Bisimulation Metric for Continuous MDPs

Pascale Gourdeau

April 21, 2017

1 Introduction

1.1 Motivation

Real-world applications of reinforcement learning abound. Usually, the most interesting applications feature very large state spaces and, sometimes, one even wants to model a continuous environment, where the number of states is infinite. Therefore, the need for state aggregation arises naturally.

A first attempt to tackle this challenge could be to discretize the state space. However, this approach poses two problems. On one hand, we may aggregate states that are "physically" close together, but that "act" very differently. For example, if we are trying to simulate an environment for a robot to navigate in, aggregating the edge of a cliff and the cliff itself could be disastrous. On the other hand, as we refine the space, we experience a blow up in the number of states.

We could also aggregate states with similar optimal values. Again, this is problematic, as we may aggregate states that require totally different policies, as showcased by the example in figure 1.

Both these ideas rely on the false assumption that the states we will aggregate have similar *behaviour*. This is where bisimulation comes in: from

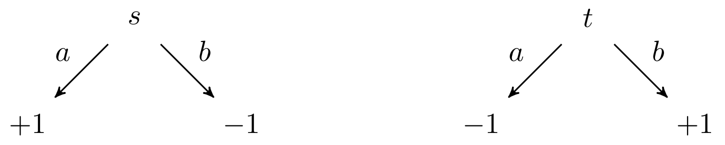

Figure 1: States s and t have the same value but they require opposite policies (the $+1$ and -1 are the rewards after taking an action, and the bottom states are absorbing).

an equivalence relation that tells us if states are behaving *exactly the same*, we build a metric that will allow us to measure behavioural distance between states in an Markov decision process (MDP) and aggregate states that are sufficiently close to each other.

The report is organized in the following way: we first introduce continuous MDPs and their relationship to partially observable MDPs. We then introduce bisimulation as a concept, and define it for MDPs. The following section tackles the construction of the bisimulation metric and shows an example on a simple MDP. An overview of an approximation algorithm is then presented and we finish with highlighting how this metric and optimal value functions are related. Most of this report will be a summary of [1] and [2].

2 Continuous MDPs

Finite MDPs are defined as in class, as the tuple $\mathcal{M} = (S, A, P, r)$. In order to introduce continuous MDPs, or rather a generalization of MDPs that would allow us to have a continuous state space, we must establish which subsets of the states we are interested in studying. This is where measure and probability theory come in: given that we are in state s and have done action a, we want to be able to know probability of reaching a certain subset S' of S (we won't review measure theory here as the reader[1] is familiar with it). S' must be *measurable*, or part of a σ-algebra on S, which is usually taken as $\mathcal{B}(S)$, or the Borel σ-algebra generated by the open subsets of S. If S is finite, we intuitively simply take the power set of S. Otherwise, most subsets of S one can think of are in $\mathcal{B}(S)$.

Definition 1 *A continuous MDP is a tuple* (S, Σ, A, P, r) *where*

- (S, Σ) *is a measurable space,*
- A *is a finite set of actions,*
- $r : S \times A \to \mathbb{R}$ *is a measurable reward function,*
- $P : S \times A \times \Sigma \to [0, 1]$ *is a labelled stochastic transition kernel:*
  - $\forall a \in A$, $\forall s \in S$, $P(s, a, \cdot) : \Sigma \to [0, 1]$ *is a probability measure,*
  - $\forall a \in A$, $\forall X \in \Sigma$, $P(\cdot, a, X) : S \to [0, 1]$ *is a measurable function.*

[1] Hi Doina!

