

---

# A Distributional Perspective on Reinforcement Learning

---

Marc G. Bellemare<sup>\*1</sup> Will Dabney<sup>\*1</sup> Rémi Munos<sup>1</sup>

## Abstract

In this paper we argue for the fundamental importance of the *value distribution*: the distribution of the random return received by a reinforcement learning agent. This is in contrast to the common approach to reinforcement learning which models the expectation of this return, or *value*. Although there is an established body of literature studying the value distribution, thus far it has always been used for a specific purpose such as implementing risk-aware behaviour. We begin with theoretical results in both the policy evaluation and control settings, exposing a significant distributional instability in the latter. We then use the distributional perspective to design a new algorithm which applies Bellman’s equation to the learning of approximate value distributions. We evaluate our algorithm using the suite of games from the Arcade Learning Environment. We obtain both state-of-the-art results and anecdotal evidence demonstrating the importance of the value distribution in approximate reinforcement learning. Finally, we combine theoretical and empirical evidence to highlight the ways in which the value distribution impacts learning in the approximate setting.

## 1. Introduction

One of the major tenets of reinforcement learning states that, when not otherwise constrained in its behaviour, an agent should aim to maximize its expected utility  $Q$ , or *value* (Sutton & Barto, 1998). Bellman’s equation succinctly describes this value in terms of the expected reward and expected outcome of the random transition  $(x, a) \rightarrow (X', A')$ :

$$Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E} Q(X', A').$$

In this paper, we aim to go beyond the notion of value and argue in favour of a distributional perspective on reinforcement

learning. Specifically, the main object of our study is the random return  $Z$  whose expectation is the value  $Q$ . This random return is also described by a recursive equation, but one of a distributional nature:

$$Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A').$$

The *distributional Bellman equation* states that the distribution of  $Z$  is characterized by the interaction of three random variables: the reward  $R$ , the next state-action  $(X', A')$ , and its random return  $Z(X', A')$ . By analogy with the well-known case, we call this quantity the *value distribution*.

Although the distributional perspective is almost as old as Bellman’s equation itself (Jaquette, 1973; Sobel, 1982; White, 1988), in reinforcement learning it has thus far been subordinated to specific purposes: to model parametric uncertainty (Dearden et al., 1998), to design risk-sensitive algorithms (Morimura et al., 2010b;a), or for theoretical analysis (Azar et al., 2012; Lattimore & Hutter, 2012). By contrast, we believe the value distribution has a central role to play in reinforcement learning.

### Contraction of the policy evaluation Bellman operator.

Basing ourselves on results by Rösler (1992) we show that, for a fixed policy, the Bellman operator over value distributions is a contraction in a maximal form of the Wasserstein (also called Kantorovich or Mallows) metric. Our particular choice of metric matters: the same operator is not a contraction in total variation, Kullback-Leibler divergence, or Kolmogorov distance.

**Instability in the control setting.** We will demonstrate an instability in the distributional version of Bellman’s optimality equation, in contrast to the policy evaluation case. Specifically, although the optimality operator is a contraction in expected value (matching the usual optimality result), it is not a contraction in any metric over distributions. These results provide evidence in favour of learning algorithms that model the effects of nonstationary policies.

**Better approximations.** From an algorithmic standpoint, there are many benefits to learning an approximate distribution rather than its approximate expectation. The distributional Bellman operator preserves multimodality in value distributions, which we believe leads to more stable learning. Approximating the full distribution also mitigates the effects of learning from a nonstationary policy. As a whole,

---

<sup>\*</sup>Equal contribution <sup>1</sup>DeepMind, London, UK. Correspondence to: Marc G. Bellemare <bellemare@google.com>.

# 一种分布视角下的强化学习

Marc G. Bellemare<sup>\*1</sup> Will Dabney<sup>\*1</sup> Rémi Munos<sup>1</sup>

## 摘要

在本文中，我们强调了 *value distribution* 的基本重要性：强化学习代理接收到的随机回报的分布。这与常见的强化学习方法形成对比，后者建模的是这种回报的期望值，或 *value*。尽管已经有一系列研究价值分布的文献，但迄今为止，它总是用于特定目的，例如实现风险意识行为。我们从策略评估和控制的理论结果开始，揭示了后者中的显著分布不稳定性。然后，我们从分布的角度设计了一个新算法，该算法将贝尔曼方程应用于近似价值分布的学习。我们使用 Arcade Learning Environment 中的游戏套件来评估我们的算法。我们获得了最先进的结果，并提供了轶事证据，证明了在近似强化学习中价值分布的重要性。最后，我们结合理论和实证证据，突出了价值分布在近似设置中影响学习的方式。

## 1. 引言

强化学习的一个主要原则是，当行为不受其他约束时，智能体应该力求最大化其预期效用  $Q$ ，或 *value* (Sutton & Barto, 1998)。贝尔曼方程简洁地描述了这一价值，即预期奖励和随机转换的预期结果  $(x, a) \rightarrow (X', A')$ ：

$$Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E} Q(X', A').$$

在本文中，我们旨在超越价值的概念，并从分布性的角度出发论证强化学习的观点。

<sup>\*</sup>Equal contribution <sup>1</sup>DeepMind, London, UK. Correspondence to: Marc G. Bellemare <bellemare@google.com>.

机器学习。具体来说，我们研究的主要对象是随机回报  $Z$ ，其期望值是价值  $Q$ 。这种随机回报还通过一个分布性的递归方程来描述：

$$Z(x, a) \stackrel{D}{=} R(x, a) + \gamma Z(X', A').$$

The *distributional Bellman equation* 表示  $Z$  的分布由三个随机变量的交互作用来表征：奖励  $R$ 、下一个状态-动作  $(X', A')$  以及其随机回报  $Z(X', A')$ 。类比于一个熟知的情况，我们将这个量称为 *value distribution*。

尽管分布视角几乎和贝尔曼方程本身一样古老 (Jaquette, 1973; Sobel, 1982; White, 1988)，在强化学习中，它迄今为止一直被从属于特定目的：用于建模参数不确定性 (Dearden 等, 1998)，设计风险敏感算法 (Morimura 等, 2010b;a)，或者进行理论分析 (Azar 等, 2012; Lattimore & Hutter, 2012)。相比之下，我们认为价值分布在强化学习中应该扮演核心角色。

收缩策略评估贝尔曼算子。基于 Rösler (1992) 的结果，我们证明，在固定策略的情况下，贝尔曼算子对于价值分布是收缩的，这是在 Wasserstein（也称为 Kantorovich 或 Mallows）度量的最大形式下。我们特别选择的度量很重要：同样的算子在总变差、Kullback-Leibler 散度或 Kolmogorov 距离下不是收缩的。

控制设置中的不稳定性。我们将展示贝尔曼最优方程在分布版本中的不稳定性，与策略评估情况不同。具体来说，尽管最优性算子在期望值意义上是一个压缩映射（符合通常的最优性结果），但它在任何分布度量下都不是压缩映射。这些结果支持了建模非稳态策略影响的学习算法。

更好的近似方法。从算法的角度来看，学习一个近似分布而不是其近似期望有许多好处。分布性的贝尔曼算子保留了价值分布的多模态性，我们认为这会导致更稳定的学习。近似整个分布还能减轻从非稳态策略中学习的影响。作为一个整体，

we argue that this approach makes approximate reinforcement learning significantly better behaved.

We will illustrate the practical benefits of the distributional perspective in the context of the Arcade Learning Environment (Bellemare et al., 2013). By modelling the value distribution within a DQN agent (Mnih et al., 2015), we obtain considerably increased performance across the gamut of benchmark Atari 2600 games, and in fact achieve state-of-the-art performance on a number of games. Our results echo those of Veness et al. (2015), who obtained extremely fast learning by predicting Monte Carlo returns.

From a supervised learning perspective, learning the full value distribution might seem obvious: why restrict ourselves to the mean? The main distinction, of course, is that in our setting there are no given targets. Instead, we use Bellman’s equation to make the learning process tractable; we must, as Sutton & Barto (1998) put it, “learn a guess from a guess”. It is our belief that this guesswork ultimately carries more benefits than costs.

## 2. Setting

We consider an agent interacting with an environment in the standard fashion: at each step, the agent selects an action based on its current state, to which the environment responds with a reward and the next state. We model this interaction as a time-homogeneous Markov Decision Process  $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$ . As usual,  $\mathcal{X}$  and  $\mathcal{A}$  are respectively the state and action spaces,  $P$  is the transition kernel  $P(\cdot | x, a)$ ,  $\gamma \in [0, 1]$  is the discount factor, and  $R$  is the reward function, which in this work we explicitly treat as a random variable. A stationary policy  $\pi$  maps each state  $x \in \mathcal{X}$  to a probability distribution over the action space  $\mathcal{A}$ .

### 2.1. Bellman’s Equations

The *return*  $Z^\pi$  is the sum of discounted rewards along the agent’s trajectory of interactions with the environment. The value function  $Q^\pi$  of a policy  $\pi$  describes the expected return from taking action  $a \in \mathcal{A}$  from state  $x \in \mathcal{X}$ , then acting according to  $\pi$ :

$$Q^\pi(x, a) := \mathbb{E} Z^\pi(x, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \right], \quad (1)$$

$$x_t \sim P(\cdot | x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot | x_t), x_0 = x, a_0 = a.$$

Fundamental to reinforcement learning is the use of Bellman’s equation (Bellman, 1957) to describe the value function:

$$Q^\pi(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_{P, \pi} Q^\pi(x', a').$$

In reinforcement learning we are typically interested in acting so as to maximize the return. The most common ap-

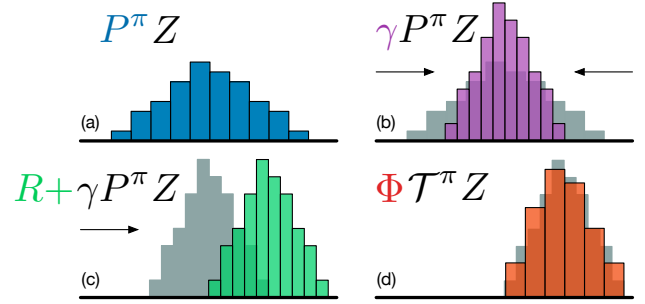


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy  $\pi$ , (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

proach for doing so involves the optimality equation

$$Q^*(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q^*(x', a').$$

This equation has a unique fixed point  $Q^*$ , the optimal value function, corresponding to the set of optimal policies  $\Pi^*$  ( $\pi^*$  is optimal if  $\mathbb{E}_{a \sim \pi^*} Q^*(x, a) = \max_a Q^*(x, a)$ ).

We view value functions as vectors in  $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ , and the expected reward function as one such vector. In this context, the *Bellman operator*  $\mathcal{T}^\pi$  and *optimality operator*  $\mathcal{T}$  are

$$\mathcal{T}^\pi Q(x, a) := \mathbb{E} R(x, a) + \gamma \mathbb{E}_{P, \pi} Q(x', a') \quad (2)$$

$$\mathcal{T} Q(x, a) := \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (3)$$

These operators are useful as they describe the expected behaviour of popular learning algorithms such as SARSA and Q-Learning. In particular they are both contraction mappings, and their repeated application to some initial  $Q_0$  converges exponentially to  $Q^\pi$  or  $Q^*$ , respectively (Bertsekas & Tsitsiklis, 1996).

## 3. The Distributional Bellman Operators

In this paper we take away the expectations inside Bellman’s equations and consider instead the full distribution of the random variable  $Z^\pi$ . From here on, we will view  $Z^\pi$  as a mapping from state-action pairs to distributions over returns, and call it the *value distribution*.

Our first aim is to gain an understanding of the theoretical behaviour of the distributional analogues of the Bellman operators, in particular in the less well-understood control setting. The reader strictly interested in the algorithmic contribution may choose to skip this section.

### 3.1. Distributional Equations

It will sometimes be convenient to make use of the probability space  $(\Omega, \mathcal{F}, \text{Pr})$ . The reader unfamiliar with mea-

我们argue认为这种approach方法使得approximate近似的reinforcement强化learning学习significantly显著地better更好地behaved表现良好。

我们将通过 Arcade Learning Environment (Bellemare 等, 2013) 的背景来说明分布视角的实际益处。通过在 DQN 代理 (Mnih 等, 2015) 中建模价值分布, 我们获得了在基准 Atari 2600 游戏范围内的显著性能提升, 并且实际上在一些游戏中达到了最先进的性能。我们的结果与 Veness 等 (2015) 的研究相呼应, 他们通过预测蒙特卡洛回报获得了非常快速的学习。

从监督学习的角度来看, 学习完整的价值分布似乎是显而易见的: 为什么要局限于均值? 当然, 主要的区别在于, 在我们的设置中没有给定的目标。相反, 我们使用贝尔曼方程使学习过程变得可行; 正如辛顿和巴托 (1998) 所说, 我们必须 “从猜测中学出一个猜测”。我们认为, 这种猜测最终带来的益处超过了成本。

## 2. 设置

我们考虑一个代理以标准方式与环境交互: 在每一步中, 代理根据其当前状态选择一个动作, 环境则对此作出响应, 给出一个奖励和下一个状态。我们将这种交互建模为时间齐次马尔可夫决策过程  $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$ 。通常,  $\mathcal{X}$  和  $\mathcal{A}$  分别是状态空间和动作空间,  $P$  是转移核  $P(\cdot | x, a)$ ,  $\gamma \in [0, 1]$  是折扣因子, 而  $R$  是奖励函数, 在这项工作中, 我们明确将其视为一个随机变量。一个平稳策略  $\pi$  将每个状态  $x \in \mathcal{X}$  映射到动作空间  $\mathcal{A}$  的一个概率分布上。

### 2.1. 贝尔曼方程

The return  $Z^\pi$  是沿代理与环境交互轨迹的折扣奖励之和。策略  $Q^\pi$  的价值函数  $\pi$  描述了从状态  $x \in \mathcal{X}$  执行动作  $a \in \mathcal{A}$  然后根据  $\pi$  行动的预期回报:

$$Q^\pi(x, a) := \mathbb{E} Z^\pi(x, a) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t R(x_t, a_t) \right], \quad (1)$$

$$x_t \sim P(\cdot | x_{t-1}, a_{t-1}), a_t \sim \pi(\cdot | x_t), x_0 = x, a_0 = a.$$

强化学习的基本原理是使用贝尔曼方程 (Bellman, 1957) 来描述价值函数:

$$Q^\pi(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_{P, \pi} Q^\pi(x', a').$$

在强化学习中, 我们通常希望行为能够最大化回报。最常见的方法-

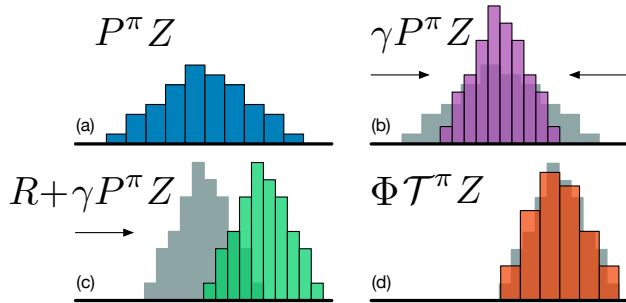


图1. 基于确定性奖励函数的分布贝尔曼算子: (a) 在策略  $\pi$  下的下一个状态分布, (b) 折扣会将分布向0收缩, (c) 奖励将其移动, (d) 投影步骤 (第4节)。

approach for doing so涉及最优方程

$$Q^*(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q^*(x', a').$$

该方程有一个唯一的不动点  $\{v^*_{23}\}$ , 即最优值函数, 对应的最优策略集  $\{v^*_{24}\}$  是最优的, 如果  $\{v^*_{25}\} \max \{v^*_{26}\}$ 。

我们视价值函数为  $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$  中的向量, 并将期望奖励函数视为其中一个向量。在此上下文中, Bellman operator  $\mathcal{T}^\pi$  和 optimality operator  $\mathcal{T}$  是

$$\mathcal{T}^\pi Q(x, a) := \mathbb{E} R(x, a) + \gamma \mathbb{E}_{P, \pi} Q(x', a') \quad (2)$$

$$\mathcal{T} Q(x, a) := \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (3)$$

这些运算符很有用, 因为它们描述了诸如SARSA和Q-Learning等流行学习算法的预期行为。特别是, 它们都是收缩映射, 并且将某些初始的  $Q_0$  反复应用后, 会以指数方式分别收敛到  $Q^\pi$  或  $Q^*$  (Bertsekas & Tsitsiklis, 1996)。

## 3. 分布式贝尔曼运算符

在本文中, 我们移除了贝尔曼方程中的期望, 并考虑了随机变量  $Z^\pi$  的完整分布。从这里开始, 我们将视  $Z^\pi$  为从状态-动作对到回报分布的映射, 并将其称为 value distribution。

我们的首要目标是理解贝尔曼算子的分布型类似物的理论行为, 特别是在控制设置中理解得较少的部分。严格对算法贡献感兴趣的读者可以选择跳过这一部分。

### 3.1. 分布方程

有时可以方便地使用概率空间  $(\Omega, \mathcal{F}, \text{Pr})$ 。不熟悉测度论的读者可以参考相关资料。

sure theory may think of  $\Omega$  as the space of all possible outcomes of an experiment (Billingsley, 1995). We will write  $\|\mathbf{u}\|_p$  to denote the  $L_p$  norm of a vector  $\mathbf{u} \in \mathbb{R}^{\mathcal{X}}$  for  $1 \leq p \leq \infty$ ; the same applies to vectors in  $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ . The  $L_p$  norm of a random vector  $U : \Omega \rightarrow \mathbb{R}^{\mathcal{X}}$  (or  $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ ) is then  $\|U\|_p := [\mathbb{E} [\|U(\omega)\|_p^p]]^{1/p}$ , and for  $p = \infty$  we have  $\|U\|_\infty = \text{ess sup } \|U(\omega)\|_\infty$  (we will omit the dependency on  $\omega \in \Omega$  whenever unambiguous). We will denote the c.d.f. of a random variable  $U$  by  $F_U(y) := \Pr\{U \leq y\}$ , and its inverse c.d.f. by  $F_U^{-1}(q) := \inf\{y : F_U(y) \geq q\}$ .

A distributional equation  $U \stackrel{D}{=} V$  indicates that the random variable  $U$  is distributed according to the same law as  $V$ . Without loss of generality, the reader can understand the two sides of a distributional equation as relating the distributions of two independent random variables. Distributional equations have been used in reinforcement learning by Engel et al. (2005); Morimura et al. (2010a) among others, and in operations research by White (1988).

### 3.2. The Wasserstein Metric

The main tool for our analysis is the Wasserstein metric  $d_p$  between cumulative distribution functions (see e.g. Bickel & Freedman, 1981, where it is called the Mallows metric). For  $F, G$  two c.d.f.s over the reals, it is defined as

$$d_p(F, G) := \inf_{U, V} \|U - V\|_p,$$

where the infimum is taken over all pairs of random variables  $(U, V)$  with respective cumulative distributions  $F$  and  $G$ . The infimum is attained by the inverse c.d.f. transform of a random variable  $\mathcal{U}$  uniformly distributed on  $[0, 1]$ :

$$d_p(F, G) = \|F^{-1}(\mathcal{U}) - G^{-1}(\mathcal{U})\|_p.$$

For  $p < \infty$  this is more explicitly written as

$$d_p(F, G) = \left( \int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p}.$$

Given two random variables  $U, V$  with c.d.f.s  $F_U, F_V$ , we will write  $d_p(U, V) := d_p(F_U, F_V)$ . We will find it convenient to conflate the random variables under consideration with their versions under the inf, writing

$$d_p(U, V) = \inf_{U, V} \|U - V\|_p.$$

whenever unambiguous; we believe the greater legibility justifies the technical inaccuracy. Finally, we extend this metric to vectors of random variables, such as value distributions, using the corresponding  $L_p$  norm.

Consider a scalar  $a$  and a random variable  $A$  independent

of  $U, V$ . The metric  $d_p$  has the following properties:

$$d_p(aU, aV) \leq |a| d_p(U, V) \quad (\text{P1})$$

$$d_p(A + U, A + V) \leq d_p(U, V) \quad (\text{P2})$$

$$d_p(AU, AV) \leq \|A\|_p d_p(U, V). \quad (\text{P3})$$

We will need the following additional property, which makes no independence assumptions on its variables. Its proof, and that of later results, is given in the appendix.

**Lemma 1** (Partition lemma). *Let  $A_1, A_2, \dots$  be a set of random variables describing a partition of  $\Omega$ , i.e.  $A_i(\omega) \in \{0, 1\}$  and for any  $\omega$  there is exactly one  $A_i$  with  $A_i(\omega) = 1$ . Let  $U, V$  be two random variables. Then*

$$d_p(U, V) \leq \sum_i d_p(A_i U, A_i V).$$

Let  $\mathcal{Z}$  denote the space of value distributions with bounded moments. For two value distributions  $Z_1, Z_2 \in \mathcal{Z}$  we will make use of a maximal form of the Wasserstein metric:

$$\bar{d}_p(Z_1, Z_2) := \sup_{x, a} d_p(Z_1(x, a), Z_2(x, a)).$$

We will use  $\bar{d}_p$  to establish the convergence of the distributional Bellman operators.

**Lemma 2.**  $\bar{d}_p$  is a metric over value distributions.

### 3.3. Policy Evaluation

In the *policy evaluation* setting (Sutton & Barto, 1998) we are interested in the value function  $V^\pi$  associated with a given policy  $\pi$ . The analogue here is the value distribution  $Z^\pi$ . In this section we characterize  $Z^\pi$  and study the behaviour of the policy evaluation operator  $\mathcal{T}^\pi$ . We emphasize that  $Z^\pi$  describes the intrinsic randomness of the agent's interactions with its environment, rather than some measure of uncertainty about the environment itself.

We view the reward function as a random vector  $R \in \mathcal{Z}$ , and define the transition operator  $P^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$

$$P^\pi Z(x, a) \stackrel{D}{=} Z(X', A') \quad (4)$$

$$X' \sim P(\cdot | x, a), \quad A' \sim \pi(\cdot | X'),$$

where we use capital letters to emphasize the random nature of the next state-action pair  $(X', A')$ . We define the distributional Bellman operator  $\mathcal{T}^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$  as

$$\mathcal{T}^\pi Z(x, a) \stackrel{D}{=} R(x, a) + \gamma P^\pi Z(x, a). \quad (5)$$

While  $\mathcal{T}^\pi$  bears a surface resemblance to the usual Bellman operator (2), it is fundamentally different. In particular, three sources of randomness define the compound distribution  $\mathcal{T}^\pi Z$ :



当然，理论中可以将  $\Omega$  看作是实验所有可能结果的空间 (Billingsley, 1995)。我们将用  $\|\mathbf{u}\|_p$  表示向量  $\mathbf{u} \in \mathbb{R}^X$  的  $L_p$  范数，其中  $1 \leq p \leq \infty$ ；同样的规则也适用于  $\mathbb{R}^{X \times A}$  中的向量。随机向量  $U: \Omega \rightarrow \mathbb{R}^X$  (或  $\mathbb{R}^{X \times A}$ ) 的  $L_p$  范数则是  $\|U\|_p := [\mathbb{E}[\|U(\omega)\|_p^p]]^{1/p}$ ，对于  $p = \infty$ ，我们有  $\|U\|_\infty = \text{ess sup } \|U(\omega)\|_\infty$  (我们将在无歧义时省略对  $\omega \in \Omega$  的依赖)。我们将用  $F_U(y) := \Pr\{U \leq y\}$  表示随机变量  $U$  的分布函数，其逆分布函数则表示为  $=: \{y \mid \inf F_U(y) \geq q\}$ 。

一个分布方程  $U \stackrel{D}{=} V$  表明随机变量  $U$  服从与  $V$  相同的分布律。不失一般性，读者可以理解分布方程的两边将两个独立随机变量的分布联系起来。分布方程已被 Engel 等人 (2005 年)、Morimura 等人 (2010 年) 以及其他研究者在强化学习中使用，并且在运筹学中被 White (1988 年) 使用。

### 3.2. 水舍夫距离

我们分析的主要工具是累积分布函数之间的 Wasserstein 度量  $d_p$  (参见例如 Bickel & Freedman, 1981，其中称为 Mallows 度量)。对于  $F$  和  $G$  两个实数上的累积分布函数，它定义为

$$d_p(F, G) := \inf_{U, V} \|U - V\|_p,$$

其中最小值是在所有具有各自的累积分布  $F$  和  $G$  的随机变量对  $(U, V)$  中取的。最小值由在  $[0, 1]$  上均匀分布的随机变量  $U$  的逆累积分布变换达到：

$$d_p(F, G) = \|F^{-1}(U) - G^{-1}(U)\|_p.$$

对于  $p < \infty$  这更明确地写为

$$d_p(F, G) = \left( \int_0^1 |F^{-1}(u) - G^{-1}(u)|^p du \right)^{1/p}.$$

给定两个随机变量  $U, V$ ，它们的分布函数分别为  $F_U, F_V$ ，我们将写成  $d_p(U, V) := d_p(F_U, F_V)$ 。我们将发现将考虑中的随机变量与其在  $\inf$  下的版本合并起来写会更方便，即

$$d_p(U, V) = \inf_{U, V} \|U - V\|_p.$$

每当没有歧义时；我们认为更好的可读性值得技术上的不准确。最后，我们将该指标扩展到随机变量的向量，例如价值分布，使用相应的  $L_p$  范数。

考虑一个标量  $a$  和一个与之独立的随机变量  $A$

of  $U, V$ 。度量  $d_p$  具有以下性质：

$$d_p(aU, aV) \leq |a| d_p(U, V) \quad (\text{P1})$$

$$d_p(A + U, A + V) \leq d_p(U, V) \quad (\text{P2})$$

$$d_p(AU, AV) \leq \|A\|_p d_p(U, V). \quad (\text{P3})$$

我们将需要一个额外的性质，它不对其变量做出独立性的假设。其证明以及后续结果的证明将在附录中给出。

引理 1 (分区引理). *Let  $A_1, A_2, \dots$  be a set of random variables describing a partition of  $\Omega$ , i.e.  $A_i(\omega) \in \{0, 1\}$  and for any  $\omega$  there is exactly one  $A_i$  with  $A_i(\omega) = 1$ . Let  $U, V$  be two random variables. Then*

$$d_p(U, V) \leq \sum_i d_p(A_i U, A_i V).$$

令  $\mathcal{Z}$  表示具有有界矩的价值分布的空间。对于两个价值分布  $Z_1, Z_2 \in \mathcal{Z}$ ，我们将使用 Wasserstein 距离的一种最大形式：

$$\bar{d}_p(Z_1, Z_2) := \sup_{x, a} d_p(Z_1(x, a), Z_2(x, a)).$$

我们将使用  $\bar{d}_p$  来建立分布型贝尔曼算子的收敛性。

引理 2.  $\bar{d}_p$  is a metric over value distributions.

