On the relation of bisimulation, model irrelevance, and corresponding policy performance bounds

Alperen Tercan
Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI 48109
tercan@umich.edu

Necmiye Ozay
Electrical and Computer Engineering
University of Michigan
Ann Arbor, MI 48109
necmiye@umich.edu

Abstract

State abstraction is a key tool for scaling reinforcement learning (RL) by reducing the complexity of the underlying Markov Decision Process (MDP). Among abstraction methods, bisimulation has emerged as a principled metric-based approach, yet its suboptimality properties remain less understood compared to model irrelevance abstractions. In this work, we clarify the relationship between these two abstraction families: while model irrelevance implies bisimulation, the converse does not hold, leading to coarser abstractions under bisimulation. We provide the first suboptimality bounds for policies derived from approximate bisimulation abstractions, analyzing both "naive" and "smart" refinement strategies for lifting abstract policies back to the original MDP. Our theoretical results show that smart refinement enjoys strictly better suboptimality guarantees, and our experiments on Garnet MDPs confirm that this advantage translates into significant performance improvements. We further explain this gap through the action gap phenomenon in RL, which helps account for why some refinement strategies yield substantially better behavior in practice.

1 Introduction

Abstraction plays a crucial role in scaling reinforcement learning (RL) to large or complex environments by simplifying the underlying Markov Decision Process (MDP). Among various abstraction techniques, the bisimulation metric introduced by [3] has emerged as a principled and widely used approach. This metric groups states based on their behavioral similarity, offering a quantifiable way to construct abstract MDPs that preserve decision-relevant structure. Its effectiveness has also been demonstrated as an auxiliary task for noise-robust representation learning [9] and as a generalization framework for goal-conditioned RL [4].

Despite its popularity, the suboptimality incurred by using policies derived from bisimulation-based abstractions remains largely unexplored. One reason for this gap may be the common assumption that bisimulation-based abstraction is a variant of model irrelevance abstraction, for which policy performance guarantees are well established. However, a key observation—and the starting point of our work—is that model irrelevance implies bisimulation equivalence, but not vice versa. As a result, bisimulation induces coarser abstractions, and existing suboptimality bounds for model irrelevance do not necessarily apply.

In this paper, we provide the first analysis of suboptimality bounds for policies derived from bisimulation-based abstract MDPs. We further investigate two natural policy refinement approaches for transferring abstract policies back to the original MDP. Interestingly, while both methods are valid, they lead to different theoretical suboptimality guarantees. Our empirical results reveal that this diver-

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

gence is not merely theoretical; in practice, one method consistently outperforms the other—often by a wide margin.

We conclude by discussing how the action gap phenomenon in RL can be used to explain this observation, offering a new perspective on why some refinement strategies yield better performance in practice. Together, our results highlight the importance of carefully considering both the abstraction method and the policy refinement strategy when using bisimulation in RL.

2 Preliminaries

In this section, we introduce the notation and concepts we use in this paper.

An MDP is defined by the tuple $\mathrm{M} = (\mathcal{S}, \mathcal{A}, \mathrm{P}, \mathrm{r}, \gamma)$. Here, $\mathcal{S}$ denotes a finite set of states, and $\mathcal{A}$ represents a finite set of actions. The transition probability function $\mathrm{P} : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$[1] gives the probability $\mathrm{P}(s'|s,a)$ of reaching state s' from state s after action a. The reward function $\mathrm{r} : \mathcal{S} \times \mathcal{A} \to [0,1]$ defines the immediate reward $r(s,a)$ received upon taking action a in state s. The discount factor $\gamma \in [0,1)$ adjusts the weight of future rewards.

A common way to abstract MDPs is state aggregation, defined by the membership function $\phi : \mathcal{S} \to \mathcal{S}_\phi$, mapping states of an MDP to abstract state. We define ϕ^{-1} to be the preimage of ϕ. Finally, we define $\mathcal{N} : \mathcal{S} \to 2^{\mathcal{S}}$ as the neighborhood function which is a shorthand for $\phi^{-1}(\phi(\cdot))$, i.e., $\mathcal{N}(s)$ returns the set of all states that share the same abstract label as s.

