Doubly Robust Augmented Transfer for Meta-Reinforcement Learning

A. Appendix

A.1. Related Work

Meta-Reinforcement Learning (Meta-RL). With the incorporation of meta-learning, meta-RL enables a fast adaptation in RL problems through the idea of "learning to learn". During meta-training, meta-RL learns an inductive bias from a set of relative training tasks for quickly adapting to some new tasks, given only a small amount of samples at the meta-test time. Current meta-RL methods can be classified in to two categories. One is the gradient-based method, which attempts to use a few number of gradient updates to implement the adaptation on a new task, such as by using policy gradient methods to directly update the policy parameters [12, 17, 18, 19]. The other is the context-based method, which builds up an inference network to infer task-specific latent context variables from the input-sampled experience (i.e., context) of the training tasks. A policy with both state and latent variable as input is also trained to maximize rewards on these training tasks, hence the adaptation is conducted by the latent context inference first, followed by the policy adjustment with the inferred latent context as input. These methods mainly differ in their ways of inference [3, 4, 20]. However, sparse reward remains a challenge in meta-RL, where the sparse reward signals provide only scarce task-relevant information and make meta-training and adaptation extremely difficult.

Sparse-Reward Meta-RL. To tackle the sparse reward problem in meta-RL, two main research lines have been developed recently. One directly generates informative samples by exploration [21, 22, 23] or by directly using the demonstration datasets [24]. For instance, training a separate exploration policy by maximizing the information gain or intrinsic rewards to collect samples. The other line follows the technique of relabeling that enables sample reuse across tasks, i.e., learning a task at hand by appropriately reusing the samples generated from other tasks. Compared with sample exploration, sample reuse has several advantages, such as no extra exploration, high sample efficiency, and low sample risk. Following the direction of relabeling, hindsight experience replay (HER) [5] has been studied as one typical method, which is originally designed for the multi-goal setting and relabels a trajectory with a lower reward under its original goal to a goal that has higher reward. Packer *et al.* apply hindsight relabeling for meta-RL, and propose hindsight task relabeling (HTR) to relabel the trajectories sampled from one task to a task which can be accomplished in these trajectories with higher rewards [14]. However, like its application on goal-conditioned tasks, this method can only cope with training tasks with different reward functions that correspond to the goals. Taking a step further than hindsight relabelling, Wan *et al.* introduce additionally foresight relabeling to meta-RL, and propose to relabel trajectories to new tasks with higher post-adaptation rewards [15].

Doubly Robust Estimator. Doubly robust (DR) is first presented in statistics [25, 26] and then brought into RL by Jiang *et al.* for policy evaluation [7], which combines the direct learning of dynamics models and importance sampling to provide an unbiased and lower-variance value estimate. The variance of DR in value evaluation can be further reduced by applying lower-variance IS estimator [8, 27] and through learning an more accurate dynamics model [28]. For policy learning in RL, Huang *et al.* derive a general form of policy gradient from DR value estimator [29], whereas a DR off-policy actor-critic method is developed by Xu *et al.* [30]. Kallus *et al.* propose the doubly robust method to find a robust policy that can achieve the near-optimum in the worst case under environment distribution shifts [31]. Similar to our work which aims at optimizing MSE of the DR estimator, Su *et al.* derive a shrinkaged importance weight of policy for bandit problem under the assumption of known importance weights, while we do not have access to the true importance weight of dynamics. Different from these works, we apply doubly robust (DR) to transfer the experience collected across a distribution of tasks, for accelerating the value function learning under a challenging sparse-reward meta-RL setting.

Transfer Learning for Meta-RL. Our problem setting partly falls into the area of transfer in RL, which aims to accelerate the learning process in a new target task by transferring knowledge learned from the source tasks. Depending on the knowledge to be transferred, these methods in RL can be roughly divided into classes including sampled transitions [32, 33], learned policies or value networks [34, 35, 36, 37], features [38, 39, 40], and skills [41, 42]. Tirinzoni *et al.* apply importance sampling (IS) to transfer samples from a set of source tasks [32], while multiple IS that has a lower variance is applied in [33]. Our method implements transition transfer by doubly robust methods, which can be proved to have a lower variance than these IS methods.

A.2. Decomposition of MSE in Eq. (4) in the main text

$$
\begin{aligned}
\text{MSE}(\hat{V}) =& \mathbb{E}_{\tau_i|t:T}\left[\left(V_j(s_t) - \hat{V}_j(s_t)\right)^2 | s_t = s\right] \\
=& \mathbb{E}_{\tau_i|t:T}\left[(V_j(s_t))^2 - 2V_j(s_t)\hat{V}_j(s_t) + (\hat{V}_j(s_t))^2 \big| s_t\right] \\
=& \mathbb{E}_{\tau_i|t:T}\left[(V_j(s_t))^2 - 2V_j(s_t)\hat{V}_j(s_t) + (\mathbb{E}_{\tau_i|t:T}\left[\hat{V}_j(s_t)\big|s_t\right)^2 | s_t\right] + \mathbb{E}_{\tau_i|t:T}\left[(V_j(s_t))^2 | s_t\right] - \left(\mathbb{E}_{\tau_i|t:T}\left[\hat{V}_j(s_t)\big|s_t\right]\right)^2 \\
=& \left(V^j(s_t) - \mathbb{E}_{\tau_i|t:T}\left[\hat{V}^j(s_t)\big|s_t\right]\right)^2 + \text{Var}(\hat{V}^j(s_t)) = \text{Bias}(\hat{V})^2 + \text{Var}(\hat{V}^j(s_t))
\end{aligned}
$$

A.3. Doubly Robust Property for Direct Use of Doubly Robust Estimator

We show the doubly robust property of the DR estimator for value function in Eq. (5) in the main text, as follows.

