Machine Learning
https://doi.org/10.1007/s10994-021-06033-3

Policy space identification in configurable environments

Alberto Maria Metelli[1] · Guglielmo Manneschi[1] · Marcello Restelli[1]

Received: 13 January 2020 / Revised: 16 February 2021 / Accepted: 21 June 2021

Abstract

We study the problem of identifying the policy space available to an agent in a learning process, having access to a set of demonstrations generated by the agent playing the optimal policy in the considered space. We introduce an approach based on frequentist statistical testing to identify the set of policy parameters that the agent can control, within a larger parametric policy space. After presenting two identification rules (combinatorial and simplified), applicable under different assumptions on the policy space, we provide a probabilistic analysis of the simplified one in the case of linear policies belonging to the exponential family. To improve the performance of our identification rules, we make use of the recently introduced framework of the Configurable Markov Decision Processes, exploiting the opportunity of configuring the environment to induce the agent to reveal which parameters it can control. Finally, we provide an empirical evaluation, on both discrete and continuous domains, to prove the effectiveness of our identification rules.

Keywords Reinforcement learning · Configurable Markov decision processes · Likelihood ratio test · Policy space identification

1 Introduction

Reinforcement Learning (RL, Sutton and Barto, 2018) deals with sequential decision–making problems in which an artificial agent interacts with an environment by sensing *perceptions* and performing *actions*. The agent's goal is to find an optimal policy, i.e., a prescription of actions that maximizes the (possibly discounted) cumulative reward collected during its interaction with the environment. The performance of an agent in an environment is constrained by its perception and its actuation possibilities, along with the

Editors: Ira Assent, Carlotta Domeniconi,Aristides Gionis, Eyke Hüllermeier.

✉ Alberto Maria Metelli
albertomaria.metelli@polimi.it

Guglielmo Manneschi
guglielmo.manneschi@mail.polimi.it

Marcello Restelli
marcello.restelli@polimi.it

[1] DEIB, Politecnico di Milano, Piazza Leonardo Da Vinci, 32, 20133 Milan, Italy

Published online: 05 September 2021

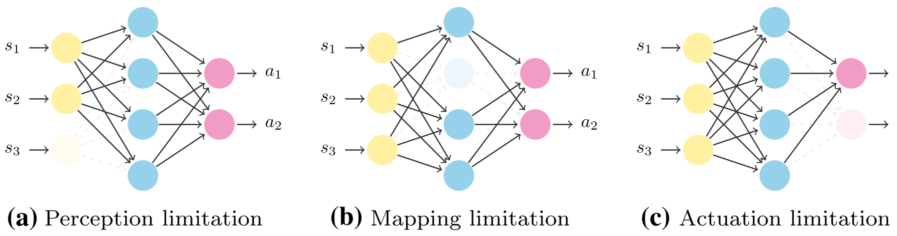
ability to *map* observations to actions. These three elements define the *policy space* available to the agent in the learning process. Agents having access to different policy spaces may exhibit different optimal behaviors, even in the same environment. Therefore, the notion of optimality is necessarily connected to the space of policies the agent can access, which we will call the *agent's policy space* in the following. While in tabular RL we typically assume access to the complete space of Markovian stationary policies, in continuous control, the policy space needs to be limited. In policy search methods (Deisenroth et al., 2013), the policies are explicitly modeled considering a parametric functional space (Sutton et al., 1999; Peters and Schaal, 2008) or a kernel space (Deisenroth and Rasmussen, 2011; Levine and Koltun, 2013); but even in value–based RL, a function approximator induces a set of representable (greedy) policies. It is important to point out that the notion of policy space is not just an algorithmic convenience. Indeed, the need to limit the policy space naturally emerges in many industrial applications, where some behaviors have to be avoided for safety reasons.

The knowledge of the agent's policy space might be useful in some subfields of RL. Recently, the framework of Configurable Markov Decision Process (Conf-MDP, Metelli et al., 2018a) has been introduced to account for the scenarios in which it is possible to configure some environmental parameters. Intuitively, the best environment configuration is intimately related to the agent's possibilities in terms of policy space. When the configuration activity is performed by an external supervisor, it might be helpful to know which parameters the agent can control in order to select the most appropriate configuration. Furthermore, in the field of Imitation Learning (IL, Osa et al., 2018), figuring out the policy space of the expert's agent can aid the learning process of the imitating policy, mitigating overfitting/underfitting phenomena.

In this paper, motivated by the examples presented above, we study the problem of *identifying* the agent's policy space in a Conf–MDP,[1] by observing the agent's behavior and, possibly, exploiting the *configuration* opportunities of the environment. We consider the case where the agent's policy space is a subset of a known super–policy space Π_{Θ} induced by a parameter space $\Theta \subseteq \mathbb{R}^d$. Thus, any policy $\pi_{\boldsymbol{\theta}}$ is determined by a d–dimensional parameter vector $\boldsymbol{\theta} \in \Theta$. However, the agent has control over a smaller number $d^{\mathrm{Ag}} < d$ of parameters (which are unknown), while the remaining ones have a fixed value, namely zero.[2] The choice of zero as a fixed value might appear arbitrary, but it is rather a common case in practice. Indeed, the formulation based on the identification of the *parameters* effectively covers the limitations of the policy space related to perception, actuation, and mapping. For instance, in a linear policy, the fact that the agent does not observe a state feature is equivalent to set the corresponding parameters to zero. Similarly, in a neural network, removing a neuron is equivalent to neglecting all of its connections, which in turn can be realized by setting the relative weights to zero. Figure 1 shows three examples of policy space limitations in the case of a 1–hidden layer neural network policy, which can be realized by setting the appropriate weights to zero.

[1] Although we assume to act in a Conf–MDP, we stress that our primary goal is to identify the policy space of the agent, rather than learning a profitable configuration in the Conf–MDP.

[2] By "controllable" parameter we mean a parameter whose value can be changed by the agent, while the "uncontrollable" parameters are those which are permanently set to zero. This is a way of modeling the limitations of the policy space.



(a) Perception limitation **(b)** Mapping limitation **(c)** Actuation limitation

Fig. 1 An example of policy space modeled as a 1-layer neural network showing a limitation in the **a** perception, **b** mapping, and **c** actuation

Our goal is to identify the parameters that the agent can control (and possibly change) by observing some demonstrations of the optimal policy π^{Ag} in the policy space Π_{Θ}.[3] To this end, we formulate the problem as deciding whether each parameter θ_i for $i \in \{1, ..., d\}$ is zero, and we address it by means of a frequentist statistical test. In other words, we check whether there is a statistically significant difference between the likelihood of the agent's behavior with the full set of parameters and the one in which θ_i is set to zero. In such a case, we conclude that θ_i is not zero and, consequently, the agent can control it. On the contrary, either the agent cannot control the parameter, or zero is the value consciously chosen by the agent.

Indeed, there could be parameters that, given the peculiarities of the environment, are useless for achieving the optimal behavior or whose optimal value is actually zero, while they could prove essential in a different environment. For instance, in a grid world where the goal is to reach the right edge, the vertical position of the agent is useless, while if the goal is to reach the upper right corner, both horizontal and vertical positions become relevant. In this spirit, configuring the environment can help the supervisor in identifying whether a parameter set to zero is actually uncontrollable by the agent or just useless in the current environment. Thus, the supervisor can change the environment configuration $\boldsymbol{\omega} \in \Omega$, so that the agent will adjust its policy, possibly by changing the parameter value and revealing whether it can control such a parameter. Consequently, the new configuration should induce an optimal policy in which the considered parameters have a value significantly different from zero. We formalize this notion as the problem of finding the new environment configuration that maximizes the *power* of the statistical test and we propose a surrogate objective for this purpose.

The paper is organized as follows. In Sect. 2, we introduce the necessary background. The *identification rules* (combinatorial and simplified) to perform parameter identification in a fixed environment are presented in Sect. 3 and the simplified one is analyzed in Sect. 4. Sect. 5 shows how to improve them by exploiting the environment configurability. The experimental evaluation, on discrete and continuous domains, is provided in Sect. 6. Besides studying the ability of our identification rules in identifying the agent's policy space, we apply them to the IL and Conf-MDP frameworks. The proofs not reported in the main paper can be found in Appendix A.

[3] We stress that, since we restrict the search to the policy space Π_{Θ}, π^{Ag} might be suboptimal compared to the optimal policy in the space of Markovian stationary policies.

2 Preliminaries

In this section, we report the essential background that will be used in the subsequent sections. For a given set $\mathcal{X}$, we denote with $\mathscr{P}(\mathcal{X})$ the set of probability distributions over $\mathcal{X}$.

(Configurable) Markov Decision Processes A discrete–time Markov Decision Process (MDP, Puterman, 2014) is defined by the tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p, \mu, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the state space and the action space respectively, $p : \mathcal{S} \times \mathcal{A} \rightarrow \mathscr{P}(\mathcal{S})$ is the transition model that provides, for every state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, a probability distribution over the next state $p(\cdot|s, a)$, $\mu \in \mathscr{P}(\mathcal{S})$ is the distribution of the initial state, $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is the reward model, defining the reward collected by the agent $r(s,\ a)$ when performing action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$, and $\gamma \in [0, 1]$ is the discount factor. The behavior of an agent is defined by means of a policy $\pi : \mathcal{S} \rightarrow \mathscr{P}(\mathcal{S})$ that provides a probability distribution over the actions $\pi(\cdot|s)$ for every state $s \in \mathcal{S}$. We limit the scope to parametric policy spaces $\Pi_\Theta = \{\pi_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is the parameter space. The goal of the agent is to find an optimal policy within Π_Θ, i.e., any policy parametrization that maximizes the *expected return*:

$$\boldsymbol{\theta}^{\mathrm{Ag}} \in \operatorname*{arg\,max}_{\boldsymbol{\theta} \in \Theta} J_{\mathcal{M}}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{\substack{s_0 \sim \mu \\ a_t \sim \pi_{\boldsymbol{\theta}}(\cdot|s_t) \\ s_{t+1} \sim p(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{+\infty} \gamma^t r(s_t, a_t) \right]. \tag{1}$$

In this paper, we consider a slightly modified version of the Conf–MDPs (Metelli et al., 2018a).

Definition 1 A Configurable Markov Decision Process (Conf–MDP) induced by the configuration space $\Omega \subseteq \mathbb{R}^p$ is defined as the set of MDPs:

$$\mathcal{C}_\Omega = \{\mathcal{M}_{\boldsymbol{\omega}} = (\mathcal{S}, \mathcal{A}, p_{\boldsymbol{\omega}}, \mu_{\boldsymbol{\omega}}, r, \gamma) \ : \ \boldsymbol{\omega} \in \Omega\}.$$

The main differences w.r.t. the original definition are: i) we allow the configuration of the initial state distribution $\mu_{\boldsymbol{\omega}}$, in addition to the transition model $p_{\boldsymbol{\omega}}$; ii) we restrict to the case of parametric configuration spaces Ω; iii) we do not consider the policy space Π_Θ as a part of the Conf–MDP.

Generalized Likelihood Ratio Test The Generalized Likelihood Ratio test (GLR, Barnard, 1959; Casella and Berger, 2002) aims at testing the goodness of fit of two statistical models. Given a parametric model having density function $p(\cdot|\boldsymbol{\theta})$ with $\boldsymbol{\theta} \in \Theta$, we aim at testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta_0$, where $\Theta_0 \subset \Theta$ is a subset of the parametric space, against the alternative $\mathcal{H}_1 : \boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta \setminus \Theta_0$. Given a dataset $\mathcal{D} = \{X_i\}_{i=1}^n$ sampled independently from $p(\cdot|\boldsymbol{\theta}^{\mathrm{Ag}})$, where $\boldsymbol{\theta}^{\mathrm{Ag}}$ is the true parameter, the GLR statistic is:

$$\Lambda = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} p(\mathcal{D}|\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} p(\mathcal{D}|\boldsymbol{\theta})} = \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})}, \tag{2}$$

where $p(\mathcal{D}|\boldsymbol{\theta}) = \widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^n p(X_i|\boldsymbol{\theta})$ is the likelihood function. We denote with $\widehat{\ell}(\boldsymbol{\theta}) =$

$-\log \widehat{\mathcal{L}}(\boldsymbol{\theta})$ the negative log–likelihood function, $\widehat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}\in\Theta}{\arg\sup}\ \widehat{\mathcal{L}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\theta}}_0 \in \underset{\boldsymbol{\theta}\in\Theta_0}{\arg\sup}\ \widehat{\mathcal{L}}(\boldsymbol{\theta})$, i.e., the maximum likelihood solutions in Θ and Θ_0 respectively. Moreover, we define the expectation of the likelihood under the true parameter: $\ell(\boldsymbol{\theta}) = \underset{X_i \sim p(\cdot|\boldsymbol{\theta}Ag)}{\mathbb{E}}[\widehat{\ell}(\boldsymbol{\theta})]$. As the maximization is carried out employing the same dataset $\mathcal{D}$ and recalling that $\Theta_0 \subset \Theta$, we have that $\Lambda \in [0,1]$. It is usually convenient to consider the logarithm of the GLR statistic: $\lambda = -2\log\Lambda = 2(\widehat{\ell}(\widehat{\boldsymbol{\theta}}_0) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}))$. Therefore, $\mathcal{H}_0$ is rejected for large values of λ, i.e., when the maximum likelihood parameter searched in the restricted set Θ_0 significantly underfits the data $\mathcal{D}$, compared to Θ. Wilk's theorem provides the asymptomatic distribution of λ when $\mathcal{H}_0$ is true (Wilks, 1938; Casella and Berger, 2002).

Theorem 1 (*Casella and Berger*, (2002), *Theorem* 10.3.3) *Let* $d = \dim(\Theta)$ *and* $d_0 = \dim(\Theta_0) < d$. *Under suitable regularity conditions* (*see* Casella and Berger, (2002) *Section* 10.6.2), if $\mathcal{H}_0$ *is true, then when* $n \to +\infty$, *the distribution of* λ *tends to a* χ^2 *distribution with* $d - d_0$ *degrees of freedom.*

The *significance* of a test $\alpha \in [0,1]$, or *type I error* probability, is the probability to reject $\mathcal{H}_0$ when $\mathcal{H}_0$ is true, while the *power* of a test $1-\beta \in [0,1]$ is the probability to reject $\mathcal{H}_0$ when $\mathcal{H}_0$ is false, β is the *type II error* probability.

3 Policy space identification in a fixed environment

As we introduced in Sect. 1, we aim at identifying the agent's policy space by observing a set of demonstrations coming from the optimal policy of the agent. We assume that the agent is playing a policy π^{Ag} belonging to a parametric policy space Π_Θ.

Assumption 1 (*Parametric Agent's Policy)* The agent's policy π^{Ag} belongs to a *known* parametric policy space Π_Θ, i.e., there exists a (maybe not unique) $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta$ such that $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}(\cdot|s) = \pi^{\mathrm{Ag}}(\cdot|s)$ almost surely for all $s \in \mathcal{S}$.

It is important to stress π^{Ag} is one of the possibly many optimal policies within the policy space Π_Θ, which, in turn, might be unable to represent the optimal Markovian stationary policy. Furthermore, we do not explicitly report the dependence on the agent's parameter $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta$ as, in the general case, there might exist multiple parameters yielding the same policy π^{Ag}.

We have access to a dataset $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$ where $s_i \sim \nu$ and $a_i \sim \pi^{\mathrm{Ag}}(\cdot|s_i)$ sampled independently.[4] ν is a sampling distribution over the states. Although we will present the method for a generic $\nu \in \mathscr{P}(\mathcal{S})$, in practice, we employ as ν the γ–discounted stationary distribution induced by π^{Ag}, i.e., $d_\mu^{\pi^{\mathrm{Ag}}}(s) = (1-\gamma)\sum_{t=0}^{+\infty} \Pr(s_t = s|\mathcal{M}, \pi^{\mathrm{Ag}})$ (Sutton et al., 1999). We assume that the agent has control over a limited number of parameters $d^{\mathrm{Ag}} < d$

[4] For exposition simplicity, we limit the presentation to the case of i.i.d. samples (Sutton et al., 2008). Nevertheless, by means of the *blocking technique* (Yu, 1994), it is possible to generalize the concentration results to β-mixing strictly stationary processes, provided that the mixing rate is exponential (e.g., Antos et al., 2008; Lazaric et al., 2012; Dai et al., 2018).

whose value can be changed during learning, while the remaining $d - d^{\mathrm{Ag}}$ are kept fixed to zero.[5] Given a set of indexes $I \subseteq \{1, ..., d\}$ we define the subset of the parameter space: $\Theta_I = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0, \forall i \in \{1, ..., d\} \setminus I\}$. Thus, the set I represents the indexes of the parameters that can be changed if the agent's parameter space were Θ_I. Our goal is to find a set of parameter indexes I^{Ag} that are *sufficient* to explain the agent's policy, i.e., $\pi^{\mathrm{Ag}} \in \Pi_{\Theta_{I^{\mathrm{Ag}}}}$ but also *necessary*, in the sense that when removing any $i \in I^{\mathrm{Ag}}$ the remaining ones are insufficient to explain the agent's policy, i.e., $\pi^{\mathrm{Ag}} \notin \Pi_{\Theta_{I^{\mathrm{Ag}} \setminus \{i\}}}$. We formalize these notions in the following definition.

Definition 2 (**Correctness**) Let $\pi^{\mathrm{Ag}} \in \Pi_\Theta$. A set of parameter indexes $I^{\mathrm{Ag}} \subseteq \{1, ..., d\}$ is *correct* w.r.t. π^{Ag} if:

$$\pi^{\mathrm{Ag}} \in \Pi_{\Theta_{I^{\mathrm{Ag}}}} \ \wedge \ \forall i \in I^{\mathrm{Ag}} : \pi^{\mathrm{Ag}} \notin \Pi_{\Theta_{I^{\mathrm{Ag}} \setminus \{i\}}}.$$

We denote with $\mathcal{I}^{\mathrm{Ag}}$ the set of all correct set of parameter indexes I^{Ag}.

Thus, there exist multiple I^{Ag} when multiple parametric representations of the agent's policy π^{Ag} are possible. The uniqueness of I^{Ag} is guaranteed under the assumption that each policy admits a unique representation in Π_Θ, i.e., under the identifiability assumption.

Assumption 2 (*Identifiability*) The policy space Π_Θ is *identifiable*, i.e., for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have that if $\pi_{\boldsymbol{\theta}}(\cdot|s) = \pi_{\boldsymbol{\theta}'}(\cdot|s)$ almost surely for all $s \in \mathcal{S}$ than $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

The identifiability property allows rephrasing Definition 2 in terms of the policy parameters only, leading to the following result.

Lemma 1 (*Correctness under Identifiability*) *Under Assumption 2, let $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta$ be the unique parameter such that $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}(\cdot|s) = \pi^{\mathrm{Ag}}(\cdot|s)$ almost surely for all $s \in \mathcal{S}$. Then, there exists a unique set of parameter indexes $I^{\mathrm{Ag}} \subseteq \{1, ..., d\}$ that is correct w.r.t. π^{Ag} defined as:*

$$I^{\mathrm{Ag}} = \left\{ i \in \{1, ..., d\} \ : \ \theta_i^{\mathrm{Ag}} \neq 0 \right\}.$$

Consequently, $\mathcal{I}^{\mathrm{Ag}} = \{I^{\mathrm{Ag}}\}$.

Proof The uniqueness of I^{Ag} is ensured by Assumption 2. Let us rewrite the condition of Definition 2 under Assumption 2:

$$\begin{aligned} &\pi^{\mathrm{Ag}} \in \Pi_{\Theta_{IAg}} \ \wedge \ \forall i \in I^{\mathrm{Ag}} : \pi^{\mathrm{Ag}} \notin \Pi_{\Theta_{IAg \setminus \{i\}}} \\ &\qquad \iff \boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta_{IAg} \ \wedge \ \forall i \in I^{\mathrm{Ag}} : \boldsymbol{\theta}^{\mathrm{Ag}} \notin \Theta_{IAg \setminus \{i\}} \end{aligned} \tag{P.1}$$

$$\begin{aligned} &\iff \forall i \in I^{\mathrm{Ag}} : \theta_i^{\mathrm{Ag}} \neq 0 \ \wedge \ \forall i \in \{1, ..., d\} \setminus I^{\mathrm{Ag}} : \theta_i^{\mathrm{Ag}} = 0 \\ &\iff I^{\mathrm{Ag}} = \left\{ i \in \{1, ..., d\} \ : \ \theta_i^{\mathrm{Ag}} \neq 0 \right\}, \end{aligned} \tag{P.2}$$

[5] The extension of the identification rules to (known) fixed values different from zero is straightforward.

where line (P.1) follows since there is a unique representation for π^{Ag} determined by parameter $\boldsymbol{\theta}^{\mathrm{Ag}}$ and line (P.2) is obtained from the definition of Θ_I. □

Remark 1 *(About the Optimality of π^{Ag})* We started this section stating that π^{Ag} is an optimal policy within the policy space Π_Θ. This is motivated by the fact that typically we start with an overparametrized policy space Π_Θ and we seek for the minimal set of parameters that allows the agent to reach an optimal policy within Π_Θ. However, in practice, we usually have access to an ϵ-optimal policy $\pi^{\mathrm{Ag}}_\epsilon$, meaning that the performance of $\pi^{\mathrm{Ag}}_\epsilon$ is ϵ-close to the optimal performance.[6] Nevertheless, the notion of correctness (Definition 2) makes no assumptions on the optimality of π^{Ag}. If we replace π^{Ag} with $\pi^{\mathrm{Ag}}_\epsilon$ we will recover a set of parameter indexes I^{Ag}_ϵ that is, in general, different from I^{Ag}_ϵ, but we can still provide some guarantees. If $I^{\mathrm{Ag}} \subseteq I^{\mathrm{Ag}}_\epsilon$, then I^{Ag}_ϵ is sufficient to explain the optimal policy π^{Ag}, but not necessary in general (it might contain useless parameters for π^{Ag}). Instead, if $I^{\mathrm{Ag}} \not\subseteq I^{\mathrm{Ag}}_\epsilon$, then I^{Ag}_ϵ is not sufficient to explain the optimal policy π^{Ag}. In any case, I^{Ag}_ϵ is necessary and sufficient to represent, at least, an ϵ-optimal policy.

