The Value Equivalence Principle: A Summary

Alireza Kazemipour

August 22, 2025

Abstract

As of August 22, 2025: In this document I intend to summarize, highlight my gaps in understanding, and my gut feelings around future directions regarding: Value-aware model learning (VAML) (Farahmand et al., 2017), the value equivalence (VE) principle (Grimm et al., 2020), proper value equivalence (PVE) (Grimm et al., 2021), and model equivalence principle for risk-sensitive reinforcement learning (Kastner et al., 2023). Especially, the first and the last references are pretty precise, so I should be able to understand them.

Notation guide

- Bold letters denote sets.
- Calligraphic letters denote distributions.
- Capital letters denote subsets (note that random variables are subsets).
- Lower case letter denote elements that belong to a set.

1 Introduction

Let our MDP be the tuple $\langle \mathbf{S}, \mathbf{A}, p^*, r^*, \gamma \rangle$, where p^* is the transition kernel and $r^* : \mathbf{S} \times \mathbf{A} \to \mathfrak{B}(\mathbb{R})$ is the immediate expected reward function which we assume is know to the agent in advance. The goal in model-based reinforcement learning (MBRL) has traditionally been learning an estimate $\hat{p}$ of p^*, and then using $\hat{p}$ for planning to produce an optimal policy. Estimating p^* by $\hat{p}$ is problem of conditional probability estimation and the goal is to make $\hat{p}$ as *close* as possible to p^*.

One approach to estimate p^* is by maximum-likelihood estimation (MLE) method. The reason MLE is an appropriate approach is because of its relation to KL divergence. We know that for two distributions p_1, and p_2, the KL divergence $\mathsf{KL}(p_1||p_2)$ is zero if and only of $p_1 = p_2$ almost surely. So, now we show that maximizing the likelihood, minimizes the KL divergence which is the goal. Suppose p_1 is the distribution we want to estimate with the true parameter θ^*, and p_2 is our estimate. We want

to find parameters θ_{min} such that

$$
\begin{aligned}
\theta_{\mathsf{min}} &= \arg\min_{\theta} \mathsf{KL}(p_1(\cdot;\theta^*)||p_2(\cdot;\theta)) \\
&= \arg\min_{\theta} \mathbb{E}_{x\sim p_1(\cdot;\theta^*)}\left[\log \frac{p_1(x;\theta^*)}{p_2(x;\theta)}\right] \\
&= \arg\min_{\theta} \mathbb{E}_{x\sim p_1(\cdot;\theta^*)}\left[\log p_1(x;\theta^*) - \log p_2(x;\theta)\right] \\
&= \arg\min_{\theta} \mathbb{E}_{x\sim p_1(\cdot;\theta^*)}\left[-\log p_2(x;\theta)\right] && (\log p_1(x;\theta^*) \text{ doesn't affect the argument of minima}) \\
&= \arg\max_{\theta} \mathbb{E}_{x\sim p_1(\cdot;\theta^*)}\left[\log p_2(x;\theta)\right] \\
&= \arg\max_{\theta} \mathbb{E}_{x\sim p_1(\cdot;\theta^*)}\left[\log p_2(x;\theta)\right] \\
&= \arg\max_{\theta} \sum_{i}^{n} p_1(x_i;\theta^*) \log p_2(x_i;\theta) && (\text{If we have a dataset of size } n) \\
&= \arg\max_{\theta} \log \Pi_i^n p_2(x_i;\theta) && (p_1(x_i;\theta^*) \text{ doesn't affect the argument of maxima}) \\
&= \arg\max_{\theta} \Pi_i^n p_2(x_i;\theta). && (\text{MLE definition})
\end{aligned}
$$

Hence, MLE is a viable option to estimate p^* by $\hat{p}$. The MLE approach has been the dominant strategy in MBRL traditionally. Making $\hat{p}$ close to p^* through MLE in MBRL is also justified by the fact that the resulting loss because of the mismatch of $\hat{p}$ and p^* in estimating action value of a policy π for state-action $(s,a) \in \mathbf{S}\times\mathbf{A}$ is upper bounded as the following:

$$
\begin{aligned}
\ell\left(\hat{p}, p^*; v_\pi\right)(s,a) &= \left|\left[p^*(\cdot \mid s,a) - \hat{p}(\cdot \mid s,a)\right] v_\pi(\cdot)\right| \\
&= \left\langle p^*(\cdot \mid s,a) - \hat{p}(\cdot \mid s,a), v_\pi \right\rangle \\
&\le \left\|p^*(\cdot \mid s,a) - \hat{p}(\cdot \mid s,a)\right\|_1 \cdot \|v_\pi\|_\infty && (\text{Hölder's inequality}) \\
&\le \sqrt{2\mathsf{KL}(p^*||\hat{p})} \cdot \|v_\pi\|_\infty. && (\text{Pinsker's inequality})
\end{aligned}
$$

Since MLE minimizes the KL divergence, the above display justifies the use of MLE. The transition kernel follows the multinomial distribution and the MLE for this distribution prescribes that if the state-action (s,a) is visited $N(s,a)$ times and the next state visited after the ith visit is S_i then

$$
\hat{p}\left(s'|s,a\right) = \frac{1}{N(s,a)} \sum_{i=1}^{N(s,a)} \mathbb{I}\left\{S_i' = s'\right\}, \quad \forall s' \in \mathbf{S}.
$$

Nonetheless, the task-agnostic model learning is wasteful. There is no need to learn an accurate model for parts of the environment that are irrelevant to the task hand. For example, for a cooking-assistant robot the dynamics of the elevator inside the building is not of interest. Hence, model learning should be tailored toward the task at hand. This argument is at the core of alternative perspectives that will come in the following sections.

2 VAML

Farahmand et al. (2017) directly considers $\ell\left(\hat{p}, p^*; v_\pi\right)$. Instead of the pointwise distance, they consider a weighted loss where the weighting $\nu \in \mathbf{\Delta}(\mathbf{S}\times\mathbf{A})$[1] is probability distribution that puts weights on important state-actions. Second, instead of the $L_1(\nu)$-norm loss, they consider the $L_2(\nu)$-norm. Third, they consider the the distance under the worse value function in their function class $\mathbf{V}$.

$$
\ell_2^2\left(\hat{p}, p^*\right) = \int d\nu(x,a) \sup_{v\in\mathbf{V}} \left| \int \left[p^*\left(ds'|s,a\right) - \hat{p}\left(ds'|s,a\right)\right] v\left(s'\right) \right|^2 \tag{1}
$$

[1] $\mathbf{\Delta}(\mathbf{X})$ represents the set of probability distributions over the set $\mathbf{X}$.

