

MDSAA

Master's Degree Program in
Data Science and Advanced Analytics

Descriptive Literature Review in Offline Reinforcement Learning
Frequency Analysis and Descriptive Literature Review in Offline
Reinforcement Learning

Rafael Felício Dinis

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Descriptive Literature Review in Offline Reinforcement Learning

Frequency Analysis and Descriptive Literature Review of the subfield of Offline
Reinforcement Learning

by

Rafael Felício Dinis

Master Thesis presented as partial requirement for obtaining the Master's degree in Data Science and Advanced Analytics, with a specialization in Data Science.

Supervised by

Professor Fernando Bação (PhD), NOVA IMS

July, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

[Lisbon, 19 November 2024]

ABSTRACT

This Master's thesis conducts a comprehensive frequency and trend analysis of 140 ORL papers, providing a descriptive review, identifying key algorithms and mathematical tools, and elucidating interrelationships between Offline Reinforcement Learning concepts using confounding matrices as a marker for the frequency of these connections. Offline Reinforcement Learning is a subfield where policies are derived from static, pre-existing datasets without the possibility of online exploration. This approach is applied in scenarios where real-time data acquisition is either impractical or risky, such as in medical treatment planning or autonomous vehicle navigation. The challenges associated with the inability to explore include data distribution shifts and policy optimization constraints. The contributions of this work include a systematic categorization of ORL literature, revealing dominant problem concepts, methodologies, and mathematical frameworks utilized in the field. By employing a descriptive analysis methodology inspired by Palvia et al. (Palvia et al., 2004), the thesis uncovers emerging trends and significant associations between problem areas and solution approaches. This analysis provides valuable insights into the current state of ORL research, highlighting areas of focus by mapping out the interconnected landscape of ORL methodologies and challenges.

KEYWORDS

Offline Reinforcement Learning; Markov Decision Processes; Frequency Analysis; Descriptive Literature Review; Confounding Matrix

Sustainable Development Goals (SDG):

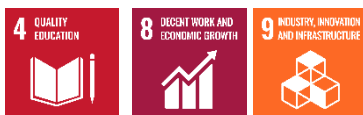


TABLE OF CONTENTS

Statement of Integrity	i
Abstract	ii
List of Abbreviations and Acronyms.....	v
List of Figures.....	viii
1. Introduction.....	1
2. Methodology	2
2.1. Descriptive Review and Frequency Analysis	2
2.2. Data Collection and Extraction	3
2.3. Data Standardization and Categorization	4
2.4. Analysis Techniques.....	4
3. Background and Preliminaries.....	6
3.1. Markov decision processes	6
3.1.1. Standard Markov decision processes.....	8
3.1.2. Constrained Markov Decision Process	9
3.1.3. Partially Observable Markov Decision Process	9
3.1.4. Goal-Conditioned Markov Decision Processes	10
3.2. Base Algorithms.....	10
3.2.1. Q-Learning Variants.....	11
3.2.1.1. Q-Learning.....	11
3.2.1.2. Q-Networks	11
3.2.1.3. Conservative Q-Learning.....	11
3.2.2. Neural Network Applications	13
3.2.2.1. Embeddings.....	13
3.2.2.2. VAEs (Variational Autoencoders).....	13
3.2.3. Transformers in RL.....	14
3.2.3.1. Transformers and Decision Transformers	14
3.2.3.2. Trajectory Transformer	14
3.3. Mathematical Artifacts.....	15
3.3.1. Optimization Frameworks and Formulations	15
3.3.1.1. Constrained Optimization	15
3.3.1.2. Lagrange Multipliers	15
3.3.2. Statistics Methods	16
3.3.2.1. Maximum Likelihood Estimation	16

3.3.2.2. Confidence Set/Regions; Hoeffding's and Bernstein's Inequalities...	17
3.3.3. Information Theory	18
3.3.3.1. Shannon Entropy.....	18
3.3.3.2. Information-Theoretic Metrics for Addressing Distribution Shift	19
4. Insights in Offline Reinforcement Learning	22
4.1. Total Frequency Histograms.....	23
4.2. Stacked Histograms	29
4.3. Confounding Matrices	32
4.3.1. Confounding Matrices – Problem Concept vs Standardized Subarea	33
4.3.2. Confounding Matrices – Problem Concept vs Base	38
4.3.3. Confounding Matrices – Standardized Subarea vs Base	41
4.3.4. Confounding Matrices – Problem Concept vs Categorized Mathematical Artifacts	44
4.3.5. Confounding Matrices – Standardized Subarea vs Categorized Mathematical	52
4.3.6. Confounding Matrices – Base vs Categorized Mathematical Artifacts.....	60
5. Conclusions and future works	65
Bibliographical References	66
Appendix A	94

LIST OF ABBREVIATIONS AND ACRONYMS

AFDT: Action-Free Decision Transformer

AWR: Advantage Weighted Regression

BAMDP: Bayes-Adaptive Markov Decision Process

BEAR: Bootstrapping Error Accumulation Reduction

CDE: Conservative Density Estimation

CEP: Contrastive Energy Prediction

CMDP: Constrained Markov Decision Process

COP: Constrained Optimization Problem

CPED: Constrained Policy Optimization with Explicit Behavior Density

CQL: Conservative Q-Learning

CROP: Conservative Reward for Model-based Offline Policy Optimization

CVAE: Conditional Variational Autoencoder

CVaR: Conditional Value at Risk

DCMRL: Decoupled Meta-Reinforcement Learning

DG: Dynamic Goal

DQN: Deep Q-Networks

DRQL: Distributionally Robust Q-Iteration

DS: Distribution Shift

EDT: Elastic Decision Transformer

FD: Fenchel Duality

GCMDP: Goal-Conditioned Markov Decision Process

GFAN: flow-GAN

IQN: Implicit Quantile Networks

KL: Kullback-Leibler (KL Divergence: Kullback-Leibler Divergence)

KSG: Kraskov-Stögbauer-Grassberger

LHS: Left Hand Side

MBORL: Model-Based Offline Reinforcement Learning

MDP: Markov Decision Process

MFF: Merge-Freeze-Finetune

MIS: Management Information Systems

ML: Machine Learning

MLE: Maximum Likelihood Estimator

MNL: Multinomial Logit

MOBILE: Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning

MO-DT: Multi-Objective Decision Transformer

MRP: Markov Reward Process

NMDP: Non-Markov Decision Process

OAP: Offline-with-Action-Preferences

OGCRL: Offline Goal-Conditioned Reinforcement Learning

O-MDP: Optimistic Markov Decision Process

OOD: Out of Distribution

ORL: Offline Reinforcement Learning

OTORL: Offline-to-Online Reinforcement Learning

PCA: Principal Component Analysis

pdf: Probability Density Function

pmf: Probability Mass Function

POMDP: Partially Observable Markov Decision Process

RHS: Right Hand Side

RL: Reinforcement Learning

RMDP: Robust Markov Decision Process

RTG: Return-to-Go

SCT: Self-Confirming Transformer

SMM: State Marginal Matching

SNR: Signal-to-Noise Ratio

SORL: Safe Offline Reinforcement Learning

SPQR: Spiked Random Model with Q-ensemble Independence Regularization

SUP: Selective Uncertainty Propagation

TATU: Trajectory Truncation with Uncertainty

t-SNE: t-distributed Stochastic Neighbor Embedding

UMAP: Uniform Manifold Approximation and Projection

VAE: Variational Auto-Encoder

VQ-VAE: Vector Quantized Variational Autoencoder

Word2Vec: Word to Vector

LIST OF FIGURES

Figure 1. Methodology Pipeline for the Literature Review in Offline Reinforcement Learning	2
Figure 2. Hierarchical Graph of the Overview of Offline Reinforcement Learning, organized into three main categories: Markov Decision Processes, Base Algorithms, and Mathematical Artifacts. The dotted lines mean interactions between different categories, ...	6
Figure 3. Flowchart of the mdp – from Sutton and Barto’s reference book (Sutton & Barto, 1998).....	8
Figure 4. Pseudo Code for the meta-algorithm encompassing the Classical Q-learning (red), the Conservative Q-learning (blue) and the Q-Networks (green)	12
Figure 5. Number of paper frequency of the Top Decision Processes	23
Figure 6. Number of paper frequency for the Standardized Problem Concepts of the papers	24
Figure 7. Frequency for the Standardized Subarea categorizing the main algorithms presented in the respective paper	25
Figure 8. Frequency for the Base Algorithm across the papers.....	26
Figure 9. Frequency for the Categorized Support Techniques used to support the main algorithm of the paper	27
Figure 10. Frequency for the Categorized Support Techniques used to support the main algorithm of the paper	28
Figure 11. Stacked Histogram of the TOP 5 Decision Processes over Time.....	29
Figure 12. Stacked Histogram of the TOP 5 Standardized Problem Concepts over Time	30
Figure 13. Stacked Histogram of the TOP 5 Standardized Subarea of the relevant algorithm mentioned in the paper	30
Figure 14. Stacked Histogram of the TOP 5 Base Algorithms over Time	31
Figure 15. Network Graph for the different categories analyzed in the paper, the Base Algorithms are in Red, the Sub-Areas are in Orange, the Problem Concepts are in Blue and the Categorized Mathematical Artifacts are in Green	32
Figure 16. Contingency Matrix of the combinations of between Problem Concepts and Standardized Subarea for the column or row total above or equal to 3.....	33
Figure 17. Contingency Matrix of the combinations of between Problems and the Base Algorithm for the column or row total above or equal to 3	38
Figure 18. Contingency Matrix of the combinations of between Subarea and the Base Algorithm for the column or row total above or equal to 3	41
Figure 19. Contingency Matrix of the combinations of between Problem Concept and Mathematical Artifacts for the column or row total above or equal to 5	44

Figure 20. Contigency Matrix of the combinations of between Subarea and Mathematical Artifacts for the column or row total above or equal to 5.....	52
Figure 21. Contigency Matrix of the combinations of between Base Algorithm and Mathematical Artifacts for the column or row total above or equal to 5	60

LIST OF FIGURES

Table 1. Table of the TOP 5 most frequent combinations between the Problems and the Sub-area.....	34
Table 2. Table of the TOP 5 combinations between the Problems and the Base Algorithm ..	39
Table 3. Table of the TOP 5 most frequent combinations between the the Sub-area and the Base Algorithm	42
Table 4. Table of the TOP 5 most frequent combinations between the Problem Concept and Mathematical Artifacts used	45
Table 5. Table of the TOP 5 most frequent combinations between the Subarea and Mathematical Artifacts used	53
Table 6. Table of the TOP 5 most frequent combinations between the Base Algorithm and Mathematical Artifacts used	60

1. INTRODUCTION

Reinforcement Learning (RL) is a machine learning approach where an agent learns to make optimal decisions by interacting with an environment (He, 2023). It is a trial-and-error method (Prudencio et al., 2024), in which RL agents interact with the environment by taking actions and receiving feedback in the form of rewards or penalties (He, 2023). Researchers Sutton and Barto distinguish Reinforcement Learning from both Supervised and Unsupervised Learning due to the unique characteristics of each.

Supervised Learning's goal is to generalize or extrapolate from labeled data, whereas in Reinforcement Learning, the complex interactive environments present a challenge in determining what is an explicitly correct behavior or set of actions. On the other hand, Unsupervised Learning, in which the goal is to identify structures from unlabeled data, differs from RL's focus on attempting to maximize the reward (Sutton & Barto, 1998). The maximization of reward signals, necessary for RL, requires active interaction with the environment to gather data for refining the agent's policies, which might end up being prohibitive in some real-world scenarios, such as autonomous driving or medical treatment planning, where mistaken actions can have severe consequences. Offline Reinforcement Learning, also known as Batch Reinforcement Learning, utilizes static datasets to train the policies. Offline RL presents challenges related to the exploitation of these fixed datasets, namely, overfitting to limited data samples, distribution shifts between the data and real-world scenarios, and the derivation of reliable policies from suboptimal data (Kostrikov et al., 2021).

The exponential growth in the number of publications on RL and ORL (Prudencio et al., 2024) might lead to a gap between systematic categorization and algorithmic/model innovation. Building upon "A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems" (Prudencio et al., 2024) and "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems" (Levine et al., 2020) as a foundation, the aim is to identify, based on the latest research, the frequent topics that constitute the field and describe them in a tutorial-like document based on Descriptive Reviews, thereby allowing newcomers to have an intuition of the field.

The contributions of this thesis include a systematic categorization of Offline Reinforcement Learning literature, revealing dominant problem concepts, methodologies, and mathematical frameworks utilized in the field. By employing a descriptive analysis methodology inspired by Palvia et al. (Palvia et al., 2004), which serves the purpose of understanding patterns in a body of research by quantifying research characteristics, such as publication time, research methodology, approach, theory, and outcomes, the research uncovers emerging trends and significant associations between problem areas and solution approaches. This analysis provides valuable insights into the current state of ORL research, highlighting areas of focus and potential gaps, and offers guidance for future research directions by mapping out the interconnected landscape of ORL methodologies and challenges.

2. METHODOLOGY

This study aims to provide a comprehensive conceptualization of the field of Offline Reinforcement Learning (ORL), focusing on identifying prominent themes, techniques in use, and the problems or objectives addressed in recent research. To achieve this objective, a descriptive analysis methodology inspired by the approach presented in the Management Information Systems (MIS) paper (Palvia et al., 2004) was employed. This methodology involves understanding the frequency of phenomena that emerge across a collection of academic papers, thereby enabling a quantitative assessment of research trends within a specific domain.

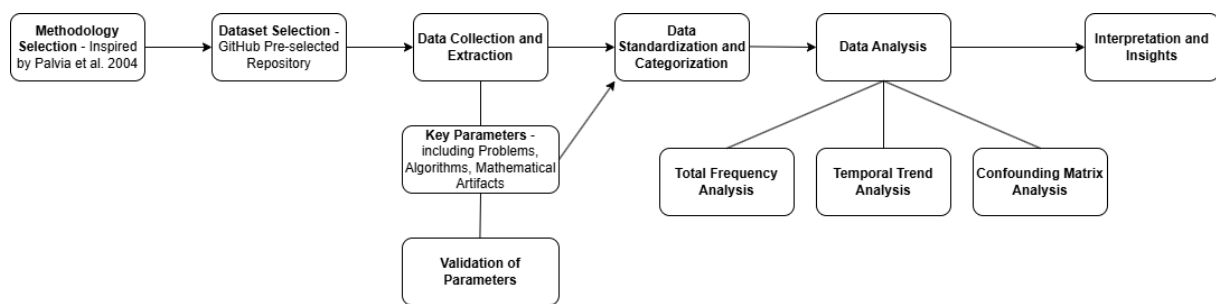


Figure 1. Methodology Pipeline for the Literature Review in Offline Reinforcement Learning

2.1. DESCRIPTIVE REVIEW AND FREQUENCY ANALYSIS

The methodology presented (Palvia et al., 2004) uses content analysis to evaluate research articles based on specific characteristics such as subject areas and research methodologies. Their approach involves coding articles from leading academic journals and analyzing trends over time to discern the evolution of research themes and methodological approaches within the MIS (Management Information Systems) field. This method allows for the systematic quantification of research patterns and the identification of emerging areas of interest.

In alignment with this approach, in this thesis a frequency analysis of research papers within the ORL domain was conducted. The parameters analyzed included:

1. Publication Dates: Extracted to understand the temporal evolution of research themes and the patterns of emergence of new techniques within ORL.
2. Problem Concepts: The specific problems or motivations addressed by the papers, providing insight into the challenges that researchers are focusing on.

3. Decision Processes: The types of Markov Decision Processes (MDPs) used, which detail how the environment progresses and the underlying assumptions about state transitions and observability.
4. Base Algorithms: The foundational algorithms upon which new approaches, algorithms, or frameworks are built, reflecting the core methodologies influencing current research.
5. Support Algorithms: Additional algorithms employed in constructing the presented approaches, indicating the integration of multiple techniques to address complex problems.
6. Mathematical Artifacts: Mathematical and statistical strategies employed within the algorithms, such as the Kullback-Leibler (KL) divergence and Mutual Information, which serve as essential tools rather than standalone algorithms.
7. Areas and Subareas: The broader and more specific categories of the papers, aiding in the classification of research topics within the ORL field.
8. Links to Sources: Archival references to each paper, enabling verification and further exploration of the research.

2.2. DATA COLLECTION AND EXTRACTION

The data for this study was collected from a curated set of ORL research papers, through a GitHub repository [<https://github.com/hanjuku-kaso/awesome-offline-rl>] by Haruka Kiyohara and Yuta Saito, from Cornell University at the time of writing. Each paper was meticulously reviewed to extract the relevant information for the parameters listed above. The extraction process was predominantly manual, involving thorough reading and interpretation to accurately capture the nuanced details of each paper.

For the publication dates, Python-based web scraping tool to automatically retrieve submission dates from the arXiv repository was employed, when applicable. In cases where the papers were not available on arXiv or required additional verification, the dates were extracted manually from the publications. This approach ensured that the data regarding publication timelines was comprehensive and accurate.

2.3. DATA STANDARDIZATION AND CATEGORIZATION

Following data extraction, a process to standardize the data was put in place to guarantee consistency across the dataset. This involved standardizing terminology so that terms referring to the same concept were represented the same way. For example, different ways of spelling Out-of-Distribution, such as OOD, or Out of Distribution were standardized into one term alone.

Additionally, I performed a categorization of terms to organize the data into meaningful clusters. This categorization aids in identifying patterns and drawing insights about the prevalence of specific techniques and concepts within the field. For instance, mathematical artifacts such as "KL divergence" and "Mutual Information" were grouped under "Information-Theoretic Measures," highlighting their role in addressing uncertainty and information estimation in ORL algorithms.

2.4. ANALYSIS TECHNIQUES

The analysis comprised three primary components:

1. **Total Frequencies:** Where the total occurrences of each concept within the dataset were calculated. This quantification provided a clear picture of the most commonly addressed problems, utilized algorithms, and employed mathematical tools in ORL research.
2. **Temporal Trends:** By plotting the frequencies over time, it was possible to observe the evolution of research themes and the emergence of new techniques. This temporal analysis highlighted shifts in focus within the field and identified areas of interest.
3. **Confounding Matrices:** To understand the relationships between pairs of concepts, Confounding matrices were computed. In this context, it allowed for the identify closely related concepts, such as specific algorithms frequently used in conjunction with mathematical artifacts or problems.

The Total Frequencies Analysis refer to the total counts for each parameter category was put into tables, such as the number of times a particular Base Algorithm or Mathematical Artifact was mentioned across all papers. This process involved aggregating the standardized data and generating frequency counts. This counting provided insights into which techniques and concepts are most prevalent in ORL research during the period that was analyzed.

For example, Transformers as a Base Algorithm had a high frequency, indicating a significant interest in leveraging sequence modeling capabilities within ORL. Similarly, mathematical artifacts like KL divergence appeared frequently, underscoring its importance in addressing distributional shifts and policy optimization challenges.

Temporal Trends Analysis serves the purpose of analyzing frequencies of key concepts against over time through their publication dates. By examining these trends, the trajectory of

research interests and the adoption of new methodologies within the ORL community was tentatively delineated, but with low conclusivity. This temporal analysis would be crucial for understanding how the field is evolving and where future research efforts may be directed, specially in the context of a much larger dataset of papers treated, where patterns would have a more noticeable impact.

Confounding Matrices were utilized to measure the strength of association between pairs of categorical variables extracted from the papers. For each pair of parameters (e.g., Base Algorithm and Mathematical Artifact), a contingency table was built. This analysis helped identify strong relationships between concepts. For instance, a high Confounding Matrix value between "Transformers" and "Sequence Modeling and Latent Representation Learning" suggests that when Transformers are used, they are frequently applied in the context of sequence modeling within ORL.

3. BACKGROUND AND PRELIMINARIES

The background section offers a thorough exploration of Offline Reinforcement Learning, beginning with foundational frameworks such as various Markov Decision Processes (including CMDP, POMDP, and GCMDP). It progresses to examine key base algorithms like Q-Learning variants, neural network applications, and transformer-based methods, and further explores crucial mathematical artifacts encompassing statistical methods, optimization frameworks, and information theory. The accompanying graph visually maps these interconnected topics into three primary categories—Markov Decision Processes, Base Algorithms, and Mathematical Artifacts—each with their respective subcategories.

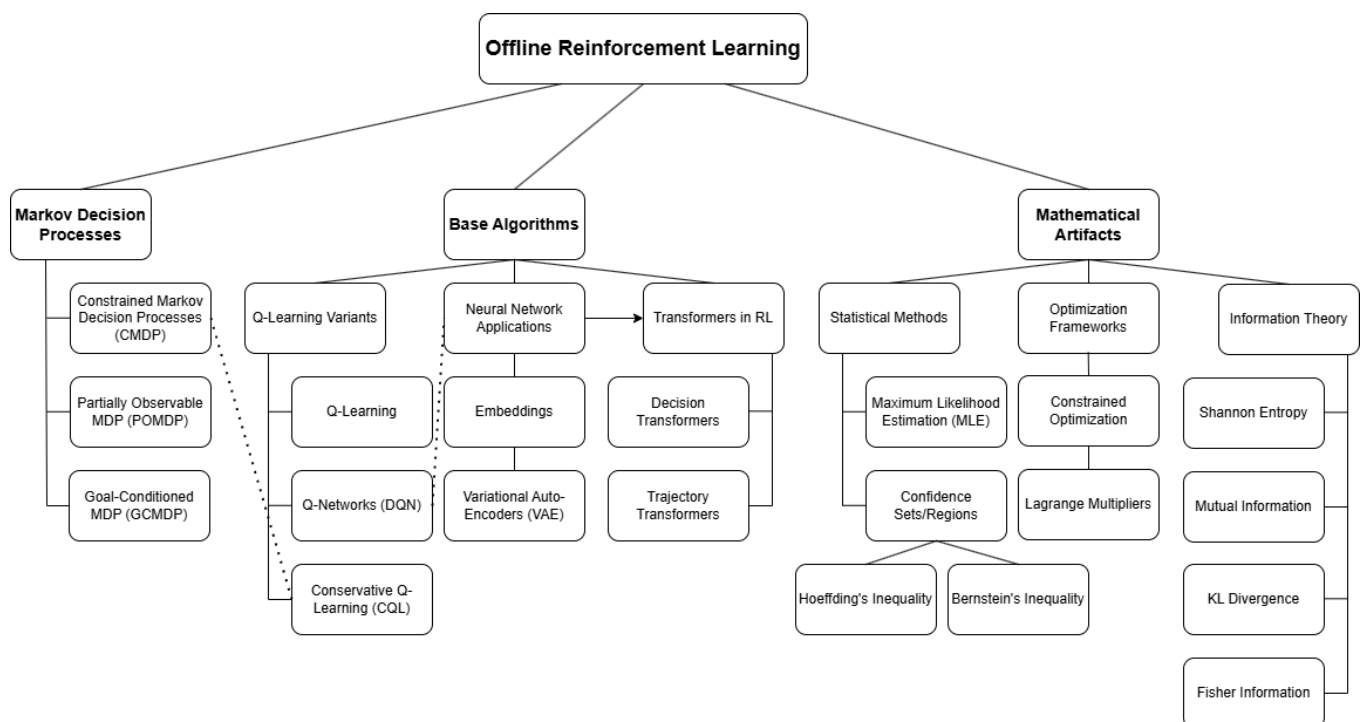


Figure 2. Hierarchical Graph of the Overview of Offline Reinforcement Learning, organized into three main categories: Markov Decision Processes, Base Algorithms, and Mathematical Artifacts. The dotted lines mean interactions between different categories,

This chapter concerns the formulations of Markov Decision Processes and frequently utilized variations, namely the Constrained, Goal-Conditioned, and the Partially Observable Markov Decision Processes. These variations add new concepts to the standard Markov Decision Process and are of interest as they relate to concepts further explored, such as the constrained optimization, in the mathematical formulation section.

In the 1950's, Richard Bellman introduced a method called dynamic programming to solve problems involving decision-making under uncertainty. This work led to the creation of what we now call Markov Decision Processes (MDPs). The key feature of MDPs is the Markov property, which means that the next state depends only on the current state and the chosen action, not on the past states. Bellman's focus was on finding the best policies—rules for

choosing actions—that maximize rewards over time. He did this by breaking down complex problems into simpler subproblems, a technique known as dynamic programming. This provided a systematic way to make optimal decisions in uncertain environments.

Building on Bellman's foundational work, K. J. Åström extended the MDP framework in 1965 to handle situations where the decision-maker doesn't have complete information about the system's state. In his paper, "Optimal Control of Markov Processes with Incomplete State Information," Åström developed methods for finding optimal control rules when only partial observations are available. This extension effectively led to the development of Partially Observable Markov Decision Processes (POMDPs).

As MDPs and POMDPs evolved, there arose a need to handle situations with multiple objectives and limitations. This led to the development of Constrained Markov Decision Processes (CMDPs). CMDPs are an extension of MDPs that include constraints on costs and resources. Early contributions by researchers like Derman in 1970 and Eitan Altman in 1999 expanded the theory to incorporate these constraints, making it possible to model complex systems more realistically. CMDPs have become important in fields like telecommunications, operations research, and engineering. They help optimize performance measures such as throughput and delay while ensuring that constraints like capacity and quality of service are met.

In the early 1990s, the concept of Goal-Conditioned Markov Decision Processes emerged to address environments with changing goals. A pivotal contribution was Leslie Pack Kaelbling's 1993 paper, "Learning to Achieve Goals," where she introduced the Dynamic Goal (DG) learning algorithm. Kaelbling extended traditional reinforcement learning by conditioning policies on goal states, enabling agents to learn strategies that generalize across different objectives within the same environment. The DG learning algorithm updated value functions for all possible goals at once, allowing for efficient knowledge transfer and quick adaptation to new tasks without starting over. This approach laid the groundwork for goal-conditioned reinforcement learning and influenced further developments like hierarchical reinforcement learning and universal value function approximators. These advancements have enhanced the ability of agents to operate effectively in complex settings where goals can change, making decision-making processes even more powerful and flexible.

Together, these developments—from Bellman's MDPs to Åström's POMDPs, the introduction of CMDPs, and Kaelbling's goal-conditioned learning—illustrate the evolution of decision-making frameworks in uncertain and dynamic environments. Each advancement built upon the last, leading to more sophisticated tools that can handle incomplete information, multiple objectives with constraints, and dynamically changing goals, all while aiming to make the best possible decisions.

3.1.1. Standard Markov decision processes

Markov Decision Processes (MDPs) provide a mathematical framework for modeling decision-making in environments that are stochastic, sequential, and discrete. The properties that constitute a Markov Decision Process are States, Actions, Rewards and Transition Probabilities (Missaoui, 2023). It helps us understand how to make a series of decisions in an environment where each choice influences what happens next. Imagine a simple game where a character navigates through a maze to find a treasure. Each position in the maze represents a state, and at each state, the character can choose to move up, down, left, or right—these are the actions. However, there's a chance the character might slip and end up in a different spot than intended, adding an element of randomness. As the character moves, it receives rewards: reaching the treasure gives positive points, stepping into a trap results in negative points, and each move incurs a small cost to encourage finding the shortest path. The goal is to develop the best strategy to reach the treasure while maximizing the total reward, despite the uncertainty in how actions will turn out. This captures the essence of an MDP, where an agent must make optimal decisions in an environment with inherent randomness to achieve the best possible outcome.

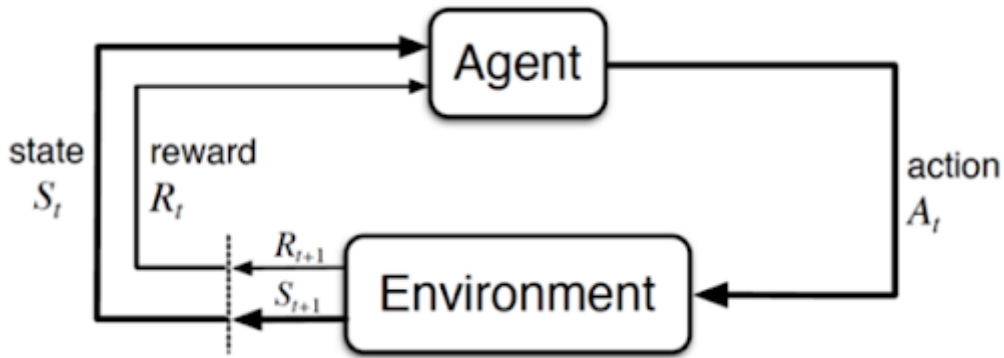


Figure 3. Flowchart of the mdp – from Sutton and Barto’s reference book (Sutton & Barto, 1998)

In that sense, an MDP is defined by a tuple $M = (S, A, R, p, \gamma)$, where:

(S) represents a finite set of states, (A) denotes the actions available in each state, (R) is the set of possible rewards, γ with $0 \leq \gamma \leq 1$ is the discount factor, impacting future rewards' present value, $p(s', R|s, a)$ defines the transition probabilities.

One important concept for Markov Decision Processes is the Markov Property, which pertains to the fact that future states in an MDP depend only on the current state and action and not on the previous states and actions taken before, in a sense, this Markov Property signifies memorylessness of the MDP, and can be formalized mathematically as (Missaoui, 2023):

$$Pr(S_{t+1}|S_t, A_t, S_{t-1}, A_{t-1}, \dots, S_1, A_1) = Pr(S_{t+1}|S_t, A_t), (1)$$

The reward function, $R(s, a)$, quantifies the immediate payoff from taking action (a) in state (s), helping guide the decision-making process: $r(s, a) = \sum_{r \in R} \sum_{s' \in S} r \cdot p(s', r | s, a)$, (2)

3.1.2. Constrained Markov Decision Process

A Constrained Markov Decision Process (CMDP) is a variant form of the Markov Decision Process that includes constraints as part of the decision-making framework. It is formally understood as a tuple (S, A, P, R, C, γ) , where (Altman, 2021):

(S) is the state space, (A) is the action space, ($P: S \times A \rightarrow \Delta(S)$) is the transition kernel that dictates state transitions, ($R: S \times A \rightarrow [0, 1]$) is the reward function, ($C: S \times A \rightarrow [-1, 1]$) is the cost function, and (γ) is the discount factor used to calculate the present value of future rewards and costs.

The goal in a CMDP is to optimize a policy ($\pi: S \rightarrow \Delta(A)$), which naturally determines the action selection in each state of the given environment it pertains to such that it maximizes the expected discounted return while at the same time adhering to the cost constraints, specific to the CMDP (Altman, 2021).

An example of this type of Markov decision process could be thought of as a delivery robot must collect parcels from various locations and deliver them to a depot. It aims to maximize the number of deliveries (reward) while ensuring it doesn't exceed its battery capacity (cost constraint). Using a CMDP, the robot plans routes that optimize deliveries without running out of power.

The challenge present in this type of constrained markov decision process lies in the balancing of the reward maximization against cost constraints.

3.1.3. Partially Observable Markov Decision Process

A Partially Observable Markov Decision Process (POMDP) is yet another variant of the classical framework of the Markov Decision Processes (MDP) that extends MDPs to environments where agents have only access to partial information about the true state of the system. A POMDP is defined as a tuple $(S, A, O, T, S_0, O, O_0, R, H, \gamma)$, where (Gu et al., 2024):

(S) is the set of states, (A) denotes the set of actions, (O) is the set of possible observations, ($T: S \times A \times S \rightarrow [0, 1]$) is the transition function that models the state dynamics, ($S_0: S \rightarrow [0, 1]$) specifies the initial state distribution, ($R: S \times A \times S \rightarrow R$) is the reward function, ($H \in N$) represents the finite time horizon of the decision process, ($\gamma \in [0, 1]$) is the discount factor, ($O: S \times O \rightarrow [0, 1]$) is the observation function that maps states to probabilities of observing each possible observation.

A POMDP's initial conditions involve sampling the start state (s_0) from ($S_0(\cdot)$) and the initial observation (o_0) from ($O_0(\cdot | s_0)$), which are the sets of States and Observations. At each time step (t), from (0) to ($H - 1$), the policy (π) selects an action (a_t) in state (s_t), the system transitions to a new state (s_{t+1}) sampled from ($T(\cdot | s_t, a_t)$), and an observation (o_{t+1}) is then generated from ($O(\cdot | s_{t+1})$). The immediate reward received is computed as $r_{t+1} = R(s_t, a_t, s_{t+1})$ (Gu et al., 2024).

An example for this type of decision process can be an agent navigating a maze environment where it can only see the immediate path ahead but cannot see what's behind walls or around corners (partial observability). The agent must find the exit without a complete map of the maze. A POMDP helps the agent make decisions based on limited input, efficiently exploring the maze despite uncertainty about its layout.

3.1.4. Goal-Conditioned Markov Decision Processes

Goal-Conditioned Markov Decision Processes (GCMDPs) adapt the standard MDP principles to contexts where the agents aim to achieve specific goals. The formal definition of a GCMDP includes the tuple (S, A, G, P, γ, r), with each term defined as follows (Wang et al., 2024):

(S) is the set of states, (A) is the set of actions, (G) is the goal space, indicating specific objectives that the agent aims to achieve, $P(S_{t+1} | s_t, a_t)$ details the transition dynamics, controlling how actions influence state changes, (γ) is the discount factor, (r) is the reward function.

In this framework, an agent employs a policy ($\pi: S \times G \rightarrow A$) to determine actions based on the current state and a designated goal. The objective of the agent is to maximize the expected discounted cumulative return, formalized as (Wang et al., 2024):

$$J(\pi) = E_{g \sim P_g, s_0 \sim P_0(s_0), a_t \sim \pi(\cdot | s_t, g), s_{t+1} \sim P(\cdot | s_t, a_t)} \left[\sum_{t=0}^T \gamma^t r(s_t, a_t, g) \right], (3)$$

where (P_g) and ($P_0(s_0)$) represent the distributions over goals and initial states, respectively, and (T) is the episode length (Wang et al., 2024).

3.2. BASE ALGORITHMS

In this section, we explore foundational algorithms in reinforcement learning (RL), focusing on Q-learning variants, neural network applications, and the integration of transformers in RL. Understanding these algorithms provides a basis for advanced topics and innovations in the field.

3.2.1. Q-Learning Variants

3.2.1.1. Q-Learning

The introductory algorithm of this section is the Q-learning which is a foundational off-policy algorithm. Developed by Watkins in 1989, the Q-learning algorithm is a model-free, off-policy Temporal Difference algorithm that learns the value action in a particular state. This is used to form a policy by updating Q-value, following the formula:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left(R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right), \quad (1)$$

This expression represents the update formula for the Q-value, which in turn estimates the utility of choosing action A_t in state S_t , by applying a learning rate α , to incorporate a reward, R at a +1-time step, and the discounted maximum future Q-value represented by $\gamma \max(Q)$ (Sutton & Barto, 1998).