The following two subsections are devoted to the presentation of the *identification rules* based on the application of Definition 2 (Sect. 3.1) and Lemma 1 (Sect. 3.2) when we only have access to a dataset of samples $\mathcal{D}$. The goal of an identification rule consists in producing a set $\widehat{\mathcal{I}}$, approximating $\mathcal{I}^{\mathrm{Ag}}$. The idea at the basis of our identification rules consists in employing the GLR test to assess the correctness (Definition 2 or Lemma 1) of a candidate set of indexes.

3.1 Combinatorial identification rule

In principle, using $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$, we could compute the maximum likelihood parameter $\widehat{\boldsymbol{\theta}} \in \operatorname*{arg\,sup}_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})$ and employ it with Definition 2. However, this approach has, at least, two drawbacks. First, when Assumption 2 is not fulfilled, it would produce a single approximate parameter, while multiple choices might be viable. Second, because of the estimation errors, we would hardly get a zero value for the parameters the agent might not control. For these reasons, we employ a GLR test to assess whether a specific set of parameters is zero. Specifically, for all $I \subseteq \{1, ..., d\}$ we consider the pair of hypotheses $\mathcal{H}_{0,I} \,:\, \pi^{\mathrm{Ag}} \in \Pi_{\Theta_I}$ against $\mathcal{H}_{1,I} \,:\, \pi^{\mathrm{Ag}} \in \Pi_{\Theta \setminus \Theta_I}$ and the GLR statistic:

$$\lambda_I = -2 \log \frac{\sup_{\boldsymbol{\theta}\in\Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})} = 2\left(\widehat{\ell}(\widehat{\boldsymbol{\theta}}_I) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})\right), \tag{3}$$

where the likelihood is defined as $\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^n \pi_{\boldsymbol{\theta}}(a_i|s_i)$, $\widehat{\boldsymbol{\theta}}_I \in \operatorname*{arg\,sup}_{\boldsymbol{\theta}\in\Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\theta}} \in \operatorname*{arg\,sup}_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})$. We are now ready to state the identification rule derived from Definition 2.

[6] We can also look at $\pi^{\mathrm{Ag}}_\epsilon$ as the optimal policy within Π_Θ for a different MDP $\mathcal{M}_\epsilon$, that is an approximation of the original MDP $\mathcal{M}$.

Identification Rule 1 The *combinatorial identification rule* with threshold function c_l selects $\widehat{\mathcal{I}}_c$ containing all and only the sets of parameter indexes $I \subseteq \{1, ..., d\}$ such that:

$$\lambda_I \leq c_{|I|} \wedge \forall i \in I : \lambda_{I \setminus \{i\}} > c_{|I|-1}. \tag{4}$$

Thus, I is defined in such a way that the null hypothesis $\mathcal{H}_{0,I}$ is not rejected, i.e., I contains parameters that are sufficient to explain the data $\mathcal{D}$, and necessary since for all $i \in I$ the set $I \setminus \{i\}$ is no longer sufficient, as $\mathcal{H}_{0,I\setminus\{i\}}$ is rejected. The threshold function c_l, which depend on the cardinality l of the tested set of indexes, controls the behavior of the tests. In practice, we recommend setting them by exploiting Wilk's asymptotic approximation (Theorem 1) to enforce (asymptotic) guarantees on the type I error. Given a significance level $\delta \in [0, 1]$, since for Identification Rule 1 we perform 2^d statistical tests by using the same dataset $\mathcal{D}$, we partition δ using Bonferroni correction and setting $c_l = \chi^2_{l,1-\delta/2^d}$, where $\chi^2_{l,\bullet}$ is the $\bullet$–quintile of a chi square distribution with l degrees of freedom. Refer to Algorithm 1 for the pseudocode of the identification procedure.

Algorithm 1 Identification Rule 1 (Combinatorial)

```
input: dataset $\mathcal{D}$, parameter space $\Theta$, threshold function $c$ (e.g., $c_l = \chi^2_{l,1-\delta/2^d}$)
  $\widehat{\mathcal{I}}_c \leftarrow \{\}$
  $\widehat{\mathcal{L}} = \max_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})$
  for $I \subseteq \{1, ..., d\}$ sorted by cardinality do
      $\widehat{\mathcal{L}}_I = \max_{\boldsymbol{\theta}\in\Theta_I} \widehat{\mathcal{L}}(\boldsymbol{\theta})$
      $\lambda_I = -2\log\frac{\widehat{\mathcal{L}}_I}{\widehat{\mathcal{L}}}$
      if $\lambda_I \leqslant c_{|I|}$ and $\forall i \in I : \lambda_{I\setminus\{i\}} > c_{|I|-1}$ then
          $\widehat{\mathcal{I}}_c \leftarrow \widehat{\mathcal{I}}_c \cup \{I\}$
      end if
  end for
  return $\widehat{\mathcal{I}}_c$
```

3.2 Simplified identification rule

Identification Rule 1 is hard to be employed in practice, as it requires performing $\mathcal{O}(2^d)$ statistical tests. However, under Assumption 2, to retrieve I^{Ag} we do not need to test all subsets, but we can just examine one parameter at a time (see Lemma 1). Thus, for all $i \in \{1, ..., d\}$ we consider the pair of hypotheses $\mathcal{H}_{0,i} : \theta_i^{\mathrm{Ag}} = 0$ against $\mathcal{H}_{1,i} : \theta_i^{\mathrm{Ag}} \neq 0$ and define $\Theta_i = \{\boldsymbol{\theta} \in \Theta : \theta_i = 0\}$. The GLR test can be performed straightforwardly, using the statistic:

$$\lambda_i = -2\log\frac{\sup_{\boldsymbol{\theta}\in\Theta_i} \widehat{\mathcal{L}}(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})} = 2\Big(\widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})\Big), \tag{5}$$

where the likelihood is defined as $\widehat{\mathcal{L}}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_{\boldsymbol{\theta}}(a_i|s_i)$, $\widehat{\boldsymbol{\theta}}_i = \arg\sup_{\boldsymbol{\theta}\in\Theta_i} \widehat{\mathcal{L}}(\boldsymbol{\theta})$ and $\widehat{\boldsymbol{\theta}} = \arg\sup_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})$.[7] In the spirit of Lemma 1, we define the following identification rule.

Identification Rule 2 The *simplified identification rule* with threshold function c_1 selects $\widehat{\mathcal{I}}_c$ containing the unique set of parameter indexes $\widehat{I}_c$ such that:

$$\widehat{I}_c = \{i \in \{1, ..., d\} : \lambda_i > c_1\}. \tag{6}$$

Therefore, the identification rule constructs $\widehat{I}_c$ by taking all the indexes $i \in \{1, ..., d\}$ such that the corresponding null hypothesis $\mathcal{H}_{0,i} : \theta_i^{\mathrm{Ag}} = 0$ is rejected, i.e., those for which there is statistical evidence that their value is not zero. Similarly to the combinatorial identification rule, we recommend setting the threshold function c_1 based on Wilk's approximation. Given a significance level $\delta \in [0, 1]$, since we perform d statistical tests, we employ Bonferroni correction and we set $c_1 = \chi^2_{1,1-\delta/d}$. Refer to Algorithm 2 for the pseudocode of the identification rule.

Algorithm 2 Identification Rule 2 (Simplified)

```
input: dataset $\mathcal{D}$, parameter space $\Theta$, threshold function $c$ (e.g., $c_1 = \chi^2_{l,1-\delta/d}$)
  $\widehat{I}_c \leftarrow \{\}$
  $\widehat{\mathcal{L}} = \max_{\boldsymbol{\theta}\in\Theta} \widehat{\mathcal{L}}(\boldsymbol{\theta})$
  for $i \in \{1, ..., d\}$ do
      $\widehat{\mathcal{L}}_i = \max_{\boldsymbol{\theta}\in\Theta_i} \widehat{\mathcal{L}}(\boldsymbol{\theta})$
      $\lambda_i = -2\log\frac{\widehat{\mathcal{L}}_i}{\widehat{\mathcal{L}}}$
      if $\lambda_i > c_1$ then
          $\widehat{I}_c \leftarrow \widehat{I}_c \cup \{i\}$
      end if
  end for
  return $\{\widehat{I}_c\}$
```

This second procedure requires a test for every parameter, i.e., $\mathcal{O}(d)$ instead of $\mathcal{O}(2^d)$ tests. However, the correctness of Identification Rule 2, in the sense of Definition 2, comes with the cost of assuming the identifiability property (Assumption 2). What happens if we employ this second procedure in a case where the assumption does not hold? Consider, for instance, the case in which two parameters θ_1 and θ_2 are exchangeable, we will include none of them in $\widehat{I}_c$ as, individually, they are not necessary to explain the agent's policy, while the pair $(\theta_1, \theta_2)^T$ is indeed necessary. We will discuss how to enforce Assumption 2, for the case of policies belonging to the exponential family, in the following section.

Remark 2 *(On Frequentist and Bayesian Statistical Tests)* In this paper, we restrict our attention to frequentist statistical tests, but, in principle, the same approaches can be extended to the Bayesian setting (Jeffreys, 1935). Indeed, the GLR test admits a Bayesian

[7] This setting is equivalent to a particular case the combinatorial rule in which $\mathcal{H}_{\star,i} \equiv \mathcal{H}_{\star,\{1,...,d\}\setminus\{i\}}$, with $\star \in \{0, 1\}$ and, consequently, $\lambda_i \equiv \lambda_{\{1,...,d\}\setminus\{i\}}$ and $\Theta_i = \Theta_{\{1,...,d\}\setminus\{i\}}$.

counterpart, known as the *Bayes Factor* (BF, Goodman, 1999; Morey et al., 2016). We consider the same setting presented in Sect. 2 in which we aim at testing the null hypothesis $\mathcal{H}_0 : \boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta_0$, against the alternative $\mathcal{H}_1 : \boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta \setminus \Theta_0$. We take the Bayesian perspective, looking at each $\boldsymbol{\theta}$ not as an unknown fixed quantity but as a realization of *prior* distributions on the parameters defined in terms of the hypothesis: $p(\boldsymbol{\theta}|\mathcal{H}_\star)$ for $\star \in \{0,1\}$. Thus, given a dataset $\mathcal{D} = \{X_i\}_{i=1}^n$, we can compute the likelihood of $\mathcal{D}$ given a parameter $\boldsymbol{\theta}$ as usual: $p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^n p(X_i|\boldsymbol{\theta})$. Combining the likelihood and the prior, we define the *Bayes Factor* as:

$$\Lambda^{\mathrm{BF}} = \frac{p(\mathcal{D}|\mathcal{H}_0)}{p(\mathcal{D}|\mathcal{H}_1)} = \frac{\int_\Theta \; p(\mathcal{D}|\boldsymbol{\theta}) \; p(\boldsymbol{\theta}|\mathcal{H}_0) \, \mathrm{d}\boldsymbol{\theta}}{\int_\Theta \; \underbrace{p(\mathcal{D}|\boldsymbol{\theta})}_{\text{likelihood}} \; \underbrace{p(\boldsymbol{\theta}|\mathcal{H}_1)}_{\text{prior}} \, \mathrm{d}\boldsymbol{\theta}}$$

The Bayesian approach has the clear advantage of incorporating additional domain knowledge by means of the prior. Furthermore, if also a prior on the hypothesis is available $p(\mathcal{H}_\star)$ for $\star \in \{0,1\}$ it is possible to compute the ratio of the *posterior* probability of each hypothesis:

$$\underbrace{\frac{p(\mathcal{H}_0|\mathcal{D})}{p(\mathcal{H}_1|\mathcal{D})}}_{\text{posterior ratio}} = \underbrace{\frac{p(\mathcal{D}|\mathcal{H}_0)}{p(\mathcal{D}|\mathcal{H}_1)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\text{prior ratio}} .$$

Compared to the GLR test, the Bayes factor provides richer information, since we can compute the likelihood of each hypothesis, given the data $\mathcal{D}$. However, like any Bayesian approach, the choice of the prior turns out to be of crucial importance. The computationally convenient prior (which might allow computing the integral in closed form) is typically not correct, leading to a biased test. In this sense, GLR replaces the integral with a single-point approximation centered in the maximum likelihood estimate. For these reasons, we leave the investigation of Bayesian approaches for policy space identification as future work.

4 Analysis for the exponential family

In this section, we provide an analysis of the Identification Rule 2 for a policy $\pi_{\boldsymbol{\theta}}$ linear in some state features $\boldsymbol{\phi}$ that belongs to the exponential family.[8] The section is organized as follows. We first introduce the exponential family, deriving a concentration result of independent interest (Theorem 2) and then we apply it for controlling the identification errors made by our identification rule (Theorem 3).

Exponential Family We refer to the definition of linear exponential family given in (Brown, 1986), that we state as an assumption.

Assumption 3 (*Exponential Family of Linear Policies*) Let $\boldsymbol{\phi} : \mathcal{S} \rightarrow \mathbb{R}^q$ be a feature function. The policy space Π_Θ is a space of *linear policies*, belonging to the exponential family, i.e., $\Theta = \mathbb{R}^d$ and all policies $\pi_{\boldsymbol{\theta}} \in \Pi_\Theta$ have form:

$$\pi_{\boldsymbol{\theta}}(a|s) = h(a) \exp\left\{\boldsymbol{\theta}^T \boldsymbol{t}(s,a) - A(\boldsymbol{\theta}, s)\right\}, \tag{7}$$

[8] We limit our analysis to Identification Rule 2 since we will show that, in the case of linear policies belonging to the exponential family, the identifiability property can be easily enforced.

where h is a positive function, $\boldsymbol{t}(s,a)$ is the *sufficient statistic* that depends on the state via the feature function $\boldsymbol{\phi}$ (i.e., $\boldsymbol{t}(s,a)=\boldsymbol{t}(\boldsymbol{\phi}(s),a)$) and $A(\boldsymbol{\theta},s)=\log\int_{\mathcal{A}}h(a)\exp\{\boldsymbol{\theta}^T\boldsymbol{t}(s,a)\}\mathrm{d}a$ is the log partition function. We denote with $\overline{\boldsymbol{t}}(s,a,\boldsymbol{\theta})=\boldsymbol{t}(s,a)-\mathop{\mathbb{E}}_{\overline{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s,\overline{a})]$ the centered sufficient statistic.

This definition allows modeling the linear policies that are a popular choice in linear time-invariant systems and a valid option for robotic control (Deisenroth et al., 2013), sometimes even competitive with complex neural network parametrizations (Rajeswaran et al., 2017). Table 1 shows how to map the Gaussian linear policy with fixed covariance, typically used in continuous action spaces, and the Boltzmann linear policy, suitable for finite action spaces, to Assumption 3 (details in Appendix A.1).

For the sake of the analysis, we enforce the following assumption concerning the tail behavior of the policy $\pi_{\boldsymbol{\theta}}$.

Assumption 4 (*Subgaussianity*) For any $\boldsymbol{\theta}\in\Theta$ and for any $s\in\mathcal{S}$ the centered sufficient statistic $\overline{\boldsymbol{t}}(s,a,\boldsymbol{\theta})$ is subgaussian with parameter $\sigma\geq 0$, i.e., for any $\boldsymbol{\alpha}\in\mathbb{R}^d$:

$$\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left\{\boldsymbol{\alpha}^T\overline{\boldsymbol{t}}(s,a,\boldsymbol{\theta})\right\}\right]\leq\exp\left\{\frac{1}{2}\|\boldsymbol{\alpha}\|_2^2\sigma^2\right\}.$$

A sufficient condition to ensure that Gaussian and Boltzmann are subgaussian is that the features $\boldsymbol{\phi}(s)$ are bounded in L_2-norm, uniformly over the state space $\mathcal{S}$ (Proposition 2). Furthermore, limited to the policies complying with Assumption 3, the identifiability (Assumption 2) can be restated in terms of the Fisher Information matrix (Rothenberg et al., 1971; Little et al., 2010).

Lemma 2 (*Rothenberg et al.*, (1971), *Theorem* 3) *Let Π_Θ be a policy space, as in Assumption* 3. *Then, under suitable regularity conditions* (*see* Rothenberg et al., (1971)), *if the Fisher Information matrix* (*FIM*) $\mathcal{F}(\boldsymbol{\theta})$:

$$\mathcal{F}(\boldsymbol{\theta})=\mathop{\mathbb{E}}_{\substack{s\sim\nu\\ a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}}\left[\overline{\boldsymbol{t}}(s,a,\boldsymbol{\theta})\overline{\boldsymbol{t}}(s,a,\boldsymbol{\theta})^T\right]\tag{8}$$

is non–singular for all $\boldsymbol{\theta}\in\Theta$, then Π_Θ is identifiable. In this case, we denote with $\lambda_{\min}=\inf_{\boldsymbol{\theta}\in\Theta}\lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))>0$.

Proposition 1 of Appendix A.2.1 shows that a sufficient condition for the identifiability in the case of Gaussian and Boltzmann linear policies is that the second moment matrix of the feature vector $\mathop{\mathbb{E}}_{s\sim\nu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right]$ is non–singular along with the fact that the policy $\pi_{\boldsymbol{\theta}}$ plays each action with positive probability for the Boltzmann policy.

Remark 3 *(How to enforce identifiability?)* Requiring that $\mathop{\mathbb{E}}_{s\sim\nu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right]$ is full rank is essentially equivalent to require that all features ϕ_i are linearly independent for all $i\in\{1,...,d\}$. This condition can be easily met with a preprocessing phase that removes the

Table 1 Action space $\mathcal{A}$, probability density function $\pi_{\widetilde{\boldsymbol{\theta}}}$, sufficient statistic $\boldsymbol{t}$, and function h for the Gaussian linear policy with fixed covariance and the Boltzmann linear policy

Policy	Gaussian	Boltzmann
$\mathcal{A}$	$\boldsymbol{a} \in \mathbb{R}^k$	$a_i \in \{a_1, ..., a_{k+1}\}$
$\pi_{\widetilde{\boldsymbol{\theta}}}$	$\frac{1}{(2\pi)^{\frac{k}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{a}-\widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{a}-\widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s))}$	$\begin{cases} \frac{e^{\widetilde{\boldsymbol{\theta}}_i^T \boldsymbol{\phi}(s)}}{1+\sum_{j=1}^k e^{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)}} & \text{if } i \leq k \\ \frac{1}{1+\sum_{j=1}^k e^{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)}} & \text{if } i = k \end{cases}$
$\boldsymbol{t}$	$\boldsymbol{\Sigma}^{-1}\boldsymbol{a} \otimes \boldsymbol{\phi}(s)$	$\begin{cases} \boldsymbol{e}_i \otimes \boldsymbol{\phi}(s) & \text{if } i \leq k \\ \boldsymbol{0} & \text{if } i = k+1 \end{cases}$
h	$\frac{1}{(2\pi)^{\frac{k}{2}} \det(\boldsymbol{\Sigma})^{\frac{1}{2}}} e^{-\frac{1}{2}\boldsymbol{a}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{a}}$	1

For convenience of representation $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{k \times q}$ is a matrix and $\boldsymbol{\theta} = \operatorname{vec}(\widetilde{\boldsymbol{\theta}}^T) \in \mathbb{R}^d$, with $d = kq$. We denote with $\boldsymbol{e}_i$ the i–th vector of the canonical basis of $\mathbb{R}^k$ and with $\otimes$ the Kronecker product

linearly dependent features, for instance, by employing Principal Component Analysis (PCA, Jolliffe, 2011). For this reason, in our experimental evaluation we will always consider the case of linearly independent features.

When working with samples, however, we need to estimate the FIM from samples, leading to the *empirical FIM*, in which the expectation over the states of Eq. (8), is replaced with the sample mean:

$$\widehat{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\bar{\boldsymbol{t}}(s_i, a, \boldsymbol{\theta})\bar{\boldsymbol{t}}(s_i, a, \boldsymbol{\theta})^T\right], \tag{9}$$

where $\{s_i\}_{i=1}^n \sim \nu$. We denote with $\widehat{\lambda}_{\min} = \inf_{\boldsymbol{\theta} \in \Theta} \lambda_{\min}(\widehat{\mathcal{F}}(\boldsymbol{\theta}))$ the minimum eigenvalue of the empirical FIM. In order to carry out the subsequent analysis, we need to require that this quantity is non-zero.

Assumption 5 (*Positive Eigenvalues of Empirical FIM*) The minimum eigenvalue of the empirical FIM $\widehat{\mathcal{F}}(\boldsymbol{\theta})$ is non-zero for all $\boldsymbol{\theta} \in \Theta$, i.e., $\widehat{\lambda}_{\min} = \inf_{\boldsymbol{\theta} \in \Theta} \lambda_{\min}(\widehat{\mathcal{F}}(\boldsymbol{\theta})) > 0$.

The condition of Assumption 5 can be enforced as long as the true FIM $\mathcal{F}(\boldsymbol{\theta})$ has a positive minimum eigenvalue $\lambda_{\min}$, i.e., under identifiability assumption (Lemma 2) and given a sufficiently large number of samples. Proposition 4 of Appendix A.2.1 provides the minimum number of samples such that with high probability it holds that $\widehat{\lambda}_{\min} > 0$.

We are now ready to present a concentration result, of independent interest, for the parameters and the negative log–likelihood that represents the central tool of our analysis.

Theorem 2 *Under Assumptions* 1, 2, 3, 4, *and* 5, *let* $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$ *be a dataset of* $n > 0$ *independent samples, where* $s_i \sim \nu$ *and* $a_i \sim \pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}(\cdot|s_i)$. *Let* $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \widehat{\ell}(\boldsymbol{\theta})$ *and* $\boldsymbol{\theta}^{\mathrm{Ag}} = \arg\min_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta})$. *Then, for any* $\delta \in [0, 1]$, *with probability at least* $1 - \delta$ *it holds that*:

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2 \leq \frac{\sigma}{\widehat{\lambda}_{\min}} \sqrt{\frac{2d}{n} \log \frac{2d}{\delta}}.$$

Furthermore, with probability at least $1 - \delta$, *it holds that individually*:

$$\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) \leq \frac{d^2\sigma^4}{\widehat{\lambda}^2_{\min} n} \log \frac{2d}{\delta} \quad \text{and} \quad \widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) \leq \frac{d^2\sigma^4}{\widehat{\lambda}^2_{\min} n} \log \frac{2d}{\delta}.$$

Proof sketch The idea of the proof is to first obtain a probabilistic bound on the parameter difference in norm $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2$. This result is given in Theorem 6. Then, we use the latter result together with Taylor expansion to bound the differences $\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})$ and $\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})$, as in Corollary 1. The full derivation can be found in Appendix A.2.3.

