Leveraging Factored Action Spaces for Off-Policy Evaluation

Aaman Rebello [1] **Shengpu Tang** [2] **Jenna Wiens** [2] **Sonali Parbhoo** [1]

Abstract

In high-stakes decision-making domains such as healthcare, off-policy evaluation (OPE) can help practitioners understand the performance of a new policy before deploying it. However, existing OPE estimators often exhibit high bias and high variance in problems involving large, combinatorial action spaces. We investigate how to mitigate this problem using factored action spaces i.e. expressing each action as a combination of independent sub-actions from smaller action spaces. We propose and study a new family of "decomposed" importance sampling (IS) estimators based on factored action spaces. Given certain assumptions on the underlying problem structure, we prove that the decomposed IS estimators have less variance than their original non-decomposed versions, while preserving the property of zero bias. This results in lower mean squared error. Through simulations, we empirically verify our theoretical results, probing the validity of various assumptions. Provided with a technique that can derive the action space factorisation for a given problem, our work shows that OPE can be improved "for free" by utilising this inherent problem structure.

1. Introduction

In recent years, reinforcement learning (RL) has made significant advances with applications in domains such as games (Silver et al., 2016), robotics (Kober et al., 2013), ride-sharing (Guo & Xu, 2020) and autonomous driving (Mueller, 2017), to mention a few. Yet in high-stakes domains such as healthcare where real-world testing is infeasible, unethical, or expensive, we must estimate the expected utility of a decision-making policy based on batch data collected under a different policy. This is known as off-policy evaluation (OPE).

Performing OPE in real-world applications is challenging; the decisions we wish to evaluate may differ from those observed in the batch of data available for evaluation, resulting in poor sample overlap. This causes high variance in the estimates of a policy's performance and low data sample efficiency i.e. more samples are needed to guarantee a particular level of variance. Meanwhile, combinatorial action spaces are common for many applications: e.g., in healthcare, an action may correspond to combinations of different drugs or treatments (Parbhoo et al., 2017; Komorowski et al., 2018; Prasad et al., 2017). For these problems, viewing each combination of "sub-actions" as a unique action results in an exponentially large action space, where only a few actions correspond to those in our batch data, thereby exacerbating the large variance problem in OPE estimates.

In this paper, we develop a new approach for OPE that leverages the idea of factored action spaces (Tang et al., 2022). The intuition is that though it may be difficult to ensure overlap between the policy we wish to evaluate and the data-generating policy, each of these policies may instead be decomposed into factors over actions, making it easier to ensure overlap among the policies with similar factors. In doing so, we improve sample efficiency and obtain OPE estimates with better bias and variance guarantees. Based on this intuition, we define a family of decomposed importance sampling (IS) estimators that account for the structures induced by action space factorisation. Our contributions are as follows:

- We introduce a family of decomposed IS estimators that leverage factored action spaces for performing off-policy evaluation (OPE).
- We prove theoretically that under certain assumptions, our decomposed estimators are unbiased and have lower variance than standard OPE estimators.
- We demonstrate empirically that decomposed OPE estimators have lower variance and similar bias to existing OPE baselines, while having better data sample efficiency, as measured by effective sample size (ESS).

[1]Department of Engineering, Imperial College London, London, UK [2]Division of Computer Science & Engineering, University of Michigan, Ann Arbor, Michigan, USA. Correspondence to: Aaman Rebello <aaman.rebello18@imperial.ac.uk>, Sonali Parbhoo <s.parbhoo@imperial.ac.uk>.

Workshop on New Frontiers in Learning, Control, and Dynamical Systems at the International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA, 2023.

2. Related Work

Off-Policy Evaluation Methods. Standard methods for OPE struggle when dealing with combinatorial action spaces. Direct methods (DM) use a model of the environment to simulate trajectories to compute the value of a policy (Paduraru, 2013; Chow et al., 2015; Hanna et al., 2017; Fonteneau et al., 2013; Liu et al., 2018); learning an accurate model of combinatorial actions from offline data is difficult because some combinations may not be observed. In general, DM can suffer from large bias where the actions chosen by the evaluation policy differ significantly from the behaviour policy. Importance sampling (IS) estimators i.e. inverse propensity score (IPS)-based estimators use reweighting to correct the sampling bias in off-policy data, such that an unbiased estimator may be obtained (e.g. Precup et al. (2000); Horvitz & Thompson (1952); Thomas & Brunskill (2016)). Unfortunately, as the size of the action space increases, these methods suffer from poor sample overlap, resulting in value estimates with high variance. Doubly robust (DR) estimators (e.g. Jiang & Li (2016); Farajtabar et al. (2018)) combine DM and IS; these can be low variance if either DM or IS is accurate. DR estimators can be used along with our approach; however, *they do not offer a way to explicitly factorise action spaces* to improve overlap as we propose here. Keramati et al. (2021) analyse how identifying subgroups with similar benefits when performing OPE can produce more reliable estimates. In contrast, we focus on explicitly *factorising combinatorial action spaces* to reduce variance of the estimate.

Factored Actions in RL. Tang et al. (2022) proposed a linear function decomposition to express the Q-function based on factored action spaces. The authors provide several theoretical guarantees of when the approach can lead to unbiased and low-variance *on-policy* estimation of the Q-function i.e. where the policy generating the data is being evaluated. They illustrated the bias-variance trade-off in offline RL. Other work, such as Tavakoli et al. (2020); Sunehag et al. (2017); Zhou et al. (2019) leveraged action factorisation for improved exploration, handling multiple agents or multiple rewards, primarily in online RL. Unlike these works, we explicitly focus on the task of *off-policy evaluation* where we are given offline data under a particular behaviour policy but would like to evaluate the performance of a different policy.

We share a similar premise with these prior works in assuming the problem comes with factored action spaces. We also utilise the theory from Tang et al. (2022) as a starting point. However, the OPE setting brings about new challenges and complexities in the theoretical guarantees and assumptions, which are the focus of this work. Since OPE methods can suffer from high variance, there is also more scope for variance reduction using factored action spaces.

3. Preliminaries

Markov Decision Processes. We consider Markov decision processes (MDPs) defined by a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, r, d_1, \gamma, T)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ and $r : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ are the transition and reward functions, $d_1 \in \Delta(\mathcal{S})$ is the initial state distribution, $\gamma \in [0, 1]$ is the discount factor, $T \in \mathbb{Z}^+$ is the fixed horizon. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies a mapping from each state to a probability distribution over actions. A T-step trajectory following policy π is denoted by $\tau = [(s_t, a_{t+1}, r_t, s_{t+1})]_{t=1}^{T}$ where $s_0 \sim d_1, a_{t+1} \sim \pi(s_t), r_t \sim r(s_t, a_{t+1}), s_{t+1} \sim p(s_t, a_{t+1})$. Here, $a \sim \pi(s)$ is short for $a \sim \pi(\cdot|s)$ and $s' \sim p(s, a)$ for $s' \sim p(\cdot|s, a)$. Let $J = \sum_{t=1}^{T} \gamma^{t-1} r_t$ denote the return of the trajectory, which is the discounted sum of rewards. The value a of policy π, denoted by $V_\pi : \mathcal{S} \to \mathbb{R}$, maps each state to the expected return starting from that state following policy π. That is, $V_\pi(s) = \mathbb{E}_\pi[J|s_1 = s]$. Similarly, the action-value function (i.e., the Q-function), $Q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined by further restricting the action taken from the starting state $Q_\pi(s, a) = \mathbb{E}_\pi[J|s_1 = s, a_1 = a]$. The performance of a policy (i.e., the value of policy π) is defined as the expected value over initial states, $V_\pi = \mathbb{E}_{s\sim d_1}[V_\pi(s)]$.

Off-Policy Evaluation. In OPE, we are given a dataset of T-step trajectories $\mathcal{D} = \{\tau^{(n)}\}_{n=1}^{N}$ each independently generated by some *behaviour policy* π_b. We aim to produce a good estimate of V_{π_e} or Q_{π_e}, the performance of a different policy, π_e, known as the *evaluation policy*. In general, the estimator $\hat{V}_{\pi_e}$ or $\hat{Q}_{\pi_e}$ is good if it achieves low mean squared error (MSE),

$$MSE(Q_{\pi_e}, \hat{Q}_{\pi_e}) = \mathbb{E}_{P^{\tau}_{\pi_b}}[(Q_{\pi_e} - \hat{Q}_{\pi_e})^2], \qquad (1)$$

where $P^{\tau}_{\pi_b}$ denotes the distribution of trajectory τ under behaviour policy π_b.

Factored Action Spaces in MDPs. While the standard MDP definition does not consider the underlying structure within action space $\mathcal{A}$, we follow Tang et al. (2022) and explicitly express a factored action space $\mathcal{A}$ as a Cartesian product of D sub-action spaces $\mathcal{A}^d$, with $d \in \{1, \cdots, D\}$. Formally, $\mathcal{A} = \otimes_{d=1}^{D} \mathcal{A}^d$. Accordingly, each action can be written as a vector of sub-actions, $a = [a^1, \cdots, a^D]$. The key insight of Tang et al. (2022) is that the Q-function (of certain policies) can be additively decomposed in terms of the factored action spaces:

$$\tilde{Q}_\pi(s, a) = \sum_{d=0}^{D} \hat{Q}^d_\pi(s, a^d) \qquad (2)$$

Sufficient conditions (on the MDP and policy π) for Equation 2 to be valid are outlined in Theorem 1. It should be noted that while not guaranteed, Equation 2 can hold even when the sufficient conditions are violated.

Theorem 1 (adapted from Theorem 1 of Tang et al. (2022))**.** *Equation 2 holds if, for the MDP, the following conditions hold:*

$$\sum_{\tilde{s} \in \phi^{-1}(\phi(s'))} p(\tilde{s}|s,a) = \prod_{d=1}^{D} p^d((z')^d|z^d,a^d) \quad (3)$$

$$r(s,a) = \sum_{d=1}^{D} r^d(z^d,a^d), \quad (4)$$

where p^d and r^d are sub-transitions and sub-rewards corresponding to sub-action space $\mathcal{A}^d$, and $\phi(s) = [z^1, \cdots, z^D]$, such that each z^d is an abstraction of s with respect to sub-action space $\mathcal{A}^d$. Additionally, for policy π, the following should hold:

$$\pi(a|s) = \prod_{d=1}^{D} \pi^d(a^d|z^d) \quad (5)$$

where each π^d is a sub-policy corresponding to sub-action space $\mathcal{A}^d$.

Below, we state the assumptions that will be used in our theoretical analyses. The first regularity assumption is a relaxed version of the standard regularity assumption in OPE, where absolute continuity is only required and assumed over factors of the action space rather than the entire action space. The second assumption states that our policies do not change with time.

Assumption 1. (Absolute Continuity). For all $(s, a^d) \in \mathcal{S} \times \mathcal{A}^d$, if $\pi_b(a^d|s) = 0$ then $\pi_e(a^d|s) = 0$.

Assumption 2. (All Policies are Stationary). The probability distributions of a policy do not change with time within the trajectory or the number of trajectories completed.

Additionally, to prove bounds on the variance of the decomposed estimators, we make the following assumptions:

Assumption 3. (Conditions for Variance Bounds on Decomposed Estimators). The following conditions hold $\forall\, d, d', t, t', n$, where $d' \neq d$ and $t' \neq t$:

$$Cov\left(r^d(z_t^{(n),d}, a_t^{(n),d}),\ r^{d'}(z_{t'}^{(n),d'}, a_{t'}^{(n),d'})\right) \geq 0 \quad (6)$$

$$Cov\left((\rho_{0:T}^{(n),d}),\ (\rho_{0:T}^{(n),d'})\right) = 0 \quad (7)$$

$$Cov\left((r^d(z_t^{(n),d}, a_t^{(n),d}),\ (\rho_{0:T}^{(n),d'})\right) = 0 \quad (8)$$

Here, $\rho_{0:T}^{(n),d} = \prod_{t'=0}^{T} \frac{\pi_e^d(a_{t'}^{(n),d}|z_{t'}^{(n),d})}{\pi_b^d(a_{t'}^{(n),d}|z_{t'}^{(n),d})}$ is the importance sampling weight specifically for the abstracted trajectory $[(z_t^{(n),d}, a_t^{(n),d}, z_{t+1}^{(n),d})]_{t=1}^{T}$ corresponding to $\mathcal{A}^d$. $Cov(X, Y)$ denotes the covariance between random variables X and Y.

These conditions require that the rewards at different times t and IS ratios in different factored action spaces d are all uncorrelated. The rewards at different t and d can be positively correlated i.e. the rewards increase together. Finally, an extra assumption is needed to prove the variance bound of one of the decomposed estimators (decomposed PDWIS):

Assumption 4. (Additional Variance Bound Condition) $\forall\, d, t, t', n$, where $t' \neq t$:

$$Cov\left(\rho_{0:t}^{(n),d},\ \rho_{0:t'}^{(n),d}\right) \geq 0 \quad (9)$$

This assumption imposes that importance sampling weights in any given factored action space and episode should be non-negatively correlated when calculated on sub-trajectories of different lengths.

4. Method

We seek to leverage factored action spaces to improve the bias and variance guarantees of OPE estimators; in this work, we focus on the IS-based estimators. To do this, we utilise our knowledge that the action space can be expressed in terms of smaller, independent sub-action spaces. The overall intuition of our approach is that while it may be difficult to ensure overlap between behaviour and evaluation policies, by decomposing both behaviour and evaluation policies into factors over actions, we may be able to ensure overlap among the policies with similar factors. In doing so, we may be able to improve sample efficiency and obtain OPE estimates with better bias and variance guarantees.

To begin, we impose that the conditions in Theorem 1 hold, i.e., Eqs. (2) to (5) hold. Applying these equations allows us to derive new versions of the IS estimators defined in Precup et al. (2000). These new estimators leverage the factorisation structure of the action spaces, and are presented in Definitions 1 and 2 below.

Definition 1. (Decomposed IS and Decomposed Per-Decision IS) The decomposed IS estimator is defined as

$$\tilde{Q}_{\pi_e}^{DecIS} = \sum_{d=0}^{D} \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \quad (10)$$

where $\rho_{0:T}^{(n),d}$ is defined in Assumption 3. The decomposed Per-Decision IS (PDIS) estimator $\tilde{Q}_{\pi_e}^{DecPDIS}$ is defined by replacing $\rho_{0:T}^{(n),d}$ with $\rho_{0:t}^{(n),d}$ i.e. IS weights up to each time step t.

By Eq. (2), we can calculate an IS estimate for each d i.e. an estimate factor and take the sum. Utilising Eq. (5), we define the IS weights ρ^d for each estimate factor. To calculate the estimate factor for each d, we utilise r^d based on Eq. (4).

Similarly, we can also define weighted variants of our new decomposed estimator:

Definition 2. (Decomposed PDWIS) The decomposed Per-Decision Weighted IS estimator based on factored action spaces is given by, where $r_t^{(n),d} \equiv r^d(z_t^{(n),d}, a_t^{(n),d})$,

$$\hat{Q}^{DecPDWIS}_{\pi_e} = \sum_{d=0}^{D} \frac{\sum_{n=1}^{N}\sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r_t^{(n),d}}{\sum_{n=1}^{N}\sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}} \quad (11)$$

The PDWIS estimator is derived from PDIS by dividing the IS weighted discounted reward sum by the sum of the IS weights instead of N. We note that the decomposed PDIS estimator is a sum of PDIS estimators, one for each d. We convert these PDIS estimators to PDWIS as discussed. The sum of these sub-PDWIS estimators gives $\hat{Q}^{DecPDWIS}_{\pi_e}$.

We next utilise Eqs. (2) to (5) to derive theoretical guarantees on the bias and variance of these decomposed estimators, set out in Theorems 2 and 3.

Theorem 2. *When the assumptions in Theorem 1 hold, the decomposed IS estimator $\hat{Q}^{DecIS}_{\pi_e}$ and decomposed PDIS estimator $\hat{Q}^{DecPDIS}_{\pi_e}$ are unbiased estimators of the true Q-function Q_{π_e}.*

Proof Sketch. The decomposed IS estimator in Eq. (10) is a sum of IS estimators, one for each Q-function factor $Q^d_{\pi_e}$ for the factored MDP and factored policy π_e^d corresponding to $\mathcal{A}^d$. Each of these IS estimators is unbiased, as this is a property of an IS estimator (Precup et al., 2000). Since Eq. (2) holds, we can sum these unbiased estimates to get an unbiased estimate of the overall Q-function. A similar argument can be applied for decomposed PDIS. The full proof is provided in Appendix C.

Theorem 3. *The decomposed IS and PDIS estimators are guaranteed to have at most the variance of their respective non-decomposed equivalent estimators i.e. $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}] \leq \mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ and $\mathbb{V}_{\pi_b}[\hat{Q}^{DecPDIS}_{\pi_e}] \leq \mathbb{V}_{\pi_b}[\hat{Q}^{PDIS}_{\pi_e}]$, provided that Theorem 1 and the conditions in Assumption 3 hold. For $\mathbb{V}_{\pi_b}[\hat{Q}^{DecPDWIS}_{\pi_e}] \leq \mathbb{V}_{\pi_b}[\hat{Q}^{PWDIS}_{\pi_e}]$, Assumption 4 must hold in addition to the mentioned conditions.*

Proof Sketch. Given the variance expressions of IS, PDIS and PDWIS estimators, and their decomposed versions, we use Eqs. (4) and (5) to write all of them in terms of the sub-actions a^d and sub-rewards r^d. Between these expressions, we identify comparable corresponding terms and we utilise guarantees from these comparisons to compare the overall variance expressions. The full proof is in Appendix D.

When Assumption 3 does not hold, there is no guarantee on the variance of the decomposed estimators, as interactions between r_t^d and ρ^d across t and d can in some cases lead to the variance of the decomposed estimator being more than its non-decomposed counterpart. It should also be noted that there is no guarantee that the variance of a decomposed estimate will always scale at a slower rate than its non-decomposed counterpart with length of the trajectory or mismatch between policies.

An important result of Theorem 3 relates to the effective sample size (ESS) of an IS (or PDIS or PDWIS) estimator $\hat{Q}^{IS}_{\pi_e}$, which is a measure of data sample efficiency:

$$ESS[\hat{Q}^{IS}_{\pi_e}] = N \times \frac{\mathbb{V}_{\pi_e}[\hat{Q}^{on\ policy}_{\pi_e}]}{\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]} \quad (12)$$

where both variances are calculated on the same number of samples and $\hat{Q}^{on\ policy}_{\pi_e}$ is a Q-function estimate based on data generated by π_e. $\mathbb{V}_{\pi_e}[\hat{Q}^{on\ policy}_{\pi_e}]$ stays the same for different IS estimators, hence the ESS is inversely proportional to $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$. Clearly then, Theorem 3 states that a decomposed estimator has a higher ESS i.e. it is more efficient than its non-decomposed version.