3.2.1.2. Q-Networks

The Q-Network evolves from the more traditional tabular Q-learning into utilizing function approximators via neural networks, which allows for the handling of larger state spaces through the mapping of state-action pairs to Q-values via an architecture of deep neural networks, similar to the base Q-learning algorithm, DQN (Deep Q-Networks) are model-free and directly learn the optimal policy from interactions with the environment without an explicit model of its dynamics (Mnih et al., 2015). DQN uses a Loss Function to update the Q-values by comparing the values predicted with the ones derived from the Bellman Equation – for iteration i , the loss is defined by:

$$L_i(\theta_i) = E_{(s,a,r,s') \sim U(D)} \left[(y_i - Q(s, a; \theta_i))^2 \right], \quad (2)$$

In which, θ_i are the Q-network parameters at iteration i , while the (s,a,r,s') represents a sample taken from the Dataset D (replay buffer), and $y_i = r + \gamma \max_{a'} Q(s', a'; \theta_i^-)$. The changes in θ_i are given by the gradient of the loss function (2)

$$\nabla_{\theta_i} L_i(\theta_i) = E_{(s,a,r,s') \sim U(D)} \left[(y_i - Q(s, a; \theta_i)) \nabla_{\theta_i} Q(s, a; \theta_i) \right], \quad (3)$$

3.2.1.3. Conservative Q-Learning

A variant of Q-learning that introduces conservative updates to prevent overestimation of Q-values, enhancing the stability and reliability of the learning process.

Standard Q-learning methods often suffer from overestimation errors, especially when out-of-distribution (OOD) actions are encountered. CQL addresses this by learning conservative Q-values that aim to lower-bound the true value, thus mitigating overestimation.

The iterative update function for the Conservative Q-Learning (Kumar et al., 2020):

$$\widehat{Q}_{k+1} \leftarrow \arg \min_Q \left[\alpha E_{s \sim D, a \sim \mu(a|s)} [Q(s, a)] + \frac{1}{2} E_{s, a \sim D} \left[\left(Q(s, a) - \widehat{B}^\pi \widehat{Q}_k(s, a) \right)^2 \right] \right], (4)$$

The Q-function $Q(s, a)$ is the expected cumulative reward of taking action a in state s ; (α) is a tradeoff factor that determines the weight of the regularization term in the objective function; (D) is the dataset of state-action-reward tuples that pertains to what was collected from the environment; ($\mu(a|s)$) is a distribution over actions given states; ($\widehat{B}^\pi \widehat{Q}_k(s, a)$) is the Bellman backup operator, which calculates the target Q-value using the current policy (π) and the Q-function ($\widehat{Q}_k$); ($\widehat{Q}_k$) is the Q-function estimate at iteration (k). Compared to the standard update of the Q-Learning algorithm, CQL introduces two core modifications, the Regularization Term, and the Bellman Error Term. The Regularization term written as: ($\alpha E_{s \sim D, a \sim \mu(a|s)} [Q(s, a)]$), penalizes the Q-values in order to guarantee conservativeness, while the Bellman Error Term is defined by: ($\frac{1}{2} E_{s, a \sim D} \left[\left(Q(s, a) - \widehat{B}^\pi \widehat{Q}_k(s, a) \right)^2 \right]$), (Kumar et al., 2020).

This simplified meta-algorithm captures the essence of all three algorithms by separating them through the usage of colors which indicate how to switch between them.

```

For each episode:
  Initialize state s
  Repeat (for each step of the episode):
    Choose action a using  $\epsilon$ -greedy policy from Q-table Q(s, a)
    Choose action a using  $\epsilon$ -greedy policy from Q-network Q(s, a;  $\theta$ )
    Execute action a, observe reward r, and next state s'
    # Update Q-table for Classical Q-learning
    Q(s, a) = Q(s, a) +  $\alpha$  * [ r +  $\gamma$  * max_a' Q(s', a') - Q(s, a) ]
    Else if using DQN or CQL:
      Store transition (s, a, r, s') in replay buffer
      Sample minibatch of transitions from replay buffer
      # Compute Loss
      L_standard = [ r +  $\gamma$  * max_a' Q(s', a';  $\theta^-$ ) - Q(s, a;  $\theta$ ) ]2
      L_conservative =  $\alpha_{cql}$  * ( Q(s, a;  $\theta$ ) - Expected Q over data )
      L = L_standard + L_conservative
      L = [ r +  $\gamma$  * max_a' Q(s', a';  $\theta^-$ ) - Q(s, a;  $\theta$ ) ]2
      Update Q-network by minimizing L
      If update_target_network:
         $\theta^- = \theta$ 
      Set s = s'
  Until s is terminal
End

```

Figure 4. Pseudo Code for the meta-algorithm encompassing the Classical Q-learning (red), the Conservative Q-learning (blue) and the Q-Networks (green)

3.2.2. Neural Network Applications

Neural Network Applications was chosen as a classification for aspects of Neural Network related algorithms and mathematical structures that could be related to Generative Models, without explicitly being Transformer Architectures.

3.2.2.1. Embeddings

An embedding in the field of ML involves converting raw data (like words, images, states, or actions) into a high-dimensional vectors that capture some aspect of the data's underlying structure or semantics.

Embeddings can be created through the weights of Neural Networks, Encoders, for textual data, words embedding models like Word2Vec or through dimensionality reduction techniques like PCA, t-SNE and UMAP.

The papers analyzed include different levels of abstractions, from high-level embeddings, such as behaviors in the trajectory (Liu et al., 2023), skill embeddings (Sudhakaran & Risi, 2023), as well as lower-level embeddings for states and actions (Shi et al., 2023)

3.2.2.2. VAEs (Variational Autoencoders)

A type of generative model that learns to encode inputs into a distribution over latent space and then decode from this space to reconstruct the input.

The VAE is a generative model composed by an Encoder, represented by $(q_\phi(z|x, y))$, with conditions on the input data and output a distribution over the latent space, where z represents the latent space and x the input data, where y is the additional input, and the Decoder $(p_\theta(x|z, y))$, which takes a sample from the latent space z and rebuilds x , influenced by the additional input y . The aim of the VAE is to maximize the marginal log-likelihood (Fujimoto et al., 2019):

$$\mathcal{L}(\theta, \phi; x) = -\text{KL}\left(q_\phi(z|x)p_\theta(z)\right) + E_{q_\phi(z|x)}[\log p_\theta(x|z)], \quad (5)$$

Where the Kullback-Leibler divergence serves the purpose of regularizing, as we will further discuss in the Mathematical Artifacts section of the thesis, and θ and ϕ are, respectively, the parameters of the decoder and encoder, x is the observed data, z the latent variable, $q_\phi(z|x)$ is the VAE's approximation of the posterior of latent variables given the input data x , while $p_\theta(z)$ is the likelihood of the data given latent variables, with the modelling coming from the Decoder (Kingma & Welling, 2022), (Ramchandran et al., 2022).

A frequent type of VAE is the Conditional Variational Auto-encoder, used in multiple surveyed papers (Fujimoto et al., 2019), (Venkatraman et al., 2023), (Liu et al., 2023) besides the mentioned θ and ϕ parameters, and q and p functions, is represented by:

$$\mathcal{L}(\theta, \phi; y, x) = E_{q_\phi(z|y, x)}[\log p_\theta(y|z, x)] - \text{KL}[q_\phi(z|y, x)|p(z|x)], \quad (6)$$

Differing with the standard VAE in that there is an additional generation conditioning based on the y , which are additional output (or observed) variables (Ramchandran et al., 2022).

3.2.3. Transformers in RL

3.2.3.1. Transformers and Decision Transformers

The Transformer Algorithm was proposed in 2017 by Vaswani in the (Vaswani et al., 2023) “Attention is all you need” paper, this algorithm is based on the Encoder and Decoder Architecture with the added component of Attention, this algorithm was built for sequence transduction (Vaswani et al., 2023).

As this algorithm allows the modeling of sequence-based data, it can be applied to not only textual data, but also to trajectory sequences, such as the known trajectory sequence used in Reinforcement Learning:

$$\tau = (s_1, a_1, r_1, s'_1, s_2, a_2, r_2, s'_2, \dots), \quad (7)$$

The Self-attention algorithm receives n -embeddings as input and output the same dimension of n -embeddings. As a matrix, the output z_i for the i -token in a sequence is calculated through the application of a weighted sum of the value vectors v_j , and these weights are derived with the softmax of the dot products between the query (q) and all keys (k) vectors.

$$z_i = \sum_{j=1}^n \text{softmax}(\{q_i \cdot k_j\}_{j=1}^n)_j \cdot v_j, \quad (8)$$

The Decision Transform corresponds to the application of the input sequencing of the trajectory displayed in the equation 7, and is similar in nature to the original Transformer architecture (Vaswani et al., 2023).

3.2.3.2. Trajectory Transformer

Similar to the Decision Transformer in the sense that it is also a sequence prediction algorithm based on the trajectories, the Trajectory Transformer focuses on beam-search-based planning instead of the Reward Conditioning focus present in the Decision Transformer (Janner & Levine, 2021).

In essence, the beam-search is a deterministic algorithm that prunes the search space based on the ranking of the prefixes, maintaining only a k number of top ranked nodes at each step (Shaham & Levy, 2022).

3.3. MATHEMATICAL ARTIFACTS

In this section, we explore foundational mathematical artifacts that underpin optimization and statistical methods used in reinforcement learning (RL). These concepts are crucial for understanding advanced algorithms and for formulating and solving complex RL problems.

3.3.1. Optimization Frameworks and Formulations

These frameworks and formulations are foundational mathematical structures and problem-solving strategies that define how optimization problems are set up and solved.

3.3.1.1. Constrained Optimization

A formulation of Constrained Optimization from Altman (Altman, 2021) defines the base constrained optimization problem (COP) as follows:

$$\text{minimize } C(u), \text{ subject to } D_k(u) \leq V_k, \text{ for } k = 1, \dots, K,$$

Where the policy is denoted by $u \in U$, $C(u)$ the cost function for policy u , this cost function can be a total expected cost, a discounted cost, or expected average cost; the $(D_k(u))$ are other cost criteria bounded by (V_k) , which is the set of limits for cost criteria.

The challenge in this optimization task is to find the optimal policy while minimizing its cost and respecting the imposed constraints (Altman, 2021).

In RL, constrained optimization is essential when agents must not only maximize rewards but also adhere to safety, resource, or performance constraints, such as in Constrained Markov Decision Processes (CMDPs).

3.3.1.2. Lagrange Multipliers

One such approach used in constrained optimization problems is the Lagrange Multipliers. In CMDPs and other constrained RL settings, Lagrange multipliers are used to handle constraints by incorporating them into the optimization objective, enabling efficient policy optimization.

Applied generically to a minimization problem of a function $f(x)$ subject to a constraint $g(x) = 0$, which can also be expressed as (Miura, 2021):

$$\min_x f(x), \quad g(x) = 0, \quad (1)$$

This minimization problem is convertible into an unconstrained problem when the dual variable, the Lagrange multiplier λ , is applied. This results in the creation of the Lagrangian:

$$L(x, \lambda) = f(x) + \lambda g(x), \quad (2)$$

which facilitates finding the solution through (Stooke et al., 2020):

$$(x^*, \lambda^*) = \arg \max_{\lambda} \min_x L(x, \lambda), \quad (3)$$

3.3.2. Statistics Methods

Statistical methods are fundamental in estimating models and parameters from data, which is critical in model-based RL and in understanding the uncertainty inherent in the learning process.

3.3.2.1. Maximum Likelihood Estimation

The maximum likelihood method is a key statistical approach for estimating the parameters of a probability distribution, tailored to both discrete and continuous stochastic variables. In the discrete case, it involves finding the parameter values that maximize the probability of observing the specific outcomes given the distribution. For continuous variables, it focuses on maximizing the probability density function across observed data points. This technique hinges on the assumption that the observed data are independent samples from the distribution, utilizing their collective likelihood to derive the most probable parameters that explain the observed data (Miura, 2021).

Two cases from the papers analyzed of how the Maximum Likelihood Estimation can be applied are presented:

Case 1: Training Dynamics Models in TATU (Zhang et al., 2023)

The purpose of using the Maximum Likelihood Estimator in this case is to predict the next state in an environment using an ensemble of dynamics models that output Gaussian distributions.

The Log Likelihood Function comes in the form:

$$L_{\Psi} = E_{(s,a,r,s') \sim D} [-\log \widehat{p}_{\Psi}(s'|s,a)], \quad (4)$$

This function is used to maximize the accuracy of predicting the next state, aiming to improve the reliability of the dynamics models through parameter optimization. And, therefore, the optimization goal in this case is the enhancing of model predictions by training each model in the ensemble independently to maximize the log likelihood of observed transitions.

Case 2: Recovering Human Policy and Human State-Action Values (Li et al., 2023)

In this second case, the Likelihood Estimator's purpose is the estimation of human behavior policies and corresponding state-action values using a softmax log likelihood approach.

Defined by:

$$L_h(Q) = \frac{1}{n} \sum_{i=1}^n \log \left(\frac{\exp(Q(s_i^h, a_i^h))}{\sum_{a'} \exp(Q(s, a'))} \right), \quad (5)$$

This expression estimates the probability of actions based on their utility, aiming to align the estimated policy with observed human decisions.

Which in turn signifies that the optimization goal of this MLE is towards maximizing this function to accurately model human behavior in decision-making processes.

In these cases, the log likelihood is pivotal for model parameter estimation, with each application leveraging this approach to optimize different aspects of system behavior or decision-making processes under uncertainty.

3.3.2.2. Confidence Set/Regions; Hoeffding's and Bernstein's Inequalities

In reinforcement learning (RL), confidence sets are used in algorithms that require uncertainty quantification, such as in exploration strategies and robust policy optimization. An example of how confidence sets have been applied in offline reinforcement learning is the Distributionally Robust Q-Iteration (DRQI) algorithm (Panaganti et al., 2023), where confidence sets are constructed around estimated transition probabilities. By defining uncertainty sets based on KL divergence, DRQI accounts for the uncertainty in transition estimates. This approach ensures that the policy optimization considers worst-case scenarios within these confidence sets, thereby mitigating the adverse effects of distribution shifts between the behavior policy and the learned policy.

A confidence set, denoted as $\Theta \subseteq R^d$, is a set that contains the true parameter θ^* with high probability. This set reflects the uncertainty surrounding the true parameter estimates obtained from offline data. The size and shape of Θ depend on the statistical properties of the estimator used to approximate θ^* and the variance inherent in the data (Li et al., 2022).

The Hoeffding's and Bernstein's Inequalities are probabilistic Inequalities that can be used in creating uncertainty sets that guide the way transitions occur in more conservative ways by providing bounds.

The Hoeffding's Inequality comes from defining independent random variables $(X_1, X_2, \dots, X_n)$ each bounded by intervals $([a_i, b_i])$, the sum $(S_n = X_1 + X_2 + \dots + X_n)$ satisfies the following bound (Carcassi et al., 2021):

$$P(|S_n - E[S_n]| \geq t) \leq 2 \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \quad (6)$$

Expected Value ($E[S_n]$): The expected value of the sum, $E[S_n]$, while the Deviation Threshold t represents the amount by which the sum S_n is allowed to deviate from its expected value.

The probability bound, L.H.S, represents the probability of the absolute deviation of S_n from its expected value being at least t . The R.H.S is the Exponential Bound, which indicates an exponential decrease in probability as t increases, controlled by the sum of the squares of the ranges.

The Bernstein's Inequality differs from the Hoeffding's Inequality in the sense that it takes into account the variances between the variables of the bounds.

With a similar formulation than For independent random variables $(X_1, X_2, \dots, X_n)$ that satisfy $(|X_i - E[X_i]| \leq M)$ for all (i) and some constant (M) , the sum $(S_n = \sum_{i=1}^n X_i)$ satisfies (Serfling, 1974):

$$P(|S_n - E[S_n]| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\left(\sigma^2 + \frac{1}{3}Mt\right)}\right), \quad (7)$$

where $\sigma^2 = \sum_{i=1}^n \text{Var}(X_i)$ is the sum of the variances of the X_i 's, the expectation $(E[X_i])$ corresponds to the expected values of the variables. M is known as the deviation constraint, which bounds the maximum deviation of each variable from its expected value, in such a way that each variable X_i doesn't have the absolute deviation from its expected value does not exceeding M .

Additionally, the Deviation Threshold t , is the threshold for the deviation of S_n from $(E[S_n])$ being considered. Standing on the R.H.S, the Probability and Exponential Bound is the probability that the absolute deviation of S_n from $s(E[S_n])$ exceeds t , which is bound by t, σ^2 and M (Serfling, 1974).

3.3.3. Information Theory

Information theory provides measures to quantify information, uncertainty, and the divergence between probability distributions, which are fundamental in RL for understanding and controlling randomness and information flow.

3.3.3.1. Shannon Entropy

A central introductory concept in Information Theory is the Shannon Entropy $(H(X))$, this concept quantifies the amount of uncertainty in a random variable (X) . Entropy is used in exploration strategies (e.g., entropy regularization) to encourage agents to explore uncertain or less-visited states.

The Shannon entropy $H(X)$ of a discrete random variable X with possible values $(\{x_1, x_2, \dots, x_n\})$ and probability mass function $(P(X = x_i) = p_i)$ is defined as (Carcassi et al., 2021):

$$H(X) = - \sum_{i=1}^n p_i \log_2(p_i), \quad (8)$$

where the sum is over all possible outcomes that have non-zero probability, and the logarithm is typically taken to base 2, making the unit of entropy bits. The Mathematical Properties include it being non-negative, meaning that the entropy is only zero if a random variable is

constant. The Maximum Entropy is given when all states are equally likely and it is Additive, which means that for two independent random variables (X) and (Y), the entropy of the joint distribution is the sum of their individual entropies: ($H(X, Y) = H(X) + H(Y)$) (Carcassi et al., 2021).

3.3.3.2. Information-Theoretic Metrics for Addressing Distribution Shift

To effectively mitigate distribution shift in offline reinforcement learning, it is essential to adjust the policy's probability distribution to remain consistent with the distribution of the offline dataset. Two key information-theoretic measures, Mutual Information and Kullback-Leibler (KL) Divergence, play crucial roles in this adjustment process. Mutual Information quantifies the dependency between different variables, enabling the policy to retain relevant information about the environment while minimizing irrelevant variations. Meanwhile, KL Divergence provides a direct metric for measuring and constraining the difference between the current policy and the behavior policy that generated the offline data. Together, these measures facilitate the correction of probability distributions, ensuring that policy updates remain aligned with the observed data and thereby effectively addressing the challenges posed by distribution shifts.

The mutual information is a general measure of the dependence between two random variables. It expresses the quantity of information one has obtained on X by observing Y (Batina et al., 2011). Mutual information is used to measure the dependence between variables, such as in mutual information maximization techniques that encourage agents to learn representations capturing relevant information about the environment. The Mutual Information can be understood in slightly different ways, here we try to understand three of these approaches. These approaches are the Expected KL Divergence between joint and marginal distributions, the Expected Divergence between Condition and Marginal Distributions, and also as a function of Entropy.

The expected Kullback–Leibler Divergence between Two Distributions (joint and marginal distributions) (Ma et al., 2024):

$$I(X; Y) = \sum_{x \in X, y \in Y} \Pr[X = x, Y = y] \cdot \log \left(\frac{\Pr[X = x, Y = y]}{\Pr[X = x] \cdot \Pr[Y = y]} \right), \quad (9)$$

The Kullback–Leibler divergence between the joint probability distribution ($\Pr[X = x, Y = y]$) and the product of the marginal distributions ($\Pr[X = x] \cdot \Pr[Y = y]$).

As the Expected Divergence of Conditional and Marginal Probabilities (Ma et al., 2024):

$$I(X; Y) = \sum_{x \in X} \Pr[X = x] \sum_{y \in Y} \Pr[Y = y | X = x] \cdot \log \left(\frac{\Pr[Y = y | X = x]}{\Pr[Y = y]} \right), \quad (10)$$

Here, mutual information is expressed as the expected value of the divergence between the conditional probability ($\Pr[Y = y|X = x]$) and the marginal probability ($\Pr[Y = y]$).

As a Function of Entropy (Batina et al., 2011):

$$[I(X; Y) = H[X] + H[Y] - H[X, Y]], \quad (11)$$

Alternatively, it can also be represented using:

$$[I(X; Y) = H[X] - H[X|Y] = H[Y] - H[Y|X] = H[X, Y] - H[X|Y] - H[Y|X]], \quad (12)$$

These different approaches to Mutual Information expressed as part of the distinct formulations serve to express how Mutual Information has a role as a measure of the amount of information that one variable has within it about another variable, or, in a statistical sense, their dependence (Batina et al., 2011). One case where mutual information is applied in the context of Offline Reinforcement Learning is present in the Ma et al. “Mutual Information Regularized Offline Reinforcement Learning” paper. In which the stated benefit of introducing the Mutual Information is to allow not only the agent to perform the updates of the policies within the dataset distribution, but also to avoid it being excessively conservative with this dataset information (Ma et al., 2024).

As seen in the definition above, Kullback-Leibler Divergence is frequently used in a multitude of algorithms, papers and mentioned throughout the thesis. KL divergence is widely used in policy optimization algorithms (e.g., Trust Region Policy Optimization, KL-regularized RL) to measure the difference between the current policy and the target or previous policies, ensuring stable updates. The Kullback-Leibler (KL) divergence serves to quantify the difference between two probability distributions (p) and (q). The Mathematical expression for the KL divergence from (q) to (p) is defined as follows:

$$[D_{KL}(p \parallel q) = \sum_i p_i \log_2 \left(\frac{p_i}{q_i} \right)], \quad (13)$$

Where ($p = \{p_i\}$) is the true probability distribution while ($q = \{q_i\}$) represents the model distribution. The properties of the KL Divergence are its Non-negativity $D_{KL}(p \parallel q) \geq 0$, meaning that a KL Divergence is zero if and only if (p) and (q) are identical distributions; Asymmetry, $D_{KL}(p \parallel q)$ is not symmetric, meaning $D_{KL}(p \parallel q) \neq D_{KL}(q \parallel p)$, And its Unboundedness, which pertains to the possibility that the divergence to be infinite (Shlens, 2014).

Similar to the concept of Mutual Information, the Fisher information is defined as a measure of the amount of information that an observable random variable (X) contains about an unknown parameter (θ) upon which the probability depends. The formal mathematical definition is given by (Liu et al., 2024):

- For discrete (X):

$$[I_X(\theta) = \sum_{x \in X} \left(\frac{d}{d\theta} \log f(x | \theta) \right)^2 p_\theta(x)], \quad (14)$$

- For continuous (X):

$$[I_X(\theta) = \int_X \left(\frac{d}{d\theta} \log f(x | \theta) \right)^2 p_\theta(x) dx], \quad (15)$$

Where ($f(x | \theta)$) is the probability function of (X), representing how the parameter (θ) is related to the outcomes (x) of (X). In the discrete case, the probability function is to be a probability mass function and in the continuous case, it is a probability density function, ($p_\theta(x) = f(x | \theta)$) denotes the pmf or pdf of X at x given (θ).

- ($\frac{d}{d\theta} \log f(x | \theta)$) is known as the score function, which quantifies the sensitivity of the log-likelihood of the observed outcome (x) to changes in the parameter (θ).

The Fisher information is a foundational concept in statistical estimation and testing, crucial for constructing confidence intervals, conducting hypothesis tests, and in defining default priors in Bayesian statistics. The Fisher information essentially captures how informative a given sample or experiment is about the parameter (θ) it intends to estimate (Ly et al., 2017).

4. INSIGHTS IN OFFLINE REINFORCEMENT LEARNING

Understanding the landscape of existing research is a crucial step in any field. This literature review aims to provide a comprehensive analysis of the prevailing themes, trends, and interrelationships within the body of literature related to Offline Reinforcement Learning. By employing frequency analysis techniques, this chapter quantitatively examines the literature to understand potential patterns that may not be immediately apparent through intuitive reading or qualitative review alone.

Frequency analysis serves as a tool to systematically assess how often certain themes, concepts, or variables appear within a set of publications. In this review, three primary analytical methods are utilized:

1. **Total Frequency Histograms:** These visual representations illustrate the distribution of specific themes or keywords across all reviewed literature. By displaying the frequency with which each theme occurs, the histograms highlight areas of significant scholarly focus and identify less-explored topics that may warrant further investigation.
2. **Trend Stacked Histograms:** Building upon the total frequencies, trend stacked histograms depict how the prominence of various themes has evolved over time or across different categories. This temporal or categorical analysis tentatively displayed some shifts in research focus, emerging topics, and the potential impact of external factors on scholarly attention.
3. **Confounding Correlation Matrices:** To explore the relationships between categorical variables within the literature, Confounding Matrices are employed as a measure of association. Confounding Matrices are used to assess the strength of association between two nominal (categorical) variables, by visually and numerically identifying significant associations between themes, methodologies, or other categorical variables, shedding light on underlying patterns and connections in the research field.

The integration of these quantitative methods aims to improve the depth of the literature review. By systematically quantifying the presence and interplay of various elements within the literature, the analysis moves beyond mere enumeration to provide insights into the structural dynamics of the research area.

In the discussion, these findings are interpreted in the context of the existing body of knowledge, highlighting their significance, implications, and the potential directions they suggest for future research.

The Frequency Analysis yielded the following results for the investigated columns present in the 140 papers analyzed. These histograms reflect how often the techniques were applied as a main concept in the development of the paper at hand.

4.1. TOTAL FREQUENCY HISTOGRAMS

Below are the total frequency histograms illustrating the top 10 most prevalent themes identified across the 140 analyzed papers on Offline Reinforcement Learning. These visualizations provide a quantitative overview of the primary concepts emphasized in the literature, highlighting the areas of significant focus.

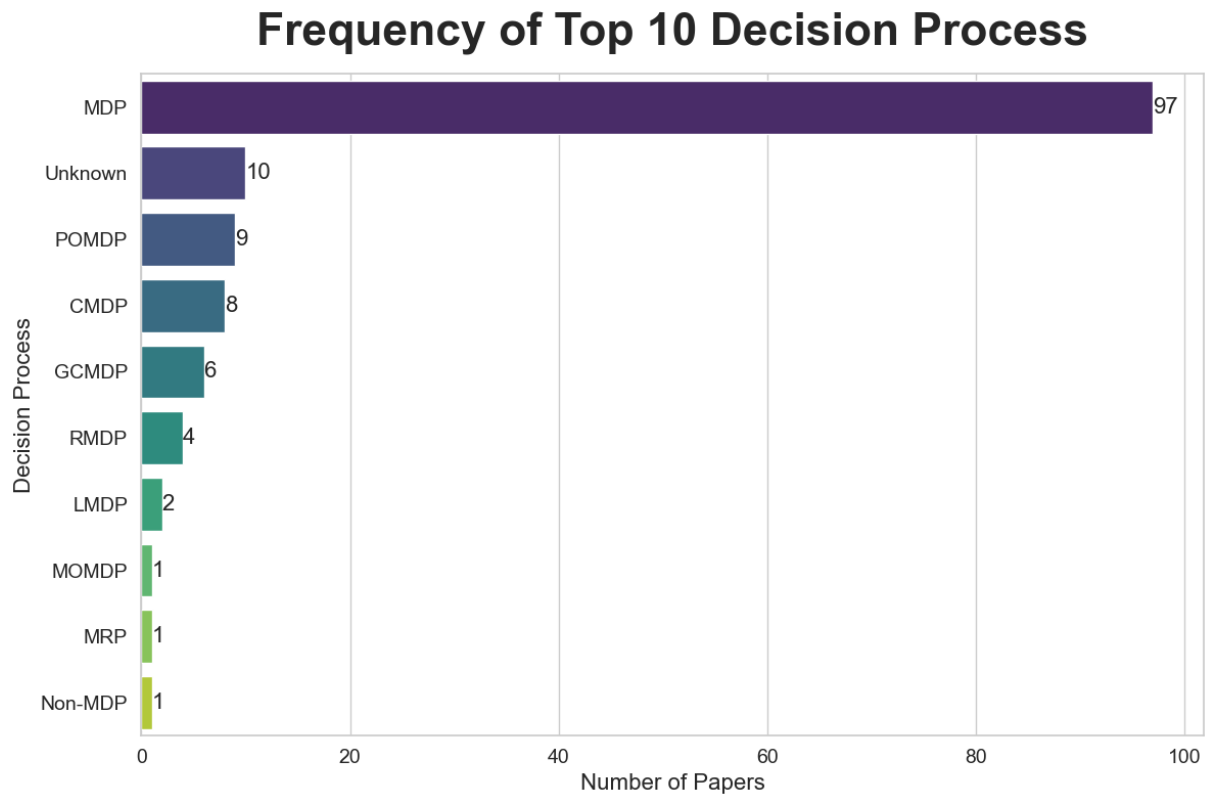


Figure 5. Number of paper frequency of the Top Decision Processes

The frequency analysis of decision process models presented in reveals a predominant emphasis on Markov Decision Processes (MDPs), which appear in 97 out of 140 papers. This highlights the foundational role of MDPs in the field. Additionally, models such as Partially Observable MDPs (POMDPs) and Constrained MDPs (CMDPs) are notably present, appearing 9 and 8 times respectively, indicating the interest in scenarios with partial observability and constraints which are particularly relevant for the Offline-Setting aspect of Offline Reinforcement Learning.

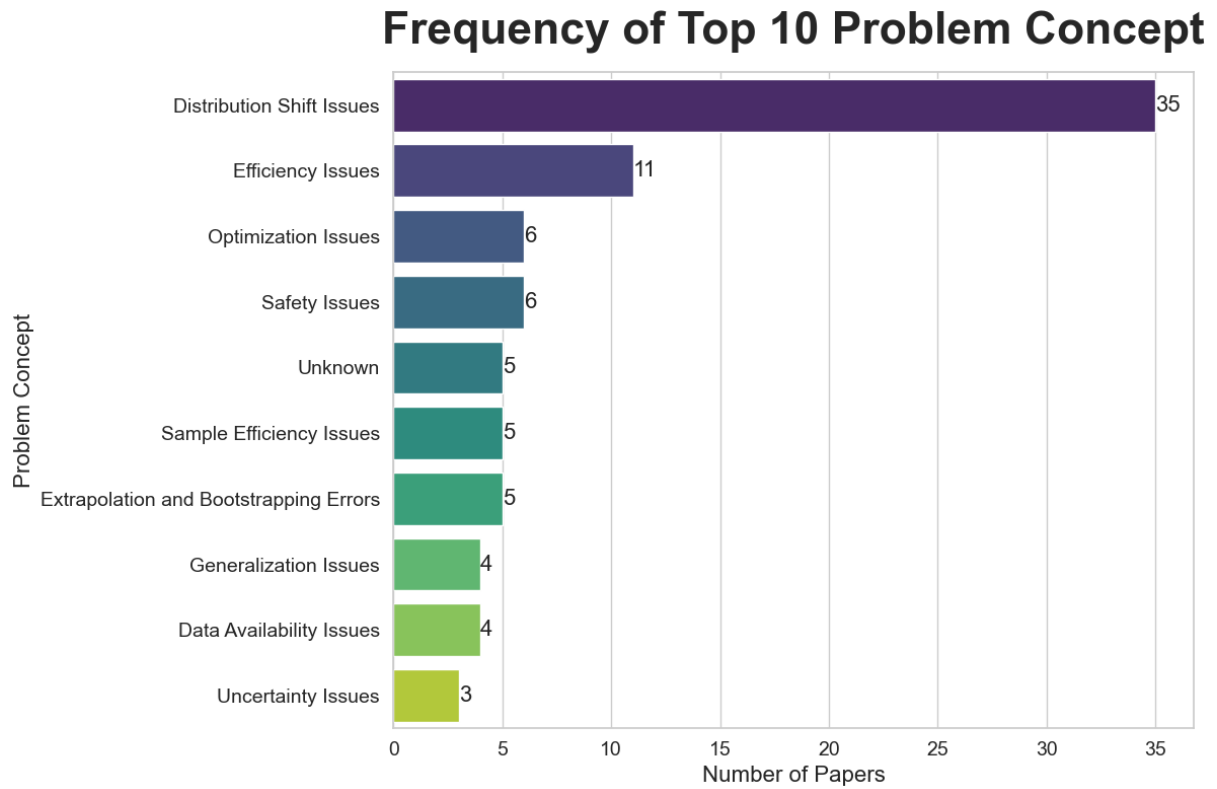


Figure 6. Number of paper frequency for the Standardized Problem Concepts of the papers

The frequency analysis of problem concepts in the Offline Reinforcement Learning literature reveals Distribution Shift (DS) Issues as the most prominent, appearing in 35 out of 140 papers. This underscores the critical challenge in ORL, where agents learn from fixed datasets without ongoing interaction, leading to discrepancies between training and deployment data distributions. Addressing DS is essential for ensuring that learned policies generalize effectively to real-world environments, maintaining performance and reliability despite these shifts. Efficiency Issues, occurring in 11 papers, reflect ongoing efforts to optimize computational efficiency, while Optimization and Safety Issues, each appearing six times, indicate a growing focus on enhancing algorithm performance and ensuring safe deployments. Additionally, lower frequencies in Generalization, Data Availability, and Uncertainty Issues point to emerging yet less central research areas.