The theorem shows that the L_2–norm of the difference between the maximum likelihood parameter $\widehat{\boldsymbol{\theta}}$ and the true parameter $\boldsymbol{\theta}^{\mathrm{Ag}}$ concentrates with rate $\mathcal{O}(n^{-1/2})$ while the likelihood $\widehat{\ell}$ and its expectation ℓ concentrate with faster rate $\mathcal{O}(n^{-1})$.

Identification Rule Analysis We are now ready to start the analysis of Identification Rule 2. The goal of the analysis is, informally, to bound the probability of an identification error as a function of the number of samples n and the threshold function c_1. For this purpose, we define the following quantities.

Definition 3 Consider an identification rule producing $\widehat{I}$ as approximate parameter index set. We define the significance α and the power $1 - \beta$ of the identification rule as:

$$\alpha = \Pr\left(\exists i \notin I^{\mathrm{Ag}} : i \in \widehat{I}\right), \quad \beta = \Pr\left(\exists i \in I^{\mathrm{Ag}} : i \notin \widehat{I}\right).$$

Thus, α represents the probability that the identification rule selects a parameter that the agent does not control, whereas β is the probability that the identification rule does not select a parameter that the agent does control.[9]

By employing the results we derived for the exponential family (Theorem 2) we can now bound α and β, under a slightly more demanding assumption on $\widehat{\lambda}_{\min}$.

Theorem 3 *Let* $\widehat{I}_c$ *be the set of parameter indexes selected by the Identification Rule 2 obtained using* $n > 0$ *i.i.d. samples collected with* $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}$, *with* $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta$. *Then, under Assumptions 1, 2, 3, 4, and 5, let* $\boldsymbol{\theta}_i^{\mathrm{Ag}} = \arg\min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta})$ *for all* $i \in \{1, ..., d\}$ *and* $\xi = \min\{1, \frac{\lambda_{\min}}{\sigma^2}\}$. *If* $\widehat{\lambda}_{\min} \geq \frac{\lambda_{\min}}{2\sqrt{2}}$ *and* $\ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - l(\boldsymbol{\theta}^{\mathrm{Ag}}) \geq c_1$, *it holds that*:

[9] We use the symbols α and β to highlight the analogy between these probabilities and the type I and type II error probabilities of a statistical test.

$$
\begin{aligned}
\alpha &\leq 2d \exp\left\{-\frac{c_1 \lambda_{\min}^2 n}{16 d^2 \sigma^4}\right\} \\
\beta &\leq (2d-1) \sum_{i \in I^{\text{Ag}}} \exp\left\{-\frac{\left(l(\boldsymbol{\theta}_i^{\text{Ag}}) - l(\boldsymbol{\theta}^{\text{Ag}}) - c_1\right) \lambda_{\min} \xi n}{16 (d-1)^2 \sigma^2}\right\}.
\end{aligned}
$$

Proof sketch Concerning $\alpha = \Pr\left(\exists i \notin I^{\text{Ag}} : i \in \widehat{I}_c\right)$, we employ a technique similar to that of Lemma 2 in (Garivier and Kaufmann, 2019) to remove the existential quantification. Instead, for $\beta = \Pr\left(\exists i \in I^{\text{Ag}} : i \notin \widehat{I}_c\right)$ we first perform a union bound over $i \in I^{\text{Ag}}$ and then we bound the individual $\Pr\left(i \notin \widehat{I}_c\right)$. The full derivation can be found in Appendix A.3. □

In principle, we could employ Theorem 3 to derive a proper value of c_1 and n, given a required value of α and β. Unfortunately, their expression depend on $\lambda_{\min}$ which is unknown in practice. As already mentioned in the previous sections, we recommend employing Wilk's asymptotic approximation to set the threshold function as $c_1 = \chi^2_{1,1-\delta/d}$. This choice allows an asymptotic control of the significance of the identification rule.

Theorem 4 *Let $\widehat{I}_c$ be the set of parameter indexes selected by the Identification Rule 2 obtained using $n > 0$ i.i.d. samples collected with $\pi_{\boldsymbol{\theta}^{\text{Ag}}}$, with $\boldsymbol{\theta}^{\text{Ag}} \in \Theta$. Then, under suitable regularity conditions (see Casella and Berger,* (2002) *Section* 10.6.2), if $c_1 = \chi^2_{1,1-\delta/d}$ *it holds that $\alpha \leq \delta$ when $n \to +\infty$.*

Proof Starting from the definition of α, we first perform a union bound over $i \notin I^{\text{Ag}}$ to remove the existential quantification.

$$
\alpha = \Pr\left(\exists i \notin I^{\text{Ag}} : i \in \widehat{I}_c\right) = \Pr\left(\bigvee_{i \notin I^{\text{Ag}}} i \in \widehat{I}_c\right) \leq \sum_{i \notin I^{\text{Ag}}} \Pr\left(i \in \widehat{I}_c\right).
$$

Now, we bound each $\Pr\left(i \in \widehat{I}_c\right)$ individually, recalling that λ_i is distributed asymptotically as a χ^2 distribution with 1 degree of freedom and that $c_1 = \chi^2_{1,1-\delta/d}$:

$$
\Pr\left(i \in \widehat{I}_c\right) = \Pr\left(\lambda_i > \chi^2_{1,1-\delta/d}\right) \to \frac{\delta}{d}, \quad n \to \infty.
$$

Thus, we have that when $n \to +\infty$:

$$
\alpha \leq \frac{d - d^{\text{Ag}}}{d} \delta \leq \delta. \tag{P.3}
$$

□

5 Policy space identification in a configurable environment

The identification rules presented so far are unable to distinguish between a parameter set to zero because the agent cannot control it or because zero is its optimal value. To overcome this issue, we employ the Conf–MDP properties to select a configuration in which the parameters we want to examine have an optimal value other than zero. Intuitively, if we want to test whether the agent can control parameter θ_i, we should place the agent in an environment $\boldsymbol{\omega}_i \in \Omega$ where θ_i is "maximally important" for the optimal policy. This intuition is justified by Theorem 3, since to maximize the *power* of the test $(1-\beta)$, all other things being equal, we should maximize the log–likelihood gap $l(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - l(\boldsymbol{\theta}^{\mathrm{Ag}})$, i.e., parameter θ_i should be essential to justify the agent's behavior. Let $I \subseteq \{1, ..., d\}$ be a set of parameter indexes we want to test, our ideal goal is to find the environment $\boldsymbol{\omega}_I$ such that:

$$\boldsymbol{\omega}_I \in \operatorname*{arg\,max}_{\boldsymbol{\omega}\in\Omega} \left\{ l(\boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega})) - l(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega})) \right\}, \tag{10}$$

where $\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}) \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta} J_{\mathcal{M}_\omega}(\boldsymbol{\theta})$ and $\boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega}) \in \operatorname*{arg\,max}_{\boldsymbol{\theta}\in\Theta_I} J_{\mathcal{M}_\omega}(\boldsymbol{\theta})$ are the parameters of the optimal policies in the environment $\mathcal{M}_{\boldsymbol{\omega}}$ considering Π_Θ and Π_{Θ_I} as policy spaces respectively. Clearly, given the samples $\mathcal{D}$ collected with a single optimal policy $\pi^{\mathrm{Ag}}(\boldsymbol{\omega}_0)$ in a single environment $\mathcal{M}_{\boldsymbol{\omega}_0}$, solving problem (10) is hard as it requires performing an off–distribution optimization both on the space of policy parameters and configurations. For these reasons, we consider a surrogate objective that assumes that the optimal parameter in the new configuration can be reached by performing a single gradient step.[10]

Theorem 5 *Let $I \in \{1, ..., d\}$ and $\overline{I} = \{1, ..., d\} \setminus I$. For a vector $\boldsymbol{v} \in \mathbb{R}^d$, we denote with $\boldsymbol{v}|_I$ the vector obtained by setting to zero the components in I. Let $\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0) \in \Theta$ the initial parameter. Let $\alpha \geq 0$ be a learning rate, $\boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega}) = \boldsymbol{\theta}_0 + \alpha \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_\omega}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))|_I$ and $\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}) = \boldsymbol{\theta}_0 + \alpha \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_\omega}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))$. Then, under Assumption 2, we have:*

$$\ell(\boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega})) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega})) \geq \frac{\lambda_{\min}\alpha^2}{2} \left\| \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_\omega}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))|_{\overline{I}} \right\|_2^2.$$

Proof By second-order Taylor expansion of ℓ and recalling that $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega})) = \boldsymbol{0}$, we have:

[10] This idea shares some analogies with the *adapted parameter* in the meta-learning setting (Finn et al., 2017).

$$
\begin{aligned}
\ell(\boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega})) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega})) &\geq \frac{\lambda_{\min}}{2} \left\| \boldsymbol{\theta}_I^{\mathrm{Ag}}(\boldsymbol{\omega}) - \boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}) \right\|_2^2 \\
&= \frac{\lambda_{\min}}{2} \left\| \boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0) + \alpha \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_{\boldsymbol{\omega}}}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))|_I - \boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0) - \alpha \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_{\boldsymbol{\omega}}}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0)) \right\|_2^2 \\
&= \frac{\lambda_{\min}\alpha^2}{2} \left\| \nabla_{\boldsymbol{\theta}} J_{\mathcal{M}_{\boldsymbol{\omega}}}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))|_{\bar{I}} \right\|_2^2.
\end{aligned}
$$

□

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 $\pi_{\boldsymbol{\theta}Ag(\boldsymbol{\omega}_0)}$ and in the current environment $\boldsymbol{\omega}_0$, we have to perform an off–distribution optimization over $\boldsymbol{\omega}$. To this end, we employ an approach analogous to that of (Metelli et al., 2018b, 2020) where we optimize the empirical version of the objective with a penalization that accounts for the distance between the distribution over trajectories:

$$
\mathcal{C}_I(\boldsymbol{\omega}/\boldsymbol{\omega}_0) = \Big\| \underbrace{\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mathcal{M}_{\boldsymbol{\omega}/\boldsymbol{\omega}_0}}(\boldsymbol{\theta}^{\mathrm{Ag}}(\boldsymbol{\omega}_0))|_{\bar{I}}}_{\text{gradient estimator}} \Big\|_2^2 - \zeta \underbrace{\sqrt{\frac{\widehat{d}_2(\boldsymbol{\omega}\|\boldsymbol{\omega}_0)}{n}}}_{\substack{\text{dissimilarity}\\ \text{penalization}}}, \tag{11}
$$

where $\zeta \geq 0$ is a regularization parameter. We assume to have access to a dataset of trajectories $\mathcal{D} = \{\tau_i\}_{i=1}^n$ independently collected using policy $\pi_{\boldsymbol{\theta}}$ in the environment $\mathcal{M}_{\boldsymbol{\omega}_0}$. Each trajectory is a sequence of triples $\{(s_{i,t}, a_{i,t}, r_{i,t})\}_{t=1}^T$, where T is the trajectory horizon. The expression of the gradient estimator is given by:

$$
\begin{aligned}
\widehat{\nabla}_{\boldsymbol{\theta}} J_{\mathcal{M}_{\boldsymbol{\omega}/\boldsymbol{\omega}_0}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=0}^{T-1} \gamma^t r_{i,t} & \underbrace{\left(\frac{\mu_{\boldsymbol{\omega}}(s_{i,0})}{\mu_{\boldsymbol{\omega}_0}(s_{i,0})} \prod_{j=0}^{t} \frac{p_{\boldsymbol{\omega}}(s_{i,j+1}|s_{i,j}, a_{i,j})}{p_{\boldsymbol{\omega}_0}(s_{i,j+1}|s_{i,j}, a_{i,j})} \right)}_{\text{importance weight}} \\
& \times \sum_{j=0}^{t} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}} \left(a_{i,j} | s_{i,j} \right).
\end{aligned}
$$

The expression is obtained starting from the well–known G(PO)MDP gradient estimator and adapting for off–distribution estimation by introducing the importance weight (Metelli et al., 2018b). The dissimilarity penalization term corresponds to the estimated 2–Rényi divergence (Rényi, 1961) is obtained from the following expression, which represents the empirical second moment of the importance weight:

$$
\widehat{d}_2(\boldsymbol{\omega}\|\boldsymbol{\omega}_0) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mu_{\boldsymbol{\omega}}(s_{i,0})}{\mu_{\boldsymbol{\omega}_0}(s_{i,0})} \prod_{t=1}^{T} \frac{p_{\boldsymbol{\omega}}(s_{i,t+1}|s_{i,t}, a_{i,t})}{p_{\boldsymbol{\omega}_0}(s_{i,t+1}|s_{i,t}, a_{i,t})} \right)^2.
$$

Refer to (Metelli et al., 2018b) for the theoretical background behind the choice of this objective function. For conciseness, we report the pseudocode of the identification procedure in a configurable environment for Identification Rule 2 only (Algorithm 3), while the pseudocode for Identification Rule 2 can be found in Appendix B.

Algorithm 3 Identification Rule 2 (Simplified) with Environment Configuration.

```
input: parameter space $\Theta$, configuration space $\Omega$, threshold function $c_l$, number of configuration attempts $N_{\mathrm{conf}}$
  Initialize $\boldsymbol{\omega}_0$ arbitrarily
  Collect $\mathcal{D}_0$ observing $\pi_0^{\mathrm{Ag}}$ in environment $\mathcal{M}_{\boldsymbol{\omega}_0}$
  Run the Identification Rule 2 on $\mathcal{D}_0$ and obtain $\widehat{I}_0$
  $\widehat{I} \leftarrow \widehat{I}_0$
  for $i \in \{1, ..., d\} : i \notin \widehat{I}$ do
      $\boldsymbol{\omega}_{i,0} \leftarrow \boldsymbol{\omega}_0$
      $\mathcal{D}_{i,0} \leftarrow \mathcal{D}$
      for $j = 1, ..., N_{\mathrm{conf}}$ do
          Optimize $\mathcal{C}_{\{i\}}(\boldsymbol{\omega}/\boldsymbol{\omega}_{i,j-1})$ getting $\boldsymbol{\omega}_{i,j}$
          Collect $\mathcal{D}_{i,j}$ observing $\pi_{i,j}^{\mathrm{Ag}}$ in environment $\mathcal{M}_{\boldsymbol{\omega}_{i,j}}$
          Run the Identification Rule 2 on $\mathcal{D}_{i,j}$ and obtain $\widehat{I}_{i,j}$
          $\widehat{I} \leftarrow \widehat{I} \cup \widehat{I}_{i,j}$
      end for
  end for
  return $\{\widehat{I}\}$
```

6 Experimental results

In this section, we present the experimental results, focusing on three aspects of policy space identification.

- In Sect. 6.1, we provide experiments to assess the quality of our identification rules in terms of the ability to correctly identifying the parameters controlled by the agent.
- In Sect. 6.2, we focus on the application of policy space identification to Imitation Learning, comparing our identification rules with commonly employed regularization techniques.
- In Sect. 6.3, we consider the Conf-MDP framework and we show how properly identifying the parameters controlled by the agent allows learning better (more specific) environment configurations.

Additional experiments together with the hyperparameter values are reported in Appendix C.

6.1 Identification rules experiments

In this section, we provide two experiments to test the ability of our identification rules in properly selecting the parameters the agent controls in different settings. We start with an experiment on a discrete grid world (Sect. 6.1.1) to highlight the beneficial effects of environment configuration in parameter identification. Then, we provide an experiment on a simulated car driving domain (Sect. 6.1.2) in which we compare the combinatorial and the simplified identification rules.

6.1.1 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.

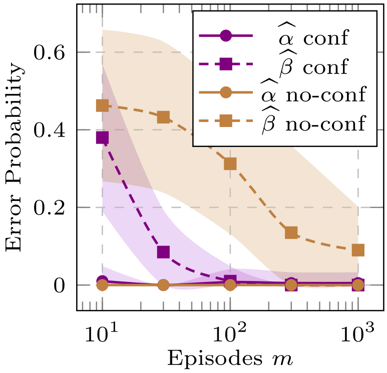


Fig. 2 *Discrete Grid World*: $\widehat{\alpha}$ and $\widehat{\beta}$ error for *conf* and *no-conf* cases varying the number of episodes. 25 runs 95% c.i

Whenever an action is performed, there is a small probability of failure (0.1) triggering a random action. 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 $\pi_{\boldsymbol{\theta}}$ with binary features $\boldsymbol{\phi}$ indicating its current row and column and the row and column of the goal.[11] For each run, the agent can control a subset I^{Ag} of the parameters $\boldsymbol{\theta}_{I^{\mathrm{Ag}}}$ associated with those features, which is randomly selected. Furthermore, the supervisor can configure the environment by changing the parameters $\boldsymbol{\omega}$ 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.

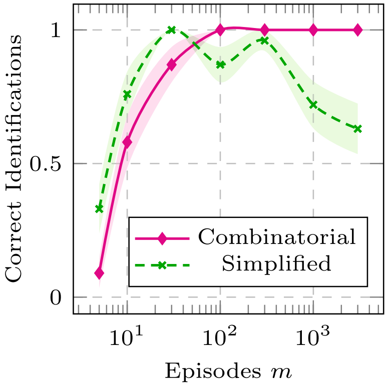
The goal of this set of experiments is to show the advantages of configuring the environment when performing the policy space identification using rule 2. Figure 2 shows the empirical $\widehat{\alpha}$ and $\widehat{\beta}$, i.e., the fraction of parameters that the agent does not control that are wrongly selected and the fraction of those the agent controls that are not selected respectively, as a function of the number m 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 Eq. (11), 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 threshold function c_1. Differently, $\widehat{\beta}$ decreases as the number of samples increases, i.e., the power of the test $1 - \widehat{\beta}$ increases with m. 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 empirically justifies our choice of objective (Eq. (11)) for selecting the new environment. Hyperparameters, further experimental results, together with experiments on a continuous version of the grid world, are reported in Appendix C.1.1–C.1.2.

6.1.2 Simulated car driving

We consider a simple version of a car driving simulator, in which the agent has to reach the end of a road in the minimum amount of time, avoiding running off-road. The agent

[11] The features are selected to fulfill Lemma 2.

Fig. 3 *Simulated Car Driving*: fraction of correct identifications varying the number of episodes. 100 runs 95% c.i



perceives its speed, four sensors placed at different angles that provide distance from the edge of the road and it can act on acceleration and steering.

The purpose of this experiment is to show a case in which the identifiability assumption (Assumption 2) may not be satisfied. The policy π_{θ} is modeled as a Gaussian policy whose mean is computed via a single hidden layer neural network with 8 neurons. Some of the sensors are not available to the agent, our goal is to identify which ones the agent can perceive.

In Fig. 3, we compare the performance of the Identification Rules 1 (Combinatorial) and 2 (Simplified), showing the fraction of runs that correctly identify the policy space. We note that, while for a small number of samples, the simplified rule seems to outperform, when the number of samples increases, the combinatorial rule displays remarkable stability, approaching the correct identification in all the runs. This is explained by the fact that, when multiple representations for the same policy are possible (like in this case when having a neural network as policy), considering one parameter at a time might induce the simplified rule to select a wrong set of parameters. Hyperparameters are reported in Appendix C.1.3.

6.2 Application to imitation learning

IL aims at recovering a policy replicating the behavior of an expert's agent. Selecting the parameters that an agent can control can be interpreted as applying a form of regularization to the IL problem (Osa et al., 2018). In the IL literature, a widely used technique is based on *entropy regularization* (Neu et al., 2017), which was employed in several successful algorithms, such as Maximum Causal Entropy IRL methods (MCE, Ziebart et al., 2008,, 2010), and Generative Adversarial IL (Ho and Ermon, 2016). Alternatively, other approaches aim at enforcing a *sparsity* constraint on the recovered policy parameters (e.g., Lee et al., 2018; Reddy et al., 2019; Brantley et al., 2020).

The goal of this experiment consists in showing that if we have appropriately identified the expert's policy space, we can mitigate overfitting/underfitting phenomena, with a general benefit on the process of learning the imitating policy. This experiment is conducted in the grid world domain, introduced in Sect. 6.1.1, using the same setting. In each run, the expert agent plays a (near) optimal Boltzmann policy $\pi_{\theta^{\mathrm{Ag}}}$ that makes use of a subset of the available parameters and provides a dataset $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{n}$ of n samples coming from m episodes.

In the IL framework knowing the policy space of the expert agent means properly tailoring the hypothesis space in which we search for the imitation policy. For this reason, we propose a comparison with common regularization techniques applied to maximum likelihood estimation. Figure 4 shows on the left the norm of the parameter difference $\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}} \right\|_2$ between the parameter recovered by the different IL methods $\widehat{\boldsymbol{\theta}}$ and the true parameter employed by the expert $\boldsymbol{\theta}^{\mathrm{Ag}}$, whereas on the right we plot the estimated expected KL-divergence between the imitation policy and the expert's policy computed as:

$$\widehat{\mathbb{D}}_{\mathrm{KL}}\left(\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}} \| \pi_{\widehat{\boldsymbol{\theta}}}\right) = \frac{1}{n}\sum_{i=1}^{n} D_{\mathrm{KL}}\left(\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}(\cdot|s_i) \| \pi_{\widehat{\boldsymbol{\theta}}}(\cdot|s_i)\right).$$

The lines *Conf* and *No-conf* refer to the results of ML estimation obtained by restricting the policy space to the parameters identified by our simplified rule with and without employing environment configurability, respectively (precisely as in Sect. 6.1.1). *ML*, *Ridge*, and *Lasso* correspond to maximum likelihood estimation in the full parameter space. Specifically, they are obtained by minimizing the objective:

$$\mathcal{Q}(\boldsymbol{\theta}; \lambda^{\mathrm{R}}, \lambda^{\mathrm{R}}) = \underbrace{-\sum_{i=1}^{n} \log \pi_{\boldsymbol{\theta}}(a_i|s_i)}_{\widehat{\ell}(\boldsymbol{\theta}) \text{ log-likelihood}} + \underbrace{\lambda^{\mathrm{R}} \|\boldsymbol{\theta}\|_2^2}_{\text{ridge}} + \underbrace{\lambda^{\mathrm{L}} \|\boldsymbol{\theta}\|_1}_{\text{lasso}}.$$

For ML we perform no regularization ($\lambda^{\mathrm{R}} = \lambda^{\mathrm{L}} = 0$), for Ridge we set $\lambda^{\mathrm{R}} = 0.001$ and $\lambda^{\mathrm{L}} = 0$, and for Lasso we have $\lambda^{\mathrm{R}} = 0$ and $\lambda^{\mathrm{L}} = 0.001$.