1) In the first case when the importance weight ρ_π and ρ_d are correctly estimated and given the state s_t at time step t, taking the expectation on the RHS of Eq. (5) in the main text w.r.t. a_t and s_{t+1}, we have

$$
\begin{aligned}
& \mathbb{E}_{\substack{\pi_\theta(a_t|s_t,z_i) \\ p_i(s_{t+1}|s_t,a_t)}}\left[V_\theta(s_t, z_j) + \rho_\pi^{ij}(t)\left[r_j(s_t, a_t) + \rho_d^{ij}(t+1)\gamma V_{ij}^{DR}(s_{t+1}) - Q_\theta(s, a, z_j)\right]\right] \\
=& V_\theta(s_t, z_j) + \mathbb{E}_{\substack{\pi_\theta(a_t|s_t,z_i) \\ p_i(s_{t+1}|s_t,a_t)}}\left[\rho_\pi^{ij}(t)\left(r_j(s_t, a_t) + \rho_d^{ij}(t+1)\gamma V_{ij}^{DR}(s_{t+1}) - Q_\theta(s, a, z_j)\right)\right] \\
=& V_\theta(s_t, z_j) + \mathbb{E}_{\substack{\pi_\theta(a_t|s_t,z_j) \\ p_j(s_{t+1}|s_t,a_t)}}\left[r_j(s_t, a_t) + \gamma V_{ij}^{DR}(s_{t+1}) - Q_\theta(s_t, a_t, z_j)\right] \\
=& \mathbb{E}_{\pi_\theta(a_t|s_t,z_j)}\left[r_j(s_t, a_t) + \gamma \mathbb{E}_{p_j(s_{t+1}|s_t,a_t)} V_{ij}^{DR}(s_{t+1})\right],
\end{aligned}
$$

where the last equality follows $V_\theta(s_t, z_j) = \mathbb{E}_{a_t \sim \pi_\theta(\cdot|s_t,z_j)}\left[Q_\theta(s_t, a_t, z_j)\right]$ and reduces to the Bellman equation, which is the correct value for state s_t's value in the target task j.

2) In the other case when $Q_\theta(s_t, a_t, z_j)$ is a correct estimate of the action-state value, namely,

$$
\hat{Q}(s_t, a_t, z_j) = r_j(s_t, a_t) + \gamma \mathbb{E}_{p_j(s_{t+1}|s_t,a_t)}\left[V_{ij}^{DR}(s_{t+1})\right],
$$

which makes the expectation of the second term in Eq. (5) in the main text become zero, then the remaining non-zero term $V_\theta(s_t, z_j)$ is a proper estimate for the state value since recursively expending V_{ij}^{DR} will result in the definition of Q-value function.

A.4. Variance of biased DR estimator using $\hat{\rho}_d$

We firstly derive the variance of biased DR estimator $\tilde{V}^{DR}$ using an arbitrary importance weight $\hat{\rho}_d$. Let $\delta = \mathbb{E}_t[\tilde{V}_{ij}^{DR}(s_t) - V^j(s_t)]$ denote the difference between $\tilde{V}_{ij}^{DR}$ and V^j, hence the bias of $\tilde{V}_{ij}^{DR}$ by using $\hat{\rho}_d$ can be denoted as $Bias(\hat{\rho}) = |\delta|$. Then, the variance $Var_t\left[V_{ij}^{DR}(s_t)\right]$ can be obtained by letting $\hat{\rho}_d = \rho_d$. Given a certain state s_t, namely, the distribution is conditioned on s_t, we thus have