Then, we construct the abstract MDP $\mathrm{M}_\phi = (\mathcal{S}_\phi, \mathcal{A}, \mathrm{P}_\phi, \mathrm{r}_\phi, \gamma)$ where:

$$\mathrm{P}_\phi(z'|z,a) = \frac{1}{|\phi^{-1}(z)|} \sum_{s \in \phi^{-1}(z)} \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|s,a), \qquad \mathrm{r}_\phi(z,a) = \frac{1}{|\phi^{-1}(z)|} \sum_{s \in \phi^{-1}(z)} \mathrm{r}(s,a). \tag{1}$$

Next, we classify these abstractions based on the properties they satisfy. There are two common state similarity notions in the literature: bisimulation and model irrelevance. In the exact case, these two notions are equivalent to each other ([8], [3]):

Definition 1. *A state abstraction ϕ is called a model-irrelevant abstraction or a bisimulation abstraction if it satisfies the following condition:*

$$\phi(s_1) = \phi(s_2) \implies (r(s_1,a) = r(s_2,a) \text{ and } \mathrm{P}(C|s_1,a) = \mathrm{P}(C|s_2,a) \ \ \forall C \in \mathcal{S}_\phi) \ \ \forall a \in \mathcal{A}, \tag{2}$$

where $\mathrm{P}(C|\cdot,a)$ is shorthand for $\sum_{s \in C} \mathrm{P}(s|\cdot,a)$.

This is an exact and "lossless" abstraction of the original MDP, i.e., the abstract MDP is enough to compute the optimal policy of the original MDP. However, this condition is very strict in practice and cannot lead to coarse enough state abstractions. Hence, in the literature, approximate relaxations of this condition have been introduced.

The first approximate abstraction is bisimulation-based and aggregates states that are close under a bisimulation metric.[2] While the bisimulation metric is not unique, we consider the bisimulation metric obtained via the fixed-point iteration introduced by [3].

Definition 2. *Bisimulation metric d_{fix} is the least-fixed point of the following fixed-point iteration:*

$$F(d)(s,s') = \max_{a \in \mathcal{A}} (1-\gamma)|r(s,a) - r(s',a)| + \gamma T_K(d)(\mathrm{P}(\cdot|s,a), \mathrm{P}(\cdot|s',a)). \tag{3}$$

where $T_K(d) : \Delta_{\mathcal{S}} \times \Delta_{\mathcal{S}} \to [0,1]$ is the Kantorovich metric under metric d.

Remark 1. *For a given d, $T_K(d)(\mathrm{P}(\cdot|s,a), \mathrm{P}(\cdot|s',a))$ can be computed via the following linear program:*

$$\begin{aligned} \max_{u_i, i=1\ldots|S|} \quad & \sum_{i=1}^{|S|} (\mathrm{P}(s_i|s,a) - \mathrm{P}(s_i|s',a)) u(s_i) \\ \text{subject to :} \quad & \forall i,j.\ u(s_i) - u(s_j) \le d(s_i,s_j) \\ & \forall i.\ 0 \le u(s_i) \le 1 \end{aligned} \tag{4}$$

[1] $\Delta_{\mathcal{S}}$ is the probability simplex over $\mathcal{S}$.

[2] While this is actually a semimetric as it does not satisfy separation property, we will call it a metric as common in literature.

Note that Definition 2 is a special case of the metric introduced in [3] with weights of reward and transition terms are set as $(1-\gamma)$ and γ, respectively.

Definition 3. *A state abstraction ϕ is an ϵ_B-approximate bisimulation abstraction if $\phi(s_1) = \phi(s_2) \implies d_{fix}(s_1, s_2) \leq \epsilon_B$ for all s_1, s_2 in $\mathcal{S}$.*

The second approximate abstraction replaces the states' similarity under a metric with conditions on their reward and transitions functions.

Definition 4. *A state abstraction ϕ is called an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction if it satisfies the following condition:*

$$\phi(s_1) = \phi(s_2) \implies \left(|r(s_1,a) - r(s_2,a)| \leq \epsilon_R \text{ and } \sum_{C \in \mathcal{S}_\phi} |\mathrm{P}(C|s_1,a) - \mathrm{P}(C|s_2,a)| \leq \epsilon_P\right) \forall a \in \mathcal{A} \tag{5}$$

3 Comparing Bisimulation and Model Irrelevance

Approximate bisimulation abstractions have long been considered an example of approximate model irrelevance abstractions as they both consider one-step reward and transition probability similarities between states and coincide in the exact (lossless) case ([8]). Potentially due to this understanding, analysis of the suboptimality of the policies obtained from bisimulation abstractions has been omitted in the literature so far to the best of our knowledge. However, in this section, we show that model irrelevance implies bisimulation but the converse does not hold. More formally, we show the following two lemmas with the first one stating that all model irrelevance abstractions are bisimulation abstractions and the second one stating that there exists bisimulation abstractions that cannot be captured by nontrivial model-irrelevance abstractions.