Similar to what we showed in Section 1, we can bound the right-hand side of Equation (1) as follows:

$$\sup_{v\in\mathbf{V}}\left|\int\left[p^*\left(ds'|s,a\right)-\hat{p}\left(ds'|s,a\right)\right]v\left(s'\right)\right|\leq\left\|p^*(\cdot\mid s,a)-\hat{p}(\cdot\mid s,a)\right\|_1\sup_{v\in\mathbf{V}}\|v\|_\infty. \tag{2}$$

But, the right-hand side of Equation (2) is quite loose. Specifically, $\sup_{v\in\mathbf{V}}\|v\|_\infty$ can be very large while the on the left-hand side the value of sup is also controlled by the mismatch between distribution. Hence, VAML aims to optimize the left-hand side of Equation (2) through Equation (1) by first gathering data using some procedure, minimize the loss and handing off the learned transition kernel to the planner. Then, VAML provides an upper bound on Equation (1). In order to state the upper bound, it is useful to revisit some concepts from the supervised learning literature.

Definition 1 (Györfi et al. (2002)[Definition 9.3])**.** Let $\epsilon>0$, let $\mathbf{G}$ be a set of function $\mathbb{R}^d\to\mathbb{R}$, $1\leq p\leq\infty$, and let ν be a probability measure on $\mathbb{R}^d$. For a function $f:\mathbb{R}^d\to\mathbb{R}$ set

$$\|f\|_{L_p(\nu)}\coloneqq\left[\int|f(z)|^p d\nu\right]^{\frac{1}{p}}.$$

(a) Every finite collection of functions $g_1,\ldots,g_N:\mathbb{R}^d\to\mathbb{R}$ with the property that for every $g\in\mathbf{G}$ there is a $j=j(g)\in\{1,\ldots,N\}$ such that

$$\|g-g_j\|_{L_p(\nu)}<\epsilon$$

is called and $\epsilon-$cover of $\mathbf{G}$ with respect to $\|\cdot\|_{L_p(\nu)}$.

(b) Let $\mathcal{N}\left(\epsilon,\mathbf{G},\|\cdot\|_{L_p(\nu)}\right)$ be the size of the smallest ϵ-cover of $\mathbf{G}$ w.r.t $\|\cdot\|_{L_p(\nu)}$. Take $\mathcal{N}\left(\epsilon,\mathbf{G},\|\cdot\|_{L_p(\nu)}\right)=\infty$ if no finite ϵ-cover exits. Then, $\mathcal{N}\left(\epsilon,\mathbf{G},\|\cdot\|_{L_p(\nu)}\right)$ is called an ϵ-covering number of $\mathbf{G}$ w.r.t $\|\cdot\|_{L_p(\nu)}$.

(c) Let $z_1^n=(z_1,\ldots,z_n)$ be n fixed points in $\mathbb{R}^d$, Let ν_n be the corresponding empirical measure, i.e.,

$$\nu_n(A)=\frac{1}{n}\sum_{i=1}^n\mathbb{I}\{z_i\in A\},\quad(A\subseteq\mathbb{R}^d),$$

then

$$\|f\|_{L_p(\nu_n)}\coloneqq\left[\frac{1}{n}\sum_{i=1}^n|f(z_i)|^p\right]^{\frac{1}{p}},$$

and any ϵ-cover of $\mathbf{G}$ w.r.t $\|\cdot\|_{L_p(\nu_n)}$ will be an L_p ϵ-cover of $\mathbf{G}$ on z_1^n an the ϵ-covering number of $\mathbf{G}$ w.r.t $\|\cdot\|_{L_p(\nu_n)}$ will be denoted by $\mathcal{N}_p\left(\epsilon,\mathbf{G},z_1^n\right)$. In other words, $\mathcal{N}_p\left(\epsilon,\mathbf{G},z_1^n\right)$ is the minimal $N\in\mathbb{N}$ such that there exists functions $g_1,\ldots,g_N:\mathbb{R}^d\to\mathbb{R}$ with the property that for every $g\in\mathbf{G}$ there is a $j=j(g)\in\{1,\ldots,N\}$ such that

$$\left[\frac{1}{n}\sum_{i=1}^n|g(z_i)-g_j(z_i)|^p\right]^{\frac{1}{p}}<\epsilon.$$

If $Z_1^n=(Z_1,\ldots,Z_n)$ is a sequence of i.i.d random variables, then $\mathcal{N}_p\left(\epsilon,\mathbf{G},Z_1^n\right)$ is a random variable whose expected value plays an important role. In summary, the covering number of a function class measure the complexity of learning it.

Definition 2 (Vector space, Introduction to Functional Analysis)**.** A vector space $\mathbf{V}$ over a field $\mathbb{K}$ (which we'll take to be either $\mathbb{R}$ or $\mathbb{C}$) is a set of vectors which comes with an addition $+:\mathbf{V}\times\mathbf{V}\to\mathbf{V}$ and scalar multiplication $\cdot:\mathbb{K}\times\mathbf{V}\to\mathbf{V}$, along with some axioms: commutativity, associativity, identity, and inverse of addition, identity of multiplication, and distributivity.

Definition 3 (Norm, Introduction to Functional Analysis)**.** A norm on a **vector space** $\mathbf{V}$ with field $\mathbb{K}$ is a function $\|\cdot\|:\mathbf{V}\to[0,\infty)$ satisfying the following three properties:

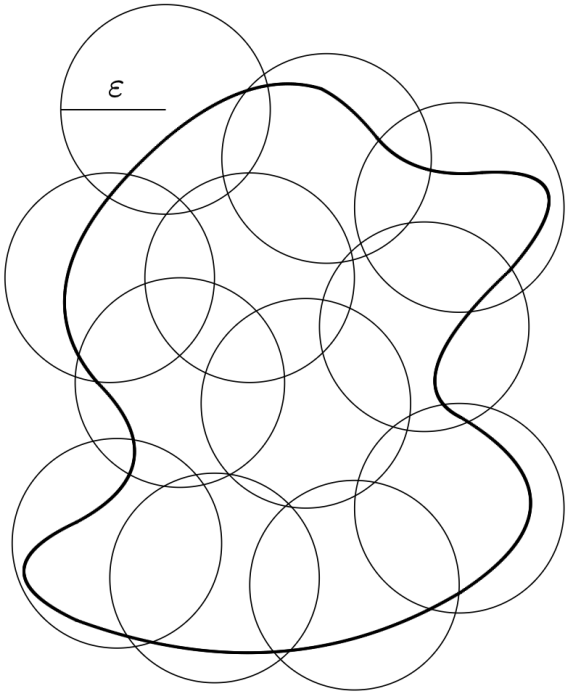

Figure 1: Example of ϵ-cover. We see that $g_1, \ldots, g_N$ are not necessarily in $\mathbf{G}$.

