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 $\mathrm{ESS}_{\mathrm{min}}$ 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:

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

□

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 $\boldsymbol{\beta}_d^*$,* $\mathbb{V}\mathrm{ar}[\widehat{\nabla}_{\boldsymbol{\theta}_d}^{\mathrm{CV}} J(\boldsymbol{\theta}, \mathcal{P})] \leq \mathbb{V}\mathrm{ar}[\widehat{\nabla}_{\boldsymbol{\theta}_d}^{\mathrm{MIS}} J(\boldsymbol{\theta}, \mathcal{P})]$.

Proof. To prove the unbiasedness, recall that

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

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),

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

□

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 $0 < \eta \le \frac{2}{L}$, and that $\mathscr{A}_{CV}$ ($\mathscr{A}_{PDCV}$) uses the optimal CV coefficients $\boldsymbol{\beta}_d^*$. Then, for all $\epsilon > 0$:*

$$\mathscr{A}_{CV} \nprec \mathscr{B}_R, \quad \mathscr{A}_{PDCV} \nprec \mathscr{B}_G.$$

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 $\mathscr{A}$,

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

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

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

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,

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

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

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

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,

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

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

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

The second equality follows from the fact that $w_0^{\text{IS}}(\boldsymbol{\tau}) := \frac{p(\boldsymbol{\tau}|\boldsymbol{\theta},f)}{p(\boldsymbol{\tau}|\boldsymbol{\theta},f)} = 1$. 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

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

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,

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

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

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

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 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

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

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

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

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,

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

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,

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

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, $kx \le x^2 + \frac{k^2}{2}$, 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

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

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

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

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

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

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 $w \le \frac{1-\alpha}{2\alpha}$ and negative for $w \ge \frac{1-\alpha}{2\alpha}$. 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 $C \le \frac{1-\alpha}{2\alpha}$, 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*

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

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 $\widetilde{f}$.

Proof. From Theorem 4.2 we know that $d_2\left(p(\cdot|\boldsymbol{\theta}, \widetilde{f}) \| q_{\boldsymbol{\alpha}}(\cdot; \widetilde{f})\right) \le 1 + u(\alpha) D_{\mathrm{KL}}\left(p(\cdot|\boldsymbol{\theta}, \widetilde{f}) \| \bar{q}_{\boldsymbol{\alpha}}(\cdot; \widetilde{f})\right)$, 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

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

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, $\boldsymbol{W} = \mathit{diag}(w_{r,t})$, $\bar{\boldsymbol{F}} = [\bar{f}(\boldsymbol{x}_{1,0}),\dots,\bar{f}(\boldsymbol{x}_{R,T-1})]^T$, $\boldsymbol{F}_{src} = \sum_{j\in\mathcal{J}_{src}}\alpha_j\boldsymbol{F}_j$, 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:

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

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

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

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 $\|g(\boldsymbol{\tau})\mathcal{R}(\boldsymbol{\tau})\|_\infty \le B$ almost surely. Then,*

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

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

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

Similarly, term (b) can be reduced to

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

Then,

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

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

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

Putting these terms together, we obtain

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

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$_{\text{min}}$	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 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$_{\text{min}}$ 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: $(A, B) \in \{(0.52, 0.6), (0.5, 0.63), (0.55, 0.55), (1.45, 1.4), (1.48, 1.46)\}$.

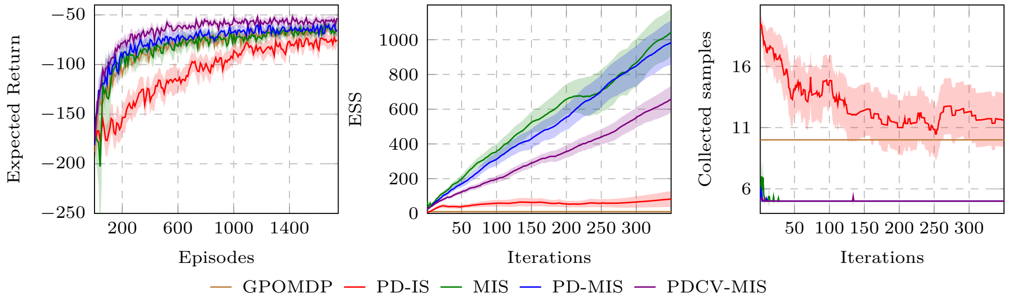

Figure 3. Additional results for the LQR experiment of Section 6.1. Expected return (left), effective sample size (center), and the number of samples collected at each iteration (right).