When the sufficient conditions hold, factored action spaces can reduce the variance of IS estimators without increasing bias. A qualitative explanation is that each factored action space has fewer possible actions than the overall action space. In each factored space, the distributions π_b and π_e are likely to overlap more i.e. greater coverage. Eq. (2) also means the Q-function can be calculated for each factored action space and then summed to give an unbiased estimate of Q_{π_e}. Thus, variance is reduced without affecting bias.

When the conditions in Theorem 1 are relaxed, Eq. (2) may not hold, causing the decomposed estimators to be biased or to have higher variance. One way to address this is to re-group the factored action spaces into larger factored action spaces that satisfy Theorem 1. For example, if we grouped factored spaces as $\mathcal{A}^{d_{\{1,2,3\}}} = \mathcal{A}^{d_1} \cup \mathcal{A}^{d_2} \cup \mathcal{A}^{d_3}$, we can define the reward factor $r^{d_{\{1,2,3\}}}(z^{d_1}, z^{d_2}, z^{d_3}, a^{d_1}, a^{d_2}, a^{d_3})$ and IS ratio $\rho_{0:T}^{d_{\{1,2,3\}}}$ based on policy factors $\pi_b^{d_{\{1,2,3\}}}(a^{d_1}, a^{d_2}, a^{d_3}|z^{d_1}, z^{d_2}, z^{d_3})$ and $\pi_e^{d_{\{1,2,3\}}}(a^{d_1}, a^{d_2}, a^{d_3}|z^{d_1}, z^{d_2}, z^{d_3})$. Thus, we can re-obtain the guarantees on bias and variance, at the cost of decreasing overlap between π_b and π_e. An illustration is in Appendix E. If all the factored action spaces were grouped together, we would get the original, non-decomposed IS estimator operating in the full action space $\mathcal{A}$.

5. Experimental Setup

We empirically validated Theorems 2 and 3 through code simulations of two MDP's specified below. Simultaneously, we investigated how varying N, T, divergence between π_b and π_e, and relaxing Theorem 1 affected the relative performance of the decomposed and non-decomposed estimators. The code is publicly available at https://github.com/ai4ai-lab/Factored-Action-Spaces-for-OPE.

MDP-1. This MDP is the 2-dimensional bandit problem from Tang et al. (2022). A diagram of the MDP is provided in Appendix F. $\mathcal{S} = \{state, terminal\}$, $\mathcal{A} = \{up\_right,\ up\_left,\ down\_right,\ down\_left\}$. Every action taken from $state$ or $terminal$ always leads to $terminal$. In terms of rewards, $r(state,\ up\_right) = 1 + \alpha + \beta$, $r(state,\ up\_left) = \alpha$, $r(state,\ down\_right) = 1$, where $\alpha = 1$ and $\beta = 0$. Otherwise, $r = 0$. The start state is always $state$. $T = 1$ always, hence γ is irrelevant. The MDP satisfies Theorem 1 for $\beta = 0$ and can be factored into two sub-action spaces as described in Appendix F.

MDP-2. This 4-state MDP is inspired by Figure 2(a) from Tang et al. (2022). A diagram is provided in Appendix G. $\mathcal{S} = \{0,0\ ,\ 0,1\ ,\ 1,0\ ,\ 1,1\}$, $\mathcal{A} = \{up\_right,\ up\_left,\ down\_right,\ down\_left\}$. P is defined such that always, up_right leads to $1,1$, $up, left$ to $0,1$, $down\_right$ to $1,0$ and $down\_left$ to $0,0$. The reward structure is complex and described in Appendix G. The start state is always $0,0$. Both T and γ are varied, as described in the training details. The MDP satsfies Theorem 1 and factors into two action spaces as described in Appendix G.

Each MDP has been constructed by combining two smaller MDP's with independent dynamics. This suggests that the rewards and IS ratios in different factored action spaces would be uncorrelated, thus satisfying Assumption 3. It is also expected that Assumption 4 is satisfied.

Metrics. The following metrics characterise experimental configurations and performance of the OPE estimators:

Bias: The bias of OPE estimator $\hat{Q}_{\pi_e}$ is the difference between the expected value of the estimator over all possible observed datasets generated by π_b and the true value Q_{π_e}. It measures how accurate the estimator is on average.

$$Bias(\hat{Q}_{\pi_e}, Q_{\pi_e}) = \mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}] - Q_{\pi_e} \tag{13}$$

Variance: The variance of estimator $\hat{Q}_{\pi_e}$ is given in Eq. (14). It measures how precise the estimator is.

$$\mathbb{V}_{\pi_b}[\hat{Q}_{\pi_e}] = \mathbb{E}_{\pi_b}[(\hat{Q}_{\pi_e})^2] - \mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}]^2 \tag{14}$$

Mean Squared Error (MSE): In addition to Eq. (1), MSE can be calculated via the expression below, which combines bias and variance into a single metric measuring the accuracy of the OPE estimates:

$$MSE(Q_{\pi_e}, \hat{Q}_{\pi_e}) = Bias(Q_{\pi_e}, \hat{Q}_{\pi_e})^2 + \mathbb{V}[\hat{Q}_{\pi_e}] \tag{15}$$

Effective Sample Size (ESS): This is defined in Eq. (12), assuming that N data samples are taken from π_b for the off-policy estimator to make its estimate. The ESS reflects the equivalent number of *on-policy* samples required from π_e to obtain the same variance. The higher the ESS, the more sample efficient the estimator is.

Policy Divergence: This metric was defined by Voloshin et al. (2019) to quantify the difference between π_b and π_e in the context of an OPE problem:

$$PD(\pi_b, \pi_e) = \Big(\sup_{a \in \mathcal{A}, s \in \mathcal{S}} \frac{\pi_e(a|s)}{\pi_b(a|s)} \Big)^T \tag{16}$$

where T is the length of each trajectory. The minimum $PD(\pi_b, \pi_e)$ of 1 indicates $\pi_b = \pi_e$, while $PD(\pi_b, \pi_e)$ increases as the overlap between π_b and π_e decreases.

Training Details. The following OPE estimators were compared: IS, PDIS, PDWIS, decomposed IS (DecIS), decomposed PDIS (DecPDIS) and decomposed PDWIS (DecPDWIS). Additionally, an on-policy estimate of the Q-function was recorded in each experiment.

For each MDP, we generated two datasets: $\mathcal{D}_{\pi_e}$ following π_e, and $\mathcal{D}_{\pi_b}$ following π_b. The $\mathcal{D}_{\pi_e}$ was used to calculate the *on-policy* estimates, which represent the true value that the OPE estimators are estimating. $\mathcal{D}_{\pi_b}$ was used by the OPE estimators. To allow experimentation with a wide range of N and (for MDP-2) T, $\mathcal{D}_{\pi_e}$ and $\mathcal{D}_{\pi_b}$ each consisted of $10,000,000$ trajectories of length 1 for MDP-1, and $100,000$ trajectories of length 1000 for MDP-2. For values of N and T smaller than these sizes, multiple subsets of $\mathcal{D}_{\pi_b}$ and $\mathcal{D}_{\pi_e}$ could be taken. Whenever the policies and MDP rewards were varied, new datasets had to be generated.

To empirically measure bias, variance, and MSE, we generated $\mathrm{R} = 100$ subsets of $\mathcal{D}_{\pi_b}$ and $\mathcal{D}_{\pi_e}$ for R estimates from each estimator in each experiment. From these, the MSE was calculated for each OPE estimator using the definition in Eq. (1), where the true value was the on-policy estimate in the experiment. The variance of the R estimates was also calculated for each estimator; this allowed us to find the bias by applying Eq. (15). To cross-check, the bias was also calculated with by applying Eq. (13), using the on-policy estimate as the true value. The bias of the on-policy estimator was always taken to be zero, hence its MSE is equal to its variance. The ESS of each estimator was calculated by Eq. (12) using the number of trajectories N, the variance of R OPE estimates and the variance of R on-policy estimates.

To generate each of our plots, we repeated the procedure in the above paragraph with five different $(\mathcal{D}_{\pi_e},\ \mathcal{D}_{\pi_b})$ pairs. All plotted values (denoted by '$\times$' in the figures) are mean values over these five trials. Our plots also display error bars at each plotted value; the one-sided size of each error bar is equal to the standard deviation over the five trials.

The tested values of each parameter in the experiments were: $N \in \{10,\ 50,\ 100,\ ...\ 100000\}$, $T \in \{1,\ 5,\ 10,\ 50,\ 100,\ 500,\ 1000\}$, $\gamma \in \{0.7,\ 0.9,\ 0.9999\}$ and $PD(\pi_b,\ \pi_e) \in \{1.44^T,\ 2.56^T,\ 3.61^T,\ 4.46^T,\ 5.64^T,\ 10.03^T,\ 22.53^T,\ 90.25^T,\ 361.0^T\}$ where we note that all the policy divergence values are raised to the value of T.

6. Results

Empirical Finding 1. *The decomposed estimators have lower variance and MSE than their corresponding non-decomposed versions for all tested values of N.*

In Figure 1, for all tested values of N, the decomposed estimators (DecPDIS, DecPDWIS) have less variance than their non-decomposed counterparts (PDIS, PDWIS, respectively). Since MDP-1 has only one transition, IS is equivalent to PDIS and is thus not shown. The variances of all estimators scale similarly with increasing N. Compared to the variance of PDIS, using the weighted version (as in PDWIS) offers a greater variance reduction than leveraging factored action spaces (as in DecPDIS), while applying both ideas (as in DecPDWIS) achieves the smallest variance, implying that these are two complementary sources of variance reduction. Finally, since all estimators have negligible bias, the MSE plot shows the same trend as the variance.

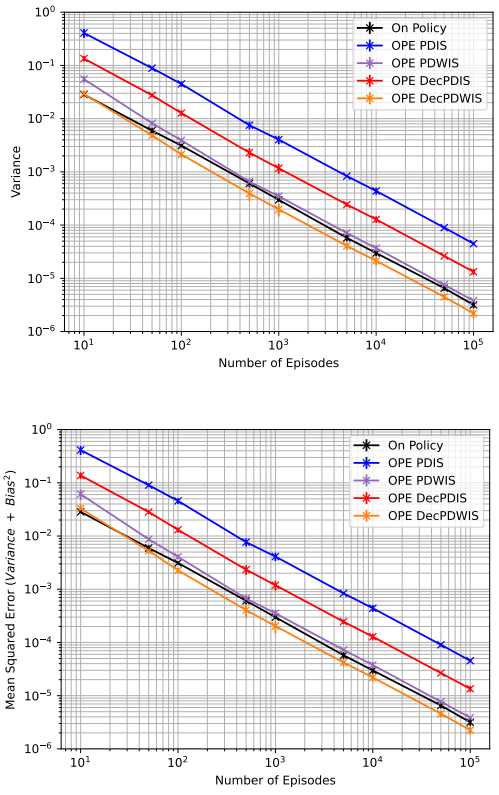

Figure 1. Variance (top) and MSE (bottom) of different OPE estimators as the number of episodes (N) varies, for MDP-1 with $PD(\pi_b, \pi_e) = 2.56$. The variances and MSEs of all estimators decrease with N, while maintaining a clear and consistent ordering among estimators.

Empirical Finding 2. *The decomposed IS and PDIS estimators have approximately zero bias as long as Theorem 1 holds and there is sufficient coverage of π_e by π_b.*

In the definition of MDP-1, β is part of the reward for taking action $up, right$ from $state$. When $\beta = 0$, it is possible to satisfy the condition in Eq. (4); and thus, satisfy Theorem 1. When $\beta \neq 0$, this condition cannot be satisfied; the more we increase $|\beta|$, the more strongly the condition is violated. In Figure 2, we compare the bias of the estimators for varying values of β. For $\beta = 0$, the bias of the decomposed estimators is comparable with the non-decomposed estimators i.e. $Bias(\hat{Q}_{\pi_e}, Q_{\pi_e}) \approx 0$. For $\beta \neq 0$, since the decomposed estimators implicitly assume $\beta = 0$ in their definitions, their bias increase rapidly as $|\beta|$ increases.

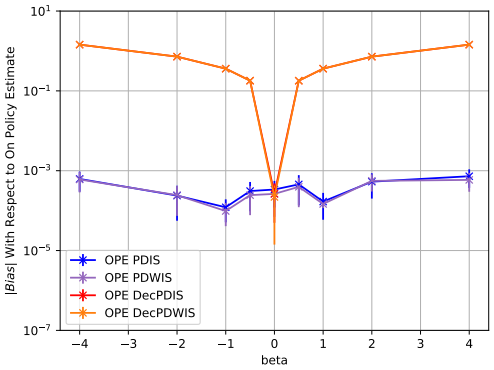

Figure 2. The magnitude of bias of different OPE estimators as β varies, for MDP-1 with $N = 100,000$, $PD(\pi_b, \pi_e) = 1.44$. While non-decomposed estimators have near-zero bias for all tested values of β as expected, decomposed estimators only have near-zero bias for $\beta = 0$ and increasing bias as $|\beta|$ increases.

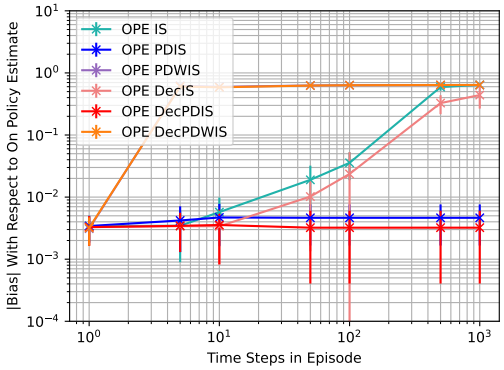

Figure 3. The magnitude of bias of different OPE estimators as the episode length (T) varies, for MDP-2 with $\gamma = 0.7$, $N = 1000$ and $PD(\pi_b, \pi_e) = 1.44^T$. Typically, the IS and PDIS estimators should always have near-zero bias, however as T increases, the divergence between π_b and π_e increases, resulting in loss of coverage of π_e by samples from π_b.

In Figure 3, the PDWIS estimators have higher bias than

IS and PDIS $\forall\, T$; this is expected, as these estimators are known to be biased (Precup et al., 2000). The PDIS estimators have $|Bias| \approx 0\ \forall\, T$, while IS has $|Bias| \approx 0$ for $T \leq 10$. For $T > 10$, the bias of IS scales rapidly. The reason is that as T increases, it is more likely that trajectories occurring under π_e would not be covered by the dataset sampled from π_b. This is particularly relevant to IS estimators, where IS weights are assigned to entire trajectories. Sachdeva et al. (2020) stated that in the absence of coverage, IS estimators are biased, with bias equal to the expected reward under π_e of following the non-covered trajectories. PDIS improves coverage by considering sub-trajectories IS weights at each time step, thereby lowering the bias. Incorporating factored action spaces (as in DecPDIS) further improves coverage and reduces bias $\forall\, T$.

Empirical Finding 3. *The variances of the decomposed estimators grow more slowly than their non-decomposed versions, with increasing T and/or $PD(\pi_b, \pi_e)$. This is because π_b and π_e overlap more in the sub-action spaces i.e. the decomposed estimators improve coverage.*

Figure 4 shows the variance and MSE for the same experiments in Figure 3. The variances of PDIS and IS estimators both scale with T, although the scaling is less pronounced for PDIS estimators. As illustrated in Appendices I.2.2 and I.2.3, the scaling becomes more pronounced when γ is increased. The decomposed estimators always have lower variance than their non-decomposed counterparts, sometimes scaling differently - this is true even for the PDWIS estimators where the variance decreases with T. The drop in variance of the IS estimator for large T is again explained by low coverage - for these cases, the IS estimator is so biased that it only estimates values close to zero, which results in low variance. Once T is large enough, all estimators would reach this state as they lose coverage of π_e.

In Figure 5, π_b and π_e were varied to adjust $PD(\pi_e, \pi_b)$. The exact policy configurations are discussed in Appendix H.2. Here, it is seen that all the variances initially scale with the policy divergence but, one-by-one, the IS and PDIS estimators lose coverage and drop in variance. This drop in variance is accompanied by a rise in bias. It is seen that the decomposed PDIS and IS scale the most slowly in bias, which indicates that they improve coverage.

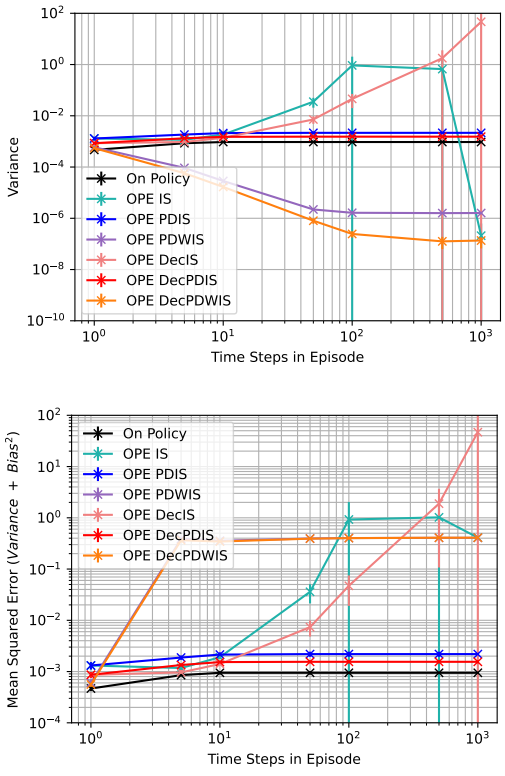

Figure 4. Variance (top) and MSE (bottom) of different OPE estimators as the episode length (T) varies, for MDP-2 with $\gamma = 0.7$, $N = 1000$ and $PD(\pi_e, \pi_b) = 1.44^T$. Comparing the speed of scaling of variance/MSE, we generally see that OPE IS $>$ OPE DecIS $>$ OPE PDIS $\approx$ OPE DecPDIS. PDWIS has the best variance scaling behavior i.e. negative scaling.

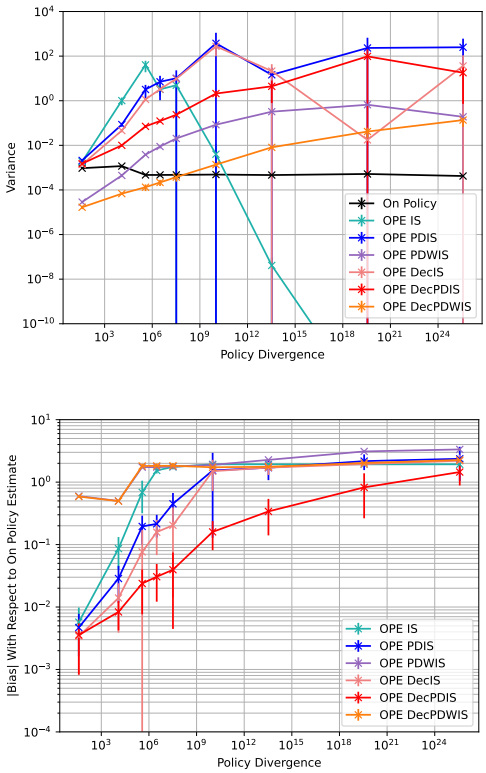

Figure 5. Variance (top) and MSE (bottom) of different OPE estimators as the policy divergence (PD) varies, for MDP-2 with $\gamma = 0.7$, $N = 1000$ and $T = 10$. As policy divergence increases, the coverage of π_e by the π_b data decreases, causing increases in bias (and reductions in variance). The PDWIS estimators were biased even when policy divergence is small.

