The Value Function Polytope in Reinforcement Learning

Robert Dadashi [1] Adrien Ali Taïga [1 2] Nicolas Le Roux [1] Dale Schuurmans [1 3] Marc G. Bellemare [1]

【强化学习 74】Value Function Polytope

张楚珩
清华大学 交叉信息院博士在读

原文传送门

Dadashi, Robert, et al. "The Value Function Polytope in Reinforcement Learning." arXiv preprint arXiv:1901.11524 (2019).

特色

一篇理论的工作，描述了在 tabular case 下策略空间到价值函数的映射关系。虽然该映射是一个非线性的映射，但是该映射中仍然包含了很多线性关系。比如所有策略能够达到的价值函数形成一个 polytope，即其边界都是线性的；如果只有一个状态上的策略不同，那么其对应的价值函数在同一条直线上。从这些这个几何的视角出发，文章还演示了常见强化学习算法在这个 polytope 上的演化规律，比如 value iteration、policy iteration、policy gradient、cross-entropy method 等。本文对于理解这些方法也会有很好的启发和帮助。

过程

1. Polytope

首先定义了 convex polytope / polyhedron，

Definition 1 (Convex Polytope). *P is a convex polytope iff there are $k \in \mathbb{N}$ points $x_1, x_2, ..., x_k \in \mathbb{R}^n$ such that $P = Conv(x_1, \ldots, x_k)$.*

Definition 2 (Convex Polyhedron). *P is a convex polyhedron iff there are $k \in \mathbb{N}$ half-spaces $\hat{H}_1, \hat{H}_2, ..., \hat{H}_k$ whose intersection is P, that is*

$$P = \cap_{i=1}^{k} \hat{H}_k.$$

然后它们的集合即为 polytope / polyhedron

Definition 3 (Polytope). *A (possibly non-convex) polytope is a finite union of convex polytopes.*

Definition 4 (Polyhedron). *A (possibly non-convex) polyhedron is a finite union of convex polyhedra.*

它们之间的等价关系如下

A celebrated result from convex analysis relates these two definitions: a *bounded*, convex polyhedron is a convex polytope (Ziegler, 2012).

接下来文中定义了 relative neighborhood、relative interior、relative boundary 和 hyperplane。

For an affine subspace $K \subseteq \mathbb{R}^n$, $V_x \subset K$ is a *relative neighbourhood* of x in K if $x \in V_x$ and V_x is open in K. For $P \subset K$, the *relative interior* of P in K, denoted $\text{relint}_K(P)$, is then the set of points in P which have a relative neighbourhood in $K \cap P$. The notion of "open in K" is key here: a point that lies on an edge of the unit square does not have a relative neighbourhood in the square, but it has a relative neighbourhood in that edge. The *relative boundary* $\partial_K P$ is defined as the set of points in P not in the relative interior of P, that is

$$\partial_K P = P \setminus \text{relint}_K(P).$$

Finally, we recall that $H \subseteq K$ is a *hyperplane* if H is an affine subspace of K of dimension $\dim(K) - 1$.

之所以要定义 relative，可以做如下理解。考虑一个空间 $\mathbb{R}^3$，考虑一个点集 $\{(x,y,0)|x^2+y^2<1\}$，它的 interior 是什么呢？考虑 interior 的定义为在其中的一个点上画一个包含这个点的小球，如果小球中的每一个点都在集合内，那么这个点就是 interior。按照这个定义，上面这个点集的 interior 就是空集，因为小球总会包含不在该平面上的点，这显然很不符合直观感受。因此会定义一个 relative interior，这个点集相对于它所在 x-y 平面的 interior 就是这个圆圈的内部。由此，还可以定义 relative neighborhood 和 relative boundary。

下面给出了 polyhedron 的一些性质。

Proposition 1. *P is a polyhedron in an affine subspace $K \subseteq \mathbb{R}^n$ if*

(i) P is closed;

(ii) There are $k \in \mathbb{N}$ hyperplanes $H_1, .., H_k$ in K whose union contains the boundary of P in K: $\partial_K P \subset \cup_{i=1}^{k} H_i$; and

(iii) For each of these hyperplanes, $P \cap H_i$ is a polyhedron in H_i.

