Using causal abstractions to accelerate decision-making in complex bandit problems

Joel Dyer[*]
University of Oxford

Nicholas Bishop[*]
University of Oxford

Anisoara Calinescu
University of Oxford

Michael Wooldridge
University of Oxford

Fabio Massimo Zennaro
University of Bergen

Abstract

Although real-world decision-making problems can often be encoded as causal multi-armed bandits (CMABs) at different levels of abstraction, a general methodology exploiting the information and computational advantages of each abstraction level is missing. In this paper, we propose AT-UCB, an algorithm which efficiently exploits shared information between CMAB problem instances defined at different levels of abstraction. More specifically, AT-UCB leverages causal abstraction (CA) theory to explore within a cheap-to-simulate and coarse-grained CMAB instance, before employing the traditional upper confidence bound (UCB) algorithm on a restricted set of potentially optimal actions in the CMAB of interest, leading to significant reductions in cumulative regret when compared to the classical UCB algorithm. We illustrate the advantages of AT-UCB theoretically, through a novel upper bound on the cumulative regret, and empirically, by applying AT-UCB to epidemiological simulators with varying resolution and computational cost.

1 Introduction

Complex decision-making problems are naturally represented and studied as causal systems at multiple levels of abstraction. Consider the problem of controlling the spread of a disease through a population. One may develop a fine-grained, computationally-expensive agent-based model (ABM) as well as a low-fidelity differential equation model (DEM) that is cheap to simulate. By experimenting with the DEM, we can quickly identify policy interventions that might be worth testing in the higher resolution ABM. But how should we divide our computational budget between the two models, and how should we use information learned from the DEM to inform our search over interventions in the ABM? A similar challenge appears in fields such as economics and ecology, that often require intervening in complex environments. More generally, integrating knowledge across different models is challenging, and practitioners often rely on ad-hoc solutions.

In this paper, we study how multiple causal models at different levels of abstraction can be formally related and jointly exploited to improve decision-making. We adopt the framework of causally abstracted multi-armed bandits (CAMABs) [Zennaro et al., 2024], where causal systems are encoded as structural causal models (SCMs), relations across SCMs are formalized via causal abstraction (CA), and decision problems are expressed as multi-armed bandits (MABs). We focus on the scenario where a decision-maker is given two models, a realistic but expensive base model and a simplified but cheap abstracted model; their aim is to take advantage of running the cheap model in order to

[*]Equal contribution. Correspondence to: `joel.dyer@cs.ox.ac.uk`; `nicholas.bishop@cs.ox.ac.uk`; `fabio.zennaro@uib.no`.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science.

gain information for discovering optimal policies in the realistic model more efficiently. We propose abstract thresholding UCB (AT-UCB), an algorithm that first performs exploration in the abstracted model in order to filter out sub-optimal interventions before running the UCB algorithm on the base model over a restricted set of actions. We provide: (i) a theoretical analysis of the regret of AT-UCB, and (ii) simulations confirming our theoretical results.

Related Work Our work adopts the CAMAB framework introduced by Zennaro et al. [2024], which also introduces methods for transferring information across causal multi-armed bandit problems (CMABs). However, whilst their work focuses on transferring information from the base model to the abstracted model, we consider the more realistic inverse setting and provide a more rigorous theoretical analysis. More generally, CAMABs are subtly related to MAB problems which assume a form of clustering between arms such as regional bandits [Wang et al., 2018], MABs with dependent arms [Singh et al., 2024], and bandits with similar arms [Pesquerel et al., 2021]. Our AT-UCB algorithm is reminiscent of approaches that warm-start bandit algorithms with offline or confounded data [Cheung and Lyu, 2024, Yang et al., 2025, Sharma et al., 2020] as well as thresholding bandits [Mason et al., 2020, Locatelli et al., 2016]. For more discussion of related MAB problems, we refer the reader to Appendix A.

2 Background

Before describing AT-UCB in detail, we provide some background on CAMABs. We first define structural causal models (SCMs) and interventions [Pearl, 2009, Peters et al., 2017].

Definition 1 (SCM)**.** *A structural causal model (SCM) $\mathcal{M}$ is a tuple $\langle \mathcal{X}, \mathcal{U}, \mathcal{F}, P \rangle$ where $\mathcal{X} = \{X_i\}_{i=1}^d$ is a collection of d endogenous variables, $\mathcal{U} = \{U_i\}_{i=1}^d$ is a collection of exogenous variables, $\mathcal{F} = \{f_{X_i}\}_{i=1}^d$ is a collection of structural functions $f_{X_i} : \mathbb{D}[\mathrm{PA}(X_i)] \times \mathbb{U}_i \to \mathbb{X}_i$, where $\mathbb{D}$ denotes the domain and with $\mathrm{PA}(X_i) \subseteq \mathcal{X} \setminus X_i$, and P is a probability distribution over $\mathcal{U}$.*

Definition 2 (Intervention)**.** *Given an SCM $\mathcal{M}$, an intervention $a = \mathrm{do}(\mathbf{X} = \mathbf{x})$ on a set of endogenous variables $\mathbf{X}$ associated with values $\mathbf{x}$, defines a new SCM $\mathcal{M}^a$ by replacing f_i with the constant x_i for each variable $X_i \in \mathbf{X}$.*

We assume that our SCMs induce directed acyclic graphs (DAGs) over the endogenous variables $\mathcal{X}$, such that PA is interpreted as graph-theoretical parents. For a given intervention a, each variable X has an interventional distribution defined by pushing P through the modified structural equations of the model $\mathcal{M}^a$. For ease of presentation, we let $X_a \sim P(X \mid a)$ denote a random variable distributed according to X under the intervened model $\mathcal{M}^a$. An SCM naturally induces a multi-armed bandit problem known as a CMAB [Bareinboim et al., 2015, Lattimore et al., 2016].

Definition 3 (CMAB)**.** *A causal multi-armed bandit (CMAB) $\mathcal{B}$ is a tuple $\langle \mathcal{M}, Y, \mathcal{A} \rangle$ where $\mathcal{M}$ is an SCM, $Y \subseteq \mathbb{R}$ is an endogenous outcome/reward variable in $\mathcal{X}$, and $\mathcal{A} = \{a_i\}_{i=1}^k$ is a set of k interventions/arms/actions on $\mathcal{M}$.*

A CMAB describes a basic sequential decision problem over a time horizon of length $n \in \mathbb{N}_+$. On each time step $t \in [n]$ the learner (who has no knowledge of P or $\mathcal{F}$) is tasked with selecting an intervention $a \in \mathcal{A}$. Upon selecting intervention $a^{(t)}$ on time step t, the learner receives a reward $y^{(t)} \sim Y_{a^{(t)}}$. The goal of the learner is to maximize their cumulative reward over the time horizon. We let $\mu_{a^{(t)}}$ describe the mean reward associated with action a and (for the sake of analysis) label actions so that $\mu_{a_1} \geq \mu_{a_2} \geq \cdots \geq \mu_{a_k}$.