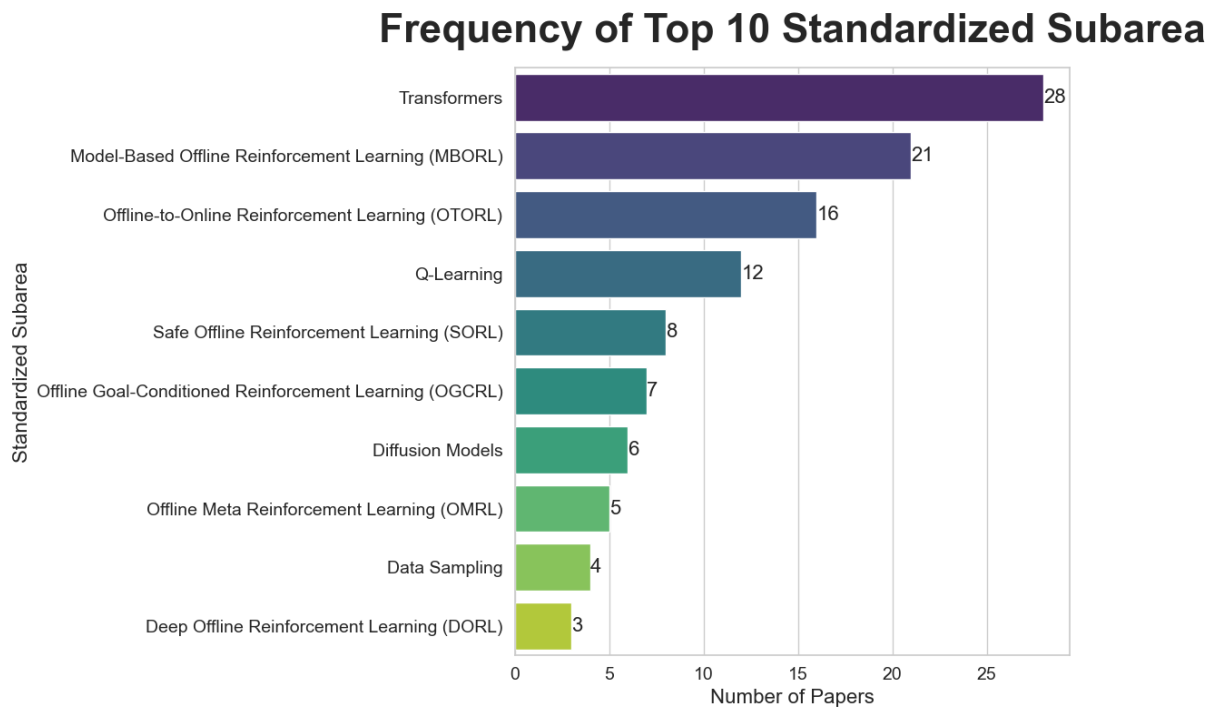


Figure 7. Frequency for the Standardized Subarea categorizing the main algorithms presented in the respective paper

This histogram reflects the frequency analysis of the subareas that were categorized in this study of Offline Reinforcement Learning (ORL), highlighting Transformers as the most studied topic, appearing in 28 papers, likely due to their powerful capability in handling complex sequential data. Model-Based Offline Reinforcement Learning (MBORL) follows with 21 mentions, reflecting a strong interest in leveraging predictive models to improve learning efficiency. Offline-to-Online Reinforcement Learning (OTORL) is featured in 16 studies, indicating efforts to seamlessly transition policies from offline training to real-world applications. Q-Learning remains a foundational method with 12 occurrences. Emerging areas like Safe Offline Reinforcement Learning (SORL) and Offline Goal-Conditioned Reinforcement Learning (OGCRL) show growing attention to safety and specific objectives.

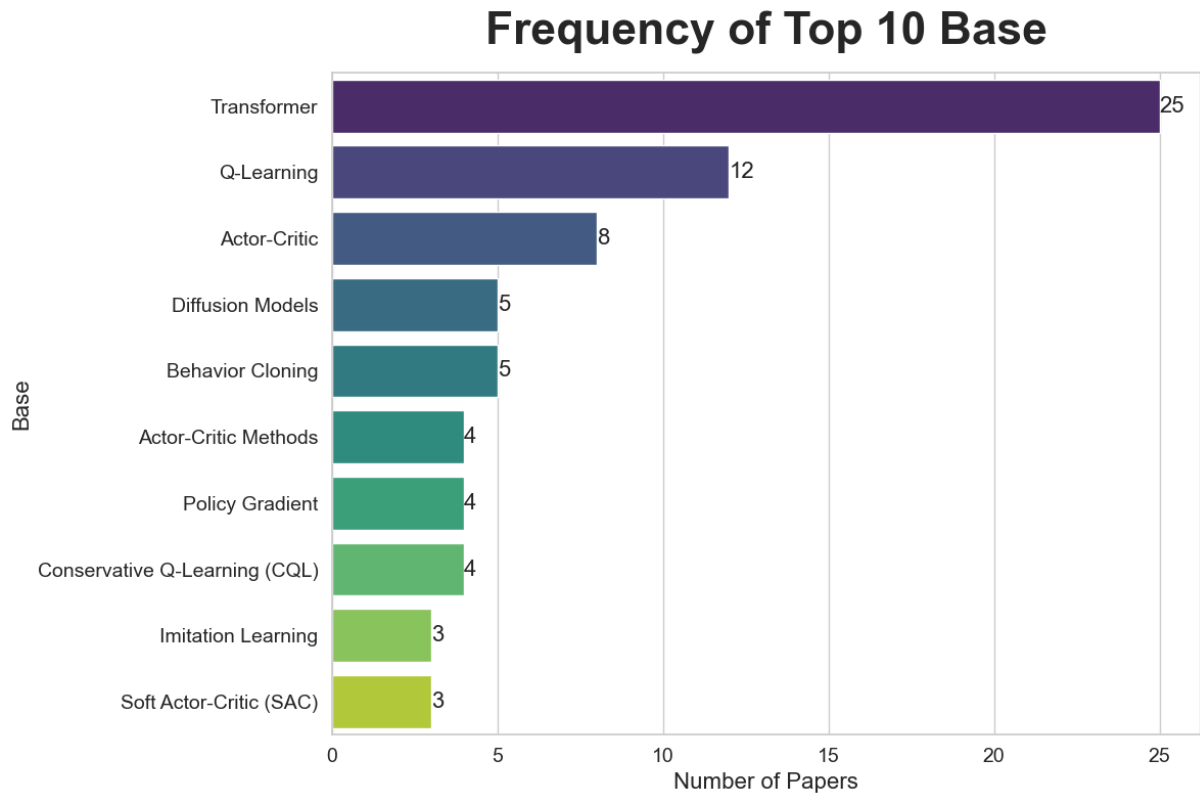


Figure 8. Frequency for the Base Algorithm across the papers

The frequency analysis of base algorithms in the Offline Reinforcement Learning (ORL) literature highlights Transformers as the most prevalent, appearing in 25 out of 140 papers. This prominence aligns with their rising popularity as a subarea, leveraging transformers' powerful capabilities in handling complex sequential data to enhance policy learning and decision-making processes. Q-Learning follows with 12 paper usages, reaffirming its foundational role in value-based methods and its continued relevance in optimizing policies. Actor-Critic methods, featured in 8 studies, demonstrate their effectiveness in balancing policy and value estimation, fostering more stable and efficient learning. Due to the relation between the Base algorithm and the Subarea of a given paper, it is expected that the frequencies align strongly between the two histograms.

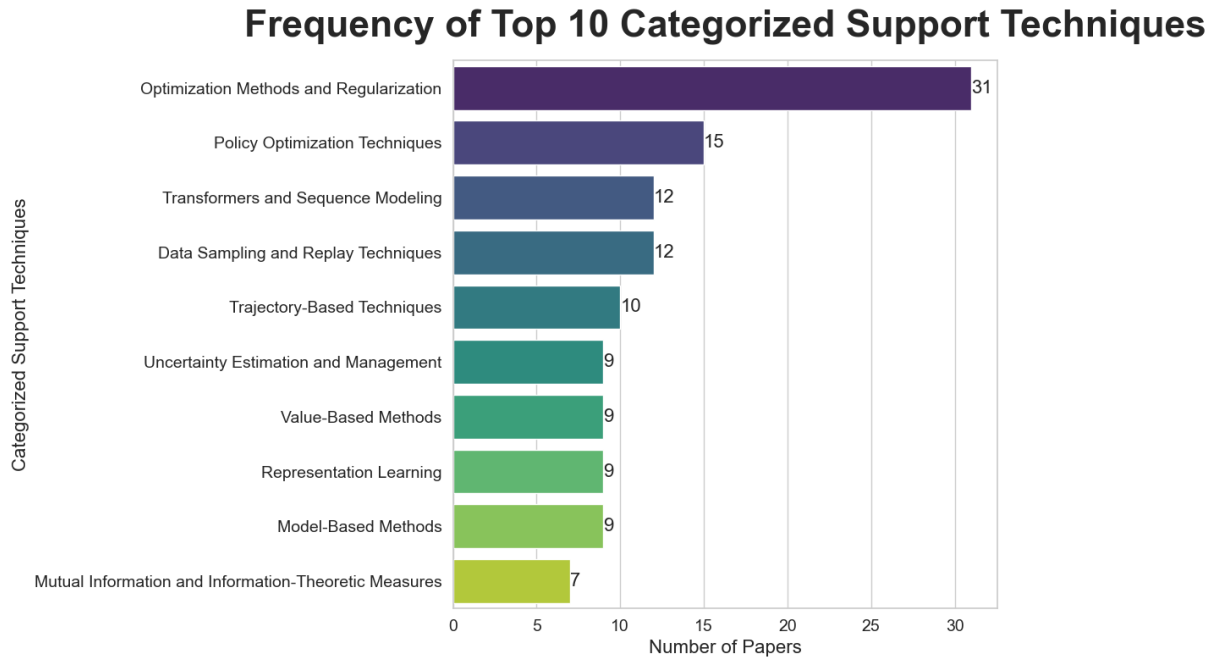


Figure 9. Frequency for the Categorized Support Techniques used to support the main algorithm of the paper

The analysis of categorized support techniques in Offline Reinforcement Learning (ORL) literature highlights Optimization Methods and Regularization as the most prevalent, appearing in 31 out of 140 papers. This emphasizes the critical role of robust optimization and regularization in enhancing algorithm performance and ensuring generalization from offline data. Policy Optimization Techniques, Transformers and Sequence Modeling, and Data Sampling and Replay Techniques each feature prominently, reflecting ongoing efforts to refine policy learning, leverage advanced model architectures, and efficiently utilize limited data. Additionally, emerging methods such as Trajectory-Based Techniques demonstrate a focus on learning from sequential data.

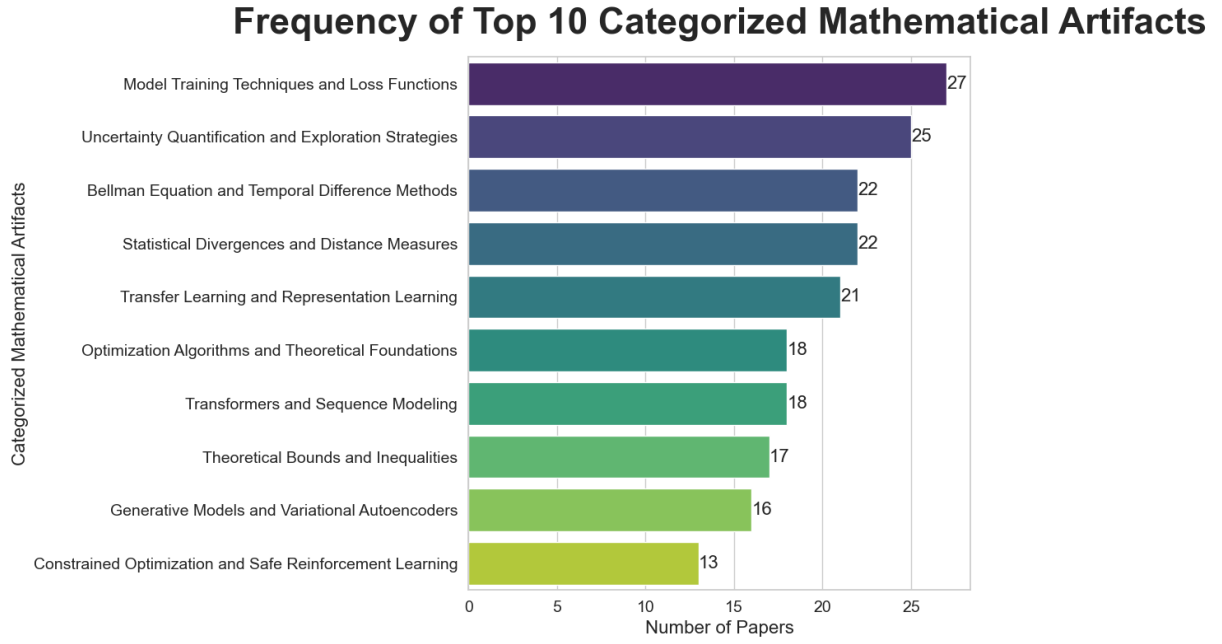


Figure 10. Frequency for the Categorized Support Techniques used to support the main algorithm of the paper

The frequency analysis of categorized mathematical artifacts in Offline Reinforcement Learning (ORL) literature highlights Statistical Inequalities and Theoretical Analysis as the most common, appearing in 37 out of 140 papers. This indicates a strong focus on establishing solid theoretical foundations for ORL algorithms. Optimization Algorithms and Gradient-Based Methods follow with 33 mentions, emphasizing their role in improving learning efficiency and performance. Fundamental concepts like the Bellman Equation and Temporal Difference Methods appear in 21 papers, underscoring their ongoing importance in reinforcement learning. Additionally, Uncertainty Estimation, Kullback-Leibler Divergence, and Sequence Modeling are frequently referenced, reflecting efforts to manage uncertainty, measure distribution differences, and utilize advanced modeling techniques. Overall, these mathematical artifacts are crucial for developing robust and effective ORL systems.

4.2. STACKED HISTOGRAMS

The stacked histograms presented below aim to illustrate the distribution and temporal evolution of key support techniques across the 140 analyzed papers on Offline Reinforcement Learning (ORL).

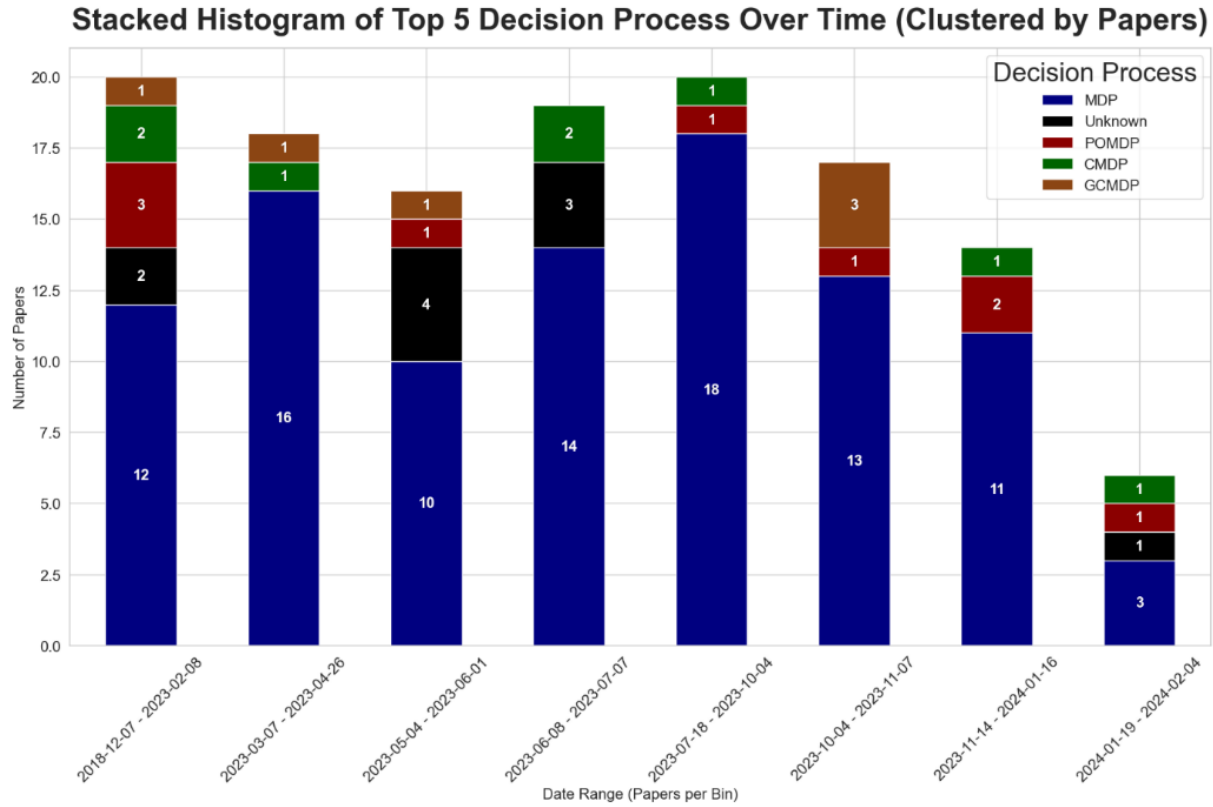


Figure 11. Stacked Histogram of the TOP 5 Decision Processes over Time

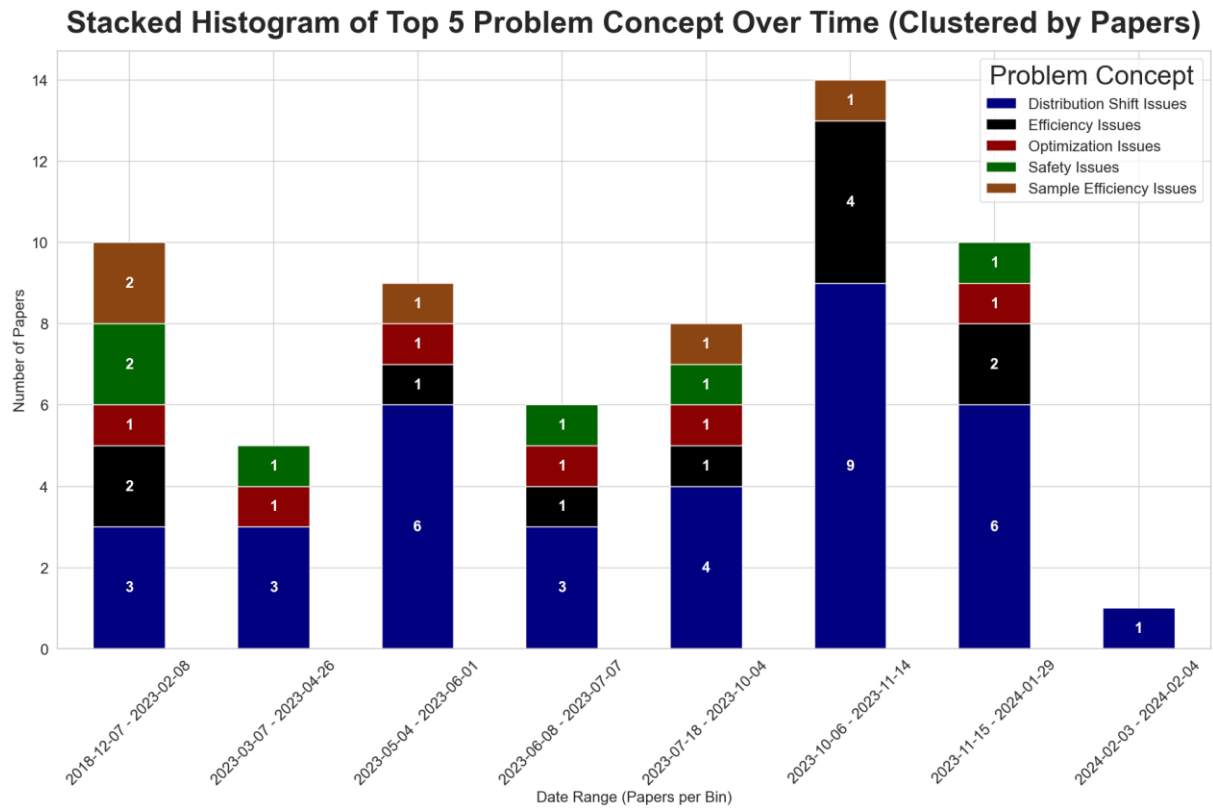


Figure 12. Stacked Histogram of the TOP 5 Standardized Problem Concepts over Time

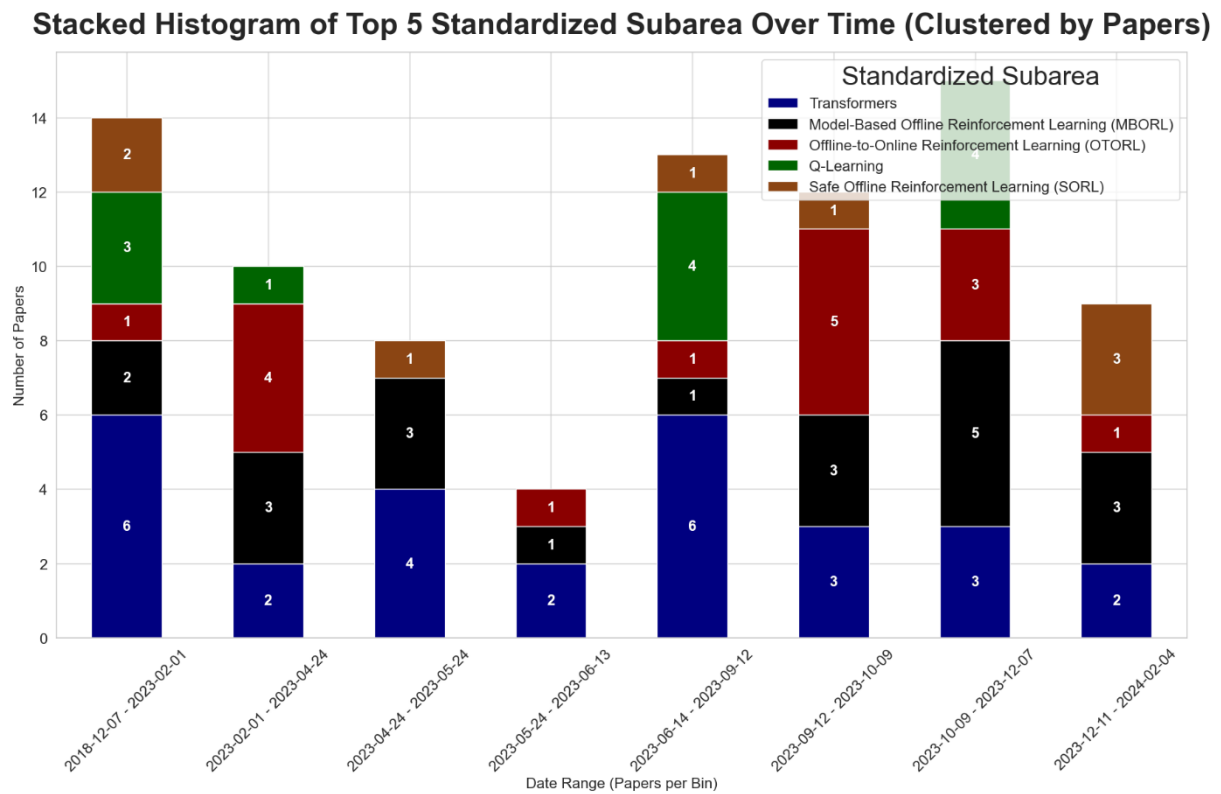


Figure 13. Stacked Histogram of the TOP 5 Standardized Subarea of the relevant algorithm mentioned in the paper

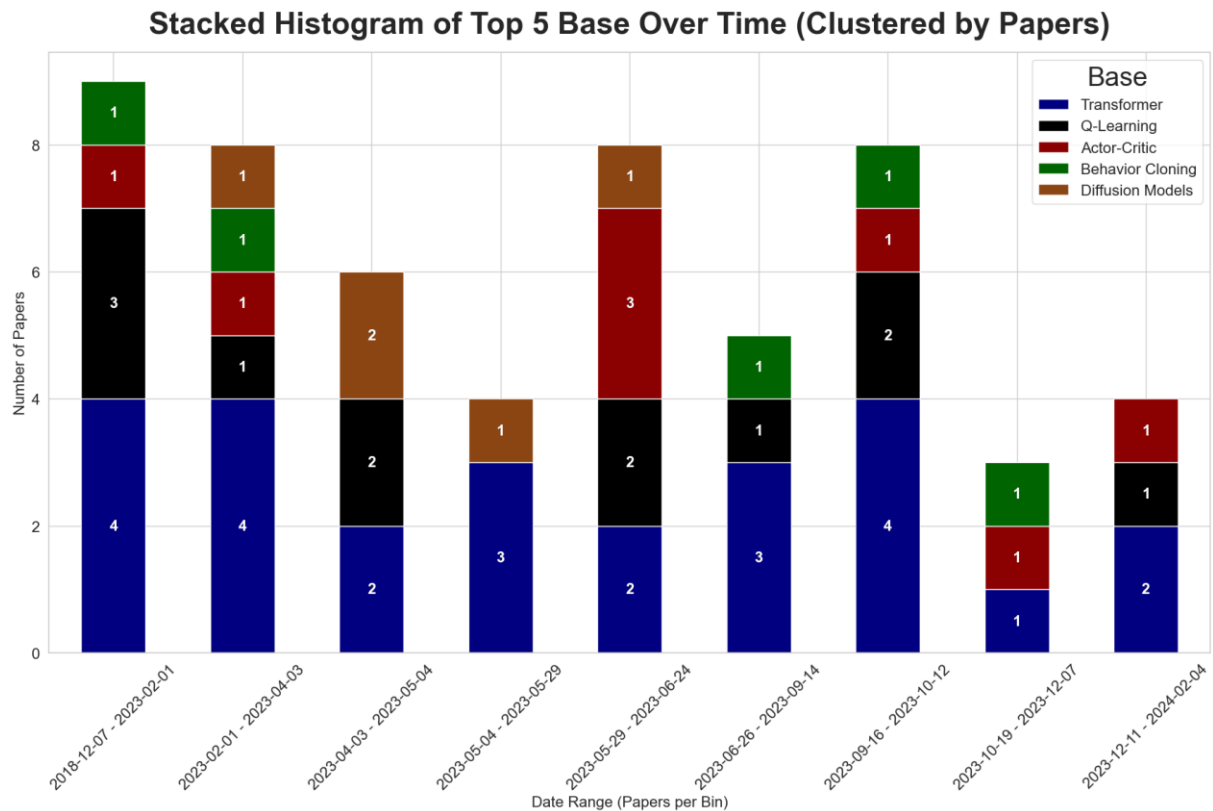


Figure 14. Stacked Histogram of the TOP 5 Base Algorithms over Time

While there is correspondence between the total frequency counts and their representation in the stacked histograms, the time-based patterns remain inconclusive due to the limited sample size. To achieve clearer insights, expanding the analysis to include a larger corpus of studies would be advantageous, enabling the identification of stronger patterns and relationships.

4.3. CONFOUNDING MATRICES

The Confounding Correlation Matrices are essential for understanding the relationships between pairs of concepts in this Thesis of the Offline Reinforcement Learning (ORL) literature. By counting the associations between high-frequency categories, these matrices allow for the comparative analyses and display the interactions between the categories studied. This study focused on prevalent concept pairs like the issue Distribution Shift paired with the Q-Learning or Transformer algorithms to establish baseline connections.

The image below consists of the network graph of the top 5 combinations between the different categories analysed in this sections, these are the Problem Concept, Sub-Area, Base Algorithm and Categories of Mathematical Artifacts, it is relevant to observe the centrality of the Distribution Shift, the Transformers, as well as Q-learning, which is reflected in the papers analyzed in this section.

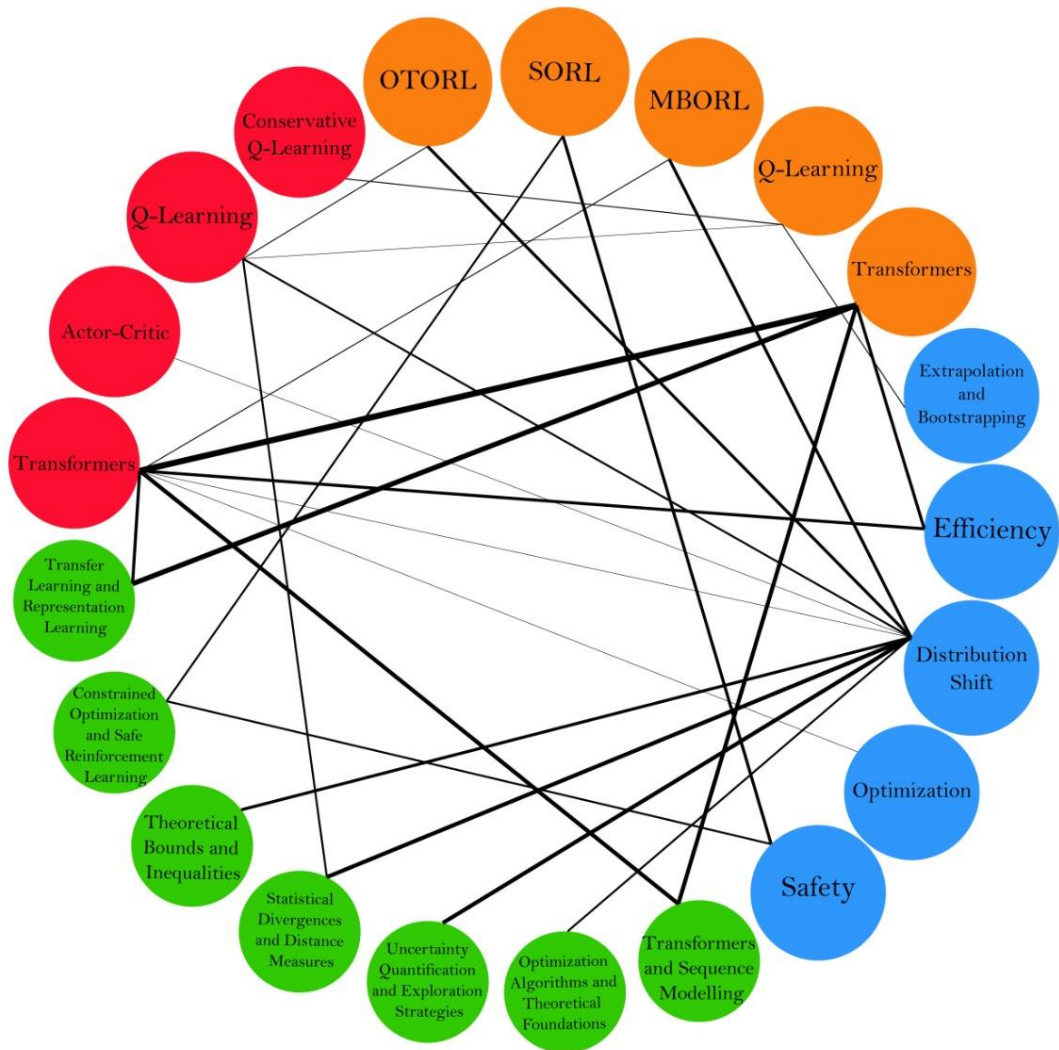


Figure 15. Network Graph for the different categories analyzed in the paper, the Base Algorithms are in Red, the Sub-Areas are in Orange, the Problem Concepts are in Blue and the Categorized Mathematical Artifacts are in Green

4.3.1. Confounding Matrices – Problem Concept vs Standardized Subarea

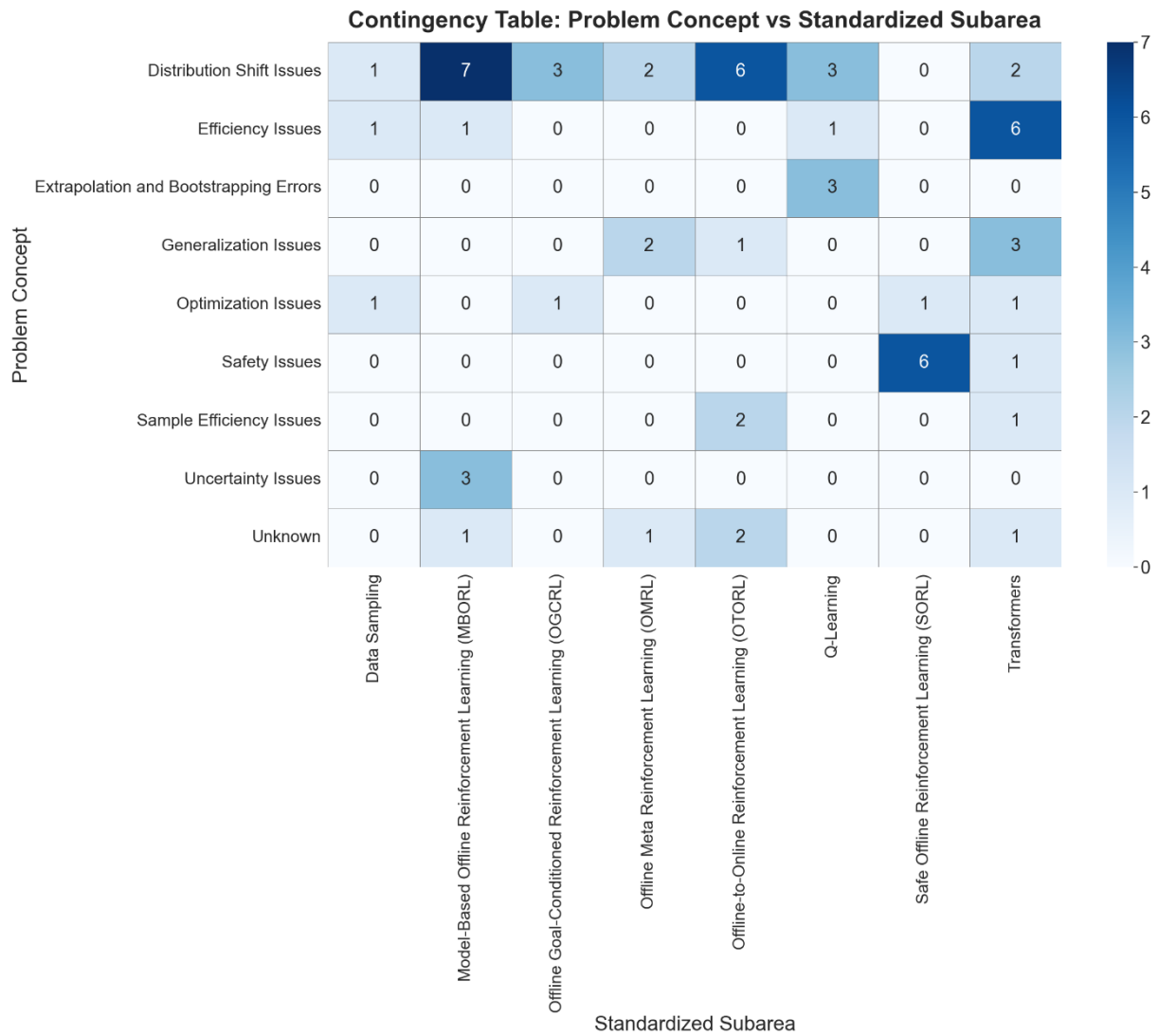


Figure 16. Contingency Matrix of the combinations of between Problem Concepts and Standardized Subarea for the column or row total above or equal to 3

Problem Concept vs Standardized Sub-Area	Frequency
Distribution Shift & Model-Based Offline Reinforcement Learning (MBORL)	7
Efficiency & Transformers	6
Safety & Safe Offline Reinforcement Learning (SORL)	6
Distribution Shift & Offline-to-Online Reinforcement Learning (OTORL)	6
Extrapolation and Bootstrapping Errors & Q-Learning	3

Table 1. Table of the TOP 5 most frequent combinations between the Problems and the Sub-area

To initiate the connections between the categories proposed in this study, the Problem Concept vs. Subarea framework is used as a starting point. As seen in the table of the figure 17, the most frequent combination between problems and subareas is the Distribution Shift with MBORL subarea, and to address this distribution shift in Model-Based Offline Reinforcement Learning (MBORL), several strategies have been introduced. Optimistic Model Rollouts for Pessimistic Offline Policy Optimization (Zhai et al., 2024) and Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning (MOBILE) (Sun et al., 2023) tackle this issue by enhancing policy learning beyond the support of offline datasets. Optimistic Model Rollouts for Pessimistic Offline Policy Optimization (Zhai et al., 2024) constructs an Optimistic Markov Decision Process (O-MDP) to generate out-of-distribution (OOD) rollouts, promoting exploration. In contrast, MOBILE (Sun et al., 2023) penalizes Bellman estimation inconsistencies across model ensembles to more closely approximate the true Bellman error. Focusing on conservatism, CROP: Conservative Reward for Model-based Offline Policy Optimization (Li et al., 2023) and Constrained Policy Optimization with Explicit Behavior Density (CPED) (Zhang et al., 2024) introduce mechanisms to ensure reliable policy evaluation and optimization while employing a flow-GAN model for behavior policy density estimation, ensuring that policy optimization remains within desired boundaries. CROP (Li et al., 2023) incorporates conservatism in reward estimation by minimizing errors and the rewards of random actions to enhance policy reliability. DiffCPS: Diffusion Model-based Constrained Policy Search for Offline Reinforcement Learning (He et al., 2024) overcomes the limitations of Gaussian-based policies by addressing Advantage Weighted Regression’s (AWR) need for exact policy probability densities, which diffusion models cannot provide. It achieves

this by integrating diffusion models into a constrained policy search framework using a primal-dual method, thereby enhancing the expressivity and reliability of the policy search process. Finally, GOPlan from the paper GOPlan: Goal-conditioned Offline Reinforcement Learning by Planning with Learned Models (Wang et al., 2024) pretrains a prior policy with an advantage-weighted GAN and employs reanalysis with planning to handle out-of-distribution (OOD) goals.