没太懂这里讲的 closed 和前面讲的 bounded 有啥区别。

2. 价值函数空间

定义：所有稳定策略形成的价值函数空间（space of value functions）为 $\mathcal{V} \subset \mathbb{R}^{\mathcal{S}}$ 。所有的稳定策略形成的空间 $\mathcal{P}(\mathcal{A})^{\mathcal{S}}$ 。价值函数的映射 f_v 把一个策略映射到其对应的价值函数空间上，同时也用它来表示集合的映射关系，即

$$\mathcal{V} = f_v(\mathcal{P}(\mathcal{A})^{\mathcal{S}}) = \left\{ f_v(\pi) \,|\, \pi \in \mathcal{P}(\mathcal{A})^{\mathcal{S}} \right\}. \qquad (2)$$

其映射关系可以写作

$$f_v(\pi) = (I - \gamma P^{\pi})^{-1} r_{\pi}.$$

下图画出了一些两个状态下的价值函数空间的形态

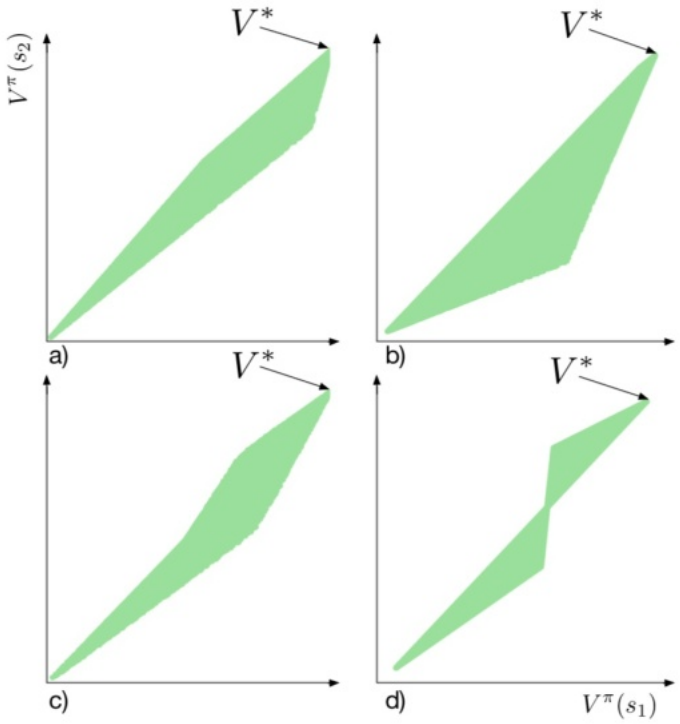

Figure 2. Space of value functions for various two-state MDPs.

它具有如下基本性质

- Dominance：最优价值函数在每一个维度上都比其他价值函数更大；
- Monotonicity：价值函数空间所有的边都和原点的夹角为正（假设了 reward > 0）；
- Continuity：价值函数空间是连续的；
- Compact and connected：价值函数的空间是 compact 并且相互连通；
- Differentiable： f_* 在策略空间上无限阶可导；

3. 关于策略的定义

下面会用到的以下两个概念：policy agreement 和 policy determinism。前者讲的是策略之间的关

系，即两个策略在一部分状态上表现一样；后者讲的是策略的性质，即策略在某个状态、一部分状态或者所以状态上是确定性的。

Definition 5 (Policy Agreement). *Two policies π_1, π_2 agree on states $s_1, .., s_k \in \mathcal{S}$ if $\pi_1(\cdot \mid s_i) = \pi_2(\cdot \mid s_i)$ for each s_i, $i = 1, \ldots, k$.*

For a given policy π, we denote by $Y^{\pi}_{s_1,\ldots,s_k} \subseteq \mathcal{P}(\mathcal{A})^{\mathcal{S}}$ the set of policies which agree with π on $s_1, \ldots, s_k$; we will also write $Y^{\pi}_{\mathcal{S}\setminus\{s\}}$ to describe the set of policies that agree with π on all states except s. Note that policy agreement does not imply disagreement; in particular, $\pi \in Y^{\pi}_{\mathscr{S}}$ for any subset of states $\mathscr{S} \subset \mathcal{S}$.