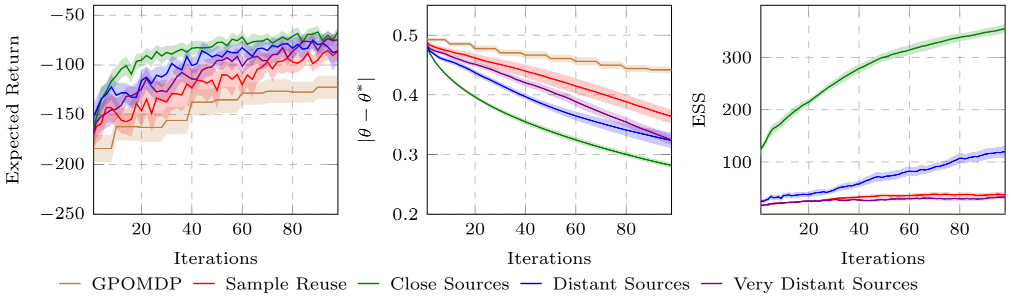

Figure 4. LQR experiment with fixed source tasks of increasing distance from the target. Expected return (left), policy parameter (center), and effective sample size (right).

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

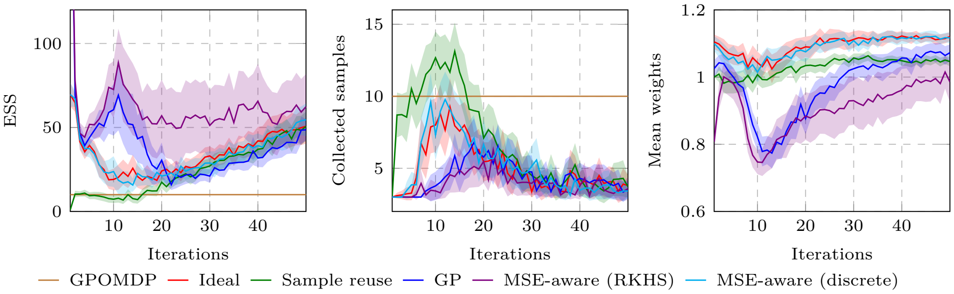

Figure 5. Some statistics on the importance weights computed by each algorithm in the Cartpole experiment of Section 6.2.

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.

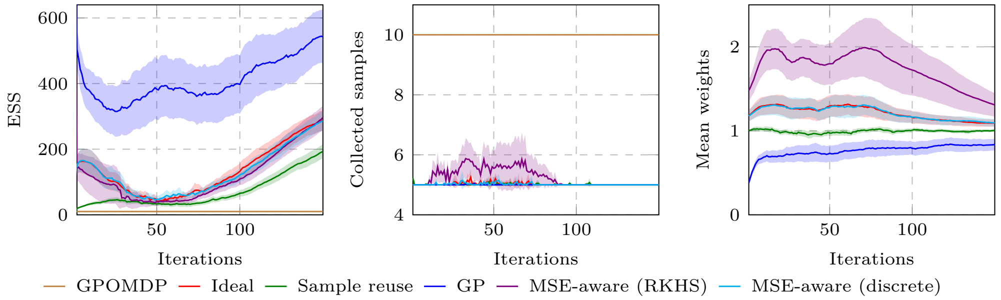

Figure 6. Some statistics on the importance weights computed by each algorithm in the Minigolf experiment of Section 6.3.

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.