Lastly, we define (τ, ω)-abstraction, which formally relates two SCMs through a state map τ and an intervention map ω. We also define two measures that capture the faithfulness of a (τ, ω)-abstraction in the bandit context; the interventional consistency (IC) error [Rubenstein et al., 2017, Beckers and Halpern, 2019, Zennaro et al., 2023] and the reward discrepancy (RD) [Zennaro et al., 2024].

Definition 4 (Abstraction)**.** *Given a base SCM $\mathcal{M}$ with a set of interventions $\mathcal{A}$ and an abstracted SCM $\mathcal{M}'$ with a set of interventions $\mathcal{A}'$, an abstraction $\boldsymbol{\alpha}$ is a tuple (τ, ω) where $\tau : \mathbb{D}[\mathcal{X}] \to \mathbb{D}[\mathcal{X}']$ maps outcomes of the base model to outcomes of the abstracted model, and $\omega : \mathcal{A} \to \mathcal{A}'$ maps base interventions to abstracted interventions.*

Definition 5 (Abstraction measures (IC,RD))**.** *Given an abstraction $\boldsymbol{\alpha}$, the interventional consistency (IC) error is:*

$$e(\boldsymbol{\alpha}) := \max_{a \in \mathcal{A}} D_{W_2}\left(\tau(P(Y|a)), P(Y'|\omega(a))\right) \tag{1}$$

and the reward discrepancy (RD) error is:

$$s(\boldsymbol{\alpha}) := \max_{a \in \mathcal{A}} D_{W_2}\left(P(Y|a), \tau(P(Y|a))\right), \tag{2}$$

where D_{W_2} is the 2-Wasserstein distance, and $\tau(P(Y|a))$ is the distribution of $\tau(Y)$ when $Y \sim P(\cdot \mid a)$.

In short, the IC error measures the difference between (i) intervening in the base model and abstracting the result, and (ii) abstracting the intervention and intervening in the abstracted model. Meanwhile, RD measures the degree to which rewards are distorted by the state map τ. A CAMAB [Zennaro et al., 2024] consists of two CMABs and an abstraction formally relating them.

Definition 6 (CAMAB)**.** *A causally abstracted MAB (CAMAB) $\mathcal{C}$ is a tuple $\langle \mathcal{B}, \mathcal{B}', \boldsymbol{\alpha} \rangle$ made up by a base CMAB $\mathcal{B}$ and an abstracted CMAB $\mathcal{B}'$ related by an abstraction $\boldsymbol{\alpha}$.*

To distinguish mathematical quantities relating to $\mathcal{B}'$ rather than $\mathcal{B}$ we append an apostrophe. For example, the total number of arms in $\mathcal{B}$ is denoted by k, whilst the total number of arms in $\mathcal{B}'$ is denoted by k'. Given a CAMAB $\mathcal{C} = \langle \mathcal{B}, \mathcal{B}', \boldsymbol{\alpha} \rangle$, we consider the following scenario. Assume that the learner will interact with the base CMAB $\mathcal{B}$ over a time horizon of length n. Prior to the start of the time horizon the learner is permitted to interact with the abstracted CMAB $\mathcal{B}'$ for as long as they wish. However, each time the learner intervenes in $\mathcal{B}'$ they pay a fixed cost $C > 0$. Note that this setting resembles the epidemiological example given in Section 1. A natural goal for the learner is to minimize their regret with respect to an oracle who does not spend any time interacting with $\mathcal{B}'$ and takes the optimal intervention a_1 on every time step when interacting with $\mathcal{B}$.

Definition 7 (Cumulative Regret)**.** *Given a CAMAB $\mathcal{C}$, the expected cumulative regret of the learner is $\mathcal{R}_n(\mathcal{C}) := n' \cdot C + \sum_{t=1}^{n} \Delta_{a^{(t)}}$, where n' denotes the number of interventions performed in $\mathcal{B}'$, and $\Delta_a = \mu_{a_1} - \mu_a$ denotes the optimality gap between action a and action a_1.*

3 Methods and theory

To propose an algorithm that can exploit the structure of a CAMAB, we first characterize the relation between: (i) the mean $\mu_{a'_1}$ of the true optimal arm in the abstracted CMAB $\mathcal{B}'$; and (ii) the mean $\mu_{\omega(a_1)}$ of the image through ω of the true optimal arm in the base CMAB $\mathcal{B}$.

Proposition 1. *Given a CAMAB $\mathcal{C}$, let $\boldsymbol{\alpha} = \langle \tau, \omega \rangle$ be an approximate abstraction with IC error $e(\boldsymbol{\alpha})$ and RD error $s(\boldsymbol{\alpha})$, and define $\epsilon(\boldsymbol{\alpha}) = 2\left(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})\right)$. Then it holds that:*

$$\mu_{a'_1} - \epsilon(\boldsymbol{\alpha}) \leq \mu_{\omega(a_1)} \leq \mu_{a'_1}. \tag{3}$$

Informally, Proposition 1 bounds the distance between the abstracted optimal arm and the image of the base optimal arm. It thus suggests that the search for the optimal base arm a_1 can be accelerated in the presence of an approximate causal abstraction $\boldsymbol{\alpha}$ by restricting attention to base interventions in an interval defined by Proposition 1 as $a \in \omega^{-1}\left(\mathcal{D}^c\right)$, where $\mathcal{D}^c$ denotes the complement of the set:

$$\mathcal{D} := \left\{ a' \in \mathcal{A}' : \mu_{a'} < \mu_{a'_1} - \epsilon(\boldsymbol{\alpha}) \right\}. \tag{4}$$

This observation informs our contribution, which is outlined in Algorithm 1 and combines the initial use of a thresholding bandit algorithm in the abstract CMAB – performed through uniform exploration of the k' abstract arms, for the sake of simplicity – with a standard upper confidence bound (UCB) [Lattimore and Szepesvári, 2020] algorithm in the base CMAB.

