# Temporally Extended Metrics for Markov Decision Processes

**Philip Amortila**
McGill University

**Marc G. Bellemare**
Google Brain*

**Prakash Panangaden**
McGill University

**Doina Precup**
McGill University / DeepMind



**Abstract**

Developing safe and efficient methods for state abstraction in reinforcement learning systems is an open research problem. We propose to address it by leveraging ideas from formal verification, namely, bisimulation. Specifically, we generalize the notion of bisimulation by considering arbitrary comparisons between states instead of strict reward matching. We further develop a notion of *temporally extended metrics*, which extend a base metric between states of an environment so as to reflect not just the current difference but the extent to which the distance is preserved through the course of transitions. We show that this property is not satisfied by bisimulation metrics, which were previously used to compare states with respect to their longterm rewards. A temporal extension can be defined for any base metric of interest, thus making the construction very flexible. The kernel of the temporally extended metrics corresponds precisely to exact bisimulation (thus these metrics form a larger class of *bisimulation metrics*). We provide bounds relating bisimulation and temporally extended metrics and also examine the couplings of state distributions which are induced.

## 1 Introduction

Building safe AI systems is crucial for a wide range of applications. This is especially difficult when the system relies on reinforcement learning, in which an agent learns from its own experience how to behave in the world. Because in most applications of reinforcement learning, the environment is very large or perhaps continuous, an agent is required to abstract over its state space. However, this operation can result in putting together states that actually have very different values, which can lead to suboptimal or even dangerous behaviour. This is especially true if an agent relies only on immediate observations to determine the similarity between states, rather than long-term predictions or behaviour. For example, a camera mounted on a car may show two very similar images, but one may lead to the car going up the curb and the other may be safe driving.

In reinforcement learning, we can leverage some structure in the problem formulation to attempt to tackle this problem. Specifically, the environment of an agent is a Markov Decision Process (MDP) in which state similarity has been studied for a couple of decades. One long-studied notion used to capture behavioral similarity of states in a Markov Decision Process (MDP) is called bisimulation. Bisimulation has originated in the fields of concurrency and formal verification for provably verifying the correctness and safety of processes and systems (Milner 1980). An extension to MDPs was proposed by (Givan, Dean, and Greig 2003). Bisimulation is a canonical tool for analyzing the behaviour of transition systems and clustering equivalent states in overly large systems. In the context of MDPs, bisimulation requires an exact match of both rewards and transition distributions. As this criterion is too brittle, a metric approach was developed by (Ferns, Panangaden, and Precup 2004), which allows one to measure "how bisimilar" two states are. These *bisimulation metrics* assign distance of zero to states if and only if they are bisimilar. Bisimulation metrics can be used for state abstraction (by aggregating states that are $\varepsilon$ away), and doing so provides one with formal (rather than statistical) safety guarantees on the difference between the true optimal value function and the approximated one. The bisimulation metric can be computed iteratively, but each step requires one to solve a linear program involving the transition distributions for every pair of states, which is not only computationally expensive but also requires a full model of the MDP. In this work, we investigate alternative metrics for behavioural equivalence, with the goal of maintaining the useful theoretical guarantees of bisimulation metrics while reducing the computational burden and the need for knowledge of the model.

Our contributions are two-fold: firstly, we propose a coupling-based generalization of bisimulation which allows for greater flexibility in the comparisons between states (instead of strict reward matching), and consequently in the properties being checked. Secondly, we consider the class of *quantitative bisimulations* and show how this defines a notion of *temporally extended (TE) metrics*. Intuitively, these metrics compute the minimum value of a chosen base metric for which the states $s$ and $s'$ can remain in that range throughout their dynamics. The TE metrics assign distance 0 to states if and only if they are bisimilar, much like the bisimulation metrics previously defined. However, both the construction and the resulting metric are quite different.

*CIFAR Fellow

The rest of the paper is organized as follows. In the next section we provide some necessary background. In Section 3 we characterize bisimulation via couplings and define the extension of this characterization to arbitrary relations. In Section 4 we define the temporally extended metrics. Section 5 compares the two metrics by providing some bounds relating them and analyzing the couplings induced by the two metrics. Lastly, we wrap up with a discussion on the benefits and disadvantages of these metrics, highlighting directions for future work.