We observe that Conf, i.e., the usage of our identification rule, together with environment configuration, outperforms the other methods. This is more evident in the expected KL-divergence plot (right), which is a more robust index compared to the norm of the parameter difference (left). Ridge and Lasso regularizations display good behavior, better than both the identification rule without configuration (No-Conf) and the plain maximum likelihood without regularization (ML). This illustrates two important points.

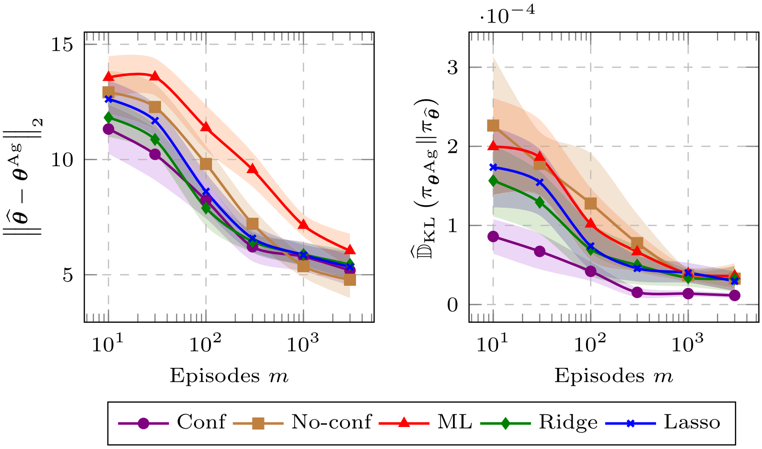


Fig. 4 *Discrete Grid World*: Norm of the difference between the expert's parameter $\boldsymbol{\theta}^{\mathrm{Ag}}$ and the estimated parameter $\widehat{\boldsymbol{\theta}}$ (*left*) and expected KL-divergence between the expert's policy $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}$ and the estimated policy $\pi_{\widehat{\boldsymbol{\theta}}}$ (*right*) as a function of the number of collected episodes m. 25 runs, 95% c.i

First, it confirms the benefits of configuring the environment for policy space identification. Second, it shows that a proper selection of the parameters controlled by the agent allows improving over standard ML, which tends to overfit.[12] We tested additional values of the regularization hyperparametrers λ^{R} and λ^{L} and other regularization techniques (Shannon and Tsallis entropy). The complete results are reported in Appendix C.2.

It is worth noting that the specific IL setting we consider, i.e., the availability of an initial dataset $\mathcal{D}$ of expert's demonstrations with no further interaction allowed[13] rules out from the comparison a large body of the literature that requires the possibility to interact with the expert or with the environment (e.g., Ho and Ermon, 2016; Lee et al., 2018). Nevertheless, these IL algorithms could be, in principle, adapted to this challenging no-interaction setting at the cost of restoring to off-policy estimation techniques (Owen, 2013), that, however, might inject further uncertainty in the learning process (see Appendix C.2 for details).

6.3 Application to configurable MDPs

The knowledge of the agent's policy space could be relevant when the learning process involves the presence of an external supervisor, as in the case of Configurable Markov Decision Process (Metelli et al., 2018a,, 2019). In a Conf-MDP, the supervisor is in charge of selecting the best configuration for the agent, i.e., the one that allows the agent to achieve the highest performance possible. As intuition suggests, the best environment configuration is closely related to the agent's capabilities. Agents with different perception and actuation possibilities might benefit from different configurations. Thus, the external supervisor should be aware of the agent's policy space to select the most appropriate configuration for the specific agent.

In the Minigolf environment (Lazaric et al., 2007), 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 2) of the distance from the hole (x) and the friction of the green (f). When an action is performed, a Gaussian noise is added whose magnitude depends on the green friction and on the action itself.

The goal of this experiment is to highlight that knowing the policy space is beneficial when learning in a Conf–MDP. We consider two agents with different perception capabilities: $\mathscr{A}_1$ 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.

Figure 5-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 $\omega^{\mathrm{Ag}}_{\mathscr{A}_1} = 5$ while for agent $\mathscr{A}_2$ is $\omega^{\mathrm{Ag}}_{\mathscr{A}_2} = 11.5$. Figure 5-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

[12] It is worth noting that the classical regularization techniques, like ridge and lasso, require choosing the regularization hyperparameter $\lambda^{\star}$ with $\star \in \{R, L\}$. In our experiments, we searched for the best parameter in $\{0.0001, 0.001, 0.01, 0.1, 1\}$.

[13] This setting was recently defined "truly batch model-free" (Ramponi et al., 2020).

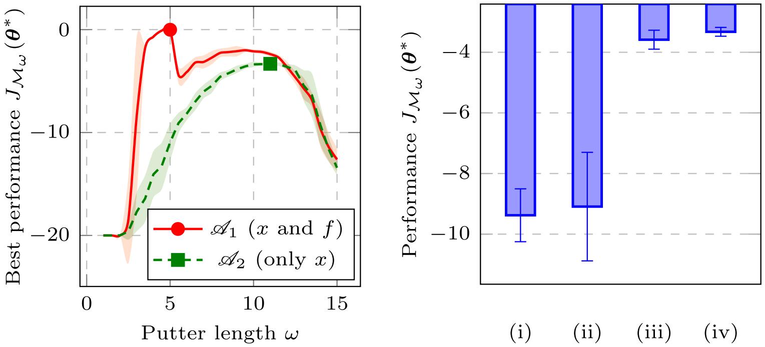


Fig. 5 *Mingolf*: Performance of the optimal policy varying the putter length ω for agents $\mathscr{A}_1$ and $\mathscr{A}_2$ (*left*) and performance of the optimal policy for agent $\mathscr{A}_2$ with four different strategies for selecting ω (*right*). 100 runs 95% c.i

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 2. 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 reported in Appendix C.3.

7 Conclusions

In this paper, we addressed the problem of identifying the policy space available to an agent in a learning process by simply observing its behavior when playing the optimal policy within such a space. We introduced two identification rules, both based on the GLR test, which can be applied to select the parameters controlled by the agent. Additionally, we have shown how to use the configurability property of the environment to improve the effectiveness of identification rules. The experimental evaluation highlights some essential points. First, the identification of the policy space brings advantages to the learning process in a Conf–MDP, helping to choose wisely the most suitable environment configuration. Second, we have shown that configuring the environment is beneficial for speeding up the identification process. Additionally, we have verified that policy space identification can improve imitation learning. Future research might investigate the usage of Bayesian statistical tests and the application of policy space identification to multi-agent RL (Busoniu et al., 2008). We believe that an agent in a multi-agent system might benefit from the knowledge of the policy space of its adversaries to understand what their action possibilities are and make decisions accordingly.

Appendix

A Proofs and derivations

In this appendix, we report the proofs and derivations of the results presented in the main paper.

A.1 Gaussian and Boltzmann linear policies as exponential family distributions

In this appendix, we show how a multivariate Gaussian with fixed covariance and a Boltzmann policy, both linear in the state features $\boldsymbol{\phi}(s)$ can be cast into Assumption 3. We are going to make use of the following identities regarding the Kronecker product (Petersen et al., 2008):

$$\operatorname{vec}(\boldsymbol{AXB}) = \left(\boldsymbol{B}^T \otimes \boldsymbol{A}\right)\operatorname{vec}(\boldsymbol{X}) \tag{12}$$

$$\boldsymbol{a}^T\boldsymbol{XBX}^T\boldsymbol{c} = \operatorname{vec}(\boldsymbol{X})^T\left(\boldsymbol{B} \otimes \boldsymbol{ca}^T\right)\operatorname{vec}(\boldsymbol{X}), \tag{13}$$

where $\operatorname{vec}(\boldsymbol{X})$ is the *vectorization* of matrix $\boldsymbol{X}$ obtained by stacking the columns of $\boldsymbol{X}$ into a single column vector.

A.1.1 Multivariate linear Gaussian policy with fixed covariance

The typical representation of a multivariate linear Gaussian policy is given by the following probability density function:

$$\pi_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{a}|s) = \frac{1}{(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{a} - \widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s))\right\},$$

where $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{k\times q}$ is a properly sized matrix. Recalling Assumption 3, we rephrase the previous equation as:

$$\pi_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{a}|s) = \frac{1}{(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\right\}\exp\left\{\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a} - \frac{1}{2}\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s)\right\}.$$

Recalling the identities at Eqs. (12) and (13) and observing that $\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}$ and $\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s)$ are scalar, we can rewrite:

$$\begin{aligned}
\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a} &= \operatorname{vec}\left(\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\right)\\
&= \left(\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)^T\right)\operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)\\
&= \operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)^T\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{a} \otimes \boldsymbol{\phi}(s)\right)\\
\boldsymbol{\phi}(s)^T\widetilde{\boldsymbol{\theta}}^T\boldsymbol{\Sigma}^{-1}\widetilde{\boldsymbol{\theta}}\boldsymbol{\phi}(s) &= \operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)^T\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right)\operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right).
\end{aligned}$$

Now, by redefining the parameter of the exponential family distribution $\boldsymbol{\theta} = \operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)$ we state the following definitions to comply with Assumption 3:

$$\begin{aligned}
&\boldsymbol{t}(s,\boldsymbol{a}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{a} \otimes \boldsymbol{\phi}(s)\\
&h(\boldsymbol{a}) = \frac{1}{(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{a}\right\}\\
&A(\boldsymbol{\theta},s) = \boldsymbol{\theta}^T\left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right)\boldsymbol{\theta}.
\end{aligned}$$

A.1.2 Boltzmann linear policy

The Boltzmann policy on a finite set of actions $\{a_1, ..., a_{k+1}\}$ is typically represented by means of a matrix of parameters $\widetilde{\boldsymbol{\theta}} \in \mathbb{R}^{k \times q}$:[14]

$$\pi_{\widetilde{\boldsymbol{\theta}}}(a_i|s) = \begin{cases} \dfrac{\exp\left\{\widetilde{\boldsymbol{\theta}}_i^T \boldsymbol{\phi}(s)\right\}}{1 + \sum_{j=1}^k \exp\left\{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)\right\}} & \text{if } i \leq k \\ \dfrac{1}{1 + \sum_{j=1}^k \exp\left\{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)\right\}} & \text{if } i = k+1 \end{cases},$$

where with $\widetilde{\boldsymbol{\theta}}_i$ we denote the i-th row of matrix $\widetilde{\boldsymbol{\theta}}$. In order to comply to Assumption 3, we rewrite the density function in the following form:

$$\pi_{\widetilde{\boldsymbol{\theta}}}(a_i|s) = \begin{cases} \exp\left\{\widetilde{\boldsymbol{\theta}}_i^T \boldsymbol{\phi}(s) - \log\left(\exp\{0\} + \sum_{j=1}^k \exp\left\{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)\right\}\right)\right\} & \text{if } i \leq k \\ \exp\left\{0 - \log\left(\exp\{0\} + \sum_{j=1}^k \exp\left\{\widetilde{\boldsymbol{\theta}}_j^T \boldsymbol{\phi}(s)\right\}\right)\right\} & \text{if } i = k+1 \end{cases}.$$

By introducing the vector $\boldsymbol{e}_i$ as the i–th vector of the canonical basis of $\mathbb{R}^k$, i.e., the vector having 1 in the i–th component and 0 elsewhere, and recalling the definition of Kronecker product, we can derive the following identity for $i \leq k$:

$$\widetilde{\boldsymbol{\theta}}_i^T \boldsymbol{\phi}(s) = \operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)^T (\boldsymbol{e}_i \otimes \boldsymbol{\phi}(s)).$$

In the case $i = k$ it is sufficient to replace the previous term with the zero vector $\boldsymbol{0}$. Therefore, by renaming $\boldsymbol{\theta} = \operatorname{vec}\left(\widetilde{\boldsymbol{\theta}}^T\right)$ we can make the following assignments in order to get the relevant quantities in Assumption 3:

$$\begin{aligned} \boldsymbol{t}(s, a_i) &= \begin{cases} \boldsymbol{e}_i \otimes \boldsymbol{\phi}(s) & \text{if } i \leq k \\ \boldsymbol{0} & \text{if } i = k+1 \end{cases} \\ h(a_i) &= 1 \\ A(\boldsymbol{\theta}, s) &= \log\left(1 + \sum_{j=1}^k \exp\left\{\boldsymbol{\theta}^T (\boldsymbol{e}_j \otimes \boldsymbol{\phi}(s))\right\}\right). \end{aligned}$$

[14] Notice that we are considering a set made of $k+1$ actions but the matrix $\widetilde{\boldsymbol{\theta}}$ has only k rows. This allows enforcing the identifiability property, otherwise if we had a row for each of the $k+1$ actions we would have multiple representation for the same policy (rescaling the rows by the same amount).

A.2 Results on exponential family

In this appendix, we derive several results that are used in Section 4, concerning policies belonging to the exponential family, as in Assumption 3.

A.2.1 Fisher information matrix

We start by providing an expression of the Fisher Information matrix (FIM) for the specific case of the exponential family, that we are going to use extensively in the derivation. We first define the FIM for a fixed state and then we provide its expectation under the state distribution ν. For any state $s \in \mathcal{S}$, we define the FIM induced by $\pi_{\boldsymbol{\theta}}(\cdot|s)$ as:

$$\mathcal{F}(\boldsymbol{\theta}, s) = \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s)^T \right]. \tag{14}$$

We can derive the following immediate result.

Lemma 3 *For a policy $\pi_{\boldsymbol{\theta}}$ belonging to the exponential family, as in Assumption 3, the FIM for state $s \in \mathcal{S}$ is given by the covariance matrix of the sufficient statistic:*

$$\mathcal{F}(\boldsymbol{\theta}, s) = \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta}) \bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})^T \right] = \mathop{\mathbb{C}\mathrm{ov}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{t}(s, a)].$$

Proof Let us first compute the gradient log-policy for the exponential family:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a|s) &= \boldsymbol{t}(s, a) - \nabla_{\boldsymbol{\theta}} A(\boldsymbol{\theta}, s) \\ &= \boldsymbol{t}(s, a) - \frac{\int_{\mathcal{A}} \boldsymbol{t}(s, \bar{a}) h(\bar{a}) \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{t}(s, \bar{a}) \right\} \mathrm{d}\bar{a}}{\int_{\mathcal{A}} h(\bar{a}) \exp\left\{ \boldsymbol{\theta}^T \boldsymbol{t}(s, \bar{a}) \right\} \mathrm{d}\bar{a}} \\ &= \boldsymbol{t}(s, a) - \mathrm{E}_{\bar{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{t}(s, \bar{a})] = \bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta}). \end{aligned} \tag{P.4}$$

Now, we just need to apply the definition given in Eq. (14) and to recall the definition of covariance matrix:

$$\begin{aligned} \mathcal{F}(\boldsymbol{\theta}, s) &= \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta}) \bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})^T \right] \\ &= \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\left(\boldsymbol{t}(s, a) - \mathop{\mathbb{E}}_{\bar{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{t}(s, \bar{a})] \right) \left(\boldsymbol{t}(s, a) - \mathop{\mathbb{E}}_{\bar{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{t}(s, \bar{a})] \right)^T \right] \\ &= \mathop{\mathbb{C}\mathrm{ov}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} [\boldsymbol{t}(s, a)]. \end{aligned}$$

□

We now define the expected FIM $\mathcal{F}(\boldsymbol{\theta})$ and its corresponding estimator $\widehat{\mathcal{F}}(\boldsymbol{\theta})$ under the γ–discounted stationary distribution induced by the agent's policy π^{Ag}:

$$\mathcal{F}(\boldsymbol{\theta}) = \mathop{\mathbb{E}}_{s\sim\nu}\left[\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\bar{\boldsymbol{t}}(s,a)\bar{\boldsymbol{t}}(s,a)^T\right]\right], \qquad \widehat{\mathcal{F}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\bar{\boldsymbol{t}}(s_i,a)\bar{\boldsymbol{t}}(s_i,a)^T\right]. \tag{15}$$

Finally, we provide a sufficient condition to ensure that the FIM $\mathcal{F}(\boldsymbol{\theta})$ is non singular in the case of Gaussian and Boltzmann linear policies.

Proposition 1 *If the second moment matrix of the feature vector* $\mathop{\mathbb{E}}_{s\sim\nu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right]$ *is non–singular, the identifiability condition of Lemma 2 is fulfilled by the Gaussian and Boltzmann linear policies for all* $\boldsymbol{\theta}\in\Theta$, *provided that each action is played with non–zero probability for the Boltzmann policy.*

Proof Let us start with the Boltzmann policy and consider the expression of $\bar{\boldsymbol{t}}(s,a_i)$ with $i\in\{1,...,k\}$:

$$\begin{aligned}
\bar{\boldsymbol{t}}(s,a_i,\boldsymbol{\theta}) &= \boldsymbol{t}(s,a_i) - \mathop{\mathbb{E}}_{\overline{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s,\overline{a})] \\
&= \boldsymbol{e}_i \otimes \boldsymbol{\phi}(s) - \sum_{j=1}^{k} \pi_{\boldsymbol{\theta}}(a_i|s)\boldsymbol{e}_i \otimes \boldsymbol{\phi}(s) \\
&= (\boldsymbol{e}_i - \boldsymbol{\pi}) \otimes \boldsymbol{\phi}(s),
\end{aligned}$$

where $\boldsymbol{\pi}$ is a vector defined as $\boldsymbol{\pi} = (\pi_{\boldsymbol{\theta}}(a_1|s),...,\pi_{\boldsymbol{\theta}}(a_k|s))^T$ and we exploited the distributivity of the Kronecker product. While for $i=k+1$, we have $(\boldsymbol{0}-\boldsymbol{\pi})\otimes\boldsymbol{\phi}(s)$. For the sake of the proof, let us define $\widetilde{\boldsymbol{e}}_i = \boldsymbol{e}_i$ if $i \le k$ and $\widetilde{\boldsymbol{e}}_{k+1} = \boldsymbol{0}$. Let us compute the FIM:

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\theta}) &= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\bar{\boldsymbol{t}}(s,a,\boldsymbol{\theta})\bar{\boldsymbol{t}}(s,a,\boldsymbol{\theta})^T\right] \\
&= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[((\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})\otimes\boldsymbol{\phi}(s))((\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})\otimes\boldsymbol{\phi}(s))^T\right] \\
&= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[(\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})(\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})^T \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right] \\
&= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[(\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})(\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi})^T\right] \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T \\
&= \left(\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\widetilde{\boldsymbol{e}}_i\widetilde{\boldsymbol{e}}_i^T\right] - \boldsymbol{\pi}\boldsymbol{\pi}^T\right) \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T \left(\operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T\right) \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T,
\end{aligned}$$

where we exploited the distributivity of the Kroneker product, observed that $\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\widetilde{\boldsymbol{e}}_i] = \boldsymbol{\pi}$ and $\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\widetilde{\boldsymbol{e}}_i\widetilde{\boldsymbol{e}}_i^T] = \operatorname{diag}(\boldsymbol{\pi})$. Let us now consider the matrix:

$$\operatorname{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^T = \begin{pmatrix}
\pi_{\boldsymbol{\theta}}(a_1|s) - \pi_{\boldsymbol{\theta}}(a_1|s)^2 & -\pi_{\boldsymbol{\theta}}(a_1|s)\pi_{\boldsymbol{\theta}}(a_2|s) & \dots & -\pi_{\boldsymbol{\theta}}(a_1|s)\pi_{\boldsymbol{\theta}}(a_k|s) \\
-\pi_{\boldsymbol{\theta}}(a_1|s)\pi_{\boldsymbol{\theta}}(a_2|s) & \pi_{\boldsymbol{\theta}}(a_2|s) - \pi_{\boldsymbol{\theta}}(a_2|s)^2 & \dots & -\pi_{\boldsymbol{\theta}}(a_2|s)\pi_{\boldsymbol{\theta}}(a_k|s) \\
\vdots & \vdots & \ddots & \vdots \\
-\pi_{\boldsymbol{\theta}}(a_1|s)\pi_{\boldsymbol{\theta}}(a_k|s) & -\pi_{\boldsymbol{\theta}}(a_2|s)\pi_{\boldsymbol{\theta}}(a_k|s) & \dots & \pi_{\boldsymbol{\theta}}(a_k|s) - \pi_{\boldsymbol{\theta}}(a_k|s)^2
\end{pmatrix}.$$

Consider a generic row $i\in\{1,...,k\}$. The element on the diagonal is $\pi_{\boldsymbol{\theta}}(a_i|s) - \pi_{\boldsymbol{\theta}}(a_i|s)^2 = \pi_{\boldsymbol{\theta}}(a_i|s)(1-\pi_{\boldsymbol{\theta}}(a_i|s))$, while the absolute sum of the elements out of the diagonal is:

$$\pi_{\boldsymbol{\theta}}(a_i|s) \sum_{j\in\{1,\dots k\}\wedge j\neq i} \pi_{\boldsymbol{\theta}}(a_j|s) = \pi_{\boldsymbol{\theta}}(a_i|s)(1 - \pi_{\boldsymbol{\theta}}(a_i|s) - \pi_{\boldsymbol{\theta}}(a_{k+1}|s)).$$

Therefore, if all actions are played with non–zero probability, i.e., $\pi_{\boldsymbol{\theta}}(a_i|s) > 0$ for all $i \in \{1, ..., k+1\}$ it follows that the matrix is strictly diagonally dominant by rows and thus it is positive definite. If also $\mathop{\mathbb{E}}_{s\sim \nu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right]$ is positive definite, for the properties of the Kroneker product, the FIM is positive definite.

