

关于加固学习的分配观点

Marc G. Bellemare^{*} 将Dabney^{*} 和Emi Munos¹

抽象的

在本文中，我们主张value distribution的基本重要性：增强学习代理人收到的随机回报的分布。这与加固学习的方法相反，该方法模拟了此回报的期望或value。尽管有一个既定的文献都在研究价值分布，但到目前为止，它一直用于特定目的，例如实施风险感知行为。我们从政策评估和控制设置的理论结果开始，暴露了后者的显着分布不稳定性。然后，我们使用分布观点来设计一种新算法，该算法将贝尔曼的方程应用于学习近似价值分布的方程。我们使用街机学习环境中的游戏套件来评估我们的算法。我们获得了最先进的资源和轶事证据，证明了价值分布在近似强化学习中的重要性。最后，我们综合理论和经验证据，以高度阐明价值分布在大概环境中学习的方式。

1. 简介

强化学习的主要原则之一指出，如果不以其他方式限制其行为，则代理应旨在最大化其预期的效用 Q 或value (Sutton & Barto, 1998)。贝尔曼的方程式用随机过渡 (x, a) 的预期奖励和表达的结果来缩短描述了这一值 $\rightarrow (X', A')$:

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

在本文中，我们旨在超越价值概念，并主张对加强的分配观点 -

^{*}Equal contribution ¹DeepMind, London, UK. Correspondence to: Marc G. Bellemare <bellemare@google.com>.

学习。特别是，我们研究的主要对象是随机返回 Z ，其期望为 Q 。递归方程也描述了这种随机回报，但分布性质之一：

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

distributional Bellman equation指出 Z 的分布的特点是三个随机变量的相互作用：奖励 R ，下一个状态 - action (X', A') 和其随机返回 $Z(X', A')$ 。通过类似于众所周知的情况，我们将此数量称为value distribution。

尽管分布观点几乎与贝尔曼方程本身一样古老 (Jaquette, 1973; Sobel, 1982; White, 1988)，但在强化学习中，迄今已将其从属于特定目的：用于模拟参数性不确定性 (Dearden等人, 1998年)，设计对风险敏感的al-gorithm (Morimura等, 2010b; a) 或用于理论分析 (Azar等, 2012; Lattimore & Hutter, 2012)。通过Concontast，我们认为价值分布在强化学习中起着核心作用。

政策评估的收缩Bellman操作员。基于R. Osler (1992) 的结果，我们表明，对于固定的策略，对价值分布的Bellman运营商是Waseserstein (也称为Kantorovich或Mallows) 指标的最大形式的收缩。我们的指标事项的特定选择：同一操作员在总变化，kullback-leibler Divergence或Kolmogorov距离上不是收缩。

在控制设置中不稳定。与政策评估案例相比，我们将证明Bellman的Opti-timelality方程式的发行版本不稳定。特别地，尽管最优运算符是预期值的违反 (与通常的最优性相匹配)，但在任何分布上的任何度量中，它都不是收缩。这些结果提供了支持学习非组织政策影响的算术算法的证据。

更好的近似值。从算法的角度来看，学习近似分布而不是其近似期望有许多好处。分布式钟楼操作员保留了价值分布的多模式，我们认为这会导致更稳定的学习。近似完整的分布还减轻了从非组织政策中学习的影响。总体而言

我们认为，这种方法使近似的强化学习表现得更好。

我们将在街机学习环境的背景下说明分布观点的实际收益（Bellemare等，2013）。通过对DQN代理中的价值划分进行建模（Mnih等，2015），我们在基准Atari 2600游戏的范围内大大提高了性能游戏数。我们的结果与Veness等人的结果相呼应。（2015年），他通过预测蒙特卡洛回报获得了非常快的学习。

从监督的学习角度来看，学习全价分布似乎很明显：为什么将我们的自我限制为卑鄙？当然，主要的区别是在我们的环境中没有给定的目标。取而代之的是，我们使用贝尔曼方程来使学习过程可进行。正如Sutton & Barto（1998）所说，我们必须“从猜测中学习猜测”。我们相信，这种猜测最终比成本更具好处。

2. 设置

我们考虑一个以标准方式与环境相互作用的代理：在每个步骤中，代理商根据其当前状态选择一个符合条件，环境以奖励和下一个状态对其进行了。我们将此反应建模为时间均匀的马尔可夫决策过程（ $\mathcal{X}, \mathcal{A}, R, P, \gamma$ ）。像往常一样， \mathcal{X} 和 \mathcal{A} 分别是状态和动作空间， P 是过渡内核 $P(\cdot | x, a)$ ， $\gamma \in [0, 1]$ 是折现因子， R 是奖励功能，在这项工作中，我们明确将其视为一个随机变量。固定策略 π 将每个状态 $x \in \mathcal{X}$ 映射到动作空间 \mathcal{A} 上的概率分布。

2.1. 贝尔曼方程

return Z^π 是沿代理与环境交互轨迹的折扣奖励之和。策略 π 的值函数 Q^π 描述了从状态 $x \in \mathcal{X}$ 采取action $a \in \mathcal{A}$ 的预期重新转向，然后根据 π 进行行动：

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

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

强化学习的基础是使用贝尔曼方程（Bellman, 1957）来描述价值的功能：

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

在加强学习中，我们通常对采取行动感兴趣，以最大程度地提高回报。最常见的ap-

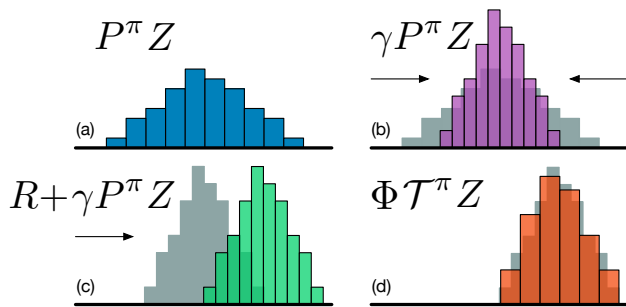


图1。具有确定性奖励函数的分配钟手操作员：（a）策略下的下一个状态分布 π ，（b）打折将分布缩小到0，（c）奖励会移动它，并且（d）投影步骤（第4节）。

这样做的过程涉及最佳方程

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

该方程具有唯一的固定点 Q^* ，即对应于最佳策略 Π^* （ π^* 集合的最佳值函数，如果 $\mathbb{E}_{a \sim \pi^*} Q^*(x, a) = \max_a Q^*(x, a)$ 是最佳的。

我们将值函数视为 $\mathbb{R}^{\mathcal{X} \times \mathcal{A}}$ 中的向量，而被指出的奖励函数则是这样的向量。在这种情况下，Bellman operator \mathcal{T}^π 和optimality operator \mathcal{T} 是

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

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

这些操作员很有用，因为它们描述了流行学习算法（例如SARSA和Q-LEARNING）的预期行为。特别是它们既是收缩映射，并且对某些初始 Q_0 的重复应用分别将指数收敛到 Q^π 或 Q^* （bert-sekas & tsitsiklis, 1996）。

3. 发行的贝尔曼操作员

在本文中，我们消除了贝尔曼方程内的期望，而是考虑随机变量 Z^π 的完整分布。从这里开始，我们将 Z^π 视为从状态行动对到分布的映射，并将其称为value distribution。

我们的第一个目的是了解Bellman操作员的分布类似物的理论行为，尤其是在不太理解的控制环境中。严格对算法贡献感兴趣的读者可能会选择跳过本节。

3.1. 分布方程

有时使用概率空间 $(\omega, \mathcal{F}, \text{pr})$ 会很方便。读者不熟悉mea-

当然，理论可以将 ω 视为实验所有可能结果的空间（Billingsley, 1995）。我们将编写 $\|u\|_p$ 来表示 $1 \leq p \leq \infty$ 的向量 $u \in \mathbb{R}^X$ 的 L_p norm; $\mathbb{R}^{X \times A}$ 中的向量也是如此。然后，随机向量 U 的 L_p 标准： $\omega \rightarrow \mathbb{R}^X$ （或 $\mathbb{R}^{X \times A}$ ）是 $\|U\|_p := [\mathbb{E} [\|U(\omega)\|_p^p]]^{1/p}$ ；对于 $p = \infty$ ，我们有 $\|U\|_\infty = \text{ess sup } \|U(\omega)\|_\infty$ （我们将省略每当明确时），对 $\omega \in \omega$ 的依赖性。我们将表示C.D.F. $F_U(y)$ 的随机变量 U ： $= \text{pr } \{U \leq y\}$ 及其倒数c.d.f.由 $F_U^{-1}(q) := \inf \{y : F_U(y) \geq q\}$ 。

