INSTANCE-OPTIMAL BEST POLICY IDENTIFICATION IN DECOUPLED CAUSAL MDPS WITH A KNOWN GRAPH

Anonymous authors
Paper under double-blind review

ABSTRACT

We quantify the value of a known causal graph for fixed-confidence best-policy identification. Actions intervene on graph variables and rewards are Gaussian with a known variance and unknown means. A graph-aware learner can ignore every intervention on a variable that cannot influence the reward, whereas a graph-blind learner with the same feedback must rule each one out. The blind characteristic time, which governs sample complexity, grows linearly in the number of reward-irrelevant actions, while the graph-aware time does not. This separation follows from a lower bound on the number of samples any correct learner needs. In decoupled layered causal Markov decision processes (MDPs), the characteristic time is the largest stage time under episodic sampling and the sum under per-stage sampling. An intervention-aware Track-and-Stop algorithm is correct and attains these times whenever no reward-ancestral intervention is worse than the null intervention. Simulations on synthetic causal-bandit and layered-MDP instances match the predicted times and the composition law. Both learners identified the best policy in every run at the target confidence, and with two or more reward-irrelevant actions the graph-aware learner needed 1.4 to 6.8 times fewer samples, a speed-up 15 to 43 percent larger than the leading-order prediction. Code is at `https://anonymous.4open.science/r/iats-crl-AC81`.

1 INTRODUCTION

A Markov decision process (MDP) has states, actions, transitions and rewards. A causal MDP places a directed acyclic graph inside each stage. Its variables are the state features, the variables an action can set and the reward. An action is a hard intervention (Pearl, 2009). It fixes one variable to a value and cuts the arrows into it. Knowing the graph means knowing which variables influence which, not how strongly. A decoupled layered causal MDP has H stages with disjoint variable sets, and each reward depends only on its own stage's intervention. Every stage is visited once per episode, so the model is H one-step problems tied only by how samples are collected. The one-stage case is a causal bandit, and Table 1 compares the two models.

Take a retail store as the running example (Figure 1). A manager sets a product's price, price drives demand, and demand drives the revenue Y, the single reward variable. After checkout, the store may send one of k loyalty notifications, and a notification descends from Y instead of feeding into it. The arms in this example are *set the price high* and *set the price low*, both ancestor arms since price drives revenue, and *send notification* j, a branch arm. In the two-stage version on the right of Figure 1, morning and afternoon each have their own price and $k = 6$ notifications, a moderate branch count fixed in our experiments. No arrow runs between the two stages, and a day is one episode whose return is the sum of the two revenues.

We ask how many samples a known causal graph saves as the number k of branch arms grows. Two learners face this example with the same arms, reward variance σ^2 and baseline mean μ_0, which is the mean revenue when no arm is pulled. Given a unit noise variance and mean revenues of 1.0 at the high price, 0.5 at the low price and $\mu_0 = 0.2$ at the baseline, the gaps are therefore 0.5 between prices and 0.8 to the baseline. A graph-aware learner knows the graph G of Figure 1, the nodes X, D, Y and Z_1 to Z_k and the arrows between them. It therefore knows that every notification arm has

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

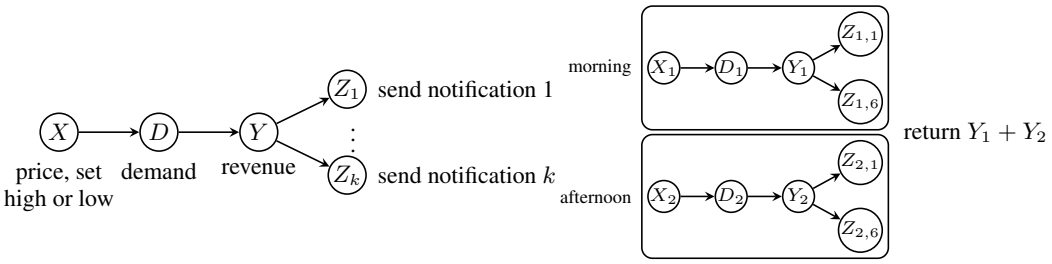


Figure 1: The running example. Left, one stage, where an arm sets the price to one of two values, high or low, or sends one of the k notifications, which lie downstream of the revenue and are therefore branches. Right, the two-stage version, with no arrow between the stages.

Table 1: From the one-stage causal bandit to the decoupled causal MDP.