$$
\begin{aligned}
& Var_t\left[\tilde{V}_{ij}^{DR}(s_t)\right] \\
=& \mathbb{E}_t[\tilde{V}_{ij}^{DR}(s_t)^2] - (\mathbb{E}_t[\tilde{V}_{ij}^{DR}(s_t)])^2 \\
=& \mathbb{E}_t[\tilde{V}_{ij}^{DR}(s_t)^2] - (\mathbb{E}_t[V^j(s_t)] + \mathbb{E}_t[\delta])^2 \\
=& \mathbb{E}_t[\tilde{V}_{ij}^{DR}(s_t)^2] - (\mathbb{E}_t[V^j(s_t)])^2 - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \\
=& \mathbb{E}_t\left[\big(\bar{V}_\theta(s_t,z_j) + \rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1}) + \rho_\pi^{ij}(t)(r(s_t,a_t) - \bar{Q}_\theta(s_t,a_t,z_j))\big)^2 - V^j(s_t)^2\right] \\
& + Var_t\left[V^j(s_t)\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \\
=& \mathbb{E}_t\Big[\big(\bar{V}_\theta(s_t,z_j) + \rho_\pi^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1}) + \rho_\pi^{ij}(t)(r(s_t,a_t) - \bar{Q}_\theta(s_t,a_t,z_j)) \\
& + \rho_\pi^{ij}(t)(\hat{\rho}_d^{ij}(t)-1)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\big)^2 - V^j(s_t)^2\Big] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \qquad \text{(A.1)} \\
=& \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)Q^j(s_t,a_t) - \rho_\pi^{ij}(t)\bar{Q}_\theta(s_t,a_t,z_j) + \bar{V}_\theta(s_t,z_j) + \rho_\pi^{ij}(t)(r(s_t,a_t) + \gamma\tilde{V}_{ij}^{DR}(s_{t+1}) - Q^j(s_t,a_t))\right. \\
& \left. + \rho_\pi^{ij}(t)(\hat{\rho}_d^{ij}(t)-1)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\Big)^2 - V^j(s_t)^2\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \qquad \text{(A.2)} \\
=& \mathbb{E}_t\left[\Big((-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)) + \rho_\pi^{ij}(t)(r(s_t,a_t) - R(s_t,a_t)) + \rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})])\right. \\
& \left. + \rho_\pi^{ij}(t)(\hat{\rho}_d^{ij}(t)-1)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\Big)^2 - V^j(s_t)^2\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \qquad \text{(A.3)} \\
=& \mathbb{E}_t\left[(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j))^2 - V^j(s_t)^2\right] + \mathbb{E}_t\left[(\rho_\pi^{ij}(t)(r(s_t,a_t) - R(s_t,a_t))^2\right] \\
& + \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)^2\right] + \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\Big(\hat{\rho}_d^{ij}(t)-1\Big)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\Big)^2\right] \\
& + 2\mathbb{E}_t\left[(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j))(\rho_\pi^{ij}(t)\Big(\hat{\rho}_d^{ij}(t)-1\Big)\gamma\tilde{V}_{ij}^{DR}(s_{t+1}))\right] \\
& + 2\mathbb{E}_t\left[\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})])\rho_\pi^{ij}(t)\Big(\hat{\rho}_d^{ij}(t)-1\Big)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \qquad \text{(A.4)} \\
=& Var_t\left[-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)|s_t\right] + \mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2 Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right] \\
& + \mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2\gamma^2 Var_{t+1}(\tilde{V}_{ij}^{DR}(s_{t+1})|a_t)|s_t\right] + \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\Big(\hat{\rho}_d^{ij}(t)-1\Big)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\Big)^2\right] \\
& + 2\mathbb{E}_t\left[\big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j) + \rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})])\big)\big(\rho_\pi^{ij}(t)(\hat{\rho}_d^{ij}(t)-1)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\big)\right] \\
& + \mathbb{E}_{a_t}\left[(\rho_\pi^{ij}(t)\gamma)^2 Var_{t+1}\left[(\tilde{V}_{ij}^{DR}(s_{t+1}))|s_t,a_t\right]\right] + \mathbb{E}_{a_t}\left[\big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\big)^2\right] \\
& - \mathbb{E}_{a_t}\left[(\rho_\pi^{ij}(t)\gamma)^2 Var_{t+1}\left[(\tilde{V}_{ij}^{DR}(s_{t+1}))|s_t,a_t\right]\right] - \mathbb{E}_{a_t}\left[\big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\big)^2\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \\
& \qquad \text{(A.5)} \\
=& Var_t\left[-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)|s_t\right] + \mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2 Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right] \\
& + \mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2\gamma^2 Var_{t+1}(\tilde{V}_{ij}^{DR}(s_{t+1})|a_t)|s_t\right] + \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\Big(\hat{\rho}_d^{ij}(t)-1\Big)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right. \\
& \left. -\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j) + \rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)^2\right] \\
& - \mathbb{E}_{a_t}\left[(\rho_\pi^{ij}(t)\gamma)^2 Var_{t+1}\left[(\tilde{V}_{ij}^{DR}(s_{t+1}))|s_t,a_t\right]\right] - \mathbb{E}_{a_t}\left[\big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\big)^2\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2 \\
& \qquad \text{(A.6)} \\
=& Var_t\left[\rho_\pi^{ij}(t)\Delta(s_t,a_t)|s_t\right] + \mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2 Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right] + \mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right. \\
& \left. -\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j) - \rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\Big)^2\right] - \mathbb{E}_{a_t}\left[\big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\big)^2\right] - 2\mathbb{E}_t V^j(s_t)\mathbb{E}[\delta] - \big(\mathbb{E}[\delta]\big)^2. \\
& \qquad \text{(A.7)}
\end{aligned}
$$

We eliminate $Var_t\left[V^j(s_t)\right]$ in Eq. (A.1) since $Var_t\left[V^j(s_t)\right]=0$ when s_t is given. The equivalence from Eq. (A.2) to Eq. (A.3) uses the fact that $Q^j(s_t,a_t)=R(s_t,a_t)+\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]$. The equivalence from Eq. (A.3) to Eq. (A.4) follows the extension of the square of sum of the four terms, namely, the square of the first parentheses in Eq. (A.3) and the following facts given s_t and a_t: 1) $r(s_t,a_t)-R(s_t,a_t)$ and $\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})]$ are random variables with zero mean and independent of each others, since $R(s_t,a_t)$ and $\mathbb{E}_{t+1}[V^j(s_{t+1})]$ are the mean of $r(s_t,a_t)$ and $\tilde{V}_{ij}^{DR}(s_{t+1})$ respectively ; 2) $r(s_t,a_t)-R(s_t,a_t)$ and $\tilde{V}_{ij}^{DR}(s_{t+1})$ are independent; 3) $(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j))$ is constant.

The equivalence from Eq. (A.5) to Eq. (A.6) follows:

$$\begin{aligned}
&\mathbb{E}_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)^2|s_t\right]\\
=&\mathbb{E}_{a_t}\left[\mathbb{E}_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)^2|s_t,a_t\right]\right]\\
=&\mathbb{E}_{a_t}\left[Var_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)|s_t,a_t\right]\right]\\
&+\mathbb{E}_{a_t}\left[\mathbb{E}_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)|s_t,a_t\right]^2\right] \qquad \text{(A.8)}\\
=&\mathbb{E}_{a_t}\left[(\rho_\pi^{ij}(t)\gamma)^2Var_t\left[(\tilde{V}_{ij}^{DR}(s_{t+1}))|s_t,a_t\right]\right]+\mathbb{E}_{a_t}\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\right], \qquad \text{(A.9)}
\end{aligned}$$

where the last step is obtained from the equivalence of the variance and the expectation in Eq. (A.8):

$$\begin{aligned}
&Var_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)|s_t,a_t\right]\\
=&Var_t\left[\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])|s_t,a_t\right]\\
=&(\rho_\pi^{ij}(t)\gamma)^2Var_t\left[(\tilde{V}_{ij}^{DR}(s_{t+1}))|s_t,a_t\right],\\
&\mathbb{E}_{a_t}\left[\mathbb{E}_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)+\rho_\pi^{ij}(t)\gamma(\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})])\Big)|s_t,a_t\right]^2\right]\\
=&\mathbb{E}_{a_t}\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\right]. \qquad \text{// use fact 1) above}
\end{aligned}$$