分布方程 $U \stackrel{D}{=} V$ 表示random variable U 是根据与 V 相同的定律分配的。在不失去一般性的情况下，读者可以理解分布方程的两个侧面，将两个独立的随机变量的分离关联。Engel等人已将分布方程用于加强学习。（2005）；Morimura等。（2010a）以及怀特（White）（1988）的运营研究中。

3.2. Wasserstein指标

我们分析的主要工具是累积分布函数之间的Wasserstein Metric d_p （参见例如Bickel & Freedman, 1981，其中称为Mallows Metric）。对于 F, G 两个c.d.f.s在真实方面，它被定义为

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

如果将所有随机变量 (U, V) 与各个累积分布 F 和 G 相应的所有对摄影。Informum由C.D.F.随机变量 U 的trans形式均匀分布在 $[0, 1]$ ：

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

对于 $p < \infty$ 这更明确地写成

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

给定两个随机变量 U, V ，带有c.d.f.s F_U, F_V ，我们将编写 $d_p(U, V) := d_p(F_U, F_V)$ 。我们将发现它可以提出相结合的随机变量，并在INF下使用其版本，写作

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

每当明确时；我们认为，更大的知名度是有理由的。最后，我们使用相应的 L_p norm将此度量扩展到随机变量的向量，例如值分布。

考虑标量 a 和一个随机变量 A 独立

U, V 的of。公制 d_p 具有以下属性：

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

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

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

我们将需要以下额外的属性，这在其变量上没有独立假设。它的证明和后来的结果是在附录中给出的。

引理1（分区引理）。Let A_1, A_2, \dots be a set of random variables describing a partition of ω , i.e. $A_i(\omega) \in \{0, 1\}$ and for any ω 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} 用有限的矩表示价值分布的空间。对于两个值分布 $Z_1, Z_2 \in \mathcal{Z}$ ，我们将使用Wasserstein Metric的最大形式：

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

我们将使用 \bar{d}_p 来建立分销钟手运营商的收敛性。

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，我们对与给定策略 π 关联的值函数 V^π 感兴趣。这里的类似物是值分布 Z^π 。在本节中，我们表征 Z^π 并研究策略评估操作员 \mathcal{T}^π 的行为。我们认为 Z^π 描述了代理与环境相互作用的固有随机性，而不是对环境本身的不确定性的某种量度。

我们将奖励函数视为随机向量 $R \in \mathcal{Z}$ ，并定义过渡操作员 P^π ： $\mathcal{Z} \rightarrow \mathcal{Z}$ 为

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

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

我们使用大写字母来强调下一个状态行动对的随机性 (X', A') 。我们定义分销的贝尔曼操作员 \mathcal{T}^π ： $\mathcal{Z} \rightarrow \mathcal{Z}$ 为

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

\mathcal{T}^π 与通常的贝尔曼操作员（2）具有表面相似之处，但这根本不同。在特定的三种随机性来源中定义化合物分配 $\mathcal{T}^\pi Z$ ：

a) 奖励 R , b) 过渡 P^π 和c) 的随机性, c) 下一态值分布 $Z(X', A')$ 。

特别是, 我们通常假设这三个数量是独立的。在本节中, 我们将证明 (5) 是一个收缩映射, 其唯一固定点是随机返回 Z^π 。

3.3.1. \bar{d}_p 中的收缩

考虑过程 $Z_{k+1} := \mathcal{T}^\pi Z_k$, 从一些 $Z_0 \in \mathcal{Z}$ 开始。我们可能希望 $\{Z_k\}$ 的限制期望能像往常一样快速地汇总为 Q^π 。正如我们现在显示的那样, 该过程从更强的意义上收敛: \mathcal{T}^π 是 \bar{d}_p 的收缩, 这意味着所有矩也很快就会成倍收敛。

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

使用引理3, 我们使用Banach的固定点定理得出结论, \mathcal{T}^π 具有独特的固定点。通过检查, 该固定点必须按 (1) 中定义为 Z^π 。正如我们假设所有矩都是有界的, 这是有足够的结论, 可以得出结论, $\{Z_k\}$ 的序列 \bar{d}_p 在 \bar{d}_p 中收敛到 $1 \leq p \leq \infty$ 。

总而言之, 我们指出, 并非所有的分布指标都是平等的。例如, Chung & Sobel (1987) 表明 \mathcal{T}^π 在总变化距离上不是收缩。可以为Kullback-Leibler Divergence和K olmogorov距离得出模拟结果。

3.3.2. 以中心的时刻收缩

观察 $d_2(U, V)$ (更普遍地, d_p) 与一个耦合 $C(\omega)$: $= U(\omega) - V(\omega)$, 从某种意义上说

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

结果, 我们不能直接使用 d_2 来绑定方差差 $|\mathbb{V}(\mathcal{T}^\pi Z(x, a)) - \mathbb{V}(Z^\pi(x, a))|$ 。但是, \mathcal{T}^π 实际上是方差的收缩 (Sobel, 1982, 另请参见附录)。通常, \mathcal{T}^π 并不是 p^{th} 居中的时刻, $p > 2$ 的收缩, 但是迭代的焦点 $\{Z_k\}$ 仍然会迅速地与 Z^π 的矩阵呈指数融合; 证明扩展了R ˆOsler (1992) 的结果。

3.4. 控制

到目前为止, 我们已经考虑了固定的策略 π , 并研究了其关联的操作员 \mathcal{T}^π 的行为。现在, 我们着手了解control设置的分布运算符 - 我们寻求最大化价值的策略 π 以及最佳价值分布的相应概念。与最佳价值函数一样, 此概念与最佳策略的概念密切相关。但是, 尽管所有最佳策略都达到相同的值 Q^* , 但在我们的情况下

很难出现: 总的来说, 有许多最佳价值分布。

在本节中, 我们表明, 贝尔曼最佳算法的分布类似物在薄弱的意义上会收敛于一组最佳价值分布。但是, 该操作员在分配之间的任何指标中都不是收缩, 并且通常比政策评估运营商更气质。我们认为, 我们在这里概述的趋势问题是贪婪更新固有不安定的症状, 例如Tsitsiklis (2002) 和最近的Harutyunyan等人。 (2016)。

令 Π^* 为最佳策略集。我们首先要描述optimal value distribution的含义。

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

我们强调的是, 并非所有具有预期 Q^* 的价值分布是最佳的: 它们必须在某些最佳策略下与退货的完整分布匹配。

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

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

回想一下, 预期的贝尔曼最佳操作员 \mathcal{T} 是

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

x' 的最大化对应于某些贪婪的策略。尽管此策略在 (6) 中是隐含的, 但我们不能在分配环境中忽略它。我们将称为distributional Bellman optimality operator任何操作员 \mathcal{T} , 该操作员将其赋予贪婪的选择规则, 即

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

与策略评估设置一样, 我们对迭代 Z_{k+1} 的行为感兴趣: $= \mathcal{T}Z_k$, $Z_0 \in \mathcal{Z}$ 。我们的第一个结果是断言 $\mathbb{E} Z_k$ 的行为如预期。

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.

通过检查引理4, 我们可能希望 Z_k 在 \bar{d}_p 中快速到 \mathcal{Z}^* 中的某个固定点。不胜非想地, 收敛既不是快速或确保达到固定点的。实际上, 我们所希望的最好的是侧重融合, 甚至不是集合 \mathcal{Z}^* , 而是与较大的nonstationary optimal value distributions集合。

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（控制设置中的收敛）。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 Π^* , 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，揭示了分布框架与通常的预期收益设置之间的显著差异。 Z_k 的平均值会迅速收敛到 Q^* ，但它的分布不一定是良好的！为了强调这种差异，我们现在提供了许多有关 \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 的唯一过渡；从 x_2 ，Action a_1 不会产生奖励，而最佳影响 a_2 产生 $1 + \epsilon$ 或 $-1 + \epsilon$ ，具有均等的可能性。这两个动作都是终端。有一个独特的最佳策略，因此有一个唯一的固定点 Z^* 。现在考虑图2（右）中给出的 Z ，其距离 Z^* ：

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

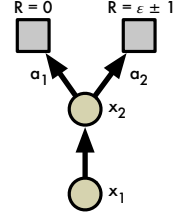
除了 (x_2, a_2) 以外，我们利用 $Z = Z^*$ 到处的事实。但是，当我们将 \mathcal{T} 应用于 Z 时，选择了贪婪的动作 a_1 并 $\mathcal{T}Z(x_1) = Z(x_2, a_1)$ 。但

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

对于有足够的小 ϵ 。这表明未估计的更新不是一个非跨性别别的： $\bar{d}_1(\mathcal{T}Z, \mathcal{T}Z^*) > \bar{d}_1(Z, Z^*)$ 。使用 $\gamma < 1$ ，相同的证明表明它不是收缩。使用更技术上的参数，我们可以将此结果扩展到将 Z 和 $\mathcal{T}Z$ 分开的任何度量。