	Causal bandit	Decoupled causal MDP
Horizon	one stage	H stages, one visit each per episode
Decision	one intervention	one intervention in each stage
Reward	one number	the sum of the H stage rewards
What the graph gives	arms that cannot beat doing nothing	the same, stage by stage
Sample unit	one pull of one arm	one episode or one stage pull
Goal, error at most δ	the best arm	the best arm of every stage
Complexity, episodes	T^*	largest stage T^*
Complexity, stage pulls	T^*	sum of the stage T^*

mean μ_0 and that none can be best, since some price arm beats μ_0. A graph-blind learner must rule each one out by sampling.

The one-stage store is a causal bandit (Lattimore et al., 2016; Sen et al., 2017) whose arms are the two price settings and the k notifications, each returning a Gaussian revenue with mean vector μ. The graph tells the learner that price is an ancestor of revenue and that the notifications are not, since no directed path leads from a notification into Y. A notification therefore leaves the law of Y unchanged (Pearl, 2009), so the notification arms are branches with the baseline mean μ_0. A learner is δ-correct if it recommends the best price with error probability at most δ, and a graph-blind learner must be correct even for a store in which a notification could raise revenue. The optimal expected number of samples grows like $T^*(\mu)\log(1/\delta)$ as the error probability δ shrinks (Garivier & Kaufmann, 2016). The constant $T^*(\mu)$ is the characteristic time of the instance, the value of a change-of-measure program over the mean vectors the learner must rule out (Section 2). The graph shrinks the set of models to rule out, since it removes every model in which a notification arm is best.

Fixed-confidence identification has an exact characteristic time and a matching Track-and-Stop strategy (Garivier & Kaufmann, 2016; Kaufmann et al., 2016), later extended to tabular MDPs (Al Marjani & Proutiere, 2021; Al Marjani et al., 2021), and none of these uses a causal graph. Causal bandits show that a known graph reduces exploration (Lattimore et al., 2016; Sen et al., 2017; Nair et al., 2021; Xiong & Chen, 2023), but the known bounds are one-sided or upper bounds only. To our knowledge, Theorem 1 is the first two-sided bound on the gap in characteristic time between a graph-aware and an otherwise optimal graph-blind learner. We make five contributions. (i) In a causal bandit with a known graph, the graph alone pins every branch mean to μ_0, so fewer models need ruling out and the graph-aware program reduces to the ancestor arms (Section 2). (ii) We prove a two-sided separation for a set A^{anc} of ancestor arms and k branches. Let Δ_0 be the gap between the best arm and μ_0, and $\underline{\Delta}$ the smallest gap over all rivals. The blind characteristic time lies between $(2\sigma^2/\Delta_0^2)\,k$ and $4\sigma^2(|A^{\mathrm{anc}}|+k)/\underline{\Delta}^2$, while the graph-aware time does not depend on k (Theorem 1, Section 3). (iii) We prove a composition law for layered causal MDPs with H stages under a decoupling assumption. The episodic time, for whole-episode sampling, is the largest stage time, and the generative time, for single-stage sampling, is their sum, with matching lower bounds (Theorem 2, Section 4). (iv) We give Intervention-Aware Track-and-Stop (IATS, Algorithm 1), which runs Track-and-Stop per stage on the ancestor arms with an explicit stopping threshold. It is δ-correct and

attains the composed times, which equal the graph-aware times whenever every reward-influencing arm, here each price arm, is at least as good as doing nothing (Theorem 3, Section 5). (v) In our simulations both learners identified the best policy in every run at the target confidence. With two or more notification arms, the graph-aware learner did so with 1.4 to 6.8 times fewer samples, 15 to 43 percent more than the characteristic-time ratio predicts. We also compute exact characteristic times and confirm the composition law (Section 6). Throughout, G is known and causally sufficient, so every variable is observed and there is no hidden confounding. This is the structural view of where to intervene (Lee & Bareinboim, 2018) and the base case for extensions to unknown graphs or unobserved confounders (Appendix B). Our method quantifies the utility of the causal graph. Applied to the store example, the graph-aware characteristic time is 32 whatever the number of notifications, the graph-blind time at $k = 6$ is 55.06, so the graph is worth a factor 1.72 in samples, and the factor grows linearly with the number of notifications.

