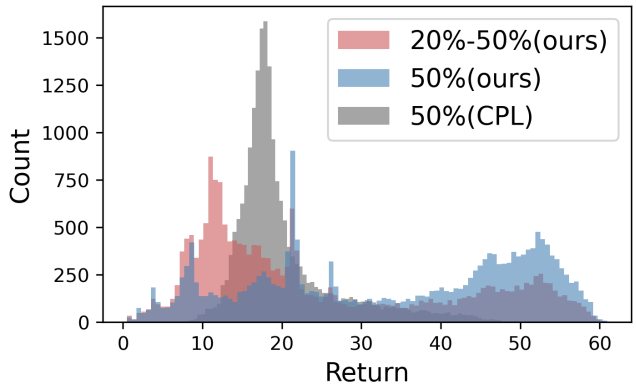


Figure 5.4: Distribution of returns in homogeneous vs heterogeneous offline dataset in `Button-Press-v2`.

geneous?

To investigate this, we examine the distribution of segment returns for both types of datasets, as shown in Figure 5.4. Compared to the homogeneous dataset, the heterogeneous dataset includes rollout data from a policy with a 20% success rate, leading to a higher density of lower-return segments.

In Table 5.2, we report the impact of diverse behavior policies on performance. `PPL` consistently outperforms other methods across various dataset conditions in the MetaWorld benchmark, particularly in challenging scenarios with preference sparsity and policy diversity. Interestingly, unlike baseline algorithms, `PPL` achieves higher performance in `Sparse` settings compared to `Dense` settings. This implies that `PPL` benefits more from datasets with broader state-action coverage rather than relying on dense pairwise comparisons across all segments. Furthermore, `PPL` exhibits greater robustness in heterogeneous datasets, outperforming or matching `P-IQL` despite utilizing only about 6.3% of its parameters. This highlights `PPL` as an efficient algorithm that maintains strong performance while incurring lower computational costs.

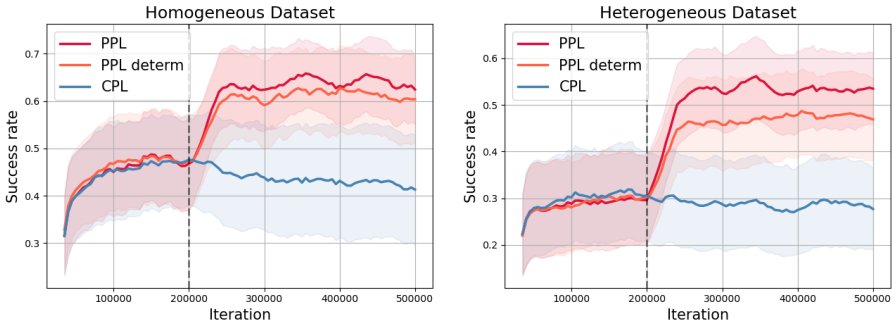


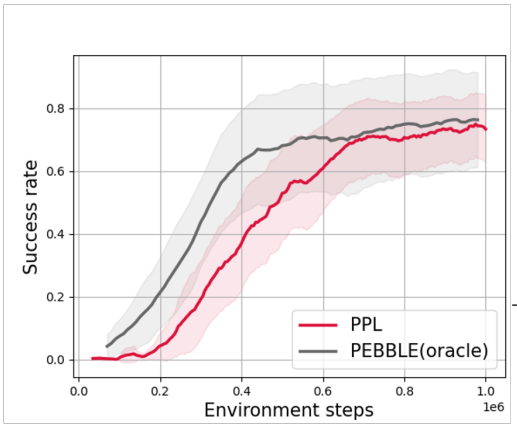
Figure 5.5: Ablation on deterministic pseudo-labeling. We compare the average performance of PPL and PPL-deterministic across six environments in MetaWorld. The dashed line indicates the point where BC pretraining stops.

One possible explanation for CPL's lower performance on our dataset is the absence of the *retraining* technique to reduce TD-error—a method uniquely applied within CPL and not commonly adopted in standard practice. However, since all algorithms were trained using the same labels, we attribute this performance gap primarily to CPL's sensitivity to label noise. This sensitivity appears to arise from an implicit assumption within CPL that all training trajectories are generated by an optimal policy.

5.3.3 Does incorporating policy labels improve learning performance?

In this experiment, we examine how the presence and accuracy of policy labels affect performance. Since the offline dataset are fixed and behavior policies are typically unknown, we ablate a pseudo-label setting, assuming each segment was executed deterministically based on the observed actions. Specifically, we introduce PPL-deterministic, where the behavior policy for each segment is assumed to be fully deterministic (See Lines 4-5 of Algorithm 4). We then compare its performance with PPL.

As shown in Figure 5.5, comparing PPL with CPL reveals that when behavior



Method	Params
PPL	76k
PEBBLE	859k

Figure 5.6: Online learning curves across five MetaWorld tasks, comparing `PPL` and `PEBBLE`.

policy information is not incorporated into learning, distinguishing environmental stochasticity from behavior policy suboptimality becomes more difficult, resulting in a significant performance gap. As an alternative, using deterministic pseudo-labels for training on offline data without policy labels proves to be a viable approach in homogeneous datasets, causing only a slight performance drop. However, in heterogeneous datasets, their effectiveness decreases, leading to a substantial performance gap. This result suggests that as the dataset becomes more diverse in behavior policies, incorporating policy labels into learning becomes increasingly important.

5.3.4 Can `PPL` be effectively applied to an online RLHF algorithm?

In the online setting, rollouts are directly executed, providing explicit access to policy labels. Leveraging this advantage, we conducted experiments to evaluate whether `PPL` can effectively serve as an online DPO algorithm. The experiments were conducted *from scratch*, without any pretraining. Unlike the offline setting, we did not apply the asymmetric regularizer in Eq. 5.1, as out-of-distribution

issues were mitigated by the iterative data collection process. We used PEBBLE [58] as an oracle because it employs a learned policy trained with unsupervised pretraining, which accelerates learning. Further implementation details are provided in Appendix C.4.5.

Figure 5.6 illustrates the average success rates across five MetaWorld tasks. Notably, despite learning from scratch, the online version of PPL achieves performance comparable to PEBBLE, which leverages unsupervised pretraining. Furthermore, since PPL does not require learning a reward model or a critic, it uses only 8.8% of the parameters compared to PEBBLE, yet still achieves comparable performance. This demonstrates that PPL can serve as a highly efficient online RLHF algorithm.

5.4 Summary

In this work, we introduced PPL, a novel DPO framework that incorporates information from the behavior policy through regret-based modeling. We highlighted the issue of likelihood mismatch and addressed it by proposing contrastive KL regularization. Furthermore, we theoretically established that minimizing regret is fundamentally equivalent to optimizing the forward KL-constrained RLHF problem. Empirically, PPL demonstrated strong performance across offline datasets containing rollouts from diverse policies, showcasing its robustness to dataset variations. In online setting, policy labels can be obtained more easily than in the offline case, and PPL effectively learned as an online DPO algorithm. However, we observed that online RLHF method is quite sensitive to the sampling of queries from preference data, suggesting that a more refined analysis is needed for future research.

Chapter 6

Conclusion

This dissertation advances the study of sequential decision-making under uncertain human feedback by integrating distributional reinforcement learning with a regret-based, human-aligned framework. To overcome the limitations of conventional reinforcement learning, which relies on expectation-based updates and explicitly defined rewards, we propose a unified approach that combines two complementary perspectives.

From the distributional perspective, the proposed framework models uncertainty and variability in returns at the level of full distributions, capturing aspects of risk and diversity that expectation-based methods overlook. From the human feedback perspective, it introduces a regret-based modeling paradigm that interprets uncertain and heterogeneous human preferences as structured feedback. By doing so, the agent can better understand human intent, reduce bias in reward estimation, and learn stable, optimal policies aligned with human objectives.

Building upon this foundation, we further develop a principled regret min-

imization framework that provides theoretical guarantees on policy learning efficiency. Through this formulation, regret quantifies the discrepancy between learned and optimal behaviors, serving as a unifying measure that connects human-aligned evaluation with algorithmic efficiency. Empirical results across diverse and high-dimensional environments demonstrate that the proposed algorithms achieve robust performance and faster convergence with fewer environment interactions, validating both their theoretical soundness and practical effectiveness.

6.1 Future Work

The findings of this dissertation offer valuable theoretical and practical insights for advancing reinforcement learning from human feedback (RLHF). Despite these contributions, several promising directions remain for future exploration.

First, while this work primarily considers feedback from a single annotator, future research could investigate methods for aggregating and reconciling heterogeneous feedback from multiple users. Such extensions are critical for developing scalable and socially aligned agents that generalize across diverse human populations.

Second, the current framework assumes that the agent's actions do not dynamically influence the user's feedback model. A natural next step is to explore interactive settings where human preferences evolve in response to the agent's behavior, requiring adaptive algorithms that jointly model human learning and agent learning dynamics.

Finally, future research may extend this framework beyond pairwise preference feedback to encompass richer modalities of human supervision, such as demonstrations, natural-language instructions, or process-level evaluations. Integrating these feedback forms will enable the development of more versatile,

interpretable, and trustworthy agents—furthering the broader goal of creating human-centered artificial intelligence.

Appendix A

Appendix of Chapter 3

A.1 Main Proof

A.1.1 Technical Lemma

Before proving our theoretical results, we present two inequalities for supremum to clear the description.

1. $\sup_{x\in X}|f(x)+g(x)| \leq \sup_{x\in X}|f(x)| + \sup_{x\in X}|g(x)|$
2. $\left|\sup_{x\in X} f(x) - \sup_{x'\in X} g(x')\right| \leq \sup_{x,x'\in X} |f(x)-g(x')|$

Proof of 1. Since $|f(x)+g(x)| \leq |f(x)|+|g(x)|$ holds for all $x \in X$,

$$\begin{aligned}\sup_{x\in X}|f(x)+g(x)| &\leq \sup_{x\in X}(|f(x)|+|g(x)|) \\ &\leq \sup_{x\in X}|f(x)| + \sup_{x\in X}|g(x)|\end{aligned}$$

■

Proof of 2. Since $\left|\|a\|-\|b\|\right| \leq \|a-b\|$ for any norm $\|\cdot\|$ and for a large enough

M,

$$\begin{aligned}\sup_{x,x'\in X}|f(x)-g(x')| &\geq \sup_{x\in X}|f(x)-g(x)| \\ &= \sup_{x\in X}|(f(x)+M)-(g(x)+M)| \\ &\geq \Big|\sup_{x\in X}(f(x)+M)-\sup_{x\in X}(g(x)+M)\Big| \\ &= \Big|\sup_{x\in X}f(x)-\sup_{x'\in X}g(x')\Big|\end{aligned}$$

■

A.1.2 Proof of Theorem A.1.3

Theorem A.1.3. If ξ_t converges to 1 in probability on Ω, then $\mathbb{E}\mathcal{T}_{\xi_t}$ converges to $\mathbb{E}\mathcal{T}$ uniformly on $\mathcal{Z}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$.

Proof. Recall that $\mathcal{Z} = \Big\{Z : \mathcal{S}\times\mathcal{A} \to \mathscr{P}(\mathbb{R})\Big|\ \mathbb{E}[|Z(s,a)|] \leq V_{\max}, \forall (s,a)\Big\}$. Then for any $Z \in \mathcal{Z}$ and $\xi \in \Xi$,

$$\mathbb{E}[|\mathcal{T}_\xi Z|] \leq R_{\max} + \gamma\frac{R_{\max}}{1-\gamma} = \frac{R_{\max}}{1-\gamma} = V_{\max}.$$

which implies PDBOO is closed in $\mathcal{Z}$, i.e. $\mathcal{T}_\xi Z \in \mathcal{Z}$ for all $\xi \in \Xi$. Hence, for any sequence ξ_t, $Z^{(n)} = \mathcal{T}_{\xi_{n:1}} Z \in \mathcal{Z}$ for any $n \geq 0$.

Since ξ_t converges to 1 in probability on Ω, there exists T such that for any $\epsilon, \delta > 0$ and $t > T$,

$$\mathbb{P}(\Omega_t) \coloneqq \mathbb{P}\left(\left\{w \in \Omega : \sup_{w\in\Omega}|\xi_t(w)-1| \geq \epsilon\right\}\right) \leq \delta$$

For any $Z \in \mathcal{Z}$, $s \in \mathcal{S}, a \in \mathcal{A}$, and $t > T$, by using Hölder's inequality,

$$\begin{aligned}&\sup_{Z\in\mathcal{Z}}\sup_{s,a}|\mathbb{E}_{\xi_t}[Z(s,a)] - \mathbb{E}[Z(s,a)]| \\ &= \sup_{Z\in\mathcal{Z}}\sup_{s,a}\left|\int_{w\in\Omega}(1-\xi_t(w))Z(s,a,w)\mathbb{P}(dw)\right| \\ &= \sup_{Z\in\mathcal{Z}}\sup_{s,a}\Big|\int_{w\in\Omega_t}(1-\xi_t(w))Z(s,a,w)\mathbb{P}(dw) \\ &\qquad\qquad + \int_{w\in\Omega\setminus\Omega_t}(1-\xi_t(w))Z(s,a,w)\mathbb{P}(dw)\Big| \\ &\leq \mathbb{P}(\Omega_t)\sup_{w\in\Omega_t}|\xi_t(w)-1|\ V_{\max} + \mathbb{P}(\Omega\setminus\Omega_t)\sup_{w\in\Omega\setminus\Omega_t}|\xi_t(w)-1|\ V_{\max} \\ &\leq \delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}\end{aligned}$$

which implies that $\mathbb{E}_{\xi_t}$ converges to $\mathbb{E}$ uniformly on $\mathcal{Z}$ for all s, a.

By using A.1.1, we can get the desired result.

$$
\begin{aligned}
&\sup_{Z\in\mathcal{Z}}\sup_{s,a}\ |\mathbb{E}[\mathcal{T}_{\xi_t}Z(s,a)] - \mathbb{E}[\mathcal{T}Z(s,a)]| \\
&\leq \sup_{Z\in\mathcal{Z}}\sup_{s,a}\ |\mathbb{E}[\mathcal{T}_{\xi_t}Z(s,a)] - \mathbb{E}_{\xi_t}[\mathcal{T}_{\xi_t}Z(s,a)]| \\
&\quad + \sup_{Z\in\mathcal{Z}}\sup_{s,a}\ |\mathbb{E}_{\xi_t}[\mathcal{T}_{\xi_t}Z(s,a)] - \mathbb{E}[\mathcal{T}Z(s,a)]| \\
&\leq (\delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}) \\
&\quad + \gamma \sup_{Z\in\mathcal{Z}}\sup_{s,a}\mathbb{E}_{s'}\left[\left|\sup_{a'}\mathbb{E}_{\xi_t}[Z(s',a')] - \sup_{a''}\mathbb{E}[Z(s',a'')]\right|\right] \\
&\leq (\delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}) \\
&\quad + \gamma \sup_{Z\in\mathcal{Z}}\sup_{s',a'}\left|\mathbb{E}_{\xi_t}[Z(s',a')] - \mathbb{E}[Z(s',a')]\right| \\
&\leq (\delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}) + \gamma(\delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}) \\
&= (1+\gamma)(\delta|B_\xi - 1|V_{\max} + \epsilon V_{\max}).
\end{aligned}
$$

■

A.1.3 Proof of Theorem 3.2.3

Theorem 3.2.3. Let ξ_n be sampled from $\bar{\mathcal{U}}_{\Delta_n}(Z^{(n-1)})$ for every iteration. If Assumption 3.2.2 holds, then the expectation of any composition of operators $\mathbb{E}\mathcal{T}_{\xi_{n:1}}$ converges, i.e. $\mathbb{E}\mathcal{T}_{\xi_{n:1}}[Z] \to \mathbb{E}[Z^*]$

Moreover, the following bound holds,

$$
\begin{aligned}
&\sup_{s,a}\ \left|\mathbb{E}[Z^{(n)}(s,a)] - \mathbb{E}[Z^*(s,a)]\right| \\
&\qquad \leq \sum_{k=n}^{\infty}\left(2\gamma^{k-1}V_{\max} + 2\sum_{i=1}^{k}\gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i})\right).
\end{aligned}
$$

Proof. We denote $a_i^*(\xi_n) = \operatorname{argmax}_{a'} \mathbb{E}_{\xi_n}[Z_i^{(n-1)}(s',a')]$ as the greedy action of $Z_i^{(n-1)}$ under perturbation ξ_n. Also, we denote $\sup_{s,a}|\cdot|$ which is the supremum norm over s and a as $\|\cdot\|_{sa}$.

Before we start from the term $\left\|\mathbb{E}[Z^{(k+1)}] - \mathbb{E}[Z^{(k)}]\right\|_{sa}$, for a given (s, a),

$$
\begin{aligned}
&\left|\mathbb{E}[Z^{(k+1)}(s,a)] - \mathbb{E}[Z^{(k)}(s,a)]\right| \\
&\leq \gamma \sup_{s'} \left|\mathbb{E}[Z^{(k)}(s', a^*(\xi_{k+1}))] - \mathbb{E}[Z^{(k-1)}(s', a^*(\xi_k))]\right| \\
&\leq \gamma \sup_{s'} \Big(\left|\mathbb{E}[Z^{(k)}(s', a^*(\xi_{k+1})] - \max_{a'} \mathbb{E}[Z^{(k)}(s', a')]\right| \\
&\quad + \left| \max_{a'} \mathbb{E}[Z^{(k)}(s', a')] - \max_{a'} \mathbb{E}[Z^{(k-1)}(s', a')]\right| \\
&\quad + \left|\max_{a'} \mathbb{E}[Z^{(k-1)}(s', a')] - \mathbb{E}[Z^{(k-1)}(s', a^*(\xi_k))]\right| \Big) \\
&\leq \gamma \sup_{s', a'} \left|\mathbb{E}[Z^{(k)}(s', a')] - \mathbb{E}[Z^{(k-1)}(s', a')]\right| \\
&\quad + \gamma \sum_{i=k-1}^{k} \sup_{s'} \left|\mathbb{E}[Z^{(i)}(s', a^*(\xi_{i+1}))] - \max_{a'} \mathbb{E}[Z^{(i)}(s', a')]\right| \\
&\leq \gamma \left\|\mathbb{E}[Z^{(k)}] - \mathbb{E}[Z^{(k-1)}]\right\|_{sa} \\
&\quad + \gamma \sum_{i=k-1}^{k} \Big[\sup_{s'} \Big(\left|\mathbb{E}[Z^{(i)}(s', a^*(\xi_{i+1}))] - \mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s', a^*(\xi_{i+1}))]\right| \\
&\quad + \left|\max_{a'} \mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s', a')] - \max_{a''} \mathbb{E}[Z^{(i)}(s', a''))]\right| \Big) \Big] \\
&\leq \gamma \left\|\mathbb{E}[Z^{(k)}] - \mathbb{E}[Z^{(k-1)}]\right\|_{sa} \\
&\quad + 2\gamma \sum_{i=k-1}^{k} \sup_{s', a'} \Big(\left|\mathbb{E}[Z^{(i)}(s', a')] - \mathbb{E}_{\xi_{i+1}}[Z^{(i)}(s', a')]\right| \Big) \\
&\leq \gamma \left\|\mathbb{E}[Z^{(k)}] - \mathbb{E}[Z^{(k-1)}]\right\|_{sa} + 2\gamma \sum_{i=k-1}^{k} \Delta_{i+1}
\end{aligned}
$$

where we use A.1.1.1 in third and fifth line and A.1.1.2 in sixth line.

Taking a supremum over s and a, then for all $k > 0$,

$$
\begin{aligned}
&\left\|\mathbb{E}[Z^{(k+1)}] - \mathbb{E}[Z^{(k)}]\right\|_{sa} \\
&\leq \gamma \left\|\mathbb{E}[Z^{(k)}] - \mathbb{E}[Z^{(k-1)}]\right\|_{sa} + 2\sum_{i=k-1}^{k} \gamma\Delta_{i+1} \\
&\leq \gamma^2 \left\|\mathbb{E}[Z^{(k-1)}] - \mathbb{E}[Z^{(k-2)}]\right\|_{sa} + 2\sum_{i=k-2}^{k-1} \gamma^2\Delta_{i+1} + 2\sum_{i=k-1}^{k} \gamma\Delta_{i+1} \\
&\quad\vdots \\
&\leq \gamma^k \left\|\mathbb{E}[Z^{(1)}] - \mathbb{E}[Z]\right\|_{sa} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i}) \\
&\leq 2\gamma^k V_{\max} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i})
\end{aligned}
$$

Since $\sum_{i=1}^{\infty} \gamma^i = \frac{\gamma}{1-\gamma} < \infty$ and $\sum_{i=1}^{\infty} \Delta_i < \infty$ by assumption, we have

$$\sum_{i=1}^{k} \gamma^i \Delta_{k+1-i} \to 0$$

which is resulted from the convergence of Cauchy product of two sequences $\{\gamma^i\}$ and $\{\Delta_i\}$. Hence, $\{\mathbb{E}[Z^{(k)}]\}$ is a Cauchy sequence and therefore converges for every $Z \in \mathcal{Z}$.

Let $\mathbb{E}[Z^*]$ be the limit point of the sequence $\{\mathbb{E}[Z^{(n)}]\}$. Then,

$$
\begin{aligned}
\left\|\mathbb{E}[Z^*] - \mathbb{E}[Z^{(n)}]\right\|_{sa} &= \lim_{l\to\infty} \left\|\mathbb{E}[Z^{(n+l)}] - \mathbb{E}[Z^{(n)}]\right\|_{sa} \\
&\leq \sum_{k=n}^{\infty} \left\|\mathbb{E}[Z^{(k+1)}] - \mathbb{E}[Z^{(k)}]\right\|_{sa} \\
&= \sum_{k=n}^{\infty} \left(2\gamma^k V_{\max} + 2\sum_{i=1}^{k} \gamma^i(\Delta_{k+2-i} + \Delta_{k+1-i})\right).
\end{aligned}
$$

■

A.1.4 Proof of Theorem 3.2.4

Theorem 3.2.4. If $\{\Delta_n\}$ follows the assumption in Theorem 3.2.3, then $\mathbb{E}[Z^*]$ is the unique solution of Bellman optimality equation.

Proof. The proof follows by linearity of expectation. Denote the Q-value based operator as $\bar{\mathcal{T}}$. Note that Δ_n converges to 0 with regularity of $\mathcal{Z}$ implies that ξ_n converges to 1 in probability on Ω, i.e.,

$$\begin{aligned}&\lim_{n\to\infty}\sup_{s,a}\left|\int_{w\in\Omega} Z^{(n)}(w;s,a)(1-\xi_n(w))\mathbb{P}(dw)\right| = 0\\ \implies &\lim_{n\to\infty}\mathbb{P}\left(\{w\in\Omega : |1-\xi_n(w)|\geq\epsilon\}\right) = 0\end{aligned}$$

By Theorem A.1.3, for a given $\epsilon > 0$, there exists a constant $K = \max(K_1, K_2)$ such that for every $k \geq K_1$,

$$\sup_{Z\in\mathcal{Z}}\|\bar{\mathcal{T}}_{\xi_k}\mathbb{E}[Z] - \bar{\mathcal{T}}\mathbb{E}[Z]\|_{sa} \leq \frac{\epsilon}{2}.$$

Since $\bar{\mathcal{T}}$ is continuous, for every $k \geq K_2$,

$$\|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa} \leq \frac{\epsilon}{2}.$$

Thus, it holds that

$$\begin{aligned}&\|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa}\\ &\leq \|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^{(k)}]\|_{sa} + \|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa}\\ &\leq \sup_{Z\in\mathcal{Z}}\|\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z] - \bar{\mathcal{T}}\mathbb{E}[Z]\|_{sa} + \|\bar{\mathcal{T}}\mathbb{E}[Z^{(k)}] - \bar{\mathcal{T}}\mathbb{E}[Z^*]\|_{sa}\\ &\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[Z^*] &= \lim_{k\to\infty}\mathbb{E}[Z^{(k)}] = \lim_{k\to\infty}\mathbb{E}[Z^{(k+1)}]\\ &= \lim_{k\to\infty}\mathbb{E}[\mathcal{T}_{\xi_{k+1}}Z^{(k)}] = \lim_{k\to\infty}\bar{\mathcal{T}}_{\xi_{k+1}}\mathbb{E}[Z^{(k)}] = \bar{\mathcal{T}}\mathbb{E}[Z^*]\end{aligned}$$

Since the standard Bellman optimality operator has a unique solution, we derived the desired result. ■

A.2 Implementation details

Except for each own hyperparameter, our algorithms and `DLTV` shares the same hyperparameter and network architecture with `QR-DQN` [31] for a fair comparison. Also, we set up `p-DLTV` by only multiplying a gaussian noise $\mathcal{N}(0, 1)$ to the coefficient of DLTV. We **do not combine** any additional improvements of Rainbow such as double Q-learning, dueling network, prioritized replay, and n-step update. Experiments on LunarLander-v2 and Atari games were performed with 3 random seeds. The training process is 0-2% slower than `QR-DQN` due to the sampling ξ and reweighting procedures.