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

要查看这一点，请考虑使用 $\epsilon = 0$ 的同一示例，以及一个贪婪的操作员 \mathcal{T} ，该 \mathcal{T} 如果 $Z(x_1) = 0$ ，则通过选择 a_2 的打破领带，否则 $Z(x_1) = 0$ ， a_1 否则。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。未交流的两态MDP，最佳操作员 \mathcal{T} 并不是一个缩写。突出显示了有助于 $\bar{d}_1(Z, Z^*)$ 和 $\bar{d}_1(\mathcal{T}Z, Z^*)$ 的条目。

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描绘了控制设置的相当黯淡的图片。在实践中实际上出现了动态化的影响，这还有待观察。一个开放的问题是，从理论上讲，是否可以使用随机策略来得出更稳定的行为，例如从保密政策迭代（Kakade & Langford, 2002）。

4. 近似分布学习

在本节中，我们提出了一种基于分歧式贝尔曼最佳操作员的算法。特别是，这将需要选择一个近似分布。尽管以前已经考虑过高斯案例（Morimura等，2010a; Tamar等，2016），据我们所知，我们首先是使用丰富的守则分布。

4.1. 参数分布

我们将使用由 $N \in \mathbb{N}$ 和 $V_{\min}, V_{\max} \in \mathbb{R}$ 参数参数的离散分布来对价值分布进行建模，其支持是原子 $\{z_i = V_{\min} + i\Delta z: 0 \leq i < N\}$ ， $\Delta z, \Delta z: = \frac{V_{\max} - V_{\min}}{N-1}$ 。从某种意义上说，这些原子是我们分布的“规范回报”。原子概率由参数模型 θ 给出： $\mathcal{X} \times \mathcal{A} \rightarrow \mathbb{R}^N$

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

离散分布具有高度表达和计算友好的优势（例如，参见van den oord等，2016）。

4.2. 投影的贝尔曼更新

使用离散分发提出了一个问题：贝尔曼更新 $\mathcal{T}Z_\theta$ ，我们的参数化 Z_θ 几乎具有不相交的支持。根据第3节的分析，将 $\mathcal{T}Z_\theta$ 和 Z_θ 之间的Wasserstein Metric（被视为损失）最小化似乎很自然，这也是

方便地对支持的差异非常强大。但是，第二个问题阻止了这一点：实际上，我们通常仅限于从样本过渡中学习，这是在瓦斯坦损失下不可能的（请参阅附录5和玩具结果）。

取而代之的是，我们将示例Bellman Update $\hat{T}Z_\theta$ 投影到 Z_θ (图1, 算法1) 的支持上，有效地将Bellman更新重新定义为多类分类。令 π 为贪婪的策略 W.R.T. $\mathbb{E} Z_\theta$ 。给定样品转移 (x, a, r, x') ，我们为每个原子 z_j 计算 Bellman Update $\hat{T}z_j \hat{T}z_j r + \gamma z_j$ ，然后将其概率 $p_j(x', \pi(x'))$ 分配给 $\hat{T}z_j$ 的直接邻居。投影更新的 i^{th} 组件 $\Phi \hat{T}Z_\theta(x, a)$ 是

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

其中 $[\cdot]_a^b$ 在 $[a, b]$ 范围内界定其参数。¹与通常一样，我们查看了由 $\hat{\theta}$ 参数化的下一个状态分布。样本损失 $\mathcal{L}_{x,a}(\theta)$ 是kl发散的跨渗透项

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

这很容易最小化，例如使用梯度下降。我们称这种分配和损失的选择为 *categorical algorithm*。当 $N = 2$ 时，一个简单的单参数替代方案是 $\Phi \hat{T}Z_\theta(x, a) := [\mathbb{E} [\hat{T}Z_\theta(x, a)] - V_{\min}] / \Delta z$ ；我们将其称为 *Bernoulli algorithm*。我们注意到，尽管这些属性似乎与Wasserstein指标无关，但最近的工作（Bellemare等，2017）暗示了更深的联系。

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. 对Atari 2600游戏的评估

为了理解复杂的环境中的方法，我们将分类算法与AR的游戏相关联

Cade学习环境（ALE; Bellemare等，2013）。虽然啤酒是确定性的，但随机性确实发生在许多伪造中：1）从状态混轴承，2）从非组织政策中学习和3）从近似错误中学习。我们使用了五个训练游戏（图3）和52场测试游戏。

在我们的研究中，我们使用DQN体系结构（Mnih等，2015），但输出原子概率 $p_i(x, a)$ 而不是动作值，然后从训练游戏中的初步实验中选择 $V_{\max} = -V_{\min} = 10$ 。我们称之为生成的体系结构 *Categorical DQN*。我们用 $\mathcal{L}_{x,a}(\theta)$ 替换平方损失 $(r + \gamma Q(x', \pi(x')) - Q(x, a))^2$ ，然后训练网络以最小化此损失。²如DQN中，我们使用的是一个简单的 ϵ -固定策略，对预期的动作值；我们将作为未来的工作离开，代理商可以根据完整的分布选择行动的许多方式。我们其余的培训制度与Mnih等人的匹配，包括将目标网络用于 $\hat{\theta}$ 。

图4说明了我们在实验中所遵循的典型价值分布。在此示例中，三个动作（包括按钮按下的动作）导致代理会过早释放其激光，并最终失去游戏。相应的分布反映了这一点：他们为0（终端值）分配了一个可能的概率。安全的动作具有相似的分布（左侧跟踪入侵者运动的左侧是有些青睐的）。这个示例有助于解释为什么我们的方法如此成功：分歧更新使低价值，“失去”事件与高价值，“生存”事件分开。³

一个令人惊讶的事实是，尽管淡啤酒的确定性ISM，但通常与高斯人相近，但并未对一个或两个值进行汇总。我们认为这是由于我们离散的 γ 引起的扩散过程。

5.1. 改变原子数

我们首先研究了与原子数量有关的算法在训练游戏中的表现（图3）。对于此实验，我们设置 $\epsilon = 0.05$ 。从数据来看，很明显，使用过多的原子会导致行为不良，并且更多地会提高性能。这并不明显，因为我们可能期望将网络容量饱和。51个原子版本和DQN之间的性能差异特别引人注目：后者在所有五场比赛中都表现出色，而在Seaquest中，我们获得了最先进的性能。作为比较的另一个点，单参数Bernoulli al-Gorithm在5场中的3场比赛中的表现优于DQN，并且最著名的是Asterix更健壮。

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

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

¹Algorithm 1 computes this projection in time linear in N .

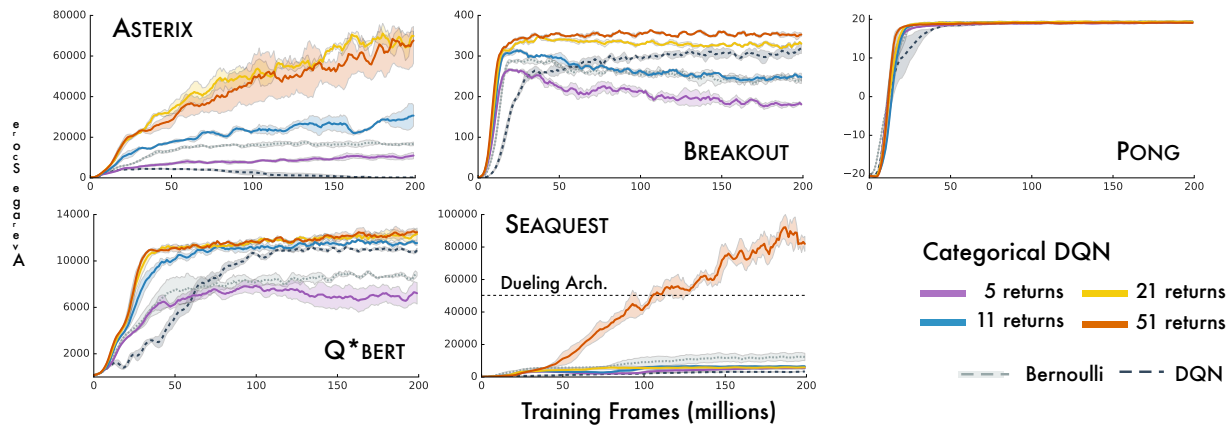


图3. 分类DQN：离散分布中的原子数量不同。分数正在移动超过500万帧的平均值。

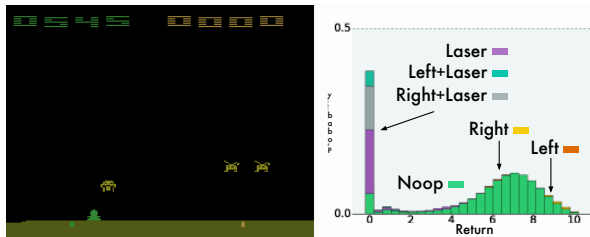


图4. 在太空入侵者的一集中学习的价值分布。不同的动作被阴影不同。重新转弯低于0（在空间入侵者中不发生），因为代理几乎没有向其分配概率。

该实验的一个有趣的结果是找出我们的方法确实在随机性方面提取。乒乓球表现出内在的随机性：重新定位的确切时机取决于内部寄存器，并且确实无法观察到。我们看到这在代理的预测中明显反映了这一点（图5）：连续五个框架，价值分布显示了两种模式，表明代理人相信它尚未获得奖励。有趣的是，由于代理商的状态不包括过去的奖励，因此在收到奖励后甚至无法占用预测，从而解释了模式的相对比例。