### 3.3. 政策评估

在 *policy evaluation* 设置 (Sutton & Barto, 1998) 中，我们关注的是与给定策略  $\pi$  相关的价值函数  $V^\pi$ 。这里的类比是价值分布  $Z^\pi$ 。在本节中，我们刻画  $Z^\pi$  并研究策略评估算子  $\mathcal{T}^\pi$  的行为。我们强调  $Z^\pi$  描述的是智能体与其环境交互的固有随机性，而不是对环境本身的不确定性的一种度量。

我们将奖励函数视为随机向量  $R \in \mathcal{Z}$ ，并定义转换算子  $P^\pi: \mathcal{Z} \rightarrow \mathcal{Z}$

$$P^\pi Z(x, a) \stackrel{D}{=} Z(X', A') \quad (4)$$

$$X' \sim P(\cdot | x, a), A' \sim \pi(\cdot | X'),$$

其中我们使用大写字母来强调下一个状态-动作对的随机性  $(X', A')$ 。我们定义分布贝尔曼运算符  $\mathcal{T}^\pi: \mathcal{Z} \rightarrow \mathcal{Z}$  为

$$\mathcal{T}^\pi Z(x, a) \stackrel{D}{=} R(x, a) + \gamma P^\pi Z(x, a). \quad (5)$$

虽然  $\mathcal{T}^\pi$  在表面上类似于通常的 Bell-man 操作符 (2)，但它本质上是不同的。特别是，三种随机性定义了复合分布  $\mathcal{T}^\pi Z$ ：

- a) The randomness in the reward  $R$ ,
- b) The randomness in the transition  $P^\pi$ , and
- c) The next-state value distribution  $Z(X', A')$ .

In particular, we make the usual assumption that these three quantities are independent. In this section we will show that (5) is a contraction mapping whose unique fixed point is the random return  $Z^\pi$ .

### 3.3.1. CONTRACTION IN $\bar{d}_p$

Consider the process  $Z_{k+1} := \mathcal{T}^\pi Z_k$ , starting with some  $Z_0 \in \mathcal{Z}$ . We may expect the limiting expectation of  $\{Z_k\}$  to converge exponentially quickly, as usual, to  $Q^\pi$ . As we now show, the process converges in a stronger sense:  $\mathcal{T}^\pi$  is a contraction in  $\bar{d}_p$ , which implies that all moments also converge exponentially quickly.

**Lemma 3.**  $\mathcal{T}^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$  is a  $\gamma$ -contraction in  $\bar{d}_p$ .

Using Lemma 3, we conclude using Banach's fixed point theorem that  $\mathcal{T}^\pi$  has a unique fixed point. By inspection, this fixed point must be  $Z^\pi$  as defined in (1). As we assume all moments are bounded, this is sufficient to conclude that the sequence  $\{Z_k\}$  converges to  $Z^\pi$  in  $\bar{d}_p$  for  $1 \leq p \leq \infty$ .

To conclude, we remark that not all distributional metrics are equal; for example, Chung & Sobel (1987) have shown that  $\mathcal{T}^\pi$  is not a contraction in total variation distance. Similar results can be derived for the Kullback-Leibler divergence and the Kolmogorov distance.

### 3.3.2. CONTRACTION IN CENTERED MOMENTS

Observe that  $d_2(U, V)$  (and more generally,  $d_p$ ) relates to a coupling  $C(\omega) := U(\omega) - V(\omega)$ , in the sense that

$$d_2^2(U, V) \leq \mathbb{E}[(U - V)^2] = \mathbb{V}(C) + (\mathbb{E} C)^2.$$

As a result, we cannot directly use  $d_2$  to bound the variance difference  $|\mathbb{V}(\mathcal{T}^\pi Z(x, a)) - \mathbb{V}(Z^\pi(x, a))|$ . However,  $\mathcal{T}^\pi$  is in fact a contraction in variance (Sobel, 1982, see also appendix). In general,  $\mathcal{T}^\pi$  is not a contraction in the  $p^{\text{th}}$  centered moment,  $p > 2$ , but the centered moments of the iterates  $\{Z_k\}$  still converge exponentially quickly to those of  $Z^\pi$ ; the proof extends the result of Rösler (1992).

## 3.4. Control

Thus far we have considered a fixed policy  $\pi$ , and studied the behaviour of its associated operator  $\mathcal{T}^\pi$ . We now set out to understand the distributional operators of the *control* setting – where we seek a policy  $\pi$  that maximizes value – and the corresponding notion of an optimal value distribution. As with the optimal value function, this notion is intimately tied to that of an optimal policy. However, while all optimal policies attain the same value  $Q^*$ , in our case

a difficulty arises: in general there are many optimal value distributions.

In this section we show that the distributional analogue of the Bellman optimality operator converges, in a weak sense, to the set of optimal value distributions. However, this operator is not a contraction in any metric between distributions, and is in general much more temperamental than the policy evaluation operators. We believe the convergence issues we outline here are a symptom of the inherent instability of greedy updates, as highlighted by e.g. Tsitsiklis (2002) and most recently Harutyunyan et al. (2016).

Let  $\Pi^*$  be the set of optimal policies. We begin by characterizing what we mean by an *optimal value distribution*.

**Definition 1** (Optimal value distribution). *An optimal value distribution is the v.d. of an optimal policy. The set of optimal value distributions is  $\mathcal{Z}^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$ .*

We emphasize that not all value distributions with expectation  $Q^*$  are optimal: they must match the full distribution of the return under some optimal policy.

**Definition 2.** *A greedy policy  $\pi$  for  $Z \in \mathcal{Z}$  maximizes the expectation of  $Z$ . The set of greedy policies for  $Z$  is*

$$\mathcal{G}_Z := \{\pi : \sum_a \pi(a|x) \mathbb{E} Z(x, a) = \max_{a' \in \mathcal{A}} \mathbb{E} Z(x, a')\}.$$

Recall that the expected Bellman optimality operator  $\mathcal{T}$  is

$$\mathcal{T}Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (6)$$

The maximization at  $x'$  corresponds to some greedy policy. Although this policy is implicit in (6), we cannot ignore it in the distributional setting. We will call a *distributional Bellman optimality operator* any operator  $\mathcal{T}$  which implements a greedy selection rule, i.e.

$$\mathcal{T}Z = \mathcal{T}^\pi Z \text{ for some } \pi \in \mathcal{G}_Z.$$

As in the policy evaluation setting, we are interested in the behaviour of the iterates  $Z_{k+1} := \mathcal{T}Z_k$ ,  $Z_0 \in \mathcal{Z}$ . Our first result is to assert that  $\mathbb{E} Z_k$  behaves as expected.

**Lemma 4.** *Let  $Z_1, Z_2 \in \mathcal{Z}$ . Then*

$$\|\mathbb{E} \mathcal{T}Z_1 - \mathbb{E} \mathcal{T}Z_2\|_\infty \leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty,$$

*and in particular  $\mathbb{E} Z_k \rightarrow Q^*$  exponentially quickly.*

By inspecting Lemma 4, we might expect that  $Z_k$  converges quickly in  $\bar{d}_p$  to some fixed point in  $\mathcal{Z}^*$ . Unfortunately, convergence is neither quick nor assured to reach a fixed point. In fact, the best we can hope for is pointwise convergence, not even to the set  $\mathcal{Z}^*$  but to the larger set of *nonstationary optimal value distributions*.

a) 奖励的随机性  $R$ , b) 转移的随机性  $P^\pi$ , 和 c) 下一个状态值分布  $Z(X', A')$ 。

特别是在此, 我们做出通常假设, 这三个量是独立的。在本节中我们将证明(5)是一个收缩映射, 其唯一的不动点是随机返回量  $Z^\pi$ 。

### 3.3.1. 收缩在 $\bar{d}_p$

考虑过程  $Z_{k+1} := \mathcal{T}^\pi Z_k$ , 从某个  $Z_0 \in \mathcal{Z}$  开始。我们通常可以期望  $\{Z_k\}$  的极限期望以指数速度收敛, 如常。如我们现在所展示的, 该过程以更强的意义收敛:  $\mathcal{T}^\pi$  在  $\bar{d}_p$  中是压缩映射, 这表明所有矩也以指数速度收敛。

引理 3.  $\mathcal{T}^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$  is a  $\gamma$ -contraction in  $\bar{d}_p$ .

使用引理3和Banach的不动点定理, 我们得出结论  $\mathcal{T}^\pi$  有一个唯一的不动点。通过检查, 这个不动点必须是根据(1)定义的  $Z^\pi$ 。由于我们假设所有矩都是有界的, 这足以得出结论序列  $\{Z_k\}$  在  $\bar{d}_p$  中收敛到  $Z^\pi$ , 对于  $1 \leq p \leq \infty$ 。

总之, 我们注意到并非所有的分布度量都是等价的; 例如, Chung & Sobel (1987) 已经证明  $\mathcal{T}^\pi$  在总体变异距离下不是收缩的。类似的结果也可以推导出对于 KL 散度和柯尔莫哥洛夫距离。

### 3.3.2. 中心矩的收缩

观察到  $d_2(U, V)$  (并且更一般地,  $d_p$ ) 与耦合  $C(\omega)$ :  $= U(\omega) - V(\omega)$  有关, 即

$$d_2^2(U, V) \leq \mathbb{E}[(U - V)^2] = \mathbb{V}(C) + (\mathbb{E}C)^2.$$

因此, 我们不能直接使用  $d_2$  来界\_variance\_difference\_  $|\mathbb{V}(\mathcal{T}^\pi Z(x, a)) - \mathbb{V}(Z^\pi(x, a))|$ 。然而,  $\mathcal{T}^\pi$  实际上是一个方差收缩 (Sobel, 1982, 参见附录)。一般来说,  $\mathcal{T}^\pi$  并不是以  $p^{th}$  为中心的矩  $p > 2$  的收缩, 但迭代  $\{Z_k\}$  的以中心化后的矩仍然以指数速度快速收敛到  $Z^\pi$  的相应矩; 证明扩展了 Rösler (1992) 的结果。

### 3.4. 控制

迄今为止, 我们考虑了一个固定策略  $\pi$ , 并研究了其相关算子  $\mathcal{T}^\pi$  的行为。现在我们将着手理解 control 设置下的分布算子——在那里我们寻求一个最大化价值的策略  $\pi$ , 以及相应的最优价值分布的概念。就像最优价值函数一样, 这种概念与最优策略的概念密切相关。然而, 虽然所有最优策略都达到相同的值  $Q^*$ , 但在我们的情况下,

一个困难出现了: 一般来说, 存在许多最优值分布。

在本节中, 我们证明了贝尔曼最优算子的分布型类比在某种弱意义下收敛到最优值分布的集合。然而, 该算子在分布之间的任何度量下都不是收缩映射, 并且通常比策略评估算子更加不稳定。我们认为我们在这里概述的收敛问题是贪婪更新固有有不确定性的一种症状, 正如 Tsitsiklis (2002) 和最近 Harutyunyan 等人 (2016) 所指出的。

令  $\Pi^*$  为最优策略的集合。我们首先描述我们所指的 *optimal value distribution* 是什么。

定义 1 (最优值分布). An *optimal value distribution* is the v.d. of an optimal policy. The set of optimal value distributions is  $\mathcal{Z}^* := \{Z^{\pi^*} : \pi^* \in \Pi^*\}$ .

我们强调, 并非所有期望值为  $Q^*$  的价值分布都是最优的: 它们必须与某个最优策略下的回报完整分布相匹配。

定义 2. A *greedy policy*  $\pi$  for  $Z \in \mathcal{Z}$  maximizes the expectation of  $Z$ . The set of greedy policies for  $Z$  is

$$\mathcal{G}_Z := \{\pi : \sum_a \pi(a|x) \mathbb{E} Z(x, a) = \max_{a' \in \mathcal{A}} \mathbb{E} Z(x, a')\}.$$

回想一下, 期望贝尔曼最优算子  $\mathcal{T}$  是

$$\mathcal{T}Q(x, a) = \mathbb{E} R(x, a) + \gamma \mathbb{E}_P \max_{a' \in \mathcal{A}} Q(x', a'). \quad (6)$$

在  $x'$  上的最大化对应于某些贪婪策略。尽管这种策略在 (6) 中是隐含的, 但在分布设置中我们不能忽视它。我们将称任何实现贪婪选择规则的操作符 *distributional* 为 *Bellman optimality operator*, 即

$$\mathcal{T}Z = \mathcal{T}^\pi Z \text{ for some } \pi \in \mathcal{G}_Z.$$

就像在政策评估设置中一样, 我们对迭代序列  $Z_{k+1} := \mathcal{T}Z_k$ ,  $Z_0 \in \mathcal{Z}$  的行为感兴趣。我们的第一个结果是断言  $\mathbb{E} Z_k$  的行为符合预期。

引理 4.  $\{v^*\}$

$$\|\mathbb{E} \mathcal{T}Z_1 - \mathbb{E} \mathcal{T}Z_2\|_\infty \leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty,$$

and in particular  $\mathbb{E} Z_k \rightarrow Q^*$  exponentially quickly.

通过检查引理4, 我们可能会期望  $Z_k$  在  $\bar{d}_p$  中快速收敛到某个固定点  $Z^*$ 。不幸的是, 收敛既不快速, 也无法保证达到固定点。事实上, 我们所能期望的最佳结果是点wise收敛, 甚至不是收敛到集合  $\mathcal{Z}^*$ , 而是收敛到更大的集合 *nonstationary optimal value distributions*。



**Definition 3.** A nonstationary optimal value distribution  $Z^{**}$  is the value distribution corresponding to a sequence of optimal policies. The set of n.o.v.d. is  $Z^{**}$ .

**Theorem 1** (Convergence in the control setting). Let  $\mathcal{X}$  be measurable and suppose that  $\mathcal{A}$  is finite. Then

$$\lim_{k \rightarrow \infty} \inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) = 0 \quad \forall x, a.$$

If  $\mathcal{X}$  is finite, then  $Z_k$  converges to  $Z^{**}$  uniformly. Furthermore, if there is a total ordering  $\prec$  on  $\Pi^*$ , such that for any  $Z^* \in \mathcal{Z}^*$ ,

$$\mathcal{T}Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*}, \pi \prec \pi' \quad \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\}.$$

Then  $\mathcal{T}$  has a unique fixed point  $Z^* \in \mathcal{Z}^*$ .

Comparing Theorem 1 to Lemma 4 reveals a significant difference between the distributional framework and the usual setting of expected return. While the mean of  $Z_k$  converges exponentially quickly to  $Q^*$ , its distribution need not be as well-behaved! To emphasize this difference, we now provide a number of negative results concerning  $\mathcal{T}$ .

**Proposition 1.** The operator  $\mathcal{T}$  is not a contraction.

Consider the following example (Figure 2, left). There are two states,  $x_1$  and  $x_2$ ; a unique transition from  $x_1$  to  $x_2$ ; from  $x_2$ , action  $a_1$  yields no reward, while the optimal action  $a_2$  yields  $1 + \epsilon$  or  $-1 + \epsilon$  with equal probability. Both actions are terminal. There is a unique optimal policy and therefore a unique fixed point  $Z^*$ . Now consider  $Z$  as given in Figure 2 (right), and its distance to  $Z^*$ :

$$\bar{d}_1(Z, Z^*) = d_1(Z(x_2, a_2), Z^*(x_2, a_2)) = 2\epsilon,$$

where we made use of the fact that  $Z = Z^*$  everywhere except at  $(x_2, a_2)$ . When we apply  $\mathcal{T}$  to  $Z$ , however, the greedy action  $a_1$  is selected and  $\mathcal{T}Z(x_1) = Z(x_2, a_1)$ . But

$$\begin{aligned} d_1(\mathcal{T}Z, \mathcal{T}Z^*) &= d_1(\mathcal{T}Z(x_1), Z^*(x_1)) \\ &= \frac{1}{2}|1 - \epsilon| + \frac{1}{2}|1 + \epsilon| > 2\epsilon \end{aligned}$$

for a sufficiently small  $\epsilon$ . This shows that the undiscounted update is not a nonexpansion:  $\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$ . With  $\gamma < 1$ , the same proof shows it is not a contraction. Using a more technically involved argument, we can extend this result to any metric which separates  $Z$  and  $\mathcal{T}Z$ .

**Proposition 2.** Not all optimality operators have a fixed point  $Z^* = \mathcal{T}Z^*$ .

To see this, consider the same example, now with  $\epsilon = 0$ , and a greedy operator  $\mathcal{T}$  which breaks ties by picking  $a_2$  if  $Z(x_1) = 0$ , and  $a_1$  otherwise. Then the sequence  $\mathcal{T}Z^*(x_1), (\mathcal{T})^2 Z^*(x_1), \dots$  alternates between  $Z^*(x_2, a_1)$  and  $Z^*(x_2, a_2)$ .

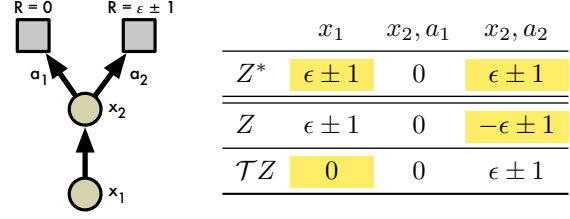


Figure 2. Undiscounted two-state MDP for which the optimality operator  $\mathcal{T}$  is not a contraction, with example. The entries that contribute to  $\bar{d}_1(Z, Z^*)$  and  $\bar{d}_1(\mathcal{T}Z, Z^*)$  are highlighted.

**Proposition 3.** That  $\mathcal{T}$  has a fixed point  $Z^* = \mathcal{T}Z^*$  is insufficient to guarantee the convergence of  $\{Z_k\}$  to  $Z^*$ .

Theorem 1 paints a rather bleak picture of the control setting. It remains to be seen whether the dynamical eccentricities highlighted here actually arise in practice. One open question is whether theoretically more stable behaviour can be derived using stochastic policies, for example from conservative policy iteration (Kakade & Langford, 2002).

## 4. Approximate Distributional Learning

In this section we propose an algorithm based on the distributional Bellman optimality operator. In particular, this will require choosing an approximating distribution. Although the Gaussian case has previously been considered (Morimura et al., 2010a; Tamar et al., 2016), to the best of our knowledge we are the first to use a rich class of parametric distributions.

### 4.1. Parametric Distribution

We will model the value distribution using a discrete distribution parametrized by  $N \in \mathbb{N}$  and  $V_{\min}, V_{\max} \in \mathbb{R}$ , and whose support is the set of atoms  $\{z_i = V_{\min} + i\Delta z : 0 \leq i < N\}$ ,  $\Delta z := \frac{V_{\max} - V_{\min}}{N-1}$ . In a sense, these atoms are the “canonical returns” of our distribution. The atom probabilities are given by a parametric model  $\theta : \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^N$

$$Z_\theta(x, a) = z_i \quad \text{w.p.} \quad p_i(x, a) := \frac{e^{\theta_i(x, a)}}{\sum_j e^{\theta_j(x, a)}}.$$

The discrete distribution has the advantages of being highly expressive and computationally friendly (see e.g. Van den Oord et al., 2016).

### 4.2. Projected Bellman Update

Using a discrete distribution poses a problem: the Bellman update  $\mathcal{T}Z_\theta$  and our parametrization  $Z_\theta$  almost always have disjoint supports. From the analysis of Section 3 it would seem natural to minimize the Wasserstein metric (viewed as a loss) between  $\mathcal{T}Z_\theta$  and  $Z_\theta$ , which is also

定义 3. A nonstationary optimal value distribution  $Z^{**}$  is the value distribution corresponding to a sequence of optimal policies. The set of n.o.v.d. is  $Z^{**}$ .

定理 1 (控制设置下的收敛性). Let  $\mathcal{X}$  be measurable and suppose that  $\mathcal{A}$  is finite. Then

$$\lim_{k \rightarrow \infty} \inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) = 0 \quad \forall x, a.$$

If  $\mathcal{X}$  is finite, then  $Z_k$  converges to  $Z^{**}$  uniformly. Furthermore, if there is a total ordering  $\prec$  on  $\Pi^*$ , such that for any  $Z^* \in \mathcal{Z}^*$ ,

$$\mathcal{T}Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*}, \pi \prec \pi' \quad \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\}.$$

Then  $\mathcal{T}$  has a unique fixed point  $Z^* \in \mathcal{Z}^*$ .

将定理1与引理4进行比较，可以发现分布框架与通常的期望回报设置之间存在显著差异。虽然 $Z_k$ 的均值以指数速度迅速收敛到 $Q^*$ ，但其分布未必会表现得那么好！为了强调这种差异，我们现在提供一些关于 $\mathcal{T}$ 的负面结果。

命题 1. The operator  $\mathcal{T}$  is not a contraction.

考虑以下示例（图2，左）。有两个状态， $x_1$  和  $x_2$ ；从  $x_1$  到  $x_2$  的唯一转移；从  $x_2$ ，动作  $a_1$  不产生奖励，而最优动作  $a_2$  以相等的概率产生  $1 + \epsilon$  或  $-1 + \epsilon$ 。两个动作都是终端动作。存在唯一的最优策略，因此存在唯一的不动点  $Z^*$ 。现在考虑图2（右）中的  $Z$ ，以及它与  $Z^*$  的距离：

$$\bar{d}_1(Z, Z^*) = d_1(Z(x_2, a_2), Z^*(x_2, a_2)) = 2\epsilon,$$

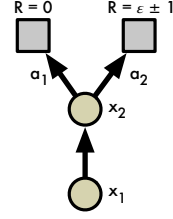
其中我们利用了除了 $(x_2, a_2)$ 处以外 $Z = Z^*$ 处处成立的事实。然而当我们把 $\mathcal{T}$ 应用到 $Z$ 时，贪婪操作 $a_1$ 被选择并且 $\mathcal{T}Z(x_1) = Z(x_2, a_1)$ 。但

$$\begin{aligned} d_1(\mathcal{T}Z, \mathcal{T}Z^*) &= d_1(\mathcal{T}Z(x_1), Z^*(x_1)) \\ &= \frac{1}{2}|1 - \epsilon| + \frac{1}{2}|1 + \epsilon| > 2\epsilon \end{aligned}$$

对于一个足够小的 $\epsilon$ 。这表明未折现更新不是非扩展性： $\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$ 。当 $\gamma < 1$ 时，同样的证明表明它不是收缩映射。使用一个更为技术上复杂的论证，我们可以将这个结果扩展到任何能够分离 $Z$ 和 $\mathcal{T}Z$ 的度量中。

命题 2. Not all optimality operators have a fixed point  $Z^* = \mathcal{T}Z^*$ .

为了说明这一点，考虑同一个例子，现在  $\epsilon = 0$ ，且贪婪操作符  $\mathcal{T}$  在打破平局时选择  $a_2$ ，如果  $Z(x_1) = 0$ ；否则选择  $a_1$ 。然后序列  $\mathcal{T}Z^*(x_1), (\mathcal{T})^2 Z^*(x_1), \dots$  在  $Z^*(x_2, a_1)$  和  $Z^*(x_2, a_2)$  之间交替。



	$x_1$	$x_2, a_1$	$x_2, a_2$
$Z^*$	$\epsilon \pm 1$	0	$\epsilon \pm 1$
$Z$	$\epsilon \pm 1$	0	$-\epsilon \pm 1$
$\mathcal{T}Z$	0	0	$\epsilon \pm 1$

图2. 一个未打折的两状态MDP，其中最优性算子 $\mathcal{T}$ 不是一个压缩映射，并附有示例。贡献于 $\bar{d}_1(Z, Z^*)$ 和 $\bar{d}_1(\mathcal{T}Z, Z^*)$ 的项已被突出显示。

命题 3. That  $\mathcal{T}$  has a fixed point  $Z^* = \mathcal{T}Z^*$  is insufficient to guarantee the convergence of  $\{Z_k\}$  to  $Z^*$ .