3.1 Regret analysis

Next, we analyse the expected cumulative regret incurred by Algorithm 1. We work under the standard assumption of 1-subgaussianity on the reward distributions and assume the abstract horizon n' is sufficiently large to bound exponentials appearing as a result of the application of Chernoff bounds when bounding the regret incurred by experimenting in $\mathcal{M}$:

Assumption 1. *With $\Delta'_{a'} := \mu_{a'_1} - \mu_{a'}$, the abstract horizon length n' satisfies,*

$$n' \geq 4k' \log(nk') \max\left\{ \left(\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right)^{-2}, \left(\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right)^{-2} \right\}. \tag{6}$$

Algorithm 1 AT-UCB

1: **Input:** Base and abstract CMABs $\mathcal{M}$, $\mathcal{M}'$; base and abstract time horizons n, n'; UCB parameter $\delta > 0$; approximate abstraction $\boldsymbol{\alpha}$; threshold $\varepsilon > 0$
2: **Output:** estimated optimal arm $\hat{a}_1$
3: Pull each abstract arm $a' \in \mathcal{A}'$ a total of n'/k' times each to construct estimates $\hat{\mu}_{a'}$.
4: Construct

$$\hat{\mathcal{D}} = \{a' \in \mathcal{A}' \mid \hat{\mu}_{a'} < \max_{\tilde{a}' \in \mathcal{A}'} \hat{\mu}_{\tilde{a}'} - \varepsilon\}. \tag{5}$$

5: Run UCB on $\hat{\mathcal{A}} := \omega^{-1}(\hat{\mathcal{D}}^c) = \cup_{a' \in \hat{\mathcal{D}}^c} \omega^{-1}(a')$ for n time steps and with parameter δ.
6: **Return:** $\hat{a}_1 = \arg\max_{a \in \hat{\mathcal{A}}} \mathrm{UCB}_a(n, \delta)$

Under these assumptions, the following upper bound on the expected cumulative regret for Algorithm 1 holds, with the proof deferred to Appendix B:

Proposition 2. *The cumulative regret $\mathcal{R}_n(\mathcal{C})$ under Algorithm 1 with $\varepsilon = \epsilon(\boldsymbol{\alpha})$ satisfies*

$$\mathcal{R}_n(\mathcal{C}) \leq n' \cdot C + \left(3 \sum_{a \in \omega^{-1}(\mathcal{D}^c)} \Delta_a + \sum_{\substack{a \in \omega^{-1}(\mathcal{D}^c), \\ \Delta_a > 0}} \frac{16 \log(n)}{\Delta_a} \right) + 2 \max_{a \in \mathcal{A}} \Delta_a. \tag{7}$$

When C is sufficiently small and $\epsilon(\boldsymbol{\alpha})$ is small enough that $|\omega^{-1}(\mathcal{D}^c)| \ll k$, Proposition 2 gives a lower regret bound than the standard upper bound for UCB.

4 Experiments

In this section, we study Algorithm 1 empirically, using two epidemic simulators. For $\mathcal{M}$, we consider a susceptible-infected-recovered-susceptible model. In this model, k spatial regions are connected in a fully-connected graph, and different regions carry and spread the hypothetical virus more easily. The epidemic is simulated over discrete time steps of size $\Delta t > 0$. Further details are provided in Appendix C.1. The learner is tasked with choosing a community $a \in \{1, \ldots, k\}$, on which a fixed-duration lockdown, starting at a prespecified time, will be imposed. The reward observed is the negation of total infections over the simulated horizon. The learner enjoys access to a high-level model $\mathcal{M}'$, comprised of large abstract communities, related to $\mathcal{M}$ via an approximate causal abstraction. Abstract communities $a' \in \{1, \ldots, k'\}$ with $k' < k$ are related to base communities via ω, and $\mathcal{M}'$ is simulated at a coarser time resolution than $\mathcal{M}$, such that $\Delta t' > \Delta t$. Further details on the abstraction used are provided in Appendix C.2.

In Figure 1, we show the difference in cumulative regret from running UCB in the base bandit and running AT-UCB with $n' = 20$, as a function of ε. Both procedures use $n = 100$, and error bars show the standard error in the mean from 10 repeats. Larger values indicate that AT-UCB achieves comparatively lower regret. For sufficiently small ε, a lower regret is achieved using the approximate causal abstraction and Algorithm 1 instead of the standard UCB algorithm. In this instance, since $\omega(a_1) = a'_1$ and the abstract reward distributions have low variance relative to the distances between arm means, the difference in regret decreases with increasing ε as suboptimal arms are included while $\omega(a_1)$ is rarely excluded even for low ε.

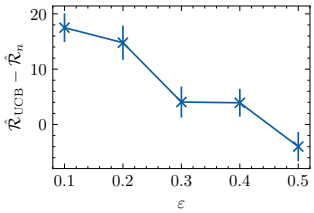

Figure 1: Difference in empirical average cumulative regret as a function of ε for the experiment in Section 4.

5 Conclusion

We have introduced AT-UCB, a CAMAB algorithm that exploits CA theory to explore an abstract CMAB so that a more complex CMAB can be solved efficiently via immediate elimination of sub-optimal arms. Whilst we have shown that AT-UCB provides tighter regret guarantees given a sufficiently accurate abstraction, AT-UCB has several limitations, discussed further in Appendix D, that we aim to address in future work. In particular, AT-UCB does not exploit causal information

provided by the SCMs $\mathcal{M}$ and $\mathcal{M}'$ to expedite exploration. Similarly, assessing the validity of Assumption 1 a priori relies on knowledge of abstract optimality gap sizes, which may be unrealistic. Moreover, the regret guarantee provided by Proposition 2 relies on setting $\varepsilon = \epsilon(\boldsymbol{\alpha})$, which may be unknown.

Acknowledgments and Disclosure of Funding

The authors are extremely grateful to Theodoros Damoulas for his feedback, helpful discussions, and his support. JD, NB, AC, and MW acknowledge funding from a UKRI AI World Leading Researcher Fellowship awarded to Wooldridge (grant EP/W002949/1). MW and AC also acknowledge funding from Trustworthy AI - Integrating Learning, Optimisation and Reasoning (TAILOR), a project funded by the European Union Horizon2020 research and innovation program under Grant Agreement 952215.

References

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28:1342–1350, 2015.

Sander Beckers and Joseph Y Halpern. Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685, 2019.

Blair Bilodeau, Linbo Wang, and Daniel M Roy. Adaptively exploiting d-separators with causal bandits. *arXiv preprint arXiv:2202.05100*, 2022.