A.2.1 Hyperparameter Setting

We report the hyperparameters for each environments we used in our experiments.

Table A.1: Table of hyperparameter setting

Hyperparameters	N-Chain	LunarLander	Atari Games
Batch size	64	128	32
Number of quantiles	200	170	200
$n-$step updates	1		
Network optimizer	Adam		
β	Grid search[0.05, 0.1, 0.5, 1] $\times \mathbf{1}^N$		
κ	1		
Memory size	1e6	1e5	1e6
Learning rate	5e-5	1.5e-3	5e-5
γ	0.9	0.99	0.99
Update interval	1	1	4
Target update interval	25	1	1e4
Start steps	5e2	1e4	5e4
ϵ (train)	LinearAnnealer(1 $\rightarrow$ 1e-2)		
ϵ (test)	1e-3	1e-3	1e-3
ϵ decay steps	2.5e3	1e5	2.5e5
Coefficient c	Grid search[1e0, 5e0, 1e1, 5e1, 1e2, 5e2, 1e3, 5e3]		
Δ_0	5e2	5e4	1e6
Number of seeds	10	3	3

A.3 Raw scores across 55 Atari games

Table A.2: Raw scores across all 55 games, starting with 30 no-op actions. We report the best scores for `DQN`, `QR-DQN`, `IQN` and `Rainbow` on 50M frames, averaged by 5 seeds. Reference values were provided by DQN_Zoo framework [76]. **Bold** are wins against `DQN`, `QR-DQN` and `IQN`, and *asterisk are wins over `Rainbow`.

GAMES	RANDOM	HUMAN	DQN(50M)	QR-DQN(50M)	IQN(50M)	RAINBOW(50M)	PQR(50M)
Alien	227.8	7127.7	1541.5	1645.7	1769.2	4356.9	**2455.8**
Amidar	5.8	1719.5	324.2	683.4	799.2	2549.2	**938.4**
Assault	222.4	742.0	2387.8	11684.2	15152.4	9737.0	10759.2
Asterix	210.0	8503.3	5249.5	18373.4	32598.2	33378.6	10490.5
Asteroids	719.1	47388.7	1106.3	1503.9	1972.6	1825.4	1662.0
Atlantis	12850.0	29028.1	283392.2	937275.0	865360.0	941740.0	897640.0
BankHeist	14.2	753.1	389.0	1223.9	1266.8	1081.7	1038.8
BattleZone	2360.0	37187.5	19092.4	26325.0	30253.9	35467.1	28470.5
BeamRider	363.9	16926.5	7133.1	12912.0	19251.4	15421.9	10224.9
Berzerk	123.7	2630.4	577.4	826.5	918.9	2061.6	***137873.1**
Bowling	23.1	160.7	34.4	45.4	41.5	54.7	***86.9**
Boxing	0.1	12.1	87.2	99.6	99.2	99.8	97.1
Breakout	1.7	30.5	316.8	426.5	468.0	335.3	380.3
Centipede	2090.9	12017.0	4935.7	7124.0	7008.3	5691.4	***7291.2**
ChopperCommand	811.0	7387.8	974.2	1187.8	1549.0	5525.1	1300.0
CrazyClimber	10780.5	35829.4	96939.0	93499.1	127156.5	160757.7	84390.9
DemonAttack	152.1	1971.0	8325.6	106401.8	110773.1	85776.5	73794.0
DoubleDunk	-18.6	-16.4	-15.7	-10.5	-12.1	-0.3	**-7.5**
Enduro	0.0	860.5	750.6	2105.7	2280.6	2318.3	***2341.2**
FishingDerby	-91.7	-38.7	8.2	25.7	23.4	35.5	**31.7**
Freeway	0.0	29.6	24.4	33.3	33.7	34.0	**34.0**
Frostbite	65.2	4334.7	408.2	3859.2	5650.8	9672.6	4148.2
Gopher	257.6	2412.5	3439.4	6561.9	26768.9	32081.3	***47054.5**
Gravitar	173.0	3351.4	180.9	548.1	470.2	2236.8	**635.8**
Hero	1027.0	30826.4	9948.3	9909.8	12491.1	38017.9	**12579.2**
IceHockey	-11.2	0.9	-11.4	-2.1	-4.2	1.9	**-1.4**
Jamesbond	29.0	302.8	486.4	1163.8	1058.0	14415.5	**2121.8**
Kangaroo	52.0	3035.0	6720.7	14558.2	14256.0	14383.6	***14617.1**
Krull	1598.0	2665.5	7130.5	9612.5	9616.7	8328.5	***9746.1**
KungFuMaster	258.5	22736.3	21330.9	27764.3	39450.1	30506.9	***43258.6**
MontezumaRevenge	0.0	4753.3	0.3	0.0	0.2	80.0	0.0
MsPacman	307.3	6951.6	2362.9	2877.5	2737.4	3703.4	**2928.9**
NameThisGame	2292.3	8049.0	6328.0	11843.3	11582.2	11341.5	10298.2
Phoenix	761.4	7242.6	10153.6	35128.6	29138.9	49138.8	20453.8
Pitfall	-229.4	6463.7	-9.5	0.0	0.0	0.0	**0.0**
Pong	-20.7	14.6	18.7	20.9	20.9	21.0	**21.0**
PrivateEye	24.9	69571.3	266.6	100.0	100.0	160.0	***372.4**
Qbert	163.9	13455.0	5567.9	12808.4	15101.8	24484.9	**15267.4**
Riverraid	1338.5	17118.0	6782.8	9721.9	13555.9	17522.9	11175.3
RoadRunner	11.5	7845.0	29137.5	54276.3	53850.9	52222.6	50854.7
Robotank	2.2	11.9	31.4	54.5	53.8	64.5	**60.3**
Seaquest	68.4	42054.7	2525.8	7608.2	17085.6	3048.9	***19652.5**
Skiing	-17098.1	-4336.9	-13930.8	-14589.7	-19191.1	-15232.3	***-9299.3**
Solaris	1236.3	12326.7	2031.5	1857.3	1301.5	2522.6	***2640.0**
SpaceInvaders	148.0	1668.7	1179.1	1753.2	2906.7	2715.3	1749.4
StarGunner	664.0	10250.0	24532.5	63717.3	78503.4	107177.8	62920.6
Tennis	-23.8	-8.3	-0.9	0.0	0.0	0.0	-1.0
TimePilot	3568.0	5229.2	2091.8	6266.8	6379.1	12082.1	**6506.4**
Tutankham	11.4	167.6	138.7	210.2	204.4	194.3	***231.3**
UpNDown	533.4	11693.2	6724.5	27311.3	35797.6	65174.2	**36008.1**
Venture	0.0	1187.5	53.3	12.5	17.4	1.1	***993.3**
VideoPinball	16256.9	17667.9	140528.4	104405.8	341767.5	465636.5	**465578.3**
WizardOfWor	563.5	4756.5	3459.9	14370.2	10612.1	12056.1	6132.8
YarsRevenge	3092.9	54576.9	16433.7	21641.4	21645.0	67893.3	**27674.4**
Zaxxon	32.5	9173.3	3244.9	9172.1	8205.2	22045.8	**10806.6**

Table A.3: Raw scores across 55 games. We report the best scores for `DQN`, `QR-DQN`, `IQN*`, and `Rainbow` on 50M frames, averaged by 5 seeds. Reference values were provided by Dopamine framework [19]. **Bolds** are wins against `DQN`, `QR-DQN`, and *asterisk are wins over `IQN*` and `Rainbow`. Note that `IQN*` and `Rainbow` implemented in Dopamine framework applied n-step updates with $n = 3$ which improves performance.

GAMES	RANDOM	HUMAN	DQN(50M)	QR-DQN(50M)	IQN*(50M)	RAINBOW(50M)	PQR(50M)
Alien	227.8	7127.7	1688.1	2754.2	4016.3	2076.2	**3173.9**
Amidar	5.8	1719.5	888.2	841.6	1642.8	1669.6	***2814.7**
Assault	222.4	742.0	1615.9	2233.1	4305.6	2535.9	***8456.5**
Asterix	210.0	8503.3	3326.1	3540.1	7038.4	5862.3	***19004.6**
Asteroids	719.1	47388.7	828.2	1333.4	1336.3	1345.1	851.8
Atlantis	12850.0	29028.1	388466.7	879022.0	897558.0	870896.0	**880303.7**
BankHeist	14.2	753.1	720.2	964.1	1082.8	1104.9	**1050.1**
BattleZone	2360.0	37187.5	15110.3	25845.6	29959.7	32862.1	***61494.4**
BeamRider	343.9	16926.5	4771.3	7143.0	7113.7	6331.9	***12217.6**
Berzerk	123.7	2630.4	529.2	603.2	627.3	697.8	***2707.2**
Bowling	23.1	160.7	38.5	55.3	33.6	55.0	***174.1**
Boxing	0.1	12.1	80.0	96.6	97.8	96.3	**96.7**
Breakout	1.7	30.5	113.5	40.7	164.4	69.8	48.5
Centipede	2090.9	12017.0	3403.7	3562.5	3746.1	5087.6	***31079.8**
ChopperCommand	811.0	7387.8	1615.3	1600.3	6654.1	5982.0	**4653.9**
CrazyClimber	10780.5	35829.4	111493.8	108493.9	131645.8	135786.1	105526.0
DemonAttack	152.1	1971.0	4396.7	3182.6	7715.5	6346.4	***19530.2**
DoubleDunk	-18.6	-16.4	-16.7	7.4	20.2	17.4	**15.0**
Enduro	0.0	860.5	2268.1	2062.5	766.5	2255.6	1765.5
FishingDerby	-91.7	-38.7	12.3	48.4	41.9	37.6	46.8
Freeway	0.0	29.6	25.8	33.5	33.5	33.2	33.0
Frostbite	65.2	4334.7	760.2	8022.8	7824.9	5697.2	***8401.5**
Gopher	257.6	2412.5	3495.8	3917.1	11192.6	7102.1	***12252.9**
Gravitar	173.0	3351.4	250.7	821.3	1083.5	926.2	703.5
Hero	1027.0	30826.4	12316.4	14980.0	18754.0	31254.8	**15655.8**
IceHockey	-11.2	0.9	-6.7	-4.5	0.0	2.3	**0.0**
Jamesbond	29.0	302.8	500.0	802.3	1118.8	656.7	***1454.9**
Kangaroo	52.0	3035.0	6768.2	4727.3	11385.4	13133.1	***13894.0**
Krull	1598	2665.5	6181.1	8073.9	8661.7	6292.5	***31927.4**
KungFuMaster	258.5	22736.3	20418.8	20988.3	33099.9	26707.0	**22040.4**
MontezumaRevenge	0.0	4753.3	2.6	300.5	0.7	501.2	0.0
MsPacman	307.3	6951.6	2727.2	3313.9	4714.4	3406.4	***5426.5**
NameThisGame	2292.3	8049.0	5697.3	7307.9	9432.8	9389.5	***9891.3**
Phoenix	761.4	7245.6	5833.7	4641.1	5147.2	8272.9	5260
Pitfall	-229.4	6463.7	-16.8	-3.4	-0.4	0.0	***0.0**
Pong	-20.7	14.6	13.2	19.2	19.9	19.4	**19.7**
PrivateEye	24.9	69571.3	1884.6	680.7	1287.3	4298.8	***12806.1**
Qbert	163.9	13455.0	8216.2	17228.0	15045.5	17121.4	15806.9
Riverraid	1338.5	17118.0	9077.8	13389.4	14868.6	15748.9	**14101.3**
RoadRunner	11.5	7845.0	39703.1	44619.2	50534.1	51442.4	**48339.7**
Robotank	2.2	11.9	25.8	53.6	65.9	63.6	48.7
Seaquest	68.4	42054.7	1585.9	4667.9	20081.3	3916.2	**5038.1**
Skiing	-17098.1	-4336.9	-17038.2	-14401.6	-13755.6	-17960.1	***-9021.2**
Solaris	1236.3	12326.7	2029.5	2361.7	2234.5	2922.2	***7145.3**
SpaceInvaders	148.0	1668.7	1361.1	940.2	3115.0	1908.0	**1602.4**
StarGunner	664.0	10250.0	1676.5	23593.3	60090.0	39456.3	**59404.6**
Tennis	-23.8	-9.3	-0.1	19.2	3.5	0.0	*15.4
TimePilot	3568.0	5229.2	3200.9	6622.8	9820.6	9324.4	5597.0
Tutankham	11.4	167.6	138.8	209.9	250.4	252.2	147.3
UpNDown	533.4	11693.2	10405.6	29890.1	44327.6	18790.7	**32155.5**
Venture	0.0	1187.5	50.8	1099.6	1134.5	1488.9	1000.0
VideoPinball	16256.9	17667.9	216042.7	250650.0	486111.5	536364.4	**460860.9**
WizardOfWor	563.5	4756.5	2664.9	2841.8	6791.4	7562.7	**5738.2**
YarsRevenge	3092.9	54576.9	20375.7	66055.9	57960.3	31864.4	***67545.8**
Zaxxon	32.5	9173.3	1928.6	8177.2	12048.6	14117.5	**9531.8**

Table A.4: Raw scores across all 49 games, starting with 30 no-op actions. We report the best scores for QR-DQN_zoo [76], QR-DQN_Zhang [111](implemented by QUOTA to evaluate the relative improvement) for a fair comparison and QUOTA [111], DLTV [64] on 40M frames, averaged by 3 seeds. **Bold** are wins against QUOTA and DLTV.

Games	Random	Human	QR-DQN_zoo(40M)	QR-DQN_Zhang(40M)	QUOTA(40M)	DLTV(40M)	PQR(40M)
Alien	227.8	7127.7	1645.7	1760.0	1821.9	2280.9	**2406.9**
Amidar	5.8	1719.5	552.9	567.9	571.4	1042.7	644.1
Assault	222.4	742	9880.4	3308.7	3511.1	5896.2	**10759.2**
Asterix	210	8503.3	13157.2	6176.0	6112.1	6336.6	**8431.0**
Asteroids	719.1	47388.7	1503.9	1305.3	1497.6	1268.7	1416.00
Atlantis	12850	29028.1	750190.1	978385.3	965193.0	845324.9	897640.0
BankHeist	14.2	753.1	1146.1	644.7	735.2	1183.7	1038.8
BattleZone	2360	37187.5	17788.4	22725.0	25321.6	23315.8	**28470.5**
BeamRider	363.9	16926.5	10684.2	5007.8	5522.6	6490.1	**10224.9**
Bowling	23.1	160.7	44.3	27.6	34.0	29.8	**86.9**
Boxing	0.1	12.1	98.2	95.0	96.1	112.8	97.1
Breakout	1.7	30.5	401.5	322.1	316.7	260.9	**357.7**
Centipede	2090.9	12017.0	6633.0	4330.3	3537.9	4676.7	**6803.6**
ChopperCommand	811.0	7387.8	1133.1	3421.1	3793.0	2586.3	1500.0
CrazyClimber	10780.5	35829.4	93499.1	107371.6	113051.7	92769.1	83900.0
DemonAttack	152.1	1971.0	98063.6	80026.6	61005.1	146928.9	73794.0
DoubleDunk	-18.6	-16.4	-10.5	-21.6	-21.5	-23.3	**-10.5**
Enduro	0.0	860.5	2105.7	1220.0	1162.3	5665.9	2252.8
FishingDerby	-91.7	-38.7	25.7	-9.6	-59.0	-8.2	**31.7**
Freeway	0.0	29.6	30.9	30.6	31.0	34.0	**34.0**
Frostbite	65.2	4334.7	3822.7	2046.3	2208.5	3867.6	**4051.2**
Gopher	257.6	2412.5	4191.2	9443.8	6824.3	10199.4	**47054.5**
Gravitar	173.0	3351.4	477.4	414.3	457.6	357.9	**583.6**
IceHockey	-11.2	0.9	-2.4	-9.8	-9.9	-14.3	**-2.1**
Jamesbond	29.0	302.8	907.1	601.7	495.5	779.8	**1747.1**
Kangaroo	52.0	3035	14171	2364.6	2555.8	4596.7	**14385.1**
Krull	1598.0	2665.5	9618.2	7725.4	7747.5	10012.21	9537.0
KungFuMaster	258.5	22736.3	27576.5	17807.4	20992.5	23078.4	**38074.1**
MontezumaRevenge	0.0	4753.3	0.0	0.0	0.0	0.0	0.0
MsPacman	307.3	6951.6	2561.0	2273.3	2423.5	3191.7	2895.6
NameThisGame	2292.3	8049.0	11770.0	7748.2	7327.5	8368.1	**10298.2**
Pitfall	-229.4	6463.7	0.0	-32.9	-30.7	-	**0.0**
Pong	-20.7	14.6	20.9	19.6	20.0	21.0	**21.0**
PrivateEye	24.9	69571.3	100.0	419.3	114.1	1358.6	**372.4**
Qbert	163.9	13455.0	8348.2	10875.3	11790.2	15856.2	14593.0
Riverraid	1338.5	17118.0	8814.1	9710.4	10169.8	10487.3	9374.7
RoadRunner	11.5	7845.0	52575.7	27640.7	27872.2	49255.7	44341.0
Robotank	2.2	11.9	50.4	45.1	37.6	58.4	53.9
Seaquest	68.4	42054.7	5854.6	1690.5	2628.6	3103.8	**16011.2**
SpaceInvaders	148.0	1668.7	1281.8	1387.6	1553.8	1498.6	**1562.6**
StarGunner	664.0	10250.0	53624.7	49286.6	52920.0	53229.5	**55475.0**
Tennis	-23.8	-8.3	0.0	-22.7	-23.7	-18.4	**-1.0**
TimePilot	3568.0	5229.2	6243.4	6417.7	5125.1	6931.1	6506.4
Tutankham	11.4	167.6	200.0	173.2	195.4	130.9	**213.3**
UpNDown	533.4	11693.2	22248.8	30443.6	24912.7	44386.7	33786.3
Venture	0.0	1187.5	12.5	5.3	26.5	1305.0	0.0
VideoPinball	16256.9	17667.9	104227.2	123425.4	44919.1	93309.6	**443870.0**
WizardOfWor	563.5	4756.5	13133.8	5219.0	4582.0	9582.0	6132.8
Zaxxon	32.5	9173.3	7222.7	6855.1	8252.8	6293.0	**10250.0**

Appendix B

Appendix of Chapter 4

B.1 Notation

Table B.1: Table of notation (Part 1: core and statistical notation)

Notation	Description
$\mathcal{S}$	State space of size S.
$\mathcal{A}$	Action space of size A.
H	Horizon length of one episode.
T	Number of episodes.
$r_h(s,a)$	Reward of (s,a) at step h.
$\mathbb{P}_h(s'\|s,a)$	Transition probability from (s,a) to s' at step h.
$\mathbb{H}_h^k$	History up to step h in episode k.
N	Number of statistical functionals.
$Q_h^\pi(s,a)$	Q-function of a given policy π at step h.
$V_h^\pi(s)$	V-function of a given policy π at step h.
$Z_h^\pi(s,a)$	Random variable of the Q-function.
$\bar{Z}_h^\pi(s)$	Random variable of the V-function.
$\eta_h^\pi(s,a)$	Probability distribution of the Q-function.
$\bar{\eta}_h^\pi(s)$	Probability distribution of the V-function.
$[\mathbb{P}_h(\cdot)]$	Expectation over transition, $[\mathbb{P}_h(\cdot)] = \mathbb{E}_{s'\sim\mathbb{P}_h}(\cdot)$.
$(\mathcal{B}_r)_\#$	Pushforward of the distribution through $\mathcal{B}_r(x) \coloneqq r + x$.
$\psi(\bar{\eta})$	Statistical functional $\mathscr{P}_\psi(\mathbb{R})^{\mathcal{S}} \to \mathbb{R}^S$.
$\psi_{1:N}(\bar{\eta})$	N-collection of statistical functionals $\mathscr{P}_{\psi_{1:N}}(\mathbb{R})^{\mathcal{S}} \to \mathbb{R}^{N\times S}$.
$\mathscr{P}_{\psi_{1:N}}(\mathbb{R})$	Domain of the sketch $\psi_{1:N}$.
$I_{\psi_{1:N}}$	Image of the sketch $\psi_{1:N}$.
$\mathcal{T}$	Distributional Bellman operator, $\mathcal{T}\bar{\eta} \coloneqq (\mathcal{B}_r)_\#[\mathbb{P}\bar{\eta}]$.
$\mathcal{T}_\psi$	Sketch Bellman operator w.r.t. ψ, $\mathcal{T}_\psi\psi(\bar{\eta}) \coloneqq \psi((\mathcal{B}_r)_\#[\mathbb{P}\bar{\eta}])$.
$\hat{\mathcal{T}}_\psi$	Empirical sketch Bellman operator w.r.t. ψ, $\hat{\mathcal{T}}_\psi\psi(\bar{\eta}) \coloneqq \psi((\mathcal{B}_r)_\#[\hat{\mathbb{P}}\bar{\eta}])$.

Table B.2: Table of notation (Part 2: functionals and dataset-related quantities)

Notation	**Description**
$f^{(n)}$	n-th element of an N-dimensional vector f.
$\|f\|_\infty$	Max norm of $f : X \to \mathbb{R}$, $\|f\|_\infty \coloneqq \max_{x \in X} \|f^{(n)}(x)\|$.
$\|f\|_{\infty,1}$	l_1-norm of max norms, $\|f\|_{\infty,1} \coloneqq \sum_{n=1}^{N} \max_{x \in X} \|f^{(n)}(x)\|$.
$\mathcal{F}^N$	Function class of N-dimensional embedding space.
$\mathcal{Z}$	Set of state–action pairs $\mathcal{Z} \coloneqq \{(s_t, a_t)\}_{t=1}^{\|\mathcal{Z}\|}$.
$\mathcal{D}$	Dataset $\mathcal{D} \coloneqq \{(s_t, a_t, [d_t^{(1)}, \ldots, d_t^{(N)}])\}_{t=1}^{\|\mathcal{D}\|}$.
$\|f\|_{\mathcal{Z}}^2$	For $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|f\|_{\mathcal{Z}}^2 \coloneqq \sum_{n=1}^{N} \sum_{(s,a) \in \mathcal{Z}} (f^{(n)}(s,a))^2$.
$\|f\|_{\mathcal{D}}^2$	For $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\|f\|_{\mathcal{D}}^2 \coloneqq \sum_{n=1}^{N} \sum_{t=1}^{\|\mathcal{D}\|} (f^{(n)}(s_t, a_t) - d_t^{(n)})^2$.
$w^{(n)}(\mathcal{F}^N, s, a)$	Width function at (s,a), $w^{(n)}(\mathcal{F}^N, s, a) \coloneqq \max_{f,g \in \mathcal{F}^N} \|f^{(n)}(s,a) - g^{(n)}(s,a)\|$.
$\tilde{f}_{h,\bar{\eta}}^k$	Solution of moment least squares regression, $\tilde{f}_{h,\bar{\eta}}^k \coloneqq \arg\min_{f \in \mathcal{F}^N} \|f\|_{\mathcal{D}_h^k}$.
$f_{\bar{\eta}}$	Target sketch of distribution $\bar{\eta}$, $f_{\bar{\eta}} \coloneqq \psi_{1:N}((\mathcal{B}_r)_\#[\mathbb{P}_h \bar{\eta}])$.
$(\mathcal{F}^N)_h^k$	Confidence region at step h, episode k, $(\mathcal{F}^N)_h^k \coloneqq \{f \in \mathcal{F}^N \mid \|f - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N, \delta)\}$.
$\mathcal{N}(\mathcal{F}^N, \epsilon)$	Covering number of $\mathcal{F}^N$ w.r.t. an ϵ-ball.
$\dim_E(\mathcal{F}^N, \epsilon)$	Eluder dimension of $\mathcal{F}^N$ w.r.t. ϵ.