Let us now focus on the Gaussian policy. Let $\boldsymbol{a} \in \mathbb{R}^d$ and denote $\boldsymbol{\mu}(s) = \mathbb{E}_{\boldsymbol{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{a}]$:

$$\bar{\boldsymbol{t}}(s, \boldsymbol{a}, \boldsymbol{\theta}) = \boldsymbol{t}(s, \boldsymbol{a}) - \mathop{\mathbb{E}}_{\bar{\boldsymbol{a}}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s, \bar{\boldsymbol{a}})] = \boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s).$$

Let us compute the FIM:

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\theta}) &= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})^T\right] \\
&= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s)\right)\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s)\right)^T\right] \\
&= \mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))(\boldsymbol{a} - \boldsymbol{\mu}(s))^T\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right] \\
&= \boldsymbol{\Sigma}^{-1}\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[(\boldsymbol{a} - \boldsymbol{\mu}(s))(\boldsymbol{a} - \boldsymbol{\mu}(s))^T\right]\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T \\
&= \boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T = \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T.
\end{aligned}$$

If $\boldsymbol{\Sigma}$ has finite values, then $\boldsymbol{\Sigma}^{-1}$ will be positive definite and additionally, considering that $\mathop{\mathbb{E}}_{s\sim\nu}\left[\boldsymbol{\phi}(s)\boldsymbol{\phi}(s)^T\right]$ is positive definite, we have that the FIM is positive definite. □

A.2.2 Subgaussianity assumption

From Assumption 4, we can prove the following result that upper bounds the maximum eigenvalue $\lambda_{\max}$ of the Fisher information matrix with the subgaussianity parameter σ.

Lemma 4 *Under Assumption 4, for any $\boldsymbol{\theta} \in \Theta$ and for any $s \in \mathcal{S}$ the maximum eigenvalue of the Fisher Information matrix $\mathcal{F}(\boldsymbol{\theta}, s)$ is upper bounded by $d\sigma^2$.*

Proof Recall that the maximum eigenvalue of a matrix $\boldsymbol{A}$ can be computed as $\sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2\leq 1}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$ and the norm of a vector $\boldsymbol{y}$ can be computed as $\sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2\leq 1}\boldsymbol{x}^T\boldsymbol{y}$. Consider now the derivation for a generic $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\|_2 \leq 1$:

$$\begin{aligned}
\boldsymbol{x}^T \mathcal{F}(\boldsymbol{\theta}, s)\boldsymbol{x} &= \boldsymbol{x}^T \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})^T\right]\boldsymbol{x} \\
&= \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\boldsymbol{x}^T\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})^T\boldsymbol{x}\right] \\
&= \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\left(\boldsymbol{x}^T\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\right)^2\right] \\
&\leq \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\left(\sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2 \leq 1} \boldsymbol{x}^T\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\right)^2\right] = \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\|\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\|_2^2\right],
\end{aligned}$$

where we employed Lemma 3 and upper bounded the right hand side. By taking the supremum over $\boldsymbol{x} \in \mathbb{R}^d$ such that $\|\boldsymbol{x}\|_2 \leq 1$ we get:

$$\lambda_{\max}(\mathcal{F}(\boldsymbol{\theta}, s)) = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|_2 \leq 1} \boldsymbol{x}^T\mathcal{F}(\boldsymbol{\theta}, s)\boldsymbol{x} \leq \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\|\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\|_2^2\right]. \tag{P.5}$$

By applying the first inequality in Remark 2.2 of Hsu et al., (2012) and setting $\boldsymbol{A} = \boldsymbol{I}$ we get that $\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)} \left[\|\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\|_2^2\right] \leq d\sigma^2$. □

We now show that the subgaussianity assumption is satisfied by the Boltzmann and Gaussian policies, as defined in Table 1, under the following assumption.

Assumption 6 (Boundedness of Features) For any $s \in \mathcal{S}$ the feature function is bounded in L_2-norm, i.e., there exists $\Phi_{\max} < \infty$ such that $\|\boldsymbol{\phi}(s)\|_2 \leq \Phi_{\max}$.

Proposition 2 *Under Assumption 6, then Assumption 4 is fulfilled by the Boltzmann linear policy with parameter* $\sigma = 2\Phi_{\max}$ *and Gaussian linear policy with parameter* $\sigma = \frac{\Phi_{\max}}{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}$.

Proof Let us start with the Boltzmann policy. From the definition of subgaussianity given in Assumption 4, requiring that the random vector $\bar{\boldsymbol{t}}(s, a_i, \boldsymbol{\theta})$ is subgaussian with parameter σ is equivalent to require that the random (scalar) variable $\frac{1}{\|\boldsymbol{\alpha}\|_2}\boldsymbol{\alpha}^T\bar{\boldsymbol{t}}(s, a_i, \boldsymbol{\theta})$ is subgaussian with parameter σ for any $\boldsymbol{\alpha} \in \mathbb{R}^d$. Thus, we now bound the term:

$$\begin{aligned}
\left|\boldsymbol{\alpha}^T\bar{\boldsymbol{t}}(s, a, \boldsymbol{\theta})\right| &= \left|\boldsymbol{\alpha}^T\left((\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi}) \otimes \boldsymbol{\phi}(s)\right)\right| \\
&= \|\boldsymbol{\alpha}\|_2\|(\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi}) \otimes \boldsymbol{\phi}(s)\|_2 \\
&= \|\boldsymbol{\alpha}\|_2\|\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi}\|_2\|\boldsymbol{\phi}(s)\|_2 \\
&\leq 2\|\boldsymbol{\alpha}\|_2\Phi_{\max},
\end{aligned}$$

where we used Cauchy–Swartz inequality, the identity $\|\boldsymbol{x} \otimes \boldsymbol{y}\|_2^2 = (\boldsymbol{x} \otimes \boldsymbol{y})^T(\boldsymbol{x} \otimes \boldsymbol{y}) = (\boldsymbol{x}^T\boldsymbol{x}) \otimes (\boldsymbol{y}^T\boldsymbol{y}) = \|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2$ and the inequality $\|\widetilde{\boldsymbol{e}}_i - \boldsymbol{\pi}\|_2^2 \leq 2$. Therefore, we have that the random variable $\frac{1}{\|\boldsymbol{\alpha}\|_2}\boldsymbol{\alpha}^T\bar{\boldsymbol{t}}(s, a_i, \boldsymbol{\theta}) \leq 2\Phi_{\max}$ is bounded. Thanks to Hoeffding's lemma we have that the subgaussianity parameter is $\sigma = 2\Phi_{\max}$.

Let us now consider the Gaussian policy. Let $\boldsymbol{a} \in \mathbb{R}^d$ and denote with $\boldsymbol{\mu}(s) = \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{a}]$:

$$\bar{\boldsymbol{t}}(s,\boldsymbol{a},\boldsymbol{\theta}) = \boldsymbol{t}(s,\boldsymbol{a}) - \mathop{\mathbb{E}}_{\bar{\boldsymbol{a}} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s,\bar{\boldsymbol{a}})] = \boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s).$$

Let us first observe that we can rewrite:

$$\begin{aligned}
\boldsymbol{\alpha}^T\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s)\right) &= \sum_{i=1}^{k}\sum_{j=1}^{q} \alpha_{ij}\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))\right)_i \phi(s)_j \\
&= \sum_{i=1}^{k}\sum_{j=1}^{q} \alpha_{ij}\phi(s)_j\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))\right)_i \\
&= \boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)),
\end{aligned}$$

where $\beta_i = \sum_j \alpha_{ij}\phi(s)_j$ for $i \in \{1, ..., k\}$. We now proceed with explicit computations:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left\{\boldsymbol{\alpha}^T\bar{\boldsymbol{t}}(s,\boldsymbol{a},\boldsymbol{\theta})\right\}\right] &= \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left\{\boldsymbol{\alpha}^T\left(\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \otimes \boldsymbol{\phi}(s)\right)\right\}\right] \\
&= \mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left\{\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))\right\}\right] \\
&= \int_{\mathbb{R}^d} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}(s))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))\right\}}{(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\exp\left\{\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s))\right\}\mathrm{d}\boldsymbol{a}.
\end{aligned}$$

Now we complete the square:

$$\begin{aligned}
&-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}(s))^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) + \boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s)) \\
&\qquad = -\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}(s) - \boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s) - \boldsymbol{\beta}) + \frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}.
\end{aligned}$$

Thus, we have:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\boldsymbol{a} \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\exp\left\{\boldsymbol{\alpha}^T\bar{\boldsymbol{t}}(s,\boldsymbol{a},\boldsymbol{\theta})\right\}\right] \\
&\quad = \exp\left\{\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right\}\int_{\mathbb{R}^d} \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{a} - \boldsymbol{\mu}(s) - \boldsymbol{\beta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{a} - \boldsymbol{\mu}(s) - \boldsymbol{\beta})\right\}}{(2\pi)^{\frac{k}{2}}\det(\boldsymbol{\Sigma})^{\frac{1}{2}}}\mathrm{d}\boldsymbol{a} \\
&\quad = \exp\left\{\frac{1}{2}\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}\right\}.
\end{aligned}$$

Now, we observe that:

$$\boldsymbol{\beta}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} \leq \|\boldsymbol{\beta}\|_2^2\left\|\boldsymbol{\Sigma}^{-1}\right\|_2 \leq \|\boldsymbol{\alpha}\|_2^2\|\boldsymbol{\phi}(s)\|_2^2\left\|\boldsymbol{\Sigma}^{-1}\right\|_2,$$

having derived from Cauchy–Swartz inequality:

$$\begin{aligned}\|\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{k}\left(\sum_{j=1}^{q}\alpha_{ij}\phi(s)_j\right)^2 &\leq \sum_{i=1}^{k}\sum_{j=1}^{q}\alpha_{ij}^2\sum_{l=1}^{q}\phi(s)_l^2\\ &= \left(\sum_{i=1}^{k}\sum_{j=1}^{q}\alpha_{ij}^2\right)\sum_{l=1}^{q}\phi(s)_l^2\\ &= \|\boldsymbol{\alpha}\|_2^2\|\boldsymbol{\phi}(s)\|_2^2.\end{aligned}$$

We get the result by setting $\sigma = \Phi_{\max}\sqrt{\left\|\boldsymbol{\Sigma}^{-1}\right\|_2} = \frac{\Phi_{\max}}{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}$. □

Furthermore, we report for completeness the standard Hoeffding concentration inequality for subgaussian random vectors.

Proposition 3 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_n$ be n i.i.d. zero–mean subgaussian d–dimensional random vectors with parameter $\sigma \geq 0$, then for any $\boldsymbol{\alpha} \in \mathbb{R}^d$ and $\epsilon > 0$ it holds that:*

$$\Pr\left(\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right) \geq \epsilon\right) \leq \exp\left\{-\frac{\epsilon^2 n}{2\|\boldsymbol{\alpha}\|_2^2\sigma^2}\right\}.$$

Proof The proof is analogous to that of the Hoeffding inequality for bounded random variables. Let $s \geq 0$:

$$\begin{aligned}&\Pr\left(\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right) \geq \epsilon\right) = \Pr\left(\exp\left\{s\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right)\right\} \quad \geq e^{s\epsilon}\right)\\ &\leq e^{-s\epsilon}\,\mathrm{E}\left[\exp\left\{s\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right)\right\}\right] = e^{-s\epsilon}\prod_{i=1}^{n}\mathrm{E}\left[\exp\left\{\frac{s}{n}\boldsymbol{\alpha}^T\boldsymbol{X}_i\right\}\right] \leq e^{-s\epsilon}\exp\left\{\frac{s^2}{2n}\|\boldsymbol{\alpha}\|_2^2\sigma^2\right\}\\ &\quad = \exp\left\{-s\epsilon + \frac{s^2}{2n}\|\boldsymbol{\alpha}\|_2^2\sigma^2\right\},\end{aligned}$$

where we employed Markov inequality, exploited the subgaussianity assumption and the independence. We minimize the last expression over s, getting the optimal $s = \frac{\epsilon n}{\|\boldsymbol{\alpha}\|_2^2\sigma^2}$, from which we get the result:

$$\Pr\left(\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\right) \geq \epsilon\right) \leq \exp\left\{-\frac{\epsilon^2 n}{2\|\boldsymbol{\alpha}\|_2^2\sigma^2}\right\}.$$

□

Under the Assumption 4, we provide the following concentration inequality for the minimum eigenvalue of the empirical FIM.

Proposition 4 *Let $\mathcal{F}(\boldsymbol{\theta})$ and $\widehat{\mathcal{F}}(\boldsymbol{\theta})$ be the FIM and its estimate obtained with $n > 0$ independent samples. Then, under Assumption 4,* for any $\epsilon > 0$ *it holds that:*

$$\Pr\left(\left|\lambda_{\min}\left(\widehat{\mathcal{F}}(\boldsymbol{\theta})\right) - \lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))\right| > \epsilon\right) \leq 2\exp\left\{-\frac{\epsilon^2 n}{\psi_\sigma d^2 \sigma^4}\right\},$$

where $\psi_\sigma > 0$ is a constant depending only on the subgaussianity parameter σ. In particular, under the following condition on n we have that, for any $\delta \in [0,1]$ and $\alpha \in [0,1)$ it holds that $\lambda_{\min}(\widehat{\mathcal{F}}(\boldsymbol{\theta})) > \alpha\lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))$ with probability at least $1-\delta$:

$$n > \frac{d^2\sigma^4\psi_\sigma \log\frac{2}{\delta}}{(1-\alpha)^2\lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))^2}.$$

Proof Let us recall that $\widehat{\mathcal{F}}(\boldsymbol{\theta})$ and $\mathcal{F}(\boldsymbol{\theta})$ are both symmetric positive semidefinite matrices, thus their eigenvalues λ_j correspond to their singular values σ_j. Let us consider the following sequence of inequalities:

$$\begin{aligned}\left|\lambda_{\min}\left(\widehat{\mathcal{F}}(\boldsymbol{\theta})\right) - \lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))\right| &= \left|\sigma_{\min}\left(\widehat{\mathcal{F}}(\boldsymbol{\theta})\right) - \sigma_{\min}(\mathcal{F}(\boldsymbol{\theta}))\right| \\ &\leq \max_{j\in\{1,\ldots,d\}}\left|\sigma_j\left(\widehat{\mathcal{F}}(\boldsymbol{\theta})\right) - \sigma_j(\mathcal{F}(\boldsymbol{\theta}))\right| \\ &\leq \left\|\widehat{\mathcal{F}}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta})\right\|_2,\end{aligned}$$

where last inequality follows from Ben-Israel and Greville, (2003). Therefore, all it takes is to bound the norm of the difference. For this purpose, we employ Corollary 5.50 and Remark 5.51 of Vershynin, (2012), having observed that the FIM is indeed a covariance matrix and its estimate is a sample covariance matrix. We obtain that with probability at least $1-\delta$:

$$\left\|\widehat{\mathcal{F}}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta})\right\|_2 \leq \|\mathcal{F}(\boldsymbol{\theta})\|_2\sqrt{\frac{\psi_\sigma \log\frac{2}{\delta}}{n}}, \tag{P.6}$$

where $\psi_\sigma \geq 0$ is a constant depending on the subgaussianity parameter σ. Recalling, from Lemma 4, that $\|\mathcal{F}(\boldsymbol{\theta})\| = \lambda_{\max}(\mathcal{F}(\boldsymbol{\theta})) \leq d\sigma^2$, we can rewrite the previous inequality as:

$$\left\|\widehat{\mathcal{F}}(\boldsymbol{\theta}) - \mathcal{F}(\boldsymbol{\theta})\right\|_2 \leq d\sigma^2\sqrt{\frac{\psi_\sigma \log\frac{2}{\delta}}{n}}. \tag{P.7}$$

By setting the right hand side equal to ϵ and solving for δ, we get the first result. The value of n can be obtained by setting the right hand side equal to $(1-\alpha)\lambda_{\min}(\mathcal{F}(\boldsymbol{\theta}))$. □

A.2.3 Concentration result

We are now ready to provide the main result of this section, that consists in a concentration result on the negative log–likelihood. Our final goal is to provide a probabilistic bound to the differences $\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})$ and $\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}})$. To this purpose, we start with a technical lemma (Lemma 5) which provides a concentration result involving a quantity that will be used later, under Assumption 4. Then, we use this result to obtain the concentration of the parameters, i.e., bounding the distance $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2$ (Theorem 6), under suitable well–conditioning properties of the involved quantities. Finally, we employ the latter result to prove the concentration of the negative log–likelihood (Corollary 1). Some parts of the derivation are inspired to Li et al., (2017).

Lemma 5 *Under Assumption 2 and Assumption 4, let $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$ be a dataset of $n > 0$ independent samples, where $s_i \sim d_\mu^{\pi_{\boldsymbol{\theta}Ag}}$ and $a_i \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s_i)$. For any $\boldsymbol{\theta} \in \Theta$, let $\boldsymbol{g}(\boldsymbol{\theta})$ be defined as:*

$$\boldsymbol{g}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s_i, a)] - \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s)}[\boldsymbol{t}(s_i, a)]\right). \tag{16}$$

Let $\widehat{\boldsymbol{\theta}} = \mathop{\arg\min}_{\boldsymbol{\theta} \in \Theta} \widehat{\ell}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \log \pi_{\boldsymbol{\theta}}(a_i|s_i)$. Then, under Assumption 4, for any $\delta \in [0, 1]$, with probability at least $1 - \delta$, it holds that:

$$\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right\|_2 \leq \sigma\sqrt{\frac{2d}{n}\log\frac{2d}{\delta}}. \tag{17}$$

Proof The negative log–likelihood of a policy complying with Assumption 3 is $\mathcal{C}^2(\mathbb{R}^d)$. Thus, since $\widehat{\boldsymbol{\theta}}$ is a minimizer of the negative log–likelihood function $\widehat{\ell}(\boldsymbol{\theta})$, it must fulfill the following first–order condition:

$$\nabla_{\boldsymbol{\theta}}\widehat{\ell}(\widehat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n}\nabla_{\boldsymbol{\theta}}\log\pi_{\widehat{\boldsymbol{\theta}}}(a_i|s_i) = \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{t}(s_i, a_i) - \mathop{\mathbb{E}}_{a \sim \pi_{\widehat{\boldsymbol{\theta}}}(\cdot|s)}[\boldsymbol{t}(s_i, a)]\right) = \boldsymbol{0}. \tag{P.8}$$

As a consequence, we can rewrite the expression of $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ exploiting this condition:

$$\begin{aligned}\boldsymbol{g}(\widehat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathop{\mathbb{E}}_{a \sim \pi_{\widehat{\boldsymbol{\theta}}}(\cdot|s)}[\boldsymbol{t}(s_i, a)] - \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s)}[\boldsymbol{t}(s_i, a)]\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{t}(s_i, a_i) - \mathop{\mathbb{E}}_{a \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s)}[\boldsymbol{t}(s_i, a)]\right)\\ &= \frac{1}{n}\sum_{i=1}^{n}\overline{\boldsymbol{t}}(s_i, a_i, \boldsymbol{\theta}^{\mathrm{Ag}}).\end{aligned}$$

By recalling that $a_i \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s_i)$ it immediately follows that $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ is a zero-mean random vector, i.e., $\mathop{\mathbb{E}}_{\substack{s_i \sim \nu \\ a_i \sim \pi_{\boldsymbol{\theta}Ag}(\cdot|s_i)}}\left[\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right] = \boldsymbol{0}$. Moreover, under Assumption 4, $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ is the sample mean of subgaussian random vectors. Our goal is to bound the probability $\Pr\left(\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right\|_2 > \epsilon\right)$; to this purpose we consider the following derivation:

$$\Pr\left(\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right\|_2 > \epsilon\right) = \Pr\left(\sqrt{\sum_{j=1}^{d} g_j(\widehat{\boldsymbol{\theta}})^2} > \epsilon\right) \le \Pr\left(\bigvee_{j=1}^{d} \left|g_j(\widehat{\boldsymbol{\theta}})\right| > \frac{\epsilon}{\sqrt{d}}\right) \tag{P.9}$$

$$\le \sum_{j=1}^{d} \Pr\left(\left|g_j(\widehat{\boldsymbol{\theta}})\right| > \frac{\epsilon}{\sqrt{d}}\right), \tag{P.10}$$

where we exploited in line (P.9) the fact that for a d-dimensional vector $\boldsymbol{x}$ if $\|\boldsymbol{x}\|_2 > \epsilon$ it must be that at least one component $j = 1, ..., d$ satisfy $x_j^2 > \frac{\epsilon^2}{d}$ and we used a union bound over the d dimensions to get line (P.10). Since for each $j = 1, ..., d$ we have that $g_j(\widehat{\boldsymbol{\theta}})$ is a zero-mean subgaussian random variable we can bound the deviation using standard results (Boucheron et al., 2013):

$$\Pr\left(\left|g_j(\widehat{\boldsymbol{\theta}})\right| > \frac{\epsilon}{\sqrt{d}}\right) \le 2\exp\left\{-\frac{\epsilon^2 n}{2d\sigma^2}\right\}. \tag{P.11}$$

Putting all together we get:

$$\Pr\left(\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right\|_2 > \epsilon\right) \le 2d\exp\left\{-\frac{\epsilon^2 n}{2d\sigma^2}\right\}. \tag{P.12}$$

By setting $\delta = 2d\exp\left\{-\frac{\epsilon^2 n}{2d\sigma^2}\right\}$ and solving for ϵ we get the result. □

We can now use the previous result to derive the concentration of the parameters, i.e., bounding the deviation $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2$.