Lemma 1. *If ϕ is an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction, it is also an ϵ_B-bisimulation abstraction with $\epsilon_B = \epsilon_R + \frac{\gamma}{1-\gamma}\frac{\epsilon_P}{2}$.*

Lemma 2. *For any $\epsilon_P < 2$, there exists an arbitrarily small ϵ_B, an MDP M, and state aggregation function ϕ such that ϕ is an ϵ_B-approximate bisimulation abstraction but there exists s_1, s_2 such that $\phi(s_1) = \phi(s_2)$ with $\sum_{C \in \mathcal{S}_\phi} |\mathrm{P}(C|s_1,a) - \mathrm{P}(C|s_2,a)| > \epsilon_P$.*

Proof of Lemma 1 Let s_1 and s_2 be any two states such that $\phi(s_1) = \phi(s_2)$. Then, $|R(s_1,a) - R(s_2,a)| \leq \epsilon_R$ and $\sum_{C \in \mathcal{S}_\phi} |P(C|s_1,a) - P(C|s_2,a)| \leq \epsilon_P$ for all actions a. Proving the lemma requires showing $d(s_1,s_2) \leq \epsilon_R + \frac{\gamma}{1-\gamma} \cdot \frac{\epsilon_P}{2}$. We establish this by induction on the number of fixed-point iterations applied on the starting metric d^0, which is zero everywhere.

The base case is trivial as $d^0(s_1,s_2) = 0$. Next, assume that

$$\phi(s_1) = \phi(s_2) \implies d^k(s_1,s_2) \leq \epsilon_R + \tfrac{\gamma}{1-\gamma} \cdot \tfrac{\epsilon_P}{2},$$

and show that the same holds for $d^{k+1}(s_1,s_2)$.

Using the definition of T_K in Remark 1,

$$T_K(d^k)(P(\cdot|s_1,a), P(\cdot|s_2,a)) = \max_{u \in U(d^k)} \sum_{i=1}^{|S|} (P(s_i|s_1,a) - P(s_i|s_2,a))\, u(s_i),$$

where $U(d^k) = \{0 \leq u \leq 1 \mid u(s_i) - u(s_j) \leq d^k(s_i,s_j)\}$.

For a fixed u, define $M_C^+ = \max_{s_i \in C} u(s_i)$, $M_C^- = \min_{s_i \in C} u(s_i)$, and $g(C) = \frac{1}{2}(M_C^+ + M_C^-)$. Then,

$$\begin{aligned}\sum_{i=1}^{|S|} (P(s_i|s_1,a) - P(s_i|s_2,a))u(s_i) = &\sum_{i=1}^{|S|} (u(s_i) - g(\phi(s_i)))(P(s_i|s_1,a) - P(s_i|s_2,a)) \\ &+ \sum_{i=1}^{|S|} (P(s_i|s_1,a) - P(s_i|s_2,a)) g(\phi(s_i)).\end{aligned}$$

For every $s_i \in C$,

$$\begin{aligned}
|u(s_i) - g(\phi(s_i))| &= \max\{g(\phi(s_i)) - u(s_i),\, u(s_i) - g(\phi(s_i))\} \\
&\leq \max\{g(\phi(s_i)) - M_C^-,\, M_C^+ - g(\phi(s_i))\} \\
&= \tfrac{M_C^+ - M_C^-}{2} \\
&= \tfrac{\max_{s_i, s_j \in C}(u(s_i) - u(s_j))}{2} \\
&\leq \tfrac{\max_{s_i, s_j \in C} d_k(s_i, s_j)}{2} \\
&\leq \tfrac{1}{2}\Big(\epsilon_R + \tfrac{\gamma}{1-\gamma}\tfrac{\epsilon_P}{2}\Big).
\end{aligned}$$

Thus,

$$\sum_{i=1}^{|S|} |u(s_i) - g(\phi(s_i))|(P(s_i|s_1, a) - P(s_i|s_2, a)) \leq \epsilon_R + \tfrac{\gamma}{1-\gamma}\tfrac{\epsilon_P}{2}. \tag{6}$$

Since $0 \leq M_{\phi(s_i)}^- \leq g(\phi(s_i)) \leq M_{\phi(s_i)}^+ \leq 1$, we have $0 \leq g(\phi(s_i)) \leq 1$. Then, rewriting the summation over clusters instead of states, we obtain:

$$\sum_{i=1}^{|S|}(P(s_i|s_1, a) - P(s_i|s_2, a))g(\phi(s_i)) = \sum_{C \in \mathcal{S}_\phi}(P(C|s_1, a) - P(C|s_2, a))g(C). \tag{7}$$

This summation is maximized by picking $g(C) = 1$ when $P(C|s_1, a) \geq P(C|s_2, a)$ and $g(C) = 0$ otherwise. Since $\sum_{C \in \mathcal{S}_\phi}(P(C|s_1, a) = \sum_{C \in \mathcal{S}_\phi}(P(C|s_2, a) = 1$, this selection of g results in value $\frac{1}{2}\sum_{C \in \mathcal{S}_\phi}|P(C|s_1, a) - P(C|s_2, a)|$. Hence,

$$\sum_{i=1}^{|S|}(P(s_i|s_1, a) - P(s_i|s_2, a))g(\phi(s_i)) \leq \tfrac{1}{2}\sum_{C \in \mathcal{S}_\phi}|P(C|s_1, a) - P(C|s_2, a)| \leq \tfrac{\epsilon_P}{2}. \tag{8}$$

