Transfer of Samples in Policy Search via Multiple Importance Sampling

Andrea Tirinzoni [1] **Mattia Salvini** [1] **Marcello Restelli** [1]

Abstract

We consider the transfer of experience samples in reinforcement learning. Most of the previous works in this context focused on value-based settings, where transferring instances conveniently reduces to the transfer of (s, a, s', r) tuples. In this paper, we consider the more complex case of reusing samples in policy search methods, in which the agent is required to transfer entire trajectories between environments with different transition models. By leveraging ideas from multiple importance sampling, we propose robust gradient estimators that effectively achieve this goal, along with several techniques to reduce their variance. In the case where the transition models are known, we theoretically establish the robustness to the negative transfer for our estimators. In the case of unknown models, we propose a method to efficiently estimate them when the target task belongs to a finite set of possible tasks and when it belongs to some reproducing kernel Hilbert space. We provide empirical results to show the effectiveness of our estimators.

1. Introduction

In many real-world problems, reinforcement learning (RL) agents (Sutton & Barto, 1998) repeatedly face environments with similar, but different, dynamics. Consider, for instance, deploying a policy on a robot with different physical parameters than the ones it was trained for, or adapting from simulation to reality, or learning under non-stationarity. In all these cases, the agent could potentially exploit knowledge acquired in the past environments to speed-up the learning process of new related tasks. How this knowledge reuse can be achieved is the fundamental problem of *transfer* in RL.

Knowledge transfer has been widely studied for RL domains (Taylor & Stone, 2009; Lazaric, 2012). In this context, the agent is supposed to reuse knowledge acquired from a set of *source* tasks to accelerate the learning process of a new *target* task. Designing a transfer algorithm poses three main questions: what element should be transferred? How should the transfer be performed? Is this procedure beneficial for the target task? An answer to the last question is particularly crucial to prevent *negative transfer*, the situation in which reused knowledge harms the learning process.

Among the many kinds of knowledge that have been successfully transferred in RL, we focus on experience samples, i.e., states, actions, and rewards that the agent collects from different tasks. Most of the previous approaches in this setting focused on value-based algorithms, where the agent attempts to reuse *single* transition and reward instances from different decision processes. Taylor et al. (2008) proposed an algorithm to transfer samples so as to augment the dataset used by a model-based RL algorithm. Almost simultaneously, Lazaric et al. (2008b) designed a model-free methodology to estimate which source samples are most likely to benefit the target and used it to transfer into a batch RL algorithm. The same settings were considered by Laroche & Barlier (2017), who showed that no measure of similarity is needed to safely transfer under the restrictive assumption that tasks do not differ in their dynamics. More recently, Tirinzoni et al. (2018b) proposed a method to transfer all given samples without carrying out any explicit selection, while reweighing their contribution to the learning process proportionally to their importance for solving the target task.

One of the drawbacks of the above-mentioned approaches is that they do not easily generalize to policy search, where the agent is required to transfer long sequences of states and actions rather than only single-step information. Since policy search algorithms, despite their recent successes, typically require large batches in order to reliably estimate the expected return or its gradient, reusing samples from different tasks would be practically very useful.

In this paper, we propose a methodology to address this limitation for families of tasks differing only in their transition models. Similarly to Tirinzoni et al. (2018b), we employ importance sampling (IS) (Owen, 2013) techniques to transfer *all* source samples, while correcting the bias introduced by their different distributions. However, instead of focusing on single transitions and rewards, we show that it is pos-

[1]Politecnico di Milano, Milan, Italy. Correspondence to: Andrea Tirinzoni <andrea.tirinzoni@polimi.it>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

sible to transfer entire trajectories from different policies and environments to improve the gradient estimates of a policy search algorithm, thus reducing its sample complexity and the number of iterations required to converge. In order to accomplish this objective, we use ideas from multiple importance sampling (MIS) (Veach & Guibas, 1995) to derive gradient estimators that automatically transfer samples from several different distributions and we propose two techniques to reduce their variance. Our estimators are simple, enjoy strong theoretical guarantees, and are of general interest even outside our transfer settings. For instance, we show that they can be successfully applied to reuse samples from past policies during the learning process (a kind of intra-environment transfer). For the ideal case in which the transition models are known, we formally establish robustness to negative transfer of the proposed methods. For the more realistic case in which the transition models are unknown, we propose a methodology to estimate them which explicitly trades off between the bias and variance these models might induce on the importance weights. Finally, we empirically demonstrate the effectiveness of our estimators in three domains of increasing difficulty.

2. Preliminaries

Policy Search We define a task as a discounted Markov decision process (Puterman, 2014), $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{P}_0, \gamma \rangle$, where $\mathcal{S} \subseteq \mathbb{R}^d$ is the state space, $\mathcal{A} \subseteq \mathbb{R}^u$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, $\mathcal{P}_0 \in \Delta(\mathcal{S})$ is the initial state distribution, and $\gamma \in [0, 1]$ is a discount factor. Here $\Delta(\Omega)$ denotes the set of probability measures over a generic Ω. At each time t, the agent is in some state $\boldsymbol{s}_t$, it takes an action $\boldsymbol{a}_t$ according to some policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, it receives a reward $r_t := \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t)$, and it transitions to a new state $\boldsymbol{s}_{t+1}$ according to $\mathcal{P}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)$. In policy search (Peters & Schaal, 2008; Sutton et al., 2000; Deisenroth et al., 2013), a class of parameterized polices $\pi_{\boldsymbol{\theta}}$ is considered and the goal is to find the parameters maximizing the expected return,

$$\arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P}) := \int p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P}) \mathcal{R}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}, \qquad (1)$$

where $\boldsymbol{\tau} = (\boldsymbol{s}_0, \boldsymbol{a}_0, \ldots, \boldsymbol{s}_T)$ denotes a trajectory of T time steps and, with some abuse of notation, we set $\mathcal{R}(\boldsymbol{\tau}) := \sum_{t=0}^{T-1} \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Common approaches for optimizing this objective function employ stochastic gradient methods. Let $g(\boldsymbol{\tau}) := \sum_{t=0}^{T-1} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)$, then the well-known REINFORCE algorithm (Williams, 1992) approximates the gradient of (1) using n i.i.d. trajectories from $\pi_{\boldsymbol{\theta}}$ as

$$\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P}) = \frac{1}{n} \sum_{i=1}^{n} g(\boldsymbol{\tau}_i) \mathcal{R}(\boldsymbol{\tau}_i). \qquad (2)$$

Transfer of Samples In the sample transfer problem (Taylor & Stone, 2009; Lazaric, 2012), a set of experience samples is provided from a small number of *source tasks* $\mathcal{M}_1, \ldots, \mathcal{M}_m$ and, given a new *target task*, the goal is to figure out how to reuse these instances in order to speed up the learning process. In our policy search settings, these samples are entire trajectories collected under arbitrary policies. We make the following assumption.

Assumption 1 (Task differences)**.** *Each task $\mathcal{M}_j$ is uniquely characterized by its transition kernel $\mathcal{P}_j$. State-action space, reward function, and initial state distribution are shared among tasks*[1].

Formally, our input is a dataset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$, where each $\mathcal{D}_j = \{\boldsymbol{\tau}_1, \ldots, \boldsymbol{\tau}_{n_j}\}$ is a set of n_j trajectories from a fixed task-policy couple $(\boldsymbol{\theta}_j, \mathcal{P}_j)$. We assume that the policies used to collect each trajectory are known.

Multiple Importance Sampling Multiple importance sampling (MIS) (Veach & Guibas, 1995; Veach, 1997) is a very effective method for estimating the expected value of a function given samples from multiple proposal distributions. Consider a measurable space $(\mathcal{X}, \mathscr{F})$, a function $f : \mathcal{X} \to \mathbb{R}$, and $m+1$ probability measures $P, Q_1, \ldots, Q_m$ absolutely continuous w.r.t. the Lebesgue measure, with $p, q_1, \ldots, q_m$ denoting their densities. A MIS estimator for $\mu := \mathbb{E}_{x \sim P}[f(x)]$ given n_j samples from each Q_j is

$$\hat{\mu} = \sum_{j=1}^{m} \frac{1}{n_j} \sum_{i=1}^{n_j} h_j(x_{i,j}) \frac{p(x_{i,j})}{q_j(x_{i,j})} f(x_{i,j}), \qquad (3)$$

where the function h is often referred to as *heuristics* and must be a partition of unity, i.e., $\sum_j h_j(x) = 1$ for all $x \in \mathcal{X}$. It is easy to show (Veach & Guibas, 1995) that the MIS estimator is unbiased. A common and convenient choice for h is the *balance heuristics*, $h_j(x) = \frac{n_j q_j(x)}{\sum_{l=1}^{m} n_l q_l(x)}$, for which the MIS estimator reduces to an IS estimator with a mixture of proposals, $\hat{\mu} = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} \frac{p(x_{i,j})}{q_\alpha(x_{i,j})} f(x_{i,j})$, where $n = \sum_{j=1}^{m} n_j$ and $q_\alpha(x) = \sum_{j=1}^{m} \alpha_j q_j(x)$, with $\alpha_j = \frac{n_j}{n}$. This estimator is also known in the literature as deterministic mixture sampling or stratification (Hesterberg, 1995; Owen & Zhou, 2000). A key property is that, when a set of n_0 samples is available from the target distribution P, the resulting weights are *bounded* by $\frac{1}{\alpha_0}$. This fact also implies that the variance of these weights is bounded, a property that rarely holds for plain IS. For these reasons, samples from P are often referred to as *defensive*. An important measure of the goodness of an IS estimator is the *effective sample size* (ESS), which is typically approximated as $\widehat{\mathrm{ESS}} := \frac{n}{1+\mathbb{V}\mathrm{ar}[p(x)/q(x)]}$ (Liu, 1996) or, equivalently, as $\widehat{\mathrm{ESS}} = \frac{n}{d_2(P||Q)}$ (Cortes et al., 2010; Ryu, 2016; Metelli et al., 2018), where $d_2(P\|Q) = \int p(x)^2 q(x)^{-1} \mathrm{d}x$ is the exponentiated second-order Renyi divergence.

[1] In practice, it is enough that the target reward function is known rather than shared. The extension to unknown rewards is simple (see, e.g., (Tirinzoni et al., 2018b)).

3. Transfer via Importance Sampling

We describe how IS techniques can be used to efficiently transfer samples in policy search in case transition models are known. We will deal with unknown models in Section 4.

3.1. Multiple Importance Sampling Estimators

Consider the transfer settings described in Section 2. Our idea is to learn the target task using standard gradient-based techniques, collecting a batch of n_0 episodes at each step, while reusing the source trajectories to augment the dataset for estimating the gradient in order to reduce its variance. Since, due to different policies and transition models, these trajectories follow different distributions than the one induced by the current policy in the target task, IS techniques can be straightforwardly employed to yield an unbiased estimator,

$$\widehat{\nabla}^{\text{IS}}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P}) = \frac{1}{n} \sum_{j=0}^{m} \sum_{i=1}^{n_j} w_j^{\text{IS}}(\boldsymbol{\tau}_{i,j}) g(\boldsymbol{\tau}_{i,j}) \mathcal{R}(\boldsymbol{\tau}_{i,j}), \quad (4)$$

where the *importance weight* $w_j^{\text{IS}}(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P})}{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)}$ can be computed in closed-form as

$$w_j^{\text{IS}}(\boldsymbol{\tau}) = \prod_{t=0}^{T-1} \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\boldsymbol{\theta}_j}(\boldsymbol{a}_t|\boldsymbol{s}_t)} \frac{\mathcal{P}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)}{\mathcal{P}_j(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)}. \quad (5)$$

Unfortunately, this IS scheme is likely to fail in most cases of practical interest. It is well known, especially from the literature on off-policy estimation (Precup, 2000; Hachiya et al., 2009; Thomas & Brunskill, 2016; Guo et al., 2017; Liu et al., 2018), that importance sampling on long trajectories is likely to give almost zero or huge weights, thus leading to estimators with very high (sometimes infinite) variance (Li et al., 2015; Jiang & Li, 2016). This drawback is even amplified in our transfer settings, where there is a model mismatch in addition to the one between the policies. Several variance reduction techniques, typically paying a small amount of bias, have been proposed (e.g., self-normalized estimators (Kong, 1992), truncation (Ionides, 2008), flattening (Hachiya et al., 2009)). Fortunately, MIS comes to the rescue in our settings, allowing us to get a low-variance estimator without introducing any bias. Using Equation (3), the MIS gradient estimator is

$$\widehat{\nabla}^{\text{MIS}}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P}) = \tfrac{1}{n} \sum_{j=0}^{m} \sum_{i=1}^{n_j} w_j^{\text{MIS}}(\boldsymbol{\tau}_{i,j}) g(\boldsymbol{\tau}_{i,j}) \mathcal{R}(\boldsymbol{\tau}_{i,j}), \quad (6)$$

where $w_j^{\text{MIS}}(\boldsymbol{\tau}) := \frac{n}{n_j} h_j(\boldsymbol{\tau}) w_j^{\text{IS}}(\boldsymbol{\tau})$ for a general heuristic function h. For brevity, we define $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) := \sum_{j=0}^{m} \alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)$, so that the weights for the particular choice of balance heuristics reduce to $w_j^{\text{MIS}}(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})}$.

Algorithm 1 outlines how these weighted estimators are used in our proposed transfer procedure. Line 1 initializes the policy parameters. A criterion for achieving a good jumpstart could be easily derived. Since our primary concern is the optimization rather than the initialization, we leave this step unspecified. At each iteration, we decide adaptively the batch size (i.e., the number of defensive samples) that guarantees a minimum ESS for the resulting dataset (line 4). The rationale is that, if the gradient of the current policy can be reliably estimated using the source samples, there is no need to collect new trajectories at all. In order to carry out this step, we derive a lower bound on the increase rate of the approximate ESS as a function of n_0. Due to space constraints, we defer this discussion to Appendix A. In practice, we impose a minimum batch size of n_{min} to avoid degenerate cases. The new batch (line 5) is then added to the current dataset (line 6). Note that this step has two key implications: (i) the gradient estimator always uses at least n_{min} defensive samples, and (ii) the number m of policy-model pairs in $\mathcal{D}$ grows with the number of iterations as trajectories from the target task (but policies potentially different from the current one) are stored. Finally, the parameters are updated using the chosen weighted estimator on the current dataset.

Algorithm 1 Transfer via Importance Sampling

Require: Target task $\mathcal{M}$, source dataset $\mathcal{D} = \{\langle(\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{n_j}), \boldsymbol{\theta}_j, \mathcal{P}_j\rangle \mid j = 1, \dots, m\}$, gradient estimator $\widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P})$, effective sample size estimator $\widehat{\text{ESS}}(n_0; \mathcal{D})$, minimum effective sample size ESS_{min}, minimum batch size n_{min}, step-size sequence η_k

1: Initialize policy: $\boldsymbol{\theta}_0 \leftarrow$ INIT-POLICY$(\mathcal{M}, \mathcal{D})$
2: Initialize iteration count: $k \leftarrow 0$
3: **while** not converged **do**
4: Find the minimum $n_0 \in \{n_{\text{min}}, \dots, \text{ESS}_{\text{min}}\}$ such that $\widehat{\text{ESS}}(n_0; \mathcal{D}) \geq \text{ESS}_{\text{min}}$
5: Sample n_0 trajectories from $\mathcal{M}$ under policy $\pi_{\boldsymbol{\theta}_k}$
6: Store samples: $\mathcal{D} \leftarrow \mathcal{D} \cup \{\langle(\boldsymbol{\tau}_1, \dots, \boldsymbol{\tau}_{n_0}), \boldsymbol{\theta}_k, \mathcal{P}\rangle\}$
7: Update parameters: $\boldsymbol{\theta}_{k+1} \leftarrow \boldsymbol{\theta}_k + \eta_k \widehat{\nabla}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_k, \mathcal{P})$
8: **end while**

The next two sections propose two techniques to further reduce the variance of the MIS estimator without introducing any bias.

3.2. Per-decision Estimators

Per-decision IS (Precup, 2000) is a common variance reduction technique from off-policy evaluation. It relies on the intuition that future actions cannot influence past rewards, i.e., each reward r_t should be weighted only by the probability of a trajectory up to that time. This technique can be easily combined with MIS, leading to the PD estimator,

$$\widehat{\nabla}^{\text{PD}}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P}) = \tfrac{1}{n} \sum_{j=0}^{m} \sum_{i=1}^{n_j} \sum_{t=0}^{T-1} \gamma^t w_{j,t}^{\text{PD}}(\boldsymbol{\tau}_{i,j}) g_t(\boldsymbol{\tau}_{i,j}) r_t, \quad (7)$$

where $g_t(\boldsymbol{\tau}) := g(\boldsymbol{\tau}_{0:t})$, with $\boldsymbol{\tau}_{0:t}$ being a trajectory up to time t, and $w^{\text{PD}}_{j,t}(\boldsymbol{\tau}) := \frac{n}{n_j} h_{j,t}(\boldsymbol{\tau}) w^{\text{IS}}_j(\boldsymbol{\tau}_{0:t})$. Using the balance heuristics, $w^{\text{PD}}_{j,t}(\boldsymbol{\tau}) = w^{\text{MIS}}_j(\boldsymbol{\tau}_{0:t})$. Notice that the heuristics is now a function of time. We show that, if this function is uniformly normalized over time, the resulting estimator remains unbiased.