定理1描绘了一个相当悲观的控制设置图景。尚需观察这里突出的动力学偏心是否会在实践中真正出现。一个待解答的问题是，是否可以利用随机策略，例如保守的策略迭代（Kakade & Langford, 2002），来推导出理论上更稳定的行为。

## 4. 近似分布学习

在本节中，我们提出了一种基于分布贝尔曼最优算子的算法。特别是，这将需要选择一个逼近分布。尽管高斯情况之前已经被考虑过（Morimura等，2010a；Tamar等，2016），据我们所知，我们是第一个使用一类丰富的参数分布的。

### 4.1. 参数分布

我们将使用由  $N \in \mathbb{N}$  和  $V_{\min}, V_{\max} \in \mathbb{R}$  参数化且其支持集为原子集  $\{z_i = V_{\min} + i\Delta z : 0 \leq i < N\}$ ,  $\Delta z := \frac{V_{\max} - V_{\min}}{N-1}$  的离散分布来建模价值分布。在某种意义上，这些原子是“我们的分布的规范返回”。原子概率由参数模型  $\theta: \mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^N$  给出。

$$Z_\theta(x, a) = z_i \quad \text{w.p.} \quad p_i(x, a) := \frac{e^{\theta_i(x, a)}}{\sum_j e^{\theta_j(x, a)}}.$$

离散分布具有高度表达性和计算友好性的优点（参见例如Van den Oord等，2016年）。

### 4.2. 预测贝尔曼更新

使用离散分布会遇到一个问题：贝尔曼更新  $\{v^*\}$  和我们的参数化  $\{v^*\}$  几乎总是具有不相交的支持。从第3节的分析来看，最小化  $\{v^*\}$  和  $\{v^*\}$  之间的 Wasserstein 距离（视为损失）似乎是自然的选择，这也是

conveniently robust to discrepancies in support. However, a second issue prevents this: in practice we are typically restricted to learning from sample transitions, which is not possible under the Wasserstein loss (see Prop. 5 and toy results in the appendix).

Instead, we project the sample Bellman update  $\hat{T}Z_\theta$  onto the support of  $Z_\theta$  (Figure 1, Algorithm 1), effectively reducing the Bellman update to multiclass classification. Let  $\pi$  be the greedy policy w.r.t.  $\mathbb{E}Z_\theta$ . Given a sample transition  $(x, a, r, x')$ , we compute the Bellman update  $\hat{T}z_j := r + \gamma z_j$  for each atom  $z_j$ , then distribute its probability  $p_j(x', \pi(x'))$  to the immediate neighbours of  $\hat{T}z_j$ . The  $i^{\text{th}}$  component of the projected update  $\Phi\hat{T}Z_\theta(x, a)$  is

$$(\Phi\hat{T}Z_\theta(x, a))_i = \sum_{j=0}^{N-1} \left[ 1 - \frac{|\hat{T}z_j|_{V_{\min}}^{V_{\max}} - z_i|}{\Delta z} \right]_0^1 p_j(x', \pi(x')), \quad (7)$$

where  $[\cdot]_a^b$  bounds its argument in the range  $[a, b]$ .<sup>1</sup> As is usual, we view the next-state distribution as parametrized by a fixed parameter  $\tilde{\theta}$ . The sample loss  $\mathcal{L}_{x,a}(\theta)$  is the cross-entropy term of the KL divergence

$$D_{\text{KL}}(\Phi\hat{T}Z_{\tilde{\theta}}(x, a) \| Z_\theta(x, a)),$$

which is readily minimized e.g. using gradient descent. We call this choice of distribution and loss the *categorical algorithm*. When  $N = 2$ , a simple one-parameter alternative is  $\Phi\hat{T}Z_\theta(x, a) := [\mathbb{E}[\hat{T}Z_\theta(x, a)] - V_{\min}]/\Delta z$ ; we call this the *Bernoulli algorithm*. We note that, while these algorithms appear unrelated to the Wasserstein metric, recent work (Bellemare et al., 2017) hints at a deeper connection.

---

**Algorithm 1** Categorical Algorithm

---

**input** A transition  $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$   
 $Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$   
 $a^* \leftarrow \arg \max_a Q(x_{t+1}, a)$   
 $m_i = 0, \quad i \in 0, \dots, N-1$   
**for**  $j \in 0, \dots, N-1$  **do**  
  # Compute the projection of  $\hat{T}z_j$  onto the support  $\{z_i\}$   
   $\hat{T}z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\min}}^{V_{\max}}$   
   $b_j \leftarrow (\hat{T}z_j - V_{\min})/\Delta z$  #  $b_j \in [0, N-1]$   
   $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$   
  # Distribute probability of  $\hat{T}z_j$   
   $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$   
   $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$   
**end for**  
**output**  $-\sum_i m_i \log p_i(x_t, a_t)$  # Cross-entropy loss

---

## 5. Evaluation on Atari 2600 Games

To understand the approach in a complex setting, we applied the categorical algorithm to games from the Ar-

cade Learning Environment (ALE; Bellemare et al., 2013). While the ALE is deterministic, stochasticity does occur in a number of guises: 1) from state aliasing, 2) learning from a nonstationary policy, and 3) from approximation errors. We used five training games (Fig 3) and 52 testing games.

For our study, we use the DQN architecture (Mnih et al., 2015), but output the atom probabilities  $p_i(x, a)$  instead of action-values, and chose  $V_{\max} = -V_{\min} = 10$  from preliminary experiments over the training games. We call the resulting architecture *Categorical DQN*. We replace the squared loss  $(r + \gamma Q(x', \pi(x')) - Q(x, a))^2$  by  $\mathcal{L}_{x,a}(\theta)$  and train the network to minimize this loss.<sup>2</sup> As in DQN, we use a simple  $\epsilon$ -greedy policy over the expected action-values; we leave as future work the many ways in which an agent could select actions on the basis of the full distribution. The rest of our training regime matches Mnih et al.’s, including the use of a target network for  $\tilde{\theta}$ .

Figure 4 illustrates the typical value distributions we observed in our experiments. In this example, three actions (those including the button press) lead to the agent releasing its laser too early and eventually losing the game. The corresponding distributions reflect this: they assign a significant probability to 0 (the terminal value). The safe actions have similar distributions (LEFT, which tracks the invaders’ movement, is slightly favoured). This example helps explain why our approach is so successful: the distributional update keeps separated the low-value, “losing” event from the high-value, “survival” event, rather than average them into one (unrealizable) expectation.<sup>3</sup>

One surprising fact is that the distributions are not concentrated on one or two values, in spite of the ALE’s determinism, but are often close to Gaussians. We believe this is due to our discretizing the diffusion process induced by  $\gamma$ .

### 5.1. Varying the Number of Atoms

We began by studying our algorithm’s performance on the training games in relation to the number of atoms (Figure 3). For this experiment, we set  $\epsilon = 0.05$ . From the data, it is clear that using too few atoms can lead to poor behaviour, and that more always increases performance; this is not immediately obvious as we may have expected to saturate the network capacity. The difference in performance between the 51-atom version and DQN is particularly striking: the latter is outperformed in all five games, and in SEAQUEST we attain state-of-the-art performance. As an additional point of the comparison, the single-parameter Bernoulli algorithm performs better than DQN in 3 games out of 5, and is most notably more robust in ASTERIX.

<sup>2</sup>For  $N = 51$ , our TensorFlow implementation trains at roughly 75% of DQN’s speed.

<sup>3</sup>Video: <http://youtu.be/yFBwyPu02Vg>.

<sup>1</sup>Algorithm 1 computes this projection in time linear in  $N$ .

方便且稳健地处理支持集的差异。然而，一个次要问题阻止了这一点：实际上我们通常只能从样本过渡中学习，而在Wasserstein损失下这是不可能的（参见附录中的Prop. 5和玩具实验结果）。

相反，我们将样本贝尔曼更新 $\hat{T}Z_\theta$ 投影到 $Z_\theta$ （的支持上，如图1、算法1所示），从而将贝尔曼更新减少为多类分类。令 $\pi$ 为相对于 $\mathbb{E}Z_\theta$ 的贪婪策略。给定一个样本过渡 $(x, a, r, x')$ ，我们为每个原子 $z_j$ 计算贝尔曼更新 $\hat{T}z_j := r + \gamma z_j$ ，然后将其概率 $p_j(x', \pi(x'))$ 分配给 $\hat{T}z_j$ 的直接邻居。投影更新 $\Phi\hat{T}Z_\theta(x, a)$ 的 $i^{th}$ 组件为

$$(\Phi\hat{T}Z_\theta(x, a))_i = \sum_{j=0}^{N-1} \left[ 1 - \frac{|\hat{T}z_j|_{V_{\min}}^{V_{\max}} - z_i|}{\Delta z} \right]_0^1 p_j(x', \pi(x')), \quad (7)$$

其中 $[\cdot]_a^b$ 将其参数限制在范围 $[a, b]$ 内。<sup>1</sup>类似地，我们将下一状态分布视为由固定参数 $\tilde{\theta}$ 参数化。样本损失 $\mathcal{L}_{x,a}(\theta)$ 是KL散度的交叉熵项。

$$D_{\text{KL}}(\Phi\hat{T}Z_{\tilde{\theta}}(x, a) \| Z_\theta(x, a)),$$

这可以通过梯度下降等方法轻松最小化。我们称这种分布和损失的选择为*categorical algorithm*。当 $N = 2$ 时，一个简单的单参数替代方案是 $\Phi\hat{T}Z_\theta(x, a) := [\mathbb{E}[\hat{T}Z_\theta(x, a)] - V_{\min}] / \Delta z$ ；我们称其为*Bernoulli algorithm*。我们注意到，尽管这些算法看似与Wasserstein度量无关，但最近的研究（Bellemare等，2017）暗示了更深层次的关系。

#### Algorithm 1 Categorical Algorithm

**input** A transition  $x_t, a_t, r_t, x_{t+1}, \gamma_t \in [0, 1]$   
 $Q(x_{t+1}, a) := \sum_i z_i p_i(x_{t+1}, a)$   
 $a^* \leftarrow \arg \max_a Q(x_{t+1}, a)$   
 $m_i = 0, \quad i \in 0, \dots, N-1$   
**for**  $j \in 0, \dots, N-1$  **do**  
  # Compute the projection of  $\hat{T}z_j$  onto the support  $\{z_i\}$   
   $\hat{T}z_j \leftarrow [r_t + \gamma_t z_j]_{V_{\min}}^{V_{\max}}$   
   $b_j \leftarrow (\hat{T}z_j - V_{\min}) / \Delta z$  #  $b_j \in [0, N-1]$   
   $l \leftarrow \lfloor b_j \rfloor, u \leftarrow \lceil b_j \rceil$   
  # Distribute probability of  $\hat{T}z_j$   
   $m_l \leftarrow m_l + p_j(x_{t+1}, a^*)(u - b_j)$   
   $m_u \leftarrow m_u + p_j(x_{t+1}, a^*)(b_j - l)$   
**end for**  
**output**  $-\sum_i m_i \log p_i(x_t, a_t)$  # Cross-entropy loss

## 5. 在Atari 2600 游戏上的评估

为了在复杂环境中理解这种方法，我们将分类算法应用于Ar-中的游戏。

<sup>1</sup>Algorithm 1 computes this projection in time linear in  $N$ .

Cade 学习环境 (ALE; Bellemare 等人, 2013 年)。虽然 ALE 是确定性的，但仍然以多种方式出现随机性：1) 从状态别名，2) 从非稳态策略学习，以及 3) 从近似误差。我们使用了五种训练游戏（图 3）和 52 种测试游戏。

对于我们的研究，我们使用了DQN架构（Mnih等，2015），但输出的是原子概率 $p_i(x, a)$ 而不是动作值，并在初步实验中选择了在训练游戏中 $V_{\max} = -V_{\min} = 10$ 。我们称这种架构为*Categorical DQN*。我们用 $\mathcal{L}_{x,a}(\theta)$ 替代了平方损失 $(r + \gamma Q(x', \pi(x')) - Q(x, a))^2$ ，并训练网络以最小化这种损失<sup>2</sup>。就像DQN一样，我们使用了简单的 $\epsilon$ -贪婪策略来选择预期的动作值；关于代理如何基于完整的分布来选择动作，我们将其留作未来的工作。我们的训练制度其余部分与Mnih等人的相匹配，包括使用目标网络 $\tilde{\theta}$ 。

图4说明了我们在实验中观察到的典型值分布。在这个例子中，三个动作（包括按钮点击）导致智能体过早释放激光并最终输掉游戏。相应的分布反映了这一点：它们赋予了0（终端值）显著的概率。安全的动作具有类似的分布（LEFT，跟踪入侵者移动的动作略占优势）。这个例子有助于解释为什么我们的方法如此成功：值的分布更新将低价值的“失败”事件与高价值的“生存”事件区分开来，而不是将它们平均为一个（不可实现的）期望值。<sup>3</sup>

一个令人惊讶的事实是，尽管ALE具有确定性，分布并不集中在一两个值上，而是往往接近高斯分布。我们相信这是由于我们将由 $\gamma$ 引起的扩散过程离散化所致。

### 5.1. 原子数量的变化

我们首先研究了算法在训练游戏中性能与原子数量的关系（图3）。为此实验，我们设定了 $\epsilon = 0.05$ 。从数据中可以看出，使用太少的原子会导致性能较差，而更多的原子总是能提高性能；这并不像我们预期的那样会饱和网络容量。51个原子版本与DQN之间的性能差异尤为明显：后者在所有五个游戏中都表现不佳，在SEAQUEST中我们达到了最先进的性能。作为比较的额外一点，单参数伯努利算法在五个游戏中有三个游戏的表现优于DQN，并且在ASTERIX中表现尤为稳健。

<sup>2</sup>For  $N = 51$ , our TensorFlow implementation trains at roughly 75% of DQN's speed.

<sup>3</sup>Video: <http://youtu.be/yFBwyPuO2Vg>.

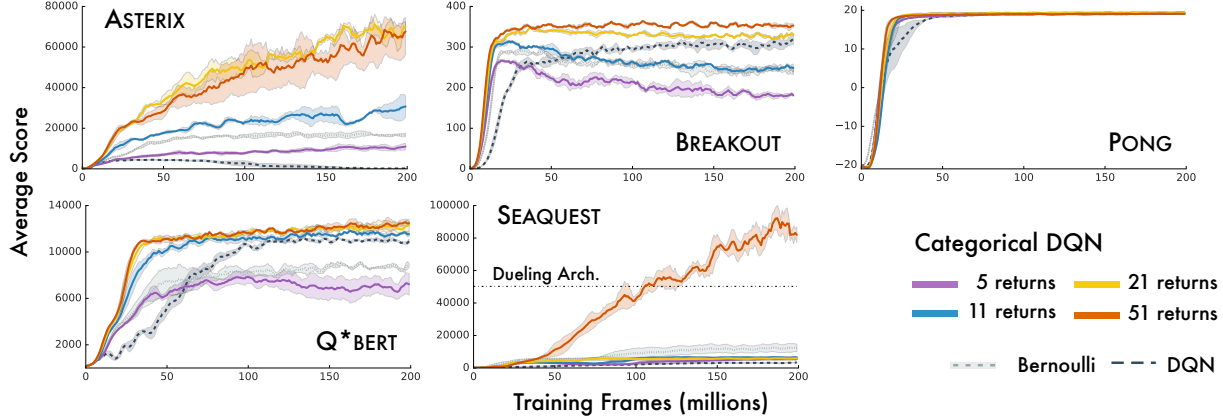


Figure 3. Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

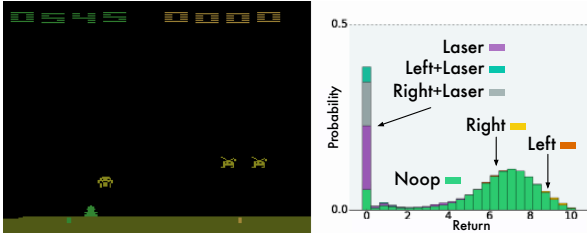


Figure 4. Learned value distribution during an episode of SPACE INVADERS. Different actions are shaded different colours. Returns below 0 (which do not occur in SPACE INVADERS) are not shown here as the agent assigns virtually no probability to them.

One interesting outcome of this experiment was to find out that our method does pick up on stochasticity. PONG exhibits intrinsic randomness: the exact timing of the reward depends on internal registers and is truly unobservable. We see this clearly reflected in the agent’s prediction (Figure 5): over five consecutive frames, the value distribution shows two modes indicating the agent’s belief that it has yet to receive a reward. Interestingly, since the agent’s state does not include past rewards, it cannot even extinguish the prediction after receiving the reward, explaining the relative proportions of the modes.

## 5.2. State-of-the-Art Results

The performance of the 51-atom agent (from here onwards, C51) on the training games, presented in the last section, is particularly remarkable given that it involved none of the other algorithmic ideas present in state-of-the-art agents. We next asked whether incorporating the most common hyperparameter choice, namely a smaller training  $\epsilon$ , could lead to even better results. Specifically, we set  $\epsilon = 0.01$  (instead of 0.05); furthermore, every 1 million frames, we

evaluate our agent’s performance with  $\epsilon = 0.001$ .

We compare our algorithm to DQN ( $\epsilon = 0.01$ ), Double DQN (van Hasselt et al., 2016), the Dueling architecture (Wang et al., 2016), and Prioritized Replay (Schaul et al., 2016), comparing the best evaluation score achieved during training. We see that C51 significantly outperforms these other algorithms (Figures 6 and 7). In fact, C51 surpasses the current state-of-the-art by a large margin in a number of games, most notably SEAQUEST. One particularly striking fact is the algorithm’s good performance on sparse reward games, for example VENTURE and PRIVATE EYE. This suggests that value distributions are better able to propagate rarely occurring events. Full results are provided in the appendix.

We also include in the appendix (Figure 12) a comparison, averaged over 3 seeds, showing the number of games in which C51’s training performance outperforms fully-trained DQN and human players. These results continue to show dramatic improvements, and are more representative of an agent’s average performance. Within 50 million frames, C51 has outperformed a fully trained DQN agent on 45 out of 57 games. This suggests that the full 200 million training frames, and its ensuing computational cost, are unnecessary for evaluating reinforcement learning algorithms within the ALE.

The most recent version of the ALE contains a stochastic execution mechanism designed to ward against trajectory overfitting. Specifically, on each frame the environment rejects the agent’s selected action with probability  $p = 0.25$ . Although DQN is mostly robust to stochastic execution, there are a few games in which its performance is reduced. On a score scale normalized with respect to the random and DQN agents, C51 obtains mean and median score improvements of 126% and 21.5% respectively, confirming the benefits of C51 beyond the deterministic setting.



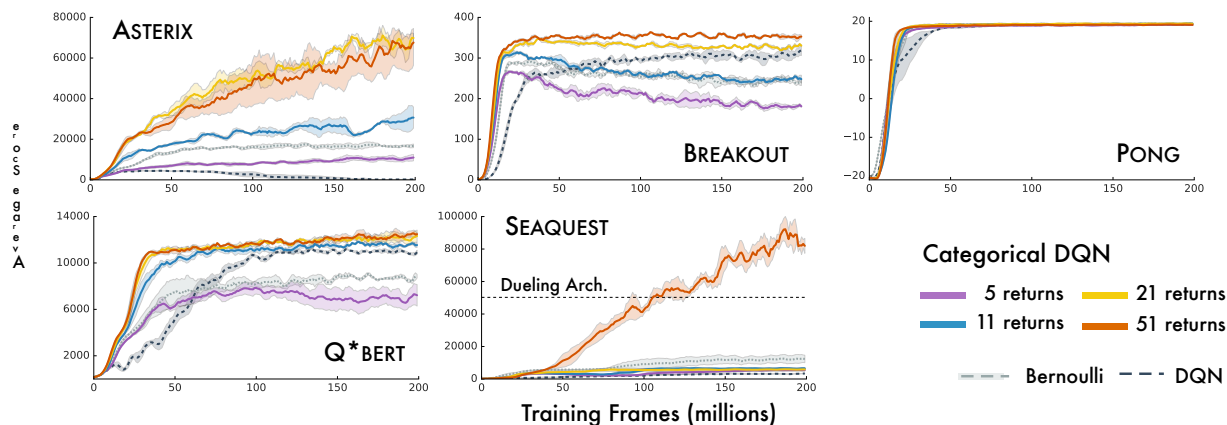


图3. 分类DQN：离散分布中原子数量的变化。得分是500万帧的移动平均值。

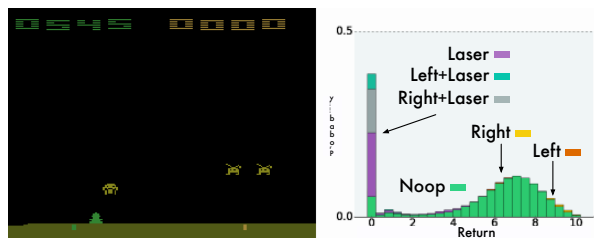


图4. 在 SPACE INVADERS 的一个回合中学习到的价值分布。不同的动作作用不同的颜色着色。低于0的回报（在SPACE INVADERS中不会出现）在这里未显示，因为代理几乎不会赋予它们任何概率。

这个实验的一个有趣结果是发现我们的方法确实能够捕捉到随机性。PONG 展现出内在的随机性：奖励的确切时机取决于内部寄存器，并且是真正不可观测的。我们在代理的预测中清楚地看到了这一点（图5）：在连续五帧中，价值分布显示出两个模式，表明代理认为它尚未收到奖励。有趣的是，由于代理的状态不包括过去的奖励，因此它甚至无法在收到奖励后消除这种预测，解释了这些模式的相对比例。

## 5.2. 最先进的结果

51个原子代理（从这里开始称为C51）在训练游戏中的表现尤其令人瞩目，因为这并未涉及当今最先进的代理中所包含的其他任何算法思想。接下来，我们询问是否将最常见的超参数选择，即较小的训练 $\epsilon$ ，纳入其中，可以带来更好的结果。具体来说，我们将 $\epsilon = 0.01$ （而不是0.05）；此外，每100万帧，我们

评估我们代理的表现使用  $\epsilon = 0.001$ 。

我们将我们的算法与DQN( $\epsilon = 0.01$ )、双DQN(van Hasselt et al., 2016)、Dueling架构(Wang et al., 2016)以及优先经验回放(Schaul et al., 2016)进行比较，比较的是训练过程中获得的最佳评估得分。我们发现C51在这些其他算法中表现显著更优（图6和图7）。实际上，在许多游戏中，C51在很大程度上超越了当前最先进的算法，尤其是在SEAQUEST游戏中。一个特别引人注目的事实是，该算法在稀疏奖励游戏中表现良好，例如VENTURE和PRIVATE EYE。这表明价值分布更能传播罕见事件。完整结果详见附录。

我们还将附录（图12）中包含一个比较内容，该比较基于3个种子的平均值，展示了C51的训练性能在多少个游戏中优于完全训练的DQN以及人类玩家。这些结果继续显示出显著的改进，并更能代表代理的平均性能。在5亿帧内，C51在45个游戏中优于完全训练的DQN代理，共有57个游戏。这表明，在评估ALE中的强化学习算法时，并不需要完整的2亿帧训练和随之而来的计算成本。

最新的ALE版本包含一个随机执行机制，旨在防止轨迹过拟合。具体来说，在每一帧中，环境以概率 $p = 0.25$ 拒绝代理选择的动作。尽管DQN对随机执行大部分是稳健的，但在少数游戏中，其性能会降低。相对于随机代理和DQN代理的得分标准化后，C51分别获得126%和21.5%的均值和中位数得分提升，证实了C51在确定性设置之外的益处。



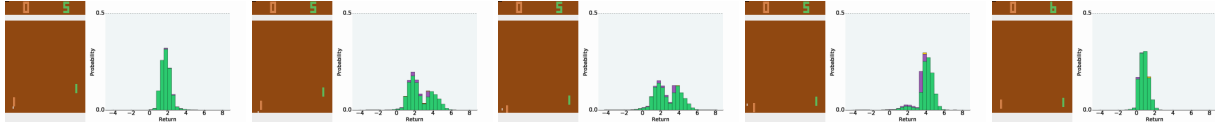


Figure 5. Intrinsic stochasticity in PONG.

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	<b>701%</b>	<b>178%</b>	<b>40</b>	<b>50</b>
UNREAL <sup>†</sup>	880%	250%	-	-

Figure 6. Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015).

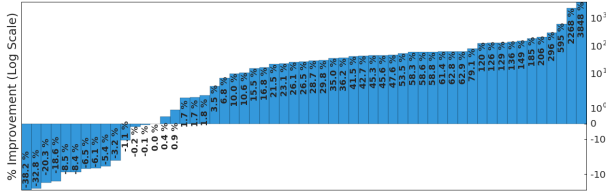


Figure 7. Percentage improvement, per-game, of C51 over Double DQN, computed using van Hasselt et al.’s method.

## 6. Discussion

In this work we sought a more complete picture of reinforcement learning, one that involves value distributions. We found that learning value distributions is a powerful notion that allows us to surpass most gains previously made on Atari 2600, without further algorithmic adjustments.