Then, combining Equation (6) and (8), we obtain:

$$\sum_{i=1}^{|S|}(P(s_i|s_1, a) - P(s_i|s_2, a))u(s_i) \leq \epsilon_R + \tfrac{1}{1-\gamma}\tfrac{\epsilon_P}{2}. \tag{9}$$

Since this holds for all $u \in U(d^k)$,

$$T_K(d^k)(P(\cdot|s_1, a), P(\cdot|s_2, a)) \leq \epsilon_R + \tfrac{1}{1-\gamma}\tfrac{\epsilon_P}{2}.$$

By definition,

$$\begin{aligned}
d^{k+1}(s_1, s_2) &= \max_a (1-\gamma)|R(s_1, a) - R(s_2, a)| + \gamma T_K(d^k)(P(\cdot|s_1, a), P(\cdot|s_2, a)) \\
&\leq (1-\gamma)\epsilon_R + \gamma\Big(\epsilon_R + \tfrac{1}{1-\gamma}\tfrac{\epsilon_P}{2}\Big) \\
&= \epsilon_R + \tfrac{\gamma}{1-\gamma}\tfrac{\epsilon_P}{2}.
\end{aligned}$$

This concludes the induction proof.

□

Proof of Lemma 2 To prove this lemma, we will construct a parametric MDP.

On the MDP shown in Figure 1, we define ϕ such that $\phi(s_1) = \phi(s_2) = \phi(s_3) \neq \phi(s_4)$. Next, we compute the bisimulation distances for all pairs:

$$\begin{aligned}
d_{\text{fix}}(s_2, s_4) &= (1-\gamma)\eta_2 \\
d_{\text{fix}}(s_2, s_3) &= (1-\gamma)\eta_1
\end{aligned}$$

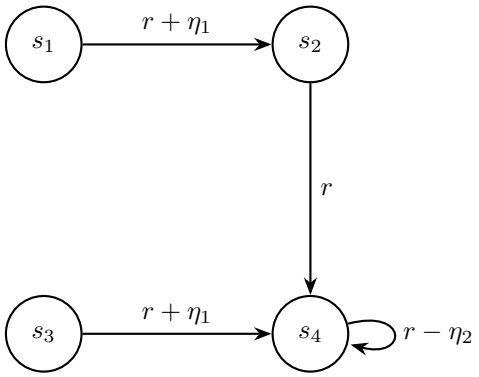

Figure 1: Counter example to be used for Lemma 2. We will show that for any ϵ, η_1 and η_2 can be chosen such that s_1, s_2, and s_3 are a single cluster, s_4 is a cluster of its own, and s_2 and s_4 are arbitrarily close.

$$
\begin{aligned}
d_{\text{fix}}(s_3, s_4) &= (1-\gamma)(\eta_1 + \eta_2) \\
d_{\text{fix}}(s_1, s_2) &= (1-\gamma)\eta_1 + \gamma(1-\gamma)\eta_2 \\
d_{\text{fix}}(s_1, s_3) &= \gamma(1-\gamma)\eta_2 \\
d_{\text{fix}}(s_1, s_4) &= (1-\gamma)(\eta_1 + \eta_2) + \gamma(1-\gamma)\eta_2
\end{aligned}
$$

First, we observe that $(1-\gamma)\eta_1 + \gamma(1-\gamma)\eta_2 < \epsilon_B$ makes ϕ an ϵ_B-approximate bisimulation abstraction. For any η_2, picking η_1 such that $\eta_1 \geq \frac{\epsilon}{(1-\gamma)} - \eta_2(\gamma+1)$ satisfies the condition. Then, we can treat η_2 as a free variable that can be chosen independently, still ensuring ϕ is an ϵ_B-approximate bisimulation abstraction.

Then, we observe that ϕ above results in $\sum_{C \in \mathcal{S}_\phi} |\mathrm{P}(C|s_1, a) - \mathrm{P}(C|s_3, a)| = 2$ but $d_{\text{fix}}(s_1, s_3) = \gamma(1-\gamma)\eta_2$, which can be taken to zero by choosing arbitrarily small η_2.

In short, there is always an MDP and an aggregation function ϕ such that ϕ is an ϵ_B-approximate bisimulation abstraction for an arbitrarily small ϵ_B but it cannot be an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction for a nontrivial (< 2) ϵ_P. This shows that ϵ_P cannot be bounded by a monotone function of ϵ_B.