(a) (Definiteness) $\|v\| = 0$ if and only if $v = 0$.

(b) (Homogeneity) $\|\lambda v\| = |\lambda|\|v\|$ for all $v \in \mathbf{V}$ and $\lambda \in \mathbb{K}$.

(c) (Triangle inequality) $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$ for all $v_1, v_2 \in \mathbf{V}$.

A seminorm does not satisfy the first property.

After our detour to revisit some background, let us state the assumptions made by Farahmand et al. (2017) and finally their master theorem. Consider a family of distribution $\mathbf{\Delta}_0$ and a *pseudo-norm* $J : \mathbf{\Delta} \to [0, \infty]$. Let set $\mathbf{\Delta}_B$ used by VAML be $\mathbf{\Delta}_B = \{p \in \mathbf{\Delta} : J(p) \leq B\}$ for some $B > 0$. The reason J is a pseudo-norm and not even a semi-norm is that the the set of probability distributions is not a vector space because it's not closed with respect to the addition and scalar multiplication. Farahmand et al. (2017) tried to give some examples of what J could be that didn't convince me. I asked ChatGPT about possible Js and here are the answers:

(A) L_p norms on densities

(B) Distances between probability measures: TV, Wasserstein, KL, etc.

Assumption 1 (Farahmand et al. (2017, Assumption A1, Capacity of the function space))**.** For $B > 0$, let $\mathbf{\Delta}_B = \{p \in \mathbf{\Delta} : J(p) \leq B\}$. There exists constants $C > 0$ and $0 < \alpha < 1$ such that for any $\epsilon > 0$, and all sequence $z_1, \ldots, z_n \in \mathbf{Z} = \mathbf{S} \times \mathbf{A}$ the following metric entropy condition is satisfied:

$$\log \mathcal{N}\left(\epsilon, \mathbf{\Delta}_B, L_2(p^*_{z_{1:n}})\right) \leq C \left(\frac{B}{\epsilon}\right)^{2\alpha}.$$

Let the value function space be $\mathbf{V} = \left\{v_\theta(s) = \phi^\top(s)\theta : \theta : \mathbb{R}^p, \|\theta\|_\theta \leq B\right\}$, with $\phi : \mathbf{S} \to \mathbb{R}^p$ being the feature map.

Theorem 1 (Farahmand et al. (2017, Theorem 2))**.** *Given a dataset $D_n = \left\{(S_i, A_i, S'_i)_{i=1}^n\right\}$ with independent and identically distributed samples $(S_i, A_i) \sim \nu$, with $S'_i \sim p^*(\cdot \mid S_i, A_i)$, let $\hat{p}$ be the minimizer of the VAML algorithm, i.e., $\hat{p} \leftarrow \arg\min_{P \in \mathbf{\Delta}(\mathbf{S})} \ell_2^2(p, p^*_n)$, with the previously specified choice of*

value function space $\mathbf{V}$. *Let Assumption 1 holds. Furthermore, assume that* $\sup_{s\in\mathbf{S}}\|\phi(s)\|_\infty \leq 1$ *and* $\sup_{s\in\mathbf{S}}\|\phi(s)\|_2 \leq 1$. *Fix* $\delta > 0$. *There exists a constant* $c > 0$ *such that*

$$\textit{Equation (1)} = \mathbb{E}\left[\sup_{v\in\mathbf{V}} |(\hat{p}_Z - p^*_Z)v|^2\right] \leq \underbrace{\inf_{p'\in\mathbf{\Delta}(\mathbf{S})} \mathbb{E}\left[\sup_{v\in\mathbf{V}} |(p'_Z - p^*_Z)v|^2\right]}_{\textit{model or function approximation error}} + \underbrace{c(1+B^\alpha)p\sqrt{\frac{\log(p/\delta)}{n}} + \frac{16\log(4/\delta)}{3n}}_{\textit{estimation error}}. \tag{3}$$

with probability at least $1-\delta$.

Note that since we're competing against values in $\mathbf{V}$ ($\sup_{v\in\mathbf{V}}$), the estimation with more data washes out and there is no residual on that part. However, since $\hat{p} \in \mathbf{\Delta}_B$ but we're competing against $p^* \in \mathbf{\Delta}$, the residual error on the model approximation persists.

2.1 What I didn't understand about VAML

I just didn't go through the proofs. The arguments are clear to me.

3 VE

VAML exclusively studies the linear value functions case. Now, we move on to a more holistic formulation of characterizing useful models. We're going to go through all of propositions and definitions given by Grimm et al. (2020). But, before doing so, we need to review the concept of an *operator* is a functional analysis sense.

You think simply think of an operator as a mapping (a function) from a set of functions to another set of functions. However, there is a more elegant and useful way of defining it. From linear algebra, we know that we should view functions as *vectors* (Strang, 2022). Also, from linear algebra we know that matrix-vector inner product results in a vector. So, by analogy, we are looking for an equivalent concept to matrices in infinite-dimensional spaces. Operators are the analog of matrices in functional analysis [hence their similar notation], so they turn a function into another function. Formally,

Definition 4 (Functional Analysis and Operator Theory, Definition C.1)**.** Let $\mathbf{X}, \mathbf{Y}$ be **vector spaces**, and let $T : \mathbf{X} \to \mathbf{Y}$ be function mapping $\mathbf{X}$ into Y. We either write $T(f)$ or Tf to denote the image under T of an element $f \in \mathbf{X}$. Some interesting properties that involve both the function-like perspective and the matrix-like perspective.

(A) T is injective of $T(f) = T(g)$ implies $f = g$.

(B) The *kernel* or *null space* of if $\mathsf{ker}(T) = \{f \in \mathbf{X} : T(f) = 0\}$

(C) The *rank* of T is the is the vector space dimension of its range. In particular T is finite-rank if its range is finite-dimensional.

Note that as indicated in Definition 4, the domain and the co-domain of an operator *must* be vector spaces. This makes an operator different from other mappings.