Lemma 2. *Consider two policies π_1, π_2 that agree on $s_1, \ldots, s_k \in \mathcal{S}$. Then the vector $r_{\pi_1} - r_{\pi_2}$ has zeros in the components corresponding to $s_1, \ldots, s_k$ and the matrix $P^{\pi_1} - P^{\pi_2}$ has zeros in the corresponding rows.*

4. 价值函数映射的线性关系

当策略只能在一部分状态上变化时，价值函数的变化范围也限定在一个子空间上

考虑到

$$f_v(\pi) = (I - \gamma P^{\pi})^{-1} r_{\pi}.$$

令矩阵 $(I-\gamma P^{\pi})^{-1} = [C^{\pi}_1 | C^{\pi}_1 | \cdots | C^{\pi}_{|\mathcal{S}|}]$ ，定义

$$H^{\pi}_{s_1,\ldots,s_k} = V^{\pi} + Span(C^{\pi}_{k+1}, \ldots, C^{\pi}_{|\mathcal{S}|}).$$

有如下性质

Lemma 3. *Consider a policy π and k states $s_1, \ldots, s_k$. Then the value functions generated by $Y^{\pi}_{s_1,..,s_k}$ are contained*

in the affine vector space $H^{\pi}_{s_1,..,s_k}$*:*

$$f_v(Y^{\pi}_{s_1,..,s_k}) = \mathcal{V} \cap H^{\pi}_{s_1,..,s_k}.$$

下面考虑只有一个状态上的策略不同的情况。

下面的引理给出，***对于任意两个策略只在某一个状态下不同，它们的价值函数形成一条直线；形成的这条直线和原点的夹角为正；并且它们之间可以单调连续地 interpolate***

Lemma 4. *Consider the ensemble* $Y^{\pi}_{\mathcal{S}\setminus\{s\}}$ *of policies that agree with a policy* π *everywhere but on* $s \in \mathcal{S}$*. For* $\pi_0, \pi_1 \in Y^{\pi}_{\mathcal{S}\setminus\{s\}}$ *define the function* $g : [0,1] \to \mathcal{V}$

$$g(\mu) = f_v(\mu\pi_1 + (1-\mu)\pi_0).$$

Then the following hold regarding g*:*

(i) g *is continuously differentiable;*

(ii) (Total order) $g(0) \preccurlyeq g(1)$ *or* $g(0) \succcurlyeq g(1)$*;*

(iii) If $g(0) = g(1)$ *then* $g(\mu) = g(0)$*,* $\mu \in [0,1]$*;*

(iv) (Monotone interpolation) If $g(0) \neq g(1)$ *there is a* $\rho : [0,1] \to \mathbb{R}$ *such that* $g(\mu) = \rho(\mu)g(1) + (1 - \rho(\mu))g(0)$*, and* ρ *is a strictly monotonic rational function of* μ*.*

那么这条直线肯定会“戳穿”价值函数空间，那么其顶点是什么呢？下面定理给出，其顶点一定是在该状态下的确定性策略。证明方法用反证法即可，即先证明存在这样的顶点，假设其对应的策略，然后证明该策略可以被分解为确定性策略的加权和，并且被确定性策略“包围”。

Theorem 1. *[Line Theorem] Let s be a state and π, a policy. Then there are two s-deterministic policies in $Y^{\pi}_{\mathcal{S}\setminus\{s\}}$, denoted π_l, π_u, which bracket the value of all other policies $\pi' \in Y^{\pi}_{\mathcal{S}\setminus\{s\}}$:*

$$f_v(\pi_l) \preccurlyeq f_v(\pi') \preccurlyeq f_v(\pi_u).$$

Furthermore, the image of f_v restricted to $Y^{\pi}_{\mathcal{S}\setminus\{s\}}$ is a line segment, and the following three sets are equivalent:

(i) $f_v\left(Y^{\pi}_{\mathcal{S}\setminus\{s\}}\right)$,

(ii) $\{f_v(\alpha\pi_l + (1-\alpha)\pi_u) \,|\, \alpha \in [0,1]\}$,

(iii) $\{\alpha f_v(\pi_l) + (1-\alpha) f_v(\pi_u) \,|\, \alpha \in [0,1]\}$.