2 Problem formulation

Lemma 1 shows that an arm on a variable with no arrow path to the reward leaves the reward law unchanged. Corollary 1 then shows that the graph-aware learner has fewer alternatives to rule out and a characteristic time free of the branch count. Lemmas 2 and 3 write both characteristic times as explicit max-min programs, the blind one over all arms and, under condition (H) below, the graph-aware one over the ancestor arms alone. The first definition fixes the causal model and what it means to intervene on one of its variables.

Definition 1 (Causal graph and hard interventions)**.** Let $G = (V, E)$ be a directed acyclic graph over a finite set V of observed random variables containing one real-valued reward variable Y. The edges are arrows $V_i \to V_j$ between variables, and V_i is then a *parent* of V_j. No chain of arrows returns to its start. Let P be a distribution of V that is Markov relative to G. This means $P(v) = \prod_{V_i \in V} P(v_i \mid \mathrm{pa}_i)$, where pa_i denotes the values of the parents of V_i in G. In words, the joint law is the product of one conditional law per variable given its parents. For a variable $X \in V \setminus \{Y\}$ and a value x, the *hard intervention* $\mathrm{do}(X{=}x)$ fixes X to x whatever its parents and keeps every other conditional law. This produces the truncated factorisation (Pearl, 2009, eq. (1.37))

$$P_{\mathrm{do}(X=x)}(v) \;=\; \prod_{V_i \in V \setminus \{X\}} P(v_i \mid \mathrm{pa}_i) \qquad \text{for all } v \text{ consistent with } X = x. \tag{1}$$

Every variable of G is observed and there is no unobserved confounding, so P is Markov relative to G over the observed variables. This is what makes (1) the interventional distribution.

The second definition gives the learner a finite menu of interventions, each returning a noisy reward.

Definition 2 (Causal bandit with a known graph)**.** Let G and P be as in Definition 1. An *arm* is a hard intervention $a = \mathrm{do}(X_a{=}x_a)$ on a single variable $X_a \in V \setminus \{Y\}$. The arm set A is a finite set of arms, $K := |A|$. Pulling arm a returns an independent draw of Y from $P_{\mathrm{do}(a)}$, and $\mu_a := \mathbb{E}_{\mathrm{do}(a)}[Y]$ is its mean reward. The *observational mean* is $\mu_0 := \mathbb{E}_P[Y]$, the mean reward when nothing is intervened on. A causal bandit with a known graph is the tuple (G, P, A).

The third definition separates the arms that can move the reward, in the example the two price arms, from those that cannot, the k notification arms.

Definition 3 (Branches and ancestor arms)**.** $\mathrm{An}(Y)$ denotes the set of *ancestors* of Y in G, the variables with a directed path to Y. A directed path is a chain of arrows leading to Y, so $Y \notin \mathrm{An}(Y)$. Arm a is a *branch* if $X_a \notin \mathrm{An}(Y)$. The set of branches is $B \subseteq A$, $k := |B|$, and $A^{\mathrm{anc}} := A \setminus B$ is the set of *ancestor arms*. For $\mu \in \mathbb{R}^A$, $\mu|_{A^{\mathrm{anc}}} := (\mu_a)_{a \in A^{\mathrm{anc}}}$ is its restriction to A^{anc}.

We work with Gaussian rewards of known variance and a unique best arm that beats doing nothing.

Assumption 1 (Instance regularity)**.** *(a)* For every $a \in A$, $Y \sim N(\mu_a, \sigma^2)$ under $P_{\mathrm{do}(a)}$, with $\sigma^2 > 0$ known. *(b)* The mean vector $\mu = (\mu_a)_{a \in A}$ has a unique best arm $a^* := \arg\max_{a \in A} \mu_a$. *(c)* $\Delta_0 := \mu_{a^*} - \mu_0 > 0$. *(d)* $|A^{\mathrm{anc}}| \geq 2$.

The gap of a rival $a \neq a^*$ is $\Delta_a := \mu_{a^*} - \mu_a$. The smallest ancestor gap is $\Delta_{\min} := \min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} \Delta_a$ and the smallest gap over all rivals is $\underline{\Delta} := \min_{a \in A \setminus \{a^*\}} \Delta_a$. The first result is standard. By the truncated factorisation (1), an intervention on a variable that is not an ancestor of Y cannot change the law of Y.