B.2 Pseudocode of SF-LSVI and Technical Remarks

Algorithm 5 Statistical Functional Least Squares Value Iteration (SF-LSVI)

Input: failure probability $\delta \in (0,1)$ and the number of episodes K

for episode $k = 1, 2, \ldots, K$ **do**
  Receive initial state s_1^k
  Initialize $\psi_{1:N}(\bar{\eta}_{H+1}^k(\cdot)) \leftarrow \mathbf{0}^N$
  for step $h = H, H-1, \ldots, 1$ **do**
    $\mathcal{D}_h^k \leftarrow \left\{ s_{h'}^\tau, a_{h'}^\tau, \psi_{1:N}\left((\mathcal{B}_{r_{h'}^\tau})_{\#} \bar{\eta}_{h+1}^k(s_{h'+1}^\tau) \right) \right\}_{(\tau,h') \in [k-1] \times [H]}$
    // Data collection
    $\tilde{f}_{h,\bar{\eta}}^k \leftarrow \arg\min_{f \in \mathcal{F}^N} \|f\|_{\mathcal{D}_h^k}$
    // Distribution estimation
    $b_h^k(\cdot,\cdot) \leftarrow w^{(1)}((\mathcal{F}^N)_h^k, \cdot, \cdot)$
    $Q_h^k(\cdot,\cdot) \leftarrow \min\{(\tilde{f}_{h,\bar{\eta}}^k)^{(1)}(\cdot,\cdot) + b_h^k(\cdot,\cdot), H\}$
    $\pi_h^k(\cdot) = \arg\max_{a \in \mathcal{A}} Q_h^k(\cdot, a)$, $V_h^k(\cdot) = Q_h^k(\cdot, \pi_h^k(\cdot))$
    // Optimistic planning
    $\psi_1\left(\eta_h^k(\cdot,\cdot)\right) \leftarrow Q_h^k(\cdot,\cdot)$
    $\psi_{2:N}\left(\eta_h^k(\cdot,\cdot)\right) \leftarrow \left(\min\{(\tilde{f}_{h,\bar{\eta}}^k)^{(n)}(\cdot,\cdot), H\} \right)_{n \in [2:N]}$
    $\psi_1\left(\bar{\eta}_h^k(\cdot)\right) \leftarrow V_h^k(\cdot)$
    $\psi_{2:N}\left(\bar{\eta}_h^k(\cdot)\right) \leftarrow \psi_{1:N}\left(\eta_h^k(\cdot, \pi_h^k(\cdot))\right)_{n \in [2:N]}$
  for $h = 1, 2, \ldots, H$ **do**
    Take action $a_h^k \leftarrow \pi_h^k(s_h^k)$
    Observe reward $r_h^k(s_h^k, a_h^k)$ and get next state s_{h+1}^k.

Remark B.2.1. For an optimistic planning, we define the bonus function as the width function $b_h^k(s,a) := w_h^k((\mathcal{F}^N)_h^k, s, a)$ where $(\mathcal{F}^N)_h^k$ denotes a confidence region at step h, episode k. When $\mathcal{F}$ is a linear function class, the width function can be evaluated by simply computing the maximal distance of weight vector. For a general function class $\mathcal{F}$, computing the width function requires to solve a set-constrained optimization problem, which is known as NP-hard [33]. However, a width function is computed simply for optimistic exploration, and approximation errors are known to have a small effect on regret [1].

B.3 Related Work and Discussion

B.3.1 Technical Clarifications on Linearity Assumption in Existing Results

Bellman Closedness and Linearity. Rowland et al. [81] proved that quantile functional is not Bellman closed by providing a specific counterexample. However, their discussion based on counterexamples can be generalized as it assumes that the sketch Bellman operator for the quantile functional needs to be linear.

They consider an discounted MDP with initial state s_0 with single action a, which transits to one of two terminal states s_1, s_2 with equal probability. Letting no reward at state s_0, $\texttt{Unif}([0,1])$ at state s_1, and $\texttt{Unif}([1/K, 1 + 1/K])$ at state s_2, the return distribution at state s_0 is computed as mixture $\frac{1}{2}\texttt{Unif}([0,\gamma]) + \frac{1}{2}\texttt{Unif}([\gamma/K, \gamma+\gamma/K])$. Then the $\frac{1}{2K}-$quantile at state s_0 is $\frac{\gamma}{K}$. They proposed a counterexample where each quantile distribution of state s_1, s_2 is represented as $\frac{1}{K}\sum_{k=1}^{K}\delta_{\frac{2k-1}{K}}$ and $\frac{1}{K}\sum_{k=1}^{K}\delta_{\frac{2k+1}{K}}$ respectively, the $\frac{1}{2K}-$quantile of state s_0 is $\psi_{q_{2K}}\left(\frac{1}{2K}\sum_{k=1}^{K}\delta_{\frac{\gamma(2k-1)}{K}} + \delta_{\frac{\gamma(2k+1)}{K}}\right) = \frac{3\gamma}{2K}$. However, this example does not consider that the mixture of quantiles is not a quantile of the mixture distribution (i.e., $\psi_q(\lambda\eta_1 + (1-\lambda)\eta_2) \neq \lambda\psi_q(\eta_1) + (1-\lambda)\psi_q(\eta_2)$), due to the nonlinearity of the quantile functional. Therefore, this does not present a valid counterexample to prove that quantile functionals are not Bellman closed.

Bellman Optimizability and Linearity. Marthe et al. [63] proposed the notion of Bellman optimizable statistical functional which redefine the Bellman update by planning with respect to statistical functionals rather than expected returns. They proved that W_1-continuous Bellman Optimizable statistical functionals are characterized by exponential utilities $\frac{1}{\lambda}\log\mathbb{E}_{Z\sim\eta}[\exp(\lambda Z)]$. However, their proof requires some technical clarification regarding the assumption that such statistical functionals are linear.

To illustrate, they define a statistical functional ψ_f and consider two probability distributions $\eta_1 = \frac{1}{2}(\delta_0 + \delta_h)$ and $\eta_2 = \delta_{\phi(h)}$ where $\phi(h) = f^{-1}\Big(\frac{1}{2}(f(0) + f(h))\Big)$. Using the translation property, they lead $\psi_f(\eta_1) = \psi_f(\eta_2)$ to $\frac{1}{2}(f(x) + f(x+h)) = f(x+\phi(h))$ for all $x \in \mathbb{R}$. However, this equality $\psi_f\Big(\frac{1}{2}(\delta_x + \delta_{x+h})\Big) = \frac{1}{2}(f(x) + f(x+h))$ holds only if ψ_f is linear, which is not necessarily a valid assumption for all statistical functionals.

B.3.2 Existence of Nonlinear Bellman Closed Sketch.

The previous two examples may not have considered the possibility that the sketch Bellman operator might not necessarily be linear. However, some statistical functionals are Bellman-closed even if they are nonlinear, so it is open question whether there is a nonlinear sketch Bellman operator that makes the quantile functional Bellman-closed. In this section, we present examples of maximum and minimum functionals that are Bellman-closed, despite being nonlinear.

In a nutshell, consider the maximum of return distribution at state s_1, s_2 is $\gamma, \gamma + \gamma/K$ respectively. Beyond linearity, the maximum of return distribution at state s_0 can be computed by taking the maximum of these values;

$$\max(\max(\bar{\eta}(s_1)), \max(\bar{\eta}(s_2))) = \max(\gamma, \gamma + \gamma/K) = \gamma + \gamma/K$$

which produces the desired result. This implies the existence of a nonlinear sketch that is Bellman closed. More precisely, by defining $\max_{s' \sim \mathbb{P}(\cdot|s,a)}$ and $\min_{s' \sim \mathbb{P}(\cdot|s,a)}$ as the maximum and minimum of the sampled sketch $\psi\big((\mathcal{B}_r)_{\#}\bar{\eta}(s')\big)$ with the distribution $\mathbb{P}(\cdot|s,a)$, we can derive the sketch Bellman operator for maximum and minimum functionals as follows;

$$\mathcal{T}_{\psi_{\max}}\Big(\psi_{\max}(\bar{\eta}(s))\Big) = \max_{s'\sim\mathbb{P}(\cdot|s,a)} \psi_{\max}\big((\mathcal{B}_r)_{\#}\bar{\eta}(s')\big) = \max_{s'\sim\mathbb{P}(\cdot|s,a)} \Big(r + \psi_{\max}\big(\bar{\eta}(s')\big)\Big)$$
$$\mathcal{T}_{\psi_{\min}}\Big(\psi_{\min}(\bar{\eta}(s))\Big) = \min_{s'\sim\mathbb{P}(\cdot|s,a)} \psi_{\min}\big((\mathcal{B}_r)_{\#}\bar{\eta}(s')\big) = \min_{s'\sim\mathbb{P}(\cdot|s,a)} \Big(r + \psi_{\min}\big(\bar{\eta}(s')\big)\Big).$$

B.3.3 Non-existence of sketch Bellman operator for quantile functional

In this section, we prove that quantile functional cannot be Bellman closed under any additional sketch. First we introduce the definition of *mixture-consistent*, which is the property that the sketch of a mixture can be computed using only the sketch of the distribution of each component.

Definition B.3.1 (mixture-consistent)**.** A sketch ψ is **mixture-consistent** if for any $\nu \in [0,1]$ and any distributions $\eta_1, \eta_2 \in \mathscr{P}_\psi(\mathbb{R})$, there exists a corresponding function h_ψ such that

$$\psi(\nu\eta_1 + (1-\nu)\eta_2) = h_\psi\Big(\psi(\eta_1), \psi(\eta_2), \nu\Big).$$

Next, we will provide some examples of determining whether a sketch is mixture-consistent or not.

Example 1. Every moment or exponential polynomial functional is mixture-consistent.

Proof. For any $n \in [N]$ and $\lambda \in \mathbb{C}$,

$$\begin{aligned}&\mathbb{E}_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z^n \exp(\lambda Z)] \\ &\quad = \nu\mathbb{E}_{Z\sim\eta_1}[Z^n\exp(\lambda Z)] + (1-\nu)\mathbb{E}_{Z\sim\eta_2}[Z^n\exp(\lambda Z)].\end{aligned}$$

■

Example 2. Variance functional is not mixture-consistent.

Proof. Let $\nu = \frac{1}{2}$ and Z, Y be the random variables where $Z \sim \frac{1}{2}\delta_0 + \frac{1}{2}\delta_2$ and $Y \sim \frac{1}{2}\delta_k + \frac{1}{2}\delta_{k+2}$. Then, $\mathrm{Var}(Z) = \mathrm{Var}(Y) = 1$. While RHS is constant for any k, LHS is not a constant for any k, i.e.,

$$\mathrm{Var}_{X\sim\frac{1}{2}(\frac{1}{2}\delta_0+\frac{1}{2}\delta_2)+\frac{1}{2}(\frac{1}{2}\delta_k+\frac{1}{2}\delta_{k+2})}(X) = \frac{1}{4}(k^2+5).$$

■

While variance functional is not mixture consistent by itself, it can be mixture consistent with another statistical functional, the mean.

Example 3. Variance functional is mixture-consistent under mean functional.

Proof. Notice that mean functional is mixture-consistent. We need to show that variance functional is mixture-consistent under mean functional.

$$\begin{aligned}
&\mathrm{Var}_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z] \\
&= \mathbb{E}_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z^2] - (\mathbb{E}_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z])^2 \\
&= \nu\mathbb{E}_{Z\sim\eta_1}[Z^2] + (1-\nu)\mathbb{E}_{Z\sim\eta_2}[Z^2] - (\nu\mathbb{E}_{Z\sim\eta_1}[Z] + (1-\nu)\mathbb{E}_{Z\sim\eta_2}[Z])^2 \\
&= \nu(\mathrm{Var}_{Z\sim\eta_1}[Z] + (\mathbb{E}_{Z\sim\eta_1}[Z])^2) + (1-\nu)(\mathrm{Var}_{Z\sim\eta_2}[Z] + (\mathbb{E}_{Z\sim\eta_2}[Z])^2) \\
&\quad - (\nu\mathbb{E}_{Z\sim\eta_1}[Z] + (1-\nu)\mathbb{E}_{Z\sim\eta_2}[Z])^2.
\end{aligned}$$

■

This means that to determine whether it is mixture-consistent or not, we should check it on a per-sketch basis, rather than on a per-statistical functional basis.

Example 4. Maximum and minimum functional are both mixture-consistent.

Proof.

$$\max_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z] = \max(\max_{Z\sim\eta_1}[Z], \max_{Z\sim\eta_2}[Z])$$

and

$$\min_{Z\sim\nu\eta_1+(1-\nu)\eta_2}[Z] = \min(\min_{Z\sim\eta_1}[Z], \min_{Z\sim\eta_2}[Z])$$

■

Since maximum and minimum functionals are mixture consistent, we can construct a nonlinear sketch bellman operator like the one in section B.3.2. This

is possible because there is a nonlinear function h_ψ that ensures the sketch is closed under mixture.

Before demonstrating that a quantile sketch cannot be mixture consistent under any additional sketch, we will first illustrate with the example of a median functional that is not mixture consistent.

Example 5. Median sketch is not mixture-consistent.

Proof. Let $\nu = \frac{1}{2}$ and Z, Y be the random variables where $Z \sim 0.2\delta_0 + 0.8\delta_1$ and $Y \sim 0.6\delta_0 + 0.4\delta_k$ for some $0 < k < 1$. Then $\psi_{\text{med}}(Z) = 1$ and $\psi_{\text{med}}(Y) = 0$. However,

$$\text{med}_{X=\frac{Z+Y}{2}}[X] = \psi_{\text{med}}(0.4\delta_0 + 0.2\delta_k + 0.4\delta_1) = k$$

which is dependent in k. ■

Lemma B.3.2. *Quantile sketch cannot be mixture-consistent, under any additional sketch.*

Proof. For a given integer $N > 0$ and a quantile level $\alpha \in (0, 1)$, let $\nu = \frac{1}{2}$ and a random variable $Y \sim p_{y_0}\delta_0 + p_{y_1}\delta_{y_1} + \cdots + p_{y_N}\delta_{y_N} (0 < y_1 < \cdots < y_N < 1)$ where $p_{y_0} > \alpha$ so that $\psi_{\alpha-\text{quantile}}[Y] = 0$. Consider another random variable $Z \sim p_{z_0}\delta_0 + p_{z_1}\delta_1$ where $p_{z_0} < \alpha$ so that $\psi_{\alpha-\text{quantile}}[Z] = 1$. Then the $\alpha-$quantile of the mixture $X = \frac{Y+Z}{2}$ is

$$\psi_{\alpha-\text{quantile}}[X] = y_n \text{ where } n = \min\left\{ n \leq N \middle| \frac{1}{2}\sum_{n'=0}^{n} p_{y_{n'}} + \frac{1}{2}p_{z_0} > \alpha \right\}.$$

Letting $p_{z_0} = 2\alpha - \sum_{n'=0}^{n} p_{y_{n'}}$, we can manipulate $\psi_{\alpha-\text{quantile}}[X]$ to be any value of y_n. Hence, $\psi_{\alpha-\text{quantile}}[X]$ is a function of all possible outcomes of Y.

If there exists a finite number of statistical functionals which make quantile sketch mixture-consistent, then such sketch would uniquely determine the distribution for any N. This results in a contradiction that infinite-dimensional distribution space can be represented by a finite number of statistical functional. ■

Lemma B.3.3. *If a sketch ψ is Bellman closed, then it is mixture-consistent.*

Proof. Consider an MDP where initial state s_0 has no reward and transits to two state s_1, s_2 with probability $\nu, 1-\nu$ and reward distribution $\bar{\eta}_1, \bar{\eta}_2$. Since ψ is Bellman closed, $\psi(\bar{\eta}(s_0))$ is a function of $\psi(\bar{\eta}(s_1))$ and $\psi(\bar{\eta}(s_2))$, (i.e., $\psi(\bar{\eta}(s_0)) = g_\psi(\psi(\bar{\eta}(s_1)), \psi(\bar{\eta}(s_2)))$ for some g_ψ). Since $\psi(\bar{\eta}(s_0)) = \psi(\nu\bar{\eta}(s_1) + (1-\nu)\bar{\eta}(s_2))$, it implies that ψ is mixture-consistent. ■

Combining the results of Lemma B.3.2 and Lemma B.3.3, we prove that a quantile sketch cannot be Bellman closed, no matter what additional sketches are provided.

B.4 Proof

Theorem (4.3.3). Quantile functional cannot be Bellman closed under any additional sketch.

Proof. See Lemma B.3.2 and Lemma B.3.3. ■

Lemma (4.3.5). Let $F_{\bar{\eta}}$ be a CDF of the probability distribution $\bar{\eta} \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$. Then a sketch is Bellman unbiased if and only if the sketch is a homogeneous of degree k, i.e., there exists some vector-valued function $h = h(x_1, \cdots, x_k) : \mathcal{X}^k \to \mathbb{R}^N$ such that

$$\psi(\bar{\eta}) = \int \cdots \int h(x_1, \cdots, x_k) dF_{\bar{\eta}}(x_1) \cdots dF_{\bar{\eta}}(x_k).$$

Proof. ($\Rightarrow$) Consider an two-stage MDP with a single action a, and an initial state s_0 which transits to one of terminal state $\{s_1, \cdots, s_K\}$ with transition kernel $\mathbb{P}(\cdot|s_0, a)$. Assume that the reward $r(s_0) = 0$. Then $\bar{\eta}(s_0) = \sum_{k=1}^{K} \mathbb{P}(s_k)\delta_{r(s_k)}$. Note that $s'_1, \cdots, s'_k$ are independent and identically distributed random variable in distribution $\mathbb{P}(\cdot|s, a)$.

$$\begin{aligned}
&\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_0,a)} \left[\phi_\psi \left(\psi\Big((\mathcal{B}_r)_\# \bar{\eta}(s'_1) \Big), \cdots, \psi\Big((\mathcal{B}_r)_\# \bar{\eta}(s'_k) \Big) \right) \right] \\
&\qquad = \psi_{1:N}\Big((\mathcal{B}_r)_\# \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_0,a)}[\bar{\eta}(s')] \Big) \\
&\implies \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_0,a)} \left[\phi_\psi \left(\psi\Big(\delta_{r(s'_1)} \Big), \cdots, \psi\Big(\delta_{r(s'_k)} \Big) \right) \right] = \psi\Big(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_0,a)}[\delta_{r(s')}] \Big) \\
&\implies \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s_0,a)} \left[\phi_\psi\Big(g(s'_1), \cdots, g(s'_k) \Big) \right] = \psi\Big(\bar{\eta}(s_0) \Big) \\
&\implies \int \cdots \int h(s'_1, \cdots, s'_k) dF_{\bar{\eta}}(s'_1) \cdots dF_{\bar{\eta}}(s'_k) = \psi\Big(\bar{\eta}(s_0) \Big).
\end{aligned}$$

($\Leftarrow$)

$$\begin{aligned}
&\psi\Big((\mathcal{B}_r)_\# \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')] \Big) \\
&= \int \cdots \int h(x_1, \cdots, x_k) dF_{(\mathcal{B}_r)_\# \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')]}(x_1), \cdots, dF_{(\mathcal{B}_r)_\# \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)}[\bar{\eta}(s')](x_k)} \\
&= \int \cdots \int h(x_1 + r, \cdots, x_k + r) d\Big(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} F_{\bar{\eta}(s')}(x_1) \Big), \cdots, d\Big(\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} F_{\bar{\eta}(s')}(x_k) \Big) \\
&= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\int \cdots \int h(x_1 + r, \cdots, x_k + r) dF_{\bar{\eta}(s')}(x_1) \cdots dF_{\bar{\eta}(s')}(x_k) \right] \\
&= \mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[\psi\Big((\mathcal{B}_r)_\# [\bar{\eta}(s')] \Big) \right]
\end{aligned}$$

■

Theorem (4.3.6). The only finite statistical functionals that are Bellman unbiased and closed are given by the collections of $\psi_1, \ldots, \psi_N$ where its linear span $\{\sum_{n=0}^{N} \alpha_n \psi_n |\ \alpha_n \in \mathbb{R}\ , \forall N\}$ is equal to the set of exponential polynomial functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l \exp(\lambda Z)] |\ l = 0, 1, \ldots, L, \lambda \in \mathbb{R}\}$, where ψ_0 is the constant functional equal to 1. In discount setting, it is equal to the linear span of the set of moment functionals $\{\eta \to \mathbb{E}_{Z\sim\eta}[Z^l] |\ l = 0, 1, \ldots, L\}$ for some $L \leq N$.

Proof. Our proof is mainly based on the proof techniques of Rowland et al. [81] and we describe in an extended form. Since their proof also considers the discounted setting, we will define $\mathcal{B}_{r,\gamma}(x) = r + \gamma x$ for discount factor $\gamma \in [0, 1)$. By assumption of Bellman closedness, $\psi_n\Big((\mathcal{B}_{r,\gamma})_{\#}\bar{\eta}(s')\Big)$ will be written as $g(r, \gamma, \psi_{1:N}(\bar{\eta}(s'))$ for some g. By assumption of Bellman unbiasedness and Lemma 4.3.5, both $\psi_{1:N}(\bar{\eta}(s'))$ and $\psi_n\Big((\mathcal{B}_{r,\gamma})_{\#}\bar{\eta}(s')\Big)$ are affine as functions of the distribution $\bar{\eta}(s')$,

$$
\begin{aligned}
&\psi_{1:N}(\alpha\bar{\eta}_1(s') + (1-\alpha)\bar{\eta}_2(s')) \\
&= \mathbb{E}_{Z_i\sim\alpha\bar{\eta}_1(s')+(1-\alpha)\bar{\eta}_2(s')}[h_{1:N}(\bar{Z}_1, \cdots, \bar{Z}_k)] \\
&= \alpha\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}_1(s')}[h_{1:N}(\bar{Z}_1, \cdots, \bar{Z}_k)] + (1-\alpha)\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}_2(s')}[h_{1:N}(\bar{Z}_1, \cdots, \bar{Z}_k)] \\
&= \alpha\psi_{1:N}(\bar{\eta}_1(s')) + (1-\alpha)\psi_{1:N}(\bar{\eta}_2(s'))
\end{aligned}
$$

and

$$
\begin{aligned}
&\psi_n\Big((\mathcal{B}_{r,\gamma})_{\#}(\alpha\bar{\eta}_1(s') + (1-\alpha)\bar{\eta}_2(s'))\Big) \\
&= \mathbb{E}_{Z_i\sim\alpha\bar{\eta}_1(s')+(1-\alpha)\bar{\eta}_2(s')}[h_n(r + \gamma\bar{Z}_1, \cdots, r + \gamma\bar{Z}_k)] \\
&= \alpha\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}_1(s')}[h_n(r + \gamma\bar{Z}_1, \cdots, r + \gamma\bar{Z}_k)] \\
&\qquad + (1-\alpha)\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}_2(s')}[h_n(r + \gamma\bar{Z}_1, \cdots, r + \gamma\bar{Z}_k)] \\
&= \alpha\psi_n\Big((\mathcal{B}_{r,\gamma})_{\#}\bar{\eta}_1(s')\Big) + (1-\alpha)\psi_n\Big((\mathcal{B}_{r,\gamma})_{\#}\bar{\eta}_2(s')\Big)
\end{aligned}
$$

Therefore, $g(r, \gamma, \cdot)$ is also affine on the convex codomain of $\psi_{1:N}$. Thus, we have

$$
\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}}[\phi_{\psi_n}(r+\gamma\bar{Z}_1, \cdots, r+\gamma\bar{Z}_k)] = a_0(r, \gamma) + \sum_{n'=1}^{N} a_{n'}(r, \gamma)\mathbb{E}_{\bar{Z}_i\sim\bar{\eta}}[\phi_{\psi_{n'}}(\bar{Z}_1, \cdots, \bar{Z}_k)]
$$

for some function $a_{0:N} : \mathbb{R} \times [0,1] \to \mathbb{R}$. By taking $\bar{\eta}(s') = \delta_x$, we obtain

$$\phi_{\psi_n}(r + \gamma x, \cdots, r + \gamma x) = a_0(r, \gamma) + \sum_{n'=1}^{N} a_{n'}(r, \gamma)\phi_{\psi_{n'}}(x, \cdots, x).$$

According to Engert [35], for any translation invariant finite-dimensional space is spanned by a set of function of the form

$$\{x \mapsto x^l \exp(\lambda_j x) |\ j \in [J], 0 \leq l \leq L\}$$

for some finite subset $\{\lambda_1, \cdots, \lambda_J\}$ of $\mathbb{C}$. Hence, each function $x \mapsto \phi_{\psi_n}(x, \cdots, x)$ is expressed as linear combination of exponential polynomial functions. In addition, the linear combination of ϕ_{ψ_n} should be closed under composition with for any discount factor $\gamma \in [0,1]$, all λ_j should be zero. Hence, the linear combination of $\phi_{\psi_1}, \cdots, \phi_{\psi_N}$ must be equal to the span of $\{x \mapsto x^l |\ 0 \leq l \leq L\}$ for some $L \in \mathbb{N}$.