The equivalence from Eq. (A.6) to Eq. (A.7) is from the fact that $\bar{V}_\theta(s_t,z_j)$ is constant given s_t. Given that $Q_\theta=0$, $\tilde{V}_{ij}^{DR}(s_t)$ will degrade to the variance of IS estimator and its variance can be written as follows:

$$\begin{aligned}
Var[\tilde{V}_{ij}^{IS}(s_t)]=&Var_t\left[\rho_\pi^{ij}(t)Q_\pi(s_t,a_t)|s_t\right]+\mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right]+\mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{IS}(s_{t+1})\right.\\
&\left.-\rho_\pi^{ij}(t)Q_\pi(s_t,a_t)+\bar{V}_\theta(s_t,z_j)-\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\Big)^2\right]-\mathbb{E}_{a_t}\left[\Big(-\rho_\pi^{ij}(t)Q_\pi(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\right].
\end{aligned} \qquad \text{(A.10)}$$

We further denote

$$\mathbb{V}(\rho_\pi)=\mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2Var_t\big[r_j(s_t,a_t)|a_t\big]\Big|s_t\right]+Var_t\Big[\rho_\pi^{ij}(t)\Delta(s_t,a_t)\Big|s_t\Big]-\mathbb{E}_t\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+V_\theta(s_t,z_j)\Big)^2\right],$$

which corresponds to the first, the second, and the third terms in Eq. (A.7).

A.5. Proof of Theorem 3.1

In this section, we derive the variance of unbiased DR estimator in Eq. (6) as shown in Theorem 3.1. Letting $\hat{\rho}_d=\rho_d$ in Eq. (A.7), we have $\delta=0$ and the variance can be obtained as:

$$\begin{aligned}
Var[V_{ij}^{DR}(s_t=s)]=&Var_t\left[\rho_\pi^{ij}(t)\Delta(s_t,a_t)|s_t\right]+\mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right]+\mathbb{E}_t\left[\Big(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right.\\
&\left.-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)-\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\Big)^2\right]-\mathbb{E}_{a_t}\left[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\right].
\end{aligned}$$

A.6. Upper bound for MSE of biased DR estimator $\tilde{V}^{DR}$

$$\begin{aligned}\text{Bias}(\hat{\rho}_d^{ij}) =&\Big|\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\tilde{V}_{ij}^{DR}(s_t=s)\big]-V^j(s_t=s)\Big|\\ =&\Big|\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma\rho_\pi^{ij}(t)\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big)\big]\Big|,\end{aligned} \tag{A.11}$$

where the second equality is obtained by the unbiasedness of V_{ij}^{DR} to V^j. Following the decomposition in Section A.2, MSE of the biased DR estimator $\tilde{V}^{DR}$ can be written as:

$$\begin{aligned}MSE(\tilde{V}_{ij}^{DR}(s_t=s)) = Var_t\left[\rho_\pi^{ij}(t)\Delta(s_t,a_t)|s_t\right]+\mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right]+\mathbb{E}_t\Big[\Big(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\\ -\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)-\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\Big)^2\Big]-\mathbb{E}_{a_t}\Big[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\Big]-2\mathbb{E}_tV^j(s_t)\mathbb{E}[\delta],\end{aligned}$$

where the last term can be bounded as

$$-2\mathbb{E}_tV^j(s_t)\mathbb{E}[\delta]\leq\big(\mathbb{E}_tV^j(s_t)\big)^2+\big(\mathbb{E}[\delta]\big)^2=\big(\mathbb{E}_tV^j(s_t)\big)^2+\big(Bias(\hat{\rho}_d^{ij})\big)^2,$$

with the bias bounded according to the Jesen's inequality

$$\text{Bias}(\hat{\rho}_d^{ij})\leq\sqrt{\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma\rho_\pi^{ij}(t)\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big)\big]^2}.$$

Hence, we can obtain an upper bound for the MSE of $\tilde{V}_{ij}^{DR}(s_t=s)$:

$$\begin{aligned}MSE(\tilde{V}_{ij}^{DR}(s_t=s)) \leq Var_t\left[\rho_\pi^{ij}(t)\Delta(s_t,a_t)|s_t\right]+\mathbb{E}_t\left[(\rho_\pi^{ij}(t))^2Var_t\left[r(s_t,a_t)|a_t\right]|s_t\right]+\mathbb{E}_t\Big[\Big(\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\\ -\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)-\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\Big)^2\Big]-\mathbb{E}_{a_t}\Big[\Big(-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\Big]\\ +\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma\rho_\pi^{ij}(t)\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big)\big]^2+\big(\mathbb{E}_tV^j(s_t)\big)^2.\end{aligned} \tag{A.12}$$

Note that $\mathbb{V}(\rho_\pi)$ also denote the terms that contains ρ_π but without $\hat{\rho}_d$ in Eq. (A.12).

