TU WIEN Informatics

Applications of Concentration Inequalities in Distributional Reinforcement Learning

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Data Science

by

Florian Mayer, BSc
Registration Number 01525689

to the Faculty of Informatics

at the TU Wien

Advisor: Associate Prof. Dr.techn. Dipl.-Ing. Clemens Heitzinger

Vienna, December 2, 2024

Florian Mayer	Clemens Heitzinger

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Florian Mayer, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. Dezember 2024

Florian Mayer

Abstract

Distributional reinforcement learning extends traditional reinforcement learning by modeling the entire distribution of returns, providing several advantages, such as insight into potential outcomes and associated risks. However, this approach results in higher computational complexity.

This thesis investigates the application of different concentration inequalities, specifically the Hoeffding, Bernstein, and Bennett inequalities to find tighter bounds on the Cramér distance between the estimated reward distributions and the true reward distribution. Tighter bounds enhance the analysis of algorithms, such as the speedy Q-learning algorithm within the distributional reinforcement learning framework.

To validate the theoretical findings, a complexity analysis is conducted to determine which inequality provides the most robust and reliable bounds under varying accuracy requirements and environmental complexities.

In addition to that, simulation studies are performed using the Taxi and FrozenLake environments from the Gymnasium library in Python. These simulations compare the performance of each inequality and observe their impact on the convergence behavior of the learning algorithms.

The tightest bound on the Cramér distance is achieved using Bennett's inequality, followed by the bound obtained through the Bernstein inequality. However, when the number of training episodes is small, the bound derived from the Hoeffding inequality exceeds the Bernstein bound in terms of tightness.

Contents

CHAPTER 1

Introduction

1.1 Motivation and Problem Statement

In reinforcement learning, an agent learns an optimal policy by observing the outcome of actions in terms of the expected returns.

Distributional reinforcement learning provides more information on the environment's dynamics and the risks associated with different actions by modeling the entire distribution of returns. This improves decision-making under uncertainty. However, the computational cost of distributional reinforcement learning is significantly higher than that of traditional reinforcement learning.

As algorithms become more complex, understanding their computational needs and how efficiently they operate becomes very important, especially if computing power is limited, or decisions need to be made quickly and reliably. In addition to that, a complexity analysis of an algorithm before employment can help to evaluate the number of training episodes required to achieve a desired accuracy.

1.2 Aim of the Thesis

The main goal of this thesis is to investigate the use of different concentration inequalities to bound the Cramér distance between the actual and estimated reward distribution. This is essential for evaluating the effectiveness of an algorithm.

The thesis will explore several specific concentration inequalities, such as the Hoeffding, Bernstein and Bennett's inequalities. In the complexity analysis we will evaluate which inequalities provide the most reliable and stable bounds under varying degrees of precision and environmental complexities.

Moreover, an experimental evaluation will be conducted to verify the theoretical results and determine the most appropriate concentration inequality for bounding the Cramér distance.

The thesis aims to deepen the understanding of evaluating speedy Q-learning algorithms in distributional reinforcement learning, which is especially important in scenarios where decisions must be made with a high degree of reliability and precision, such as in autonomous driving, financial forecasting, or complex control systems.

1.3 Related Work

The concept of distributional reinforcement learning was initially introduced by Bellemare et al. (2017) in "*A Distributional Perspective on Reinforcement Learning*" [BDM17].

Q-learning, was developed by Christopher J.C.H. Watkins in 1989 in his PhD thesis "*Learning from Delayed Rewards*" [Wat89], which he further detailed in [WD92]. Moreover, Ghavamzadeh et al. proposed speedy Q-learning in 2011 [AMGK11] as an enhancement to the original Q-learning algorithm, intended to exalerate convergence.

The concentration inequalities referenced in this thesis are:

Hoeffding inequality: Introduced by Wassily Hoeffding in his work "*Probability Inequalities for Sums of Bounded Random Variables*" in 1963 [Hoe94].

Bernstein inequality: Introduced by Sergei N. Bernstein in 1946 [Ber46].

Bennett's inequality: Introduced by George Bennett in his work "*Probability Inequalities for the Sum of Independent Random Variables*" in 1962 [Ben62].

Rowland et al. (2018)[RBD$^+$18a] presented the categorical method as a practical application of the distributional reinforcement learning concept and demonstrated its convergence.

The paper "*Speedy Categorical Distributional Reinforcement Learning and Complexity Analysis*" by Markus Böck and Clemens Heitzinger [BH22] provided an complexity analysis of speedy categorical distributional reinforcement learning utilizing the Hoeffding inequality.

1.4 Structure of the Thesis

Firstly we discuss the important concepts and terms regarding distributional reinforcement learning and Q-learning, especially speedy Q-learning in Chapter 2.

Following that, in Chapter 3, we will introduce bounds on the Cramér distance for the Hoeffding, Bernstein, and Bennett concentration inequalities. Afterwards, in Chapter 4, we numerically analyze those bounds, which includes evaluating the number of training episodes needed to achieve the desired accuracy under varying accuracy requirements and different levels of environmental complexity.

In Chapter 5, we validate our findings by performing an experimental evaluation using the Gymnasium environments in Python, specifically utilizing the Taxi and FrozenLake environments to investigate the convergence properties of the different bounds on the Cramér distance.

We conclude the thesis in Chapter 6 with a summary and discussion of our findings.

CHAPTER 2

Preliminaries

In this section, we present the idea of distributional reinforcement learning, particularly categorical reinforcement learning and its integration with Q-learning.

We start by explaining the standard reinforcement learning framework as described in Sutton and Barto (2018) [SB18].

2.1 Fundamentals of Reinforcement Learning

The objective of reinforcement learning is to learn a policy that maximizes the cumulative reward an agent receives over time through interacting with an environment. The agent chooses actions based on the current policy, receives rewards or penalties based on those actions, and updates the policy accordingly to improve future rewards.

Definition 2.1.1. In RL, we define *policy* π as a probability distribution over actions, which denotes the probability of choosing action a when in state x. $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$, where $\sum_{a \in \mathcal{A}} \pi(a|x) = 1$. We denote the *optimal policy* that maximizes the cumulative reward as π^*.

This framework can be formally expressed using a Markov Decision Process.

2.1.1 Markov Decision Process (MDP)

We only consider a finite state space $\mathcal{X}$ and a finite action space $\mathcal{A}$.

At each discrete time step $t = 0, 1, 2, 3, ...$, the agent observes the current state of the environment, X_t, from the set of possible states $\mathcal{X}$. Based on this information and the policy, the agent selects an action, A_t, from the set of possible actions $\mathcal{A}$. After that, the agent receives a reward R_{t+1} and transitions to a new state S_{t+1}. This sequence of states, actions, and rewards is a trajectory of the form $(X_0, A_0, R_1, X_1, A_1, R_2, X_2, A_2, ...)$.

Definition 2.1.2. The finite Markov decision process is characterized by the tuple $(\mathcal{X}, \mathcal{A}, r, p)$, where the set of states and actions are finite $|\mathcal{X}| < \infty$, $|\mathcal{A}| < \infty$. The probability of possible values for X_t and R_t satisfies the Markov property and depends only on the state and action immediately preceding X_{t-1} and A_{t-1}, i.e.,

$$\mathbb{P}\left[R_t = s, X' = x' | X_t = x, A_t = a, X_{t-1} = x_{t-1}, ...\right] = r(s|x, a, x')p(x'|x, a),$$

where $p(\cdot|x, a)$ is the state transition probability and the kernel $r(\cdot|x, a, x')$ represents the immediate reward when transitioning from state x to state x' with action a.

2.1.2 Bellman equation

The Bellman equation as defined by [Bel66] as

$$\begin{aligned} Q(x, a)^\pi &= \mathbb{E}[R_{t+1} + \gamma Q^\pi(x', a') \,|\, X_t = x, A_t = a], \\ x' \sim P(\cdot|x, a), \quad a' \sim \pi(\cdot|x'), \quad R \sim r(\cdot|x, a, x') \end{aligned} \tag{2.1}$$

is essential in reinforcement learning as it provides a recursive formula for calculating value functions. These functions estimate the expected return from a given state or state-action pair under a specific policy. It simplifies the optimization of long-term rewards into manageable, iterative updates, making it foundational for algorithms such as Q-learning.

Definition 2.1.3. The state-action value function $Q^\pi(x, a)$ defined as

$$Q(x, a)^\pi \coloneqq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \,|\, x_0 = x, a_0 = a\right] \tag{2.2}$$

is the expected return of taking action $a \in \mathcal{A}$ in state $x \in \mathcal{X}$, then following the policy π. The optimal state-value function Q^* is defined as

$$Q^*(x, a) \coloneqq Q^{\pi^*}(x, a) = \sup_\pi Q^\pi(x, a),$$

where π^* is an optimal policy.

Definition 2.1.4. The *Bellman operator* $\mathcal{T}^\pi$ and the *Bellman optimality operator* $\mathcal{T}$ are defined as

$$\begin{aligned} \mathcal{T}^\pi Q(x, a) &= \mathbb{E}[R(x, a) + \gamma Q(x', a'), \\ \mathcal{T} Q(x, a) &= \mathbb{E}\left[R(x, a) + \gamma \max_{a' \in \mathcal{A}} Q(x', a')\right], \\ X' \sim p(\cdot|x, a), \quad A' \sim \pi(\cdot|X'), \end{aligned} \tag{2.3}$$

where $x' \sim p(\cdot|x, a)$ and $a' \sim \pi(\cdot|A')$. These operators both fulfill the following properties:

- Q^π is a fixed point of $\mathcal{T}^\pi$ and Q^* is a fixed point of $\mathcal{T}$. That means that $Q^\pi = \mathcal{T}^\pi Q^\pi$ and $Q^* = \mathcal{T} Q^*$ hold.

- Both $\mathcal{T}^\pi$ and $\mathcal{T}$ are γ-contraction in the supremum norm $\|.\|_\infty$, e.g. $\|\mathcal{T}^\pi Q_1 - \mathcal{T}^\pi Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$.

This operator, which is a contraction mapping, ensures convergence to the optimal value function through its repeated application as shown in [Tsi94].

2.2 Distributional Reinforcement Learning

Distributional reinforcement learning deviates from conventional reinforcement learning by modeling the full distribution of possible returns.

The return distribution function is defined by [BDM17] as follows.

Definition 2.2.1. The return at $(x, a) \in \mathcal{X} \times \mathcal{A}$ is the sum of discounted rewards along the trajectory of the agents in interactions with the environment, when following the policy π. The return distribution function Z^π is the mapping of state-action pairs to random variables, i.e.,

$$\begin{aligned} Z^\pi(x,a) &= \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t), \\ x_t \sim P(\cdot|x_{t-1}, a_{t-1}), \quad a_t &\sim \pi(\cdot|x_t), \end{aligned} \tag{2.4}$$

where the parameter $x = X_0$ and $a = A_0$ are the starting point of the discounted sum of the cumulative reward and $\mathcal{Z}$ denotes the set of all return distribution functions.

In distributional reinforcement learning, the value function Q^π (see definition 2.1.3) can be interpreted as the expected value of the return distribution function Z^π, i.e.,

$$Q^\pi(x,a) = \mathbb{E} Z^\pi(x,a). \tag{2.5}$$

The Bellman equation, as specified in the subsection 2.1, can be extended to the distributional case.

$$\begin{aligned} Z^\pi(x,a) &\overset{D}{=} R + \gamma Z^\pi(x', a'), \\ x' \sim P(\cdot|x,a), \quad a' \sim \pi(\cdot|x'), \quad R &\sim r(\cdot|x,a,x') \end{aligned} \tag{2.6}$$

where the equality sign $\overset{D}{=}$ indicates that the random variables are equally distributed.