We know that the Bellman equation for policy evaluation for a policy π is written as

$$v_\pi(s) \coloneqq \mathbb{E}_\pi\left[r^*(s,a) + \gamma\sum_{s'} p^*\left(s'|s,a\right) v_\pi\left(s'\right)\right], \quad s \in \mathbf{S}. \tag{4}$$

Let us fix π and define r_π and p_π as

$$r_\pi(s) = \sum_a \pi(a \mid s) r^*(s,a), \quad p_\pi(\cdot \mid s) = \sum_a \pi(a \mid s) p^*(\cdot|s,a), \quad \forall s \in \mathbf{S}.$$

Then, we can rewrite Equation (5) as

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi\left(s'|s\right) v_\pi\left(s'\right), \quad \forall s \in \mathbf{S}.$$

Now, we can define the Bellman operator for policy evaluation as

$$\left(T_\pi v\right)(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi\left(s'|s\right) v_\pi\left(s'\right), \quad \forall s \in \mathbf{S},$$

or by viewing r_π as a vector in $\mathbb{R}^{|\mathbf{S}|}$ and p_π a matrix in $\mathbb{R}^{|\mathbf{S}|\times|\mathbf{S}|}$, compactly as $T_\pi : v \mapsto r_\pi + \gamma p_\pi v$.

The Bellman optimality equation is written as

$$v^*(s) \coloneqq \max_a \left\{ r^*(s,a) + \gamma \sum_{s'} p^*\left(s'|s,a\right) v^*\left(s'\right) \right\}, \quad s \in \mathbf{S}. \tag{5}$$

Hence, by a similar procedure for p_π, we can define the Bellman optimality operator T^* as

$$\left(Tv\right)(s) = \max_a \left\{ r^*(s,a) + \gamma \sum_{s'} p^*\left(s'|a,s\right) v^*\left(s'\right) \right\}, \quad \forall s \in \mathbf{S},$$

or compactly as $T : v \mapsto \max_\pi \left\{r_\pi + \gamma p_\pi v\right\}$. Now, we have all the tools needed to dive into VE.

Let $\mathbf{\Pi}$ be set of all stationary Markov policies[2], i.e., $\mathbf{\Pi} = \mathbf{\Delta}(\mathbf{A})^{\mathbf{S}} = \{\pi \mid \pi : \mathbf{S} \to \mathbf{\Delta}(\mathbf{A})\}$, and let $\mathbf{V} = \mathbb{R}^{\mathbf{S}} = \{v \mid v : \mathbf{S} \to \mathbb{R}\}$ be set of all value functions. Given state and action spaces, model approximation in MBRL consists an approximation of the expected immediate reward r^* and an approximation of the transition kernel p^*. So, we if represent a model by $m = (r, p)$, where r and p are some arbitrary approximation, then we can represent the environment itself as the true model by $m^* = (r^*, p^*)$. Now, we can state the value equivalence principle and its associated propositions.

Definition 5 (Grimm et al. (2020, Definition 1))**.** Let $\Pi \subseteq \mathbf{\Pi}$ be a set of policies and let $V \subseteq \mathbf{V}$ be a set of value functions. We say that two models m and $\hat{m}$ are value equivalent with respect to Π and V if and only if

$$T_\pi v = \hat{T}_\pi v, \quad \forall \pi \in \Pi, \forall v \in V.$$

> "Two models are value equivalent with respect to Π and V if the effect of the Bellman operator induced by any policy $\pi \in \Pi$ on any function $v \in V$ is the same for both models. Thus, if we are only interested in Π and V, value-equivalent models are functionally identical (Grimm et al., 2020)."

Definition 6 (Grimm et al. (2020)[Definition 2])**.** Let Π and V be defined as in Definition 5. Let M be a set of models. Given a model m, $M(\Pi, V; m)$ the set of value-equivalent models to m with respect to Π and V that are in M, is a subset of M, i.e., $M(\Pi, V; m) \subseteq M$.

Let $\mathbf{M}^*$ be a set of models containing at least one model m^* [I'd call it the realizable setting]. Given a set of models $M \in \mathbf{M}^*$ [that doesn't necessarily contain m^*], often one is interested in models $m \in M$ that are value equivalent to m^*. We simplify the notation by defining $M^*(\Pi, V) = M(\Pi, V; m^*)$.

> "The set $M^*(\Pi, V)$ contains all the models in M that are value equivalent to the true model m^* with respect to Π and V. Since any two models $m_1, m_2 \in M^*(\Pi, V)$ are equally suitable for value-based planning using Π and V, we are free to use other criteria to choose between them. For example, if m_1 is much simpler to represent or learn than m_2, it can be preferred without compromises (Grimm et al., 2020)."

Property 1 (Grimm et al. (2020)[Property 1])**.** *Given $M_1 \subseteq M_2$, we have that $M_1^*(\Pi, V) \subseteq M_2^*(\Pi, V)$.*

Property 1 is an elementary set topology argument. It makes sense.

[2]Throughout this document, we'll only focus on stationary Markov policies, and for consciousness, we refer to them simply as policies.

Property 2 (Grimm et al. (2020)[Property 2]). *$M^*(\mathbf{\Pi}, \mathbf{V})$ either contains m^* or is the empty set.*

Property 2 says that only the environment itself is value equivalent with respect to all policies and values. In other words, any other model that is value equivalent with respect to all policies and values is simply equivalent to the environment itself. Now, it the set of models that has been chosen M doesn't contain the model of the environment, then the set of models equivalent to the environment is empty which makes sense.

Property 3 (Grimm et al. (2020)[Property 3]). *Given that $\Pi_1 \subseteq \Pi_2$ and $V_1 \subseteq V_2$, we have that $M^*(\Pi_2, V_2) \subseteq M^*(\Pi_1, V_1)$.*

Intuitively Property 3 makes sense. The bigger you make the set of policies and values you want to be equivalent to the environment, the more accurate your models should be which eventually collapses into only being the environment itself capable of achieving.

Property 4 (Grimm et al. (2020)[Property 4]). *If $m^* \in M$, then $m^* \in M^*(\Pi, V)$ for all Π and V.*

Property 4 is immediate from the definition of $M^*(\Pi, V)$.

3.1 Controlling the set of equivalent models' size

How much does $M^*(\Pi, V)$ decrease in size when we, say, add one function to V? In this section we address this and similar questions. Grimm et al. (2020) introduces the concept of p-span which initially I found completely unnecessary given necessary definitions have already been given in functional analysis, but since Π is not a vector space, it seems the approach of Grimm et al. (2020) is inevitable. We review the definition of span that suffices. Also, Grimm et al. (2020) mentions *discrete* sets not countable or finite. So, we must revisit what discrete means and how it is different than countable.