■

Lemma (4.5.3). Consider a fixed $k \in [K]$ and a fixed $h \in [H]$. Let $\mathcal{Z}_h^k = \{(s_h^\tau, a_h^\tau)\}_{\tau \in [k-1]}$ and $\mathcal{D}_{h,\bar{\eta}}^k = \left\{\left(s_h^\tau, a_h^\tau, \psi_{1:N}\left((\mathcal{B}_{r_{h'}}^\tau)_{\#}\bar{\eta}(s_{h'+1}^\tau)\right)\right)\right\}_{\tau \in [k-1]}$ for any $\bar{\eta} : \mathcal{S} \to \mathscr{P}([0,H])$. Define $\tilde{f}_{h,\bar{\eta}}^k = \arg\min_{f \in \mathcal{F}^N} \|f\|_{\mathcal{D}_{h,\bar{\eta}}^k}^2$. For any $\bar{\eta}$ and $\delta \in (0,1)$, there is an event $\mathcal{E}(\bar{\eta}, \delta)$ such that conditioned on $\mathcal{E}(\bar{\eta}, \delta)$, with probability at least $1 - \delta$, for any $\bar{\eta}' : \mathcal{S} \to \mathscr{P}([0,H])$ with $\|\psi_{1:N}(\bar{\eta}') - \psi_{1:N}(\bar{\eta})\|_{\infty,1} \leq 1/T$ or $\sum_{n=1}^N \|\psi_n(\bar{\eta}') - \psi_n(\bar{\eta})\|_\infty \leq 1/T$, we have

$$\begin{aligned}
&\left\|\tilde{f}_{h,\bar{\eta}'}(\cdot,\cdot) - \psi_{1:N}\left((\mathcal{B}_{r(\cdot,\cdot)})_{\#}[\mathbb{P}\bar{\eta}'](\cdot,\cdot)\right)\right\|_{\mathcal{Z}_h^k} \\
&\quad \leq c'\left(N^{\frac{1}{2}} H \sqrt{\log(1/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T)}\right)
\end{aligned}$$

for some constant $c' > 0$.

Proof. Define the sketch of target $f_{\bar{\eta}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$,

$$f_{\bar{\eta}}(\cdot,\cdot) \coloneqq \psi_{1:N}\left((\mathcal{B}_{r(\cdot,\cdot)})_{\#}[\mathbb{P}\bar{\eta}](\cdot,\cdot)\right)$$

for all $i \in [N]$.

For any $f \in \mathcal{F}$,

$$
\begin{aligned}
&\|f\|^2_{\mathcal{D}^k_{h,\bar{\eta}'}} - \|f_{\bar{\eta}'}\|^2_{\mathcal{D}^k_{h,\bar{\eta}'}} \\
&= \sum_{n=1}^{N} \sum_{s_h^\tau, a_h^\tau \in \mathcal{Z}^k_{h,\bar{\eta}'}} \Big(f^{(n)}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}'(s_{h+1}^\tau)\Big)\Big)^2 \\
&\qquad\qquad - \Big(f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}'(s_{h+1}^\tau)\Big)\Big)^2 \\
&= \sum_{n=1}^{N} \sum_{s_h^\tau, a_h^\tau \in \mathcal{Z}^k_{h,\bar{\eta}'}} (f^{(n)}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau))^2 \\
&\quad + 2(f^{(n)}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau))\Big(f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}'(s_{h+1}^\tau)\Big)\Big) \\
&\geq \|f - f_{\bar{\eta}'}\|^2_{\mathcal{Z}^k_h} - 4\sum_{n=1}^{N} \|f^{(n)}_{\bar{\eta}} - f^{(n)}_{\bar{\eta}'}\|_\infty (H+1)|\mathcal{Z}^k_h| \\
&+ \sum_{n=1}^{N} \sum_{s_h^\tau, a_h^\tau \in \mathcal{Z}^k_{h,\bar{\eta}'}} \Big[\underbrace{2(f^{(n)}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau))\Big(f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}(s_{h+1}^\tau)\Big)\Big)}_{\chi_h^\tau(f^{(n)})}\Big] \\
&\geq \|f - f_{\bar{\eta}'}\|^2_{\mathcal{Z}^k_h} - 4N(H+1) - \Big|\sum_{n=1}^{N} \sum_{s_h^\tau, a_h^\tau \in \mathcal{Z}^k_{h,\bar{\eta}'}} \chi_h^\tau(f^{(n)})\Big|.
\end{aligned}
$$

For the first inequality, we change the second term from $\bar{\eta}'$ to $\bar{\eta}$ which are the ϵ-covers. Notice that $AC - BC' \geq -|AC - BC'| \geq -|(A-B)C| - |(A-B)C'| \geq -2|A-B|\max(C, C')|$.

$$
\begin{aligned}
&(f^{(n)}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau))\Big(f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}'(s_{h+1}^\tau)\Big)\Big) \\
&\quad - (f^{(n)}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau))\Big(f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}(s_{h+1}^\tau)\Big)\Big) \\
&\geq -2\|f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau)\| \\
&\times \max\Big(\Big|f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}'(s_{h+1}^\tau)\Big)\Big|, \Big|f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar{\eta}(s_{h+1}^\tau)\Big)\Big|\Big) \\
&\geq -2\|f^{(n)}_{\bar{\eta}'}(s_h^\tau, a_h^\tau) - f^{(n)}_{\bar{\eta}}(s_h^\tau, a_h^\tau)\|(H+1)
\end{aligned}
$$

For the second inequality, consider $\bar{\eta}' : \mathcal{S} \to \mathscr{P}([0, H])$ with $\sum_{n=1}^{N} \|\psi_n(\bar{\eta}') -$

$\psi_n(\bar{\eta})\|_\infty \le 1/T$. We have

$$\begin{aligned}\|f^{(n)}_{\bar{\eta}} - f^{(n)}_{\bar{\eta}'}\|_\infty &= \max_{s,a}\Big|\sum_{n'=1}^{n} H^{n'}[\psi_{n'}([\mathbb{P}\bar{\eta}](s,a)) - \psi_{n'}([\mathbb{P}\bar{\eta}'](s,a))]r^{n-n'}/H^{n-1}\Big| \\ &\le \sum_{n'=1}^{n}\max_{s'}\Big|\psi_{n'}(\bar{\eta}(s')) - \psi_{n'}(\bar{\eta}'(s'))\Big| \\ &\le 1/T.\end{aligned}$$

Defining $\mathbb{F}^k_h$ as the filtration induced by the sequence $\{(s^\tau_{h'}, a^\tau_{h'})\}_{\tau,h'\in[k-1]\times[H]} \cup \{(s^k_1,a^k_1),(s^k_2,a^k_2),\ldots,(s^k_h,a^k_h)\}$, notice that

$$\begin{aligned}&\mathbb{E}\Big[\sum_{n=1}^{N}\chi^\tau_h(f^{(n)})\Big|\mathbb{F}^\tau_h\Big] \\ &= \sum_{n=1}^{N} 2(f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h))(f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h) - \mathbb{E}\Big[\psi_n\Big((\mathcal{B}_{r^\tau_h})_\#\bar{\eta}(s^\tau_{h+1})\Big)\Big|\mathbb{F}^\tau_h\Big]) \\ &= \sum_{n=1}^{N} 2(f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h))(f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h) - \mathbb{E}_{s^\tau_{h+1}\sim\mathbb{P}_h(\cdot|s^\tau_h,a^\tau_h)}\Big[\psi_n\Big((\mathcal{B}_{r^\tau_h})_\#\bar{\eta}(s^\tau_{h+1})\Big)\Big]) \\ &= \sum_{n=1}^{N} 2(f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h))(f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h) - \psi_n\Big((\mathcal{B}_{r^\tau_h})_\#\mathbb{E}_{s^\tau_{h+1}\sim\mathbb{P}_h(\cdot|s^\tau_h,a^\tau_h)}[\bar{\eta}(s^\tau_{h+1})]\Big)) \\ &= 0\end{aligned}$$

and

$$\begin{aligned}&\Big|\sum_{n=1}^{N}\chi^\tau_h(f^{(n)})\Big| \\ &= \Big|\sum_{n=1}^{N} 2(f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h))(f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h) - \psi_n\Big((\mathcal{B}_{r^\tau_h})_\#\bar{\eta}(s^\tau_{h+1})\Big))\Big| \\ &\le \max_{n\in[N]}\Big\{2(f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h) - \psi_n\Big((\mathcal{B}_{r^\tau_h})_\#\bar{\eta}(s^\tau_{h+1})\Big))\Big\}\sum_{n=1}^{N}\Big|f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h)\Big| \\ &\le 2(H+1)\sum_{n=1}^{N}\Big|f^{(n)}(s^\tau_h,a^\tau_h) - f^{(n)}_{\bar{\eta}}(s^\tau_h,a^\tau_h)\Big|\end{aligned}$$

In third equality, we emphasize that only Bellman unbiased sketch can derive the martingale difference sequence which induce the concentration result. Since

every moment functional is commutable with mixing operation, the transformation ϕ_{ψ_n} in Definition 4.3.4 is identity for all $n \in [N]$. Hence, we choose the sketch as moment which already knows ϕ_ψ.

By Azuma-Hoeffding inequality,

$$\begin{aligned}
&\mathbb{P}\Big[\Big|\sum_{(\tau,h)\in[k-1]\times[H]}\sum_{n=1}^{N}\chi_h^\tau(f^{(n)})\Big| \geq \epsilon\Big] \\
&\leq 2\exp\Big(-\frac{\epsilon^2}{2(2(H+1))^2\sum_{(\tau,h)\in[k-1]\times[H]}\Big(\sum_{n=1}^{N}|f^{(n)}-f_{\bar{\eta}}^{(n)}|\Big)^2}\Big) \\
&\leq 2\exp\Big(-\frac{\epsilon^2}{2(2(H+1))^2\sum_{(\tau,h)\in[k-1]\times[H]}\Big(N\sum_{n=1}^{N}|f^{(n)}-f_{\bar{\eta}}^{(n)}|^2\Big)}\Big) \\
&= 2\exp\Big(-\frac{\epsilon^2}{2N(2(H+1))^2\|f-f_{\bar{\eta}}\|_{\mathcal{Z}_h^k}^2}\Big)
\end{aligned}$$

where the second inequality follows from the Cauchy-Schwartz inequality.

We set

$$\epsilon = \sqrt{8N(H+1)^2\|f-f_{\bar{\eta}}\|_{\mathcal{Z}_h^k}^2\log\Big(\frac{\mathcal{N}(\mathcal{F}^N,1/T)}{\delta}\Big)}$$

With union bound for all $f \in \mathcal{C}(\mathcal{F}^N, 1/T)$, with probability at least $1-\delta$,

$$\Big|\sum_{(\tau,h)\in[k-1]\times[H]}\sum_{n=1}^{N}\chi_h^\tau(f^{(n)})\Big| \leq c'N^{\frac{1}{2}}(H+1)\|f-f_{\bar{\eta}}\|_{\mathcal{Z}_h^k}\sqrt{\log\Big(\frac{\mathcal{N}(\mathcal{F}^N,1/T)}{\delta}\Big)}$$

for some constant $c' > 0$.

For all $f \in \mathcal{F}^N$, there exists $g \in \mathcal{C}(\mathcal{F}^N, 1/T)$, such that $\|f-g\|_{\infty,1} \leq 1/T$

or $\sum_{n=1}^{N} \|f^{(n)} - g^{(n)}\|_\infty \le 1/T$ for all $n \in [N]$,

$$
\begin{aligned}
&\Big| \sum_{(\tau,h)\in[k-1]\times[H]} \sum_{n=1}^{N} \chi_h^\tau(f^{(n)}) \Big| \\
&\le \Big| \sum_{(\tau,h)\in[k-1]\times[H]} \sum_{n=1}^{N} \chi_h^\tau(g^{(n)}) \Big| + 2(H+1)|\mathcal{Z}_h^k| \sum_{n=1}^{N} \frac{1}{T} \\
&\le c' N^{\frac{1}{2}} (H+1) \|g - f_{\bar\eta}\|_{\mathcal{Z}_h^k} \sqrt{\log\Big(\frac{\mathcal{N}(\mathcal{F}^N, 1/T)}{\delta}\Big)} + 2N(H+1) \\
&\le c' N^{\frac{1}{2}} (H+1) (\|f - f_{\bar\eta}\|_{\mathcal{Z}_h^k} + 1) \sqrt{\log\Big(\frac{\mathcal{N}(\mathcal{F}^N, 1/T)}{\delta}\Big)} + 2N(H+1) \\
&\le c' N^{\frac{1}{2}} (H+1) (\|f - f_{\bar\eta'}\|_{\mathcal{Z}_h^k} + 2) \sqrt{\log\Big(\frac{\mathcal{N}(\mathcal{F}^N, 1/T)}{\delta}\Big)} + 2N(H+1)
\end{aligned}
$$

where the third inequality follows from,

$$
\begin{aligned}
\|f - g\|_{\mathcal{Z}_h^k}^2 &\le \sum_{n=1}^{N} \sum_{(\tau,h)\in[k-1]\times[H]} |f^{(n)}(s_h^\tau, a_h^\tau) - g^{(n)}(s_h^\tau, a_h^\tau)|^2 \\
&\le NT\Big(\frac{1}{T}\Big)^2 \\
&\le 1.
\end{aligned}
$$

Recall that $\tilde f_{h,\eta'}^k = \arg\min_{f\in\mathcal{F}} \|f\|_{\mathcal{D}_{h,\eta'}^k}^2$. We have $\|\tilde f_{h,\eta'}^k\|_{\mathcal{D}_{h,\eta'}^k}^2 - \|f_{\bar\eta'}\|_{\mathcal{D}_{h,\eta'}^k}^2 \le 0$, which implies,

$$
\begin{aligned}
0 &\ge \|\tilde f_{h,\bar\eta'}^k\|_{\mathcal{D}_{h,\bar\eta'}^k}^2 - \|f_{\bar\eta'}\|_{\mathcal{D}_{h,\bar\eta'}^k}^2 \\
&= \|\tilde f_{h,\bar\eta'}^k - f_{\bar\eta'}\|_{\mathcal{Z}_h^k}^2 + 2\sum_{n=1}^{N} \sum_{(\tau,h)\in[k-1]\times[H]} \Big[\Big((\tilde f_{h,\bar\eta'}^k)^{(n)}(s_h^\tau, a_h^\tau) - f_{\bar\eta'}^{(n)}(s_h^\tau, a_h^\tau)\Big) \\
&\qquad \Big((f_{\bar\eta'}^{(n)}(s_h^\tau, a_h^\tau) - \psi_n\Big((\mathcal{B}_{r_h^\tau})_{\#}\bar\eta'(s_{h+1}^\tau)\Big)\Big)\Big] \\
&\ge \|\tilde f_{h,\bar\eta'}^k - f_{\bar\eta'}\|_{\mathcal{Z}_h^k}^2 - c' N^{\frac{1}{2}} (H+1)(\|\hat f_{h,\bar\eta'}^k - f_{\bar\eta'}\|_{\mathcal{Z}_h^k} + 2) \\
&\qquad \sqrt{\log(2/\delta) + \log\mathcal{N}(\mathcal{F}^N, 1/T)} - 6N(H+1).
\end{aligned}
$$

Recall that if $x^2 - 2ax - b \le 0$ holds for constant $a, b > 0$, then $x \le a + \sqrt{a^2 + b} \le c' \cdot a$ for some constant $c' > 0$.

Hence,

$$\|\tilde{f}^k_{h,\eta'} - f_{\bar{\eta}'}\|_{\mathcal{Z}^k_h} \le c'(N^{\frac{1}{2}} H \sqrt{\log(1/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T)})$$

for some constant $c' > 0$. ■

Lemma (4.5.4). Let $(\mathcal{F}^N)^k_h = \{f \in \mathcal{F}^N | \|f - \tilde{f}^k_{h,\bar{\eta}}\|^2_{\mathcal{Z}^k_h} \le \beta(\mathcal{F}^N, \delta)\}$, where

$$\beta(\mathcal{F}^N, \delta) \ge c' \cdot NH^2(\log(T/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T))$$

for some constant $c' > 0$. Then with probability at least $1 - \delta/2$, for all $k, h \in [K] \times [H]$, we have

$$\psi_n\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h \bar{\eta}^k_{h+1}](\cdot,\cdot)\Big) \in (\mathcal{F}^N)^k_h$$

Proof. For all $(k, h) \in [K] \times [H]$,

$$\mathbf{S} \coloneqq \begin{cases} \left\{ \left(\min\{f^{(1)}(\cdot,\cdot) + b^k_{h+1}(\cdot,\cdot), H\} \right) \middle| \ f \in \mathcal{C}(\mathcal{F}^N, 1/T) \right\} \cup \left\{0\right\} & n = 1 \\ \left\{ \left(\min\{f^{(n)}(\cdot,\cdot), H\} \right) \middle| \ f \in \mathcal{C}(\mathcal{F}^N, 1/T) \right\} \cup \left\{0\right\} & 2 \le n \le N \end{cases}$$

is a $(1/T)$-cover of $\psi_{1:N}(\eta^k_{h+1}(\cdot,\cdot))$ where

$$\psi_{1:N}(\eta^k_{h+1}(\cdot,\cdot)) = \begin{cases} \min\{(f^k_{h+1})^{(1)}(\cdot,\cdot) + b^k_{h+1}(\cdot,\cdot), H\} & n = 1 \text{ and } h < H \\ \min\{(f^k_{h+1})^{(n)}(\cdot,\cdot), H\} & 2 \le n \le N \text{ and } h < H \\ \mathbf{0}^N & h = H \end{cases},$$

i.e., there exists $\psi_{1:N}(\eta) \in \mathbf{S}$ such that $\|\psi_{1:N}(\eta) - \psi_{1:N}(\eta^k_{h+1})\|_{\infty,1} \le 1/T$. This implies

$$\bar{\mathbf{S}} \coloneqq \left\{ \psi_{1:N}\Big(\eta(\cdot, \arg\max_{a \in \mathcal{A}} \psi_1(\eta(\cdot, a)))\Big) \mid \psi_{1:N}(\eta) \in \mathbf{S} \right\}$$

is a $(1/T)$-cover of $\psi_{1:N}(\bar{\eta}^k_{h+1})$ with $\log(|\bar{\mathbf{S}}|) \le \log \mathcal{N}(\mathcal{F}^N, 1/T)$.

For each $\psi_{1:N}(\bar{\eta}) \in \bar{\mathbf{S}}$, let $\mathcal{E}(\bar{\eta}, \delta/2|\bar{\mathbf{S}}|T)$ be the event defined in Lemma 4.5.3. By union bound for all $\psi_{1:N}(\bar{\eta}) \in \bar{\mathbf{S}}$, we have $\Pr[\bigcap_{\psi_{1:N}(\bar{\eta}) \in \bar{\mathbf{S}}} \mathcal{E}(\bar{\eta}, \delta/2|\bar{\mathbf{S}}|T)] \ge 1 - \delta/2T$.

Let $\psi_{1:N}(\bar{\eta}) \in \bar{\mathbf{S}}$ such that $\|\psi_{1:N}(\bar{\eta}) - \psi_{1:N}(\bar{\eta}^k_{h+1})\|_{\infty,1} \le 1/T$. Conditioned on $\bigcap_{s_N(\bar{\eta}) \in \bar{\mathbf{S}}} \mathcal{E}(\bar{\eta}, \delta/2|\bar{\mathbf{S}}|T)$ and by Lemma 4.5.3, we have

$$\begin{aligned} &\left\| \tilde{f}^k_{h,\bar{\eta}}(\cdot,\cdot) - \psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h \bar{\eta}^k_{h+1}](\cdot,\cdot)\Big) \right\|^2_{\mathcal{Z}^k_h} \\ &\qquad \le c'\Big(NH^2(\log(T/\delta) + \log \mathcal{N}(\mathcal{F}^N, 1/T))\Big) \end{aligned}$$

for some constant $c' > 0$.

By union bound for all $(k,h) \in [K] \times [H]$, we have the target sketch $\psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h \bar{\eta}^k_{h+1}](\cdot,\cdot)\Big) \in (\mathcal{F}^N)^k_h$ with probability $1-\delta/2$. ■

Lemma B.4.1. *Let $Q^k_h(s,a) := \min\{H, \tilde{f}^k_h(s,a) + b^k_h(s,a)\}$ for some bonus function $b^k_h(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. If $b^k_h(s,a) \geq w^{(1)}((\mathcal{F}^N)^k_h, s, a)$, then with probability at least $1-\delta/2$,*

$$Q^*_h(s,a) \leq Q^k_h(s,a) \text{ and } V^*_h(s) \leq V^k_h(s)$$

for all $(k,h) \in [K] \times [H]$, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Proof. We use induction on h from $h = H$ to 1 to prove the statement. Let $\mathcal{E}$ be the event that for $(k,h) \in [K] \times [H]$, $\psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h \bar{\eta}^k_{h+1}](\cdot,\cdot)\Big) \in (\mathcal{F}^N)^k_h$. By Lemma 4.5.4, $\Pr|\mathcal{E}| \geq 1-\delta/2$. In the rest of the proof, we condition on $\mathcal{E}$.

When $h = H+1$, the desired inequality holds as $Q^*_{H+1}(s,a) = V^*_{H+1}(s) = Q^k_{H+1}(s,a) = V^k_{H+1}(s) = 0$. Now, assume that $Q^*_{h+1}(s,a) \leq Q^k_{h+1}(s,a)$ and $V^*_{h+1}(s) \leq V^k_{h+1}(s)$ for some $h \in [H]$. Then, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}
Q^*_h(s,a) &= \min\{H, r_h(s,a) + [\mathbb{P}_h V^*_{h+1}](s,a)\} \\
&\leq \min\{H, r_h(s,a) + [\mathbb{P}_h V^k_{h+1}](s,a)\} \\
&\leq \min\{H, \tilde{f}^k_h(s,a) + w^{(1)}(\mathcal{F}^k_h, s, a)\} \\
&= \min\{H, Q^k_h(s,a) - b^k_h(s,a) + w^{(1)}(\mathcal{F}^k_h, s, a)\} \\
&\leq Q^k_h(s,a)
\end{aligned}$$

■

Lemma B.4.2 (Regret decomposition)**.** *With probability at least $1-\delta/4$, we have*

$$Reg(K) \leq \sum_{k=1}^{K}\sum_{h=1}^{H}(2b^k_h(s^k_h, a^k_h) + \xi^k_h),$$

where $\xi^k_h = [\mathbb{P}_h(V^k_{h+1} - V^{\pi^k}_{h+1})](s^k_h, a^k_h) - (V^k_{h+1}(s^k_{h+1}) - V^{\pi^k}_{h+1}(s^k_{h+1}))$ is a martingale difference sequence with respect to the filtration $\mathbb{F}^k_h$ induced by the history $\mathbb{H}^k_h$.

Proof. We condition on the above event $\mathcal{E}$ in the rest of the proof. For all $(k,h) \in [K] \times [H]$, we have

$$\left\| \tilde{f}^k_{h,\bar{\eta}}(\cdot,\cdot) - \psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_{\#}[\mathbb{P}_h \bar{\eta}^k_{h+1}](\cdot,\cdot)\Big) \right\|^2_{\mathcal{Z}^k_h} \leq \beta(\mathcal{F}^N, \delta).$$

Recall that $(\mathcal{F}^N)_h^k = \{f \in \mathcal{F}^N \mid \|f - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N, \delta)\}$ is the confidence region. Since $\psi_{1:N}\Big((\mathcal{B}_{r_h(\cdot,\cdot)})_\#[\mathbb{P}_h\bar{\eta}_{h+1}^k](\cdot,\cdot)\Big) \in (\mathcal{F}^N)_h^k$, then by the definition of width function $w^{(1)}(\mathcal{F}_h^k, s, a)$, for $(k,h) \in [K] \times [H]$, we have

$$\begin{aligned} w^{(1)}(\mathcal{F}_h^k, s, a) &\geq \Big|\psi_1\Big((\mathcal{B}_{r_h(s,a)})_\#[\mathbb{P}_h\bar{\eta}_{h+1}^k](s,a)\Big) - (\tilde{f}_{h,\bar{\eta}}^k)^{(1)}(s,a)\Big| \\ &= \Big|r_h(s,a) + [\mathbb{P}_h V_{h+1}^k](s,a) - (\tilde{f}_{h,\bar{\eta}}^k)^{(1)}(s,a)\Big|. \end{aligned}$$

Recall that $Q_h^*(\cdot,\cdot) \leq Q_h^k(\cdot,\cdot)$.

$$\begin{aligned} \mathrm{Reg}(K) &= \sum_{k=1}^K V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k) \\ &\leq \sum_{k=1}^K V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k) \\ &= \sum_{k=1}^K Q_1^k(s_1^k, a_1^k) - Q_1^{\pi^k}(s_1^k, a_1^k) \\ &= \sum_{k=1}^K Q_1^k(s_1^k, a_1^k) - (r_1(s_1^k, a_1^k) + [\mathbb{P}_1 V_2^k](s_1^k, a_1^k)) + (r_1(s_1^k, a_1^k) \\ &\qquad + [\mathbb{P}_1 V_2^k](s_1^k, a_1^k)) - Q_1^{\pi^k}(s_1^k, a_1^k) \\ &\leq \sum_{k=1}^K w^{(1)}((\mathcal{F}^N)_1^k, s_1^k, a_1^k) + b_1^k(s_1^k, a_1^k) + [\mathbb{P}_1(V_2^k - V_2^{\pi^k})](s_1^k, a_1^k) \\ &\leq \sum_{k=1}^K w^{(1)}((\mathcal{F}^N)_1^k, s_1^k, a_1^k) + b_1^k(s_1^k, a_1^k) + (V_2^k(s_2^k) - V_2^{\pi^k}(s_2^k)) + \xi_1^k \\ &\quad \vdots \\ &\leq \sum_{k=1}^K \sum_{h=1}^H (w^{(1)}((\mathcal{F}^N)_h^k, s_h^k, a_h^k) + b_h^k(s_h^k, a_h^k) + \xi_h^k) \\ &\leq \sum_{k=1}^K \sum_{h=1}^H (2b_h^k(s_h^k, a_h^k) + \xi_h^k) \end{aligned}$$

■

It remains to bound $\sum_{k=1}^K \sum_{h=1}^H b_h^k(s_h^k, a_h^k)$, for which we will exploit fact that $\mathcal{F}^N$ has bounded eluder dimension.