5.2. 最先进的结果

上一节中介绍的51个原子代理（从此处开始，C51）在培训游戏中的性能尤其引人注目，鉴于它没有涉及到最先进的代理商中其他算法的想法。接下来，我们询问合并最常见的超级参数选择，即较小的培训 ϵ 是否会带来更好的结果。具体而言，我们设置 $\epsilon = 0.01$ （而不是 0.05 ）；此外，每100万帧，我们

用 $\epsilon = 0.001$ 评估我们的代理商的性能。

我们将算法与DQN（ $\epsilon = 0.01$ ），Double DQN（Van Hasselt等，2016），决斗结构（Wang等，2016）和优先级重播（Schaul等，2016年），比较培训期间达到的最佳评估评分。我们看到C51显著胜过这些其他算法（图6和7）。实际上，在许多游戏中，C51超过了当前的最新水平，最著名的是Seaquest。一个特别惊人的事实是该算法在稀疏奖励游戏中的良好表现，例如冒险和私人眼睛。这表明价值分布能够更好地提出很少发生的事件。附录中提供了完整的结果。

我们还包括一个比较的附录（图12），平均有3种种子，显示了C51的训练性能优于训练有素的DQN和人类玩家的游戏数量。这些结果继续显示出巨大的改进，并且更代表了代理商的平均表现。在5000万帧之内，C51在57场比赛中的45场比赛中表现出了全面训练的DQN代理。这表明，整个200万个训练框架及其随之而来的计算成本对于评估啤酒中的强化学习算法是不必要的。

ALE的最新版本包含一种随机执行机制，旨在防止轨迹上的拟合轨迹。确切地，在每个帧中，环境都以概率 $p = 0.25$ 的概率重新启用了代理所选的动作。尽管DQN大多对随机执行非常强大，但在一些游戏中，其性能会降低。在相对于随机和DQN代理的得分量表上，C51分别获得了平均值和中位数得分，分别为126%和21.5%，确认了超出确定性环境的C51的好处。

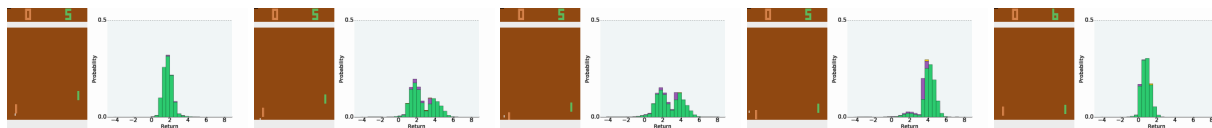


图5。乒乓球中的内在随机性。

	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	701%	178%	40	50
UNREAL [†]	880%	250%	-	-

图6。在57场Atari游戏中的平均值和中位数得分为人基线的百分比（H.B., Nair等, 2015）。

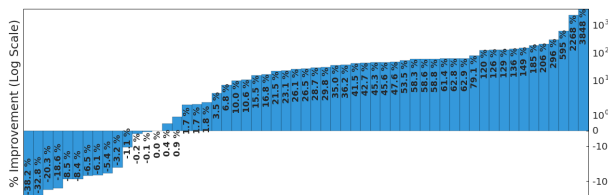


图7。使用van Hasselt等人的方法计算的C51比DQN的C51的百分比提高百分比。

6. 讨论

在这项工作中，我们寻求更完整的强化学习图片，其中涉及价值分布。我们发现，学习价值分布是一个强大的疾病，它使我们能够超过以前在Atari 2600上获得的大多数收益，而无需进一步的算法调整。

6.1. 为什么学习分布很重要？

令人惊讶的是，当我们使用旨在最大程度地提高预期回报的策略时，我们应该看到绩效的任何差异。我们希望做出的区别是

learning distributions matters in the presence of approximation. 我们现在概述了一些可能的原因。

减少chat不休。我们的结果是第3.4节高点，在Bellman最佳操作员中显着不稳定性。当与函数大约结合使用时，这种不稳定可能会阻止策略收敛，Gordon（1995）称为*chattering*。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辨率来减轻这些影响

[†] 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，类似于保守的政策迭代（Kakade & Langford, 2002）。尽管聊天持续存在，但它与近似解决方案有关。

国家的混叠。即使在确定性的环境中，状态混叠也可能导致有效的随机性。例如，麦卡勒姆（McCallum, 1995）表明，在部分可用的领域中，耦合表示与政策学习的重要性。我们看到了一个在乒乓球的国家别名的例子，在那里代理人无法准确预测重新定时。同样，通过明确建模产生的分布，我们提供了一个更稳定的学习目标。

一组更丰富的预测。人工智能中的一个反复出现的主题是代理从预测的程度学习的想法（Caruana 1997; Utgoff & Straczuzi 2002; Sutton et al. 2011; Jaderberg et al. 2017）。分歧的方法自然为我们提供了一系列辅助预测，即回报会带有特定价值的概率。但是，与以前提出的方法不同，这些预写的准确性与代理商的性能紧密结合。

感应偏见的框架。强化学习的分布表现出了一个更自然的框架，在该框架中，我们可以对域或学习问题本身施加假设。在这项工作中，我们使用了 $[V_{\min}, V_{\max}]$ 中的支持分布。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} ）视为等效来改变优化问题。令人惊讶的是，在大多数游戏中，DQN中类似的价值剪辑显着降低。举一个例子：正如一些作者所说的那样，将折现因子 γ 作为适当的概率导致了不同的算法。

行为良好的优化。众所周知，分类分布之间的KL差异是最小化的损失。这可以解释我们的一些经验表现。然而，早期的实验损失的早期实验，例如持续密度之间的KL差异并不富有成果，部分原因是Kl di-di-Gerge nce对其结果的值不敏感。Wasserstein度量的近距离最小化应该比我们在这里提出的更好的结果更好。

在结束时，我们相信我们的结果凸显了需要在设计，理论或其他方面的算法中分布的必要性。

致谢

作者承认，在整个工作的整个发展过程中，他们的同事在DeepMind上发挥了重要作用。特别感谢Yee Whye Teh, Alex Graves, Joel Veness, Guillaume Desjardins, Tom Schaul, David Silver, Andre Barreto, Max Jadeberg, Mohammad Azar, Georg Ostrovski, Bernardo 尤其是克里斯·麦迪逊 (Chris Maddison) 对早期草案的全面看法。还要感谢Marek Petrik对相关文献的指导，并在最终版本中提供了详细的细节。

勘误

本文的相机就绪副本不正确地报告了C51的平均得分为1010%。校正后的数字为701%，比其他可比基线高。中位数得分保持不变，为178%。

错误是由于一场比赛 (ATLANTIS) 持续30分钟的评估发作。相比之下，此处呈现的其他结果在30分钟内 (标准) 呈现。先前报道的亚特兰蒂斯分数为370万；我们的30分钟分数是841,075，我们的得分接近了这个时间范围的最大值。30分钟时的盖帽使我们对亚特兰蒂斯的人体范围分数从22824%提高到仅 (!) 5199%，不幸的是，足以明显影响平均得分，其对离群值的敏感性已得到充分纪录。

参考

- Azar, Mohammad Gheslghi, Munos, R EMI和Hilbert的K appen。关于使用生成模型的增强学习的样本复杂性。在 *Proceedings of the International Conference on Machine Learning* 中，2012年。
- Bellemare, Marc G, Naddaf, Yavar, Veness, Joel和Bowling, Michael。街机学习环境：普通代理的评估平台。 *Journal of Artificial Intelligence Research*, 47: 253–279, 2013。
- Bellemare, Marc G., Danihelka, Ivo, Dabney, Will, Mohamed, Shakir, Lakshminarayanan, Balaji, Hoyer, Stephan, Stephan和Munos, R EMI。cramer距离作为偏见的瓦斯坦梯度的解决方案。 *arXiv*, 2017年。
- 贝尔曼, 理查德.E. *Dynamic programming*。普林斯顿大学出版社, 新泽西州普林斯顿, 1957年。
- Bertsekas, Dimitri P.和Tsitsiklis, John N. *Neuro-Dynamic Programming*。雅典娜科学, 1996年。
- Bickel, Peter J.和Freedman, David A. *The Annals of Statistics*, 第1196–1217页, 1981年。
- Billingsley, 帕特里克。 *Probability and measure*。约翰·威利 (John Wiley & Sons), 1995年。

