A Distributional Perspective on Reinforcement Learning

Marc G. Bellemare [*1] Will Dabney [*1] Rémi Munos [1]

【强化学习 47】Distributional RL

张楚珩
清华大学 交叉信息院博士在读

一篇很有启发性的工作，目前主流的强化学习方法主要关注价值函数的均值，这里提出把价值函数的分布也考虑进来。

原文传送门

Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017.

特色

这篇工作很有启发性，讲述了当价值函数不在仅仅是一个期望值而是是一个分布会怎样？相应的Bellman Operator是否还具有较好的性质？是否能够形成一个有效的算法？把分布考虑进来是否能够带来算法性能的提升？

过程

1. 考虑价值函数的分布有什么好处？

- 分布相比于均值能够对决策提供更多的信息，比如对于某些risk aware的场景，我们可能会更倾向于选择方差较小或者最坏情况较好的行动，而不是一味地选择均值较高的行动；
- 对于某些具有简并状态（state aliasing）的MDP或者POMDP，表征上看起来一样的状态可能具有完全不一样的两个价值函数，如果仅仅考虑均值，这部分信息就会被完全混淆；
- 考虑价值函数的分布能够缓解奖励稀疏的问题，较为稀疏的奖励会在通常的迭代中慢慢“稀释”，如果像文中的做法，稀疏的奖励在传播过程中更容易被留存下来（看到具体的算法之后才容易体会到这一点）

2. 价值函数的分布表示和距离度量

用分布来表示价值函数只需要把通常的价值函数 $V(x)$ 和 $Q(x,a)$ ，替换成相应的分布 $Z(x)$ 和 $Z(x,a)$ 即可。

距离度量使用Wasserstein距离

$$d_p(U,V) := \inf_{U,V} ||U-V||_p := \inf_{U,V} [\mathbb{E}_{\omega\in\Omega} ||U(\omega)-V(\omega)||^p]^{1/p}$$

对于两个随机变量 U,V （由于本文讨论价值函数的分布，可以认为它们都是一维随机变量），第一

个等式中的infimum是对于所有符合 U,V 各自边缘分布的 (U,V) 联合分布（参见本专栏另外的文章Wasserstein距离）。 $\omega\in\Omega$ 表示对于所有可能的实验结果取期望。

该距离度量有以下性质

$$
\begin{aligned}
d_p(aU, aV) &\le |a| d_p(U, V) \\
d_p(A + U, A + V) &\le d_p(U, V) \\
d_p(AU, AV) &\le \|A\|_p d_p(U, V).
\end{aligned}
$$

Lemma 1 (Partition lemma). *Let $A_1, A_2, \ldots$ 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*

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

（观察到，该引理左边可以在所有的联合分布中寻找最小值，而该定理的右边要求所寻找的联合分布协方差矩阵还需要按照A的划分是分块矩阵，这显然会使得找到的最小值更大，证明按照此思路易得）

注意到以上的定义都只针对两个表征价值的随机变量，当考虑价值函数的时候，还有 x 或者 (x,a) 的自变量输入，下面定义***两个分布价值函数之间的距离度量***。

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

可以证明它是一个距离度量，即满足三角不等式。

3. Policy Evaluation

有了分布价值函数的定义之后，我们考虑的第一个问题是对于一个给定的策略 π ，是否存在一个类似Bellman算子 $\mathcal{T}^\pi$ ，使得对于任意的初始分布价值函数，都能够在上面所定义的分布距离度量下，收敛到其真实分布价值函数？即常说的policy evaluation或者prediction问题（Sutton书）。

定义Bellman算子

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

$$
\begin{aligned}
P^\pi Z(x,a) &\overset{D}{:=} Z(X', A') \\
&\quad X' \sim P(\cdot \,|\, x, a),\ A' \sim \pi(\cdot \,|\, X'),
\end{aligned}
$$

对于这样的算子有很好的结果，即对于任意的初始分布价值函数，都能够在Wasserstein距离度量下，收敛到其真实分布价值函数。要证明这一点只需要证明contraction。

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

（要证明这一点即证明 $\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \le \gamma \bar{d}_p(Z_1, Z_2)$ ，把Bellman算子展开并且利用前面关于Wasserstein距离的性质就可以得到）

另外，Bellman算子 $\mathcal{T}^\pi$ 对于分布的前两阶中心矩也是contraction。

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*

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

（第一个式子的证明只需要利用期望算子的线性，这样就转化为了普通Bellman算子的contraction了；第二个式子还是把Bellman算子展开之后证明）

4. Control

前面讨论了prediction问题，现在讨论control的情形，即是否存在一个类似Bellman算子 $\mathcal{T}$ ，使得对于任意的初始分布价值函数，都能够在Wasserstein距离度量下，收敛到最优分布价值函数？