While addressing distribution shifts in MBORL is crucial, another significant challenge in reinforcement learning is improving computational efficiency. To tackle this, there have been recent advancements focused on improving efficiency by integrating transformer architectures. Switch Trajectory Transformer with Distributional Value Approximation for Multi-Task Reinforcement Learning (Lin et al., 2022) introduces a multi-task extension to the Trajectory Transformer using a sparsely activated switch transformer architecture. This approach increases model capacity without proportional computational costs and improves policy performance with a distributional trajectory value estimator, especially in sparse reward settings. Similarly, hierarchical sequence modeling frameworks in Hierarchical Decision Transformer (Correia & Alexandre, 2022) and Rethinking Decision Transformer via Hierarchical Reinforcement Learning (Ma et al., 2023) enhance efficiency by structuring the learning process through a high-level mechanism that guides a low-level controller via sub-goal selection or prompts. This replaces traditional returns-to-go and enables the stitching of sub-optimal trajectories for better performance. PDiT: Interleaving Perception and Decision-making Transformers for Deep Reinforcement Learning (Ma et al., 2023) cascades two transformers to separately handle environmental perception and decision-making. PASTA: Pretrained Action-State Transformer Agents (Boige et al., 2023) improves resource usage by utilizing component-level tokenization for actions and states and employing fundamental pre-training objectives like next-token prediction. This results in policy learning with fewer parameters across multiple domains. Addressing inefficiencies due to the forgetting phenomenon in Decision Transformer-based agents, Think Before You Act: Decision Transformers with Internal Working Memory (Kang et al., 2024) proposes a working memory module inspired by the human brain's distributed memory storage to enhance training efficiency and generalization by preventing performance deterioration on previous tasks when learning new ones. These approaches demonstrate that integrating transformers with strategies like sparse activation, hierarchical modeling, and memory modules enhances efficiency in reinforcement learning.

Beyond efficiency improvements, ensuring the safety of reinforcement learning agents is also a significant combination in the confounding matrix results. The area of Safe Offline Reinforcement Learning (SORL) has focused on balancing safety constraints with reward maximization while addressing the challenges posed by limited offline data and the risk of over-conservatism. Researchers have proposed approaches that leverage techniques such as reachability analysis, policy distillation, boundary construction, conditional sequence modeling, and multi-objective optimization. For example, Safe Offline Reinforcement Learning

with Feasibility-Guided Diffusion Model (Zheng et al., 2024) introduces a feasibility-guided framework that translates hard safety constraints into identifying the largest feasible region via reachability analysis, enabling safe policy learning through weighted behavior cloning with a guided diffusion model. Similarly, Guided Online Distillation: Promoting Safe Reinforcement Learning by Offline Demonstration (Li et al., 2023) mitigates over-conservatism by extracting expert policies from offline data and distilling them into policy networks through guided online safe RL training, enhancing both performance and computational efficiency. Other methods, like the approach in Safe Reinforcement Learning with Dead-Ends Avoidance and Recovery (Zhang et al., 2023), construct boundaries to discriminate between safe and unsafe states by learning separate task and recovery policies, ensuring safe exploration with minimal limitations on performance. SaFormer: A Conditional Sequence Modeling Approach to Offline Safe Reinforcement Learning (Zhang et al., 2023) addresses offline safe RL by employing conditional sequence modeling with cost-related tokens and posterior safety verification, achieving adaptable and generalizable constraint satisfaction without retraining. Focusing on worst-case scenarios, Risk Sensitive Dead-end Identification in Safety-Critical Offline Reinforcement Learning (Killian et al., 2023) estimates distributions of expected returns to identify dead-ends earlier, allowing tunable risk tolerance in safety-critical environments. Lastly, Constrained Decision Transformer for Offline Safe Reinforcement Learning (Liu et al., 2023) approaches the trade-off between safety and task performance from a multi-objective optimization perspective, enabling dynamic adjustment of constraints during deployment to learn adaptive, safe, and high-reward policies. These approaches demonstrate efforts to enhance safety in offline RL by integrating modeling techniques and optimization strategies, aiming to develop policies that are safe and high-performing in safety-critical scenarios.

In addition to safety considerations, combining offline and online reinforcement learning introduces new challenges related to distribution shifts that occur during the transition from pre-training to fine-tuning. To address the challenges posed by distribution shift in combining offline and online reinforcement learning, several approaches have been proposed to enhance policy performance and stability during the transition from offline pre-training to online fine-tuning. Using Curiosity for an Even Representation of Tasks in Continual Offline Reinforcement Learning (Pathmanathan et al., 2024) investigates the use of curiosity-driven replay buffers to detect task boundaries and prioritize the retention of informative transitions, thereby mitigating catastrophic forgetting in multi-task continual learning where tasks are non-stationary and unlabeled. Recognizing that directly fine-tuning pre-trained policies can lead to sub-optimal performance due to distribution shifts, Planning to Go Out-of-Distribution in Offline-to-Online Reinforcement Learning (McInroe et al., 2024) introduces Q-Conditioned State Entropy (QCSE) as an intrinsic reward to encourage exploration of low-frequency states while penalizing high-frequency ones, implicitly achieving state marginal matching and improving online fine-tuning performance. Similarly, Towards Robust Offline-to-Online Reinforcement Learning via Uncertainty and Smoothness (Wen et al., 2023) reframes offline-to-online reinforcement learning as an exploration problem and proposes the Planning To Go Out-Of-Distribution (PTGOOD) algorithm, which employs a non-myopic planning procedure to

explore high-reward regions unlikely visited by the behavior policy, thus maximizing the benefit of online data collection without introducing instability. To mitigate performance drops caused by distribution shifts during online adaptation, Boosting Offline Reinforcement Learning with Action Preference Query (Yang et al., 2023) introduces Offline-with-Action-Preferences (OAP), an interaction-free training scheme that queries preferences between pre-collected and learned actions to adaptively adjust policy constraints, improving the evaluation of unseen data and achieving higher performance without online fine-tuning.

Extrapolation errors in Q-learning present challenges in offline reinforcement learning, where agents must learn from fixed datasets without additional environment interactions. Off-policy Deep Reinforcement Learning without Exploration (Fujimoto et al., 2019) demonstrates that standard off-policy deep RL algorithms like DQN and DDPG are ineffective in this setting because they suffer from extrapolation errors when the data distribution does not align with the agent's current policy. To address this, the paper introduces batch-constrained reinforcement learning, which restricts the action space to ensure the agent behaves on-policy with respect to a subset of the data, enabling effective learning from arbitrary fixed datasets. Building upon the notion that bootstrapping errors accumulate due to actions outside the training data distribution, Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction (Kumar et al., 2019) identifies this as a key source of instability in Q-learning and proposes the Bootstrapping Error Accumulation Reduction (BEAR) algorithm. BEAR mitigates bootstrapping errors by constraining action selection during the Bellman backup, allowing robust learning from diverse off-policy distributions. Advancing this line of work, Reasoning with Latent Diffusion in Offline Reinforcement Learning (Venkatraman et al., 2023) tackles the challenge of stitching suboptimal trajectories without incurring extrapolation errors in offline RL by leveraging latent diffusion models to represent in-support trajectory sequences as compressed latent skills. This approach facilitates Q-function learning while avoiding extrapolation errors through batch constraining and handles multi-modal data. The learned latent space enhances credit assignment and accelerates reward propagation in Q-learning, leading to state-of-the-art performance on benchmark tasks. These approaches underscore the importance of constraining action selection and utilizing expressive models to mitigate extrapolation errors in Q-learning within offline RL, enabling stable and effective policy learning from fixed datasets.

4.3.2. Confounding Matrices – Problem Concept vs Base

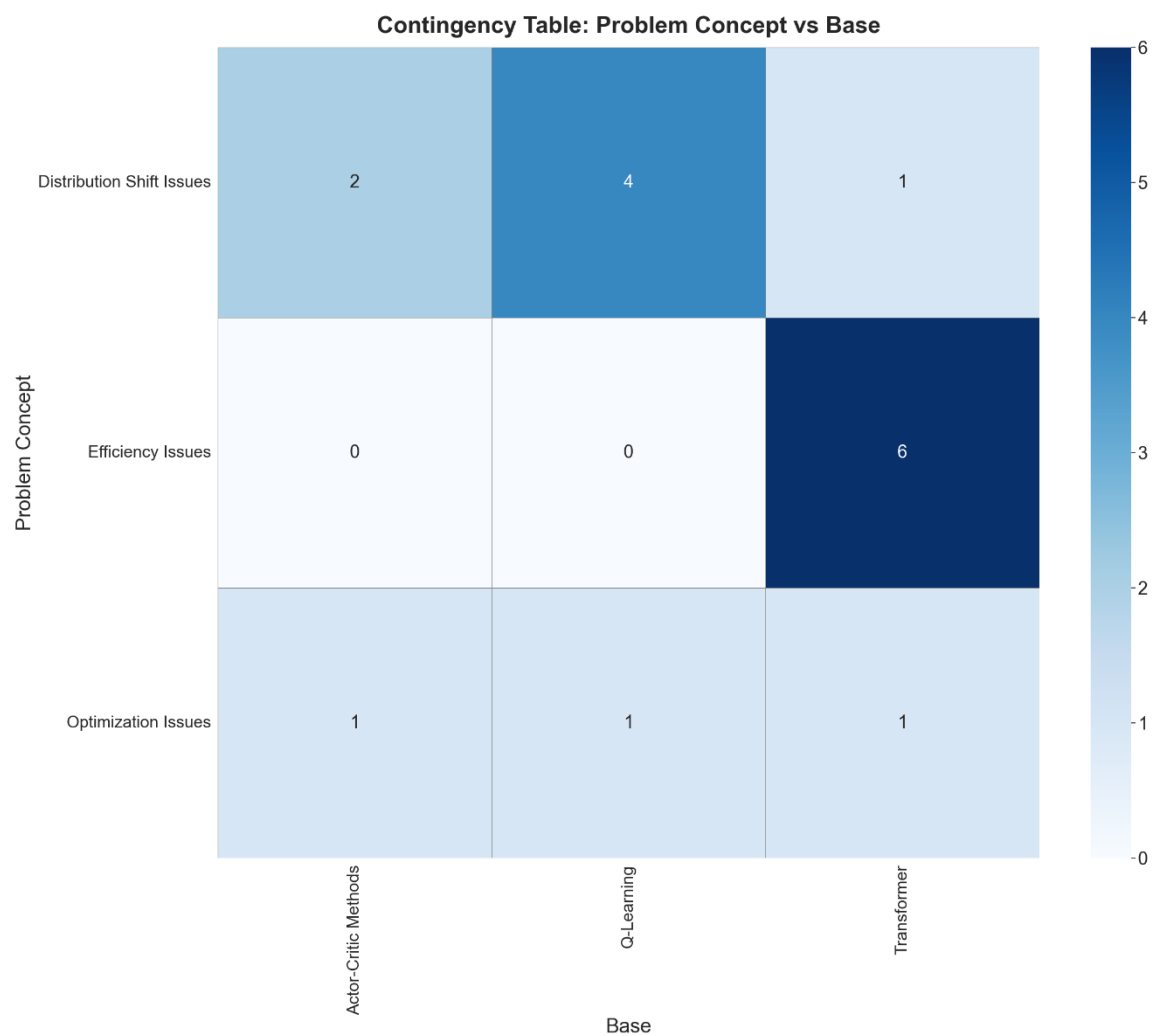


Figure 17. Contingency Matrix of the combinations of between Problems and the Base Algorithm for the column or row total above or equal to 3

Problem Concept vs Base	Frequency
Efficiency & Transformers	6
Distribution Shift & Q-Learning	4
Distribution Shift & Actor-Critic Methods	2
Distribution Shift & Transformers	1
Optimization & Transformers	1

Table 2. Table of the TOP 5 combinations between the Problems and the Base Algorithm

The sections on Efficiency – Transformers (6) and Distribution Shift – Q-Learning (4) exhibit a substantial overlap with the discussions presented in the preceding chapter. Given the high degree of intersection between these problem concepts and the foundational material already covered, providing detailed explanations in this context would result in unnecessary redundancy.

Adding to the last chapter’s discussion on Transformers, this section examines transformer-based methods for addressing distribution shifts and optimizing reinforcement learning. Self-Confirming Transformer for Locally Consistent Online Adaptation in Multi-Agent Reinforcement Learning (Guevara et al., 2023) introduces SCT, which uses a self-confirming loss to locally predict and adapt to nonstationary opponents, outperforming standard transformers and offline MARL baselines. Multi-Objective Decision Transformers for Offline Reinforcement Learning (Ghanem et al., 2023) presents offline RL as a multi-objective problem, enabling transformers to simultaneously predict actions, states, and returns. By integrating action space regions, it smooths action distributions and enhances the attention mechanism, achieving performance comparable to or better than state-of-the-art methods on D4RL locomotion benchmarks.

Beyond addressing out-of-distribution (OOD) through Transformers, several studies explore strategies to handle distribution shifts through actor-critic methodologies. The papers analysed: "Beyond OOD State Actions: Supported Cross-Domain Offline Reinforcement Learning," and "Offline Policy Optimization in RL with Variance Regularization"— explore these actor-critic strategies to address distribution shifts. "Beyond OOD State Actions: Supported Cross-Domain Offline Reinforcement Learning" (Liu et al., 2023) introduces the BOSA method, which simultaneously tackles two primary sources of distribution shift: out-of-distribution (OOD) state-action pairs and varying transition dynamics across multiple environments. By implementing supported policy optimization, BOSA constrains the policy's actions to remain within the support of the offline dataset, effectively preventing the selection of actions that lie outside the training distribution. Concurrently, supported value

optimization filters out transitions that do not align with the target environment's dynamics, thereby reducing extrapolation errors during value estimation. This dual constraint ensures that both the policy and value functions are grounded within the data-supported regions, enhancing the reliability of policy evaluations within the actor-critic framework. Meanwhile, "Offline Policy Optimization in RL with Variance Regularization" (Islam et al., 2022) approaches distribution shift by addressing the variance introduced by distributional discrepancies in importance sampling corrections during off-policy evaluation. This paper proposes incorporating a variance regularizer into the policy optimization objective to minimize the variance of returns under stationary distribution corrections. Leveraging Fenchel duality, the authors transform the variance minimization into a tractable optimization problem, enabling the actor-critic framework to not only maximize expected returns but also ensure that the policy operates within a stationary distribution closely matching the offline dataset. This variance regularization leads to more stable and reliable policy learning by penalizing actions that could result in high-variance returns due to distribution shifts, thereby enhancing the robustness of value function estimations.

Together, these papers offer complementary methodologies that enhance the robustness and reliability of offline RL by addressing different facets of distributional discrepancies within the actor-critic framework.

4.3.3. Confounding Matrices – Standardized Subarea vs Base



Figure 18. Contingency Matrix of the combinations of between Subarea and the Base Algorithm for the column or row total above or equal to 3

Sub-Area vs Base	Frequency
Transformers (Sub-Area) & Transformer (Algorithm)	17
Model-Based Offline Reinforcement Learning (MBORL) & Transformer	3
Offline-to-Online Reinforcement Learning (OTORL) & Q-Learning	3
Q-Learning & Conservative Q-Learning (CQL)	3
Q-Learning (Sub-Area) & Q-Learning (Algorithm)	2

Table 3. Table of the TOP 5 most frequent combinations between the the Sub-area and the Base Algorithm

Due to significant overlap between certain sub-areas and their base algorithms, where the algorithm is essentially part of the sub-area, the majority of combinations present in this subchapter, such as "Transformers - Transformer" and "Q-Learning - Conservative Q-Learning (CQL)" have high redundancy. As they offer little insight because they restate the same concept in both the sub-area and the base algorithm categories. Consequently, only the combinations "Model-Based Offline Reinforcement Learning (MBORL) - Transformer" and "Offline-to-Online Reinforcement Learning (OTORL) - Q-Learning," each with three instances, provide meaningful distinctions, and therefore, is mentioned in this subchapter.

Model-Based Offline Reinforcement Learning (MBORL) has significantly benefited from transformer architectures, as demonstrated by Environment Transformer and Policy Optimization for Model-Based Offline Reinforcement Learning (Wang et al., 2023) and Graph Decision Transformer (Hu et al., 2023). Environment Transformer and Policy Optimization for Model-Based Offline Reinforcement Learning (Wang et al., 2023) leverages transformers to model environment dynamics and reward functions as sequence-to-sequence tasks, enabling the prediction of future state-reward distributions from historical state-action data. By integrating both aleatoric and epistemic uncertainties, it enhances modeling accuracy without requiring online interactions and reduces error accumulation in long-term predictions, thereby improving policy learning outcomes.

On the other hand, Graph Decision Transformer (Hu et al., 2023) integrates graph representations with transformers to refine environment dynamics modeling further. By

representing trajectories as graphs, it captures temporal and causal relationships between states and actions. The Graph Transformer employs relation-enhanced attention mechanisms and optionally incorporates a Sequence Transformer to aggregate fine-grained spatial information, ensuring robust environment modeling and superior policy performance without the need for online interactions.

Together, these sequence-to-sequence and graph-based transformer models highlight the versatility of transformers in capturing complex dependencies and uncertainties within offline datasets. By effectively modeling environment dynamics and incorporating comprehensive uncertainty estimations, these transformer-enhanced MBORL frameworks achieve substantial improvements in policy learning and stability, paving the way for more reliable and efficient reinforcement learning applications.

When it comes to how Q-learning based algorithms address the problem of Offline to Online Reinforcement Learning (OTORL), all the papers have been mentioned in previous subchapters, therefore a quick recapitulation is presented. When it comes to the Constrained Policy Optimization with Explicit Behavior Density (CPED) paper (Zhang et al., 2024), it CPED integrates Q-learning within a constrained optimization framework, using a Flow-GAN to estimate behavior policy density and identify safe regions. This mitigates distribution shifts and enables effective offline policy learning without online interactions. While Improving Offline-to-Online Reinforcement Learning with Q Conditioned State Entropy Exploration (QCSE) (Zhang et al., 2024) uses QCSE to incorporate Q-learning by using Q-conditioned state entropy as an intrinsic reward to encourage exploration of valuable states during online fine-tuning. This aligns the state distribution with the optimal policy, enhancing performance. Finally, Boosting Offline Reinforcement Learning with Action Preference Query (OAP) (Yang et al., 2023) uses OAP to adjust policy constraints based on action preferences evaluated via Q-values within the Q-learning framework. This allows adaptive policy improvement without online interactions, addressing distributional shift in offline RL.

4.3.4. Confounding Matrices – Problem Concept vs Categorized Mathematical Artifacts

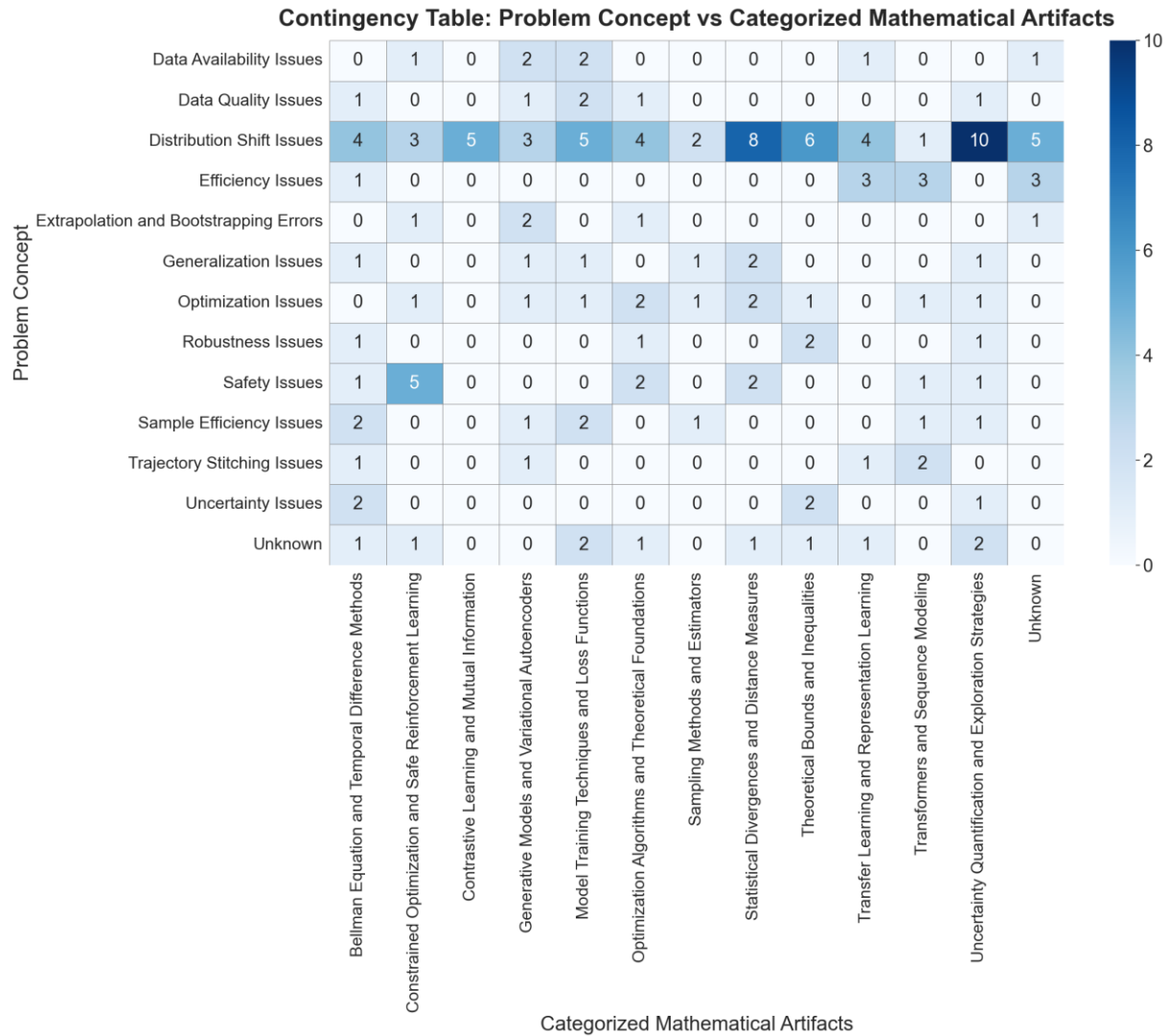


Figure 19. Contingency Matrix of the combinations of between Problem Concept and Mathematical Artifacts for the column or row total above or equal to 5

Problem Concept vs Mathematical Artifacts	Frequency
Distribution Shift – Uncertainty Quantification and Exploration Strategies	10
Distribution Shift – Statistical Divergences and Distance Measures	8
Distribution Shift – Theoretical Bounds and Inequalities	6
Safety Issues - Constrained Optimization and Safe Reinforcement Learning	5
Distribution Shift – Optimization Algorithms and Theoretical Foundations	5

Table 4. Table of the TOP 5 most frequent combinations between the Problem Concept and Mathematical Artifacts used

This section serves to explore the key challenges in offline reinforcement learning, specifically focusing on distribution shift and safety issues, and how recent research addresses these problems using mathematical frameworks such as uncertainty quantification, exploration strategies, statistical divergences, distance measures, constrained optimization, and safe reinforcement learning techniques. The following discussions are organized around these combinations to highlight the strategies employed in tackling the issues.

To address distribution shift issues with uncertainty quantification and exploration strategies, several studies have proposed methods that enhance policy evaluation and guide exploration in uncertain regions. The distribution shift leads to uncertainty in policy evaluation and can result in suboptimal or unsafe actions when deploying the policy. Various research efforts have addressed this issue through different mathematical frameworks, the analyzed papers can be broadly grouped into three categories: robust optimization via uncertainty sets, policy constraints to prevent out-of-distribution (OOD) actions, and leveraging model uncertainty for guided exploration.

Multiple papers tackle distribution shift by constructing uncertainty sets around the estimated models or transition probabilities and then optimizing policies that perform well under worst-case scenarios within these sets. “MICRO: Model-Based Offline Reinforcement Learning with a Conservative Bellman Operator” (Liu et al., 2024) uses the Wasserstein distance to define an uncertainty set $\mathcal{M}_\varepsilon$ of possible Markov Decision Processes (MDPs). The conservative Bellman operator incorporates this uncertainty:

$$(TQ)(s, a) = \min_{M \in \mathcal{M}_\varepsilon} \left[rM(s, a) + \gamma \sum_{s'} s' PM(s' | s, a) Q(s', a') \right].$$

A conservative Bellman operator minimizes Q-values within the uncertainty set, enforcing policy conservatism. This approach ensures the policy remains cautious in uncertain regions

by avoiding overly optimistic estimates (Liu et al., 2024). The paper “Bridging Distributionally Robust Learning and Offline RL: An Approach to Mitigate Distribution Shift and Partial Data Coverage” (Panaganti et al., 2023) introduces the Distributionally Robust Q-Iteration (DRQI) algorithm, defining an uncertainty set P using divergence measures such as total variation, Wasserstein, Kullback-Leibler (KL), and chi-square distances. The robust Bellman operator is:

$$(TQ)(s, a) = r(s, a) + \gamma \inf_{P_{s,a} \in \mathcal{P}_{s,a}} E_{s' \sim P_{s,a}} \left[\max_{a'} Q(s', a') \right].$$

By optimizing over P , the policy becomes robust to discrepancies between estimated and true dynamics. “Achieving the Asymptotically Optimal Sample Complexity of Offline Reinforcement Learning: A DRO-Based Approach” (Wang et al., 2023) refines this approach by using the chi-square divergence to construct less conservative uncertainty sets, improving sample efficiency:

$$\widehat{\mathcal{P}}_s^a = \{q \in \Delta(S) : \chi^2(q \parallel \widehat{P}_s^a) \leq R_s^a\},$$

with R_s^a dependent on sample size and confidence level. (Wang et al., 2023) “Offline Policy Optimization in RL with Variance Regularization” (Islam et al., 2022) addresses distribution shift by regularizing the variance in the policy optimization objective. The method connects the variance of importance sampling ratios to the Rényi divergence between the policy's distribution and the dataset's distribution. The authors introduce the Offline Variance Regularizer (OVR), which leverages Fenchel duality to avoid the double sampling issue, allowing for scalable and stable policy optimization. The policy improvement bound established in the paper is:

$$J(\pi) \geq J(\pi') + E_{(s,a) \sim d^\pi} [A^{\pi'}(s, a)] - \text{Var}_{(s,a) \sim d^\pi} [f(s, a)]$$

where $f(s, a)$ arises from the variational representation of divergence measures. By minimizing this variance, the method stabilizes policy optimization under distribution shift.

To avoid unsafe actions due to distribution shift, the papers titles Learning from Sparse Offline Datasets via Conservative Density Estimation (Cen et al., 2024) and Policy Regularization with Dataset Constraint for Offline Reinforcement Learning (Ran et al., 2023) impose constraints directly on the policy. The former (Cen et al., 2024) introduces Conservative Density Estimation (CDE), constraining the policy's stationary distribution in OOD regions:

$$d_\pi(s, a) \leq \epsilon \mu(s, a),$$

where $d_\pi(s, a)$ is the policy's stationary distribution and $\mu(s, a)$ is a predefined distribution over state-action pairs. This constraint, enforced via f-divergence regularization, prevents the policy from assigning high probability to poorly represented actions. While the latter presents Policy Regularization with Dataset Constraint (Ran et al., 2023) (PRDC), which regularizes the policy to stay close to the entire offline dataset using a point-to-set distance:

$$d_D^\beta(s, a) = \min_{(\hat{s}, \hat{a}) \in D} |(\beta s) \oplus a - (\beta \hat{s}) \oplus \hat{a}|,$$

with β balancing state and action differences. Under Lipschitz continuity assumptions, they bound the difference between the Q-values of the learned policy and the behavior policy, quantifying uncertainty and allowing the adjustment of β to balance conservatism and generalization.

Another batch of papers in this section mentioned ahead, handle distribution shift by utilizing exploration strategies, encouraging the agent to focus on uncertain or underrepresented areas of the state-action space. “Using Curiosity for an Even Representation of Tasks in Continual Offline Reinforcement Learning” (Pathmanathan et al., 2024) uses curiosity-driven exploration, quantifying curiosity as a measure of surprise based on prediction errors. A signal-to-noise ratio (SNR) filter detects significant changes in the curiosity signal:

$$\text{SNR}(t) = \frac{|\delta(t)|}{\sigma(t)} < m\mu(t),$$

where $\delta(t)$ is the prediction error, allowing the agent to adapt to new tasks by focusing on transitions with high curiosity.

“Optimistic Model Rollouts for Pessimistic Offline Policy Optimization” (Zhai et al., 2024), which has been mentioned before in the Problem Concept vs Standardized Subarea section adds a reward bonus proportional to an uncertainty quantifier $u(s, a)$ derived from model ensemble predictions:

$$\max \pi E s, a \sim \pi[r(s, a) + \gamma E s' \mid s, a[V(s')] + \alpha u(s, a)],$$

where α balances optimism and pessimism. Incorporating $u(s, a)$ encourages exploration into less certain regions. Additionally, the framework decouples the training of an optimistic rollout policy in an Optimistic MDP (O-MDP) from the pessimistic optimization of the output policy in a Pessimistic MDP (P-MDP). This two-phase approach allows ORPO to leverage the dynamics model’s generalization capabilities while maintaining conservatism to prevent unsafe actions due to distribution shift. Planning to Go Out-of-Distribution in Offline-to-Online Reinforcement Learning (McInroe et al., 2024) takes a proactive approach by deliberately planning to explore OOD regions during online fine-tuning. It computes a rate $R(x)$ for state-action pairs, quantifying their novelty:

$$R(x) = \log e(z_X \mid x) - \log m(z_X),$$

where e is an encoder and $m(z_X)$ is a marginal distribution in the latent space. Maximizing $R(x)$ guides exploration toward less familiar regions with potential high rewards. Offline Meta Reinforcement Learning with In-Distribution Online Adaptation (Wang et al., 2023) paper addresses distribution shift in offline meta-RL by incorporating behavior policies into the belief updates of a transformed Bayesian MDP, ensuring the agent remains in in-distribution hyper-states. An uncertainty function $Q(\tau)$ measures the confidence that online episodes τ are in-

distribution, and the agent updates its belief only with episodes where $Q(\tau) \leq \delta$, leveraging model uncertainty to guide safe exploration.

Together, these studies address distribution shift in reinforcement learning by implementing diverse exploration strategies that direct the agent toward uncertain or underrepresented regions of the state-action space. “Using Curiosity for an Even Representation of Tasks in Continual Offline Reinforcement Learning” (Pathmanathan et al., 2024) leverages curiosity-driven exploration, using prediction errors to quantify surprise and adaptively focus on novel transitions. “Optimistic Model Rollouts for Pessimistic Offline Policy Optimization” introduces an optimism-driven reward bonus based on uncertainty quantifiers from model ensembles, encouraging exploration in less certain areas while maintaining policy conservatism through a dual-phase approach. “Planning to Go Out-of-Distribution in Offline-to-Online Reinforcement Learning” (McInroe et al., 2024) employs a proactive planning method that calculates a novelty rate for state-action pairs, guiding the agent to explore unfamiliar regions with potential high rewards. Lastly, “Offline Meta Reinforcement Learning with In-Distribution Online Adaptation” (Wang et al., 2023) integrates behavior policies into belief updates within a Bayesian MDP framework, using an uncertainty function to ensure exploration remains within safe, in-distribution hyper-states.

Another approach to handling distribution shift involves using statistical divergences and distance measures to quantify and mitigate discrepancies between the learned policy's distribution and that of the dataset.