### 6.1. Why does learning a distribution matter?

It is surprising that, when we use a policy which aims to maximize expected return, we should see any difference in performance. The distinction we wish to make is that *learning distributions matters in the presence of approximation*. We now outline some possible reasons.

**Reduced chattering.** Our results from Section 3.4 highlighted a significant instability in the Bellman optimality operator. When combined with function approximation, this instability may prevent the policy from converging, what Gordon (1995) called *chattering*. We believe the gradient-based categorical algorithm is able to mitigate these effects by effectively averaging the different distri-

<sup>†</sup> The UNREAL results are not altogether comparable, as they were generated in the asynchronous setting with per-game hyperparameter tuning (Jaderberg et al., 2017).

butions, similar to conservative policy iteration (Kakade & Langford, 2002). While the chattering persists, it is integrated to the approximate solution.

**State aliasing.** Even in a deterministic environment, state aliasing may result in effective stochasticity. McCallum (1995), for example, showed the importance of coupling representation learning with policy learning in partially observable domains. We saw an example of state aliasing in PONG, where the agent could not exactly predict the reward timing. Again, by explicitly modelling the resulting distribution we provide a more stable learning target.

**A richer set of predictions.** A recurring theme in artificial intelligence is the idea of an agent learning from a multitude of predictions (Caruana 1997; Utgoff & Stracuzzi 2002; Sutton et al. 2011; Jaderberg et al. 2017). The distributional approach naturally provides us with a rich set of auxiliary predictions, namely: the probability that the return will take on a particular value. Unlike previously proposed approaches, however, the accuracy of these predictions is tightly coupled with the agent’s performance.

**Framework for inductive bias.** The distributional perspective on reinforcement learning allows a more natural framework within which we can impose assumptions about the domain or the learning problem itself. In this work we used distributions with support bounded in  $[V_{\min}, V_{\max}]$ . Treating this support as a hyperparameter allows us to change the optimization problem by treating all extremal returns (e.g. greater than  $V_{\max}$ ) as equivalent. Surprisingly, a similar value clipping in DQN significantly degrades performance in most games. To take another example: interpreting the discount factor  $\gamma$  as a proper probability, as some authors have argued, leads to a different algorithm.

**Well-behaved optimization.** It is well-accepted that the KL divergence between categorical distributions is a reasonably easy loss to minimize. This may explain some of our empirical performance. Yet early experiments with alternative losses, such as KL divergence between continuous densities, were not fruitful, in part because the KL divergence is insensitive to the values of its outcomes. A closer minimization of the Wasserstein metric should yield even better results than what we presented here.

In closing, we believe our results highlight the need to account for distribution in the design, theoretical or otherwise, of algorithms.

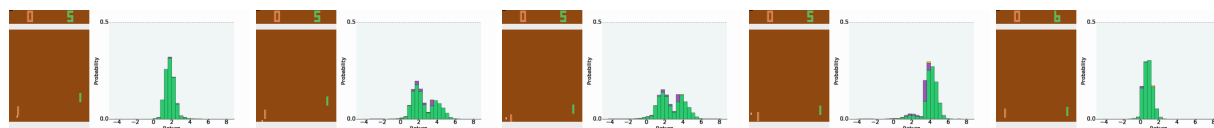


图5. 弹球游戏中固有的随机性。

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	<b>701%</b>	<b>178%</b>	<b>40</b>	<b>50</b>
UNREAL <sup>†</sup>	880%	250%	-	-

图6. 在57个 Atari 游戏中的平均分和中位数得分，以人类基准 (H.B., Nair等, 2015年) 的百分比表示。

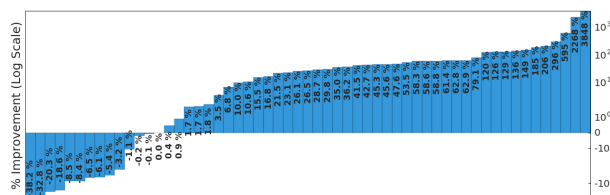


图7. C51相对于Double DQN的每场比赛百分比改进，使用van Hasselt等人方法计算得出。

## 6. 讨论

在本工作中，我们寻求获得关于强化学习更加完整的图景，一个涉及价值分布的图景。我们发现学习价值分布是一个强大的概念，它使我们能够在不进行进一步算法调整的情况下超越之前在Atari 2600上取得的大多数进展。

### 6.1. 为什么学习分布matter? (注：在这里“matter”

令人惊讶的是，当我们使用一个旨在最大化预期回报的策略时，我们应该看到任何性能上的差异。我们想要区分的是

*learning distributions matters in the presence of approximation.* 我们现在概述一些可能的原因。

减少颤振。如第3.4节所示，我们的结果突出显示了贝尔曼最优性算子中的显著不稳定性。当与函数逼近结合时，这种不稳定性可能会阻止策略收敛，正如戈登（1995年）所称的 $\{v^*\}$ 。我们认为基于梯度的分类算法能够通过有效平均不同的分布来减轻这些影响。

但这些操作类似于保守的策略迭代(Kakade & Langford, 2002)。在发散持续存在时，它被集成到近似解中。

状态混同。即使在确定性环境中，状态混同也可能导致有效的随机性。例如，McCallum（1995年）展示了在部分可观测域中，将表示学习与策略学习耦合的重要性。我们在PONG中看到了状态混同的一个例子，其中代理无法准确预测奖励的时间。同样，通过明确建模由此产生的分布，我们可以提供一个更稳定的学习目标。

更丰富的预测集。人工智能中的一个 recurring 主题是代理从多种预测中学习的想法 (Caruana 1997; Utgoff & Stracuzzi 2002; Sutton et al. 2011; Jaderberg et al. 2017)。分布方法自然地为我们提供了一组丰富的辅助预测，即：回报取特定值的概率。然而，与之前提出的其他方法不同，这些预测的准确性与代理的性能紧密相关。

框架中的归纳偏置。强化学习的分布视角为我们提供了一个更自然的框架，可以在其中对领域或学习问题本身做出假设。在本文中，我们使用了支持在 $[V_{\min}, V_{\max}]$ 内的分布。将此支持视为超参数，使我们能够通过将所有极端回报（例如，大于 $V_{\max}$ ）视为等价来改变优化问题。令人惊讶的是，在大多数游戏中，DQN中的类似值剪裁显著降低了性能。再举一个例子：

将折扣因子解释为一个适当的概率，如一些作者所保持不变的，因为它在中文中没有直接对应的含义，保持原文意图。）主张的，会导致不同的算法。

良好的优化。人们普遍认为，类别分布之间的KL散度是一个相对容易最小化的损失函数。这或许可以解释我们的一些实证表现。然而，早期使用其他损失函数（如连续密度之间的KL散度）的实验并未取得成功，部分原因是KL散度对其实现值不够敏感。更接近最小化Wasserstein度量应该能比我们这里展示的获得更好的结果。

总之，我们认为我们的结果强调了在算法的设计（无论是理论上的还是其他方式）中考虑分布的必要性。

<sup>†</sup> The UNREAL results are not altogether comparable, as they were generated in the asynchronous setting with per-game hyperparameter tuning (Jaderberg et al., 2017).

## Acknowledgements

The authors acknowledge the important role played by their colleagues at DeepMind throughout the development of this work. Special thanks to Yee Whye Teh, Alex Graves, Joel Veness, Guillaume Desjardins, Tom Schaul, David Silver, Andre Barreto, Max Jaderberg, Mohammad Azar, Georg Ostrovski, Bernardo Avila Pires, Olivier Pietquin, Audrunas Gruslys, Tom Stepleton, Aaron van den Oord; and particularly Chris Maddison for his comprehensive review of an earlier draft. Thanks also to Marek Petrik for pointers to the relevant literature, and Mark Rowland for fine-tuning details in the final version.

## Erratum

The camera-ready copy of this paper incorrectly reported a mean score of 1010% for C51. The corrected figure stands at 701%, which remains higher than the other comparable baselines. The median score remains unchanged at 178%.

The error was due to evaluation episodes in one game (Atlantis) lasting over 30 minutes; in comparison, the other results presented here cap episodes at 30 minutes, as is standard. The previously reported score on Atlantis was 3.7 million; our 30-minute score is 841,075, which we believe is close to the achievable maximum in this time frame. Capping at 30 minutes brings our human-normalized score on Atlantis from 22824% to a mere (!) 5199%, unfortunately enough to noticeably affect the mean score, whose sensitivity to outliers is well-documented.

## References

- Azar, Mohammad Gheshlaghi, Munos, Rémi, and Kappen, Hilbert. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the International Conference on Machine Learning*, 2012.
- Bellemare, Marc G, Naddaf, Yavar, Veness, Joel, and Bowling, Michael. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.
- Bellemare, Marc G., Danihelka, Ivo, Dabney, Will, Mohamed, Shakir, Lakshminarayanan, Balaji, Hoyer, Stephan, and Munos, Rémi. The cramer distance as a solution to biased wasserstein gradients. *arXiv*, 2017.
- Bellman, Richard E. *Dynamic programming*. Princeton University Press, Princeton, NJ, 1957.
- Bertsekas, Dimitri P. and Tsitsiklis, John N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
- Bickel, Peter J. and Freedman, David A. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, pp. 1196–1217, 1981.
- Billingsley, Patrick. *Probability and measure*. John Wiley & Sons, 1995.
- Caruana, Rich. Multitask learning. *Machine Learning*, 28(1): 41–75, 1997.
- Chung, Kun-Jen and Sobel, Matthew J. Discounted mdps: Distribution functions and exponential utility maximization. *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.
- Dearden, Richard, Friedman, Nir, and Russell, Stuart. Bayesian Q-learning. In *Proceedings of the National Conference on Artificial Intelligence*, 1998.
- Engel, Yaakov, Mannor, Shie, and Meir, Ron. Reinforcement learning with gaussian processes. In *Proceedings of the International Conference on Machine Learning*, 2005.
- Geist, Matthieu and Pietquin, Olivier. Kalman temporal differences. *Journal of Artificial Intelligence Research*, 39:483–532, 2010.
- Gordon, Geoffrey. Stable function approximation in dynamic programming. In *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.
- Harutyunyan, Anna, Bellemare, Marc G., Stepleton, Tom, and Munos, Rémi.  $Q(\lambda)$  with off-policy corrections. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2016.
- Hoffman, Matthew D., de Freitas, Nando, Doucet, Arnaud, and Peters, Jan. An expectation maximization algorithm for continuous markov decision processes with arbitrary reward. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009.
- Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David, and Kavukcuoglu, Koray. Reinforcement learning with unsupervised auxiliary tasks. *Proceedings of the International Conference on Learning Representations*, 2017.
- Jaquette, Stratton C. Markov decision processes with a new optimality criterion: Discrete time. *The Annals of Statistics*, 1(3): 496–505, 1973.
- Kakade, Sham and Langford, John. Approximately optimal approximate reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2002.
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *Proceedings of the International Conference on Learning Representations*, 2015.
- Lattimore, Tor and Hutter, Marcus. PAC bounds for discounted MDPs. In *Proceedings of the Conference on Algorithmic Learning Theory*, 2012.
- Mannor, Shie and Tsitsiklis, John N. Mean-variance optimization in markov decision processes. 2011.
- McCallum, Andrew K. *Reinforcement learning with selective perception and hidden state*. PhD thesis, University of Rochester, 1995.
- Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Riedmiller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

## 致谢

作者们感谢在整个工作中发挥重要作用的DeepMind同事。特别感谢叶卫艺、亚历克斯·格雷夫斯、约尔·文尼斯、Guillaume Desjardins、汤姆·沙乌尔、大卫·银、安德烈·巴雷托、马克斯·雅德贝格、穆罕默德·阿扎尔、乔治·奥斯特洛夫斯基、伯纳多·阿维拉·皮雷斯、奥利维尔·皮埃吉恩、奥杜纳斯·格鲁斯利斯、汤姆·斯蒂普莱顿、Aaron van den Oord；特别是克里斯·马迪森对早期草稿的全面审阅表示感谢。同时也要感谢马雷克·佩特里克提供的相关文献指针，以及Mark Rowland在最终版本中对细节的精细调整。

## 勘误

这篇论文的 camera-ready 版本错误地报告了 C51 的平均得分为 1010%。修正后的数字是 701%，这仍然高于其他可比基准。中位数得分保持不变，仍为 178%。

错误是由于在一款游戏（亚特兰蒂斯）中评估episode持续时间超过30分钟造成的；相比之下，这里呈现的其他结果将episode限制在30分钟，这是标准做法。之前报告的亚特兰蒂斯得分为370万；我们的30分钟得分为841,075，我们认为这接近在这个时间范围内可达到的最大值。将限制在30分钟将我们的亚特兰蒂斯的人类标准化得分为22824%降低到仅仅（！）5199%，不幸的是这足以明显影响平均得分，而平均得分对异常值的敏感性是众所周知的。

## 参考文献

Azar, Mohammad Gheshlaghi, Munos, Rémi, 和 Kappen, Hilbert. 关于生成模型下的强化学习样本复杂性。在 *Proceedings of the International Conference on Machine Learning*, 2012年。

Bellemare, Marc G, Naddaf, Yavar, Veness, Joel, 和 Bowling, Michael. 游戏学习环境：通用代理的评估平台。 *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

Bellemare, 马克·G., 丹尼赫卡, 伊沃, 达宾尼, 威尔, 莫罕梅德, 沙基尔, 拉克希米纳拉亚南, 巴拉吉, 赖耶, 斯蒂芬, 和 摩努斯, Rémi. 作为有偏Wasserstein梯度解决方案的Cramer距离。 *arXiv*, 2017.

Bellman, Richard E. *Dynamic programming*. 喷气岭大学出版社, 喷气岭, 新泽西, 1957年。

Bertsekas, Dimitri P. 和 Tsitsiklis, John N. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

Bickel, Peter J. 和 Freedman, David A. 一些渐近理论及自助法。 *The Annals of Statistics*, 第1196-1217页, 1981年。

比尔林赛利, 帕特里克. *Probability and measure*. 约翰威利 Sons, 1995.

Caruana, Rich. 多任务学习. *Machine Learning*, 28(1): 41–75, 1997.

Chung, Kun-Jen 和 Sobel, Matthew J. 折扣MDP：分布函数和指数效用最大化。 *SIAM Journal on Control and Optimization*, 25(1):49–62, 1987.

Dearden, Richard, Friedman, Nir, 和 Russell, Stuart. 信念Q学习。在 *Proceedings of the National Conference on Artificial Intelligence*, 1998.

Engel, Yaakov, Mannor, Shie, 和 Meir, Ron. 使用高斯过程的强化学习。在 *Proceedings of the International Conference on Machine Learning*, 2005. Geist, Matthieu 和 Pietquin, Olivier. Kalman 时间差分。在 *Journal of Artificial Intelligence Research*, 39:483–532, 2010. Gordon, Geoffrey. 动态规划中稳定的函数逼近。在 *Proceedings of the Twelfth International Conference on Machine Learning*, 1995.

Harutyunyan, Anna, Bellemare, Marc G., Stepleton, Tom, 和 Munos, Rémi.  $Q(\lambda)$  伴有离策校正。在 *Proceedings of the Conference on Algorithmic Learning Theory*, 2016. Hoffman, Matthew D., de Freitas, Nando, Doucet, Arnaud, 和 Peters, Jan. 连续马尔可夫决策过程的期望最大化算法，带有任意奖励。在 *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009.

Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David, 和 Kavukcuoglu, Koray. 使用未监督辅助任务的强化学习。 *Proceedings of the International Conference on Learning Representations*, 2017.

Jaquette, Stratton C. 1973年离散时间的马尔可夫决策过程的新最优性准则： *The Annals of Statistics*, 1(3): 496–505.

Kakade, Sham 和 Langford, John. 大致最优的近似强化学习。在 *Proceedings of the International Conference on Machine Learning*, 2002.

Kingma, Diederik 和 Ba, Jimmy. Adam：一种随机优化方法。 *Proceedings of the International Conference on Learning Representations*, 2015.

Lattimore, 托尔和Hutter, 马库斯. 折扣MDP的PAC边界。在 *Proceedings of the Conference on Algorithmic Learning Theory*, 2012.

Mannor, Shie 和 Tsitsiklis, John N. 关于马尔可夫决策过程中的均值-方差优化, 2011.

McCallum, Andrew K. *Reinforcement learning with selective perception and hidden state*. 博士学位论文, 罗切斯特大学, 1995年。

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, An-drei A, Veness, Joel, Bellemare, Marc G, Graves, Alex, Ried-miller, Martin, Fidjeland, Andreas K, Ostrovski, Georg, 等. 通过深度强化学习达到人类水平的控制能力。 *Nature*, 518(7540):529–533, 2015.

- Morimura, Tetsuro, Hachiya, Hirotaka, Sugiyama, Masashi, Tanaka, Toshiyuki, and Kashima, Hisashi. Parametric return density estimation for reinforcement learning. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010a.
- Morimura, Tetsuro, Sugiyama, Masashi, Kashima, Hisashi, Hachiya, Hirotaka, and Tanaka, Toshiyuki. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 799–806, 2010b.
- Nair, Arun, Srinivasan, Praveen, Blackwell, Sam, Alcicek, Cagdas, Fearon, Rory, De Maria, Alessandro, Panneershelvam, Vedavyas, Suleyman, Mustafa, Beattie, Charles, and Petersen, Stig et al. Massively parallel methods for deep reinforcement learning. In *ICML Workshop on Deep Learning*, 2015.
- Prashanth, LA and Ghavamzadeh, Mohammad. Actor-critic algorithms for risk-sensitive mdps. In *Advances in Neural Information Processing Systems*, 2013.
- Puterman, Martin L. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.
- Rösler, Uwe. A fixed point theorem for distributions. *Stochastic Processes and their Applications*, 42(2):195–214, 1992.
- Schaul, Tom, Quan, John, Antonoglou, Ioannis, and Silver, David. Prioritized experience replay. In *Proceedings of the International Conference on Learning Representations*, 2016.
- Sobel, Matthew J. The variance of discounted markov decision processes. *Journal of Applied Probability*, 19(04):794–802, 1982.
- Sutton, Richard S. and Barto, Andrew G. *Reinforcement learning: An introduction*. MIT Press, 1998.
- Sutton, R.S., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., and Precup, D. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the International Conference on Autonomous Agents and Multiagents Systems*, 2011.
- Tamar, Aviv, Di Castro, Dotan, and Mannor, Shie. Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36, 2016.
- Tieleman, Tijmen and Hinton, Geoffrey. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4 (2), 2012.
- Toussaint, Marc and Storkey, Amos. Probabilistic inference for solving discrete and continuous state markov decision processes. In *Proceedings of the International Conference on Machine Learning*, 2006.
- Tsitsiklis, John N. On the convergence of optimistic policy iteration. *Journal of Machine Learning Research*, 3:59–72, 2002.
- Utgoff, Paul E. and Straczuzi, David J. Many-layered learning. *Neural Computation*, 14(10):2497–2529, 2002.
- Van den Oord, Aaron, Kalchbrenner, Nal, and Kavukcuoglu, Koray. Pixel recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, 2016.
- van Hasselt, Hado, Guez, Arthur, and Silver, David. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016.
- Veness, Joel, Bellemare, Marc G., Hutter, Marcus, Chua, Alvin, and Desjardins, Guillaume. Compress and control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015.
- Wang, Tao, Lizotte, Daniel, Bowling, Michael, and Schuurmans, Dale. Dual representations for dynamic programming. *Journal of Machine Learning Research*, pp. 1–29, 2008.
- Wang, Ziyu, Schaul, Tom, Hessel, Matteo, Hasselt, Hado van, Lanctot, Marc, and de Freitas, Nando. Dueling network architectures for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, 2016.
- White, D. J. Mean, variance, and probabilistic criteria in finite markov decision processes: a review. *Journal of Optimization Theory and Applications*, 56(1):1–29, 1988.

Morimura, Tetsuro, Hachiyu, Hirotaka, Sugiyama, Masashi, Tanaka, Toshiyuki, 和 Kashima, Hisashi. 基于参数的回报密度估计在强化学习中的应用. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010a. Morimura, Tetsuro, Sugiyama, Masashi, Kashima, Hisashi, Hachiyu, Hirotaka, 和 Tanaka, Toshiyuki. 基于非参数的回报分布近似在强化学习中的应用. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 799–806, 2010b. Nair, Arun, Srinivasan, Parveen, Blackwell, Sam, Alcicek, Cagdas, Fearon, Rory, De Maria, Alessandro, Panneershelvam, Vedavyas, Suleyman, Mustafa, Beattie, Charles, 和 Petersen, Stig 等. 大规模并行方法在深度强化学习中的应用. In *ICML Workshop on Deep Learning*, 2015. Prashanth, LA 和 Ghavamzadeh, Mohammad. 风险敏感MDP的演员-评论者算法. In *Advances in Neural Information Processing Systems*, 2013. Puterman, Martin L. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994. Rösler, Uwe. 分布的不动点定理. *Stochastic Processes and their Applications*, 42(2):195–214, 1992. Schaul, Tom, Quan, John, Antonoglou, Ioannis, 和 Silver, David. 优先经验回放. In *Proceedings of the International Conference on Learning Representations*, 2016. Sobel, Matthew J. 折扣马尔可夫决策过程的方差. *Journal of Applied Probability*, 19(04):794–802, 1982. Sutton, Richard S. 和 Barto, Andrew G. *Reinforcement learning: An introduction*. MIT Press, 1998. Sutton, R.S., Modayil, J., Delp, M., Degris, T., Pilarski, P.M., White, A., 和 Precup, D. Horde: 一种可扩展的实时架构, 用于从无监督的传感器-运动交互中学习知识. In *Proceedings of the International Conference on Autonomous Agents and Multiagents Systems*, 2011. Tamar, Aviv, Di Castro, Dotan, 和 Mannor, Shie. 学习回报到去的方差. *Journal of Machine Learning Research*, 17(13):1–36, 2016. Tieleman, Tijmen 和 Hinton, Geoffrey. 讲座 6.5-RMSprop: 将梯度除以其最近幅度的运行平均值. *COURSERA: Neural networks for machine learning*, 4 (2), 2012. Toussaint, Marc 和 Storkey, Amos. 概率推理在解决离散和连续状态马尔可夫决策过程中的应用. In *Proceedings of the International Conference on Machine Learning*, 2006. Tsitsiklis, John N. 乐观策略迭代的收敛性. *Journal of Machine Learning Research*, 3:59–72, 2002. Utgoff, Paul E. 和 Stracuzzi, David J. 多层学习. *Neural Computation*, 14(10):2497–2529, 2002. Van den Oord, Aaron, Kalchbrenner, Nal, 和 Kavukcuoglu, Koaray. 像素递归神经网络. In *Proceedings of the International Conference on Machine Learning*, 2016.

van Hasselt, Hado, Guez, Arthur, 和 Silver, David. 使用双Q学习的深度强化学习. 在 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. Veness, Joel, Bellemare, Marc G., Hutter, Marcus, Chua, Alvin, 和 Desjardins, Guillaume. 压缩与控制. 在 *Proceedings of the AAAI Conference on Artificial Intelligence*, 2015. Wang, Tao, Lizotte, Daniel, Bowling, Michael, 和 Schuurmans, Dale. 动态规划的对偶表示. *Journal of Machine Learning Research*, 页码: 1–29, 2008. Wang, Ziyu, Schaul, Tom, Hessel, Matteo, Hasselt, Hado van, Lanctot, Marc, 和 de Freitas, Nando. 深度强化学习中的对弈网络架构. 在 *Proceedings of the International Conference on Machine Learning*, 2016. White, D. J. 有限马尔可夫决策过程中的均值、方差和概率准则: 一个回顾. *Journal of Optimization Theory and Applications*, 56(1):1–29, 1988.



## A. Related Work

To the best of our knowledge, the work closest to ours are two papers (Morimura et al., 2010b;a) studying the distributional Bellman equation from the perspective of its cumulative distribution functions. The authors propose both parametric and nonparametric solutions to learn distributions for risk-sensitive reinforcement learning. They also provide some theoretical analysis for the policy evaluation setting, including a consistency result in the nonparametric case. By contrast, we also analyze the control setting, and emphasize the use of the distributional equations to improve approximate reinforcement learning.

The variance of the return has been extensively studied in the risk-sensitive setting. Of note, Tamar et al. (2016) analyze the use of linear function approximation to learn this variance for policy evaluation, and Prashanth & Ghavamzadeh (2013) estimate the return variance in the design of a risk-sensitive actor-critic algorithm. Mannor & Tsitsiklis (2011) provides negative results regarding the computation of a variance-constrained solution to the optimal control problem.

The distributional formulation also arises when modelling uncertainty. Dearden et al. (1998) considered a Gaussian approximation to the value distribution, and modelled the uncertainty over the parameters of this approximation using a Normal-Gamma prior. Engel et al. (2005) leveraged the distributional Bellman equation to define a Gaussian process over the unknown value function. More recently, Geist & Pietquin (2010) proposed an alternative solution to the same problem based on unscented Kalman filters. We believe much of the analysis we provide here, which deals with the intrinsic randomness of the environment, can also be applied to modelling uncertainty.