Theorem 6 (Parameter concentration) *Under Assumption 2 and Assumption 4, let* $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^{n}$ *be a dataset of* $n > 0$ *independent samples, where* $s_i \sim \nu$ *and* $a_i \sim \pi_{\boldsymbol{\theta}^{Ag}}(\cdot|s_i)$. *Let* $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \widehat{\ell}(\boldsymbol{\theta})$. *If the empirical FIM* $\widehat{\mathcal{F}}(\boldsymbol{\theta})$ *has a positive minimum eigenvalue* $\widehat{\lambda}_{\min} > 0$ *for all* $\boldsymbol{\theta} \in \Theta$, *for any* $\delta \in [0, 1]$, *with probability at least* $1 - \delta$, *it holds that*:

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2 \le \frac{\sigma}{\widehat{\lambda}_{\min}}\sqrt{\frac{2d}{n}\log\frac{2d}{\delta}}. \tag{18}$$

Proof Recalling that $\boldsymbol{g}(\boldsymbol{\theta}^{\mathrm{Ag}}) = \boldsymbol{0}$, we employ the mean value theorem to rewrite $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ centered in $\boldsymbol{\theta}^{\mathrm{Ag}}$:

$$\boldsymbol{g}(\widehat{\boldsymbol{\theta}}) = \boldsymbol{g}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{g}(\boldsymbol{\theta}^{\mathrm{Ag}}) = \widehat{\mathcal{F}}(\overline{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right), \tag{P.13}$$

where $\overline{\boldsymbol{\theta}} = t\widehat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta}^{\mathrm{Ag}}$ for some $t \in [0,1]$ and $\widehat{\mathcal{F}}(\overline{\boldsymbol{\theta}})$ is defined as:

$$\begin{aligned}\widehat{\mathcal{F}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{g}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)\boldsymbol{t}(s_i,a)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{a\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}\left[\left(\boldsymbol{t}(s_i,a) - \mathop{\mathbb{E}}_{\overline{a}\sim\pi_{\boldsymbol{\theta}}(\cdot|s)}[\boldsymbol{t}(s_i,\overline{a})]\right)\boldsymbol{t}(s_i,a)\right] = \widehat{\mathcal{F}}(\boldsymbol{\theta}),\end{aligned}$$

where we exploited the expression of $\nabla_{\boldsymbol{\theta}}\log\pi_{\boldsymbol{\theta}}(a|s)$ and the definition of Fisher information matrix given in Eq. (14). Under the hypothesis of the statement, we can derive the following lower bound:

$$\left\|\boldsymbol{g}(\widehat{\boldsymbol{\theta}})\right\|_2^2 = \left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right)^T\widehat{\mathcal{F}}(\overline{\boldsymbol{\theta}})^T\widehat{\mathcal{F}}(\overline{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right) \geq \widehat{\lambda}_{\min}^2\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2^2. \tag{P.14}$$

By solving for $\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2$ and applying Lemma 5 we get the result. □

Finally, we can get the concentration result for the negative log–likelihood.

Corollary 1 (Negative log–likelihood concentration) *Under Assumption 2 and Assumption 4, let* $\mathcal{D} = \{(s_i,a_i)\}_{i=1}^n$ *be a dataset of* $n > 0$ *independent samples, where* $s_i \sim \nu$ *and* $a_i \sim \pi_{\boldsymbol{\theta}^{Ag}}(\cdot|s_i)$. *Let* $\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta}\widehat{\ell}(\boldsymbol{\theta})$. *If* $\lambda_{\min}(\widehat{\mathcal{F}}(\boldsymbol{\theta})) = \widehat{\lambda}_{\min} > 0$ *for all* $\boldsymbol{\theta} \in \Theta$, *for any* $\delta \in [0,1]$, *with probability at least* $1-\delta$, *it holds that*:

$$\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) \leq \frac{d^2\sigma^4}{\widehat{\lambda}_{\min}^2 n}\log\frac{2d}{\delta}, \tag{19}$$

and also:

$$\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) \leq \frac{d^2\sigma^4}{\widehat{\lambda}_{\min}^2 n}\log\frac{2d}{\delta}. \tag{20}$$

Proof Let us start with $\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})$. We consider the first order Taylor expansion of the negative log–likelihood centered in $\boldsymbol{\theta}^{\mathrm{Ag}}$:

$$\begin{aligned}&\ell(\widehat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) \\ &\quad = \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{\mathrm{Ag}})^T\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right) + \frac{1}{2}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right)^T\mathcal{H}_{\boldsymbol{\theta}}\ell(\overline{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right),\end{aligned} \tag{P.15}$$

where $\overline{\boldsymbol{\theta}} = t\widehat{\boldsymbol{\theta}} + (1-t)\boldsymbol{\theta}^{\mathrm{Ag}}$ for some $t \in [0,1]$. We first observe that $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{\mathrm{Ag}}) = \boldsymbol{0}$ being $\boldsymbol{\theta}^{\mathrm{Ag}}$ the true parameter and we develop $\mathcal{H}_{\boldsymbol{\theta}}\ell(\overline{\boldsymbol{\theta}})$:

$$\begin{aligned}
\mathcal{H}_{\boldsymbol{\theta}}\ell(\overline{\boldsymbol{\theta}}) &= \underset{\substack{s\sim\nu \\ a\sim\pi_{\boldsymbol{\theta}Ag}(\cdot|s)}}{\mathbb{E}}\left[\mathcal{H}_{\boldsymbol{\theta}}\log\pi_{\overline{\boldsymbol{\theta}}}(a|s)\right] \\
&= \underset{\substack{s\sim\nu \\ a\sim\pi_{\boldsymbol{\theta}Ag}(\cdot|s)}}{\mathbb{E}}\left[\nabla_{\boldsymbol{\theta}}\left(\boldsymbol{t}(s,a)-\underset{\overline{a}\sim\pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}{\mathbb{E}}[\boldsymbol{t}(s,\overline{a})]\right)\right] \\
&= \underset{s\sim\nu}{\mathbb{E}}\left[\nabla_{\boldsymbol{\theta}}\underset{\overline{a}\sim\pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}{\mathbb{E}}[\boldsymbol{t}(s,\overline{a})]\right] \\
&= \underset{s\sim\nu}{\mathbb{E}}\left[\underset{\overline{a}\sim\pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}{\mathbb{E}}\left[\left(\boldsymbol{t}(s,\overline{a})-\underset{\widetilde{a}\sim\pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}{\mathbb{E}}[\boldsymbol{t}(s,\widetilde{a})]\right)\boldsymbol{t}(s,\overline{a})^T\right]\right] = \underset{s\sim\nu}{\mathbb{E}}\left[\mathcal{F}(\overline{\boldsymbol{\theta}},s)\right].
\end{aligned}$$

By using Lemma 4 to bound the maximum eigenvalue of $\mathcal{F}(\overline{\boldsymbol{\theta}},s)$, we can state the inequality:

$$\frac{1}{2}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{\mathrm{Ag}}\right)^T\mathcal{H}_{\boldsymbol{\theta}}\ell(\overline{\boldsymbol{\theta}})\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{\mathrm{Ag}}\right)\le\frac{d\sigma^2}{2}\left\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2^2. \tag{P.16}$$

Using the concentration result of Theorem 6, we get the result. Concerning $\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}})-\widehat{\ell}(\widehat{\boldsymbol{\theta}})$, the derivation is analogous with the only difference that the Taylor expansion has to be centered in $\widehat{\boldsymbol{\theta}}$ instead of $\boldsymbol{\theta}^{\mathrm{Ag}}$. □

To conclude this appendix, we present the following technical lemma.

Theorem 7 *Under Assumption 2 and Assumption 4, let $\mathcal{D}=\{(s_i,a_i)\}_{i=1}^n$ be a dataset of $n>0$ independent samples, where $s_i\sim\nu$ and $a_i\sim\pi_{\boldsymbol{\theta}Ag}(\cdot|s_i)$. Let $\boldsymbol{\theta},\boldsymbol{\theta}'\in\Theta$, then for any $\epsilon>0$, it holds that:*

$$\Pr\left(\left[\ell(\boldsymbol{\theta})-\widehat{\ell}(\boldsymbol{\theta})\right]-\left[\ell(\boldsymbol{\theta}')-\widehat{\ell}(\boldsymbol{\theta}')\right]>\epsilon\right)\le\exp\left\{-\frac{\epsilon^2 n}{2\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|_2^2\sigma^2}\right\}.$$

Proof We write explicitly the involved expression, using Assumption 3 and perform some algebraic manipulations:

$$
\begin{aligned}
&\left[\ell(\boldsymbol{\theta}) - \widehat{\ell}(\boldsymbol{\theta})\right] - \left[\ell(\boldsymbol{\theta}') - \widehat{\ell}(\boldsymbol{\theta}')\right] \\
&= \mathop{\mathbb{E}}_{\substack{s \sim \nu \\ a \sim \pi_{\boldsymbol{\theta} Ag}(\cdot|s)}} \left[\boldsymbol{\theta}^T \boldsymbol{t}(s,a) - A(\boldsymbol{\theta}, s)\right] - \frac{1}{n}\sum_{i=1}^{n}\left(\boldsymbol{\theta}^T \boldsymbol{t}(s_i,a_i) - A(\boldsymbol{\theta}, s_i)\right) \\
&\quad - \mathop{\mathbb{E}}_{\substack{s \sim \nu \\ a \sim \pi_{\boldsymbol{\theta} Ag}(\cdot|s)}} \left[(\boldsymbol{\theta}')^T \boldsymbol{t}(s,a) - A(\boldsymbol{\theta}', s)\right] + \frac{1}{n}\sum_{i=1}^{n}\left((\boldsymbol{\theta}')^T \boldsymbol{t}(s_i,a_i) - A(\boldsymbol{\theta}', s_i)\right) \\
&= \mathop{\mathbb{E}}_{\substack{s \sim \nu \\ a \sim \pi_{\boldsymbol{\theta} Ag}(\cdot|s)}} \left[(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{t}(s,a) - (A(\boldsymbol{\theta}, s) - A(\boldsymbol{\theta}', s))\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\left((\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{t}(s_i,a_i) - (A(\boldsymbol{\theta}, s_i) - A(\boldsymbol{\theta}', s_i))\right).
\end{aligned}
$$

Essentially, we are comparing the mean and the sample mean of the random variable $(\boldsymbol{\theta} - \boldsymbol{\theta}')^T \boldsymbol{t}(s,a) - (A(\boldsymbol{\theta}, s) - A(\boldsymbol{\theta}', s))$. Let us now focus on $A(\boldsymbol{\theta}, s) - A(\boldsymbol{\theta}', s)$. From the mean value theorem we know that, for some $t \in [0,1]$ and $\overline{\boldsymbol{\theta}} = t\boldsymbol{\theta} + (1-t)\boldsymbol{\theta}'$, we have:

$$
A(\boldsymbol{\theta}, s) - A(\boldsymbol{\theta}', s) = \nabla_{\boldsymbol{\theta}} A(\overline{\boldsymbol{\theta}}, s)^T (\boldsymbol{\theta} - \boldsymbol{\theta}'). \tag{P.17}
$$

From Eq. (P.4), we know that $\nabla_{\boldsymbol{\theta}} A(\overline{\boldsymbol{\theta}}, s) = \mathop{\mathbb{E}}_{\overline{a} \sim \pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}[\boldsymbol{t}(s,\overline{a})]$. The random variable $\overline{\boldsymbol{t}}(s,a,\overline{\boldsymbol{\theta}}) = \boldsymbol{t}(s,a) - \mathop{\mathbb{E}}_{\overline{a} \sim \pi_{\overline{\boldsymbol{\theta}}}(\cdot|s)}[\boldsymbol{t}(s,\overline{a})]$ is a subgaussian random variable for any $\overline{\boldsymbol{\theta}} \in \Theta$. Thus, under Assumption 4 we have:

$$
\left[\ell(\boldsymbol{\theta}) - \widehat{\ell}(\boldsymbol{\theta})\right] - \left[\ell(\boldsymbol{\theta}') - \widehat{\ell}(\boldsymbol{\theta}')\right] = (\boldsymbol{\theta} - \boldsymbol{\theta}')^T \left(\mathop{\mathbb{E}}_{\substack{s \sim \nu \\ a \sim \pi_{\boldsymbol{\theta} Ag}(\cdot|s)}} \left[\overline{\boldsymbol{t}}(s,a,\overline{\boldsymbol{\theta}})\right] - \frac{1}{n}\sum_{i=1}^{n} \overline{\boldsymbol{t}}(s_i,a_i,\overline{\boldsymbol{\theta}}) \right).
$$

If we apply Proposition 3, we get the result. □

A.3 Results on significance and power of the tests

Theorem 3 Let $\widehat{I}_c$ be the set of parameter indexes selected by the Identification Rule 2 obtained using $n > 0$ i.i.d. samples collected with $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}$, with $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta$. Then, under Assumptions 1, 2, 3, 4, and 5, let $\boldsymbol{\theta}_i^{\mathrm{Ag}} = \arg\min_{\boldsymbol{\theta} \in \Theta_i} \ell(\boldsymbol{\theta})$ for all $i \in \{1, ..., d\}$ and $\xi = \min\{1, \frac{\lambda_{\min}}{\sigma^2}\}$. If $\widehat{\lambda}_{\min} \ge \frac{\lambda_{\min}}{2\sqrt{2}}$ and $\ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - l(\boldsymbol{\theta}^{\mathrm{Ag}}) \ge c_1$, it holds that:

$$\alpha \leq 2d \exp\left\{ -\frac{c_1 \lambda_{\min}^2 n}{16 d^2 \sigma^4} \right\}$$

$$\beta \leq (2d-1) \sum_{i \in I^{\mathrm{Ag}}} \exp\left\{ -\frac{\left(l(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - l(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1 \right) \lambda_{\min} \xi n}{16 (d-1)^2 \sigma^2} \right\}.$$

Proof We start considering $\alpha = \Pr\left(\exists i \notin I^{\mathrm{Ag}} : i \in \widehat{I}_c \right)$. We employ an argument analogous to that of (Garivier and Kaufmann, 2019):

$$\begin{aligned}
\Pr\left(\exists i \notin I^{\mathrm{Ag}} : i \in \widehat{I}_c \right) &= \Pr\left(\exists i \notin I^{\mathrm{Ag}} : \lambda_i > c_1 \right) \\
&= \Pr\left(\exists i \notin I^{\mathrm{Ag}} : \widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) > \frac{c_1}{2} \right) \\
&\leq \Pr\left(\exists i \notin I^{\mathrm{Ag}} : \widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) > \frac{c_1}{2} \right) \\
&= \Pr\left(\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) > \frac{c_1}{2} \right) \leq 2d \exp\left\{ -\frac{c_1 \lambda_{\min}^2 n}{16 d^2 \sigma^4} \right\},
\end{aligned}$$

where we observed that $\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) \geq \widehat{\ell}(\widehat{\boldsymbol{\theta}}_i)$ as $\boldsymbol{\theta}^{\mathrm{Ag}} \in \Theta_i$ under $\mathcal{H}_0$ and we applied Corollary 1 in the last line, recalling that $\widehat{\lambda}_{\min} \geq \frac{\lambda_{\min}}{2\sqrt{2}}$. For the second inequality, the derivation is a little more articulated. Concerning $\beta = \Pr\left(i \in I^{\mathrm{Ag}} : i \notin \widehat{I} \right)$, we first perform a union bound:

$$\Pr\left(\exists i \in I^{\mathrm{Ag}} : i \notin \widehat{I}_c \right) = \Pr\left(\bigvee_{i \in I^{\mathrm{Ag}}} i \notin \widehat{I}_c \right) \leq \sum_{i \in I^{\mathrm{Ag}}} \Pr\left(i \notin \widehat{I}_c \right).$$

Let us now focus on the single terms $\Pr\left(i \notin \widehat{I}_c \right)$. We now perform the following manipulations:

$$\begin{aligned}
\Pr\left(i \notin \widehat{I}_c \right) &= \Pr\left(\widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) \leq \frac{c_1}{2} \right) \\
&= \Pr\left(\left[\widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] + \left[\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) \right] + \left[\widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) \right] \leq \frac{c_1}{2} \right)
\end{aligned} \tag{P.18}$$

$$\begin{aligned}
&\leq \Pr\left(\left[\widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] + \left[\widehat{\ell}(\boldsymbol{\theta_i}^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) \right] \leq \frac{c_1}{2} \right) \\
&= \Pr\Big(\left[\widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) - \widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] + \left[\widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] + \left[\ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) \right] \\
&\leq \frac{c_1}{2} + \left[\ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] \Big) \\
&= \Pr\Big(\left[\widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) \right] + \left[\ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}_i^{\mathrm{Ag}}) \right] + \left[\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) \right] \\
&\geq \left[\ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) \right] - \frac{c_1}{2} \Big).
\end{aligned} \tag{P.19}$$

where line (P.18) is obtained by observing that $\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}) \geq 0$. Thus, we have:

$$\begin{aligned}
\Pr\left(i \notin \widehat{I}_c\right) \leq{} & \Pr\left(\widehat{\ell}(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) \geq \frac{1}{2}\left[\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right] - \frac{c_1}{2}\right) \\
& + \Pr\left(\left[\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}_i{}^{\mathrm{Ag}})\right] + \left[\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right] \geq \right. \\
& \left. \frac{1}{2}\left[\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right]\right)
\end{aligned} \tag{P.20}$$

$$\begin{aligned}
\leq{} & \Pr\left(\widehat{\ell}(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \widehat{\ell}(\widehat{\boldsymbol{\theta}}_i) \geq \frac{1}{2}\left[\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right] - \frac{c_1}{2}\right) \\
& + \Pr\left(\left[\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \widehat{\ell}(\boldsymbol{\theta}_i{}^{\mathrm{Ag}})\right] + \left[\widehat{\ell}(\boldsymbol{\theta}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right] \geq \right. \\
& \left. \frac{1}{2}\left[\frac{1}{2}\lambda_{\min}\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right)\left\|\boldsymbol{\theta}_i^{\mathrm{Ag}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2^2\right]^{\frac{1}{2}}\right)
\end{aligned} \tag{P.21}$$

$$\begin{aligned}
\leq{} & 2(d-1)\exp\left\{-\frac{\left(\ell(\boldsymbol{\theta}_i^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1\right)\lambda_{\min}^2 n}{16(d-1)^2\sigma^4}\right\} \\
& + \exp\left\{-\frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right)\lambda_{\min} n}{16\sigma^2}\right\}
\end{aligned} \tag{P.22}$$

$$\begin{aligned}
\leq{} & 2(d-1)\exp\left\{-\frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1\right)\lambda_{\min} n\xi}{16(d-1)^2\sigma^2}\right\} \\
& + \exp\left\{-\frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1\right)\lambda_{\min} n\xi}{16(d-1)^2\sigma^2}\right\} \\
\leq{} & (2d-1)\exp\left\{-\frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1\right)\lambda_{\min} n\xi}{16(d-1)^2\sigma^2}\right\}.
\end{aligned} \tag{P.23}$$

where line (P.20) derives from the inequality $\Pr(X + Y \geq c) \leq \Pr(X \geq a) + \Pr(Y \geq b)$ with $c = a + b$, line (P.21) is obtained by the following second order Taylor expansion, recalling that $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{\mathrm{Ag}}) = \boldsymbol{0}$:

$$\begin{aligned}
\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) &= \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}^{\mathrm{Ag}})^T\left(\boldsymbol{\theta}_i{}^{\mathrm{Ag}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right) + \frac{1}{2}\left(\boldsymbol{\theta}_i{}^{\mathrm{Ag}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right)^T\mathcal{H}_{\boldsymbol{\theta}}\ell(\overline{\boldsymbol{\theta}})\left(\boldsymbol{\theta}_i{}^{\mathrm{Ag}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right) \\
&\geq \frac{\lambda_{\min}}{2}\left\|\boldsymbol{\theta}_i{}^{\mathrm{Ag}} - \boldsymbol{\theta}^{\mathrm{Ag}}\right\|_2^2,
\end{aligned}$$

where $\overline{\boldsymbol{\theta}} = t\boldsymbol{\theta}^{\mathrm{Ag}} + (1-t)\boldsymbol{\theta}_i^{\mathrm{Ag}}$ for some $t \in [0,1]$. Line (P.22) is obtained by applying Corollary 1, recalling that $\widehat{\lambda}_{\min} \geq \frac{\lambda_{\min}}{2\sqrt{2}}$ and Theorem 7. Finally, line (P.23) derives by introducing the term $\xi = \min\left\{1, \frac{\lambda_{\min}}{\sigma^2}\right\}$ and observing that:

$$\frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1\right)\xi}{(d-1)^2} \leq \frac{\left(\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}})\right)n}{16}.$$

Clearly, this result is meaningful as long as $\ell(\boldsymbol{\theta}_i{}^{\mathrm{Ag}}) - \ell(\boldsymbol{\theta}^{\mathrm{Ag}}) - c_1 \geq 0$. □

B Detail on identification rules with configurable environment

In the following, we report the pseudocode for the environment configuration procedure in the case of application of Identification Rule 1 (Algorithm 4) which was omitted in the main text.