Stéphane Caron, Branislav Kveton, Marc Lelarge, and Smriti Bhagat. Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*, 2012.

Wang Chi Cheung and Lixing Lyu. Leveraging (biased) information: Multi-armed bandits with offline data. *arXiv preprint arXiv:2405.02594*, 2024.

Joel Dyer, Nicholas Bishop, Yorgos Felekis, Fabio Massimo Zennaro, Anisoara Calinescu, Theodoros Damoulas, and Michael Wooldridge. Interventionally consistent surrogates for complex simulation models. *Advances in Neural Information Processing Systems*, 37:21814–21841, 2024.

Finnian Lattimore, Tor Lattimore, and Mark D Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems*, pages 1181–1189, 2016.

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

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4164–4172, 2019.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence*, pages 141–150. PMLR, 2020.

Shie Mannor and Ohad Shamir. From bandits to experts: On the value of side-observations. *Advances in neural information processing systems*, 24, 2011.

Blake Mason, Lalit Jain, Ardhendu Tripathy, and Robert Nowak. Finding all \epsilon-good arms in stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 33, pages 20707–20718, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Fabien Pesquerel, Hassan Saber, and Odalric-Ambrym Maillard. Stochastic bandits with groups of similar arms. *Advances in Neural Information Processing Systems*, 34:19461–19472, 2021.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: Foundations and learning algorithms*. MIT Press, 2017.

Han Qi, Guo Fei, and Li Zhu. Graph feedback bandits with similar arms. *arXiv preprint arXiv:2405.11171*, 2024.

Eigil Fjeldgren Rischel. The category theory of causal models. Master's thesis, University of Copenhagen, 2020.

Paul K Rubenstein, Sebastian Weichwald, Stephan Bongers, Joris M Mooij, Dominik Janzing, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Causal consistency of structural equation models. In *33rd Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, pages 808–817. Curran Associates, Inc., 2017.

Willem Schooltink and Fabio Massimo Zennaro. Aligning graphical and functional causal abstractions. *arXiv preprint arXiv:2412.17080*, 2024.

Nihal Sharma, Soumya Basu, Karthikeyan Shanmugam, and Sanjay Shakkottai. Warm starting bandits with side information from confounded data. *arXiv preprint arXiv:2002.08405*, 2020.

Rahul Singh, Fang Liu, Yin Sun, and Ness Shroff. Multi-armed bandits with dependent arms. *Machine Learning*, 113(1):45–71, 2024.

Zhiyang Wang, Ruida Zhou, and Cong Shen. Regional multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 510–518. PMLR, 2018.

Le Yang, Vincent YF Tan, and Wang Chi Cheung. Best arm identification with possibly biased offline data. *arXiv preprint arXiv:2505.23165*, 2025.

Fabio Massimo Zennaro, Paolo Turrini, and Theodoros Damoulas. Quantifying consistency and information loss for causal abstraction learning. In *Proceedings of the Thrity-Second International Conference on International Joint Conferences on Artificial Intelligence*, 2023.

Fabio Massimo Zennaro, Nicholas Bishop, Joel Dyer, Yorgos Felekis, Anisoara Calinescu, Michael Wooldridge, and Theodoros Damoulas. Causally abstracted multi-armed bandits. In *Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence*, pages 4109–4139, 2024.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.

A Further literature review

In this appendix, we discuss further relations between our work and the literature.

CA. Differently from Zennaro et al. [2024] which relies on α-abstractions [Rischel, 2020], we adopt the formalism of τ-ω abstraction [Rubenstein et al., 2017, Beckers and Halpern, 2019]. Although the two frameworks are related [Schooltink and Zennaro, 2024, Beckers and Halpern, 2019], the τ-ω abstraction allows us to deal with the intervention map ω separately from the variable map τ. The τ-ω formalism was also used to motivate a procedure for learning interventionally consistent surrogate models of complex simulators in Dyer et al. [2024]; the current work differs in that we consider how to exploit knowledge of an approximate causal abstraction to accelerate decision-making, rather than the problem of learning approximate causal abstractions in the first place.

MABs. Like regional MABs [Wang et al., 2018] and MABs with dependent arms [Singh et al., 2024], CAMABs implicitly cluster base actions through abstraction; however, in general, CA cannot provide guarantees on the functional form or the parameters of the actions grouped together. As actions are formally related via abstractions in CAMABs, taking a single action (base or abstract) provides information about several other actions in a fashion similar to MABs with side information [Mannor and Shamir, 2011, Qi et al., 2024, Caron et al., 2012]. However, exploiting this information is less trivial in CAMABs, as abstractions may introduce arbitrary biases, as captured through

abstraction measures such as the interventional consistency error and reward discrepancy error. AT-UCB uses abstract data to rapidly eliminate arms before interacting with a base model, in a fashion reminiscent of MABs with offline information [Cheung and Lyu, 2024, Yang et al., 2025, Sharma et al., 2020], which use offline data to warm start MAB algorithms. Employing such warm start algorithms in the CAMAB setting is complicated by the biases introduced by abstractions discussed previously. The arm elimination procedure adopted by AT-UCB is also similar in nature to thresholding bandit algorithms [Locatelli et al., 2016], which attempt to identify a subset of arms with mean reward above given a threshold. Unfortunately, many thresholding bandit algorithms assume a fixed threshold rather than a relative one that depends of the mean reward of the optimal arm, making them incompatible AT-UCB. To the best of our knowledge, the only exception to this is Mason et al. [2020], who aim to identify a set of ϵ-satisficing arms.

CMABs. A wide range of algorithms have been proposed for the CMAB setting which exploit knowledge of the causal graph associated with the underlying SCM to expedite exploration [Lattimore et al., 2016, Lu et al., 2020, Bilodeau et al., 2022]. Similar to [Lee and Bareinboim, 2019], we also perform arm elimination to efficiently solve a CMAB; however, where Lee and Bareinboim [2019] filters out actions through the identification of a possibly optimal minimal intervention set via causal analysis, we rely on CA to estimate a similar set of possibly optimal actions. Finally, CAs allow for information transfer akin to transfer learning in CMABs [Zhang and Bareinboim, 2017]; however, where transfer learning in Zhang and Bareinboim [2017] requires knowing the causal graphs of both SCMs (and their overlap), a τ-ω abstraction requires only knowledge of a map between variables and between interventions.

B Further mathematical results and proofs

In this appendix, we provide further theoretical results to prove the claims in the main body. For completeness, we recall the following result from Zennaro et al. [2024], which we utilize in our proofs.