□

Remark 2. *Lemma 2 states that there are no "non-trivial" model-irrelevant abstractions. However, it is easy to see that if ϕ is ϵ_B-approximate bisimulation abstraction, it is also a $(\frac{\epsilon_B}{1-\gamma}, 2)$-approximate model-irrelevant abstraction. Since 2 is the largest possible ϵ_P anyway, we consider this a "trivial" abstraction because it only needs to account for the reward distances.*

4 Suboptimality Bounds for Abstract Policy Refinement

A natural question for a state abstraction ϕ is the usefulness of solution of M_ϕ in approximating the optimal policy of M. We consider two possible refinements of the M_ϕ's solution. Assuming $V^*_\phi : \mathcal{S}_\phi \to \mathbb{R}$ is computed, the refined policy can be either:

$$\pi_{\phi,N}(s) = \arg\max_a \mathrm{r}_\phi(\phi(s), a) + \gamma \sum_{z' \in \mathcal{S}_\phi} \mathrm{P}_\phi(z'|\phi(s), a) V^*_\phi(z'), \tag{10}$$

or

$$\pi_{\phi,S}(s) = \arg\max_a \mathrm{r}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathrm{P}(s'|s, a) V^*_\phi(\phi(s')). \tag{11}$$

Equation (10) does not make any use of M and is the same as setting $\pi_{\phi,N}(s) = \pi^*_\phi(\phi(s))$. As it does not leverage any additional information about M, we refer to it as the "naive refinement".

In contrast, Equation (11) utilizes the M state the agent is in, instead of the abstract state it is mapped to. This avoids errors due to aggregation in greedy action selection. We refer to this strategy as the "smart refinement".

One desired property for a state abstraction ϕ is that its refined policies $\pi_{\phi,N}$ and $\pi_{\phi,S}$ have bounded suboptimality, i.e., $\|V^* - V^{\pi}\|_\infty$ is bounded for $\pi \in \{\pi_{\phi,N}, \pi_{\phi,S}\}$. Note that when ϕ is an exact abstraction as in Definition 1, $\pi_{\phi,N}$ and $\pi_{\phi,S}$ coincide with each other and result in zero suboptimality.

Suboptimality bound of $\pi_{\phi,N}$ when ϕ is an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction is well-known in the literature.

Theorem 1. *[5] Let ϕ be an (ϵ_R, ϵ_P)-approximate model-irrelevant abstraction on* M *and M_ϕ is defined as in Equation* (1)*. Then:*

$$\|V^* - V^{\pi_{\phi,N}}\| \leq \frac{2\epsilon_R}{1-\gamma} + \frac{\gamma\epsilon_P}{(1-\gamma)^2}. \tag{12}$$

While the theoretical properties of bisimulation metric have been studied in the literature and several value function bounds for the gap between the abstract and original optimal value functions have been given ([3], [7]), suboptimality of the abstract policy on the original MDP has not been shown so far. In this section, we will derive such bounds for $\pi_{\phi,N}$ and $\pi_{\phi,S}$.

Theorem 2. *Let ϕ be an ϵ_B-approximate bisimulation abstraction on* M *and let M_ϕ be defined as in Equation* (1)*. Then:*

$$\|V^* - V^{\pi_{\phi,N}}\| \leq \frac{2\epsilon_B}{(1-\gamma)^3}, \tag{13}$$

and

$$\|V^* - V^{\pi_{\phi,S}}\| \leq \frac{2\gamma\epsilon_B}{(1-\gamma)^3}, \tag{14}$$

Proof of Equation (14) Performance Difference Lemma (PDL) ([6]) states that for any policy π and for any state $s_0 \in S$:

$$V^{\pi^*}(s_0) - V^{\pi}(s_0) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{s_0}} [V^*(s) - Q^*(s,\pi(s))] = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{s\sim d^{\pi}_{s_0}} [\max_a Q^*(s,a) - Q^*(s,\pi(s))], \tag{15}$$

where $d^{\pi}_{s_0}$ is the discounted occupancy measure induced by π when starting from s_0. So, we need to bound $\max_a Q^*(s,a) - Q^*(s,\pi(s))$. Take π to be the policy:

$$\pi_{\phi,S}(s) = \arg\max_a \mathrm{r}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathrm{P}(s'|s,a) V^*_\phi(\phi(s')). \tag{16}$$

