The Value Equivalence Principle in Infinite Domains and in Policy Search

Alireza Kazemipour

July 2025

Abstract

As of August 22, 2025: I want to pose the research question (the gap in the literature) that I'd like to answer. The atomic focus will be on Section 1. The gap that I have found is the lack of value equivalence principle in the function approximation regime and also policy search approaches.

1 Motivation

Let A_t be the action that the agent took at state S_t at time step t while following policy π and let R_{t+1} be the reward received. Also, let s_0 be the initial state of the interaction. The performance metric in the discounted setting is the expected cumulative discounted return defined as

$$J(\pi) = \mathbb{E}_\pi \left[\sum_t \gamma^t R_{t+1} \middle| S_t = s_0 \right]. \tag{1}$$

The goal is find a policy π^* such that

$$J(\pi^*) = \max_\pi J(\pi). \tag{2}$$

First, let us model the interaction as an MDP $\langle \mathcal{S}, \mathcal{A}, P, r \rangle$. There are two general approaches to find π^* known as *policy search* and *value-based* methods. Policy search methods tries to solve the optimization problem of Equation (2) directly. Value-based methods break down Equation (1) using dynamic programming for each state s that the agent experiences during the interaction, known as the state-value function, as the following

$$V_\pi(s) \coloneqq \mathbb{E}_\pi \left[\sum_t \gamma^t R_{t+1} \middle| S_t = s \right] = \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{s'} P\left(s'|s,a\right) V_\pi\left(s'\right) \middle| S_t = s, a \sim \pi(\cdot \mid s) \right]. \tag{3}$$

Then, value-based methods find π^* through

$$\pi^* \in \arg\max_\pi V_\pi(s), \quad \forall s \quad (\text{if } \arg\max \text{ exists}).$$

Note that the goal is solely finding π^* and although we modeled the interaction as an MDP, nowhere there exits a need to learn anything about the mean reward r and the transition kernel P. All that is useful are the quadruples $(S_t, A_t, R_{t+1}, S_{t+1})_t$ that agent experiences through its interaction with environment. The independence from learning r and P creates a contention on how to find π^*. On one hand, if learning r and P is hard (or even impossible), it does not exacerbate the learning of the agent. On the other hand, interaction with the environment can be expensive and learning internal models of r, and p can significantly accelerates the agent's learning. These two perspectives have resulted on two downstream approaches employed by both policy search and value-based methods called model-free and model-based.

Model-free approaches do not try to estimate r, nor P, so are not the focus of this work. In model-based methods, the agent tries to learn r, and P. For the sake of simplicity, let us assume that the agent knows r in advance and only needs to learn P, and let us focus on only the value-based methods to unravel our argument (**AK**: how do model-based policy search use models? Answer in.)

If $\mathcal{S}$ and $\mathcal{A}$ are finite, $V_\pi(s) \in \mathbb{R}$ is a point-wise function for any policy, hence we can use a vector representation to write Equation (3) as

$$T^\pi(V) = \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{s'} P\left(s'|s,a\right) V_\pi\left(s'\right) \middle| S_t = s, a \sim \pi(\cdot \mid s) \right].$$

So, if we want to find and estimate $\widehat{P}$ such that

$$\widehat{T}^\pi(V) = \mathbb{E}_\pi \left[r(s,a) + \gamma \sum_{s'} \widehat{P}\left(s'|s,a\right) V^\pi\left(s'\right) \middle| S_t = s, a \sim \pi(\cdot \mid s) \right],$$

and $T^\pi(V^\pi) - \widehat{T}^\pi(V^\pi)$ is zero, one way is to make $\widehat{P}$ as close as possible to P in some measure of closeness. One objective to learn P is to minimize the ℓ_1 distance between $\widehat{P}$ and P for all state-actions (s,a),

$$\widehat{P} = \arg\min_{P' \in \mathcal{P}} \|P'(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1, \tag{4}$$

where $\mathcal{P}$ is the set of transition kernels. The solution to Equation (4) coincides with the maximum-likelihood estimation of P, meaning that if (s,a) has been visited v times, and the next visited state is S'_i after the ith visit, then

$$\widehat{P}(\cdot \mid s,a) = \frac{1}{v} \sum_{i=1}^{v} \mathbb{I}\left\{S'_i = s'\right\}, \quad \forall s' \in \mathcal{S},$$

where $\mathbb{I}$ is the indicator function. The obtained $\widehat{P}$ in this way is **accurate**, i.e., $\widehat{P} \approx P$ (for large enough v) for all state-actions, and also **useful** because $T^\pi(V^\pi) - \widehat{T}^\pi(V^\pi) = 0$, which consequently enables us to use $\widehat{T}(V)$ to find π^*.