Theorem 3.1 (Unbiasedness of PD estimator)**.** *Let $h_{j,t}(\boldsymbol{\tau})$ be a function such that, for all $t \in \{0, \dots, T-1\}$ and $\boldsymbol{\tau}$, $\sum_{j=0}^{m} h_{j,t}(\boldsymbol{\tau}) = 1$. Then, the per-decision MIS estimator in (7) is unbiased.*

3.3. Regression-based Control Variates

Control variates (CVs) are a widely applied variance reduction technique for general Monte Carlo estimators (Hammersley & Handscomb, 1964). The key idea is that a random variable with known expectation could be used to reduce the variance of a mean estimator for another random variable. Owen & Zhou (2000) have popularized the usage of CVs for IS and MIS. For the d-th dimension of the gradient, consider a vector of functions $\boldsymbol{\psi}_d(\boldsymbol{\tau}) := [\psi_{0,d}(\boldsymbol{\tau}), \dots, \psi_{m+1,d}(\boldsymbol{\tau})]$ such that $\mathbb{E}_{\boldsymbol{\tau} \sim q_{\boldsymbol{\alpha}}}[\psi_{j,d}(\boldsymbol{\tau})] = 0$ for every j. Then, our MIS estimator (6) with $\boldsymbol{\psi}_d$ as CVs becomes

$$\widehat{\nabla}^{\text{CV}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P}) = \widehat{\nabla}^{\text{MIS}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P}) - \frac{1}{n} \sum_{j=0}^{m} \sum_{i=1}^{n_j} \boldsymbol{\beta}_d^T \boldsymbol{\psi}_d(\boldsymbol{\tau}_{i,j}), \quad (8)$$

where $\boldsymbol{\beta}_d \in \mathbb{R}^{m+1}$ is the vector of CV coefficients. As shown by Owen & Zhou (2000), the proposal distributions composing the mixture of a MIS estimator, whose integral is known to be 1, can be used as very effective control variates. Thus, we consider $\psi_{j,d}(\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \mathcal{P}_j)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})} - 1$ for $j = 0, \dots, m$. Furthermore, $g(\boldsymbol{\tau})$ is a widely adopted control variate in policy search (a.k.a. baseline) since its expectation is known to be 0. Therefore, we consider $\psi_{m+1,d}(\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P}) g_d(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})}$. Finally, the vectors $\boldsymbol{\beta}^*_d$ minimizing the variance of (8) can be approximated by solving the following regression problem (Owen & Zhou, 2000):

$$\arg\min_{\boldsymbol{\beta}_d} \sum_{j=0}^{m} \sum_{i=1}^{n_j} \left(\frac{p(\boldsymbol{\tau}_{i,j})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_{i,j})} g(\boldsymbol{\tau}_{i,j}) \mathcal{R}(\boldsymbol{\tau}_{i,j}) - \boldsymbol{\beta}_d^T \boldsymbol{\psi}_d(\boldsymbol{\tau}_{i,j}) \right)^2.$$

In practice, we can fit the CVs and estimate the gradient using different partitions of the current dataset to keep an unbiased estimator.

Proposition 3.1. *The estimator (8) is unbiased for any $\boldsymbol{\beta}_d$. Furthermore, under the optimal coefficients $\boldsymbol{\beta}^*_d$, $\mathbb{V}\text{ar}[\widehat{\nabla}^{\text{CV}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P})] \leq \mathbb{V}\text{ar}[\widehat{\nabla}^{\text{MIS}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P})]$.*

Note that, when the number of dimensions d of the parameter space is large, it is common to fit a unique $\boldsymbol{\beta}$ for all d. In this case, simply taking $\psi_{m+1}(\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, \mathcal{P}) \sum_d g_d(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})}$ and solving the regression problem is therefore equivalent to (approximately) minimizing $\text{Tr}(\mathbb{C}\text{ov}[\widehat{\nabla}^{\text{CV}}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, \mathcal{P})])$. Finally, we note that the CVs can be combined straightforwardly with the PD estimator of Section 3.2.

3.4. Robustness to Negative Transfer

We now show that our MIS estimator with CVs enjoys safety guarantees against negative transfer. We first propose a definition of negative transfer for policy gradient algorithms in terms of convergence to ϵ-optimal stationary points [2].

Definition 3.1 (Negative transfer)**.** *Let $\mathscr{A}$ and $\mathscr{B}$ be two policy gradient algorithms. Fix an initial parameter $\boldsymbol{\theta}_0$, a learning rate η, a batch size n, and an accuracy $\epsilon > 0$. Then, $\mathscr{A}$ negatively transfers w.r.t. $\mathscr{B}$ ($\mathscr{A} \prec \mathscr{B}$) if there exists an iteration number $k \geq 1$ such that we can guarantee that $\frac{1}{k} \sum_{l=0}^{k-1} \mathbb{E}_{\mathscr{B}}[\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_l)\|^2] \leq \epsilon$ but $\frac{1}{k} \sum_{l=0}^{k-1} \mathbb{E}_{\mathscr{A}}[\|\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_l)\|^2] > \epsilon$.*

Let $\mathscr{B}_{\text{R}}$ be the REINFORCE algorithm and $\mathscr{B}_{\text{G}}$ be the GPOMDP algorithm (Baxter & Bartlett, 2001). The next result shows that Algorithm 1 using CVs and the MIS estimator ($\mathscr{A}_{\text{CV}}$) or the PD estimator ($\mathscr{A}_{\text{PDCV}}$) cannot be worse than their no-transfer counterparts.

Theorem 3.2. *Assume the return J is L-smooth (i.e., its gradient is L-Lipschitz). Let $n_{min} > 0$ be the minimum batch size for $\mathscr{A}_{CV}$ ($\mathscr{A}_{PDCV}$) and the fixed batch size for $\mathscr{B}_R$ ($\mathscr{B}_G$). Assume all algorithms start from the same parameter $\boldsymbol{\theta}_0$, use a learning rate $0 < \eta \leq \frac{2}{L}$, and that $\mathscr{A}_{CV}$ ($\mathscr{A}_{PDCV}$) uses the optimal CV coefficients $\boldsymbol{\beta}^*_d$. Then, for all $\epsilon > 0$:*

$$\mathscr{A}_{CV} \not\prec \mathscr{B}_R, \quad \mathscr{A}_{PDCV} \not\prec \mathscr{B}_G.$$

We remark that, according to our definition, the fact that $\mathscr{A}$ is robust against $\mathscr{B}$ does not necessarily imply that $\mathscr{A}$ converges faster than $\mathscr{B}$ but only that, whenever we can prove that $\mathscr{B}$ converged, we can also prove that $\mathscr{A}$ converged. Hence, we might say that $\mathscr{A}$ cannot be much worse than $\mathscr{B}$, the standard (weaker) notion of negative transfer that is often considered in the literature (Taylor & Stone, 2009).

4. The Case of Unknown Models

The transfer algorithm presented in Section 3 requires full knowledge of the transition models of each task, an assumption that rarely holds in practice. Here we consider the more realistic case in which the models (equivalently, the importance weights) are unknown and have to be estimated from data. In general, this goal can be efficiently achieved by directly estimating density ratios (Sugiyama et al., 2012). Unfortunately, in our settings, this is not a good approach for at least two reasons: (i) density ratios are not transferable, i.e., they must be recomputed for each new policy and/or task; (ii) our weights are defined over entire trajectories, high-dimensional random variables whose distributions have some structure that would not be exploited by a direct estimator. Therefore, we decide to take a more indirect ap-

[2] As a standard in non-convex optimization, convergence to stationary points could be replaced with convergence to local maxima at the cost of a more complicated analysis.

proach and estimate only the missing components, namely the transition models. However, instead of naively plugging in any density estimator or probabilistic model of the uncertain dynamics, we propose an estimator that is aware of the MIS scheme in which these models will be adopted.

From now on, we consider system dynamics of the form $\boldsymbol{s}_{t+1} = f(\boldsymbol{s}_t, \boldsymbol{a}_t) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_{\mathcal{P}}^2 \boldsymbol{I})$, and we suppose each task to be uniquely identified by its transition function f. Our main assumption is that the source tasks f_j are known, while the target task f is uncertain according to a distribution $\varphi \in \Delta(\mathcal{F})$, where $\mathcal{F}$ is the set of all possible transition functions in the family of tasks under consideration.[3] Let us fix a policy parameter $\boldsymbol{\theta}$ and a dataset $\mathcal{D}$. Let $\nabla J := \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, f)$ denote the true gradient at $\boldsymbol{\theta}$ and $\widehat{\nabla} J(\widetilde{f}) := \widehat{\nabla}_{\boldsymbol{\theta}}^{(\mathrm{MIS})} J(\boldsymbol{\theta}, \widetilde{f})$ denote its MIS estimate (6) using the balance heuristics where an arbitrary function $\widetilde{f}$ is used to compute the importance weights instead of the unknown target model f. Note that, although we assume to know the source distributions, the denominator $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) = \sum_{j=0}^{m} \alpha_j p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ is still unknown since at least one of its components depends on f. We write $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; f)$ to make this dependence explicit in the remainder. Given the set $\mathcal{J} = \{0, 1, \dots, m\}$, we use $\mathcal{J}_{\mathrm{tgt}} = \{j \in \mathcal{J} | f_j = f\}$ to denote the indexes of proposals from the target task, and define $\mathcal{J}_{\mathrm{src}} = \mathcal{J} \setminus \mathcal{J}_{\mathrm{tgt}}$. Moreover, $\alpha_{\mathrm{tgt}} = \sum_{j \in \mathcal{J}_{\mathrm{tgt}}} \alpha_j$ denotes the proportion of target samples in $\mathcal{D}$, and similarly for $\alpha_{\mathrm{src}} = 1 - \alpha_{\mathrm{tgt}}$. We begin by deriving an upper bound on the mean square error (MSE) of the MIS estimator.

Theorem 4.1. *Let $\widetilde{f} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ be any function and $p_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) = \sum_{j \in \mathcal{J}_{tgt}} \frac{\alpha_j}{\alpha_{tgt}} p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f})$. Suppose that $\|g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\|_\infty \leq B$ almost surely. Then, for $f \sim \varphi$,*

$$\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}) - \nabla J\|^2\right] \leq \frac{dB^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}, \widetilde{f}) \| q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f})\right) + c_1 dB^2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\alpha}}}\left[\|\bar{f}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2\right] + c_1 dB^2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\alpha}}}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + \mathcal{O}(n^{-1}), \quad (9)$$

where the expectation is w.r.t. $\boldsymbol{\tau}_{i,j} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ and $f \sim \varphi$. Here $\bar{f}(\boldsymbol{s}, \boldsymbol{a}) := \mathbb{E}_{f \sim \varphi}[f(\boldsymbol{s}, \boldsymbol{a})]$, $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{C}\mathrm{ov}_{f \sim \varphi}[f(\boldsymbol{s}, \boldsymbol{a})]$, and c_1 is a constant.

Let $\mathcal{L}(\widetilde{f})$ denote the value of this bound as a function of $\widetilde{f}$. Then, we look for the function $f^* \in \mathcal{F}$ that minimizes $\mathcal{L}$. Intuitively, we seek for a model that trades off between three different objectives: (i) when few trajectories are available and the target model is highly uncertain, it should stay close to the mixture of source distributions in order to reduce the variance of the resulting estimator (first term); (ii) as the number of samples grows, it should move towards $\bar{f}$, our best guess for the true model (second term); finally, it should give priority to the regions of the state-action space where the target model is more accurate (third term). Our MSE-aware approach to model estimation is summarized in Algorithm 2. Although appealing, optimizing the bound for the optimal transition function is non-trivial. Now we show two cases in which this can be done efficiently.

[3] A discussion on how our results can be generalized to unknown source tasks is deferred at the end of this section.

Algorithm 2 MSE-aware Model Estimation

Require: Model space $\mathcal{F}$, target trajectories $\mathcal{D}_{\mathrm{tgt}} = \{\mathcal{D}_j | \, j \in \mathcal{J}_{\mathrm{tgt}}\}$, uncertainty model φ
1: Update φ using $\mathcal{D}_{\mathrm{tgt}}$
2: Compute $\widetilde{f} \in \mathcal{F}$ minimizing the bound of Theorem 4.1
3: **return** $\widetilde{f}$

4.1. Discrete Task Family

We start by considering the simple setting in which $\mathcal{F} = \{f_1, f_2, \dots, f_H\}$ is a finite set of possible transition functions. Our uncertainty model is therefore a discrete distribution over this set. Assuming a uniform prior $\varphi_0(f) = \frac{1}{|\mathcal{F}|}$, this distribution can be updated iteratively for every $f \in \mathcal{F}$ given a batch of target trajectories $\mathcal{D}_{\mathrm{tgt}}$ under policy $\boldsymbol{\theta}_k$ as

$$\varphi_{k+1}(f) \propto \varphi_k(f) \prod_{\boldsymbol{\tau} \in \mathcal{D}_{\mathrm{tgt}}} \prod_{t=0}^{T-1} \pi_{\boldsymbol{\theta}_k}(\boldsymbol{a}_t|\boldsymbol{s}_t) \mathcal{P}_f(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t).$$

Given φ_k, the bound of Theorem 4.1 can be easily approximated for every $f \in \mathcal{F}$. Notice that all expectations in (9) are under distributions induced by the model $\widetilde{f}$ and, therefore, they can be approximated by simulating trajectories without interacting with the true environment.

4.2. Reproducing Kernel Hilbert Spaces

We now consider a more general functional space for our transition models. Let $\mathcal{X} = \mathcal{S} \times \mathcal{A}$ and consider a positive semi-definite kernel function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We suppose that $\mathcal{F}$ is the unique reproducing kernel Hilbert space (RKHS) induced by $\mathcal{K}$. In an RKHS, the reproducing property implies that every function f can be written as $f(\boldsymbol{x}) = \langle f, \mathcal{K}(\boldsymbol{x}, \cdot) \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the dot product on $\mathcal{F}$. For simplicity, we consider each dimension of our transition models separately. We refer the reader to (Micchelli & Pontil, 2005) for the extension to vector-valued RKHS.

We represent the uncertainty over the target model as a Gaussian process (GP) (Williams & Rasmussen, 2006), $f \sim \mathcal{GP}(0, \mathcal{K})$, with a zero-mean prior and $\mathcal{K}$ as covariance function. As usual in model-based RL (e.g., (Deisenroth & Rasmussen, 2011)), we train conditionally independent GPs for each dimension of the state space. Suppose that we get a set of l training inputs $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_l]^T$ and l

training targets $\boldsymbol{Y} = [\boldsymbol{y}_1, \dots, \boldsymbol{y}_l]^T$ from our dataset $\mathcal{D}_{\text{tgt}}$ of trajectories from the target task, where $\boldsymbol{x}_i = (\boldsymbol{s}_i, \boldsymbol{a}_i)$ and $\boldsymbol{y}_i = f(\boldsymbol{x}_i) + \mathcal{N}(0, \sigma_{\mathcal{P}}^2 \boldsymbol{I})$. Then, the posterior mean function can be evaluated at any point $\boldsymbol{x}$ as $\bar{f}(\boldsymbol{x}) = \boldsymbol{k}(\boldsymbol{x})^T (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y}$ (Williams & Rasmussen, 2006), where $\boldsymbol{k}(\boldsymbol{x})$ is the vector with entries $\boldsymbol{k}_i(\boldsymbol{x}) = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x})$ and $\boldsymbol{K}$ is the Gram matrix, $K_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Now that we have an uncertainty model for our target transition function, let us move to optimize our bound $\mathcal{L}$ on the MSE. Unfortunately, minimizing $\mathcal{L}(\widetilde{f})$ w.r.t. $\widetilde{f} \in \mathcal{F}$ is not as simple as in the finite-model case since (i) $\widetilde{f}$ controls the distributions under which expectations are taken, and (ii) it appears as a product over several time steps in the Renyi divergence term. For these reasons, we now introduce some simplifications that will lead to a convenient closed-form solution. First, we approximate the two expectations by drawing a small number of trajectories from our last hypothesized model, so that their dependence on the function to be computed is removed. Secondly, we further bound our objective in a more convenient way. In order to carry out this last step, we derive an upper bound on the exponentiated Renyi divergence w.r.t. the Kullback-Leibler (KL) divergence, which could be of independent interest.