Lemma B.4.3. *If* $b_h^k(s,a) \geq w^{(1)}((\mathcal{F}^N)_h^k, s, a)$ *for all* $(s,a) \in \mathcal{S} \times \mathcal{A}$ *and* $k \in [K]$ *where*

$$(\mathcal{F}^N)_h^k = \{f \in \mathcal{F}^N |\ \|f - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N, \delta)\},$$

then

$$\sum_{k=1}^{K}\sum_{h=1}^{H} \mathbf{1}\{b_h^k(s_h^k, a_h^k) > \epsilon\} \leq \left(\frac{4\beta(\mathcal{F}^N,\delta)}{\epsilon^2} + 1\right) dim_E(\mathcal{F}^N, \epsilon)$$

for some constant $c > 0$.

Proof. We first want to show that for any sequence $\{(s_1,a_1),\ldots,(s_\kappa,a_\kappa)\} \subseteq \mathcal{S}\times\mathcal{A}$, there exists $j \in [\kappa]$ such that (s_j,a_j) is ϵ-dependent on at least $L = \lceil(\kappa-1)/\dim_E(\mathcal{F}^N,\epsilon)\rceil$ disjoint subsequences in $\{(s_1,a_1),\ldots,(s_{j-1},a_{j-1})\}$ with respect to $\mathcal{F}^N$. We demonstrate this by using the following procedure. Start with L disjoint subsequences of $\{(s_1,a_1),\ldots,(s_{j-1},a_{j-1})\}$, $\mathcal{B}_1,\mathcal{B}_2,\ldots,\mathcal{B}_L$, which are initially empty. For each j, if (s_j,a_j) is ϵ-dependent on every $\mathcal{B}_1,\ldots,\mathcal{B}_L$, we achieve our goal so we stop the process. Else, we choose $i \in [L]$ such that (s_j,a_j) is ϵ-independent on $\mathcal{B}_i$ and update $\mathcal{B}_i \leftarrow \mathcal{B}_i \cup \{(s_j,a_j)\}$, $j \leftarrow j+1$. Since every element of $\mathcal{B}_i$ is ϵ-independent on its predecessors, $|\mathcal{B}_i|$ cannot get bigger than $\dim_E(\mathcal{F}^N,\epsilon)$ at any point in this process. Therefore, the process stops at most step $j = L\dim_E(\mathcal{F}^N,\epsilon)+1 \leq \kappa$.

Now we want to show that if for some $j \in [\kappa]$ such that $b_h^k(s_j,a_j) > \epsilon$, then (s_j,a_j) is ϵ-dependent on at most $4\beta(\mathcal{F}^N,\delta)/\epsilon^2$ disjoint subsequences in $\{(s_1,a_1),\ldots,(s_{j-1},a_{j-1})\}$ with respect to $\mathcal{F}^N$. If $b_h^k(s_j,a_j) > \epsilon$ and (s_j,a_j) is ϵ-dependent on a subsequence of $\{(s_1',a_1'),\ldots,(s_l',a_l')\} \subseteq \{(s_1,a_1),\ldots,(s_\kappa,a_\kappa)\}$, it implies that there exists $f,g \in \mathcal{F}^N$ with $\|f - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N,\delta)$ and $\|g - \tilde{f}_{h,\bar{\eta}}^k\|_{\mathcal{Z}_h^k}^2 \leq \beta(\mathcal{F}^N,\delta)$ such that $f^{(1)}(s_t',a_t') - g^{(1)}(s_t',a_t') \geq \epsilon$. By triangle inequality, $\|f-g\|_{\mathcal{Z}_h^k}^2 \leq 4\beta(\mathcal{F}^N,\delta)$. On the other hand, if (s_j,a_j) is ϵ-dependent on L disjoint subsequences in $\{(s_1,a_1),\ldots,(s_\kappa,a_\kappa)\}$, then

$$4\beta(\mathcal{F}^N,\delta) \geq \|f-g\|_{\mathcal{Z}^k}^2 \geq \|f^{(1)} - g^{(1)}\|_{\mathcal{Z}^k}^2 \geq L\epsilon^2$$

resulting in $L \leq 4\beta(\mathcal{F}^N,\delta)/\epsilon^2$. Therefore, we have $(\kappa/\dim_E(\mathcal{F}^N,\epsilon)) - 1 \leq 4\beta(\mathcal{F}^N,\delta)/\epsilon^2$ which results in

$$\kappa \leq \left(\frac{4\beta(\mathcal{F},\delta)}{\epsilon^2} + 1\right)\dim_E(\mathcal{F}^N,\epsilon)$$

■

Lemma B.4.4 (Refined version of Lemma 10 in Wang et al. [101]). *If $b_h^k(s,a) \geq w^{(1)}((\mathcal{F}^N)_h^k, s, a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $k \in [K]$, then*

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq H\dim_E(\mathcal{F}^N, 1/T).$$

Proof. We first sort the sequence $\{b_h^k(s_h^k, a_h^k)\}_{(k,h)\in[K]\times[H]}$ in a decreasing order and denote it by $\{e_1, \ldots, e_T\}(e_1 \geq e_2 \geq \cdots \geq e_T)$. By Lemma B.4.3, for any constant $M > 0$ and $e_t \geq 1/\sqrt{MT}$, we have

$$t \leq \Big(\frac{4\beta(\mathcal{F}^N, \delta)}{Me_t^2} + 1\Big)\dim_E(\mathcal{F}^N, \sqrt{M}e_t) \leq \Big(\frac{4\beta(\mathcal{F}^N, \delta)}{Me_t^2} + 1\Big)\dim_E(\mathcal{F}^N, 1/T)$$

which implies

$$e_t \leq \Big(\frac{t}{\dim_E(\mathcal{F}^N, 1/T)} - 1\Big)^{-1/2}\sqrt{\frac{4\beta(\mathcal{F}^N, \delta)}{M}},$$

for $t \geq \dim_E(\mathcal{F}^N, 1/T)$. Since we have $e_t \leq H$,

$$\begin{aligned}
\sum_{t=1}^{T} e_t &= \sum_{t=1}^{T} e_t\mathbf{1}\{e_t < 1/\sqrt{MT}\} + \sum_{t=1}^{T} e_t\mathbf{1}\{e_t \geq 1/\sqrt{MT}, t < \dim_E(\mathcal{F}^N, 1/T)\} \\
&\quad + \sum_{t=1}^{T} e_t\mathbf{1}\{e_t \geq 1/\sqrt{MT}, t \geq \dim_E(\mathcal{F}^N, 1/T)\} \\
&\leq \frac{1}{\sqrt{M}} + H\dim_E(\mathcal{F}^N, 1/T) \\
&\quad + \sum_{\dim_E(\mathcal{F}^N,1/T)\leq t\leq T} \Big(\frac{t}{\dim_E(\mathcal{F}^N, 1/T)} - 1\Big)^{-1/2}\sqrt{\frac{4\beta(\mathcal{F}^N, \delta)}{M}} \\
&\leq \frac{1}{\sqrt{M}} + H\dim_E(\mathcal{F}^N, 1/T) \\
&\quad + 2\Big(\frac{T}{\dim_E(\mathcal{F}^N, 1/T)} - 1\Big)^{1/2}\dim_E(\mathcal{F}^N, 1/T)\sqrt{\frac{4\beta(\mathcal{F}^N, \delta)}{M}} \\
&= \frac{1}{\sqrt{M}} + H\dim_E(\mathcal{F}^N, 1/T) + \sqrt{16 \cdot \dim_E(\mathcal{F}^N, 1/T) \cdot T \cdot \beta(\mathcal{F}^N, \delta)/M}.
\end{aligned}$$

Taking $M \to \infty$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) \leq H\dim_E(\mathcal{F}^N, 1/T).$$

■

Theorem (4.5.5). Under Assumption 4.3.7, with probability at least $1-\delta$, `SF-LSVI` achieves a regret bound of

$$\mathrm{Reg}(K) \leq 2H\dim_E(\mathcal{F}^N, 1/T) + 4H\sqrt{KH\log(2/\delta)}.$$

Proof. Recall that $\xi_h^k = [\mathbb{P}_h(V_{h+1}^k - V_{h+1}^{\pi^k})](s_h^k, a_h^k) - (V_{h+1}^k(s_{h+1}^k) - V_{h+1}^{\pi^k}(s_{h+1}^k))$ is a martingale difference sequence where $\mathbb{E}[\xi_h^k|\mathbb{F}_h^k] = 0$ and $|\xi_h^k| \leq 2H$. By Azuma-Hoeffding's inequality, with probability at least $1-\delta/2$,

$$\sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k \leq 4H\sqrt{KH\log(2/\delta)}.$$

Conditioning on the above event and Lemma B.4.4, we have

$$\begin{aligned}\mathrm{Reg}(K) &\leq 2\sum_{k=1}^{K}\sum_{h=1}^{H} b_h^k(s_h^k, a_h^k) + \sum_{k=1}^{K}\sum_{h=1}^{H}\xi_h^k \\ &\leq 2H\dim_E(\mathcal{F}^N, 1/T) + 4H\sqrt{KH\log(2/\delta)}\end{aligned}$$

■

Appendix C

Appendix of Chapter 5

C.1 Main Proof

Lemma (5.2.2). (Structural Condition for α-optimality) A reward function and a soft optimal Q-function where $\pi^*(\cdot|s)$ is α-optimal have a one-to-one correspondence with a state-dependent function $\beta : \mathcal{S} \to \mathbb{R}$ as follows,

$$\mathcal{R}_{\alpha,\pi^*} = \{r_*(s,a) = \alpha \log \pi^*(a|s) + \beta(s) - \gamma \mathbb{E}_{\mathbb{P}}[\beta(s')],\ \forall s, a|\ \alpha \geq 0, \beta : \mathcal{S} \to \mathbb{R}\}$$
$$\mathcal{Q}^{\pi^*}_{\alpha,\pi^*} = \{Q^{\pi^*}_*(s,a) = \alpha \log \pi^*(a|s) + \beta(s),\ \forall s, a|\ \alpha \geq 0, \beta : \mathcal{S} \to \mathbb{R}\}$$

Proof. (π^* is α-optimal $\iff Q^{\pi^*}_*(s,a) = \alpha \log \pi^*(a|s) + \beta(s)$ for some $\beta : \mathcal{S} \to \mathbb{R}$.)

Remark that the policy π^* is α-optimal, if and only if there exists the optimal soft Q-function satisfies the following relation:

$$\pi^*(a|s) = \exp\Big(\frac{1}{\alpha}(Q^{\pi^*}(s,a) - V^{\pi^*}(s))\Big),$$
$$V^{\pi^*}(s) = \alpha \log \int_{a \in \mathcal{A}} \exp\Big(\frac{1}{\alpha} Q^{\pi^*}(s,a) da\Big).$$

Since V^{π^*} is merely a partition function, letting $X(s,a) = \exp\Big(\frac{1}{\alpha} Q^{\pi^*}(s,a)\Big)$,

we can derive

$$\begin{aligned}
&\pi^*(a|s) = \frac{X(s,a)}{\int_{a\in\mathcal{A}} X(s,a)da} \\
&\iff X(s,a) = d(s)\pi^*(a|s) \text{ for some } d:\mathcal{S}\to\mathbb{R} \\
&\iff Q^{\pi^*}(s,a) = \alpha\log\pi^*(a|s) + \beta(s) \text{ for some } \beta:\mathcal{S}\to\mathbb{R},
\end{aligned}$$

where β is defined as $\beta(s) = \log d(s)$.

Using the soft Bellman equation, consider a reward function for any state-dependent function $\beta:\mathcal{S}\to\mathbb{R}$ and substitute the expression of $Q^{\pi^*}(s,a)$. Then, we have:

$$\begin{aligned}
r(s,a) &\coloneqq Q^{\pi^*}(s,a) - \gamma\mathbb{E}_{\mathbb{P}}[V^{\pi^*}(s')] \\
&= \alpha\Big(\log\pi^*(a|s) + \beta(s) - \gamma\mathbb{E}_{\mathbb{P}}[\beta(s')]\Big)
\end{aligned}$$

where π^* and $\mathbb{P}$ are given. By the definition of optimal soft Q-function, we recursively substitute the soft Bellman equation and sum over timesteps:

$$\begin{aligned}
Q^{\pi^*}(s,a) &= r(s,a) + \mathbb{E}_{\tau\sim\mathbb{P}^{\pi^*}}\Big[\sum_{t>0}\gamma^t(r(s_t,a_t) + \alpha\mathcal{H}^{\pi^*}(\cdot|s_t))\Big|s_0=s,a_0=a\Big] \\
&= \alpha\log\pi^*(a|s) + \beta(s) - \gamma\mathbb{E}_{\mathbb{P}}[\beta(s_1)] \\
&\quad + \mathbb{E}_{\tau\sim\mathbb{P}^{\pi^*}}\Big[\sum_{t>0}\gamma^t(\beta(s_t) - \gamma\mathbb{E}_{\mathbb{P}}[\beta(s_{t+1})])\Big|s_0=s,a_0=a\Big] \\
&= \alpha\log\pi^*(a|s) + \beta(s) - \gamma^2\mathbb{E}_{\tau\sim\mathbb{P}^{\pi^*}}[\beta(s_2)] \\
&\quad + \mathbb{E}_{\tau\sim\mathbb{P}^{\pi^*}}\Big[\sum_{t>1}\gamma^t(\beta(s_t) - \gamma\mathbb{E}_{\mathbb{P}}[\beta(s_{t+1})])\Big|s_0=s,a_0=a\Big] \\
&\qquad\vdots \\
&= \alpha\log\pi^*(a|s) + \beta(s).
\end{aligned}$$

■

Lemma (5.2.3). (Unique Fixed Point of Soft Bellman π-operator) Let π^* is α-optimal. For a given policy π and Q-function $Q_A^\pi \in \mathcal{Q}^\pi$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, define the Bellman π-operator $\mathcal{T}_*^\pi : \mathcal{Q}^\pi \to \mathcal{Q}^\pi$ where

$$\begin{aligned}
\mathcal{T}_*^\pi Q_A^\pi(s,a) \coloneqq Q_*^{\pi^*}(s,a) - \gamma\mathbb{E}_{\mathbb{P}}\Big[&\alpha\Big(\mathcal{H}^{\pi^*}(\cdot|s') - \mathcal{H}^{\pi}(\cdot|s')\Big) \\
&+\mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a')] - \mathbb{E}_{\pi}[Q_A^\pi(s',a')]\Big].
\end{aligned}$$

Then, $\mathcal{T}_*^\pi$ has a unique fixed point Q_*^π.

Proof. Consider $Q_A^\pi, Q_B^\pi \in \mathcal{Q}^\pi$. Then

$$\begin{aligned}\sup_{s,a}\left|\mathcal{T}_*^\pi Q_A^\pi(s,a) - \mathcal{T}_*^\pi Q_B^\pi(s,a)\right| &\le \sup_{s,a}\left|\gamma\mathbb{E}_{\mathbb{P}}\Big[\mathbb{E}_\pi[Q_A^\pi(s',a')] - \mathbb{E}_\pi[Q_B^\pi(s',a')]\Big]\right| \\ &= \gamma\sup_{s',a'}\left|Q_A^\pi(s',a') - Q_B^\pi(s',a')\right|\end{aligned}$$

Hence, $\mathcal{T}_*^\pi$ is a γ-contraction for any $Q_A^\pi, Q_B^\pi \in \mathcal{Q}^\pi$. Since $\mathcal{Q}^\pi$ is a complete metric space, by using Banach fixed point theorem, $\mathcal{T}_*^\pi$ has a unique fixed point.

Notice that $Q_*^{\pi^*}$ and Q_*^π satisfies soft Bellman equation respectively, i.e.,

$$\begin{aligned}Q_*^{\pi^*}(s,a) &= \mathbb{E}_{\mathbb{P}}\Big[r_*(s,a) + \gamma\mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a') + \alpha\mathcal{H}^{\pi^*}(\cdot|s')]\Big], \\ Q_*^\pi(s,a) &= \mathbb{E}_{\mathbb{P}}\Big[r_*(s,a) + \gamma\mathbb{E}_\pi[Q_*^\pi(s',a') + \alpha\mathcal{H}^\pi(\cdot|s')]\Big] \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}.\end{aligned}$$

Then,

$$\begin{aligned}&\mathcal{T}_*^\pi Q_*^\pi(s,a) \\ &= Q_*^{\pi^*}(s,a) - \gamma\mathbb{E}_{\mathbb{P}}\left[\alpha\Big(\mathcal{H}^{\pi^*}(\cdot|s') - \mathcal{H}^\pi(\cdot|s')\Big) + \mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a')] - \mathbb{E}_\pi[Q_*^\pi(s',a')]\right] \\ &= Q_*^{\pi^*}(s,a) - \gamma\mathbb{E}_{\mathbb{P}}\left[\alpha\mathcal{H}^{\pi^*}(\cdot|s') + \mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a')] - \Big(\alpha\mathcal{H}^\pi(\cdot|s') + \mathbb{E}_\pi[Q_*^\pi(s',a')]\Big)\right] \\ &= \mathbb{E}_{\mathbb{P}}\Big[r_*(s,a) + \gamma\mathbb{E}_\pi[Q_*^\pi(s',a') + \alpha\mathcal{H}^\pi(\cdot|s')]\Big] \\ &= Q_*^\pi(s,a) \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}.\end{aligned}$$

Hence, Q_*^π is a unique fixed point of $\mathcal{T}_*^\pi$. ■

Theorem (5.2.4). If a policy π^* is α-optimal, then for any policy π,

$$Q_*^{\pi^*}(s,a) - Q_*^\pi(s,a) = \alpha\bar{D}_{KL}(\pi||\pi^*;s,a)$$

where the **sequential forward KL divergence** is defined as

$$\bar{D}_{KL}(\pi||\pi';s,a) \coloneqq \mathbb{E}_{\tau\sim\mathbb{P}^\pi_{s,a}}\Big[\sum_{l>0}\gamma^l D_{KL}(\pi(\cdot|s_l)||\pi'(\cdot|s_l))\Big].$$

Here, $\mathbb{P}^\pi_{s,a}$ is the distribution of trajectories $\tau = (s_0,a_0,\cdots,s_l,a_l,\cdots)$ generated by policy π and the transition $\mathbb{P}$, starting at $(s_0,a_0) = (s,a)$.

Proof. Let $\tilde{Q}_*^\pi(s,a) = Q_*^{\pi^*}(s,a) - \alpha\sum_{t>0}\gamma^t\mathbb{E}_{\tau\sim\mathbb{P}^\pi}\Big[D_{KL}(\pi(\cdot|s_t)||\pi^*(\cdot|s_t))\Big|s_0 =$

$s, a_0 = a\Big]$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Then

$$
\begin{aligned}
&\mathcal{T}_*^{\pi} \tilde{Q}_*^{\pi}(s,a) \\
&= Q_*^{\pi^*}(s,a) - \gamma \mathbb{E}_{\mathbb{P}} \Big[\alpha \Big(\mathcal{H}^{\pi^*}(\cdot|s') - \mathcal{H}^{\pi}(\cdot|s') \Big) + \mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s',a')] - \mathbb{E}_{\pi}[\tilde{Q}_*^{\pi}(s',a')] \Big] \\
&= Q_*^{\pi^*}(s,a) - \gamma \mathbb{E}_{\mathbb{P}} \Bigg[\alpha \Big(\mathcal{H}^{\pi^*}(\cdot|s') - \mathcal{H}^{\pi}(\cdot|s') \Big) + \mathbb{E}_{\pi^*}[\alpha \log \pi^*(s',a') + \beta(s')] \\
&\quad - \mathbb{E}_{\pi} \Big[Q_*^{\pi^*}(s',a') - \alpha \sum_{t>0} \gamma^t \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \big[D_{KL}(\pi(\cdot|s_t) || \pi^*(\cdot|s_t)) \big] \Big| s_1 = s', a_1 = a' \Big] \Bigg] \\
&= Q_*^{\pi^*}(s,a) - \gamma \mathbb{E}_{\mathbb{P}} \Bigg[\beta(s') - \alpha \mathcal{H}^{\pi}(\cdot|s') - \mathbb{E}_{\pi}[Q_*^{\pi^*}(s',a')] \\
&\quad + \alpha \mathbb{E}_{\pi} \Big[\sum_{t>0} \gamma^t \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \big[D_{KL}(\pi(\cdot|s_t) || \pi^*(\cdot|s_t)) \big] \Big| s_1 = s', a_1 = a' \Big] \Bigg] \\
&= Q_*^{\pi^*}(s,a) - \alpha \gamma \mathbb{E}_{\mathbb{P}} \Big[D_{KL}(\pi(\cdot|s') || \pi^*(\cdot|s')) \Big] \\
&\quad - \alpha \sum_{t>1} \gamma^t \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \Big[D_{KL}(\pi(\cdot|s_t) || \pi^*(\cdot|s_t)) \Big| s_0 = s, a_0 = a \Big] \\
&= Q_*^{\pi^*}(s,a) - \alpha \sum_{t>0} \gamma^t \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \Big[D_{KL}(\pi(\cdot|s_t) || \pi^*(\cdot|s_t)) \Big| s_0 = s, a_0 = a \Big] \\
&= \tilde{Q}_*^{\pi}(s,a)
\end{aligned}
$$

which implies that $\tilde{Q}_*^{\pi}$ is a unique fixed point of $\mathcal{T}_*^{\pi}$. In Lemma 5.2.3, we observe that $\mathcal{T}_*^{\pi}$ has a unique fixed point Q_*^{π}. Hence,

$$
Q_*^{\pi}(s,a) = Q_*^{\pi^*}(s,a) - \alpha \sum_{t>0} \gamma^t \mathbb{E}_{\tau \sim \mathbb{P}^{\pi}} \Big[D_{KL}(\pi(\cdot|s_t) || \pi^*(\cdot|s_t)) \Big| s_0 = s, a_0 = a \Big]
$$

■

C.2 Further Theoretical Analysis & Discussion

C.2.1 Mathematical derivation of PPL framework

We recall the PPL model and objective:

$$
P_{\pi_\psi}^{(\pi^+, \pi^-)}[\zeta^+ \succ \zeta^-] = \sigma \Big(- \sum_{t \geq 0} \mathrm{Reg}_{\pi_\psi}^{\pi^+}(s_t^+, a_t^+) - \mathrm{Reg}_{\pi_\psi}^{\pi^-}(s_t^-, a_t^-) \Big),
$$

$$
\mathcal{L}_{\mathrm{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{(\zeta^+, \zeta^-, y, p) \sim \mathcal{D}} \Bigg[\log \sigma \Big(- \sum_{t \geq 0} \mathrm{Reg}_{\pi_\psi}^{\pi^+}(s_t^+, a_t^+) - \mathrm{Reg}_{\pi_\psi}^{\pi^-}(s_t^-, a_t^-) \Big) \Bigg]
$$

where

$$-\mathrm{Reg}^{\pi}_{\pi^*}(s_t, a_t) \coloneqq -(V_*^{\pi^*}(s_t) - Q_*^{\pi}(s_t, a_t)).$$