先说结论，对于value iteration类算法，一般关心两件事情，即

- 【问题1】是否每迭代一轮，都离最优分布函数更近，即在Wasserstein距离度量下有contraction？答案是***不能保证每轮迭代，Wasserstein距离都缩小***。在此问题上有一个更弱一点的结论，***其期望在 $\|\cdot\|_\infty$ 度量下有contraction***。
- 【问题2】多轮迭代之后，是否能够收敛到最优分布价值函数？该问题通常分为两部分，即***是否有不动点***和***是否能收敛到不动点***。对于第一个问题，***在排除掉一些琐碎的简并情况后，它具有不动点，并且不动点属于某个最优分布价值函数。***对于第二个问题，***它能够在Wasserstein距离度量下收敛到一族nonstationary的最优价值函数***。*（由于我们考虑的“最优”仍然是相对于期望值的，因此把分布引入进来的时候，期望相同的不同分布都同等地“最优”，这会造成一些混淆，这样的混淆造成前面所说的“琐碎的简并情况”）*

下面来具体说。

首先，最优策略的定义是按照均值来定义的，即均值最大的策略。相应的最优分布价值函数也是这种最优策略下对应的分布价值函数。

其次，Bellman算子定义如下

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

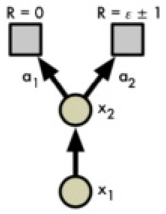
其中相对于价值函数的贪心策略是相对于分布的期望来定义的（注意正是这样只考虑均值的定义造

成了后面的琐碎）

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

【问题1】

为了说明不能保证每轮迭代都是的Wasserstein距离减小，文中举了一个反例。如图所示，做 a_1 的时候确定性得到奖励0；做 a_2 的时候各一半的概率得到图上所示奖励。最优分布价值函数在表中 Z^* ，从任意一个分布价值函数 Z 出发，做一次迭代，得到 $\mathcal{T}Z$ 。观察到 $d_1(\mathcal{T}Z, \mathcal{T}Z^*) > d_1(Z, Z^*)$ 是可能发生的，即 $\mathcal{T}$ 不是contraction。



	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	$\epsilon \pm 1$

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.

但是，其一阶矩是contraction。

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

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

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

（证明只需要利用期望的线性，转化为普通的Bellman算子contraction证明即可）

【问题2】

其收敛到的是一族不稳定的最优分布价值函数 Z^{**} ，这是啥意思呢？对于一族最优策略，按照任意序列去执行这个最优策略，所对应的价值函数就是***不稳定的最优分布价值函数***。

只考虑期望的时候，最优价值函数只可能是一个（Banach’s fixed point theorem），但是考虑分布的时候，由于没有contraction，因此就很可能出现一族最优价值函数。可以看到之所以会出现一族最优价值函数，是因为会出现一族最优策略，如果对于所有最优策略排个序，只允许有一个最优策略，那么就会产生一个唯一的不动点（定理第二部分）。

为什么在考虑分布的情况下会出现不稳定的情况呢？个人认为，对于同样的价值函数 $Q(x,\cdot)$ ，不论是greedy还是 ϵ-greedy策略都是没有歧义的；而对于分布价值函数 $Z(x,\cdot)$ ，就算是对均值的greedy策略，在均值相同的情况下，还能够根据不同分布的其他高阶矩做出不同的行动，这就产生了一族

在原本语义下同等最优的策略，这些策略造成了不稳定。

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*

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

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

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

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

在不考虑对最优策略排序情况下，会产生以下不稳定的情形。

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

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

个人认为，定理一已经挺好的了，给出的不稳定情形的例子过于极端，都是分布均值相同，而分布不同的情况，这种情况实际数值计算中出现概率较小，不稳定可以被缓解。

如果上面的定理看迷糊了，这里提供一个易于理解的版本

> If the optimal policy is unique, then the iterates $Z \leftarrow \mathcal{T}Z$ converge to Z^{π^*} .

5. 算法

首先第一个问题是如何表示一个分布，之前已经有算法来使用高斯分布来表示价值函数的分布，这种做法的缺点在于不能够表示多模的分布。本文把价值函数值的取值范围 $[V_{\min}, V_{\max}]$ 分为 N 个格子，

每个格子代表范围为 $\Delta z = \frac{V_{\max} - V_{\min}}{N-1}$ 的价值函数值，然后分别估计价值函数值落在每个格子内的概率。

使用神经网络 $\theta : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^N$ ，使用Boltzmann分布来表示价值函数的分布

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

这样表示之后还会存在一个问题，就是本来在同一个格子的值，在通过Bellman算子的更新之后，不能保证还落在同一个格子里面（因为有 γ 产生收缩，同时奖励值也不能保证正好是 Δz 的整数倍）。因此还需要做一个投影的操作，即按照线性比例投影到最近的格子中。

$$(\Phi \hat{\mathcal{T}} Z_\theta(x,a))_i = \sum_{j=0}^{N-1} \left[1 - \frac{\left| [\hat{\mathcal{T}} z_j]_{V_{\mathrm{MIN}}}^{V_{\mathrm{MAX}}} - z_i \right|}{\triangle z} \right]_0^1 p_j(x', \pi(x')),$$