Lemma 1 (Proposition 4.1 of Zennaro et al. [2024])**.** *Given a CAMAB $\mathcal{C}$, for any $a \in \mathcal{A}$ and approximate abstraction $\boldsymbol{\alpha} = \langle \tau, \omega \rangle$,*

$$|\mu_a - \mu_{\omega(a)}| \leq s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha}). \tag{8}$$

Lemma 1 bounds the absolute difference between the expected reward associated with intervening with an action $a \in \mathcal{A}$ in the base bandit $\mathcal{B}$ and the expected reward associated with intervening with the corresponding abstracted action $\omega(a)$ in $\mathcal{B}'$. That is, Lemma 1 bounds how much the expected reward associated with a base action changes after abstracting. The next lemma bounds the optimality gap of an action in $\mathcal{B}$ in terms of the IC and RD error of $\boldsymbol{\alpha}$ and the optimality gap of the corresponding action in $\mathcal{B}'$.

Lemma 2. *Given a CAMAB $\mathcal{C}$, for any $a_k \in \mathcal{A}$ and approximate abstraction $\boldsymbol{\alpha} = \langle \tau, \omega \rangle$,*

$$\Delta_{a_k} \leq \Delta'_{\omega(a_k)} + 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) - \Delta'_{\omega(a_1)}. \tag{9}$$

Proof of Lemma 2. For any $a_k \in \mathcal{A}$,

$$\begin{aligned}
\Delta_{a_k} &= \mu_{a_1} - \mu_{a_k} \\
&= \mu_{a_1} + \mu_{a'_1} - \mu_{a'_1} - \mu_{\omega(a_1)} + \mu_{\omega(a_1)} - \mu_{a_k} + \mu_{\omega(a_k)} - \mu_{\omega(a_k)} \\
&\leq |\mu_{a_1} - \mu_{\omega(a_1)} - \mu_{a_k} + \mu_{\omega(a_k)}| + \mu_{a'_1} - \mu_{a'_1} + \mu_{\omega(a_1)} - \mu_{\omega(a_k)} \\
&= |\mu_{a_1} - \mu_{\omega(a_1)} - \mu_{a_k} + \mu_{\omega(a_k)}| + \Delta'_{\omega(a_k)} - \Delta'_{\omega(a_1)} \\
&\leq |\mu_{a_1} - \mu_{\omega(a_1)}| + |\mu_{a_k} - \mu_{\omega(a_k)}| + \Delta'_{\omega(a_k)} - \Delta'_{\omega(a_1)} && \text{(triangle inequality)} \\
&\leq 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) + \Delta'_{\omega(a_k)} - \Delta'_{\omega(a_1)}. && \text{(Lemma 1)}
\end{aligned}$$

□

We now prove the bound on the mean of the abstract arm corresponding to the optimal arm in the base bandit instance, provided in Proposition 1:

Proposition 1. *Given a CAMAB $\mathcal{C}$, let $\boldsymbol{\alpha} = \langle \tau, \omega \rangle$ be an approximate abstraction with IC error $s(\boldsymbol{\alpha})$ and RD error $e(\boldsymbol{\alpha})$, and define $\epsilon(\boldsymbol{\alpha}) = 2\left(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})\right)$. Then it holds that:*

$$\mu_{a'_1} - \epsilon(\boldsymbol{\alpha}) \leq \mu_{\omega(a_1)} \leq \mu_{a'_1}. \tag{10}$$

Proof of Proposition 1. To prove the right side of the inequality, $\mu_{\omega(a_1)} \leq \mu_{a'_1}$, it is sufficient to recall the definition of the optimal abstract arm a'_1, according to which $\mu_{a'_1} \geq \mu_{a_k}$ for all $a_k \in \mathcal{A}'$, including $\omega(a_1)$.

To prove the left side of the inequality, $\mu_{a'_1} - \epsilon(\boldsymbol{\alpha}) \leq \mu_{\omega(a_1)}$, we apply Lemma 2 with $a_k \in \omega^{-1}(a'_1)$

$$\begin{aligned}
\Delta_{a_k} &\leq \Delta'_{\omega(a_k)} + 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) - \Delta'_{\omega(a_1)} && \text{(Lemma 2)} \\
&\leq \Delta'_{a'_1} + 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) - \Delta'_{\omega(a_1)} && \text{(mapping } a_k \text{ to the optimal } a'_1\text{)} \\
&\leq 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) - \Delta'_{\omega(a_1)} && \text{(optimality gap for } a'_1 \text{ is zero)} \\
&= 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) - \mu_{a'_1} + \mu_{\omega(a_1)}. && \text{(definition of optimality gap)}
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
\mu_{\omega(a_1)} &\geq \mu_{a'_1} - 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})) + \Delta_{a_k} \\
&\geq \mu_{a'_1} - 2(s(\boldsymbol{\alpha}) + e(\boldsymbol{\alpha})), && (\Delta_{a_k} \geq 0) \\
&\geq \mu_{a'_1} - \epsilon(\boldsymbol{\alpha}), && \text{(definition of } \epsilon(\boldsymbol{\alpha})\text{)}
\end{aligned}$$

which proves the left side of the inequality. □

Next, we state and prove an auxiliary lemma, used in the proof Proposition 2. In short, Lemma 3 shows that Assumption 1 ensures AT-UCB obtains enough abstract samples to correctly filter abstract actions with probability sublinear in n.

Lemma 3. *Under Assumption 1,*

$$\sum_{a' \in \mathcal{A}'} \exp\left(-\frac{n'}{4k'}\left(\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)}\right)^2\right) + \sum_{a' \in \mathcal{D}} \exp\left(-\frac{n'}{4k'}\left(\Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right)^2\right) \leq \frac{2}{n}. \tag{11}$$

Proof. By Assumption 1,

$$\frac{n'}{4k'} \min\left\{\left(\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right)^2, \left(\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right)^2\right\} \geq \log(nk'). \tag{12}$$

Taking the exponent of both sides yields,

$$\exp\left(\min\left\{\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right]^2, \frac{n'}{4k'}\left[\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right\}\right) \geq nk'. \tag{13}$$

Reciprocating both sides and reversing the inequality yields,

$$\frac{1}{n} \geq k' \max\left\{\exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right]^2\right), \exp\left(-\frac{n'}{4k'}\left[\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right)\right\}. \tag{14}$$

Recalling that $k' := |\mathcal{A}'|$ gives,

$$\frac{1}{n} \geq \sum_{a' \in \mathcal{A}'} \max\left\{\exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right]^2\right), \exp\left(-\frac{n'}{4k'}\left[\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right)\right\}. \tag{15}$$