In “Improving Offline-to-Online Reinforcement Learning with Q Conditioned State Entropy Exploration” (Zhang et al., 2024) addresses distribution shift in offline-to-online (OtO) reinforcement learning by introducing Q Conditioned State Entropy (QCSE) as an intrinsic reward. Instead of directly optimizing statistical divergences, QCSE maximizes the state entropy of each sample conditioned on its Q-value:

$$QCSE(s, a) = \lambda \cdot \tanh \left(H \left(s \mid \min \left(Q_{\phi_1}(s, a), Q_{\phi_2}(s, a) \right) \right) \right) + r(s, a),$$

where $H(s \mid Q)$ represents the state entropy conditioned on the Q-values, Q_{ϕ_1} and Q_{ϕ_2} are double Q-networks, and λ is a scaling factor. By maximizing QCSE, the agent is encouraged to explore underrepresented or high-reward states, effectively achieving State Marginal Matching (SMM). This approach minimizes the KL divergence between the policy's state distribution $p_\pi(s)$ and the target distribution $p^*(s)$, thereby mitigating distribution shift.

In “Federated Offline Policy Learning” (Carranza & Athey, 2023) the authors handle distribution shift across clients by quantifying skewness using the chi-squared divergence and the KL divergence between the client sampling distribution λ and the empirical distribution $\bar{n}$:

$$s(\lambda \parallel \bar{n}) = E_{c \sim \lambda} \left[\frac{\lambda_c}{\bar{n}_c} \right] = 1 + \chi^2(\lambda \mid \bar{n}).$$

bound the total variation (TV) distance between the global distribution and each local client distribution using KL divergences among the marginal context distributions p_{X^c} , propensity functions e^c , and conditional reward distributions $\{\vec{Y}^c|X^c\}$:

$$\text{TV}(p_{\bar{D}_c}, p_{\bar{D}_\lambda}) \leq E_{k \sim \lambda} \left[\sqrt{\text{KL}(p_{X^c} \parallel p_{X^k})} + \sqrt{\text{KL}(e^c \parallel e^k)} + \sqrt{\text{KL}(p_{\vec{Y}^c|X^c} \parallel p_{\vec{Y}^k|X^k})} \right]$$

By incorporating these divergence measures into their regret bounds, the authors effectively account for client heterogeneity and mitigate distribution shifts in federated settings. This ensures that the personalized decision policies learned are robust across diverse and heterogeneous data sources without requiring the central aggregation of raw data.

"Supported Trust Region Optimization for Offline Reinforcement Learning" (Mao et al., 2023) uses KL divergence and Total Variation (TV) distance to constrain policy updates:

$$\pi_{i+1} = \arg \max_{\pi} E_{a \sim \pi} [\widehat{A}^{\pi_{\text{pe}}}(s, a)] \quad \text{s.t.} \quad D_{\text{KL}}(\pi \parallel \pi_{\text{base}})[s] \leq \epsilon,$$

ensuring updates remain close to the baseline policy π_{base} . The solution:

$$\pi_{i+1}(a|s) = \frac{1}{Z(s)} \pi_{\text{base}}(a|s) \exp\left(\frac{\widehat{A}^{\pi_{\text{pe}}}(s, a)}{\lambda^*(s)}\right),$$

maintains safety by limiting divergence from the behavior policy, thus addressing distribution shift.

In "Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning" (Sun et al., 2023) the authors penalize value estimation using the standard deviation of Bellman estimates from an ensemble of models:

$$U(s, a) = \text{Std}_i [T_i^{\pi} Q_{\psi}(s, a)],$$

defining a penalized Bellman operator:

$$T^{\text{MOBIP}} Q_{\psi}(s, a) = T^{\pi} Q_{\psi}(s, a) - \beta U(s, a).$$

This uncertainty measure discourages actions in regions with high model disagreement, effectively handling distribution shift without traditional divergence measures.

"Bayesian Reparameterization of Reward-Conditioned Reinforcement Learning with Energy-based Models" (Ding et al., 2023) reparametrizes the policy using energy-based models, ensuring the KL divergence between the policy π and the behavior policy β is bounded:

$$\text{KL}(\pi(T) \parallel \beta(T)) \leq N \log\left(\frac{1}{\eta}\right),$$

where N is the trajectory length. They express the posterior policy via Bayes' theorem:

$$\pi(a | s, R) = \pi(a | s) \cdot \frac{P(R | s, a)}{P(R | s)},$$

modeling dependencies between actions and rewards, mitigating distribution shift by keeping the policy within the dataset's support.

Finally, "Constrained Policy Optimization with Explicit Behavior Density for Offline Reinforcement Learning" (Zhang et al., 2024) estimates the behavior policy's density $L_{\pi_{\beta}, \theta}(s, a)$ using a flow-based model and defines a safe set:

$$\tilde{S} \times \tilde{A} = \left\{ (s, a) \in S \times A \mid -\log L_{\pi_{\beta}, \theta}(s, a) < \epsilon \right\}.$$

The policy optimization is constrained to this set:

$$\max_{\psi^{E_{S \sim P_D, a \sim \pi_{\psi}(\cdot | S), (s, a) \in \tilde{S} \times \tilde{A}}}} [Q_{\eta}(s, a)].$$

By excluding low-density actions, the method avoids out-of-distribution actions, addressing distribution shift.

In comparison, all these papers use methods in statistical divergences or distance measures to handle distribution shift while differing in implementation. While KL divergence is common due to its effectiveness, other measures like chi-squared divergence and total variation distance are also used.

In the analyzed papers below, safety issues in offline reinforcement learning are addressed through constrained optimization and safe reinforcement learning techniques, ensuring that learned policies adhere to safety constraints while maximizing performance.

In the first paper, "Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model" (Zheng et al., 2024) the authors introduce hard state-wise constraints, replacing traditional soft constraints on expected cumulative costs with strict conditions $h(s_t) \leq 0$ for all time steps t . They define an optimal feasible value function:

$$V_h^*(s) = \min_{\pi} \max_t h(s_t),$$

where $s_0 = s$ and actions a_t are drawn from policy π . The optimization is split into feasible regions (where $V_h^*(s) \leq 0$) and infeasible regions (where $V_h^*(s) > 0$). In feasible regions, they maximize expected rewards while ensuring safety; in infeasible regions, they minimize constraint violations. They derive a closed-form solution for the optimal policy that balances reward maximization and safety considerations.

The second paper, "Safe Reinforcement Learning with Dead-Ends Avoidance and Recovery" (Zhang et al., 2023) proposes a decoupled safety mechanism by separating the safety critic $Q_c^{\pi_{\text{rec}}}$ and the recovery policy π_{rec} from the task policy π_{task} . The safety critic is trained using expectile regression to avoid overestimation caused by out-of-distribution actions. During

execution, a behavior correction mechanism checks if $Q_c^{\pi_{\text{rec}}}(s_t, a^{\pi_{\text{task}}}) < \epsilon_{\text{safe}}$; if so, the agent executes the task action $a^{\pi_{\text{task}}}$, otherwise it switches to the recovery action $a^{\pi_{\text{rec}}}$.

In the third paper, "SaFormer: A Conditional Sequence Modeling Approach to Offline Safe Reinforcement Learning" (Zhang et al., 2023) the authors extend the Decision Transformer framework by introducing cost-related tokens: the cost-to-go $C_t = \sum_{t'=t}^T c_{t'}$ and a cost limit DDD. The policy generates actions conditioned on these tokens and past trajectories. The training objective minimizes the negative log-likelihood of these tokens to align future actions with safety constraints. During execution, the actor samples multiple action candidates based on the conditioned policy, and a critic evaluates these candidates to ensure the selected action does not exceed the remaining cost budget C_t .

The fourth paper, "Risk Sensitive Dead-end Identification in Safety-Critical Offline Reinforcement Learning" (Killian et al., 2023) employs distributional reinforcement learning with risk measures to identify and avoid unsafe actions. They use Implicit Quantile Networks (IQN) to model two separate return distributions $Z_D^{\pi}(s, a)$ and $Z_R^{\pi}(s, a)$ and estimate the Conditional Value at Risk (CVaR) to assess action risk:

$$\text{CVaR}_{\alpha}(Z^{\pi}(s, a)) = E[Z^{\pi}(s, a) \mid Z^{\pi}(s, a) \leq \text{VaR}_{\alpha}(Z^{\pi}(s, a))],$$

where VaR_{α} is the value-at-risk at confidence level α . Actions with both CVaR measures below certain thresholds are considered risky and are avoided. They incorporate a conservative Q-learning penalty to prevent overestimation of unobserved actions, enhancing the reliability of risk assessments.

In summary, these methods enhance safety in reinforcement learning by enforcing hard safety constraints during policy optimization, using recovery policies to avoid unsafe states, conditioning policies on cost-related information to guide safe action selection, and evaluating risks using distributional approaches with measures like CVaR. Each approach offers a different mathematical framework to ensure that the agent avoids unsafe actions or states, thereby improving the safety of reinforcement learning systems.

4.3.5. Confounding Matrices – Standardized Subarea vs Categorized Mathematical Artifacts

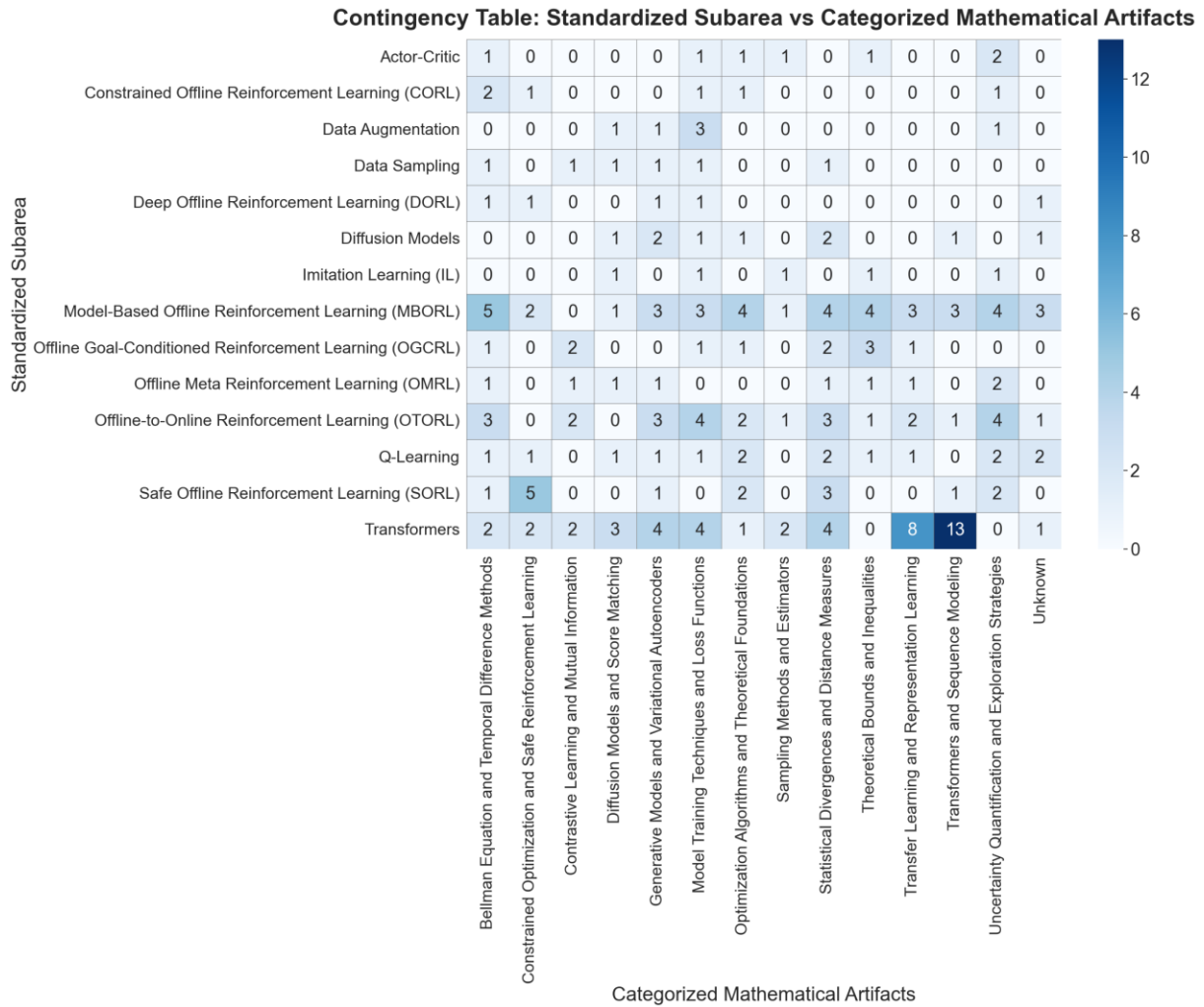


Figure 20. Contingency Matrix of the combinations of between Subarea and Mathematical Artifacts for the column or row total above or equal to 5

Standardized Subarea vs Mathematical Artifacts	Frequency
Transformers - Transformers and Sequence Modeling	13
Transformers - Transfer Learning and Representation Learning	8
Safe Offline Reinforcement Learning (SORL) & Constrained Optimization and Safe Reinforcement Learning	5
Model-Based Offline Reinforcement Learning (MBORL) - Bellman Equation and Temporal Difference Methods	5
Transformers – Generative Models and Variational Auto-encoders	4

Table 5. Table of the TOP 5 most frequent combinations between the Subarea and Mathematical Artifacts used

The landscape of Offline Reinforcement Learning (ORL) has seen advancements through the incorporation of Transformer-based architectures aided by optimization techniques. Among these, the Switch Trajectory Transformer (Lin et al., 2022) (SwitchTT) is an example of integrating a Switch Transformer architecture, which employs multiple feed-forward networks or experts. The routing mechanism in SwitchTT directs tokens to the most relevant experts using softmax-derived gate values, mathematically represented as

$$p_i(x) = \frac{e^{h(x)_i}}{\sum_{j=1}^n e^{h(x)_j}}$$

This allows SwitchTT to effectively manage computational resources while enhancing policy performance. Additionally, SwitchTT introduces a distributional value estimator that models return-to-go (RTG) values using categorical distributions, formulated as $V(x) = \sum_{i=0}^{N-1} p_i(x) \cdot z_i$, where each z_i represents discretized RTG values. This nuanced value approximation guides beam search during planning, distinguishing SwitchTT from traditional Decision Transformers.

In contrast, the Multi-Objective Decision Transformer (Ghanem et al., 2023) (MO-DT) broadens the Decision Transformer framework by incorporating additional objectives for state and return predictions within a multi-objective optimization (MOO) framework. The optimization objective in MO-DT is given by $\theta^* = \arg \min_{\theta} (\lambda_0 \mathcal{L}_{\mathcal{DT}}(\theta) + \lambda_1 \mathcal{L}_s(\theta) + \lambda_2 \mathcal{L}_g(\theta))$, where each λ term balances different loss components. This approach ensures that the model not only predicts actions based on past states and RTGs but also accurately forecasts future states and returns, thereby enhancing the robustness of the learned policy.

Elastic Decision Transformer (EDT) (Mao et al., 2023) introduces a novel mechanism by dynamically adjusting the history length based on trajectory performance. Utilizing expectile regression, EDT estimates the maximum achievable return for varying history lengths with the

objective $\widetilde{R}_t \approx \arg \min_{\widetilde{R}_t} E \left[L_\alpha \left(\widetilde{R}_t - \widehat{R}_t(\tau_T) \right) \right]$, where L_α is an asymmetric least squares loss function. This allows EDT to prioritize higher-return trajectories by effectively performing trajectory stitching—forgetting sub-optimal past experiences and focusing on more promising paths. Such dynamic history management sets EDT apart from both SwitchTT and MO-DT, which do not adapt the history length based on performance.

Architecturally, methods like PDiT (Ma et al., 2023), Decision ConvFormer (Kim et al., 2024), and Transformer in Transformer (TIT) (Mao et al., 2023) introduce significant modifications to the traditional Transformer architecture to better capture spatial and temporal dependencies. PDiT interleaves perception and decision-making Transformer blocks, treating each state and action component as separate tokens. This component-level tokenization, expressed as $\widehat{z}_0 = [E_p(o_p^1); E_p(o_p^2); \dots; E_p(o_p^N)]$, allows for fine-grained processing of environmental observations, enhancing the model's ability to generate accurate actions based on complex state representations.

Decision ConvFormer diverges from conventional self-attention mechanisms by incorporating 1D depthwise convolution modules tailored for different token types—RTG, state, and action. The convolution operations are mathematically defined as

$$C_t[p, q] = \sum_{l=0}^{L-1} \widehat{w}_q^R[l] \cdot X_t[p-l, q]$$

for RTG tokens, and similarly for state and action tokens with respective weights. This convolution-based token mixing captures local temporal dependencies more efficiently than self-attention, reducing computational complexity while maintaining performance.

TIT, on the other hand, employs a dual-transformer architecture with an inner transformer dedicated to spatial processing within individual observations and an outer transformer handling temporal dependencies across multiple observations. The inner transformer processes observation patches, resulting in embeddings like z_L , which are then fed into the outer transformer to model the sequence over time. This hierarchical processing enhances the model's capacity to understand both spatial and temporal aspects of the environment, leading to more informed decision-making.

In the realm of multi-task and skill-based learning, approaches such as Merging Decision Transformers, PASTA (Boige et al., 2023), and Skill Decision Transformer (Skill DT) (Sudhakaran & Risi, 2023) offer distinct methodologies. Merging Decision Transformers (Lawson & Qureshi, 2023) is an algorithm that explores the combination of multiple task-specific models into a unified multi-task model without additional training. This is achieved through direct weight averaging $\theta_M = \frac{1}{m} \sum_{i=1}^m \theta_i$ or Fisher merging $\theta_M = \frac{\sum_{i=1}^m F_i \theta_i}{\sum_{i=1}^m F_i}$, where F_i represents the Fisher information for each model. These techniques allow the creation of a consolidated model that retains the capabilities of individual task-specific models.

PASTA formulates the offline assortment optimization problem as a robust max-min optimization task, aiming to maximize expected revenue while minimizing the worst-case estimation errors within a confidence set Θ_n . Mathematically, PASTA merges multiple task-specific Decision Transformers by either direct weight averaging, expressed as $\theta_M = \frac{1}{m} \sum_{i=1}^m \theta_i$, or Fisher merging, defined by $\theta_M = \frac{\sum_{i=1}^m F_i \theta_i}{\sum_{i=1}^m F_i}$, where F_i represents the Fisher information matrix for each model. Additionally, PASTA introduces the Merge-Freeze-Finetune (MFF) strategy, which involves freezing the merged parameters and independently finetuning the remaining ones to enhance multi-task performance. Theoretical guarantees are provided through finite-sample regret bounds under the Multinomial Logit (MNL) model, demonstrating that PASTA effectively identifies the optimal assortment by requiring data coverage only at the optimal set. Furthermore, PASTA employs gradient descent with line search (GDLS) to efficiently solve the inner minimization problem, ensuring computational feasibility and robustness in diverse assortment optimization scenarios.

Skill Decision Transformer (Skill DT) employs a Vector Quantized Variational Autoencoder (VQ-VAE) to encode states into discrete skill embeddings z_t , which are then quantized to the nearest codebook entry $z_t = \text{quantize}(e(s_t))$. Future skill distributions are represented as histograms $Z_t = \sum_{k=t}^T \text{one_hot}(z_k)$, enabling the model to predict actions based on both current skill embeddings and their future distributions. This focus on skill embeddings and their temporal distributions allows Skill DT to effectively generalize across different tasks by leveraging learned skills.

Guiding Online Reinforcement Learning with Action-Free Offline Pretraining (Zhu et al., 2023) introduces a novel approach by leveraging action-free offline datasets to guide online policy learning. It utilizes an Action-Free Decision Transformer (AFDT) (Zhu et al., 2023) to predict state changes based on past states and desired returns, formulated as $s_{t+1} = s_t + \Delta s_t = s_t + \text{Preds}(f_{\text{output}})$. The discrepancy between predicted and actual state changes generates a guiding reward $r_g^t = \frac{1}{D} |s_{t+1}^{\text{pred}} - s_{t+1}|_2^2$, which is integrated into the policy learning process through a combined Q function $Q(s_t, a_t) = Q_e(s_t, a_t) + Q_g(s_t, a_t)$. This guiding mechanism enhances exploration and policy robustness by providing additional feedback based on the model's predictions, even in the absence of action labels in the offline data.

Overall, these approaches leverage strategies like enhancing value estimation through distributional and multi-objective frameworks to optimizing Transformer architectures for better spatial-temporal representation, and from facilitating multi-task learning through model merging and skill embeddings to integrating language and vision models for more intuitive reasoning, each method contributes uniquely to improving policy reliability and performance.

When it comes to how transformers used statistical divergences and distance measures multiple approaches are worthy of mention. One notable approach, Decoupled Meta-Reinforcement Learning (DCMRL) (He et al., 2023), leverages Gaussian task contexts and skills

to improve meta-training and meta-testing phases. In the meta-training phase, DCMRL employs the Kullback-Leibler (KL) divergence to align skill distributions derived from a skill encoder with those predicted by a skill prior. This alignment is mathematically expressed as minimizing the divergence $\text{DKL}(s_g[q(Z|s, a)], p(Z|s_0))$, ensuring that the skill prior accurately reflects the encoded skills from various tasks. Additionally, DCMRL introduces a contrastive learning mechanism to enhance task context representations. By utilizing a triplet loss $L_{\text{triplet}}(\tilde{C}, \tilde{C}^+, \tilde{C}^-) = \max(0, \text{sim}(\tilde{C}, \tilde{C}^-) - \text{sim}(\tilde{C}, \tilde{C}^+) + \epsilon)$, the model distinguishes between anchor, positive, and negative task contexts, thereby promoting more discriminative and generalizable task embeddings.

From the paper ‘‘Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning’’ (Lu et al., 2023) comes the Contrastive Energy Prediction (CEP) method which introduces an exact formulation for energy-guided sampling in offline reinforcement learning. CEP addresses the challenge of estimating the gradient of an intermediate energy function $\nabla_{x_t} E_t(x_t)$ by employing a contrastive loss that compares energy predictions across multiple samples. The objective function

$$\min_{\phi} E \left[- \sum_{i=1}^K e^{-\beta E(x_i^o)} \log \frac{e^{-f_{\phi}(x_i^t, t)}}{\sum_{j=1}^K e^{-f_{\phi}(x_j^t, t)}} \right]$$

ensures that the energy model accurately captures the desired distribution by normalizing across sampled actions. This contrastive approach not only stabilizes the training process by mitigating the impact of outliers but also provides an unbiased estimation of the energy guidance necessary for effective sampling.

In the realm of safe reinforcement learning, the Constrained Decision Transformer (Liu et al., 2023) (CDT) integrates statistical measures to enforce safety constraints while optimizing performance. CDT extends the traditional Decision Transformer by incorporating cost return tokens alongside reward returns, enabling the model to condition its actions on both reward and cost information. The training objective $\min_{\theta} E_{\mathcal{O} \sim \mathbb{T}} [-\log \pi_{\theta}(a|\mathcal{O}) - \lambda H[\pi_{\theta}(\cdot|\mathcal{O})]]$ combines negative log-likelihood with an entropy regularizer, where $H[\pi_{\theta}]$ represents the Shannon entropy. This formulation encourages the policy to maintain stochasticity, thereby preventing deterministic behaviors that could lead to constraint violations. Furthermore, CDT employs a data augmentation strategy based on the Pareto frontier to relabel trajectories, ensuring that the augmented data respects the trade-offs between rewards and costs. This is achieved through an algorithm that identifies Pareto-optimal trajectories and adjusts their reward and cost returns to align with desired target pairs, effectively using statistical measures to enhance the dataset’s quality and the policy’s safety adherence.

Comparing these methods reveals a common reliance on statistical divergences to guide policy learning, yet each employs these measures in distinct ways to address specific challenges in ORL. DCMRL utilizes the KL divergence to align skill distributions, facilitating the decoupling of task contexts and skills, which enhances generalization across diverse tasks. In contrast, CEP leverages a contrastive loss to accurately estimate energy guidance, crucial for sampling

actions that conform to desired distributions without relying on action labels. CDT, on the other hand, integrates entropy regularization and contrastive data augmentation to balance reward maximization with safety constraints, demonstrating how statistical measures can enforce safety without compromising performance.

Building upon the foundational modifications to Bellman’s equations and temporal difference (TD) learning presented in previous methodologies, the subsequent papers introduce nuanced adjustments tailored to enhance robustness, uncertainty estimation, and constraint adherence within Offline Reinforcement Learning (RL) frameworks.

The MICRO (Liu et al., 2024) framework innovates by integrating a conservative Bellman operator within a model-based paradigm to mitigate extrapolation errors inherent in offline datasets. The core mathematical alteration lies in redefining the Bellman operator T to account for model uncertainty:

$$TQ(s, a) = \begin{cases} r(s, a) + \gamma \mathbb{E}_{s' \sim T_M(s'|s, a)} [\max_{a'} Q(s', a')], & \text{if data is accurate} \\ r(s, a) + \gamma \inf_{M \in \mathcal{M}_\epsilon} \mathbb{E}_{s' \sim T_M(s'|s, a)} [\max_{a'} Q(s', a')], & \text{if data is from the model} \end{cases}$$

This dual-condition operator selectively applies the standard Bellman update for reliable offline data while invoking a robust, pessimistic update for model-generated data. The introduction of the infimum over an uncertainty set $\mathcal{M}_\epsilon$ ensures that the policy remains conservative, thereby safeguarding against inaccuracies in the learned model. Additionally, MICRO leverages an ensemble of dynamics models to approximate $\mathcal{M}_\epsilon$, facilitating efficient uncertainty quantification without incurring prohibitive computational costs.

The approach in Model-Based Epistemic Variance of Values for Risk-Aware Policy Optimization (Luis et al., 2023) extends the conventional Bellman framework by incorporating epistemic uncertainty into the value function estimation. Instead of modifying the temporal difference update, the variance is estimated separately using an uncertainty Bellman equation. The standard value function is estimated as:

$$V^\pi(s) = E_{a \sim \pi(\cdot|s)} \left[r(s, a) + \gamma E_{s' \sim p(\cdot|s, a)} [V^\pi(s')] \right]$$

To incorporate epistemic variance, the variance in value estimates induced by model uncertainty is estimated separately, and then used in the policy optimization objective:

$$\pi = \arg \max_{\pi} (\overline{Q}^\pi(s, a) + \lambda U^\pi(s, a))$$

Here, λ serves as a risk-aversion coefficient, modulating the influence of the variance term. This modification ensures that policies are not only optimized for expected returns but also account for the uncertainty in value estimates, thereby promoting risk-aware decision-making.

Uncertainty-driven Trajectory Truncation for Data Augmentation in Offline Reinforcement Learning (Zhang et al., 2023) presents a methodology that improves data augmentation by incorporating a trajectory truncation mechanism based on accumulated uncertainty. The method selectively truncates imagined trajectories when the accumulated uncertainty $U_t(s, a)$ surpasses a predefined threshold ϵ , ensuring that only reliable synthetic data is used for policy learning. The accumulated uncertainty is calculated along the trajectory, and the trajectory is truncated if $U_t(s, a) > \epsilon$.

Here, $U_t(s, a)$ quantifies the accumulated uncertainty along a trajectory, and λ and κ are coefficients used in the modified reward function within the ϵ -Pessimistic MDP to penalize high-uncertainty actions:

$$r_p(s_t, a_t) = \begin{cases} r(s_t, a_t) - \lambda u(s_t, a_t), & \text{if } U_t(s_t, a_t) \leq \epsilon \\ r(s_t, a_t) - \lambda u(s_t, a_t) - \kappa, & \text{if } U_t(s_t, a_t) > \epsilon \end{cases}$$

This selective truncation ensures that the policy remains within regions of the state-action space where the model is confident, thereby enhancing the reliability of the learned policy. By integrating trajectory truncation into the data augmentation process, the approach focuses on generating and utilizing trustworthy synthetic samples, fostering more robust policy learning.

Selective Uncertainty Propagation (SUP) is a method presented in the "Selective Uncertainty Propagation in Offline RL" (Krishnamurthy et al., 2024) paper that refines the value function updates by selectively propagating uncertainty based on the estimated shift in next-state distributions. The modified value function estimation incorporates adjustments to account for uncertainty:

In SUP, the algorithm adjusts the Q-values by incorporating shift estimates and uncertainty in future value estimates into the value iteration process. Specifically, the adjusted Q-value is computed as:

$$\widehat{Q}^{(h)}(x, a) \leftarrow \widehat{R}^{(h)}(x, a) + \int_{x'} \widehat{V}^{(h+1)}(x') \widehat{P}^{(h)}(x' | x, a) dx'$$

Here, $\widehat{Q}^{(h)}(x, a)$ is the estimated Q-value at step h , $\widehat{R}^{(h)}(x, a)$ represents the estimation uncertainty bonus for the state-action pair (x, a) at step h , $\widehat{\Delta}^{(h)}(x' | x, a)$ is the estimated shift in the next-state distribution due to action a at state x , and $\widehat{V}_o^{(h+1)}(x')$ and $\widehat{V}_p^{(h+1)}(x')$ are the optimistic and pessimistic estimates of the value function at the next step $h+1$.

In SUP, the algorithm adjusts the Q-values by incorporating shift estimates and uncertainty in future value estimates into the value iteration process. Specifically, the adjusted Q-value is computed as:

$$\begin{aligned} \widehat{Q}^{(h)}(x, a) \leftarrow & \widehat{R}^{(h)}(x, a) + \int_{x'} \widehat{V}^{(h+1)}(x') \widehat{P}^{(h)}(x' | x, a) dx' - b^{(h)}(x, a) \\ & - \int_{x'} \left| \widehat{\Delta}^{(h)}(x' | x, a) \right| \left(\widehat{V}_{o^{(h+1)}}(x') - \widehat{V}_{p^{(h+1)}}(x') \right) dx' \end{aligned}$$

Here, $\widehat{Q}^{(h)}(x, a)$ is the estimated Q-value at step h , $b^{(h)}(x, a)$ represents the estimation uncertainty bonus for the state-action pair (x, a) at step h , $\widehat{\Delta}^{(h)}(x' | x, a)$ is the estimated shift in the next-state distribution due to action a at state x , and $\widehat{V}_{o^{(h+1)}}(x')$ and $\widehat{V}_{p^{(h+1)}}(x')$ are the optimistic and pessimistic estimates of the value function at the next step $h+1$.

This selective propagation allows the algorithm to adaptively focus on areas with higher uncertainty, thereby balancing exploration and exploitation more effectively. Additionally, SUP employs a dual-value function approach, maintaining both optimistic and pessimistic estimates to guide the policy towards regions of higher confidence while avoiding overestimation in uncertain areas.

In summary, these MBORL techniques enhance policy learning by modifying the Bellman equations and temporal difference updates to incorporate uncertainty estimation and robustness. By integrating conservative operators, uncertainty-driven adjustments, and selective uncertainty propagation into the Bellman framework, they focus on reliable regions of the state-action space. These modifications to temporal difference learning allow the methods to mitigate extrapolation errors, promote risk-aware decision-making, and adhere to constraints within offline datasets. Ultimately, these changes enable the development of policies that are both effective and robust in offline reinforcement learning settings.

4.3.6. Confounding Matrices – Base vs Categorized Mathematical Artifacts

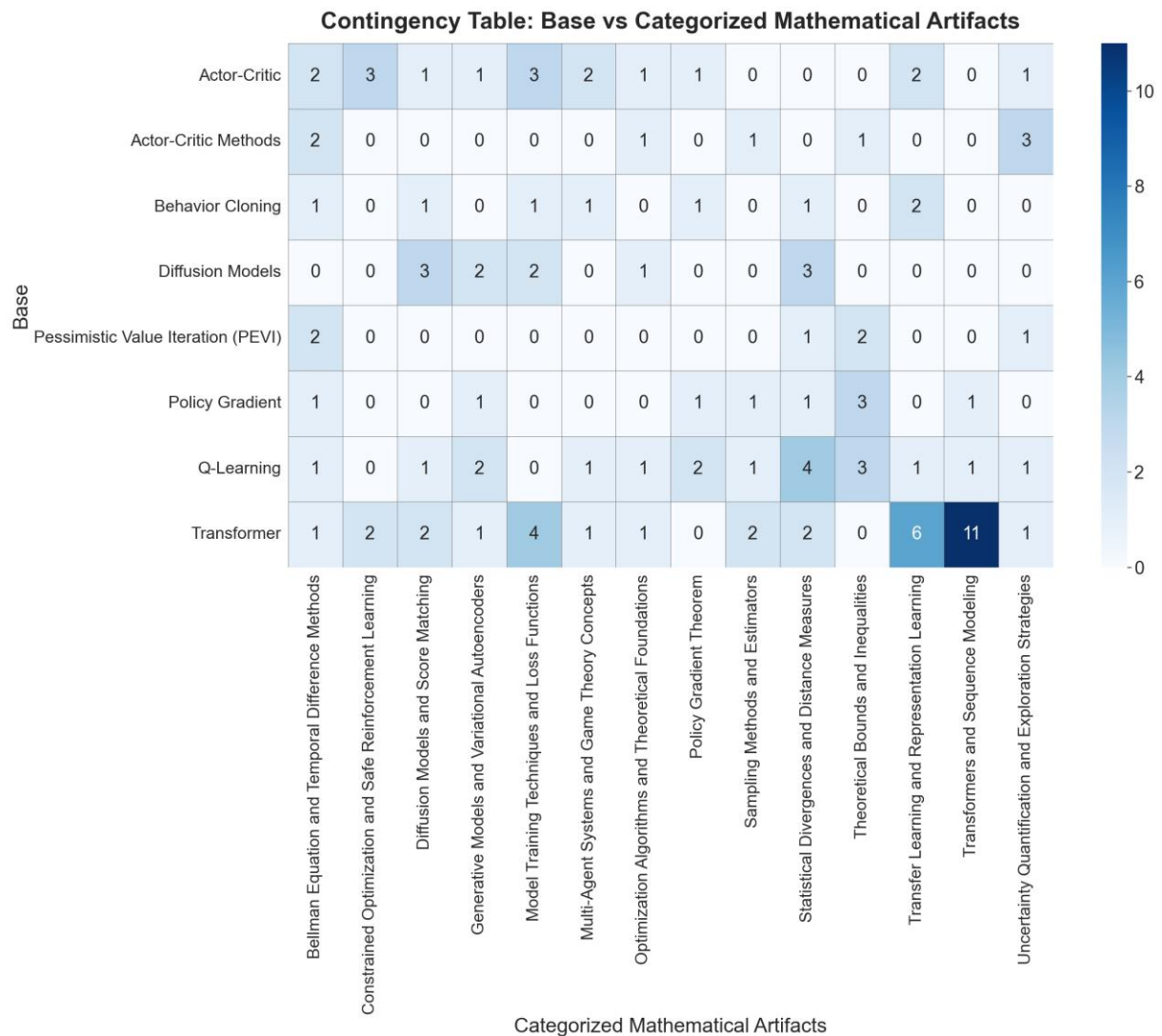


Figure 21. Contingency Matrix of the combinations of between Base Algorithm and Mathematical Artifacts for the column or row total above or equal to 5

Base vs Mathematical Artifacts	Frequency
Transformer & Transformers and Sequence Modeling	11
Transformer & Transfer Learning and Representation Learning	6
Q-Learning & Statistical Divergences and Distance Measures	4
Transformer & Model Training Techniques and Loss Functions	4
Policy Gradient & Theoretical Bounds and Inequalities	3

Table 6. Table of the TOP 5 most frequent combinations between the Base Algorithm and Mathematical Artifacts used

This section explores how statistical divergences and distance measures are employed in Q-Learning to address key challenges in reinforcement learning. The following analyses focus on combinations of base algorithms and mathematical artifacts, highlighting recent advancements in the field as demonstrated by selected papers.