Definition 7 (Lax (2014)[Theorem 2]). Given a vector space $\mathbf{V}$ over a field $\mathbb{K}$, the span of set $V \subseteq \mathbf{V}$ is the set of all finite linear combinations of elements V. Formally,

$$\mathsf{span}(V) = \{a_1 v_1 + \cdots + a_n v_n \mid v_1, \ldots, v_n \in V, a_1, \ldots, a_n \in \mathbb{K}, \text{ for any } n \in \mathbb{N}\}.$$

In order to understand what *discrete* mean, we need to understand what: a metric space, an open set, a topology mean in order. I'll use this post of Terrance Tao to understand these concepts.

Definition 8 (Metric spaces). A metric space is a set $\mathbf{X}$ together with a distance function $d : \mathbf{X} \times \mathbf{X} \to [0, \infty)$, $(\mathbf{X} = (\mathbf{X}, d))$ which obeys the following properties:

(A) (Non-degeneracy) For any $x_1, x_2 \in \mathbf{X}$, we have $d(x_1, x_2) \geq 0$, with equality if and only if $x_1 = x_2$.

(B) (Symmetry) For any $x_1, x_2 \in \mathbf{X}$, we have $d(x_1, x_2) = d(x_2, x_1)$.

(C) (Triangle inequality) For any $x_1, x_2, x_3 \in \mathbf{X}$, we have $d(x_1, x_3) \leq d(x_1, x_2) + d(x_2, x_3)$.

Definition 9 (An open set). Let (X, d) be a metric space. Given any $x \in \mathbf{X}$ and $r > 0$, define the open ball $B(x, r)$ centered at x with radius r to be the set of all $y \in \mathbf{X}$ such that $d(x, y) < r$. Given a set $\mathbf{E}$, we say that x is an interior point of $\mathbf{E}$ if there is some open ball centered at x which is contained in $\mathbf{E}$. The set of all interior points is called the interior $\mathbf{E}$. A set is open if every point is an interior point.

Definition 9 is like our usual open interval in one dimension.

Definition 10 (A topological space). A topological is a set $\mathbf{X}$, together with a collection $\mathscr{F}$ of $\mathbf{X}$'s subsets, known as *open sets*, which follow the following *axioms*:

(A) $\varnothing$ and $\mathbf{X}$ are open. $(\varnothing, \mathbf{X} \in \mathscr{F})$

(B) The intersection of any finite number of open sets is open.

(C) The union of any arbitrary number of open sets is open.

The collection $\mathscr{F}$ is called a topology on $\mathbf{X}$.

Note that the definition of open sets in Definition 10 is different than Definition 9 and is by construction. So, the difference between countable and discrete is quite substantial. Countable focuses on assigning a natural number to each element, but discrete focuses of having open sets.

Example. The finest (or strongest) topology on any set $\mathbf{X}$ is the discrete topology $2^{\mathbf{X}} = \{E : E \subseteq \mathbf{X}\}$, in which every set is open; this is the topology generated by the discrete metric[3]. The coarsest (or weakest) topology is the trivial topology $\{\varnothing, \mathbf{X}\}$, in which only the empty set and the full set are open.

Proposition 1 (Grimm et al. (2020, Proposition 1))**.** *For discrete [finite is correct]* Π *and* V*, we have that* $M^*(\Pi, V) = M^*(p\text{-}\mathsf{span}(\Pi) \cap \mathbf{\Pi}, \mathsf{span}(V))$*.*

From our aforementioned definitions, it is evident that by *discrete*, Grimm et al. (2020) meant *finite*. Because Π for example cannot be a discrete topology because $\{\pi_1\} \cup \{\pi_2\} = \{\pi_1, \pi_2\} \notin \Pi$. Actually, they in fact meant Π and V are finite and linearly independent, however I see no point on why limiting to this case, why not using the definition of span in general even for infinite-dimensional vector spaces. Proposition 1's proof in the original work is correct.

> "Proposition 1 provides one possible answer to the question posed at the beginning of this section. the contraction of $M^*(\Pi, V)$ resulting from the addition of one policy to Π or one function to V. For instance, if a function v can be obtained as a linear combination of the functions in V, adding it to this set will have no effect on the space of equivalent models $M^*(\Pi, V)$ (Grimm et al., 2020)."

Let $\mathbf{P}$ be the set of all transition kernels, $P \subset \mathbf{P}$ be a set of transitions, $P^*(\Pi, V)$ the set of transition kernels that are that value equivalent to p^*.

Definition 11 (Grimm et al. (2020))**.** The dimension of a set $\mathbf{X}$ is the Hamel dimension of a vector-space that encloses some translated version of $\mathbf{X}$.

$$\mathsf{dim}[\mathbf{X}] = \min_{\mathbf{W}, \vec{c} \in W(\mathbf{X})} \mathsf{H\text{-}dim}[\mathbf{W}],$$

where $W(\mathbf{X}) = \{(\mathbf{W}, \vec{c}) : \mathbf{X} + \vec{c} \subseteq \mathbf{W}\}$, $\mathbf{W}$ is a vector space, and $\vec{c}$ is an offset.

I think Definition 11 is just trying to say that $\mathbf{X}$ can become subspace but making sure it contains the zero vector neutralizing by $\vec{c}$.

Remark 1 (Grimm et al. (2020))*.* $\mathsf{dim}[\mathbf{P}] = (|\mathbf{S}| - 1)|\mathbf{S}||\mathbf{A}|$.

Proof. By ChatGPT: For each (s, a) the transition kernel defines a probability simplex over over $\mathbf{S}$. Since the probabilities have to sum up to one there are $|\mathbf{S}| - 1$ free parameters. Therefore, the total free parameters is $|\mathbf{S}||\mathbf{A}|(|\mathbf{S}| - 1)$. □

Proposition 2 (Grimm et al. (2020, Proposition 2))**.** *Let* Π *be set of* m *linearly independent policies* $\pi_i \in \mathbb{R}^{|\mathbf{S}||\mathbf{A}|}$ *and let* V *be the set of* k *linearly independent vectors* $v_i \in \mathbb{R}^{|\mathbf{S}|}$*, Then,*

$$\mathsf{dim}[P^*(\Pi, V)] \leq |\mathbf{S}|(|\mathbf{S}||\mathbf{A}| - mk).$$

To prove Proposition 2. We need four lemmas. I skip the first lemma as I understood what it was, though the original work had made it really convoluted. They could used an argument involving the Kronecker product instead.

Lemma 1 (Grimm et al. (2020, Lemma 2))**.** *For any vector* $\vec{c}$ *and any set* $\mathbf{Y} + \vec{c} = \{y + \vec{c} : y \in \mathbf{Y}\}$*, it follows that* $\mathsf{dim}[\mathbf{Y} + \vec{c}] = \mathsf{dim}[\mathbf{Y}]$.