Theorem 4.2. *Let $(\mathcal{X}, \mathscr{F})$ be a measurable space, P and Q be two probability measures on $\mathcal{X}$ such that $P \ll Q$, and $Q_\alpha = \alpha P + (1-\alpha) Q$ denotes their convex combination with coefficient $\alpha \in (0,1)$. Suppose there exists a finite constant $C > 0$ such that $\operatorname{ess\,sup} \frac{dP}{dQ} \leq C$. Then,*

$$d_2(P||Q_\alpha) \leq 1 + u(\alpha) D_{KL}(P||Q), \tag{10}$$

where

$$u(\alpha) = \begin{cases} \frac{2C(1-\alpha)^2}{(\alpha C + 1 - \alpha)^3} & \text{if } C \leq \frac{1-\alpha}{2\alpha} \\ \frac{8}{27\alpha} & \text{otherwise.} \end{cases}$$

Using Theorem 4.2, our objective $\mathcal{L}(\widetilde{f})$ can be bounded in a very convenient way.

Proposition 4.1. *The objective $\mathcal{L}(\widetilde{f})$ given in (9) can be upper bounded by*

$$\mathcal{L}(\widetilde{f}) \leq k_1 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}, \widetilde{f})} \left[\sum_{j \in \mathcal{J}_{src}} \alpha_j \|\widetilde{f}(\boldsymbol{x}_t) - f_j(\boldsymbol{x}_t)\|_2^2 \right] + k_2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\alpha}}} \left[\|\widetilde{f}(\boldsymbol{x}_t) - \bar{f}(\boldsymbol{x}_t)\|_2^2 \right] + k_3,$$

where $k_1 = \frac{u(\alpha) d B^2}{2\sigma_{\mathcal{P}}^2 n (1-\alpha_0)}$, $k_2 = \frac{4\alpha_{tgt} d B^2}{\alpha_0^2 \sigma_{\mathcal{P}}^2}$, and k_3 is a constant independent of $\widetilde{f}$.

Our new bound is quite appealing. While the bias term remains unchanged, the variance is now a mixture of expected l_2 distances between $\widetilde{f}$ and the known source models f_j. In practice, we optimize a regularized version of this objective so that the representer theorem of RKHS applies. Furthermore, as mentioned above, we approximate the two expectations by drawing R trajectories from the mixture $p_{\boldsymbol{\alpha}}$ using our last hypothesized model. The resulting objective reduces to a regularized least-squares problem,

$$\arg\min_{\widetilde{f} \in \mathcal{F}} \frac{1}{R} \sum_{r=1}^{R} \sum_{t=0}^{T-1} \left(k_1 w_{r,t} \sum_{j \in \mathcal{J}_{\text{src}}} \alpha_j \|\widetilde{f}(\boldsymbol{x}_{r,t}) - f_j(\boldsymbol{x}_{r,t})\|_2^2 + k_2 \|\bar{f}(\boldsymbol{x}_{r,t}) - \widetilde{f}(\boldsymbol{x}_{r,t})\|_2^2 \right) + \lambda \|f\|_{\mathcal{K}}^2, \tag{11}$$

where $\lambda > 0$ is the regularization parameter and $w_{r,t} = \prod_{l=0}^{t} \frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_l|\boldsymbol{s}_l)}{\sum_{j \in \mathcal{J}_{\text{tgt}}} \frac{\alpha_j}{\alpha_{\text{tgt}}} \pi_{\boldsymbol{\theta}_j}(\boldsymbol{a}_l|\boldsymbol{s}_l)}$ is an importance weight to correct the distribution mismatch in the first expectation. Most importantly, its solution is available in closed form.

Proposition 4.2. *The function f^* minimizing (11) is*

$$f^*(\boldsymbol{x}) = \boldsymbol{A}^T \boldsymbol{k}(\boldsymbol{x}),$$

where $\boldsymbol{k}(\boldsymbol{x})$ is the RT-dimensional vector with entries $\mathcal{K}(\boldsymbol{x}_{r,t}, \boldsymbol{x})$ and

$$\boldsymbol{A} = (k_1 \alpha_{tgt} \boldsymbol{W} \boldsymbol{K} + k_2 \boldsymbol{K} + \lambda R \boldsymbol{I})^{-1} (k_1 \boldsymbol{W} \boldsymbol{F}_{src} + k_2 \bar{\boldsymbol{F}}),$$

with $\boldsymbol{K}$ being the Gram matrix, $\boldsymbol{W} = diag(w_{r,t})$, $\bar{\boldsymbol{F}} = [\bar{f}(\boldsymbol{x}_{1,0}), \dots, \bar{f}(\boldsymbol{x}_{R,T-1})]^T$, $\boldsymbol{F}_{src} = \sum_{j \in \mathcal{J}_{src}} \alpha_j \boldsymbol{F}_j$, and $\boldsymbol{F}_j = [f_j(\boldsymbol{x}_{1,0}), \dots, f_j(\boldsymbol{x}_{R,T-1})]^T$.

Discussion One might be wondering why the estimated transition models are not directly used for planning in a standard model-based RL algorithm instead of estimating the importance weights for a model-free approach. It is well-known that even small errors can lead to disastrous performances when the optimal policy is computed under the estimated models. Since in our case the learned models are only used to re-weight samples from the true environment, we argue that the impact of such errors is much more contained. In fact, as far as the weights keep reasonable values, the learning process could potentially be carried out effectively. Furthermore, note that our estimators are consistent as the number of target samples goes to infinity for *any* hypothesized model, not necessarily the true one.

Remark 4.1 (Unknown source models)**.** *The extension of Theorem 4.1 to unknown source models is reported in Appendix B. Unfortunately, optimizing the resulting bound over all models becomes considerably more difficult due to the combinatorial growth of the problem. However, in practice, we typically update online only the target model. In fact, we cannot require additional source samples and, thus, we can estimate the source models in batch before learning starts. Therefore, our algorithm can be straightforwardly applied, e.g.,, by fitting GPs and then plugging the MAP estimates into Theorem 4.1 instead of the true source models (hence our simplification of known sources).*

5. Related Works

The closest work to ours is the recent paper of Tirinzoni et al. (2018b). The authors propose an IS-based algorithm to transfer samples in a value-based batch RL setting, together with a methodology to estimate the weights using Gaussian processes. Besides extending these algorithms to policy search, our MIS estimators naturally combine multiple distributions and are much more robust to source tasks that are very different from the target, a drawback of plain IS.

Among the existing transfer approaches for policy search, Ammar et al. (2014) and Ammar et al. (2015) focus on a lifelong learning scenario, where the agent continually faces new tasks sampled from a common distribution and must quickly adapt to each of them. Similarly to our work, several recent papers provide theoretical guarantees on the transfer procedure (e.g., Brunskill & Li, 2013; Zhan et al., 2016; Barreto et al., 2017; Abel et al., 2018; Tirinzoni et al., 2018a). Our assumption of environments with different dynamics is also related to Hidden-parameter MDPs (Doshi velez & Konidaris, 2013; Killian et al., 2017), although we do not require transitions to be parameterized. Crucially, while we both learn transition models, we do not use these for planning but only for computing the weights, which is typically more robust to estimation errors.

Finally, our MIS estimators are related to, and might be of interest for, different RL settings such as off-policy evaluation (Precup, 2000; Hachiya et al., 2009; Thomas & Brunskill, 2016; Guo et al., 2017; Liu et al., 2018), off-policy learning (Precup et al., 2001; Mahmood et al., 2014; Geist & Scherrer, 2014; Munos et al., 2016; Metelli et al., 2018), and sample reuse (Zhao et al., 2013; Hachiya et al., 2011).

6. Experiments

We analyze the performance of our estimators with known models in Section 6.1 and with model estimation in Sections 6.2 and 6.3. Due to space constraints, we only provide the high-level configuration of each experiment, while referring the reader to Appendix D for the specific hyperparameters that were adopted (see Table 1 for a quick summary).

6.1. Linear-Quadratic Regulator

Our first test domain is the one-dimensional linear-quadratic regulator (LQR) (Dorato et al., 1995; Peters & Schaal, 2008), a well-known benchmark from the control literature. The system has linear dynamics, $s_{t+1} = As_t + Ba_t + \epsilon$, with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_{\mathcal{P}}^2)$, and quadratic rewards.

We begin by evaluating the MIS estimators proposed in Section 3. Besides our proposed estimators, we compare to per-decision IS (PD-IS), which is widely adopted in the literature and can be straightforwardly adapted to our case, and to GPOMDP (Baxter & Bartlett, 2001) as our no-transfer baseline. The source tasks are randomly generated by uniformly sampling A in $[0.6, 1.4]$ and B in $[0.8, 1.2]$, while the target task is fixed with $A = 1$ and $B = 1$. We employ linear Gaussian policies and consider 8 source parameters, $\theta \in \{-0.1, -0.2, \dots, -0.8\}$, generating 20 episodes from each model-policy couple. To have a fair comparison, we use the same learning rate and initialization for all algorithms. Figure 1*(left)* shows the distance to the optimal target parameter as a function of the number of episodes. As expected, PD-IS shows a significant amount of negative transfer w.r.t. GPOMDP. This is due to the fact that the huge variance of the importance weights forces the algorithm to collect large batches to guarantee the required ESS. MIS and PD-MIS achieve an improvement over the no-transfer baseline, with the latter having smaller variance. When introducing CVs, the algorithm enjoys much better gradient estimates and significantly outperforms all alternatives.

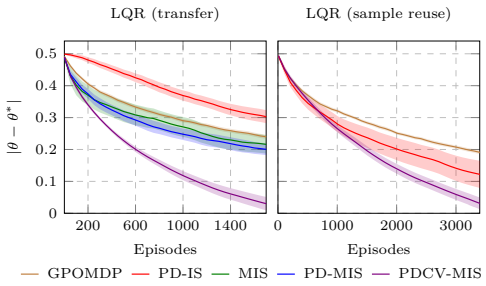

Figure 1. Comparison of the proposed gradient estimators in the LQR domain under known models. (left) transfer experiment, and (right) sample reuse experiment. Each curve is the average of 40 independent runs, each re-sampling the source tasks, with Student's t 95% confidence intervals

Sample Reuse in Policy Gradients Our estimators can be successfully adopted to reuse samples generated by previous policies in standard (no-transfer) policy gradients. Figure 1*(right)* shows the result of learning the same target task as before from scratch (i.e., without any source sample), where each algorithm uses the same (fixed) batch size, learning rate, and initialization. We can appreciate that both the per-decision and our estimator enjoy a speedup over GPOMDP, but the former suffers a much higher variance.

6.2. Cart-pole Balancing

Our second domain is the well-known Cartpole problem (Sutton & Barto, 1998), where the goal is to balance a pole on a moving cart. We generate source tasks by uniformly sampling the mass of the cart in the interval $[0.8, 1.2]$ and the length of the pole in the interval $[0.3, 0.7]$. The target task is the standard Cartpole with cart mass 1.0 and pole length 0.5. For each of 5 source tasks, we consider a sequence of 10 linear policies generated by GPOMDP during its learning

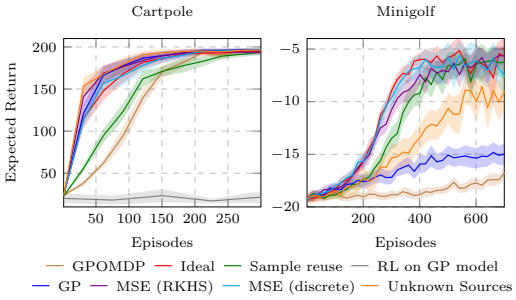

Figure 2. Comparison of our model estimation approach in the Cartpole (left) and minigolf (right) domains. Each curve is the average of 40 independent runs, each re-sampling the source tasks, with Student's t 95% confidence intervals.

process and collect 10 episodes from each. While the LQR domain was relatively short-horizon ($T = 20$), here we allow trajectories up to $T = 200$ time steps, which allows us to verify the transferability of long state-action sequences.

We now test our model estimation approaches. Besides GPOMDP, we report the performance of the sample reuse (SR) variant of our algorithm where we ignore the source tasks and transfer only past target trajectories. All transfer algorithms use the PD-MIS estimator. Figure 2*(left)* shows the results. Interestingly, all transfer approaches outperform both GPOMDP and SR, which confirms that reusing trajectories from different environments can significantly improve the quality of the learning process. Both our model estimation approaches perform comparably to the ideal estimators. While this should be expected for the discrete estimator, the continuous one works well since an accurate GP model of the Cartpole dynamics can be obtained with a relatively small amount of samples. This fact can be verified from the GP curve, where the weights have been estimated by directly plugging in the GP predictions instead of optimizing our bound. Not surprisingly, the algorithm that estimates the source models as in Remark 4.1 performs comparably to the best alternatives. However, the fact that the fitted GP is accurate for estimating the weights does not imply that it is an accurate model of the system dynamics which can be used for planning. The gray curve in Figure 2*(left)*, which shows the performance achieved by optimal policies for the estimated dynamics, confirms this statement.

6.3. Minigolf

In this last domain, we want to study how much the proposed transfer algorithms can speed up the learning process of an agent playing a minigolf game by reusing the experience made on different minigolf courses. The various tasks may differ in the length of the putter (between $70cm$ and $100cm$), in the hole size (between $10cm$ and $15cm$), and in the dynamic friction coefficient (whose range was measured empirically (Penner, 2002) between 0.065 and 0.196). The minigolf domain was originally introduced in the RL field by Lazaric et al. (2008a); here, we change the dynamics following the modeling developed by Penner (2002) in order to make the problem more realistic. A detailed description of the minigolf domain is available in Appendix D.3.

In this experiment, we adopt Gaussian policies with a fourth-order polynomial basis function. We generated 5 source tasks by randomly sampling dynamic friction coefficient, hole size, and putter length from the realistic ranges defined above. Furthermore, we considered 10 source policies of increasing quality and generated 40 episodes from each model-policy pair. The target task is fixed with a friction of 0.131, a putter of $100cm$, and a hole of diameter $10cm$. All transfer algorithms use the PDCV estimator. The results are shown in Figure 2*(right)*. Unlike the simpler Cartpole domain, GPOMDP is not able to learn the task in such a small number of episodes. Interestingly, due to the high level of noise present in this environment, direct estimation of weights using the GP predictions leads to unsatisfactory results. On the other hand, both our model estimation approaches solve the task with performance comparable to the ideal estimator. We note that the speed-up over the SR algorithm is not as remarkable as in the previous experiment due to the little amount of knowledge transfer that can be achieved in this more complicated setting. We finally point out that most of the simplifications introduced in the previous sections do not hold in this domain. In fact, the transition models are not Gaussian, while the noise is heteroscedastic and changes between tasks. Despite these relaxations, our approach can be applied without suffering any considerable performance degradation.

7. Conclusion

In this paper, we introduced a methodology which employs multiple importance sampling for transferring samples (i.e., entire trajectories) to reduce the variance of the estimated gradients in policy search. We showed that our estimators are general, in the sense that they can be of interest outside the transfer literature, and they enjoy strong theoretical properties. We proposed a methodology to estimate the unknown task models in a principled way by directly minimizing an upper bound on the MSE that these models induce on the resulting importance-weighted estimator. Finally, we empirically demonstrated the effectiveness of our algorithm in different domains, both in the ideal and estimated cases.

An interesting question is whether our method could be generalized to the policy gradient theorem (Sutton et al., 2000), where the importance weights would be on the stationary distributions of single states. This could dramatically increase the amount of transferred knowledge, as shown by Liu et al. (2018) for the case of off-policy evaluation.

References

Abel, D., Jinnai, Y., Guo, S. Y., Konidaris, G., and Littman, M. Policy and value transfer in lifelong reinforcement learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 20–29, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than sgd. *arXiv preprint arXiv:1708.08694*, 2017.

Ammar, H. B., Eaton, E., Ruvolo, P., and Taylor, M. Online multi-task learning for policy gradient methods. In *International Conference on Machine Learning*, pp. 1206–1214, 2014.

Ammar, H. B., Eaton, E., Luna, J.-M., and Ruvolo, P. Autonomous cross-domain knowledge transfer in lifelong policy gradient reinforcement learning. In *IJCAI*, 2015.

Barreto, A., Dabney, W., Munos, R., Hunt, J. J., Schaul, T., van Hasselt, H. P., and Silver, D. Successor features for transfer in reinforcement learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4055–4065. Curran Associates, Inc., 2017.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. *arXiv preprint arXiv:1309.6821*, 2013.

Cortes, C., Mansour, Y., and Mohri, M. Learning bounds for importance weighting. In *Advances in neural information processing systems*, pp. 442–450, 2010.

Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *studia scientiarum Mathematicarum Hungarica*, 2:229–318, 1967.

Deisenroth, M. and Rasmussen, C. E. Pilco: A model-based and data-efficient approach to policy search. In *Proceedings of the 28th International Conference on machine learning (ICML-11)*, pp. 465–472, 2011.