Algorithm 4 Identification Rule 1 (Combinatorial) with Environment Configuration.

```
input: parameter space $\Theta$, configuration space $\Omega$, threshold function $c$, number of configuration attempts $N_{\mathrm{conf}}$
  Initialize $\boldsymbol{\omega}_0$ arbitrarily
  Collect $\mathcal{D}_0$ observing $\pi_0^{\mathrm{Ag}}$ in environment $\mathcal{M}_{\boldsymbol{\omega}_0}$
  Run the Identification Rule 1 on $\mathcal{D}_0$ with $\delta'$ and get $\widehat{\mathcal{I}}_0$
  $\widehat{\mathcal{I}} \leftarrow \widehat{\mathcal{I}}_0$
  for $I \subseteq \{1, ..., d\} : I \notin \widehat{\mathcal{I}}$ do
      $\boldsymbol{\omega}_{i,0} \leftarrow \boldsymbol{\omega}_0$
      $\mathcal{D}_{i,0} \leftarrow \mathcal{D}$
      for $j = 1, ..., N_{\mathrm{conf}}$ do
          Optimize $\mathcal{C}_I(\boldsymbol{\omega}/\boldsymbol{\omega}_{i,j-1})$ getting $\boldsymbol{\omega}_{i,j}$
          Collect $\mathcal{D}_{i,j}$ observing $\pi_{i,j}^{\mathrm{Ag}}$ in environment $\mathcal{M}_{\boldsymbol{\omega}_{i,j}}$
          Run the Identification Rule 1 on $\mathcal{D}_{i,j}$ and obtain $\widehat{\mathcal{I}}_{i,j}$
          $\widehat{\mathcal{I}} \leftarrow \widehat{\mathcal{I}} \cup \widehat{\mathcal{I}}_{i,j}$
      end for
  end for
  return $\widehat{\mathcal{I}}$
```

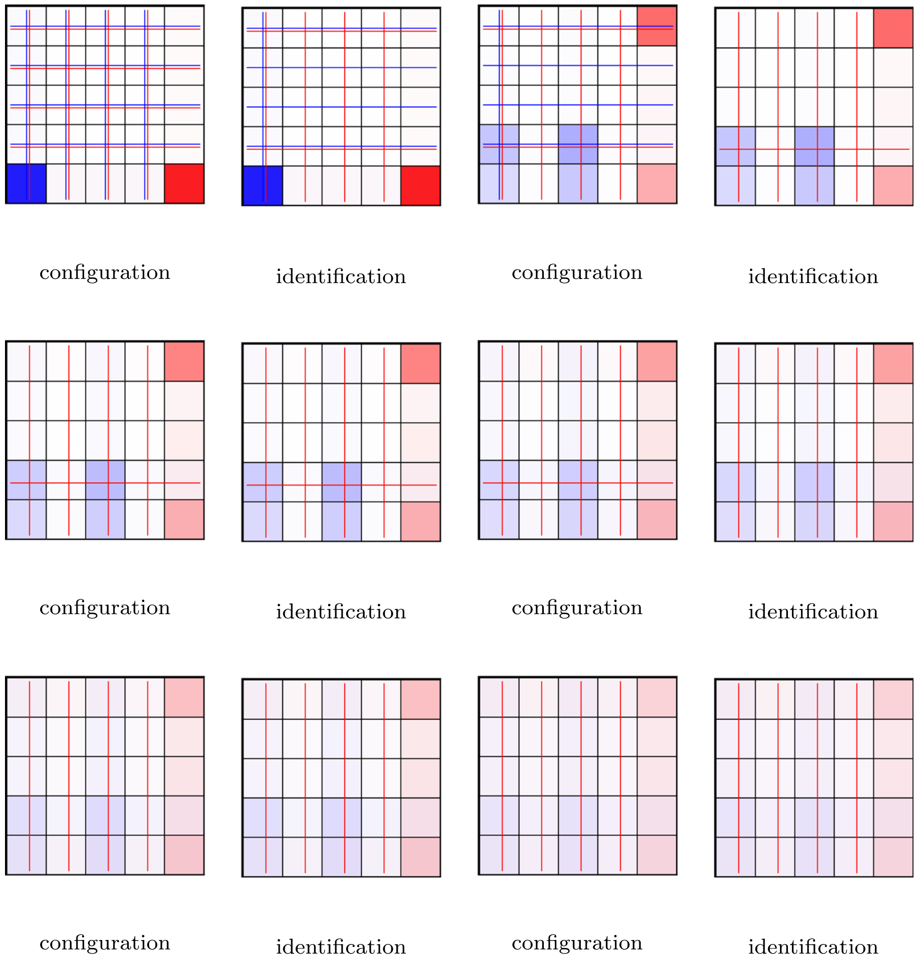
C Experimental details

In this appendix, we report the full experimental results, along with the hyperparameters employed.

C.1 Experimental details for section 6.1

C.1.1 Discrete grid world

Hyperparameters In the following, we report the hyperparameters used for the experiments on the discrete grid world:



configuration identification configuration identification

configuration identification configuration identification

configuration identification configuration identification

Fig. 6 Example of configuration and identification in the discrete grid world

- Horizon (T): 50
- Discount factor (γ): 0.98
- Learning steps with G(PO)MDP: 200
- Batch size: 250
- Max-likelihood maximum update steps: 1000
- Max-likelihood learning rate (using Adam): 0.03
- Number of configuration attempts per feature (N_{conf}): 3
- Environment configuration update steps: 150
- Regularization parameter of the Rényi divergence (ζ): 0.125
- Significance of the likelihood-ratio tests (δ): 0.01

Example of configuration and identification in the discrete grid world In Fig. 6, we show a graphical representation of a single experiment with the grid world environment using its configurability to better identify the policy space. The colors inside the squares indicate the probability mass function associated to the initial state distribution, consisting of the agent's position (blue) and the goal position (red), where sharper colors mean higher probabilities. The colored lines represent the features the agent has access to, they are binary features indicating if the agent is on a certain row or column (blue lines) and if the goal is on a certain row or column (red lines). Note that, to avoid redundancy of representation (and so enforcing the identifiability), the last row and column are not explicitly encoded, but they can be represented by the absence of the other rows and columns. When a line is not shown anymore, it means that it has been rejected, i.e., we think the agent has access to that feature. The agent has access to every feature except for the goal columns, i.e., only to its own position and to the goal row are known.

The configuration of the environment is updated in the images at even position, the identification step is performed at even positions. The environment is configured in order to maximize the influence on the gradient of the first – not rejected – feature, considering the blue features first and then the red ones. After the model was configured three times for a feature, and the feature has not been rejected, the model was configured for the next one.

We can see that the general trend of this configuration is to change the parameters in order to spread the initial value of the mass probability functions across a greater number of grid cells. This is an expected behavior since with the initial model configuration, very often an episode starts with the agent in the bottom-left of the grid and the goal in the bottom-right, causing the policy to depend mostly on the position of the agent. In fact, only blue column features are rejected at the first iteration, as we can see in the third image. Instead, distributing the probabilities across the whole grid let an episode starts with the two positions extracted almost uniformly. Eventually, the correct policy space is identified. It is interesting to observe that such is can hardly be obtained without the configuration of the environment, given the initial state distribution shown in the first image.

C.1.2 Continuous grid world

In this appendix, we report the experiments performed on the continuous version of the grid world. In this environment, the agent has to reach a goal region, delimited by a circle, starting from an initial position. Both initial position and center of the goal are sampled at the beginning of the episode from a Gaussian distribution with fixed covariance μ_{ω}. The supervisor is allowed to change, via the parameters $\boldsymbol{\omega}$, the mean of this distribution. The agent specifies, at each time step, the speed in the vertical and horizontal direction, by means of a bivariate Gaussian policy with fixed covariance, linear in a set of radial basis functions (RBF) for representing both the current position of the agent and the position of

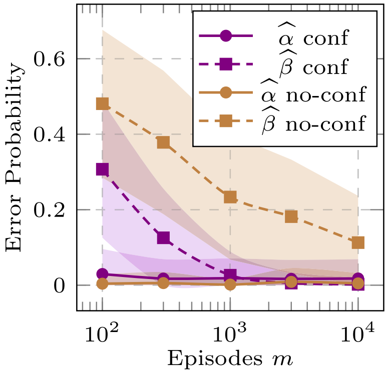


Fig. 7 $\widehat{\alpha}$ and $\widehat{\beta}$ error for *conf* and *no-conf* cases in the continuous grid world varying the number of episodes m. 25 runs 95% c.i

the goal (5×5 both for the agent position and the goal). The feature, and consequently the parameters, that the agent can control are randomly selected at the beginning. In Fig. 7, we show the results of an experiment analogous to that of the discrete grid world, by comparing $\widehat{\alpha}$ and $\widehat{\beta}$ for the case in which we do not perform environment configuration (no-conf) and the case in which the configuration is performed (conf). Once again, we confirm our findings that configuring the environment allows speeding up the identification process by inducing the agent chaining its policy and, as a consequence, revealing which parameters it can actually control.

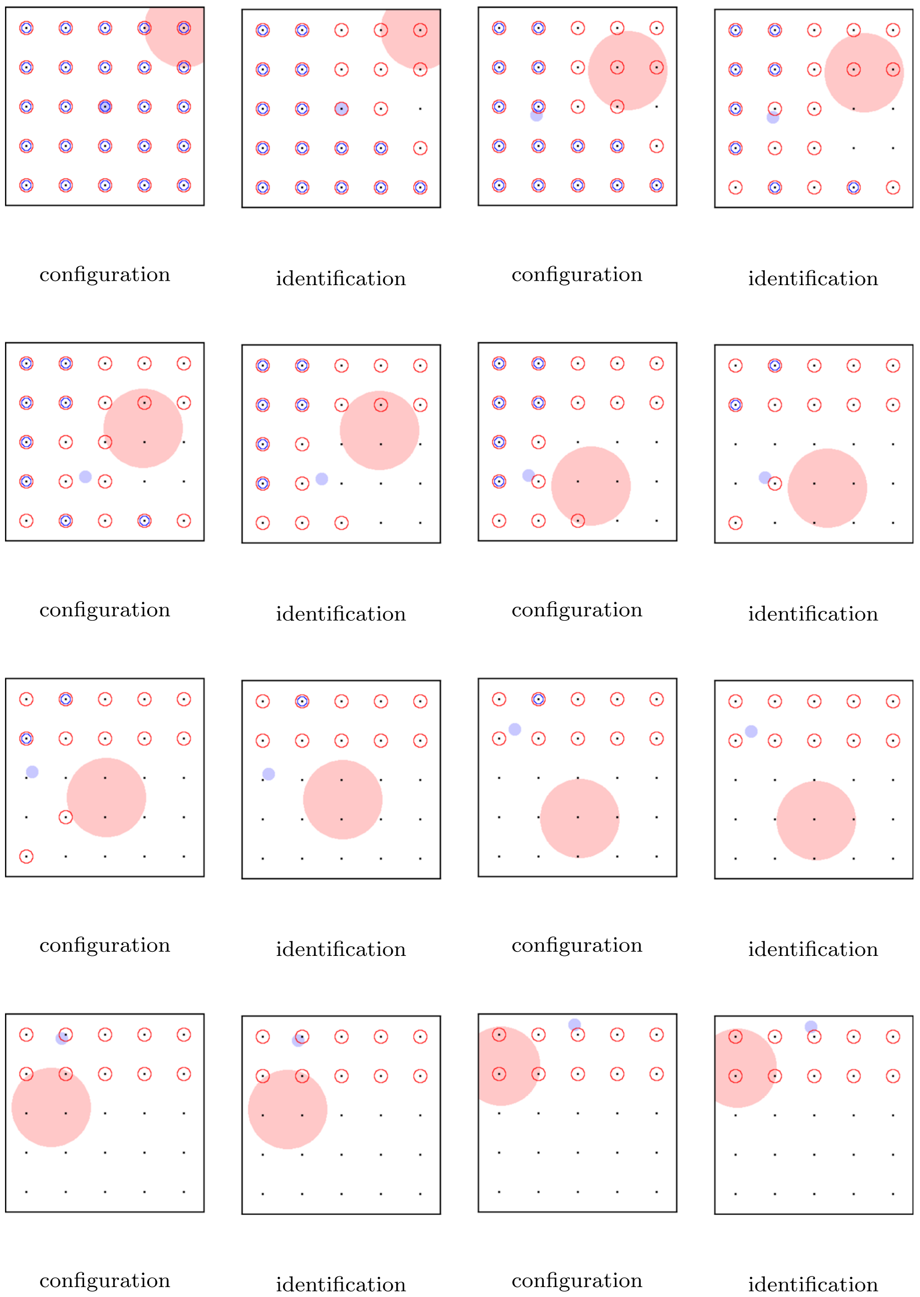
Hyperparameters In the following, we report the hyperparameters used for the experiments on the continuous grid world:

- Horizon (T): 50
- Discount factor (γ): 0.98
- Policy covariance ($\boldsymbol{\Sigma}$): $0.02^2\boldsymbol{I}$
- Learning steps with G(PO)MDP: 100
- Batch size: 100
- Max-likelihood maximum update steps: closed form
- Number of configuration attempts per feature (N_{conf}): 3
- Environment configuration update steps: 100
- Regularization parameter of the Rényi divergence (ζ): $1e-6$
- Significance of the likelihood-ratio tests (δ): 0.01

Example of configuration and identification in the continuous grid world In Fig. 8, we show an example of model configuration in the continuous grid world environment. The two filled circles are a graphical representation of the normal distributions from which the initial position of the agent (light blue) and the position of the goal (pink) are sampled at the beginning of each episode. The circumferences correspond to the set of features (RBF) to which the agent has access, among which we want to discover the ones accessible by the agent. Since the policy space is composed by Gaussian policies with mean specified by a linear combination of these features, each one is associated to a parameter. If a circumference is not shown anymore at an iteration step, it means that the hypothesis associated to that feature was rejected, i.e., we believe that the agent has access to that feature.

The group of images is an alternated sequence of new environment configurations and parameter identifications. In the first image we can see the initial model with no rejected features. The identification with the initial model yields to the rejection of a certain set of features, which can be seen in the second image. The third image shows the new configuration of the model, in which the mean of the two initial state distributions are moved in order to investigate the remaining features. Then a new test is performed and the result is shown in the fourth image, and so on. In this experiment, the environment was configured in order to maximize the influence of one feature at a time, starting from the blue ones from bottom-left to top-right in row order, and then with the red ones in the same order. Each feature is used to configure the model for a maximum of three times, after that point the next feature is considered.

The only features that were not actually in the agent's set are the red ones on the two top rows. We can see that the mean of the initial position of the agent (a configurable parameter of the environment) always tracked the first available feature yet to be tested, as expected from this experiment. In fact, when the initial position is close enough to those features, the agent often moves around those blue circumferences to reach the goal, making them more important in the definition of the optimal policy. Eventually, the tests reject all



configuration identification configuration identification

configuration identification configuration identification

configuration identification configuration identification

configuration identification configuration identification

Fig. 8 Example of configuration and identification in the continuous grid world

the features that are actually accessible by the agent, and only them, yielding to a correct identification of the policy space. The rest of the configurations are not shown, since no more features were rejected. In this experiment, similarly to the discrete grid world case, the use of Conf–MDPs was crucial to obtain this result.

C.1.3 Simulated car driving

In this environment, an agent has to drive a car to reach the end of the track without running off the road. The control directives are the acceleration and the steering, and are expressed through a two dimensional bounded action space. The car has four sensors oriented in different directions: $-\frac{\pi}{4}$, $-\frac{\pi}{6}$, $\frac{\pi}{6}$, $\frac{\pi}{4}$ w.r.t. the axis pointing toward the front of the car. The values of these sensors are the normalized distances from the car to the nearest road margin along the direction of the sensor, or the maximum value if the margin is outside the range of the sensor. The complete set of state features is made up by the normalized car speed and the values of the four sensors. In the experiments, the agent has access to the speed and the sensor at angles $\frac{\pi}{6}$ and $\frac{\pi}{4}$. The track consists in a single road segment with a fixed curvature. The rewards are given proportionally to the speed of the car, i.e., greater speeds yield higher rewards. The episode finishes when the car goes outside the road, and a negative reward is given in this case, when the track is completed, or when a maximum number of time steps is elapsed.

Hyperparameters In the following, we report the hyperparameters used for the experiments on the simulated car driving:

- Horizon (T): 250
- Discount factor (γ): 0.996
- Policy covariance ($\boldsymbol{\Sigma}$): $0.1\boldsymbol{I}$
- Learning steps with G(PO)MDP: 100
- Batch size: 50
- Max-likelihood maximum update steps: 200
- Max-likelihood learning rate (using Adam): 0.1
- Significance of the likelihood-ratio tests (δ): 0.1 rescaled by 0.1/5 for the simplified identification rule and 0.1/32 for the combinatorial identification rule

C.2 Experimental details of section 6.2

Hyperparameters In the following, we report the hyperparameters used for the experiments on the discrete grid world:

- Horizon (T): 50
- Discount factor (γ): 0.98
- Learning steps with G(PO)MDP: 200
- Batch size: 250
- Max-likelihood maximum update steps: 1000
- Max-likelihood learning rate (using Adam): 0.03
- Number of configuration attempts per feature (N_{conf}): 3
- Environment configuration update steps: 150
- Regularization parameter of the Rényi divergence (ζ): 0.125
- Significance of the likelihood-ratio tests (δ): 0.01

Additional Results In the following, we report the complete results about the imitation learning experiments. These results extend the ones presented in the main paper providing additional algorithms and additional metrics for comparison.

Concerning the additional algorithms, we include other two regularization techniques for the maximum likelihood estimation: Shannon and Tsallis entropy. Given a policy π, the Shannon $\mathbb{H}(\pi)$ and Tsallis $\mathbb{W}(\pi)$ entropies are defined as follows (Ho and Ermon, 2016; Lee et al., 2018):

$$\mathbb{H}(\pi) = \mathop{\mathbb{E}}_{s \sim d_{\mu}^{\pi}}[H(\pi(\cdot|s))] = \mathop{\mathbb{E}}_{\substack{s \sim d_{\mu}^{\pi} \\ a \sim \pi(\cdot|s)}} [-\log \pi(a|s)],$$

$$\mathbb{W}(\pi) = \mathop{\mathbb{E}}_{s \sim d_{\mu}^{\pi}}[W(\pi(\cdot|s))] = \frac{1}{2} \mathop{\mathbb{E}}_{\substack{s \sim d_{\mu}^{\pi} \\ a \sim \pi(\cdot|s)}} [1 - \pi(a|s)].$$

It is worth noting that, differently from the other regularizers (like ridge and lasso), Shannon and Tsallis entropies require to compute an expectation w.r.t. to the policy π we are optimizing. Since samples are collected with a policy that is, in general, different and unknown (the expert's policy) those expectations are approximated, in our experiments, with *self-normalized* importance weighting (Owen, 2013). Thus, the complete loss function that is optimized, ignoring the ridge and lasso regularizers for brevity, is the following:

$$\mathcal{Q}(\boldsymbol{\theta}; \lambda^{\mathrm{S}}, \lambda^{\mathrm{W}}) = -\sum_{i=1}^{n} \log \pi_{\boldsymbol{\theta}}(a_i|s_i) + \underbrace{\lambda^{\mathrm{S}} \sum_{i=1}^{n} \widetilde{\omega}_i(\boldsymbol{\theta}) \log \pi_{\boldsymbol{\theta}}(a_i|s_i)}_{\text{Shannon entropy}} - \underbrace{\lambda^{\mathrm{W}} \sum_{i=1}^{n} \widetilde{\omega}_i(\boldsymbol{\theta}) \left(1 - \pi_{\boldsymbol{\theta}}(a_i|s_i)\right)}_{\text{Tsallis entropy}},$$

where $\widetilde{\omega}_i(\boldsymbol{\theta}) = \frac{n\pi_{\boldsymbol{\theta}}(a_i|s_i)}{\sum_{j=1}^{n} \pi_{\boldsymbol{\theta}}(a_j|s_j)}$ is the self-normalized importance weight.

Furthermore, we have tested other IL methods that require natively the interaction with the environment. In our truly batch model-free setting, we replaced again the interaction with the environment with off-policy estimation. These algorithms are based on the notion of *feature expectation*, i.e., the expectation of a feature function $\boldsymbol{\phi}(s,a)$ under the γ-discounted stationary distribution induced by a policy π:

$$\boldsymbol{\phi}(\pi) = \mathop{\mathbb{E}}_{\substack{s \sim d_{\mu}^{\pi} \\ a \sim \pi(\cdot|s)}} [\boldsymbol{\phi}(s,a)].$$

The goal consists in finding a policy $\pi_{\widehat{\boldsymbol{\theta}}}$ that *matches* the feature expectations induced by the expert's policy $\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}$, i.e., $\boldsymbol{\phi}(\pi_{\widehat{\boldsymbol{\theta}}}) \simeq \boldsymbol{\phi}(\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}})$ and applying a regularization on $\pi_{\widehat{\boldsymbol{\theta}}}$. If the regularization is the Shannon entropy we have the *Maximum Causal Entropy Inverse Reinforcement Learning* (MCE, Ziebart et al., 2010):

$$\begin{aligned} &\max_{\boldsymbol{\theta} \in \Theta} \alpha^{S} \mathbb{H}(\pi_{\boldsymbol{\theta}}) \\ &\text{s.t. } \boldsymbol{\phi}(\pi_{\boldsymbol{\theta}}) = \boldsymbol{\phi}(\pi_{\boldsymbol{\theta}^{\mathrm{Ag}}}), \end{aligned}$$

where α^{S} is a scale parameter. Instead, if we employ Tsallis entropy we obtain the *Maximum Tsallis Entropy Imitation Learning* (MTE, Lee et al., 2018):

Table 2 Norm of the parameter difference for the IL algorithms tested. We report sample mean ± sample std

Algorithm	Episodes m					
	10	30	100	300	1000	3000
True	**10.826 ± 3.026**	**10.265 ± 2.373**	**7.773 ± 2.093**	**6.061 ± 1.696**	**5.328 ± 1.175**	**4.715 ± 1.934**
No-Conf	12.918 ± 2.405	12.280 ± 2.554	9.808 ± 2.022	7.212 ± 1.729	**5.373 ± 1.248**	**4.785 ± 2.022**
Conf	**11.323 ± 2.735**	**10.227 ± 2.857**	**8.232 ± 2.022**	**6.215 ± 1.594**	**5.790 ± 1.638**	**5.189 ± 1.705**
ML	13.561 ± 2.346	13.582 ± 2.053	11.381 ± 2.506	9.563 ± 1.893	7.143 ± 1.227	6.042 ± 1.886
Ridge ($\lambda^{\mathrm{R}} = 0.001$)	**11.820 ± 2.151**	**10.872 ± 1.637**	**7.896 ± 2.003**	**6.448 ± 1.200**	**5.880 ± 1.318**	**5.449 ± 1.627**
Ridge ($\lambda^{\mathrm{R}} = 0.01$)	**11.052 ± 2.493**	**10.623 ± 1.859**	8.936 ± 2.127	8.630 ± 1.686	9.067 ± 1.838	8.155 ± 1.444
Lasso ($\lambda^{\mathrm{L}} = 0.001$)	12.623 ± 2.081	11.686 ± 1.807	**8.613 ± 2.036**	**6.581 ± 1.584**	**5.858 ± 1.154**	**5.341 ± 1.715**
Lasso ($\lambda^{\mathrm{L}} = 0.01$)	**11.624 ± 2.320**	**10.669 ± 2.075**	8.896 ± 2.078	8.480 ± 1.473	8.950 ± 1.698	8.185 ± 1.541
Shannon ($\lambda^{\mathrm{S}} = 0.1$)	13.460 ± 2.309	13.461 ± 2.023	11.261 ± 2.485	9.447 ± 1.848	7.084 ± 1.217	6.012 ± 1.880
Shannon ($\lambda^{\mathrm{S}} = 1$)	12.313 ± 2.230	12.172 ± 1.598	10.354 ± 1.755	8.972 ± 0.943	8.665 ± 0.918	8.251 ± 1.218
Tsallis ($\lambda^{\mathrm{W}} = 0.1$)	13.491 ± 2.319	13.512 ± 2.033	11.318 ± 2.503	9.494 ± 1.861	7.094 ± 1.218	6.006 ± 1.887
Tsallis ($\lambda^{\mathrm{W}} = 1$)	12.435 ± 2.291	12.386 ± 1.625	10.682 ± 1.839	9.221 ± 1.040	8.806 ± 0.971	8.331 ± 1.258
FE	19.682 ± 3.240	20.307 ± 2.986	19.372 ± 3.255	16.951 ± 2.599	15.582 ± 2.387	13.797 ± 1.958
MCE ($\alpha^{\mathrm{S}} = 0.01$)	24.040 ± 5.123	26.554 ± 5.029	25.822 ± 5.408	24.733 ± 4.076	24.157 ± 4.569	20.809 ± 3.297
MCE ($\alpha^{\mathrm{S}} = 0.1$)	24.196 ± 5.329	26.163 ± 4.837	26.195 ± 5.532	24.980 ± 4.025	24.095 ± 4.350	20.767 ± 3.279
MTE ($\alpha^{\mathrm{W}} = 0.01$)	23.264 ± 5.211	25.537 ± 4.139	24.430 ± 5.184	23.857 ± 4.365	24.232 ± 4.721	20.824 ± 3.281
MTE ($\alpha^{\mathrm{W}} = 0.1$)	23.988 ± 5.130	26.403 ± 4.857	25.681 ± 5.323	24.626 ± 4.093	24.142 ± 4.574	20.805 ± 3.286