We start by showing $\bar{Q}(s,a) = \mathrm{r}(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathrm{P}(s'|s,a) V^*_\phi(\phi(s'))$ satisfies $|\bar{Q}(s,a) - Q^*(s,a)| \leq \frac{\gamma\epsilon_B}{(1-\gamma)^2}$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$:

$$\begin{aligned}
|\bar{Q}(s,a) - Q^*(s,a)| &= |\gamma \sum_{s'\in\mathcal{S}} \mathrm{P}(s'|s,a)(V^*(s') - V^*_\phi(\phi(s'))| \\
&\leq \gamma \sum_{s'\in\mathcal{S}} \mathrm{P}(s'|s,a)|V^*(s') - V^*_\phi(\phi(s')| \\
&\leq \gamma \sum_{s'\in\mathcal{S}} \mathrm{P}(s'|s,a)\frac{\epsilon_B}{(1-\gamma)^2} \qquad \text{[3], for } c_T \geq \gamma \\
&= \frac{\gamma\epsilon_B}{(1-\gamma)^2}.
\end{aligned} \tag{17}$$

Next, we observe that this implies $|\max_a Q^*(s,a) - Q^*(s,\pi_{\phi,S}(s))| \leq \frac{2\gamma\epsilon_B}{(1-\gamma)^2}$. Let $a^* = \arg\max_a Q^*(s,a)$ and $\bar{a}^* = \arg\max_a \bar{Q}(s,a)$. Then,

$$Q^*(s,\bar{a}^*) + \frac{\gamma\epsilon_B}{(1-\gamma)^2} \geq \bar{Q}(s,\bar{a}^*) \geq \bar{Q}(s,a^*) \geq Q^*(s,a^*) - \frac{\gamma\epsilon_B}{(1-\gamma)^2},$$

which implies $|\max_a Q^*(s,a) - Q^*(s,\pi_{\phi,S}(s))| \leq \frac{2\gamma\epsilon_B}{(1-\gamma)^2}$.

Then, from Equation (15):

$$V^*(s_0) - V^{\pi_{\phi,S}}(s_0) \leq \frac{2\gamma\epsilon_B}{(1-\gamma)^3}. \tag{18}$$

□

Proof of Equation (13) We can again use the approach shown above, we just need a new bound on $\max_a Q^*(s,a) - Q^*(s, \pi_{\phi,N}(s))$ for the state-agnostic $\pi_{\phi,N}$:

$$\pi_{\phi,N}(s) = \arg\max_a \mathrm{r}_\phi(\phi(s),a) + \gamma \sum_{z' \in \mathcal{S}_\phi} \mathrm{P}_\phi(z'|\phi(s),a) V^*_\phi(z'). \tag{19}$$

Define $\hat{Q}(s,a) = \mathrm{r}_\phi(\phi(s),a) + \gamma \sum_{z' \in \mathcal{S}_\phi} \mathrm{P}_\phi(z'|\phi(s),a) V^*_\phi(z')$. Then,

$$\begin{aligned}
&|\hat{Q}(s,a) - Q^*(s,a)| \\
=&\left| (\mathrm{r}_\phi(\phi(s),a) - \mathrm{r}(s,a)) + \gamma \sum_{z' \in \mathcal{S}_\phi} \left(\mathrm{P}_\phi(z'|\phi(s),a) V^*_\phi(z') - \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|s,a) V^*(s') \right) \right| \\
\leq& |\mathrm{r}_\phi(\phi(s),a) - \mathrm{r}(s,a)| + \gamma \left| \sum_{z' \in \mathcal{S}_\phi} \left(V^*_\phi(z') \mathrm{P}_\phi(z'|\phi(s),a) - \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|s,a) V^*(s') \right) \right|
\end{aligned}$$

Then, using the definition of abstract MDP, this is equal to:

$$\begin{aligned}
&\left| \frac{1}{|\mathcal{N}(s)|} \sum_{\bar{s} \in \mathcal{N}(s)} \mathrm{r}(\bar{s},a) - \mathrm{r}(s,a) \right| \\
&\quad + \gamma \left| \sum_{z' \in \mathcal{S}_\phi} \left(\left(V^*_\phi(z') \frac{1}{|\mathcal{N}(s)|} \sum_{\bar{s} \in \mathcal{N}(s)} \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|\bar{s},a) \right) - \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|s,a) V^*(s') \right) \right|.
\end{aligned}$$

Reorganizing the summations in the second term:

$$\begin{aligned}
&\frac{1}{|\mathcal{N}(s)|} \left| \sum_{\bar{s} \in \mathcal{N}(s)} \mathrm{r}(\bar{s},a) - \mathrm{r}(s,a) \right| \\
&\quad + \gamma \left| \frac{1}{|\mathcal{N}(s)|} \sum_{\bar{s} \in \mathcal{N}(s)} \sum_{z' \in \mathcal{S}_\phi} \left(\sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|\bar{s},a) V^*_\phi(z') - \sum_{s' \in \phi^{-1}(z')} \mathrm{P}(s'|s,a) V^*(s') \right) \right|.
\end{aligned}$$