The original proof is absolutely wrong. Let's see if we can prove it ourselves.

Proof.

$$\mathsf{dim}[\mathbf{Y} + \vec{c}] = \min_{\mathbf{W}, \vec{b} \in W(\mathbf{Y} + \vec{c})} \mathsf{H\text{-}dim}[\mathbf{W}].$$

Also,

$$W(\mathbf{Y} + \vec{c}) = \left\{ \left(\mathbf{W}, \vec{b}\right) : \mathbf{Y} + \underbrace{\vec{c} + \vec{b}}_{\vec{d}} \in \mathbf{W} \right\} = \left\{ \left(\mathbf{W}, \vec{b}\right) : \mathbf{Y} + \vec{d} \in \mathbf{W} \right\} = W(\mathbf{Y}).$$

[3] Discrete metric $d : \mathbf{X} \times \mathbf{X} \to [0, \infty)$, defined by setting $d(x, y) = 0$ when $x = y$ and $d(x, y) = 1$ otherwise

Hence,

$$\mathsf{dim}[\mathbf{Y}+\vec{c}] = \min_{\mathbf{W},\vec{b}\in W(\mathbf{Y}+\vec{c})} \mathsf{H\text{-}dim}[\mathbf{W}] = \min_{\mathbf{W},\vec{b}\in W(\mathbf{Y})} \mathsf{H\text{-}dim}[\mathbf{W}] = \mathsf{dim}[\mathbf{Y}].$$

□

Lemma 2 (Grimm et al. (2020, Lemma 3)). *If* $\mathbf{Y}$ *is a vector-space, then* $\mathsf{H\text{-}dim}[\mathbf{Y}] = \mathsf{dim}[\mathbf{Y}]$.

The original proof is correct but unnecessarily complicated. By definition,

Proof.

$$\mathsf{dim}[\mathbf{Y}] = \min_{\mathbf{W},\vec{c}\in W(\mathbf{Y})} \mathsf{H\text{-}dim}[\mathbf{W}].$$

Since $\mathbf{Y}$ is a vector space and $\mathbf{W}$ is also a vector space, then $\vec{c}$ must be zero (to include the the zero vector). Since $\mathbf{Y}$ is a vector space, it must be that $\mathbf{Y} = \mathbf{W}$ and $\mathsf{H\text{-}dim}[\mathbf{W}] = \mathsf{H\text{-}dim}[\mathbf{Y}]$. □

Lemma 3 (Grimm et al. (2020, Lemma 4)). *If* $\mathbf{X} \subseteq \mathbf{Y}$, *then* $\mathsf{dim}[\mathbf{X}] \leq \mathsf{dim}[\mathbf{Y}]$.

The original proof is sound. The Proposition 2's original proof is sound except of $\min\{|\mathbf{S}||\mathbf{A}|, |\mathbf{S}|m\} = |\mathbf{S}|m$, that needs a justification that authors didn't give.

Proposition 3 (Grimm et al. (2020, Proposition 3)). *Let* $\hat{P}$ *be set of approximation models. The maximum-likelihood estimate of* p^* *in* $\hat{P}$ *might not belong to* $\hat{P}^*(\Pi, V) \neq \varnothing$.

Proposition 3 is saying that MLE gives an estimate that might not be useful for planning with respect to policies and values we care.

Definition 12 (Likelihood function). Let $X_1, \ldots, X_n$ have a joint density function $f(X_1, \ldots, X_n \mid \theta)$. Given $X_1 = x_1, \ldots, X_n = x_n$ is observed, the likelihood function is defined by

$$L(\theta) = L(\theta \mid x_1, \ldots, x_n) = f(x_1, \ldots, x_n \mid \theta)\Big|_{\text{countable setting}} = \mathbb{P}(x_1, \ldots, x_n \mid \theta)\Big|_{\text{i.i.d}} = \Pi_{i=1}^n \mathbb{P}(x_i \mid \theta)$$

Proof. I couldn't understand how the authors had computed the log-likelihood below. So, I asked ChatGPT and use its answer to make the (left out) calculation clearer.

Suppose we are trying to estimate a transition matrix $\Theta \in \mathbb{R}^{n\times n}$ and choose to use one parameter $\theta_i \in R$ per row. Specifically, we parametrize the distribution on the i-th row as

$$\Theta_{ii} = \theta_i, \quad \theta_{ij} = \frac{1-\theta_i}{n-1}, \quad \text{for } j \neq j, \text{ and } \theta_i \in [0,1]: \quad \Theta = \begin{pmatrix} \theta_1 & \frac{1-\theta_1}{n-1} & \ldots & \frac{1-\theta_1}{n-1} \\ \frac{1-\theta_2}{n-1} & \theta_2 & \ldots & \frac{1-\theta_2}{n-1} \\ \vdots & \vdots & \ldots & \vdots \\ \frac{1-\theta_n}{n-1} & \frac{1-\theta_n}{n-1} & \ldots & \theta_n \end{pmatrix}$$

Now, we compute the *expected log*-likelihood function of $\boldsymbol{\theta} \in \mathbb{R}^n$. Let N_{ij} denote the number of times that transition happened from s_i to s_j and p_{ij} be the true transition probability.

$$L(\boldsymbol{\theta}) = \Pi_{i=1}^n \Pi_{j=1}^n \mathbb{P}(s_i \to s_j \mid \theta_{ij})^{N_{ij}}, \quad \text{hence} \quad \log L(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^n N_{ij} \log \mathbb{P}(s_i \to s_j \mid \theta_{ij}).$$

Since we are looking for the expected log-likelihood and $\frac{N_{ij}}{N_i} \xrightarrow{N_i\to\infty} p^*_{ij}$, we can replace the empirical frequencies with the true probability.