Similarly, the Bellman operators can also be extended to the distributional Bellman operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ and the distributional Bellman optimality operator $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$.

$$\begin{aligned} \mathcal{T}^\pi Z(x,a) &= R(x,a) + \gamma Z(X', A'), \quad X' \sim p(\cdot|x,a), \quad A' \sim \pi(\cdot|X') \\ \mathcal{T} Z &= R(x,a) + \gamma Z(X', A^*), \quad X' \sim p(\cdot|x,a), \quad A^* = \arg\max_{a \in \mathcal{A}} \mathbb{E}[Z(X', a)]. \end{aligned} \tag{2.7}$$

The underlying probability distribution of the random variable $Z^\pi(x,a)$

$$\begin{aligned} \eta_\pi^{(x,a)}((-\infty,z]) &= P[Z^\pi(x,a) \leq z], \\ Z^\pi(x,a) &\sim \eta_\pi^{(x,a)} \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A} \end{aligned} \tag{2.8}$$

satisfies the distributional variant of the Bellman equation as shown in [BDM17]. From this statement, it follows that the equation

$$\eta_\pi^{(x,a)} = (\mathcal{T}^\pi \eta_\pi)^{(x,a)} \tag{2.9}$$

is satisfied for $\forall (x,a) \in \mathcal{X} \times \mathcal{A}$, where $\mathcal{T}^\pi : \mathcal{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}} \to \mathcal{P}(\mathbb{R})^{\mathcal{X}\times\mathcal{A}}$ is the distributional Bellman operator that can be written in terms of cumulative distribution functions as

$$F_{\mathcal{T}^\pi \eta_\pi^{(x,a)}}(z) = \mathbb{E}\left[F_{\eta^{(X',A')}}\left(\frac{z-R}{\gamma}\right)\right]. \tag{2.10}$$

2.2.1 Categorical Distributional Reinforcement Learning

A fundamental challenge in distributional reinforcement learning is accurately approximating the return distributions. Approximating these distributions is complex because it is not feasible to represent the entire space of probability distributions $\mathbb{P}(\mathbb{R})$ by a finite collection of parameters.

We use the categorical approach presented in [RBD+18a], where we utilize the parametric family of categorical distributions $\mathcal{P} \subset \mathbb{P}(\mathbb{R})$ with set bounds for the return $V_{\min}$, $V_{\max}$ over a fixed set of N equally spaced supports $z_1 < \cdots < z_N$, $\Delta z = (V_{\max} - V_{\min})/(N-1)$, i.e.,

$$\mathcal{P}_z = \left\{ \sum_{i=1}^N p_i \delta_{z_i} \ : \ p_i \geq 0, \sum_{i=1}^N p_i = 1 \right\} \tag{2.11}$$

where δ_{z_i} is the Dirac measure at z_i.

The Bellman operator modifies the return distribution by scaling it with γ and adding the reward, causing categorical distributions to lose stability in this operation. As a result, it is necessary to define the categorical projection operator $\Pi_C : \mathbb{P}(\mathbb{R}) \to \mathcal{P}$ to project the distribution back on the fixed support.

The operator is defined by [BDM17] as follows.

Definition 2.2.2. The categorical projection operator $\Pi_C : \mathbb{P}(\mathbb{R}) \to \mathcal{P}$ is defined by

$$\begin{aligned} \Pi_C(\delta_y) &:= \begin{cases} \delta_{z_1}, & y \leq z_1, \\ \frac{z_{i+1}-y_j}{z_{i+1}-z_i}\delta_{z_i} + \frac{y_j - z_i}{z_{i+1}-z_i}\delta_{z_{i+1}}, & z_i \leq y_j \leq z_{i+1}, \\ \delta_{z_N}, & y \geq z_N, \end{cases} \\ \Pi_C\left(\sum_{i=1}^N p_i \delta_{y_i}\right) &= \sum_{i=1}^N p_i \Pi_C(\delta_{y_i}). \end{aligned} \tag{2.12}$$

The operator redistributes the probability of a point across the two adjacent fixed atoms.

The Cramér distance defined in [RBD+18b] for this operator is used to quantify the difference between two probability distributions. It is based on the squared differences of their cumulative distribution functions (CDFs).

Definition 2.2.3. The Cramér distance l_2 between two distributions $v_1, v_2 \in \mathbb{P}[\mathbb{R}]$, with cumulative distribution functions F_{v_1}, F_{v_2}, is defined as

$$l_2(v_1, v_2) \coloneqq \left(\int_{\mathbb{R}} (F_{v_1}(x) - F_{v_2}(x))^2 dx \right)^{1/2}. \tag{2.13}$$

Moreover the supremum-Cramér metric $\bar{l}_2$ between two distribution functions $\eta, \mu \in \mathbb{P}[\mathbb{R}]^{\mathcal{X}\times\mathcal{A}}$ is calculated as

$$\begin{aligned} \bar{l}_2(\eta, \mu) &= \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} l_2(\eta^{(x,a)}, \mu^{(x,a)}) \\ &= \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \left(\int_{\mathbb{R}} (F_{\eta^{(x,a)}}(x) - F_{\mu^{(x,a)}}(x))^2 dx \right)^{1/2}. \end{aligned} \tag{2.14}$$

The operator $\Pi_C \mathcal{T}^\pi$, which is the composition of the distributional Bellman operator $\mathcal{T}^\pi$ 2.1.4 and the projection operator Π_C 2.2.2 is a $\sqrt{\gamma}$-contraction in $\bar{l}_2$.

In addition, there is a unique distribution function $\eta_C \in \mathcal{P}^{\mathcal{X}\times\mathcal{A}}$, such that for any $\eta_0 \in \mathbb{P}(\mathbb{R}^{\mathcal{X}\times\mathcal{A}})$, the convergence $(\Pi_C \mathcal{T}^\pi)^m \eta_0 \to \eta_C$ holds for $\bar{l}_2$ as $m \to \infty$ converges.

However, because we can only approximate distributions, $\eta_C \neq \eta_\pi$.

Increasing the number of atoms $[z_1, z_N]$ that support the return distribution $\eta_\pi^{(x,a)}$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ improves the precision of the approximation, i.e.,

$$\bar{l}_2(\eta_C, \eta_\pi) \leq \frac{1}{1-\gamma} \max_{1\leq i\leq N} (z_{i+1} - z_i). \tag{2.15}$$

2.3 Q-Learning

Q-learning, first introduced by [WD92], is a fundamental reinforcement learning algorithm. It iteratively updates the Q-values, which estimates the expected rewards for actions taken in specific states, allowing the agent to make decisions that maximize future rewards. Notably, Q-learning operates without requiring a model of the environment, enhancing its applicability across diverse and complex scenarios.

The update rule in Q-learning is defined as follows.

Definition 2.3.1. After observing the current state x in episode k, performing an action a that leads to observing the subsequent state x' and receiving the immediate reward r', the $Q(x, a)$ value gets adjusted according to the update rule

$$Q_{k+1}(x_k, a_k) = Q_k(x, a) + \alpha_k(x, a)[r_k + \gamma \max_{a\in\mathcal{A}} Q_k(x', a) - Q_k(x, a)], \tag{2.16}$$

where α is the learning rate that balances the weight given to recent rewards versus past experiences, influencing the convergence speed and stability of the learning process.

We denote $Q^*(x,a)$ as the unique optimal value function.

If the reward is bounded $|r_k| \leq R_{\max}$ for each episode and the learning rate $0 \leq \alpha_k < 1$ fulfills the conditions

$$\sum_{k=1}^{\infty} \alpha_k(x,a) = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k(x,a)^2 < \infty \tag{2.17}$$

for $\forall x, a$, then

$$Q_k(x,a) \to Q^*(x,a) \qquad \text{as } n \to \infty \text{ with probability 1} \tag{2.18}$$

according to [WD92].

The optimal learning rate for standard Q-learning is defined as $\alpha_k = 1/(k+1)^\omega$, where $\omega \in (0.5, 1]$, as described in [EDMB03].

The distributional counterpart of the Q-learning update rule 2.3.1 was introduced by [RBD+18a].

Definition 2.3.2. We define the categorical distributional update rule for the sample (x_k, a_k, r_k, x_{k+1}) and the action chosen by a policy a_{t+1} as

$$\eta_{k+1}^{(x_k,a_k)} := (1 - \alpha_k(x_k,a_k))\eta_k^{(x_k,a_k)} + \alpha_k(x_k,a_k))\Pi_C \mathcal{T}_k^\pi \eta_k^{(x_k,a_k)}, \tag{2.19}$$

where $\mathcal{T}_k^\pi$ is the distributional Bellman operator as defined in 2.1.4 at episode k and Π_C is the categorical projection operator as defined in 2.2.2.

As demonstrated by [LBC19], the same policies are obtained by the distributional version 2.3.2 and the value-based version of the update rule 2.3.1.

Proposition 2.3.3. *Let $\eta_0^{(x,a)} \in P_z$, $Q_0(x,a) = \mathbb{E}_{Z \sim \eta_0^{(x,a)}}(Z)$, η_t result from the update rule as described in 2.3.2, $z_1 \leq -R_{max}/(1-\gamma)$ and $z_N \geq R_{max}/(1-\gamma)$. Then the state action value functions are updated as*

$$Q(x,a)_k := \begin{cases} (1-\alpha_k(x,a))Q_{k-1}(x,a) + \alpha_k(x,a)[r' + \gamma \max_{a \in \mathcal{A}} Q_k(x',a')], \forall (x,a) = (x_k,a_k) \\ Q_{k-1}(x,a), \quad \forall (x,a) \neq (x_k,a_k) \end{cases} \tag{2.20}$$

and satisfy

$$Q_k(x,a) = \mathbb{E}_{Z \sim \eta_t^{(x,a)}}(Z), \quad \forall (x,a) \in \mathcal{X} \times \mathcal{A}, \forall k \geq 0. \tag{2.21}$$

Speedy Q-Learning

In this thesis we will focus on speedy Q-learning, which was introduced by [AMGK11], since it shows faster convergence compared to standard Q-learning.

Definition 2.3.4. After observing the state x_k during episode k, and taking action a_k that results in observing the new state x' and obtaining the immediate reward r', the $Q(x,a)$ value is updated by the rule

$$\begin{aligned} Q_{k+1}(x,a) \coloneqq Q_k(x,a) + \alpha_k[r_k + \gamma \max_{a\in\mathcal{A}} Q_k(x',a) - Q_k(x,a)] \\ +(1-\alpha_k)(Q_k(s,a) - Q_{k-1}(s,a)), \end{aligned} \tag{2.22}$$

where α is the learning rate and the third term $(1-\alpha_k)(Q_{k+1}(s,a) - Q_k(s,a))$ is the "speedy"-adjustment, which uses the previous and new Q-values to accelerate convergence by correcting the update with the difference between $Q_{k+1}(s,a)$ and $Q_k(s,a)$.

Speedy Q-learning as just described in 2.3.4 can be translated to categorical distributional reinforcement learning similar to standard Q-learning and the update rule can be defined as

$$\eta_{k+1}^{(x,a)} \coloneqq \eta_k^{(x,a)} + \alpha_k(\Pi_C \mathcal{T}_k^\pi \eta_k^{(x,a)} - \eta_k^{(x,a)}) + (1-\alpha_k)(\Pi_C \mathcal{T}_k^\pi \eta_k^{(x,a)} - \Pi_C \mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}). \tag{2.23}$$

The update rule for speedy Q-learning in algorithm format can be expressed as follows:

Algorithm 2.1: Speedy Q-learning in Categorical Distributional RL

Input: initial distribution η_0, policy π, discount factor γ, number of iterations T

$\eta_{-1} \leftarrow \eta_0$
$a_0 \leftarrow 1$
for $i \leftarrow 2$ ***to*** n **do**
  Update learning rate:
  $\alpha_k \leftarrow 1/k^\omega$ Sample $x'_k \sim p(\cdot|x,a)$, $a'_k \sim \pi(\cdot|x'_n)$, $r_k \sim (\cdot|x,a,x'_k)$
  Bellman update:
  $\mathcal{T}_k^\pi \eta_k^{(x,a)} \leftarrow \sum_{i=1}^N p_{k,i}^{(x'_k,a'_k)} \delta_{r_k}$
  Update η:
  $\eta_{k+1}^{(x,a)} \leftarrow \eta_k^{(x,a)} + \alpha_k(\Pi_C \mathcal{T}_k^\pi \eta_k^{(x,a)} - \eta_k^{(x,a)}) + (1-\alpha_k)(\Pi_C \mathcal{T}_k^\pi \eta_k^{(x,a)} - \Pi_C \mathcal{T}_k^\pi \eta_{k-1}^{(x,a)})$
end

The formula 2.23 can be written as

$$\eta_{k+1}^{(x,a)} = \frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}, \tag{2.24}$$

where $\mathcal{D}_k$ is defined as

$$\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \Pi_C \mathcal{T}_k^\pi \eta_k^{(x,a)} - (k-1)\Pi_C \mathcal{T}_k^\pi \eta_{k-1}^{(x,a)}. \tag{2.25}$$

We choose the learning rate $\alpha_k = 1/(k+1)$ because it can been shown that the convergence of speedy Q-learning is optimized with a linear learning rate. This is different from standard Q-learning for which a polynomial learning rate is optimal for convergence as shown in [EDMB03].