A.7. Reduction of MSE by optimizing upper bound w.r.t. $\hat{\rho}_d$

Optimization of the upper bound in Eq. (A.12) w.r.t. $\hat{\rho}_d$ can be formulated as:

$$\begin{aligned}\min_{\hat{\rho}_d}\mathbb{E}_t\big[\gamma\rho_\pi^{ij}(t)\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_t)-\rho_d^{ij}(t)V_{ij}^{DR}(s_t)\big)\big]^2+\mathbb{E}_t\Big[\Big(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})]\big)\\ -\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)^2\Big].\end{aligned}$$

This optimization problem is convex w.r.t. $\hat{\rho}_d$. By letting the first-order derivative of the objective function be zero, we have:

$$\begin{aligned}&2\mathbb{E}_t\big[(\gamma\rho_\pi^{ij}(t))^2\tilde{V}_{ij}^{DR}(s_{t+1})\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big)\big]\\ &+2\mathbb{E}_t\Big[\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\Big(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})-\mathbb{E}_{t+1}[V^j(s_{t+1})]\big)-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Big)\Big]=0.\end{aligned}$$

By eliminating the constant of 2 and merging the two expectations on the left-hand side into one expectation, we have:

$$\begin{aligned}\mathbb{E}_t\Bigg[\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\cdot\Bigg(&2\gamma\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\\ &-\gamma\rho_\pi^{ij}(t)\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})-\gamma\rho_\pi^{ij}(t)\mathbb{E}_{t+1}[V^j(s_{t+1})]-\rho_\pi^{ij}(t)\Delta(s_t,a_t)+\bar{V}_\theta(s_t,z_j)\Bigg)\Bigg]=0.\end{aligned}$$

Note that inside expectation on the left-hand side is the multiplication of $\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$ with a summation enclosed by the parentheses, which can be rewritten as:

$$\left(2\gamma\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \gamma\rho_\pi^{ij}(t)\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1}) - \gamma\rho_\pi^{ij}(t)\mathbb{E}_{t+1}[V^j(s_{t+1})] - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)$$
$$=\gamma\rho_\pi^{ij}(t)\left(2\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})]\right) - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j).$$

On the right-hand side of this equation, the first three terms are closely correlated to $\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$, while the last two terms are loosely correlated to it. Furthermore, given that $\hat{\rho}_d^{ij}(t)$ is inside an interval upper-bounded by its true value $\rho_d^{ij}(t)$, the value of $\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$ is comparable to those of $\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})$ and $\mathbb{E}_{t+1}[V^j(s_{t+1})]$. Thus, the first three terms will be compensated by each other, while value of the last two terms $-\rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)$ will dominate, which is loosely correlated to $\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$.

Since $\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$ cannot dominate the value in the parentheses, we make assumption that it is loosely correlated to the term in the parentheses and have

$$2\mathbb{E}_t\left[\gamma\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\right] - \mathbb{E}_t\left[\gamma\rho_\pi^{ij}(t)\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\right] + \mathbb{E}_t\left[\left(-\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})] - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)\right] = 0,$$
$$2\mathbb{E}_t\left[\gamma\rho_\pi^{ij}(t)\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\right] - \mathbb{E}_{a_t\sim\pi_j,s_{t+1}\sim p_j}\left[\gamma V_{ij}^{DR}(s_{t+1})\right] - \gamma\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{t+1}\left[V^j(s_{t+1})\right] + V^j(s_t) = 0,$$
$$2\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\left[\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right]\hat{\rho}_d^{ij}(t) + \mathbb{E}_{a_t\sim\pi_j}\left[r(s_t,a_t)\right] - \gamma\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\left[V^j(s_{t+1})\right] = 0.$$

Hence, we can get the optimal dynamics importance weight, as follows:

$$\hat{\rho}_d^{ij}(t) = (\gamma V_j(s_{t+1}) - r_j(s_t,a_t)) \Big/ \left(2\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\right).$$

A.8. Proof of optimal dynamic weight that minimizing the variance

The variance can be rewritten as

$$\begin{aligned} Var_t\left[\tilde{V}_{ij}^{DR}(s_t)\right] =& \mathbb{E}_t\left[\left(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})]\big) - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)^2\right] \\ &- (\mathbb{E}[\tilde{V}_{ij}^{DR}])^2 + (V(s_t))^2 + \mathbb{V}(\rho_\pi), \end{aligned} \tag{A.13}$$

where the second term is always negative and will be zero under $\hat{\rho}_d^{ij} = \frac{-r_j(s_t,a_t)}{\gamma\tilde{V}_{ij}^{DR}(s_{t+1})}$, which is nearly zero especially under the sparse-reward setting. Since the rest terms contain no $\hat{\rho}_d^{ij}(t)$, we consider the optimization of the first term in Eq. (A.13) w.r.t. $\hat{\rho}_d^{ij}(t)$, which is also the third term in Eq. (6) of Theorem 3.1 in the main text. We formulate the following optimization problem

$$\min_{\hat{\rho}_d^{ij}(t)} \quad \mathbb{E}_t\left[\left(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})]\big) - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)^2\right], \tag{A.14}$$

whose first-order derivative can be given as

$$2\mathbb{E}_t\left[\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\left(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})]\big) - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)\right] = 0.$$

Under the assumption that $\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$ is loosely correlated to the term in the parentheses, we have:

$$\begin{aligned} \mathbb{E}_t\left[\left(\rho_\pi^{ij}(t)\gamma\big(\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \mathbb{E}_{t+1}[V^j(s_{t+1})]\big) - \rho_\pi^{ij}(t)\Delta(s_t,a_t) + \bar{V}_\theta(s_t,z_j)\right)\right] = 0, \\ \mathbb{E}_t\left[\rho_\pi^{ij}(t)\gamma\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})]\right] + V_j(s_t) = 0, \\ \mathbb{E}_t\left[\rho_\pi^{ij}(t)\gamma\hat{\rho}_d^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1}) - \rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})] + V_j(s_t)\right] = 0. \end{aligned}$$