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^n \sum_{j=1}^n p_{ij} \log \theta_{ij}$$

The maximum log-likelihood estimation $\frac{\partial \log L(\boldsymbol{\theta})}{\theta_i} = 0$ leads to $\theta_i = p^*_{ii}$. This means that the solution provided by MLE will not be exact if and only if $p_{ij} \neq p_{ik}$ for all $i \neq j \neq k$. Now, suppose we have $V = \{v\}$ with $v_i = 1$ for some i and $v_j = 0$ for $j \neq i$. In this case it is possible to get an exact value-equivalent solution by making $\theta_i = p^*_{ii}$ and $\theta_j = 1 - (n-1)p^*_{ji}$ for $j \neq i$, regardless what MLE says =, which in case, since $\theta_j \neq p^*_{jj}$ is contrasting it. □

Now we have shown that MLE is not the most appropriate way of finding a value-equivalent model, Grimm et al. (2020) proposes the following objective. We ensure the imprecision of Grimm et al. (2020)'s notation is not repeated here.

$$\ell_{\Pi,V}(m^*, \hat{m}) = \sum_{\Pi} \sum_{V} \left\| Tv - \hat{T}v \right\|,$$

where $\|\cdot\|$ is a norm. Since we do not have access to T, we use the empirical version of it. Let $\nu \in \mathbf{\Delta}(\mathbf{S})$, and Let $D_\pi = \left\{(S_i, A_i, R_i, S_i')_{i=1}^n\right\}$ be dataset of n transitions corresponding to policy $\pi \in \Pi$, where $S_i \sim \nu(\cdot)$ and $A_i \sim \pi(\cdot \mid S_t)$. Then, the empirical value-equivalent loss is defined by

$$\ell_{\Pi,V,\nu}(m^*, \hat{m}) = \sum_{\pi\in\Pi} \sum_{v\in V} \sum_{s_0 \in D_\pi} \left[\left| \frac{\sum_{i=1}^n \mathbb{I}\left[S_i = s_0\right] \left(R_i + \gamma v\left(S_i'\right)\right)}{\sum_{i=1}^n \mathbb{I}\left[S_i = s_0\right]} - \hat{T}v \right|^p \right]^{\frac{1}{p}}.$$

3.2 How to choose the subset of policies Π and values V?

Proposition 4 (Grimm et al. (2020, Proposition 4))**.** *Suppose $v \in V' \Rightarrow T_\pi v \in V'$, for all $\pi \in \Pi$. Let $\mathbf{\Pi} \subseteq$ p-span(Π) and* span$(V) = V'$*. Then, starting form any $v' \in V'$, any $\hat{m} \in M^*(\Pi, V)$ yields the same solution as m^*.*

The original proof is sound.

The rest of the main body of the paper is extremely ambiguous to me. I may write someday why, but I wanna move one. Honestly, I just got bored to decipher their intuitions. I'll get back to these skipped parts later.

3.3 What I didn't understand

(A) Really the word *discrete* in Proposition 1. To me they meant finite and linearly independent ad they have alluded to in the passage after Remark 1. Proposition 1 is really suspicious.

(B) MLE in Proposition 3 was sneaky and approximate. the counter example inside was also wrong and I changed it.

(C) Their empirical value-equivalent loss uses a wrong norm formulation.

(D) Page 6 to 9 and 17 to the end were not clear to me a t all.

4 PVE

> " A fundamental question underlying the VE principle is thus how to select the smallest sets of policies and functions that are sufficient for planning. In this paper we take an important step towards answering this question. We start by generalizing the concept of VE to order-k counterparts defined with respect to k applications of the Bellman operator. Unlike VE, the PVE class may contain multiple models even in the limit when all value functions are used. Crucially, all these models are sufficient for planning, meaning that they will yield an optimal policy despite the fact that they may ignore many aspects of the environment (Grimm et al., 2021)."

PVE wants to show that only value functions are enough to specify the equivalence. Since, every policy is associated with a value function, in contrast to VE that we needed to choose Π and V, now we only need to specify Π and V would naturally be their corresponding values. The main advantage of PVE over VE is that even if all value functions are considered, the class of equivalent models doesn't shrink to a singleton.

It is crucial: Grimm et al. (2021) uses T_π^n notation as the repeated application of T_π such that $\lim_{n\to\infty} T_\pi^n v = v_\pi$.

Definition 13 (Order-k VE class)**.**

$$M_k^*(\Pi, V) = \left\{\hat{m} \in M : \hat{T}_\pi^k v = T_\pi^k v,\ \forall \pi \in \Pi, \forall v \in V\right\}.$$

Grimm et al. (2021) claims that in contrast to Grimm et al. (2020)'s argument that $M_1^*(\mathbf{\Pi}, \mathbf{V})$ only contains the environment, this not true for $M_k^*(\mathbf{\Pi}, \mathbf{V})$ when $k > 1$.

Proposition 5 (Grimm et al. (2021, Proposition 1))**.** *Let V be a set of functions such that if $v \in V$, then $T_\pi v \in V$ for all $\pi \in \Pi$. Then, for k, K in $\mathbb{N}$ such that k divides K ($k|K$, or $K = mk, m \in \mathbb{N}$) , it follows that*

(A) For any $M \subseteq \mathbf{M}$ and any $\Pi \in \mathbf{\Pi}$, we have that $M_k^*(\Pi, V) \subseteq M_K^*(\Pi, V)$.

(B) If Π is non-empty and V contains at least one constant function, then there exits environments such that $\mathbf{M}_k^*(\mathbf{\Pi}, \mathbf{V}) \subset \mathbf{M}_K^*(\mathbf{\Pi}, \mathbf{V})$.

The proof the original work for Proposition 5's part (A) is sound. For part (B), Grimm et al. (2021) should show that all $m \in \mathbf{M}_k^*(\mathbf{\Pi}, \mathbf{V})$ are also in $\mathbf{M}_K^*(\mathbf{\Pi}, \mathbf{V})$ (which is proved by part (A)), but there exists a model $m_0 \in \mathbf{M}_K^*(\mathbf{\Pi}, \mathbf{V})$ such that $m_0 \notin \mathbf{M}_k^*(\mathbf{\Pi}, \mathbf{V})$. There is a **HUGE** subtlety for the proof of part (B) in Grimm et al. (2021). Specifically, they have assumed that the k application of the Bellman operator $\hat{T}_\pi^k$ is the same as one application of the Bellman operator on the k-step return $T_\pi^{(k)}$. Using ChatGPT, now we dive into the relationship between $\hat{T}_\pi^k$ and $\hat{T}_\pi^{(k)}$.