Lemma 1 (Branches are point-identified)**.** *Let (G, P, A) be as in Definition 2 and let $a = \mathrm{do}(X{=}x) \in B$. Then $P_{\mathrm{do}(X=x)}(Y \in \cdot) = P(Y \in \cdot)$. In particular $\mu_a = \mu_0$, and under Assumption 1(a), $Y \sim N(\mu_0, \sigma^2)$ under P whenever $k \geq 1$.*

The proof is in Appendix C. The graph gives $\mu_a = \mu_0$ for every branch but not the value of μ_0, which the next assumption supplies and Remark 2 in Appendix C shows how to drop.

Assumption 2 (Known observational baseline)**.** The observational mean μ_0 is known to the learner.

We compare two learners that differ only in whether they know the graph. Each must be correct on its own *model class* of mean vectors, so their sample-complexity gap is the value of knowing G.

Definition 4 (Learners and model classes)**.** The *graph-aware* learner and the *graph-blind* learner both know the arm set A, including which variable and value each arm sets, the variance σ^2 and, by Assumption 2, the value μ_0. Neither knows $(\mu_a)_{a \in A^{\mathrm{anc}}}$. The graph-aware learner is in addition given G, and therefore B and A^{anc}, while the graph-blind learner is not. Their model classes are $\mathcal{M}_{\mathrm{G}} := \{\mu \in \mathbb{R}^A : \ \mu_j = \mu_0 \text{ for all } j \in B, \ \max_{a \in A^{\mathrm{anc}}} \mu_a > \mu_0\}$ for the graph-aware learner and $\mathcal{M}_{\mathrm{blind}} := \mathbb{R}^A$ for the graph-blind learner. We require each learner to be δ-correct relative to its model class (Definition 5 below).

The next definition fixes the information available after t rounds, what a strategy is and what counts as success. Remark 1 in Appendix C discusses the class $\mathcal{M}_{\mathrm{G}}$.

Definition 5 (δ-correct strategy relative to a class)**.** *(a) History.* Let $(U_t)_{t \geq 0}$ be independent uniform random variables on $[0, 1]$, independent of all rewards (internal randomisation). Let $a_s \in A$ be the arm pulled at round s and Y_s the reward observed. Let $\mathcal{F}_t := \sigma(U_0, a_1, Y_1, U_1, \ldots, a_t, Y_t, U_t)$ be the sigma-algebra generated by these variables up to round t, the information available after t rounds. *(b) Strategy.* A *strategy* consists of a sampling rule, a stopping time τ with respect to $(\mathcal{F}_t)_{t \geq 0}$ and a recommendation $\hat{a} \in A$ that is $\mathcal{F}_\tau$-measurable. The sampling rule makes a_{t+1} $\mathcal{F}_t$-measurable for every $t \geq 0$, so the next arm is decided from the information available after t rounds. The stopping time property means that whether $\tau \leq t$ is decided by $\mathcal{F}_t$. *(c) Correctness.* For $\mu \in \mathbb{R}^A$, $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ denote law and expectation of the observations when pulling arm a returns an independent $N(\mu_a, \sigma^2)$ draw. Let $\mathcal{S}_A := \{\mu \in \mathbb{R}^A : \ \mu \text{ has a unique best arm}\}$ and write $a^*(\mu)$ for that arm. For a class $\mathcal{M} \subseteq \mathbb{R}^A$ and $\delta \in (0, 1)$, a strategy is *δ-correct relative to $\mathcal{M}$* if for every $\mu \in \mathcal{M} \cap \mathcal{S}_A$, $\mathbb{P}_\mu(\tau < \infty) = 1$ and $\mathbb{P}_\mu(\hat{a} \neq a^*(\mu)) \leq \delta$.

By Assumption 1(a) the true instance is $\mathbb{P}_\mu$ for its mean vector. The last definition measures how hard an instance is for a learner with class $\mathcal{M}$. Hardness depends on the mean vectors the learner must rule out and on the best way to spread its pulls.