Empirical Finding 4. *The decomposed estimators have higher ESS than their non-decomposed versions for most tested values of N and T. This implies that they have higher data sample efficiency.*

The ESS of an IS estimator, as defined in Eq. (12), is inversely proportional to the variance of the estimator. While it is also directly proportional to the number of trajectories N, most of the ESS graphs plotted appear as inverted variance graphs, as seen in Figures 6 and 7, which respectively correspond to Figures 1 and 4.

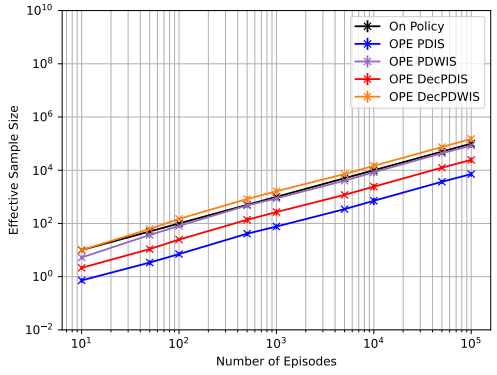

Figure 6. The effective sample size (ESS) of different OPE estimators as the number of episodes (N) varies, for MDP-1 with $PD(\pi_b, \pi_e) = 2.56$. The graph is almost an inverted version of the variance graph in Figure 1, showing the same precedence in estimators.

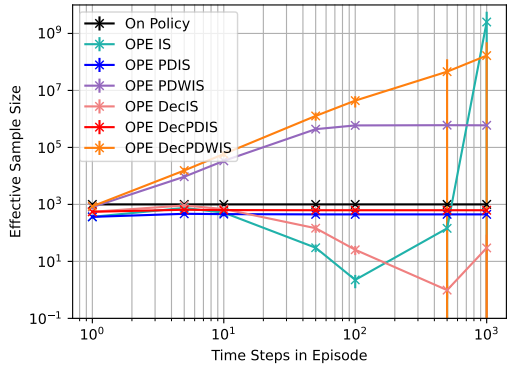

Figure 7. The effective sample size (ESS) of different OPE estimators as the episode length (T) varies, for MDP-2 with $\gamma = 0.7$, $N = 1000$ and $PD(\pi_b, \pi_e) = 1.44^T$. The graph is almost an inverted version of Figure 4. The IS estimators demonstrate how the ESS decreases with increasing T due to greater mismatch in trajectory trajectories from π_b and π_e.

An interesting insight is the high efficiency of the PDWIS estimators, especially the decomposed version, compared to others. For most values of N and T in Figure 6 and Figure 7, the decomposed PDWIS estimator has higher ESS than the on-policy estimate itself - this may be due to the advantages of factored actions and weighting, which are not available to the latter. Due to lower variance, the decomposed estimators have higher ESS than their non-decomposed equivalent estimators, which implies they use available data samples more efficiently.

7. Conclusion

In this paper, we study the role of factored action spaces in improving the variance, mean squared error and data sample efficiency of off-policy evaluation (OPE) without any increase in estimator bias. By proposing a new family of decomposed importance sampling (IS) based estimators that leverage the factorisation structure of actions, we have demonstrated theoretically and empirically the potential of leveraging factored action spaces to improve IS.

Of note, unlike Tang et al. (2022) where only *implicit* MDP factorisation is required, our decomposed IS estimators require *explicit* knowledge of the state abstractions, the reward factorisation and policy factorisations (for both behavior and evaluation policies), as seen in the definitions of DecPDIS and DecPDWIS (interestingly, we do not require knowledge of transition factorisations). Admittedly, an action space factorisation (along with the reward/policy factorisations) that satisfies the sufficient conditions we have outlined for variance reduction and/or zero bias may not always exist, or may be challenging to derive in practice. Future work should investigate the practicality of our approach; for example, using offline data to design the state abstraction, policy and reward factorisations based on action factorisations for general and/or specific OPE problems. We believe these are important avenues of future research before the decomposed estimators can be applied to real-life OPE problems. By enhancing OPE with domain knowledge about problem structures - in this case, how the action space can be factorised - our approach represents an important step to improving OPE for high-stakes decision-making domains such as healthcare, which are often characterized by complex and combinatorial action spaces.

References

Chow, Y., Petrik, M., and Ghavamzadeh, M. Robust policy optimization with baseline guarantees. *arXiv preprint arXiv:1506.04514*, 2015.

Farajtabar, M., Chow, Y., and Ghavamzadeh, M. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Fonteneau, R., Murphy, S. A., Wehenkel, L., and Ernst, D. Batch mode reinforcement learning based on the synthesis of artificial trajectories. *Annals of operations research*, 208:383–416, 2013.

Guo, G. and Xu, Y. A deep reinforcement learning approach to ride-sharing vehicle dispatching in autonomous mobility-on-demand systems. *IEEE Intelligent Transportation Systems Magazine*, 14(1):128–140, 2020.

Hanna, J., Stone, P., and Niekum, S. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Horvitz, D. G. and Thompson, D. J. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685, 1952.

Jiang, N. Notes on importance sampling and policy gradient, 2022. URL https://nanjiang.cs.illinois.edu/files/cs542f22/note6.pdf.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR, 2016.

Keramati, R., Gottesman, O., Celi, L. A., Doshi-Velez, F., and Brunskill, E. Identification of subgroups with similar benefits in off-policy policy evaluation. *arXiv preprint arXiv:2111.14272*, 2021.

Kober, J., Bagnell, J. A., and Peters, J. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Liu, Y., Gottesman, O., Raghu, A., Komorowski, M., Faisal, A. A., Doshi-Velez, F., and Brunskill, E. Representation balancing mdps for off-policy policy evaluation. *Advances in Neural Information Processing Systems*, 31, 2018.

Mueller, M. A. Reinforcement learning: Mdp applied to autonomous navigation. *Machine Learning and Applications: An International Journal*, 4(4):01–10, 2017.

Paduraru, C. *Off-policy evaluation in Markov decision processes*. PhD thesis, 2013.

Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V., and Doshi-Velez, F. Combining kernel and model based learning for hiv therapy selection. *AMIA Summits on Translational Science Proceedings*, 2017:239, 2017.

Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M., and Engelhardt, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. *arXiv preprint arXiv:1704.06300*, 2017.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pp. 759–766, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072.

Sachdeva, N., Su, Y., and Joachims, T. Off-policy bandits with deficient support. *CoRR*, abs/2006.09438, 2020. URL https://arxiv.org/abs/2006.09438.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Tang, S., Makar, M., Sjoding, M., Doshi-Velez, F., and Wiens, J. Leveraging factored action spaces for efficient offline reinforcement learning in healthcare. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Jd70afzIvJ4.

Tavakoli, A., Fatemi, M., and Kormushev, P. Learning to represent action values as a hypergraph on the action vertices. *arXiv preprint arXiv:2010.14680*, 2020.

Thomas, P. and Brunskill, E. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *CoRR*, abs/1911.06854, 2019. URL http://arxiv.org/abs/1911.06854.

Zhou, M., Chen, Y., Wen, Y., Yang, Y., Su, Y., Zhang, W., Zhang, D., and Wang, J. Factorized q-learning for large-scale multi-agent systems. In *Proceedings of the first international conference on distributed artificial intelligence*, pp. 1–7, 2019.

A. Derivation of the Decomposed IS Estimator

Let us consider the definition of the importance sampling (IS) estimator, which was introduced in Precup et al. (2000):

$$\hat{Q}^{IS}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}_{0:T}\sum_{t=0}^{T}\gamma^t \cdot r(s^{(n)}_t, a^{(n)}_t) \tag{17}$$

We first note that the number of trajectories N, the discount factor γ and the number of transitions T in each trajectory are all constants. They do not change with i, t or d. The remaining terms: r and ρ, do depend on i and/or t. Next, we apply Equation 4 to write:

$$\hat{Q}^{IS}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}_{0:T}\sum_{t=0}^{T}\gamma^t \sum_{d=1}^{D} r^d(z^{(n),d}_t, a^{(n),d}_t) \tag{18}$$

Since i, t and d do not depend on each other, we may make the summation over d the outer summation:

$$\hat{Q}^{IS}_{\pi_e} = \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}_{0:T}\sum_{t=0}^{T}\gamma^t \cdot r^d(z^{(n),d}_t, a^{(n),d}_t) \tag{19}$$

At this point, we may write:

$$\hat{Q}^{IS}_{\pi_e} = \sum_{d=1}^{D}\hat{Q}^{IS,d} \tag{20}$$

where:

$$\hat{Q}^{IS,d}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}_{0:T}\sum_{t=0}^{T}\gamma^t \cdot r^d(z^{(n),d}_t, a^{(n),d}_t) \tag{21}$$

In essence, we have decomposed $\hat{Q}^{IS}_{\pi_e}$ based on the factored action spaces, in the form of Equation 2. However, each estimate $\hat{Q}^{IS,d}_{\pi_e}$ has the full $\rho^{(n)}_{0:T}$ and we now examine whether this is necessary. To do this we observe the definition of $\rho^{(n)}_{0:T}$ in Equation 17:

$$\rho^{(n)}_{0:T} = \prod_{t=0}^{T}\frac{\pi_e(a^{(n)}_{t+1}|s^{(n)}_t)}{\pi_b(a^{(n)}_{t+1}|s^{(n)}_t)} = \frac{\pi_e(a^{(n)}_1|s^{(n)}_0)\cdot\pi_e(a^{(n)}_2|s^{(n)}_1)\ ...\ \pi_e(a^{(n)}_{T+1}|s^{(n)}_T)}{\pi_b(a^{(n)}_1|s^{(n)}_0)\cdot\pi_b(a^{(n)}_2|s^{(n)}_1)\ ...\ \pi_b(a^{(n)}_{T+1}|s^{(n)}_T)} \tag{22}$$

We may apply Equation 5 to the numerator and denominator to write:

$$\rho^{(n)}_{0:T} = \prod_{t=0}^{T}\frac{\pi_e(a^{(n)}_{t+1}|s^{(n)}_t)}{\pi_b(a^{(n)}_{t+1}|s^{(n)}_t)} = \prod_{t=0}^{T}\frac{\prod_{d=1}^{D}\pi^d_e(a^{(n),d}_{T+1}|z^{(n),d}_T)}{\prod_{d=1}^{D}\pi^d_b(a^{(n),d}_{T+1}|z^{(n),d}_T)} \tag{23}$$

$$= \frac{\pi^1_e(a^{(n),1}_1|z^{(n),1}_0)...\pi^D_e(a^{(n),D}_1|z^{(n),D}_0)...\pi^1_e(a^{(n),1}_{T+1}|z^{(n),1}_T)...\pi^D_e(a^{(n),D}_{T+1}|z^{(n),D}_T)}{\pi^1_b(a^{(n),1}_1|z^{(n),1}_0)...\pi^D_b(a^{(n),D}_1|z^{(n),D}_0)...\pi^1_b(a^{(n),1}_{T+1}|z^{(n),1}_T)...\pi^D_b(a^{(n),D}_{T+1}|z^{(n),D}_T)}$$

The continued products in the numerator and denominator contain terms generated by the policies π^e_d and π^b_d corresponding to each action space where $\mathcal{A}^d$.

In Equation 21, we see that $\hat{Q}^{IS,d}$ contains only reward terms $r^d(z_t^{(n),d}, a_t^{(n),d})$ corresponding to the d^{th} action space. The purpose of importance sampling is to account for the fact that each sampled reward term is generated from a behaviour policy, but used to evaluate an evaluation policy. In the context of $\hat{Q}^{IS,d}$, the sampled reward term is $r^d(z_t^{(n),d}, a_t^{(n),d})$, the behaviour policy is π_b^d and the evaluation policy is π_e^d. Thus we propose $\rho_{0:T}^{(n),d}$ such that:

$$\rho_{0:T}^{(n),d} = \prod_{t=0}^{T} \frac{\pi_e^d(a_{t+1}^{(n),d}|z_t^{(n),d})}{\pi_b^d(a_{t+1}^{(n),d}|z_t^{(n),d})} = \frac{\pi_e^d(a_1^{(n),d}|z_0^{(n),d}) \cdot \pi_e^d(a_2^{(n),d}|z_1^{(n),d}) \dots \pi_e^d(a_{T+1}^{(n),d}|z_T^{(n),d})}{\pi_b^d(a_1^{(n),d}|z_0^{(n),d}) \cdot \pi_b^d(a_2^{(n),d}|z_1^{(n),d}) \dots \pi_b^d(a_{T+1}^{(n),d}|z_T^{(n),d})} \tag{24}$$

This only contains the terms relevant to the d^{th} action space. We note that we have essentially decomposed $\rho_{0:T}^{(n)}$:

$$\rho_{0:T}^{(n)} = \prod_{d=1}^{D} \rho_{0:T}^{(n),d} \tag{25}$$

We thus propose a new version of $\hat{Q}_{\pi_e}^{IS,d}$ as:

$$\hat{Q}_{\pi_e}^{DecIS,\ d} = \frac{1}{N} \sum_{n=1}^{N} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \tag{26}$$

and the overall decomposed IS estimator would be given by:

$$\hat{Q}_{\pi_e}^{DecIS} = \sum_{d=1}^{D} \hat{Q}_{\pi_e}^{DecIS,\ d} = \sum_{d=1}^{D} \frac{1}{N} \sum_{n=1}^{N} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \tag{27}$$

The decomposed Per-Decision IS (PDIS) estimator is derived in the same way as above. The expression for the PDIS estimator from Precup et al. (2000) is given by:

$$\hat{Q}_{\pi_e}^{PDIS} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T} \rho_{0:t}^{(n)} \cdot \gamma^t \cdot r(s_t^{(n)}, a_t^{(n)}) \tag{28}$$

where

$$\rho_{0:t}^{(n)} = \prod_{t'=0}^{t} \frac{\pi_e(a_{t'+1}^{(n)}|s_{t'}^{(n)})}{\pi_b(a_{t'+1}^{(n)}|s_{t'}^{(n)})} \tag{29}$$

We would apply equations 2, 4 and 5 in the same way to obtain the decomposed PDIS estimator.

B. Derivation of the Decomposed PDWIS Estimator

The Per-Decision Weighted IS (PDWIS) estimator (Precup et al., 2000) is defined as:

$$\hat{Q}^{IS} = \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n)} \cdot r(s_t^{(n)}, a_t^{(n)})}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n)}} \tag{30}$$

Observing the expression for the PDIS estimator in Equation 28, we see that instead of dividing by N, we have divided by the discounted sum of IS weights $\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n)}$ in Equation 30. This is how we convert the PDIS estimator to the PDWIS estimator.

The decomposed PDIS estimator is given by:

$$\hat{Q}_{\pi_e}^{DecPDIS} = \sum_{d=1}^{D} \hat{Q}_{\pi_e}^{DecPDIS,\ d} = \sum_{d=1}^{D} \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \tag{31}$$

We can convert each $\hat{Q}_{\pi_e}^{DecPDIS,\ d}$ into a $\hat{Q}_{\pi_e}^{DecPDWIS,\ d}$ given by:

$$\hat{Q}_{\pi_e}^{DecPDWIS,\ d} = \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}}$$

Then, the decomposed PDWIS estimator is given by:

$$\hat{Q}_{\pi_e}^{DecPDWIS} = \sum_{d=1}^{D} \hat{Q}_{\pi_e}^{DecPDWIS,\ d} = \sum_{d=1}^{D} \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}} \tag{32}$$

C. Bias of the Decomposed Estimators

We are finding the bias with respect to the true Q_{π_e}, which is given by:

$$Q_{\pi_e} = \mathbb{E}_{\pi_e}[\sum_{t=0}^{T} \gamma^t \cdot r(s_t^{(n)}, a_t^{(n)})] \tag{33}$$

For each estimator, we calulate the bias using Equation 13. For all subsequent derivations, we define $\mathcal{T}(T)$ as the set of all possible trajectories τ, each of length T, that can be generated by π_b (where a trajectory is as defined in part 3).

C.1. Decomposed IS Estimator

We first write:

$$\mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}^{DecIS}] = \mathbb{E}_{\pi_b}[\sum_{d=1}^{D} \frac{1}{N} \sum_{n=1}^{N} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})] = \mathbb{E}_{\pi_b}[\sum_{d=1}^{D} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})]$$

where we make use of the fact that $\mathbb{E}[\frac{1}{N}\sum_{i=1}^{N} X_i] = \mathbb{E}[X_i]$. We then say:

$$\mathbb{E}_{\pi_b}[\sum_{d=1}^{D} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})] = \sum_{\tau\ \in\ \mathcal{T}(T)} (\prod_{t=0}^{T} \pi_b(a_t^{(n)}|s_t^{(n)})) \cdot \sum_{d=1}^{D} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})$$

Here, $(\prod_{t=0}^{T} \pi_b(s_t^{(n)}, a_t^{(n)}))$ denotes the probability of this entire trajectory occurring under π_b. Given its definition in Equation 24, the denominator of the $\rho_{0:T}^{(n),d}$ cancels with $(\prod_{t=0}^{T} \pi_b(s_t^{(n)}, a_t^{(n)}))$ to give:

$$\sum_{\tau\ \in\ \mathcal{T}(T)} \sum_{d=1}^{D} (\prod_{t=0}^{T} \prod_{d' \neq d} \pi_b^{d'}(a_t^{(n),d'}|z_t^{(n),d'})) \cdot (\prod_{t=0}^{T} \pi_e^{d}(a_t^{(n),d}|z_t^{(n),d})) \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})$$

The term $(\prod_{t=0}^{T} \prod_{d' \neq d} \pi_b^{d'}(a_t^{(n),d'}|z_t^{(n),d'}))$ is the probability under π_b of all trajectories occurring in all factored action spaces except d. This probability is independent of the other terms that are related to d, and the probabilities sum to 1 over all trajectories, hence we get:

$$\sum_{\tau\ \in\ \mathcal{T}(T)} 1 \cdot \sum_{d=1}^{D} (\prod_{t=0}^{T} \pi_e^{d}(a_t^{(n),d}|z_t^{(n),d})) \sum_{t=0}^{T} \gamma^t \cdot r^d(z_t^{(n),d}, a_t^{(n),d})$$

Since $(\prod_{t=0}^{T}\prod_{d'\neq d}\pi_e^{d'}(a_t^{(n),d'}|z_t^{(n),d'}))$ would also sum to 1 over all trajectories as it is independent of d, we can write:

$$= \sum_{\tau\,\in\,\mathcal{T}(T)}\sum_{d=1}^{D}(\prod_{t=0}^{T}\prod_{d'\neq d}\pi_e^{d'}(a_t^{(n),d'}|z_t^{(n),d'}))\cdot(\prod_{t=0}^{T}\pi_e^{d}(a_t^{(n),d}|z_t^{(n),d}))\sum_{t=0}^{T}\gamma^t\cdot r^d(z_t^{(n),d},a_t^{(n),d})$$

$$= \sum_{\tau\,\in\,\mathcal{T}(T)}(\prod_{t=0}^{T}\pi_e(a_t^{(n)}|s_t^{(n)}))\cdot\sum_{d=1}^{D}\sum_{t=0}^{T}\gamma^t\cdot r^d(z_t^{(n),d},a_t^{(n),d})$$

$$= \mathbb{E}_{\pi_e}[\sum_{d=1}^{D}\sum_{t=0}^{T}\gamma^t\cdot r^d(z_t^{(n),d},a_t^{(n),d})] = \mathbb{E}_{\pi_e}[\sum_{t=0}^{T}\gamma^t\cdot r(s_t^{(n)},a_t^{(n)})]$$

Where we have applied Equation 4 i.e. $r(s,a)=\sum_{d=1}^{D}r^d(z^d,a^d)$. Thus it is clear that:

$$\mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}^{DecIS}] = \mathbb{E}_{\pi_b}[\sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\rho_{0:T}^{(n),d}\sum_{t=0}^{T}\gamma^t\cdot r^d(z_t^{(n),d},a_t^{(n),d})] = \mathbb{E}_{\pi_e}[\sum_{t=0}^{T}\gamma^t\cdot r(s_t^{(n)},a_t^{(n)})] = Q_{\pi_e}$$

That is:

$$Bias[\tilde{Q}^{IS}] = \mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}^{DecIS}] - Q_{\pi_e} = 0$$

This means that the decomposed IS estimator is unbiased with respect to Q_{π_e}.

