NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World

Pure Exploration under Mediators' Feedback

Riccardo Poiani RICCARDO.POIANI@POLIMI.IT
DEIB, Politecnico di Milano
Milano, Italy

Alberto Maria Metelli ALBERTOMARIA.METELLI@POLIMI.IT
DEIB, Politecnico di Milano
Milano, Italy

Marcello Restelli MARCELLO.RESTELLI@POLIMI.IT
DEIB, Politecnico di Milano
Milano, Italy

Abstract

Stochastic multi-armed bandits are a sequential-decision-making framework, where, at each interaction step, the learner selects an arm and observes a stochastic reward. Within the context of best-arm identification (BAI) problems, the goal of the agent lies in finding the optimal arm, i.e., the one with the highest expected reward, as accurately and efficiently as possible. Nevertheless, the sequential interaction protocol of classical BAI problems, where the agent has complete control over the arm being pulled at each round, does not effectively model several decision-making problems of interest (e.g., off-policy learning, human feedback). For this reason, in this work, we propose a novel strict generalization of the classical BAI problem that we refer to as best-arm identification under mediators' feedback (BAI-MF). More specifically, we consider the scenario in which the learner has access to a set of *mediators*, each of which selects the arms on the agent's behalf according to a stochastic and possibly *unknown* policy. The mediator, then, communicates back to the agent the pulled arm together with the observed reward. In this setting, the agent's goal lies in sequentially choosing which mediator to query to identify with high probability the optimal arm while minimizing the identification time, i.e., the sample complexity. To this end, we first derive and analyze a statistical lower bound on the sample complexity specific to our general mediator feedback scenario. Then, we propose a sequential decision-making strategy for discovering the best arm; as our theory verifies, this algorithm matches the lower bound both almost surely and in expectation.

Keywords: Pure Exploration, Stochastic Bandits

1. Introduction

Stochastic multi-armed bandits (Lattimore and Szepesvári, 2020) are a sequential decision-making framework where, during each interaction round, the learner selects an arm and observes a sample drawn from its reward distribution. Contrary to regret minimization problems, where the agent aims at maximizing the cumulative reward, in *best-arm identification* (BAI) scenarios (Even-Dar et al., 2002), the agent's primary focus lies in computing

A concurrent study of this setting is presented in Reddy et al. (2023). Our findings were derived independently and followed by different algorithmic choices and theoretical analyses.

the arm with the highest expected reward (i.e., the optimal arm) as accurately and efficiently as possible. More specifically, in the *fixed-confidence* setting, given a maximal risk parameter δ, the agent's primary focus is on identifying, with probability at least $1-\delta$, the optimal arm with a minimum number of samples. Nevertheless, the sequential interaction protocol of classical BAI settings, in which the agent has complete control of the arm being pulled at each round (i.e., at each step, the agent chooses which arm to query), fails to adequately represent various decision-making problems that are of importance. In fact, in some relevant scenarios, the agent possesses only partial or no control over the arms being played. Consider, indeed, the following examples.

- **Off-Policy Learning.** Off-policy learning is a crucial aspect of decision-making theory that has gathered significant attention, especially within the Reinforcement Learning (RL) community (Sutton and Barto, 2018). Here, the agent continuously observes, at each round, actions sampled from a fixed behavioral policy, together with the corresponding rewards. The goal, here, consequently, lies in exploiting these off-policy interactions to identify the best arm with high probability.

- **Active Off-Policy Learning.** This scenario generalizes the off-policy setting previously presented. In this case, multiple behavioral policies are available to the agent. The learner can decide which behavioral policy to query to quickly identify the optimal arm. In practice, these behavioral policies can be, for instance, those of experts with the skill necessary to perform a subset of actions within the arm set. Another relevant example might arise in scenarios with human feedback (Li et al., 2019), where multiple humans can perform actions on the agent's behalf according to some private and personal policy.

As we can see, these scenarios cannot be properly modeled with the usual bandit interaction protocol as the agent has limited or no control on the arms being pulled during each interaction round. For this reason, in this work, we study a strict generalization of the classical BAI framework that circumvents the limits of complete controllability that is typical of bandit frameworks. To this end, we introduce the best-arm identification problem under *mediators' feedback*, where the learner has access to a set of mediators, each of which will query arms on the agent's behalf according to some stochastic, *possibly unknown* and *fixed* behavioral policy. The mediator will then communicate back to the agent which action it has played, together with the observed reward realization. In this setting, the agent's goal lies in sequentially choosing which mediator to query to identify with high probability the optimal arm while minimizing the sample complexity. As one can verify, such formalism decouples the arms' pulls from the agent's choices, thus allowing to properly model all the scenarios depicted above.

2. Preliminaries and Backgrounds

2.1 Fixed-Confidence Best-Arm Identification

In fixed-confidence best-arm identification (BAI) problems (Even-Dar et al., 2002), the agent interacts with a set of K probability distributions $\boldsymbol{\nu} = (\nu_1, \dots \nu_K)$ with respective means $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$. For simplicity, we assume that there is a unique optimal arm, and, w.l.o.g., $\mu_1 > \mu_2 \geq \dots \geq \mu_K$. In the rest of this work, we consider distributions within the one-dimensional canonical exponential family (Cappé et al., 2013), which are directly parameterized by their mean.[1] For this reason, with a little abuse of notation, we will often refer to the bandit model $\boldsymbol{\nu}$ using the means of its arms $\boldsymbol{\mu}$. We use the symbol $\mathcal{M}$ to denote this class of bandit models with unique optimal arms. Given two distributions $p, q \in \mathcal{M}$, we denote with $d(p, q)$ the KL divergence between p and q.

We now proceed by formalizing the interaction scheme between the agent and the bandit model. At every interaction step $t \in \mathbb{N}$, the agent selects an arm $A_t \in [K]$ and receives a new and independent reward $X_t \sim \nu_{A_t}$. The procedure that defines how arms A_t are selected is often referred to as *sampling rule*. Given a maximal risk parameter $\delta \in (0, 1)$, the goal of the agent is to output the optimal arm $\hat{a}_{\tau_\delta} = \{1\}$ with probability at least $1 - \delta$, while minimizing the *sample complexity* $\tau_\delta \in \mathbb{N}$. More formally, τ_δ is a stopping time that controls the end of the data acquisition phase, after which a decision $\hat{a}_{\tau_\delta}$ is made. We refer to algorithms that satisfy $\mathbb{P}\left(\hat{a}_{\tau_\delta} \in \operatorname{argmax}_{a \in [K]} \mu_a\right) \leq \delta$ as δ-correct strategies.

We now describe in detail the statistical complexity of fixed-confidence BAI problems (Garivier and Kaufmann, 2016). Given a bandit model $\boldsymbol{\mu} \in \mathcal{M}$, let $a^*(\boldsymbol{\mu}) = \operatorname{argmax}_{a \in [K]} \mu_a$. We introduce the set $\operatorname{Alt}(\boldsymbol{\mu})$ as the set of problems where the optimal arm is different w.r.t. to $\boldsymbol{\mu}$, namely $\operatorname{Alt}(\boldsymbol{\mu}) \coloneqq \{\boldsymbol{\lambda} \in \mathcal{M} : a^*(\boldsymbol{\lambda}) \neq a^*(\boldsymbol{\mu})\}$. Let $\operatorname{kl}(x, y) = x \log(x/y) + (1-x) \log((1-x)/(1-y))$. Then, for any δ-correct algorithm it holds that $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq T^*(\boldsymbol{\mu}) \operatorname{kl}(\delta, 1-\delta)$, where $T^*(\boldsymbol{\mu})^{-1}$ is given by:

$$T^*(\boldsymbol{\mu})^{-1} = \sup_{\omega \in \Delta_K} \inf_{\boldsymbol{\lambda} \in \operatorname{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{K} \omega_a d(\mu_a, \lambda_a) \right). \tag{1}$$

We remark that, when $\delta \to 0$, $T^*(\boldsymbol{\mu})$ fully describes the statistical complexity of each problem $\boldsymbol{\mu}$. More specifically, it is possible to derive the following result: $\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \geq T^*(\boldsymbol{\mu})$. For this reason, $T^*(\boldsymbol{\mu})$ has played a crucial role in several BAI studies (e.g., Garivier and Kaufmann, 2016; Wang et al., 2021; Tirinzoni and Degenne, 2022). From Equation (1), we can see that $T^*(\boldsymbol{\mu})^{-1}$ can be seen as a max-min game where the first player chooses a pull proportion among the different arms, and the second player chooses a hard-to-identify alternative problem where the optimal arm is different (Degenne et al., 2019). In this sense, the unique maximizer of Equation (1), which we denote as $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$, can be interpreted as the optimal proportion with which arms should be queried in order to identify $a^*(\boldsymbol{\mu})$. Since solving Equation (1) requires access to quantities unknown to the learner, $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ often takes the name of *oracle weights*.

2.2 Best-Arm Identification under Mediators' Feedback

In this work, we study the following generalization of the best-arm identification problem. Given a bandit model $\boldsymbol{\nu}$ with K arms, the learner cannot directly sample rewards from each arm ν_a, but, instead it can query a set of E mediators, each of which is described by a *possibly unknown* and *fixed* behavioral policy $\boldsymbol{\pi_e} \in \Delta_K$. More specifically, at each interaction step $t \in \mathbb{N}$, the agent will select a mediator $E_t \in [E]$, which, on the agent's

[1] The reader who is not familiar with the subject may consider Bernoullian or Gaussian distributions with known variance.

behalf, will pull an arm $A_t \sim \boldsymbol{\pi_{E_t}}$ and will observe a reward $X_t \sim \nu_{A_t}$. The mediator E_t will then communicate back to the agent both the action A_t and observed reward X_t. For brevity, we adopt the symbol $\boldsymbol{\pi}$ as a shortcut for the set of mediators' policies $(\boldsymbol{\pi_e})_{e=1}^E$. Given a maximal risk parameter δ, the goal of the agent remains identifying with high-probability the optimal arm within $\boldsymbol{\mu}$ while minimizing the sample complexity τ_δ. To this end, we restict our study to the following scenarios.