CHAPTER 3

Bounding the Cramér Distance in Categorical Distributional Reinforcement Learning

In this chapter we use different concentration inequalities to calculate the bounds for the Cramér distance in categorical distributional reinforcement learning.

The construction of the bounds follows the outline of [BH22]. In the following, we present assumptions necessary for deriving bounds on the Cramér distance.

Assumption 3.0.1. Consider a finite state-action space with $n := |\mathcal{X} \times \mathcal{A}|$ elements. The categorical distribution η_c is established as the unique fixed point of the operator $\Pi_C \mathcal{T}^\pi$. Rewards are bounded by a maximum value of $R_{\max} > 0$. We introduce the variable $\hat{\beta} = 1/(1-\sqrt{\gamma})$ that is defined by $\gamma < 1$. We calculate the upper limit of possible returns, $V_{\max}$, using $R_{\max}/(1-\gamma)$. For the N discretely defined atoms in this setting, the bounds are set as $z_1 = -V_{\max}$ and $z_N = V_{\max}$. Initially, the return distributions η_{-1} and η_0 are identical. Updates to η_k are according to the formula 2.24.

Prior to bounding the Cramér distance, we establish the martingale representation of the errors. Additionally, we formulate several lemmas that provide the theoretical foundation necessary to apply martingale inequalities.

3.0.1 Error Martingal

The history of the Q-learning algorithm at time k can be mathematically defined as a filtration, which consists of σ-fields generated by the trajectories $r_1, x'_1, a'_1, \cdots, r_k, x'_k, a'_k$, where $(x, a) \in \mathcal{X} \times \mathcal{A}$.

Filtrations are defined by the following properties:

- Filtrations contain the empty set $\emptyset$.
- Filtrations are closed under complementation, meaning that if $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$ also holds.
- Filtrations are closed under countable unions, meaning that if $A_1, A_2, \cdots$ are sets in $\mathcal{F}$ then the union $\bigcup_{i=1}^{\infty} A_i$ is also in $\mathcal{F}$.
- Filtrations consist of a sequence of σ-fields, indexed by the number of iterations k, that is non-decreasing. This means that for any $k_i \leq k_j$ the σ-field $\mathcal{F}_{k_i}$ is a subset of $\mathcal{F}_{k_j}$, ensuring that the information grows or remains the same.

Filtrations grow over time as the agent interacts with the environment, but are capable of handling the structure of all possible events that could potentially be observed.

The expected update is defined as

$$\mathcal{D}[\eta_k, \eta_{k-1}]^{(x,a)} \coloneqq \mathbb{E}\left[\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \mid \mathcal{F}_{k-1}\right] = k\Pi_C \mathcal{T}^\pi \eta_k^{(x,a)} - (k-1)\Pi_C \mathcal{T}^\pi \eta_{k-1}^{(x,a)}. \quad (3.1)$$

Now we can formulate the error $\epsilon_k^{(x,a)}$ and cumulative error of the sample update $E_k^{(x,a)}$ as

$$\begin{aligned} \epsilon_k^{(x,a)} &\coloneqq \mathcal{D}[\eta_k, \eta_{k-1}]^{(x,a)} - \mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)}, \\ E_k^{(x,a)} &\coloneqq \sum_{i=0}^{k} \epsilon_i^{(x,a)}. \end{aligned} \quad (3.2)$$

Now we can also write the sample update as

$$\eta_{k+1}^{(x,a)} \coloneqq \frac{k}{k+1}\eta_k^{(x,a)} + \frac{1}{k+1}\left(\mathcal{D}[\eta_k, \eta_{k-1}]^{(x,a)} - \epsilon_k^{(x,a)}\right). \quad (3.3)$$

The error can be turned into a martingale as described in [BH22].

A martingale is defined by the characteristic that the conditional expected value of the next observation, given all past observations, is equivalent to the most recent observation. This property implies that the expected future value of the process remains consistent with the present value, assuming knowledge of all prior values.

Mathematically, a stochastic process (X_n) is a martingale with the properties

- X_n is integratable, its expected value is well-defined and finite $\mathbb{E}[|X_n|] \leq \infty$.
- X_n is $\mathcal{F}_n$-measurable, which means that X_n takes all information up to time n into account.

- The conditional expectation of X_{n+1} given the σ-field $\mathcal{F}_n$ is equal to X_n, e.i.,

$$\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n. \tag{3.4}$$

Before we formulate the lemmas that show that the error can be defined as a martingal, we have to define $\mathcal{L}$ as the set of finite signed Borel measures

$$\begin{aligned}\mathcal{L} = \{v \text{ signed measure } \mid \exists F_v : \mathbb{R} \to \mathbb{R}, v((a,b]) = F_v(a),\\ |v(\mathbb{R})| < \infty, \quad \lim_{z\to-\infty} F_v(z) = 0, \quad |\lim_{z\to\infty} F_v(z)| < \infty\}.\end{aligned} \tag{3.5}$$

We can extend the categorical distributions to a subspace of the signed measures by

$$P_z \subseteq \mathcal{L}_z = \left\{ \sum_{i=1}^{N} c_i \delta_{z_i} \mid c_i \in \mathbb{R} \subseteq \mathcal{L} \right\}. \tag{3.6}$$

As described in [BH22]

$$\begin{aligned}\mathcal{D}_k[\eta_k, \eta_{k-1}]^{(x,a)} \in \mathcal{P}(P_z),\\ \eta_k^{(x,a)} \in \mathcal{P}(P_z)\end{aligned} \tag{3.7}$$

holds for all $k \geq 0$, where $\mathcal{P}(P_z)$ is the set of random measures with values in P_z. Moreover [BH22] introduced the following lemmas.

Lemma 3.0.2. *The inclusions $\epsilon_k^{(x,a)} \in \mathcal{P}(\mathcal{L}_z)$ and $E_k^{(x,a)} \in \mathcal{P}(\mathcal{L}_z)$ hold for all $k \geq 0$. For each atom z_i, it holds that the cumulative distribution functions of the error ϵ_k evaluated at z_i form a uniformly bounded martingale different sequence, i.e.,*

$$\forall k \geq 0 : \quad \mathbb{E}\left[F_{\epsilon_k^{(x,a)}}(z_i) \mid \mathcal{F}_{k-1} \right] = 0 \wedge \mid F_{\epsilon_k^{(x,a)}}(z_i) \mid \leq 1. \tag{3.8}$$

As $\mathcal{L}_z$ is a vector space we introduce the l_2-norm for the distributional case as

$$\|x\|_{l_2} \coloneqq \left(\sum_{i=1}^{N-1} (z_{i+1} - z_i) F_v(z_i)^2 + F_v(z_N)^2 \right)^{\frac{1}{2}} \tag{3.9}$$

for all $v \in \mathcal{L}_z$.

We can also define $\bar{l}_2$ as a norm by taking the supremum over all state-action pairs

$$\|v\|_{\bar{l}_\infty} \coloneqq \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|v\|_{l_\infty} = \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \max_{1\leq i\leq N} |F_v(z_i)|. \tag{3.10}$$

Due to the definition of the l_2-norm, the l_∞-norm and the assumption 3.0.1 the following inequalities hold for all $\mu, v \in P_z$:

$$\begin{aligned}l_2(\mu, v) = \|\mu - v\|_{l_2} \leq \sqrt{2V_{\max}}\|\mu - v\|_{l_\infty} \leq \sqrt{2V_{\max}}\\ \|E_k\|_{\bar{l}_2} \leq \sqrt{2V_{\max}}\|E_k\|_{\bar{l}_\infty}\end{aligned} \tag{3.11}$$

Lemma 3.0.3. *For all $k \geq 1$, the equality*

$$\eta_k = \frac{1}{k}(\Pi_C \mathcal{T}^\pi \eta_0 + (k-1)\Pi_C \mathcal{T}^\pi \eta_{k-1} - E_{k-1}) \tag{3.12}$$

holds.

If lemma 3.0.3 holds, the statement

$$\eta_k \approx \Pi_C \mathcal{T}^\pi \eta_{k-1} \tag{3.13}$$

also holds, because the influence of the initial state and the error term decreases as k increases. As k grows larger, the term $\Pi_C \mathcal{T}^\pi \eta_{k-1}$ becomes the dominant component in the expression. With the help of the equation 3.11, the lemmas 3.0.3 and 3.0.2 and the fact that $\Pi_C \mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in $\bar{l}_2$ [BH22] showed the following lemma.

Lemma 3.0.4. *For all $k \geq 1$, the inequality*

$$\|\eta_C - \eta_k\|_{\bar{l}_2} \leq \frac{\sqrt{\gamma}\bar{\beta}}{k}\sqrt{2V_{max}} + \frac{1}{k}\sum_{j=1}^{k}\sqrt{\gamma}^{k-j}\|E_{j-1}\|_{\bar{l}_2} \tag{3.14}$$

holds.

Now we can bound the Cramér distance in categorical distributional reinforcement learning.

3.0.2 Concentration Inequalities

Concentration inequalities are used for bounding the probability that a random variable deviates from some value.

Definition 3.0.5. Let X be a random variable with expected value $\mathbb{E}[X]$. A concentration inequality gives a bound of the form

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq f(\epsilon), \tag{3.15}$$

where $\epsilon > 0$ is a deviation threshold and $f(\epsilon)$ is a function that decays as ϵ decreases.

We now utilize concentration inequalities alongside the equation 3.11 and the lemmas 3.0.3 and 3.0.4, to derive bounds on the Cramér distance $\bar{l}_2(\eta_C, \eta_T)$. This distance is a measure of the difference between two probability distributions, initially described in [Cra28].

This method provides a foundation for the complexity analysis later and helps us to understand the convergence behaviors of distributional RL algorithms.

3.1 Maximal Hoeffding–Azuma Inequality

Hoeffding's inequality was introduced by Wassily Hoeffding (1963) [Hoe94] and provides an exponential bound on the tail probabilities of the sum of independent random variables within a specified range.

Definition 3.1.1. Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables such that $a_i \leq X_i \leq b_i$ holds for all $1 \leq i \leq n$. The sum is defined as $S_n = \sum_{i=1}^{n} X_i$. Then the inequality

$$\mathbb{P}\left[S_n - \mathbb{E}[S_n] \geq \epsilon\right] \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{3.16}$$

holds for any $\epsilon > 0$.

The maximal Hoeffding–Azuma inequality extends the classic Hoeffding inequality to handle the maximum of partial sums of a martingale difference sequence and is defined as follows.

Definition 3.1.2. Let $\mathcal{V} = V_1, V_2, ..., V_n$ be a bounded martingale difference sequence $|V_i| \leq L$ with respect to the filtration $\mathcal{F}_k(\mathbb{E}[\mathcal{V}_k|\mathcal{F}_{k-1}] = 0)$. Let the associated martingale be defined as the sum

$$S_i = \sum_{j=1}^{i} V_j. \tag{3.17}$$

Then the inequality

$$\mathbb{P}\left[\max_{i=1,...,n} S_i > \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2TL}\right) \tag{3.18}$$

holds vor all constants ϵ, $\nu > 0$, where T is the number of time steps.

With the help of this inequality another lemma can be shown.

Lemma 3.1.3. *For all $\epsilon > 0$ and all time steps T, under the assumption 3.0.1 the inequality*

$$\mathbb{P}\left[\max_{i \leq k \leq T} \|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq 2nN\exp\left(\frac{-\epsilon^2}{2T}\right) \tag{3.19}$$

holds.

The paper [BH22] showed that under assumption 3.0.1 and with 3.1.3 the following inequality

$$\bar{l}_2^H(\eta_C, \eta_T) \leq \sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T} + \sqrt{\frac{2\ln\frac{2nN}{\delta}}{T}}\right) \tag{3.20}$$

holds with probability of at least $1 - \delta$.

3.2 Bernstein Inequality

The Bernstein inequality introduced by Sergei N. Bernstein (1946) [Ber46] can offer tighter bounds by incorporating information about both the variance and the range of the random variables.