注意到上面 (ii) 和 (iii) 两条线相同，但是其中的元素的一一对应关系并不是线性的，可以参考如下例子。

Example 1. *Suppose $\mathcal{S} = \{s_1, s_2\}$, with s_2 terminal with no reward associated to it, $\mathcal{A} = \{a_1, a_2\}$. The transitions and rewards are defined by $P(s_2|s_1, a_2) = 1, P(s_1, |s_1, a_1) = 1, r(s_1, a_1) = 0, r(s_1, a_2) = 1$. Define two deterministic policies π_1, π_2 such that $\pi_1(a_1|s_1) = 1, \pi_2(a_2|s_1) = 1$. We have*

$$f_v((1-\mu)\pi_1 + \mu\pi_2) = \begin{bmatrix} \frac{\mu}{1-\gamma(1-\mu)} \\ 0 \end{bmatrix}.$$

同时，以上结论只适用于两个只相差一个状态的策略，对于任意的两个策略，其策略内插形成的价值函数不是一条直线，例如下面的例子。

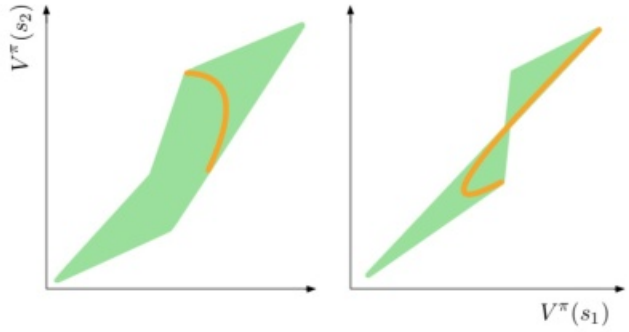

Figure 4. Value functions of mixtures of two policies in the general case. The orange points describe the value functions of mixtures of two policies.

5. 价值函数映射的其他结论以及价值函数空间的形态

其实有了定理 1 之后其实就能够对 f_* 的映射关系有了很好的认识了，由此这里直接形象地总结一下相应的结论，具体的定理见原文。

- 价值函数空间在确定性策略（可能是对于某一些状态的确定性策略）张成的一个 convex hull 中。如下图所示所示。

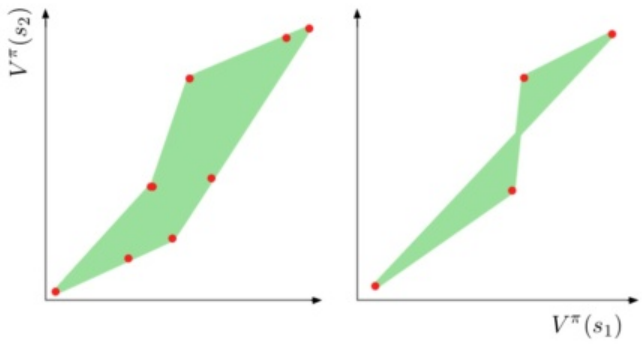

Figure 5. Visual representation of Corollary 1. The space of value functions is included in the convex hull of value functions of deterministic policies (red dots).

- 由于一共有 $|S|$ 个状态，从任意策略出发，每次变化一个状态上的策略，相应地在价值函数空间上会形成一条直线。因此从任意一个策略对应的价值函数出发，都可以经过 $k<|S|$ 条直线（相应的策略只在一个状态上变化）到达另一个策略对应的价值函数。
- 考虑一族策略，它们在状态 $\{s_1,\cdots,s_k\}$ 上都和 π 一样，这一族策略形成的价值函数族的边界一定是由至少某个状态上是确定性的策略构成的，边界上的策略不仅在状态 $\{s_1,\cdots,s_k\}$ 上都和 π 一样，而且在某个除状态 $\{s_1,\cdots,s_k\}$ 外的状态上是确定性的。如下图所示

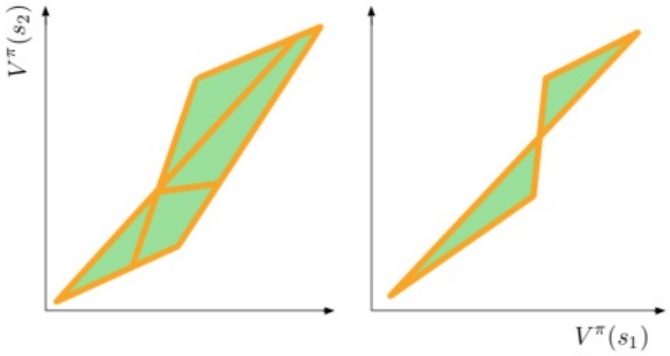

Figure 6. Visual representation of Corollary 3. The orange points are the value functions of semi-deterministic policies.