Definition 6 (Alternative set and characteristic time)**.** For $\mathcal{M} \subseteq \mathbb{R}^A$ and $\mu \in \mathcal{M} \cap \mathcal{S}_A$, the *alternative set* is $\mathrm{Alt}_{\mathcal{M}}(\mu) := \{\lambda \in \mathcal{M} \cap \mathcal{S}_A : \ a^*(\lambda) \neq a^*(\mu)\}$, the mean vectors of the class with a different best arm. Let Σ_S denote the probability simplex on a finite set S, so that $\Sigma_A = \{w \in \mathbb{R}_+^A : \sum_{a \in A} w_a = 1\}$. Its elements w are called allocations, and w_a is the fraction of pulls given to arm a. For $w \in \Sigma_A$ and $\lambda \in \mathbb{R}^A$ let $c_w(\mu, \lambda) := \sum_{a \in A} w_a(\mu_a - \lambda_a)^2/(2\sigma^2) = \sum_{a \in A} w_a \, \mathrm{KL}(N(\mu_a, \sigma^2), N(\lambda_a, \sigma^2))$, where KL is the relative entropy. In words, c_w is the information per pull that w collects for telling μ from λ. The *characteristic time of μ relative to $\mathcal{M}$* is

$$T^*_{\mathcal{M}}(\mu)^{-1} := \sup_{w \in \Sigma_A} \inf_{\lambda \in \mathrm{Alt}_{\mathcal{M}}(\mu)} c_w(\mu, \lambda), \tag{2}$$

with the convention $\inf \emptyset = +\infty$. Let $T^*_{\mathrm{G}} := T^*_{\mathcal{M}_{\mathrm{G}}}$, $T^*_{\mathrm{blind}} := T^*_{\mathcal{M}_{\mathrm{blind}}}$, $\mathrm{Alt}_{\mathrm{G}} := \mathrm{Alt}_{\mathcal{M}_{\mathrm{G}}}$ and $\mathrm{Alt}_{\mathrm{blind}} := \mathrm{Alt}_{\mathcal{M}_{\mathrm{blind}}}$. A bare $T^*(\mu)$ denotes $T^*_{\mathbb{R}^A}(\mu)$.

In the example the characteristic time equals 32 for the graph-aware learner. Since $\mathcal{M}_{\mathrm{G}}$ pins every branch mean to μ_0, it has fewer alternatives than $\mathcal{M}_{\mathrm{blind}}$, so the branch count drops out of T^*_{G}.

Corollary 1 (of Lemma 1)**.** *Under Definitions 1 to 6 and Assumptions 1 and 2, the following hold.* (i) $\mathrm{Alt}_{\mathrm{G}}(\mu) \subseteq \mathrm{Alt}_{\mathrm{blind}}(\mu)$*, and the inclusion is strict when $k \geq 1$.* (ii) $T^*_{\mathrm{G}}(\mu) \leq T^*_{\mathrm{blind}}(\mu)$. (iii) *If $k \geq 1$, then $T^*_{\mathrm{G}}(\mu)$ is a function of A^{anc}, $(\mu_a)_{a \in A^{\mathrm{anc}}}$ and μ_0 only, and in particular does not depend on k. In that case the supremum in* (2) *may be restricted to allocations supported on A^{anc}.*

The proof is in Appendix C. For $w \in \Sigma_A$ and $a \neq a^*$ define the *rival term*

$$h_a(w) := \frac{\Delta_a^2}{2\sigma^2} \cdot \frac{w_{a^*} w_a}{w_{a^*} + w_a} = \frac{\Delta_a^2}{2\sigma^2 \, (1/w_{a^*} + 1/w_a)}, \qquad h_a(w) := 0 \ \text{ if } w_{a^*} w_a = 0, \tag{3}$$

the rate at which allocation w separates rival a from a^*.

Lemma 2 (Gaussian inner infimum of the blind program)**.** *Let $\mu \in \mathcal{S}_A$ with $K \geq 2$ and $\Delta_a > 0$ for all $a \neq a^*$.* (a) *For every $w \in \Sigma_A$, $\inf_{\lambda \in \mathrm{Alt}_{\mathrm{blind}}(\mu)} c_w(\mu, \lambda) = \min_{a \neq a^*} h_a(w)$, and hence $T^*_{\mathrm{blind}}(\mu)^{-1} = \sup_{w \in \Sigma_A} \min_{a \neq a^*} h_a(w)$.*

The proof is in Appendix E.1. Condition (H) below says that every ancestor arm is at least as good as doing nothing. Under (H), the graph-aware program is that of the ancestor bandit.