Definition 3.2.1. Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables, each bounded such that $|X_i| \leq M$ holds almost surely for all $1 \leq i \leq n$. The sum and total variance are defined as $S_n = \sum_{i=1}^{n} X_i$ and $\sigma^2 = \sum_{i=1}^{n} \mathbb{V}\text{ar}(X_i)$ Then the inequality

$$\mathbb{P}\left[S_n - \mathbb{E}[S_n] \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon/2}{\sigma^2 + M\epsilon/3}\right) \tag{3.21}$$

holds for any $\epsilon > 0$.

We utilize a specific form of the Bernstein inequality, which is modified for martingale difference sequences.

The Bernstein inequality for martingales which is also known as the Freedman inequality is defined by [Fre75] and [CBL06] as follows

Definition 3.2.2. Let $\mathcal{V} = V_1, V_2, ..., V_T$ be a bounded martingale difference sequence $|V_i| \leq L$ with respect to the filtration $\mathcal{F}_k(\mathbb{E}[\mathcal{V}_k|\mathcal{F}_{k-1}] = 0)$. The associated martingale is defined as the sum

$$S_i = \sum_{j=1}^{i} X_j, \tag{3.22}$$

and the sum of the conditional variances by

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathbb{E}[X_t^2|\mathcal{F}_{t-1}]. \tag{3.23}$$

Then the inequality

$$\mathbb{P}\left[\max_{i=1,...,T} S_i > \epsilon \wedge \Sigma_n^2 \leq \nu\right] \leq \exp\left(-\frac{\epsilon^2}{2(\nu + L\epsilon/3)}\right) \tag{3.24}$$

holds for all constants ϵ, $\nu > 0$. Therefore the inequality

$$\mathbb{P}\left[\max_{i=1,...,T} S_i > \sqrt{2\nu\epsilon} + (\sqrt{2}/3)L\epsilon \wedge \Sigma_T^2 \leq \nu\right] \leq e^{-\epsilon} \tag{3.25}$$

also holds.

We now want to define an inequality to bound the Cramér distance $\bar{l}_2(\eta_C, \eta_T)$ similar to 3.20.

Lemma 3.2.3. *Under the assumption 3.0.1 the inequality*

$$\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq nN\exp\left(-\frac{\epsilon^2}{2(T+\epsilon/3)}\right) \tag{3.26}$$

holds for all $\epsilon > 0$ and time steps T.

Proof. We define

$$E_k^i = F_{E_k^{(x,a)}}(z_i) = \sum_{j=0}^{k} F_{\epsilon_j^{(x,a)}}(z_i) \tag{3.27}$$

for $(x,a) \in \mathcal{X}\times\mathcal{A}$. By Lemma 3.0.2 the martingal difference sequence $V_j = F_{\epsilon_j^{(x,a)}}(z_i), j = 0,\dots,T$ with regards to the filtration $\mathcal{F}_j$ is uniformly bounded by 1. We can therefore use the Bernstein inequality 3.2.2 on E_k^j and Σ_T^2 is at most T. We therefore get the inequality

$$\mathbb{P}\left[\max_{i=1,\dots,n}|E_{k-1}^i| > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2(T+\epsilon/3)}\right). \tag{3.28}$$

To get an inequality for E_{k-1}, we first must take the union over all atoms

$$\begin{aligned}
&\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \\
&= \mathbb{P}\left[\max_{i\leq k\leq T}\max_{1\leq i\leq n}|E_{k-1}^i| > \epsilon\right] = \mathbb{P}\left[\bigcup_{i=1}^{N}\max_{1\leq i\leq n}|E_{k-1}^i| > \epsilon\right] \\
&\leq N\exp\left(-\frac{\epsilon^2}{2(T+\epsilon/3)}\right).
\end{aligned} \tag{3.29}$$

When extending this to the n state-action pairs $(x,a) \in \mathcal{X}\times\mathcal{A}$ we get

$$\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq nNexp\left(-\frac{\epsilon^2}{2(T+\epsilon/3)}\right). \tag{3.30}$$

□

With the help of those inequalities we can show the Cramér distance using the Bernstein inequality.

Lemma 3.2.4. *Under the assumption 3.0.1 the inequality*

$$\bar{l}_2^B(\eta_C,\eta_T) \leq \sqrt{2V_{max}}\beta\left(\frac{\sqrt{\gamma}}{T} + \frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right) + \sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2 + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T}\right) \tag{3.31}$$

holds for all $\epsilon > 0$ and all time steps T.

Proof. Firstly we set the right hand side of the inequality in 3.2.3

$$\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq nN\exp\left(-\frac{\epsilon^2}{2(T+\epsilon/3)}\right) \tag{3.32}$$

to δ and get the ϵ. Starting with the inequality

$$nN\exp\left(-\frac{\epsilon^2}{(2(T+\epsilon/3)}\right) = \delta \tag{3.33}$$

we isolate the exponent leading to

$$-\frac{\epsilon^2}{2(T+L\epsilon/3)} = \ln\left(\frac{\delta}{nN}\right). \tag{3.34}$$

Now simplifying this equation results in the quadratic equation

$$\epsilon^2 + \frac{2\epsilon}{3}\ln\left(\frac{\delta}{nN}\right) + 2T\ln\left(\frac{\delta}{nN}\right). \tag{3.35}$$

Solving this quadratic equation yields

$$\epsilon = \frac{-\frac{2}{3}\ln\left(\frac{\delta}{nN}\right) \pm \sqrt{(\frac{2}{3}\ln\left(\frac{\delta}{nN}\right))^2 - 8T\ln\left(\frac{\delta}{nN}\right)}}{2}. \tag{3.36}$$

We will use the positive quadratic solution

$$\epsilon = \frac{-\frac{2}{3}\ln\left(\frac{\delta}{nN}\right) + \sqrt{(\frac{2}{3}\ln\left(\frac{\delta}{nN}\right))^2 - 8T\ln\left(\frac{\delta}{nN}\right)}}{2}. \tag{3.37}$$

Because it the only solution that yields a guaranteed positive ϵ. We now rewrite the inequality as

$$\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} \leq \frac{-\frac{2}{3}\ln\left(\frac{\delta}{nN}\right) + \sqrt{(\frac{2}{3}\ln\left(\frac{\delta}{nN}\right))^2 - 8T\ln\left(\frac{\delta}{nN}\right)}}{2}\right] \leq 1-\delta. \tag{3.38}$$

The inequality as defined in 3.0.4 can now be modified to find

$$\begin{aligned}
\|\eta_C - \eta_T\|_{\bar{l}_2} &\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{1}{T}\sum_{j=1}^{k}\sqrt{\gamma}^{T-j}\|E_{j-1}\|_{\bar{l}_2} \\
&\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{\bar{\beta}}{T}\sqrt{2V_{max}}\max_{1\leq j\leq T}\|E_j - 1\|_{\bar{l}_\infty} \\
&\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{\bar{\beta}}{T}\sqrt{2V_{max}}\frac{-\frac{2}{3}\ln\left(\frac{\delta}{nN}\right) + \sqrt{(\frac{2}{3}\ln\left(\frac{\delta}{nN}\right))^2 - 8T\ln\left(\frac{\delta}{nN}\right)}}{2} \\
&= \sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T} + \frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right) + \sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2 + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T}\right).
\end{aligned} \tag{3.39}$$

□

3.3 Bennett's Inequality

Bennett's inequality as introduced by George Bennett (1962) [Ben62] provides a refined bound under certain conditions, especially when the variance is relatively small compared to the maximum possible deviation. This approach often leads to tighter bounds in practice.

Definition 3.3.1. Let $X_1, X_2, ..., X_n$ be a sequence of independent real-valued random variables such that they have zero mean and $X_i \leq 1$ holds with probability 1. The sum and total variance are defined as $S_n = \sum_{i=1}^{n} X_i$ and $\sigma^2 = \sum_{i=1}^{n} \mathbb{V}\text{ar}(X_i)$. Then the inequality

$$\mathbb{P}\left[S_n > \epsilon\right] \leq \exp\left(-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)\right), \tag{3.40}$$

holds for any $\epsilon > 0$, where $h(u) = (1+u)\log(1+u) - u$ for $u \geq 0$.

The martingale difference sequence $\mathcal{V} = V_1, V_2, ..., V_T$ fulfills conditions for Bennett's inequality.

We can now extend Bennett's inequality to martingales which originally has been shown by David Pollard [Pol].

Definition 3.3.2. Let $\mathcal{V} = V_1, V_2, ..., V_T$ be a bounded martingale difference sequence $|V_i| \leq L$ for $L > 0$ with respect to the filtration $\mathcal{F}_k(\mathbb{E}[\mathcal{V}_k|\mathcal{F}_{k-1}] = 0)$. The associated martingale is defined as the sum

$$S_i = \sum_{j=1}^{i} V_j, \tag{3.41}$$

and the sum of the conditional variances by

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathbb{E}[V_t^2|\mathcal{F}_{t-1}]. \tag{3.42}$$

Then the inequality

$$\mathbb{P}(S_T > \epsilon) \leq \mathbb{P}(\Sigma_T^2 > W) + \exp\left(-\frac{\epsilon^2}{2W}\psi\left(\frac{L\epsilon}{W}\right)\right) \tag{3.43}$$

holds for all constants $\epsilon, \nu > 0$, where $\psi(t)$ denotes the function

$$\psi(t) \coloneqq \frac{(1+t)\log(1+t) - t}{t^2/2}. \tag{3.44}$$

With the assumption defined in 3.0.1, we can now formulate a lemma similiar to 3.20 and 3.2.4.

Lemma 3.3.3. *Let $\mathcal{V} = V_1, V_2, ..., V_T$ be a bounded martingale difference sequence $|V_i| \leq 1$ under the assumption 3.0.1 with respect to the filtration $\mathcal{F}_k(\mathbb{E}[\mathcal{V}_k|\mathcal{F}_{k-1}] = 0)$. The associated martingale is defined as the sum*

$$S_i = \sum_{j=1}^{i} V_j, \tag{3.45}$$

and the sum of the conditional variances by

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathbb{E}[V_t^2|\mathcal{F}_{t-1}]. \tag{3.46}$$

Then the following inequality holds for all constants ϵ,

$$\mathbb{P}\left[\max_{i=1,...,T} S_i > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right) \tag{3.47}$$

where T is the number of iterations and $\psi(t)$ is defined as

$$\psi(t) \coloneqq \frac{(1+t)\log(1+t) - t}{t^2/2}. \tag{3.48}$$

Proof. It is trivial that we can use the inequality from 3.3.2 with $\mathbb{P}\left[\max_{i=1,...,T} S_i > \epsilon\right]$.

We choose W from the inequality in 3.3.2 to be T, therefore the term $\mathbb{P}(\Sigma_T^2 > W)$ disappears as $\Sigma_T^2 \leq T$ holds due to the martingale difference being bounded by 1. □

Now we want to define an inequality to bound the Cramér distance $\bar{l}_2(\eta_C, \eta_T)$ similar to 3.20 and 3.2.4.

Therefore, we define the following lemma.

Lemma 3.3.4. *Under the assumption 3.0.1 the inequality*

$$\mathbb{P}\left[\max_{i\leq k\leq T} \|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq nN \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right) \tag{3.49}$$

holds for all $\epsilon > 0$ and time steps T, where $\psi(t)$ denotes the function

$$\psi(t) \coloneqq \frac{(1+t)\log(1+t) - t}{t^2/2}. \tag{3.50}$$

Proof. We define

$$E_k^i = F_{E_k^{(x,a)}}(z_i) = \sum_{j=0}^{k} F_{\epsilon_j^{(x,a)}}(z_i) \tag{3.51}$$

for $(x, a) \in \mathcal{X} \times \mathcal{A}$.

By lemma 3.0.2 the martingal difference sequence $V_j = F_{\epsilon_j^{(x,a)}}(z_i), j = 0, ..., T$ with regards to the filtration $\mathcal{F}_j$ is uniformly bounded by 1. We can therefore use Bennett's inequality for martingales 3.3.3 on E_k^j and Σ_T^2 is at most T. We define the inequality

$$\mathbb{P}\left[\max_{i=1,...,n} |E_{k-1}^i| > \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right). \tag{3.52}$$

To get an inequality for E_{k-1}, we first must take the union over all atoms

$$\begin{aligned} &\mathbb{P}\left[\max_{i\leq k\leq T} \|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \\ &= \mathbb{P}\left[\max_{i\leq k\leq T} \max_{1\leq i\leq n} |E_{k-1}^i| > \epsilon\right] = \mathbb{P}\left[\bigcup_{i=1}^{N} \max_{1\leq i\leq n} |E_{k-1}^i| > \epsilon\right] \\ &\leq N \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right). \end{aligned} \tag{3.53}$$

When extending this to the n state-action pairs $(x, a) \in \mathcal{X} \times \mathcal{A}$, we get

$$\mathbb{P}\left[\max_{i\leq k\leq T} \|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] \leq nN \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right). \tag{3.54}$$

□

Now we can bound the Cramér distance using the Bennett inequality.