## 2 Background

Let $\mathcal{D}(\mathcal{S})$ denote the set of probability distributions on $\mathcal{S}$. A *Markov decision process* (MDP) is a 5-tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \{p_a : \mathcal{S} \to \mathcal{D}(\mathcal{S})\}_{a\in\mathcal{A}}, r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{\geq 0}, \gamma \rangle$ where we emphasize that $p_a(s) \in \mathcal{D}(\mathcal{S})$ is to be read as a probability distribution over the states (one for each action). The transition probability from $s$ to a set of states $X$ is written $p_a(s)(X) = \sum_{x\in X} p_a(s)(x)$. In many practical applications, the state space of the MDP is simply too large to allow one to compute the value functions exactly without the use of *state abstraction*. Bisimulation is a canonical example of a safe abstraction, in that the optimal policy and optimal value functions will be preserved in the aggregated MDP (Li, Walsh, and Littman 2006). Bisimulation is defined in terms of equivalence classes, we write $S/\mathcal{R}$ for the set of equivalence classes of an equivalence relation $\mathcal{R}$.

**Definition 2.1** (Bisimulation)**.** A *bisimulation relation* $\mathcal{U}$ on $\mathcal{S}$ is an equivalence relation such that $s\mathcal{U}s'$ implies:

1. $\forall a \in \mathcal{A}, r(s,a) = r(s',a)$ and
2. $\forall a \in \mathcal{A}, \forall C \in S/\mathcal{U}, p_a(s)(C) = p_a(s')(C)$.

We say that $s$ and $s'$ are *bisimilar* and write $s \sim s'$ if there is some bisimulation relation $\mathcal{U}$ relating them.

Thus bisimilar states will have equal rewards and thereafter transition with equal probability to more bisimilar states.

Given a relation $\mathcal{R}$ between states, the *lifting* $(\mathcal{R})^{\#}$ of that relation allows one to naturally extend $\mathcal{R}$ to *distributions* on states. Liftings are defined in terms of couplings, we will later see that this notion will allow us to generalize bisimulation.

**Definition 2.2** (Couplings and Liftings)**.** A coupling of two distributions $(\mu, \nu)$ on $\mathcal{S}$ is a joint distribution $\lambda$ on $\mathcal{S} \times \mathcal{S}$ such that the marginals are $\mu$ and $\nu$:

$$\lambda(s, \mathcal{S}) = \mu(s) \quad \& \quad \nu(s') = \lambda(\mathcal{S}, s').$$

Moreover, a coupling of distributions $(\mu, \nu)$ is a *lifting* of $\mathcal{R}$ if:

$$\lambda(s,s') > 0 \Rightarrow s\mathcal{R}s', \text{ i.e. } \text{support}(\lambda) \subseteq \mathcal{R}.$$

When there exists a lifting of $\mu$ and $\nu$ as above we write $\mu(\mathcal{R})^{\#}\nu$.

**A simple example: bisimulation and liftings** We consider the example of the bisimilar states in Figure 1. To see that $s_0$ and $t_0$ are bisimilar, take the equivalence classes $\mathcal{S}/\mathcal{U} = \{\{s_0, t_0\}, \{s_1, t_1\}, \{s_2, t_2, t_3\}\}$. All states in each class receive the same rewards and transition with equal probability to all other classes, so $\mathcal{U}$ is indeed a bisimulation relation. Now consider the coupling $\lambda$ of $(p(s_0), p(t_0))$ given by the dashed arrows, i.e. $\lambda(s_1, t_1) = \lambda(s_2, t_2) = \lambda(s_2, t_3) = \frac{1}{3}$. This coupling is a *lifting* of $\mathcal{U}$, since all supported states are $\mathcal{U}$-related. This is not a coincidence: the fact that bisimulation is equivalent to the existence of a lifting is the subject of Section 3. Not all couplings are liftings: the trivial coupling $\omega(s_i, t_j) = p(s_0)(s_i)p(t_0)(t_j)$ is not a lifting of $\mathcal{U}$.