Using Equation (15), we may bound both terms on the LHS of Equation (11). Starting with the first term, since $\Delta'_{a'} \geq 0$ and $\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)} \geq 0$ (by Proposition 1), it follows that,

$$\exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) - \Delta'_{\omega(a_1)}\right]^2\right) \geq \exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)}\right]^2\right). \tag{16}$$

Combining Equation (15) with Equation (16) yields,

$$\frac{1}{n} \geq \sum_{a' \in \mathcal{A}'} \exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)}\right]^2\right). \tag{17}$$

Moving onto the second term, since $\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha}) > 0$ by definition of $\mathcal{D}$, it follows that,

$$\exp\left(-\frac{n'}{4k'}\left[\min_{a' \in \mathcal{D}} \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right) \geq \exp\left(-\frac{n'}{4k'}\left[\Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right) \qquad \forall a' \in \mathcal{D}. \tag{18}$$

Combining Equation (15) with Equation (18) yields,

$$\frac{1}{n} \geq \sum_{a' \in \mathcal{A}'} \exp\left(-\frac{n'}{4k'}\left[\min_{\tilde{a}' \in \mathcal{D}} \Delta'_{\tilde{a}'} - \epsilon(\boldsymbol{\alpha})\right]^2\right) \geq \sum_{a' \in \mathcal{D}} \exp\left(-\frac{n'}{4k'}\left[\Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right). \tag{19}$$

Finally, adding Equation (17) and Equation (19) gives the result. □

We now prove our main result, Proposition 2:

Proposition 2. *The expected cumulative regret $\mathcal{R}_{n',n}(\boldsymbol{\alpha})$ for Algorithm 1 has the upper bound*

$$\mathcal{R}_n(\mathcal{C}) \leq n' \cdot C + \left(3 \sum_{a \in \omega^{-1}(\mathcal{D}^c)} \Delta_a + \sum_{\substack{a \in \omega^{-1}(\mathcal{D}^c), \\ \Delta_a > 0}} \frac{16 \log(n)}{\Delta_a}\right) + 2 \max_{a \in \mathcal{A}} \Delta_a. \tag{20}$$

Proof of Proposition 2. The expected cumulative regret is given by,

$$\mathcal{R}_n(\mathcal{C}) = n' \cdot C + \mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_{a_1} - Y_{a^{(t)}}\right)\right]. \tag{21}$$

We focus on bounding the cumulative regret incurred by AT-UCB during the UCB phase of AT-UCB, which is represented by the second term. With this goal in mind, consider the event,

$$G = \left\{\hat{\mu}_{\omega(a_1)} \geq \max_{a' \in \mathcal{A}'} \hat{\mu}_{a'} - \epsilon(\boldsymbol{\alpha})\right\} \cap \left(\bigcap_{a' \in \mathcal{D}} \left\{\hat{\mu}_{a'} < \max_{\tilde{a}' \in \mathcal{A}'} \hat{\mu}_{\tilde{a}'} - \epsilon(\boldsymbol{\alpha})\right\}\right) =: G_1 \cap \left(\bigcap_{a' \in \mathcal{D}} G_{a'}\right).$$

Note that G_1 corresponds to the event wherein the empirical mean reward $\hat{\mu}_{\omega(a_1)}$ associated with $\omega(a_1)$ (the intervention to which the optimal action a_1 in $\mathcal{B}$ maps), is *larger* than the highest observed empirical mean reward $\hat{\mu}_{a'}$ minus the threshold hyperparameter $\epsilon(\boldsymbol{\alpha})$. In other words, when G_1 occurs $\omega(a_1)$ is not included in $\hat{\mathcal{D}}$ and thus action a_1 is available when UCB is ran in AT-UCB.

Meanwhile, $G_{a'}$ corresponds to the event where the empirical mean reward $\hat{\mu}_{a'}$ of action a' is *smaller* than the highest observed empirical mean reward $\hat{\mu}_{\tilde{a}'}$ minus the threshold hyperparameter $\epsilon(\boldsymbol{\alpha})$. In other words, when $G_{a'}$ occurs action a' is included $\hat{\mathcal{D}}$, and thus any base action $a \in \omega^{-1}(a')$ mapping to a' is not considered once UCB is ran. When the events $G_{a'}$ for all $a' \in \mathcal{D}$ simultaneously occur, $\mathcal{D} \subseteq \hat{\mathcal{D}}$. That is, all actions $a' \in \mathcal{D}$ are filtered out by AT-UCB and any corresponding base actions $a \in \omega^{-1}(\mathcal{D})$ are not available when UCB is ran.

In summary, the event G_1 ensures that $\mu_{\omega(a_1)}$ is not *relatively underestimated* so that the optimal base arm a_1 is not filtered out and available when UCB is ran whilst $\bigcap_{a' \in \mathcal{D}} G_{a'}$ ensures that all abstract arms in $\mathcal{D}$ are not *relatively overestimated* so that all base arms corresponding to an abstract arm $a' \in \mathcal{D}$ are filtered out before UCB is ran. Thus when the event G occurs $\mathcal{D} \subseteq \hat{\mathcal{D}}$ and $\omega(a_1) \notin \hat{\mathcal{D}}$. That is, the optimal base arm a_1 is not filtered out and all base arms in $\omega^{-1}(\mathcal{D})$ are.

Using the law of total probability, we may re-express cumulative regret incurred during the UCB phase of AT-UCB by conditioning on the event G:

$$\mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_{a_1} - Y_{a^{(t)}}\right)\right] = \mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_{a_1} - Y_{a^{(t)}}\right) \middle| G\right] \cdot \mathbb{P}(G) + \mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_{a_1} - Y_{a^{(t)}}\right) \middle| G^c\right] \cdot \mathbb{P}(G^c) \tag{22}$$

$$\leq \mathbb{E}\left[\sum_{t=1}^{n} \left(\mu_{a_1} - Y_{a^{(t)}}\right) \middle| G\right] \cdot 1 + n \max_{a \in \mathcal{A}} \Delta_a \cdot \mathbb{P}(G^c), \tag{23}$$

where the inequality is justified by that fact that $\mathbb{P}(G) \leq 1$ and the fact that the expected regret on each time step is upper bounded by $\max_{a \in \mathcal{A}} \Delta_a$.