C.2. Decomposed PDIS Estimator

The proof is similar to that of the decomposed IS estimator.

$$\mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}^{DecPDIS}] = \mathbb{E}_{\pi_b}\Big[\sum_{d=0}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t\cdot\rho_{0:t}^{(n),d}\cdot r^d(z_t^d,a_{t+1}^d)\Big] = \mathbb{E}_{\pi_b}\Big[\sum_{d=0}^{D}\sum_{t=0}^{T}\gamma^t\cdot\rho_{0:t}^{(n),d}\cdot r^d(z_t^d,a_{t+1}^d)\Big]$$

$$= \sum_{\tau\,\in\,\mathcal{T}(T)}(\prod_{t=0}^{T}\pi_b(a_t^{(n)}|s_t^{(n)}))\sum_{t=0}^{T}\gamma^t\cdot\sum_{d=0}^{D}\rho_{0:t}^{(n),d}\cdot r^d(z_t^d,a_{t+1}^d)$$

$$= \sum_{\tau\,\in\,\mathcal{T}(T)}\sum_{t=0}^{T}\gamma^t\cdot\sum_{d=0}^{D}\Big(\prod_{t=0}^{T}\prod_{d'\neq d}\pi_b^{d'}(a_t^{(n),d'}|z_t^{(n),d'})\Big)\cdot\Big(\prod_{t=t+1}^{T}\pi_b^{d}(a_t^{(n),d}|z_t^{(n),d})\Big)$$
$$\cdot\Big(\prod_{t=0}^{t}\pi_e^{d}(a_t^{(n),d}|z_t^{(n),d})\Big)\cdot r^d(z_t^d,a_{t+1}^d)$$

The term $\Big(\prod_{t=0}^{T}\prod_{d'\neq d}\pi_b^{d'}(a_t^{(n),d'}|z_t^{(n),d'})\Big)\cdot\Big(\prod_{t=t+1}^{T}\pi_b^{d}(a_t^{(n),d}|z_t^{(n),d})\Big)$ is a probability value that varies independently of the reward $r^d(z_t^d,a_{t+1}^d)$ because it corresponds to unrelated factored action spaces and timesteps. Hence, we can expect that it will sum to 1 under the nested summations to give:

$$= \sum_{\tau\,\in\,\mathcal{T}(T)}\sum_{t=0}^{T}\gamma^t\cdot\sum_{d=0}^{D}1\cdot\Big(\prod_{t=0}^{t}\pi_e^{d}(a_t^{(n),d}|z_t^{(n),d})\Big)\cdot r^d(z_t^d,a_{t+1}^d)$$

$$= \sum_{t=0}^{T} \gamma^t \cdot \sum_{d=0}^{D} \sum_{\tau \in \mathcal{T}(T)} \Big(\prod_{t=0}^{t} \pi_e^d(a_t^{(n),d} | z_t^{(n),d}) \Big) \cdot r^d(z_t^d, a_{t+1}^d) = \sum_{t=0}^{T} \gamma^t \cdot \sum_{d=0}^{D} \mathbb{E}_{\pi_e} \Big[r^d(z_t^d, a_{t+1}^d) \Big]$$

Then, since expectation is a linear operator, we get:

$$\mathbb{E}_{\pi_e} \Big[\sum_{t=0}^{T} \gamma^t \cdot \sum_{d=0}^{D} \cdot r^d(z_t^d, a_{t+1}^d) \Big] = \mathbb{E}_{\pi_e} \Big[\sum_{t=0}^{T} \gamma^t \cdot r(s_t^{(n)}, a_t^{(n)}) \Big] = Q_{\pi_e}$$

i.e. the true value of the Q function. Thus, the decomposed PDIS estimator is unbiased with respect to the true Q-value.

C.3. Decomposed PDWIS Estimator

Here, we find that the PDWIS estimator has non-zero bias:

$$\mathbb{E}_{\pi_b}[\hat{Q}_{\pi_e}^{DecPDIS}] = \mathbb{E}_{\pi_b} \Big[\sum_{d=0}^{D} \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r^d(z_t^d, a_{t+1}^d)}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}} \Big]$$

$$= \sum_{\tau \in \mathcal{T}(T)} (\prod_{t=0}^{T} \pi_b(a_t^{(n)} | s_t^{(n)})) \sum_{d=0}^{D} \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d} \cdot r^d(z_t^d, a_{t+1}^d)}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}}$$

$$= \sum_{\tau \in \mathcal{T}(T)} \sum_{d=0}^{D} \frac{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \Big(\prod_{t=0}^{T} \prod_{d' \neq d} \pi_b^{d'}(z_t^{(n),d'}, a_t^{(n),d'}) \Big) \cdot \Big(\prod_{t=t+1}^{T} \pi_b^d(z_t^{(n),d}, a_t^{(n),d}) \Big)}{\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}}$$

$$\times \frac{\Big(\prod_{t=0}^{t} \pi_e^d(z_t^{(n),d}, a_t^{(n),d}) \Big) \cdot r^d(z_t^d, a_{t+1}^d)}{1}$$

We can no longer say that $\Big(\prod_{t=0}^{T} \prod_{d' \neq d} \pi_b^{d'}(z_t^{(n),d'}, a_t^{(n),d'}) \Big) \cdot \Big(\prod_{t=t+1}^{T} \pi_b^d(z_t^{(n),d}, a_t^{(n),d}) \Big)$ will add up to 1 over $\tau \in \mathcal{T}(T)$ because each term is weighted differently due to the denominator $\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}$ varying based on the trajectory. There is then no way to simplify the above expression, hence it is clear that the bias of the decomposed PDWIS estimator is non-zero. The bias would approach zero if the denominator term $\sum_{n=1}^{N} \sum_{t=0}^{T} \gamma^t \cdot \rho_{0:t}^{(n),d}$ is consistently ≈ 1.

D. Variance Guarantees for Decomposed Estimators

In this section, we compare the variances of the decomposed IS, PDIS and PDWIS estimators.

D.1. Decomposed IS Estimator

We utilise Equation 14 to write out for the decomposed IS estimator:

$$\mathbb{V}_{\pi_b}[\hat{Q}_{\pi_e}^{DecIS}] = \sum_{d=1}^{D} \sum_{n=1}^{N} \sum_{t=0}^{T} \sum_{t'=0}^{T} \frac{\gamma^{t+t'}}{N^2} \Big(\mathbb{E}_{\pi_b} \Big[\Big(\rho_{0:T}^{(n),d} \Big)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d}) \Big]$$

$$- \mathbb{E}_{\pi_b} \Big[\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \Big] \cdot \mathbb{E}_{\pi_b} \Big[\rho_{0:T}^{(n),d} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d}) \Big] \Big)$$

For the IS estimator, we have:

$$\mathbb{V}_{\pi_b}[\hat{Q}_{\pi_e}^{IS}] = \sum_{n=1}^{N} \sum_{t=0}^{T} \sum_{t'=0}^{T} \frac{\gamma^{t+t'}}{N^2} \Big(\mathbb{E}_{\pi_b} \Big[\Big(\rho_{0:T}^{(n)} \Big)^2 \cdot r(s_t^{(n)}, a_t^{(n)}) \cdot r(s_{t'}^{(n)}, a_{t'}^{(n)}) \Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\rho^{(n)}_{0:T}\cdot r(s^{(n)}_t,a^{(n)}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n)}_{0:T}\cdot r(s^{(n)}_{t'},a^{(n)}_{t'})\Big]\Big)$$

Applying equations 5 (decomposition of policy) and 4 (decomposition of reward), we get:

$$\mathbb{V}_{\pi_b}[\hat{Q}^{IS}]=\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho^{(n),d}_{0:T}\Big)^2\cdot\Big(\sum_{d=0}^{D}r^d(z^{(n),d}_t,a^{(n),d}_t)\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z^{(n),d}_{t'},a^{(n),d}_{t'})\Big)\Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho^{(n),d}_{0:T}\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z^{(n),d}_t,a^{(n),d}_t)\Big)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho^{(n),d}_{0:T}\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z^{(n),d}_{t'},a^{(n),d}_{t'})\Big)\Big]$$

D.1.1. Simplified Case of D=2

We consider a simplified case where the number of factored action spaces $D=2$; we let the labels $d=\{0,1\}$. In this case:

$$\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]=\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[(\rho^{(n),0}_{0:T})^2\cdot(\rho^{(n),1}_{0:T})^2\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]+$$

$$2\cdot\mathbb{E}_{\pi_b}\Big[(\rho^{(n),0}_{0:T})^2\cdot(\rho^{(n),1}_{0:T})^2\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]+$$

$$\mathbb{E}_{\pi_b}\Big[(\rho^{(n),0}_{0:T})^2\cdot(\rho^{(n),1}_{0:T})^2\cdot r^1(z^{(n),1}_t,a^{(n),1}_t)\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]-$$

$$\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]-$$

$$2\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]-$$

$$\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^1(z^{(n),1}_t,a^{(n),0}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot\rho^{(n),1}_{0:T}\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]\Big)$$

And for the decomposed IS estimator $\hat{Q}^{DecIS}_{\pi_e}$, we get:

$$\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]=\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[\Big(\rho^{(n),0}_{0:T}\Big)^2\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),0}_{0:T}\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]$$

$$+\mathbb{E}_{\pi_b}\Big[\Big(\rho^{(n),1}_{0:T}\Big)^2\cdot r^1(z^{(n),1}_t,a^{(n),1}_t)\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\rho^{(n),1}_{0:T}\cdot r^1(z^{(n),1}_t,a^{(n),1}_t)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho^{(n),1}_{0:T}\cdot r^1(z^{(n),1}_{t'},a^{(n),1}_{t'})\Big]\Big)$$

The following terms in $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ and $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$ correspond to each other and are similar but slightly different:

- $\mathbb{E}_{\pi_b}\Big[(\rho^{(n),0}_{0:T})^2\cdot(\rho^{(n),1}_{0:T})^2\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]$ and $\mathbb{E}_{\pi_b}\Big[\Big(\rho^{(n),0}_{0:T}\Big)^2\cdot r^0(z^{(n),0}_t,a^{(n),0}_t)\cdot r^0(z^{(n),0}_{t'},a^{(n),0}_{t'})\Big]$

- $\mathbb{E}_{\pi_b}\left[(\rho_{0:T}^{(n),0})^2 \cdot (\rho_{0:T}^{(n),1})^2 \cdot r^1(z_t^{(n),1}, a_t^{(n),1}) \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$ and $\mathbb{E}_{\pi_b}\left[\left(\rho_{0:T}^{(n),1}\right)^2 \cdot r^1(z_t^{(n),1}, a_t^{(n),1}) \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$
- $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_t^{(n),0}, a_t^{(n),0})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_{t'}^{(n),0}, a_{t'}^{(n),0})\right]$ and $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot r^0(z_t^{(n),0}, a_t^{(n),0})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot r^0(z_{t'}^{(n),0}, a_{t'}^{(n),0})\right]$
- $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_t^{(n),1}, a_t^{(n),1})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$ and $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),1} \cdot r^1(z_t^{(n),1}, a_t^{(n),1})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$

Additionally, The expression for $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ has the following extra terms:

- $2 \cdot \mathbb{E}_{\pi_b}\left[(\rho_{0:T}^{(n),0})^2 \cdot (\rho_{0:T}^{(n),1})^2 \cdot r^0(z_t^{(n),0}, a_t^{(n),0}) \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$
- $-2 \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_t^{(n),0}, a_t^{(n),0})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$

First, we would like to conclude something about whether the sum of the extra terms is greater than or less than zero:

$$2 \cdot \mathbb{E}_{\pi_b}\left[(\rho_{0:T}^{(n),0})^2 \cdot (\rho_{0:T}^{(n),1})^2 \cdot r^0(z_t^{(n),0}, a_t^{(n),0}) \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right] -$$
$$2 \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_t^{(n),0}, a_t^{(n),0})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})\right]$$

The above expression is a covariance expression i.e. it is equal to $Cov(\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_t^{(n),0}, a_t^{(n),0}),\ \rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1}))$. This is the covariance between the IS weighted rewards in dimension 0 and dimension 1 at times t and t' respectively. If this covariance is negative, it would mean that increasing the weighted reward in one action space would cause a (possibly delayed) decrease in weighted reward in another action space, which means that optimising these rewards in different action spaces is like trying to achieve contrasting goals.

Here, we will assume that we have factored the action spaces in such a way that all of them are "aimed" towards a common goal i.e. optimising reward in one action space will either not affect the reward in another action space or increase the reward in this action space. In this case, we would have $Cov(r^0(z_t^{(n),0}, a_t^{(n),0}),\ r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})) \geq 0$. This relates to condition 6 from Assumption 3. The product of importance sampling weights i.e. $\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1}$ is common to each expression, hence the overall covariance depends only on the reward terms. Therefore, we can assume that $Cov(\rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^0(z_t^{(n),0}, a_t^{(n),0}),\ \rho_{0:T}^{(n),0} \cdot \rho_{0:T}^{(n),1} \cdot r^1(z_{t'}^{(n),1}, a_{t'}^{(n),1})) \geq 0$ ie. *the sum of the extra terms is never less than zero*. We hold on to this result.

Second, we want to compare $\mathbb{E}_{\pi_b}\left[(\rho_{0:T}^{(n),d})^2 \cdot (\rho_{0:T}^{(n),d'})^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ and $\mathbb{E}_{\pi_b}\left[\left(\rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ for $(d, d') = (0,1), (1,0)$. Let $X = (\rho_{0:T}^{(n),d})^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})$ and $A = (\rho_{0:T}^{(n),d'})^2$. Here, we assume that $Cov(X, A) = 0$ due to the two terms being in different factored action spaces d and d' - this is condition 7 from Assumption 3. However, $Cov(X, A) = E(XA) - E(X)E(A)$, which implies $E(XA) = E(X)E(A) + Cov(X, A)$. Let us consider $E(A) = \mathbb{E}_{\pi_b}[(\rho_{0:T}^{(n),d})^2]$. As mentioned in Jiang (2022), we know that the expected value of an IS weight is 1 i.e. $\mathbb{E}_{\pi_b}[(\rho_{0:T}^{(n),d})] = 1$. Since we know that $Var(\rho_{0:T}^{(n)}) \geq 0$ (true for all variances), we have $\mathbb{E}_{\pi_b}[(\rho_{0:T}^{(n),d})^2] - \mathbb{E}_{\pi_b}[(\rho_{0:T}^{(n),d})]^2 \geq 0$. Thus, $E((\rho_{0:T}^{(n),d})^2) - 1^2 \geq 0$, which implies $E((\rho_{0:T}^{(n),d})^2) \geq 1$, and thus $E(A) \geq 1$. This gives us $E(XA) = \lambda \cdot E(X) + 0$, where $\lambda \geq 1$. Clearly, $E(XA) \geq E(X)$ i.e. $\mathbb{E}_{\pi_b}\left[(\rho_{0:T}^{(n),d})^2 \cdot (\rho_{0:T}^{(n),d'})^2 \cdot r^0(z_t^{(n),d}, a_t^{(n),0}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right] \geq \mathbb{E}_{\pi_b}\left[\left(\rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ for $(d, d') = (0,1), (1,0)$.

Third, we want to compare $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot \rho_{0:T}^{(n),d'} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot \rho_{0:T}^{(n),d'} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ and $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$. Here, we assume that $Cov(\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d}),\ \rho_{0:T}^{(n),d'}) = 0$ due to the difference in the factored action space - this utilises conditions 7 and 8 from assumption 3. As a result,

$\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d}\cdot\rho_{0:T}^{(n),d'}\cdot r^d(z_t^{(n),d},a_t^{(n),d})]-\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d'}]\cdot\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})]=0$. Knowing that $\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d'}]=1$ gives us: $\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d}\cdot\rho_{0:T}^{(n),d'}\cdot r^d(z_t^{(n),d},a_t^{(n),d})]=\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})]$. This means that we can say $\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot\rho_{0:T}^{(n),d'}\cdot r^d(z_t^{(n),d},a_t^{(n),d})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot\rho_{0:T}^{(n),d'}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big] = \mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_t^{(n),d},a_t^{(n),d})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$.

Overall, the corresponding terms in $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ are larger or equal to those in $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_b}]$ and the extra terms in the former also add up to a value ≥ 0. Thus, clearly in this case, $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]\geq\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_b}]$.

D.1.2. Extension to $D>2$

We extend the findings from the previous section to the general case. Consider the general expressions for $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$ and $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$:

$$\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]=\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[\Big(\rho_{0:T}^{(n),d}\Big)^2\cdot r^d(z_t^{(n),d},a_t^{(n),d})\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_t^{(n),d},a_t^{(n),d})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]\Big)$$

$$\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]=\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)^2\cdot\Big(\sum_{d=0}^{D}r^d(z_t^{(n),d},a_t^{(n),d})\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big)\Big]$$

$$-\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z_t^{(n),d},a_t^{(n),d})\Big)\Big]\cdot\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot\Big(\sum_{d=0}^{D}r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big)\Big]$$

A general positive term in $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$ is $PT_1=\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\Big(\rho_{0:T}^{(n),d}\Big)^2\cdot r^d(z_t^{(n),d},a_t^{(n),d})\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$, while a general positive term in $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ is $PT_2=\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)^2\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1})\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big]$. Looking at PT_2, when $d_1=d_2=d$, PT_2 corresponds to PT_1 in that the terms are similar. When $d_1\neq d_2$, PT_2 is extra i.e. has no corresponding term in $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$.