Lemma 3 (Reduction of the graph-aware program under (H))**.** *Under Definitions 1 to 6 and Assumptions 1 and 2, suppose in addition that*

$$\mu_a \;\geq\; \mu_0 \qquad \textit{for every } a \in A^{\mathrm{anc}}. \tag{H}$$

(i) *For every $w \in \Sigma_A$, $\inf_{\lambda \in \mathrm{Alt}_{\mathrm{G}}(\mu)} c_w(\mu, \lambda) = \min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} h_a(w)$. Therefore, for every $k \geq 0$, $T^*_{\mathrm{G}}(\mu) = T^*_{\mathrm{anc}}(\mu) := \left[\max_{u \in \Sigma_{A^{\mathrm{anc}}}} \min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} h_a(u)\right]^{-1}$, the characteristic time of the ancestor bandit $(\mu_a)_{a \in A^{\mathrm{anc}}}$ relative to $\mathbb{R}^{A^{\mathrm{anc}}}$. If $|A^{\mathrm{anc}}| = 2$ then $T^*_{\mathrm{anc}}(\mu) = 8\sigma^2/\Delta^2_{\min}$.*

The proof is in Appendix E.2. In the example the graph-aware program is that of the two price arms and $T^*_{\mathrm{G}} = 8\sigma^2/\Delta^2_{\min} = 32$ for every k. The difficulty lies in which arm is best, not in the reward level. Knowing that some arm beats the baseline leaves the blind program unchanged under (H) (Lemma 5 in Appendix C). Knowing which arms cannot be best, the content of the graph, contracts the program.

3 Main results for causal bandits

We analyse both learners with the lower bound of Garivier & Kaufmann (2016, Theorem 1). We state it as a fact, and Appendix E.3 restates its proof in our notation.

Fact 1 (Lower bound of Garivier & Kaufmann, 2016, Theorem 1)**.** *Let $\mathcal{M} \subseteq \mathbb{R}^A$, $\mu \in \mathcal{M} \cap \mathcal{S}_A$ and $\delta \in (0, 1/2)$, and let a strategy be δ-correct relative to $\mathcal{M}$ (Definition 5). Then*

$$\mathbb{E}_\mu[\tau] \;\geq\; T^*_{\mathcal{M}}(\mu)\, \mathrm{kl}(\delta, 1-\delta),$$

where $\mathrm{kl}(p, q) := p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$ is the binary relative entropy. Since $\mathrm{kl}(\delta, 1-\delta)/\log(1/\delta) \to 1$ as $\delta \to 0$, any family $(\tau_\delta)_{\delta \in (0,1/2)}$ of strategies that are δ-correct relative to $\mathcal{M}$ satisfies $\liminf_{\delta \to 0} \mathbb{E}_\mu[\tau_\delta]/\log(1/\delta) \geq T^*_{\mathcal{M}}(\mu)$.

The next definition names what Theorem 1 proves. The blind time grows linearly in k and no faster.

Definition 7 (Two-sided separation)**.** Let $\boldsymbol{\mu} := (\mu^{(k)})_{k \geq 1}$ be a family of instances satisfying Definitions 1 to 3 and Assumptions 1 and 2, where $\mu^{(k)}$ has k branches. The family $\boldsymbol{\mu}$ *exhibits a two-sided linear separation* between the graph-blind and the graph-aware learner if there exist constants $0 < c_1 \leq c_2 < \infty$ such that

$$c_1\, k \;\leq\; \frac{T^*_{\mathrm{blind}}(\mu^{(k)})}{T^*_{\mathrm{G}}(\mu^{(k)})} \;\leq\; c_2\, k \qquad \text{for all } k \geq 1.$$

Two-sided means the blind characteristic time lies between two linear functions of k, so its linear growth is exact in order and not only a lower bound. With six notification arms, part (i) below brackets the exact blind time 55.06 between 18.75 and 128, while the graph-aware time stays at 32.

Theorem 1 (Two-sided separation)**.** *Fix a finite set A^{anc} with $|A^{\mathrm{anc}}| \geq 2$, a vector $(\mu_a)_{a \in A^{\mathrm{anc}}}$ with a unique maximiser a^*, and $\mu_0 < \mu_{a^*}$. Write Δ_0, $\Delta_{\min}$ and $\underline{\Delta} = \min(\Delta_{\min}, \Delta_0)$ as in Section 2. For $k \geq 0$ let $B^{(k)}$ be a set of k branches disjoint from A^{anc}, so $|B^{(k)}| = k$. Let $\mu^{(k)} \in \mathbb{R}^{A^{\mathrm{anc}} \sqcup B^{(k)}}$ be the mean vector on the disjoint union, with $\mu^{(k)}_a = \mu_a$ for $a \in A^{\mathrm{anc}}$ and $\mu^{(k)}_j = \mu_0$ for $j \in B^{(k)}$. Then the following hold for every $k \geq 1$.*