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

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

Dearden, Richard, Friedman, Nir和Stuart的Russell。贝叶斯Q学习。在 *Proceedings of the National Conference on Artificial Intelligence*, 1998年。

恩格尔, 雅科夫, 曼诺, 史和梅尔, 罗恩。使用高斯过程的增强学习。在 *Proceedings of the International Conference on Machine Learning* 中, 2005年。Geist, Matthieu和Pietquin, Olivier。卡尔曼时间差异。 *Journal of Artificial Intelligence Research*, 39: 483–532, 2010。

戈登, 杰弗里。动态程序中的稳定函数近似。在 *Proceedings of the Twelfth International Conference on Machine Learning*, 1995年。Harutyunyan, Anna, Bellemare, Marc G., Stepleton, Tom和Munos, R EMI。Q (λ) 带有额外校正。在 *Proceedings of the Conference on Algorithmic Learning Theory* 中, 2016年。

在 *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2009年。

Jaderberg, Max, Mnih, Volodymyr, Czarnecki, Wojciech Marian, Schaul, Tom, Leibo, Joel Z, Silver, David和Kavukcuoglu, Koray。通过无监督的辅助任务进行强化学习。 *Proceedings of the International Conference on Learning Representations*, 2017年。

Jaquette, Stratton C. Markov的决策过程具有新的选择标准：离散时间。 *The Annals of Statistics*, 1 (3) : 496–505, 1973。

Kakade, Sham和Langford, John。大约最佳的强化学习。在 *Proceedings of the International Conference on Machine Learning*, 2002年。

Kingma, Diederik和Ba, Jimmy。Adam：一种用于优化的方法。 *Proceedings of the International Conference on Learning Representations*, 2015年。

Lattimore, Tor和Hutter, Marcus。PAC的折扣MDP界限。在 *Proceedings of the Conference on Algorithmic Learning Theory* 中, 2012年。

Mannor, Shie和Tsitsiklis, John N. Markov决策过程中的均值优化。2011。

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

Mnih, Volodymyr, Kavukcuoglu, Koray, Silver, David, Rusu, Andrei A, Veness, Joel, Joel, Bellemare, Marc G, Marc G, Graves, Alex, Ried-Miller, Martin, Martin, Fidjeland, Andreas K, Andreas K, Ostrovski, Georg, Georg等。通过深入的强化学习来控制人类水平的控制。 *Nature*, 518 (7540) : 529–533, 2015。

Morimura, Tetsuro, Hachiya, Hirotaka, Sugiyama, Masashi, Tanaka, Toshiyuki和Kashima, Hisashi. 加固学习的参数回流密度估计。在*Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2010a中。

Morimura, Tetsuro, Sugiyama, Masashi, Kashima, Hisashi, Hachiya, Hirotaka和Toshiyuki. 用于加固学习的非参数重新分布近似。在*Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 第799–806页, 2010b中。

Nair, Arun, Srinivasan, Praveen, Blackwell, Sam, Alcicek, Cagdas, Fearon, Rory, De Maria, De Maria, Alessandro, Panneershelvam, Vedavyas, Suleyman, Mustafa, Beattie, Charles, Charles和Petersen, Petersen, Stig等。深入增强学习的大量平行方法。在*ICML Workshop on Deep Learning*, 2015年。风险敏感的MDP的参与者批评算法。在*Advances in Neural Information Processing Systems*, 2013年。

Puterman, Martin L. *Markov Decision Processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994. R'Osler, UWE. 分布的固定点定理。 *Stochastic Processes and their Applications*, 42 (2) : 195–214, 1992。Schaul, Tom, Quan, John, John, Antonoglou, Ioannis和Silver, David. 优先经验重播。在*Proceedings of the International Conference on Learning Representations*, 2016年。Sobel, Matthew J.折扣马尔可夫决策过程的差异。 *Journal of Applied Probability*, 19 (04) : 794–802, 1982。Sutton, Richard S.和Barto, Andrew G. *Reinforcement learning: An introduction*. 麻省理工学院出版社, 1998年。来自无监督的传感器相互作用的知识。在*Proceedings of the International Conference on Autonomous Agents and Multiagents Systems*, 2011年。

Tamar, Aviv, di Castro, Dotan和Mannor, Shie. 学习奖励前进的差异。 *Journal of Machine Learning Research*, 17 (13) : 1–36, 2016。讲座6.5-RMSPROP: 将梯度除以其近期宏伟的平均值。 *COURSERA: Neural networks for machine learning*, 4 (2), 2012年。图森, 马克和斯托基, 阿莫斯。解决离散和连续状态马尔可夫决策过程的概率推论。在*Proceedings of the International Conference on Machine Learning*, 2006年。

Tutgoff, Paul E.和Stracuzzi, David J.多层学习。 *Journal of Machine Learning Research*, 3: 59–72, 2002。Van den Oord, Aaron, Kalchbrenner, Nal和Ko-Ray的Kavukcuoglu. 像素复发性神经网络。在*Proceedings of the International Conference on Machine Learning*中, 2016年。

Van Hasselt, Hado, Guez, Arthur和Silver, David. 通过双重Q学习的深度重新学习。在*Proceedings of the AAAI Conference on Artificial Intelligence*, 2016年。压缩和控制。在*Proceedings of the AAAI Conference on Artificial Intelligence*, 2015年。Wang, Tao, Lizotte, Daniel, Bowling, Michael和Schuurmans, Dale. 动态编程的双重表示。 *Journal of Machine Learning Research*, 第1–29页, 2008年。决斗网络档案, 用于深入增强学习。在*Proceedings of the International Conference on Machine Learning*, 2016年。White, D. J. Mean, 差异, 方差和概率标准中有限的马尔可夫决策过程: 审查。 *Journal of Optimization Theory and Applications*, 56 (1) : 1–29, 1988。

A.相关工作

据我们所知，最接近我们的工作的是两篇论文（Morimura等，2010b；a）从其作用分布函数的角度研究了分散的Bellman方程。作者提出了参数和非参数解决方案，以学习对风险敏感的增强学习的分布。他们还政策评估设置提供了一些理论分析，包括在非参数案例中的一致性结果。相比之下，我们还分析了控制设置，并强调使用分布方程来证明近似强化学习。

回报的差异已在风险敏感的环境中广泛介绍。值得注意的是，Tamar等人。（2016年）分析了线性函数近似来学习这种差异进行策略评估，Prashanth & Ghavamzadeh（2013）估计了对风险敏感的参与者 - 批判算法的设计回报差异。Mannor & Tsitsiklis（2011）就计算差异解决方案的计算对光控制问题的解决方案提供了负面结果。

在建模不确定性时，也会出现分布式配方。Dearden等。（1998年）考虑了对值分布的高斯近似，并建模了与正常 γ 先验的近似参数的不确定性。Engel等。（2005年）利用分布的钟形方程将高斯过程定义为未知的值函数。最近，Geist & Pietquin（2010）提出了基于无味的卡尔曼过滤器的同一问题的替代解决方案。我们认为，我们在这里提供的大部分分析，这些分析涉及环境的固有随机性，也可以应用于建模不确定性。

我们在这里的工作是基于许多基础知识，尤其是关于替代性最佳症状的基础。一开始，Jaquette（1973）表明moment optimality标准是可以实现的，该标准是可以实现的，它是可以实现，并定义了一个固定的最佳策略，回应了定理的第二部分。）通常将其作为对较高时刻（但不是分布）转弯的钟声方程的第一个引用。Chung & Sobel（1987）提供了有关在总变化距离中分配钟形操作员收敛的结果。White（1988）从优化国家行动对占用的角度研究了“非标准MDP标准”。

近年来，已经提出了许多用于加强学习的概率框架。plan-ning as inference方法（Toussaint & Storkey，2006；Hoffman等，2009）将返回嵌入到图形模型中，并应用概率推理来确定

一系列行动，导致最大的预期奖励。Wang等。（2008年）考虑了对潜在学习的双重表述，其中一种优化的固定分布受到过渡函数给出的约束（Puterman，1994），特别是其与线性近似的关系。与此双重相关的是压缩和控制算法Veness等。（2015年），它通过使用DENSITY模型来学习返回分布来描述价值函数。这项工作的目的之一是解决他们是否可以根据钟声方程设计的实用分配算法而不是蒙特卡洛估计的实用分配算法所留下的问题。