A general negative term in $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$ is $NT_1=-\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_t^{(n),d},a_t^{(n),d})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\rho_{0:T}^{(n),d}\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$, while a general negative term in $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ is $NT_2=-\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big]$. Again, when $d_1=d_2=d$, then NT_2 corresponds to NT_1 and when $d_1\neq d_2$, NT_2 is extra.

Combining positive and negative terms, a general extra term in $\mathbb{V}_{\pi_b}[\hat{Q}^{DecIS}_{\pi_e}]$ that has no counterpart in $\mathbb{V}_{\pi_b}[\hat{Q}^{IS}_{\pi_e}]$ is $\frac{\gamma^{t+t'}}{N^2}\Big(\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)^2\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1})\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big]-\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1})\Big]\cdot\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big]\Big)$, which is equal to $\frac{\gamma^{t+t'}}{N^2}\cdot Cov\Big(\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1}),\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big)$. In part D.1.1, we argued that $Cov(\rho_{0:T}^{(n),0}\cdot\rho_{0:T}^{(n),1}\cdot r^0(z_t^{(n),0},a_t^{(n),0}),\ \rho_{0:T}^{(n),0}\cdot\rho_{0:T}^{(n),1}\cdot r^1(z_{t'}^{(n),1},a_{t'}^{(n),1}))\geq 0.$, based on the assumption that $Cov(r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1}),\ r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2}))\geq 0$. We can use this assumption to argue that $Cov\Big(\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_1}(z_t^{(n),d_1},a_t^{(n),d_1}),\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)\cdot r^{d_2}(z_{t'}^{(n),d_2},a_{t'}^{(n),d_2})\Big)\geq 0$.

We now consider the positive corresponding terms i.e. PT_1 and PT_2 where $d_1=d_2$. These terms are $PT_1=\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\Big(\rho_{0:T}^{(n),d}\Big)^2\cdot r^d(z_t^{(n),d},a_t^{(n),d})\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$ and $PT_2=\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\Big[\Big(\prod_{d=0}^{D}\rho_{0:T}^{(n),d}\Big)^2\cdot r^d(z_t^{(n),d},a_t^{(n),d})\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]$. In part Appendix D.1.1, we argued that $\mathbb{E}_{\pi_b}\Big[(\rho_{0:T}^{(n),d})^2\cdot(\rho_{0:T}^{(n),d'})^2\cdot r^0(z_t^{(n),d},a_t^{(n),0})\cdot r^d(z_{t'}^{(n),d},a_{t'}^{(n),d})\Big]\geq$

$\mathbb{E}_{\pi_b}\left[\left(\rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ for $(d, d') = (0,1),(1,0)$, using the assumption that an importance sampling ratio in any factored action space is independent of the reward terms and IS ratios in other factored action spaces. Using this assumption, we can say that $\left(\rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})$ and $\left(\prod_{d' \neq d} \rho_{0:T}^{(n),d'}\right)^2$ are uncorrelated. We can also use this assumption to say that $\mathbb{E}_{\pi_b}\left[\left(\prod_{d' \neq d} \rho_{0:T}^{(n),d'}\right)^2\right] \geq 1$ because $Var\left[\prod_{d' \neq d} \rho_{0:T}^{(n),d'}\right] \geq 0$ and $\mathbb{E}_{\pi_b}\left[\prod_{d' \neq d} \rho_{0:T}^{(n),d'}\right]^2 = \left(\prod_{d' \neq d} \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d'}\right]\right)^2 = \prod_{d' \neq d} 1$. Here, we have also used the result that $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d'}\right] = 1$ (Jiang, 2022). Therefore, $\mathbb{E}_{\pi_b}\left[\left(\prod_{d=0}^{D} \rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right] \geq \mathbb{E}_{\pi_b}\left[\left(\rho_{0:T}^{(n),d}\right)^2 \cdot r^d(z_t^{(n),d}, a_t^{(n),d}) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$.

Finally we consider the negative corresponding terms i.e. NT_1 and NT_2 where $d_1 = d_2$. These terms are $NT_1 = -\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$ and $NT_2 = -\frac{\gamma^{t+t'}}{N^2}\mathbb{E}_{\pi_b}\left[\left(\prod_{d=0}^{D} \rho_{0:T}^{(n),d}\right) \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\left(\prod_{d=0}^{D} \rho_{0:T}^{(n),d}\right) \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$. In part D.1.1, we argued that $\mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot \rho_{0:T}^{(n),d'} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot \rho_{0:T}^{(n),d'} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right] = \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_{t'}^{(n),d}, a_{t'}^{(n),d})\right]$, using the assumption that an importance sampling ratio in any factored action space is independent of the reward terms and IS ratios in other factored action spaces. Based on this assumption, we can say that $Cov(\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d}), \prod_{d' \neq d} \rho_{0:T}^{(n),d'}) = 0$. Furthermore, assuming that $\mathbb{E}_{\pi_b}[\rho_{0:T}^{(n),d}] = 1$, we can say that $\mathbb{E}_{\pi_b}\left[\prod_{d' \neq d} \rho_{0:T}^{(n),d'}\right] = \prod_{d' \neq d} \mathbb{E}_{\pi_b}\left[\rho_{0:T}^{(n),d'}\right] = \prod_{d' \neq d} 1 = 1$. Based on this, we can conclude that $E_{\pi_b}\left[\rho_{0:T}^{(n),d} \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right] = \mathbb{E}_{\pi_b}\left[\left(\prod_{d=0}^{D} \rho_{0:T}^{(n),d}\right) \cdot r^d(z_t^{(n),d}, a_t^{(n),d})\right]$, and thus $NT_1 = NT_2$.

Since the negative terms are equal, the positive terms in the variance of the original IS estimator are larger and the extra terms are all no less than zero, it is clear that even for $D > 2$, we have that $\mathbb{V}_{\pi_b}[\hat{Q}_{\pi_e}^{DecIS}] \leq \mathbb{V}_{\pi_b}[\hat{Q}_{\pi_e}^{IS}]$.

D.2. Decomposed PDIS

Given the highly similar natures of the IS and PDIS estimators, the Theorem 3 variance guarantee for the decomposed PDIS estimator may be proven in the exact same way as for decomposed IS in part D.1. The only difference, is that at all steps, we would replace every occurrence of an episode-based IS ratio of the form $\rho_{0:T}^{(n),d}$ with a per-decision IS ratio of the form $\rho_{0:t}^{(n),d}$.

D.3. Decomposed PDWIS

Using Equation 14, the variance of the decomposed PDWIS estimator can be derived as below. Note that for brevity, we have replaced $r^d(z_t^{(n),d}, a_t^{(n),d})$ with $r_t^{(n),d}$:

$$\mathbb{V}[\hat{Q}_{\pi_b}^{DecPDWIS}] = \sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{t=0}^{T}\sum_{t'=0}^{T} \frac{\gamma^{t+t'}}{N^2}\left(\frac{\mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n),d} \cdot \rho_{0:t'}^{(n),d} \cdot r_t^{(n),d} \cdot r_{t'}^{(n),d}\right]}{\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T} \gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n),d} \cdot \rho_{0:t'}^{(n),d}\right]}\right.$$
$$\left. - \frac{\mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n),d} \cdot r_t^{(n),d}\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:t'}^{(n),d} \cdot r_{t'}^{(n),d}\right]}{\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T} \gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n),d}\right] \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:t'}^{(n),d}\right]}\right)$$

This is similar to the expression for the decomposed IS estimator - the difference is the presence of denominator terms. The expression for the variance of the PDWIS estimator can be similarly obtained as:

$$Var[\hat{Q}_{\pi_b}^{DecPDWIS}] = \sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T} \frac{\gamma^{t+t'}}{N^2}\left(\frac{\mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n)} \cdot \rho_{0:t'}^{(n)} \cdot r(s_t^{(n)}, a_t^{(n)}) \cdot r(s_{t'}^{(n)}, a_{t'}^{(n)})\right]}{\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T} \gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho_{0:t}^{(n)} \cdot \rho_{0:t'}^{(n)}\right]}\right.$$

$$- \frac{\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t} \cdot r(s^{(n)}_t, a^{(n)}_t)\right]\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t'} \cdot r(s^{(n)}_{t'}, a^{(n)}_{t'})\right]}{\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t}\right] \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t'}\right]}\Bigg)$$

We have already shown in the variance comparison between decomposed and non-decomposed IS in part D.1 that, since the expected value of any IS ratio is 1, then $\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t} \cdot \rho^{(n)}_{0:t'} \cdot r(s^{(n)}_t, a^{(n)}_t) \cdot r(s^{(n)}_{t'}, a^{(n)}_{t'})\right] \geq \sum_{d=1}^{D}\mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t} \cdot \rho^{(n),d}_{0:t'} \cdot r^d(z^{(n),d}_t, a^{(n),d}_t) \cdot r^d(z^{(n),d}_{t'}, a^{(n),d}_{t'})\right]$. We also showed that $\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t} \cdot r(s^{(n)}_t, a^{(n)}_t)\right]\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t'} \cdot r(s^{(n)}_{t'}, a^{(n)}_{t'})\right] = \sum_{d=1}^{D}\mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t} \cdot r^d(z^{(n),d}_t, a^{(n),d}_t)\right]\mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t'} \cdot r^d(z^{(n),d}_{t'}, a^{(n),d}_{t'})\right]$. This, so far, has allowed us to compare the numerators of the corresponding pairs of terms in the variance expressions of the PDWIS estimators.

What remains is the denominators. Since we have assumed $\mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t}\right] = \mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t}\right] = 1$, this means that $\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t}\right] \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t'}\right] = \sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t}\right] \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t'}\right]$ i.e. the denominators of the second terms in each variance expression are equal. This only leaves the denominators of the first terms to compare.

We can rewrite $\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\rho^{(n)}_{0:t} \cdot \rho^{(n)}_{0:t'}\right]$ as $\sum_{n=1}^{N}\sum_{t=0}^{T}\sum_{t'=0}^{T}\gamma^{t+t'} \cdot \mathbb{E}_{\pi_b}\left[\left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t}\right) \cdot \left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t'}\right)\right]$. Thus, we guarantee that the decomposed PDWIS estimator has at most the variance of the PDWIS estimator if $\mathbb{E}_{\pi_b}\left[\left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t}\right) \cdot \left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t'}\right)\right] \leq \mathbb{E}_{\pi_b}\left[\rho^{(n),d}_{0:t} \cdot \rho^{(n),d}_{0:t'}\right]$. This could be the case if $Cov\Big(\left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\right) \cdot \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right),\ \rho^{(n),d}_{0:t} \cdot \rho^{(n),d}_{0:t'}\Big) = \mathbb{E}_{\pi_b}\left[\left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t}\right) \cdot \left(\prod_{d=1}^{D}\rho^{(n),d}_{0:t'}\right)\right] - \mathbb{E}\Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\Big) \cdot \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big] \cdot \mathbb{E}\left[\rho^{(n),d}_{0:t} \cdot \rho^{(n),d}_{0:t'}\right] \leq 0$. This is in fact related to the condition in Equation 7 in Assumption 3. We additionally require that $Cov\Big(\left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\right),\ \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big) \geq 0$, which is related to condition 9 in Assumption 4 to ensure that $\mathbb{E}\Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\Big) \cdot \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big] \geq 1$. This is because $Cov\Big(\left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\right),\ \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big) = \mathbb{E}\Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\Big) \cdot \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big] - \mathbb{E}\Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\Big] \cdot \Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\Big] = \mathbb{E}\Big[\prod_{d'\neq d}\rho^{(n),d'}_{0:t}\Big) \cdot \left(\prod_{d'\neq d}\rho^{(n),d'}_{0:t'}\right)\Big] - 1$.

Thus, when necessary conditions are satisfied, we can show that the decomposed PDWIS estimator has lower or equal variance, compared to non-decomposed PDWIS.

E. Grouping Action Spaces Together to Create Unbiased Estimators

Here we illustrate how actions may be grouped in cases where equations 5 and 4 do not apply. As a running example, we take the decomposed IS estimator, but the same transformations can be applied to the decomposed PDIS and PDWIS estimators.

When Equation 5 does not apply i.e. the policy probabilities cannot be product-wise separated for the current action space factorisation, we need to adjust the IS ratios. We group the action space indices $d \in \{1, 2...D\}$ into sets $S_{k,e}$ for policy π_e , such that $k \in \{1, 2...K_e\}$, and $S_{k,b}$ for policy π_b, such that $k \in \{1, 2...K_b\}$. We can then express Equation 5 as:

$$\pi(a|s) = \prod_{k=1}^{K}\pi^k(a^k|z^k) \tag{34}$$

where a^k is a composite action from the union of action spaces in set S_k, π^k is a policy on actions from this composite space and z^k is a a corresponding state abstraction. The expression for the decomposed IS estimator is similar to before:

$$\tilde{Q}^{DecIS}_{\pi_e} = \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\rho^{(n),d}_{0:T}\sum_{t=0}^{T}\gamma^t \cdot r^d(z^{(n),d}_t, a^{(n),d}_t) \tag{35}$$

However $\rho_{0:T}^{(n),d}$ is defined as follows if $d \in S_{k_1,e}$ and $d \in S_{k_2,b}$:

$$\rho_{0:T}^{(n),d} = \prod_{t=0}^{T} \frac{\pi_e^{k_1}(a_t^{k_1}|z_t^{k_1})}{\pi_b^{k_2}(a_t^{k_2}|z_t^{k_2})}$$

This version of the decomposed IS estimator is unbiased provided that $S_{k,e} = S_{k,b},\ \forall k$ i.e. we have the exact same groupings of action spaces for π_b and π_e. This is because in this case, we can also group the rewards as:

$$r^k(z^k, a^k) = \sum_{d \in S_{k,e}} r^d(z^d, a^d)$$

and thus we are satisfying Theorem 1 with $K_b = K_e$ factored action spaces. The expression for the estimator would be:

$$\tilde{Q}_{\pi_e}^{DecIS} = \sum_{k=1}^{K_e} \frac{1}{N} \sum_{n=1}^{N} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^k(z_t^{(n),k}, a_t^{(n),k}) \tag{36}$$

Since the estimator is a decomposed estimator with theorem 1 satisfied, by theorem 3 it has at most the variance of the non-decomposed IS estimator. Even if the condition $S_{k,e} = S_{k,b},\ \forall k$ is not met, and thus theorem 1 is not satisfied, the decomposed nature of of the estimator still allows us to guarantee that it will have at most the variance of the corresponding non-decomposed estimator. The proof for this is similar to part D and is omitted.

When Equation 4 does not apply either i.e. the rewards cannot be decomposed as a summation, we would again group the the action spaces $d \in \{1, 2...D\}$ into sets $S_{k,r}$ such that $k \in \{1, 2...K_r\}$. This allows Equation 4 to be expressed as:

$$r(s, a) = \sum_{k=1}^{K_r} r^k(z^k, a^k), \tag{37}$$

The new expression for the decomposed IS estimator would be:

$$\tilde{Q}_{\pi_e}^{DecIS} = \sum_{k=1}^{K_r} \frac{1}{N} \sum_{n=1}^{N} \rho_{0:T}^{(n),d} \sum_{t=0}^{T} \gamma^t \cdot r^k(z_t^{(n),k}, a_t^{(n),k}) \tag{38}$$

where $\rho_{0:T}^{(n),d}$ is as defined above. It can again be shown that if $S_{k,r} = S_{k,e} = S_{k,b},\ \forall k$, then the above estimator is unbiased, because we can satisfy Theorem 1 with $K_b = K_e = K_r$ factored action spaces. Again, since it is a decomposed estimator, we can show that it has at most the variance of the IS estimator, regardless of whether the condition $S_{k,r} = S_{k,e} = S_{k,b},\ \forall k$ is met.

We can view the grouping sets $S_{k,r}, S_{k,e}, S_{k,b}$ as the merging of action spaces to form new larger ones that satisfy Theorem 1 - we can define how exactly this grouping is done. In the most general case that $S_{1,r} = S_{1,e} = S_{1,b} = \mathcal{A}$ i.e. we group all the spaces together into one space, we get the original IS estimator back, and then the bias and variance of our "decomposed estimator" is equal to that of the original non-decomposed IS estimator.

F. Diagram of MDP 1 and its Factorisation

In Figure 8, states are represented as labelled circles, while the actions are represented by directed arrows. For example, an arrow pointing up and right represents the action up_right and an arrow pointing down and right represents $down\_right$. There is only one transition possible, from $state$ to $terminal$, and only the action can vary.

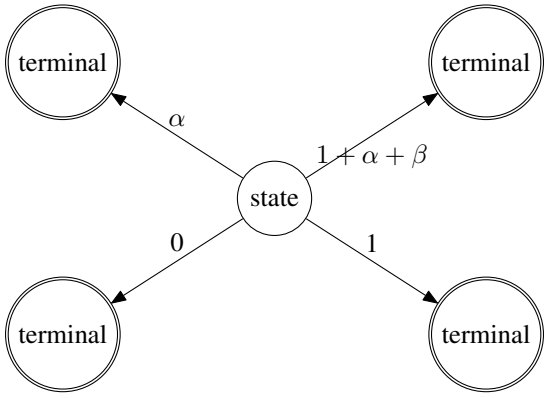

Figure 8. Factored MDP-1 in first factored action space (horizontal).

As discussed, throughout experiments $\alpha = 1$. Furthermore, unless explicitly stated otherwise $\beta = 0$. The only time β is varied was in a set of experiments explicitly designed to observe the effect of varying β, as $\beta \neq 0$ results in Theorem 1 no longer applying to the MDP, due to violation of the condition in Equation 4. Thus, in the default configuration, $r(state,\ up\_right) = 2$, $r(state,\ up\_left) = 1$, $r(state,\ down\_right) = 1$ and $r(state,\ down\_left) = 0$. The MDP action space can be factored into $\mathcal{A}^1 = \{left, right\}$ and $\mathcal{A}^2 = \{up, down\}$. The MDP based on $\mathcal{A}^1$, assuming $\beta = 0$, is in Figure 10. Here, there is no need to abstract the states according to the action as every action fully affects the state.

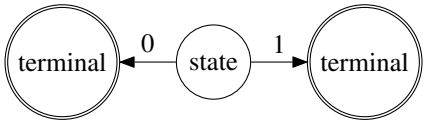

Figure 9. Factored MDP-1 in first factored action space (horizontal).