(i) $\dfrac{2\sigma^2}{\Delta_0^2}\, k \;\leq\; T^*_{\mathrm{blind}}(\mu^{(k)}) \;\leq\; \dfrac{4\sigma^2\,(|A^{\mathrm{anc}}| + k)}{\underline{\Delta}^2}.$

(ii) $T^*_{\mathrm{G}}(\mu^{(k)}) = T^*_{\mathrm{G}}(\mu^{(1)}) =: \bar{T}_{\mathrm{G}} \in (0, \infty).$

(iii) Consequently $\dfrac{2\sigma^2}{\Delta_0^2\,\bar{T}_{\mathrm{G}}}\,k \;\le\; \dfrac{T^*_{\mathrm{blind}}(\mu^{(k)})}{T^*_{\mathrm{G}}(\mu^{(k)})} \;\le\; \dfrac{4\sigma^2\,(1+|A^{\mathrm{anc}}|)}{\underline{\Delta}^{\,2}\,\bar{T}_{\mathrm{G}}}\,k$, *so the family* $\boldsymbol{\mu} = (\mu^{(k)})_{k\ge 1}$ *exhibits a two-sided linear separation (Definition 7).*

If moreover (H) *holds, then* $\bar{T}_{\mathrm{G}} = T^*_{\mathrm{anc}}(\mu)$ *and* $T^*_{\mathrm{G}}(\mu^{(k)}) = T^*_{\mathrm{anc}}(\mu)$ *for every* $k \ge 0$ *(Lemma 3).*

Appendix E.4 gives the proof and Appendix C a proof sketch, with Remark 4 on the mechanism, Corollary 2 on the sample-complexity separation and Remark 5 on its size in the example.

4 Extension to layered causal MDPs

The first definition stacks H one-stage causal bandits into one graph and one episode whose return is the sum of the stage rewards. In MDP terms the state at step h is the stage index, the transition is the move to stage $h+1$ and the reward is Y_h.

Definition 8 (Layered causal MDP)**.** Let $H \ge 1$ and let $G^{(H)}$ be a directed acyclic graph over the disjoint union of finite variable sets $V_1, \dots, V_H$, with a real-valued reward variable $Y_h \in V_h$ for each h. Let P be Markov relative to $G^{(H)}$. Interventions on one or several variables follow (1), so the factors of all intervened variables are removed. For each h, A_h is a finite set of hard interventions on single variables of $V_h \setminus \{Y_h\}$, and $P_{\mathrm{do}(a)}$ is the law under the single intervention $a \in A_h$ in $G^{(H)}$. The stage means are $\mu_h(a) := \mathbb{E}_{\mathrm{do}(a)}[Y_h]$ and $\mu_{h,0} := \mathbb{E}_P[Y_h]$. An *episode* chooses $a_h \in A_h$ for every h. The joint intervention $\mathrm{do}(a_1, \dots, a_H)$ is the hard intervention on the H chosen variables, and the episode return is $\sum_{h=1}^{H} Y_h$. The *branches* of stage h are $B_h := \{a \in A_h : X_a \notin \mathrm{An}_{G^{(H)}}(Y_{h'})$ for every $h'\}$, with $k_h := |B_h|$ and $A_h^{\mathrm{anc}} := A_h \setminus B_h$. A (deterministic) *policy* is a tuple $\pi = (a_1, \dots, a_H) \in \prod_h A_h$.

Decoupling says that no stage influences another's reward and that every episode visits every stage.

Assumption 3 (Decoupling)**.** *(a) Stage-local reward ancestry.* $\mathrm{An}_{G^{(H)}}(Y_h) \subseteq V_h$ for every h, so that no variable of another stage is an ancestor of Y_h. *(b) Rectangular visitation.* Every stage is visited exactly once in every episode, whatever the actions. *(c) Per-stage regularity.* For every h and $a \in A_h$, $Y_h \sim N(\mu_h(a), \sigma^2)$ under $P_{\mathrm{do}(a)}$ with σ^2 known. The vector $\mu_h := (\mu_h(a))_{a\in A_h}$ has a unique best arm a_h^* with $\Delta_{0,h} := \mu_h(a_h^*) - \mu_{h,0} > 0$. Moreover $|A_h^{\mathrm{anc}}| \ge 2$, and $\mu_{h,0}$ is known.