Here, a negative regret at (s_t, a_t) can be decomposed into two components:

$$-\mathrm{Reg}^{\pi}_{\pi^*}(s_t, a_t) = \alpha\Big(\underbrace{\log \pi^*(a_t|s_t)}_{\text{increase likelihood}} \underbrace{- \mathbb{E}_{\tau\sim\mathbb{P}^{\pi}_{s_t,a_t}}\Big[\sum_{l>0}\gamma^l D_{\mathrm{KL}}(\pi(\cdot|s_l)||\pi^*(\cdot|s_l))\Big]}_{\text{decrease sequential forward KL divergence}} \Big)$$

Proof. By the definition of regret,

$$\begin{aligned}
&- \mathrm{Reg}^{\pi}_{\pi^*}(s_t, a_t) \\
&\coloneqq -(V_*^{\pi^*}(s_t) - Q_*^{\pi}(s_t, a_t)). \\
&= -\Big(\mathbb{E}_{\pi^*}[Q_*^{\pi^*}(s_t, a) - \alpha \log \pi^*(\cdot|s_t)]\Big) + Q_*^{\pi^*}(s_t, a_t) - \alpha \bar{D}_{KL}(\pi||\pi^*; s_t, a_t) \\
&= \cancel{-\beta(s_t)} + \alpha \log \pi^*(a_t|s_t) + \cancel{\beta(s_t)} - \alpha \bar{D}_{KL}(\pi||\pi^*; s_t, a_t) \\
&= \alpha\Big(\log \pi^*(a_t|s_t) - \bar{D}_{KL}(\pi||\pi^*; s_t, a_t)\Big)
\end{aligned}$$

■

C.3 Variants of PPL and Baselines

BC: BC (Behavior Cloning) is the initial stage in RLHF, where the policy is trained to maximize the likelihood of the demonstrated actions given the corresponding states:

$$\mathcal{L}_{\mathrm{BC}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\zeta\sim\mathcal{D}}\Big[\sum_{t\geq 0} \log \pi_\psi(a_t|s_t)\Big]$$

SFT: SFT (Supervised Fine Tuning) is trained to maximize the likelihood of the demonstrated actions given the corresponding states in preferred segments:

$$\mathcal{L}_{\mathrm{SFT}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{\zeta^+\sim\mathcal{D}}\Big[\sum_{t\geq 0} \log \pi_\psi(a_t^+|s_t^+)\Big]$$

CPL: CPL [44] is our primary baseline, where the optimal advantage is defined as the score function:

$$S_{\mathrm{CPL}}(\pi_\psi; \zeta^+) - S_{\mathrm{CPL}}(\pi_\psi; \zeta^-) = \sum_{t\geq 0} \log \frac{\pi_\psi(a_t^+|s_t^+)}{\pi_\psi(a_t^-|s_t^-)}.$$

The objective is to minimize the following loss function:

$$\mathcal{L}_{\mathrm{CPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{(\zeta^+,\zeta^-)\sim\mathcal{D}}\Big[\log \sigma\big(S_{\mathrm{CPL}}(\pi_\psi; \zeta^+) - S_{\mathrm{CPL}}(\pi_\psi; \zeta^-)\big)\Big]$$

A key issue raised in CPL is assigning high weights to OOD actions while still maintaining the same optimal policy. This leads to extrapolation too much into unseen states, ultimately degrading performance. To mitigate this, an asymmetric regularizer is introduced:

$$\begin{aligned} S_{\mathrm{CPL}(\lambda)}(\pi_\psi; \zeta^+) - S_{\mathrm{CPL}(\lambda)}(\pi_\psi; \zeta^-) &= S_{\mathrm{CPL}}(\pi_\psi; \zeta^+) - \lambda S_{\mathrm{CPL}}(\pi_\psi; \zeta^-) \\ &= \sum_{t\geq 0} \log \frac{\pi_\psi(a_t^+|s_t^+)}{\pi_\psi(a_t^-|s_t^-)^\lambda} \end{aligned}$$

PPL: Based on policy deviation lemma in Theorem 5.2.4, PPL extends CPL by incorporating entropy regularization and KL divergence-based constraints, making preference learning more structured. The score function includes multiple terms:

$$
\begin{aligned}
&S_{\mathrm{PPL}}(\pi_\psi; \zeta^+, \pi^+) - S_{\mathrm{PPL}}(\pi_\psi; \zeta^-, \pi^-) \\
&= \sum_{t \geq 0} \Bigg[\log \frac{\pi_\psi(a_t^+ | s_t^+)}{\pi_\psi(a_t^- | s_t^-)} \\
&\quad + \frac{1}{L} \sum_{l=1}^{L} \Big(- D_{KL}(\pi^+(\cdot|s_{t+l}^+) || \pi_\psi(\cdot|s_{t+l}^+)) + D_{KL}(\pi^-(\cdot|s_{t+l}^-) || \pi_\psi(\cdot|s_{t+l}^-)) \Big) \Bigg],
\end{aligned}
$$

and the objective function is:

$$
\mathcal{L}_{\mathrm{PPL}}(\pi_\psi; \mathcal{D}) = -\mathbb{E}_{(\zeta^+, \zeta^-) \sim \mathcal{D}} \Big[\log \sigma \big(S_{\mathrm{PPL}}(\pi_\psi; \zeta^+) - S_{\mathrm{PPL}}(\pi_\psi; \zeta^-) \big) \Big]
$$

The score function of PPL with the same asymmetric regularizer as CPL is given by:

$$
\begin{aligned}
&S_{\mathrm{PPL}(\lambda)}(\pi_\psi; \zeta^+, \pi^+) - S_{\mathrm{PPL}(\lambda)}(\pi_\psi; \zeta^-, \pi^-) \\
&= S_{\mathrm{PPL}}(\pi_\psi; \zeta^+, \pi^+) - \lambda S_{\mathrm{PPL}}(\pi_\psi; \zeta^-, \pi^-) \\
&= \sum_{t \geq 0} \Bigg[\log \frac{\pi_\psi(a_t^+ | s_t^+)}{\pi_\psi(a_t^- | s_t^-)^\lambda} \\
&\quad + \frac{1}{L} \sum_{l=1}^{L} \Big(- D_{KL}(\pi^+(\cdot|s_{t+l}^+) || \pi_\psi(\cdot|s_{t+l}^+)) + \lambda D_{KL}(\pi^-(\cdot|s_{t+l}^-) || \pi_\psi(\cdot|s_{t+l}^-)) \Big) \Bigg],
\end{aligned}
$$

PPL-deterministic: If policy-label is unknown, we apply deterministic pseudo-labels by assuming that each segment was generated by a deterministic policy that executed the observed action.

$$
\begin{aligned}
&S_{\text{PPL-d}}(\pi_\psi; \zeta^+) - S_{\text{PPL-d}}(\pi_\psi; \zeta^-) \\
&= \sum_{t \geq 0} \left[\log \frac{\pi_\psi(a_t^+ | s_t^+)}{\pi_\psi(a_t^- | s_t^-)} + \frac{1}{L} \sum_{l=1}^{L} \log \frac{\pi_\psi(a_{t+l}^+ | s_{t+l}^+)}{\pi_\psi(a_{t+l}^- | s_{t+l}^-)} \right]
\end{aligned}
$$

C.4 Implementation Details

C.4.1 Hyperparameter Setting

Table C.1: Hyperparameter settings for offline implementation.

Hyperparameter	**State**
Total Training Steps	500k
Pre-training Steps (except `P-IQL`)	200k
Batch Size	96
Segment Size	64
Fixed log std	-1.5
Actor Dropout	0.0 (0.25 for `CPL` reproduce)
Architecture	[256, 256] MLP Gaussian

Table C.2: Hyperparameters for online implementation

Hyperparameter	**State**
Total Environment Steps	1m
Segment Size	32
Fixed log std	-1.0
Query Frequency(steps)	1000
Policy update Frequency(steps)	1000
Episode Length	250
Learning rates	3e-4
Temperature α	0.1
Asymmetric regularizer λ	1.0
BC weights	0
γ	1
Actor Dropout	0.0
Architecture	[256, 256] MLP Gaussian

Table C.3: Hyperparameters for PPL, CPL, SFT, and P-IQL

Hyperparameter	PPL	CPL	SFT	P-IQL
Learning rates	1e-4	1e-4	1e-4	1e-4
Temperature α	0.1	0.1	0.1	0.1
Asymmetric regularizer λ	0.5	0.5	-	-
BC weights	0	0	0	0
γ	1	1	1	1
Number of Parameters	76k	76k	76k	859k

C.4.2 MetaWorld Benchmark

Our experiments were conducted on six MetaWorld environments: `Bin-Picking`, `Button-Press`, `Door-Open`, `Drawer-Open`, `Plate-Slide`, and `Sweep-Into`.

Each task requires precise control of a robotic arm to interact with objects in a structured environment. The diverse task set includes object relocation, pushing, pulling, and fine-grained manipulation, making it a suitable testbed for reinforcement learning from preference-based feedback.

Each environment is designed with a handcrafted reward function tailored to its objective. Instead of human annotations, we trained a critic using SAC to assign labels. During our experiments, we observed that return did not always align well with success rates. In case of `Door-Open`, despite achieving the highest return, PPL exhibited a relatively low success rate. This implies that the environment allows reward exploitation due to the imprecise design of the reward function.

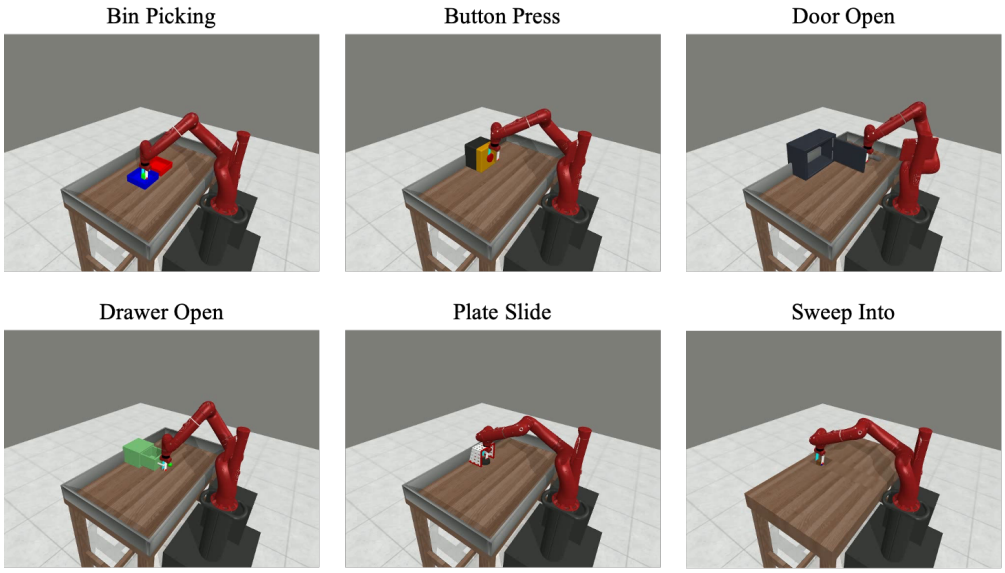


Figure C.1: Visualization of the MetaWorld Benchmark Tasks.

C.4.3 Reproducibility Check

For a fair comparison, we first verified the reproducibility of `CPL` using the Metaworld `State Dense` and `State Sparse` datasets provided by Hejna et al. [44] and evaluated the performance of `PPL` on these datasets. We used the official `CPL` implementation (`https://github.com/jhejna/cpl`) without modifications and ensured reproducibility by fixing the random seed ([`123,231,312,321`]). The figure below presents the PPL performance alongside the reproduced `CPL` results. The horizontal dashed line represents the scores reported in CPL, confirming the reproducibility of the algorithm. The vertical dashed line indicates the point where behavior cloning (`BC`) training stops.

In all environments except `Plate-Slide-v2`, the reproduced `CPL` performance closely matches the reported values, with deviations attributed to seed variability. Across the provided datasets, `PPL` exhibits comparable overall performance to `CPL`.

Table C.4: Success rates of all methods on six tasks from the MetaWorld across different datasets from Hejna et al. [44]. Each score is reported as the highest average performance across four seeds over a 200-episode evaluation window.

		Bin Picking	Button Press	Door Open	Drawer Open	Plate Slide	Sweep Into
State 2.5k Dense	CPL(Reported)	80.0 ± 2.5	24.5 ± 2.1	80.0 ± 6.8	83.6 ± 1.6	61.1 ± 3.0	70.4 ± 3.0
	CPL(Reproduced)	76.0 ± 4.1	24.9 ± 4.7	75.5 ± 6.0	87.6 ± 2.8	45.3 ± 10.4	74.5 ± 3.4
	PPL	77.7 ± 2.6	30.2 ± 7.8	76.7 ± 7.1	84.2 ± 2.4	41.7 ± 3.2	79.2 ± 5.5
State 20k Sparse	CPL(Reported)	83.2 ± 3.5	29.8 ± 1.8	77.9 ± 9.3	79.1 ± 5.0	56.4 ± 3.9	81.2 ± 1.6
	CPL(Reproduced)	69.1 ± 21.4	25.5 ± 5.3	74.4 ± 3.5	80.9 ± 4.5	41.1 ± 4.9	80.5 ± 2.8
	PPL	83.0 ± 3.7	25.4 ± 2.8	72.2 ± 1.7	79.0 ± 4.0	42.9 ± 1.6	76.0 ± 2.0

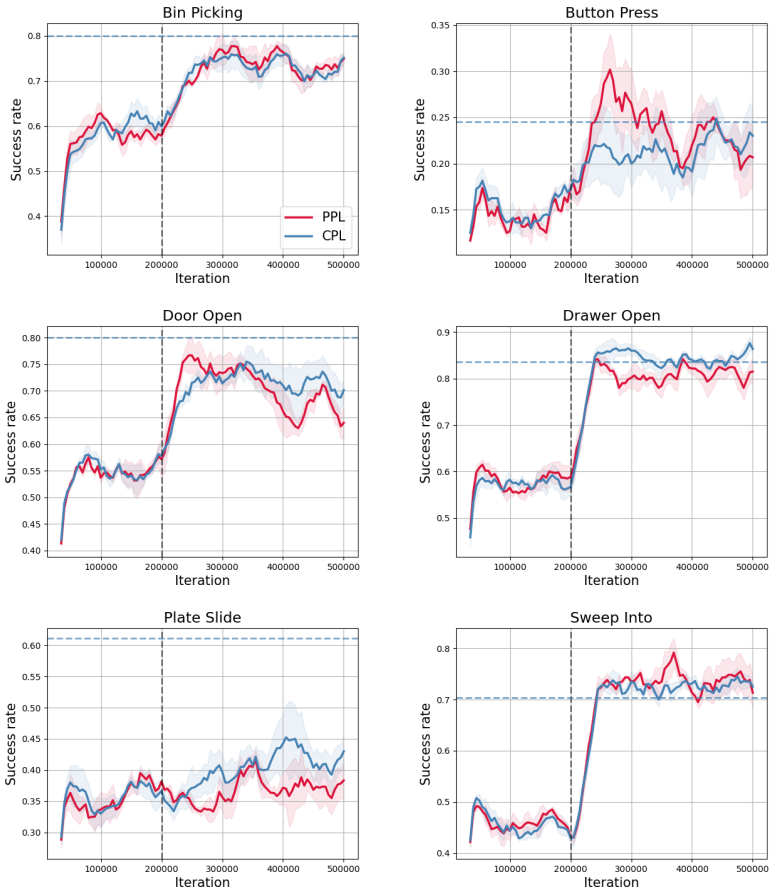


Figure C.2: Reproducibility check on `State Dense` dataset

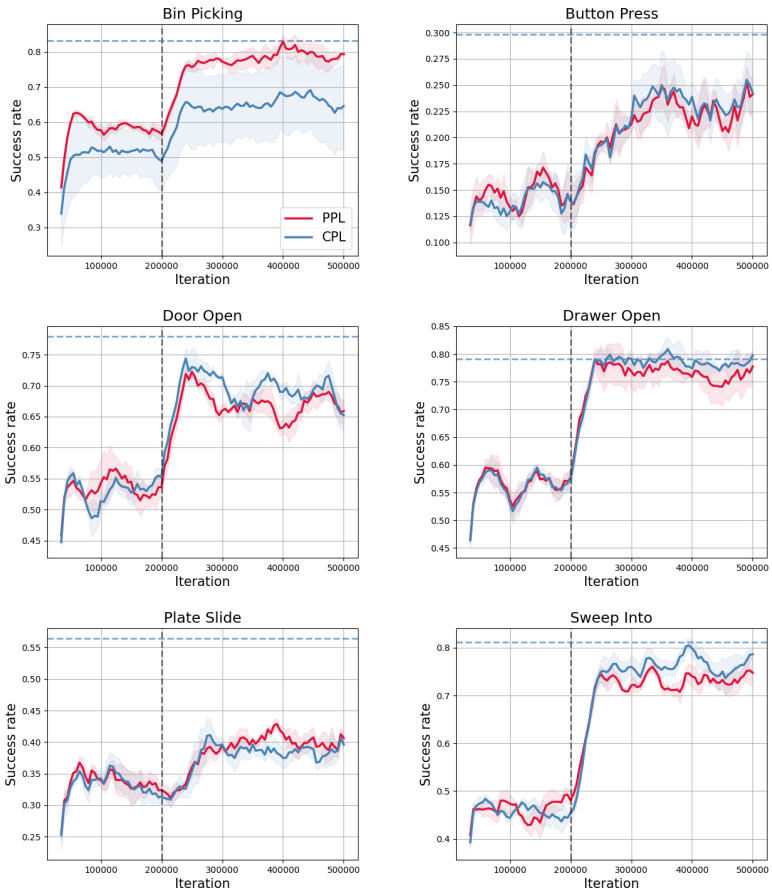


Figure C.3: Reproducibility check on `State Sparse` dataset

C.4.4 Offline dataset generation and its distribution

We construct a heterogeneous dataset by incorporating various policies, following the dataset generation method used in Hejna et al. [44]. Specifically, we load suboptimal SAC checkpoints with success rates of 20% and 50% using the same approach. During rollouts, we introduce Gaussian noise with a standard deviation of 0.3 and rolling out 20,000 episodes, each lasting 250 steps, using their suboptimal soft actor-critic (SAC) [41] checkpoints, which achieved an approximate 50% success rate.

While following this data generation procedure, we found a step in the reference code where transitions following a success signal were explicitly truncated. This truncation was intended to prevent segments from being overly dominated by successful transitions. However, we opted to retain the raw data without truncation. As a result, the distribution of our 50% success rate dataset differs from that of Hejna et al. [44]. To highlight this difference, we provide a visualization of the data distribution across environments.

For our experiments, we generated the following four datasets:

- `Homogeneous Dense`
- `Homogeneous Sparse`
- `Hetereogeneous Dense`
- `Heterogeneous Sparse`

In the additional experimental setting, we kept all aspects—such as the hyperparameters of all algorithms, the SAC critic, and the label generation method—identical to the original setup, modifying only the dataset. Interestingly, `CPL` exhibited significant performance variations depending on the dataset, whereas `PPL` demonstrated robust performance across diverse datasets.

The robustness of PPL's performance can be attributed to its ability to adjust the magnitude of feedback for diverse policies and accurately reflect the likelihood of each segment.

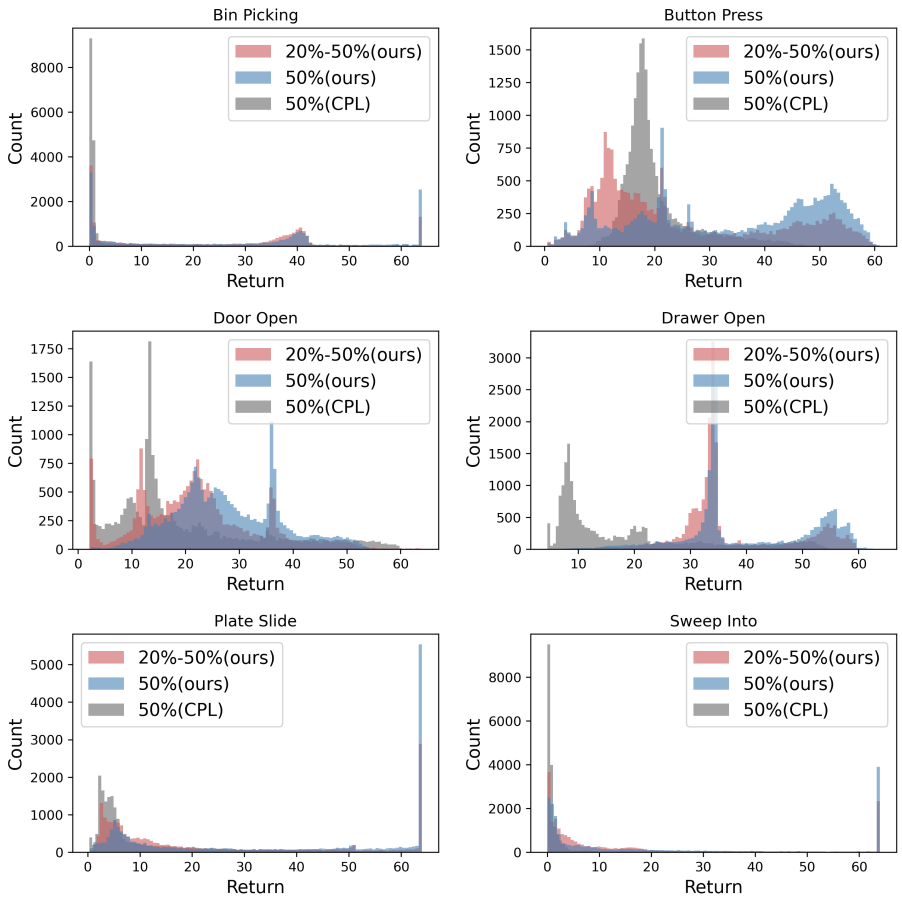


Figure C.4: Comparison of return distributions across environments for different dataset configurations. The histograms illustrate the distribution of the partial returns for segments with 20% and 50% success rates generated using our method **(red and blue)** and the 50% success rate dataset from Hejna et al. [44] **(gray)**.

C.4.5 Online Implementation

In the online setting, we use a Gaussian actor with a fixed standard deviation to maintain consistency with the offline setting. The model is trained from scratch without any pretraining. The online learning process consists of three phases. First, rollouts are conducted in the environment for a fixed number of steps to generate trajectory data. Next, preference queries and labels are constructed from segments of these trajectories. Finally, the policy is updated using the generated preference query data.

During the rollout phase, actions are sampled from a stochastic policy without additional exploration strategies. In the query generation phase, two policies are selected for comparison, with one always being the most recent and the other randomly chosen from the last 25 policies. Segments from the most recent policy are first over-sampled at three times the required number, then ranked based on their regret scores relative to the current policy. The top-ranked segments are retained, while segments from the other policy are sampled uniformly at random. Preference labels are assigned according to the method described in Appendix D.2 of Hejna et al. [44].

In the policy update phase, stochastic gradient updates are applied over a fixed number of epochs using all preference query data collected up to that point. Unlike reward-based preference learning methods, which predominantly generate preference queries early in training and subsequently optimize policies using a learned reward function and an RL algorithm, the online `PPL` algorithm continuously collects preference queries throughout the entire training process. This ensures sustained policy improvement over time.

To reproduce the online baseline `PEBBLE` algorithm, we utilized the official `B-Pref` implementation (`https://github.com/rll-research/BPref`) and ad-

hered to the hyperparameter settings and random seeds reported in the original paper. Our online experiments were performed on five tasks from the Meta-World benchmark: `Button Press`, `Door Open`, `Drawer Open`, `Plate Slide`, and `Sweep Into`. All hyperparameters were kept consistent across tasks, except for the total number of preference queries, which was set to match the values specified for each environment in the `PEBBLE` paper.