The MDP based on $\mathcal{A}^2$, assuming $\beta = 0$, is in Figure 10. It is clear that putting the actions from corresponding states together from the factored MDPs together and summing the coinciding rewards would give the overall MDP - this is possible because Theorem 1 is satisfied. This would not be the case if $\beta \neq 0$ - a situation which we investigate in the main paper.

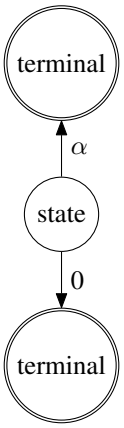

Figure 10. Factored MDP-1 in second factored action space (vertical).

G. Diagram of MDP 2 and its Factorisation

In Figure 11, the states are labelled circles, while actions from states are arrows that leave the state in a particular direction. For example, an arrow leaving a state in the upward-right direction represents the action up_right from that state. Each arrow is annotated with a number representing the reward of the *(state, action)* pair it represents i.e. $r(s, a)$. To clarify in case there is ambiguity for the reader the rewards are: $r(0, 0,\ up\_right) = 2$, $r(0, 0,\ up\_left) = 1$, $r(0, 0,\ down\_right) = 1$, $r(0, 0,\ down\_left) = 0$, $r(0, 1,\ up\_right) = 1$, $r(0, 1,\ up\_left) = -1$, $r(0, 1,\ down\_right) = 1$, $r(0, 1,\ down\_left) = 0$, $r(1, 0,\ up\_right) = 1$, $r(1, 0,\ up\_left) = 1$, $r(1, 0,\ down\_right) = -1$, $r(1, 0,\ down\_left) = 0$, $r(1, 1,\ up\_right) = -2$, $r(1, 1,\ up\_left) = 0$, $r(1, 1,\ down\_right) = 0$, $r(1, 1,\ down\_left) = 0$.

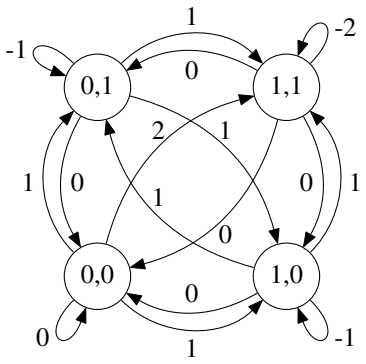

Figure 11. MDP-2 with corresponding rewards used in experiments.

This MDP has no terminal states, which enables the agent to follow trajectories of unlimited length. The MDP action space may be factored into two spaces: $\mathcal{A}^1 = \{left,\ right\}$ and $\mathcal{A}^2 = \{up,\ down\}$. The MDP based on $\mathcal{A}^1$ is in Figure 12. If Figure 11 is observed, it is seen that only the first number in the state label is relevant to the actions in $\mathcal{A}^1$; the other number is independent. Thus every state can be abstracted to only its first number e.g. $0, 0$ becomes $0, ?$ and $1, 0$ becomes $1, ?$.

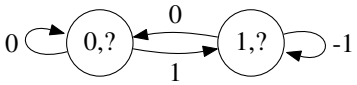

Figure 12. Factored MDP-2 in first factored action space (horizontal).

The MDP based on $\mathcal{A}^2$ is in Figure 13. Here, only the second number in the state label is relevant to the actions being taken. Thus every state is abstracted to its second number e.g. $0, 0$ becomes $?, 0$ and $0, 1$ becomes $?, 1$. It is clear to see that combinatorially putting the actions and state abstractions of the factored MDPs together and summing the coinciding rewards would give the overall MDP - this is possible because Theorem 1 is satisfied.

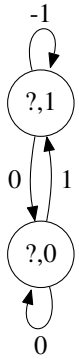

Figure 13. Factored MDP-2 in second factored action space (vertical).

H. Policies Used in Experiments

Throughout experiments, the policy divergence, as defined in Equation 16, is varied by specifically designing the behaviour policy π_b and evaluation policy π_e. In this section, we discuss the exact policies utilised for each MDP, in order to generate each recorded policy divergence value. Note that all policy divergence values used throughout this paper have been rounded to 2 decimal places.

H.1. In MDP 1

The states of MDP 1 are $\mathcal{S} = \{state, terminal\}$. The action space is $\mathcal{A} = \{up\_right,\ up\_left,\ down\_right,\ down\_left\}$. There is only one transition, hence any policy would only be defined at $state$.

The first factored action space $\mathcal{A}^1 = \{right,\ left\}$, while if $\phi_d : \mathcal{S} \to \mathcal{Z}^d$ maps a state to its abstraction with respect to the d^{th} factored action space, then $\phi^1(state) = state$ and $\phi^1(terminal) = terminal$.

The second factored action space $\mathcal{A}^2 = \{up,\ down\}$, while $\phi^2(state) = state$ and $\phi^2(terminal) = terminal$.

H.1.1. Policy Divergence = 1.44

Behaviour Policy:

$\pi_b(up\_right|state) = \pi_b(up\_left|state) = \pi_b(down\_right|state) = \pi_b(down\_left|state) = 0.25.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = \pi_b^1(left|state) = 0.5$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = \pi_b^2(down|state) = 0.5$.

Evaluation Policy:

$\pi_e(up\_right|state) = 0.36,\ \pi_e(up\_left|state) = 0.24,\ \pi_e(down\_right|state) = 0.24,\ \pi_e(down\_right|state) = 0.16.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.6,\ \pi_e^1(left|state) = 0.4$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.6,\ \pi_e^2(down|state) = 0.4$.

H.1.2. Policy Divergence = 2.56

Behaviour Policy:

$\pi_b(up\_right|state) = \pi_b(up\_left|state) = \pi_b(down\_right|state) = \pi_b(down\_left|state) = 0.25.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = \pi_b^1(left|state) = 0.5$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = \pi_b^2(down|state) = 0.5$.

Evaluation Policy:

$\pi_e(up\_right|state) = 0.64,\ \pi_e(up\_left|state) = 0.16,\ \pi_e(down\_right|state) = 0.16,\ \pi_e(down\_right|state) = 0.04.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.8,\ \pi_e^1(left|state) = 0.2$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.8,\ \pi_e^2(down|state) = 0.2$.

H.1.3. Policy Divergence = 3.61

Behaviour Policy:

$\pi_b(up\_right|state) = \pi_b(up\_left|state) = \pi_b(down\_right|state) = \pi_b(down\_left|state) = 0.25.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = \pi_b^1(left|state) = 0.5$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = \pi_b^2(down|state) = 0.5$.

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) =$

$0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.4. POLICY DIVERGENCE = 4.46

Behaviour Policy:

$\pi_b(up\_right|state) = 0.2025,\ \pi_b(up\_left|state) = 0.2475,\ \pi_b(down\_right|state) = 0.2475,\ \pi_b(down\_left|state) = 0.3025.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.45,\ \pi_b^1(left|state) = 0.55.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.45,\ \pi_b^2(down|state) = 0.55.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.5. POLICY DIVERGENCE = 5.64

Behaviour Policy:

$\pi_b(up\_right|state) = 0.16,\ \pi_b(up\_left|state) = 0.24,\ \pi_b(down\_right|state) = 0.24,\ \pi_b(down\_left|state) = 0.36.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.4,\ \pi_b^1(left|state) = 0.6.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.4,\ \pi_b^2(down|state) = 0.6.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.6. POLICY DIVERGENCE = 10.03

Behaviour Policy:

$\pi_b(up\_right|state) = 0.09,\ \pi_b(up\_left|state) = 0.21,\ \pi_b(down\_right|state) = 0.21,\ \pi_b(down\_left|state) = 0.49.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.3,\ \pi_b^1(left|state) = 0.7.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.3,\ \pi_b^2(down|state) = 0.7.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.7. POLICY DIVERGENCE = 22.56

Behaviour Policy:

$\pi_b(up\_right|state) = 0.04,\ \pi_b(up\_left|state) = 0.16,\ \pi_b(down\_right|state) = 0.16,\ \pi_b(down\_left|state) = 0.64.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.2,\ \pi_b^1(left|state) = 0.8.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.2,\ \pi_b^2(down|state) = 0.8.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.8. POLICY DIVERGENCE = 90.25

Behaviour Policy:

$\pi_b(up\_right|state) = 0.01,\ \pi_b(up\_left|state) = 0.09,\ \pi_b(down\_right|state) = 0.09,\ \pi_b(down\_left|state) = 0.81.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.1,\ \pi_b^1(left|state) = 0.9.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.1,\ \pi_b^2(down|state) = 0.9.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.1.9. POLICY DIVERGENCE = 361.0

Behaviour Policy:

$\pi_b(up\_right|state) = 0.0025,\ \pi_b(up\_left|state) = 0.0475,\ \pi_b(down\_right|state) = 0.0475,\ \pi_b(down\_left|state) = 0.9025.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|state) = 0.05,\ \pi_b^1(left|state) = 0.95.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|state) = 0.05,\ \pi_b^2(down|state) = 0.95.$

Evaluation Policy:

$\pi_e(up\_right|state) = 0.9025,\ \pi_e(up\_left|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0475,\ \pi_e(down\_right|state) = 0.0025.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|state) = 0.95,\ \pi_e^1(left|state) = 0.05.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|state) = 0.95,\ \pi_e^2(down|state) = 0.05.$

H.2. In MDP 2

The states of MDP 1 are $\mathcal{S} = \{0,0,\ 0,1,\ 1,0,\ 1,1\}$. The policy must be defined at all these states. The action space is $\mathcal{A} = \{up\_right,\ up\_left,\ down\_right,\ down\_left\}$. There can be unlimited transitions, we denote the number of transitions i.e. trajectory length as T. The policy divergence is thus represented as P^T, where P is the policy divergence for 1 transition.

The first factored action space $\mathcal{A}^1 = \{right,\ left\}$, while if $\phi_d : \mathcal{S} \to \mathcal{Z}_d$ maps a state to its abstraction with respect to the d^{th} factored action space, then $\phi^1(0,0) = \phi^1(0,1) = 0, ?$ and $\phi^1(1,0) = \phi^1(1,1) = 1, ?$.

The second factored action space $\mathcal{A}^2 = \{up,\ down\}$, while $\phi^2(0,0) = \phi^2(1,0) = ?, 0$ and $\phi^2(0,1) = \phi^2(1,1) = ?, 1$.

H.2.1. Policy Divergence = 1.44^T

Behaviour Policy:

$\pi_b(up\_right|s) = \pi_b(up\_left|s) = \pi_b(down\_right|s) = \pi_b(down\_left|s) = 0.25,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = \pi_b^1(left|z^1) = 0.5,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = \pi_b^2(down|z^2) = 0.5,\ \forall\ z^2 \in \mathcal{Z}^2$.

Evaluation Policy:

$\pi_e(up\_right|s) = 0.36,\ \pi_e(up\_left|s) = 0.24,\ \pi_e(down\_right|s) = 0.24,\ \pi_e(down\_right|s) = 0.16,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.6,\ \pi_e^1(left|z^1) = 0.4,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.6,\ \pi_e^2(down|z^2) = 0.4,\ \forall\ z^2 \in \mathcal{Z}^2$.

H.2.2. Policy Divergence = 2.56^T

Behaviour Policy:

$\pi_b(up\_right|s) = \pi_b(up\_left|s) = \pi_b(down\_right|s) = \pi_b(down\_left|s) = 0.25,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = \pi_b^1(left|z^1) = 0.5,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = \pi_b^2(down|z^2) = 0.5,\ \forall\ z^2 \in \mathcal{Z}^2$.

Evaluation Policy:

$\pi_e(up\_right|s) = 0.64,\ \pi_e(up\_left|s) = 0.16,\ \pi_e(down\_right|s) = 0.16,\ \pi_e(down\_right|s) = 0.04,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.8,\ \pi_e^1(left|z^1) = 0.2,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.8,\ \pi_e^2(down|z^2) = 0.2,\ \forall\ z^2 \in \mathcal{Z}^2$.

H.2.3. Policy Divergence = 3.61^T

Behaviour Policy:

$\pi_b(up\_right|s) = \pi_b(up\_left|s) = \pi_b(down\_right|s) = \pi_b(down\_left|s) = 0.25,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = \pi_b^1(left|z^1) = 0.5,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = \pi_b^2(down|z^2) = 0.5,\ \forall\ z^2 \in \mathcal{Z}^2$.

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2$.

H.2.4. POLICY DIVERGENCE = 4.46^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.2025,\ \pi_b(up\_left|s) = 0.2475,\ \pi_b(down\_right|s) = 0.2475,\ \pi_b(down\_left|s) = 0.3025,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.45,\ \pi_b^1(left|z^1) = 0.55,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.45,\ \pi_b^2(down|z^2) = 0.55,\ \forall\ z^2 \in \mathcal{Z}^2.$

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2.$

H.2.5. POLICY DIVERGENCE = 5.64^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.16,\ \pi_b(up\_left|s) = 0.24,\ \pi_b(down\_right|s) = 0.24,\ \pi_b(down\_left|s) = 0.36,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.4,\ \pi_b^1(left|z^1) = 0.6,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.4,\ \pi_b^2(down|z^2) = 0.6,\ \forall\ z^2 \in \mathcal{Z}^2.$

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2.$

H.2.6. POLICY DIVERGENCE = 10.03^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.09,\ \pi_b(up\_left|s) = 0.21,\ \pi_b(down\_right|s) = 0.21,\ \pi_b(down\_left|s) = 0.49,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.3,\ \pi_b^1(left|z^1) = 0.7,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.3,\ \pi_b^2(down|z^2) = 0.7,\ \forall\ z^2 \in \mathcal{Z}^2.$

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2.$

H.2.7. POLICY DIVERGENCE = 22.56^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.04,\ \pi_b(up\_left|s) = 0.16,\ \pi_b(down\_right|s) = 0.16,\ \pi_b(down\_left|s) = 0.64,\ \forall\ s \in \mathcal{S}.$

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.2,\ \pi_b^1(left|z^1) = 0.8,\ \forall\ z^1 \in \mathcal{Z}^1.$

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.2,\ \pi_b^2(down|z^2) = 0.8,\ \forall\ z^2 \in \mathcal{Z}^2.$

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2$.

H.2.8. POLICY DIVERGENCE = 90.25^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.04,\ \pi_b(up\_left|s) = 0.16,\ \pi_b(down\_right|s) = 0.16,\ \pi_b(down\_left|s) = 0.81,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.1,\ \pi_b^1(left|z^1) = 0.9,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.1,\ \pi_b^2(down|z^2) = 0.9,\ \forall\ z^2 \in \mathcal{Z}^2$.

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2$.

H.2.9. POLICY DIVERGENCE = 361.0^T

Behaviour Policy:

$\pi_b(up\_right|s) = 0.0025,\ \pi_b(up\_left|s) = 0.0475,\ \pi_b(down\_right|s) = 0.0475,\ \pi_b(down\_left|s) = 0.9025,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_b^1(right|z^1) = 0.05,\ \pi_b^1(left|z^1) = 0.95,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_b^2(up|z^2) = 0.05,\ \pi_b^2(down|z^2) = 0.95,\ \forall\ z^2 \in \mathcal{Z}^2$.

Evaluation Policy:

$\pi_e(up\_right|s) = 0.9025,\ \pi_e(up\_left|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0475,\ \pi_e(down\_right|s) = 0.0025,\ \forall\ s \in \mathcal{S}$.

With respect to $\mathcal{A}^1$, $\pi_e^1(right|z^1) = 0.95,\ \pi_e^1(left|z^1) = 0.05,\ \forall\ z^1 \in \mathcal{Z}^1$.

With respect to $\mathcal{A}^2$, $\pi_e^2(up|z^2) = 0.95,\ \pi_e^2(down|z^2) = 0.05,\ \forall\ z^2 \in \mathcal{Z}^2$.

I. Additional Plots From Experiments

I.1. MDP 1

I.1.1. Varying the Number of Trajectories in MDP 1 With Policy Divergence 1.44

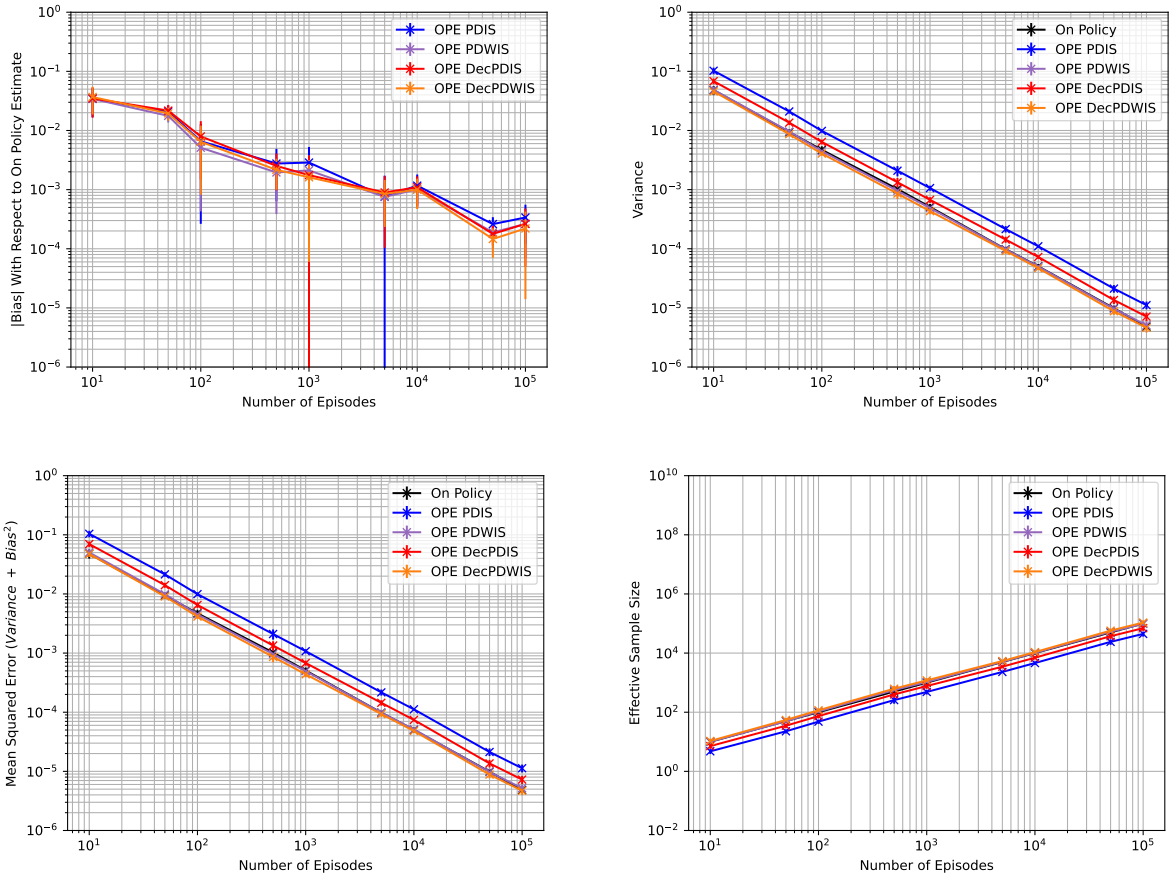

Figure 14. The biases of all estimators are very low and gradually converge to a lower limit of zero bias. Since this is a 1-step problem, the bias is low even for the PDWIS estimators. The variance shows similar behaviour to Figure 1, except that the graphs are closer together i.e. less difference between the variances of the policies due to lower policy divergence. MSE is dominated by the variance behaviour and is thus similar to Figure 1. The ESS graphs look similar to the variance graphs inverted about the x-axis, and have similar but inverted trends.