Our work here is based on a number of foundational results, in particular concerning alternative optimality criteria. Early on, Jaquette (1973) showed that a *moment optimality* criterion, which imposes a total ordering on distributions, is achievable and defines a stationary optimal policy, echoing the second part of Theorem 1. Sobel (1982) is usually cited as the first reference to Bellman equations for the higher moments (but not the distribution) of the return. Chung & Sobel (1987) provides results concerning the convergence of the distributional Bellman operator in total variation distance. White (1988) studies “nonstandard MDP criteria” from the perspective of optimizing the state-action pair occupancy.

A number of probabilistic frameworks for reinforcement learning have been proposed in recent years. The *planning as inference* approach (Toussaint & Storkey, 2006; Hoffman et al., 2009) embeds the return into a graphical model, and applies probabilistic inference to determine the

sequence of actions leading to maximal expected reward. Wang et al. (2008) considered the dual formulation of reinforcement learning, where one optimizes the stationary distribution subject to constraints given by the transition function (Puterman, 1994), in particular its relationship to linear approximation. Related to this dual is the Compress and Control algorithm Veness et al. (2015), which describes a value function by learning a return distribution using density models. One of the aims of this work was to address the question left open by their work of whether one could be design a practical distributional algorithm based on the Bellman equation, rather than Monte Carlo estimation.

## B. Proofs

**Lemma 1** (Partition lemma). *Let  $A_1, A_2, \dots$  be a set of random variables describing a partition of  $\Omega$ , i.e.  $A_i(\omega) \in \{0, 1\}$  and for any  $\omega$  there is exactly one  $A_i$  with  $A_i(\omega) = 1$ . Let  $U, V$  be two random variables. Then*

$$d_p(U, V) \leq \sum_i d_p(A_i U, A_i V).$$

*Proof.* We will give the proof for  $p < \infty$ , noting that the same applies to  $p = \infty$ . Let  $Y_i \stackrel{D}{=} A_i U$  and  $Z_i \stackrel{D}{=} A_i V$ , respectively. First note that

$$\begin{aligned} d_p^p(A_i U, A_i V) &= \inf_{Y_i, Z_i} \mathbb{E} [|Y_i - Z_i|^p] \\ &= \inf_{Y_i, Z_i} \mathbb{E} [\mathbb{E} [|Y_i - Z_i|^p | A_i]]. \end{aligned}$$

Now,  $|A_i U - A_i V|^p = 0$  whenever  $A_i = 0$ . It follows that we can choose  $Y_i, Z_i$  so that also  $|Y_i - Z_i|^p = 0$  whenever  $A_i = 0$ , without increasing the expected norm. Hence

$$d_p^p(A_i U, A_i V) = \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p | A_i = 1]. \quad (8)$$

Next, we claim that

$$\begin{aligned} \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|A_i U - A_i V|^p | A_i = 1] \\ \leq \inf_{Y_1, Y_2, \dots, Z_1, Z_2, \dots} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p | A_i = 1]. \end{aligned} \quad (9)$$

Specifically, the left-hand side of the equation is an infimum over all r.v.’s whose cumulative distributions are  $F_U$  and  $F_V$ , respectively, while the right-hand side is an infimum over sequences of r.v.’s  $Y_1, Y_2, \dots$  and  $Z_1, Z_2, \dots$  whose cumulative distributions are  $F_{A_i U}, F_{A_i V}$ , respectively. To prove this upper bound, consider the c.d.f. of  $U$ :

$$\begin{aligned} F_U(y) &= \Pr\{U \leq y\} \\ &= \sum_i \Pr\{A_i = 1\} \Pr\{U \leq y | A_i = 1\} \\ &= \sum_i \Pr\{A_i = 1\} \Pr\{A_i U \leq y | A_i = 1\}. \end{aligned}$$

## A. 相关工作

据我们所知，与我们工作最接近的是两篇论文（Mori-mura等，2010b；a），它们从累积分布函数的角度研究了分布性的贝尔曼方程。作者们提出了参数和非参数两种方法来学习风险敏感强化学习中的分布。他们还策略评估设置提供了部分理论分析，包括非参数情况的一致性结果。相比之下，我们还分析了控制设置，并强调使用分布性方程来改进近似强化学习。

波动率在风险敏感设置中已经被广泛研究。值得注意的是，Tamar等（2016）分析了使用线性函数逼近来学习这种波动率以进行策略评估，而Prashanth & Ghavamzadeh（2013）则在风险敏感的演员-评论家算法设计中估计了回报的波动率。Mannor & Tsitsiklis（2011）提供了关于计算波动率约束下的最优控制问题解的负面结果。

分布式的表述在建模不确定性时也会出现。Dearden等人（1998年）考虑了价值分布的高斯近似，并使用Normal-Gamma先验来建模对该近似的参数的不确定性。Engel等人（2005年）利用分布式的贝尔曼方程定义了一个未知价值函数的高斯过程。最近，Geist与Pietquin（2010年）提出了基于无迹卡尔曼滤波的另一种解决方案。我们认为，我们在这里提供的大部分分析，即处理环境固有的随机性，也可以应用于建模不确定性。

我们的工作基于一些基础结果，特别是关于替代最优标准的结果。早期，Jaquette（1973）展示了总排序分布的moment optimality标准是可以实现的，并定义了一个稳态最优策略，这与定理1的第二部分相呼应。Sobel（1982）通常被认为是首次引用贝尔曼方程来描述回报的高阶矩（但不是分布）。Chung & Sobel（1987）提供了关于分布贝尔曼算子在总体变差距离下的收敛性的结果。White（1988）从优化状态-动作对占用率的角度研究了“非标准MDP标准”。

近年来，已经提出了许多强化学习的概率框架。planning as inference方法（Toussaint & Storkey, 2006; Hoffman et al., 2009）将回报嵌入到图形模型中，并应用概率推理来确定

序列导致最大期望奖励的一系列动作。王等（2008年）考虑了强化学习的对偶形式，其中在转移函数（Puterman, 1994）给出的约束条件下优化平稳分布，特别是其与线性逼近的关系。与此对偶相关的是Veness等（2015年）提出的Compress and Control算法，该算法通过学习回报分布来描述价值函数，使用密度模型。这项工作的目标之一是解决他们工作中留下的问题，即是否可以基于贝尔曼方程设计一种实用的分布算法，而不是使用蒙特卡洛估计。

## B. 证明

引理1 (分区引理). *Let  $A_1, A_2, \dots$  be a set of random variables describing a partition of  $\Omega$ , i.e.  $A_i(\omega) \in \{0, 1\}$  and for any  $\omega$  there is exactly one  $A_i$  with  $A_i(\omega) = 1$ . Let  $U, V$  be two random variables. Then*

$$d_p(U, V) \leq \sum_i d_p(A_i U, A_i V).$$

*Proof.* 我们将给出  $p < \infty$  的证明，注意到同样的证明也适用于  $p = \infty$ 。令  $Y_i \stackrel{D}{:=} A_i U$  和  $Z_i \stackrel{D}{:=} A_i V$  分别。首先注意到

$$\begin{aligned} d_p^p(A_i U, A_i V) &= \inf_{Y_i, Z_i} \mathbb{E} [|Y_i - Z_i|^p] \\ &= \inf_{Y_i, Z_i} \mathbb{E} [\mathbb{E} [|Y_i - Z_i|^p | A_i]]. \end{aligned}$$

现在， $|A_i U - A_i V|^p = 0$  whenever  $A_i = 0$ 。因此我们可以选择  $Y_i, Z_i$  使得当  $A_i = 0$  时  $|Y_i - Z_i|^p = 0$ ，而不增加期望范数。因此

$$d_p^p(A_i U, A_i V) = \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p | A_i = 1]. \quad (8)$$

接下来，我们断言 rằng

$$\begin{aligned} \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|A_i U - A_i V|^p | A_i = 1] \\ \leq \inf_{Y_1, Y_2, \dots, Z_1, Z_2, \dots} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p | A_i = 1]. \end{aligned} \quad (9)$$

具体地，方程的左边是对所有累积分布分别为  $F_U$  和  $F_V$  的随机变量 r.v.'s 的下确界，而右边是对累积分布分别为  $F_{A_i U}, F_{A_i V}$  的随机变量序列  $Y_1, Y_2, \dots$  和  $Z_1, Z_2, \dots$  的下确界。为了证明这个上界，考虑  $U$  的分布函数：

$$\begin{aligned} F_U(y) &= \Pr\{U \leq y\} \\ &= \sum_i \Pr\{A_i = 1\} \Pr\{U \leq y | A_i = 1\} \\ &= \sum_i \Pr\{A_i = 1\} \Pr\{A_i U \leq y | A_i = 1\}. \end{aligned}$$

Hence the distribution  $F_U$  is equivalent, in an almost sure sense, to one that first picks an element  $A_i$  of the partition, then picks a value for  $U$  conditional on the choice  $A_i$ . On the other hand, the c.d.f. of  $Y_i \stackrel{D}{=} A_i U$  is

$$\begin{aligned} F_{A_i U}(y) &= \Pr\{A_i = 1\} \Pr\{A_i U \leq y \mid A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \Pr\{A_i U \leq y \mid A_i = 0\} \\ &= \Pr\{A_i = 1\} \Pr\{A_i U \leq y \mid A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \mathbb{I}[y \geq 0]. \end{aligned}$$

Thus the right-hand side infimum in (9) has the additional constraint that it must preserve the conditional c.d.f.s, in particular when  $y \geq 0$ . Put another way, instead of having the freedom to completely reorder the mapping  $U : \Omega \rightarrow \mathbb{R}$ , we can only reorder it within each element of the partition. We now write

$$\begin{aligned} d_p^p(U, V) &= \inf_{U, V} \|U - V\|_p \\ &= \inf_{U, V} \mathbb{E} [|U - V|^p] \\ &\stackrel{(a)}{=} \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|U - V|^p \mid A_i = 1] \\ &= \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|A_i U - A_i V|^p \mid A_i = 1], \end{aligned}$$

where (a) follows because  $A_1, A_2, \dots$  is a partition. Using (9), this implies

$$\begin{aligned} d_p^p(U, V) &= \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|A_i U - A_i V|^p \mid A_i = 1] \\ &\leq \inf_{\substack{Y_1, Y_2, \dots \\ Z_1, Z_2, \dots}} \sum_i \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p \mid A_i = 1] \\ &\stackrel{(b)}{=} \sum_i \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \mathbb{E} [|Y_i - Z_i|^p \mid A_i = 1] \\ &\stackrel{(c)}{=} \sum_i d_p(A_i U, A_i V), \end{aligned}$$

because in (b) the individual components of the sum are independently minimized; and (c) from (8).  $\square$

**Lemma 2.**  $\bar{d}_p$  is a metric over value distributions.

*Proof.* The only nontrivial property is the triangle inequality. For any value distribution  $Y \in \mathcal{Z}$ , write

$$\begin{aligned} \bar{d}_p(Z_1, Z_2) &= \sup_{x, a} d_p(Z_1(x, a), Z_2(x, a)) \\ &\stackrel{(a)}{\leq} \sup_{x, a} [d_p(Z_1(x, a), Y(x, a)) + d_p(Y(x, a), Z_2(x, a))] \\ &\leq \sup_{x, a} d_p(Z_1(x, a), Y(x, a)) + \sup_{x, a} d_p(Y(x, a), Z_2(x, a)) \\ &= \bar{d}_p(Z_1, Y) + \bar{d}_p(Y, Z_2), \end{aligned}$$

where in (a) we used the triangle inequality for  $d_p$ .  $\square$

**Lemma 3.**  $\mathcal{T}^\pi : \mathcal{Z} \rightarrow \mathcal{Z}$  is a  $\gamma$ -contraction in  $\bar{d}_p$ .

*Proof.* Consider  $Z_1, Z_2 \in \mathcal{Z}$ . By definition,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \sup_{x, a} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)). \quad (10)$$

By the properties of  $d_p$ , we have

$$\begin{aligned} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)) &= d_p(R(x, a) + \gamma P^\pi Z_1(x, a), R(x, a) + \gamma P^\pi Z_2(x, a)) \\ &\leq \gamma d_p(P^\pi Z_1(x, a), P^\pi Z_2(x, a)) \\ &\leq \gamma \sup_{x', a'} d_p(Z_1(x', a'), Z_2(x', a')), \end{aligned}$$

where the last line follows from the definition of  $P^\pi$  (see (4)). Combining with (10) we obtain

$$\begin{aligned} \bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{x, a} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)) \\ &\leq \gamma \sup_{x', a'} d_p(Z_1(x', a'), Z_2(x', a')) \\ &= \gamma \bar{d}_p(Z_1, Z_2). \quad \square \end{aligned}$$

**Proposition 1** (Sobel, 1982). Consider two value distributions  $Z_1, Z_2 \in \mathcal{Z}$ , and write  $\mathbb{V}(Z_i)$  to be the vector of variances of  $Z_i$ . Then

$$\begin{aligned} \|\mathbb{E} \mathcal{T}^\pi Z_1 - \mathbb{E} \mathcal{T}^\pi Z_2\|_\infty &\leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty, \text{ and} \\ \|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty &\leq \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty. \end{aligned}$$

*Proof.* The first statement is standard, and its proof follows from  $\mathbb{E} \mathcal{T}^\pi Z = \mathcal{T}^\pi \mathbb{E} Z$ , where the second  $\mathcal{T}^\pi$  denotes the usual operator over value functions. Now, by independence of  $R$  and  $P^\pi Z_i$ :

$$\begin{aligned} \mathbb{V}(\mathcal{T}^\pi Z_i(x, a)) &= \mathbb{V}(R(x, a) + \gamma P^\pi Z_i(x, a)) \\ &= \mathbb{V}(R(x, a)) + \gamma^2 \mathbb{V}(P^\pi Z_i(x, a)). \end{aligned}$$

And now

$$\begin{aligned} \|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty &= \sup_{x, a} |\mathbb{V}(\mathcal{T}^\pi Z_1(x, a)) - \mathbb{V}(\mathcal{T}^\pi Z_2(x, a))| \\ &= \sup_{x, a} \gamma^2 |\mathbb{V}(P^\pi Z_1(x, a)) - \mathbb{V}(P^\pi Z_2(x, a))| \\ &= \sup_{x, a} \gamma^2 |\mathbb{E} [\mathbb{V}(Z_1(X', A')) - \mathbb{V}(Z_2(X', A'))]| \\ &\leq \sup_{x', a'} \gamma^2 |\mathbb{V}(Z_1(x', a')) - \mathbb{V}(Z_2(x', a'))| \\ &\leq \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty. \quad \square \end{aligned}$$

**Lemma 4.** Let  $Z_1, Z_2 \in \mathcal{Z}$ . Then

$$\|\mathbb{E} \mathcal{T} Z_1 - \mathbb{E} \mathcal{T} Z_2\|_\infty \leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty,$$

and in particular  $\mathbb{E} Z_k \rightarrow Q^*$  exponentially quickly.

因此分布  $F_U$  在几乎必然的意义上等价于先从分区中选择元素  $A_i$ , 然后在该选择的基础上为  $U$  选择一个条件值  $A_i$ 。另一方面,  $Y_i \stackrel{D}{=} A_i U$  的分布函数是

$$\begin{aligned} F_{A_i U}(y) &= \Pr\{A_i = 1\} \Pr\{A_i U \leq y \mid A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \Pr\{A_i U \leq y \mid A_i = 0\} \\ &= \Pr\{A_i = 1\} \Pr\{A_i U \leq y \mid A_i = 1\} \\ &\quad + \Pr\{A_i = 0\} \mathbb{I}[y \geq 0]. \end{aligned}$$

因此, 在公式(9)中右端的下确界还具有额外的约束, 即它必须保持条件分布函数, 特别是在  $y \neq 0$  时。换句话说, 我们不再有完全重新排列映射  $U: \Omega \rightarrow \mathbb{R}$  的自由, 而是只能在每个分区的元素内部重新排列。我们现在写出

$$\begin{aligned} d_p^p(U, V) &= \inf_{U, V} \|U - V\|_p \\ &= \inf_{U, V} \mathbb{E}[|U - V|^p] \\ &\stackrel{(a)}{=} \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E}[|U - V|^p \mid A_i = 1] \\ &= \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E}[|A_i U - A_i V|^p \mid A_i = 1], \end{aligned}$$

其中 (a) 成立是因为  $A_1, A_2, \dots$  是一个划分。使用 (9) 可知, 这 implies 意味着

$$\begin{aligned} d_p^p(U, V) &= \inf_{U, V} \sum_i \Pr\{A_i = 1\} \mathbb{E}[|A_i U - A_i V|^p \mid A_i = 1] \\ &\leq \inf_{\substack{Y_1, Y_2, \dots \\ Z_1, Z_2, \dots}} \sum_i \Pr\{A_i = 1\} \mathbb{E}[|Y_i - Z_i|^p \mid A_i = 1] \\ &\stackrel{(b)}{=} \sum_i \inf_{Y_i, Z_i} \Pr\{A_i = 1\} \mathbb{E}[|Y_i - Z_i|^p \mid A_i = 1] \\ &\stackrel{(c)}{=} \sum_i d_p(A_i U, A_i V), \end{aligned}$$

因为 (b) 中求和的各个分量是独立最小化的; 以及 (c) 来自 (8)。□

引理 2.  $\bar{d}_p$  is a metric over value distributions.

*Proof.* 唯一的非平凡性质是三角不等式。对于任何值分布  $Y \in \mathcal{Z}$ , 写

$$\begin{aligned} \bar{d}_p(Z_1, Z_2) &= \sup_{x, a} d_p(Z_1(x, a), Z_2(x, a)) \\ &\stackrel{(a)}{\leq} \sup_{x, a} [d_p(Z_1(x, a), Y(x, a)) + d_p(Y(x, a), Z_2(x, a))] \\ &\leq \sup_{x, a} d_p(Z_1(x, a), Y(x, a)) + \sup_{x, a} d_p(Y(x, a), Z_2(x, a)) \\ &= \bar{d}_p(Z_1, Y) + \bar{d}_p(Y, Z_2), \end{aligned}$$

其中在 (a) 中我们使用了三角不等式对于  $d_p$ 。□

引理 3.  $\mathcal{T}^\pi: \mathcal{Z} \rightarrow \mathcal{Z}$  is a  $\gamma$ -contraction in  $\bar{d}_p$ .

*Proof.* 考虑  $Z_1, Z_2 \in \mathcal{Z}$ 。根据定义,

$$\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) = \sup_{x, a} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)). \quad (10)$$

由  $d_p$  的性质, 我们有

$$\begin{aligned} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)) &= d_p(R(x, a) + \gamma P^\pi Z_1(x, a), R(x, a) + \gamma P^\pi Z_2(x, a)) \\ &\leq \gamma d_p(P^\pi Z_1(x, a), P^\pi Z_2(x, a)) \\ &\leq \gamma \sup_{x', a'} d_p(Z_1(x', a'), Z_2(x', a')), \end{aligned}$$

其中最后一行来自于  $P^\pi$  (的定义, 参见 (4))。结合 (10) 我们得到

$$\begin{aligned} \bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) &= \sup_{x, a} d_p(\mathcal{T}^\pi Z_1(x, a), \mathcal{T}^\pi Z_2(x, a)) \\ &\leq \gamma \sup_{x', a'} d_p(Z_1(x', a'), Z_2(x', a')) \\ &= \gamma \bar{d}_p(Z_1, Z_2). \quad \square \end{aligned}$$

命题 1 (Sobel, 1982). Consider two value distributions  $Z_1, Z_2 \in \mathcal{Z}$ , and write  $\mathbb{V}(Z_i)$  to be the vector of variances of  $Z_i$ . Then

$$\begin{aligned} \|\mathbb{E} \mathcal{T}^\pi Z_1 - \mathbb{E} \mathcal{T}^\pi Z_2\|_\infty &\leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty, \text{ and} \\ \|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty &\leq \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty. \end{aligned}$$

*Proof.* 第一个命题是标准的, 其证明来自于  $\mathbb{E} \mathcal{T}^\pi Z = \mathcal{T}^\pi \mathbb{E} Z$ , 其中第二个  $\mathcal{T}^\pi$  表示通常的价值函数运算符。现在, 由于  $R$  和  $P^\pi Z_i$  的独立性:

$$\begin{aligned} \mathbb{V}(\mathcal{T}^\pi Z_i(x, a)) &= \mathbb{V}(R(x, a) + \gamma P^\pi Z_i(x, a)) \\ &= \mathbb{V}(R(x, a)) + \gamma^2 \mathbb{V}(P^\pi Z_i(x, a)). \end{aligned}$$

现在

$$\begin{aligned} \|\mathbb{V}(\mathcal{T}^\pi Z_1) - \mathbb{V}(\mathcal{T}^\pi Z_2)\|_\infty &= \sup_{x, a} |\mathbb{V}(\mathcal{T}^\pi Z_1(x, a)) - \mathbb{V}(\mathcal{T}^\pi Z_2(x, a))| \\ &= \sup_{x, a} \gamma^2 |\mathbb{V}(P^\pi Z_1(x, a)) - \mathbb{V}(P^\pi Z_2(x, a))| \\ &= \sup_{x, a} \gamma^2 |\mathbb{E}[\mathbb{V}(Z_1(X', A')) - \mathbb{V}(Z_2(X', A'))]| \\ &\leq \sup_{x', a'} \gamma^2 |\mathbb{V}(Z_1(x', a')) - \mathbb{V}(Z_2(x', a'))| \\ &\leq \gamma^2 \|\mathbb{V} Z_1 - \mathbb{V} Z_2\|_\infty. \quad \square \end{aligned}$$

引理 4. Let  $Z_1, Z_2 \in \mathcal{Z}$ . Then

$$\|\mathbb{E} \mathcal{T} Z_1 - \mathbb{E} \mathcal{T} Z_2\|_\infty \leq \gamma \|\mathbb{E} Z_1 - \mathbb{E} Z_2\|_\infty,$$

and in particular  $\mathbb{E} Z_k \rightarrow Q^*$  exponentially quickly.

*Proof.* The proof follows by linearity of expectation. Write  $\mathcal{T}_D$  for the distributional operator and  $\mathcal{T}_E$  for the usual operator. Then

$$\begin{aligned} \|\mathbb{E} \mathcal{T}_D Z_1 - \mathbb{E} \mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E \mathbb{E} Z_1 - \mathcal{T}_E \mathbb{E} Z_2\|_\infty \\ &\leq \gamma \|Z_1 - Z_2\|_\infty. \quad \square \end{aligned}$$

**Theorem 1** (Convergence in the control setting). *Let  $Z_k := \mathcal{T} Z_{k-1}$  with  $Z_0 \in \mathcal{Z}$ . Let  $\mathcal{X}$  be measurable and suppose that  $\mathcal{A}$  is finite. Then*

$$\lim_{k \rightarrow \infty} \inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) = 0 \quad \forall x, a.$$

*If  $\mathcal{X}$  is finite, then  $Z_k$  converges to  $\mathcal{Z}^{**}$  uniformly. Furthermore, if there is a total ordering  $\prec$  on  $\Pi^*$ , such that for any  $Z^* \in \mathcal{Z}^*$ ,*

$$\mathcal{T} Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*}, \pi \prec \pi' \quad \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\},$$

*then  $\mathcal{T}$  has a unique fixed point  $Z^* \in \mathcal{Z}^*$ .*

The gist of the proof of Theorem 1 consists in showing that for every state  $x$ , there is a time  $k$  after which the greedy policy w.r.t.  $Q_k$  is mostly optimal. To clearly expose the steps involved, we will first assume a unique (and therefore deterministic) optimal policy  $\pi^*$ , and later return to the general case; we will denote the optimal action at  $x$  by  $\pi^*(x)$ . For notational convenience, we will write  $Q_k := \mathbb{E} Z_k$  and  $\mathcal{G}_k := \mathcal{G}_{Z_k}$ . Let  $B := 2 \sup_{Z \in \mathcal{Z}} \|Z\|_\infty < \infty$  and let  $\epsilon_k := \gamma^k B$ . We first define the set of states  $\mathcal{X}_k \subseteq \mathcal{X}$  whose values must be sufficiently close to  $Q^*$  at time  $k$ :

$$\mathcal{X}_k := \left\{ x : Q^*(x, \pi^*(x)) - \max_{a \neq \pi^*(x)} Q^*(x, a) > 2\epsilon_k \right\}. \quad (11)$$

Indeed, by Lemma 4, we know that after  $k$  iterations

$$|Q_k(x, a) - Q^*(x, a)| \leq \gamma^k |Q_0(x, a) - Q^*(x, a)| \leq \epsilon_k.$$

For  $x \in \mathcal{X}$ , write  $a^* := \pi^*(x)$ . For any  $a \in \mathcal{A}$ , we deduce that

$$Q_k(x, a^*) - Q_k(x, a) \geq Q^*(x, a^*) - Q^*(x, a) - 2\epsilon_k.$$

It follows that if  $x \in \mathcal{X}_k$ , then also  $Q_k(x, a^*) > Q_k(x, a')$  for all  $a' \neq \pi^*(x)$ : for these states, the greedy policy  $\pi_k(x) := \arg \max_a Q_k(x, a)$  corresponds to the optimal policy  $\pi^*$ .