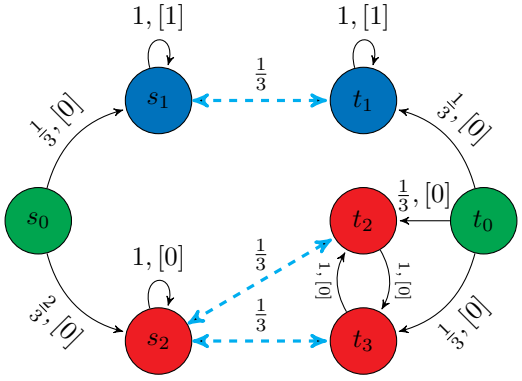



Figure 1: Rewards indicated in square brackets. Dashed arrows give the weights of the coupling of $p(s_0)$ and $p(t_0)$. Colours represent different equivalence classes of $\sim$.

Previous work on metric extensions of bisimulation is built on the Kantorovich metric $K(d)(\mu,\nu) = \min_{\lambda\in\Lambda(\mu,\nu)}\langle\lambda, d\rangle$ (also called the Wasserstein metric). The bisimulation metric $d_\sim$ is the fixed point of the operator $F(d)(s,s') = \max_a\{(1-\gamma)|r(s,a) - r(s',a)| + \gamma K(d)(p_a(s), p_a(s'))\}$, we refer the reader to (Ferns, Panangaden, and Precup 2004) for further background.

## 3 Generalized Bisimulation via Liftings

In bisimulation, the rewards match at the first step and thereafter the states transition such that the bisimulation relation is preserved. This definition can also be captured in terms of couplings: our first result is that bisimulation is equivalent to the existence of a particular lifting of the states.

**Theorem 3.1.** A relation $\mathcal{U}$ is a bisimulation relation $\iff$ $s\mathcal{U}s'$ implies:

1. $\forall a\ r(s,a) = r(s',a)$
2. $\forall a\ p_a(s)(\mathcal{U})^{\#}p_a(s')$

The proofs of this result and later results are given in the Appendix. The backward implication follows from the remarkable Strassen's theorem on couplings (see e.g. (Lindvall 1999)), which implies that for any $\mathcal{R}$, $\mu(\mathcal{R})^{\#}\nu \iff \forall A \subseteq \mathcal{S},\ \mu(A) \leq \nu(\mathcal{R}(A))$. Applied to an equivalence class $C$ and using that $\mathcal{U}$ is symmetric gives the bisimulation property.

Building on the previous result, one can readily generalize the first condition by using a generic relation between states instead of demanding that the rewards match.

**Definition 3.1** ($\mathcal{R}$-bisimulation)**.** Given a *base* relation $\mathcal{R} \subseteq \mathcal{S} \times \mathcal{S}$, an *$\mathcal{R}$-bisimulation relation* $\mathcal{U} \subseteq \mathcal{S} \times \mathcal{S}$ is a new relation where the states are $\mathcal{R}$-related and their transition distributions are $\mathcal{U}$-lifted. Formally, $s\mathcal{U}s'$ implies:

1. $s\mathcal{R}s'$
2. $\forall a\ p_a(s)(\mathcal{U})^{\#}p_a(s')$

We define *$\mathcal{R}$-bisimulation* $\overset{\mathcal{R}}{\sim}$ to be the largest $\mathcal{R}$-bisimulation relation.

This allows one to define arbitrary properties that are preserved by the dynamics of the MDP in a systematic way. We remark, firstly, that the second condition is much stronger than merely requiring that the lifting is a coupling supported by $\mathcal{R}$, that is, $p_a(s)(\mathcal{R})^{\#}p_a(s')$. Requiring $\mathcal{U}$ to be lifted also demands (in a coinductive manner) that the successor states after a transition are $\mathcal{R}$-related *and can themselves exhibit an appropriate coupling*. Secondly, we note that $\overset{\mathcal{R}}{\sim}$ is well-defined, since the union of $\mathcal{R}$-bisimulation relations is itself an $\mathcal{R}$-bisimulation relation. The well-behavedness of $\mathcal{R}$-bisimulations depends on their base relations, i.e. $\overset{\mathcal{R}}{\sim}$ is reflexive, symmetric, and transitive whenever $\mathcal{R}$ has the same property (see Appendix).