It is clear from previous discussion that the cumulative regret conditioned on the event G is upper-bounded by the cumulative regret of UCB applied to the arm set $\omega^{-1}(\mathcal{D}^c) = \cup_{a' \in \mathcal{D}^c} \omega^{-1}(a')$. Substituting the standard regret bound for UCB into the previous inequality yields

$$\mathbb{E}\left[\sum_{t=1}^{n} (\mu_{a_1} - Y_{a^{(t)}})\right] \leq \left(3 \sum_{a \in \omega^{-1}(\mathcal{D}^c)} \Delta_a + \sum_{\substack{a \in \omega^{-1}(\mathcal{D}^c), \\ \Delta_a > 0}} \frac{16 \log(n)}{\Delta_a}\right) + n \max_{a \in \mathcal{A}} \Delta_a \cdot \mathbb{P}(G^c). \quad (24)$$

All that remains is to upper bound $\mathbb{P}(G^c)$. From the union bound it follows that:

$$\mathbb{P}(G^c) \leq \mathbb{P}(G_1^c) + \sum_{a' \in \mathcal{D}} \mathbb{P}(G_{a'}^c). \quad (25)$$

We will proceed by bounding $\mathbb{P}(G_1^c)$ and $\mathbb{P}(G_{a'}^c)$ for arbitrary $a' \in \mathcal{D}$. For $\mathbb{P}(G_1^c)$, note that

$$\begin{aligned}
\mathbb{P}(G_1^c) &= \mathbb{P}(\hat{\mu}_{\omega(a_1)} < \max_{a' \in \mathcal{A}'} \hat{\mu}_{a'} - \epsilon(\boldsymbol{\alpha})) && (26)\\
&\leq \mathbb{P}\left(\hat{\mu}_{\omega(a_1)} \leq \max_{a' \in \mathcal{A}'} \hat{\mu}_{a'} - \epsilon(\boldsymbol{\alpha})\right) && (27)\\
&\overset{(a)}{\leq} \sum_{a' \in \mathcal{A}'} \mathbb{P}\left(\hat{\mu}_{\omega(a_1)} \leq \hat{\mu}_{a'} - \epsilon(\boldsymbol{\alpha})\right) && (28)\\
&= \sum_{a' \in \mathcal{A}'} \mathbb{P}\left(\epsilon(\boldsymbol{\alpha}) - \mu_{a'} + \mu_{\omega(a_1)} \leq \hat{\mu}_{a'} - \mu_{a'} + \mu_{\omega(a_1)} - \hat{\mu}_{\omega(a_1)}\right) && (29)\\
&= \sum_{a' \in \mathcal{A}'} \mathbb{P}\left(\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)} \leq \hat{\mu}_{a'} - \mu_{a'} + \mu_{\omega(a_1)} - \hat{\mu}_{\omega(a_1)}\right) && (30)\\
&\overset{(b)}{\leq} \sum_{a' \in \mathcal{A}'} \exp\left(-\frac{n'}{4k'}\left[\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)}\right]^2\right), && (31)
\end{aligned}$$

where (a) follows from the union bound. Meanwhile, step (b) follows from the fact that $\hat{\mu}_{a'} - \mu_{a'}$ and $\mu_{\omega(a_1)} - \hat{\mu}_{\omega(a_1)}$ are 1-subgaussian, which in turn implies that $\hat{\mu}_{a'} - \mu_{a'} + \mu_{\omega(a_1)} - \hat{\mu}_{\omega(a_1)}$ is $\sqrt{2k'/n'}$-subgaussian. Note that in order to apply the definition of subgaussian variables we must first ensure that $\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} - \Delta'_{\omega(a_1)}$ is nonnegative. However this follows from the below chain of inequalities, justified by the nonnegativity of Δ'_a for all $a' \in \mathcal{A}'$ and Proposition 1:

$$\epsilon(\boldsymbol{\alpha}) + \Delta'_{a'} \geq \epsilon(\boldsymbol{\alpha}) \geq \Delta'_{\omega(a_1)}. \quad (32)$$

Next, we bound $\mathbb{P}(G_{a'}^c)$ as follows:

$$\begin{aligned}
\mathbb{P}(G_{a'}^c) &= \mathbb{P}\left(\hat{\mu}_{a'} \geq \max_{\tilde{a}' \in \mathcal{A}'} \hat{\mu}_{\tilde{a}'} - \epsilon(\boldsymbol{\alpha})\right) && (33)\\
&\overset{(a)}{\leq} \mathbb{P}\left(\hat{\mu}_{a'} \geq \hat{\mu}_{a'_1} - \epsilon(\boldsymbol{\alpha})\right) && (34)\\
&\leq \mathbb{P}\left(\hat{\mu}_{a'} - \mu_{a'} + \mu_{a'_1} - \hat{\mu}_{a'_1} \geq \Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right) && (35)\\
&\overset{(b)}{\leq} \exp\left(-\frac{n'}{4k'}\left[\Delta'_{a'} - \epsilon(\boldsymbol{\alpha})\right]^2\right) && (36)
\end{aligned}$$

where (a) follows from the fact that $\hat{\mu}_{a'_1} \leq \max_{a' \in \mathcal{A}'} \hat{\mu}_{a'}$ as implied by $a'_1 \in \mathcal{A}'$. Similar to before, (b) follows by appealing to the definition of subgaussian random variables. Note that the positivity of $\Delta'_{a'} - \epsilon(\boldsymbol{\alpha})$ follows from the definition $\mathcal{D}$ which implies $\Delta'_{a'} > \epsilon(\boldsymbol{\alpha})$ for all $a' \in \mathcal{D}$.

Substituting Equation (31) and Equation (36) into Equation (25) and applying Lemma 3 in conjunction with Assumption 1 yields,

$$\mathbb{P}(G^c) \leq \sum_{a' \in \mathcal{A}'} \exp\left(-\frac{n'}{4k'}\left(\epsilon + \Delta'_{a'} - \Delta'_{\omega(a_1)}\right)^2\right) + \sum_{a' \in \mathcal{D}} \exp\left(-\frac{n'}{4k'}\left(\Delta'_{a'} - \epsilon\right)^2\right) \leq \frac{2}{n} \quad (37)$$

Finally, substituting Equation (37) into Equation (24) yields the desired regret bound:

$$\mathbb{E}\left[\sum_{t=1}^{n}\left(\mu_{a_1} - Y_{a^{(t)}}\right)\right] \leq \left(3 \sum_{a \in \omega^{-1}(\mathcal{D}^c)} \Delta_a + \sum_{\substack{a \in \omega^{-1}(\mathcal{D}^c), \\ \Delta_a > 0}} \frac{16 \log(n)}{\Delta_a}\right) + 2 \max_{a \in \mathcal{A}} \Delta_a. \tag{38}$$