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 ω , i.e. $A_i(\omega) \in \{0, 1\}$ and for any ω 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. 我们将提供 $p < \infty$ 的证明，并指出相同的 $p = \infty$ 。令 $Y_i = A_i U$ 和 $Z_i = A_i V$ 。第一注意

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

现在， $|A_i U - A_i V|^p = 0$ 时，每当 $A_i = 0$ 时。随之而来的是，我们可以选择 Y_i, Z_i ，以便每当 $|Y_i - Z_i|^p = 0$ 时 $|Y_i - Z_i|^p = 0$ $A_i = 0$ ，而无需增加预期的规范。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，我们声称

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

特定地，方程的左侧是所有r.v.的含量，其累积分布分别为 F_U 和 F_V ，而右侧则是 r 序列的内部。 V 's Y_1, Y_2, \dots 和 Z_1, Z_2, \dots 的累积分布是 $F_{A_i U}, F_{A_i V}$ ，分别为 $F_{A_i U}, F_{A_i V_0}$ 。要证明这种上限，请考虑C.D.F. U :

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

因此，在几乎可以肯定的是，分布 F_U 是等于第一个选择分区的元素 A_i 的分布，然后选择 U 有条件的 A_i 条件的值。另一方面，C.D.F. $Y_i \stackrel{D}{=} A_i U$ 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，(9) 中的右侧右侧具有附加的约束，即必须保留条件的c.d.fs，特别是当 $y \geq 0$ 时。换句话说，而不是自由地完全重新安排映射 $U: \omega \rightarrow \mathbb{R}$ ，我们只能在分区的每个元素中重新排序。我们现在写

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

其中 (a) 之所以跟随，是因为 A_1, A_2, \dots 是一个分区。使用 (9)，这意味着

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

因为在 (b) 中，总和的各个组成部分被独立最小化；(c) 来自 (8)。□

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

Proof. 唯一的非平凡特性是三角形不等。对于任何值分发 $Y \in \mathcal{Z}$ ，写

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

在 (a) 中，我们将三角形不等式用于 d_p 。□

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

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

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

根据 d_p 的属性，我们有

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

最后一行是根据 P^π (的定义，请参见 (4))。结合 (10) 我们获得

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

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. 第一个语句是标准的，其证明来自 $\mathbb{E} \mathcal{T}^\pi Z = \mathcal{T}^\pi \mathbb{E} Z$ ，其中第二个 \mathcal{T}^π 表示通常的operator在值函数上。现在，通过 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. 证明是根据期望的线性性。为分布运算符编写 \mathcal{T}_D ，为通常的操作器编写 \mathcal{T}_E 。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 (控制设置中的收敛)。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 Π^* , 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证明的要旨包括表明，对于每个状态 x ，贪婪的策略 W.R.T. 都有一个时间 k 。 Q_k 主要是最佳的。为了清楚地揭示所涉及的步骤，我们将首先假设一个独特的 (又是确定性的) 最佳策略 π^* ，然后返回一般情况；我们将用 $\pi^*(x)$ 在 x 上表示最佳动作。为了方便起见，我们将编写 $Q_k := \mathbb{E} Z_k$ 和 $\mathcal{G}_k := \mathcal{G}_{Z_k}$ 。令 $B := 2 \sup_{Z \in \mathcal{Z}} \|Z\|_\infty < \infty$ ，让 $\epsilon_k := \gamma^k B$ 。我们首先定义一组状态 $\mathcal{X}_k \subseteq \mathcal{X}$ ，其值必须足够接近 Q^* 时 k :

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

确实，通过引理4，我们知道 k 迭代之后

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

对于 $x \in \mathcal{X}$ ，写 $a^* := \pi^*(x)$ 。对于任何 $a \in \mathcal{A}$ ，我们推论

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

因此，如果 $x \in \mathcal{X}_k$ ，则所有 $Q_k(x, a^*) > Q_k(x, a')$ 也为所有 $a' \neq \pi^*(x)$ ：对于这些状态，贪婪的策略 $\pi_k(x) := \arg \max_a Q_k(x, a)$ 对应于最佳策略 π^* 。

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，所以差距

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

通过定义，有一些严格的阳性 $\Delta(x) > 0$ 。存在一个 k ，这样

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

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

这种引理使我们能够保证存在迭代 k 的存在，然后有足够的表现良好，从某种意义上说，这些州的贪婪政策选择了最佳行动。我们将这些国家称为“解决”。实际上，我们不仅要求解决这些国家，还要求其大多数继任者以及这些州的大多数继任者，等等。我们将此概念形式化如下： $x \in \mathcal{X}$ 且 $\delta > 0$ ，让 $\mathcal{X}_{k,0} := \mathcal{X}_k$ ，并为 $i > 0$ 定义。

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

如以下引理所示，对于任何 i ，最终都包含在递归确定的集合 $\mathcal{X}_{k,i}$ 的集合中。

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. 修复 i ，让我们假设 $\mathcal{X}_{k,i} \uparrow \mathcal{X}$ 。通过引理5，对于 $i = 0$ 。我们在 \mathcal{X} 上的任何概率度量 P ， $P(\mathcal{X}_{k,i}) \rightarrow P(\mathcal{X}) = 1$ 。尤其是给定的 $x \in \mathcal{X}_k$ 上推断出来，这意味着这意味着它

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

因此，对于任何 x ，存在一段时间，然后仍然是 $\mathcal{X}_{k,i+1}$ 的成员， $P(\mathcal{X}_{k-1,i} | x, \pi^*(x)) \geq 1 - \delta$ 的一组状态。我们得出结论， $\mathcal{X}_{k,i+1} \uparrow \mathcal{X}$ 。该声明以归纳为由。 \square

Proof of Theorem 1. 该证明类似于政策迭代类型的结果，但需要更多的护理来处理指标和可能的国家空间。我们将同样地编写 $W_k(x) := Z_k(x, \pi_k(x))$ ，定义 W^* ，并在一定的符号中写入 $\mathcal{T} W_k(x)$ ：

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

修复 $i > 0$ 和 $x \in \mathcal{X}_{k+1,i+1} \subseteq \mathcal{X}_k$ 。我们首先使用引理1 将 x 的过渡分为求解的术语和未解决的术语：

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

其中 X' 是采取动作 $\pi_k(x)$ 的随机后继： $= \pi^*(x)$ ，然后我们编写 $S_i^k = S_i^k(X')$ ， $\bar{S}_i^k = \bar{S}_i^k(X')$ 来简化符号。相似地，

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

现在

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

在 (a) 中, 我们使用了 Wasser-Stein Metric 的属性 P1 和 P2, 在 (B) 中, 我们使用 Lemma 1 ($\{S_i^k, \bar{S}_i^k\}$ 形成 parti-partimi- ω)。令 $\delta_i := \Pr\{X' \notin \mathcal{X}_{k,i}\} = \mathbb{E}\{\bar{S}_i^k(X')\} = \|\bar{S}_i^k(X')\|_p$ 。从 Wasserstein 公制的财产 P3 中, 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 $B < \infty$ 是最大的可达到的 $\|Z\|_\infty$ 。由于 $\delta_i < \delta$ 也可以选择 $x \in \mathcal{X}_{k+1,i+1}$, 因此我们可以在 (12) 中通过 $\gamma\delta B$ 在 (12) 中绑定第二项。这会产生

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

通过 $i > 0$ 的归纳, 我们得出结论, 对于 $x \in \mathcal{X}_{k+i,i}$ 和一些随机状态 X'' 向前迈进,

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

因此, 对于任何 $x \in \mathcal{X}$, $\epsilon > 0$, 我们可以采用 δ, i , 并且最终 k 足以使 $d_p(W_k(x), W^*(x)) < \epsilon$ 。然后, 通过考虑 \mathcal{T} 的另一种副本来扩展到 $Z_k(x, a)$ 。

现在, 我们考虑有多种最佳政策的更普遍的情况。我们扩展了 $\mathcal{X}_{k,i}$ 的定义, 如下所示:

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

因为有很多行动, 因此引理 6 也适用于这一新的定义。和以前一样, 请 $x \in \mathcal{X}_{k,i}$, 但现在考虑通过 \mathcal{T} 连续应用程序选择的贪婪政策的顺序, 然后写入

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

这样

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

现在用 Z^{**} 非组织最佳策略的集合表示。如果我们采用任何 $Z^* \in \mathcal{Z}^*$, 我们将其推断

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

因为 Z^* 对应于某些最佳策略 π^* , 并且 $\bar{\pi}_k$ 沿 (x, a) 的大多数轨迹是最佳的。实际上, $\mathcal{T}^{\bar{\pi}_k} Z^*$ 接近非局部 - 最佳策略 $\pi_k \pi^*$ 的价值分布。现在为此 Z^* ,