Hence, we can get the optimal importance weight as:

$$\begin{aligned}\hat{\rho}_d^{var}(t) =&\big(\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})] - V^j(s_t)\big)/\big(\gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})\big)\\ =&\big(\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})] - V^j(s_t)/\rho_\pi^{ij}(t)\big)/\big(\gamma\tilde{V}_{ij}^{DR}(s_{t+1})\big).\end{aligned}$$

We now review the variance in Eq. (A.13). When the increase of $\rho_\pi^{ij}(t)$ results in that $\hat{\rho}_d^{var}(t) > \frac{-r_j(s_t,a_t)}{\gamma\tilde{V}_{ij}^{DR}(s_{t+1})}$, continuously reducing $\hat{\rho}_d^{ij}(t)$ from $\hat{\rho}_\pi^{var}(t)$ to $\frac{-r_j(s_t,a_t)}{\gamma\tilde{V}_{ij}^{DR}(s_{t+1})}$ will enlarge the variance, since the first and the second terms in Eq. (A.13) will both increase. And reducing $\hat{\rho}_d^{ij}$ to near $\hat{\rho}_\pi^{var}(t)$ will reduce the variance.

A.9. Proof of the upper bound for error of MSE

Following the decomposition of MSE, we have

$$MSE(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) = Bias(\hat{\rho}_d^{ij*})^2 + Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}),$$

$$MSE(\tilde{V}^{DR}, \hat{\rho}_d^{ij}) = Bias(\hat{\rho}_d^{ij})^2 + Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij}).$$

Computing the bias of DR estimator using $\hat{\rho}_d^*$ and $\hat{\rho}_d^{var}$ separately and letting $\epsilon = \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big] - \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\big[\gamma V^j(s_{t+1})\big]$, we have

$$\begin{aligned}Bias(\hat{\rho}_d^{ij*}) &= \Big|\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\rho_\pi^{ij}(t)\big(\gamma V^j(s_{t+1}) - r_j(s_t,a_t)\big)/2 - \gamma\rho_\pi^{ij}(t)\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big]\Big|\\ &= \Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big]/2 - \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t,a_t)\big]/2 - \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\big[\gamma V^j(s_{t+1})\big]\Big|\\ &= \Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big]/2 + \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t,a_t)\big]/2 - V^j(s_t)\Big|\\ &= \Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big]/2 + \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t,a_t)\big]/2 - V^j(s_t)\Big|\\ &= \Big|\frac{\epsilon}{2} + \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\big[\gamma V^j(s_{t+1})\big]/2 + \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t,a_t)\big]/2 - V^j(s_t)\Big| = \Big|\frac{\epsilon}{2} - \frac{V^j(s_t)}{2}\Big|,\end{aligned}$$

$$\begin{aligned}Bias(\hat{\rho}_d^{var}) =&\Big|\mathbb{E}_{a_t\sim\pi_i}\mathbb{E}_{s_{t+1}\sim p_i}\big[\rho_\pi^{ij}(t)\gamma\mathbb{E}_{t+1}[V^j(s_{t+1})] - V^j(s_t) - \gamma\rho_\pi^{ij}(t)\rho_d^{ij}(t)V_{ij}^{DR}(s_{t+1})\big]\Big|\\ =&\Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big] - \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\big[\gamma V^j(s_{t+1})\big] - V^j(s_t)\Big| = \Big|\epsilon - V^j(s_t)\Big|.\end{aligned}$$

Hence, we have

$$\Big|Bias(\hat{\rho}_d^{ij*})^2 - Bias(\hat{\rho}_d^{var})^2\Big| = \frac{3}{4}\Big|\epsilon - V^j(s_t)\Big|^2, \quad \Big|Bias(\hat{\rho}_d^{ij*})^2 - Bias(\rho_d^{ij})^2\Big| = \frac{1}{4}\Big|\epsilon - V^j(s_t)\Big|^2,$$

and

$$\begin{aligned}\Big|\epsilon - V^j(s_t)\Big| &= \Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_i}\big[\gamma V^j(s_{t+1})\big] - \mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\big[\gamma V^j(s_{t+1})\big] - V^j(s_t)\Big|\\ &= \Big|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\Big[\left(\frac{1}{\rho_d^{ij}(t)} - 2\right)\gamma V^j(s_{t+1})\Big] - \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t,a_t)\big]\Big|.\end{aligned}$$

For the difference of variance, we have an upper bound as:

$$\begin{aligned}&\Big|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij})\Big|\\ \leq&\max\Big(\Big|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{var})\Big|, \Big|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(\tilde{V}^{DR}, \rho_d^{ij})\Big|\Big)\\ =&\max\Big(\mathbb{E}_t\Big[\big(A\hat{\rho}_d^{ij*}(t) - 2B\big)^2\Big], \Big|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(V^{DR}, \rho_d^{ij})\Big|\Big).\end{aligned}$$

Table A.1. Values set for the constant pair (A, B) to generate random environment parameters.