Assumption 1 *For any $a \in [K]$ there exists $e \in [E]$ such that $\pi_e(a) > 0$.*

Assumption 1 states that the mediators' policies explores with positive probability each action $a \in [K]$. In other words, the agent should be able to gather information on each arm within the arm set. [2]

To conclude, we notice that the proposed interaction protocol is a strict generalization w.r.t. to the usual BAI framework. Indeed, whenever (i) the mediators' policies are known, (ii) $E = K$ and, (iii) for all action $a \in [K]$, $\boldsymbol{\pi_{E_a}}$ is a Dirac distribution on action a, we recover the usual best-arm identification problem. In the rest of this document, we refer to this peculiar set of mediators' policies as $\bar{\boldsymbol{\pi}}$.

3. On the Statistical Complexity

This section discusses the intrinsic statistical complexity of the best-arm identification problems under mediators' feedback. More specifically, we provide and analyze a lower bound on the sample complexity that is necessary to identify the optimal arm with high-probability.

Theorem 1 *Let $\delta \in (0,1)$. For any δ-correct strategy, any bandit model $\boldsymbol{\mu}$, and any set of mediators $\boldsymbol{\pi}$ it holds that $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta] \geq \mathrm{kl}(\delta, 1-\delta) T^*(\boldsymbol{\mu}, \boldsymbol{\pi})$, where $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ is defined as:*

$$\sup_{\boldsymbol{\omega} \in \Sigma_E} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{e=1}^{E} \omega_e \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right). \tag{2}$$

Theorem 1 deserves some comments. First of all, as we can appreciate from Equation (2), $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ reports the typical max-min game that describes lower bounds for standard best-arm identification problems. More specifically, the max-player determines the proportion with which each mediator should be queried, while the min-player decides an alternative (and hard) alternative instance in which the optimal arm is modified. It has to be remarked that $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$,and, consequently, the oracle weights $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$, directly depends on the set of mediators' policies $\boldsymbol{\pi}$. In other words, $\boldsymbol{\pi}$ plays a crucial role in the statistical complexity of the problem. To further investigate this dipendency, let us introduce some additional notation. Given $\boldsymbol{\omega} \in \Sigma_E$, we define $\tilde{\boldsymbol{\pi}}(\boldsymbol{\omega}) \in \Sigma_K$, where $\tilde{\pi}_a(\boldsymbol{\omega}) = \sum_{e=1}^E \omega_e \pi_e(a)$ denotes the probability of playing an arm a when sampling mediators according to $\boldsymbol{\omega}$. Then, let $\widetilde{\Sigma}_K \subseteq \Sigma_K$ be the set of all the possible $\tilde{\boldsymbol{\pi}}$ that can be obtained starting from any $\boldsymbol{\omega} \in \Sigma_E$. Given this notation, it is possible to rewrite $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ as:

$$\sup_{\tilde{\boldsymbol{\pi}} \in \widetilde{\Sigma}_K} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{a=1}^{K} \tilde{\pi}_a d(\mu_a, \lambda_a). \tag{3}$$

[2] We argue that this is a very mild requirement. Indeed, as we shall see in the appendix, Assumption 1 is necessary for finite sample complexity results.

At this point, we notice that Equation (3) shares significant similarities with the definition of $T^*(\boldsymbol{\mu})^{-1}$ for classical BAI problems; i.e., Equation (1). The only difference, indeed, stands in the fact that, under mediators' feedback, the max-player can only act on the restricted set $\widetilde{\Sigma}_K$ rather than the entire simplex Σ_K. In this sense, the max-min game is between the proportion of arm pulls that is *possible* to play according to the mediators $\boldsymbol{\pi}$, and the alternative hard instance. In the rest of this document, we denote maximizers of Equation (3) with $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. Given this interpreation of Theorem 1 we now proceed by further investigating the comparison with classical BAI problems.[3]

3.1 Comparison with classical BAI

First of all, it is worth noting that Theorem 1 effectively generalizes existing statistical complexity results of the typical BAI problem, thus offering a broader perspective. Indeed, whenever the set of mediators' policies is equal to $\bar{\boldsymbol{\pi}}$, Theorem 1 directly reduces to the usual BAI lower bound. In other words, $T^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}})^{-1}$ is exactly $T^*(\boldsymbol{\mu})^{-1}$. Furthermore, for a general set of mediators $\boldsymbol{\pi}$, it is possible to derive the following result.

Proposition 2 *For any bandit model $\boldsymbol{\mu}$ and mediators' policies $\boldsymbol{\pi}$ it holds that:*

$$T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1} \leq T^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}})^{-1}. \tag{4}$$

Furthermore, $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1} < T^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}})^{-1}$ holds if and only if $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}}) \notin \widetilde{\Sigma}_K$.

From Equation 4, we can see that the mediators' feedback problem is always at least as difficult as the classical BAI setting. From an intuitive perspective, this result is expected. Indeed, from Equation (3), we know that the only difference between $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ and $T^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}})^{-1}$ lies in the definition of $\widetilde{\Sigma}_K$, that, as previously discussed, encodes the partial controllability on the arm space that is introduced by the mediators $\boldsymbol{\pi}$. Furthermore, Proposition 2 fully characterizes the set of instances in which the mediators' feedback introduces additional challenges in identifying the optimal arm. More precisely, the lower bound of Theorem 1 separates from the one of classical BAI whenever the max-player cannot pull, in expectation, arms according to the proportion $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \bar{\boldsymbol{\pi}})$ that results from the lower bound of the classical BAI problem.

4. Track and Stop under Mediators' feedback

In this section, we continue by providing our algorithm for the best-arm identification problem under mediators feedback. Here, we focus on the case in which the mediators policies $\boldsymbol{\pi}$ are known to the learner.[4] As algorithm, we cast the Track and Stop (TaS) framework (Garivier and Kaufmann, 2016) to our interaction setting in the following way.[5]

[3] Further analysis on the statistical complexity are provided in the appendix.

[4] We refer the reader to the appendix for the case in which $\boldsymbol{\pi}$ is unknown. Nevertheless, we remark that, with a slight modification to the algorithm, it is possible to obtain identical theoretical results.

[5] We report a short description of the classical TaS algorithm in the appendix; for further details see Garivier and Kaufmann (2016).

Sampling Rule As a sampling rule, we adopt C-tracking (Garivier and Kaufmann, 2016) of the oracle mediator proportions $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. More formally, let $\hat{\boldsymbol{\mu}}(t)$ be the vector of estimates of the mean of each arm at time t. We then compute any maximizers of the empirical version of Equation (2) (i.e., $\boldsymbol{\mu}$ is replaced with $\hat{\boldsymbol{\mu}}$), and we L^∞ project it onto $\Sigma_E^{\epsilon_t} = \{\boldsymbol{\omega} \in \Sigma_E : \forall i \; \omega_i \geq \epsilon_t\}$, where ϵ_t is given by $\epsilon_t = (E^2 + t)^{-1/2}/2$. We notice that, in the original version of TaS, C-Tracking was applied to track optimal proportions between arms. Our algorithmic choice (i.e., tracking mediator proportions) is a direct consequence of the fact that we cannot directly track arm proportions (e.g., $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$), but, instead, the learner can only decide which mediator will be queried at time t.

Stopping Rule Since the goal lies in identifying the optimal arm $a \in [K]$, we stick to the successful Generalized Likelihood Ratio (GLR) statistic to decide when enough information has been gathered to confidently reccomend which arm has the highest mean (Garivier and Kaufmann, 2016).

The reccomendation For the same reasons of the stopping rule, we rely on the reccomendation rule of Garivier and Kaufmann (2016). Namely, our algorithm reccomend the arm with the highest empirical mean $\hat{a}_{\tau_\delta} = \operatorname{argmax}_{a \in [K]} \hat{\mu}_a(\tau_\delta)$.

4.1 Theoretical Results

At this point, we are ready to present our theoretical analysis on the performance of our algorithm. We begin by providing the following almost surely convergence result.

Theorem 3 *Consider any $\boldsymbol{\mu} \in \mathcal{M}$ and any $\boldsymbol{\pi}$ such that Assumption 1 is satisfied. Let $\alpha \in (1, e/2]$. It holds that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\limsup_{\delta \to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}) \right) = 1. \tag{5}$$

Similarly, it is possible to derive a result that directly controls the expectation of the stopping time τ_δ. More specifically, we prove the following result.

Theorem 4 *Consider any $\boldsymbol{\mu} \in \mathcal{M}$ and any $\boldsymbol{\pi}$ such that Assumption 1 is satisfied. Let $\alpha \in (1, e/2]$. It holds that:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}). \tag{6}$$

In other words, Theorem 4 shows that in the asymptotic regime of $\delta \to 0$, our algorithm matches the lower bound presented in Theorem 1.

Acknowledgments

This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

References

Shubhada Agrawal, Sandeep Juneja, and Peter Glynn. Optimal -correct best-arm selection for heavy-tailed distributions. In *Algorithmic Learning Theory*, pages 61–110. PMLR, 2020.

Jason M Altschuler, Victor-Emmanuel Brunel, and Alan Malek. Best arm identification for contaminated bandits. *J. Mach. Learn. Res.*, 20(91):1–39, 2019.

Jean-Pierre Aubin and Hélène Frankowska. *Set-valued analysis.* Springer Science & Business Media, 2009.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *COLT*, pages 41–53. Citeseer, 2010.

Francesco Bacchiocchi, Francesco Emanuele Stradi, Matteo Papini, Alberto Maria Metelli, and Nicola Gatti. Online adversarial mdps with off-policy feedback and known transitions. In *Sixteenth European Workshop on Reinforcement Learning*, 2023.

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, pages 23–37. Springer, 2009.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pages 1516–1541, 2013.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

Rémy Degenne and Wouter M Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Wouter M Koolen, and Pierre Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

Rémy Degenne, Pierre Ménard, Xuedong Shang, and Michal Valko. Gamification of pure exploration for linear bandits. In *International Conference on Machine Learning*, pages 2432–2442. PMLR, 2020.

Khaled Eldowa, Emmanuel Esposito, Tommaso Cesari, and Nicolò Cesa-Bianchi. On the minimax regret for online learning with feedback graphs. *arXiv preprint arXiv:2305.15383*, 2023.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and Markov decision processes. In *International Conference on Computational Learning Theory*, pages 255–270. Springer, 2002.