In the pursuit of improving offline-to-online reinforcement learning, both Q Conditioned State Entropy Exploration (QCSE) (Zhang et al., 2024) and Spiked Random Model with Q-ensemble Independence Regularization (SPQR) (Lee et al., 2024) address challenges such as distribution shift and overestimation bias by employing statistical divergence measures, specifically the Kullback-Leibler (KL) divergence. The QCSE method aims to alleviate distribution shift between offline and online stages by implicitly realizing State Marginal Matching (SMM). This is achieved by minimizing the KL divergence between the marginal state distribution $\rho_\pi(s)$ of the policy and a target state distribution $p^*(s)$:

$$\min_{\pi} D_{\text{KL}}[\rho_\pi(s)|p^*(s)],$$

(Zhang et al., 2024), where the marginal state distribution is defined as:

$$\rho_\pi(s) = E_{\tau \sim \pi} \left[\frac{1}{N} \sum_{t=1}^T I(s_t = s) \right].$$

To approximate this minimization, QCSE maximizes the state entropy $H_\pi[s]$, encouraging exploration and helping align $\rho_\pi(s)$ with $p^*(s)$. The relationship is established by expressing state entropy maximization as:

$$\max_{\pi} [-\log p(s)] \leq - \int \rho_\pi(s) \log \rho_\pi(s) ds + \int p^*(s) \log p^*(s) ds,$$

indicating that maximizing the entropy of $\rho_\pi(s)$ implicitly minimizes $D_{\text{KL}}[\rho_\pi(s)|p^*(s)]$.

Since $\rho_\pi(s)$ is not directly accessible, QCSE approximates state entropy using the Kraskov-Stögbauer-Grassberger (KSG) estimator, which relies on nearest-neighbor distances in state space. The intrinsic reward is computed as:

$$r_{\text{QCSE}}(s, a) = \frac{1}{d_s} [\psi(\widehat{n_Q}(i) + 1) + \log 2 \cdot \max(|s_i - s_i^{\text{knn}}|, |\widehat{Q}(s, a) - \widehat{Q}(s, a)^{\text{knn}}|)],$$

where ψ is the digamma function, d_s is the dimensionality of the state space, and $\widehat{Q}(s, a)$ is the minimum of the double Q-networks. This reward, conditioned on Q-values, guides the policy to explore states that balance task performance and reduce distribution shift.

Conversely, the SPQR method focuses on controlling the independence of the Q-ensemble in ensemble Q-learning algorithms to prevent overestimation bias. It recognizes that the assumption of independent and identically distributed (i.i.d.) biases in ensemble methods is

often violated due to high correlations among Q-functions resulting from shared Bellman targets.

To address this, SPQR employs Random Matrix Theory and models the ensemble of Q-functions using a spiked Wishart model. The key idea is to detect and reduce "spikes" in the eigenvalue distribution of the Q-ensemble matrix, which indicate correlations among the Q-functions. The eigenvalue distribution $p_\psi(\lambda)$ of the Q-ensemble matrix Y is compared to the Wigner's semicircle distribution $p_{sc}(\lambda)$, representing the eigenvalue distribution of a random matrix with independent entries.

The divergence between these two distributions is quantified using the KL divergence (Lee et al., 2024):

$$L_{\text{SPQR}} = D_{\text{KL}}(p_\psi(\lambda) \| p_{sc}(\lambda)),$$

where L_{SPQR} serves as a regularization term in the loss function, encouraging the Q-ensemble matrix to have an eigenvalue distribution closer to that of an independent random matrix. Minimizing this divergence promotes independence among the Q-functions in the ensemble, reducing overestimation bias and improving learning stability.

In practice, the SPQR loss is computed by constructing a symmetric matrix YYY from the Q-values of the ensemble and normalizing it. The eigenvalues $\{\lambda_i\}$ are then calculated, and the empirical spectral density $p_\psi(\lambda)$ is formed. The KL divergence between $p_\psi(\lambda)$ and $p_{sc}(\lambda)$ is minimized:

$$L_{\text{SPQR}} = \frac{1}{N} \sum_{i=1}^N p_\psi(\lambda_i) \log \left(\frac{p_\psi(\lambda_i)}{p_{sc}(\lambda_i)} \right).$$

By integrating this regularization term into the training of the Q-functions, SPQR effectively controls the independence of the Q-ensemble, leading to better performance in both online and offline reinforcement learning tasks.

Both QCSE and SPQR leverage statistical divergence measures to address core challenges in offline reinforcement learning. QCSE minimizes the KL divergence between state distributions to mitigate distribution shift, while SPQR minimizes the KL divergence between spectral densities to promote independence among Q-functions. Despite targeting different aspects—state distribution matching versus ensemble independence—both methods use the KL divergence as a fundamental tool to guide the learning process toward more desirable behaviors, ultimately improving policy performance in offline-to-online reinforcement learning settings.

Similarly, integrating policy gradient methods with theoretical bounds and inequalities has resulted in novel techniques that improve policy optimization, offering theoretical guarantees and enhancing learning stability. The three papers present distinct approaches to improving

policy gradient methods by leveraging theoretical bounds and inequalities, each introducing novel techniques to enhance policy optimization.

In Value Enhancement of Reinforcement Learning via Efficient and Robust Trust Region Optimization (Shi et al., 2023), the authors propose a policy optimization algorithm that employs an efficient value difference estimator, denoted as η_1 . This estimator is constructed through a cross-fitting procedure, which leverages multiple nuisance functions estimated from separate data subsets to mitigate bias and overfitting. The core of their approach lies in solving a constrained problem:

$$\max_{\pi \in \Pi} \eta_1(\pi) \quad \text{subject to} \quad \frac{1}{L} \sum_{\ell=1}^L E_{S^* \sim \hat{d}^{(\ell)}, \nu} D_{\text{KL}}(\pi_{\text{old}}^{(\ell)}(\cdot | S^*), \pi(\cdot | S^*)) \leq \delta,$$

where Π is the policy class, D_{KL} denotes the Kullback-Leibler divergence, and δ is a threshold parameter controlling the trust region.

Goal-conditioned Offline Reinforcement Learning through State Space Partitioning (Wang et al., 2024) introduces a goal-conditioned offline reinforcement learning algorithm that utilizes state space partitioning to define a target region advantage function. The state space is partitioned based on the goal-conditioned value function $V^\pi(s, g)$ into regions $B_k(g)$, and a target region $G(s, g)$ is defined as the next higher-valued region. The algorithm then defines an auxiliary reward function that provides a positive reward when transitioning into the target region. The core idea is to combine the original goal-conditioned advantage $A^\pi(s, a, g)$ with the target region advantage $\tilde{A}^\pi(s, a, G(s, g))$ to form a dual-advantage weighting scheme:

$$w_t = \exp\left(\beta A^\pi(s_t, a_t, g) + \tilde{\beta} \tilde{A}^\pi(s_t, a_t, G(s_t, g))\right),$$

where β and $\tilde{\beta}$ are hyperparameters controlling the influence of each advantage term. This weighting is used in the policy optimization objective to prioritize actions that are advantageous both in terms of reaching the goal and progressing towards higher-valued regions. The theoretical analysis provides policy improvement guarantees by showing that the learned policy minimizes the Kullback-Leibler divergence from an implicit target policy that is biased towards higher advantages, ensuring that the policy is improved in expectation.

The Optimal Goal-Reaching Reinforcement Learning via Quasimetric Learning (Wang et al., 2023) paper presents a novel approach by framing the optimal value function V^* in goal-reaching tasks as a quasimetric, a generalization of a metric that does not require symmetry. The authors establish a value-quasimetric equivalence by proving that the negative optimal value function $-V^*$ belongs to the space of quasimetrics over the state space S :

$$\text{Qmet}(S) \equiv \{-V^*: V^* \text{ is the optimal value function of an MDP on } S\}.$$

Using this equivalence, they propose Quasimetric Reinforcement Learning (QRL), which directly learns the optimal value function by optimizing a quasimetric model d_θ that satisfies the triangle inequality and aligns with observed local transitions. The optimization objective is formulated as:

$$\max_{\theta} E_{s \sim p_{\text{state}}, g \sim p_{\text{goal}}} [d_\theta(s, g)] \quad \text{subject to} \quad E_{(s, a, s', r) \sim p_{\text{transition}}} [\text{ReLU}(d_\theta(s, s') + r)] \leq \varepsilon^2,$$

where $\varepsilon > 0$ is a small constant ensuring that the learned distances do not exceed the negative rewards (costs) of transitions. This approach leverages the triangle inequality inherent in quasimetrics to propagate local cost constraints globally, ensuring that the learned value function is consistent with both the observed data and the optimality criteria.

Comparatively, all three papers focus on improving policy optimization by incorporating theoretical bounds and leveraging mathematical structures to guide learning. The first paper uses a trust region constraint based on the KL divergence to ensure that the policy update does not deviate significantly from the behavior policy, with theoretical bounds on the value difference providing guarantees on policy improvement. The second paper employs state space partitioning and a dual-advantage weighting scheme to create a structured advantage function that emphasizes both immediate and long-term progress, with policy improvement guarantees derived from properties of the weighting scheme. The third paper fundamentally redefines the learning of the value function through the lens of quasimetrics, using the triangle inequality as a global constraint to ensure that the learned value function approximates the optimal value function across the entire state space, with theoretical results guaranteeing exact recovery under ideal conditions.

In conclusion, these combinations demonstrate the effective use of mathematical frameworks such as statistical divergences, distance measures, and theoretical bounds in advancing Q-Learning and policy gradient methods. By addressing challenges like distribution shift and overestimation bias, these approaches contribute to the development of more robust and efficient reinforcement learning algorithms.

5. CONCLUSIONS AND FUTURE WORKS

This master's thesis explores systematically the domain of Offline Reinforcement Learning (ORL), identifying key algorithms and mathematical tools through a frequency analysis and literature review. The study reinforces that despite the challenges inherent in Offline Reinforcement Learning, such as the lack of interactive data gathering, there are substantial opportunities for innovation through the application of various algorithmic techniques and mathematical innovations.

The frequency analysis clarifies which techniques are most effective for addressing challenges in Offline Reinforcement Learning. Techniques such as the various forms of Q-learning, neural network-based approaches, Transformers, and others have shown promise in improving the reliability and performance of ORL-based systems. Additionally, exploring the mathematical tools used in these algorithms aims to deepen the theoretical understanding of the fundamental concepts frequently employed in algorithm construction.

ORL seeks to derive effective policies from static datasets without further environmental interaction, yet it faces significant challenges due to distribution shifts between the offline data and the policy's state-action visitation. To address these challenges, interconnected approaches have been developed, including uncertainty quantification, statistical divergences, theoretical bounds, constrained optimization, and advanced architectures like Transformers. Uncertainty quantification and exploration strategies help guide policies towards underrepresented or high-value regions, while avoiding overestimation in out-of-distribution areas. Techniques such as model ensembles estimate uncertainty, penalize unreliable actions, and enhance robustness. Statistical divergences, such as Kullback-Leibler, quantify and minimize discrepancies between the learned policy's distribution and the offline data, ensuring that the policy remains within the data support, thereby mitigating distribution shift and preventing unsafe actions. Theoretical bounds and inequalities offer guarantees on policy improvement and stability by formulating optimization problems with trust regions or leveraging properties like the triangle inequality in quasimetric learning, ensuring consistent performance gains. Constrained optimization and safe reinforcement learning techniques enforce hard safety constraints or incorporate recovery mechanisms to avoid unsafe states, thereby maintaining safety while optimizing performance through defining feasible regions and making policy adjustments. Moreover, integrating transformer architectures enhances the modeling of environment dynamics and policy representations, thereby improving efficiency and generalization in ORL. Collectively, these interconnected methods address the core challenges of ORL, ensuring that despite the limitations of offline data, the learned policies are both effective and reliable, thus advancing the field.

Future work includes creating comprehensive tutorials on algorithms and mathematical techniques, with the aim of transforming this research into an introductory book for new researchers in the field.

BIBLIOGRAPHICAL REFERENCES

- Altman, E. (2021). *Constrained Markov Decision Processes: Stochastic Modeling* (1st ed.). Routledge. <https://doi.org/10.1201/9781315140223>
- Ball, P. J., Smith, L., Kostrikov, I., & Levine, S. (2023). *Efficient Online Reinforcement Learning with Offline Data* (arXiv:2302.02948). arXiv. <http://arxiv.org/abs/2302.02948>
- Batina, L., Gierlichs, B., Prouff, E., Rivain, M., Standaert, F.-X., & Veyrat-Charvillon, N. (2011). Mutual Information Analysis: A Comprehensive Study. *Journal of Cryptology*, 24(2), 269–291. <https://doi.org/10.1007/s00145-010-9084-8>
- Bhardwaj, M., Xie, T., Boots, B., Jiang, N., & Cheng, C.-A. (2023). *Adversarial Model for Offline Reinforcement Learning* (arXiv:2302.11048). arXiv. <http://arxiv.org/abs/2302.11048>
- Blanchet, J., Lu, M., Zhang, T., & Zhong, H. (2023). *Double Pessimism is Provably Efficient for Distributionally Robust Offline Reinforcement Learning: Generic Algorithm and Robust Partial Coverage* (arXiv:2305.09659). arXiv. <http://arxiv.org/abs/2305.09659>
- Boige, R., Flet-Berliac, Y., Flajolet, A., Richard, G., & Pierrot, T. (2023). *PASTA: Pretrained Action-State Transformer Agents* (arXiv:2307.10936). arXiv. <http://arxiv.org/abs/2307.10936>
- Carcassi, G., Aidala, C. A., & Barbour, J. (2021). Variability as a better characterization of Shannon entropy. *European Journal of Physics*, 42(4), 045102. <https://doi.org/10.1088/1361-6404/abe361>
- Carranza, A. G., & Athey, S. (2023). *Federated Offline Policy Learning with Heterogeneous Observational Data* (arXiv:2305.12407). arXiv. <http://arxiv.org/abs/2305.12407>

- Cen, Z., Liu, Z., Wang, Z., Yao, Y., Lam, H., & Zhao, D. (2024). *Learning from Sparse Offline Datasets via Conservative Density Estimation* (arXiv:2401.08819). arXiv. <http://arxiv.org/abs/2401.08819>
- Chebotar, Y., Vuong, Q., Irpan, A., Hausman, K., Xia, F., Lu, Y., Kumar, A., Yu, T., Herzog, A., Pertsch, K., Gopalakrishnan, K., Ibarz, J., Nachum, O., Sontakke, S., Salazar, G., Tran, H. T., Peralta, J., Tan, C., Manjunath, D., ... Levine, S. (2023). *Q-Transformer: Scalable Offline Reinforcement Learning via Autoregressive Q-Functions* (arXiv:2309.10150). arXiv. <http://arxiv.org/abs/2309.10150>
- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). *Decision Transformer: Reinforcement Learning via Sequence Modeling* (arXiv:2106.01345). arXiv. <http://arxiv.org/abs/2106.01345>
- Chen, L., Yan, J., Shao, Z., Wang, L., Lin, Q., Rajmohan, S., Moscibroda, T., & Zhang, D. (2023). *Conservative State Value Estimation for Offline Reinforcement Learning* (arXiv:2302.06884). arXiv. <http://arxiv.org/abs/2302.06884>
- Chen, Y., Li, H., & Zhao, D. (2024). *Boosting Continuous Control with Consistency Policy* (arXiv:2310.06343). arXiv. <http://arxiv.org/abs/2310.06343>
- Cheng, P., Zhan, X., Wu, Z., Zhang, W., Song, S., Wang, H., Lin, Y., & Jiang, L. (2023). *Look Beneath the Surface: Exploiting Fundamental Symmetry for Sample-Efficient Offline RL* (arXiv:2306.04220). arXiv. <http://arxiv.org/abs/2306.04220>

- Cipollone, R., Ronca, A., Jonsson, A., & Talebi, M. S. (2024). *Provably Efficient Offline Reinforcement Learning in Regular Decision Processes*. NeurIPS Proceedings. https://proceedings.neurips.cc/paper_files/paper/2023/file/7bf3e93543a612b75b6373178ba1faa4-Paper-Conference.pdf
- Corrado, N. E., Qu, Y., Balis, J. U., Labiosa, A., & Hanna, J. P. (2024). *Guided Data Augmentation for Offline Reinforcement Learning and Imitation Learning* (arXiv:2310.18247). arXiv. <http://arxiv.org/abs/2310.18247>
- Correia, A., & Alexandre, L. A. (2022). *Hierarchical Decision Transformer* (arXiv:2209.10447). arXiv. <http://arxiv.org/abs/2209.10447>
- Cousins, C., Petrik, M., Lobo, E. A., & Zick, Y. (2024). *Percentile Criterion Optimization in Offline Reinforcement Learning* (arXiv:2404.05055). arXiv. <https://arxiv.org/abs/2404.05055>
- Di Palo, N., Byravan, A., Hasenclever, L., Wulfmeier, M., Heess, N., & Riedmiller, M. (2023). *Towards A Unified Agent with Foundation Models* (arXiv:2307.09668). arXiv. <http://arxiv.org/abs/2307.09668>
- Di, Q., Zhao, H., He, J., & Gu, Q. (2023). *Pessimistic Nonlinear Least-Squares Value Iteration for Offline Reinforcement Learning* (arXiv:2310.01380). arXiv. <http://arxiv.org/abs/2310.01380>
- Ding, W., Che, T., Zhao, D., & Pavone, M. (2023). *Bayesian Reparameterization of Reward-Conditioned Reinforcement Learning with Energy-based Models* (arXiv:2305.11340). arXiv. <http://arxiv.org/abs/2305.11340>
- Ding, W., Majcherczyk, N., Deshpande, M., Qi, X., Zhao, D., Madhivanan, R., & Sen, A. (2023). *Learning to View: Decision Transformers for Active Object Detection* (arXiv:2301.09544). arXiv. <http://arxiv.org/abs/2301.09544>

- Ding, Z., & Jin, C. (2024). *Consistency Models as a Rich and Efficient Policy Class for Reinforcement Learning* (arXiv:2309.16984). arXiv. <http://arxiv.org/abs/2309.16984>
- Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., & Zhang, T. (2023). *RAFT: Reward rAnked FineTuning for Generative Foundation Model Alignment* (arXiv:2304.06767). arXiv. <http://arxiv.org/abs/2304.06767>
- Dong, J., Mo, W., Qi, Z., Shi, C., Fang, E. X., & Tarokh, V. (2023). *PASTA: Pessimistic Assortment Optimization*. Proceedings in Machine Learning Research. <https://proceedings.mlr.press/v202/dong23e/dong23e.pdf>
- Eysenbach, B., Geist, M., Levine, S., & Salakhutdinov, R. (2023). *A Connection between One-Step Regularization and Critic Regularization in Reinforcement Learning* (arXiv:2307.12968). arXiv. <http://arxiv.org/abs/2307.12968>
- Formanek, C., Jeewa, A., Shock, J., & Pretorius, A. (2023). *Off-the-Grid MARL: Datasets with Baselines for Offline Multi-Agent Reinforcement Learning* (arXiv:2302.00521). arXiv. <http://arxiv.org/abs/2302.00521>
- Fu, J., Kumar, A., Nachum, O., Tucker, G., & Levine, S. (2021). *D4RL: Datasets for Deep Data-Driven Reinforcement Learning* (arXiv:2004.07219). arXiv. <http://arxiv.org/abs/2004.07219>
- Fujimoto, S., Meger, D., & Precup, D. (2019). *Off-Policy Deep Reinforcement Learning without Exploration* (arXiv:1812.02900). arXiv. <http://arxiv.org/abs/1812.02900>
- Gabbianelli, G., Neu, G., Okolo, N., & Papini, M. (2023). *Offline Primal-Dual Reinforcement Learning for Linear MDPs* (arXiv:2305.12944). arXiv. <http://arxiv.org/abs/2305.12944>

- Gao, Y., Zhang, R., Guo, J., Wu, F., Yi, Q., Peng, S., Lan, S., Chen, R., Du, Z., Hu, X., Guo, Q., Li, L., & Chen, Y. (2023). *Context Shift Reduction for Offline Meta-Reinforcement Learning* (arXiv:2311.03695). arXiv. <http://arxiv.org/abs/2311.03695>
- Ghanem, A., Ciblat, P., & Ghogho, M. (2023). *Multi-Objective Decision Transformers for Offline Reinforcement Learning* (arXiv:2308.16379). arXiv. <http://arxiv.org/abs/2308.16379>
- Ghosh, D., Bhateja, C., & Levine, S. (2023). *Reinforcement Learning from Passive Data via Latent Intentions* (arXiv:2304.04782). arXiv. <http://arxiv.org/abs/2304.04782>
- Gu, P., Cai, X., Xing, D., Wang, X., Zhao, M., & An, B. (2024). *Offline RL with Discrete Proxy Representations for Generalizability in POMDPs*. NeurIPS Proceedings. https://proceedings.neurips.cc/paper_files/paper/2023/file/ce1c1ff5d94079dea348a2317a889281-Paper-Conference.pdf
- Guan, J., Chen, G., & Ji, J. (2023). *VOCE: Variational Optimization with Conservative Estimation for Offline Safe Reinforcement Learning*. NeurIPS Proceedings. https://proceedings.neurips.cc/paper_files/paper/2023/file/6a7c2a320f5f36bb98f8eb878c6f1180-Paper-Conference.pdf
- Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor* (arXiv:1801.01290). arXiv. <http://arxiv.org/abs/1801.01290>
- Hansen-Estruch, P., Kostrikov, I., Janner, M., Kuba, J. G., & Levine, S. (2023). *IDQL: Implicit Q-Learning as an Actor-Critic Method with Diffusion Policies* (arXiv:2304.10573). arXiv. <http://arxiv.org/abs/2304.10573>

- He, H. (2023). *A Survey on Offline Model-Based Reinforcement Learning* (arXiv:2305.03360). arXiv. <http://arxiv.org/abs/2305.03360>
- He, H., Bai, C., Xu, K., Yang, Z., Zhang, W., Wang, D., Zhao, B., & Li, X. (2023). *Diffusion Model is an Effective Planner and Data Synthesizer for Multi-Task Reinforcement Learning* (arXiv:2305.18459). arXiv. <http://arxiv.org/abs/2305.18459>
- He, H., Zhu, A., Liang, S., Chen, F., & Shao, J. (2023). *Decoupling Meta-Reinforcement Learning with Gaussian Task Contexts and Skills* (arXiv:2312.06518). arXiv. <http://arxiv.org/abs/2312.06518>
- He, L., Shen, L., Zhang, L., Tan, J., & Wang, X. (2024). *DiffCPS: Diffusion Model based Constrained Policy Search for Offline Reinforcement Learning* (arXiv:2310.05333). arXiv. <http://arxiv.org/abs/2310.05333>
- Hejna, J., Gao, J., & Sadigh, D. (2023). *Distance Weighted Supervised Learning for Offline Interaction Data* (arXiv:2304.13774). arXiv. <http://arxiv.org/abs/2304.13774>
- Hong, J., Dragan, A., & Levine, S. (2023). *Offline RL with Observation Histories: Analyzing and Improving Sample Complexity* (arXiv:2310.20663). arXiv. <http://arxiv.org/abs/2310.20663>
- Hong, J., Levine, S., & Dragan, A. (2023). *Learning to Influence Human Behavior with Offline Reinforcement Learning* (arXiv:2303.02265). arXiv. <http://arxiv.org/abs/2303.02265>
- Hong, K., Li, Y., & Tewari, A. (2023). *A Primal-Dual-Critic Algorithm for Offline Constrained Reinforcement Learning* (arXiv:2306.07818). arXiv. <http://arxiv.org/abs/2306.07818>
- Hong, M., Zhang, Z., Wu, Y., & Xu, Y. (2024). *MoMA: Model-based Mirror Ascent for Offline Reinforcement Learning* (arXiv:2401.11380). arXiv. <http://arxiv.org/abs/2401.11380>

- Hong, Z.-W., Agrawal, P., Combes, R. T. des, & Laroche, R. (2023). *Harnessing Mixed Offline Reinforcement Learning Datasets via Trajectory Weighting* (arXiv:2306.13085). arXiv. <http://arxiv.org/abs/2306.13085>
- Hong, Z.-W., Kumar, A., Karnik, S., Bhandwaldar, A., Srivastava, A., Pajarinen, J., Laroche, R., Gupta, A., & Agrawal, P. (2023). *Beyond Uniform Sampling: Offline Reinforcement Learning with Imbalanced Datasets* (arXiv:2310.04413). arXiv. <http://arxiv.org/abs/2310.04413>
- Hu, J., Sun, Y., Huang, S., Guo, S., Chen, H., Shen, L., Sun, L., Chang, Y., & Tao, D. (2023). *Instructed Diffuser with Temporal Condition Guidance for Offline Reinforcement Learning* (arXiv:2306.04875). arXiv. <http://arxiv.org/abs/2306.04875>
- Hu, S., Shen, L., Zhang, Y., & Tao, D. (2023a). *Graph Decision Transformer* (arXiv:2303.03747). arXiv. <http://arxiv.org/abs/2303.03747>
- Hu, S., Shen, L., Zhang, Y., & Tao, D. (2023b). *Prompt-Tuning Decision Transformer with Preference Ranking* (arXiv:2305.09648). arXiv. <http://arxiv.org/abs/2305.09648>
- Hu, X., & Leung, H. (2023). *Provably (More) Sample-Efficient Offline RL with Options*. NeurIPS Proceedings. https://papers.nips.cc/paper_files/paper/2023/file/d91b532a76ea98ac1ef5226b862bfc49-Paper-Conference.pdf
- Hu, X., Ma, Y., Xiao, C., Zheng, Y., & Hao, J. (2023). *Iteratively Refined Behavior Regularization for Offline Reinforcement Learning* (arXiv:2306.05726). arXiv. <http://arxiv.org/abs/2306.05726>
- Huang, K., Shen, L., Zhao, C., Yuan, C., & Tao, D. (2024). *Solving Continual Offline Reinforcement Learning with Decision Transformer* (arXiv:2401.08478). arXiv. <http://arxiv.org/abs/2401.08478>

- Islam, R., Sinha, S., Bharadhwaj, H., Arnob, S. Y., Yang, Z., Garg, A., Wang, Z., Li, L., & Precup, D. (2022). *Offline Policy Optimization in RL with Variance Regularization* (arXiv:2212.14405). arXiv. <http://arxiv.org/abs/2212.14405>
- Islam, R., Tomar, M., Lamb, A., Efroni, Y., Zang, H., Didolkar, A., Misra, D., Li, X., Seijen, H. V., Combes, R. T. D., & Langford, J. (2023). *Principled Offline RL in the Presence of Rich Exogenous Information*. Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v202/islam23a.html>
- Jain, V., & Ravanbakhsh, S. (2024). *Learning to Reach Goals via Diffusion* (arXiv:2310.02505). arXiv. <http://arxiv.org/abs/2310.02505>
- Jang, Y., Kim, G.-H., Lee, J., Sohn, S., Kim, B., Lee, H., & Lee, M. (2023). *SafeDICE: Offline Safe Imitation Learning with Non-Preferred Demonstrations*. NeurIPS Proceedings. https://proceedings.neurips.cc/paper_files/paper/2023/file/ed2fb79f2664c3d9ba878be7e575b2af-Paper-Conference.pdf
- Janner, M., Li, Q., & Levine, S. (2021). *Offline Reinforcement Learning as One Big Sequence Modeling Problem* (arXiv:2106.02039). arXiv. <http://arxiv.org/abs/2106.02039>
- Jiang, K., Yao, J., & Tan, X. (2023). *Recovering from Out-of-sample States via Inverse Dynamics in Offline Reinforcement Learning*. OpenReview. <https://openreview.net/forum?id=4gLWjSaw4o¬eId=J97B9oi5jz>
- Jiang, L., Cheng, S., Qiu, J., Xu, H., Chan, W. K., & Ding, Z. (2024). *Offline Reinforcement Learning with Imbalanced Datasets* (arXiv:2307.02752). arXiv. <http://arxiv.org/abs/2307.02752>

- Kang, B., Ma, X., Du, C., Pang, T., & Yan, S. (2023). *Efficient Diffusion Policies for Offline Reinforcement Learning* (arXiv:2305.20081). arXiv. <http://arxiv.org/abs/2305.20081>
- Kang, J., Laroché, R., Yuan, X., Trischler, A., Liu, X., & Fu, J. (2024). *Think Before You Act: Decision Transformers with Working Memory* (arXiv:2305.16338). arXiv. <http://arxiv.org/abs/2305.16338>
- Killian, T. W., Parbhoo, S., & Ghassemi, M. (2023). *Risk Sensitive Dead-end Identification in Safety-Critical Offline Reinforcement Learning* (arXiv:2301.05664). arXiv. <http://arxiv.org/abs/2301.05664>
- Kim, B., & Oh, M. (2023). *Model-based Offline Reinforcement Learning with Count-based Conservatism* (arXiv:2307.11352). arXiv. <https://arxiv.org/abs/2307.11352>
- Kim, J., Lee, S., Kim, W., & Sung, Y. (2024a). *Decision ConvFormer: Local Filtering in MetaFormer is Sufficient for Decision Making* (arXiv:2310.03022). arXiv. <http://arxiv.org/abs/2310.03022>
- Kim, J., Lee, S., Kim, W., & Sung, Y. (2024b). *Value-Aided Conditional Supervised Learning for Offline RL* (arXiv:2402.02017). arXiv. <http://arxiv.org/abs/2402.02017>
- Kingma, D. P., & Welling, M. (2022). *Auto-Encoding Variational Bayes* (arXiv:1312.6114). arXiv. <http://arxiv.org/abs/1312.6114>
- Kostrikov, I., Nair, A., & Levine, S. (2021). *Offline Reinforcement Learning with Implicit Q-Learning* (arXiv:2110.06169). arXiv. <http://arxiv.org/abs/2110.06169>
- Krishnamurthy, S. K., Modi, S., Gangwani, T., Katariya, S., Kveton, B., & Rangi, A. (2024). *Selective Uncertainty Propagation in Offline RL* (arXiv:2302.00284). arXiv. <http://arxiv.org/abs/2302.00284>

- Kumar, A., Fu, J., Tucker, G., & Levine, S. (2019). *Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction* (arXiv:1906.00949). arXiv. <http://arxiv.org/abs/1906.00949>
- Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). *Conservative Q-Learning for Offline Reinforcement Learning* (arXiv:2006.04779). arXiv. <http://arxiv.org/abs/2006.04779>
- Lawson, D., & Qureshi, A. H. (2023). *Merging Decision Transformers: Weight Averaging for Forming Multi-Task Policies* (arXiv:2303.07551). arXiv. <http://arxiv.org/abs/2303.07551>
- Lee, D., Han, S., Cho, T., & Lee, J. (2024). *SPQR: Controlling Q-ensemble Independence with Spiked Random Model for Reinforcement Learning* (arXiv:2401.03137). arXiv. <http://arxiv.org/abs/2401.03137>
- Lee, J. N., Xie, A., Pacchiano, A., Chandak, Y., Finn, C., Nachum, O., & Brunskill, E. (2023). *Supervised Pretraining Can Learn In-Context Reinforcement Learning* (arXiv:2306.14892). arXiv. <http://arxiv.org/abs/2306.14892>
- Lei, K., He, Z., Lu, C., Hu, K., Gao, Y., & Xu, H. (2024). *Uni-O4: Unifying Online and Offline Deep Reinforcement Learning with Multi-Step On-Policy Optimization* (arXiv:2311.03351). arXiv. <http://arxiv.org/abs/2311.03351>
- Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems* (arXiv:2005.01643). arXiv. <http://arxiv.org/abs/2005.01643>
- Li, A., Boots, B., & Cheng, C.-A. (2023). *MAHALO: Unifying Offline Reinforcement Learning and Imitation Learning from Observations* (arXiv:2303.17156). arXiv. <http://arxiv.org/abs/2303.17156>

- Li, A., Misra, D., Kolobov, A., & Cheng, C.-A. (2023). *Survival Instinct in Offline Reinforcement Learning* (arXiv:2306.03286). arXiv. <http://arxiv.org/abs/2306.03286>
- Li, C., Tang, C., Nishimura, H., Mercat, J., Tomizuka, M., & Zhan, W. (2024). *Residual Q-Learning: Offline and Online Policy Customization without Value* (arXiv:2306.09526). arXiv. <http://arxiv.org/abs/2306.09526>
- Li, D., Huang, R., Shen, C., & Yang, J. (2023). *Near-optimal Conservative Exploration in Reinforcement Learning under Episode-wise Constraints* (arXiv: 2306.06265). arXiv. <https://arxiv.org/abs/2306.06265>
- Li, G., Ma, C., & Srebro, N. (2022). *Pessimism for Offline Linear Contextual Bandits using ℓ_p Confidence Sets* (arXiv:2205.10671). arXiv. <http://arxiv.org/abs/2205.10671>
- Li, G., Shan, Y., Zhu, Z., Long, T., & Zhang, W. (2024). *DiffStitch: Boosting Offline Reinforcement Learning with Diffusion-based Trajectory Stitching* (arXiv:2402.02439). arXiv. <http://arxiv.org/abs/2402.02439>
- Li, H., Zhou, X.-H., Xie, X.-L., Liu, S.-Q., Feng, Z.-Q., Liu, X.-Y., Gui, M.-J., Xiang, T.-Y., Huang, D.-X., Yao, B.-X., & Hou, Z.-G. (2023). *CROP: Conservative Reward for Model-based Offline Policy Optimization* (arXiv:2310.17245). arXiv. <http://arxiv.org/abs/2310.17245>
- Li, J., Liu, X., Zhu, B., Jiao, J., Tomizuka, M., Tang, C., & Zhan, W. (2023). *Guided Online Distillation: Promoting Safe Reinforcement Learning by Offline Demonstration* (arXiv:2309.09408). arXiv. <http://arxiv.org/abs/2309.09408>
- Li, J., Liu, X., Zhu, B., Jiao, J., Tomizuka, M., Tang, C., & Zhan, W. (2023). *Guided Online Distillation: Promoting Safe Reinforcement Learning by Offline Demonstration* (arXiv:2309.09408). arXiv. <http://arxiv.org/abs/2309.09408>