□

C Further experimental details

C.1 Base model definition

In our experiment, we consider as our base model $\mathcal{M}$ a discrete population model of an epidemic. In particular, we consider k communities labelled by $c = 1, \ldots, k$, within which the number of individuals who are susceptible to the virus, infected by the virus, and recovered from the virus at time $t \in [0, T]$ are denoted, respectively, by $S_{c,t}$, $I_{c,t}$, and $R_{c,t}$. Each of the k communities consists of an equal number of N individuals, and the simulation progresses in time steps of size $\Delta t = 0.5$ until time $T = 50$. Between times t and $t + \Delta t$, the number of individuals in community c that transition from a state of susceptibility to a state of infection is modelled as

$$N_{c,S \to I,t} \sim \text{Binomial}\left(S_{c,t}, l_{c,t}\left[1 - \exp\left\{-\beta_c \Delta t \cdot \frac{\sum_{c'=1}^{k} I_{c',t}}{kN}\right\}\right]\right), \tag{39}$$

where $\beta_c > 0$ is a community-dependent parameter governing the ease with which individuals within that community become infected, and $l_{c,t} \in \{0, 1\}$ assumes the value 0 if a lockdown is in force at time t and is 1 otherwise. Secondly, the number of individuals transitioning from a state of infection to a recovered state is modelled as

$$N_{c,I \to R,t} \sim \text{Binomial}\left(I_{c,t}, 1 - \exp\left\{-\gamma_c \Delta t\right\}\right), \tag{40}$$

where $\gamma_c > 0$ is another community-dependent parameter. Finally, the number of individuals transitioning from a recovered state back to a state of susceptibility is modelled as

$$N_{c,R \to S,t} \sim \text{Binomial}\left(R_{c,t}, 1 - \exp\left\{-\zeta_c \Delta t\right\}\right), \tag{41}$$

where $\zeta_c > 0$ is a final community-dependent parameter.

In our experiment, we take $k = 10$ communities/regions who are initialised to have $S_{c,0} = 0.9 \cdot N$, $I_{c,0} = 0.1 \cdot N$, $R_{c,0} = 0$. Interventions take the form of an action $a \in \{1, \ldots, k\}$ such that taking action a sets $l_{a,t} = 0$ for $t \in [0.1 \cdot T, 0.5 \cdot T]$. In our experiment, we further take $N = 100$, and take the target variable to be the negative of the total number of individuals that are infected over time, i.e.:

$$Y = -\sum_{c=1}^{k} \sum_{j=0}^{T/\Delta t} I_{c,j \cdot \Delta t}. \tag{42}$$

C.2 Approximate causal abstraction and abstracted model

To construct a high-level model $\mathcal{M}'$ and an approximate causal abstraction between $\mathcal{M}$ and $\mathcal{M}'$, we define a model in the same way as above using $k' = 4$ communities of $N' = 50$ individuals each, and group communities into composite communities via the τ map. The abstract model $\mathcal{M}'$ is also simulated with a larger time step size of $\Delta t' = 2$. Analogosuly to $\mathcal{M}$, we take the abstract target variable Y' to be the negative of the total number of infections over time across all communities in the abstract model:

$$Y' = -\sum_{c'=1}^{k'} \sum_{j=0}^{T/\Delta t'} I'_{c',j\Delta t'}. \tag{43}$$

C.3 Additional experimental details

For both AT-UCB and UCB, we use $\delta = 0.1$ in our experiments.

Code for reproducing the experiments will be released upon acceptance.

All experiments were run on a Macbook Pro 2022 model with M2 chip, requiring ~ 1 hour of CPU time.

D Limitations of AT-UCB

As discussed in Section 5, AT-UCB has several limitations that we now discuss in more detail. Recall that Assumption 1 provides a sufficient lower bound on n', the number of exploratory actions taken in $\mathcal{B}'$, which ensures the set $\mathcal{D}$ can be identified with high probability. Unfortunately the lower bound provided by Assumption 1 depends on the optimality gap associated with each abstract arm $a' \in \mathcal{A}'$. As a result, setting n' in practice requires deep knowledge of the abstract CMAB $\mathcal{B}'$, that is unlikely to be available a priori.

Note that the AT-UCB algorithm bears resemblance to the traditional explore-then-commit (ETC) algorithm [Lattimore and Szepesvári, 2020]. Whilst ETC starts with an initial exploration phase comprised of uniform sampling before committing to the best empirical arm, AT-UCB performs uniform exploration at the abstract level before committing to a subset of of near-optimal arms. Similarly to AT-UCB, tuning the length of the exploration phase in ETC is challenging, with the optimal choice depending on the reward gaps of each arm.

Observe that setting n' in accordance with Assumption 1 also requires knowledge of $\epsilon(\boldsymbol{\alpha})$. Similarly, the regret bound provided by Proposition 2 holds only when the hyperparameter ε is set to $\epsilon(\boldsymbol{\alpha})$. In other words, the theoretical guarantees presented rely on the learner knowing $\epsilon(\boldsymbol{\alpha})$. In many scenarios, this requirement is not problematic. For example, the learner may be able to consult with a domain expert who can provide a trustworthy upper bound on $\epsilon(\boldsymbol{\alpha})$. However, there are settings where it is unreasonable to expect the learner to know $\epsilon(\boldsymbol{\alpha})$, such as when $\mathcal{M}$ and $\mathcal{M}'$ are hard-to-interpret black-box models. Note that $\epsilon(\boldsymbol{\alpha})$ relies on the IC and RD errors, which both provide bounds on the divergence between reward distributions that hold uniformly over pairs of interventions. As a result, actively exploring both $\mathcal{B}$ and $\mathcal{B}'$ to learn $\epsilon(\boldsymbol{\alpha})$ is costly, and developing CAMAB algorithms that do not rely on knowledge of $\epsilon(\boldsymbol{\alpha})$ is non-trivial.

Finally, note that AT-UCB does not exploit any of the causal information encoded in the SCMs $\mathcal{M}$ and $\mathcal{M}'$. As discussed in Appendix A, many algorithms have been proposed for the CMAB setting that exploit partial knowledge of the underlying SCM to facilitate exploration. The fundamental ideas behind such algorithms could be used to augment the uniform exploration phase of AT-UCB when given appropriate partial knowledge of the abstract SCM $\mathcal{M}'$.