**Lemma 5.** *For each  $x \in \mathcal{X}$  there exists a  $k$  such that, for all  $k' \geq k$ ,  $x \in \mathcal{X}_{k'}$ , and in particular  $\arg \max_a Q_{k'}(x, a) = \pi^*(x)$ .*

*Proof.* Because  $\mathcal{A}$  is finite, the gap

$$\Delta(x) := Q^*(x, \pi^*(x)) - \max_{a \neq \pi^*(x)} Q^*(x, a)$$

is attained for some strictly positive  $\Delta(x) > 0$ . By definition, there exists a  $k$  such that

$$\epsilon_k = \gamma^k B < \frac{\Delta(x)}{2},$$

and hence every  $x \in \mathcal{X}$  must eventually be in  $\mathcal{X}_k$ .  $\square$

This lemma allows us to guarantee the existence of an iteration  $k$  after which sufficiently many states are well-behaved, in the sense that the greedy policy at those states chooses the optimal action. We will call these states “solved”. We in fact require not only these states to be solved, but also most of their successors, and most of the successors of those, and so on. We formalize this notion as follows: fix some  $\delta > 0$ , let  $\mathcal{X}_{k,0} := \mathcal{X}_k$ , and define for  $i > 0$  the set

$$\mathcal{X}_{k,i} := \{x : x \in \mathcal{X}_k, P(\mathcal{X}_{k-1,i-1} | x, \pi^*(x)) \geq 1 - \delta\},$$

As the following lemma shows, any  $x$  is eventually contained in the recursively-defined sets  $\mathcal{X}_{k,i}$ , for any  $i$ .

**Lemma 6.** *For any  $i \in \mathbb{N}$  and any  $x \in \mathcal{X}$ , there exists a  $k$  such that for all  $k' \geq k$ ,  $x \in \mathcal{X}_{k',i}$ .*

*Proof.* Fix  $i$  and let us suppose that  $\mathcal{X}_{k,i} \uparrow \mathcal{X}$ . By Lemma 5, this is true for  $i = 0$ . We infer that for any probability measure  $P$  on  $\mathcal{X}$ ,  $P(\mathcal{X}_{k,i}) \rightarrow P(\mathcal{X}) = 1$ . In particular, for a given  $x \in \mathcal{X}_k$ , this implies that

$$P(\mathcal{X}_{k,i} | x, \pi^*(x)) \rightarrow P(\mathcal{X} | x, \pi^*(x)) = 1.$$

Therefore, for any  $x$ , there exists a time after which it is and remains a member of  $\mathcal{X}_{k,i+1}$ , the set of states for which  $P(\mathcal{X}_{k-1,i} | x, \pi^*(x)) \geq 1 - \delta$ . We conclude that  $\mathcal{X}_{k,i+1} \uparrow \mathcal{X}$  also. The statement follows by induction.  $\square$

*Proof of Theorem 1.* The proof is similar to policy iteration-type results, but requires more care in dealing with the metric and the possibly infinite state space. We will write  $W_k(x) := Z_k(x, \pi_k(x))$ , define  $W^*$  similarly and with some overload of notation write  $\mathcal{T} W_k(x) := W_{k+1}(x) = \mathcal{T} Z_k(x, \pi_{k+1}(x))$ . Finally, let  $S_i^k(x) := \mathbb{I}[x \in \mathcal{X}_{k,i}]$  and  $\bar{S}_i^k(x) = 1 - S_i^k(x)$ .

Fix  $i > 0$  and  $x \in \mathcal{X}_{k+1,i+1} \subseteq \mathcal{X}_k$ . We begin by using Lemma 1 to separate the transition from  $x$  into a solved term and an unsolved term:

$$P^{\pi_k} W_k(x) = S_i^k W_k(X') + \bar{S}_i^k W_k(X'),$$

where  $X'$  is the random successor from taking action  $\pi_k(x) := \pi^*(x)$ , and we write  $S_i^k = S_i^k(X')$ ,  $\bar{S}_i^k = \bar{S}_i^k(X')$  to ease the notation. Similarly,

$$P^{\pi_k} W^*(x) = S_i^k W^*(X') + \bar{S}_i^k W^*(X').$$

*Proof.* 证明通过期望的线性性得出。记  $\mathcal{T}_D$  为分布算子，记  $\mathcal{T}_E$  为通常的算子。然后

$$\begin{aligned}\|\mathbb{E} \mathcal{T}_D Z_1 - \mathbb{E} \mathcal{T}_D Z_2\|_\infty &= \|\mathcal{T}_E \mathbb{E} Z_1 - \mathcal{T}_E \mathbb{E} Z_2\|_\infty \\ &\leq \gamma \|Z_1 - Z_2\|_\infty.\end{aligned}\quad \square$$

定理 1 (控制设置下的收敛). *Let  $Z_k$  :  $= \mathcal{T} Z_{k-1}$  with  $Z_0 \in \mathcal{Z}$ . Let  $\mathcal{X}$  be measurable and suppose that  $\mathcal{A}$  is finite. Then*

$$\lim_{k \rightarrow \infty} \inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) = 0 \quad \forall x, a.$$

*If  $\mathcal{X}$  is finite, then  $Z_k$  converges to  $Z^{**}$  uniformly. Furthermore, if there is a total ordering  $\prec$  on  $\Pi^*$ , such that for any  $Z^* \in \mathcal{Z}^*$ ,*

$$\mathcal{T} Z^* = \mathcal{T}^\pi Z^* \text{ with } \pi \in \mathcal{G}_{Z^*}, \pi \prec \pi' \quad \forall \pi' \in \mathcal{G}_{Z^*} \setminus \{\pi\},$$

*then  $\mathcal{T}$  has a unique fixed point  $Z^* \in \mathcal{Z}^*$ .*

定理1的证明要点在于展示对于每一个状态 $x$ ，存在一个时间 $k$ ，之后相对于 $Q_k$ 的贪婪策略几乎是最佳的。为了清晰地展示涉及的步骤，我们首先假设一个唯一的（因此是确定性的）最优策略 $\pi^*$ ，之后再回到一般情况；我们将最优动作在 $x$ 时记为 $\pi^*(x)$ 。为了方便记号，我们写 $Q_k := \mathbb{E} Z_k$ 和 $\mathcal{G}_k := \mathcal{G}_{Z_k}$ 。令 $B := 2 \sup_{Z \in \mathcal{Z}} \|Z\|_\infty < \infty$ ，令 $\epsilon_k := \gamma^k B$ 。我们首先定义状态集合 $\mathcal{X}_k \subseteq \mathcal{X}$ ，这些状态在时间 $k$ 时的值必须足够接近 $Q^*$ ：

$$\mathcal{X}_k := \left\{ x : Q^*(x, \pi^*(x)) - \max_{a \neq \pi^*(x)} Q^*(x, a) > 2\epsilon_k \right\}. \quad (11)$$

确实，由引理4可知，在经过 $k$ 次迭代后

$$|Q_k(x, a) - Q^*(x, a)| \leq \gamma^k |Q_0(x, a) - Q^*(x, a)| \leq \epsilon_k.$$

对于 $x \in \mathcal{X}$ ，写出 $a^* := \pi^*(x)$ 。对于任意 $a \in \mathcal{A}$ ，我们得出“对于 $x \in \mathcal{X}$ ，写出 $a^* := \pi^*(x)$ 。对于任意 $a \in \mathcal{A}$ ，我们得出”

$$Q_k(x, a^*) - Q_k(x, a) \geq Q^*(x, a^*) - Q^*(x, a) - 2\epsilon_k.$$

它意味着如果 $x \in \mathcal{X}_k$ ，那么对于所有

$Q_k(x, a^*) > Q_k(x, a')$ ：对于这些状态，贪婪策略 $\pi_k(x) := \arg \max_a Q_k(x, a)$ 对应于最优策略 $\pi^*$ 。

引理 5. *For each  $x \in \mathcal{X}$  there exists a  $k$  such that, for all  $k' \geq k$ ,  $x \in \mathcal{X}_{k'}$ , and in particular  $\arg \max_a Q_k(x, a) = \pi^*(x)$ .*

*Proof.* 因为 $\mathcal{A}$ 是有限的，所以空隙

$$\Delta(x) := Q^*(x, \pi^*(x)) - \max_{a \neq \pi^*(x)} Q^*(x, a)$$

在某些严格正的 $\Delta(x) > 0$ 时达到最大值。根据定义，存在一个 $k$ 使得

$$\epsilon_k = \gamma^k B < \frac{\Delta(x)}{2},$$

因此，每项 $x \in \mathcal{X}$ 最终都必须在 $\mathcal{X}_k$ 中。  $\square$

这个引理允许我们保证在某个迭代 $k$ 之后，足够多的状态是行为良好的，也就是说，在这些状态下，贪婪策略会选择最优动作。我们将这些状态称为“已解决”。事实上，我们不仅要求这些状态是已解决的，还要求它们的大多数后继状态也是已解决的，这些后继状态的大多数后继状态也是已解决的，依此类推。我们将这种概念形式化如下：固定某个 $\delta > 0$ ，令 $\mathcal{X}_{k,0} := \mathcal{X}_k$ ，并定义对于 $i > 0$ 的集合

$$\mathcal{X}_{k,i} := \{x : x \in \mathcal{X}_k, P(\mathcal{X}_{k-1,i-1} | x, \pi^*(x)) \geq 1 - \delta\},$$

如以下引理所示，任何 $x$ 最终包含在递归定义的集合 $\mathcal{X}_{k,i}$ 中，对于任何 $i$ 。

引理 6. *For any  $i \in \mathbb{N}$  and any  $x \in \mathcal{X}$ , there exists a  $k$  such that for all  $k' \geq k$ ,  $x \in \mathcal{X}_{k',i}$ .*

*Proof.* Fix  $i$  并且假设 $\mathcal{X}_{k,i} \uparrow \mathcal{X}$ 。由引理 5，对于 $i = 0$ ，这是正确的。因此我们可以推断，对于 $P$ 上的任何概率测度 $\mathcal{X}$ ， $P(\mathcal{X}_{k,i}) \rightarrow P(\mathcal{X}) = 1$ 。特别是对于给定的 $x \in \mathcal{X}_k$ ，这表明

$$P(\mathcal{X}_{k,i} | x, \pi^*(x)) \rightarrow P(\mathcal{X} | x, \pi^*(x)) = 1.$$

因此，对于任何 $x$ ，存在一个时间点，在此之后它一直是且继续是 $\mathcal{X}_{k,i+1}$ 的一个成员，该集合的元素满足 $P(\mathcal{X}_{k-1,i} | x, \pi^*(x)) \geq 1 - \delta$ 。我们得出结论 $\mathcal{X}_{k,i+1} \uparrow \mathcal{X}$ 也满足。该陈述通过归纳得出。  $\square$

*Proof of Theorem 1.* 证明类似于政策迭代的结果，但需要在处理度量和可能的无限状态空间时更加小心。我们将写出 $W_k(x) := Z_k(x, \pi_k(x))$ ，类似地定义 $W^*$ 并使用一些符号重载写出 $\mathcal{T} W_k(x)$ ：

$$\begin{aligned}&= W_{k+1}(x) = \mathcal{T} Z_k(x, \pi_{k+1}(x)). \text{ 最后, 令 } S_i^k(x) : \\ &= \mathbb{I}[x \in \mathcal{X}_{k,i}] \text{ 和 } \bar{S}_i^k(x) = 1 - S_i^k(x).\end{aligned}$$

Fix  $i > 0$  和  $x \in \mathcal{X}_{k+1,i+1} \subseteq \mathcal{X}_{k,i}$  我们首先使用引理 1 将 $x$ 的转移分解为已解项和未解项：

$$P^{\pi_k} W_k(x) = S_i^k W_k(X') + \bar{S}_i^k W_k(X'),$$

其中 $X'$ 是采取动作 $\pi_k(x) := \pi^*(x)$ 的随机后继，我们用 $S_i^k = S_i^k(X')$ ,  $\bar{S}_i^k = \bar{S}_i^k(X')$ 来简化表示。类似地，

$$P^{\pi_k} W^*(x) = S_i^k W^*(X') + \bar{S}_i^k W^*(X').$$



Now

$$\begin{aligned}
 d_p(W_{k+1}(x), W^*(x)) &= d_p(\mathcal{T}W_k(x), \mathcal{T}W^*(x)) \\
 &\stackrel{(a)}{\leq} \gamma d_p(P^{\pi_k}W_k(x), P^{\pi^*}W^*(x)) \\
 &\stackrel{(b)}{\leq} \gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) \\
 &\quad + \gamma d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')), \tag{12}
 \end{aligned}$$

where in (a) we used Properties P1 and P2 of the Wasserstein metric, and in (b) we separate states for which  $\pi_k = \pi^*$  from the rest using Lemma 1 ( $\{S_i^k, \bar{S}_i^k\}$  form a partition of  $\Omega$ ). Let  $\delta_i := \Pr\{X' \notin \mathcal{X}_{k,i}\} = \mathbb{E}\{\bar{S}_i^k(X')\} = \|\bar{S}_i^k(X')\|_p$ . From property P3 of the Wasserstein metric, we have

$$\begin{aligned}
 d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')) &\leq \sup_{x'} d_p(\bar{S}_i^k(X')W_k(x'), \bar{S}_i^k(X')W^*(x')) \\
 &\leq \|\bar{S}_i^k(X')\|_p \sup_{x'} d_p(W_k(x'), W^*(x')) \\
 &\leq \delta_i \sup_{x'} d_p(W_k(x'), W^*(x')) \\
 &\leq \delta_i B.
 \end{aligned}$$

Recall that  $B < \infty$  is the largest attainable  $\|Z\|_\infty$ . Since also  $\delta_i < \delta$  by our choice of  $x \in \mathcal{X}_{k+1,i+1}$ , we can upper bound the second term in (12) by  $\gamma\delta B$ . This yields

$$\begin{aligned}
 d_p(W_{k+1}(x), W^*(x)) &\leq \\
 &\gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) + \gamma\delta B.
 \end{aligned}$$

By induction on  $i > 0$ , we conclude that for  $x \in \mathcal{X}_{k+i,i}$  and some random state  $X''$   $i$  steps forward,

$$\begin{aligned}
 d_p(W_{k+i}(x), W^*(x)) &\leq \\
 &\gamma^i d_p(S_0^k W_k(X''), S_0^k W^*(X'')) + \frac{\delta B}{1-\gamma} \\
 &\leq \gamma^i B + \frac{\delta B}{1-\gamma}.
 \end{aligned}$$

Hence for any  $x \in \mathcal{X}$ ,  $\epsilon > 0$ , we can take  $\delta, i$ , and finally  $k$  large enough to make  $d_p(W_k(x), W^*(x)) < \epsilon$ . The proof then extends to  $Z_k(x, a)$  by considering one additional application of  $\mathcal{T}$ .

We now consider the more general case where there are multiple optimal policies. We expand the definition of  $\mathcal{X}_{k,i}$  as follows:

$$\mathcal{X}_{k,i} := \{x \in \mathcal{X} : \forall \pi^* \in \Pi^*, \mathbb{E}_{a^* \sim \pi^*(x)} P(\mathcal{X}_{k-1,i-1} | x, a^*) \geq 1-\delta\},$$

Because there are finitely many actions, Lemma 6 also holds for this new definition. As before, take  $x \in \mathcal{X}_{k,i}$ , but now consider the sequence of greedy policies  $\pi_k, \pi_{k-1}, \dots$  selected by successive applications of  $\mathcal{T}$ , and write

$$\mathcal{T}^{\bar{\pi}_k} := \mathcal{T}^{\pi_k} \mathcal{T}^{\pi_{k-1}} \dots \mathcal{T}^{\pi_{k-i+1}},$$

such that

$$Z_{k+1} = \mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}.$$

Now denote by  $Z^{**}$  the set of nonstationary optimal policies. If we take any  $Z^* \in \mathcal{Z}^*$ , we deduce that

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \leq \frac{\delta B}{1-\gamma},$$

since  $Z^*$  corresponds to some optimal policy  $\pi^*$  and  $\bar{\pi}_k$  is optimal along most of the trajectories from  $(x, a)$ . In effect,  $\mathcal{T}^{\bar{\pi}_k} Z^*$  is close to the value distribution of the nonstationary optimal policy  $\bar{\pi}_k \pi^*$ . Now for this  $Z^*$ ,

$$\begin{aligned}
 \inf_{Z^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) &\leq d_p(Z_k(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) \\
 &\quad + \inf_{Z^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \\
 &\leq d_p(\mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) + \frac{\delta B}{1-\gamma} \\
 &\leq \gamma^i B + \frac{2\delta B}{1-\gamma},
 \end{aligned}$$

using the same argument as before with the newly-defined  $\mathcal{X}_{k,i}$ . It follows that

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) \rightarrow 0.$$

When  $\mathcal{X}$  is finite, there exists a fixed  $k$  after which  $\mathcal{X}_k = \mathcal{X}$ . The uniform convergence result then follows.

To prove the uniqueness of the fixed point  $Z^*$  when  $\mathcal{T}$  selects its actions according to the ordering  $\prec$ , we note that for any optimal value distribution  $Z^*$ , its set of greedy policies is  $\Pi^*$ . Denote by  $\pi^*$  the policy coming first in the ordering over  $\Pi^*$ . Then  $\mathcal{T} = \mathcal{T}^{\pi^*}$ , which has a unique fixed point (Section 3.3).  $\square$

**Proposition 4.** *That  $\mathcal{T}$  has a fixed point  $Z^* = \mathcal{T}Z^*$  is insufficient to guarantee the convergence of  $\{Z_k\}$  to  $Z^*$ .*

We provide here a sketch of the result. Consider a single state  $x_1$  with two actions,  $a_1$  and  $a_2$  (Figure 8). The first action yields a reward of  $1/2$ , while the other either yields 0 or 1 with equal probability, and both actions are optimal. Now take  $\gamma = 1/2$  and write  $R_0, R_1, \dots$  for the received rewards. Consider a stochastic policy that takes action  $a_2$  with probability  $p$ . For  $p = 0$ , the return is

$$Z_{p=0} = \frac{1}{1-\gamma} \frac{1}{2} = 1.$$

For  $p = 1$ , on the other hand, the return is random and is given by the following fractional number (in binary):

$$Z_{p=1} = R_0.R_1R_2R_3\dots$$

现在

$$\begin{aligned}
 d_p(W_{k+1}(x), W^*(x)) &= d_p(\mathcal{T}W_k(x), \mathcal{T}W^*(x)) \\
 &\stackrel{(a)}{\leq} \gamma d_p(P^{\pi_k}W_k(x), P^{\pi^*}W^*(x)) \\
 &\stackrel{(b)}{\leq} \gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) \\
 &\quad + \gamma d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')), \quad (12)
 \end{aligned}$$

其中在(a)中我们使用了Wasserstein度量的性质P1和P2，而在(b)中我们使用引理1将状态分为两类，使得 $\pi_k = \pi^*$ 与其他状态分开（ $\{S_i^k, \bar{S}_i^k\}$ 构成 $\Omega$ 的一个划分）。令 $\delta_i := \Pr\{X' \notin \mathcal{X}_{k,i}\} = \mathbb{E}\{\bar{S}_i^k(X')\} = \|\bar{S}_i^k(X')\|_p$ 。根据Wasserstein度量的性质P3，我们有

$$\begin{aligned}
 d_p(\bar{S}_i^k W_k(X'), \bar{S}_i^k W^*(X')) &\leq \sup_{x'} d_p(\bar{S}_i^k(X')W_k(x'), \bar{S}_i^k(X')W^*(x')) \\
 &\leq \|\bar{S}_i^k(X')\|_p \sup_{x'} d_p(W_k(x'), W^*(x')) \\
 &\leq \delta_i \sup_{x'} d_p(W_k(x'), W^*(x')) \\
 &\leq \delta_i B.
 \end{aligned}$$

回想一下， $B < \infty$  是可达到的最大值 $\|Z\|_\infty$ 。由于我们选择 $\delta_i < \delta$ ，因此我们可以用 $\gamma\delta B$ 来上界估计(12)中的第二项。这给出了

$$\begin{aligned}
 d_p(W_{k+1}(x), W^*(x)) &\leq \\
 &\gamma d_p(S_i^k W_k(X'), S_i^k W^*(X')) + \gamma\delta B.
 \end{aligned}$$

通过归纳 $i > 0$ ，我们得出结论，在 $x \in \mathcal{X}_{k+i,i}$ 和某些随机状态 $X''$   $i$ 步进后，

$$\begin{aligned}
 d_p(W_{k+i}(x), W^*(x)) &\leq \\
 &\gamma^i d_p(S_0^k W_k(X''), S_0^k W^*(X'')) + \frac{\delta B}{1-\gamma} \\
 &\leq \gamma^i B + \frac{\delta B}{1-\gamma}.
 \end{aligned}$$

因此对于任何 $x \in \mathcal{X}$ ， $\epsilon > 0$ ，我们可以取 $\delta, i$ ，最后取 $k$ 足够大以使 $d_p(W_k(x), W^*(x)) < \epsilon$ 。然后通过考虑 $\mathcal{T}$ 的额外一次应用，证明可以扩展到 $Z_k(x, a)$ 。

我们现在考虑更一般的情况，其中存在多个最优策略。我们将 $\{v^*\}$ 的定义扩展如下：

$$\mathcal{X}_{k,i} := \{x \in \mathcal{X} : \forall \pi^* \in \Pi^*, \mathbb{E} P(\mathcal{X}_{k-1,i-1} | x, a^*) \geq 1-\delta\},$$

因为有限多的动作，引理6也适用于这个新定义。如之前一样，取 $x \in \mathcal{X}_{k,i}$ ，但现在考虑通过依次应用 $\mathcal{T}$ 选择的贪婪策略序列 $\pi_k, \pi_{k-1}, \dots$ ，并写出

$$\mathcal{T}^{\bar{\pi}_k} := \mathcal{T}^{\pi_k} \mathcal{T}^{\pi_{k-1}} \dots \mathcal{T}^{\pi_{k-i+1}},$$

使得

$$Z_{k+1} = \mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}.$$

现在用 $Z^{**}$ 表示非稳态最优策略集。如果我们取任何一个 $Z^* \in \mathcal{Z}^*$ ，我们可以推导出

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \leq \frac{\delta B}{1-\gamma},$$

自从 $Z^*$ 对应于某个最优策略 $\pi^*$ ，并且 $\bar{\pi}_k$ 在 $(x, a)$ 的大多数轨迹上是最优的。实际上， $\mathcal{T}^{\bar{\pi}_k} Z^*$ 接近于非稳态最优策略 $\bar{\pi}_k \pi^*$ 的值分布。现在对于这个 $Z^*$ ，

$$\begin{aligned}
 \inf_{Z^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) &\leq d_p(Z_k(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) \\
 &\quad + \inf_{Z^{**}} d_p(\mathcal{T}^{\bar{\pi}_k} Z^*(x, a), Z^{**}(x, a)) \\
 &\leq d_p(\mathcal{T}^{\bar{\pi}_k} Z_{k-i+1}(x, a), \mathcal{T}^{\bar{\pi}_k} Z^*(x, a)) + \frac{\delta B}{1-\gamma} \\
 &\leq \gamma^i B + \frac{2\delta B}{1-\gamma},
 \end{aligned}$$

使用之前相同的论证方法，基于新定义的 $\mathcal{X}_{k,i}$ 。因此有

$$\inf_{Z^{**} \in \mathcal{Z}^{**}} d_p(Z_k(x, a), Z^{**}(x, a)) \rightarrow 0.$$

当 $\mathcal{X}$ 有限时，存在一个固定的 $k$ ，使得在 $\mathcal{X}_k = \mathcal{X}$ 之后成立。随后可以得出一致收敛的结果。

为了证明当 $\mathcal{T}$ 根据顺序 $\prec$ 选择其行动时，固定点 $Z^*$ 的唯一性，我们注意到对于任何最优值分布 $Z^*$ ，其贪心策略集为 $\Pi^*$ 。记在 $\Pi^*$ 的顺序中第一个出现的策略为 $\pi^*$ 。然后 $\mathcal{T} = \mathcal{T}^{\pi^*}$ 具有唯一的固定点（第3.3节）。□

命题4. *That  $\mathcal{T}$  has a fixed point  $Z^* = \mathcal{T}Z^*$  is insufficient to guarantee the convergence of  $\{Z_k\}$  to  $Z^*$ .*

我们提供了一个结果的大致描述。考虑一个单一状态 $x_1$ ，它有两个动作， $a_1$ 和 $a_2$ （图8）。第一个动作产生奖励1/2，而另一个动作以相等的概率产生0或1的奖励，且这两个动作都是最优的。现在取 $\gamma = 1/2$ ，记 $R_0, R_1, \dots$ 为收到的奖励。考虑一个随机策略，以概率 $p$ 采取动作 $a_2$ 。对于 $p = 0$ ，回报是

$$Z_{p=0} = \frac{1}{1-\gamma} \frac{1}{2} = 1.$$

对于 $p = 1$ ，回报是随机的，并由以下二进制分数给出：

$$Z_{p=1} = R_0.R_1R_2R_3\cdots.$$

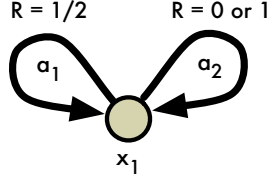


Figure 8. A simple example illustrating the effect of a nonstationary policy on the value distribution.