- Li, L., Zhang, H., Zhang, X., Zhu, S., Zhao, J., & Heng, P.-A. (2024). *Towards an Information Theoretic Framework of Context-Based Offline Meta-Reinforcement Learning* (arXiv:2402.02429). arXiv. <http://arxiv.org/abs/2402.02429>
- Li, T., Guevara, J., Xie, X., & Zhu, Q. (2023). *Self-Confirming Transformer for Locally Consistent Online Adaptation in Multi-Agent Reinforcement Learning* (arXiv:2310.04579). arXiv. <http://arxiv.org/abs/2310.04579>
- Li, W., Wang, X., Jin, B., & Zha, H. (2023). *Hierarchical Diffusion for Offline Decision Making*. Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v202/li23ad.html>
- Li, Z., Xu, K., Liu, L., Li, L., Ye, D., & Zhao, P. (2023). *Deploying Offline Reinforcement Learning with Human Feedback* (arXiv:2303.07046). arXiv. <http://arxiv.org/abs/2303.07046>
- Li, Z., Yang, Z., & Wang, M. (2023). *Reinforcement Learning with Human Feedback: Learning Dynamic Choices via Pessimism* (arXiv:2305.18438). arXiv. <http://arxiv.org/abs/2305.18438>
- Lian, S., Ma, Y., Liu, J., Zheng, Y., & Meng, Z. (2023). *HIPODE: Enhancing Offline Reinforcement Learning with High-Quality Synthetic Data from a Policy-Decoupled Approach* (arXiv:2306.06329). arXiv. <http://arxiv.org/abs/2306.06329>
- Lin, Q., Liu, H., & Sengupta, B. (2022). *Switch Trajectory Transformer with Distributional Value Approximation for Multi-Task Reinforcement Learning* (arXiv:2203.07413). arXiv. <http://arxiv.org/abs/2203.07413>
- Lin, Q., Liu, H., & Sengupta, B. (2022). *Switch Trajectory Transformer with Distributional Value Approximation for Multi-Task Reinforcement Learning* (arXiv:2203.07413). arXiv. <http://arxiv.org/abs/2203.07413>

- Lin, Q., Tang, B., Wu, Z., Yu, C., Mao, S., Xie, Q., Wang, X., & Wang, D. (2024). *Safe Offline Reinforcement Learning with Real-Time Budget Constraints* (arXiv:2306.00603). arXiv. <http://arxiv.org/abs/2306.00603>
- Lin, Q., Yu, C., Liu, Z., & Wu, Z. (2024). *Policy-regularized Offline Multi-objective Reinforcement Learning* (arXiv:2401.02244). arXiv. <http://arxiv.org/abs/2401.02244>
- Liu, J., Ma, Y., Hao, J., Hu, Y., Zheng, Y., Lv, T., & Fan, C. (2023). *Prioritized Trajectory Replay: A Replay Memory for Data-driven Reinforcement Learning* (arXiv:2306.15503). arXiv. <http://arxiv.org/abs/2306.15503>
- Liu, J., Zhang, H., Zhuang, Z., Kang, Y., Wang, D., & Wang, B. (2023). *Design from Policies: Conservative Test-Time Adaptation for Offline Policy Optimization* (arXiv:2306.14479). arXiv. <http://arxiv.org/abs/2306.14479>
- Liu, J., Zhang, Z., Wei, Z., Zhuang, Z., Kang, Y., Gai, S., & Wang, D. (2023). *Beyond OOD State Actions: Supported Cross-Domain Offline Reinforcement Learning* (arXiv:2306.12755). arXiv. <http://arxiv.org/abs/2306.12755>
- Liu, J., Zu, L., He, L., & Wang, D. (2023). *CLUE: Calibrated Latent Guidance for Offline Reinforcement Learning* (arXiv:2306.13412). arXiv. <http://arxiv.org/abs/2306.13412>
- Liu, X., Liu, A., Broeck, G. V. den, & Liang, Y. (2024). *A Tractable Inference Perspective of Offline RL* (arXiv:2311.00094). arXiv. <http://arxiv.org/abs/2311.00094>
- Liu, X.-Y., Zhou, X.-H., Li, G., Li, H., Gui, M.-J., Xiang, T.-Y., Huang, D.-X., & Hou, Z.-G. (2024). *MICRO: Model-Based Offline Reinforcement Learning with a Conservative Bellman Operator* (arXiv:2312.03991). arXiv. <http://arxiv.org/abs/2312.03991>

- Liu, X.-Y., Zhou, X.-H., Xie, X.-L., Liu, S.-Q., Feng, Z.-Q., Li, H., Gui, M.-J., Xiang, T.-Y., Huang, D.-X., & Hou, Z.-G. (2024). *DOMAIN: Mildly Conservative Model-BAsed OfflIne Reinforcement Learning* (arXiv:2309.08925). arXiv. <http://arxiv.org/abs/2309.08925>
- Liu, Y., Chaudhari, P., & Fakoor, R. (2024). *Budgeting Counterfactual for Offline RL* (arXiv:2307.06328). arXiv. <http://arxiv.org/abs/2307.06328>
- Liu, Z., Guo, Z., Yao, Y., Cen, Z., Yu, W., Zhang, T., & Zhao, D. (2023). *Constrained Decision Transformer for Offline Safe Reinforcement Learning* (arXiv:2302.07351). arXiv. <https://arxiv.org/abs/2302.07351>
- Lu, C., Ball, P. J., Teh, Y. W., & Parker-Holder, J. (2023). *Synthetic Experience Replay* (arXiv:2303.06614). arXiv. <http://arxiv.org/abs/2303.06614>
- Lu, C., Chen, H., Chen, J., Su, H., Li, C., & Zhu, J. (2023). *Contrastive Energy Prediction for Exact Energy-Guided Diffusion Sampling in Offline Reinforcement Learning* (arXiv:2304.12824). arXiv. <http://arxiv.org/abs/2304.12824>
- Luis, C. E., Bottero, A. G., Vinogradska, J., Berkenkamp, F., & Peters, J. (2023). *Model-Based Epistemic Variance of Values for Risk-Aware Policy Optimization* (arXiv:2312.04386). arXiv. <http://arxiv.org/abs/2312.04386>
- Luo, F.-M., Xu, T., Cao, X., & Yu, Y. (2023). *Reward-Consistent Dynamics Models are Strongly Generalizable for Offline Reinforcement Learning* (arXiv:2310.05422). arXiv. <http://arxiv.org/abs/2310.05422>
- Lv, O., & Zhou, B. (2023). *LLQL: Logistic Likelihood Q-Learning for Reinforcement Learning* (arXiv:2307.02345). arXiv. <http://arxiv.org/abs/2307.02345>

- Ly, A., Marsman, M., Verhagen, J., Grasman, R., & Wagenmakers, E.-J. (2017). *A Tutorial on Fisher Information* (arXiv:1705.01064). arXiv. <http://arxiv.org/abs/1705.01064>
- Ma, X., Kang, B., Xu, Z., Lin, M., & Yan, S. (2024). *Mutual Information Regularized Offline Reinforcement Learning* (arXiv:2210.07484). arXiv. <http://arxiv.org/abs/2210.07484>
- Ma, Y., Tang, H., Li, D., & Meng, Z. (n.d.). *Reining Generalization in Offline Reinforcement Learning via Representation Distinction*. NeurIPS Proceedings. https://proceedings.neurips.cc/paper_files/paper/2023/file/802a4350ca4fced76b13b8b320af1543-Paper-Conference.pdf
- Ma, Y., Xiao, C., Liang, H., & Hao, J. (2023). *Rethinking Decision Transformer via Hierarchical Reinforcement Learning* (arXiv:2311.00267). arXiv. <http://arxiv.org/abs/2311.00267>
- Mao, C. (2023). *Offline Reinforcement Learning with Additional Covering Distributions* (arXiv:2305.12679). arXiv. <http://arxiv.org/abs/2305.12679>
- Mao, H., Zhao, R., Chen, H., Hao, J., Chen, Y., Li, D., Zhang, J., & Xiao, Z. (2023). *Transformer in Transformer as Backbone for Deep Reinforcement Learning* (arXiv:2212.14538). arXiv. <http://arxiv.org/abs/2212.14538>
- Mao, H., Zhao, R., Li, Z., Xu, Z., Chen, H., Chen, Y., Zhang, B., Xiao, Z., Zhang, J., & Yin, J. (2023). *PDiT: Interleaving Perception and Decision-making Transformers for Deep Reinforcement Learning* (arXiv:2312.15863). arXiv. <http://arxiv.org/abs/2312.15863>
- Mao, Y., Zhang, H., Chen, C., Xu, Y., & Ji, X. (2023). *Supported Trust Region Optimization for Offline Reinforcement Learning* (arXiv:2311.08935). arXiv. <http://arxiv.org/abs/2311.08935>

Mao, Y., Zhang, H., Chen, C., Xu, Y., & Ji, X. (2023). *Supported Value Regularization for Offline Reinforcement Learning*. OpenReview.

<https://openreview.net/forum?id=fze7P9oy6l¬elid=SI50qnxg1o>

Mark, M. S., Sharma, A., Tajwar, F., Rafailov, R., Levine, S., & Finn, C. (2023). *Offline Retraining for Online RL: Decoupled Policy Learning to Mitigate Exploration Bias* (arXiv:2310.08558). arXiv.

<http://arxiv.org/abs/2310.08558>

Mathieu, M., Ozair, S., Srinivasan, S., Gulcehre, C., Zhang, S., Jiang, R., Paine, T. L., Powell, R., Źořna, K., Schrittwieser, J., Choi, D., Georgiev, P., Toyama, D., Huang, A., Ring, R., Babuschkin, I., Ewalds, T., Bordbar, M., Henderson, S., ... Vinyals, O. (2023). *AlphaStar Unplugged: Large-Scale Offline Reinforcement Learning* (arXiv:2308.03526). arXiv. <http://arxiv.org/abs/2308.03526>

Matsunaga, D. E., Lee, J., Yoon, J., Leonardos, S., Abbeel, P., & Kim, K.-E. (2023). *AlberDICE: Addressing Out-Of-Distribution Joint Actions in Offline Multi-Agent RL via Alternating Stationary Distribution Correction Estimation* (arXiv:2311.02194). arXiv.

<http://arxiv.org/abs/2311.02194>

Mazouze, B., Bruce, J., Precup, D., Fergus, R., & Anand, A. (2023). *Accelerating exploration and representation learning with offline pre-training* (arXiv:2304.00046). arXiv.

<http://arxiv.org/abs/2304.00046>

McInroe, T., Jelley, A., Albrecht, S. V., & Storkey, A. (2024). *Planning to Go Out-of-Distribution in Offline-to-Online Reinforcement Learning* (arXiv:2310.05723). arXiv.

<http://arxiv.org/abs/2310.05723>

- Missaoui, O. (2023). Markov Decision Processes: A Gentle Tutorial. *SSRN Electronic Journal*.
<https://doi.org/10.2139/ssrn.4535241>
- Mittal, D., & Lee, W. S. (2024). *Differentiable Tree Search in Latent State Space* (arXiv:2401.11660).
 arXiv. <http://arxiv.org/abs/2401.11660>
- Miura, K. (2011). An Introduction to Maximum Likelihood Estimation and Information Geometry. *Interdisciplinary Information Sciences*, 17(3), 155–174. <https://doi.org/10.4036/iis.2011.155>
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533. <https://doi.org/10.1038/nature14236>
- Modhe, N., Gao, Q., Kalyan, A., Batra, D., Thattai, G., & Sukhatme, G. (2023). *Exploiting Generalization in Offline Reinforcement Learning via Unseen State Augmentations* (arXiv:2308.03882). arXiv. <http://arxiv.org/abs/2308.03882>
- Nakamoto, M., Zhai, Y., Singh, A., Mark, M. S., Ma, Y., Finn, C., Kumar, A., & Levine, S. (2024). *Cal-QL: Calibrated Offline RL Pre-Training for Efficient Online Fine-Tuning* (arXiv:2303.05479).
 arXiv. <http://arxiv.org/abs/2303.05479>
- Nguyen-Tang, T., & Arora, R. (2024). *On Sample-Efficient Offline Reinforcement Learning: Data Diversity, Posterior Sampling, and Beyond* (arXiv:2401.03301). arXiv.
<http://arxiv.org/abs/2401.03301>
- Ni, F., Hao, J., Mu, Y., Yuan, Y., Zheng, Y., Wang, B., & Liang, Z. (2023). *MetaDiffuser: Diffusion Model as Conditional Planner for Offline Meta-RL* (arXiv:2305.19923). arXiv.
<http://arxiv.org/abs/2305.19923>

- Nikulin, A., Kurenkov, V., Tarasov, D., & Kolesnikov, S. (2023). *Anti-Exploration by Random Network Distillation* (arXiv:2301.13616). arXiv. <https://arxiv.org/abs/2301.13616>
- Niu, H., Ji, T., Liu, B., Zhao, H., Zhu, X., Zheng, J., Huang, P., Zhou, G., Hu, J., & Zhan, X. (2023). *H2O+: An Improved Framework for Hybrid Offline-and-Online RL with Dynamics Gaps* (arXiv:2309.12716). arXiv. <http://arxiv.org/abs/2309.12716>
- Pace, A., Yèche, H., Schölkopf, B., Rätsch, G., & Tennenholtz, G. (2023). *Delphic Offline Reinforcement Learning under Nonidentifiable Hidden Confounding* (arXiv:2306.01157). arXiv. <http://arxiv.org/abs/2306.01157>
- Palvia, P., Leary, D., Mao, E., Midha, V., Pinjani, P., & Salam, A. F. (2004). Research Methodologies in MIS: An Update. *Communications of the Association for Information Systems*, 14. <https://doi.org/10.17705/1CAIS.01424>
- Panaganti, K., Xu, Z., Kalathil, D., & Ghavamzadeh, M. (2023). *Bridging Distributionally Robust Learning and Offline RL: An Approach to Mitigate Distribution Shift and Partial Data Coverage* (arXiv:2310.18434). arXiv. <http://arxiv.org/abs/2310.18434>
- Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2), 183–199. <https://doi.org/10.1016/j.im.2014.08.008>
- Park, S., Ghosh, D., Eysenbach, B., & Levine, S. (2024). *HIQL: Offline Goal-Conditioned RL with Latent States as Actions* (arXiv:2307.11949). arXiv. <http://arxiv.org/abs/2307.11949>
- Pathmanathan, P., Díaz-Rodríguez, N., & Del Ser, J. (2024). Using Curiosity for an Even Representation of Tasks in Continual Offline Reinforcement Learning. *Cognitive Computation*, 16(1), 425–453. <https://doi.org/10.1007/s12559-023-10213-9>

- Pinneri, C., Bechtle, S., Wulfmeier, M., Byravan, A., Zhang, J., Whitney, W. F., & Riedmiller, M. (2023). *Equivariant Data Augmentation for Generalization in Offline Reinforcement Learning* (arXiv:2309.07578). arXiv. <http://arxiv.org/abs/2309.07578>
- Prudencio, R. F., Maximo, M. R. O. A., & Colombini, E. L. (2024). A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE Transactions on Neural Networks and Learning Systems*, 1–0. <https://doi.org/10.1109/TNNLS.2023.3250269>
- Qiao, D., & Wang, Y.-X. (2023). *Offline Reinforcement Learning with Differential Privacy* (arXiv:2206.00810). arXiv. <http://arxiv.org/abs/2206.00810>
- Ramchandran, S., Tikhonov, G., Lönnroth, O., Tiikkainen, P., & Lähdesmäki, H. (2022). *Learning Conditional Variational Autoencoders with Missing Covariates* (arXiv:2203.01218). arXiv. <http://arxiv.org/abs/2203.01218>
- Ran, Y., Li, Y.-C., Zhang, F., Zhang, Z., & Yu, Y. (2023). *Policy Regularization with Dataset Constraint for Offline Reinforcement Learning* (arXiv:2306.06569). arXiv. <http://arxiv.org/abs/2306.06569>
- Rengarajan, D., Ragothaman, N., Kalathil, D., & Shakkottai, S. (2023). *Federated Ensemble-Directed Offline Reinforcement Learning* (arXiv:2305.03097). arXiv. <http://arxiv.org/abs/2305.03097>
- Rigter, M., Lacerda, B., & Hawes, N. (2023). *One Risk to Rule Them All: A Risk-Sensitive Perspective on Model-Based Offline Reinforcement Learning* (arXiv:2212.00124). arXiv. <http://arxiv.org/abs/2212.00124>
- Roderick, M., Manek, G., Berkenkamp, F., & Kolter, J. Z. (2023). *Projected Off-Policy Q-Learning (POP-QL) for Stabilizing Offline Reinforcement Learning* (arXiv:2311.14885). arXiv. <http://arxiv.org/abs/2311.14885>

- Schmied, T., Hofmarcher, M., Paischer, F., Pascanu, R., & Hochreiter, S. (2023). *Learning to Modulate pre-trained Models in RL* (arXiv:2306.14884). arXiv. <http://arxiv.org/abs/2306.14884>
- Serfling, R. J. (1974). Probability Inequalities for the Sum in Sampling without Replacement. *The Annals of Statistics*, 2(1), 39–48. <https://doi.org/10.1214/aos/1176342611>
- Shaham, U., & Levy, O. (2022). *What Do You Get When You Cross Beam Search with Nucleus Sampling?* (arXiv:2107.09729). arXiv. <http://arxiv.org/abs/2107.09729>
- Shao, J., Qu, Y., Chen, C., Zhang, H., & Ji, X. (2023). *Counterfactual Conservative Q Learning for Offline Multi-agent Reinforcement Learning* (arXiv:2309.12696). arXiv. <http://arxiv.org/abs/2309.12696>
- Shi, C., Qi, Z., Wang, J., & Zhou, F. (2023). *Value Enhancement of Reinforcement Learning via Efficient and Robust Trust Region Optimization* (arXiv:2301.02220). arXiv. <http://arxiv.org/abs/2301.02220>
- Shi, C., Xiong, W., Shen, C., & Yang, J. (2023). *Provably Efficient Offline Reinforcement Learning with Perturbed Data Sources* (arXiv:2306.08364). arXiv. <http://arxiv.org/abs/2306.08364>
- Shi, R., Liu, Y., Ze, Y., Du, S. S., & Xu, H. (2023). *Unleashing the Power of Pre-trained Language Models for Offline Reinforcement Learning* (arXiv:2310.20587). arXiv. <http://arxiv.org/abs/2310.20587>
- Shlens, J. (2014). *Notes on Kullback-Leibler Divergence and Likelihood* (arXiv:1404.2000). arXiv. <http://arxiv.org/abs/1404.2000>

- Sikchi, H., Chitnis, R., Touati, A., Geramifard, A., Zhang, A., & Niekum, S. (2024). *SMORE: Score Models for Offline Goal-Conditioned Reinforcement Learning* (arXiv:2311.02013). arXiv. <http://arxiv.org/abs/2311.02013>
- Singh, A., Kumar, A., Vuong, Q., Chebotar, Y., & Levine, S. (2022). *Offline RL With Realistic Datasets: Heteroskedasticity and Support Constraints* (arXiv:2211.01052). arXiv. <http://arxiv.org/abs/2211.01052>
- Stooke, A., Achiam, J., & Abbeel, P. (2020). *Responsive Safety in Reinforcement Learning by PID Lagrangian Methods* (arXiv:2007.03964). arXiv. <http://arxiv.org/abs/2007.03964>
- Sudhakaran, S., & Risi, S. (2023). *Skill Decision Transformer* (arXiv:2301.13573). arXiv. <http://arxiv.org/abs/2301.13573>
- Sun, C., Qian, H., & Miao, C. (2023). *CUDC: A Curiosity-Driven Unsupervised Data Collection Method with Adaptive Temporal Distances for Offline Reinforcement Learning* (arXiv:2312.12191). arXiv. <http://arxiv.org/abs/2312.12191>
- Sun, H., Hüyük, A., Jarrett, D., & van der Schaar, M. (2023). *Accountability in Offline Reinforcement Learning: Explaining Decisions with a Corpus of Examples* (arXiv:2310.07747). arXiv. <http://arxiv.org/abs/2310.07747>
- Sun, Y., Zhang, J., Jia, C., Lin, H., Ye, J., & Yu, Y. (2023). *Model-Bellman Inconsistency for Model-based Offline Reinforcement Learning*. Proceedings of Machine Learning Research. <https://proceedings.mlr.press/v202/sun23g/sun23g.pdf>
- Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT Press. <https://mitpress.mit.edu/9780262039246/reinforcement-learning/>

- Tarasov, D., Kurenkov, V., Nikulin, A., & Kolesnikov, S. (2023). *Revisiting the Minimalist Approach to Offline Reinforcement Learning* (arXiv:2305.09836). arXiv. <http://arxiv.org/abs/2305.09836>
- Uehara, M., Shi, C., & Kallus, N. (2022). *A Review of Off-Policy Evaluation in Reinforcement Learning* (arXiv:2212.06355). arXiv. <http://arxiv.org/abs/2212.06355>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. <http://arxiv.org/abs/1706.03762>
- Venkatraman, S., Khaitan, S., Akella, R. T., Dolan, J., Schneider, J., & Berseth, G. (2023). *Reasoning with Latent Diffusion in Offline Reinforcement Learning* (arXiv:2309.06599). arXiv. <http://arxiv.org/abs/2309.06599>
- Wang, H., Lin, S., & Zhang, J. (2023). *Warm-Start Actor-Critic: From Approximation Error to Sub-optimality Gap* (arXiv:2306.11271). arXiv. <http://arxiv.org/abs/2306.11271>
- Wang, J., Zhang, J., Jiang, H., Zhang, J., Wang, L., & Zhang, C. (2023). *Offline Meta Reinforcement Learning with In-Distribution Online Adaptation* (arXiv:2305.19529). arXiv. <http://arxiv.org/abs/2305.19529>
- Wang, M., Jin, Y., & Montana, G. (2024). *Goal-conditioned Offline Reinforcement Learning through State Space Partitioning* (arXiv:2303.09367). arXiv. <http://arxiv.org/abs/2303.09367>
- Wang, M., Yang, R., Chen, X., Sun, H., Fang, M., & Montana, G. (2024). *GOPlan: Goal-conditioned Offline Reinforcement Learning by Planning with Learned Models* (arXiv:2310.20025). arXiv. <http://arxiv.org/abs/2310.20025>

- Wang, P., Zhu, M., & Shen, S. (2023). *Environment Transformer and Policy Optimization for Model-Based Offline Reinforcement Learning* (arXiv:2303.03811). arXiv. <http://arxiv.org/abs/2303.03811>
- Wang, Q., McCarthy, R., Bulens, D. C., McGuinness, K., O'Connor, N. E., Gürtler, N., Widmaier, F., Sanchez, F. R., & Redmond, S. J. (2023). *Improving Behavioural Cloning with Positive Unlabeled Learning* (arXiv:2301.11734). arXiv. <http://arxiv.org/abs/2301.11734>
- Wang, S., Yang, Q., Gao, J., Lin, M. G., Chen, H., Wu, L., Jia, N., Song, S., & Huang, G. (2023). *Train Once, Get a Family: State-Adaptive Balances for Offline-to-Online Reinforcement Learning* (arXiv:2310.17966). arXiv. <http://arxiv.org/abs/2310.17966>
- Wang, T., Torralba, A., Isola, P., & Zhang, A. (2023). *Optimal Goal-Reaching Reinforcement Learning via Quasimetric Learning* (arXiv:2304.01203). arXiv. <http://arxiv.org/abs/2304.01203>
- Wang, X., Xu, H., Zheng, Y., & Zhan, X. (2023). *Offline Multi-Agent Reinforcement Learning with Implicit Global-to-Local Value Regularization* (arXiv:2307.11620). arXiv. <http://arxiv.org/abs/2307.11620>
- Wang, Y., Xiong, J., & Zou, S. (2023). *Achieving the Asymptotically Optimal Sample Complexity of Offline Reinforcement Learning: A DRO-Based Approach* (arXiv:2305.13289). arXiv. <http://arxiv.org/abs/2305.13289>
- Wang, Y., Yang, C., Wen, Y., Liu, Y., & Qiao, Y. (2023). *Critic-Guided Decision Transformer for Offline Reinforcement Learning* (arXiv:2312.13716). arXiv. <http://arxiv.org/abs/2312.13716>
- Wei, H., Peng, X., Liu, X., & Ghosh, A. (2023). *Adversarially Trained Actor Critic for offline CMDPs* (arXiv:2401.00629). arXiv. <http://arxiv.org/abs/2401.00629>

- Wen, X., Yu, X., Yang, R., Bai, C., & Wang, Z. (2023). *Towards Robust Offline-to-Online Reinforcement Learning via Uncertainty and Smoothness* (arXiv:2309.16973). arXiv. <http://arxiv.org/abs/2309.16973>
- Wu, C., Tang, P., Yang, J., Hu, Y., Lv, T., Fan, C., & Zhang, C. (2023). *Conservative Offline Policy Adaptation in Multi-Agent Games*. NeurIPS Proceedings. [a31253f4871694f09541122d6b6f5ad1-Paper-Conference.pdf](https://arxiv.org/abs/2312.11863)
- Wu, D., Jiao, Y., Shen, L., Yang, H., & Lu, X. (2023). *Neural Network Approximation for Pessimistic Offline Reinforcement Learning* (arXiv:2312.11863). arXiv. <http://arxiv.org/abs/2312.11863>
- Wu, P., Majumdar, A., Stone, K., Lin, Y., Mordatch, I., Abbeel, P., & Rajeswaran, A. (2023). *Masked Trajectory Models for Prediction, Representation, and Control* (arXiv:2305.02968). arXiv. <http://arxiv.org/abs/2305.02968>
- Wu, Y.-H., Wang, X., & Hamaya, M. (2023). *Elastic Decision Transformer* (arXiv:2307.02484). arXiv. <http://arxiv.org/abs/2307.02484>
- Xie, Z., Lin, Z., Ye, D., Fu, Q., Yang, W., & Li, S. (2023). *Future-conditioned Unsupervised Pretraining for Decision Transformer* (arXiv:2305.16683). arXiv. <http://arxiv.org/abs/2305.16683>
- Xu, Y., & Singh, G. (2023). *Efficient Two-Phase Offline Deep Reinforcement Learning from Preference Feedback* (arXiv:2401.00330). arXiv. <http://arxiv.org/abs/2401.00330>
- Yang, Q., Wang, S., Lin, M. G., Song, S., & Huang, G. (2023). *Boosting Offline Reinforcement Learning with Action Preference Query* (arXiv:2306.03362). <https://arxiv.org/abs/2306.03362>

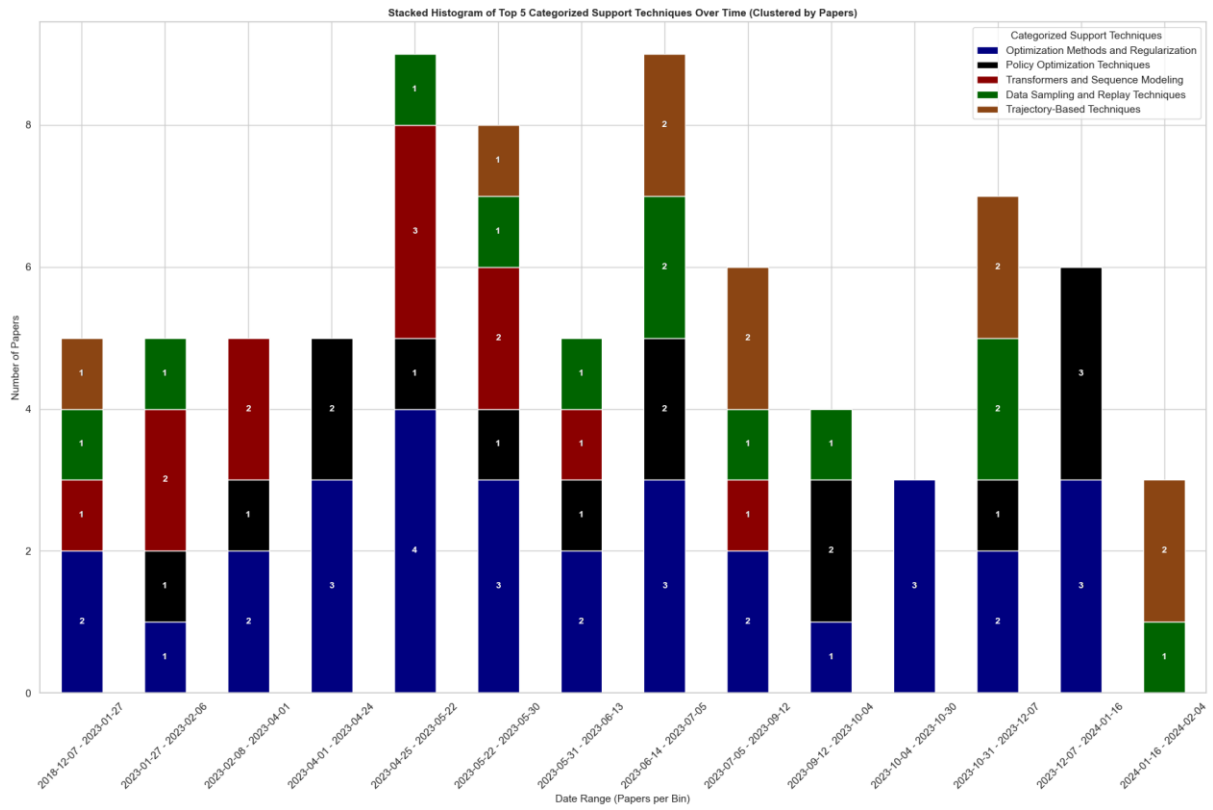
- Yang, Q., Wang, S., Zhang, Q., Huang, G., & Song, S. (2023). *Hundreds Guide Millions: Adaptive Offline Reinforcement Learning with Expert Guidance* (arXiv:2309.01448). arXiv. <http://arxiv.org/abs/2309.01448>
- Yang, R., Lin, Y., Ma, X., Hu, H., Zhang, C., & Zhang, T. (2023). *What is Essential for Unseen Goal Generalization of Offline Goal-conditioned RL?* (arXiv:2305.18882). arXiv. <http://arxiv.org/abs/2305.18882>
- Yang, R., Zhong, H., Xu, J., Zhang, A., Zhang, C., Han, L., & Zhang, T. (2024). *Towards Robust Offline Reinforcement Learning under Diverse Data Corruption* (arXiv:2310.12955). arXiv. <http://arxiv.org/abs/2310.12955>
- Ye, C., Yang, R., Gu, Q., & Zhang, T. (2024). *Corruption-Robust Offline Reinforcement Learning with General Function Approximation* (arXiv:2310.14550). arXiv. <http://arxiv.org/abs/2310.14550>
- Yue, Y., Kang, B., Ma, X., Yang, Q., Huang, G., Song, S., & Yan, S. (2024). *Decoupled Prioritized Resampling for Offline RL* (arXiv:2306.05412). arXiv. <http://arxiv.org/abs/2306.05412>
- Yue, Y., Lu, R., Kang, B., Song, S., & Huang, G. (2023). *Understanding, Predicting and Better Resolving Q-Value Divergence in Offline-RL* (arXiv:2310.04411). arXiv. <http://arxiv.org/abs/2310.04411>
- Zang, H., Li, X., Zhang, L., Liu, Y., Sun, B., & Islam, R. (2023). *Understanding and Addressing the Pitfalls of Bisimulation-based Representations in Offline Reinforcement Learning*(arXiv:2310.17139). arXiv. <http://arxiv.org/abs/2310.17139>
- Zeng, Z., Zhang, C., Wang, S., & Sun, C. (2023). *Goal-Conditioned Predictive Coding for Offline Reinforcement Learning* (arXiv:2307.03406). arXiv. <http://arxiv.org/abs/2307.03406>