For each number of episodes m in bold the algorithm with the smallest sample mean together with those not statistically significantly different from that one (Welch's t-test with $p < 0.05$)

Table 3 Estimated expected KL divergence, multiplied by 1000 for easiness of visualization, for the IL algorithms tested. We report sample mean ± sample std

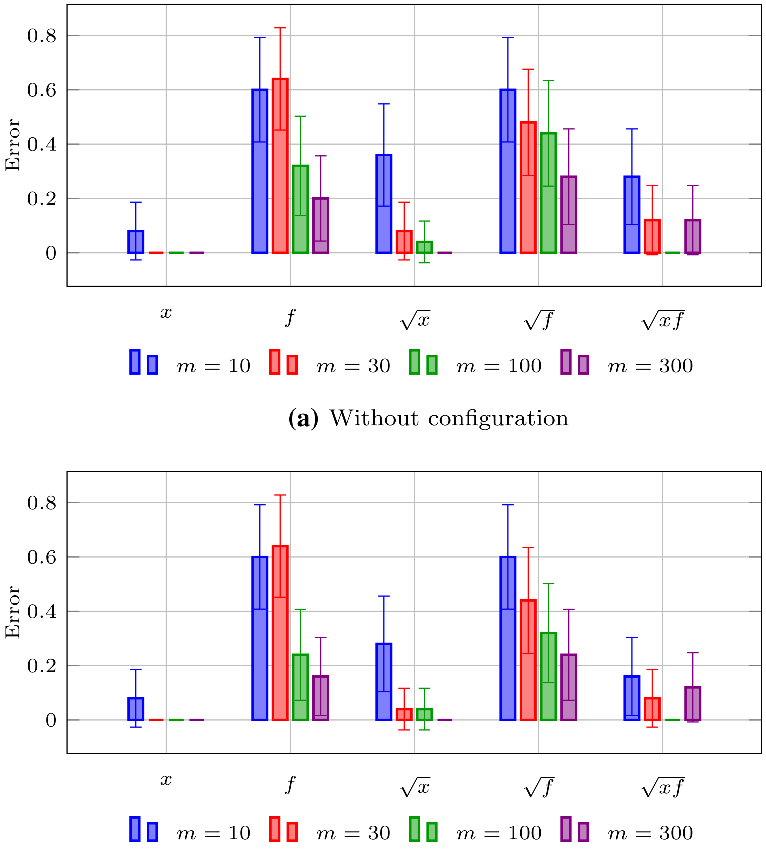
Algorithm	Episodes m					
	10	30	100	300	1000	3000
True	0.141 ± 0.131	0.134 ± 0.110	**0.066** ± **0.085**	0.042 ± 0.045	0.036 ± 0.021	0.031 ± 0.040
No-Conf	0.226 ± 0.223	0.178 ± 0.104	0.128 ± 0.161	0.078 ± 0.094	0.037 ± 0.022	0.033 ± 0.042
Conf	**0.086** ± **0.055**	**0.067** ± **0.058**	**0.042** ± **0.031**	**0.015** ± **0.016**	**0.014** ± **0.009**	**0.011** ± **0.008**
ML	0.200 ± 0.157	0.186 ± 0.122	0.102 ± 0.117	0.067 ± 0.063	0.039 ± 0.021	0.036 ± 0.040
Ridge ($\lambda^{\mathrm{R}} = 0.001$)	0.157 ± 0.115	0.129 ± 0.102	**0.070** ± **0.086**	0.050 ± 0.062	0.034 ± 0.029	0.032 ± 0.041
Ridge ($\lambda^{\mathrm{R}} = 0.01$)	0.127 ± 0.107	0.104 ± 0.084	0.088 ± 0.124	0.106 ± 0.158	0.075 ± 0.063	0.071 ± 0.079
Lasso ($\lambda^{\mathrm{L}} = 0.001$)	0.174 ± 0.130	0.154 ± 0.107	0.074 ± 0.077	0.046 ± 0.042	0.040 ± 0.032	0.030 ± 0.032
Lasso ($\lambda^{\mathrm{L}} = 0.01$)	0.148 ± 0.114	0.121 ± 0.109	0.081 ± 0.074	0.076 ± 0.083	0.074 ± 0.086	0.058 ± 0.050
Shannon ($\lambda^{\mathrm{S}} = 0.1$)	0.196 ± 0.152	0.184 ± 0.121	0.100 ± 0.114	0.066 ± 0.062	0.040 ± 0.021	0.038 ± 0.042
Shannon ($\lambda^{\mathrm{S}} = 1$)	0.202 ± 0.182	0.188 ± 0.124	0.121 ± 0.090	0.138 ± 0.116	0.110 ± 0.060	0.116 ± 0.076
Tsallis ($\lambda^{\mathrm{W}} = 0.1$)	0.197 ± 0.153	0.185 ± 0.122	0.101 ± 0.115	0.066 ± 0.062	0.040 ± 0.022	0.037 ± 0.042
Tsallis ($\lambda^{\mathrm{W}} = 1$)	0.210 ± 0.191	0.197 ± 0.131	0.125 ± 0.091	0.137 ± 0.111	0.112 ± 0.061	0.116 ± 0.074
FE	0.622 ± 0.331	0.658 ± 0.362	0.484 ± 0.329	0.446 ± 0.394	0.380 ± 0.288	0.336 ± 0.284
MCE ($\alpha^{\mathrm{S}} = 0.01$)	1.008 ± 0.784	1.042 ± 0.512	0.784 ± 0.412	0.830 ± 0.615	0.743 ± 0.444	0.529 ± 0.327
MCE ($\alpha^{\mathrm{S}} = 0.1$)	0.916 ± 0.568	0.929 ± 0.441	0.755 ± 0.406	0.802 ± 0.600	0.734 ± 0.390	0.467 ± 0.267
MTE ($\alpha^{\mathrm{W}} = 0.01$)	1.013 ± 0.642	1.209 ± 0.703	0.900 ± 0.585	0.843 ± 0.713	0.845 ± 0.534	0.555 ± 0.372
MTE ($\alpha^{\mathrm{W}} = 0.1$)	0.994 ± 0.724	1.045 ± 0.523	0.808 ± 0.416	0.854 ± 0.666	0.748 ± 0.449	0.530 ± 0.326

For each number of episodes m in bold the algorithm with the smallest sample mean together with those not statistically significantly different from that one (Welch's t-test with $p < 0.05$)

Table 4 Estimated difference of feature expectation for the IL algorithms tested. We report sample mean ± sample std

Algorithm	Episodes m					
	10	30	100	300	1000	3000
True	2.687 ± 1.822	2.139 ± 1.306	0.943 ± 0.478	$\mathbf{0.794 \pm 0.429}$	0.802 ± 0.554	0.763 ± 0.456
No-Conf	16.381 ± 16.492	8.867 ± 8.356	4.339 ± 6.707	0.969 ± 0.680	$\mathbf{0.685 \pm 0.401}$	0.743 ± 0.398
Conf	12.263 ± 7.021	6.627 ± 3.259	2.823 ± 2.264	1.857 ± 0.698	1.576 ± 0.674	1.657 ± 0.647
ML	3.032 ± 1.642	2.340 ± 1.249	1.012 ± 0.417	0.911 ± 0.398	0.763 ± 0.352	$\mathbf{0.709 \pm 0.325}$
Ridge ($\lambda^{\mathrm{R}} = 0.001$)	2.840 ± 1.541	2.297 ± 1.367	1.052 ± 0.446	0.822 ± 0.415	0.705 ± 0.308	0.783 ± 0.425
Ridge ($\lambda^{\mathrm{R}} = 0.01$)	$\mathbf{2.616 \pm 1.730}$	2.532 ± 1.736	1.651 ± 0.711	1.301 ± 0.705	1.284 ± 0.646	1.149 ± 0.600
Lasso ($\lambda^{\mathrm{L}} = 0.001$)	2.870 ± 1.700	$\mathbf{2.051 \pm 1.073}$	$\mathbf{0.851 \pm 0.436}$	0.899 ± 0.507	0.687 ± 0.363	0.804 ± 0.335
Lasso ($\lambda^{\mathrm{L}} = 0.01$)	2.866 ± 1.739	2.219 ± 1.232	1.216 ± 0.524	1.140 ± 0.598	0.891 ± 0.413	0.926 ± 0.457
Shannon ($\lambda^{\mathrm{S}} = 0.1$)	2.910 ± 1.692	2.373 ± 1.481	0.984 ± 0.451	0.900 ± 0.445	0.808 ± 0.500	0.771 ± 0.361
Shannon ($\lambda^{\mathrm{S}} = 1$)	4.095 ± 2.995	3.864 ± 3.701	2.483 ± 1.782	1.704 ± 1.470	1.725 ± 0.765	1.584 ± 0.753
Tsallis ($\lambda^{\mathrm{W}} = 0.1$)	2.929 ± 1.666	2.293 ± 1.308	1.077 ± 0.472	0.969 ± 0.545	0.708 ± 0.379	0.765 ± 0.345
Tsallis ($\lambda^{\mathrm{W}} = 1$)	4.380 ± 3.503	3.875 ± 4.068	2.492 ± 1.620	1.694 ± 1.148	1.706 ± 0.754	1.850 ± 0.819
FE	25.106 ± 22.279	25.300 ± 22.170	27.079 ± 24.628	8.702 ± 12.867	15.799 ± 18.865	19.666 ± 25.388
MCE ($\alpha^{\mathrm{S}} = 0.01$)	29.586 ± 25.318	33.387 ± 25.642	32.952 ± 26.407	6.926 ± 6.279	16.468 ± 19.983	20.573 ± 26.272
MCE ($\alpha^{\mathrm{S}} = 0.1$)	26.397 ± 24.818	25.932 ± 24.781	31.217 ± 26.770	6.970 ± 10.155	13.845 ± 18.148	18.507 ± 24.950
MTE ($\alpha^{\mathrm{W}} = 0.01$)	33.981 ± 25.209	38.497 ± 26.502	37.628 ± 25.751	8.153 ± 10.906	18.014 ± 21.473	20.743 ± 26.215
MTE ($\alpha^{\mathrm{W}} = 0.1$)	29.498 ± 25.283	33.346 ± 26.013	34.226 ± 26.064	8.212 ± 11.459	17.120 ± 20.806	20.696 ± 26.201

For each number of episodes m in bold the algorithm with the smallest sample mean together with those not statistically significantly different from that one (Welch's t-test with $p < 0.05$)



(a) Without configuration

(b) With configuration

Fig. 9 Experiment with randomly chosen features on the minigolf domain for different number of episodes *m*. 100 runs, 95% c.i

$$\max_{\theta\in\Theta} \alpha^W \mathbb{W}(\pi_{\theta})$$
$$\text{s.t. } \boldsymbol{\phi}(\pi_{\theta}) = \boldsymbol{\phi}(\pi_{\theta^{\mathrm{Ag}}}),$$

where α^W is a scale parameter.

In both cases, similarly to the regularizers presented above, the computation of the objective requires to perform an off-policy estimation via importance sampling. In these cases, we have the additional complexity that also the constraint, i.e., matching the feature expectation, requires off-policy estimation for the left hand side.

The tables reported in the following pages present the complete results. As comparison metrics, we employed the norm of the parameter difference (Table 2), the estimated expected KL-divergence (Table 3), as defined in Section 6.2, and the norm of the estimated difference in the feature expectations (Table 4). In each table we report, as an *oracle* baseline, the results of ML assuming to have the knowledge of the parameters actually controlled by the agent (*True*). *FE* is a feature matching baseline, obtained by looking for the policy that better explains the feature expectations induced by the expert's data:

$$\min_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{n}\sum_{i=1}^{n}\widetilde{\omega}_i(\boldsymbol{\theta})\boldsymbol{\phi}(s_i,a_i)-\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\phi}(s_i,a_i)\right\|_2^2,$$

where $\widetilde{\omega}_i(\boldsymbol{\theta})$ is the self-normalized importance weight, as defined before. Finally, MCE and MCT are Maximum Causal Entropy and Maximum Tsallis Entropy, adapted with importance sampling. As a general trend, we can see that all algorithms that employ importance weighting do not perform well. This can be explained by the fact that expert's policy, which is likely (near) optimal, does not provide good information across the state-action space. As a consequence, the importance weighting procedure injects a large uncertainty (Owen, 2013; Metelli et al., 2018b). This also highlights how this no-interaction setting makes the IL problem challenging.

C.3 Experimental details of section 6.3

In the minigolf experiment, the polynomial features obtained from the distance from the goal x and the friction f are the following:

$$\boldsymbol{\phi}(x,f)=\left(1,\,x,f,\,\sqrt{x},\,\sqrt{f},\,\sqrt{xf}\right)^T.$$

While agent $\mathscr{A}_1$ perceives all the features, agent $\mathscr{A}_2$ has access to $(1,\,x,\,\sqrt{x})^T$ only.

Hyperparameters In the following, we report the hyperparameters used for the experiments on the minigolf:

- Horizon (T): 20
- Discount factor (γ): 0.99
- Policy covariance ($\boldsymbol{\Sigma}$): 0.01
- Learning steps with G(PO)MDP: 100
- Batch size: 100
- Max-likelihood maximum update steps: closed form
- Number of configuration attempts per feature (N_{conf}): 10
- Environment configuration update steps: 100
- Regularization parameter of the Rényi divergence (ζ): 0.25
- Significance of the likelihood-ratio tests (δ): 0.01

C.3.1 Experiment with randomly chosen features

In the following, we report an additional experiment in the minigolf domain in which the features that the agent can perceive are randomly selected at the beginning, comparing the case in which we do not configure the environment and the case in which environment configuration is performed, and for different number of episodes collected. Although, less visible w.r.t. to the grid world case, we can see that for some features (e.g., $\sqrt{x}$ and $\sqrt{xf}$) the environment configurability is beneficial.

Funding Open access funding provided by Politecnico di Milano within the CRUI-CARE Agreement.

References

Antos, A., Szepesvári, C., & Munos, R. (2008). Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning, 71*(1), 89–129. https://doi.org/10.1007/s10994-007-5038-2.

Barnard, G. A. (1959). Control charts and stochastic processes. Journal of the Royal Statistical Society: Series B (Methodological)

Ben-Israel, A., Greville, T.N. (2003). Generalized inverses: theory and applications, vol 15. Berlin: Springer Science & Business Media

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities—A nonasymptotic theory of independence*. Oxford: Oxford University Press.

Brantley, K., Sun, W., Henaff, M. (2020). Disagreement-regularized imitation learning. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia. April 26-30, 2020. OpenReview.net

Brown, L.D. (1986). Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims

Busoniu, L., Babuska, R., & De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 38*(2), 156–172.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (Vol. 2). Duxbury: Pacific Grove.

Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., Song, L. (2018). SBEED: convergent reinforcement learning with nonlinear function approximation. In: Dy JG, Krause A (eds) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden. July 10-15, 2018, PMLR, Proceedings of Machine Learning Research, vol. 80, pp. 1133–1142

Deisenroth, M.P., Rasmussen, C.E. (2011). PILCO: A model-based and data-efficient approach to policy search. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28–July 2, 2011, pp. 465–472

Deisenroth, M.P., Neumann, G., Peters, J. (2013). A survey on policy search for robotics. Foundations and Trends in Robotics

Finn, C., Abbeel, P., Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In: Proceedings of the 34th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol. 70, pp. 1126–1135

Garivier, A., Kaufmann, E. (2019). Non-asymptotic sequential tests for overlapping hypotheses and application to near optimal arm identification in bandit models. arXiv preprint arXiv:190503495

Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of internal medicine, 130*(12), 995–1004.

Ho, J., Ermon, S. (2016). Generative adversarial imitation learning. In: Lee DD, Sugiyama M, von Luxburg U, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, pp. 4565–4573

Hsu, D., Kakade, S., Zhang, T., et al. (2012). A tail inequality for quadratic forms of subgaussian random vectors. Electronic Communications in Probability, 17

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society, 31,* 203–222.

Jolliffe, I.T. (2011). Principal component analysis. In: Lovric M (ed) International Encyclopedia of Statistical Science. Springer, pp. 1094–1096. https://doi.org/10.1007/978-3-642-04898-2_455

Lazaric, A., Restelli, M., Bonarini, A. (2007). Reinforcement learning in continuous action spaces through sequential monte carlo methods. In: Platt JC, Koller D, Singer Y, Roweis ST (eds) Advances in Neural Information Processing Systems 20. In: Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007, Curran Associates, Inc., pp. 833–840

Lazaric, A., Ghavamzadeh, M., & Munos, R. (2012). Finite-sample analysis of least-squares policy iteration. *Journal of Machine Learning Research, 13,* 3041–3074.

Lee, K., Choi, S., Oh, S. (2018). Maximum causal Tsallis entropy imitation learning. In: Bengio S, Wallach HM, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada, pp. 4408–4418

Levine, S., Koltun, V. (2013). Guided policy search. In: Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16–21 June 2013, JMLR.org, JMLR Workshop and Conference Proceedings, vol. 28, pp. 1–9

Li, L., Lu, Y., Zhou, D. (2017). Provably optimal algorithms for generalized linear contextual bandits. In: Proceedings of the 34th International Conference on Machine Learning, PMLR, Proceedings of Machine Learning Research, vol. 70, pp. 2071–2080

Little, M. P., Heidenreich, W. F., & Li, G. (2010). Parameter identifiability and redundancy: theoretical considerations. *PloS ONE, 5*(1), e8915.

Metelli, A.M., Mutti, M., Restelli, M. (2018a). Configurable Markov decision processes. In: Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, PMLR, Proceedings of Machine Learning Research, vol. 80, pp. 3488–3497

Metelli, A. M., Papini, M., Faccio, F., & Restelli, M. (2018b). Policy optimization via importance sampling. *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3–8 December 2018* (pp. 5447–5459). Canada.: Montréal.

Metelli, A.M., Ghelfi, E., Restelli, M. (2019). Reinforcement learning in configurable continuous environments. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, PMLR, Proceedings of Machine Learning Research, vol. 97, pp. 4546–4555

Metelli, A. M., Papini, M., Montali, N., & Restelli, M. (2020). Importance sampling techniques for policy optimization. *Journal of Machine Learning Research, 21,* 141:1-141:75.

Morey, R. D., Romeijn, J. W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72,* 6–18.

Neu, G., Jonsson, A., Gómez, V. (2017). A unified view of entropy-regularized Markov decision processes. arXiv preprint arXiv:170507798

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J.A., Abbeel, P., Peters, J. (2018). An algorithmic perspective on imitation learning. Foundations and Trends in Robotics

Owen, A. B. (2013). *Monte Carlo theory, methods and examples*. Methods and Examples Art Owen: Monte Carlo Theory.

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

Petersen, K. B., Pedersen, M. S., et al. (2008). The matrix cookbook. *Technical University of Denmark, 7*(15), 510.

Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. London: John Wiley & Sons.

Rajeswaran, A., Lowrey, K., Todorov, E., Kakade, S. M., & (2017) Towards generalization and simplicity in continuous control. In: Guyon I, von Luxburg U, Bengio S, Wallach HM, Fergus R, Vishwanathan SVN, Garnett R (eds) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017(December), pp. 4–9, . (2017). Long Beach, CA, USA (pp. 6550–6561).

Ramponi, G., Likmeta, A., Metelli, A.M., Tirinzoni, A., Restelli, M. (2020). Truly batch model-free inverse reinforcement learning about multiple intentions. In: Chiappa S, Calandra R (eds) Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, PMLR, Online, Proceedings of Machine Learning Research, vol. 108, pp. 2359–2369

Reddy, S., Dragan, A.D., Levine, S. (2019). Sqil: Imitation learning via regularized behavioral cloning. arXiv preprint arXiv:190511108

Rényi, A. (1961). *On measures of entropy and information*. Hungarian Academy of Sciences Budapest Hungary: Technical report.

Rothenberg, T. J., et al. (1971). Identification in parametric models. *Econometrica, 39*(3), 577–591.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction. Adaptive computation and machine learning*. Cambridge: MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. In S. A. Solla, T. K. Leen, & K. Müller (Eds.), *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]* (pp. 1057–1063). Cambridge: The MIT Press.

Sutton, R.S., Szepesvári, C., Geramifard, A., Bowling, M.H. (2008). Dyna-style planning with linear function approximation and prioritized sweeping. In: McAllester DA, Myllymäki P (eds) UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9–12, 2008, AUAI Press, pp. 528–536

Vershynin, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In: Compressed Sensing, Cambridge University Press, pp. 210–268

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics, 9*(1), 60–62.

Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. The Annals of Probability, pp. 94–116

Ziebart, B.D., Maas, A.L., Bagnell, J.A., Dey, A.K. (2008). Maximum entropy inverse reinforcement learning. In: Fox D, Gomes CP (eds) Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13–17, 2008, AAAI Press, pp. 1433–1438

Ziebart, B.D., Bagnell, J.A., Dey, A.K. (2010). Modeling interaction via the principle of maximum causal entropy. In: Fürnkranz J, Joachims T (eds) Proceedings of the 27th International Conference on Machine Learning (ICML-10), June 21–24, 2010, Haifa, Israel, Omnipress, pp. 1255–1262