## 4 Temporally Extended Metrics

Although the framework presented in Section 3 is agnostic with respect to the base relation $\mathcal{R}$, we will be focusing on the setting of *quantitative relations*. These are relations parametrized by the use of a real number $\varepsilon$, which arise from a *base pseudometric* $\delta : S \times S \to \mathbb{R}^{\geq 0}$ on states (or state-action pairs). More formally, given a base metric $\delta : S \times S \to \mathbb{R}^{\geq 0}$, a quantitative relation $\delta_\varepsilon$ is the relation $\delta_\varepsilon := \delta^{-1}([0, \varepsilon]) = \{(s, s') \mid \delta(s, s') \leq \varepsilon\}$. We note the distinction between the metric $\delta$ and the relations $\delta_\varepsilon$ derived from the metric. We call a bisimulation arising from such a quantitative relation a *quantitative bisimulation*, and will abuse notation by writing $\overset{\delta}{\sim}_\varepsilon$ rather than $\overset{\delta_\varepsilon}{\sim}$. An example of quantitative relations is *approximate* reward equality $s\rho_\varepsilon s' := \max_a |r(s, a) - r(s', a)| \leq \varepsilon$, derived from the base metric $\rho(s, s') = \max_a |r(s, a) - r(s', a)|$.

In the context of quantitative bisimulations, we can define a new metric by taking an infimum over the $\varepsilon$ parameter. We call these the *temporally extended (TE) metrics*. The TE metric finds the minimum $\varepsilon$ such that the states are $\delta_\varepsilon$-bisimilar. That is, the two states are a distance of $\varepsilon$ away (in the base metric $\delta$) and can be coupled corecursively so that future states are $\varepsilon$ away and can themselves be coupled. A temporal extension can be defined for any base metric.

**Definition 4.1** (TE metric)**.** Given a base metric $\delta$ and a corresponding collection of quantitative relations $\{\delta_\varepsilon\}_{\varepsilon \geq 0}$, the *TE metric* for $\delta$ is defined by

$$d_\tau^\delta(s, s') = \inf \left\{ \varepsilon \mid s \overset{\delta}{\sim}_\varepsilon s' \right\}.$$

This construction gives well-defined pseudometrics. The proof follows from the symmetry, transitivity, and additivity[1] of the relations $\overset{\delta}{\sim}_\varepsilon$ (see Appendix for details).

**Theorem 4.1.** Given a base pseudometric $\delta$, the TE metric $d_\tau^\delta$ is indeed a pseudometric on $\mathcal{S}$.

Moreover, the temporally extended metrics assign distance 0 to states if and only if they are perfectly bisimilar in the base metric (i.e. they are $\delta_0$-bisimilar). In the context of reward differences, this implies:

**Theorem 4.2.** Classical bisimulation corresponds exactly to the kernel of the temporal extension of $\rho$, i.e.

$$s \sim s' \iff d_\tau^\rho(s, s') = 0.$$

For reward differences, the bisimulation metrics share this property, although our metrics are more general. Furthermore, despite the kernels matching, the TE metrics are *not the same* as the bisimulation metrics, both in construction and in the distances they assign.

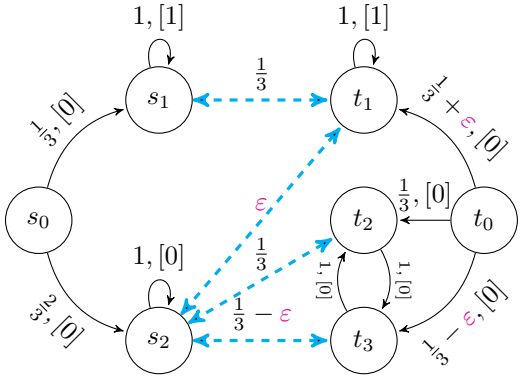



Figure 2: Rewards indicated in square brackets. Dashed arrows give the weights of the coupling of $p(s_0)$ and $p(t_0)$.

**A simple example revisited** We consider the almost-bisimilar states in Figure 2, examined with $\rho$ as the base metric. In this example, all states are $\rho_1$-bisimilar, but not $\rho_0$-bisimilar. This is captured by the metric: one needs to couple $(s_2, t_1)$, since the marginal onto $t_1$ has to equal $1/3 + \varepsilon$ and $s_1$ only has $1/3$ to spare. Since $(s_2, t_1) \in \text{support}(\lambda)$ and $|r(s_2) - r(t_1)| = 1$, then $d_\tau^\rho(s_0, t_0) = 1$. This example highlights the discontinuous behaviour of the TE metric – the states can either be coupled with rewards 0 or 1.