I.1.2. Varying the Number of Trajectories in MDP 1 With Policy Divergence 2.56

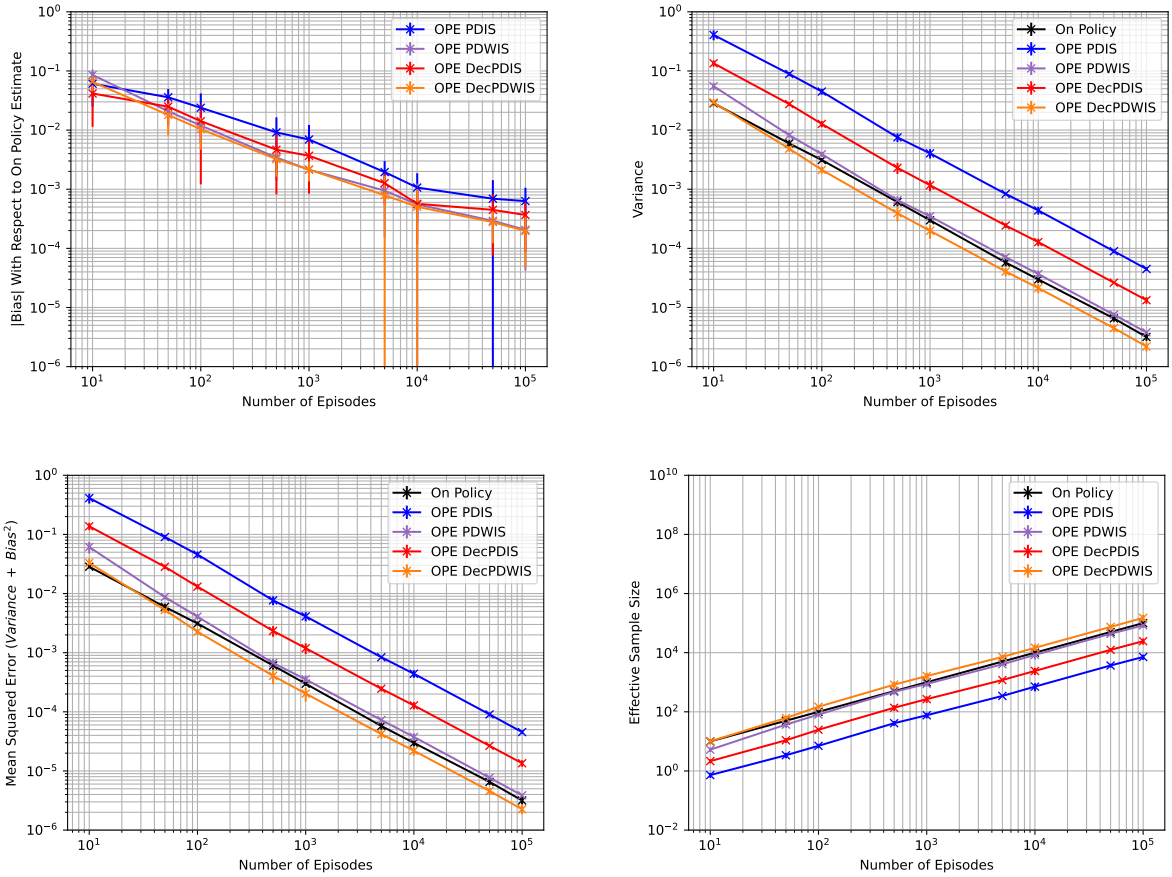

Figure 15. The overall bias, variance, MSE and ESS behaviour of the estimators is highly similar to Figure 14. However, the graphs are more spread apart; this could due to the increased policy divergence. This makes OPE estimation more difficult and results in the distinction between estimators becoming more clear. In the graphs, it is clear that the PDWIS estimators are performing better in this problem. Note that the variance and MSE graphs are the same as Figure 1 in the main paper, while the ESS graph is the same as Figure 6.

I.1.3. Varying the Number of Trajectories in MDP 1 With Policy Divergence 3.61

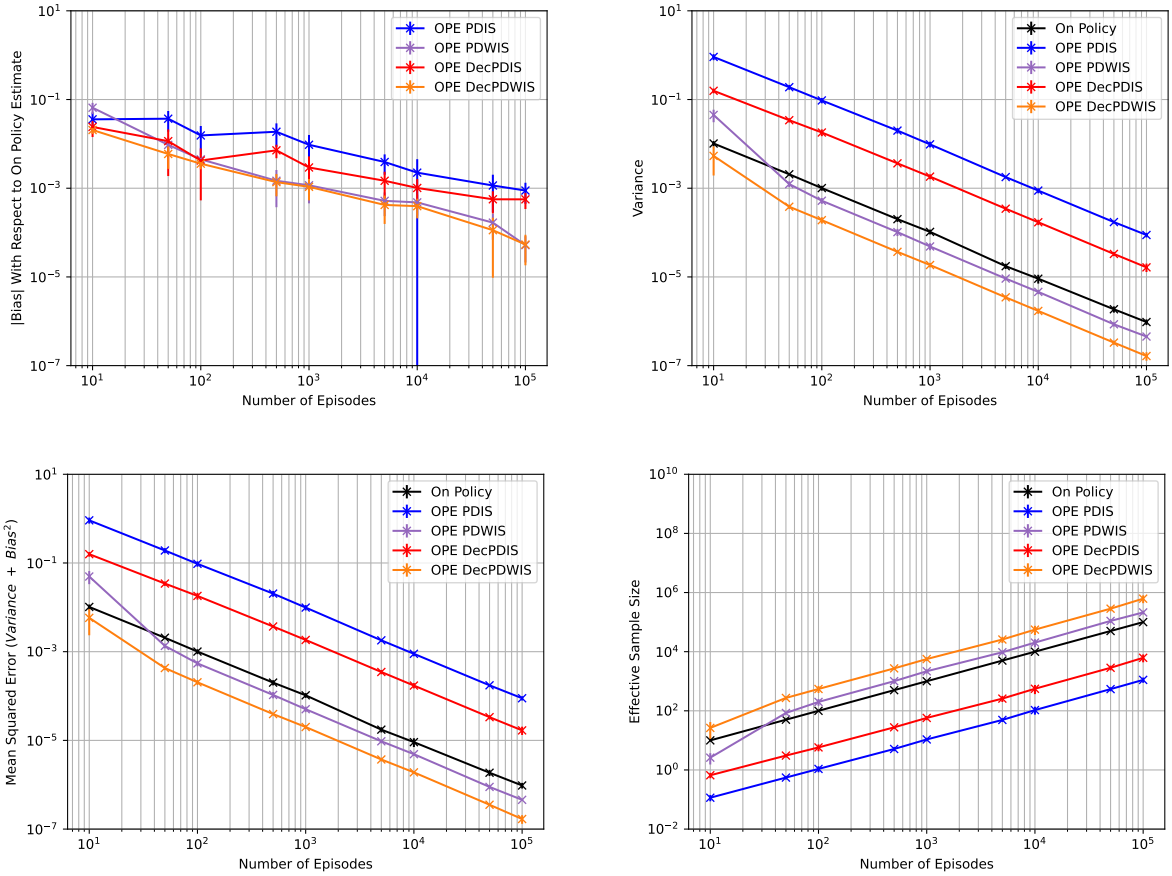

Figure 16. The bias, variance, MSE and ESS graphs again show almost identical trends to Figures 14 and 15, with the graphs being more spread out due to greater policy divergence. The superiority of PDWIS estimators over all other OPE methods, and over on-policy estimation, is clearer.

I.1.4. VARYING β IN MDP-1 WITH POLICY DIVERGENCE 1.44 AND 1000 TRAJECTORIES

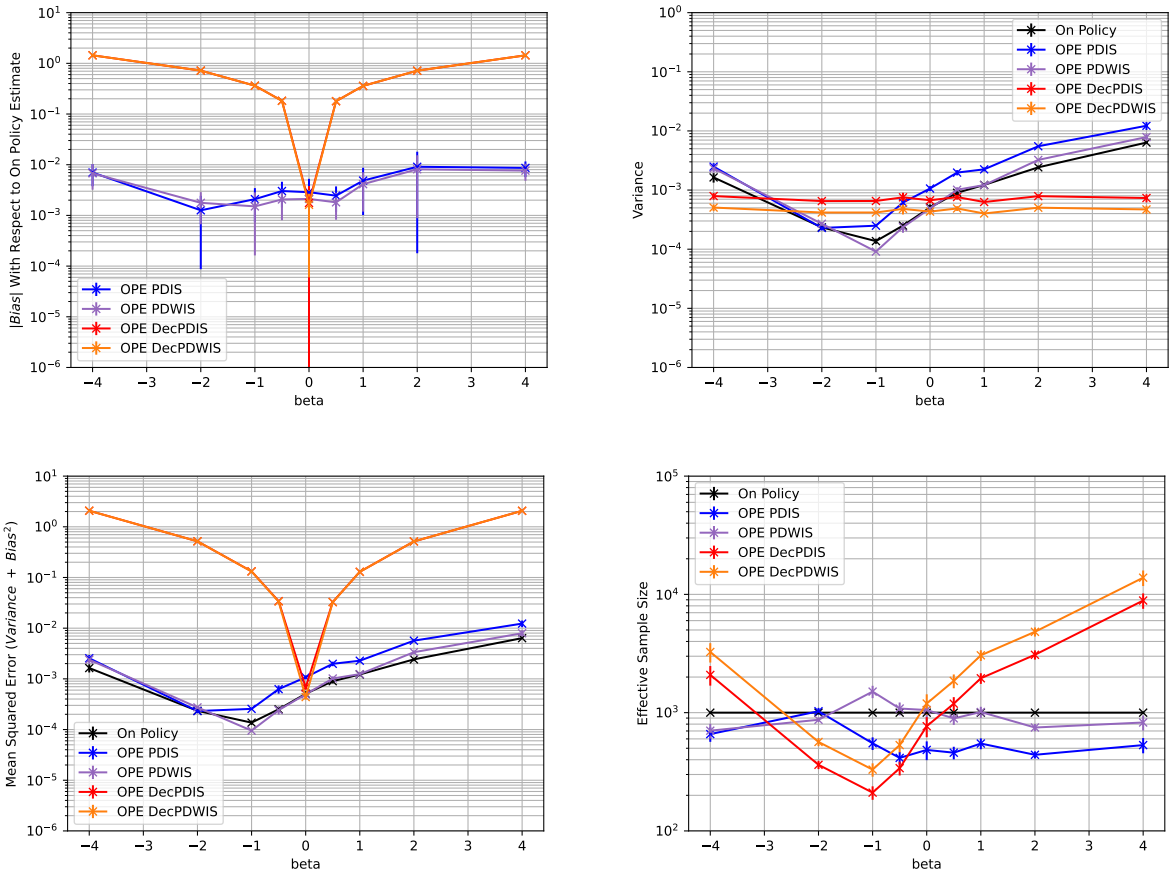

Figure 17. The bias graphs are similar to Figure 2 in the main paper. Similar to that figure, we observe that for $\beta = 0$, the bias of the decomposed estimators is comparable to that of the non-decomposed estimators. However, as we increase $|\beta|$, the bias of the decomposed estimators scales so dramatically that this behaviour dominates the MSE performance of these estimators. This is because for $|\beta| \neq 0$, Theorem 1 no longer applies and thus it is not assured that the MDP can be factored into sub-MDP's based on factored action spaces without introducing bias. It is notable that the MSE of non-decomposed estimators is dominated by variance, rather than bias. Meanwhile, β affects the variance of the on-policy and non-decomposed estimators, and does not affect that of the decomposed estimators. The presence of a variance minimum at $\beta = -1$ is due to the equal reward in the MDP for three different actions - see Appendix F. The further we depart from this equal-reward scenario, the more the variance of the on-policy and non-decomposed estimators increase. The relative trajectories of the OPE and on-policy estimator variance graphs determine the ESS graph.

I.1.5. Varying β in MDP-1 With Policy Divergence 1.44 and 100,000 Trajectories

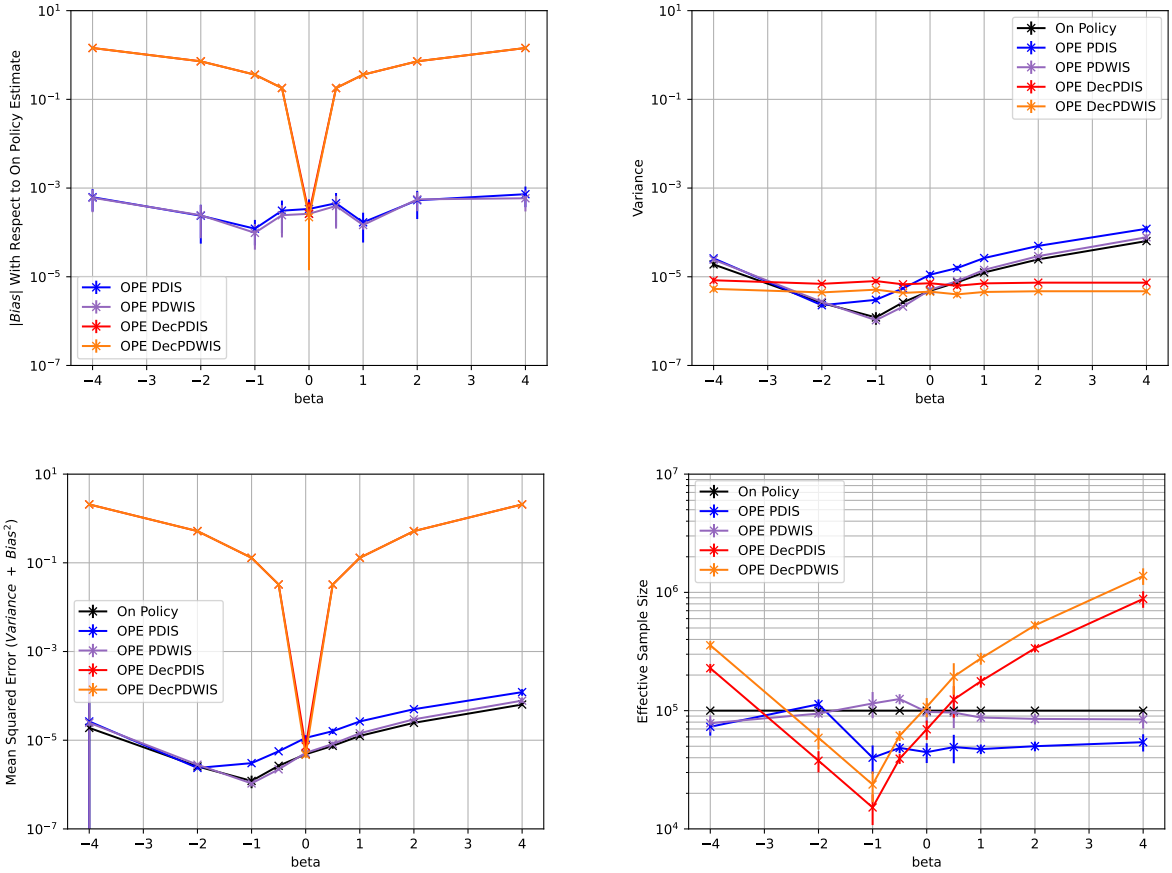

Figure 18. These graphs are highly similar to Figure 17, except that the variance and MSE are at lower scales and the ESS is at a higher scale.

I.1.6. Varying Policy Divergence in MDP-1 With 1000 Trajectories

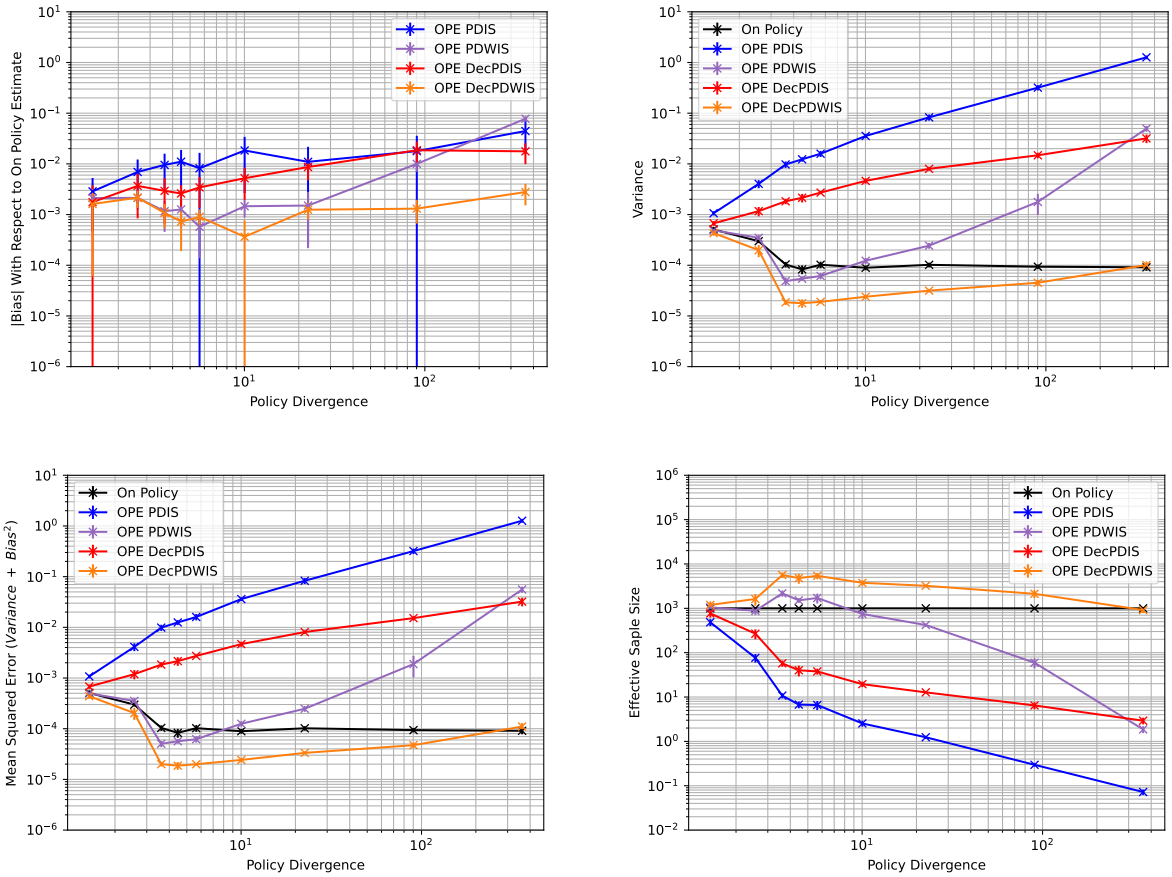

Figure 19. As the policy divergence increases, bias, with some random anomalies, increases due to decreasing coverage of the evaluation policy by the behaviour policy dataset. The variances of PDIS and decomposed PDIS also increase with policy divergence for the same reason, however decomposed PDIS always has lower variance and also scales slower than non-decomposed PDIS. Clearly, decomposed PDIS has ensured better coverage here. On the other hand, the variance of the PDWIS estimators first fall and then increase - this may be due to the sum of IS weights in the denominator initially becoming larger due to policy divergence then becoming smaller due to lower coverage. The ESS again inverts the variance graph, while the MSE is primarily affected by the variance.