Deisenroth, M. P., Neumann, G., Peters, J., et al. A survey on policy search for robotics. *Foundations and Trends® in Robotics*, 2(1–2):1–142, 2013.

Dorato, P., Abdallah, C. T., Cerone, V., and Jacobson, D. H. *Linear-quadratic control: an introduction*. Prentice Hall Englewood Cliffs, NJ, 1995.

Doshi velez, F. and Konidaris, G. Hidden parameter markov decision processes: A semiparametric regression approach for discovering latent task parametrizations. *IJCAI : proceedings of the conference*, 2016, 08 2013.

Elvira, V., Martino, L., and Robert, C. P. Rethinking the effective sample size. *arXiv preprint arXiv:1809.04129*, 2018.

Geist, M. and Scherrer, B. Off-policy learning with eligibility traces: A survey. *The Journal of Machine Learning Research*, 15(1):289–333, 2014.

Guo, Z., Thomas, P. S., and Brunskill, E. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pp. 2492–2501, 2017.

Hachiya, H., Akiyama, T., Sugiayma, M., and Peters, J. Adaptive importance sampling for value function approximation in off-policy reinforcement learning. *Neural Networks*, 22(10):1399–1410, 2009.

Hachiya, H., Peters, J., and Sugiyama, M. Reward-weighted regression with sample reuse for direct policy search in reinforcement learning. *Neural Computation*, 23(11): 2798–2832, 2011.

Hammersley, J. and Handscomb, D. *Monte Carlo Methods*. Methuen's monographs on applied probability and statistics. Methuen, 1964. ISBN 9780416523409. URL https://books.google.it/books?id=Kk4OAAAAQAAJ.

Hesterberg, T. Weighted average importance sampling and defensive mixture distributions. *Technometrics*, 37(2): 185–194, 1995.

Ionides, E. L. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008.

Jiang, N. and Li, L. Doubly robust off-policy value evaluation for reinforcement learning. In Balcan, M. F. and Weinberger, K. Q. (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR.

Killian, T. W., Daulton, S., Konidaris, G., and Doshi-Velez, F. Robust and efficient transfer learning with hidden parameter markov decision processes. In *Advances in Neural Information Processing Systems*, pp. 6250–6261, 2017.

Kong, A. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

Laroche, R. and Barlier, M. Transfer reinforcement learning with shared dynamics. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Lazaric, A. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pp. 143–173. Springer, 2012.

Lazaric, A., Restelli, M., and Bonarini, A. Reinforcement learning in continuous action spaces through sequential monte carlo methods. In *Advances in neural information processing systems*, pp. 833–840, 2008a.

Lazaric, A., Restelli, M., and Bonarini, A. Transfer of samples in batch reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pp. 544–551. ACM, 2008b.

Le Cam, L. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 1986.

Li, L., Munos, R., and Szepesvari, C. Toward Minimax Off-policy Value Estimation. In Lebanon, G. and Vishwanathan, S. V. N. (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 608–616, San Diego, California, USA, 09–12 May 2015. PMLR.

Liu, J. S. Metropolized independent sampling with comparisons to rejection sampling and importance sampling. *Statistics and Computing*, 6(2):113–119, 1996.

Liu, Q., Li, L., Tang, Z., and Zhou, D. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5361–5371, 2018.

Mahmood, A. R., van Hasselt, H. P., and Sutton, R. S. Weighted importance sampling for off-policy learning with linear function approximation. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 27*, pp. 3014–3022. Curran Associates, Inc., 2014.

Martino, L., Elvira, V., and Louzada, F. Effective sample size for importance sampling based on discrepancy measures. *Signal Processing*, 131:386–401, 2017.

Metelli, A. M., Papini, M., Faccio, F., and Restelli, M. Policy optimization via importance sampling. In *Advances in Neural Information Processing Systems*, pp. 5447–5459, 2018.

Micchelli, C. A. and Pontil, M. On learning vector-valued functions. *Neural computation*, 17(1):177–204, 2005.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 1054–1062, 2016.

Owen, A. and Zhou, Y. Safe and effective importance sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Owen, A. B. *Monte Carlo theory, methods and examples*. 2013.

Pearson, K. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

Penner, A. The physics of putting. *Canadian Journal of Physics*, 80(2):83–96, 2002.

Peters, J. and Schaal, S. Reinforcement learning of motor skills with policy gradients. *Neural networks*, 21(4):682–697, 2008.

Precup, D. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Precup, D., Sutton, R., and Dasgupta, S. Off-policy temporal-difference learning with function approximation. *Proceedings of the 18th International Conference on Machine Learning*, 06 2001.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Ryu, E. K. *Convex optimization for Monte Carlo: Stochastic optimization for importance sampling*. PhD thesis, Stanford University, 2016.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pp. 1057–1063, 2000.

Taneja, I. J. Generalized relative information and information inequalities. *Journal of Inequalities in Pure and Applied Mathematics*, 5(1):1–19, 2004.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

Taylor, M. E., Jong, N. K., and Stone, P. Transferring instances for model-based reinforcement learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 488–505. Springer, 2008.

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

Tirinzoni, A., Rodriguez Sanchez, R., and Restelli, M. Transfer of value functions via variational methods. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6179–6189. Curran Associates, Inc., 2018a.

Tirinzoni, A., Sessa, A., Pirotta, M., and Restelli, M. Importance weighted transfer of samples in reinforcement learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4936–4945, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018b. PMLR.

Veach, E. *Robust monte carlo methods for light transport simulation*. Number 1610. Stanford University PhD thesis, 1997.

Veach, E. and Guibas, L. J. Optimally combining sampling techniques for monte carlo rendering. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 419–428. ACM, 1995.

Williams, C. K. and Rasmussen, C. E. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

Zhan, Y., Ammar, H. B., et al. Theoretically-grounded policy advice from multiple teachers in reinforcement learning settings with applications to negative transfer. *arXiv preprint arXiv:1604.03986*, 2016.

Zhao, T., Hachiya, H., Tangkaratt, V., Morimoto, J., and Sugiyama, M. Efficient sample reuse in policy gradients with parameter-based exploration. *Neural computation*, 25(6):1512–1547, 2013.

A. Computing the Number of Defensive Samples

Algorithm 1 requires a measure of ESS in order to evaluate the quality of a gradient estimate and, consequently, to adapt the batch size. Although several ESS measures for IS have been studied (see, e.g., (Martino et al., 2017)), to the best of our knowledge no measure specifically designed for MIS estimators has been proposed. A recent work by Elvira et al. (2018) has analyzed the classical ESS measure introduced in Section 2 and has empirically demonstrated its effectiveness in MIS. Thus, we have decided to apply it to our context as well. However, since for our application we are satisfied with a lower bound on the ESS, instead of taking the variance of the importance weights under the given proposals, we take it w.r.t. the mixture of these. This is motivated by the following proposition, which follows directly from the fact that the former variance is always smaller than the latter (see (Owen & Zhou, 2000) or Lemma C.1 in Appendix C.5).

Proposition A.1. *Under the balance heuristics, we have that* $\widehat{ESS} := \frac{n}{1+\mathbb{V}\mathrm{ar}[w^{MIS}(\boldsymbol{\tau})]} \geq \frac{n}{d_2(p(\cdot|\boldsymbol{\theta},f)\|q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}))}$.

To estimate the Renyi divergence, we use the fact that the expected value of importance weights under trajectory distribution $q_{\boldsymbol{\alpha}}$ is equal to one, so that: $d_2(p(\cdot|\boldsymbol{\theta},f)\|q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})) = 1 + \mathbb{V}\mathrm{ar}_{q_{\boldsymbol{\alpha}}}[\frac{p(\cdot|\boldsymbol{\theta},f)}{q_{\boldsymbol{\alpha}}}] \simeq 1 + \frac{1}{n}\sum_{i=1}^{n}(w_i - 1)^2$, where the sum is over the trajectories from all the proposals and w_i are their importance weights. This is in practice much better than using a naïve estimate of the second moment, $\frac{1}{n}\sum_{i=1}^{n} w_i^2$, which would lead to an infinite ESS when the target distribution p gets too far from $q_{\boldsymbol{\alpha}}$, and better than taking the sample variance of the weights, $1 + \frac{1}{n-1}\sum_{i=1}^{n}(w_i - \bar{w})^2$, which would result in an ESS of n in such case. Since these degenerate cases are not uncommon in our context (recall the changes in target distribution during learning), it is extremely important to have a guard against them.

Finally, computing the number n_0 of defensive samples to be collected to guarantee a minimum ESS of $\mathrm{ESS}_{\min}$ requires the analysis of the increase rate of the function of Proposition A.1 when adding the target trajectories. Although, in the asymptotic case, this rate is 1 (i.e., every new target sample increases the ESS by 1), this may not hold when a finite sample is considered. However, we show that, for a given weight vector $\boldsymbol{w}$, this rate cannot be worse than $c := \frac{\bar{\boldsymbol{w}}_3 + 3(1-\bar{\boldsymbol{w}})}{(1+\widehat{\mathbb{V}\mathrm{ar}}[\boldsymbol{w}])^2}$, where $\bar{\boldsymbol{w}}_3 = \frac{1}{n}\sum_{i=1}^{n} w_i^3$, $\bar{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^{n} w_i$, and $\widehat{\mathbb{V}\mathrm{ar}}[\boldsymbol{w}] = \frac{1}{n}\sum_{i=1}^{n}(w_i - 1)^2$. This leads to the following proposition, whose proof can be found in Appendix C.

Proposition A.2. *The number* n_0 *of defensive samples to guarantee an ESS greater than or equal to* ESS_{min} *can be computed as* $n_0 = \max\{n_{min}, \min\{ESS_{min}, n_0'\}\}$, *where*

$$n_0' = \left\lceil \frac{ESS_{min} - \frac{n}{1+\widehat{\mathbb{V}\mathrm{ar}}[\boldsymbol{w}]}}{\min\{1, c\}} \right\rceil . \tag{12}$$

B. Estimating the Source Models

The model-estimation algorithms of Section 4, starting from Theorem 4.1, are derived only for the case where the source tasks (i.e., their transition models) are fully known. Here we show how the proposed methods can be straightforwardly generalized when such an assumption does not hold.

We first extend the upper bound on the MSE of Theorem 4.1 to account for inexact source models. For each $j \in \mathcal{J}$, we now have an arbitrary function $\widetilde{f}_j \in \mathcal{F}$, while all true models $f_j \sim \varphi_j$ are uncertain according to distributions $\varphi_j \sim \Delta(\mathcal{F})$. Similarly to Section 4, we use ∇J to denote the true gradient at $\boldsymbol{\theta}$ and $\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m)$ to denote the MIS estimator with arbitrarily chosen models.

Theorem B.1. *Let $\widetilde{f}_0, \dots, \widetilde{f}_m : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ be arbitrary functions and suppose that $\|g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\|_\infty \leq B$ almost surely. Then,*

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|^2\right] &\leq \frac{dB^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}_0, \widetilde{f}_0) \| q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f}_0, \dots, \widetilde{f}_m)\right) \\ &+ c_1 dB^2 \sum_{l=0}^{m} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[\|\bar{f}_l(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}_l(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2\right] \\ &+ c_1 dB^2 \sum_{l=0}^{m} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_l(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + \mathcal{O}(1), \end{aligned} \tag{13}$$

where the expectation is w.r.t. $\boldsymbol{\tau}_{i,j} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ and $f_j \sim \varphi_j$. Here $\bar{f}_l(\boldsymbol{s}, \boldsymbol{a}) := \mathbb{E}_{f_l \sim \varphi_l}[f_l(\boldsymbol{s}, \boldsymbol{a})]$, $\boldsymbol{\Sigma}_l(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{C}\mathrm{ov}_{f_l \sim \varphi_l}[f_l(\boldsymbol{s}, \boldsymbol{a})]$, and c_1 is a constant.

As mentioned in Remark 4.1, the only component that really needs to be estimated online is the target model. Furthermore, if we plug the mean estimate of each source model $\bar{f}_l$ into 13, the resulting bound reduces, without considering the covariance terms, to the one for the case of known sources (Theorem 4.1). Therefore, the algorithms derived in the main paper can be easily adopted when the source tasks are estimated. One simply needs to plug these estimates, obtained in batch before learning starts, into the bound of Theorem 4.1 as if they were the true source models.

C. Proofs

C.1. Proof of Theorem 3.1

Theorem 3.1 (Unbiasedness of PD estimator)**.** *Let $h_{j,t}(\boldsymbol{\tau})$ be a function such that, for all $t \in \{0, \dots, T-1\}$ and $\boldsymbol{\tau}$, $\sum_{j=0}^{m} h_{j,t}(\boldsymbol{\tau}) = 1$. Then, the per-decision MIS estimator in (7) is unbiased.*

Proof. The proof simply starts from the definition of PDMIS and shows that, by leveraging the assumption that the heuristic function is a partition of unity, its expected value is the policy gradient:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\nabla}^{PD}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, f)\right] &= \sum_{j=0}^{m} \mathbb{E}_{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)} \left[\sum_{t=0}^{T-1} h_{j,t}(\boldsymbol{\tau}) \frac{p(\boldsymbol{\tau}_{0:t}|\boldsymbol{\theta}, f)}{p(\boldsymbol{\tau}_{0:t}|\boldsymbol{\theta}_j, f_j)} \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) g_t(\boldsymbol{\tau})\right] \\
&= \sum_{j=0}^{m} \int \sum_{t=0}^{T-1} h_{j,t}(\boldsymbol{\tau}) p(\boldsymbol{\tau}_{0:t}|\boldsymbol{\theta}, f) \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) g_t(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} \\
&= \int \sum_{t=0}^{T-1} p(\boldsymbol{\tau}_{0:t}|\boldsymbol{\theta}, f) \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) g_t(\boldsymbol{\tau}) \underbrace{\sum_{j=0}^{m} h_{j,t}(\boldsymbol{\tau})}_{=1} \mathrm{d}\boldsymbol{\tau} \\
&= \int \sum_{t=0}^{T-1} p(\boldsymbol{\tau}_{0:t}|\boldsymbol{\theta}, f) \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) g_t(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} \\
&= \mathbb{E}_{p(\boldsymbol{\tau}|\boldsymbol{\theta}, f)} \left[\sum_{t=0}^{T-1} \gamma^t \mathcal{R}(\boldsymbol{s}_t, \boldsymbol{a}_t) \sum_{l=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(\boldsymbol{a}_l|\boldsymbol{s}_l)\right] \\
&= \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}, f).
\end{aligned}$$

□

C.2. Proof of Proposition 3.1

Proposition 3.1. *The estimator (8) is unbiased for any $\boldsymbol{\beta}_d$. Furthermore, under the optimal coefficients $\boldsymbol{\beta}^*_d$, $\mathbb{V}\mathrm{ar}[\widehat{\nabla}^{\mathrm{CV}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P})] \leq \mathbb{V}\mathrm{ar}[\widehat{\nabla}^{\mathrm{MIS}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, \mathcal{P})]$.*

Proof. To prove the unbiasedness, recall that

$$\mathbb{E}_{\boldsymbol{\tau}_{i,j} \sim p(\cdot|\boldsymbol{\theta}_j, f_j)} \left[\widehat{\nabla}^{\mathrm{MIS}}_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, f)\right] = \nabla_{\boldsymbol{\theta}_d} J(\boldsymbol{\theta}, f).$$

Thus, we only need to prove that the second term has expected value equal to zero. Hence,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\tau}_{i,j} \sim p(\cdot|\boldsymbol{\theta}_j, f_j)} \left[\frac{1}{n} \sum_{j=0}^{m} \sum_{i=1}^{n_j} \boldsymbol{\beta}_d^T \boldsymbol{\psi}_d(\boldsymbol{\tau}_{i,j})\right] &= \frac{1}{n} \sum_{j=0}^{m} n_j \boldsymbol{\beta}_d^T \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_j, f_j)} \left[\boldsymbol{\psi}_d(\boldsymbol{\tau})\right] \\
&= \frac{1}{n} \sum_{j=0}^{m} n_j \sum_{l=0}^{m+1} \beta_{l,d} \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_j, f_j)} \left[\psi_{l,d}(\boldsymbol{\tau})\right].
\end{aligned}$$

Recall that we consider $\psi_{j,d}(\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})} - 1$ for $j = 0, \dots, m$ and $\psi_{m+1,d}(\boldsymbol{\tau}) = \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) g_d(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})}$. The first m CVs are well-known to have zero expectation (Owen & Zhou, 2000). In fact,

$$\begin{aligned}
\frac{1}{n} \sum_{j=0}^{m} n_j \sum_{l=0}^{m} \beta_{l,d} \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_j, f_j)} \left[\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}_l, f_l)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})} - 1\right] &= \sum_{l=0}^{m} \beta_{l,d} \int \sum_{j=0}^{m} \frac{n_j}{n} p(\cdot|\boldsymbol{\theta}_j, f_j) \left(\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}_l, f_l)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})} - 1\right) \mathrm{d}\boldsymbol{\tau} \\
&= \sum_{l=0}^{m} \beta_{l,d} \left(\int p(\boldsymbol{\tau}|\boldsymbol{\theta}_l, f_l) \mathrm{d}\boldsymbol{\tau} - \int q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}\right) = 0.
\end{aligned}$$

The $(m+1)$-th term is the well-known baseline commonly adopted in policy gradient methods. It's expectation can be

easily verified to be zero:

$$\beta_{m+1,d}\frac{1}{n}\sum_{j=0}^{m} n_j \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_j,f_j)}\left[\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},f)g_d(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})}\right] = \beta_{m+1,d}\int p(\boldsymbol{\tau}|\boldsymbol{\theta},f)g_d(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau}$$
$$= \beta_{m+1,d}\int p(\boldsymbol{\tau}|\boldsymbol{\theta},f)\sum_{t=0}^{T-1}\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)\mathrm{d}\boldsymbol{\tau}.$$

The last integral can be rewritten as

$$\sum_{t=0}^{T-1}\int\mathcal{P}_0(\boldsymbol{s}_0)\int\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_0|\boldsymbol{s}_0)\cdots\int\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)\mathrm{d}\boldsymbol{s}_0\ldots\mathrm{d}\boldsymbol{a}_t,$$

which is equal to zero since $\int\nabla_{\boldsymbol{\theta}}\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})\mathrm{d}\boldsymbol{a}=0$ for any state $\boldsymbol{s}$. This concludes the proof of the first statement.