Flattening the nested summations in the second term and reorganizing terms:

$$= \frac{1}{|\mathcal{N}(s)|} \left| \sum_{\bar{s} \in \mathcal{N}(s)} \mathrm{r}(\bar{s},a) - \mathrm{r}(s,a) \right| + \gamma \left| \frac{1}{|\mathcal{N}(s)|} \sum_{\bar{s} \in \mathcal{N}(s)} \left(\sum_{s' \in \mathcal{S}} \mathrm{P}(s'|\bar{s},a) V^*_\phi(\phi(s')) - \mathrm{P}(s'|s,a) V^*(s') \right) \right|$$

$$\begin{aligned}
=& \frac{1}{|\mathcal{N}(s)|} \left| \sum_{\bar{s} \in \mathcal{N}(s)} \mathrm{r}(\bar{s},a) - \mathrm{r}(s,a) \right| \\
&\quad + \gamma \left| \frac{1}{|\mathcal{N}(s)|} \sum_{\bar{s} \in \mathcal{N}(s)} \left(\sum_{s' \in \mathcal{S}} \mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s')) - V^*(s')) + \sum_{s' \in \mathcal{S}} (\mathrm{P}(s'|\bar{s},a) - \mathrm{P}(s'|s,a)) V^*(s') \right) \right|
\end{aligned}$$

Using triangular inequality:

$$\leq \frac{1}{|\mathcal{N}(s)|}\left(\left|\sum_{\bar{s}\in\mathcal{N}(s)} \mathrm{r}(\bar{s},a)-\mathrm{r}(s,a)\right| + \gamma\left|\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}(\mathrm{P}(s'|\bar{s},a)-\mathrm{P}(s'|s,a))V^*(s')\right|\right)$$
$$+\gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s'))-V^*(s'))\right|$$
$$= \frac{1}{|\mathcal{N}(s)|}\left(\left|\sum_{\bar{s}\in\mathcal{N}(s)} \mathrm{r}(\bar{s},a)-\mathrm{r}(s,a)\right| + \frac{\gamma}{1-\gamma}\left|\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}(\mathrm{P}(s'|\bar{s},a)-\mathrm{P}(s'|s,a))(1-\gamma)V^*(s')\right|\right)$$
$$+\gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s'))-V^*(s'))\right|$$

$(1-\gamma)V^*(\cdot)$ is a feasible solution to the optimization problem in Equation (4)[3]:

$$\leq \frac{1}{|\mathcal{N}(s)|}\left(\left|\sum_{\bar{s}\in\mathcal{N}(s)} \mathrm{r}(\bar{s},a)-\mathrm{r}(s,a)\right| + \frac{\gamma}{1-\gamma}\left|\sum_{\bar{s}\in\mathcal{N}(s)} T_K(d_{\text{fix}})(\mathrm{P}(\cdot|\bar{s},a),\mathrm{P}(\cdot|s,a))\right|\right)$$
$$+\gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s'))-V^*(s'))\right|$$

Using triangular inequality and the definition of d_{fix}:

$$\leq \frac{1}{|\mathcal{N}(s)|(1-\gamma)}\left(\sum_{\bar{s}\in\mathcal{N}(s)}(1-\gamma)\left|\mathrm{r}(\bar{s},a)-\mathrm{r}(s,a)\right| + \gamma\left|T_K(d_{\text{fix}})(\mathrm{P}(\cdot|\bar{s},a),\mathrm{P}(\cdot|s,a))\right|\right)$$
$$+\gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s'))-V^*(s'))\right|$$
$$\leq \frac{1}{|\mathcal{N}(s)|(1-\gamma)}\sum_{\bar{s}\in\mathcal{N}(s)} d_{\text{fix}}(s,\bar{s}) + \gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)(V^*_\phi(\phi(s'))-V^*(s'))\right|$$
$$\leq \frac{1}{|\mathcal{N}(s)|(1-\gamma)}\sum_{\bar{s}\in\mathcal{N}(s)}\epsilon_B + \gamma\left|\frac{1}{|\mathcal{N}(s)|}\sum_{\bar{s}\in\mathcal{N}(s)}\sum_{s'\in\mathcal{S}}\mathrm{P}(s'|\bar{s},a)\frac{\epsilon_B}{(1-\gamma)^2}\right|$$
$$= \frac{\epsilon_B}{1-\gamma}+\frac{\epsilon_B\gamma}{(1-\gamma)^2} = \frac{\epsilon_B}{(1-\gamma)^2}$$

Then, $\max_a Q^*(s,a) - Q^*(s,\pi_{\phi,N}(s)) \leq \frac{2\epsilon_B}{(1-\gamma)^2}$. Applying PDL gives:

$$V^*(s_0) - V^{\pi_{\phi,N}}(s_0) \leq \frac{2\epsilon_B}{(1-\gamma)^3} \tag{20}$$