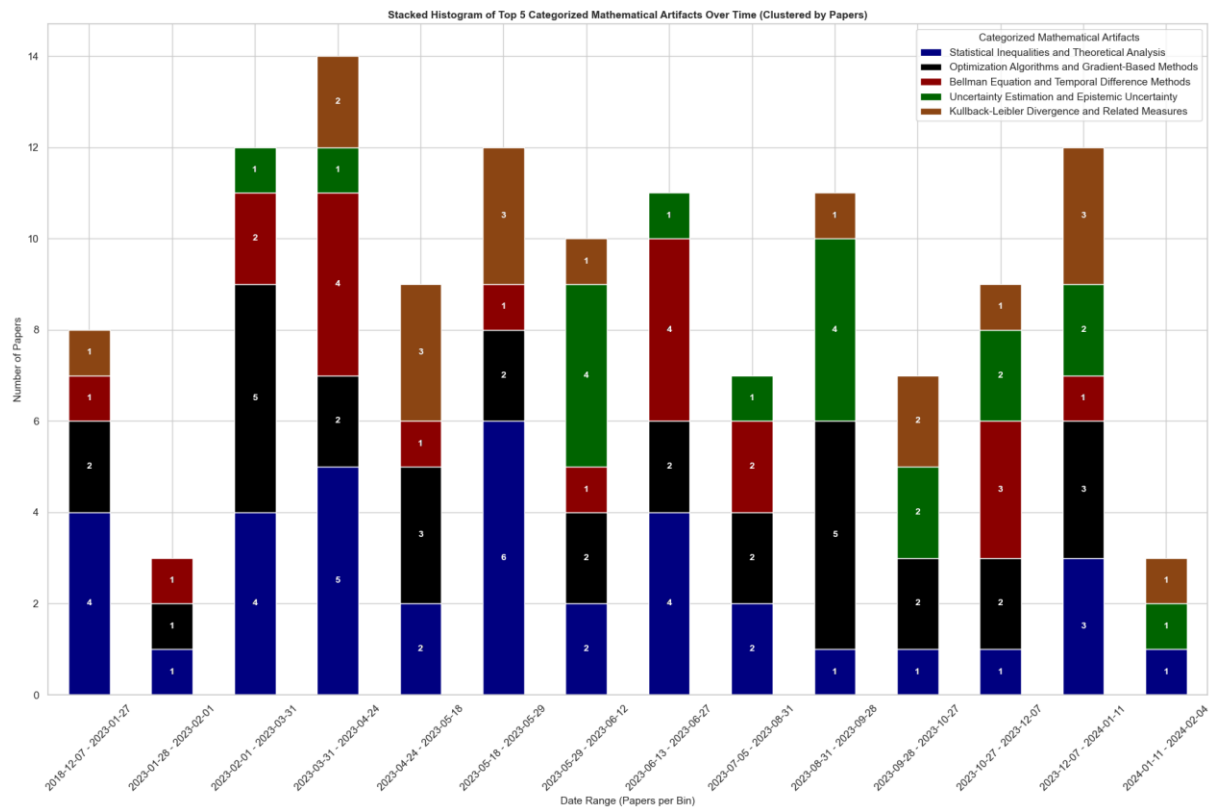
- Zhai, Y., Li, Y., Gao, Z., Gong, X., Xu, K., Feng, D., Bo, D., & Wang, H. (2024). *Optimistic Model Rollouts for Pessimistic Offline Policy Optimization* (arXiv:2401.05899). arXiv. <http://arxiv.org/abs/2401.05899>
- Zhan, W., Uehara, M., Kallus, N., Lee, J. D., & Sun, W. (2023). *Provable Offline Preference-Based Reinforcement Learning* (arXiv:2305.14816). arXiv. <http://arxiv.org/abs/2305.14816>
- Zhang, B., Mao, H., Li, L., Xu, Z., Li, D., Zhao, R., & Fan, G. (2023). *Stackelberg Decision Transformer for Asynchronous Action Coordination in Multi-Agent Systems* (arXiv:2305.07856). arXiv. <http://arxiv.org/abs/2305.07856>
- Zhang, J., Lyu, J., Ma, X., Yan, J., Yang, J., Wan, L., & Li, X. (2023). *Uncertainty-driven Trajectory Truncation for Data Augmentation in Offline Reinforcement Learning* (arXiv:2304.04660). arXiv. <http://arxiv.org/abs/2304.04660>
- Zhang, J., Zhang, C., Wang, W., & Jing, B.-Y. (2024). *Constrained Policy Optimization with Explicit Behavior Density for Offline Reinforcement Learning* (arXiv:2301.12130). arXiv. <http://arxiv.org/abs/2301.12130>
- Zhang, Q., Zhang, L., Xu, H., Shen, L., Wang, B., Chang, Y., Wang, X., Yuan, B., & Tao, D. (2023). *SaFormer: A Conditional Sequence Modeling Approach to Offline Safe Reinforcement Learning* (arXiv:2301.12203). arXiv. <http://arxiv.org/abs/2301.12203>
- Zhang, R., & Zanette, A. (2023). *Policy Finetuning in Reinforcement Learning via Design of Experiments using Offline Data* (arXiv:2307.04354). arXiv. <http://arxiv.org/abs/2307.04354>
- Zhang, X., Zhang, H., Zhou, H., Huang, C., Zhang, D., Ye, C., & Zhao, J. (2023). *Safe Reinforcement Learning with Dead-Ends Avoidance and Recovery* (arXiv:2306.13944). arXiv. <http://arxiv.org/abs/2306.13944>

- Zhang, Y., Liu, J., Li, C., Niu, Y., Yang, Y., Liu, Y., & Ouyang, W. (2023). *A Perspective of Q-value Estimation on Offline-to-Online Reinforcement Learning* (arXiv:2312.07685). arXiv. <http://arxiv.org/abs/2312.07685>
- Zhang, Z., Xiong, X., Zhuang, Z., Liu, J., & Wang, D. (2024). *Improving Offline-to-Online Reinforcement Learning with Q Conditioned State Entropy Exploration* (arXiv:2310.19805). arXiv. <http://arxiv.org/abs/2310.19805>
- Zhang, Z., Xu, J., Liu, J., Zhuang, Z., Wang, D., Liu, M., & Zhang, S. (2024). *Context-Former: Stitching via Latent Conditioned Sequence Modeling* (arXiv:2401.16452). arXiv. <http://arxiv.org/abs/2401.16452>
- Zhao, C., Zhou, Z., & Liu, B. (2023). *On Context Distribution Shift in Task Representation Learning for Offline Meta RL* (arXiv:2304.00354). arXiv. <http://arxiv.org/abs/2304.00354>
- Zhao, K., Ma, Y., Hao, J., Liu, J., Zheng, Y., & Meng, Z. (2023). *Improving Offline-to-Online Reinforcement Learning with Q-Ensembles* (arXiv:2306.06871). arXiv. <http://arxiv.org/abs/2306.06871>
- Zheng, Y., Li, J., Yu, D., Yang, Y., Li, S. E., Zhan, X., & Liu, J. (2024). *Safe Offline Reinforcement Learning with Feasibility-Guided Diffusion Model* (arXiv:2401.10700). arXiv. <http://arxiv.org/abs/2401.10700>
- Zhou, B., Li, K., Jiang, J., & Lu, Z. (2023). *Learning from Visual Observation via Offline Pretrained State-to-Go Transformer* (arXiv:2306.12860). arXiv. <http://arxiv.org/abs/2306.12860>
- Zhou, W. (2023). *Bi-Level Offline Policy Optimization with Limited Exploration* (arXiv:2310.06268). arXiv. <http://arxiv.org/abs/2310.06268>

- Zhou, W., & Qu, A. (2023). *Stackelberg Batch Policy Learning* (arXiv:2309.16188). arXiv. <http://arxiv.org/abs/2309.16188>
- Zhou, Y., Sekhari, A., Song, Y., & Sun, W. (2023). *Offline Data Enhanced On-Policy Policy Gradient with Provable Guarantees* (arXiv:2311.08384). arXiv. <http://arxiv.org/abs/2311.08384>
- Zhu, D., Wang, Y., Schmidhuber, J., & Elhoseiny, M. (2023). *Guiding Online Reinforcement Learning with Action-Free Offline Pretraining* (arXiv:2301.12876). arXiv. <http://arxiv.org/abs/2301.12876>
- Zhu, H., Rashidinejad, P., & Jiao, J. (2023). *Importance Weighted Actor-Critic for Optimal Conservative Offline Reinforcement Learning* (arXiv:2301.12714). arXiv. <http://arxiv.org/abs/2301.12714>
- Zhu, Z., Liu, M., Mao, L., Kang, B., Xu, M., Yu, Y., Ermon, S., & Zhang, W. (2024). *MADiff: Offline Multi-agent Learning with Diffusion Models* (arXiv:2305.17330). arXiv. <http://arxiv.org/abs/2305.17330>

APPENDIX A

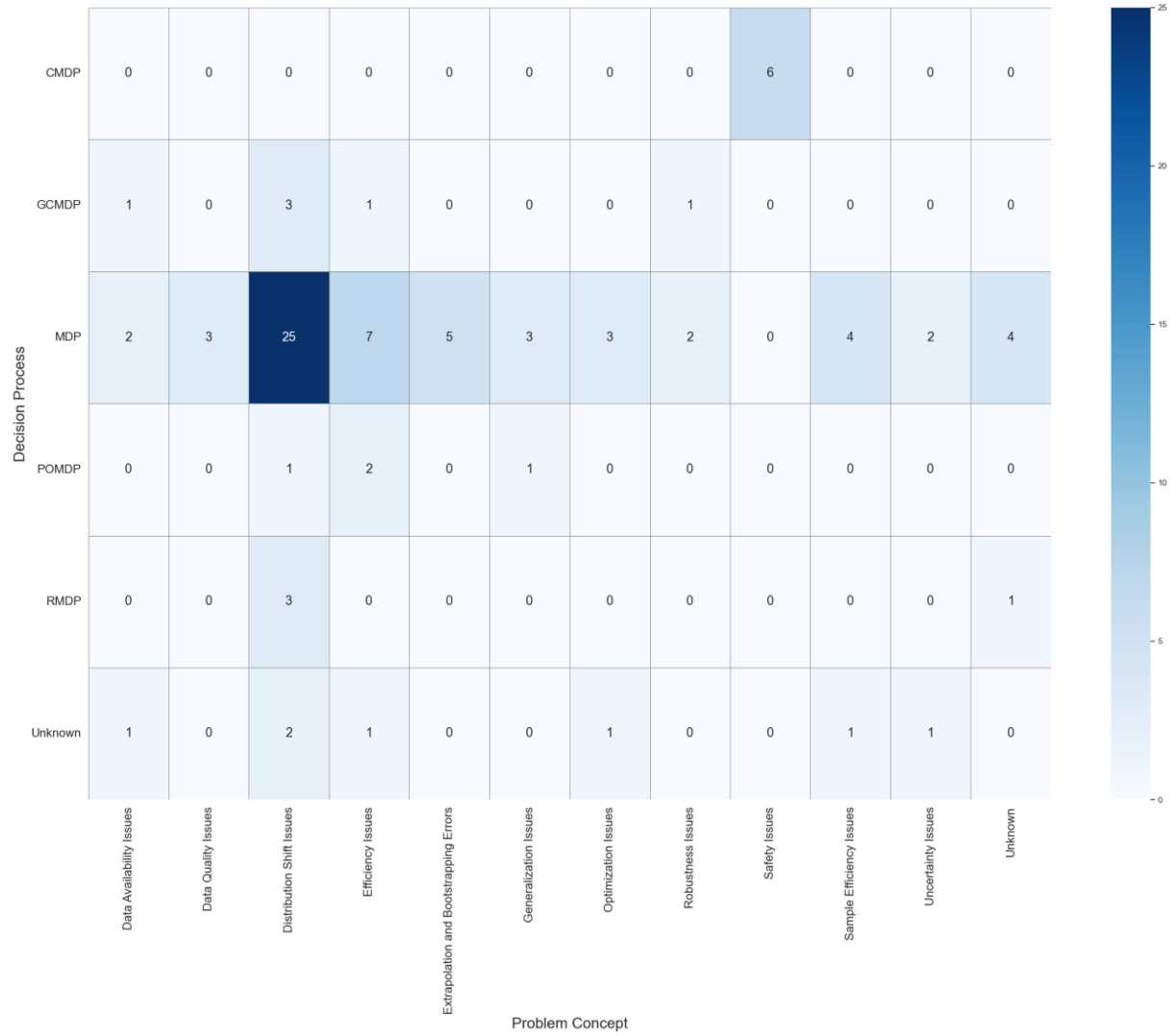




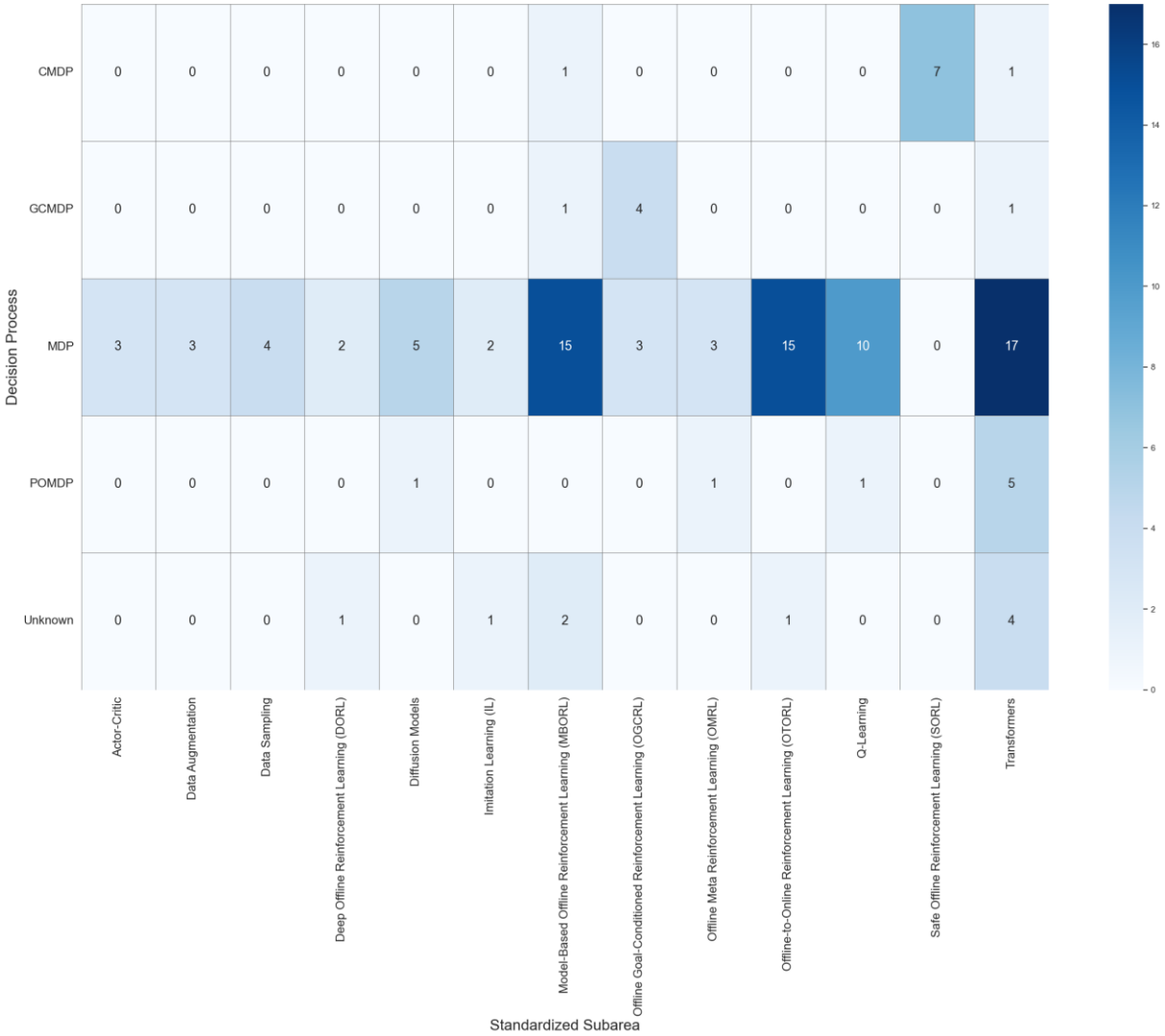
Contingency Table: Decision Process vs Categorized Mathematical Artifacts

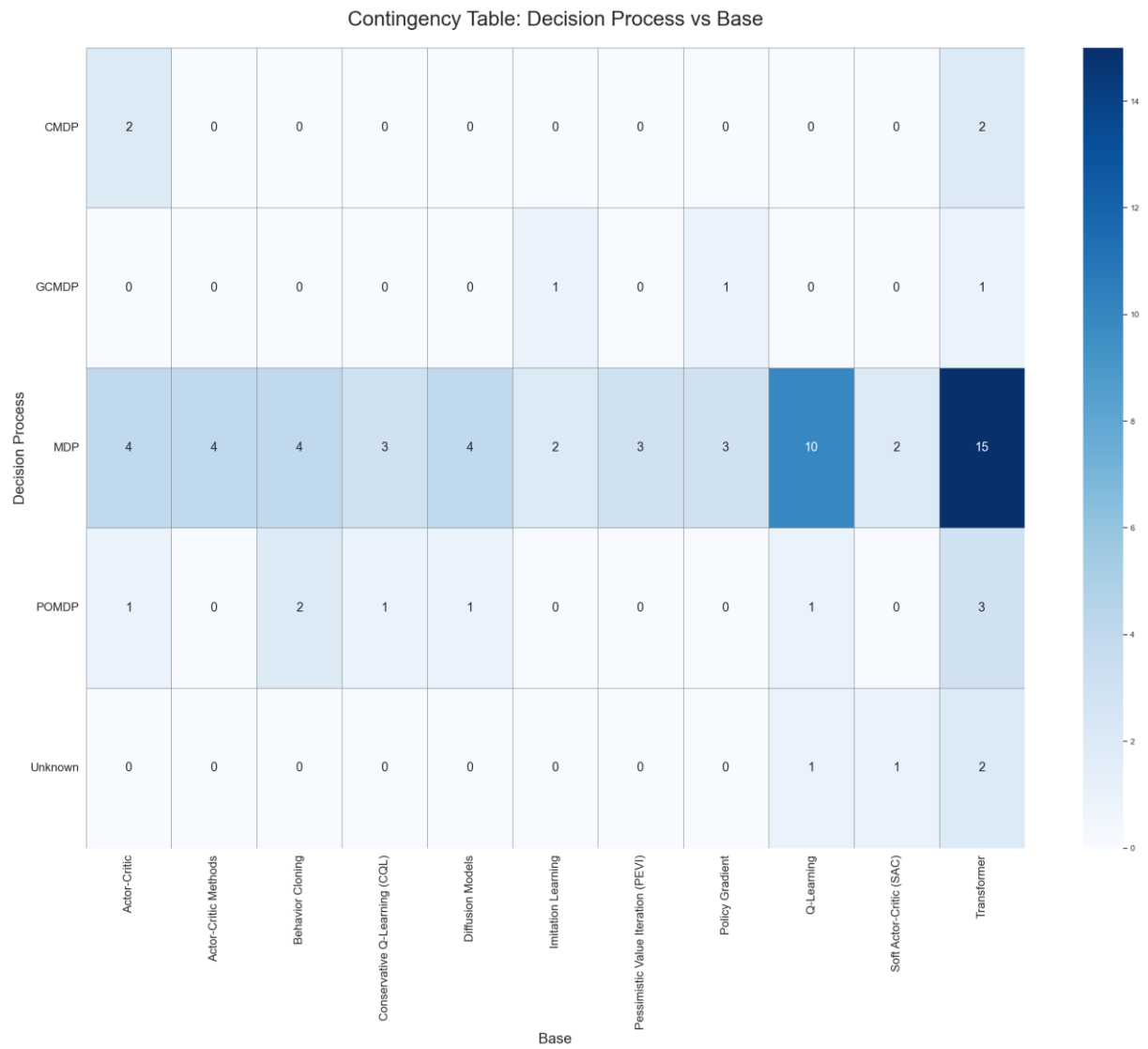
Decision Process	Categorized Mathematical Artifacts																										Value	
	Bellman Equation and Temporal Difference Methods	Bellman Residual and Error Minimization	Diffusion Models and Probabilistic Processes	Ensemble Methods	Fine-Tuning and Transfer Learning	Game Theory Concepts	Generative Models	Imitation Learning and Behavior Cloning	Kullback-Leibler Divergence and Related Measures	Lagrangian Methods and Convex Optimization	Multi-Agent Reinforcement Learning	Mutual Information and Contrastive Learning	Optimization Algorithms and Gradient-Based Methods	Others	Others (Quantum Computing)	Policy Evaluation and Improvement Methods	Policy Gradient Methods	Primal-Dual Methods and Optimization	Regularization Methods	Reinforcement Learning Algorithms	Sampling Techniques and Data Augmentation	Sequence Modeling and Latent Representation Learning	Statistical Inequalities and Theoretical Analysis	Transformers and Attention Mechanisms	Uncertainty Estimation and Epistemic Uncertainty	Variational Autoencoders (VAE) and Related Techniques		
multi-objective Markov decision process (MOMDP)	BAMDP	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
	CMDP	3	0	0	0	0	1	0	0	2	5	0	0	1	0	0	0	0	0	1	0	0	2	1	1	0	0	
	Conservative MDP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
	GCMDP	1	0	0	0	0	0	0	0	1	1	0	1	0	0	0	1	0	0	0	1	0	1	2	0	0	0	
	LMDP	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	
	MCP	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	MDP	15	2	6	3	2	0	2	1	11	6	0	8	29	10	1	3	3	1	3	10	11	9	25	8	15	9	1
	MRP	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
	Markov Game	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
	Non-MDP	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
	Optimistic MDP	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	POMDP	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	1	0	0	2	1	2	1	2	0	2	0
	Pessimistic MDP	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0
	RMDP	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	4	0	2	0	0
	Unknown	0	0	1	0	1	0	0	0	1	1	0	0	1	1	0	1	0	0	0	1	0	3	3	2	1	0	0
	incomplete MDP	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0
partially observable Markov game (POMG)	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	

Contingency Table: Decision Process vs Problem Concept

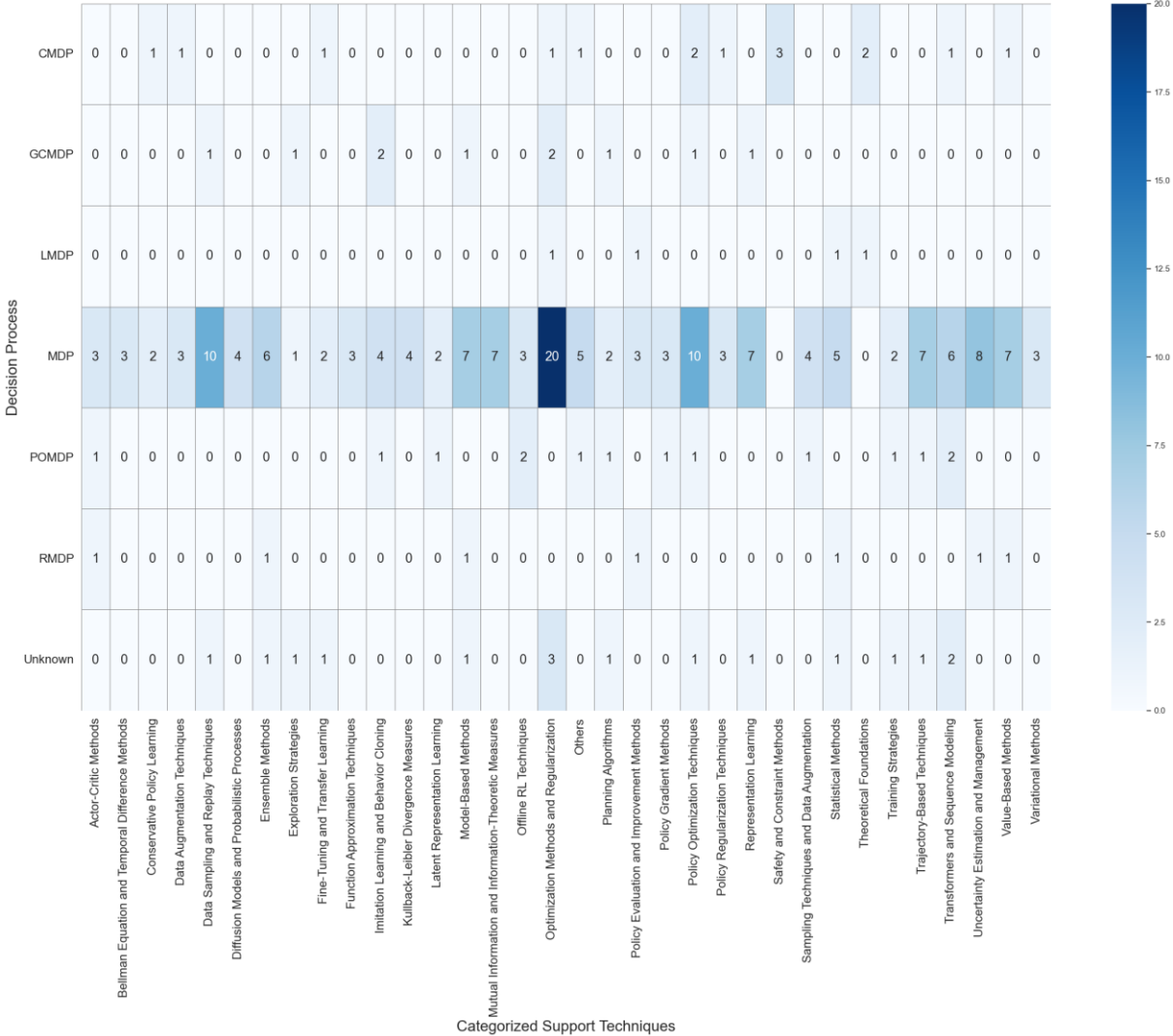


Contingency Table: Decision Process vs Standardized Subarea





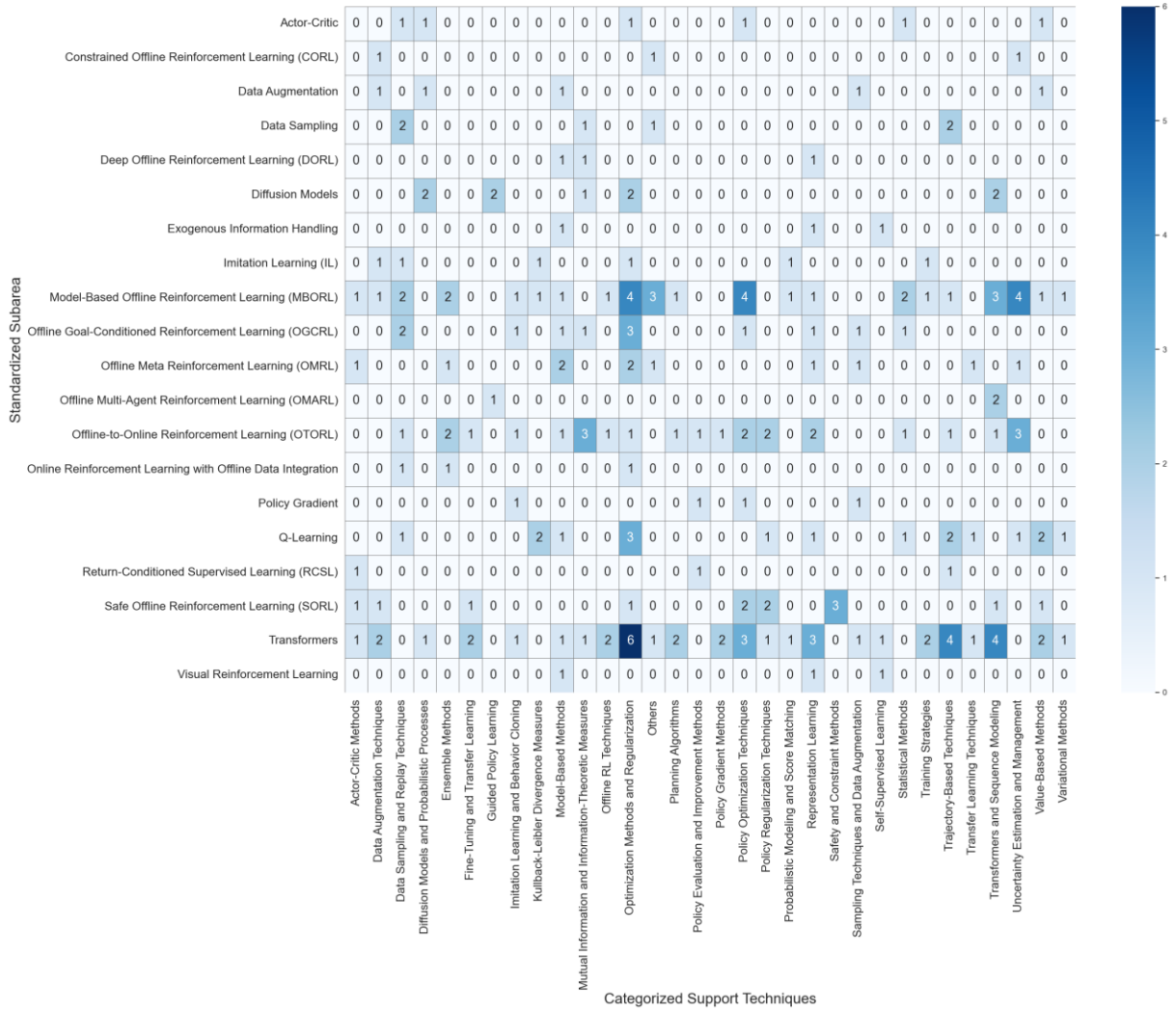
Contingency Table: Decision Process vs Categorized Support Techniques



Contingency Table: Decision Process vs Categorized Mathematical Artifacts

Decision Process	CMDP	3	0	0	0	2	5	0	1	0	0	0	0	1	0	0	2	1	1	0
	GCMDP	1	0	0	0	1	1	1	0	0	1	0	0	1	0	1	2	0	0	0
	LMDP	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0
	MDP	15	6	3	2	11	6	8	29	10	3	3	3	10	11	9	25	8	15	9
	Optimistic MDP	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
	POMDP	0	0	0	0	1	1	0	0	0	0	1	0	2	1	2	1	2	0	2
	Pessimistic MDP	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1
	RMDP	2	0	0	0	1	0	0	0	0	0	0	0	1	0	0	4	0	2	0
	Unknown	0	1	0	1	1	1	0	1	1	1	0	0	1	0	3	3	2	1	0
	incomplete MDP	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0
	multi-objective Markov decision process (MOMDP)	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	1
		Bellman Equation and Temporal Difference Methods	Diffusion Models and Probabilistic Processes	Ensemble Methods	Fine-Tuning and Transfer Learning	Kullback-Leibler Divergence and Related Measures	Lagrangian Methods and Convex Optimization	Mutual Information and Contrastive Learning	Optimization Algorithms and Gradient-Based Methods	Others	Policy Evaluation and Improvement Methods	Policy Gradient Methods	Regularization Methods	Reinforcement Learning Algorithms	Sampling Techniques and Data Augmentation	Sequence Modeling and Latent Representation Learning	Statistical Inequalities and Theoretical Analysis	Transformers and Attention Mechanisms	Uncertainty Estimation and Epistemic Uncertainty	Variational Autoencoders (VAE) and Related Techniques
		Categorized Mathematical Artifacts																		

Contingency Table: Standardized Subarea vs Categorized Support Techniques



Contingency Table: Problem Concept vs Categorized Support Techniques

Problem Concept	Categorized Support Techniques																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																														
	Actor-Critic Methods				Data Sampling and Replay Techniques		Ensemble Methods		Fine-Tuning and Transfer Learning		Imitation Learning and Behavior Cloning		Kullback-Leibler Divergence Measures		Model-Based Methods		Multi-Objective Optimization		Mutual Information and Information-Theoretic Measures		Offline RL Techniques		Optimization Methods and Regularization		Others		Planning Algorithms		Policy Evaluation and Improvement Methods		Policy Gradient Methods		Policy Optimization Techniques		Policy Regularization Techniques		Representation Learning		Safety and Constraint Methods		Sampling Techniques and Data Augmentation		Statistical Methods		Trajectory-Based Techniques		Transformers and Sequence Modeling		Uncertainty Estimation and Management		Value-Based Methods		Variational Methods																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																																										
Benchmarking Issues	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0</



Contingency Table: Base vs Categorized Support Techniques



Contingency Table: Categorized Support Techniques vs Categorized Mathematical Artifacts

Categorized Support Techniques	Categorized Mathematical Artifacts																																
	Bellman Equation and Temporal Difference Methods										Bayesian Methods										Diffusion Models and Probabilistic Processes												
	Ensemble Methods										Fine-Tuning and Transfer Learning										Game Theory Concepts												
	Generative Models										Imitation Learning and Behavior Cloning										Kullback-Leibler Divergence and Related Measures												
Actor-Critic Methods	1	0	1	0	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	0	0	1	1	0	0	2	0	3	1	1			
Bayesian Methods	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	1	1	1			
Bellman Equation and Temporal Difference Methods	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0	0			
Causal Inference Techniques	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0			
Conservative Policy Learning	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	0			
Contrastive Learning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
Data Augmentation Techniques	2	0	1	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0		
Data Sampling and Replay Techniques	2	0	0	0	0	0	0	0	0	3	1	1	1	0	2	1	0	1	1	3	0	0	5	0	2	0	0	0	0	0	0		
Diffusion Models and Probabilistic Processes	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0		
Energy-Based Models	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1		
Ensemble Methods	1	0	0	1	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	2	0	0	2	0	3	1	0	0	0	1		
Error Correction Techniques	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0		
Evaluation and Analysis Techniques	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
Exploration Strategies	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0		
Federated Learning Techniques	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
Fine-Tuning and Transfer Learning	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0		
Function Approximation Techniques	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0		
Generative Models	0	0	0	0	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	
Guided Policy Learning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	
Human-in-the-Loop Learning	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
Imitation Learning and Behavior Cloning	1	0	0	0	1	0	0	0	1	1	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
Intrinsic Motivation Techniques	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Kullback-Leibler Divergence Measures	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Model-Based Methods	2	0	1	0	0	0	0	0	0	0	0	0	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
Multi-Objective Optimization	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Mutual Information and Information-Theoretic Measures	0	0	1	0	0	0	0	0	0	0	3	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Offline RL Techniques	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Optimization Methods and Regularization	2	0	3	1	1	2	0	0	0	6	5	2	12	2	2	1	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0	1
Others	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Planning Algorithms	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Policy Evaluation and Improvement Methods	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Policy Gradient Methods	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Policy Optimization Techniques	5	0	0	0	0	0	0	0	0	0	3	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Policy Regularization Techniques	1	0	0	0	0	0	0	0	0	0	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Reinforcement Learning Algorithms	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Representation Learning	2	0	2	0	0	0	0	0	0	1	0	1	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Safety and Constraint Methods	0	0	0	0	0	0	0	0	0	1	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Sampling Techniques and Data Augmentation	1	0	0	0	0	0	0	0	0	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Self-Supervised Learning	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Statistical Methods	2	0	0	0	0	0	0	0	0	0	2	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Supervised Learning Techniques	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Theoretical Foundations	0	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Training Strategies	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Trajectory-Based Techniques	3	0	0	1	0	0	0	0	0	1	1	0	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Transfer Learning Techniques	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Transformers and Sequence Modeling	1	0	1	0	0	0	0	0	0	0	1	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Uncertainty Estimation and Management	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Value-Based Methods	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Variational Methods	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Bellman Equation and Temporal Difference Methods																																	
Bellman Residual and Error Minimization																																	
Diffusion Models and Probabilistic Processes																																	
Ensemble Methods																																	
Fine-Tuning and Transfer Learning																																	
Game Theory Concepts																																	
Generative Models																																	
Imitation Learning and Behavior Cloning																																	
Kullback-Leibler Divergence and Related Measures																																	
Lagrangian Methods and Convex Optimization																																	
Mutual Information and Contrastive Learning																																	
Optimization Algorithms and Gradient-Based Methods																																	
Others																																	
Policy Evaluation and Improvement Methods																																	
Policy Gradient Methods																																	
Primal-Dual Methods and Optimization																												</					