ALGORITHM	BODY MASS	BODY INERTIA	JOINT DAMPING	FRICTION
POINT-ROBOT-PARAMS-SPARSE	$(1.5, 1.0)$	$(1.5, 1.0)$	$(1.3, 1.0)$	$(1.5, 1.0)$
ANT-PARAMS-SPARSE	$(1.5, 3.0)$	$(1.5, 3.0)$	$(1.3, 3.0)$	$(1.5, 3.0)$
HUMANOID-PARAMS-SPARSE	$(1.5, 3.0)$	$(1.5, 3.0)$	$(1.3, 3.0)$	$(1.5, 3.0)$
HOPPER-PARAMS	$(1.5, 3.0)$	$(1.5, 3.0)$	$(1.3, 3.0)$	$(1.5, 3.0)$
WALKER-2D-PARAMS	$(1.5, 3.0)$	$(1.5, 3.0)$	$(1.3, 3.0)$	$(1.5, 3.0)$
POINT-ROBOT-PARAMS	$(1.5, 1.0)$	$(1.5, 1.0)$	$(1.3, 1.0)$	$(1.5, 1.0)$
SAWYER-PUSH-PARAMS-SPARSE	$(1.5, 2.5)$	$(1.5, 2.5)$	$(1.3, 2.5)$	$(1.5, 2.5)$

Let $A = \gamma\rho_\pi^{ij}(t)\tilde{V}_{ij}^{DR}(s_{t+1})$ and $B = -\gamma\rho_\pi^{ij}(t)\mathbb{E}_{t+1}[V^j(s_{t+1})] - \rho_\pi^{ij}(t)\Delta(s_t, a_t) + V_\theta(s_t, z_j)$. Hence, we have the upper bound for the MSE difference between biased DR estimators using $\hat{\rho}_d^*$ and $\hat{\rho}_d$ as

$$\begin{aligned}
&\left|MSE(\tilde{V}_{ij}^{DR}, \hat{\rho}_d^{ij*}) - MSE(\tilde{V}_{ij}^{DR}, \hat{\rho}_d^{ij})\right| \\
=&\left|bias(\hat{\rho}_d^{ij*})^2 + Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - bias(\hat{\rho}_d^{ij})^2 - Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij})\right| \\
\leq&\left|bias(\hat{\rho}_d^{ij*})^2 - bias(\hat{\rho}_d^{ij})^2\right| + \left|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij})\right| \\
\leq&\frac{3}{4}\left|\mathbb{E}_{a_t\sim\pi_j}\mathbb{E}_{s_{t+1}\sim p_j}\Big[\left(\frac{1}{\rho_d^{ij}(t)} - 2\right)\gamma V^j(s_{t+1})\Big] - \mathbb{E}_{a_t\sim\pi_j}\big[r_j(s_t, a_t)\big]\right|^2 \\
&+ \max\left(\mathbb{E}_t\Big[\big(A\hat{\rho}_d^{ij*}(t) - 2B\big)^2\Big], \left|Var_t(\tilde{V}^{DR}, \hat{\rho}_d^{ij*}) - Var_t(V^{DR}, \rho_d^{ij})\right|\right).
\end{aligned}$$

B. Experimental Setup

B.1. Hyper-Parameters and Implementation Details

In our experiments, we utilize the same hyper-parameters for meta-training as in the open-sourced code of the baseline meta-RL approach, PEARL [4]. For our proposed DRT, we build up networks of predicting state transition for each task that has two layers with 500 units at each layer. The learning rate for the prediction network is $1e^{-3}$. We update the lower bound $\hat{\rho}_d^l$ at the end of each training epoch. Since negative transfer brought by reusing samples that may be inappropriately chosen by strategy $\mathcal{S}_I$ could result in a significantly lower target value as computed by DRaE for training, in practice, we take the maximum state value $\hat{V}(s)$ estimated by the value network V_θ and our DRaE $\tilde{V}^{DR}$ to further alleviate this issue.

B.2. Generation of Varying Dynamics

The randomization of dynamics on all the environments in our experiments is implemented by generating different environment parameters through:

$$param_{ij} = \beta_j * init\_param_i, \tag{A.15}$$

where $\beta_j = A^{x_j}, x_j \sim Uniform(-B, B)$, A and B are the constants which control the generation of β_j for each environment parameter of task j, and $init\_param_i$ is the initial value of the i-th environment parameter. Overall, these randomly sampled environment parameters include the body mass, body inertia, joint damping, and body component's friction, for which the values of constant pair (A, B) are listed in Table A.1. With the initial values $init\_param_i$ loaded directly from the original file of "mujoco_py", the randomized environment parameter $param_{ij}$ is then obtained and set on the MuJoCo simulation engine to generate various environment dynamics.

B.3. Reward Functions

For the dense-reward environments in Section 5.2, we use the same implementation as in PEARL's open-sourced code. For the sparse-reward environments with varying rewards and dynamics in Section 5.1, we modify their reward functions as:

$$reward = \begin{cases} -dist(robot, goal) + C & if\ dist(robot, goal) < D, \\ 0 & otherwise. \end{cases} \tag{A.16}$$

In the Point-Robot-Params-Sparse environment, we generate the goals in a cubic space, where we uniformly sample the goal coordinates in $(0.2, 0.5)$, $(-0.4, 0.4)$, and $(0.5, 1.5)$, and set $C = 1.0$ and $D = 0.2$. In the Ant-Params-Sparse environment, we uniformly generate the goals on a semi-circle with radius 2.0, and set $C = 4.0$ and $D = 0.8$. In the Humanoid-Params-Sparse environment, we uniformly generate the goals on a semi-circle with radius 3.0, and set $C = 3$ and $D = 0.8$. For all the environments, we keep the additional proprioceptive reward signal, which are control cost, contact cost, and survive bonus.

References

[1] Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6672–6679, 2020.

[2] Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science Robotics*, 7(62):eabk2822, 2022.

[3] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. RL2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv*:1611.02779, 2016.

[4] Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *International conference on machine learning*, pages 5331–5340. PMLR, 2019.

[5] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, OpenAI Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in neural information processing systems*, 30, 2017.