Yifan Feng, René Caldentey, and Christopher Thomas Ryan. Robust learning of consumer preferences. *Operations Research*, 70(2):918–962, 2022.

Germano Gabbianelli, Gergely Neu, and Matteo Papini. Online learning with off-policy feedback. In *International Conference on Algorithmic Learning Theory*, pages 620–641. PMLR, 2023.

Aurélien Garivier, Pierre Ménard, Laurent Rossi, and Pierre Menard. Thresholding bandit for dose-ranging: The impact of monotonicity. *arXiv preprint arXiv:1711.04454*, 2017.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

Kevin Jamieson and Robert Nowak. Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting. In *2014 48th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2014.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. *Advances in Neural Information Processing Systems*, 33:10007–10017, 2020.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, pages 1238–1246. PMLR, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 17(1):1–42, 2016. Publisher: JMLR. org.

Tomáš Kocák and Aurélien Garivier. Best arm identification in spectral bandits. *arXiv preprint arXiv:2005.09841*, 2020.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems*, 49(4):337–349, 2019.

Alberto Maria Metelli, Matteo Papini, Pierluca D'Oro, and Marcello Restelli. Policy optimization as online learning with mediator feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8958–8966, 2021.

Vrettos Moulos. Optimal best markovian arm identification with fixed confidence. *Advances in Neural Information Processing Systems*, 32, 2019.

Arpan Mukherjee, Ali Tajer, Pin-Yu Chen, and Payel Das. Mean-based best arm identification in stochastic bandits under reward contamination. *Advances in Neural Information Processing Systems*, 34:9651–9662, 2021.

Matteo Papini, Alberto Maria Metelli, Lorenzo Lupo, and Marcello Restelli. Optimistic policy optimization via multiple importance sampling. In *International Conference on Machine Learning*, pages 4989–4999. PMLR, 2019.

Kota Srinivas Reddy, P. N. Karthik, Nikhil Karamchandani, and Jayakrishnan Nair. Best arm identification in bandits with limited precision sampling, 2023.

Yoan Russac, Christina Katsimerou, Dennis Bohle, Olivier Cappé, Aurélien Garivier, and Wouter M Koolen. A/b/n testing with control in the presence of subpopulations. *Advances in Neural Information Processing Systems*, 34:25100–25110, 2021.

Rajat Sen, Karthikeyan Shanmugam, and Sanjay Shakkottai. Contextual bandits with stochastic experts. In *International Conference on Artificial Intelligence and Statistics*, pages 852–861. PMLR, 2018.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction.* MIT press, 2018.

Andrea Tirinzoni and Rémy Degenne. On elimination strategies for bandit fixed-confidence identification. *Advances in Neural Information Processing Systems*, 35:18586–18598, 2022.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821, 2021.

Junwen Yang and Yifan Feng. Nested elimination: A simple algorithm for best-item identification from choice-based feedback. *International Conference on Machine Learning*, 2023.

Appendix A: Related Works

In this section, we provide in-depth discussion with works that are related to ours.

Best-Arm Identification Since the seminal work of Even-Dar et al. (2002), the fixed-confidence BAI setting has gathered increasing attention within the community. In particular, considerable efforts have been dedicated to refining algorithms and statistical lower-bounds, all with the ultimate objective of constructing optimal identification strategies (e.g., Bubeck et al., 2009; Audibert et al., 2010; Karnin et al., 2013; Jamieson et al., 2014; Jamieson and Nowak, 2014; Kaufmann et al., 2016). In this context, and of particular relevance for our work, Garivier and Kaufmann (2016) have proposed the celebrated Track and Stop (TaS) algorithm, which attains *optimal* statistical complexity in the asymptotic regime, i.e., $\delta \to 0$. Building upon this work, numerous studies have been conducted to propose improvements and generalizations upon the TaS algorithm (e.g., Degenne and Koolen, 2019; Wang et al., 2021; Degenne et al., 2020; Tirinzoni and Degenne, 2022).

Structured Best-Arm Identification One of the key factors that contributed to the success of TaS is its ability to emerge as a versatile framework that can be meticulously adapted to several variants of the BAI problem, such as linear (Jedra and Proutiere, 2020) and spectral bandits (Kocák and Garivier, 2020), multiple answers problems (Degenne and Koolen, 2019), and many others (e.g., Garivier et al., 2017; Moulos, 2019; Agrawal et al., 2020). Among problems with additional structure, our work is related to BAI under choice-based feedback (Feng et al., 2022; Yang and Feng, 2023), where a company sequentially shows sets of items to a population of customers and collects their choices. The objective is to identify the most preferred item with the least number of samples and with high-probability. Another relevant work is Russac et al. (2021), where the authors study the BAI problem in the presence of sub-populations. In more precise terms, the authors make the assumption that a population can be divided into distinct and similar subgroups. During each time step, one of these subgroups is sampled and an action (i.e., arm) is chosen . The observed outcome is a random draw from the selected arm, considering the characteristics of the current subgroup. To evaluate the effectiveness of each arm, a weighted average of its subpopulation means is used. Finally, our feedback structure is also related to BAI in contaminated bandits (Altschuler et al., 2019; Mukherjee et al., 2021), where each arm pull has a probability ϵ of generating a sample from an arbitrary distribution, rather than the true one. Nevertheless, we remark that none of these settings can be mapped to the mediators' feedback one and viceversa.

Mediators' Feedback The mediator feedback terminology was introduced by Metelli et al. (2021) in the context of Policy Optimization (PO) in RL. Similar to the previous studies of Papini et al. (2019), the authors deal with the PO problem as a bandit where each policy in a given set is mapped to a distinct arm, whose reward is given by the usual cumulative RL return. Notice that, in this setting, the ability to perform actions in the environment is mediated by the policy set of the agent. For this reason, in our work, we adopt their terminology to disentangle the arms' pull from the agent's choices. Among this line of works, we notice that, recently, a variant of this problem has also been studied in the context of non-stochastic bandits with expert advice (Eldowa et al., 2023). Here, during each round, the learner selects an expert that will perform an action on the agent's behalf according to some fixed distribution. Similar ideas have also been investigated in Sen et al. (2018) for regret minimization in contextual bandits. More specifically, they assume access to a class of stochastic experts, where each expert is a conditional distribution over the arms given the context. Compared to our work both Sen et al. (2018) and Eldowa et al. (2023) consider the problem of minimizing the regret against the best expert.

Other Related Works Off-policy learning plays a vital role in decision-making theory and has garnered considerable interest, particularly in RL (Sutton and Barto, 2018). In particular, the off-policy feedback has received extensive research in the offline RL literature (Levine et al., 2020), where the agent lacks the ability to directly interact with the environment and is instead limited to utilizing a fixed dataset gathered by possible multiple and unknown behaviorial policies. Finally, related to our work, Gabbianelli et al. (2023) have studied regret minimization in adversarial bandits with off-policy feedback. More specifically, the authors assume that the learner cannot directly observe its rewards, but instead sees the ones obtained by a behavioral and unknown policy that runs in parallel. Similar ideas have also been extended to the MDP setting with known transitions in Bacchiocchi et al. (2023).

Appendix B: Further Details on Track and Stop

The seminal work of Garivier and Kaufmann (2016) has presented the Track and Stop (TaS) algorithm, which is the first asymptotically optimal approach for the fixed confidence BAI scenario; i.e., when $\delta \to 0$, it guarantees to stop with sample complexity that matches the lower bound. The core idea behind TaS lies in solving an empirical version of Equation (1) to estimate the optimal oracle weights. Then, in order to match the optimal proportions $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ (which guarantees to achieve optimality), the sampling rule will allocate samples by *tracking* this empirical estimation. We remark that this is combined with a forced exploration sampling strategy that ensures that the estimate of the mean of each arm, and consequently the estimate of $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$, is sufficiently accurate. Lastly, as a stopping criterion, TS employs the Generalized Likelihood Ratio (GLR) statistic to determine if enough information has been collected to infer, with a risk not exceeding δ, whether the mean of one arm is greater than that of all the others. More specifically, the algorithm stops whenever the following condition is verified:

$$Z(t) \coloneqq t \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}})} \sum_{a=1}^{K} \frac{N_a(t)}{t} d(\hat{\mu}_a(t), \lambda_a) \geq \beta(t, \delta), \tag{7}$$

where $N_a(t)$ denotes the number of pulls to arm a at time t, and $\beta(t, \delta)$ represents an exploration rate that is commonly set to $\log\left(\frac{Ct^\alpha}{\delta}\right)$, for some $\alpha > 1$ and appropriate constanct C.[6]

[6] We refer the reader to the original work of Garivier and Kaufmann (2016) for a formal exposition of the GLR statistic and its use within pure exploration problems.

Appendix C: Further Analysis on the Statistical Complexity

In this section, we provide additional analysis on the statistical complexity of best arm identification under mediators' feedback. All statements will be formally proven in Appendix E.

On the Action Covering Assumption

We now continue with a formal justification behind the action covering assumption, i.e., Assumption 1. To this end, we analyze the behavior of Theorem 1 under the peculiar single mediator setting, that is $E = 1$. More specifically, let us focus on the case in which there are two different actions, a_1 and a_2, associated to Gaussian reward distributions with unitary variance and means $\mu_{a_1} > \mu_{a_2}$. In this case, $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ reduces to:

$$\frac{1}{2} \frac{\pi_e(a_1)\pi_e(a_2)}{\pi_e(a_1) + \pi_e(a_2)} \Delta^2, \tag{8}$$

where $\Delta = \mu_{a_1} - \mu_{a_2}$. In this context, it is easy to see that, as soon as $\pi_e(a) \to 1$ for any of the two actions, $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ tends to 0, and, consequently, $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta] \to +\infty$. In this sense, we can appreciate as Assumption 1 turns out being a necessary assumption for finite sample complexity result. This should come as no surprise: if we cannot observe any realization from a certain arm $a \in [K]$, we are unable to conclude whether a is optimal or not.