□

Remark 3. *As a consequence of Remark 2, we can use Theorem 1 to derive an alternative to the bound in Equation* (13)*:*

$$\|V^* - V^{\pi_{\phi,N}}\| \leq \frac{2(\gamma+\epsilon_B)}{(1-\gamma)^2}. \tag{21}$$

Note that Equation (13) *is the tighter bound if and only if:*

$$\begin{aligned} &\frac{2(\gamma+\epsilon_B)}{(1-\gamma)^2} > \frac{2\epsilon_B}{(1-\gamma)^3} \\ \iff &(1-\gamma)\gamma + (1-\gamma)\epsilon_B > \epsilon_B \\ \iff &1-\gamma > \epsilon_B. \end{aligned} \tag{22}$$

This suggests for fine-grained abstractions with smaller ϵ_B, Equation (13) *is the tighter bound. Finally, note that while γ and ϵ_B are not independent for a fixed ϕ, their relation depends on the MDP and how the reward and transition terms relate to each other.*

4.1 Numerical Experiments and Discussion

As can be seen from Theorem 2, the upper bound on suboptimality for the "smart refinement" is smaller than the "naive refinement" by a factor of γ. A natural question is whether this relation between suboptimality bounds carry over to the comparison of actual suboptimalities.

In order to test this question, we randomly construct 50 Garnet MDPs [1] each with 30 states, 3 actions, branching factor of 12, and discount factor $\gamma = 0.95$. Then, we compute a 0.07-approximate bisimulation abstraction for each, reducing the number of states by a factor of ~ 4. Finally, we compare value of refined abstract policies between smart and naive refinements.

Our results indicate that the naive refinement yields a mean suboptimality of 1.553, corresponding to a $\sim 10\%$ deterioration in performance. In contrast, the mean suboptimality of the smart refinement is 0.056, corresponding to only $\sim 0.4\%$ decrease in the performance compared to the true optimal policy.

Notice that the gap between the performance of two refinement strategies is much larger than the γ factor suggested by the theoretical bounds. We attribute the large performance gap between refinement policies to the *action gap phenomenon* [2], which is often used to explain why RL agents can perform well despite imperfect Q-value estimates. The phenomenon highlights that the value of the optimal action is typically much larger than that of the second-best action. As a result, even substantial errors in action-value estimates often lead to the same greedy action. Under smart refinement, the error stems primarily from value function approximation—specifically, using the abstract optimal value function instead of the true optimal one. In contrast, naive refinement compounds both dynamics errors and value function errors, leading to significantly worse performance.

5 Conclusions

In this paper, we studied the relation between bisimulation and model irrelevance abstractions and their implications for suboptimality in reinforcement learning. We showed that model irrelevance always induces bisimulation abstractions, but not vice versa, highlighting that suboptimality bounds for the former cannot be directly applied to the latter. We then derived the first suboptimality guarantees for policies obtained from approximate bisimulation abstractions, distinguishing between naive and smart refinement strategies. Our analysis revealed that smart refinement achieves tighter bounds—by a factor of γ—and our experiments confirmed that this improvement is not only theoretical but also leads to markedly smaller suboptimality in practice. Finally, we argued that the action gap phenomenon provides an intuitive explanation for these results, shedding light on why RL agents often perform well despite approximation errors. Overall, our preliminary results underscore the importance of both the abstraction method and the refinement strategy in achieving reliable performance when leveraging bisimulation in RL. For future work, we are interested in understanding the practical implications of these insights not only for bisimulations but other model reduction and approximation techniques.

6 Acknowledgement

This work was supported in part by ONR CLEVR-AI MURI (#N00014-21-1-2431). We also thank Professor Nan Jiang for helpful feedback and comments on an earlier version of this work.

References

[1] Thomas Welsh Archibald, Ken I.M. McKinnon, and Lyn C. Thomas. On the generation of markov decision processes. *Journal of the Operational Research Society*, 46(3):354–361, 1995.

[2] Amir-massoud Farahmand. Action-gap phenomenon in reinforcement learning. *Advances in neural information processing systems*, 24, 2011.

[3] Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. 2004.

[4] Philippe Hansen-Estruch, Amy Zhang, Ashvin Nair, Patrick Yin, and Sergey Levine. Bisimulation makes analogies in goal-conditioned reinforcement learning. In *International Conference on Machine Learning*, pages 8407–8426. PMLR, 2022.

[5] Nan Jiang. Notes on tabular methods, 2020.

[6] Sham M. Kakade. *On the sample complexity of reinforcement learning*. PhD thesis, University College London, 2003.

[7] Mete Kemertas and Tristan Aumentado-Armstrong. Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems*, 34:4764–4777, 2021.

[8] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006.

[9] Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020.