As a result,  $Z_{p=1}$  is uniformly distributed between 0 and 2! In fact, note that

$$Z_{p=0} = 0.11111 \dots = 1.$$

For some intermediary value of  $p$ , we obtain a different probability of the different digits, but always putting some probability mass on all returns in  $[0, 2]$ .

Now suppose we follow the nonstationary policy that takes  $a_1$  on the first step, then  $a_2$  from there on. By inspection, the return will be uniformly distributed on the interval  $[1/2, 3/2]$ , which does not correspond to the return under any value of  $p$ . But now we may imagine an operator  $\mathcal{T}$  which alternates between  $a_1$  and  $a_2$  depending on the exact value distribution it is applied to, which would in turn converge to a nonstationary optimal value distribution.

**Lemma 7** (Sample Wasserstein distance). *Let  $\{P_i\}$  be a collection of random variables,  $I \in \mathbb{N}$  a random index independent from  $\{P_i\}$ , and consider the mixture random variable  $P = P_I$ . For any random variable  $Q$  independent of  $I$ ,*

$$d_p(P, Q) \leq \mathbb{E}_{i \sim I} d_p(P_i, Q),$$

and in general the inequality is strict and

$$\nabla_Q d_p(P_I, Q) \neq \mathbb{E}_{i \sim I} \nabla_Q d_p(P_i, Q).$$

*Proof.* We prove this using Lemma 1. Let  $A_i := \mathbb{I}[I = i]$ . We write

$$\begin{aligned} d_p(P, Q) &= d_p(P_I, Q) \\ &= d_p\left(\sum_i A_i P_i, \sum_i A_i Q\right) \\ &\leq \sum_i d_p(A_i P_i, A_i Q) \\ &\leq \sum_i \Pr\{I = i\} d_p(P_i, Q) \\ &= \mathbb{E}_I d_p(P_i, Q). \end{aligned}$$

where in the penultimate line we used the independence of  $I$  from  $P_i$  and  $Q$  to appeal to property P3 of the Wasserstein metric.

To show that the bound is in general strict, consider the mixture distribution depicted in Figure 9. We will simply

consider the  $d_1$  metric between this distribution  $P$  and another distribution  $Q$ . The first distribution is

$$P = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2. \end{cases}$$

In this example,  $i \in \{1, 2\}$ ,  $P_1 = 0$ , and  $P_2 = 1$ . Now consider the distribution with the same support but that puts probability  $p$  on 0:

$$Q = \begin{cases} 0 & \text{w.p. } p \\ 1 & \text{w.p. } 1 - p. \end{cases}$$

The distance between  $P$  and  $Q$  is

$$d_1(P, Q) = |p - \frac{1}{2}|.$$

This is  $d_1(P, Q) = \frac{1}{2}$  for  $p \in \{0, 1\}$ , and strictly less than  $\frac{1}{2}$  for any other values of  $p$ . On the other hand, the corresponding expected distance (after sampling an outcome  $x_1$  or  $x_2$  with equal probability) is

$$\mathbb{E}_I d_1(P_i, Q) = \frac{1}{2}p + \frac{1}{2}(1 - p) = \frac{1}{2}.$$

Hence  $d_1(P, Q) < \mathbb{E}_I d_1(P_i, Q)$  for  $p \in (0, 1)$ . This shows that the bound is in general strict. By inspection, it is clear that the two gradients are different.  $\square$

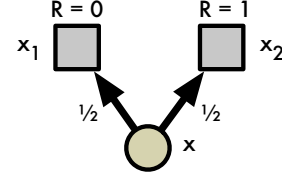


Figure 9. Example MDP in which the expected sample Wasserstein distance is greater than the Wasserstein distance.

**Proposition 5.** *Fix some next-state distribution  $Z$  and policy  $\pi$ . Consider a parametric value distribution  $Z_\theta$ , and and define the Wasserstein loss*

$$\mathcal{L}_W(\theta) := d_p(Z_\theta(x, a), R(x, a) + \gamma Z(X', \pi(X'))).$$

Let  $r \sim R(x, a)$  and  $x' \sim P(\cdot | x, a)$  and consider the sample loss

$$L_W(\theta, r, x') := d_p(Z_\theta(x, a), r + \gamma Z(x', \pi(x'))).$$

Its expectation is an upper bound on the loss  $\mathcal{L}_W$ :

$$\mathcal{L}_W(\theta) \leq \mathbb{E}_{R, P} L_W(\theta, r, x'),$$

in general with strict inequality.

The result follows directly from the previous lemma.

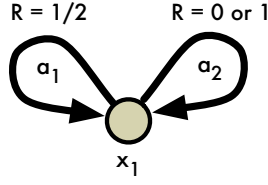


图 8. 一个简单的例子，说明非平稳策略对价值分布的影响。

因此， $Z_{p=1}$  在 0 和 2 之间均匀分布！实际上，注意 rān g

$$Z_{p=0} = 0.11111 \dots = 1.$$

对于某些中介值  $p$ ，我们获得不同的各位数字的概率，但总是将概率质量分配到所有在  $[0, 2]$  范围内的回报中。

现在假设我们遵循一个非平稳策略，在第一步采取  $a_1$ ，然后从那以后采取  $a_2$ 。通过检查，回报将在区间  $[1/2, 3/2]$  上均匀分布，这并不对应于任何  $p$  的值。但是现在我们可以想象一个操作符  $\mathcal{T}$ ，它在应用到不同的确切值分布时交替使用  $a_1$  和  $a_2$ ，这反过来会收敛到一个非平稳的最优值分布。

引理 7 (样本 Wasserstein 距离)。Let  $\{P_i\}$  be a collection of random variables,  $I \in \mathbb{N}$  a random index independent from  $\{P_i\}$ , and consider the mixture random variable  $P = P_I$ . For any random variable  $Q$  independent of  $I$ ,

$$d_p(P, Q) \leq \mathbb{E}_{i \sim I} d_p(P_i, Q),$$

and in general the inequality is strict and

$$\nabla_Q d_p(P_I, Q) \neq \mathbb{E}_{i \sim I} \nabla_Q d_p(P_i, Q).$$

*Proof.* 我们使用引理 1 来证明这一点。令  $A_i := \mathbb{I}[I = i]$ 。我们写出

$$\begin{aligned} d_p(P, Q) &= d_p(P_I, Q) \\ &= d_p\left(\sum_i A_i P_i, \sum_i A_i Q\right) \\ &\leq \sum_i d_p(A_i P_i, A_i Q) \\ &\leq \sum_i \Pr\{I = i\} d_p(P_i, Q) \\ &= \mathbb{E}_I d_p(P_i, Q). \end{aligned}$$

在倒数第二行，我们利用了  $I$  与  $P_i$  和  $Q$  的独立性，引用了 Wasserstein 距离的性质 P3。

为了说明该界一般是严格的，考虑图 9 中所示的混合分布。我们将简单地

考虑该分布  $d_1$  与另一个分布  $P$  之间的  $Q$  度量。第一个分布是

$$P = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2. \end{cases}$$

在这个例子中， $i \in \{1, 2\}$ ， $P_1 = 0$ ，和  $P_2 = 1$ 。现在考虑具有相同支持但将概率  $p$  放在 0 上的分布：

$$Q = \begin{cases} 0 & \text{w.p. } p \\ 1 & \text{w.p. } 1 - p. \end{cases}$$

$P$  和  $Q$  之间的距离是

$$d_1(P, Q) = |p - \frac{1}{2}|.$$

这是  $d_1(P, Q) = \frac{1}{2}$  对于  $p \in \{0, 1\}$ ，并且对于  $p$  的其他任何值都严格小于  $\frac{1}{2}$ 。另一方面，相应的期望距离（在以相等概率采样结果  $x_1$  或  $x_2$  后）是

$$\mathbb{E}_I d_1(P_i, Q) = \frac{1}{2}p + \frac{1}{2}(1 - p) = \frac{1}{2}.$$

因此  $d_1(P, Q) < \mathbb{E}_I d_1(P_i, Q)$  对于  $p \in (0, 1)$ 。这表明边界通常是严格的。通过检查，很明显两个梯度是不同的。  $\square$

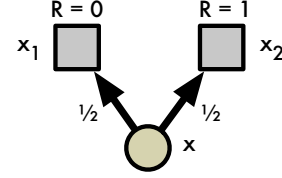


图 9. 一个示例 MDP，其中期望采样 Wasserstein 距离大于 Wasserstein 距离。

命题 5. Fix some next-state distribution  $Z$  and policy  $\pi$ . Consider a parametric value distribution  $Z_\theta$ , and and define the Wasserstein loss

$$\mathcal{L}_W(\theta) := d_p(Z_\theta(x, a), R(x, a) + \gamma Z(X', \pi(X'))).$$

Let  $r \sim R(x, a)$  and  $x' \sim P(\cdot | x, a)$  and consider the sample loss

$$L_W(\theta, r, x') := d_p(Z_\theta(x, a), r + \gamma Z(x', \pi(x'))).$$

Its expectation is an upper bound on the loss  $\mathcal{L}_W$ :

$$\mathcal{L}_W(\theta) \leq \mathbb{E}_{R, P} L_W(\theta, r, x'),$$

in general with strict inequality.

结果直接来自于之前的引理。

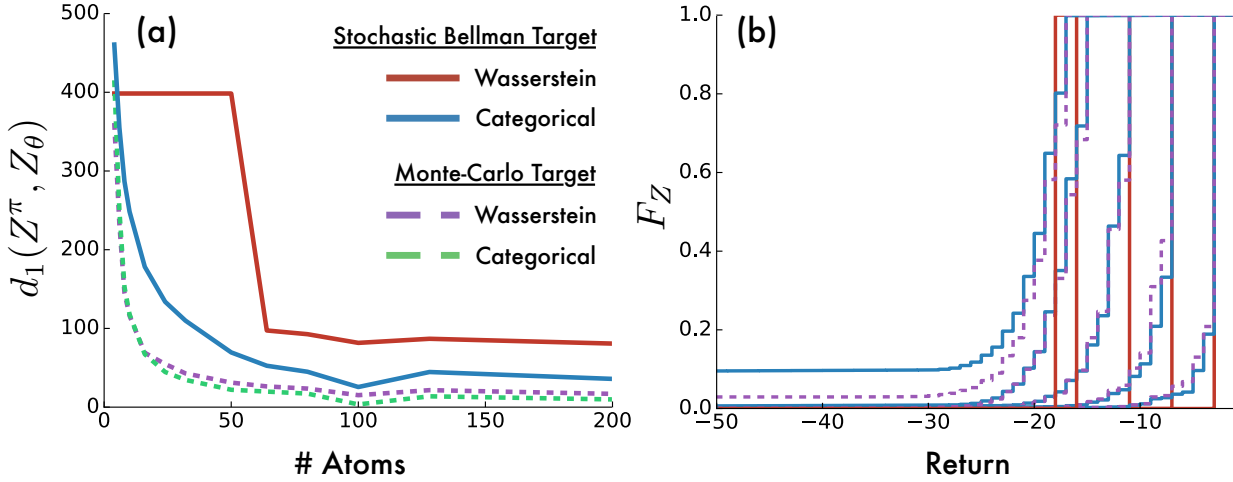


Figure 10. (a) Wasserstein distance between ground truth distribution  $Z^\pi$  and approximating distributions  $Z_\theta$ . Varying number of atoms in approximation, training target, and loss function. (b) Approximate cumulative distributions for five representative states in CliffWalk.

### C. Algorithmic Details

While our training regime closely follows that of DQN (Mnih et al., 2015), we use Adam (Kingma & Ba, 2015) instead of RMSProp (Tieleman & Hinton, 2012) for gradient rescaling. We also performed some hyperparameter tuning for our final results. Specifically, we evaluated two hyperparameters over our five training games and choose the values that performed best. The hyperparameter values we considered were  $V_{\text{MAX}} \in \{3, 10, 100\}$  and  $\epsilon_{\text{adam}} \in \{1/L, 0.1/L, 0.01/L, 0.001/L, 0.0001/L\}$ , where  $L = 32$  is the minibatch size. We found  $V_{\text{MAX}} = 10$  and  $\epsilon_{\text{adam}} = 0.01/L$  performed best. We used the same step-size value as DQN ( $\alpha = 0.00025$ ).

Pseudo-code for the categorical algorithm is given in Algorithm 1. We apply the Bellman update to each atom separately, and then project it into the two nearest atoms in the original support. Transitions to a terminal state are handled with  $\gamma_t = 0$ .

### D. Comparison of Sampled Wasserstein Loss and Categorical Projection

Lemma 3 proves that for a fixed policy  $\pi$  the distributional Bellman operator is a  $\gamma$ -contraction in  $\bar{d}_p$ , and therefore that  $\mathcal{T}^\pi$  will converge in distribution to the true distribution of returns  $Z^\pi$ . In this section, we empirically validate these results on the CliffWalk domain shown in Figure 11. The dynamics of the problem match those given by Sutton & Barto (1998). We also study the convergence of the distributional Bellman operator under the sampled Wasserstein loss and the categorical projection (Equation 7) while fol-

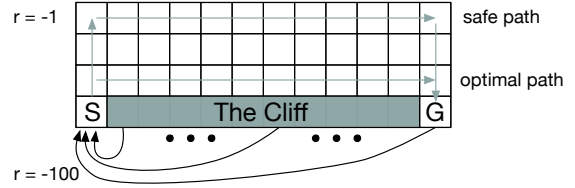


Figure 11. CliffWalk Environment (Sutton & Barto, 1998).

lowing a policy that tries to take the safe path but has a 10% chance of taking another action uniformly at random.

We compute a ground-truth distribution of returns  $Z^\pi$  using 10000 Monte-Carlo (MC) rollouts from each state. We then perform two experiments, approximating the value distribution at each state with our discrete distributions.

In the first experiment, we perform supervised learning using either the Wasserstein loss or categorical projection (Equation 7) with cross-entropy loss. We use  $Z^\pi$  as the supervised target and perform 5000 sweeps over all states to ensure both approaches have converged. In the second experiment, we use the same loss functions, but the training target comes from the one-step distributional Bellman operator with sampled transitions. We use  $V_{\text{MIN}} = -100$  and  $V_{\text{MAX}} = -1.4$ . For the sample updates we perform 10 times as many sweeps over the state space. Fundamentally, these experiments investigate how well the two training regimes

<sup>4</sup>Because there is a small probability of larger negative returns, some approximation error is unavoidable. However, this effect is relatively negligible in our experiments.

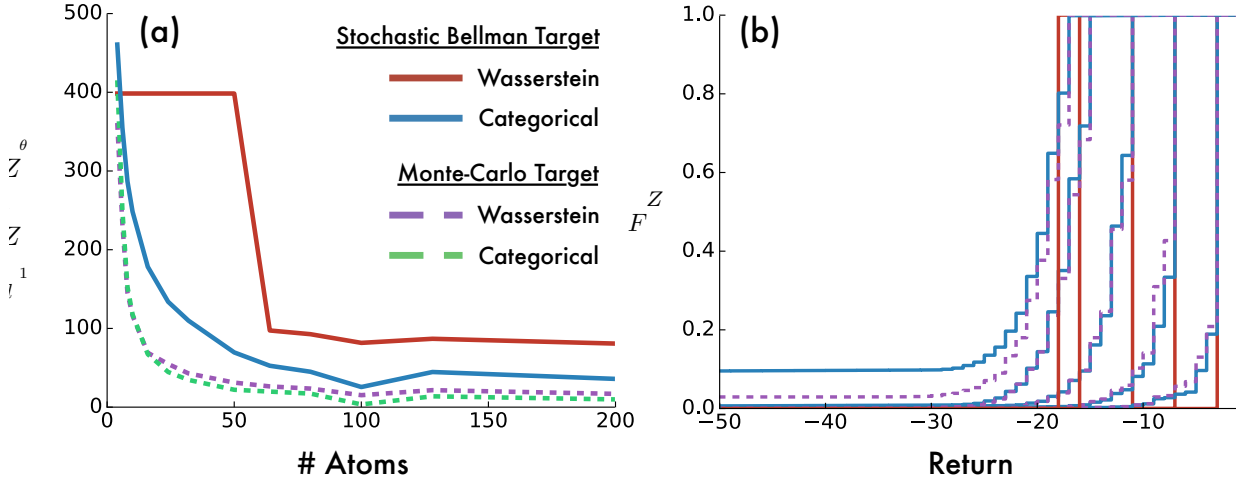


图10. (a) 地面真实分布  $Z^\pi$  与逼近分布  $Z_\theta$  之间的 Wasserstein 距离。变化逼近中的原子数、训练目标和损失函数。(b) CliffWalk 中五个代表性状态的逼近累积分布。

### C. 算法细节

虽然我们的训练制度与DQN (Mnih等, 2015) 相近, 但我们使用Adam (Kingma & Ba, 2015) 进行梯度重新缩放, 而不是RMSProp (Tieleman & Hinton, 2012)。我们还对最终结果进行了超参数调整。具体来说, 我们在五种训练游戏中评估了两个超参数, 并选择了表现最好的值。我们考虑的超参数值为  $V_{\text{MAX}} \in \{3, 10, 100\}$  和  $\epsilon_{\text{adam}} \in \{1/L, 0.1/L, 0.01/L, 0.001/L, 0.0001/L\}$ , 其中  $L = 32$  是小批量大小。我们发现  $V_{\text{MAX}} = 10$  和  $\epsilon_{\text{adam}} = 0.01/L$  表现最好。我们使用与DQN相同的步长值 ( $\alpha = 0.00025$ )。

分类算法的伪代码如算法1所示。我们分别对每个原子应用贝尔曼更新, 然后将其投影到原始支持的两个最近的原子中。转移到终端状态使用  $\gamma_t = 0$  处理。

### D. 样本 Wasserstein 损失与分类投影的比较

引理 3 证明, 在固定策略  $\pi$  的情况下, 分布 Bellman 运算符在  $\bar{d}_p$  中是  $\gamma$ -收缩的, 因此  $\mathcal{T}^\pi$  将以分布收敛到真实的回报分布  $Z^\pi$ 。在本节中, 我们通过图 11 所示的 CliffWalk 域上的实证结果验证了这些结果。该问题的动力学与 Sutton & Barto (1998) 给出的一致。我们还研究了在采样 Wasserstein 损失和分类投影 (方程 7) 下分布 Bellman 运算符的收敛性。

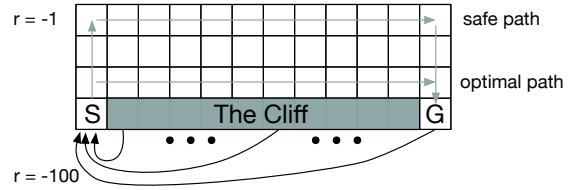


图11. 悬崖行走环境 (萨顿与巴托, 1998)。

遵循一个政策, 试图走安全路径, 但有10%的机会采取另一种随机行动。

我们使用每个状态进行10000次蒙特卡洛(MC)滚动来计算一个真实回报分布  $Z^\pi$ 。然后我们进行两个实验, 在每个状态下用我们的离散分布来逼近价值分布。

在第一个实验中, 我们使用Wasserstein损失或类别投影 (方程7) 与交叉熵损失进行监督学习。我们使用  $Z^\pi$  作为监督目标, 并对所有状态进行5000次迭代以确保两种方法均已收敛。在第二个实验中, 我们使用相同的损失函数, 但训练目标来自采样过渡的一步分布贝尔曼运算符。我们使用  $V_{\text{MIN}} = -100$  和  $V_{\text{MAX}} = -1$ 。<sup>4</sup> 对于样本更新, 我们在状态空间中进行10倍的迭代次数。从根本上说, 这些实验研究了两种训练制度的性能如何。

<sup>4</sup>Because there is a small probability of larger negative returns, some approximation error is unavoidable. However, this effect is relatively negligible in our experiments.



(minimizing the Wasserstein or categorical loss) minimize the Wasserstein metric under both ideal (supervised target) and practical (sampled one-step Bellman target) conditions.

In Figure 10a we show the final Wasserstein distance  $d_1(Z^\pi, Z_\theta)$  between the learned distributions and the ground-truth distribution as we vary the number of atoms. The graph shows that the categorical algorithm does indeed minimize the Wasserstein metric in both the supervised and sample Bellman setting. It also highlights that minimizing the Wasserstein loss with stochastic gradient descent is in general flawed, confirming the intuition given by Proposition 5. In repeat experiments the process converged to different values of  $d_1(Z^\pi, Z_\theta)$ , suggesting the presence of local minima (more prevalent with fewer atoms).

Figure 10 provides additional insight into why the sampled Wasserstein distance may perform poorly. Here, we see the cumulative densities for the approximations learned under these two losses for five different states along the safe path in CliffWalk. The Wasserstein has converged to a fixed-point distribution, but not one that captures the true (Monte Carlo) distribution very well. By comparison, the categorical algorithm captures the variance of the true distribution much more accurately.

## E. Supplemental Videos and Results

In Figure 13 we provide links to supplemental videos showing the C51 agent during training on various Atari 2600 games. Figure 12 shows the relative performance of C51 over the course of training. Figure 14 provides a table of evaluation results, comparing C51 to other state-of-the-art agents. Figures 15–18 depict particularly interesting frames.

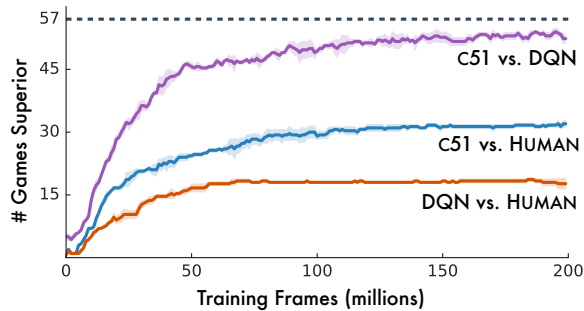


Figure 12. Number of Atari games where an agent’s training performance is greater than a baseline (fully trained DQN & human). Error bands give standard deviations, and averages are over number of games.

GAMES	VIDEO URL
Freeway	<a href="http://youtu.be/97578n9kFIk">http://youtu.be/97578n9kFIk</a>
Pong	<a href="http://youtu.be/vIz5P6s80qA">http://youtu.be/vIz5P6s80qA</a>
Q*Bert	<a href="http://youtu.be/v-RbNX4uETw">http://youtu.be/v-RbNX4uETw</a>
Seaquest	<a href="http://youtu.be/d1yz4PNFUjI">http://youtu.be/d1yz4PNFUjI</a>
Space Invaders	<a href="http://youtu.be/yFBwyPuO2Vg">http://youtu.be/yFBwyPuO2Vg</a>

Figure 13. Supplemental videos of C51 during training.