In order to prove the second statement, let $\boldsymbol{z}=[0,0,\ldots,0]$ be the $(m+1)$-th dimensional vector of zeros. Then, under the coefficients $\boldsymbol{\beta}_d^*$ minimizing the variance of the CV estimator (8),

$$\mathbb{V}\mathrm{ar}[\widehat{\nabla}^{\mathrm{CV}}_{\boldsymbol{\theta}_d}J(\boldsymbol{\theta},f)] = \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathrm{MIS}}_{\boldsymbol{\theta}_d}J(\boldsymbol{\theta},f)-\frac{1}{n}\sum_{j=0}^{m}\sum_{i=1}^{n_j}\boldsymbol{\beta}_d^T\boldsymbol{\psi}_d(\boldsymbol{\tau}_{i,j})\right]$$
$$\leq \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathrm{MIS}}_{\boldsymbol{\theta}_d}J(\boldsymbol{\theta},f)-\frac{1}{n}\sum_{j=0}^{m}\sum_{i=1}^{n_j}\boldsymbol{z}^T\boldsymbol{\psi}_d(\boldsymbol{\tau}_{i,j})\right]$$
$$= \mathbb{V}\mathrm{ar}[\widehat{\nabla}^{\mathrm{MIS}}_{\boldsymbol{\theta}_d}J(\boldsymbol{\theta},f)].$$

□

C.3. Proof of Theorem 3.2

Theorem 3.2. *Assume the return J is L-smooth (i.e., its gradient is L-Lipschitz). Let $n_{min}>0$ be the minimum batch size for $\mathscr{A}_{CV}$ ($\mathscr{A}_{PDCV}$) and the fixed batch size for $\mathscr{B}_R$ ($\mathscr{B}_G$). Assume all algorithms start from the same parameter $\boldsymbol{\theta}_0$, use a learning rate $0<\eta\leq\frac{2}{L}$, and that $\mathscr{A}_{CV}$ ($\mathscr{A}_{PDCV}$) uses the optimal CV coefficients $\boldsymbol{\beta}_d^*$. Then, for all $\epsilon>0$:*

$$\mathscr{A}_{CV}\not\prec\mathscr{B}_R,\quad \mathscr{A}_{PDCV}\not\prec\mathscr{B}_G.$$

Proof. Let $\tilde{J}(\boldsymbol{\theta}):=-J(\boldsymbol{\theta},f)$ and consider the equivalent problem $\arg\min_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta})$. Following standard proofs of convergence for non-convex stochastic optimization (see, e.g., Appendix B of (Allen-Zhu, 2017)), we can show that, for a fixed parameter $\boldsymbol{\theta}_k$ and algorithm $\mathscr{A}$,

$$\tilde{J}(\boldsymbol{\theta}_k)-\mathbb{E}_{\mathscr{A}}\left[\tilde{J}(\boldsymbol{\theta}_{k+1})\right]\geq\left(\eta-\frac{\eta^2L}{2}\right)\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_k)\|_2^2-\frac{\eta^2L}{2}\mathbb{V}\mathrm{ar}_{\mathscr{A}}\left[\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}(\tilde{J}(\boldsymbol{\theta}_k))\right], \tag{14}$$

where $\boldsymbol{\theta}_{k+1}=\boldsymbol{\theta}_k-\eta\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_k)$ and the expectations are taken w.r.t. the stochasticity in the estimation of the gradient. Rearranging (14), we obtain

$$\left(\eta-\frac{\eta^2L}{2}\right)\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_k)\|_2^2\leq\tilde{J}(\boldsymbol{\theta}_k)-\mathbb{E}_{\mathscr{A}}\left[\tilde{J}(\boldsymbol{\theta}_{k+1})\right]+\frac{\eta^2L}{2}\mathbb{V}\mathrm{ar}_{\mathscr{A}}\left[\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_k)\right]. \tag{15}$$

Let us now take the expectation under the whole stochastic process $\boldsymbol{\theta}_{0:k}$, with $\boldsymbol{\theta}_0$ being deterministic and fixed, and sum over iterations the first k iterations. Then,

$$\left(\eta-\frac{\eta^2L}{2}\right)\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}}\left[\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\|_2^2\right]\leq\tilde{J}(\boldsymbol{\theta}_0)-\mathbb{E}_{\mathscr{A}}\left[\tilde{J}(\boldsymbol{\theta}_{k+1})\right]+\frac{\eta^2L}{2}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}}\left[\mathbb{V}\mathrm{ar}_{\mathscr{A}}\left[\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\mid\boldsymbol{\theta}_l\right]\right]$$
$$\leq\tilde{J}(\boldsymbol{\theta}_0)-\tilde{J}(\boldsymbol{\theta}^*)+\frac{\eta^2L}{2}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}}\left[\mathbb{V}\mathrm{ar}_{\mathscr{A}}\left[\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\mid\boldsymbol{\theta}_l\right]\right],$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \tilde{J}(\boldsymbol{\theta})$. Rearranging,

$$\frac{1}{k}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}}\left[\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\|_2^2\right] \leq \frac{1}{k\left(\eta - \frac{\eta^2 L}{2}\right)}\left(\tilde{J}(\boldsymbol{\theta}_0) - \tilde{J}(\boldsymbol{\theta}^*)\right) + \frac{\eta L}{2-\eta L}\frac{1}{k}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}}\left[\mathbb{V}\mathrm{ar}_{\mathscr{A}}\left[\widehat{\nabla}^{\mathscr{A}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l) \mid \boldsymbol{\theta}_l\right]\right]. \tag{16}$$

Let us now compare $\mathscr{B}_{\mathrm{R}}$ and $\mathscr{A}_{\mathrm{CV}}$. From Theorem 2 of Owen & Zhou (2000), we know that, for every $j = 1, \dots, m$, a mixture IS estimator with proportions $\boldsymbol{\alpha}$ using the optimal CV parameter $\boldsymbol{\beta}^*$ has a variance that is upper bounded by that of an IS estimator using only the j-th proposal divided by the proportion α_j of samples from such proposal. Furthermore, this property holds for a MIS estimator as well since its variance is always smaller than the one of the corresponding mixture estimator. In our context, the algorithm $\mathscr{A}_{\mathrm{CV}}$ uses the MIS estimator (8) with the optimal CV coefficients. Thus, for any $\boldsymbol{\theta}$ and dimension d,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})\right] \leq \min_{j=0,\dots,m}\frac{\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\text{IS-}j}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})\right]}{\alpha_j}, \tag{17}$$

where $\widehat{\nabla}^{\text{IS-}j}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})$ denotes an IS estimator using n samples from the j-th proposal $p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ only. Recalling that the 0-th proposal corresponds to the current target distribution, $p(\boldsymbol{\tau}|\boldsymbol{\theta}, f)$, and that, by assumption, the minimum number of trajectories that $\mathscr{A}_{\mathrm{CV}}$ collects at each step is n_{min}, we obtain

$$\begin{aligned}
\min_{j=0,\dots,m}\frac{\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\text{IS-}j}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})\right]}{\alpha_j} &\leq \frac{\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\text{IS-}0}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})\right]}{\alpha_0} \\
&= \frac{\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n_0} w_0^{\mathrm{IS}}(\boldsymbol{\tau}_i) g(\boldsymbol{\tau}_i)\mathcal{R}(\boldsymbol{\tau}_i)\right]}{\alpha_0} \\
&= \frac{\frac{1}{n}\mathbb{V}\mathrm{ar}\left[g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\right]}{\alpha_0} \\
&\leq \frac{1}{n_{\mathrm{min}}}\mathbb{V}\mathrm{ar}\left[g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\right] \\
&= \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{B}_{\mathrm{R}}}_{\boldsymbol{\theta}_d}\tilde{J}(\boldsymbol{\theta})\right].
\end{aligned}$$

The second equality follows from the fact that $w_0^{\mathrm{IS}}(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}, f)}{p(\boldsymbol{\tau}|\boldsymbol{\theta}, f)} = 1$. Since this holds for all the policy dimensions d and $\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}}(\tilde{J}(\boldsymbol{\theta}))\right] = \mathrm{Tr}(\mathbb{C}\mathrm{ov}[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}}(\tilde{J}(\boldsymbol{\theta}))])$, we obtain

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta})\right] \leq \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{B}_{\mathrm{R}}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta})\right], \tag{18}$$

i.e., the variance of the transfer algorithm in estimating the gradients is always smaller than the one of the no-transfer baseline. Furthermore, by assumption, the variance of the no-transfer baseline is bounded. Let $C_{\mathscr{B}_{\mathrm{R}}}$ be its bound. Then,

$$\frac{1}{k}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{B}_{\mathrm{R}}}\left[\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\|_2^2\right] \leq \frac{1}{k\left(\eta - \frac{\eta^2 L}{2}\right)}\left(\tilde{J}(\boldsymbol{\theta}_0) - \tilde{J}(\boldsymbol{\theta}^*)\right) + \frac{\eta L}{2-\eta L}C_{\mathscr{B}_{\mathrm{R}}}. \tag{19}$$

Suppose now that the upper bound (19) is less or equal then ϵ, which implies that $\mathscr{B}_{\mathrm{R}}$ converged. Since we showed that $\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta})\right] \leq C_{\mathscr{B}_{\mathrm{R}}}$,

$$\begin{aligned}
\frac{1}{k}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}_{\mathrm{CV}}}\left[\|\nabla_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l)\|_2^2\right] &\leq \frac{1}{k\left(\eta - \frac{\eta^2 L}{2}\right)}\left(\tilde{J}(\boldsymbol{\theta}_0) - \tilde{J}(\boldsymbol{\theta}^*)\right) + \frac{\eta L}{2-\eta L}\frac{1}{k}\sum_{l=0}^{k-1}\mathbb{E}_{\mathscr{A}_{\mathrm{CV}}}\left[\mathbb{V}\mathrm{ar}_{\mathscr{A}_{\mathrm{CV}}}\left[\widehat{\nabla}^{\mathscr{A}_{\mathrm{CV}}}_{\boldsymbol{\theta}}\tilde{J}(\boldsymbol{\theta}_l) \mid \boldsymbol{\theta}_l\right]\right] \\
&\leq \frac{1}{k\left(\eta - \frac{\eta^2 L}{2}\right)}\left(\tilde{J}(\boldsymbol{\theta}_0) - \tilde{J}(\boldsymbol{\theta}^*)\right) + \frac{\eta L}{2-\eta L}C_{\mathscr{B}_{\mathrm{R}}} \leq \epsilon.
\end{aligned}$$

Hence, whenever we are able to prove that $\mathscr{B}_{\mathrm{R}}$ converged, we are also able to prove that $\mathscr{A}_{\mathrm{CV}}$ converged, which is exactly our definition of robustness against negative transfer.

The proof for $\mathscr{A}_{\mathrm{PDCV}}$ and $\mathscr{B}_{\mathrm{G}}$ proceeds analogously by noticing that the variance of the former is always less or equal than the variance of the latter. □

C.4. Proof of Proposition A.2

Proposition A.2. *The number n_0 of defensive samples to guarantee an ESS greater than or equal to ESS_{min} can be computed as $n_0 = \max\{n_{min}, \min\{ESS_{min}, n_0'\}\}$, where*

$$n_0' = \left\lceil \frac{ESS_{min} - \frac{n}{1+\widehat{\mathbb{V}\mathrm{ar}}[\boldsymbol{w}]}}{\min\{1, c\}} \right\rceil. \tag{12}$$

Suppose our current dataset $\mathcal{D}$ contains n trajectories, with the target distribution being $p(\boldsymbol{\tau}) = p(\boldsymbol{\tau}|\boldsymbol{\theta}, f)$ and the mixture of source proposals being $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) = \sum_{j=1}^{m} \alpha_j p_j(\boldsymbol{\tau})$, with $p_j(\boldsymbol{\tau}) = p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$. Notice that $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau})$ does not necessarily contain the target distribution p since the number of defensive samples has still to be computed. If we add n_0 defensive samples, the resulting ESS (according to Proposition A.1) is

$$\begin{aligned}
\mathrm{ESS}(n_0; \mathcal{D}) &= \frac{n + n_0}{\int \frac{p(\boldsymbol{\tau})^2}{\sum_{j=1}^{m} \frac{n_j}{n+n_0} p_j(\boldsymbol{\tau}) + \frac{n_0}{n+n_0} p(\boldsymbol{\tau})} \mathrm{d}\boldsymbol{\tau}} \\
&= \frac{n + n_0}{\mathbb{E}_{\boldsymbol{\tau} \sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} \left[\frac{p(\boldsymbol{\tau})^2}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)^2} \right]} \\
&= \frac{n + n_0}{1 + \mathbb{V}\mathrm{ar}_{\boldsymbol{\tau} \sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} \left[\frac{p(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} \right]} \\
&= \frac{n + n_0}{1 + \int q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0) \left(\frac{p(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} - 1 \right)^2},
\end{aligned}$$

where we use $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)$ to denote the updated mixture after collecting n_0 defensive trajectories from p. Let us now approximate the variance term using our current samples. To simplify the notation, let us index the samples with $i = 1, \dots, n$, while dropping the index j of the proposal which generated the sample itself (under the balance heuristics, the weight does not depend on j). We have

$$\begin{aligned}
\mathbb{V}\mathrm{ar}_{\boldsymbol{\tau} \sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} \left[\frac{p(\boldsymbol{\tau})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; n_0)} \right] &\simeq \frac{1}{n} \sum_{i=1}^{n} \frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i; n_0)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i)} \left(\frac{p(\boldsymbol{\tau}_i)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i; n_0)} - 1 \right)^2 \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{p(\boldsymbol{\tau}_i)^2}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i) q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i; n_0)} + \frac{1}{n} \sum_{i=1}^{n} \frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i; n_0)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i)} - \frac{2}{n} \sum_{i=1}^{n} \frac{p(\boldsymbol{\tau}_i)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i)} \\
&= (n + n_0) \sum_{i=1}^{n} \frac{a_i^2}{b_i(b_i + a_i n_0)} + \frac{n}{n + n_0} + \frac{n_0}{n + n_0} \sum_{i=1}^{n} \frac{a_i}{b_i} - 2 \sum_{i=1}^{n} \frac{a_i}{b_i},
\end{aligned}$$

where we defined the constants $a_i := p(\boldsymbol{\tau}_i)$ and $b_i := \sum_{j=1}^{m} n_j p_j(\boldsymbol{\tau}_i)$. Thus, the ESS improvement as a function of n_0 can be approximated as

$$\begin{aligned}
\widehat{\mathrm{ESS}}(n_0; \mathcal{D}) &= \frac{n + n_0}{1 + (n + n_0) \sum_{i=1}^{n} \frac{a_i^2}{b_i(b_i + a_i n_0)} + \frac{n}{n+n_0} + \frac{n_0}{n+n_0} \sum_{i=1}^{n} \frac{a_i}{b_i} - 2 \sum_{i=1}^{n} \frac{a_i}{b_i}} \\
&= \frac{1}{\sum_{i=1}^{n} \frac{a_i^2}{b_i(b_i + a_i n_0)} + \frac{n}{(n+n_0)^2} + \frac{n_0}{(n+n_0)^2} \sum_{i=1}^{n} \frac{a_i}{b_i} + \frac{1}{n+n_0} \left(1 - 2 \sum_{i=1}^{n} \frac{a_i}{b_i} \right)}.
\end{aligned}$$

Note that $\widehat{\mathrm{ESS}}(0; \mathcal{D}) = \frac{n}{1 + \frac{1}{n} \sum_{i=1}^{n} \left(\frac{p(\boldsymbol{\tau}_i)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i)} - 1 \right)^2}$, which is exactly our ESS measure. Furthermore, this function is strictly increasing for $n_0 \geq 0$ and $\lim_{n_0 \to \infty} \frac{\widehat{\mathrm{ESS}}(n_0; \mathcal{D})}{n_0} = 1$, i.e., the asymptotic increase rate is linear with slope 1. Therefore, $\widehat{\mathrm{ESS}}(n_0; \mathcal{D}) \geq \widehat{\mathrm{ESS}}(0; \mathcal{D}) + n_0 \inf_{x \in (0, +\infty)} \widehat{\mathrm{ESS}}'(x; \mathcal{D})$. It is easy to check that $\inf_{x \in (0, +\infty)} \widehat{\mathrm{ESS}}'(x)$ is either 1, when the function grows at a rate that is always grater than the asymptotic one, or c, when the initial rate is smaller. Thus,

$$\widehat{\mathrm{ESS}}(n_0; \mathcal{D}) \geq \frac{n}{1 + \frac{1}{n} \sum_{i=1}^{n} \left(\frac{p(\boldsymbol{\tau}_i)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_i)} - 1 \right)^2} + \min\{1, c\} n_0.$$

Using this equation and setting the right-hand side to $\mathrm{ESS}_{\mathrm{min}}$, we get that the number n_0 of defensive samples to guarantee

an ESS of at least ESS_{min} can be approximated as

$$n_0 = \left\lceil \frac{\text{ESS}_{\text{min}} - \frac{n}{1+\widehat{\mathbb{V}\text{ar}}[\boldsymbol{w}]}}{\min\{1, c\}} \right\rceil.$$

Finally, we clip this value to n_{min} below, as required by our algorithm, and to ESS_{min} above since the collecting ESS_{min} samples is sufficient to guarantee an ESS of at least such value. In fact, if we have ESS_{min} samples from the target p in our dataset, $d_2(p\|q_{\boldsymbol{\alpha}}) \leq \frac{n}{\text{ESS}_{\text{min}}}$ since the importance weights are bounded by $\frac{1}{\alpha_0} = \frac{n}{\text{ESS}_{\text{min}}}$. Hence, $\frac{n}{d_2(p\|q_{\boldsymbol{\alpha}})} \geq \text{ESS}_{\text{min}}$.