However, if $\mathcal{S}$, or $\mathcal{A}$ are infinite, then finding an accurate model is impossible in general. Because the agent needs to approximate P using a model class $\widetilde{\mathcal{P}}$ that does not necessarily contain P. Hence inevitably, the agent has approximate P using the best model available in $\widetilde{\mathcal{P}}$. However, now that the approximation is evitable, accuracy of models are rendered obsolete and the main goal should be on the usefulness of models. The preference of usefulness over accuracy has already be argued by (Grimm et al., 2020) for finite MDPs as *the value equivalence principle*.

Definition 1 (Grimm et al. (2020), Definition 1)**.** Let $\widetilde{\Pi} \subseteq \Pi$ be a set of policies and let $\widetilde{\mathcal{V}} \subseteq \mathcal{V}$ be a set of state-value functions. We say that models P_1, and P_2 are value equivalent with respect to $\widetilde{\Pi}$ and $\widetilde{\mathcal{V}}$ if and only if

$$T_1^\pi(V) = T_2^\pi(V), \quad \text{for all } V \in \widetilde{\mathcal{V}}, \text{ and for all } \pi \in \widetilde{\Pi}.$$

Since in the infinite spaces the agent employs function approximation, the full equally in Definition 1 is not achievable. Since the agent's approximation is constrained to the model class and the class of state-value functions it can represent, we define the *constrained* value equivalence principle.

Definition 2. Let $\widetilde{\Pi} \subseteq \Pi$ be a set of policies, $\mathcal{V}_w$ be the class of state-value functions parametrized by w, and $\mathcal{P}_\theta$ be the model class parametrized by θ. Two models P_{θ_1} and P_{θ_2} are constrained ϵ-value equivalent with respect to $\widetilde{\Pi}$, and $\mathcal{V}_w$ if and only if

$$|T_1^\pi(V_w)_s - T_2^\pi(V_w)_s| \leq \epsilon, \quad \forall s \in \mathcal{S}, V_w \in \widetilde{\mathcal{V}}_w, \pi \in \widetilde{\Pi}, \text{ and } \epsilon > 0,$$

where we define

$$T_i^\pi(V_w)_s \coloneqq \mathbb{E}_\pi \left[r(s,a) + \gamma \int_{\mathcal{S}} P_{\theta_i}\left(ds'|s,a\right) V_w^\pi\left(s'\right) \middle| S_t = s, a \sim \pi(\cdot \mid s) \right],$$

and $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is the feature-vector.

We also emphasize again that the goal is finding π^* and not even an accurate estimate of V^*. The ϵ approximation of Definition 2 lets the agent benefit from the action-gap phenomenon (Farahmand, 2011) to find an ϵ-optimal state-values that are enough to find the optimal policy.

ϵ in Definition 2 accounts for two possibilities: the error in estimating the model, and the error in estimating the value functions. (**AK**: continue)

2 Next Steps

1. Connect the new definition to the special case of linear function approximation with exponential family of distributions for P, thoroughly studied by Farahmand et al. (2017).
2. Use projected Bellman operator in the definition instead. But with an infinite dimensional Φ, Banach spaces? Functional analysis definitely helps here.
3. ODE definition.

References

Farahmand, A.-m. (2011). Action-Gap Phenomenon in Reinforcement Learning. In *Advances in Neural Information Processing Systems*.

Farahmand, A.-M., Barreto, A., and Nikovski, D. (2017). Value-Aware Loss Function for Model-based Reinforcement Learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 1486–1494.

Grimm, C., Barreto, A., Singh, S., and Silver, D. (2020). The value equivalence principle for model-based reinforcement learning. *Advances in neural information processing systems*, 33:5541–5552.