Off-Policy Learning

Given the significant importance of off-policy learning within the sequential decision-making community, we now provide additional details on the lower bound for the case in which $E = 1$. We notice, indeed, that whenever $E = 1$, our setting reduces to the off-policy best-arm identification problem, where the learner continuously observes actions and rewards from another agent (i.e, the mediator). In this case, assuming for the sake of exposition Gaussian distributions with unitary variance, $T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1}$ can be rewritten as:

$$T^*(\boldsymbol{\mu}, \boldsymbol{\pi})^{-1} = \min_{a \neq 1} \frac{1}{2} \frac{\pi_e(1)\pi_e(a)}{\pi_e(1) + \pi_e(a)} \Delta_a^2, \tag{9}$$

where $\Delta_a = \mu_1 - \mu_a$. Equation (9) expresses the lower bound *only* in term of the most difficult to identify alternative arm (i.e., the minimum over the different sub-optimal actions). Furthermore, contrary to what usually happens in classical BAI problems, this difficult to identify alternative arm is not the one with the smallest gap Δ_a, but there is a trade-off between Δ_a and how easy it is to observe the mediator playing action a, namely $\pi_e(a)$.

On $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$

Finally, we conclude with some more technical considerations on $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. It has to be noticed that, compared to standard BAI problems, $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ are, in general, not unique. In other words, the mappings $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ are set-valued. [7] As shown by previous works (Degenne and Koolen, 2019), this sort of feature

[7] As a simple example, it is sufficient to consider the case in which linearly-dependent policies are present in $\boldsymbol{\pi}$.

introduces significant challenges within the algorithmic design/analysis of, e.g., Track and Stop inspired algorithms. The following proposition shows significant and technical relevant properties that help overcoming these challenges.

Proposition 5 *The sets $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ are convex. Furthermore, the mappings $(\boldsymbol{\mu}, \boldsymbol{\pi}) \to \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ and $(\boldsymbol{\mu}, \boldsymbol{\pi}) \to \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ are upper hemicontinuous.*

The unfamiliar reader might think of upper hemicontinuity as a generalization of the continuity property for set-valued mappings (Aubin and Frankowska, 2009). [8] As our analysis will reveal, Proposition 5 will play a crucial role for the analysis of our algorithmic solution.

Appendix D: Unknown Mediators' policies

In this section, we extend our results to the case in which the agent does not know $\boldsymbol{\pi}$, but instead it has learn it directly from data. All theoretical results are formally proven in the Appendix E.

Before detailing our theoretical findings, we notice that Theorem 1 still represents a valid lower bound for this more intricated setting. For this reason, one might be tempted to extend our algorithm to track the optimal mediators proportions $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ to the case in which the set $\boldsymbol{\pi}$ is unknown to the learner. To this end, let $\hat{\boldsymbol{\pi}}(t)$ be the matrix containing empirical estimates for each mediator policy $\boldsymbol{\pi_e}$. Then, it is sufficient to modify the C-tracking sampling rule presented in Section 4 by computing any maximizer of the empirical version of Equation (2) where *both* $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$ are replaced with $\hat{\boldsymbol{\mu}}(t)$ and $\hat{\boldsymbol{\pi}}(t)$ respectively. [9] As we shall now see, this simple modification allows us to derive results that are equivalent to Theorem 3 and 4. More specifically, we begin by showing the following almost surely convergence result.

Theorem 6 *Consider any $\boldsymbol{\mu} \in \mathcal{M}$ and any $\boldsymbol{\pi}$ such that Assumption 1 is satisfied. Let $\boldsymbol{\pi}$ be unknown to the learner prior to interacting with the environment. Let $\alpha \in (1, e/2]$. It holds that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\limsup_{\delta \to 0} \frac{\tau_\delta}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}) \right) = 1. \tag{10}$$

Furthermore, as done in the previous section, it is possible to derive a result that directly controls the expectation of the stopping time τ_δ.

Theorem 7 *Consider any $\boldsymbol{\mu} \in \mathcal{M}$ and any $\boldsymbol{\pi}$ such that Assumption 1 is satisfied. Let $\boldsymbol{\pi}$ be unknown to the learner prior to interacting with the environment. Let $\alpha \in (1, e/2]$. It holds that:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}). \tag{11}$$

[8] Further technical details on this point are deferred to the appendix.

[9] We notice that, in the previous section, $\boldsymbol{\mu}$ was replaced with $\hat{\boldsymbol{\mu}}(t)$ while $\boldsymbol{\pi}$ was used directly since it was available to the learner.

We now proceed by analyzing the results of Theorems 6 and 7. First of all, as we can appreciate, they fully extend the results of Theorems 3 and 4 to the unknown policies setting. Notice, in particular, that the theoretical results of the unknown policy setting, i.e., Equations (10) and (11), are *completely equivalent* to the ones previously presented for the case in which $\boldsymbol{\pi}$ is available to the learner, i.e., Equations (5) and (6). Furthermore, as a direct consequence of the fact that Theorem 1 represents a lower bound to the problem, it follows that the simple modification that we presented at the beginning of this section, is sufficient to derive an *asymptotically optimal* algorithm even in the case in which $\boldsymbol{\pi}$ is not available to the learner. Most importantly, these considerations implies that not knowing $\boldsymbol{\pi}$ does not affect the statistical complexity of the problem, at least in the asymptotic regime $\delta \to 0$. As a direct consequence, all the analysis and discussion we presented in the lower bound section hold equivalently both for the known and the unknown policy settings.

Appendix E: Proofs and derivations

Statistical Complexity

In this section, we derive claims concerning the statistical complexity of the problem. We begin by proving Theorem 1.

Theorem 1 *Let $\delta \in (0,1)$. For any δ-correct strategy, any bandit model $\boldsymbol{\mu}$, and any set of mediators $\boldsymbol{\pi}$ it holds that $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\tau_\delta\right] \geq \mathrm{kl}(\delta, 1-\delta) T^*(\boldsymbol{\mu},\boldsymbol{\pi})$, where $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1}$ is defined as:*

$$\sup_{\boldsymbol{\omega} \in \Sigma_E} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \left(\sum_{e=1}^{E} \omega_e \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right). \tag{2}$$

Proof Consider an instance $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$. It is easy to see that, from Lemma 1 in Kaufmann et al. (2016) that:

$$\sum_{e=1}^{E} \left(\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[N_e(\tau_\delta)\right] \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right) \geq \mathrm{kl}(\delta, 1-\delta),$$

where $N_e(t)$ denotes the number of pulls to mediator e at time t. Following Garivier and Kaufmann (2016), we notice that the previous Equation holds for all $\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})$. Therefore, we have that:

$$\begin{aligned} \mathrm{kl}(\delta, 1-\delta) &\leq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta] \sum_{e=1}^{E} \left(\frac{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[N_e(\tau_\delta)\right]}{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta]} \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right) \\ &\leq \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta] \sup_{\boldsymbol{\omega} \in \Sigma_E} \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{e=1}^{E} \left(\omega_e \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right), \end{aligned}$$

thus concluding the proof. ■

Given Theorem 1, we notice that, from Lemma 3 in Garivier and Kaufmann (2016), we have that $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1}$ can be rewritten as:

$$\min_{a\neq 1}\left(\sum_{e=1}^{E}(\pi_e(1)+\pi_e(a))\right) I_{\frac{\sum_{e=1}^{E}\pi_e(1)}{\sum_{e=1}^{E}(\pi_e(1)+\pi_e(a))}}(\mu_1,\mu_a), \tag{12}$$

where $I_\alpha(\mu_1,\mu_2) \coloneqq \alpha d(\mu_1,\alpha\mu_1+(1-\alpha)\mu_2)+(1-\alpha)d(\mu_2,\alpha\mu_1+(1-\alpha)\mu_2)$ denotes a generalized version of the Jensen-Shannon divergence. We notice that, for Gaussian distributions with unitary variance, Equation (12), reduces to (see, e.g., Appendix A.4 in Garivier and Kaufmann, 2016):

$$\min_{a\neq 1}\frac{1}{2}\frac{\left(\sum_{e=1}^{K}\pi_e(1)\right)\left(\sum_{e=1}^{K}\pi_e(a)\right)}{\left(\sum_{e=1}^{K}\pi_e(1)+\pi_e(a)\right)}\Delta_a^2, \tag{13}$$

from which the proof of Equation (8) and (9) follows directly.

At this point, we continue by proving Proposition 2.

Proposition 8 *For any bandit model $\boldsymbol{\mu}$ and mediators' policies $\boldsymbol{\pi}$ it holds that:*

$$T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} \leq T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}. \tag{4}$$

Furthermore, $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} < T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$ holds if and only if $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}}) \notin \widetilde{\Sigma}_K$.

Proof By begin by proving Equation (4). From Equation (3), we have that, for any mediators' policies $\boldsymbol{\pi}$, the following equality holds:

$$T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} = \sup_{\tilde{\boldsymbol{\pi}}\in\widetilde{\Sigma}_K}\inf_{\boldsymbol{\lambda}\in \mathrm{Alt}(\boldsymbol{\mu})}\sum_{a=1}^{K}\tilde{\pi}_a d(\mu_a,\lambda_a).$$

At this point, we notice that, whenever we consider $\bar{\boldsymbol{\pi}}$, $\widetilde{\Sigma}_K$ is equal to Σ_K. The result follows by noticing that $\widetilde{\Sigma}_K \subseteq \Sigma_K$ for any mediators' policies $\boldsymbol{\pi}$.

We now continue by showing that $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} < T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$ holds if and only if $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}}) \notin \widetilde{\Sigma}_K$. First of all, suppose that $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} < T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$ holds. However, if $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}}) \in \widetilde{\Sigma}_K$ holds as well, then we would have that $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} \geq T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$ by definition of $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})$. Therefore, $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}}) \notin \widetilde{\Sigma}_K$. On the other hand, if $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}}) \notin \widetilde{\Sigma}_K$ holds, than $T^*(\boldsymbol{\mu},\boldsymbol{\pi})^{-1} < T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$ follows from the fact that $\boldsymbol{\omega}^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})$ is the unique maximizer of $T^*(\boldsymbol{\mu},\bar{\boldsymbol{\pi}})^{-1}$. ■

Finally, before proving Proposition 5, we first note that the corresponding $\boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ are, in general, not unique. As a simple example, it is sufficient to consider the case in which linearly-dependent policies are present in $\boldsymbol{\pi}$. At this point, we continue with the proof of Proposition 5.