C.5. Proof of Theorem 4.1

To prove Theorem 4.1 we need to introduce the following Lemma about the variance of the sample mean estimator.

Lemma C.1. *Let $Q_1, \dots, Q_m$ be probability measures over $(\mathcal{X}, \mathscr{F})$, $Q_\alpha = \sum_{j=1}^m \alpha_j Q_j$ be a mixture of these measures with coefficients $\alpha_j \geq 0$ such that $\sum_{j=1}^m \alpha_j = 1$, and $f : \mathcal{X} \to \mathbb{R}$ be any measurable function. Consider $\hat{\mu} = \frac{1}{n}\sum_{j=1}^m \sum_{i=1}^{n_j} f(x_{i,j})$ where $x_{i,j}$ are i.i.d. samples and $n = \sum_{j=1}^m n_j$. Then, choosing $\alpha_j = \frac{n_j}{n}$, for each $j \in \{1, \dots, m\}$, $\mathbb{V}\text{ar}_{x_{i,j}\sim Q_j}[\hat{\mu}] \leq \mathbb{V}\text{ar}_{x_{i,j}\sim Q_\alpha}[\hat{\mu}]$.*

Proof. Let $\mu = \mathbb{E}_{x\sim Q_\alpha}[f(x)]$ and $\mu_j = \mathbb{E}_{x\sim Q_j}[f(x)]$. Then,

$$\begin{aligned}
\mathbb{V}\text{ar}_{x_{i,j}\sim Q_\alpha}[\hat{\mu}] &= \frac{1}{n^2}\sum_{j=1}^m n_j \int q_j(x)(f(x)-\mu)^2 dx \\
&= \frac{1}{n^2}\sum_{j=1}^m n_j \int q_j(x)(f(x)-\mu\pm\mu_j)^2 dx \\
&= \frac{1}{n^2}\sum_{j=1}^m n_j \int q_j(x)(f(x)-\mu_j)^2 dx + \frac{1}{n^2}\sum_{j=1}^m n_j \int q_j(x)(\mu_j-\mu)^2 dx \\
&= \mathbb{V}\text{ar}_{x_{i,j}\sim Q_j}[\hat{\mu}] + \frac{1}{n}\sum_{j=1}^m \alpha_j(\mu_j-\mu)^2 \\
&\geq Var_{x_{i,j}\sim Q_j}[\hat{\mu}].
\end{aligned}$$

□

Theorem 4.1. *Let $\widetilde{f} : \mathcal{S}\times\mathcal{A} \to \mathcal{S}$ be any function and $p_{\boldsymbol{\alpha}}(\boldsymbol{\tau}) = \sum_{j\in\mathcal{J}_{tgt}} \frac{\alpha_j}{\alpha_{tgt}} p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f})$. Suppose that $\|g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\|_\infty \leq B$ almost surely. Then, for $f \sim \varphi$,*

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\nabla}J(\widetilde{f}) - \nabla J\|^2\right] &\leq \frac{dB^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}, \widetilde{f})\|q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f})\right) \\
&+ c_1 dB^2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\alpha}}}\left[\|\bar{f}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2\right] \\
&+ c_1 dB^2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\alpha}}}\left[\text{Tr}\left(\boldsymbol{\Sigma}(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + \mathcal{O}(n^{-1}),
\end{aligned} \tag{9}$$

where the expectation is w.r.t. $\boldsymbol{\tau}_{i,j} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ and $f \sim \varphi$. Here $\bar{f}(\boldsymbol{s}, \boldsymbol{a}) := \mathbb{E}_{f\sim\varphi}[f(\boldsymbol{s}, \boldsymbol{a})]$, $\boldsymbol{\Sigma}(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{C}\text{ov}_{f\sim\varphi}[f(\boldsymbol{s}, \boldsymbol{a})]$, and c_1 is a constant.

Proof. Since $\mathbb{E}_{\boldsymbol{\tau}_{i,j}\sim p(\cdot|\boldsymbol{\theta}_j, f_j), f\sim\varphi}\left[\|\widehat{\nabla}J(\widetilde{f}) - \nabla J\|_2^2\right] = \mathbb{E}_{f\sim\varphi}\left[\mathbb{E}_{\boldsymbol{\tau}_{i,j}\sim p(\cdot|\boldsymbol{\theta}_j, f_j)}\left[\|\widehat{\nabla}J(\widetilde{f}) - \nabla J\|_2^2 \middle| f\right]\right]$, let us start by bounding the inner expectation given a fixed f. Thus, whenever not explicitly stated, expectations are taken under

the trajectories $\boldsymbol{\tau}_{i,j}$ distributed according to $p(\cdot|\boldsymbol{\theta}_j, f_j)$. We start by decomposing the MSE into variance and bias squared:

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\nabla}J(\widetilde{f}) - \nabla J\|_2^2\right] &= \sum_d \mathbb{E}\left[\left(\widehat{\nabla}_d J(\widetilde{f}) - \nabla_d J\right)^2\right] \\ &= \sum_d \mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_d J(\widetilde{f})\right] + \sum_d \left(\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right] - \nabla_d J\right)^2,\end{aligned}$$

where we use $\nabla_d J$ and $\widehat{\nabla}_d J$ to denote the d-th component of the gradient. We now analyze the two terms separately. Regarding the variance, we have

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_d J(\widetilde{f})\right] &= \mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{j=0}^{m}\sum_{i=1}^{n_j}\frac{p(\boldsymbol{\tau}_{i,j}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}_{i,j};\widetilde{f})} g(\boldsymbol{\tau}_{i,j})\mathcal{R}(\boldsymbol{\tau}_{i,j})\right] \\ &= \frac{1}{n^2}\sum_{j=0}^{m} n_j \mathbb{V}\mathrm{ar}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_j,f_j)}\left[\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\right] \\ &\leq \frac{1}{n}\mathbb{V}\mathrm{ar}_{\boldsymbol{\tau}\sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f)}\left[\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\right] \\ &\leq \frac{1}{n}\mathbb{E}_{\boldsymbol{\tau}\sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f)}\left[\frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q^2_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} g^2(\boldsymbol{\tau})\mathcal{R}^2(\boldsymbol{\tau})\right] \\ &\leq \frac{B^2}{n}\mathbb{E}_{\boldsymbol{\tau}\sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f)}\left[\frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q^2_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})}\right],\end{aligned}$$

where the first equality leverages trajectory independence and the first inequality follows from Lemma C.1. The last expectation can be further decomposed as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\tau}\sim q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f)}\left[\frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q^2_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})}\right] &= \int \left(q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f) \pm q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})\right)\frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q^2_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})}\mathrm{d}\boldsymbol{\tau} \\ &= \int \frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})}\mathrm{d}\boldsymbol{\tau} + \int \sum_{j\in\mathcal{J}_{\text{tgt}}}\alpha_j\left(p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,\widetilde{f})\right)\frac{p^2(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q^2_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})}\mathrm{d}\boldsymbol{\tau} \\ &\leq d_2\left(p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})||q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})\right) + \frac{1}{\alpha_0^2}\int\sum_{j\in\mathcal{J}_{\text{tgt}}}\alpha_j\left|p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,\widetilde{f})\right|\mathrm{d}\boldsymbol{\tau} \\ &= d_2\left(p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})||q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})\right) + \frac{2}{\alpha_0^2}\sum_{j\in\mathcal{J}_{\text{tgt}}}\alpha_j D_{\text{TV}}\left(p(\cdot|\boldsymbol{\theta}_j,f)\|p(\cdot|\boldsymbol{\theta}_j,\widetilde{f})\right),\end{aligned}$$

where D_{TV} is the total variation divergence. Note that the last inequality is valid since $\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} \leq \frac{1}{\alpha_0}$ thank to the defensive component in $q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})$ (see Section 2). Thus,

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_d J(\widetilde{f})\right] \leq \frac{B^2}{n} d_2\left(p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})||q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})\right) + \frac{B^2}{n}\frac{2}{\alpha_0^2}\sum_{j\in\mathcal{J}_{\text{tgt}}}\alpha_j D_{\text{TV}}\left(p(\cdot|\boldsymbol{\theta}_j,f)\|p(\cdot|\boldsymbol{\theta}_j,\widetilde{f})\right). \tag{20}$$

Let us now consider the bias term. First note that $\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right]$ can be written as

$$\begin{aligned}\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right] &= \frac{1}{n}\sum_{j=0}^{m} n_j \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_j,f_j)}\left[\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} g_d(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\right] \\ &= \int \frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};f)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau};\widetilde{f})} p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f}) g_d(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\mathrm{d}\boldsymbol{\tau},\end{aligned}$$

while $\nabla_d J = \int p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) g_d(\boldsymbol{\tau}) \mathcal{R}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau}$. Then,

$$\begin{aligned}
\left| \mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right] - \nabla_d J \right| &= \left| \mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right] \pm \int p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) g_d(\boldsymbol{\tau}) \mathcal{R}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} - \nabla_d J \right| \\
&\leq \left| \int \left(\frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; f)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; \widetilde{f})} - 1 \right) p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) g_d(\boldsymbol{\tau}) \mathcal{R}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} \right| + \left| \int \left(p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) - p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) \right) g_d(\boldsymbol{\tau}) \mathcal{R}(\boldsymbol{\tau}) \mathrm{d}\boldsymbol{\tau} \right| \\
&\leq B \int \left| \frac{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; f)}{q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; \widetilde{f})} - 1 \right| p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) \mathrm{d}\boldsymbol{\tau} + B \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) - p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) \right| \mathrm{d}\boldsymbol{\tau} \\
&\leq \frac{B}{\alpha_0} \int \left| q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; f) - q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}; \widetilde{f}) \right| \mathrm{d}\boldsymbol{\tau} + B \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) - p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) \right| \mathrm{d}\boldsymbol{\tau} \\
&= \frac{B}{\alpha_0} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f}) \right| \mathrm{d}\boldsymbol{\tau} + B \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f}) - p(\boldsymbol{\tau}|\boldsymbol{\theta}, f) \right| \mathrm{d}\boldsymbol{\tau}.
\end{aligned}$$

Since the first addendum contains the second one (for $j = 0$), this equation can be upper bounded by $2\frac{B}{\alpha_0} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f}) \right| \mathrm{d}\boldsymbol{\tau}$. Thus,

$$\begin{aligned}
\left(\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f})\right] - \nabla_d J \right)^2 &\leq 4 \frac{B^2}{\alpha_0^2} \left(\sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f}) \right| \mathrm{d}\boldsymbol{\tau} \right)^2 \\
&\leq 4 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \left(\int \left| p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f) - p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f}) \right| \mathrm{d}\boldsymbol{\tau} \right)^2 \\
&= 8 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j D_{\text{TV}} \left(p(\cdot|\boldsymbol{\theta}_j, f) \| p(\cdot|\boldsymbol{\theta}_j, \widetilde{f}) \right)^2.
\end{aligned}$$

where in the second inequality we used Jensen's inequality. Summing the last term with the corresponding one in (20), we obtain

$$8 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \left(\frac{1}{4n} D_{\text{TV}} \left(p(\cdot|\boldsymbol{\theta}_j, f) \| p(\cdot|\boldsymbol{\theta}_j, \widetilde{f}) \right) + D_{\text{TV}}^2 \left(p(\cdot|\boldsymbol{\theta}_j, f) \| p(\cdot|\boldsymbol{\theta}_j, \widetilde{f}) \right) \right).$$

Since, $kx \leq x^2 + \frac{k^2}{2}$, this equation can be upper bounded by

$$16 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j D_{\text{TV}}^2 \left(p(\cdot|\boldsymbol{\theta}_j, f) \| p(\cdot|\boldsymbol{\theta}_j, \widetilde{f}) \right) + \frac{B^2 \alpha_{\text{tgt}}}{4 \alpha_0^2 n^2}$$

The first term can be upper bounded using Pinsker's inequality as

$$\begin{aligned}
16 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j D_{\text{TV}}^2 \left(p(\cdot|\boldsymbol{\theta}_j, f) \| p(\cdot|\boldsymbol{\theta}_j, \widetilde{f}) \right) &\leq 8 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j D_{\text{KL}} \left(p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f}) \| p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f) \right) \\
&= 8 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_j, \widetilde{f})} \left[\log \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, \widetilde{f})}{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f)} \right] \\
&= 8 \frac{B^2}{\alpha_0^2} \sum_{j \in \mathcal{J}_{\text{tgt}}} \alpha_j \mathbb{E}_{\boldsymbol{\tau} \sim p(\cdot|\boldsymbol{\theta}_j, \widetilde{f})} \left[\sum_{t=0}^{T-1} \log \frac{\mathcal{P}_{\widetilde{f}}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)}{\mathcal{P}_f(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t, \boldsymbol{a}_t)} \right] \\
&= 8 \frac{B^2 \alpha_{\text{tgt}}}{\alpha_0^2} \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau} \sim p_{\boldsymbol{\alpha}}} \left[D_{\text{KL}} \left(\mathcal{P}_{\widetilde{f}}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t) \| \mathcal{P}_f(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t) \right) \right].
\end{aligned}$$

By combining the bounds on variance and bias and summing over the gradient dimensions, we obtain

$$\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}) - \nabla J\|^2\right] \leq d\frac{B^2}{n} d_2\left(p(\boldsymbol{\tau}|\boldsymbol{\theta}, \widetilde{f})||q_{\boldsymbol{\alpha}}(\boldsymbol{\tau}|\widetilde{f})\right) + 8d\frac{B^2\alpha_{\text{tgt}}}{\alpha_0^2}\sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\alpha}}}\left[D_{\text{KL}}\left(\mathcal{P}_{\widetilde{f}}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\|\mathcal{P}_f(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + d\frac{B^2\alpha_{\text{tgt}}}{4\alpha_0^2 n^2}. \tag{21}$$

If we now consider the outer expectation over $f \sim \varphi$, we note that only the bias term depends on f. Since $D_{\text{KL}}\left(\mathcal{P}_{\widetilde{f}}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\|\mathcal{P}_f(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\right) = \frac{1}{2\sigma_{\mathcal{P}}^2}\|f(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2$, the expected KL divergence is

$$\begin{aligned}\mathbb{E}_{f\sim\varphi}\left[D_{\text{KL}}\left(\mathcal{P}_{\widetilde{f}}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\|\mathcal{P}_f(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] &= \frac{1}{2\sigma_{\mathcal{P}}^2}\mathbb{E}_{f\sim\varphi}\left[\|f(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2\right] \\ &= \frac{1}{2\sigma_{\mathcal{P}}^2}\|\bar{f}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2 + \frac{1}{2\sigma_{\mathcal{P}}^2}\text{Tr}\left(\mathbb{C}\text{ov}_{f\sim\varphi}[f(\boldsymbol{s}, \boldsymbol{a})]\right).\end{aligned}$$