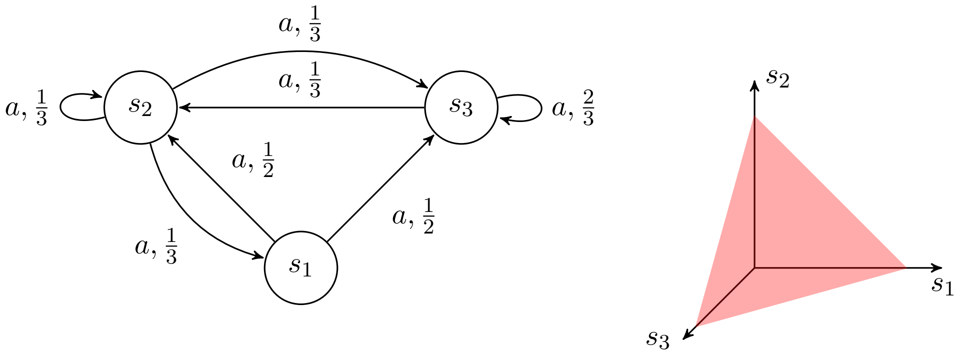

Figure 2: POMDP (left) and its belief-MDP (right). If all the states emit observations o_1, o_2, o_3 with different probabilities, after each action, we move on the simplex in the belief MDP.

Continuous MDPs arise naturally as the belief-MDP of a partially observable MDP (POMDP). A POMDP is a tuple $(S, A, P, r, \Omega, \mathcal{O})$ where

- S is a (finite) set of states and A is a finite set of actions,
- $P : S \times A \times S \to [0, 1]$ is a probabilistic transition map between states,
- $r : S \times A \to \mathbb{R}$ is a reward function,
- Ω is a set of observations,
- $\mathcal{O}$ is a set of conditional observation probabilities.

Indeed, a belief is a distribution on the states of the POMDP that corresponds to where we think we are. It is in fact exactly the simplex on S, as shown in figure 2 for $|S| = 3$.

3 Bisimulation

3.1 General Idea

Bisimulation captures the idea of *behavioural equivalence* between systems. This equivalence relation, first defined by Larsen and Skou [3], allows us to partition a state space into equivalence classes. Intuitively, two states are bisimilar if they act the same (for example, they have the same reward distribution given the action taken) and if they transition with the same probabilities to the same equivalence classes.

3.2 Bisimulation for MDPs

In order to define the bisimulation metric, we need to impose certain restrictions on the MDP to be studied (a review of metric spaces is provided in the appendix):

- S is a Polish space with its Borel σ-algebra Σ.
- $\text{img}(r) \subseteq [0, 1]$.
- For each $a \in A$, $r(\cdot, a)$ is continuous on S.
- For each $a \in A$, P_s^a is continuous as a function of s.

Definition 2 *Let (S, Σ, A, P, r) be an MDP satisfying the above assumptions. An equivalence relation R on S is a* bisimulation relation *if and only if it satisfies*

$$sRs' \iff \begin{array}{l} \textit{for every } a \in A,\ r_s^a = r_{s'}^a \textit{ and} \\ \textit{for every } X \in \Sigma(R),\ P_s^a(X) = P_{s'}^a(X). \end{array}$$

Here, $\Sigma(R)$ refers to Σ-measurable subsets of S that are R-closed. This means that if $X \in \Sigma(R)$, then $R(X) = \{s' \in S | \exists s \in X \text{ s.t. } sRs'\} \subseteq X$. Since R is an equivalence relation, $\Sigma(R)$ is the set of Σ-measurable subsets of S that can be partitioned into R-equivalence classes.

Now, we say that two states are *bisimilar* ($s \sim s'$) if and only if there exists a bisimulation relation R such that sRs'. In this sense, $\sim$ is the largest bisimulation relation. Finally, what is a bisimulation metric? Simply, a pseudometric $\rho : S \times S \to [0, +\infty)$ on the states of an MDP is a *bisimulation metric* if it satisfies

$$\rho(s, s') = 0 \iff s \sim s' \ ,$$

i.e. two states have a distance of zero if and only if they are bisimilar.

4 Bisimulation Metric Construction

4.1 Kantorovich Metric

There are multiple ways to construct a bisimulation metric. A common way to do so in the literature is to define a map on pseudometrics to pseudometrics, take its fixed point and prove that it is a bisimulation metric. Of course, this is easier said than done. The map must capture the notion that, as it is applied repeatedly, we get closer to a metric that expresses bisimilarity. In this paper, this is done by using the Kantorovich metric.