We know that

$$
\begin{aligned}
T_\pi v(s) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma v(S_{t+1}) | S_t = s \right], \text{ and} \\
G_t^{(k)} &= \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{k-1} R_{t+k} + \gamma^k v(S_{t+k}) \middle| S_t = s \right].
\end{aligned}
$$

By induction, we'll show that

$$
T_\pi^k v(s) = \mathbb{E}_\pi \left[\sum_{i=0}^{k-1} \gamma^i R_{t+i+1} + \gamma^k v(S_{t+k}) \middle| S_t = s \right] = \mathbb{E}_\pi \left[G_t^{(k)} \middle| S_t = s \right].
$$

Proof. **Base case**: $k = 1$ is immediate. Induction step: Suppose the the assumption for step k, i.e., $T_\pi^k v(s) = \mathbb{E}_\pi \left[\sum_{i=0}^{k-1} \gamma^i R_{t+i+1} + \gamma^k v(S_{t+k}) \middle| S_t = s \right]$, now we show that it holds for step $k+1$. We have

$$
\begin{aligned}
T_\pi^{k+1} v(s) = T_\pi \left(T_\pi^k v \right)(s) &= \mathbb{E}_\pi \left[R_{t+1} + \gamma T_\pi^k v(S_{t+1}) \middle| S_t = s \right] \\
&= \mathbb{E}_\pi \left[R_{t+1} + \gamma \mathbb{E}_\pi \left[\sum_{i=0}^{k-1} \gamma^i R_{t+1+i+1} + \gamma^k v(S_{t+1+k}) \middle| S_{t+1} \right] \middle| S_t = s \right] \\
&= \mathbb{E}_\pi \left[\sum_{i=1}^{k} \gamma^i R_{t+i+1} + \gamma^{k+1} v(S_{t+1+k}) \middle| S_t = s \right]. \qquad \text{(tower rule)}
\end{aligned}
$$

□

So, as long as the environment and the policy are deterministic (which is the case in Grimm et al. (2021)'s proof of Proposition 5) $T_\pi^k = T_\pi^{(k)}$. With this important consideration, Grimm et al. (2021)'s proof for part (B) is sound.

Definition 14 (Grimm et al. (2021, Definition 1))**.** Given a set of policies $\Pi \subset \mathbf{\Pi}$, let

$$
M_\infty^* = \lim_{k \to \infty} M_k^*(\Pi, \mathbf{V}) = \{ \hat{m} \in M : \hat{v}_\pi = v_\pi, \forall \pi \in \Pi \}.
$$

We say that each $\hat{m} \in M_\infty^*$ is proper value-equivalent to the environment with respect to Π.

Since in the limit of infinite applications of the Bellman operator all value functions converge to the state-value function of a given policy, PVE only needs defining Π in contrast to VE that needed V as well.

Proposition 6 (Grimm et al. (2021, Definition 2))**.** *For any $\Pi \in \mathbf{\Pi}$ and $k \in \mathbb{N}$ it follows that*

$$
M_\infty^*(\Pi) = \bigcap_{\pi \in \Pi} M_k^*(\{\pi\}, \{v_\pi\}).
$$

The original proof of Proposition 6 is sound.

> "We showing how irrelevant aspects of the environment that are eventually captured by order-one VE are always ignored by PVE in Proposition 7 Grimm et al. (2021)."

Proposition 7 (Grimm et al. (2021, Proposition 3))**.** *Let* $\Pi \in \mathbf{\Pi}$*. If the environment state can be factored as* $\mathbf{S} = \mathbf{X} \times \mathbf{Y}$*, where* $|\mathbf{Y}| > 1$ *and* $v_\pi(s) = v_\pi((x, y)) = v_\pi(x)$ *for all* $\pi \in \Pi$*, then* $\mathbf{M}_1^*(\Pi, \mathbf{V}) \subset \mathbf{M}_\infty^*(\Pi)$*.*

The Proposition 7's original proof is okay (with some corrections in the notation and description).

Proposition 8 (Grimm et al. (2021, Proposition 4))**.** *An optimal policy for any* $\hat{m} \in M_\infty^*(\mathbf{\Pi})$ *is an optimal policy in the environment.*

Proposition 8 seems immediate by definition and Grimm et al. (2021) didn't prove it either.

Corollary 1 (Grimm et al. (2021, Corollary 1))**.** *Let* $\mathbf{\Pi}_{\text{det}}$ *be the set of all deterministic policies. An optimal policy for any* $\hat{m} \in M_\infty^*(\mathbf{\Pi}_{\text{det}})$ *is also optimal in the environment.*

The Corollary 1's proof seems immediate given that one optimal policy in an MDP is deterministic (Puterman, 2014). The proof of Corollary 1 by Grimm et al. (2021) is fine, in their proof they used the assumption of focusing on only deterministic policies that was unnecessary compared to my explanation that related it the infamous optimality of deterministic polices in Puterman (2014).

Proposition 9 (Grimm et al. (2021, Proposition 5))**.** *There exits environments and model classes for which* $\mathbf{M}_\infty^*(\mathbf{\Pi}) \subset M_\infty^*(\mathbf{\Pi}_{\text{det}})$*.*

Proposition 9's original proof is okay-ish.

Proposition 10 (Grimm et al. (2021, Proposition 6))**.** *For any* $\pi \in \mathbf{\Pi}, v \in \mathbf{V}$*, and* $k, n \in \mathbb{N}$*, we have that*

$$\|v_\pi - \hat{T}_\pi^k v_\pi\|_\infty \leq (\gamma^n + \gamma^k)\|v_\pi - v\|_\infty + \|T_\pi^n v - \hat{T}_\pi^k v\|_\infty.$$

Proposition 10's original proof is sound except that the authors should have made it clear that the model belongs to $M_k^*(\mathbf{\Pi}, \mathbf{V})$.

4.1 Promises

(A) Unlike VE, the PVE class *may* contain multiple models even in the limit when all value functions are used. *They kept it. Good job!*

(B) We construct a loss function for learning PVE models and argue that popular algorithms such as MuZero can be understood as minimizing an upper bound for this loss (with mild assumptions). *Nope! The authors had confused* T^n *and* $T^(n)$ *in the MuZero part.*

(C) We leverage this connection to propose a modification to MuZero and show that it can lead to improved performance in practice

4.2 What I didn't understand

(A) On page 2 they cite Rich's and Csaba's book which were unnecessary. One of them was enough.

(B) I didn't understand the upper and lower bound on the MuZero loss. The derivation is way too sloppy.

5 Model equivalence for risk-sensitive

5.1 What I didn't understand

6 Future work

(A) These papers never talked about policy search methods. What can we say about usefulness of models for those methods? In another words, if the the action space is not finite, these methods cannot explain the equivalence.

(B) These papers never talked about average reward. What can we say about usefulness of models for those methods?

References

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., Farquhar, G., Silver, D., and Singh, S. (2021). Proper value equivalence. *Advances in neural information processing systems*, 34:7773–7786.

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.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.

Kastner, T., Erdogdu, M. A., and Farahmand, A.-m. (2023). Distributional model equivalence for risk-sensitive reinforcement learning. *Advances in Neural Information Processing Systems*.

Lax, P. D. (2014). *Functional analysis*. John Wiley & Sons.

Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Strang, G. (2022). *Introduction to linear algebra*. SIAM.