Proposition 9 *The sets $\boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ and $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ are convex. Furthermore, the mappings $(\boldsymbol{\mu},\boldsymbol{\pi}) \to \boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ and $(\boldsymbol{\mu},\boldsymbol{\pi}) \to \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ are upper hemicontinuous.*

Proof First of all, we prove the convexity of the sets. We begin by recalling the definition of $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$:

$$\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi}) = \operatorname*{argmax}_{\boldsymbol{\omega} \in \Sigma_E} \inf_{\boldsymbol{\lambda} \in \Lambda(\boldsymbol{\mu})} \left(\sum_{e=1}^{E} \omega_e \sum_{a=1}^{K} \pi_e(a) d(\mu_a, \lambda_a) \right).$$

In other words $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ is the set of maximizers of an infimum over linear functions (which is well-known to be concave). For this reason $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ is convex. A similar reasoning can be applied for $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. [10] At this point, we proceed with the upper-hemicontinuity. First of all, consider:

$$f\left(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\omega}\right) = \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{e} \omega_e \sum_{a} \pi_e(a) d(\mu_a, \lambda_a).$$

The function $f : \mathcal{M} \times (\Delta_K)^E \times \Delta_E \to \mathbb{R}$ is continuous. To see this, from Equation (12), we can rewrite f as:

$$f\left(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\omega}\right) = \min_{a \neq 1} \left(\sum_{e} \omega_e (\pi_e(1) + \pi_e(a)) \right) I_{\frac{\sum_e \omega_e \pi_e(1)}{\sum_e \omega_e(\pi_e(1) + \pi_e(a))}} \left(\mu_1, \mu_a\right).$$

Therefore, f can be expressed as a minimum over continous functions. It follows that f is continuous as well. At this point, the proof follows from an application of the Berge's Theorem (Aubin and Frankowska, 2009) (see e.g., Theorem 22 in Degenne and Koolen (2019)). Adopting the same notation as in Degenne and Koolen (2019), consider $\mathbb{X} = \mathcal{M} \times (\Delta_K)^E$, $\mathbb{Y} = \Delta_E$, $\phi(\boldsymbol{\mu}, \boldsymbol{\pi}) = \Delta_E$, and $u((\boldsymbol{\mu}, \boldsymbol{\pi}), \boldsymbol{\omega}) = f(\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\omega})$. At this point, we notice that ϕ is compact valued and continuous (since it is constant), while u is continuous. Therefore, due to Berge's Theorem $(\boldsymbol{\mu}, \boldsymbol{\pi}) \to \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ is upper hemicontinous and compact-valued. An identical reasoning can be applied for $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ replacing $\mathbb{Y}$ and $\phi(\boldsymbol{\mu}, \boldsymbol{\pi})$ with $\widetilde{\Sigma}_K$. ■

Helper Lemmas

Before diving into the details of our analysis, we report a known result on C-tracking.

Lemma 10 (Lemma 7 in Garivier and Kaufmann (2016)) *For all $t > 1$ and $e \in [E]$, the C-tracking rules ensures that $N_e(t) \geq \sqrt{t + E^2} - 2E$. Furthermore, consider a sequence $(\boldsymbol{\omega}_t)$ such $\boldsymbol{\omega}_t \in \boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(t), \boldsymbol{\pi})$ for all t. Then, C-tracking ensures that:*

$$||N_t^E - \sum_{s=0}^{t-1} \boldsymbol{\omega}_s||_\infty \leq E(1 + \sqrt{t}),$$

where $N_t^E = (N_1(t), \dots, N_E(t))$ denotes the number of pulls to each mediator at time t.

Furthermore, we notice that the fact $\mathbb{P}_{\boldsymbol{\mu}, \boldsymbol{\pi}}\left(\tau_\delta < +\infty, \hat{a}_{\tau_\delta} \neq a^*\right) \leq \delta$ holds, is a direct consequence of Assumption 1, forced-exploration, and Proposition 12 in Garivier and Kaufmann (2016).

[10] This argument was used, for instance, in Degenne and Koolen (2019).

Algorithm analysis (Known Mediators' Policies; Almost Surely Convergence)

At this point, we proceed by deriving the almost surely convergence result (i.e, Theorem 3).

Lemma 11 *Consider a sequence of $(\hat{\boldsymbol{\mu}}(t))_{t\in\mathbb{N}}$ that converges almost surely to $\boldsymbol{\mu}$. For all $t \in \mathbb{N}$, let $\boldsymbol{\omega}_t \in \boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(t), \boldsymbol{\pi})$ be arbitrary oracle weights for $\hat{\boldsymbol{\mu}}(t)$ and $\boldsymbol{\pi}$. Then, the following holds:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left(\lim_{t\to+\infty} \inf_{\boldsymbol{\omega}\in\boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty = 0\right) = 1$$

Proof The proof follows the one of Lemma 6 of Degenne and Koolen (2019).

Let $\mathcal{E}$ be the following event:

$$\mathcal{E} = \{\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}\}.$$

Event $\mathcal{E}$ holds by assumption with probability 1. Due to Proposition 5, we also have that there for all $\epsilon > 0$, there exists $\xi > 0$ such that if $\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_\infty \le \xi$ holds, then for all $\boldsymbol{\omega}_t \in \boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(t), \boldsymbol{\pi})$, $\inf_{\boldsymbol{\omega}\in\boldsymbol{\omega}^*(\boldsymbol{\omega},\boldsymbol{\pi})} \|\boldsymbol{\omega}_t - \boldsymbol{\omega}\|_\infty$ holds as well. At this point, we also notice that, on $\mathcal{E}$, for any $\xi > 0$, there exists t_0 such that for all $t \ge t_0$, $\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_\infty \le \xi$ holds.

At this point, for any $\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\omega})$, we have that, for all $t \ge t_0$:

$$\left\| \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty \le \frac{t_0}{t} + \frac{t-t_0}{t} \left\| \frac{1}{t-t_0}\sum_{t=t_0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|.$$

Taking infimums and using the convexity of $\boldsymbol{\omega}^*(\boldsymbol{\mu})$ (Proposition 5), together with Lemma 33 of Degenne and Koolen (2019), we have that:

$$\inf_{\boldsymbol{\omega}\in\boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty \le \frac{t_0}{t} + \epsilon, \tag{14}$$

which concludes the proof. ■

A direct consequence of Lemma 11 is the following one.

Lemma 12 *Consider a sequence of $(\hat{\boldsymbol{\mu}}(t))_{t\in\mathbb{N}}$ generated while following the C-tracking sampling strategy. Then, it holds that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left(\lim_{t\to+\infty} \inf_{\boldsymbol{\omega}\in\boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty = 0\right) = 1.$$

Proof Consider any $\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. Then, for any sequence $(\boldsymbol{\omega}_t)_{t\in\mathbb{N}}$ such that $\boldsymbol{\omega}_t \in \boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}, \boldsymbol{\pi})$, we have that:

$$\begin{aligned}\left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty &\le \left\| \frac{N_t^E}{t} - \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s \right\|_\infty + \left\| \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty \\ &\le \frac{E(1+\sqrt{t})}{t} + \left\| \frac{1}{t}\sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty,\end{aligned}$$

where in the second inequality we have used Lemma 10. Then, taking infimums and applying Lemma 11 concludes the proof. ■

At this point, we are ready to state the main Lemma that allows us to match, almost surely, the expected lower bound on the sample complexity.

Lemma 13 *Consider a sequence of $(\hat{\boldsymbol{\mu}}(t))_{t \in \mathbb{N}}$ generated while following the C-tracking sampling strategy. Then, it holds that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\lim_{t \to +\infty} \inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}} \right\|_\infty = 0 \right) = 1.$$

Proof We need to show that, with probability 1, for all $\epsilon > 0$, there exists t_ϵ such that for all $t \geq t_\epsilon$ the following holds:

$$\inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}} \right\|_\infty \leq \epsilon. \tag{15}$$

We notice that a sufficient condition for Equation (15) to hold is that it holds for some policy $\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ (i.e., not necessary the one that attains the infimum).

At this point, we proceed with some considerations. First of all, we know that, due to the law of large numbers, and the fact that $N_e(t) \to +\infty$, we have that $\frac{N_(e,a)(t)}{N_e(t)} \to \pi_e(a)$ with probability 1. More precisely, with probability 1, for all $\epsilon_1 > 0$, there exists t_{ϵ_1} such that for all $t \geq t_{\epsilon_1}$ the following holds:

$$\frac{N_{e,a}(t)}{N_e(t)} \in \left[\pi_e(a) - \epsilon_1, \pi_e(a) + \epsilon_1 \right].$$

Furthermore, due to Lemma 12, we have that, with probability 1, for all $\epsilon_2 > 0$, there exists t_{ϵ_2}, such that for all $t \geq t_{\epsilon_2}$ the following holds:

$$\inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_e(t)}{t} - \omega_e \right\|_\infty \in \left[-\epsilon_2, \epsilon_2 \right].$$

In other words, let $\boldsymbol{\omega}_t \in \operatorname{argmin}_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_e(t)}{t} - \omega_e \right\|_\infty$. The following holds:

$$\frac{N_e(t)}{t} \in \left[\omega_{t,e} - \epsilon_2, \omega_{t,e} + \epsilon_2 \right].$$

At this point, focus on Equation (15). Let $\tilde{\boldsymbol{\pi}}$ be any policy within $\tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$. Then, Equation (15) is satisfied whenever for all actions $a \in \mathcal{A}$ the following holds:

$$\frac{N_a(t)}{t} - \tilde{\pi}(a) \leq \epsilon,$$

and

$$-\frac{N_a(t)}{t} + \tilde{\pi}(a) \leq \epsilon,$$

holds. Let us focus on $\frac{N_a(t)}{t} - \tilde{\pi}(a) \le \epsilon$ (the case of $-\frac{N_a(t)}{t} + \tilde{\pi}(a) \le \epsilon$ is almost identical):

$$\frac{N_a(t)}{t} - \tilde{\pi}(a) = \sum_e \frac{N_{a,e}(t)}{N_e(t)} \frac{N_e(t)}{t} - \sum_e \tilde{\omega}_e \pi_e(a),$$

where $\tilde{\boldsymbol{\omega}}$ is any oracle weight that induces $\tilde{\boldsymbol{\pi}}$. With probability 1, however, we have that:

$$\begin{aligned}\frac{N_a(t)}{t} - \tilde{\pi}(a) &= \sum_e \frac{N_{a,e}(t)}{N_e(t)} \frac{N_e(t)}{t} - \sum_e \tilde{\omega}_e \pi_e(a) \\ &\le \sum_e (\pi_e(a) + \epsilon_1)(\omega_{t,e} + \epsilon_2) - \sum_e \tilde{\omega}_e \pi_e(a).\end{aligned}$$

At this point, we notice that the previous equation holds for any oracle weight $\tilde{\boldsymbol{\omega}}$ that induces $\tilde{\boldsymbol{\pi}}$, and furthermore, it also holds for any $\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$. It thus sufficies to pick $\tilde{\boldsymbol{\omega}}$ equal to $\boldsymbol{\omega}_t$ to obtain the following:

$$\frac{N_a(t)}{t} - \tilde{\pi}_t(a) \le \sum_e \pi_e(a)\epsilon_2 + \epsilon_1 \omega_{t,e} + \epsilon_1 \epsilon_2,$$

which concludes the proof.