C.5 Experimental Results on Homogeneous/ Heterogeneous Datasets (Section 5.3.2)

C.5.1 `Homogeneous Dense` Offline Dataset

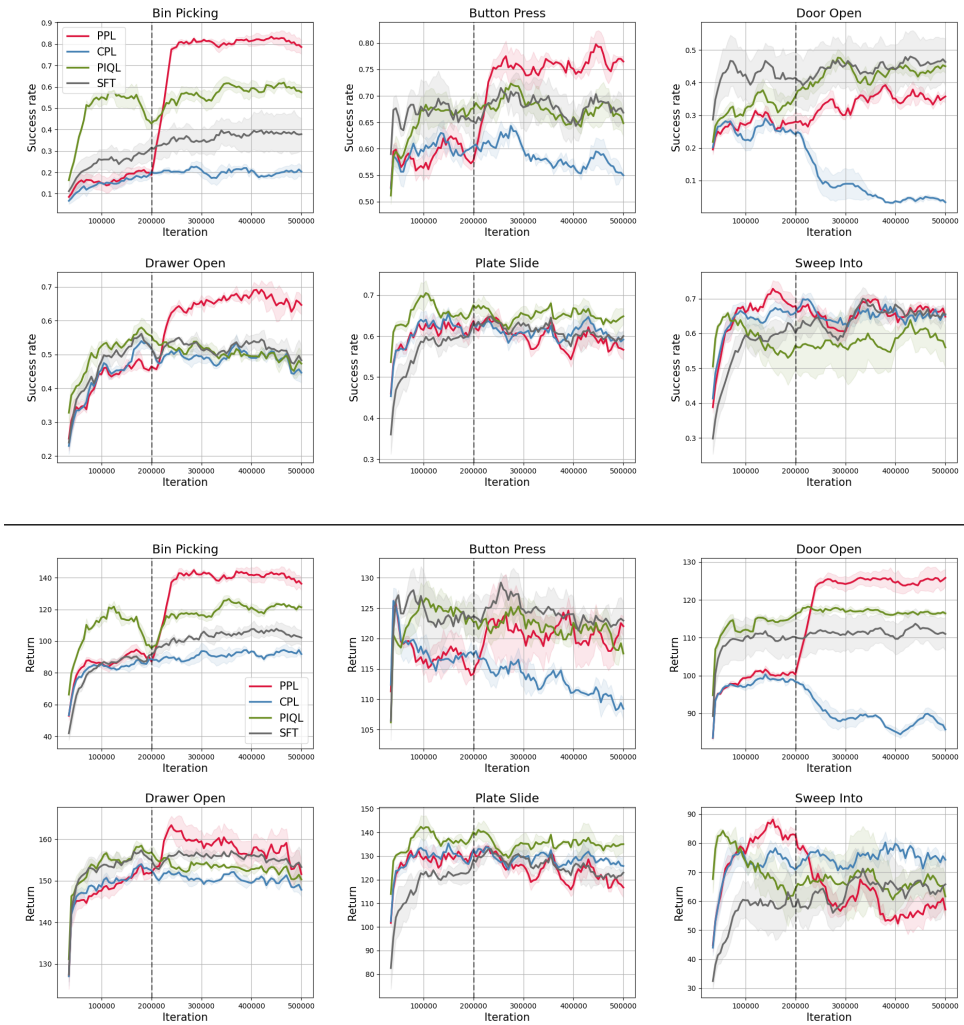


Figure C.5: Performance comparison of different methods on the `Homogeneous Dense` dataset across six MetaWorld tasks. The top row shows the success rate over training iterations, while the bottom row presents the corresponding return values.

C.5.2 `Homogeneous Sparse` Offline Dataset

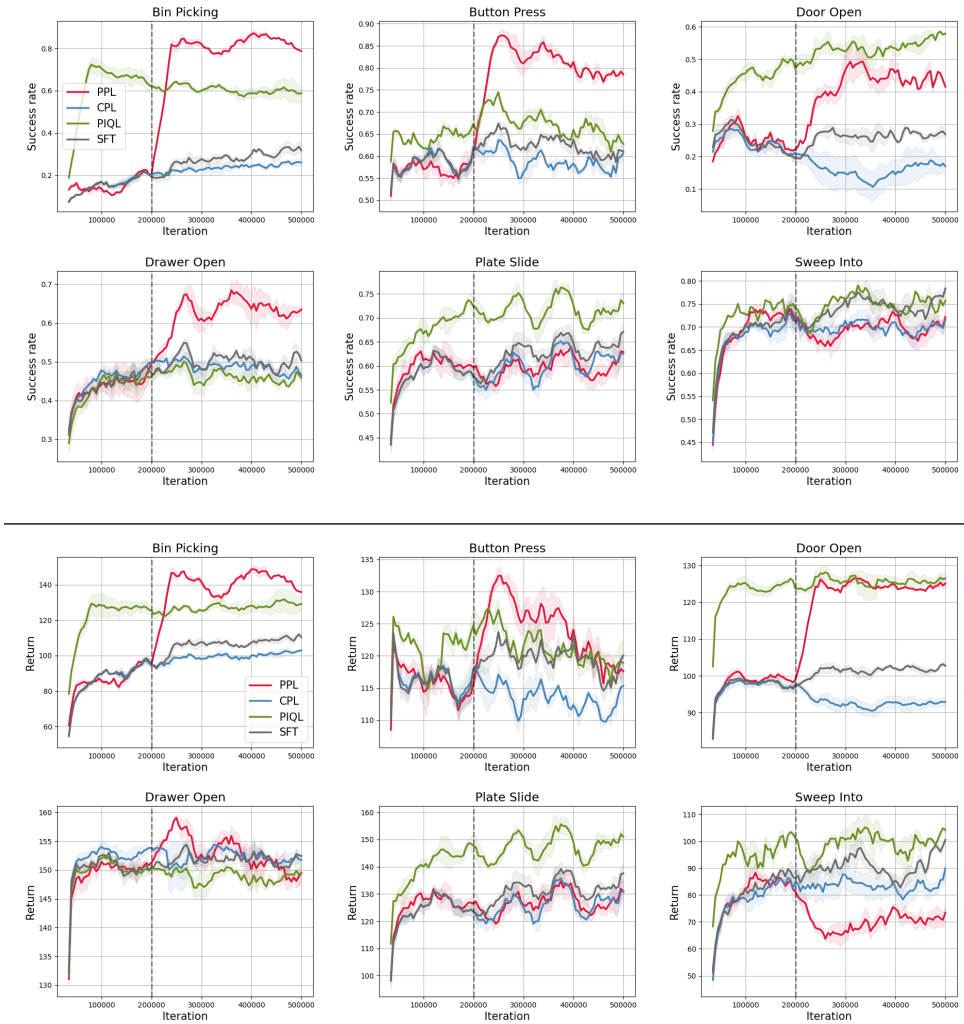


Figure C.6: Performance comparison of different methods on the `Homogeneous Sparse` dataset across six MetaWorld tasks.

C.5.3 `Heterogeneous Dense` Offline Dataset

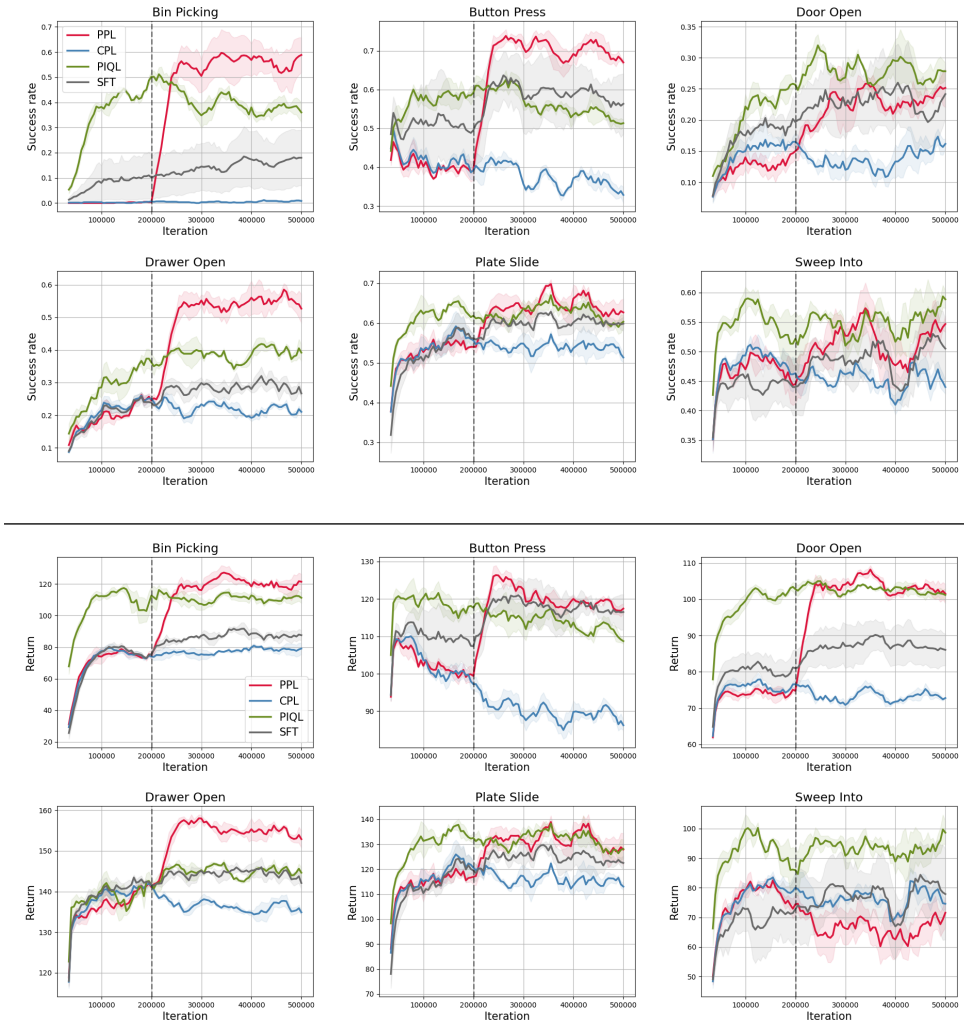


Figure C.7: Performance comparison of different methods on the `Heterogeneous Dense` dataset across six MetaWorld tasks.

C.5.4 `Heterogeneous Sparse` Offline Dataset

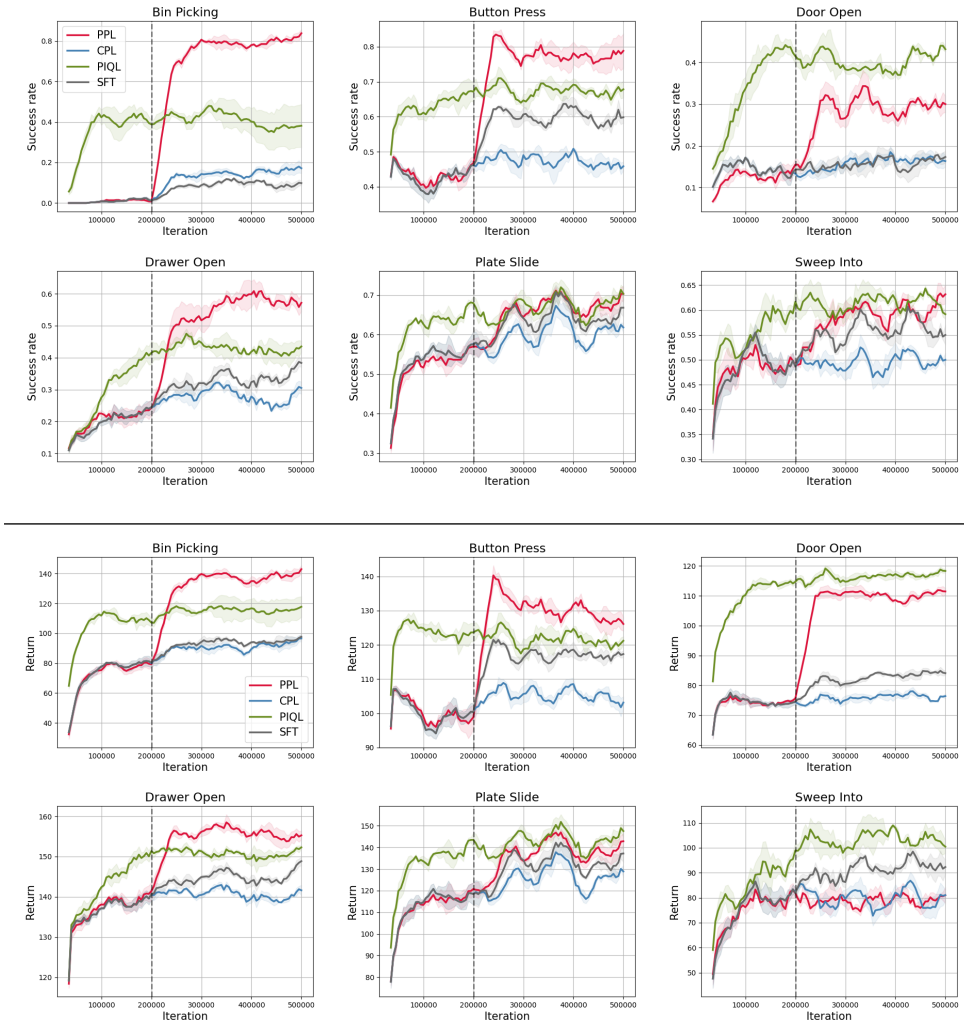


Figure C.8: Performance comparison of different methods on the `Heterogeneous Sparse` dataset across six MetaWorld tasks.

C.6 Comparison with Deterministic Pseudo-labels (Section 5.3.3)

C.6.1 `Homogeneous Dense` Offline Dataset

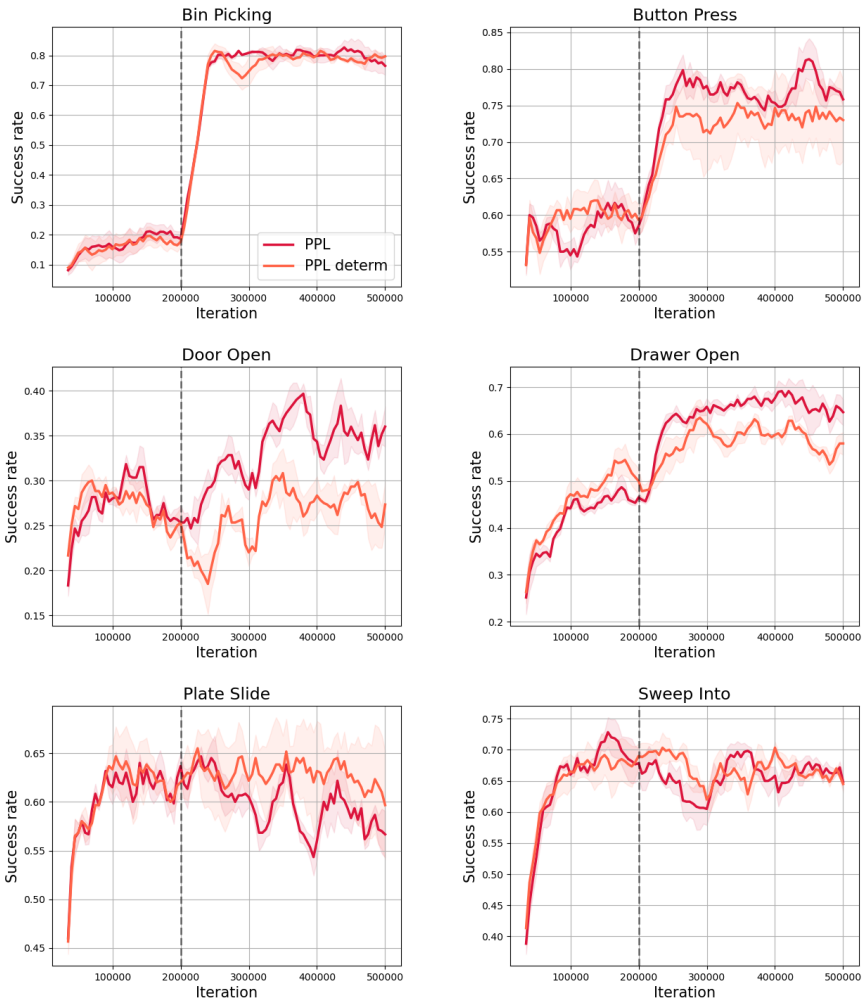


Figure C.9: Comparison of `PPL` and `PPL-deterministic` on the `Homogeneous Dense` Offline Dataset.

C.6.2 `Heterogeneous Dense` Offline Dataset

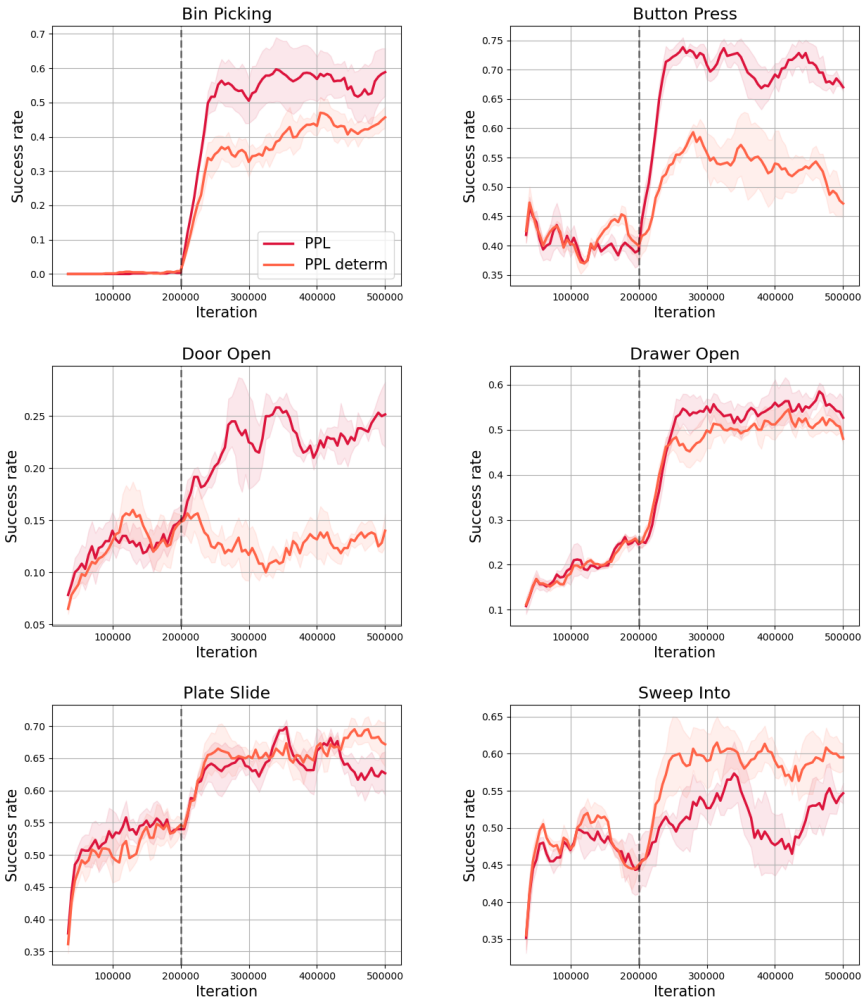


Figure C.10: Comparison of `PPL` and `PPL-deterministic` on the `Heterogeneous Dense` Offline Dataset.

C.7 Experimental Results on Online Implementation (Section 5.3.4)

C.7.1 Online Learning Curves

We evaluated the performance of PPL in an online setting across five MetaWorld tasks. The number of preference queries (#Pref) varied for each environment based on the quantities used in PEBBLE, and these differences are illustrated in each plot.

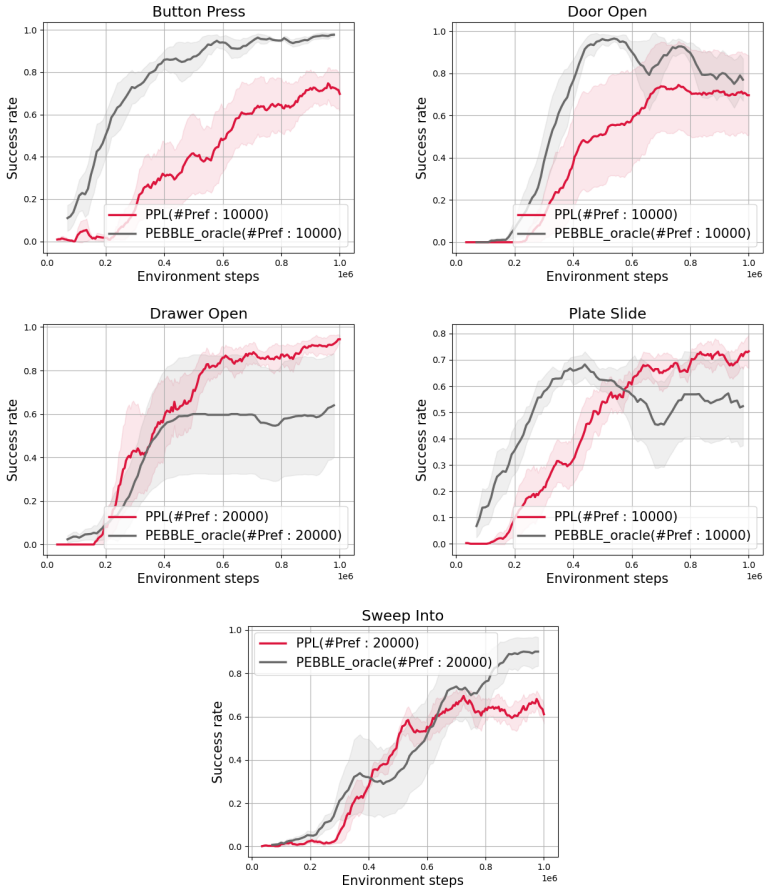


Figure C.11: PPL and PEBBLE learning curves in online learning.

C.7.2 Ablation on Preference Query Count

We evaluate the performance of PPL over iterations with different numbers of preference queries (#Pref). Overall, increasing the number of preference queries leads to improved performance, demonstrating the benefit of richer preference feedback in online learning.

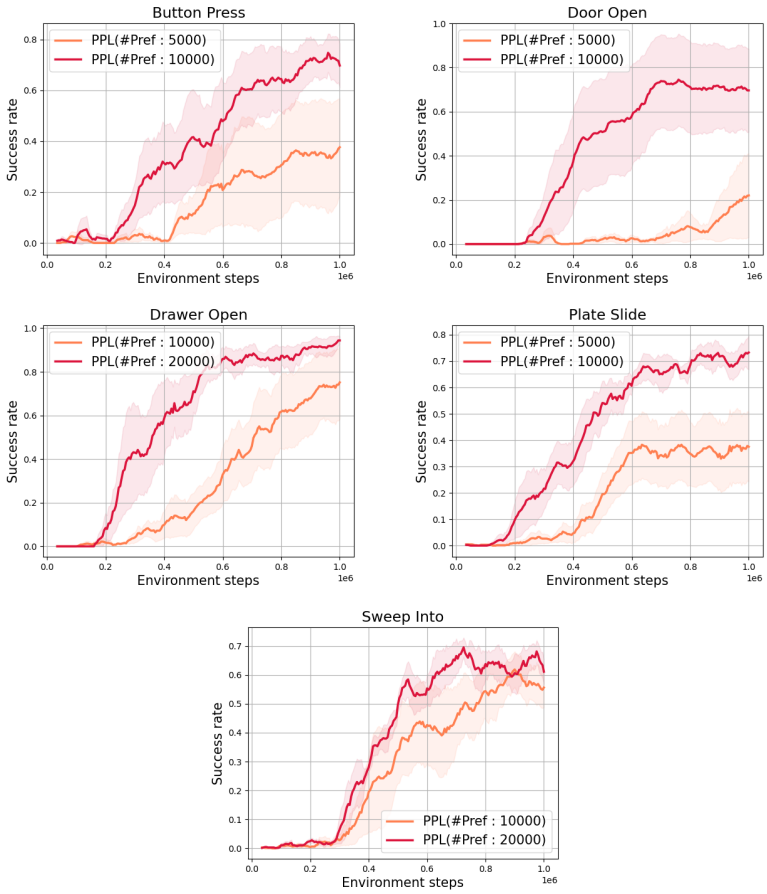


Figure C.12: Effect of preference query count in online learning.

C.7.3 Ablation on Rollout Length

We analyze the impact of different rollout lengths L on the performance of `PPL` in an online RLHF setting across five MetaWorld tasks. Each plot compares the success rate over training iterations for three rollout lengths: $L = \{5, 10, 20\}$.

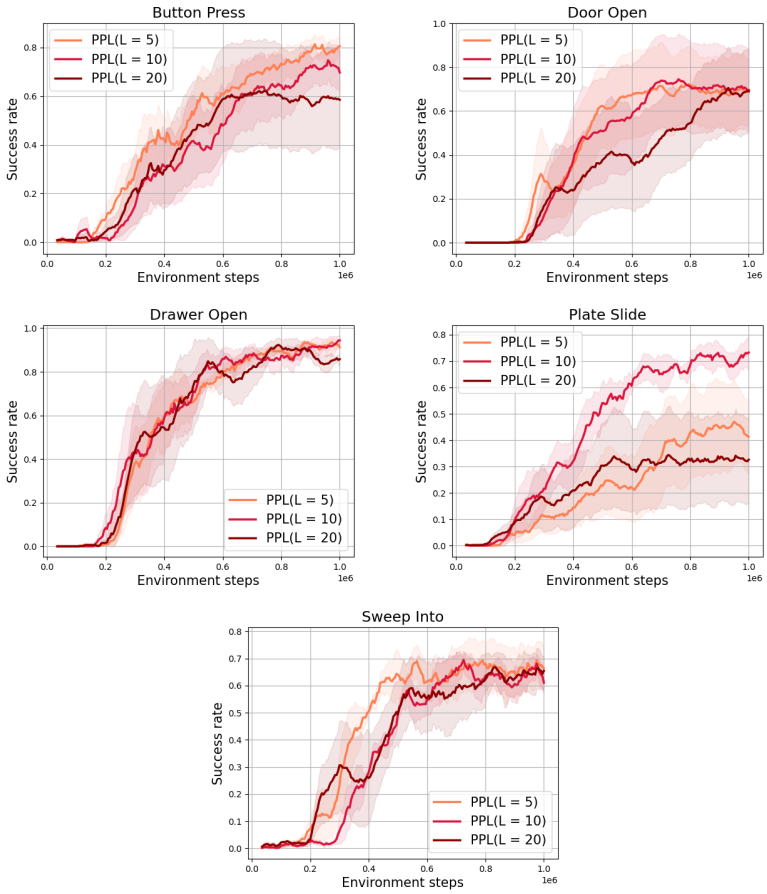


Figure C.13: Effect of rollout length in online learning.