- 同样考虑一族策略，它们在状态 $\{s_1,\cdots,s_k\}$ 上都和 π 一样，这一族策略形成的价值函数族是一个 polytope。作为一个特殊情况， ν 也是一个 polytope。

6. 不同算法在价值函数空间中的演化

注意，一下所有的分析中的更新都去掉了 stochastic 的成分，相当于每一步都更新的是更新公式的均值。同时所有涉及到 function approximation 的部分也都等价地改为 tabular case 的情形，即 state embedding 为 one-hot vector，function approximation 为线性拟合。

Value iteration

Value iteration 是做 optimality Bellman operator 的迭代，可以看出，它迭代过程中产生的价值函数并不保证对应实际的策略（即会超出 ν 的范围），不过最终它会迭代到最优价值函数位置。

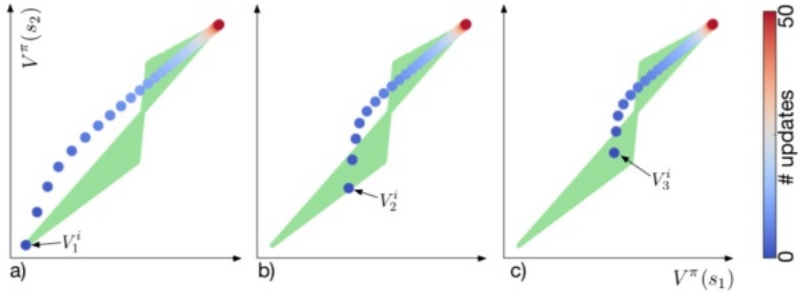

Figure 7. Value iteration dynamics for three initialization points.

Policy iteration

Policy iteration 分为 policy evaluation 和 policy improvement 两个步骤。考虑每次 policy improvement 都是一个相对于前一步价值函数的 greedy 策略，那么每次的 policy 都会是一个确定性策略，即会对应 polytope 的一个顶点。

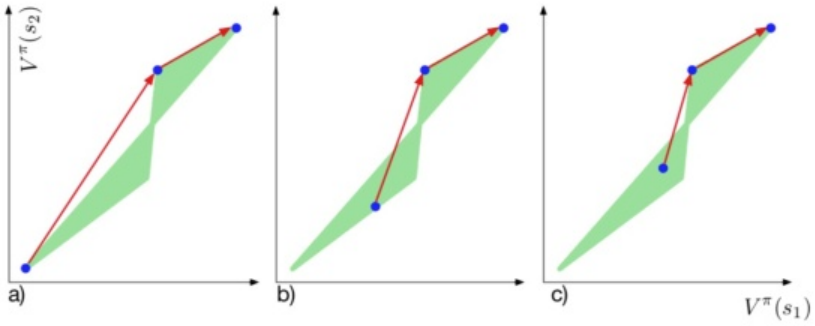

Figure 8. Policy iteration. The red arrows show the sequence of value functions (blue) generated by the algorithm.

Policy gradient

策略梯度方法的收敛速率对于初始值十分敏感，比如从图 a 可以看出，它在原始状态附近过了很久才开始往最优策略方向更新。同时它还可能会陷入局部极小值点，在该点上，策略逐渐变为确定性策略，这样会缺乏探索，使得它不知道其实在某状态上选择另外一个行动会更好。

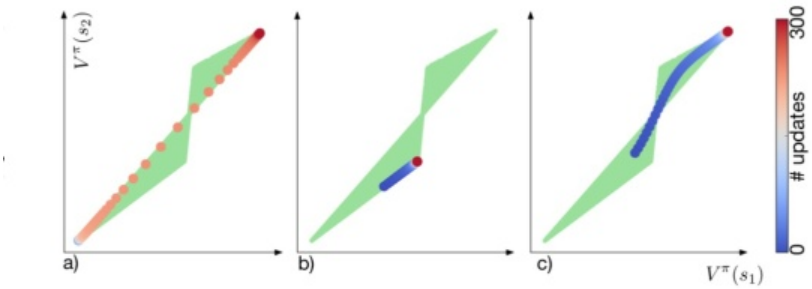

Figure 9. Value functions generated by policy gradient.