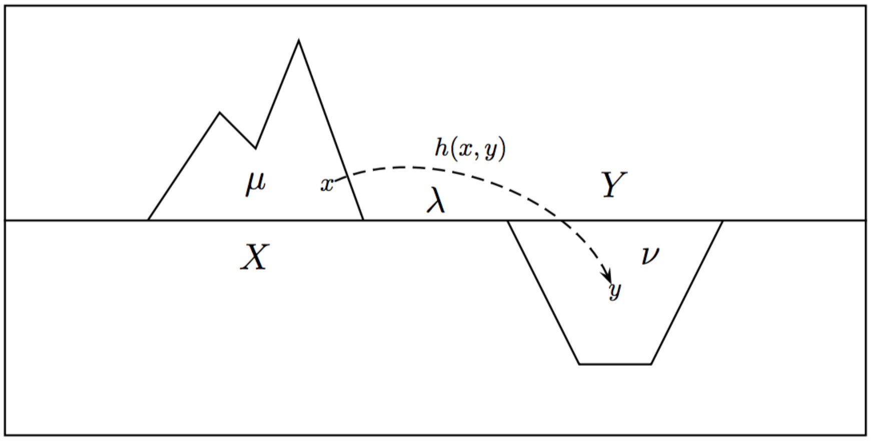

Figure 3: The transportation problem.

Definition 3 *Let (S, d) be a Polish metric space, h a bounded pseudo-metric on S that is lower semi-continuous on $S \times S$ and $Lip(h)$ the set of all bounded functions $f : S \to \mathbb{R}$ that are measurable w.r.t. $\mathcal{B}(S)$ and satisfy the Lipschitz condition $f(x) - f(y) \leq h(x, y)$ for every $x, y \in S$. Given two probability measures P and Q, the* Kantorovich distance *$T_K(h)$ is defined by*

$$T_K(h)(P, Q) = \sup_{f \in Lip(h)} (P(f) - Q(f)) = \sup_{f \in Lip(h)} \left(\int f dP - \int f dQ \right)$$

The Kantorovich metric has a dual definition, which can be modelled as the transportation problem. We start with a measurable space (S, Σ), and make two copies of it, (X, Σ_X), representing a pile of sand, and (Y, Σ_Y), representing a hole in the ground. Associated to each $x \in X$ and $y \in Y$, we have a function $h : X \times Y \to \mathbb{R}$ that represents the cost of transferring a unit of mass from x to y. There are also two measures P and Q, on Σ_X and Σ_Y, respectively, where $P(A)$ is how much sand occupies $A \in \Sigma_X$ and $Q(B)$ is how much sand can be put in $B \in \Sigma_Y$. We want to determine a plan for transferring all the mass from X to Y while minimizing the cost. The collection of all such plans is denoted $\Lambda(P, Q)$, as they represent a measure on the product space, and must have marginals P and Q.

Theorem 1 (Kantorovich-Rubinstein Duality Theorem)

$$T_K(h)(P, Q) = \sup_{f \in Lip(h)} (P(f) - Q(f)) = \inf_{\lambda \in \Lambda(P,Q)} h(\lambda)$$

The Kantorovich metric is used to construct the bisimulation metric, as it has the following important property.

Lemma 1 *Let $\mathfrak{lsc}_{\mathfrak{m}}$ be the set of bounded pseudometrics on S which are lower semi-continuous on $S \times S$, $h \in \mathfrak{lsc}_{\mathfrak{m}}$ and $Rel(h)$ be the kernel of h. Then*

$$T_K(h)(P,Q) = 0 \iff P(X) = Q(X)\ \forall X \in \Sigma(Rel(h)) \ \ .$$

In the context of a metric on MDPs, this means that two probability measures on states have Kantorovich distance 0 with respect to the bisimulation metric ρ^* if and only if they agree on all measurable sets that are closed with respect to bisimilarity. This is another of saying that it matches distributions, which is exactly what we want.