## 5 Comparing Bisimulation Metrics and Temporally Extended Metrics

In this section, we compare the temporally extended metrics with the bisimulation metrics. Results in this section are given in terms of reward metric $\rho$ but can be generalized to arbitrary base metrics. The proofs (see Appendix) elucidate the very useful properties of liftings.

[1] $s \overset{\delta}{\sim}_\alpha t \ \&\ t \overset{\delta}{\sim}_\beta w \Rightarrow s \overset{\delta}{\sim}_{\alpha+\beta} w$

### 5.1 Bounds

Our first result relates the TE metric and the bisimulation metric with a bound.

**Theorem 5.1.** The temporal extension of $\rho$ upper bounds the bisimulation metric: $\forall s, s' \in \mathcal{S}$,

$$d_{\sim}(s, s') \leq d^{\rho}_{\tau}(s, s').$$

The bound is tight, and equality need not hold, as Figure 2 shows. Consequently, using bounds from (Ferns, Panangaden, and Precup 2004), the TE metric gives a guarantee on the difference in optimal value functions and on the approximation error for state-abstraction.

**Corollary 5.1.1.** Let $\hat{V}$ be the value function in the abstract MDP of any abstraction $\phi$, not necessarily a bisimulation.[2] Then, $\forall s, s' \in \mathcal{S}$:

$$|V^*(s) - V^*(s')| \leq \frac{1}{1-\gamma} d^{\rho}_{\tau}(s, s') \quad \text{and}$$

$$|\hat{V}^*(\phi(s)) - V^*(s)| \leq \frac{\gamma}{(1-\gamma)^2} \max_{\phi(s')} \max_{s' \in \phi(s')} \frac{1}{|\phi(s')|} \sum_{s'' \in \phi(s)} d^{\rho}_{\tau}(s', s'')$$

### 5.2 Optimal Couplings

In Figure 2, the same coupling minimized both the bisimulation metric and the TE metric. Interestingly, the couplings chosen need not be the same in general. This is the content of the next theorem.

**Theorem 5.2.** A minimum coupling $\lambda \in \operatorname{argmin}_{\Lambda(p_a(s), p_a(s'))} K(d_{\sim})(p_a(s), p_a(s'))$ of the bisimulation metric need not be a lifting of the optimal bisimulation $\overset{\rho}{\sim}_{d^{\rho}_{\tau}(s,s')}$. Conversely, $\lambda$ which lifts the optimal bisimulation $\overset{\rho}{\sim}_{d^{\rho}_{\tau}(s,s')}$ need not be a minimizer of $K(d_{\sim})(p_a(s), p_a(s'))$.

This highlights the different behaviours of the two metrics – the TE metric aims to minimize the reward difference between coupled states at every step so as to ensure that a single bisimulation relation holds, whereas the bisimulation metric is not preserving a single relation and is willing to couple large differences at an initial step. The couplings chosen by the bisimulation metric do not give a (generalized) bisimulation relation, and the best that one can do with a (generalized) bisimulation relation is given by the temporal extension.

## 6 Discussion

We have introduced the *temporally extended metrics*, a novel class of metrics for behavioural equivalence, which are based on a generalized notion of bisimulation. We have established bounds and other connections with the more familiar bisimulation metric, and seen that they neither compute the same values nor pick out the same couplings of state distributions. After completing this work we discovered that similar ideas for quantitative bisimulations have successfully appeared in the control-theory literature (Girard and Pappas 2007), although in the different setting of non-deterministic (rather than probabilistic) transition systems without rewards.