Lemma 3.3.5. *For all $\epsilon > 0$ and all time steps T , under the assumption 3.0.1 the inequality*

$$\bar{l}_2^{Be}(\eta_C, \eta_T) \leq \sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T} + \sqrt{\frac{\ln(\frac{nN}{\delta})}{T}}\right) \tag{3.55}$$

holds.

Proof. Firstly we set the right hand side of the inequality in 3.2.3 to find

$$\begin{aligned} \mathbb{P}\left[\max_{i\leq k\leq T} \|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right] &\leq nN \exp\left(-\frac{\epsilon^2}{2T}\psi\left(\frac{\epsilon}{T}\right)\right) \\ &\leq nN \exp\left(-\frac{\epsilon^2}{2T}\frac{(1+\frac{\epsilon}{T})\log(1+\frac{\epsilon}{T}) - \frac{\epsilon}{T}}{\epsilon^2/2T^2}\right) \\ &\leq nN \exp\left(-T\left(1+\frac{\epsilon}{T}\right)\log\left(1+\frac{\epsilon}{T}\right) - \frac{\epsilon}{T}\right). \end{aligned} \tag{3.56}$$

We simplify the term $\log(1 + \epsilon/T)$ to ϵ/T because solving the equation later for ϵ is not feasible as it is a transcendental equation.

It is possible to simplify the term, as $\log(1+\epsilon/T) < x$ holds for $x > 0$. In addition it is a good approximation when ϵ/T is small, as supported by the Taylor series expansion of the logarithm function around zero. We find

$$\begin{aligned}\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} > \epsilon\right[&\leq nN\exp\left(-T\left(1+\frac{\epsilon}{T}\right)\frac{\epsilon}{T}-\frac{\epsilon}{T}\right)\\ &\leq nN\exp\left(-\frac{\epsilon^2}{T}\right).\end{aligned} \tag{3.57}$$

To get δ and ϵ we do some transformations. Starting with the inequality

$$nN\exp\left(-\frac{\epsilon^2}{T}\right) = \delta \tag{3.58}$$

we isolate ϵ to get

$$\epsilon = \sqrt{-T\ln\left(\frac{\delta}{nN}\right)}. \tag{3.59}$$

The inequality $0 < \delta/nN < 1$ always holds and therefore the expression inside the square root is positive and provides a real solution. We now substitute ϵ and get the equation

$$\mathbb{P}\left[\max_{i\leq k\leq T}\|E_{k-1}\|_{\bar{l}_\infty} \leq \sqrt{-T\ln\left(\frac{\delta}{nN}\right)}\right] \leq 1-\delta. \tag{3.60}$$

As a final step we modify the inequality as defined in 3.0.4

$$\begin{aligned}\|\eta_C - \eta_T\|_{\bar{l}_2} &\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{1}{T}\sum_{j=1}^{k}\sqrt{\gamma}^{T-j}\|E_{j-1}\|_{\bar{l}_2}\\ &\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{\bar{\beta}}{T}\sqrt{2V_{max}}\max_{1\leq j\leq T}\|E_j - 1\|_{\bar{l}_\infty}\\ &\leq \frac{\sqrt{\gamma}\bar{\beta}}{T}\sqrt{2V_{max}} + \frac{\bar{\beta}}{T}\sqrt{2V_{max}}\sqrt{-T\ln\left(\frac{\delta}{nN}\right)}\\ &= \frac{\bar{\beta}}{T}\sqrt{2V_{max}}\left(\sqrt{\gamma} + \sqrt{-T\ln\left(\frac{\delta}{nN}\right)}\right)\\ &= \sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T} + \sqrt{\frac{\ln(\frac{nN}{\delta})}{T}}\right).\end{aligned} \tag{3.61}$$

□

CHAPTER 4

Convergence and Complexity Analysis

In this section, we analyze the convergence properties and computational complexity of the algorithm, based on the bounds of the Cramér distance established in the preceding section.

Our primary objective is to compare the minimum number of time steps T required for the algorithm to achieve a specified precision ϵ. This analysis is crucial to understanding the practical applicability of the algorithm in real-world scenarios, where both accuracy and computational resources are of great importance.

Lemma 4.0.1. *Under the assumption 3.0.1 the inequality*

$$\begin{aligned} &\sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T}+\sqrt{\frac{\ln(\frac{nN}{\delta})}{T}}\right) \\ \leq \sqrt{2V_{max}}\bar{\beta}&\left(\frac{\sqrt{\gamma}}{T}+\frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right)+\sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2+8T\ln\left(\frac{nN}{\delta}\right)}}{2T}\right) \\ &\leq \sqrt{2V_{max}}\bar{\beta}\left(\frac{\sqrt{\gamma}}{T}+\sqrt{\frac{2\ln\left(\frac{2nN}{\delta}\right)}{T}}\right) \end{aligned} \tag{4.1}$$

holds for $\ln^2(nN/\delta)/18\ln(2nN/\delta) \leq T$.

Proof. After simplifying the first term and dividing by $\sqrt{2V_{max}}\bar{\beta}$, it is sufficient to show that the inequality

$$\sqrt{\frac{\ln(\frac{nN}{\delta})}{T}} \leq \frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right) + \sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2 + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T} \leq \sqrt{\frac{2\ln\left(\frac{2nN}{\delta}\right)}{T}} \tag{4.2}$$

holds. We square the inequality and now have to show that the following two inequalities

$$\begin{aligned} \frac{\ln(\frac{nN}{\delta})}{T} &\leq \frac{\frac{8}{9}\ln^2\left(\frac{nN}{\delta}\right) + 8T\ln\left(\frac{nN}{\delta}\right) + \frac{4}{3}\ln\left(\frac{nN}{\delta}\right)\sqrt{\ln\left(\frac{nN}{\delta}\right)\left(\frac{4}{9}\ln\left(\frac{nN}{\delta}\right) + 8T\right)}}{4T^2} \\ \frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right) + \sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2 + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T} &\leq \sqrt{\frac{2\ln\left(\frac{2nN}{\delta}\right)}{T}} \end{aligned} \tag{4.3}$$

hold. We begin by demonstrating that the initial inequality is satisfied. Because $\ln(nN/\delta) > 0$ always holds, it suffices to show that the inequality

$$\begin{aligned} \frac{\ln(\frac{nN}{\delta})}{T} &\leq \frac{8T\ln(\frac{nN}{\delta})}{4T^2} \\ \ln\left(\frac{nN}{\delta}\right) &\leq 2\ln\left(\frac{nN}{\delta}\right) \end{aligned} \tag{4.4}$$

holds. which is the case for all $T > 0$.

We show the second inequality by showing that the following inequalities

$$\begin{aligned} \frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right)}{2T} &\leq \sqrt{\frac{2\ln\left(\frac{2nN}{\delta}\right)}{T}} \\ \frac{\sqrt{(-\frac{2}{3}\ln\left(\frac{nN}{\delta}\right))^2 + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T} &\leq \sqrt{\frac{2\ln\left(\frac{2nN}{\delta}\right)}{T}} \end{aligned} \tag{4.5}$$

hold. The first inequality holds for

$$T \geq \frac{\ln^2(\frac{nN}{\delta})}{18\ln(\frac{2nN}{\delta})}. \tag{4.6}$$

We show this by squaring both sides

$$\frac{\frac{4}{9}\ln^2\left(\frac{nN}{\delta}\right)}{4T^2} \leq \frac{2\ln\left(\frac{2nN}{\delta}\right)}{T} \tag{4.7}$$

and isolating T

$$T \geq \frac{\ln^2(\frac{nN}{\delta})}{18\ln(\frac{2nN}{\delta})}. \tag{4.8}$$

$$\frac{\frac{2}{3}\ln\left(\frac{nN}{\delta}\right)}{2T} = \frac{\ln^2(\frac{nN}{\delta})}{9T^2} \leq \frac{\ln(\frac{2nN}{\delta})}{T}. \tag{4.9}$$

The second inequality

$$\begin{gathered}\frac{\sqrt{\frac{4}{9}\ln^2(\frac{nN}{\delta}) + 8T\ln\left(\frac{nN}{\delta}\right)}}{2T} \leq \sqrt{\frac{\left(\ln(2) + \ln(\frac{nN}{\delta})\right)}{T}} \\ T \geq \frac{\ln^2(\frac{nN}{\delta})}{9(\ln(2) - \ln(\frac{nN}{\delta})}\end{gathered} \tag{4.10}$$

holds for all $T > 0$. This is valid for any $T > 0$, since $\ln(nN/\delta) > \ln(2)$ always holds, due to $n \geq 2$, $N \geq 1$ and $\delta < 1$. □

We demonstrated in lemma 4.0.1 that the bound on the Cramér distance utilizing the Bennett inequality for martingales 3.3.5 is the tightest bound followed by the bound utilizing the Bernstein inequality for martingales.

Both bounds are tighter than the bound utilizing the Maximal Hoeffding–Azuma inequality 3.20 for $\ln^2(nN/\delta)/18\ln(2nN/\delta) \leq T$. It is important to note that $\bar{l}_2^{Be}(\eta_C, \eta_T) \leq \bar{l}_2^H(\eta_C, \eta_T)$ holds for all $T > 0$.

For the inequality 3.20 the following corollary has been shown in [BH22].

Corollary 4.0.2. *Under the assumption 3.0.1, η_T converges to η_C almost surely in $\bar{l}_2^H(\eta_C, \eta_T)$.*

Consequently, η_T also almost surely converges to η_C for both bounds $\bar{l}_2^B(\eta_C, \eta_T)$ and $\bar{l}_2^{Be}(\eta_C, \eta_T)$, given their increased tightness as demonstrated in 4.0.1.

Now we can proceed with the complexity analysis to find and compare the minimum number of time steps T required to get the corresponding bound that has the specified precision ϵ.

4.1 Analysis of $\bar{l}_2^H(\eta_C, \eta_T)$

We start of by examining the number of iteration T for the Maximal Hoeffding–Azuma Cramér Distance bound to have the precision ϵ as this serves as our baseline.

Lemma 4.1.1. *Under the assumption 3.0.1, the inequality $\bar{l}_2^H(\eta_C, \eta_T) \leq \epsilon$ holds for any $0 \leq \epsilon \leq \sqrt{V_{max}}$ with the probability of $1 - \delta$ after*

$$T \geq \frac{\left(\sqrt{V_{max}\ln\left(\frac{2nN}{\delta}\right)}\bar{\beta} + \sqrt{V_{max}\ln\left(\frac{2nN}{\delta}\right)\bar{\beta}^2 + \epsilon\sqrt{2V_{max}}\gamma\bar{\beta}}\right)^2}{\epsilon^2} \tag{4.11}$$

iterations of speedy Q-learning.

Proof. We define $\kappa = \sqrt{2V_{max}\gamma}\bar{\beta}$ and $\tau = \sqrt{V_{max}\ln(2nN/\delta)}\bar{\beta}$ and can therefore write (3.20) as

$$\begin{aligned} \frac{\kappa}{T} + \frac{2\tau}{\sqrt{T}} &\leq \epsilon, \\ \kappa + 2\tau\sqrt{T} &\leq \epsilon T, \\ \epsilon T - 2\tau\sqrt{T} - \kappa &\geq 0. \end{aligned} \tag{4.12}$$

By setting $x = \sqrt{T}$, we get

$$\begin{aligned} \epsilon x^2 - 2\tau x - \kappa &= 0 \\ x_{1,2} = \frac{2\tau \pm \sqrt{4\tau^2 + 4\epsilon\kappa}}{2\epsilon} &= \frac{\tau \pm \sqrt{\tau^2 + \epsilon\kappa}}{\epsilon} \end{aligned} \tag{4.13}$$

by using the quadratic formula. We choose the larger solution because it provides the T that satisfies the inequality, and then substitute the terms back in and get

$$\begin{aligned} T \geq x_2^2 &= \frac{(\tau + \sqrt{\tau^2 + \epsilon\kappa})^2}{\epsilon^2} \\ &= \frac{\left(\sqrt{V_{max}\ln\left(\frac{2nN}{\delta}\right)}\bar{\beta} + \sqrt{V_{max}\ln\left(\frac{2nN}{\delta}\right)\bar{\beta}^2 + \epsilon\sqrt{2V_{max}\gamma}\bar{\beta}}\right)^2}{\epsilon^2}. \end{aligned} \tag{4.14}$$

□

4.2 Analysis of $\bar{l}_2^B(\eta_C, \eta_T)$

We proceed with analysing the Bernstein Cramér distance in more detail, which is tighter than the Maximal Hoeffding–Azuma Cramér Distance bound if the number of iteration exceeds $\ln^2(nN/\delta)/18\ln(2nN/\delta)$ as shown in lemma 4.0.1.

