components in I. Let $\theta^*(\omega_0) \in \Theta$ the initial parameter. Let $\alpha \geqslant 0$, $\theta^*_I(\omega) = \theta_0 + \alpha \nabla_\theta J_{\mathcal{M}_\omega}(\theta^*(\omega_0))|_I$ and $\theta^*(\omega) = \theta_0 + \alpha \nabla_\theta J_{\mathcal{M}_\omega}(\theta^*(\omega_0))$. Then, under Assumption, we have:

$$\ell(\theta^*_I(\omega)) - \ell(\theta^*(\omega)) \geqslant \frac{\lambda_{\min}\alpha^2}{2} \left\| \nabla_\theta J_{\mathcal{M}_\omega}(\theta^*(\omega_0))|_{\overline{I}} \right\|_2^2 .$$

Thus, we maximize the L^2–norm of the gradient components that correspond to the parameters we want to test. Since we have at our disposal only samples $\mathcal{D}$ collected with the current policy $\pi_{\theta^*(\omega_0)}$ and in the current environment ω_0, we have to perform an off–distribution optimization over ω. To this end, we employ an approach analogous to that of where we optimize the empirical version of the objective with a penalization that accounts for the distance between the distribution over trajectories:

$$\mathcal{C}_I(\omega/\omega_0) = \left\| \widehat{\nabla}_\theta J_{\mathcal{M}_{\omega/\omega_0}}(\theta^*(\omega_0))|_{\overline{I}} \right\|_2^2 - \zeta \sqrt{\frac{\widehat{d}_2(\omega \| \omega_0)}{n}}, \quad (1)$$

where $\widehat{\nabla}_\theta J_{\mathcal{M}_{\omega/\omega_0}}(\theta^*(\omega_0))$ is an off-distribution estimator of the gradient $\nabla_\theta J_{\mathcal{M}_\omega}(\theta^*(\omega_0))$ using samples collected with ω_0, $\widehat{d}_2$ is the estimated 2-Renyi divergence that works as a penalization to discourage too large updates and $\zeta \geqslant 0$ is a regularization parameter. The expression of the estimated gradient, 2-Renyi divergence and the pseudocode are reported in Appendix.

Experimental Evaluation

In this section, we present the experimental evaluation of the identification rules in three RL domains. To set the values of $c(l)$ we resort to the Wilk's asymptotic approximation (Theorem) to enforce (asymptotic) guarantees on the type I error. For Identification Rule we perform 2^d statistical tests by using the same dataset $\mathcal{D}$. Thus, we partition δ using Bonferroni correction and setting $c(l) = \chi^2_{l,1-\delta/2^d}$, where $\chi^2_{l,\xi}$ is the ξ–quintile of a chi square distribution with l degrees of freedom. Instead, for Identification Rule, we perform d statistical test, and thus, we set $c(1) = \chi^2_{1,1-\delta/d}$.

Discrete Grid World

The grid world environment is a simple representation of a two-dimensional world (5×5 cells) in which an agent has to reach a target position by moving in the four directions. The goal of this set of experiments is to show the advantages of configuring the environment when performing the policy space identification using rule. The initial position of the agent and the target position are drawn at the beginning of each episode from a Boltzmann distribution μ_ω. The agent plays a Boltzmann linear policy π_θ with binary features ϕ indicating its current row and column and the row and column of the goal. For each run, the agent can control a subset I^* of the parameters θ_{I*} associated with those features, which is randomly selected. Furthermore, the supervisor can configure the environment by changing the parameters ω of the initial state distribution μ_ω. Thus, the supervisor can induce the agent to explore certain regions of the grid world and, consequently, change the relevance of the corresponding parameters in the optimal policy.

Figure shows the empirical $\widehat{\alpha}$ and $\widehat{\beta}$, i.e. the fraction of parameters that the agent does not control that are wrongly selected and the fraction of those the agent controls that are not selected respectively, as a function of the number n of episodes used to perform the identification. We compare two cases: *conf* where the identification is carried out by also configuring the environment, i.e. optimizing Equation (1), and *no-conf* in which the identification is performed in the original environment only. In both cases, we can see that $\widehat{\alpha}$ is almost independent of the number of samples, as it is directly controlled by the critical value $c(1)$. Differently, $\widehat{\beta}$ decreases as the number of samples increases, i.e. the power of the test $1 - \widehat{\beta}$ increases with n. Remarkably, we observe that configuring the environment gives a significant advantage in understanding the parameters controlled by the agent w.r.t using a fixed environment, as $\widehat{\beta}$ decreases faster in the *conf* case. This phenomenon also justifies empirically our choice of objective (Equation (1)) for selecting the new environment. Hyperparameters, further experimental results, together with experiments on a continuous version of the grid world, are reported in Appendix–.

Minigolf

In the Minigolf environment, an agent hits a ball using a putter with the goal of reaching the hole in the minimum number of attempts. Surpassing the hole causes the termination of the episode and a large penalization. The agent selects the force applied to the putter by playing a Gaussian policy linear in some polynomial features (complying to Lemma) of the distance from the hole (x) and the friction of the green (f). We consider two agents: $\mathscr{A}_1$ has access to both the x and f whereas $\mathscr{A}_2$ knows only x. Thus, we expect that $\mathscr{A}_1$ learns a policy that allows reaching the hole in a smaller number of hits, compared to $\mathscr{A}_2$, as it can calibrate force according to friction; whereas $\mathscr{A}_2$ has to be more conservative, being unaware of f. There is also a supervisor in charge of selecting, for the two agents, the best putter length ω, i.e. the configurable parameter of the environment. In this experiment, we want to highlight that knowing the policy space might be of crucial importance when learning in a Conf–MDP.

Figure-left shows the performance of the optimal policy as a function of the putter length ω. We can see that for agent $\mathscr{A}_1$ the optimal putter length is $\omega^*_{\mathscr{A}_1} = 5$ while for agent $\mathscr{A}_2$ is $\omega^*_{\mathscr{A}_2} = 11.5$. Figure-right compares the performance of the optimal policy of agent $\mathscr{A}_2$ when the putter length ω is chosen by the supervisor using four different strategies. In (i) the configuration is sampled uniformly in the interval $[1, 15]$. In (ii) the supervisor employs the optimal configuration for agent $\mathscr{A}_1$ ($\omega = 5$), i.e. assuming the agent is aware of the friction. (iii) is obtained by selecting the optimal configuration of the policy space produced by using our identification rule. Finally, (iv) is derived by employing an oracle that knows the true agent's policy space ($\omega = 11.5$). We can see that the performance of the identification procedure (iii) is comparable with that of the oracle (iv) and notably higher than the performance when employing an incorrect policy space (ii). Hyperparameters and additional experiments are