[6] Alexander Li, Lerrel Pinto, and Pieter Abbeel. Generalized hindsight for reinforcement learning. *Advances in neural information processing systems*, 33:7754–7767, 2020.

[7] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning* , pages 652–661. PMLR, 2016.

[8] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[9] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(11), 2013.

[10] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, pages 6005–6014. PMLR, 2019.

[11] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off- policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[12] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adap- tation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[13] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In 2012 IEEE/RSJ *International Conference on Intelligent Robots and Systems*, pages 5026–5033, 2012.

[14] Charles Packer, Pieter Abbeel, and Joseph E Gonzalez. Hindsight task relabelling: Experience replay for sparse reward meta-RL. *Advances in Neural Information Processing Systems*, 34:2466–2477, 2021.

[15] Michael Wan, Jian Peng, and Tanmay Gangwani. Hindsight foresight relabeling for meta- reinforcement learning. In *International Conference on Learning Representations*, 2022.

[16] Rasool Fakoor, Pratik Chaudhari, Stefano Soatto, and Alexander J Smola. Meta-q-learning. In *International Conference on Learning Representations*, 2020.

[17] Bradly C Stadie, Ge Yang, Rein Houthooft, Xi Chen, Yan Duan, Yuhuai Wu, Pieter Abbeel, and Ilya Sutskever. Some considerations on learning to explore via meta-reinforcement learning. *arXiv preprint arXiv*:1803.01118, 2018.

[18] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv*:1803.02999, 2018.

[19] Jonas Rothfuss, Dennis Lee, Ignasi Clavera, Tamim Asfour, and Pieter Abbeel. Promp: Proximal meta-policy search. In *International Conference on Learning Representations*, 2019.

[20] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. Varibad: A very good method for bayes-adaptive deep RL via meta- learning. 2020.

[21] Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine. Meta- reinforcement learning of structured exploration strategies. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[22] Swaminathan Gurumurthy, Sumit Kumar, and Katia Sycara. Mame: Model-agnostic meta- exploration. In *Conference on Robot Learning*, pages 910–922. PMLR, 2020.

[23] Jin Zhang, Jianhao Wang, Hao Hu, Tong Chen, Yingfeng Chen, Changjie Fan, and Chongjie Zhang. Metacure: Meta reinforcement learning with empowerment-driven exploration. In *International Conference on Machine Learning*, pages 12600–12610. PMLR, 2021.

[24] Desik Rengarajan, Sapana Chaudhary, Jaewon Kim, Dileep Kalathil, and Srinivas Shakkottai. Enhanced meta reinforcement learning via demonstrations in sparse reward environments. In *Advances in Neural Information Processing Systems*, 2022.

[25] James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.

[26] Heejung Bang and James M Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

[27] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

[28] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pages 1447–1456. PMLR, 2018.

[29] Jiawei Huang and Nan Jiang. From importance sampling to doubly robust policy gradient. In *International Conference on Machine Learning*, pages 4434–4443. PMLR, 2020.

[30] Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pages 11581–11591. PMLR, 2021.

[31] Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. Doubly robust distribu- tionally robust off-policy evaluation and learning. In Proceedings of the 39th *International Conference on Machine Learning*, volume 162 of Proceedings of Machine Learning Research. PMLR, 2022.

[32] Andrea Tirinzoni, Andrea Sessa, Matteo Pirotta, and Marcello Restelli. Importance weighted transfer of samples in reinforcement learning. In *International Conference on Machine Learning*, pages 4936–4945. PMLR, 2018.

[33] Andrea Tirinzoni, Mattia Salvini, and Marcello Restelli. Transfer of samples in policy search via multiple importance sampling. In *International Conference on Machine Learning*, pages 6264–6274. PMLR, 2019.

[34] Andrea Tirinzoni, Rafael Rodriguez Sanchez, and Marcello Restelli. Transfer of value functions via variational methods. *Advances in Neural Information Processing Systems*, 31, 2018.

[35] Andrea Tirinzoni, Riccardo Poiani, and Marcello Restelli. Sequential transfer in reinforcement learning with a generative model. In *International Conference on Machine Learning*, pages 9481–9492. PMLR, 2020.

[36] Giuseppe Canonaco, Andrea Soprani, Matteo Giuliani, Andrea Castelletti, Manuel Roveri, and Marcello Restelli. Time-variant variational transfer for value functions. In *Uncertainty in Artificial Intelligence*, pages 876–886. PMLR, 2021.

[37] Safa Alver and Doina Precup. Constructing a good behavior basis for transfer using generalized policy updates. In

International Conference on Learning Representations, 2022.

[38] André Barreto, Will Dabney, Rémi Munos, Jonathan J Hunt, Tom Schaul, Hado P van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[39] Andre Barreto, Diana Borsa, John Quan, Tom Schaul, David Silver, Matteo Hessel, Daniel Mankowitz, Augustin Zidek, and Remi Munos. Transfer in deep reinforcement learning using successor features and generalised policy improvement. In *International Conference on Machine Learning*, pages 501–510. PMLR, 2018.

[40] Jaekyeom Kim, Seohong Park, and Gunhee Kim. Constrained gpi for zero-shot transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2022.

[41] Tianpei Yang, Weixun Wang, Hongyao Tang, Jianye Hao, Zhaopeng Meng, Hangyu Mao, Dong Li, Wulong Liu, Yingfeng Chen, Yujing Hu, et al. An efficient transfer learning framework for multiagent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17037–17048, 2021.

[42] Sasha Salter, Kristian Hartikainen, Walter Goodwin, and Ingmar Posner. Priors, hierar- chy, and information asymmetry for skill transfer in reinforcement learning. *arXiv preprint arXiv*:2201.08115, 2022.