The following lemma demonstrates the number of iterations T necessary for the bound to be tighter than ϵ.

Lemma 4.2.1. *Under the assumption 3.0.1, the inequality $\bar{l}_2^B(\eta_C, \eta_T) \leq \epsilon$ holds for any $0 \leq \epsilon \leq \sqrt{V_{max}}$ with the probability of $1-\delta$ after*

$$\begin{aligned} T \geq &\frac{\sqrt{2V_{max}}\bar{\beta}\left(\epsilon(\sqrt{\gamma} + \frac{1}{3}\ln(\frac{nN}{\delta})) + \sqrt{2V_{max}}\bar{\beta}\ln(\frac{nN}{\delta})\right)}{\epsilon^2} \\ &+ \frac{\sqrt{2V_{max}}\bar{\beta}\sqrt{2\epsilon(\sqrt{\gamma} + \frac{1}{3}\ln(\frac{nN}{\delta}))\sqrt{2V_{max}}\bar{\beta}\ln(\frac{nN}{\delta}) + (2V_{max}\bar{\beta}^2 + \frac{\epsilon^2}{9})\ln^2(\frac{nN}{\delta})}}{\epsilon^2} \end{aligned} \tag{4.15}$$

iterations of speedy Q-learning.

Proof. We start off by defining constants

$$A = \sqrt{2V_{max}}\bar{\beta}, \quad B = \sqrt{\gamma}, \quad C = \ln(\frac{nN}{\delta}) \tag{4.16}$$

to simplify the inequality 3.2.4. Then the inequality 3.2.4 becomes

$$A\left(\frac{B}{T} + \frac{\frac{2}{3}C + \sqrt{(-\frac{2}{3}C)^2 + 8TC}}{2T}\right) \leq \epsilon. \tag{4.17}$$

By simplifying and multiplying both sides by $2T$ we have

$$A\left(2B + \frac{2}{3}C + \sqrt{\frac{4}{9}C^2 + 8TC}\right) \leq 2\epsilon T. \tag{4.18}$$

Next we isolate the square root term

$$\sqrt{\frac{4}{9}C^2 + 8TC} \leq \frac{2\epsilon T}{A} - 2\left(B + \frac{1}{3}C\right). \tag{4.19}$$

Let $D \coloneqq B + 1/3C$, then

$$\sqrt{\frac{4}{9}C^2 + 8TC} \leq \frac{2\epsilon T}{A} - 2D \tag{4.20}$$

holds. After expanding the right side, bringing all terms to one side and dividing both sides by 4 we get

$$0 \leq \frac{\epsilon^2 T^2}{A^2} - \left(\frac{2\epsilon D}{A} + 2C\right)T + D^2 - \frac{1}{9}C^2. \tag{4.21}$$

We now multiply both sides by A^2 to eliminate denominators and set the quadratic equal to zero to apply the quadratic formula and get the solution

$$\begin{gathered} \epsilon^2 T^2 - (2\epsilon DA + 2CA^2)T + A^2(D^2 - \frac{1}{9}C^2) = 0 \\ T = \frac{2\epsilon DA + 2CA^2 \pm \sqrt{(2\epsilon DA + 2CA^2)^2 - 4\epsilon^2 A^2(D^2 - \frac{1}{9}C\check{\mathrm{S}}}}{2\epsilon^2}. \end{gathered} \tag{4.22}$$

After simplifying and focusing on the solution corresponding to the positive root, we get

$$T = \frac{A\left(\epsilon D + AC + \sqrt{2\epsilon DAC + (A^2 + \frac{\epsilon^2}{9})C^2}\right)}{\epsilon^2}. \tag{4.23}$$

Therefore after simplifying, the minimum T satisfying the inequality after substituting is

$$\begin{gathered} T \geq \frac{\sqrt{2V_{max}}\bar{\beta}\left(\epsilon(\sqrt{\gamma} + \frac{1}{3}\ln(\frac{nN}{\delta})) + \sqrt{2V_{max}}\bar{\beta}\ln(\frac{nN}{\delta})\right)}{\epsilon^2} \\ + \frac{\sqrt{2V_{max}}\bar{\beta}\sqrt{2\epsilon(\sqrt{\gamma} + \frac{1}{3}\ln(\frac{nN}{\delta}))\sqrt{2V_{max}}\bar{\beta}\ln(\frac{nN}{\delta}) + (2V_{max}\bar{\beta}^2 + \frac{\epsilon^2}{9})\ln^2(\frac{nN}{\delta})}}{\epsilon^2}. \end{gathered} \tag{4.24}$$

□

4.3 Analysis of $\bar{l}_2^{Be}(\eta_C, \eta_T)$

We continue with analysing the Bennett Cramér distance in more detail, which is the tightest of the bounds according to lemma 4.0.1.

Now we define a new lemma similiar to the one regarding the Maximal Hoeffding–Azuma Cramér Distance 3.20 and the Bernstein Cramér Distance 3.2.4.

Lemma 4.3.1. *Under the assumption 3.0.1, the inequality* $\bar{l}_2^{Be}(\eta_C, \eta_T) \leq \epsilon$ *holds for any* $0 \leq \epsilon \leq \sqrt{V_{max}}$ *with the probability of* $1 - \delta$ *after*

$$T \geq \frac{\left(\sqrt{V_{max} \ln\left(\frac{nN}{\delta}\right)}\bar{\beta} + \sqrt{V_{max} \ln\left(\frac{nN}{\delta}\right)\bar{\beta}^2 + 4\epsilon\sqrt{2V_{max}\gamma}\bar{\beta}}\right)^2}{4\epsilon^2} \tag{4.25}$$

iterations of speedy Q-learning.

Proof. The proof is structured similiar to the proof of 4.1.1. We start off by defining the following variables $\kappa := \sqrt{2V_{max}\gamma}\bar{\beta}$ and $\tau := \sqrt{V_{max} \ln\left(\frac{nN}{\delta}\right)}\bar{\beta}$ and can therefore write (3.20) as follows

$$\begin{aligned} \frac{\kappa}{T} + \frac{\tau}{\sqrt{T}} &\leq \epsilon, \\ \epsilon T - \tau\sqrt{T} - \kappa &\geq 0. \end{aligned} \tag{4.26}$$

By setting $x := \sqrt{T}$, we can solve this using the quadratic formula to get

$$\begin{aligned} \epsilon x^2 - \tau x - \kappa &= 0 \\ x_{1,2} &= \frac{\tau \pm \sqrt{\tau^2 + 4\epsilon\kappa}}{2\epsilon}. \end{aligned} \tag{4.27}$$

We choose the larger solution because it provides the T that satisfies the inequality, and then substitute the terms back in to get

$$\begin{aligned} T \geq x_2^2 &= \frac{\left(\tau + \sqrt{\tau^2 + 4\epsilon\kappa}\right)^2}{4\epsilon^2} \\ &= \frac{\left(\sqrt{V_{max} \ln\left(\frac{nN}{\delta}\right)}\bar{\beta} + \sqrt{V_{max} \ln\left(\frac{nN}{\delta}\right)\bar{\beta}^2 + 4\epsilon\sqrt{2V_{max}\gamma}\bar{\beta}}\right)^2}{4\epsilon^2}. \end{aligned} \tag{4.28}$$

□

4.4 Complexity Analysis and Comparison

We derived theroretical bounds on the number of iteration T required for the speedy Q-learning algorithm to achieve a specified precision ϵ in the Cramér distance between the

estimated and true return distributions. In addition to that we analyzed the convergence properties under the Maximal Hoeffding–Azuma inequality, the Bernstein inequality, and the Bennett inequality, resulting in progressively tighter bounds as established in Lemma 4.0.1.

In this subsection, we aim to confirm our theoretical findings with numerical simulations, by plotting the required number of updates T against varying levels of the precision ϵ and varying levels of complexity of the environment $n = \|\mathcal{X} \times \mathcal{A}\|$.

The code for the plots can be found on GitHub.

4.4.1 Plotting the Number of Updates T against the Precision ϵ

We start by examining how the required number of updates T varies with the desired precision ϵ for each of the convergence bounds. By fixing the complexity of the environment, we can isolate the impact of precision on the convergence rate of the algorithm under different inequalities.

To accurately reflect the framework, we integrate the specified parameters and conditions from our assumptions 3.0.1 into the numerical simulation as follows:

- The maximal reward R_{max} is set to 1, satisfying $R_{max} > 0$.
- V_{max} is calculated as $R_{max}/(1-\gamma) = 10$.
- The discount factor is set at $\gamma = 0.9$, which is considered a standard choice in reinforcement learning literature.
- The complexity of the environment n is set to 3000, reflecting the complexity of the taxi environment as described by [Die00].
- The number of atoms of the categorical distributions N is fixed at 51 in accordance with the C51 algorithm detailed in [BDM17].
- $\bar{\beta}$ is calculated using $\bar{\beta} = 1/(1-\sqrt{\gamma})$.
- δ is set to 0.05 to represent a 95% confidence lvel.

We implement the convergence bounds derived from the Hoeffding inequality, the Bernstein inequality, and the Bennett inequality as Python functions. The simulations involves determining the required number of updates T for achieving a specified precision level ϵ, where ϵ ranges between 0.01 and 0.1.

The precision parameter ϵ is plotted on a logarithmic scale to enhance the visibility of the differences between the convergence bounds, escpecially between $\bar{l}_2^H(\eta_C, \eta_T)$ and $\bar{l}_2^{Be}(\eta_C, \eta_T)$.

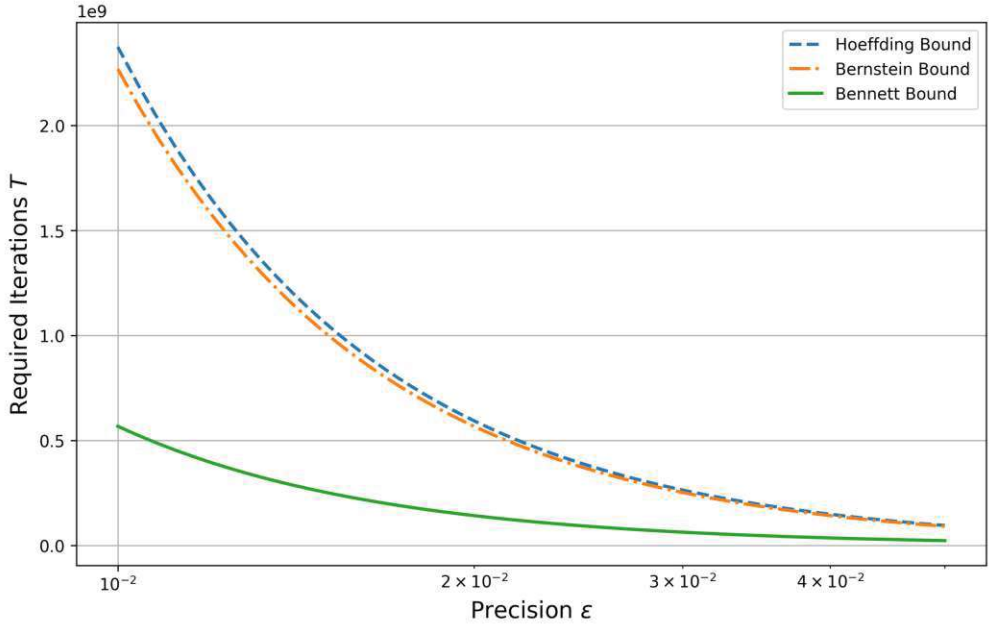

Figure 4.1: Updates vs precision requirement

We observe that $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) requires significantly fewer updates compared to the other bounds to achieve the same level of precision. This advantage becomes even more significant as ϵ decreases, which suggests that $\bar{l}_2^{Be}(\eta_C, \eta_T)$ is more effective in high-precision scenarios. This makes the Bennett bound a more practical choice when minimizing the number of updates T is crucial, especially in cases where reducing computational cost is a priority.