■

We are now ready to prove Theorem 3.

Theorem 3 *Consider any $\boldsymbol{\mu} \in \mathcal{M}$ and any $\boldsymbol{\pi}$ such that Assumption 1 is satisfied. Let $\alpha \in (1, e/2]$. It holds that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\limsup_{\delta \to 0} \frac{\tau_\delta}{\log(1/\delta)} \le \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}) \right) = 1. \tag{5}$$

Proof Consider the following event:

$$\mathcal{E} = \left\{ \lim_{t \to +\infty} \inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}} \right\|_\infty = 0 \text{ and } \hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu} \right\}$$

Due to Lemma 13, the sampling strategy, the assumption on the mediators's policies, and the law of large numbers we know that $\mathcal{E}$ is of probability 1. Therefore, there exists t_0 such that for all $t \ge t_0$, $\hat{\mu}_1(t) > \max_{a \neq 1} \hat{\mu}_a(t)$ and, consequently:

$$Z(t) = t \left[\min_{a \neq 1} \left(\frac{N_1(t)}{t} + \frac{N_a(t)}{t} \right) I_{\frac{N_1(t)/t}{N_1(t)/t + N_a(t)/t}} (\hat{\mu}_1(t)), \hat{\mu}_a(t)) \right].$$

Let $\tilde{\boldsymbol{\pi}}_t \in \operatorname{arginf}_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}} \right\|_\infty$. For all $\epsilon > 0$, there exists $t_1 \ge t_0$ such that for all $t \ge t_1$ and all action $a \in \mathcal{A} \setminus \{1\}$ the following holds:

$$\left(\frac{N_1(t)}{t} + \frac{N_a(t)}{t} \right) I_{\frac{N_1(t)/t}{N_1(t)/t + N_a(t)/t}} (\hat{\mu}_1(t)), \hat{\mu}_a(t)) \ge \frac{\tilde{\pi}_t(1) + \tilde{\pi}_t(a)}{1 + \epsilon} I_{\frac{\tilde{\pi}_t(1)}{\tilde{\pi}_t(1) + \tilde{\pi}_t(a)}} (\mu_1, \mu_a).$$

Therefore, since $\tilde{\boldsymbol{\pi}}_t \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$, for all $t \geq t_1$ we have that:

$$Z(t) \geq \frac{t}{(1+\epsilon)T^*(\boldsymbol{\mu}, \boldsymbol{\pi})}.$$

The rest of the proof follows unchanged w.r.t. Proposition 13 in Garivier and Kaufmann (2016).

■

Algorithm analysis (Known Mediators' Policies; Expectation)

In order to prove Theorem 4, we begin with some concentration events analysis.

First of all, let $h(T) = T^{1/4}$, and $\epsilon > 0$. Define:

$$\mathcal{E}_T = \bigcap_{t=h(T)}^{T} \left(\|\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}\|_\infty \leq \xi \right),$$

where ξ is such that:

$$\left\|\boldsymbol{\mu}' - \boldsymbol{\mu}\right\|_\infty \leq \xi \implies \forall \boldsymbol{\omega}' \in \boldsymbol{\omega}^*(\boldsymbol{\mu}', \boldsymbol{\pi})\ \exists \boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi}), \left\|\boldsymbol{\omega}' - \boldsymbol{\omega}\right\|_\infty \leq \epsilon.$$

Lemma 14 *There exists two constants B and C such that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8}).$$

Proof Let T be such that $h(T) \geq E^2$. Let $p_{\min}(a) = \min_e \pi_e(a)$, and define the the event $\mathcal{J}_T$ as

$$\mathcal{J}_T = \bigcap_{t=h(T)}^{T} \bigcap_{a=1}^{K} \left\{ N_a(t) \geq \frac{1}{4} p_{\min}(a) \min_e N_e(t) \right\}.$$

At this point, from the tower rule, we have that:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T^c) &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}\right] \\
&= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}|\mathcal{J}_T\right] \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T) + \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}|\mathcal{J}_T^c\right] \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c) \\
&\leq \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}|\mathcal{J}_T\right] + \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c).
\end{aligned}$$

At this point, we first focus on $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}|\mathcal{J}_T\right]$. Due to the forced tracking (Lemma 10), we know that, under $\mathcal{J}_T$, the following holds:

$$N_a(t) \geq \frac{1}{4} p_{\min}(a) \min_e N_e(t) \geq \frac{1}{4} p_{\min}(a) \left(\sqrt{t + E^2} - 2E \right).$$

Therefore, from Lemma 19 of Garivier and Kaufmann (2016), we have that:

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\mathbf{1}\left\{\mathcal{E}_T^c\right\}|\mathcal{J}_T\right] \leq B_1 T \exp\left(-C_1 T^{1/8}\right).$$

We now continue with bounding $\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c)$. From Boole's inequality we have that:

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c) \leq \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(N_a(t) \leq \frac{1}{4} p_{\min}(a) \min_e N_e(t)) \right).$$

For each time t, let E_t be the mediator selected at time t by the algorithm. Then,

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c) \leq \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(N_a(t) - \frac{1}{2} \sum_{s=0}^{t-1} \pi_{E_s}(a) \leq \frac{1}{4} p_{\min}(a) \min_e N_e(t) - \frac{1}{2} \sum_{s=0}^{t-1} \pi_{E_s}(a)) \right).$$

Notice that, by definition we have that:

$$\frac{1}{2} \sum_{s=0}^{t-1} \pi_{E_s}(a) = \frac{1}{2} \sum_{s=0}^{t-1} \sum_{e=1}^{E} \mathbf{1} \left\{ E_s = e \right\} \pi_e(a) = \frac{1}{2} \sum_{e=1}^{E} N_e(t) \pi_e(a) \geq \frac{1}{2} p_{\min}(a) \min_e N_e(t).$$

Therefore, we have that:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c) \leq & \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(N_a(t) - \frac{1}{2} \sum_{s=0}^{t-1} \pi_{E_s}(a) \leq -\frac{1}{4} p_{\min}(a) \min_e N_e(t) \right) \\ & \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(N_a(t) - \frac{1}{2} \sum_{s=0}^{t-1} \pi_{E_s}(a) \leq -\frac{1}{4} p_{\min}(a) \left(\sqrt{t} - E \right) \right), \end{aligned}$$

where in the last inequality we have used Lemma 10 together with $h(T) \geq E^2$. At this point, applying Lemma F.4 of Dann et al. (2017), we obtain:

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{J}_T^c) &\leq \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \exp \left(-\frac{1}{2} p_{\min}(a) \left(\sqrt{t} - E \right) \right) \\ &\leq \sum_{t=h(T)}^{T} B_2 \exp \left(-C_2 \sqrt{t} \right) \\ &\leq B_2 T \exp \left(-C_2 T^{1/8} \right), \end{aligned}$$

which concludes the proof. ■

We now continue by defining another event that is crucial to our analysis. For all $t \geq 0$, we denote with E_t the mediator that is played at time t by the algorithm. Then, for some $\gamma \in (\frac{1}{2}, 1)$ and some constant c, we define:

$$\mathcal{E}'_T(\gamma) = \bigcap_{t=h(T)}^{T} \bigcap_{a=1}^{K} \left\{ |N_a(t) - \sum_{s=0}^{t-1} \pi_{E_s}(a)| \leq c t^{\gamma} \right\}.$$

Lemma 15 *Let $\gamma = \frac{3}{4}$. There exists two constants B and C such that:*

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}'^c_T(\gamma)) \leq B T \exp(-C T^{1/8}).$$

Proof We have that:

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T'^c) &\leq \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(|N_a(t) - \sum_{s=0}^{t-1} \pi_{E_s}(a)| \leq ct^{\gamma} \right) \\
&\leq \sum_{t=h(T)}^{T} \sum_{a=1}^{K} \exp\left(\frac{-c^2 t^{2\gamma-1}}{2} \right) \\
&= \sum_{t=h(T)}^{T} K \exp\left(\frac{-c^2 \sqrt{t}}{2} \right) \\
&= \sum_{t=h(T)}^{T} B \exp\left(-C\sqrt{t} \right) \\
&\leq BT \exp\left(-CT^{1/8} \right),
\end{aligned}$$

where, in the first step we have used Boole's inequality, in the second one Azuma-Hoeefding, in the third one have used $\gamma = \frac{3}{4}$, and in the fourth one we have redefined the constants. ■

At this point, given our events, we directly inherit Lemma from Degenne and Koolen (2019).

Lemma 16 (Lemma 35 in Degenne and Koolen (2019)) *There exists a constant T_ϵ such that for $T \geq T_\epsilon$ it holds that on $\mathcal{E}_T$, C-tracking verifies:*

$$\forall t \geq \sqrt{T}, \quad \inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty \leq 3\epsilon.$$

We are now ready to state the equivalent of Lemma 13 for the analysis of $\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta]$.