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

使用与以前相同的参数与新固定的 $\mathcal{X}_{k,i}$ 使用相同的参数。遵循

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

当 \mathcal{X} 是有限的时, 存在一个固定的 k , 然后 $\mathcal{X}_k = \mathcal{X}$ 。然后随后均匀的收敛结果。

为了证明当 \mathcal{T} 根据订购 \prec 进行动作时, 固定点 Z^* 的唯一性, 我们注意到, 对于任何最佳价值分布 Z^* , 其贪婪的派别集为 Π^* 。用 π^* 表示策略在 Π^* 上首次提出。然后 $\mathcal{T} = \mathcal{T}^{\pi^*}$, 它具有唯一的固定点 (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^* .*

我们在这里提供结果的草图。考虑一个单个状态 x_1 , 其中两个动作 a_1 和 a_2 (图 8)。第一个动作产生的奖励为 $1/2$, 而另一个动作则以同等的概率产生 0 或 1 的奖励, 并且两种动作都是最佳的。现在, 以 $\gamma = 1/2$ 为 2, 并为接收到的奖励编写 R_0, R_1, \dots 。考虑一种随机策略, 该策略将动作 a_2 带有概率 p 。对于 $p = 0$, 返回是

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

另一方面, 对于 $p = 1$, 返回是随机的, 由以下分数编号 (以二进制为单位) 给出:

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

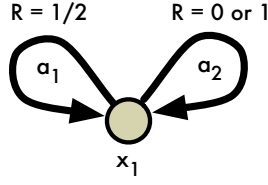


图8。一个简单的示例说明了非机构策略对价值分布的影响。

结果， $Z_{p=1}$ 在0到2之间均匀分布！实际上，请注意

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

对于 p 的某些中间值，我们获得了不同数字的不同概率，但始终在 $[0, 2]$ 中的所有返回上都放置一些概率质量。

现在假设我们遵循第一个步骤 a_1 的非组织策略，然后从那里开始 a_2 。通过检查，返回将在间隔 $[1/2, 3/2]$ 的间隔上均匀分布，这与 p 的任何值下的返回不符。但是现在我们可以想象一个操作员 \mathcal{T} ，该操作员在 a_1 和 a_2 之间交替，具体取决于它应用于的外部值分布，这又会收敛到非平稳的最佳价值分布。

引理7（样本瓦斯坦距离）。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。Let $A_i := \mathbb{I}[I = i]$ 证明了这一点。我们写

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

在倒数第二行中，我们使用 I 和 Q 的 I 的独立性吸引了Wasserstein Metric的属性P3。

为了表明界限通常是严格的，请考虑图9中描述的混合物分布。我们将简单地

考虑此分布 P 和 A_n -其他分布 Q 之间的 d_1 度量。第一个分布是

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

在此示例中， $i \in \{1, 2\}$ ， $P_1 = 0$ 和 $P_2 = 1$ 。现在考虑具有相同支持的分布，但在 $0: 0: \text{在} 0: v39\}$ 的情况下。

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

P 和 Q 之间的距离是

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

对于 $p \in \{0, 1\}$ ，这是 $d_1(P, Q) = \frac{1}{2}$ ，严格小于 $\frac{1}{2}$ 的任何其他值 p 。另一方面，相关的预期距离（在采样结果 x_1 或 x_2 以均等概率为

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

因此， $d_1(P, Q) < \mathbb{E}_I d_1(P_i, Q)$ 对于 $p \in (0, 1)$ 。这表明界限通常是严格的。通过检查，很明显，这两个梯度不同。□

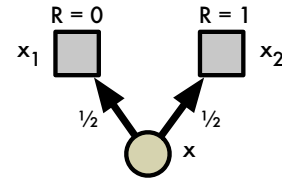


图9。示例MDP，其中预期的样品遗漏距离大于Wasserstein距离。

命题5。Fix some next-state distribution Z and policy π . Consider a parametric value distribution Z_θ , 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.

结果直接来自先前的引理。

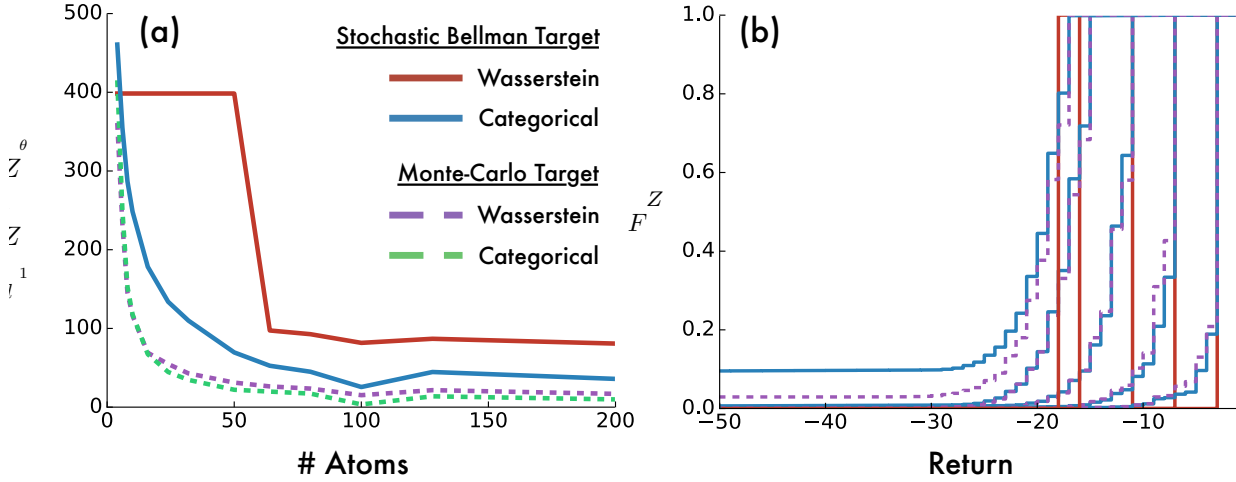


图10。(a) 地面真相分布 Z^π 和近似分布 Z_θ 之间的距离。近似，训练目标和损失功能的原子数量有所不同。(b) 在悬崖上五个代表状态的近似累积分布。

C. 算法细节

尽管我们的培训制度紧密遵循DQN (Mnih等人, 2015年), 但我们使用Adam (Kingma & Ba, 2015年) 而不是RMSPROP (Tieleman & Hinton, 2012) 进行续订。我们还为最终结果进行了一些HyperParam-ETER调整。具体而言, 我们在五个训练游戏中评估了两个超参数, 并选择表现最佳的值。我们认为的HyperParameter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MiniBatch大小。我们发现 $V_{\text{MAX}} = 10$ 和 $\epsilon_{\text{adam}} = 0.01/L$ 执行最佳。我们使用了与DQN ($\alpha = 0.00025$) 相同的步长值。

分类算法的伪代码在算法1中给出。我们将Bellman更新应用于每个原子sepa, 然后将其投影到原始支持中的两个最近的原子中。 $\gamma_t = 0$ 处理到终端状态的过渡。

D. 采样的瓦斯林损失和分类投影的比较

引理3证明, 对于固定策略 π , 分布式贝尔曼操作是 \bar{d}_p 中的 γ -contraction, 因此 \mathcal{T}^π 将在分布中收敛到返回的真实分布 Z^π 。在本节中, 我们从图11所示的悬崖域上进行了经验验证这些结果。问题的动力学匹配了Sutton & Barto (1998) 给出的动力学。我们还研究了在采样的瓦斯林损失和分类投影下 (等式7) 下的分散式贝尔曼操作员的收敛性 (方程7)

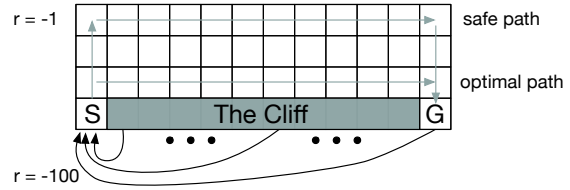


图11。悬崖环境 (Sutton & Barto, 1998)。

降低试图采取安全道路但有10%的机会随机采取另一项行动的政策。

我们使用每个状态的10000 Monte-Carlo (MC) 推出, 计算回报 Z^π 的基础真相分布。然后, 我们执行两个实验, 以我们的离散分布近似于每个状态的值分布。

在第一个实验中, 我们进行了监督的学习, 以学习瓦斯汀损失或分类投影 (方程7), 并具有跨膜损失。我们使用 Z^π 作为监督目标, 并对所有状态进行5000次扫描, 以确保两种方法都收敛。在第二个实验中, 我们使用相同的损失函数, 但是训练目标来自带有采样过渡的一步分配钟形操作器。我们使用 $V_{\text{MIN}} = -100$ 和 $V_{\text{MAX}} = -1$ 。用于示例更新, 我们在状态空间上执行的扫描的10倍。从根本上讲, 这些实验研究了两个培训制度