This work marks the beginning of an investigation into formally safe, computationally tractable, and model-free metrics for behavioural equivalence. There are many interesting avenues for future work that we intend to pursue. For computational aspects, the TE metric involves the computation of a bisimulation relation rather than a bisimulation metric, which can be done exactly in $\mathcal{O}(|\mathcal{A}||\mathcal{S}|^3)$ via partition refinements as opposed to approximated upto a degree of accuracy $\varepsilon$ in $\mathcal{O}(|\mathcal{A}||\mathcal{S}|^4 \log |\mathcal{S}| \log \varepsilon)$ (Ferns, Panangaden, and Precup 2004). Deriving an exact algorithm, however, is left for future work. The possibility of model-free computation is hypothesized since the metric requires only the *existence* of a lifting, as opposed to the exact weights of a coupling as does the Kantorovich metric, thus should be easier to estimate from samples.

A slightly disconcerting aspect of the TE metrics is their discontinuity with respect to the transition distributions, observed in Figure 2. This is because we require exact couplings - we are currently investigating the use of *approximate couplings* to remedy this, which have recently surfaced in the study of differential privacy (Barthe et al. 2016).

Finally, as the general notions of $\mathcal{R}$-bisimulations and $\delta$-temporal extensions can be considered for arbitrary relations and metrics, examining the interplay between different notions of bisimulation (e.g. for optimal value functions or policy value functions) promises to be a fruitful direction.

[2] we refer the reader to (Li, Walsh, and Littman 2006) for background on abstract MDPs

## References


Barthe, G.; Fong, N.; Gaboardi, M.; Grégoire, B.; Hsu, J.; and Strub, P.-Y. 2016. Advanced probabilistic couplings for differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 55–67. ACM.

Ferns, N.; Panangaden, P.; and Precup, D. 2004. Metrics for finite Markov decision precesses. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 162–169.

Girard, A., and Pappas, G. J. 2007. Approximation metrics for discrete and continuous systems. *IEEE Transactions on Automatic Control* 52(5):782–798.

Givan, R.; Dean, T.; and Greig, M. 2003. Equivalence notions and model minimization in Markov decision processes. *Artificial Intelligence* 147(1-2):163–223.

Li, L.; Walsh, T. J.; and Littman, M. L. 2006. Towards a unified theory of state abstraction for mdps.

Lindvall, T. 1999. On Strassen's theorem on stochastic domination. *Electronic communications in probability* 4:51–59.

Milner, R. 1980. *A Calculus for Communicating Systems*, volume 92 of *Lecture Notes in Computer Science*. Springer-Verlag.

# A Proofs

**Theorem A.1.** A relation $\mathcal{U}$ is a bisimulation relation $\iff s\mathcal{U}s'$ implies: 1. $\forall a\ r(s,a) = r(s',a)$ and 2. $\forall a\ p_a(s)(\mathcal{U})^{\#}p_a(s')$

*Proof.* The first condition is immediate in both cases. For the forward implication, we pick the coupling $\lambda_a(s',t') = \mathbb{1}_{[s'\mathcal{U}t']}p_a(s)(s')p_a(t)(t')/p_a(s)([s'])$, where $[s'] := \{t' \mid s'\mathcal{U}t'\}$ is the equivalence class of $s'$. The marginals match since: $\lambda_a(s',\mathcal{S}) = \sum_{t':s'\mathcal{U}t'} p_a(s)(s')p_a(t)(t')/p_a(s)([s']) = p_a(s)(s')p_a(t)([s'])/p_a(s)([s']) = p_a(s)(s')$, as $p_a(s)([s']) = p_a(t)([s'])$ by the second condition of bisimulation. Similarly for $\lambda_a(\mathcal{S},t')$. To make $\lambda_a$ a lifting of $\mathcal{U}$ we still need to check that $\text{support}(\lambda_a) \subseteq \mathcal{U}$, which is evident since $\lambda_a(s',t')$ is only non-zero when $s'\mathcal{U}t'$. For the converse, we use Strassen's theorem. Let $C \in \mathcal{S}/\mathcal{U}$, note that we have $p_a(s)(C) \leq p_a(t)(C)$ since $\mathcal{U}(C) = C$. By symmetry of $\mathcal{U}$, we also have $p_a(t)(\mathcal{U})^{\#}p_a(s)$, so that $p_a(t)(C) \leq p_a(s)(C)$. So $s \sim t$. □

**Theorem A.2.** Given a base pseudometric $\delta$, the TE metric $d_\tau^\delta$ is indeed a pseudometric on $\mathcal{S}$.