Lemma 17 *There exists a constant T_ϵ such that for $T \geq T_\epsilon$ it holds that on $\mathcal{E}_T \cap \mathcal{E}_T'(3/4)$, C-tracking verifies:*

$$\forall t \geq \sqrt{T}, \quad \inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})} \left\| \frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}} \right\|_\infty \leq 2E\epsilon.$$

Proof Consider T such that the condition that defines Lemma 16 is satisfied. At this point, focus on $t \geq h(T)^2$. For any action $a \in \mathcal{A}$ and $\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})$ we have that:

$$\begin{aligned}
\left| \frac{N_a(t)}{t} - \tilde{\pi}(a) \right| &\leq \left| \frac{N_a(t)}{t} - \frac{1}{t} \sum_{s=0}^{t-1} \pi_{E_s}(a) \right| + \left| \frac{1}{t} \sum_{s=0}^{t-1} \pi_{E_s}(a) - \tilde{\pi}(a) \right| \\
&\leq \frac{ct^{3/4}}{t} + \frac{h(T)}{t} + \left| \frac{1}{t} \sum_{s=h(T)}^{t-1} (\pi_{E_s}(a) - \tilde{\pi}(a)) \right| \\
&\leq \frac{1}{T^{1/4}} + \frac{c}{T^{1/4}} + \left| \frac{1}{t} \sum_{s=h(T)}^{t-1} (\pi_{E_s}(a) - \tilde{\pi}(a)) \right|,
\end{aligned}$$

where the first inequality follows from triangular decomposition, the second one, instead, by definition of $\mathcal{E}'_T$, and the third one by $t \geq h(T)^2$. At this point, however, we notice that:

$$\Big|\frac{1}{t}\sum_{s=0}^{t-1}(\pi_{E_s}(a) - \tilde{\pi}(a))\Big| = \Big|\sum_{e=1}^{E}\pi_e(a)\frac{N_e(t)}{t} - \tilde{\pi}(a)\Big|.$$

Therefore, we have that:

$$\begin{aligned}\left\|\frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}}\right\|_\infty &\leq \frac{c+1}{T^{1/4}} + \max_a \Big|\sum_e \pi_e(a)\frac{N_e(t)}{t} - \sum_e \pi_e(a)\tilde{\omega}_e\Big| \\ &\leq \frac{c+1}{T^{1/4}} + E\left\|\frac{N_t^E}{t} - \tilde{\boldsymbol{\omega}}\right\|_\infty,\end{aligned}$$

where $\tilde{\boldsymbol{\omega}}$ is any weights vector that induces $\tilde{\boldsymbol{\pi}}$. Taking infimums, we obtain:

$$\inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})}\left\|\frac{N_t^A}{t} - \tilde{\boldsymbol{\pi}}\right\|_\infty \leq \frac{c+1}{T^{1/4}} + E \inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu},\boldsymbol{\pi})}\left\|\frac{N_t^E}{t} - \boldsymbol{\omega}\right\|_\infty.$$

Applying Lemma 16 concludes the proof. ■

At this point, we are ready to analyze the sample complexity of Track and Stop within our peculiar setting. First of all, we begin a known result widely adopted within the TaS literature (e.g., Lemma 13 of Degenne and Koolen (2019)).

Lemma 18 *Suppose there exists $T_0 \in \mathbb{N}$ such that, for all $T \geq T_0$, $\mathcal{E}_T \cap \mathcal{E}'_T \subset \tau_\delta \leq T$. Then,*

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}\left[\tau_\delta\right] \leq T_0 + \sum_{t=T_0}^{+\infty}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T^c) + \sum_{t=T_0}^{+\infty}\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}'^c_T).$$

Before proceeding within the analysis, we notice that the sums that depends on the events $\mathcal{E}_T^c$ and $\mathcal{E}'^c_T$ are finite. This is a direct consequence of Lemmas 14 and 15.

We proceed by providing a suitable T_0 that can be used within Lemma 18. We begin with some definition. For any $\boldsymbol{\mu}$ and $\tilde{\boldsymbol{\pi}}$, we define:

$$g(\boldsymbol{\mu}, \tilde{\boldsymbol{\pi}}) = \min_{a \neq 1}(\tilde{\pi}_1 + \tilde{\pi}_a) I_{\frac{\tilde{\pi}_1}{\tilde{\pi}_1 + \tilde{\pi}_a}}(\mu_1, \mu_a).$$

Then, for any $\epsilon, \xi > 0$, we introduce the quantity:

$$C^*_{\epsilon,\xi}(\boldsymbol{\mu}) = \inf_{\substack{\boldsymbol{\mu}' : \|\boldsymbol{\mu} - \boldsymbol{\mu}'\|_\infty \leq \xi \\ \tilde{\boldsymbol{\pi}}' : \inf_{\tilde{\boldsymbol{\pi}} \in \tilde{\boldsymbol{\pi}}^*(\boldsymbol{\mu},\boldsymbol{\pi})}\|\tilde{\boldsymbol{\pi}}' - \tilde{\boldsymbol{\pi}}\|_\infty \leq 3E\epsilon}} g(\boldsymbol{\mu}', \tilde{\boldsymbol{\pi}}').$$

Lemma 19 *Suppose there exists $T_1 \in \mathbb{N}$ such that, for all $T \geq T_1$, if $\mathcal{E}_T \cap \mathcal{E}'_T$ holds, then $g(\hat{\boldsymbol{\mu}}(t), \frac{N_t^A}{t}) \geq tC^*_{\epsilon,\xi}(\boldsymbol{\mu})$ holds as well for $t \geq \sqrt{T}$. Then, under that event:*

$$\tau_\delta \leq T_0 = \max\left\{T_1, \inf\left\{T \in \mathbb{N} : \sqrt{T} + \frac{\beta(T,\delta)}{C^*_{\epsilon,\xi}(\boldsymbol{\mu})} \leq T\right\}\right\}.$$

Proof Set $T \geq T_1$ and suppose that $\mathcal{E}_T \cap \mathcal{E}'_T$ holds.

$$\begin{aligned}
\min\{\tau_\delta, T\} &\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbf{1}\{\tau_\delta > t\} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbf{1}\{Z(t) \leq \beta(t,\delta)\} \\
&\leq \sqrt{T} + \sum_{t=\sqrt{T}}^{T} \mathbf{1}\left\{t \leq \frac{\beta(t,\delta)}{C^*_{\epsilon,\xi}(\boldsymbol{\mu})}\right\} \\
&\leq \sqrt{T} + \frac{\beta(T,\delta)}{C^*_{\epsilon,\xi}(\boldsymbol{\mu})},
\end{aligned}$$

where in the third inequality we have used the fact that, under $\mathcal{E}_T \cap \mathcal{E}'_T$, $g(\hat{\boldsymbol{\mu}}(t), \frac{N_t^A}{t}) \geq tC^*_{\epsilon,\xi}(\boldsymbol{\mu})$ holds for $t \geq \sqrt{T}$. The other steps follows from simple algebraic manipulations.

The statement follows directly from the definition of T_0. ■

At this point, we can prove Theorem 4.

Theorem 4 *Consider any* $\boldsymbol{\mu} \in \mathcal{M}$ *and any* $\boldsymbol{\pi}$ *such that Assumption 1 is satisfied. Let* $\alpha \in (1, e/2]$*. It holds that:*

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta]}{\log(1/\delta)} \leq \alpha T^*(\boldsymbol{\mu}, \boldsymbol{\pi}). \tag{6}$$

Proof Using Lemma 17, for $T \geq T_\epsilon$, on the event $\mathcal{E}_T \cap \mathcal{E}'_T$ it holds that for every $t \geq \sqrt{T}$:

$$Z(t) \geq tC^*_{\epsilon,\xi}(\boldsymbol{\mu}).$$

Therefore, the condition of Lemma 19 are satisfied, and we can use T_0 defined as in Lemma 19 to apply Lemma 18. Therefore, we obtain:

$$\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\pi}}[\tau_\delta] \leq T_0 + \sum_{T=1}^{+\infty} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T^c) + \sum_{T=1}^{+\infty} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}'^c_T).$$

At this point, the rest of the proof follows as the original proof of TaS (Theorem 14 in Garivier and Kaufmann (2016)). ■

Algorithm analysis (Unknown Mediators' Policies)

For the case in which the mediators' policies are unknown to the learner, few mofidications are needed for the proof.

Concerning the analysis of Theorem 6, it is sufficient to notice that Lemma 11 holds unchanged for sequences of arbitrary oracle weights $\boldsymbol{\omega}_t$ that belongs to $\boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{\pi}}(t)$. Indeed, consider a sequence $(\hat{\boldsymbol{\pi}}(t))_{t \in \mathbb{N}}$ that converges to $\boldsymbol{\pi}$ almost surely (which is guaranteed by C-tracking and the law of large numbers). Then, it is sufficient to notice that the mapping $(\boldsymbol{\omega}, \boldsymbol{\pi}) \to \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\omega})$ is upper hemicontinuous in both arguments. The rest of the proofs follow directly.

The problem is slightly more subtle for the analysis of Theorem 7. More specifically, $\mathcal{E}_T$ needs to be redefined in the following way.

$$\mathcal{E}_T = \bigcap_{t=h(T)}^{T} \left(\{ \| \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu} \|_\infty \leq \xi \} \cap \{ \| \hat{\boldsymbol{\pi}}(t) - \boldsymbol{\pi} \|_\infty \leq \xi \} \right),$$

where ξ is such that, $\| \boldsymbol{\mu}' - \boldsymbol{\mu} \|_\infty \leq \xi$ and $\| \boldsymbol{\pi}' - \boldsymbol{\pi} \|_\infty \leq \xi$ implies that

$$\forall \boldsymbol{\omega}' \in \boldsymbol{\omega}^*(\boldsymbol{\mu}', \boldsymbol{\pi}') \ \exists \boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi}), \left\| \boldsymbol{\omega}' - \boldsymbol{\omega} \right\|_\infty \leq \epsilon.$$

At this point, we need to control the probability of $\mathcal{E}_T$. The following Lemma, combined with the proof of Lemma 14, directly implies that $\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{E}_T^c) \leq BT \exp(-CT^{1/8})$ holds.