The theorem follows by plugging this last equation into (21) and noticing that the constant term $d\frac{B^2\alpha_{\text{tgt}}}{4\alpha_0^2 n^2}$ decreases as $\mathcal{O}(n^{-1})$. □

C.6. Proof of Theorem 4.2

Theorem 4.2. *Let $(\mathcal{X}, \mathscr{F})$ be a measurable space, P and Q be two probability measures on $\mathcal{X}$ such that $P \ll Q$, and $Q_\alpha = \alpha P + (1-\alpha)Q$ denotes their convex combination with coefficient $\alpha \in (0, 1)$. Suppose there exists a finite constant $C > 0$ such that* $\operatorname{ess\,sup}\frac{dP}{dQ} \leq C$. *Then,*

$$d_2(P||Q_\alpha) \leq 1 + u(\alpha)D_{KL}(P||Q), \tag{10}$$

where

$$u(\alpha) = \begin{cases} \frac{2C(1-\alpha)^2}{(\alpha C + 1 - \alpha)^3} & \text{if } C \leq \frac{1-\alpha}{2\alpha} \\ \frac{8}{27\alpha} & \text{otherwise.} \end{cases}$$

Proof. Since the proof relies on the theory of f-divergences (Csiszár, 1967), let us first recall some basic definitions. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function such that $f(1) = 0$. Then, the f-divergence between P and Q is defined as:

$$D_f(P\|Q) = \int f\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}Q. \tag{22}$$

An example of f-divergence, which we will adopt in the remaining, is the chi-square divergence (Pearson, 1900), which is given by

$$D_{\chi^2}(P||Q) = \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2\mathrm{d}Q - 1, \tag{23}$$

or, equivalently, as the f-divergence with $f_{\chi^2}(w) = (w-1)^2$. Then,

$$d_2(P\|Q_\alpha) = \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2\mathrm{d}Q = D_{\chi^2}(P||Q_\alpha) + 1. \tag{24}$$

Let us now introduce an f-divergence $\Delta_\alpha(P||Q)$ defined by the function

$$f_{\Delta_\alpha}(w) = \frac{(w-1)^2}{\frac{\alpha}{1-\alpha}w + 1}. \tag{25}$$

This diverge can be seen as a skewed version of the triangular discrimination $\Delta(P||Q)$ (Le Cam, 1986), which is given by (25) for the particular case $\alpha = \frac{1}{2}$. Furthermore, it is easy to check that $D_{\chi^2}(P||Q_\alpha) = (1-\alpha)\Delta_\alpha(P||Q)$. Thus, in order to bound $d_2(P\|Q_\alpha)$ we only need to bound $\Delta_\alpha(P||Q)$.

Note that f_{Δ_α} is twice differential on $(0, \infty)$ and that, by assumption, $\operatorname{ess\,sup}\frac{dP}{dQ} \leq C$. Then, we can apply Theorem 3.1 of

Taneja (2004)[4] to obtain

$$\Delta_\alpha(P||Q) \leq D_{\mathrm{KL}}(P||Q) \sup_{w\in(0,C)} w f''_{\Delta_\alpha}(w). \tag{26}$$

Let us now compute the constant multiplying the KL divergence on the right-hand side. A simple algebra shows that the first derivative of f_{Δ_α} is

$$f'_{\Delta_\alpha}(w) = \frac{(w-1)\left(\frac{\alpha}{1-\alpha}w + \frac{\alpha}{1-\alpha} + 2\right)}{\left(\frac{\alpha}{1-\alpha}w + 1\right)^2},$$

while the second derivative is

$$f''_{\Delta_\alpha}(w) = \frac{2\left(\frac{\alpha}{1-\alpha} + 1\right)^2}{\left(\frac{\alpha}{1-\alpha}w + 1\right)^3}.$$

Let us define $g_{\Delta_\alpha}(w) := w f''_{\Delta_\alpha}(w)$. Then,

$$g'_{\Delta_\alpha}(w) = \frac{2\left(\frac{\alpha}{1-\alpha} + 1\right)^2\left(1 - 2\frac{\alpha}{1-\alpha}w\right)}{\left(\frac{\alpha}{1-\alpha}w + 1\right)^4}.$$

This function is positive for $w \leq \frac{1-\alpha}{2\alpha}$ and negative for $w \geq \frac{1-\alpha}{2\alpha}$. Thus,

$$\sup_{w\in(0,C)} g_{\Delta_\alpha}(w) = g_{\Delta_\alpha}(C) = \frac{2C\left(\frac{\alpha}{1-\alpha} + 1\right)^2}{\left(\frac{\alpha}{1-\alpha}C + 1\right)^3}$$

for $C \leq \frac{1-\alpha}{2\alpha}$, while

$$\sup_{w\in(0,C)} g_{\Delta_\alpha}(w) = g_{\Delta_\alpha}\left(\frac{1-\alpha}{2\alpha}\right) = \frac{8}{27}\frac{1}{\alpha(1-\alpha)}$$

in the opposite case. The theorem follows after multiplying these two constants by $1-\alpha$ and plugging everything into (24).

□

C.7. Proof of Proposition 4.1

Proposition 4.1. *The objective $\mathcal{L}(\widetilde{f})$ given in (9) can be upper bounded by*

$$\mathcal{L}(\widetilde{f}) \leq k_1 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta},\widetilde{f})}\left[\sum_{j\in\mathcal{J}_{src}} \alpha_j \|\widetilde{f}(\boldsymbol{x}_t) - f_j(\boldsymbol{x}_t)\|_2^2\right]$$
$$+ k_2 \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p_{\boldsymbol{\alpha}}}\left[\|\widetilde{f}(\boldsymbol{x}_t) - \bar{f}(\boldsymbol{x}_t)\|_2^2\right] + k_3,$$

where $k_1 = \frac{u(\alpha)dB^2}{2\sigma_{\mathcal{P}}^2 n(1-\alpha_0)}$, $k_2 = \frac{4\alpha_{tgt}dB^2}{\alpha_0^2\sigma_{\mathcal{P}}^2}$, and k_3 is a constant independent of $\widetilde{f}$.

Proof. From Theorem 4.2 we know that $d_2\left(p(\cdot|\boldsymbol{\theta},\widetilde{f})\|q_{\boldsymbol{\alpha}}(\cdot;\widetilde{f})\right) \leq 1 + u(\alpha)D_{\mathrm{KL}}\left(p(\cdot|\boldsymbol{\theta},\widetilde{f})\|\bar{q}_{\boldsymbol{\alpha}}(\cdot;\widetilde{f})\right)$, where we write $\bar{q}$ to denote the normalized mixture of proposals without the defensive component $p(\cdot|\boldsymbol{\theta},\widetilde{f})$. Neglecting the constant term

[4]Technically speaking, Taneja (2004) consider only discrete spaces. However, their result generalizes straightforwardly to general probability measures.

(which does not depend on $\widetilde{f}$), we have

$$
\begin{aligned}
D_{\mathrm{KL}}\left(p(\cdot|\boldsymbol{\theta},\widetilde{f})\|\bar{q}_{\boldsymbol{\alpha}}(\cdot;\widetilde{f})\right) &\leq \frac{1}{1-\alpha_0}\sum_{j=1}^{m}\alpha_j D_{\mathrm{KL}}\left(p(\cdot|\boldsymbol{\theta},\widetilde{f})\|p(\cdot|\boldsymbol{\theta}_j,\widetilde{f}_j)\right)\\
&= \frac{1}{1-\alpha_0}\sum_{j=1}^{m}\alpha_j \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta},\widetilde{f})}\left[\log\frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},\widetilde{f})}{p(\boldsymbol{\tau}|\boldsymbol{\theta}_j,\widetilde{f}_j)}\right]\\
&= \frac{1}{1-\alpha_0}\sum_{j=1}^{m}\alpha_j \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta},\widetilde{f})}\left[\sum_{t=0}^{T-1}\log\frac{\mathcal{P}_{\widetilde{f}}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)}{\mathcal{P}_{\widetilde{f}_j}(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,\boldsymbol{a}_t)}\right] + \frac{1}{1-\alpha_0}\sum_{j=1}^{m}\alpha_j \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta},\widetilde{f})}\left[\sum_{t=0}^{T-1}\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{a}_t|\boldsymbol{s}_t)}{\pi_{\boldsymbol{\theta}_j}(\boldsymbol{a}_t|\boldsymbol{s}_t)}\right]\\
&= \frac{1}{1-\alpha_0}\sum_{j\in\mathcal{J}_{\mathrm{src}}}\alpha_j\sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\pi_{\boldsymbol{\theta}},\widetilde{f})}\left[D_{\mathrm{KL}}(\mathcal{P}_{\widetilde{f}}(\cdot|\boldsymbol{s}_t,\boldsymbol{a}_t)\|\mathcal{P}_{f_j}(\cdot|\boldsymbol{s}_t,\boldsymbol{a}_t))\right] + \text{const}\\
&= \frac{1}{1-\alpha_0}\frac{1}{2\sigma_{\mathcal{P}}^2}\sum_{t=0}^{T-1}\mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\pi_{\boldsymbol{\theta}},\widetilde{f})}\left[\sum_{j\in\mathcal{J}_{\mathrm{src}}}\alpha_j\|\widetilde{f}(\boldsymbol{x}_t)-f_j(\boldsymbol{x}_t)\|_2^2\right] + \text{const},
\end{aligned}
$$

where the first inequality follows from the convexity of the function $1/x$ and Jensen's inequality. Note that the expected KL divergence between policies can be considered constant since, according to the approximation introduced in Section 4.2, the expectation is not computed under the current model $\widetilde{f}$. Furthermore, in the penultimate equality we dropped all the components from the target task since their KL divergence w.r.t. $\mathcal{P}_{\widetilde{f}}$ is zero.

The last term in (9) can be safely regarded as a constant since the integrand does not depend on $\widetilde{f}$. Noting that the bias term remained unchanged, the proposition is obtained after renaming the constants. □

C.8. Proof of Proposition 4.2

Proposition 4.2. *The function f^* minimizing (11) is*

$$f^*(\boldsymbol{x}) = \boldsymbol{A}^T\boldsymbol{k}(\boldsymbol{x}),$$

where $\boldsymbol{k}(\boldsymbol{x})$ is the RT-dimensional vector with entries $\mathcal{K}(\boldsymbol{x}_{r,t},\boldsymbol{x})$ and

$$\boldsymbol{A} = (k_1\alpha_{tgt}\boldsymbol{W}\boldsymbol{K} + k_2\boldsymbol{K} + \lambda R\boldsymbol{I})^{-1}(k_1\boldsymbol{W}\boldsymbol{F}_{src} + k_2\bar{\boldsymbol{F}}),$$

with $\boldsymbol{K}$ being the Gram matrix, $\boldsymbol{W} = diag(w_{r,t})$, $\bar{\boldsymbol{F}} = [\bar{f}(\boldsymbol{x}_{1,0}),\dots,\bar{f}(\boldsymbol{x}_{R,T-1})]^T$, $\boldsymbol{F}_{src} = \sum_{j\in\mathcal{J}_{src}}\alpha_j\boldsymbol{F}_j$, and $\boldsymbol{F}_j = [f_j(\boldsymbol{x}_{1,0}),\dots,f_j(\boldsymbol{x}_{R,T-1})]^T$.

Proof. For the objective (11) the representer theorem of RKHS holds. Then, for each dimension d of the state space, the solution has the form

$$f_d^*(\boldsymbol{x}) = \sum_{r=1}^{R}\sum_{t=0}^{T-1}a_{r,t}^{(d)}\mathcal{K}(\boldsymbol{x}_{r,t},\boldsymbol{x}) = \boldsymbol{a}_d^T\boldsymbol{k}(\boldsymbol{x}).$$

Let us define the matrix $\boldsymbol{A} = [\boldsymbol{a}_1,\dots,\boldsymbol{a}_d]$ of coefficients and rewrite the objective in matrix form:

$$\frac{1}{R}k_1\sum_{j\in\mathcal{J}_{\mathrm{src}}}\alpha_j\mathrm{Tr}\left((\boldsymbol{F}_j-\boldsymbol{K}\boldsymbol{A})^T\boldsymbol{W}(\boldsymbol{F}_j-\boldsymbol{K}\boldsymbol{A})\right) + \frac{1}{R}k_2\mathrm{Tr}\left((\bar{\boldsymbol{F}}-\boldsymbol{K}\boldsymbol{A})^T(\bar{\boldsymbol{F}}-\boldsymbol{K}\boldsymbol{A})\right) + \lambda\mathrm{Tr}\left(\boldsymbol{A}^T\boldsymbol{K}\boldsymbol{A}\right).$$

Taking the derivative with respect to $\boldsymbol{A}$, we obtain

$$-\frac{2}{R}k_1\sum_{j\in\mathcal{J}_{\mathrm{src}}}\alpha_j\boldsymbol{K}\boldsymbol{W}(\boldsymbol{F}_j-\boldsymbol{K}\boldsymbol{A}) - \frac{2}{R}k_2\boldsymbol{K}(\bar{\boldsymbol{F}}-\boldsymbol{K}\boldsymbol{A}) + 2\lambda\boldsymbol{K}\boldsymbol{A}.$$

The result follows by equating to zero and solving for $\boldsymbol{A}$.

□

C.9. Proof of Theorem B.1

Theorem B.1. *Let $\widetilde{f}_0, \dots, \widetilde{f}_m : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ be arbitrary functions and suppose that $\|g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\|_\infty \leq B$ almost surely. Then,*

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|^2\right] &\leq \frac{dB^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}_0, \widetilde{f}_0)\|q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f}_0, \dots, \widetilde{f}_m)\right) \\ &+ c_1 dB^2 \sum_{l=0}^{m} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[\|\bar{f}_l(\boldsymbol{s}_t, \boldsymbol{a}_t) - \widetilde{f}_l(\boldsymbol{s}_t, \boldsymbol{a}_t)\|_2^2\right] \\ &+ c_1 dB^2 \sum_{l=0}^{m} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[\mathrm{Tr}\left(\boldsymbol{\Sigma}_l(\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + \mathcal{O}(1), \end{aligned} \tag{13}$$

where the expectation is w.r.t. $\boldsymbol{\tau}_{i,j} \sim p(\boldsymbol{\tau}|\boldsymbol{\theta}_j, f_j)$ and $f_j \sim \varphi_j$. Here $\bar{f}_l(\boldsymbol{s}, \boldsymbol{a}) := \mathbb{E}_{f_l\sim\varphi_l}[f_l(\boldsymbol{s}, \boldsymbol{a})]$, $\boldsymbol{\Sigma}_l(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{C}\mathrm{ov}_{f_l\sim\varphi_l}[f_l(\boldsymbol{s}, \boldsymbol{a})]$, and c_1 is a constant.