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

(最大程度地减少Wasserstein或分类损失)在理想（监督目标）和实用（采样一步的Bellman目标）条件下，将Wasserstein度量最小化。

在图10a中，我们显示了最终的瓦斯汀距离 $d_1(Z^\pi, Z_\theta)$ ，因为我们在改变原子的数量时，学到的分布与地面真相分布之间的最终距离 $d_1(Z^\pi, Z_\theta)$ 。该图表明，在监督和样本的贝尔曼设置中，分类算法确实确实使Wasserstein指标最小化。它还强调说，将瓦斯汀损失与随机梯度下降最小化是一般的浮躁，并确定了预言5所给出的直觉。在重复实验中，该过程融合到了 $d_1(Z^\pi, Z_\theta)$ 的不同值，表明存在局部最小值（更普遍（更普遍）原子更少）。

图10提供了有关为何采样的瓦斯汀距离可能表现不佳的进一步见解。在这里，我们看到沿悬崖旁的安全路径的五个不同状态在这两个损失下学到的近似值的累积密度。Wasserstein已融合到固定点分布，但没有很好地捕获真实（Monte Carlo）分布的分布。相比之下，分类算法更准确地捕获了真实分布的方差。

E. 补充视频和结果

在图13中，我们提供了在各种Atari 2600游戏培训期间显示C51代理的补充视频的链接。图12显示了C51在训练过程中的相对性能。图14提供了评估结果表，将C51与其他最先进的代理进行了比较。图15–18描绘了特别有趣的帧。

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

图13。培训期间C51的补充视频。

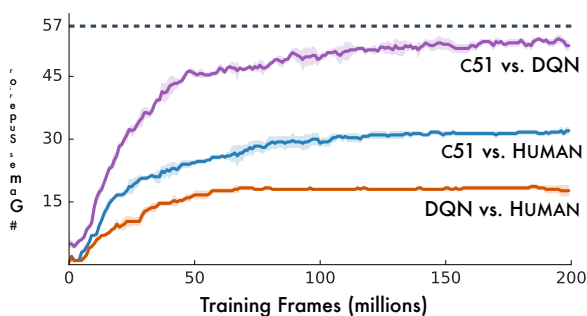


图12。代理商的训练能力大于基线（完全训练的DQN&Human）的Atari游戏数量。错误频段会产生标准偏差，平均值超过了游戏。

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

图14。从30个无操作动作开始，所有游戏的原始分数。Wang等人的参考值。（2016）。

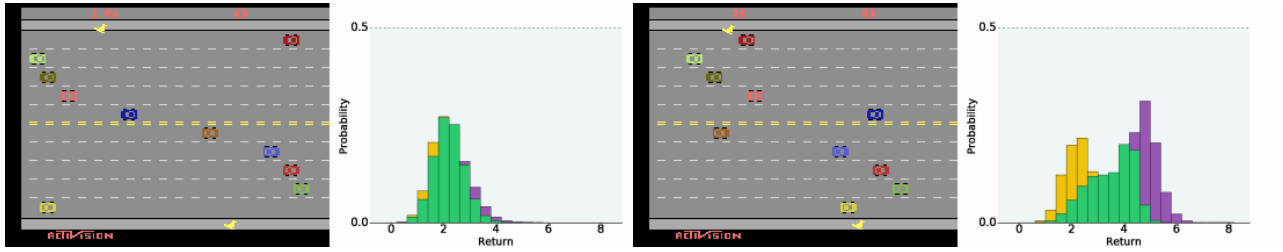


图15. 高速公路：代理在压力下区分了动作值分布。

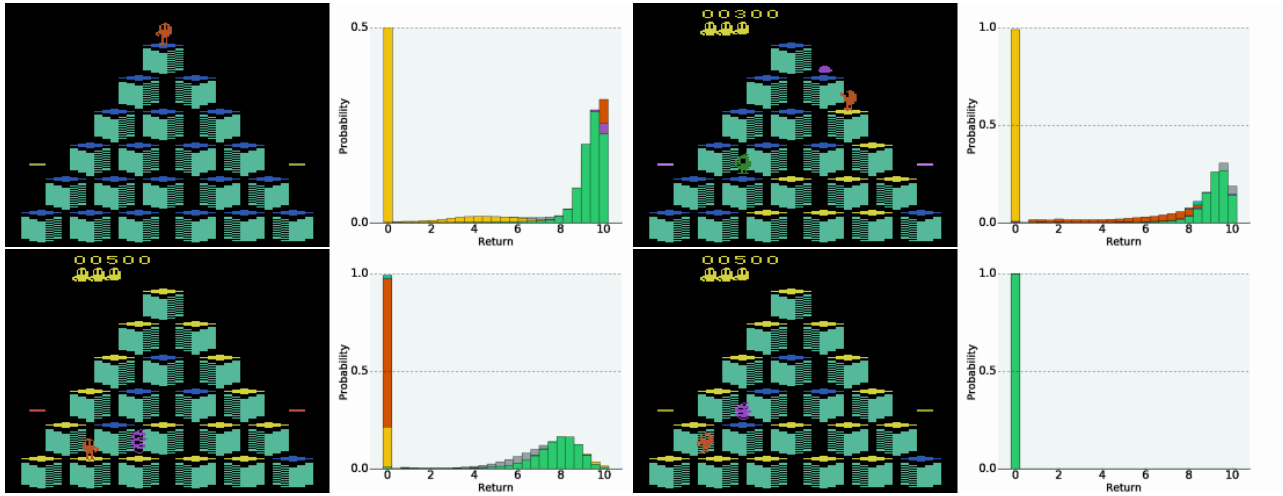


图16. Q*BERT：顶部，左右：预测哪些行动是不可恢复的致命。左下：价值分布显示出对错误动作的严重后果。右下角：代理商犯了一个巨大的错误。

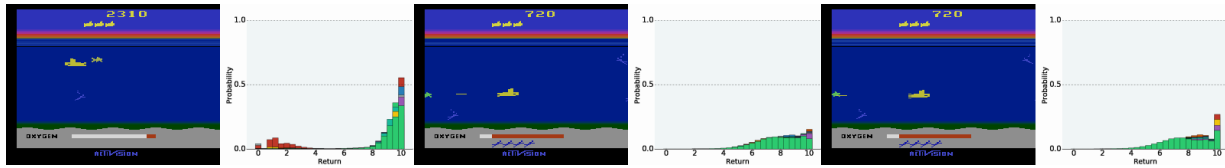


图17. Seaquest：左：双峰分布。中间：可能会击中鱼。右：绝对要击中鱼。

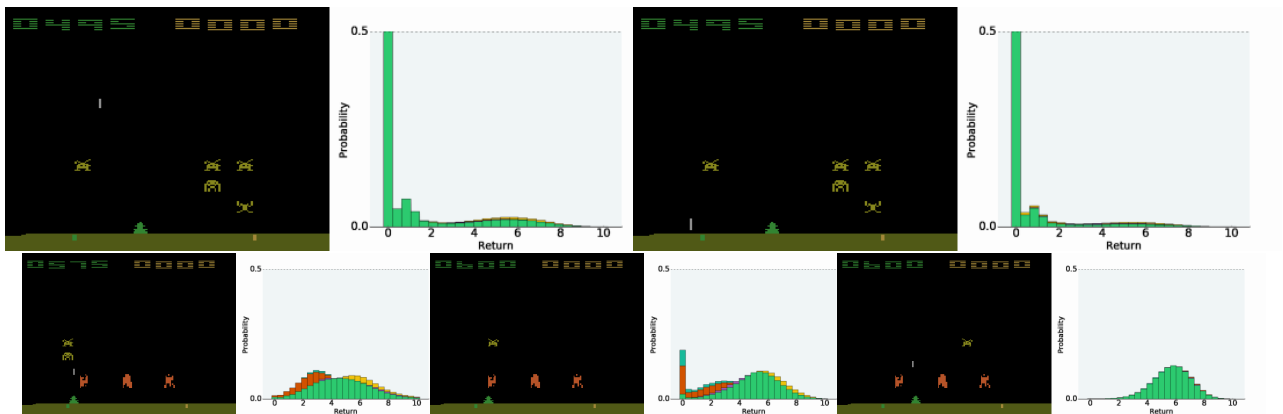


图18. 空间入侵者：左上：多模式分布，不确定性高。右翼：随后的框架，更确定的灭亡。左下：动作之间的明显差异。底部中间：不确定的生存。右右：一定的成功。