I.1.7. Varying Policy Divergence in MDP-1 With 100,000 Trajectories

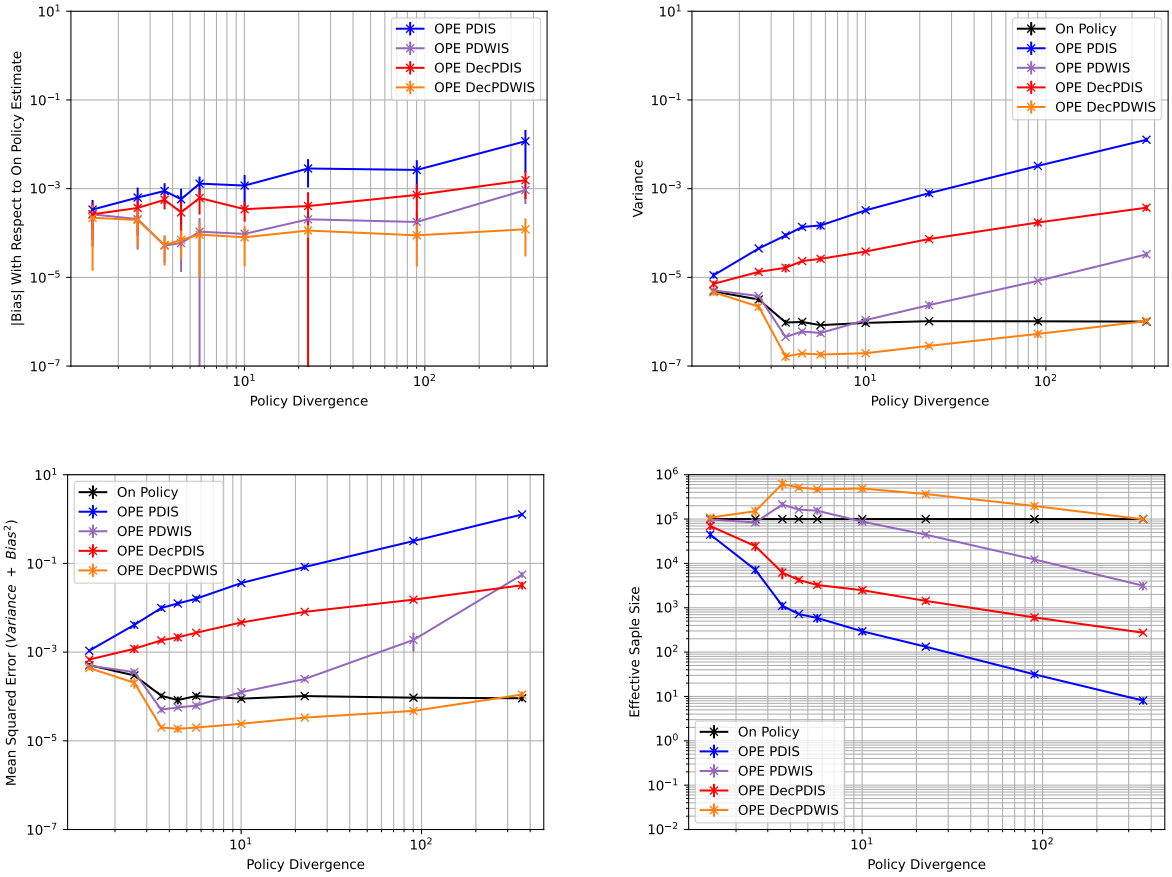

Figure 20. These graphs are highly similar to those in Figure 19, except that the variance and ESS are on different scales i.e. variance is less and ESS is more.

I.2. MDP 2

I.2.1. Varying the Trajectory Length T in MDP 2 With Policy Divergence 1.44^T, 1000 Trajectories and Discount Factor 0.7

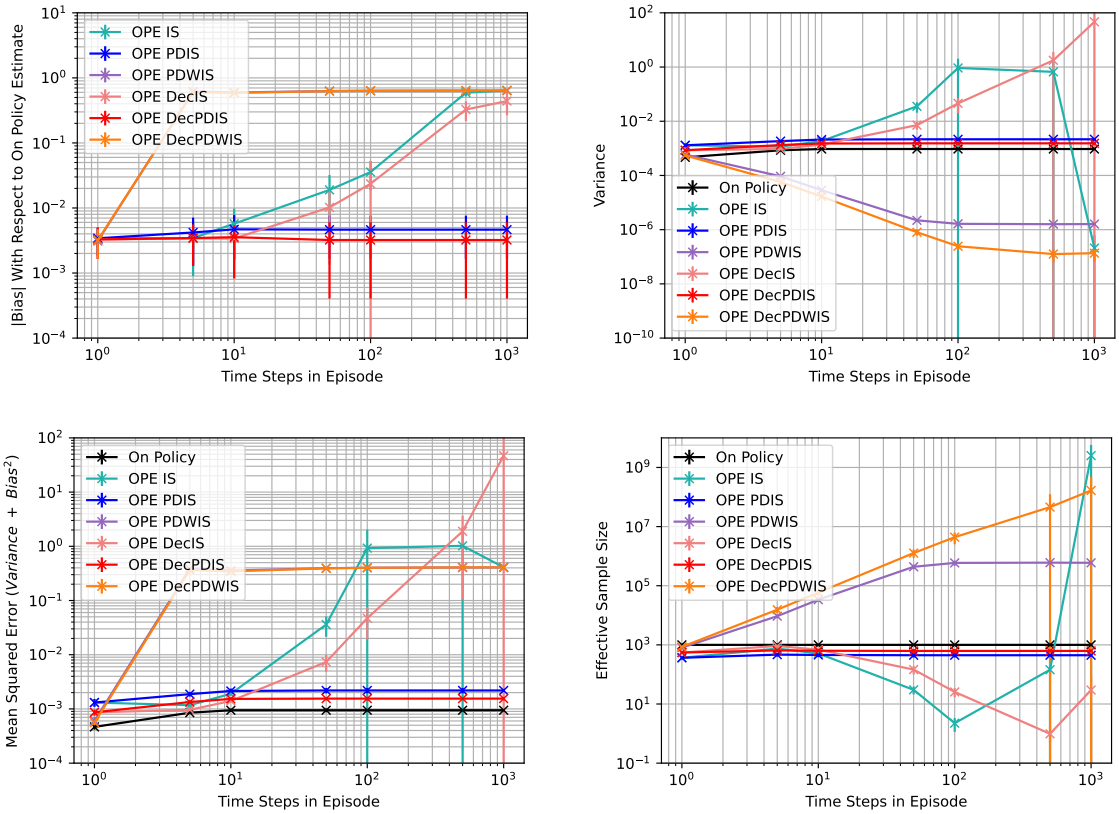

Figure 21. These graphs are the same as those in figures 3, 4 and 7 in the main report. They are discussed at length in the main report. The decrease in variance of the PDWIS estimators with T may be due to the IS weights in the denominator becoming larger in magnitude.

I.2.2. Varying the Trajectory Length T in MDP 2 With Policy Divergence 1.44^T, 1000 Trajectories and Discount Factor 0.9

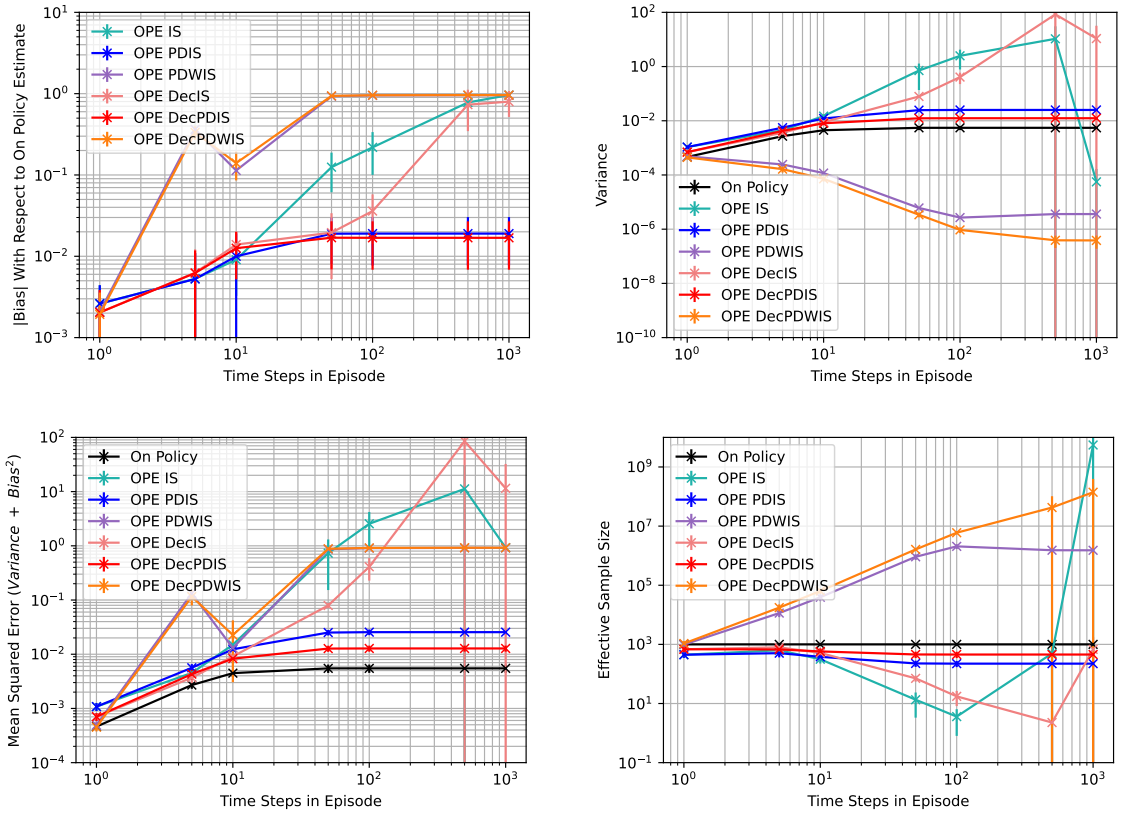

Figure 22. These graphs are similar to Figure 21 with some differences. These differences are observed because the larger discount factor enables transitions further into the future to contribute to the Q-function estimate. The bias and variance of the PDIS estimators increase more before reaching their constant values. The variances of the IS estimators also scale faster and reach the point where the estimate is a low-variance, high bias value. The biases of the PDWIS estimators actually scale more slowly; this may be because the IS weight sum in the denominator is larger in magnitude for larger γ, which scales down the variance. Finally the gap between the MSE of the estimators and that of the on-policy estimate has increased.

I.2.3. Varying the Trajectory Length T in MDP 2 With Policy Divergence 1.44^T, 1000 Trajectories and Discount Factor 0.9999

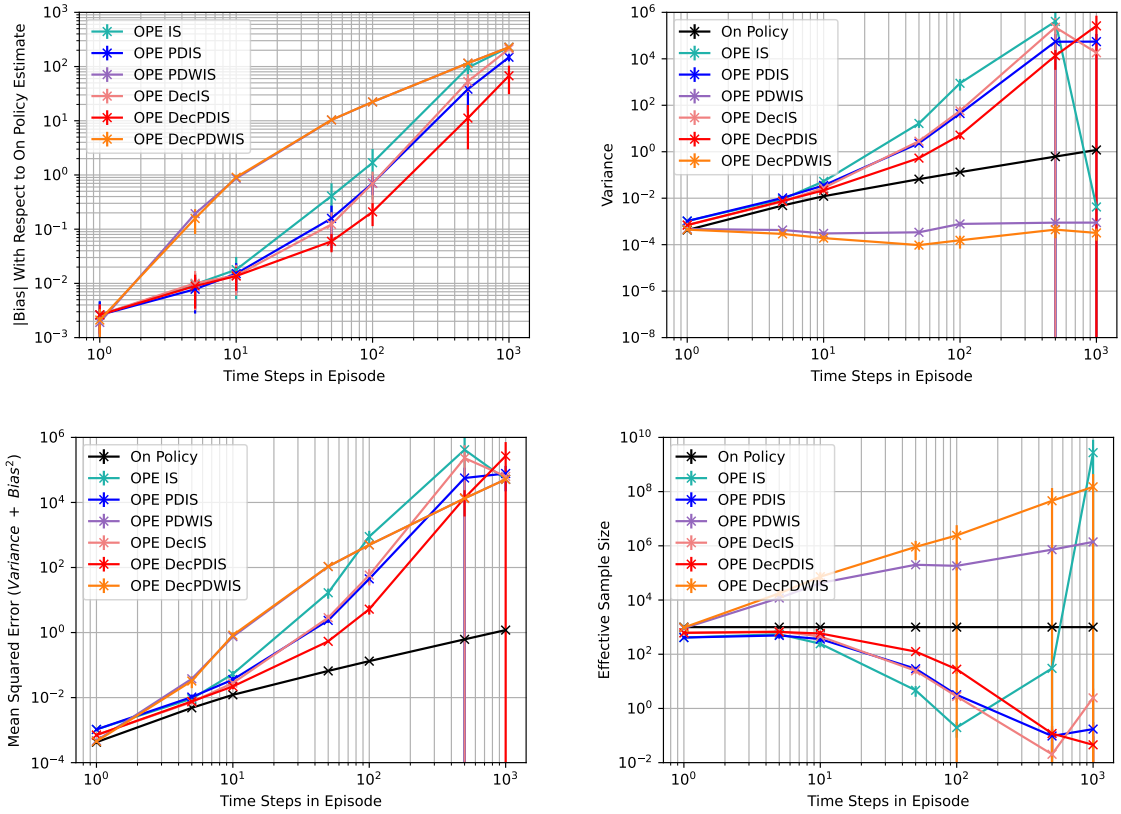

Figure 23. Here, nearly all transitions, even those far into the future, have a high contribution to the value function estimate. It is clear that a high discount factor causes rapid scaling in estimator variance and loss of coverage which leads to scaling in estimator bias. The MSE's of the estimators thus scale faster than the on-policy MSE. Interestingly, however, the variances of the PDWIS estimators seem relatively unaffected by variation in the discount factor; perhaps the effect of loss in coverage is counterbalanced by the effect of denominator weighting.

I.2.4. Varying Policy Divergence in MDP 2 With 1000 Trajectories of Length 10 and Discount Factor 0.7

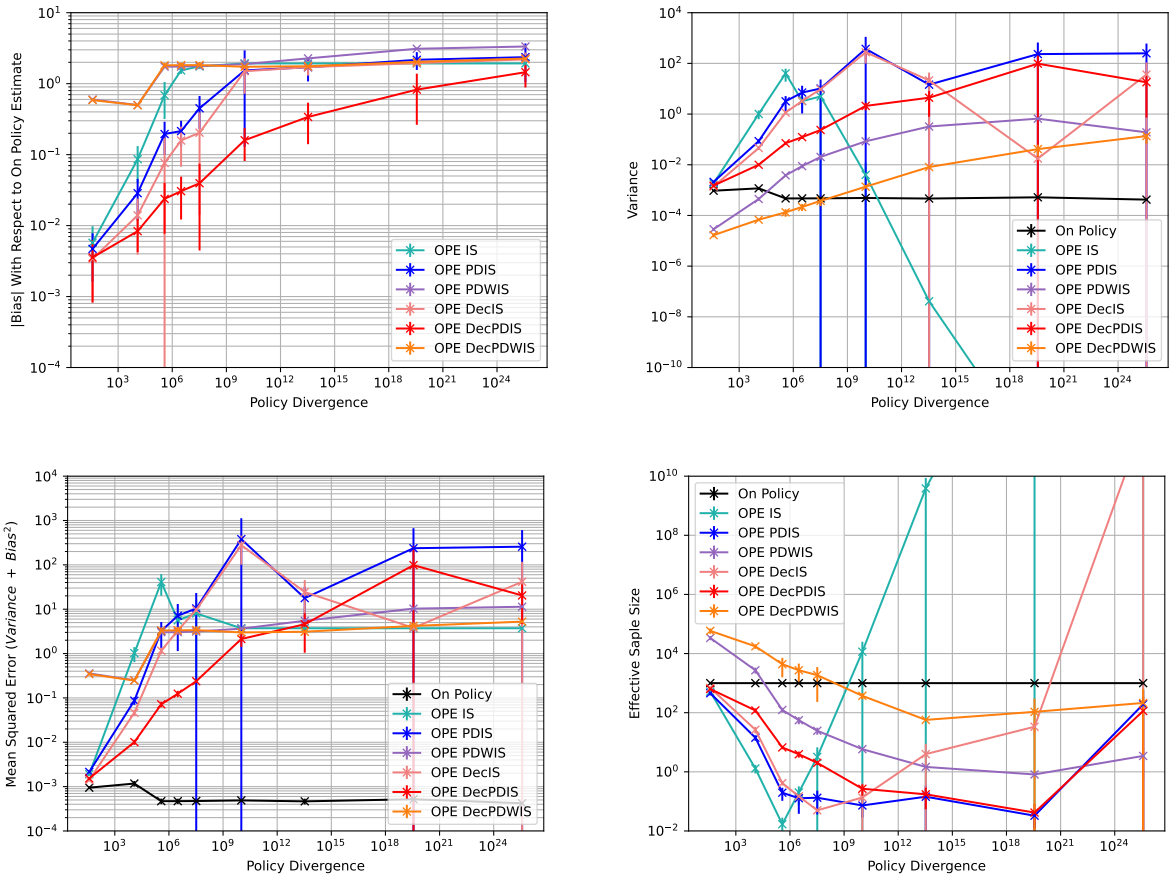

Figure 24. The bias and variance graphs are the same as those in Figure 5 from the main report and have been discussed there in detail. A notable insight is how rapidly coverage of the evaluation policy decreases as policy divergence increases. This can be seen from the rapid scaling in bias, and the variance graphs turning downwards after initial growth. The MSE behaviour is dominated by bias, while ESS seems almost identical to the variance graph but inverted.