(最小化Wasserstein或分类损失)在理想（监督目标）和实际（采样的一步贝尔曼目标）条件下最小化Wasserstein度量。

在图10a中，我们展示了随原子数量变化时学习分布与真实分布之间的最终Wasserstein距离 $d_1(Z^\pi, Z_\theta)$ 。图表显示，分类算法确实能够在监督和样本贝尔曼设置中最小化Wasserstein度量。它还强调了使用随机梯度下降最小化Wasserstein损失通常是不完善的，这证实了命题5给出的直觉。在重复实验中，该过程收敛到不同的 $d_1(Z^\pi, Z_\theta)$ 值，表明存在局部极小值（尤其是在原子数量较少时更为常见）。

图10提供了关于为什么采样的Wasserstein距离可能表现不佳的额外见解。在这里，我们看到了在这些两种损失下学习的近似值沿CliffWalk安全路径的五种不同状态的累积密度。Wasserstein已经收敛到一个固定的分布，但并不是一个能够很好地捕捉真实（蒙特卡洛）分布的分布。相比之下，分类算法更准确地捕捉了真实分布的方差。

## E. 补充视频和结果

在图13中，我们提供了补充视频的链接，展示了C51代理在各种Atari 2600游戏中训练的过程。图12显示了C51在整个训练过程中的相对性能。图14提供了一个比较C51与其他先进代理的评估结果表格。图15-18展示了特别有趣的帧。

GAMES	VIDEO URL
Freeway	<a href="http://youtu.be/97578n9kFIk">http://youtu.be/97578n9kFIk</a>
Pong	<a href="http://youtu.be/vIz5P6s80qA">http://youtu.be/vIz5P6s80qA</a>
Q*Bert	<a href="http://youtu.be/v-RbNX4uETw">http://youtu.be/v-RbNX4uETw</a>
Seaquest	<a href="http://youtu.be/d1yz4PNFUjI">http://youtu.be/d1yz4PNFUjI</a>
Space Invaders	<a href="http://youtu.be/yFBwyPuO2Vg">http://youtu.be/yFBwyPuO2Vg</a>

图13. C51在训练期间的补充视频。

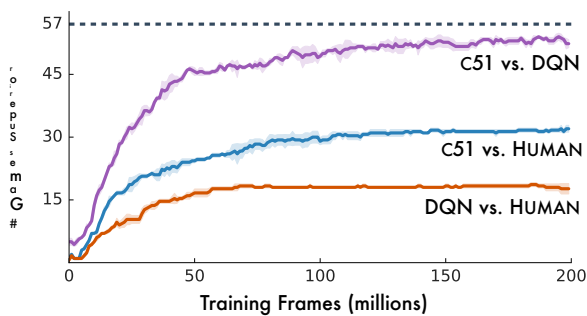


图12. 在 Atari 游戏中，代理训练性能超过基线（完全训练的DQN 和人类）的游戏数量。误差带给出标准差，平均值是基于游戏数量计算的。

# A Distributional Perspective on Reinforcement Learning

GAMES	RANDOM	HUMAN	DQN	DDQN	DUEL	PRIOR. DUEL.	C51
Alien	227.8	7,127.7	1,620.0	3,747.7	4,461.4	3,941.0	3,166
Amidar	5.8	1,719.5	978.0	1,793.3	2,354.5	2,296.8	1,735
Assault	222.4	742.0	4,280.4	5,393.2	4,621.0	11,477.0	7,203
Asterix	210.0	8,503.3	4,359.0	17,356.5	28,188.0	375,080.0	406,211
Asteroids	719.1	47,388.7	1,364.5	734.7	2,837.7	1,192.7	1,516
Atlantis	12,850.0	29,028.1	279,987.0	106,056.0	382,572.0	395,762.0	841,075
Bank Heist	14.2	753.1	455.0	1,030.6	1,611.9	1,503.1	976
Battle Zone	2,360.0	37,187.5	29,900.0	31,700.0	37,150.0	35,520.0	28,742
Beam Rider	363.9	16,926.5	8,627.5	13,772.8	12,164.0	30,276.5	14,074
Berzerk	123.7	2,630.4	585.6	1,225.4	1,472.6	3,409.0	1,645
Bowling	23.1	160.7	50.4	68.1	65.5	46.7	81.8
Boxing	0.1	12.1	88.0	91.6	99.4	98.9	97.8
Breakout	1.7	30.5	385.5	418.5	345.3	366.0	748
Centipede	2,090.9	12,017.0	4,657.7	5,409.4	7,561.4	7,687.5	9,646
Chopper Command	811.0	7,387.8	6,126.0	5,809.0	11,215.0	13,185.0	15,600
Crazy Climber	10,780.5	35,829.4	110,763.0	117,282.0	143,570.0	162,224.0	179,877
Defender	2,874.5	18,688.9	23,633.0	35,338.5	42,214.0	41,324.5	47,092
Demon Attack	152.1	1,971.0	12,149.4	58,044.2	60,813.3	72,878.6	130,955
Double Dunk	-18.6	-16.4	-6.6	-5.5	0.1	-12.5	2.5
Enduro	0.0	860.5	729.0	1,211.8	2,258.2	2,306.4	3,454
Fishing Derby	-91.7	-38.7	-4.9	15.5	46.4	41.3	8.9
Freeway	0.0	29.6	30.8	33.3	0.0	33.0	33.9
Frostbite	65.2	4,334.7	797.4	1,683.3	4,672.8	7,413.0	3,965
Gopher	257.6	2,412.5	8,777.4	14,840.8	15,718.4	104,368.2	33,641
Gravitar	173.0	3,351.4	473.0	412.0	588.0	238.0	440
H.E.R.O.	1,027.0	30,826.4	20,437.8	20,130.2	20,818.2	21,036.5	38,874
Ice Hockey	-11.2	0.9	-1.9	-2.7	0.5	-0.4	-3.5
James Bond	29.0	302.8	768.5	1,358.0	1,312.5	812.0	1,909
Kangaroo	52.0	3,035.0	7,259.0	12,992.0	14,854.0	1,792.0	12,853
Krull	1,598.0	2,665.5	8,422.3	7,920.5	11,451.9	10,374.4	9,735
Kung-Fu Master	258.5	22,736.3	26,059.0	29,710.0	34,294.0	48,375.0	48,192
Montezuma's Revenge	0.0	4,753.3	0.0	0.0	0.0	0.0	0.0
Ms. Pac-Man	307.3	6,951.6	3,085.6	2,711.4	6,283.5	3,327.3	3,415
Name This Game	2,292.3	8,049.0	8,207.8	10,616.0	11,971.1	15,572.5	12,542
Phoenix	761.4	7,242.6	8,485.2	12,252.5	23,092.2	70,324.3	17,490
Pitfall!	-229.4	6,463.7	-286.1	-29.9	0.0	0.0	0.0
Pong	-20.7	14.6	19.5	20.9	21.0	20.9	20.9
Private Eye	24.9	69,571.3	146.7	129.7	103.0	206.0	15,095
Q*Bert	163.9	13,455.0	13,117.3	15,088.5	19,220.3	18,760.3	23,784
River Raid	1,338.5	17,118.0	7,377.6	14,884.5	21,162.6	20,607.6	17,322
Road Runner	11.5	7,845.0	39,544.0	44,127.0	69,524.0	62,151.0	55,839
Robotank	2.2	11.9	63.9	65.1	65.3	27.5	52.3
Seaquest	68.4	42,054.7	5,860.6	16,452.7	50,254.2	931.6	266,434
Skiing	-17,098.1	-4,336.9	-13,062.3	-9,021.8	-8,857.4	-19,949.9	-13,901
Solaris	1,236.3	12,326.7	3,482.8	3,067.8	2,250.8	133.4	8,342
Space Invaders	148.0	1,668.7	1,692.3	2,525.5	6,427.3	15,311.5	5,747
Star Gunner	664.0	10,250.0	54,282.0	60,142.0	89,238.0	125,117.0	49,095
Surround	-10.0	6.5	-5.6	-2.9	4.4	1.2	6.8
Tennis	-23.8	-8.3	12.2	-22.8	5.1	0.0	23.1
Time Pilot	3,568.0	5,229.2	4,870.0	8,339.0	11,666.0	7,553.0	8,329
Tutankham	11.4	167.6	68.1	218.4	211.4	245.9	280
Up and Down	533.4	11,693.2	9,989.9	22,972.2	44,939.6	33,879.1	15,612
Venture	0.0	1,187.5	163.0	98.0	497.0	48.0	1,520
Video Pinball	16,256.9	17,667.9	196,760.4	309,941.9	98,209.5	479,197.0	949,604
Wizard Of Wor	563.5	4,756.5	2,704.0	7,492.0	7,855.0	12,352.0	9,300
Yars' Revenge	3,092.9	54,576.9	18,098.9	11,712.6	49,622.1	69,618.1	35,050
Zaxxon	32.5	9,173.3	5,363.0	10,163.0	12,944.0	13,886.0	10,513

Figure 14. Raw scores across all games, starting with 30 no-op actions. Reference values from Wang et al. (2016).

GAMES	RANDOM	HUMAN	DQN	DDQN	DUEL	PRIOR. DUEL.	C51
Alien	227.8	<b>7,127.7</b>	1,620.0	3,747.7	4,461.4	3,941.0	3,166
Amidar	5.8	1,719.5	978.0	1,793.3	<b>2,354.5</b>	2,296.8	1,735
Assault	222.4	742.0	4,280.4	5,393.2	4,621.0	<b>11,477.0</b>	7,203
Asterix	210.0	8,503.3	4,359.0	17,356.5	28,188.0	375,080.0	<b>406,211</b>
Asteroids	719.1	<b>47,388.7</b>	1,364.5	734.7	2,837.7	1,192.7	1,516
Atlantis	12,850.0	29,028.1	279,987.0	106,056.0	382,572.0	395,762.0	<b>841,075</b>
Bank Heist	14.2	753.1	455.0	1,030.6	<b>1,611.9</b>	1,503.1	976
Battle Zone	2,360.0	<b>37,187.5</b>	29,900.0	31,700.0	37,150.0	35,520.0	28,742
Beam Rider	363.9	16,926.5	8,627.5	13,772.8	12,164.0	<b>30,276.5</b>	14,074
Berzerk	123.7	2,630.4	585.6	1,225.4	1,472.6	<b>3,409.0</b>	1,645
Bowling	23.1	<b>160.7</b>	50.4	68.1	65.5	46.7	81.8
Boxing	0.1	12.1	88.0	91.6	<b>99.4</b>	98.9	97.8
Breakout	1.7	30.5	385.5	418.5	345.3	366.0	<b>748</b>
Centipede	2,090.9	<b>12,017.0</b>	4,657.7	5,409.4	7,561.4	7,687.5	9,646
Chopper Command	811.0	7,387.8	6,126.0	5,809.0	11,215.0	13,185.0	<b>15,600</b>
Crazy Climber	10,780.5	35,829.4	110,763.0	117,282.0	143,570.0	162,224.0	<b>179,877</b>
Defender	2,874.5	18,688.9	23,633.0	35,338.5	42,214.0	41,324.5	<b>47,092</b>
Demon Attack	152.1	1,971.0	12,149.4	58,044.2	60,813.3	72,878.6	<b>130,955</b>
Double Dunk	-18.6	-16.4	-6.6	-5.5	0.1	-12.5	<b>2.5</b>
Enduro	0.0	860.5	729.0	1,211.8	2,258.2	2,306.4	<b>3,454</b>
Fishing Derby	-91.7	-38.7	-4.9	15.5	<b>46.4</b>	41.3	8.9
Freeway	0.0	29.6	30.8	33.3	0.0	33.0	<b>33.9</b>
Frostbite	65.2	4,334.7	797.4	1,683.3	4,672.8	<b>7,413.0</b>	3,965
Gopher	257.6	2,412.5	8,777.4	14,840.8	15,718.4	<b>104,368.2</b>	33,641
Gravitar	173.0	<b>3,351.4</b>	473.0	412.0	588.0	238.0	440
H.E.R.O.	1,027.0	30,826.4	20,437.8	20,130.2	20,818.2	21,036.5	<b>38,874</b>
Ice Hockey	-11.2	<b>0.9</b>	-1.9	-2.7	0.5	-0.4	-3.5
James Bond	29.0	302.8	768.5	1,358.0	1,312.5	812.0	<b>1,909</b>
Kangaroo	52.0	3,035.0	7,259.0	12,992.0	<b>14,854.0</b>	1,792.0	12,853
Krull	1,598.0	2,665.5	8,422.3	7,920.5	<b>11,451.9</b>	10,374.4	9,735
Kung-Fu Master	258.5	22,736.3	26,059.0	29,710.0	34,294.0	<b>48,375.0</b>	48,192
Montezuma's Revenge	0.0	<b>4,753.3</b>	0.0	0.0	0.0	0.0	0.0
Ms. Pac-Man	307.3	<b>6,951.6</b>	3,085.6	2,711.4	6,283.5	3,327.3	3,415
Name This Game	2,292.3	8,049.0	8,207.8	10,616.0	11,971.1	<b>15,572.5</b>	12,542
Phoenix	761.4	7,242.6	8,485.2	12,252.5	23,092.2	<b>70,324.3</b>	17,490
Pitfall!	-229.4	<b>6,463.7</b>	-286.1	-29.9	0.0	0.0	0.0
Pong	-20.7	14.6	19.5	20.9	<b>21.0</b>	20.9	20.9
Private Eye	24.9	<b>69,571.3</b>	146.7	129.7	103.0	206.0	15,095
Q*Bert	163.9	13,455.0	13,117.3	15,088.5	19,220.3	18,760.3	<b>23,784</b>
River Raid	1,338.5	17,118.0	7,377.6	14,884.5	<b>21,162.6</b>	20,607.6	17,322
Road Runner	11.5	7,845.0	39,544.0	44,127.0	<b>69,524.0</b>	62,151.0	55,839
Robotank	2.2	11.9	63.9	65.1	<b>65.3</b>	27.5	52.3
Seaquest	68.4	42,054.7	5,860.6	16,452.7	50,254.2	931.6	<b>266,434</b>
Skiing	-17,098.1	<b>-4,336.9</b>	-13,062.3	-9,021.8	-8,857.4	-19,949.9	-13,901
Solaris	1,236.3	<b>12,326.7</b>	3,482.8	3,067.8	2,250.8	133.4	8,342
Space Invaders	148.0	1,668.7	1,692.3	2,525.5	6,427.3	<b>15,311.5</b>	5,747
Star Gunner	664.0	10,250.0	54,282.0	60,142.0	89,238.0	<b>125,117.0</b>	49,095
Surround	-10.0	6.5	-5.6	-2.9	4.4	1.2	<b>6.8</b>
Tennis	-23.8	-8.3	12.2	-22.8	5.1	0.0	<b>23.1</b>
Time Pilot	3,568.0	5,229.2	4,870.0	8,339.0	<b>11,666.0</b>	7,553.0	8,329
Tutankham	11.4	167.6	68.1	218.4	211.4	245.9	<b>280</b>
Up and Down	533.4	11,693.2	9,989.9	22,972.2	<b>44,939.6</b>	33,879.1	15,612
Venture	0.0	1,187.5	163.0	98.0	497.0	48.0	<b>1,520</b>
Video Pinball	16,256.9	17,667.9	196,760.4	309,941.9	98,209.5	479,197.0	<b>949,604</b>
Wizard Of Wor	563.5	4,756.5	2,704.0	7,492.0	7,855.0	<b>12,352.0</b>	9,300
Yars' Revenge	3,092.9	54,576.9	18,098.9	11,712.6	49,622.1	<b>69,618.1</b>	35,050
Zaxxon	32.5	9,173.3	5,363.0	10,163.0	12,944.0	<b>13,886.0</b>	10,513

图14. 所有游戏的原始分数，从30个noop动作开始。参考值来自王等（2016年）。

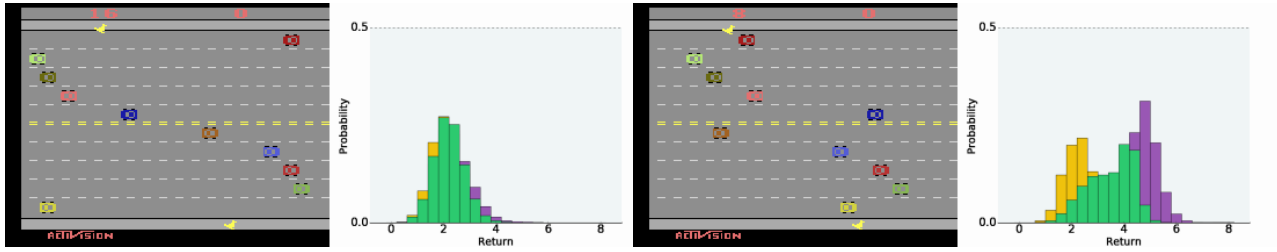


Figure 15. FREEWAY: Agent differentiates action-value distributions under pressure.

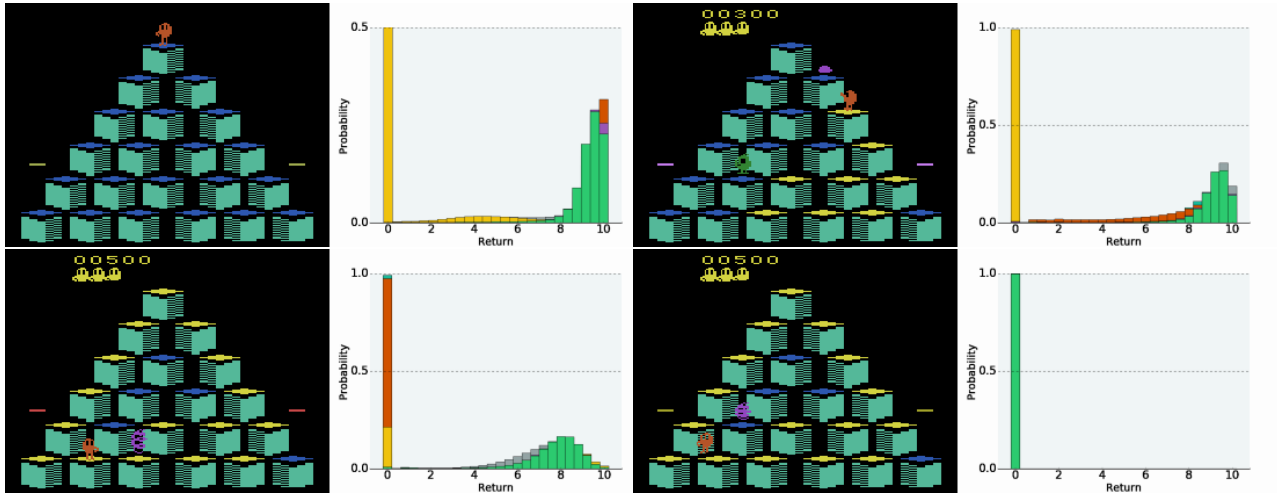


Figure 16. Q\*BERT: Top, left and right: Predicting which actions are unrecoverably fatal. Bottom-Left: Value distribution shows steep consequences for wrong actions. Bottom-Right: The agent has made a huge mistake.

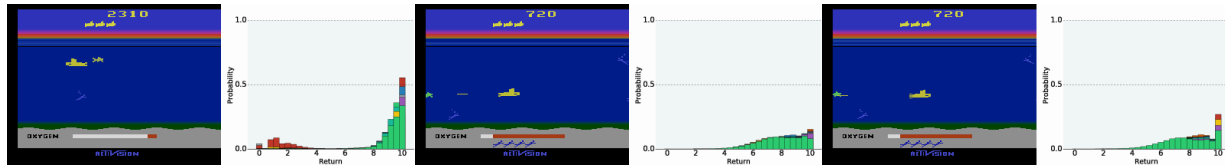


Figure 17. SEAQUEST: Left: Bimodal distribution. Middle: Might hit the fish. Right: Definitely going to hit the fish.

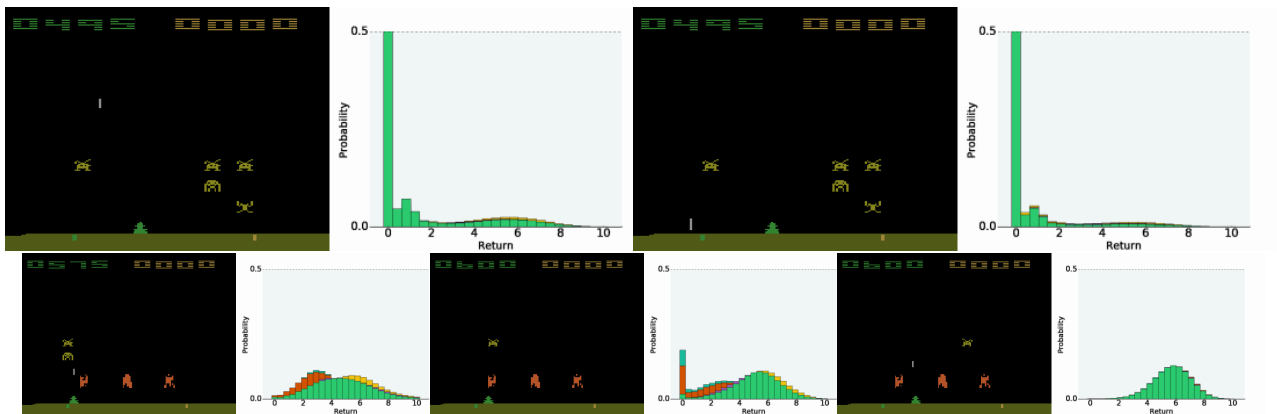


Figure 18. SPACE INVADERS: Top-Left: Multi-modal distribution with high uncertainty. Top-Right: Subsequent frame, a more certain demise. Bottom-Left: Clear difference between actions. Bottom-Middle: Uncertain survival. Bottom-Right: Certain success.

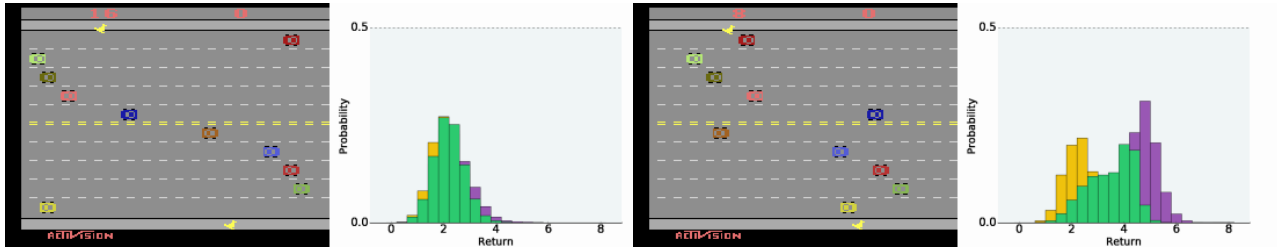


图15. FREEWAY: 代理在压力下区分动作值分布。

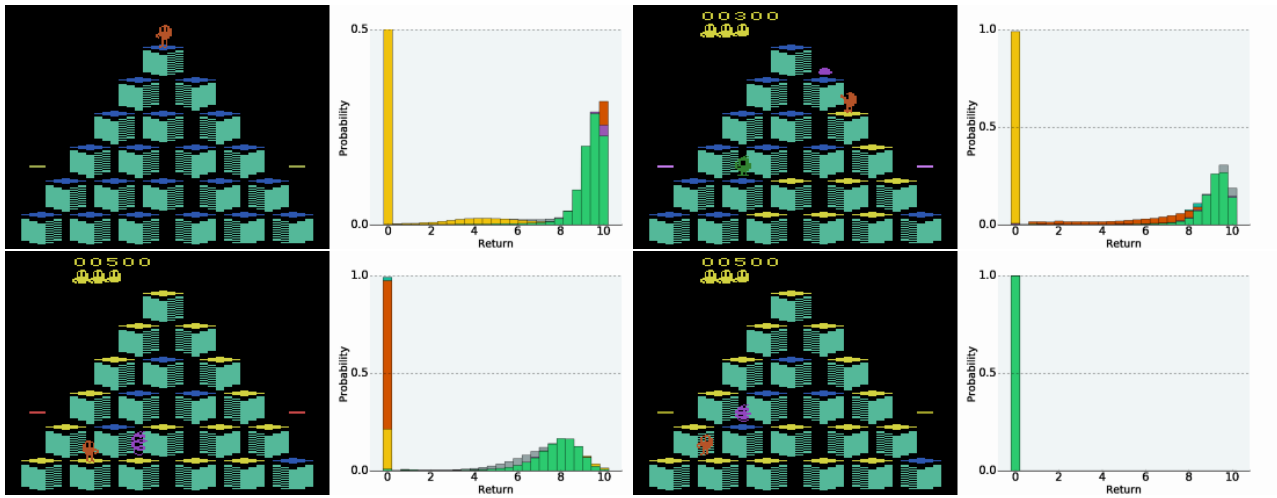


图16. Q\*BERT: 最上面、左起第一个和第二个: 预测哪些动作是不可逆转的致命动作。最下面左: 价值分布显示错误动作的后果非常严重。最下面右: 智能体犯了一个大错误。

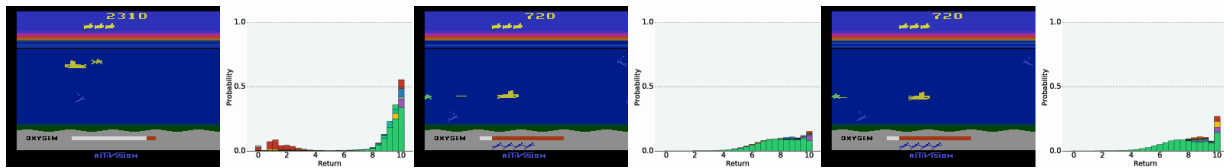


图17. SEAQUEST: 左: 双模分布。中: 可能会撞到鱼。右: 肯定会撞到鱼。

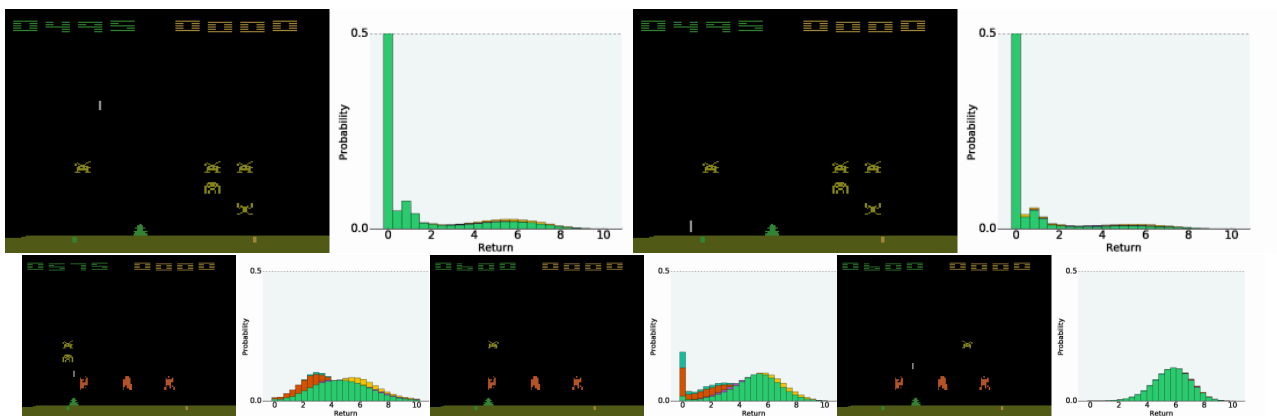


图18. SPACE INVADERS: 左上: 具有高不确定性多模态分布。右上: 后续帧, 更确定的失败。左下: 动作之间的明显差异。中下: 不确定的生存。右下: 确定的成功。