*Proof.* 1. Note that $s \overset{\delta}{\sim}_0 s$ (via the identity coupling $\lambda(s',s'') = \mathbb{1}_{[s=s']}p_a(s)(s')$), thus $d_\tau^\delta(s,s) = \inf_{\varepsilon\geq 0}\{s \overset{\delta}{\sim}_\varepsilon s\} = 0$.

2. Note that $s \overset{\delta}{\sim}_\varepsilon t \Rightarrow t \overset{\delta}{\sim}_\varepsilon s$ (via the mirror coupling $\psi(t',s') = \lambda(s',t')$), thus $d_\tau^\delta(s,t) = \inf_\varepsilon\{s \overset{\delta}{\sim}_\varepsilon t\} = \inf_\varepsilon\{t \overset{\delta}{\sim}_\varepsilon s\} = d_\tau^\delta(t,s)$.

3. Let $A = A_1 + A_2 = \{\varepsilon_1+\varepsilon_2 | s \overset{\delta}{\sim}_{\varepsilon_1} w \ \&\ w \overset{\delta}{\sim}_{\varepsilon_2} t\}$, and $B = \{\varepsilon | s \overset{\delta}{\sim}_\varepsilon t\}$. Note $A \subseteq B$ since $s \overset{\delta}{\sim}_{\varepsilon_1+\varepsilon_2} t$ (via the transitive coupling $\lambda_a(s',t') = \sum_{w'\in\text{support}(p_a(w))} \frac{\lambda_{a,1}(s',w')\lambda_{a,2}(w',t')}{p_a(w)(w')}$). So $d_\tau^\delta(s,t) = \inf(B) \leq \inf(A) = \inf(A_1) + \inf(A_2) = d_\tau^\delta(s,w) + d_\tau^\delta(w,t)$. □

**Theorem A.3.** The temporal extension of $\rho$ upper bounds the bisimulation metric: $\forall s,s' \in \mathcal{S}$,

$$d_\sim(s,s') \leq d_\tau^\rho(s,s').$$

*Proof.* We proceed by induction, showing that $\forall s,t,\ d_n(s,s') \leq (1-\gamma)\sum_{i=0}^n \gamma^i d_\tau^\rho(s,s')$. The base case is $d_1(s,s') = \max_a\{(1-\gamma)|r(s,a) - r(s',a)|\} \leq (1-\gamma)d_\tau^\rho(s,s')$, using $\max_a|r(s,a) - r(s',a)| \leq d_\tau^\rho(s,s')$. For the induction step we upper bound the min-cost coupling from the minimization problem with the liftings $\lambda_a \in \Lambda\left(p_a(s),p_a(s')\right)$ given by $\overset{\rho}{\sim}_{d_\tau^\rho(s,s')}$ (one can show the infs are always attained).

$$\begin{aligned}
d_{n+1}(s,s') &= \max_a\{1-\gamma)|r(s,a) - r(s',a)| \\
&\quad + \gamma K(d_n)(p_a(s),p_a(s'))\} \\
&\leq (1-\gamma)d_\tau^\rho(s,s') + \gamma\sum_{k,j}\lambda_a(s_k,s_j)d_n(s_k,s_j) \\
&\leq (1-\gamma)d_\tau^\rho(s,s') \\
&\quad + \gamma\sum_{k,j}\lambda_a(s_k,s_j)\left((1-\gamma)\sum_{i=0}^n\gamma^i d_\tau^\rho(s_k,s_j)\right) \\
&\qquad \text{(induction hypothesis)}
\end{aligned}$$

Now we use the lifting property: the only non-zero terms in the summation are those for which $(s_k,s_j) \in \text{support}(\lambda_a) \subseteq \overset{\rho}{\sim}_{d_\tau^\rho(s,s')}$. Thus $(s_k,s_j) \in \overset{\rho}{\sim}_{d_\tau^\rho(s,s')}$, and we conclude that $d_\tau^\rho(s_k,s_j) = \inf_\varepsilon s_k \overset{\rho}{\sim}_\varepsilon s_j \leq d_\tau^\rho(s,s'), \forall s_k,s_j$.