4.2 Fixed Point Construction

Given a discount factor $0 < c < 1$, the Kantorovich metric is used explicitly in the following map $F_c : \mathfrak{lsc}_{\mathfrak{m}} \to \mathfrak{lsc}_{\mathfrak{m}}$:

$$F_c(h)(s,s') = \max_{a \in A} \left[(1-c)|r_s^a - r_{s'}^a| + cT_K(h)(P_s^a, P_{s'}^a)\right] \ \ .$$

Intuitively, this means that, given a pseudometric $h \in \mathfrak{lsc}_{\mathfrak{m}}$, the new pseudo metric assigns a distance between $s, s' \in S$ with the action that will maximize over actions the sum of

- the discounted difference between rewards starting in s and s' and after taking action a and
- the discounted Kantorovich metric between the probability measures of taking action a on s versus s' with respect to the pseudometric h.

The greater the discount factor, the more importance we place on transitioning to similar states (versus being in states that give similar rewards).

By showing that F_c is a contraction on $\mathfrak{lsc}_{\mathfrak{m}}$ (endowed with the uniform norm), and that $\mathfrak{lsc}_{\mathfrak{m}}$ is a complete metric space, we can use Banach's fixed point theorem and get that F_c has a unique fixed point ρ^*. Using the monotonicity of F_c, it is also possible to show that ρ^* is indeed a bisimulation metric and that it scales with rewards (so the metric is defined for rewards in $[0,k]$ for $k \in \mathbb{R}^+$).

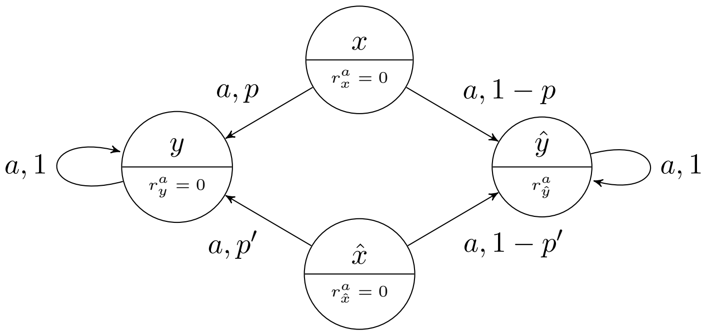

Figure 4: An MDP on which we calculate $\rho^*(s, s')$ for all $s, s' \in S$.

4.3 Example

We will now look at an example of an explicit computation of the bisimulation metric. This example is shown in figure 4 and is also in [1], although the distances are only stated. We include the full derivation here. First note that:

1. There is only one action,
2. $T_K(\rho^*)(\delta_x, \delta_y) = \rho^*(x, y)$, δ_x being the measure where all the mass is at $x \in X$,
3. $F(\rho^*)(s, s') = (\rho^*)(s, s')$ and ρ^* is unique.

We also assume that $r_{\hat{y}}^a > 0$, otherwise all the states are bisimilar.

Let us first start with y and $\hat{y}$:

$$\begin{aligned}
\rho^*(y, \hat{y}) &= \max_{a \in A} \left[(1-c)|r_y^a - r_{\hat{y}}^a| + cT_K(\rho^*)(P_y^a, P_{\hat{y}}^a)\right] \\
&= (1-c)r_{\hat{y}}^a + cT_K(\rho^*)(\delta_y, \delta_{\hat{y}}) \\
&= (1-c)r_{\hat{y}}^a + c\rho^*(y, \hat{y}) \\
\implies \rho^*(y, \hat{y}) &= r_{\hat{y}}^a \ .
\end{aligned}$$

Now, since x and $\hat{x}$ both have zero rewards, and we only care that they transition to y and $\hat{y}$ with different probabilities if $p \neq p'$.

