Proposition of Appendix 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 $\mathbb{E}_{s\sim d_\mu^{\pi}*}\left[\phi(s)\phi(s)^T\right]$ is non–singular along with the fact that the policy π_θ plays each action with positive probability for the Boltzmann policy.

Concentration Result 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 (details and derivation in Appendix).

Under Assumption and Assumption, let $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$ be a dataset of $n > 0$ independent samples, where $s_i \sim d_\mu^{\pi_{\theta^*}}$ and $a_i \sim \pi_{\theta^*}(\cdot|s_i)$. Let $\widehat{\theta} = arg\,min_{\theta\in\Theta}\widehat{\ell}(\theta)$ and $\theta^* = arg\,min_{\theta\in\Theta}\ell(\theta)$. If the empirical FIM:

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

has a positive minimum eigenvalue $\widehat{\lambda}_{\min} > 0$ for all $\theta \in \Theta$, then, for any $\delta \in [0,1]$, with probability at least $1-\delta$:

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

Furthermore, with probability at least $1-\delta$, individually:

$$\ell(\widehat{\theta})-\ell(\theta^*) \leqslant \frac{d^2\sigma^4}{\widehat{\lambda}_{\min}^2 n}\log\frac{2d}{\delta}$$
$$\widehat{\ell}(\theta^*)-\widehat{\ell}(\widehat{\theta}) \leqslant \frac{d^2\sigma^4}{\widehat{\lambda}_{\min}^2 n}\log\frac{2d}{\delta}.$$

The theorem shows that the L^2–norm of the difference between the maximum likelihood parameter $\widehat{\theta}$ and the true parameter θ^* concentrates with rate $\mathcal{O}(n^{-1/2})$ while the likelihood $\widehat{\ell}$ and its expectation ℓ concentrate with faster rate $\mathcal{O}(n^{-1})$. Note that the result assumes that the empirical FIM $\widehat{\mathcal{F}}(\theta)$ has a strictly positive eigenvalue $\widehat{\lambda}_{\min} > 0$. This condition can be enforced as long as the true Fisher matrix $\mathcal{F}(\theta)$ has a positive minimum eigenvalue $\lambda_{\min}$, i.e. under identifiability assumption (Lemma) and given a sufficiently large number of samples. Proposition of Appendix provides the minimum number of samples such that with probability at least $1-\delta$ it holds that $\widehat{\lambda}_{\min} > 0$.

Identification Rule Analysis The goal of the analysis of the identification rule is to find the critical value $c(1)$ so that the following probabilistic requirement is enforced.

Let $\delta \in [0,1]$. An identification rule producing $\widehat{I}$ is δ–*correct* if: $\Pr\left(\widehat{I} \neq I^*\right) \leqslant \delta$.

We denote with $\alpha = \frac{1}{d-d^*}\mathbb{E}\left[\left|\{i \notin I^* : i \in \widehat{I}_c\}\right|\right]$ the expected fraction of parameters that the agent does not control selected by the identification rule and with $\beta = \frac{1}{d^*}\mathbb{E}\left[\left|\{i \in I^* : i \notin \widehat{I}_c\}\right|\right]$ the expected fraction of parameters that the agent does control not selected by the identification rule. We now provide a result that bounds α and β and employs them to derive δ–correctness.

Let $\widehat{I}_c$ be the set of parameter indexes selected by the Identification Rule obtained using $n > 0$ i.i.d. samples collected with π_{θ^*}, with $\theta^* \in \Theta$. Then, under Assumption and Assumption, let $\theta_i^* = arg\,min_{\theta\in\Theta_i}\ell(\theta)$ for all $i \in \{1,...,d\}$ and $\nu = \min\left\{1, \frac{\lambda_{\min}}{\sigma^2}\right\}$. If $\widehat{\lambda}_{\min} \geqslant \frac{\lambda_{\min}}{2\sqrt{2}}$ and $\ell(\theta_i^*) - l(\theta^*) \geqslant c(1)$, it holds that:

$$\alpha \leqslant 2d\exp\left\{-\frac{c(1)\lambda_{\min}^2 n}{16d^2\sigma^4}\right\}$$
$$\beta \leqslant \frac{2d-1}{d^*}\sum_{i\in I^*}\exp\left\{-\frac{\left(l(\theta_i^*)-l(\theta^*)-c(1)\right)\lambda_{\min}\nu n}{16(d-1)^2\sigma^2}\right\}.$$

Furthermore, the Identification Rule is $((d-d^*)\alpha + d^*\beta)$–correct.

Since α and β are functions of $c(1)$, we could, in principle, employ Theorem to enforce a value δ, as in Definition, and derive $c(1)$. However, Theorem is not very attractive in practice as it holds under an assumption regarding the minimum eigenvalue of the FIM and the corresponding estimate, i.e. $\widehat{\lambda}_{\min} \geqslant \frac{\lambda_{\min}}{2\sqrt{2}}$, that cannot be verified in practice since $\lambda_{\min}$ is unknown. Similarly, the constants d^*, $l(\theta_i^*)$ and $l(\theta^*)$ are typically unknown. We will provide in Section a heuristic for setting $c(1)$.

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 $\omega_i \in \Omega$ where θ_i is maximally important for the optimal policy. This intuition is justified by Theorem, since to maximize the *power* of the test $(1-\beta)$, all other things being equal, we should maximize the log–likelihood gap $l(\theta_i^*) - l(\theta^*)$, i.e. parameter θ_i should be essential to justify the agent's behavior. Let $I \in \{1,...,d\}$ be a set of parameter indexes we want to test, our ideal goal is to find the environment ω_I such that:

$$\omega_I \in arg\,max_{\omega\in\Omega}\left\{l(\theta_I^*(\omega)) - l(\theta^*(\omega))\right\}, \quad (2)$$

where $\theta^*(\omega) \in arg\,max_{\theta\in\Theta}J_{\mathcal{M}_\omega}(\theta)$ and $\theta_I^*(\omega) \in arg\,max_{\theta\in\Theta_I}J_{\mathcal{M}_\omega}(\theta)$ are the parameters of the optimal policies in the environment $\mathcal{M}_\omega$ in Π_Θ and Π_{Θ_I} respectively. Clearly, given the samples $\mathcal{D}$ collected with a single optimal policy $\pi^*(\omega_0)$ in a single environment $\mathcal{M}_{\omega_0}$, solving problem (2) 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

Let $I \in \{1,...,d\}$ and $\overline{I} = \{1,...,d\}\backslash I$. For a vector $\mathbf{v}$, we denote with $\mathbf{v}|_I$ the vector obtained by setting to zero the