$$\begin{aligned}
d_{n+1}(s,s') &\leq (1-\gamma)d_\tau^\rho(s,s') \\
&\quad + \gamma\sum_{k,j}\lambda_a(s_k,s_j)\left((1-\gamma)\sum_{i=0}^n\gamma^i d_\tau^\rho(s,s')\right) \\
&= (1-\gamma)d_\tau^\rho(s,s')\sum_{i=0}^{n+1}\gamma^i
\end{aligned}$$

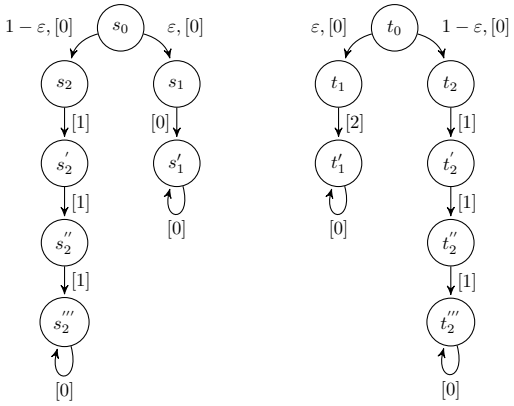
Thus the inequality holds for all $n$. Taking limits finishes the proof:

$$d_\sim(s,s') \leq \frac{1-\gamma}{1-\gamma}d_\tau^\rho(s,s') = d_\tau^\rho(s,s'),$$

as desired. □

**Theorem A.4.** A minimum coupling $\lambda \in \text{argmin}_{\Lambda(p_a(s),p_a(s'))} K(d_\sim)(p_a(s),p_a(s'))$ of the bisimulation metric need not be a lifting of the optimal bisimulation $\overset{\rho}{\sim}_{d_\tau^\rho(s,s')}$. Conversely, $\lambda$ which lifts the optimal bisimulation $\overset{\rho}{\sim}_{d_\tau^\rho(s,s')}$ need not be a minimizer of $K(d_\sim)(p_a(s),p_a(s'))$.

*Proof.* Consider the following MDP, taking $\rho$ as our base metric.




And consider the following two couplings $\omega_\sim, \lambda_\tau \in \Lambda(p(s_0),p(t_0))$.

| $\omega_\sim$ | $s_1$ | $s_2$ |
|---|---|---|
| $t_1$ | $\varepsilon$ | 0 |
| $t_2$ | 0 | $1-\varepsilon$ |

| $\lambda_\tau$ | $s_1$ | $s_2$ |
|---|---|---|
| $t_1$ | 0 | $\varepsilon$ |
| $t_2$ | $\varepsilon$ | $1-2\varepsilon$ |

One can verify that $\omega_\sim$ minimizes the bisimulation distance and that $\lambda_\tau$ minimizes the temporally extended metric. Indeed, $\sum_{u,v\in\mathcal{S}}\omega_\sim(u,v)d_\sim(u,v) = \varepsilon d_\sim(s_1,t_1) + (1-\varepsilon)d_\sim(s_2,t_2) = 2\varepsilon\{2(1-\gamma)\}$ and this coupling is optimal for $K(d_\sim)$. Meanwhile $\sum_{u,v\in\mathcal{S}}\lambda_\tau(u,v)d_\sim(u,v) = \varepsilon d_\sim(s_1,t_2) + \varepsilon d_\sim(s_2,t_1) + (1-2\varepsilon)d_\sim(s_2,t_2) = \varepsilon d_\sim(s_1,t_2) + \varepsilon d_\sim(s_2,t_1) = 2\varepsilon\{(1-\gamma)(1+\gamma+\gamma^2)\}$.

On the other hand, using $\omega_\sim$, the best relation that can be lifted is $\overset{\rho}{\sim}_2$, since $(s_1,t_1) \in \text{support}(\omega_\sim)$ and $r(s_1) - r(t_1) = 2$. Meanwhile, $\lambda_\tau$ lifts $\overset{\rho}{\sim}_1$ and thus achieves the minimum lifting, since $(s_1,t_2),(s_2,t_1),(s_2,t_2)$ all have reward differences of 1. Thus $\omega_\sim$ minimizes the bisimulation metric but not the temporal extension metric. Conversely, $\lambda_\tau$ minimizes the temporal extension metric since it achieves the minimum lifting, but not the bisimulation metric. □