Bibliography

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

[2] Alekh Agarwal, Yujia Jin, and Tong Zhang. Vo ql: Towards optimal regret in model-free rl with nonlinear function approximation. In *Conference on Learning Theory*, pages 987–1063. PMLR, 2023.

[3] Riad Akrour, Marc Schoenauer, and Michèle Sebag. April: Active preference learning-based reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 116–131. Springer, 2012.

[4] Gaon An, Junhyeok Lee, Xingdong Zuo, Norio Kosaka, Kyung-Min Kim, and Hyun Oh Song. Direct preference-based policy optimization without reward modeling. In *Advances in Neural Information Processing Systems*, volume 36, pages 70247–70266, 2023.

[5] Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. *Advances in Neural Information Processing Systems*, 21, 2008.

[6] Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pages 463–474. PMLR, 2020.

[7] Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

[8] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[9] Osbert Bastani, Jason Yecheng Ma, Estelle Shen, and Wanqiao Xu. Regret bounds for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*, 35:36259–36269, 2022.

[10] Syrine Belakaria, Joshua Kazdan, Charles Marx, et al. Sharpe ratio-guided active learning for preference optimization in rlhf. In *arXiv preprint arXiv:2503.22137*, 2025.

[11] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

[12] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional

perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

[13] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

[14] Han Bleichrodt and Peter P Wakker. Regret theory: A bold alternative to the alternatives. *The Economic Journal*, 125(583):493–532, 2015.

[15] Han Bleichrodt, Andrea Cillo, and Enrico Diecidue. A quantitative measurement of regret theory. *Management Science*, 56(1):161–175, 2010.

[16] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4): 324–345, 1952.

[17] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[18] Daniele Calandriello, Daniel Guo, Remi Munos, et al. Human alignment of large language models through online preference optimisation. In *arXiv preprint arXiv:2403.08635*, 2024.

[19] Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G Bellemare. Dopamine: A research framework for deep reinforcement learning. *arXiv preprint arXiv:1812.06110*, 2018.

[20] Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

[21] Yu Chen, Xiangcheng Zhang, Siwei Wang, and Longbo Huang. Provable risk-sensitive distributional reinforcement learning with general function approximation. *arXiv preprint arXiv:2402.18159*, 2024.

[22] Zheng Chen, Shuang Li, Xiaoyang Qiu, and Chi Jin. Regret analysis for reinforcement learning from preferences. *arXiv preprint arXiv:2402.11487*, 2024.

[23] Taehyun Cho, Seungyub Han, Heesoo Lee, Kyungjae Lee, and Jungwoo Lee. Pitfall of optimism: Distributional reinforcement learning by randomizing risk criterion. *arXiv preprint arXiv:2310.16546*, 2023.

[24] Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in Neural Information Processing Systems*, 27, 2014.

[25] Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in Neural Information Processing Systems*, 28, 2015.

[26] Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

[27] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[28] Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor-critic. *arXiv preprint arXiv:1910.12807*, 2019.

[29] William R Clements, Bastien Van Delft, Benoît-Marie Robaglia, Reda Bahi Slaoui, and Sébastien Toth. Estimating risk and uncertainty in deep reinforcement learning. *arXiv preprint arXiv:1905.09638*, 2019.

[30] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pages 1096–1105. PMLR, 2018.

[31] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[32] Will Dabney, Zeb Kurth-Nelson, Naoshige Uchida, Clara Kwon Starkweather, Demis Hassabis, Rémi Munos, and Matthew Botvinick. A distributional code for value in dopamine-based reinforcement learning. *Nature*, 577(7792):671–675, 2020.

[33] Christoph Dann, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. On oracle-efficient pac rl with rich observations. *Advances in Neural Information Processing Systems*, 31, 2018.

[34] Jan Dhaene, Alexander Kukush, Daniël Linders, and Qihe Tang. Remarks on quantiles and distortion risk measures. *European Actuarial Journal*, 2 (2):319–328, 2012.

[35] Martin Engert. Finite dimensional translation invariant subspaces. *Pacific Journal of Mathematics*, 32(2):333–343, 1970.

[36] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance

guarantees and tractable reformulations. *Mathematical Programming*, 171 (1):115–166, 2018.

[37] Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *International Conference on Machine Learning*, pages 3198–3207. PMLR, 2021.

[38] Mohammad Gheshlaghi Azar, Raphael Marinier, Lihong Li, Yiding Luo, and Mohammad Norouzi. A general theoretical paradigm to understand learning from human preferences. *arXiv preprint arXiv:2310.12036*, 2023.

[39] Thomas Gilovich and Victoria H Medvec. The temporal pattern to the experience of regret: What, when, and why. *Journal of Personality and Social Psychology*, 70(3):357–384, 1995.

[40] Adam Gleave, Michael Dennis, Shane Legg, Stuart Russell, and Jan Leike. Quantifying differences in reward functions. *arXiv preprint arXiv:2006.13900*, 2020.

[41] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[42] Paul R Halmos. The theory of unbiased estimation. *The Annals of Mathematical Statistics*, 17(1):34–43, 1946.

[43] Joey Hejna and Dorsa Sadigh. Inverse preference learning: Preference-based rl without a reward function. *Advances in Neural Information Processing Systems*, 36, 2024.

[44] Joey Hejna, Rafael Rafailov, Harshit Sikchi, Chelsea Finn, Scott Niekum, and W Bradley Knox. Contrastive preference learning: Learning from human feedback without rl. *arXiv preprint arXiv:2310.13639*, 2023.

[45] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2018.

[46] Jiwoo Hong, Noah Lee, and James Thorne. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

[47] Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

[48] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in Neural Information Processing Systems*, 31, 2018.

[49] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pages 2137–2143. PMLR, 2020.

[50] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in Neural Information Processing Systems*, 34:13406–13418, 2021.

[51] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.

[52] Sham Machandranath Kakade. *On the sample complexity of reinforcement learning*. University College London, 2003.

[53] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023. v2.

[54] Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

[55] W Bradley Knox, Stephane Hatgis-Kessell, Serena Booth, Scott Niekum, Peter Stone, and Alessandro Allievi. Models of human preference for learning reward functions. *arXiv preprint arXiv:2206.02231*, 2022.

[56] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021.

[57] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[58] Kimin Lee, Laura Smith, and Pieter Abbeel. Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. *arXiv preprint arXiv:2106.05091*, 2021.

[59] Hao Liang and Zhi-Quan Luo. Bridging distributional and risk-sensitive

reinforcement learning with provable regret bounds. *arXiv preprint arXiv:2210.14051*, 2022.

[60] Graham Loomes and Robert Sugden. Regret theory: An alternative theory of rational choice under uncertainty. *The economic journal*, 92(368): 805–824, 1982.

[61] Graham Loomes, Chris Starmer, and Robert Sugden. Observing violations of transitivity by experimental methods. *Econometrica*, 59(2):425–439, 1991.

[62] Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018.

[63] Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in markov decision processes. *arXiv preprint arXiv:2310.20266*, 2023.

[64] Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International Conference on Machine Learning*, pages 4424–4434. PMLR, 2019.

[65] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[66] Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. The

potential of the return distribution for exploration in rl. *arXiv preprint arXiv:1806.04242*, 2018.

[67] Ted Moskovitz, Jack Parker-Holder, Aldo Pacchiano, Michael Arbel, and Michael Jordan. Tactical optimism and pessimism for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 34:12849–12863, 2021.

[68] W. Muldrew et al. Active preference learning for large language models. *arXiv preprint arXiv:2402.08114*, 2024.

[69] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, volume 99, pages 278–287, 1999.

[70] Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9144–9152, 2021.

[71] Jihwan Oh, Joonkee Kim, and Se-Young Yun. Risk perspective exploration in distributional reinforcement learning. *arXiv preprint arXiv:2206.14170*, 2022.

[72] Ian Osband and Benjamin Van Roy. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

[73] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in Neural Information Processing Systems*, 29:4026–4034, 2016.

[74] Ian Osband, Benjamin Van Roy, Daniel J Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 20(124):1–62, 2019.

[75] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[76] John Quan and Georg Ostrovski. DQN Zoo: Reference implementations of DQN-based agents, 2020. URL `http://github.com/deepmind/dqn_zoo`.

[77] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[78] Marc Rigter, Bruno Lacerda, and Nick Hawes. Risk-averse bayes-adaptive reinforcement learning. *arXiv preprint arXiv:2102.05762*, 2021.

[79] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–42, 2000.

[80] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[81] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional

reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

[82] Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *arXiv preprint arXiv:2301.04462*, 2023.

[83] Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26, 2013.

[84] Konrad Schmüdgen et al. *The moment problem*, volume 9. Springer, 2017.

[85] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[86] Nihar B Shah, Sivaraman Balakrishnan, Martin J Wainwright, and Kannan Ramchandran. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Conference on Learning Theory*, pages 856–898. PMLR, 2016.

[87] Omar Shaikh, Michelle Lam, Joey Hejna, Yijia Shao, Michael Bernstein, and Diyi Yang. Show, don't tell: Aligning language models with demonstrated feedback. *arXiv preprint arXiv:2406.00888*, 2024.

[88] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[89] James Alexander Shohat and Jacob David Tamarkin. *The problem of moments*. Number 1. American Mathematical Society, 1943.

[90] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[91] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419): 1140–1144, 2018.

[92] Elena Smirnova, Elvis Dohmatob, and Jérémie Mary. Distributionally robust reinforcement learning. *arXiv preprint arXiv:1902.08708*, 2019.

[93] Silvestr Stanko and Karel Macek. Risk-averse distributional reinforcement learning: A cvar optimization approach. In *International Joint Conference on Computational Intelligence (IJCCI)*, pages 412–423, 2019.

[94] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[95] Robert Sugden. An axiomatic foundation for regret theory. *Journal of Economic Theory*, 60(1):159–180, 1993.

[96] Yunhao Tang and Shipra Agrawal. Exploration by distributional reinforcement learning. *arXiv preprint arXiv:1805.01907*, 2018.

[97] Amos Tversky and Daniel Kahneman. Advances in prospect theory:

Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5(4):297–323, 1992.

[98] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

[99] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.

[100] Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. *arXiv preprint arXiv:2305.15703*, 2023.

[101] Ruosong Wang, Russ R Salakhutdinov, and Lin Yang. Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems*, 33:6123–6135, 2020.

[102] Shaun S Wang. A class of distortion operators for pricing financial and insurance risks. *Journal of Risk and Insurance*, pages 15–36, 2000.

[103] Yining Wang, Ruosong Wang, Simon S Du, and Akshay Krishnamurthy. Optimism in reinforcement learning with generalized linear function approximation. *arXiv preprint arXiv:1912.04136*, 2019.

[104] Yuting Xu, Kamalika Chaudhuri, Arnaud Durand, and Peter L Bartlett. Information-theoretic lower bounds for preference learning. In *Interna-*

tional Conference on Artificial Intelligence and Statistics, pages 3068–3078. PMLR, 2020.

[105] Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 32:6193–6202, 2019.

[106] Insoon Yang. Wasserstein distributionally robust stochastic control: A data-driven approach. *IEEE Transactions on Automatic Control*, 2020.

[107] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning*, pages 1094–1100. PMLR, 2020.

[108] Andrea Zanette, Alessandro Lazaric, Mykel Kochenderfer, and Emma Brunskill. Learning near optimal policies with low inherent bellman error. In *International Conference on Machine Learning*, pages 10978–10989. PMLR, 2020.

[109] Marcel Zeelenberg and Rik Pieters. A theory of regret regulation 1.0. *Organizational Behavior and Human Decision Processes*, 104(1):3–20, 2007.

[110] Yongcheng Zeng, Guoqing Liu, Weiyu Ma, Ning Yang, Haifeng Zhang, and Jun Wang. Token-level direct preference optimization. *arXiv preprint arXiv:2404.11999*, 2024.

[111] Shangtong Zhang and Hengshuai Yao. Quota: The quantile option architecture for reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5797–5804, 2019.

[112] Shangtong Zhang, Bo Liu, and Shimon Whiteson. Mean-variance policy iteration for risk-averse reinforcement learning. *arXiv preprint arXiv:2004.10888*, 2020.

[113] Yuxuan Zhao, Ananya Joshi, et al. Slic-hf: Sequence likelihood calibration with human feedback. In *arXiv preprint arXiv:2305.10425*, 2023.

[114] Dongruo Zhou, Quanquan Gu, and Csaba Szepesvári. Nearly minimax optimal reinforcement learning for linear mixture markov decision processes. In *Conference on Learning Theory*, pages 4532–4576. PMLR, 2021.

[115] Fan Zhou, Zhoufan Zhu, Qi Kuang, and Liwen Zhang. Non-decreasing quantile function network with efficient exploration for distributional reinforcement learning. *arXiv preprint arXiv:2105.06696*, 2021.

[116] W. Zhou et al. Enhancing rlhf with weighted preference optimization (wpo). In *EMNLP 2024*, pages 475–491. Association for Computational Linguistics, 2024.

[117] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

[118] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

초록

불확실성 속 인간 피드백 기반 순차적 의사결정을 다루는 본 논문은 분포 강화학습과 인간 피드백 기반 강화학습이라는 두 가지 핵심 연구 분야에 초점을 맞춘다. 분포 강화학습은 위험에 민감한 제어에서, 인간 피드백 기반 강화학습은 인간 선호도 정렬에서 출발했지만, 두 분야 모두 불확실성과 불완전한 정보가 필연적인 환경에서 원칙적인 의사결정을 가능하게 하는 알고리즘을 설계해야 한다는 공통된 과제를 공유한다. 본 논문은 이러한 도전 과제들을 후회 최소화라는 통일된 관점에서 해석하고, 에이전트의 행동을 인간의 의사결정 구조에 정렬하기 위한 이론적 프레임워크와 실용적인 알고리즘 원리를 제시한다.

논문의 첫 번째 부분에서는 분포 강화학습 분야에서의 탐색 문제를 재조명한다. 기존의 '불확실성에 대한 낙관주의'는 반환 분포의 분산 추정치를 활용하지만, 이는 인식론적 불확실성과 내재적 불확실성을 혼동하여 지속적인 위험 추구 편향과 편향된 데이터 수집을 야기하는 문제점이 발생함을 확인하였다. 이를 해결하기 위해 우리는 외란된 퀀타일 정규화 알고리즘(Perturbed Quantile Regression)을 제안한다. 이는 왜곡된 위험 척도에 무작위 외란이 적용된 척도를 도입하여 행동을 선택하는 방식으로, 이론적으로 편향된 탐색을 피하면서 본래의 최적점에 도달하는 것을 증명하며, 55개의 아타리 게임을 포함한 다양한 벤치마크에서 기존의 분산 기반 탐색 방법보다 우수한 성능을 달성함을 보였다.

두 번째 부분은 분포 강화학습에서 분포의 무한 차원성이라는 근본적인 난제를 다룬다. 기존 연구들은 '벨만 닫힘(Bellman closedness)' 개념을 도입했으나, 이는 온라인 학습에서 유한 개의 표본만으로 통계적 함수들이 편향 없이 업데이트될 수 있음을 보장하지 못하는 한계가 존재한다. 이에 우리는 벨만 업데이트에서 보존될 뿐만 아니라 유한 개의 샘플로부터 편향 없이 추정 가능한 기능을 특징짓는 '벨만 비편향성(Bellman Unbiasedness)' 개념을 제안한다. 우리의 분석은 오직 모멘트

함수족만이 이 두 가지 특성을 만족함을 밝히고, 이를 바탕으로 일반적인 가치 함수 근사에서도 이론적으로 효율성을 갖춘 최초의 분포형 강화학습 알고리즘인 '통계적 함수 기반 최소제곱 가치 반복 알고리즘(Statistical Functional Least-Squares Value Iteration)'을 설계하였다. 이는 이전 연구들보다 향상된 $\tilde{O}(d_E H^{3/2}\sqrt{K})$라는 타이트한 후회 상한선을 달성한다.

논문의 세 번째 부분은 인간의 수작업 보상 대신 선호도 피드백으로부터 학습하는 인간 피드백 기반 강화학습을 다룬다. 직접 선호도 최적화(Direct Preference Optimization)와 같은 최근 프레임워크는 보상 모델 없이 정책을 직접 최적화하지만, 모든 데이터가 최적의 정책에 의해 생성되었다고 가정하는 '우도 불일치(likelihood mismatch)' 문제를 내재적으로 전제하고 있음을 밝힌다. 이를 해결하기 위해 우리는 후회 개념을 활용하여 인간 선호도를 재해석하고 행동 정책 레이블을 학습 과정에 명시적으로 통합하는 '정책 레이블 기반 선호학습(Policy-labeled Preference Learning)'을 제안한다. 제안하는 알고리즘은 선호되는 데이터에 정책을 맞추고 덜 선호되는 데이터와 대조하는 '대조적 KL 정규화'를 도입한다. 이론적으로 주어진 최적 정책에 대해 보상체계의 등가 클래스를 제공하며, 후회가 유일하게 정의됨에 따른 통계적 강건성을 입증하였다. 실험적으로 로봇 조작 작업에서 오프라인 학습 환경에서의 인간 피드백 기반 강화학습의 성능을 크게 향상시키고 온라인 학습 환경에서 강건함을 입증하였다.

요약하자면, 본 논문은 (1) 분포 강화학습의 편향된 탐색 문제를 해결하는 알고리즘, (2) 일반적인 가치 함수 근사에서 온라인 분포 업데이트를 위한 최초의 증명 가능한 효율적 프레임워크를 제공하는 벨만 비편향성 개념과 이에 기반한 알고리즘, (3) 후회 기반 선호도 모델링을 통해 인간 피드백 기반 강화학습의 우도 불일치 문제를 해결하는 알고리즘을 제시한다. 이러한 연구들은 후회 최소화를 이론과 실무를 아우르는 통일된 원리로 정립하며, 불확실성 속에서도 신뢰성 높고 인간과 정렬된 인공지능 의사결정을 가능하게 하는 기반을 제공한다.

주요어: 강화학습, 분포 강화학습, 인간 피드백 기반 강화학습, 후회 최소화
학번: 2020-24770

감사의 글

6년간의 대학원 생활을 마치며, 마침내 공학박사라는 하나의 매듭을 짓게 되었습니다. 정해진 길 없이 스스로 미지의 영역을 개척해야 하기에, 연구자의 길은 때로 막막했지만 동시에 참 낭만적이면서도 저를 단단하게 만들어주는 모험이었습니다. 비록 그 과정을 걷는 동안 수많은 고민과 방황의 순간들이 있었지만, 함께 이야기를 나눌 수 있는 소중한 분들이 곁에 있었기에 길을 잃지 않고 끝까지 나아가며 마무리 할 수 있었습니다. 돌이켜보면 저는 큰 인복을 누린 사람이라 생각합니다. 이제 학위 과정을 함께해주신 소중한 인연들을 하나하나 되짚어보며, 마음 깊은 곳의 감사를 전하고자 합니다.

무엇보다도, 긴 학위 과정 동안 언제나 흔쾌히 면담에 응해 주시고 아낌없는 지도를 해주신 이정우 교수님께 깊이 감사드립니다. 다양한 시행착오와 방황을 겪던 제게 연구자로서의 자긍심을 심어 주시고, 제 가능성을 믿고 원하는 연구 방향을 존중하며 지지해 주신 덕분에 스스로 자랑스럽다고 여길 수 있는 성과로 이어질 수 있었다고 생각합니다. 교수님의 지도를 통해 연구를 대하는 태도뿐만 아니라, 연구자이자 교육자로서 가져야 할 자세에 대해서도 깊이 고민하고 배울 수 있었습니다. 그 가르침에 다시 한 번 진심으로 감사드립니다.

또한 바쁘신 와중에도 흔쾌히 학위 심사위원을 맡아 주신 오성회 교수님, 정교민 교수님, 문태섭 교수님, 그리고 이경재 교수님께도 감사의 말씀을 드립니다. 심사 과정에서 주신 조언들을 되짚어보며, 박사 학위 과정이 제게 어떤 의미였는

지 스스로 정리할 수 있었고, 논문을 보다 완성도 있게 마무리할 수 있었습니다. 특히 이경재 교수님께서는 항상 열린 마음으로 다가와 주셔서 연구실 선배처럼 편안하게 소통할 수 있었고, 깊은 논의를 긴말없이 공유하는 것으로 공동 연구의 즐거움을 느낄 수 있도록 해주셨습니다. 연구에 대한 열정을 더욱 북돋아 주신 점에 대해 진심으로 감사드립니다.

함께 동고동락한 연구실 동료들과 주변 소중한 사람들에게도 고마움을 전합니다. 먼저, 많은 시간을 함께하고 든든한 버팀목이 되어주신 승엽이 형에게 깊은 감사를 드립니다. 연구실이 안정적이고 활기찬 공간이 될 수 있도록 조성해주신 노력과 연구에 대한 열정을 옆에서 지켜보며, 저 또한 그러한 모습을 본받고자 부단히 노력하게 된 것 같습니다. 또한 석훈, 수환, 도혁, 그리고 김도형 연구원과 함께한 시간들은 연구자로서 누릴 수 있는 가장 큰 행복이자 행운이었습니다. 서로의 연구를 진지하게 논의하고, 다양한 아이디어를 나누며 함께 성장할 수 있었던 경험은 연구자로서 가장 소중한 기억으로 남을 것 같습니다.

연구뿐만 아니라 지친 일상의 무게를 나누며 큰 힘이 되어준 상우, 지윤, 주환, 형근, 형준, 세환이를 비롯해 뉴미연 생활을 함께한 승찬이 형, 연군, 정은 누나, 302동에서 늘 반갑게 맞이해주는 지민, 민해, 재인, 진우, 나경, 정민 등 연구실의 모든 분들에게 고맙다는 말을 전합니다. 여러 어려운 상황 속에서도 각자 맡은 역할에 책임감을 다하며 묵묵히 연구를 병행해 나가는 분들과 즐겁게 연구실 생활할 수 있었던 건 되돌아보면 큰 인복이라 생각합니다. 아울러 연구실 밖에서도 변치 않는 우정으로 곁을 지켜준 준홍이 형, 시현이 형, 용기, 영현, 연훈, 재우, 재현, 솔찬, 원모에게도 고마움을 전합니다. 연구가 풀리지 않아 답답할 때마다 세상과 연결된 숨구멍이 되어준 친구들 덕분에, 박사 과정이라는 긴 터널을 끝까지 웃으며 지날 수 있었습니다.

그 누구보다, 저를 믿어주시고 끝까지 든든히 응원해 주신 부모님과 동생에게 가장 깊은 감사를 전하고 싶습니다. 부족함 없이 연구에 몰두할 수 있도록 가족의 끝없는 사랑과 격려가 있었기에, 고된 박사 과정을 묵묵히 버텨낼 수 있었습니다. 이제 사회로 나아가 그 은혜에 조금씩 보답하고자 하니, 다들 건강한 모습으로

오래도록 제 곁을 지켜봐 주셨으면 합니다.

마지막으로, 긴 여정을 마치고 새로운 시작을 앞둔 제 자신에게도 작은 다짐을 남기고 싶습니다. 졸업의 기쁨도 잠시, 불확실한 미래에 대한 또다른 두려움과 걱정이 밀려오지만, 지금까지 그래왔던 것처럼 제 자신을 믿고 후회 없이 최선을 다하며 한 걸음씩 나아가고자 합니다. 인공지능의 시대에 학문의 의미가 희미해져 간다 하지만, 오히려 그렇기에 진리를 추구하고 인간적인 가치를 전하는 연구의 길은 더욱 소중하다고 믿습니다. 연구하며 느끼는 모든 감정을 불행이 아닌 축복으로 여기며, 이 과정을 오롯이 즐길 줄 아는 사람이 되겠습니다. 훗날 이 글을 다시 읽을 때, 지금의 열정과 다짐이 헛되지 않았음을 증명하는 삶을 살아가도록 노력하겠습니다.