$$\begin{aligned}
\rho^*(x, \hat{x}) &= \max_{a \in A} [(1-c)|r_x^a - r_{\hat{x}}^a| + cT_K(\rho^*)(P_x^a, P_{\hat{x}}^a)] \\
&= cT_K(\rho^*)(P_x^a, P_{\hat{x}}^a) \\
&= c|p - p'|\rho^*(y, \hat{y}) \\
&= c|p - p'|r_{\hat{y}}^a \ .
\end{aligned}$$

The cases of x and y are similar to $\hat{x}$ and y so we will only show one.

$$\begin{aligned}
\rho^*(x,y) &= \max_{a\in A}\left[(1-c)|r^a_x - r^a_y| + cT_K(\rho^*)(P^a_x, P^a_y)\right] \\
&= cT_K(\rho^*)(P^a_x, P^a_y) \\
&= c(1-p)\rho^*(y,\hat{y}) \\
&= c(1-p)r^a_{\hat{y}} \ \ .
\end{aligned}$$

Similarly, $\rho^*(x,\hat{y})$ is similar to $\rho^*(\hat{x},\hat{y})$ so we only show the first one.

$$\begin{aligned}
\rho^*(x,\hat{y}) &= \max_{a\in A}\left[(1-c)|r^a_x - r^a_{\hat{y}}| + cT_K(\rho^*)(P^a_x, P^a_{\hat{y}})\right] \\
&= (1-c)r^a_{\hat{y}} + cp(\rho^*)(y,\hat{y}) \\
&= (1-c)r^a_{\hat{y}} + cpr^a_{\hat{y}} \\
&= (1-c+cp)r^a_{\hat{y}} \ \ .
\end{aligned}$$

5 Approximation Algorithm

The full version of Banach's fixed point theorem give us an algorithm to get arbitrarily close to the unique fixed point and a bound on how good an approximation is with respect to the number of iterations and the initial element on which we applied the contraction. This naturally leads to the following approximation algorithm for a *finite* MDP:
Given tolerance δ,

- Initialize $\rho(s,s') = 0$ for all state pairs (s,s').
- For $\lceil \frac{\ln\delta}{\ln c} \rceil$ iterations:
  - For each tuple (s,s',a): $TK_a(s,s') = Hungarian\_alg(\rho, P^a_s, P^a_{s'})$
- For each state pair (s,s')
  - $\rho(s,s') = \max_a[(1-c)|r^a_s - r^a_{s'}| + cTK_a(s,s')]$

Here, the Hungarian algorithm is a polynomial-time algorithm that solves the minimum-cost assignment problem. For discrete distributions P and Q, there is a one-to-one correspondence between solving this problem and finding the Kantorovich metric between P and Q for cost ρ.

If we have a continuous MDP, we need a finite set of states to approximate the result. We however can't pick any set. Intuitively, we want a set that is representative of the whole MDP. Assume we are provided with $U \subseteq S$, where U is a countable and dense in S, and a metric d on $U \times U$. We construct a subset X of U such that X is an ϵ-net, which means that all the points are at distance ϵ apart and $U \subseteq \bigcup_{x \in X} B(x, \epsilon)$. We start with $X = \{s\}$ for some $s \in U$, then pick $s' \in U \setminus X$ such that maximizes $\min\{d(x, s') : x \in X\}$ repeatedly until we get an ϵ-net. In practice, we use sampling to find such an s' (if U is countably infinite).

Given a number of samples i, $|X| = n$ and $|A| = m$, the worst-case running time is $O\left(\frac{\ln \delta}{\ln c} mn^2 i^3\right)$. This algorithm turns out to be impractical to use, but presents interesting theoretical ideas and intricate proofs.

6 Optimal Value Functions

6.1 An interesting bound

For a continuous MDP, we have the following value function for a given $s \in S$:

$$V^*(s) = \max_a \left(R(s,a) + \gamma \int_S P(s,a,s') V^*(s') \right) ds'$$

Ferns et al. [1] show that, for an MDP with discount factor $\gamma \in (0,1)$ such that $\gamma \leq c$,

$$|V^*(s) - V^*(s')| \leq \frac{1}{1-c} \rho_c^*(s,s') \ .$$

Earlier, when motivating the need for a bisimulation metric, we dismissed the solution of aggregating states with similar optimal values, but it turns out that states that are close (with respect to bisimilarity) are more likely to share optimal values functions and hence policies. Therefore, aggregating states that are close in behaviour implies aggregating states with similar value functions.