Proof. We only sketch the main steps since the proofs is very similar to the one of Theorem 4.1. We start by writing $\mathbb{E}_{\boldsymbol{\tau}_{i,j}\sim p(\cdot|\boldsymbol{\theta}_j, f_j), f_j\sim\varphi_j}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|_2^2\right] = \mathbb{E}_{f_j\sim\varphi_j}\left[\mathbb{E}_{\boldsymbol{\tau}_{i,j}\sim p(\cdot|\boldsymbol{\theta}_j, f_j)}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|_2^2 | f\right]\right]$. Let us fix $f_0, \dots, f_m$ and focus on the inner expectation. Using a bias-variance decomposition,

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|_2^2\right] &= \sum_d \mathbb{E}\left[\left(\widehat{\nabla}_d J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla_d J\right)^2\right] \\ &= \sum_d \underbrace{\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_d J(\widetilde{f}_0, \dots, \widetilde{f}_m)\right]}_{(a)} + \sum_d \underbrace{\left(\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f}_0, \dots, \widetilde{f}_m)\right] - \nabla_d J\right)^2}_{(b)}.\end{aligned}$$

Term (a) can be easily bounded as in the proof of Theorem 4.1, obtaining

$$\mathbb{V}\mathrm{ar}\left[\widehat{\nabla}_d J(\widetilde{f}_0, \dots, \widetilde{f}_m)\right] \leq \frac{B^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}_0, \widetilde{f}_0)\|q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f}_0, \dots, \widetilde{f}_m)\right) + \underbrace{\frac{2B}{\alpha_0^2 n} \sum_{l\in\mathcal{J}} \alpha_l D_{\mathrm{TV}}\left(p(\cdot|\boldsymbol{\theta}_l, f_l)\|p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)\right)}_{(c)}.$$

Similarly, term (b) can be reduced to

$$\left(\mathbb{E}\left[\widehat{\nabla}_d J(\widetilde{f}_0, \dots, \widetilde{f}_m)\right] - \nabla_d J\right)^2 \leq \underbrace{\frac{8B^2}{\alpha_0^2} \sum_{l\in\mathcal{J}} \alpha_l D_{\mathrm{TV}}\left(p(\cdot|\boldsymbol{\theta}_l, f_l)\|p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)\right)^2}_{(d)}.$$

Then,

$$(c) + (d) \leq \underbrace{\frac{16B^2}{\alpha_0^2} \sum_{l\in\mathcal{J}} \alpha_l D_{\mathrm{TV}}\left(p(\cdot|\boldsymbol{\theta}_l, f_l)\|p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)\right)^2}_{(e)} + \underbrace{\frac{B^2}{4\alpha_0^2 n^2}}_{(f)}.$$

(f) is the $\mathcal{O}(1)$ term in the final bound, while (e) can be upper bounded using Pinsker's inequality as

$$(e) \leq \frac{8B^2}{\alpha_0^2} \sum_{l\in\mathcal{J}} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[D_{\mathrm{KL}}\left(\mathcal{P}_{\widetilde{f}_l}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\|\mathcal{P}_{f_l}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right].$$

Putting these terms together, we obtain

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\nabla} J(\widetilde{f}_0, \dots, \widetilde{f}_m) - \nabla J\|_2^2\right] &\leq \frac{dB^2}{n} d_2\left(p(\cdot|\boldsymbol{\theta}_0, \widetilde{f}_0)\|q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f}_0, \dots, \widetilde{f}_m)\right) \\ &+ \frac{8B^2}{\alpha_0^2} \sum_{l\in\mathcal{J}} \alpha_l \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\tau}\sim p(\cdot|\boldsymbol{\theta}_l, \widetilde{f}_l)}\left[D_{\mathrm{KL}}\left(\mathcal{P}_{\widetilde{f}_l}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\|\mathcal{P}_{f_l}(\cdot|\boldsymbol{s}_t, \boldsymbol{a}_t)\right)\right] + \frac{B^2}{4\alpha_0^2 n^2}.\end{aligned}$$

Then, the theorem follows after taking the expectation under $f_j \sim \varphi_j$ and decomposing the KL between Gaussian models as in Theorem 4.1. □

Parameter	LQR (transfer)	LQR (sample reuse)	Cartpole	Minigolf
Policy space	Linear	Linear	Linear	Polynomial (4-th order)
Optimizer	SGD	SGD	SGD	ADAM
Learning rate	1e-5	8e-6	1e-3	1e-2
Horizon	20	20	200	20
Adaptive	Yes	No	Yes	Yes
n_{min} or fixed batch size	5	10	3	5
ESS_{min}	20	–	20	20
Number of source tasks	5	–	5	5
Source samples per configuration	20	–	10	40
Maximum number of samples for GPs	–	–	250	1000
Maximum number of samples for $\mathcal{L}$	–	–	20	50

Table 1. Summary of the hyperparameters adopted in all experiments for the transfer algorithms.

D. Additional Details on the Experiments

In this section, we provide the details of the experiments presented in the main paper, together with some additional results. The hyperparameters used in all experiments are compactly summarized in Table 1.

D.1. Linear Quadratic Regulator

The system has linear dynamics, $s_{t+1} = As_t + Ba_t + \epsilon$, with Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2_{\mathcal{P}})$, and quadratic rewards $\mathcal{R}(s_t, a_t) = -Us_t^2 - Va_t^2$. Due to its simplicity and the fact that the optimal policy is available in closed-form, the LQR is a suitable benchmark for testing the properties of different gradient estimators.

Parameters We used Gaussian policies with a linearly parameterized mean and fixed variance σ^2_π, $\pi_\theta(a|s) = \mathcal{N}(a|\theta s, \sigma^2_\pi)$. We set the maximum horizon to $T = 20$. For each run, we randomly generated 5 source tasks by uniformly sampling A in $[0.6, 1.4]$ and B in $[0.8, 1.2]$, while the target task was fixed with $A = 1$ and $B = 1$. We considered 8 policies with parameters $\{-0.1, -0.2, \dots, -0.8\}$ and generated 20 episodes from each model-policy couple to built our initial source dataset. We used the same learning rate of 1e-5 and the same initialization $\theta_0 = -0.1$ for all algorithms. We set the batch size of GPOMDP to 10 and used the adaptive version of Algorithm 1 with $n_{\text{min}} = 5$ and $\text{ESS}_{\text{min}} = 20$. We learned the target task using standard SGD.

For the sample reuse experiment, all algorithms used a learning rate of 8e-6, a fixed batch size of 10, and the same initialization as before.

Additional Results We provide additional insights into the performance of each estimator. Figure 3*(left)* shows the expected return achieved by all alternatives as a function of the number of episodes. The results are coherent with those presented in the main paper, although the differences between the algorithms' performances are harder to appreciate. Figure 3(center) shows how the ESS changes at each iteration. As expected, the ESS of PD-IS remains almost constant, which is due to the fact that general IS estimators highly depend on the chosen proposal distributions. On the other hand, the MIS estimators do not suffer this problem and their ESS linearly increases with the number of iterations. Finally, Figure 3 shows the number of samples collected by each algorithm at each iteration. Coherently with the plot of the ESS, PD-IS needs to collect a high number of samples to meet the ESS_{min} requirement. On the other hand, all transfer algorithms manage to learn while sampling the minimum number of trajectories allowed.

In order to better demonstrate the benefits of transfer using our estimators, we repeat the LQR experiments using three fixed sets of source tasks of increasing distance from the target:

- Close sources: $(A, B) \in \{(0.92, 0.96), (0.95, 0.93), (0.98, 0.99), (1.02, 1.04), (1.05, 1.08)\}$;
- Distant sources: $(A, B) \in \{(0.78, 0.85), (0.85, 0.88), (0.9, 0.9), (1.1, 1.15), (1.12, 1.2)\}$;
- Very distant sources: $(A, B) \in \{(0.52, 0.6), (0.5, 0.63), (0.55, 0.55), (1.45, 1.4), (1.48, 1.46)\}$.

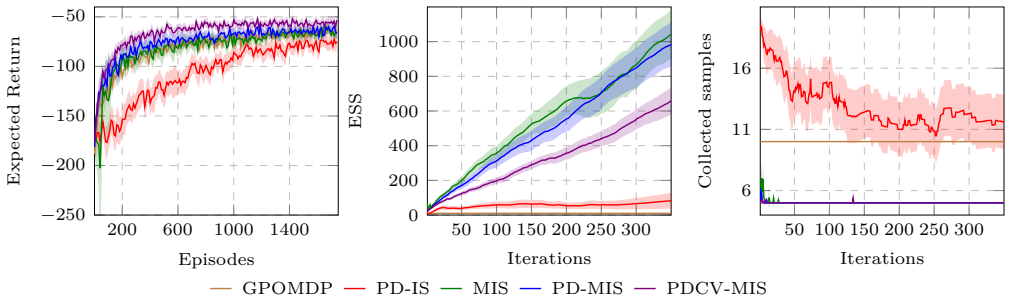

Figure 3. Additional results for the LQR experiment of Section 6.1. Expected return (left), effective sample size (center), and the number of samples collected at each iteration (right).

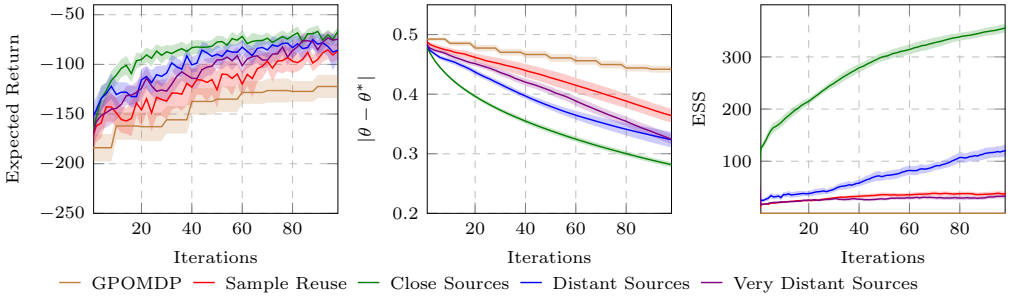

Figure 4. LQR experiment with fixed source tasks of increasing distance from the target. Expected return (left), policy parameter (center), and effective sample size (right).

The target task is again fixed with $A = B = 1$. For the sake of conciseness, we report only the results of the PDCV estimator. Figure 4 shows the results. We may notice that, as expected, the learning performance improves as the source tasks get closer to the target, i.e.,, when more information can be transferred. Interestingly, the performance using very distant source tasks almost reduces to the one achieved by sample reuse. That is, the algorithm is not able to transfer any information from the sources but still shows robustness to negative transfer.

D.2. Cartpole

Parameters Similarly to the previous experiments, we used Gaussian policies with linearly parameterized mean and fixed variance. We considered different tasks by varying the mass m of the cart and the length l of the pole. For each run, we generated 5 source tasks by uniformly sampling m in the interval $[0.8, 1.2]$ and l in $[0.3, 0.7]$. The target task used the standard Cartpole parameters, with $m = 1.0$ and $l = 0.5$. For each source task, we considered a sequence of 10 policies generated by GPOMDP during its learning process and collected 10 episodes from each. We set the maximum horizon to $T = 200$.

For the transfer algorithm with discrete model estimation, we considered the fixed set of possible tasks $(m, l) \in \{(1.0, 0.5), (0.8, 0.3), (1.2, 0.7), (1.1, 0.6), (0.9, 0.4), (0.9, 0.6), (1.1, 0.4), (1.5, 1.0)\}$ and used $R = 40$ simulated trajectories to approximate the bound of Theorem 4.1 for each of them.

For the transfer algorithm with continuous model estimation, we use the squared exponential kernel,

$$\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \rho^2 \exp\left(\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T \boldsymbol{\Lambda}^{-1}(\boldsymbol{x} - \boldsymbol{x}')\right). \tag{27}$$

The hyperparameters were not tuned and set to the default values of $\rho = 1$ and $\boldsymbol{\Lambda} = \mathrm{diag}(1, \dots, 1)$. The objective 11 were approximated by simulating 20 trajectories from the previously estimated model. Furthermore, since the constants in our bound highly favor the bias term when a high number of source samples is available, we rebalanced these coefficients by starting with equal values and decreasing the one multiplying the variance linearly with n.

For the gray curve in Figure 2, we logged the GP models learned by the RKHS estimator at fixed iterations and computed

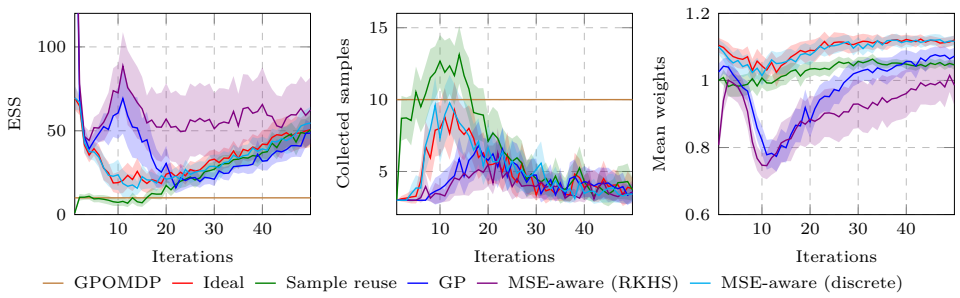

Figure 5. Some statistics on the importance weights computed by each algorithm in the Cartpole experiment of Section 6.2.

the optimal policies under these models using GPOMDP.

Additional Results Similarly to the LQR experiment, we investigate the quantities related to the importance weights computed by the different algorithms (Figure 5). It is particularly interesting to notice that the continuous model estimation approaches achieve similar performance to the other transfer algorithms while collecting less samples. It is also worth noticing that, due to estimation errors, the mean of their weights is much lower than the one of the ideal weights. However, this fact seems not to have only a negligible impact on the learning process.

D.3. Minigolf

In the minigolf game, the agent has to shoot a ball with radius r inside a hole of diameter D with the minimum number of strokes. We assume that the ball moves along a level surface with a constant deceleration $d = \frac{5}{7}\rho g$, where ρ is the dynamic friction coefficient between the ball and the ground and g is the gravitational acceleration. Given the distance x_0 of the ball from the hole, the agent must determine the angular velocity ω of the putter that determines the initial velocity $v_0 = \omega l$ (where l is the length of the putter) to put the ball in the hole in one strike. For each distance x_0, the ball falls in the hole if its initial velocity v_0 ranges from $v_{min} = \sqrt{2dx_0}$ to $v_{max} = \sqrt{(2D - r)^2 \frac{g}{2r} + v_{min}^2}$. v_{max} is the maximum allowed speed o the edge of the hole to let the ball enter the hole and not to overcome it. At the beginning of each trial the ball is placed at random, between 2000cm and 0cm far from the hole. At each step, the agent chooses an action that determines the initial velocity v_0 of the ball. When the ball enters the hole the episode ends with reward 0. If $v_0 > v_{max}$ the ball is lost and the episode ends with reward 100. Finally, if $v_0 < v_{min}$ the episode goes on and the agent can try another hit with reward 1 from position $x = x_0 - \frac{(v_0)^2}{2d}$. The angular speed of the putter is determined by the action a selected by the agent as follows: $\omega = al(1 + \epsilon)$, where $\epsilon \sim \mathcal{N}(0, 0.3)$. This implies that the stronger the action chosen the more uncertain its outcome will be. As a result, the agent is disencumbered by trying to make a hole in one shot when it is away from the hole and will prefer to perform a sequence of approach shots.

Parameters In this experiment, we adopted Gaussian policies with a linearly parameterized mean in a fourth-order polynomial basis, $\pi_{\boldsymbol{\theta}}(a|s) = \mathcal{N}(a|\boldsymbol{\theta}^T\boldsymbol{\phi}(s), \sigma_\pi^2)$, where $\boldsymbol{\phi}(s) = [1, s, s^2, s^3, s^4]$. Our source tasks were generated by varying dynamic friction coefficient, hole size, and putter length from the realistic ranges defined above. Each run uniformly sampled a set of 5 source tasks in these intervals. Furthermore, we considered 10 (fixed) source policies of increasing quality, from those achieving very considered behavior to those overshooting the hole. We generated 40 episodes from each model-policy pair. The target task was fixed with a friction of 0.131, a putter of $100cm$, and a hole of diameter $10cm$. The maximum horizon was set to 20 time steps, which are sufficient for safely reaching the hole when starting from any position.

We used a fixed batch size of 10 episodes for GPOMDP, while the transfer algorithms were adaptive with $n_{\text{min}} = 5$ and $\text{ESS}_{\text{min}} = 20$.

For the discrete model estimator, we consider the source tasks of each run as the set of possible environments and generate 40 trajectories to approximate the bound on the MSE.

For the continuous model estimator, we used the squared exponential kernel (27) and we tuned the hyperparameters using the source samples. The objective (11) was approximated by collecting 50 episodes under the previously learned model, and the maximum number of samples to train the GPs was limited to 1000.

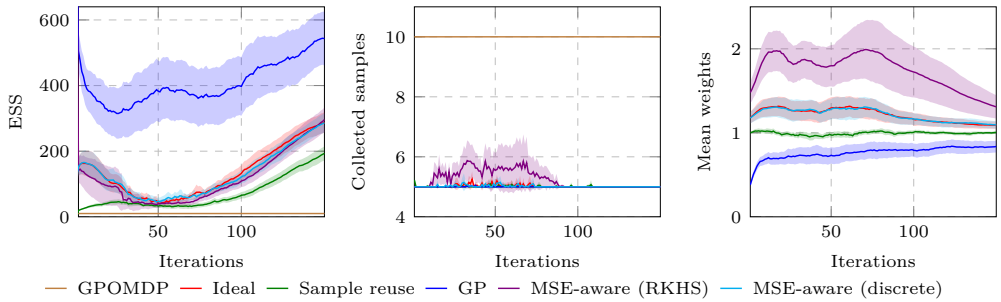

Figure 6. Some statistics on the importance weights computed by each algorithm in the Minigolf experiment of Section 6.3.

Additional Results Figure 6 shows the usual statistics on the importance weights. Here it is worth noticing that the very high ESS achieved when using GPs predictions to directly estimated the importance weights is actually due to a drawback of the ESS estimators. In fact, the errors due to the very imprecise GP models make all weights small (the mean is significantly below one) and with low variance, a situation in which the ESS is typically overestimated. This leads the algorithm to collect the minimum allowed amount of samples at each iteration, while it should actually collect many more. The MSE-aware estimator, on the other hand, keeps an ESS which is very close to that of the ideal and discrete estimators. The higher mean of the importance weights also implies that more information is transferred, which leads to the good empirical performance showed in Section 6.3.