Next, we plot the difference in the number of updates required by $\bar{l}_2^{H}(\eta_C, \eta_T)$ (Hoeffding bound) and $\bar{l}_2^{B}(\eta_C, \eta_T)$ (Bernstein bound) to more closely examine their differences. This allows us to better understand the performance gap between the two bounds across varying levels of precision.

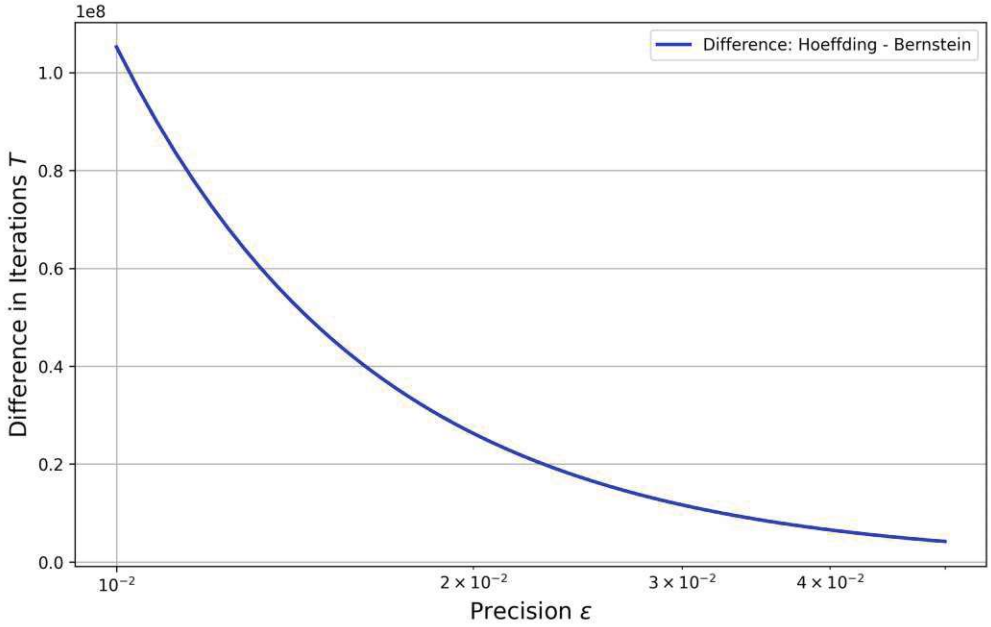

Figure 4.2: Updates vs environment complexity, difference between Hoeffding bound and Bernstein bound

As we observe, the difference decreases as the precision increases, meaning the performance gap between the two bounds narrows for lower precision requirements.

Overall, it can be concluded that both $\overline{l}_2^B(\eta_C, \eta_T)$ (Bernstein bound) and $\overline{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) outperform $\overline{l}_2^H(\eta_C, \eta_T)$ (Hoeffding bound), particularly in high-precision scenarios.

This demonstrates the superiority of these bounds when it is crucial to minimize the number of updates while maintaining a high degree of precision.

4.4.2 Plotting the Number of Updates T against the Complexity of the Environment n

Next, we investigate how the required number of updates T scales with the complexity of the environment $n = \|\mathcal{X} \times \mathcal{A}\|$, for a fixed precision ϵ.

Our goal is to assess the scalability of the algorithm in more complex environments.

We utilize mostly the same parameters as in the numerical simulation with the exception that we now fixate the value of ϵ at 0.1 and vary the complexity of the environment n between 10 and 1×10^6.

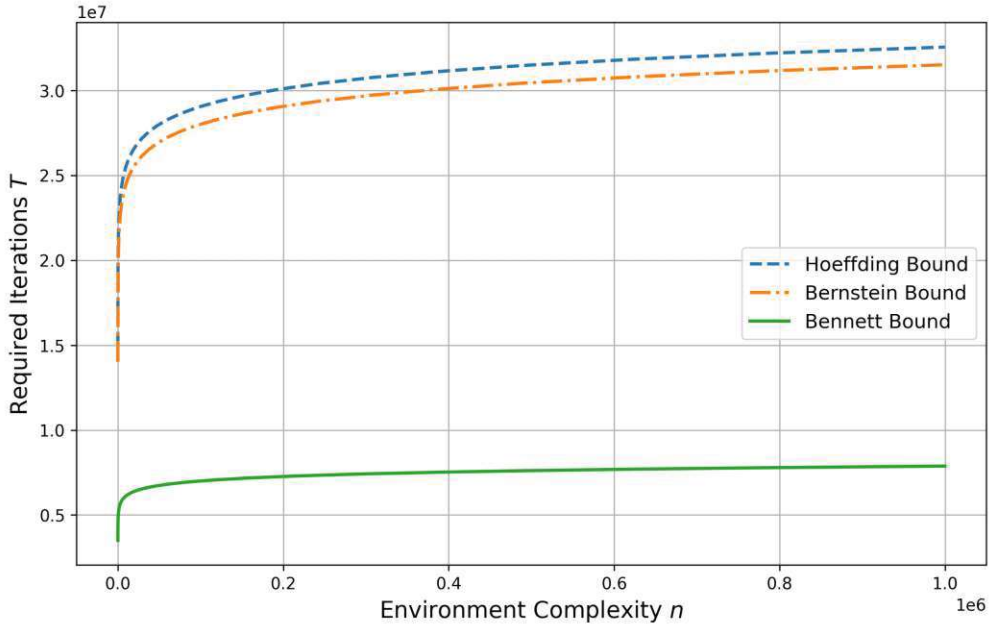

Figure 4.3: Updates vs environment complexity

Similar to the previous results in 4.4, we can observe that $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) consistently outperforms the other bounds. Additionally, we can investigate that $\bar{l}_2^{B}(\eta_C, \eta_T)$ (Bernstein bound) needs less updates across increasing environment complexity than $\bar{l}_2^{H}(\eta_C, \eta_T)$ (Hoeffding bound).

To illustrate how steep the number of updates increases with higher environment complexity we employ a semi-logarithmic scale (scaled x-axis).

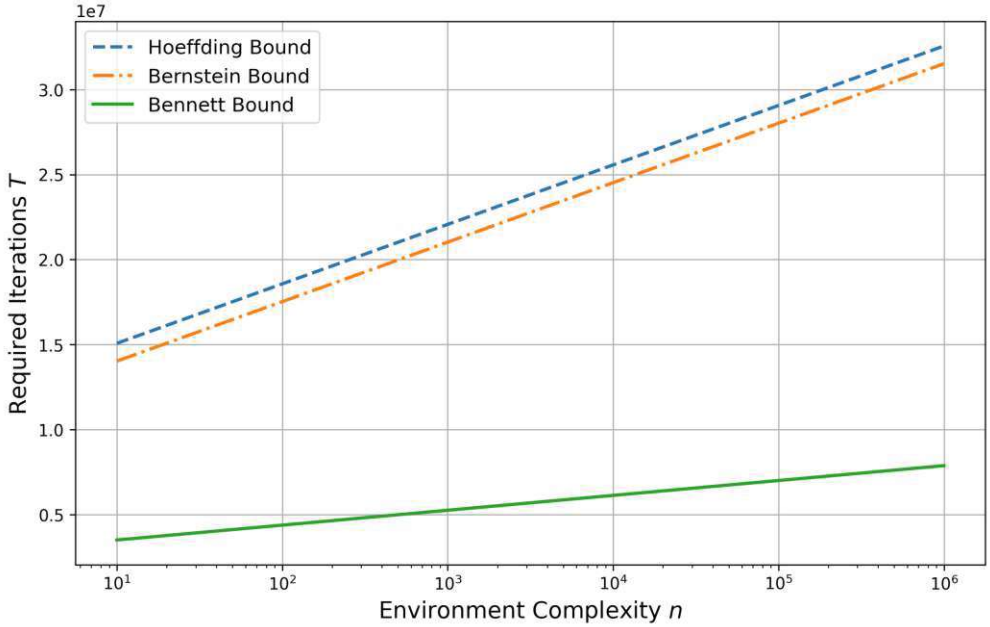

Figure 4.4: Updates vs environment complexity with semi-logarithmic scale

It is clear that the divergence between $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) and the other bounds grows with increasing environmental complexity.

This further underlines the benefits of the Bennett's bound, making it especially suitable for complex scenarios requiring precision while also being computationally efficient.

4.4.3 Summary of Findings

The numerical analysis confirms that $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) provides the tightest bound among the bounds defined in 3. It requires significantly fewer total updates to achieve the specified precision, especially in high-precision scenarios and complex environments. The Bernstein bound also outperforms the Hoeffding bound but to a lesser extent than the Bennett bound.

These results illustrate the benefit of employing the Bennett's inequality to optimize the convergence rate of the speedy Q-learning algorithm to reduce computational expenses.

CHAPTER 5

Experimental Evaluation

In this chapter we want to analyse and compare the bounds defined in 3 under the given assumption 3.0.1 using Gymnasium environments [TTK+23] to validate the performance of these bounds in practical reinforcement learning settings.

For this experiment we select two distinct environments from the Gymnasium library: *Taxi-v3* and *FrozenLake-v1*. We choose these environments due to their compatability with our assumptions and both having a deterministic action-state space, as they are both grid-based problems. This means that the environment is represented as a grid of tiles. Each tile represents a state, and the agent can navigate through the grid by selecting actions and moving from one state to another.

Additionally, these environments vary in complexity, with the *FrozenLake-v1* environment introducing a stochastic element, allowing us to evaluate the bounds under varying conditions.

Each environment undergoes training for 1000 episodes, with each episode being capped at a maximum of 1000 steps. The code can be found on GitHub.

5.1 Taxi Environment

In the *Taxi-v3* environment the agent must navigate a gridworld to pick up and drop off passengers. The taxi can be in any of 25 distinct positions in the 5×5 grid world, while the passengers can start from one of five specific locations and have one of four designated destinations. Together with the 6 possible actions available to an agent (moving north, south, east, or west, picking up the passenger, dropping off the passenger), this results in a state-action space of 3000.

Successfully picking up and dropping off a passengers results in a positive reward. The agents receives a negative reward for each time step it takes in order to achieve the fastest routes possible, as well as for incorrectly picking up and or dropping off passengers.

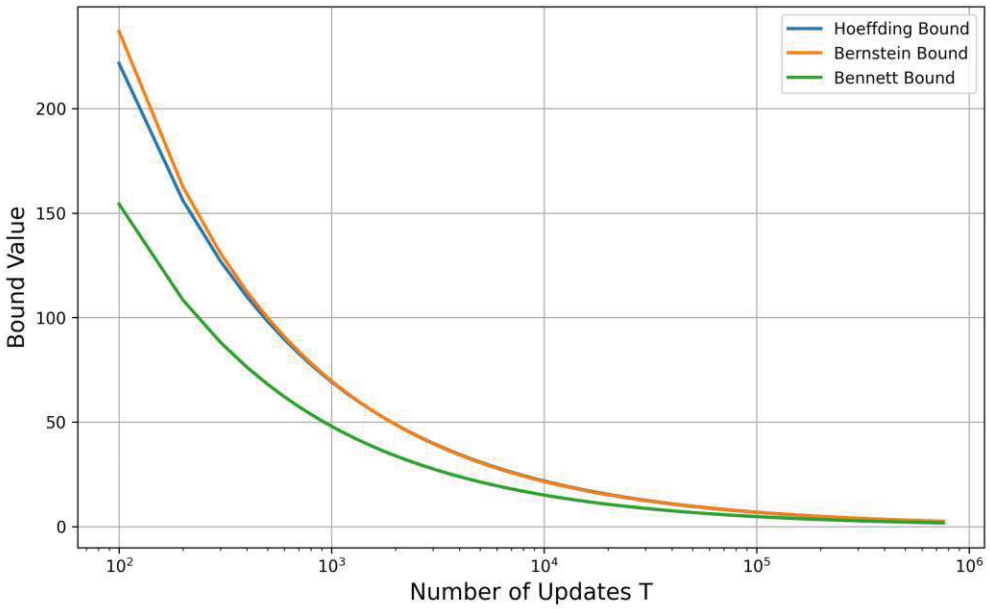

Figure 5.1: Bounds comparison for the *Taxi-v3* environment

We can investigate in figure 5.1 that $\overline{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) clearly performs best, however the difference between the bounds shrinks as the number of updates increases.