整个Bellman算子的更新操作可以由下图表示

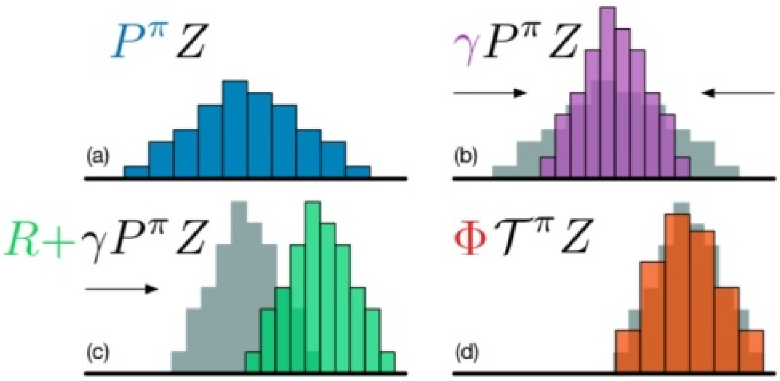


Figure 1. A distributional Bellman operator with a deterministic reward function: (a) Next state distribution under policy π, (b) Discounting shrinks the distribution towards 0, (c) The reward shifts it, and (d) Projection step (Section 4).

最后，相对于输出 N 个数值的神经网络来说，对分布拟合的目标就可以自然转化为cross-entropy

的损失函数，接下来使用梯度下降类方法就可以对该目标进行优化。最后得到下面的算法。

Algorithm 1 Categorical Algorithm

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

实验结果

文章在ALE上做实验，毕竟Q-learning相关算法要求行动空间是离散的。个人认为实验结果里面有以下看点。

首先，该算法能够学习到非平庸的情况，而不是全是看起来类似Gaussian的分布。比如对于一些致命的操作，能够在分布中很明确地反映出来。

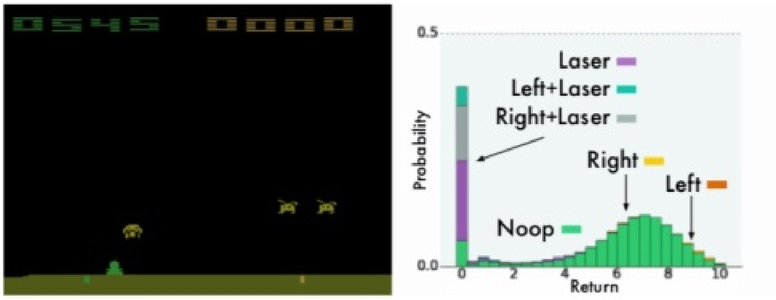


Figure 4. Learned value distribution during an episode of SPACE INVADERS. Different actions are shaded different colours. Returns below 0 (which do not occur in SPACE INVADERS) are not shown here as the agent assigns virtually no probability to them.

一些致命的操作，能够在分布中很明确地反映出来

其次，实验显示对于一些简并的状态，能够学习到多模的分布（两种可能的情况的加和），之前的很多算法是不能够表示出来这样的结果的。

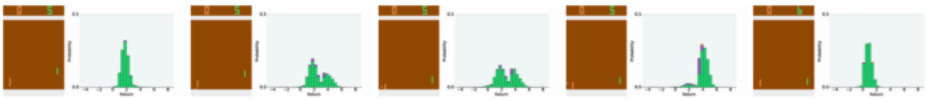


Figure 5. Intrinsic stochasticity in PONG.

对于一些不确定的状态，能够学习到多模的分布

最后，该算法对于奖励十分稀疏的任务提升较大，主要是由于稀疏的奖励在分布的传播中相对不容易lost。

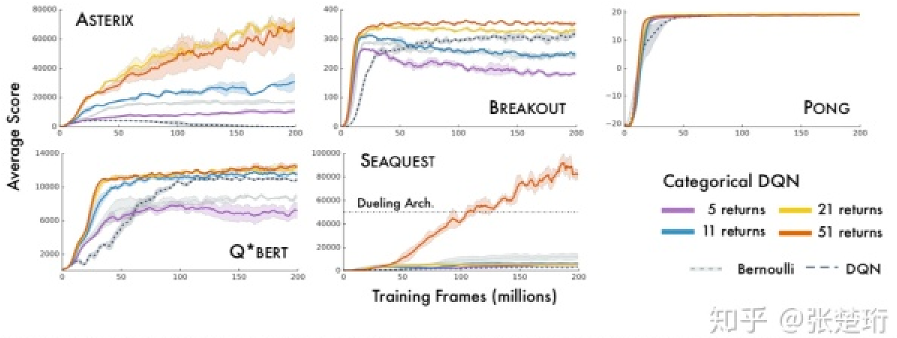


Figure 3. Categorical DQN: Varying number of atoms in the discrete distribution. Scores are moving averages over 5 million frames.

注意到最后一个任务奖励稀疏，相比于之前的算法提升较大

强化学习里面考虑价值函数的分布/深度学习里面考虑预测数值的分布，在金融中有比较实际的意义，很多情况我们希望能够在提高期望收益的时候同时控制风险，这样给出一个分布有用的信息就比单纯一个期望值的信息大很多。比如可以优化一个lower confidence bound而不仅仅是一个均值。

编辑于 2019-03-29