Lemma 20 *Consider the event:*

$$\mathcal{K}_T = \bigcap_{t=h(T)}^{T} \{ \| \hat{\boldsymbol{\pi}}(t) - \boldsymbol{\pi} \|_\infty \leq \xi \},$$

then, $\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{K}_T^c) \leq BT \exp(-CT^{1/8})$ *holds for some constants* B *and* C.

Proof From Boole's inequality we obtain:

$$\mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}}(\mathcal{K}_T^c) \leq \sum_{t=h(T)}^{T} \sum_{e=1}^{E} \sum_{a=1}^{K} \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\frac{N_{e,a}(t)}{N_e(t)} \leq \pi_e(a) - \xi \right) + \mathbb{P}_{\boldsymbol{\mu},\boldsymbol{\pi}} \left(\frac{N_{e,a}(t)}{N_e(t)} \geq \pi_e(a) + \xi \right),$$

where $N_{e,a}(t)$ denotes the number of pulls to action a when querying mediator e at time t. At this point, let T be such that $h(T) \geq E^2$, then due to Lemma 10, $N_e(t) \geq \sqrt{t} - E$ for every mediators' e. The result than follows from Lemma 19 in Garivier and Kaufmann (2016). ■

The new definition of $\mathcal{E}_T$ plays a role in the equivalent of Lemma 16 that needs to be derived for the unknown policy setting (for which we report the proof below for completeness). The rest of the proof of Theorem 7 follows directly from the ones of Theorem 4.

Lemma 21 *There exists a constant* T_ϵ *such that for* $T \geq T_\epsilon$ *it holds that on* $\mathcal{E}_T$*, C-tracking with unknown mediators' policies verifies:*

$$\forall t \geq \sqrt{T}, \quad \inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu})} \left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty \leq 3\epsilon.$$

Proof Consider any $\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\pi}, \boldsymbol{\omega})$, and consider a sequence $\boldsymbol{\omega}_s \in \boldsymbol{\omega}^*(\hat{\boldsymbol{\mu}}(s), \hat{\boldsymbol{\pi}}(s))$. At this point, we can write:

$$\begin{aligned}
\left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty &\leq \left\| \frac{N_t^E}{t} - \frac{1}{t} \sum_{s=0}^{t-1} \boldsymbol{\omega}_s \right\|_\infty + \left\| \frac{1}{t} \sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty \\
&\leq \frac{E(1+\sqrt{t})}{t} + \left\| \frac{1}{t} \sum_{s=0}^{t-1} \boldsymbol{\omega}_s - \boldsymbol{\omega} \right\|_\infty \\
&\leq \frac{E(1+\sqrt{t})}{t} + \frac{h(T)}{t} + \left\| \frac{1}{t} \sum_{s=h(T)}^{t-1} (\boldsymbol{\omega}_s - \boldsymbol{\omega}) \right\|_\infty ,
\end{aligned}$$

where we combined simple algebraic manipulations with Lemma 10. However, under event $\mathcal{E}_T$, we have that ξ is such that, $\|\boldsymbol{\mu'} - \boldsymbol{\mu}\|_\infty \leq \xi$ and $\|\boldsymbol{\pi'} - \boldsymbol{\pi}\|_\infty \leq \xi$ implies that:

$$\forall \boldsymbol{\omega'} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}', \boldsymbol{\pi}') \ \exists \boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi}), \left\| \boldsymbol{\omega'} - \boldsymbol{\omega} \right\|_\infty \leq \epsilon.$$

The convexity of $\boldsymbol{\omega}^*(\boldsymbol{\mu}, \boldsymbol{\pi})$ then ensures that $\inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^*(\boldsymbol{\omega}, \boldsymbol{\pi})} \left\| \frac{1}{t} \sum_{s=h(T)}^{t-1} (\boldsymbol{\omega}_s - \boldsymbol{\omega}) \right\|_\infty \leq \epsilon$ holds as well (Lemma 33 in Garivier and Kaufmann (2016)). Therefore, under $\mathcal{E}_T$ we obtain:

$$\inf_{\boldsymbol{\omega} \in \boldsymbol{\omega}^* \boldsymbol{\mu}, \boldsymbol{\pi}} \left\| \frac{N_t^E}{t} - \boldsymbol{\omega} \right\|_\infty \leq \frac{E(1+\sqrt{t})}{t} + \frac{h(T)}{t} + \epsilon,$$

which concludes the proof.

■

	TaS	**TaS-MF-k**	**TaS-MF-u**	**Uniform Sampling**
$\delta = 0.4$	389.12 ± 3.79	863.87 ± 13.14	883.52 ± 18.44	1689.71 ± 25.98
$\delta = 0.1$	450.12 ± 3.45	990.27 ± 14.11	1013.14 ± 15.89	1891.62 ± 29.76
$\delta = 0.01$	553.02 ± 3.45	1179.11 ± 13.59	1205.61 ± 16.95	2245.46 ± 33.03
$\delta = 0.001$	655.03 ± 3.97	1376.72 ± 15.94	1378.14 ± 18.91	2619.26 ± 37.53

Table 1: Experiment results for 100 runs. The table report mean and 95% confidence intervals of the empirical stopping time.

Appendix E: Experiments

To conclude, we report simple numerical experiments.

Experiment 1

First of all, we analyze empirically the effect of the partial controllability that arises from the mediators' feedback. More specifically, we consider a Gaussian bandit model with $K = 4$. The different arms have mean $\boldsymbol{\mu} = (1.5, 1.0, 0.7, 0.5)$. We then compare the following baseliens:

- Track and Stop (TaS)
- Track and Stop with mediators' feedback and known policies (TaS-MF-k)
- Track and Stop with mediators' feedback and unknown policies (TaS-MF-u)
- Uniform Sampling (US) over the set of mediator

Both for TaS-MF-k, TaS-MF-u and Uni we assume access to the a set of $E = 4$ mediators, whose policies are given by $\boldsymbol{\pi}_{e_1} = [0.1, 0.8, 0.1, 0]$, $\boldsymbol{\pi}_{e_2} = [0, 0.1, 0.8, 0.1]$, $\boldsymbol{\pi}_{e_3} = [0, 0.1, 0.1, 0.8]$, and $\boldsymbol{\pi}_{e_4} = [0.2, 0.0, 0.4, 0.4]$. We have tested our algorithm on 4 different values of δ, namely $\delta = [0.4, 0.1, 0.01, 0.001]$.

We have run 100 for each algorithm using 100 Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz cpus and 256GB of RAM. The rime required for obtaining all results is moderate (less than a day).

Table 1 shows the result. As we can see, TaS outperforms TaS-MF-k and TaS-MF-u due to the presence of the mediators that limit the controllability over the arm space. Nevertheless, TaS-MF-k and TaS-MF-u are still able to outperform the Uniform Sampling baseline, thus showing the fact that our sequential-decision making strategy is effective at exploiting the set of mediators. Finally, TaS-MF-k and TaS-MF-u performs similar but having access to the knowledge of $\boldsymbol{\pi}$, in practice, leads to some advantages (especially for moderate regimes of δ).

Experiment 2

We now propose a second experiment that analyzes more deeply the difference in performance between TaS-MF-k and TaS-MF-u. Indeed, it has to be noticed that, whenever the

	TaS-MF-k	**TaS-MF-u**
$\delta = 0.4$	254.75 ± 15.01	1120.97 ± 63.77
$\delta = 0.1$	277.83 ± 14.63	1105.42 ± 60.53
$\delta = 0.01$	307.87 ± 14.01	1154.54 ± 65.54
$\delta = 0.001$	343.43 ± 16.66	1121.26 ± 56.74
$\delta = 0.0001$	395.66 ± 16.21	1219.50 ± 60.55
$\delta = 0.00001$	405.05 ± 14.89	1228.79 ± 63.32
$\delta = 0.000001$	408.33 ± 15.33	1264.96 ± 60.18
$\delta = 0.0000001$	484.49 ± 17.70	1330.48 ± 61.40
$\delta = 0.00000001$	493.63 ± 16.10	1366.83 ± 61.08
$\delta = 0.000000001$	506.81 ± 16.94	1347.70 ± 62.64
$\delta = 0.0000000001$	569.99 ± 18.65	1514.64 ± 67.12
$\delta = 0.00000000001$	609.26 ± 17.10	1434.92 ± 59.57

Table 2: Experiment results for 1000 runs. The table report mean and 95% confidence intervals of the empirical stopping time.

mediators' policies are known to the agent, if identical policies are present within the set, the algorithm can actually avoid querying identical copies of the mediators. [11] In other words, TaS-MF-k can be easily modified (without affecting its theoretical guarantees) to remove all copies of identical policies $\boldsymbol{\pi_e} \in \boldsymbol{\pi}$. In pratice, in scenarios such as the one that we will describe in a moment, this turns out to significantly affect the sample complexity, especially for moderate regime of δ. More precisely, we consider the case where $\boldsymbol{\mu} = [5.0, 1.0]$ and $\boldsymbol{\pi}$ contains $E = 10$ mediators. We consider $\boldsymbol{\pi}_{e_1} = [0.01, 0.99]$ and $\boldsymbol{\pi}_{e_i} = [0.0, 1.0]$ for all $i \neq 1$. We test both algorithms in the regimes of δ: [0.4, 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001, 0.0000001, 0.00000001, 0.000000001, 0.0000000001, 0.00000000001, 0.000000000001].

We run 1000 runs for both algorithms using 100 Intel(R) Xeon(R) Gold 6238R CPU @ 2.20GHz cpus and 256GB of RAM. The rime required for obtaining all results is moderate (less than a day).

Table 2 reports the result. As we can see, especially for moderate regime of δ there is a significant difference between the two algorithms. This is due to the fact TaS-MF-u pays a significant amount of samples for the forced exploration to various mediators that are equivalent, namely $\boldsymbol{\pi}_{e_i}$ for all $i \neq 1$. The difference between TaS-MF-k and TaS-MF-u, however, tends to shrink as soon as we decrease the values of δ (as predicted by our theory).

[11] More generally, the algorithm can remove some $\boldsymbol{\pi_e} \in \boldsymbol{\pi}$ whenever the set $\widetilde{\Sigma}_k$ does not change.