Additionally we can investigate that for a lower number of updates $\overline{l}_2^{H}(\eta_C, \eta_T)$ (Hoeffding bound) outperforms $\overline{l}_2^{B}(\eta_C, \eta_T)$ (Bernstein bound), which further highlights our findings from 4.0.1, where we showed that $\overline{l}_2^{B}(\eta_C, \eta_T) \leq \overline{l}_2^{H}(\eta_C, \eta_T)$ holds for $\ln^2(nN/\delta)/18\ln(2nN/\delta) \leq T$, while $\overline{l}_2^{Be}(\eta_C, \eta_T) \leq \overline{l}_2^{H}(\eta_C, \eta_T)$ for all $T > 0$.

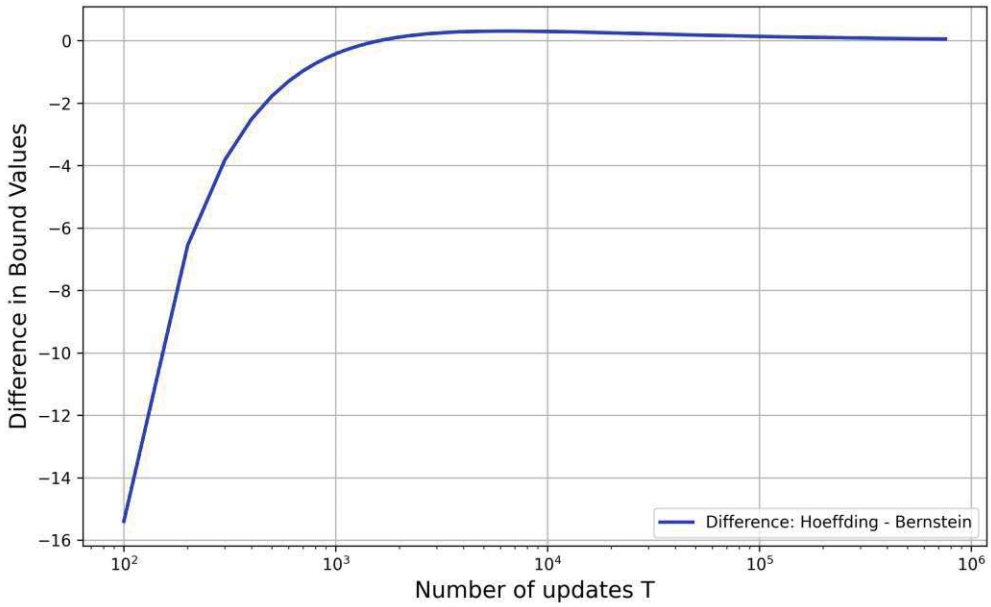

Figure 5.2: Difference between Hoeffding and Bernstein bound for the *Taxi-v3* environment

Evaluating the expression $\ln^2(51 \times 3000/0.05)/18 \ln(2 \times 51 \times 3000/0.05)$ yields approximately 0.8. This result indicates that each distribution $\eta^{(x,a)}$ must be updated around 0.8 times. This finding is supported by the total number of updates ranging between 1×10^3 and 1×10^4 in our experiment, before $\overline{l}_2^B(\eta_C, \eta_T) \leq \overline{l}_2^H(\eta_C, \eta_T)$ holds true, considering the size of the state-action space of 3000.

5.2 FrozenLake Environment

The *FrozenLake-v1* environment is a stochastic gridworld environment, where the agent must navigate across a frozen lake (4×4 tiles big) to reach a goal. Each tile is either frozen (safe) or a hole (terminal state). The agent can choose from four different actions: moving left, right, up, or down, making the state-action space equal to 64. Furthermore, there is a 33% probability that the agent will slip and move in an unintended direction upon taking an action. The reward function is quite simple: the agent earns a reward of 1 for reaching the goal and 0 otherwise. Moreover, the episode terminates if the agent falls into a hole.

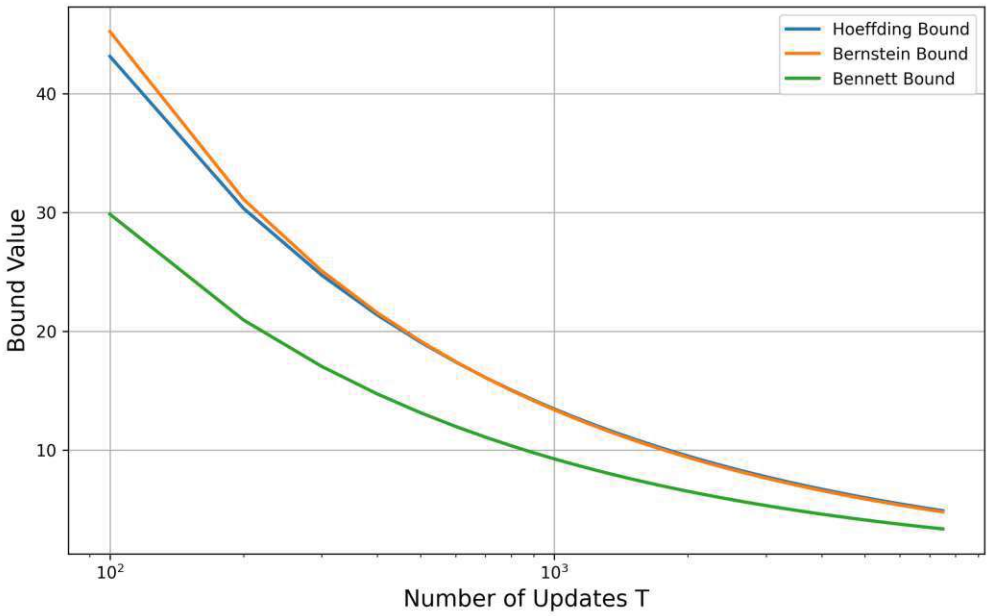

Figure 5.3: Bounds Ccmparison for the *FrozenLake-v1* Environment

This experiment yields the same results as the experiment above and further highlights them.

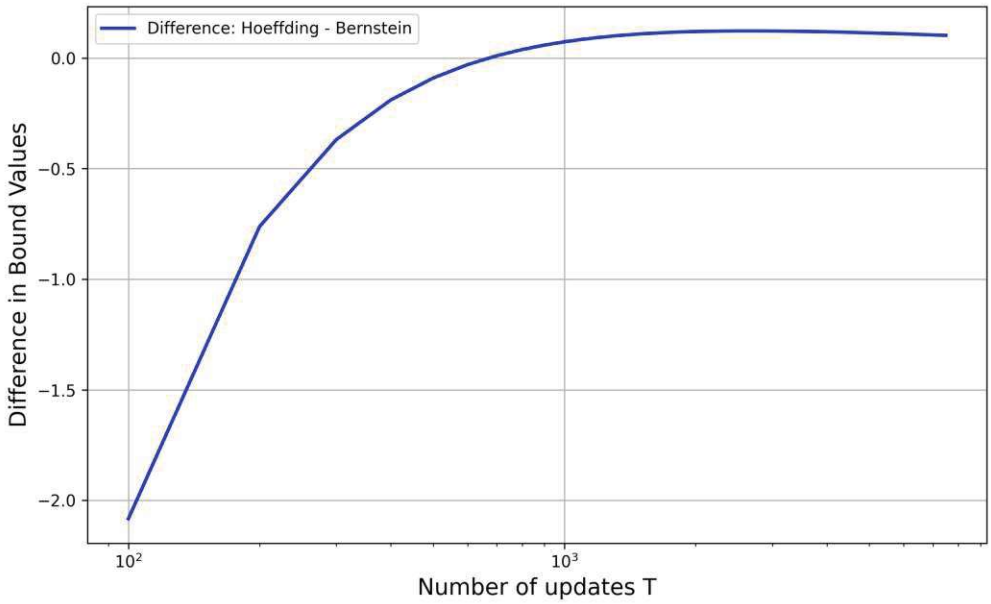

Figure 5.4: Difference between Hoeffding and Bernstein bound for the *FrozenLake-v1* Environment

Due to the smaller state-action space of the *FrozenLake-v1* environment we can observe that the inequality $\overline{l}_2^B(\eta_C, \eta_T) \leq \overline{l}_2^H(\eta_C, \eta_T)$ already holds for the number of total updates being $< 1 \times 10^3$.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we investigated the application of different concentration inequalities, specifically the Hoeffding, Bernstein, and Bennett inequalities, to enhance the analysis of the speedy Q-learning algorithm within the framework of distributional reinforcement learning (RL).

We demonstrated in 4.0.1 that the bound on the Cramér distance obtained by using the Bennett inequality is the tightest, followed by the Bernstein inequality. Specifically, the Bennett bound was shown to be tighter than the Hoeffding bound for all number of updates $T > 0$, while the Bernstein bound becomes tighter when the number of updates T exceeds $\ln^2(nN/\delta)/18\ln(2nN/\delta)$.

In the complexity analysis, we examined how the number of iterations T necessary for the algorithm to reach a given precision ϵ varies with differing levels of precision and environmental complexity $n = \|\mathcal{X} \times \mathcal{A}\|$. Those simulations demonstrated that $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) consistently requires fewer iterations compared to the $\bar{l}_2^{B}(\eta_C, \eta_T)$ (Bernstein bound) and $\bar{l}_2^{B}(\eta_C, \eta_T)$ (Hoeffding bounds). This was even more significant in scenarios with higher required precision and in environments with greater environmental complexity. The results also indicated a lower rise in required iterations as environmental complexity increased for $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound), suggesting that the gap between the bounds widens with increasing environmental complexity.

In the experimental evaluations using the Gymnasium environments *Taxi-v3* and *FrozenLake-v1* the $\bar{l}_2^{Be}(\eta_C, \eta_T)$ (Bennett bound) also performed best.

6.2 Future Work

Future work could explore applying these concentration inequalities to other distributional reinforcement learning algorithms. In addition to that the analysis could also be extended to environments with continuous state-action spaces.

Moreover, additional concentration inequalities could be assessed to determine if even tighter bounds on the Cramér distance can be found.

Developing adaptive algorithms that dynamically select the most appropriate bound on the Cramér distance based on the specific characteristics of the environment or the data observed during the learning process could also be an interesting topic for future research.

List of Figures

List of Algorithms

Bibliography

[AMGK11] Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy q-learning. In *Advances in neural information processing systems*, 2011.

[BDM17] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

[Bel66] Richard Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

[Ben62] George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

[Ber46] Sergei N. Bernstein. The theory of probabilities. *Annals of the Moscow University*, 1946.

[BH22] Markus Böck and Clemens Heitzinger. Speedy categorical distributional reinforcement learning and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 4(2):675–693, 2022.

[CBL06] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge university press, 2006.

[Cra28] Harald Cramér. On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

[Die00] Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

[EDMB03] Eyal Even-Dar, Yishay Mansour, and Peter Bartlett. Learning rates for q-learning. *Journal of machine learning Research*, 5(1), 2003.

[Fre75] David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118, 1975.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *The collected works of Wassily Hoeffding*, pages 409–426, 1994.

[LBC19] Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

[Pol] David Pollard. *Probability Tools, Tricks, and Miracles*. Unpublished. The utilized proof can be found on `https://web.archive.org/web/20200109010923/http://www.stat.yale.edu/~pollard/Books/Mini/BasicMG.pdf`.

[RBD+18a] Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, 09–11 Apr 2018.

[RBD+18b] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[SB18] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[Tsi94] John N Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine learning*, 16:185–202, 1994.

[TTK+23] Mark Towers, Jordan K. Terry, Ariel Kwiatkowski, John U. Balis, Gianluca de Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Arjun KG, Markus Krimmel, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Andrew Tan Jin Shen, and Omar G. Younis. Gymnasium, March 2023.

[Wat89] Christopher John Cornish Hellaby Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge United Kingdom, 1989.

[WD92] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.