6.2 Coupling an MDP with itself

In fact, bisimulation metric are even more closely tied with optimal value functions. Ferns and Precup show in [2] that for a given MDP $\mathcal{M}$, there exists a coupling K^* of M with itself, such that the bisimulation metric for $\mathcal{M}$ and the optimal value function of the coupling are the same.

Formally, a coupling is defined on the product space of two measurable spaces: let $(X, \mathcal{B}(X))$ and $(Y, \mathcal{B}(Y))$ be Borel spaces, μ a probability measure on X, ν a probability measure on Y and λ a probability measure on the product space $(X \times Y, \mathcal{B}(X) \otimes \mathcal{B}(Y))$. We say that λ is a coupling of μ and ν if and only if its marginals on X and Y are μ and ν, respectively. The set of all couplings of μ and ν is denoted $\Lambda(\mu, \nu)$.

To extend this idea to MDPs, we let $K = (K_i)_{i \in I}$, $L = (L_i)_{i \in I}$ and $M = (M_i)_{i \in I}$ (for some index set I) be labelled Markov kernels on X, Y and $X \times Y$, respectively. We say that M is a coupling of K and L if and only if for all $i \in I, x \in X$ and $y \in Y$, $M_i(x, y)$ is a coupling of $K_i(x)$ and $L_i(y)$, as in the previous paragraph.

Now, to couple MDPs $\mathcal{M}$ and $\mathcal{M}'$ together, we use $I = A$, the set of actions, and take a coupling $K \in \Lambda(P, P')$, where P is the set of probability transitions indexed by actions $a \in A$. This translates into a new MDP where a state is a pair (s, s') such that $s \in S$ and $s' \in S'$. This MDP also has an optimal value function $V^*(s, s')_c(K)$ (c is the discount factor used for the bisimulation metric). We define the new reward function $\mathfrak{r} : A \times S \times S'$ as $(1 - c)|r_s^a - r_{s'}^a|$.

We are now ready to formally couple $\mathcal{M}$ with itself: $\mathcal{M}(K) = (S \times S, \mathcal{B}_{S \times S}, A, (K_a)_{a \in A}, \mathfrak{r})$ for $K \in \Lambda(P, P)$. It turns out, as shown in [2], that there exists an optimal $K* \in \Lambda(P, P)$ such that $\rho_c^* = V_c^*(K^*)$.

7 Conclusion

To conclude, this report outlined the construction of a bisimulation metric between states in MDPs that can be extended to the continuous case, and that captures the notion of behavioural distance between states. While the algorithm outlined is impractical, the theory behind this bisimulation metric is quite fascinating and relies on very complex proof techniques.

References

[1] Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. *SIAM J. Comput.*, 40(6):1662–1714, 2011.

[2] Norm Ferns and Doina Precup. Bisimulation metrics are optimal value functions. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*, 2014. Article 67.

[3] K. G. Larsen and A. Skou. Bisimulation through probablistic testing. *Information and Computation*, 94:1–28, 1991.

8 Appendix

8.1 Metric Spaces Refresh

Definition 4 *A* metric *on a set X is a map $d : X \times X \to [0, \infty)$ such that for all $x, y, z \in X$:*

1. $x = y \iff d(x, y) = 0$
2. $d(x, y) = d(y, x)$
3. $d(x, y) \leq d(x, z) + d(z, y)$

We say that the tuple (X,d) where X is a set with a metric $d : X \times X \to [0, \infty)$ is a metric space*.*

A metric space (X, d) is said to be *separable* if it has some countable dense subset, *complete* if every Cauchy sequence converges and *Polish* if it is both separable and complete.