可以看出上述问题的主要原因都在于其策略太过于贴近 polytope 边界（缺乏探索），因此可以加入 entropy 正则项来缓解这一问题。

Entropy regularized policy gradient

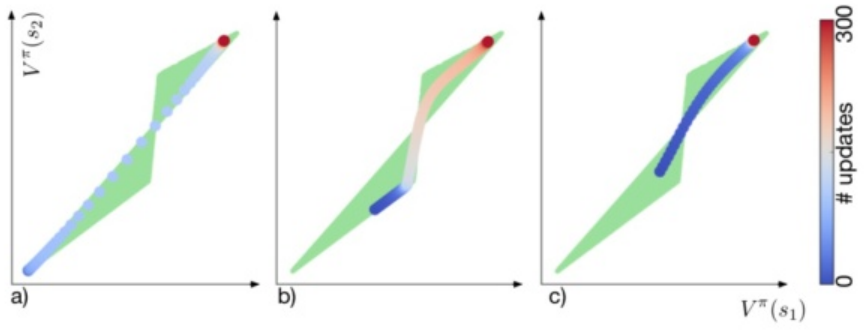

Figure 10. Value functions generated by policy gradient with entropy, for three different initialization points.

Natural policy gradient

$$\theta_{k+1} := \theta_k + \eta F^{-1} \nabla_\theta J(\theta_k).$$

从下面的实验结果可以看出，使用这样的自然梯度可以避免像之前 policy gradient 那样多次迭代仍然陷在某个区域附近；同时，它不加 entropy regularization 也没有出现（在这个例子中）陷入局部极小的情况。它的更新轨迹和 policy iteration 比较类似。

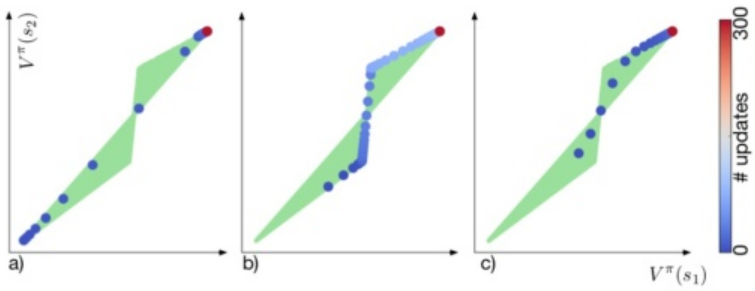

Figure 11. Natural policy gradient.

Cross-entropy method

这是一种 gradient-free 的方法，可以见专栏之前的文章（CEM）。这里发现如果不加噪声，它比较容易陷入局部极小值。

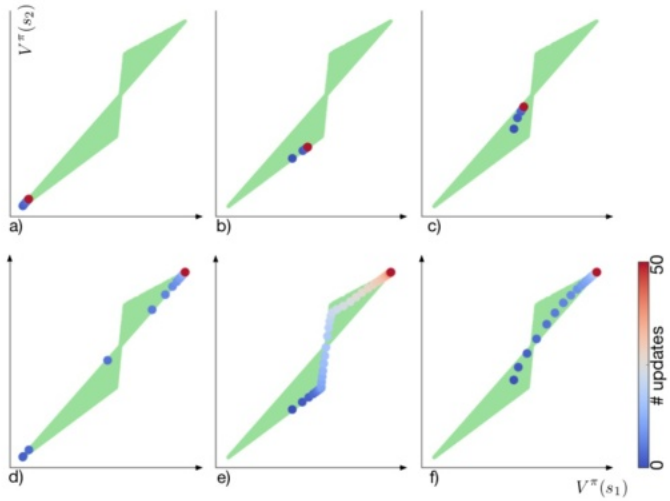

Figure 12. The cross-entropy method without noise (CEM) (a, b, c); with constant noise (CEM-CN) (d, e, f).

编辑于 2019-06-21

强化学习 (Reinforcement Learning)　优化