Under Assumption 3(a) the stage rewards are independent for every joint intervention. The law of Y_h depends on a_h only, and every branch of stage h has mean $\mu_{h,0}$ (Lemma 6 in Appendix C, proved in Appendix E.5). The value of a policy π under μ is then $V_\mu(\pi) = \sum_h \mu_h(a_h)$, and the optimal policy $\pi^*(\mu)$ is unique and equals $(a_h^*)_{h\le H}$. The number of stages H is the *horizon*. Let $\Sigma_H := \Sigma_{\{1,\dots,H\}}$ be the probability simplex over the H stages and $\mathcal{S}^{(H)} := \prod_h \mathcal{S}_{A_h}$ the set of mean vectors $\mu = (\mu_h)_{h\le H}$ with a unique optimal policy. The product model classes from Definition 4 are $\mathcal{M}_{\mathrm{G}}^{(H)} := \prod_h \mathcal{M}_{\mathrm{G},h}$ and $\mathcal{M}_{\mathrm{blind}}^{(H)} := \prod_h \mathbb{R}^{A_h}$. For $\mathcal{M} = \prod_h \mathcal{M}_h$, $T^*_{\mathcal{M},h} := T^*_{\mathcal{M}_h}(\mu_h)$ is the stage-h characteristic time of Definition 6, written $T^*_{\mathrm{G},h}$ and $T^*_{\mathrm{blind},h}$ for the two classes.

Definition 9 (Sampling protocols, δ-correctness and characteristic times)**.** The *episodic* protocol runs whole episodes $n = 1, 2, \dots$ and the *generative* protocol queries one stage per round $t = 1, 2, \dots$. Here $\mathbf{1}\{\cdot\}$ is the indicator function and $N_{h,a}$ counts the pulls of arm a at stage h.

	episodic	generative (generative-model access)
chosen	$(a_{1,n}, \dots, a_{H,n}) \in \prod_h A_h$ at the start of episode n, from past data and internal randomisation	a pair (h_t, a_t) with $a_t \in A_{h_t}$ in round t
observed	$(Y_{h,n})_{h\le H} \sim \prod_h P_{\mathrm{do}(a_{h,n})}(Y_h \in \cdot)$ (Lemma 6(i)), independent across episodes	$Y_t \sim P_{\mathrm{do}(a_t)}$, the law of Y_{h_t} under a_t
counter	$N_{h,a}(n) := \sum_{m\le n} \mathbf{1}\{a_{h,m} = a\}$, so that $\sum_a N_{h,a}(n) = n$ for every h	$N_{h,a}(t) := \sum_{s\le t} \mathbf{1}\{(h_s, a_s) = (h,a)\}$, so that $\sum_{h,a} N_{h,a}(t) = t$
allocations	$\Omega_{\mathrm{epi}} := \prod_h \Sigma_{A_h}$	$\Omega_{\mathrm{gen}} := \{w \in \prod_h \mathbb{R}_+^{A_h} : \sum_{h,a} w_{h,a} = 1\}$

The number of episodes N is a stopping time, decided at an episode boundary from the data of the completed episodes and internal randomisation. Likewise τ is the generative stopping time.

The learner chooses its arms only at episode boundaries. A choice within an episode that depends on $Y_{1,n}, \dots, Y_{h-1,n}$ would be a conditional intervention, which Definition 8 does not define. By Lemma 6(i) and Assumption 3(c) the true instance is $\mathbb{P}_\mu$ for its mean vector. *δ-correctness.* Here $\mathbb{P}_\mu$ is the law under which the stage-h observation of arm a is an independent $N(\mu_h(a), \sigma^2)$ draw. A strategy (sampling rule, stopping rule, recommended policy $\hat{\pi}$) is δ-correct relative to $\mathcal{M} \subseteq \prod_h \mathbb{R}^{A_h}$ if for every $\mu \in \mathcal{M} \cap \mathcal{S}^{(H)}$ the stopping time is almost surely finite and $\mathbb{P}_\mu(\hat{\pi} \neq \pi^*(\mu)) \leq \delta$. *Alternative set and characteristic times.* For $\mu \in \mathcal{M} \cap \mathcal{S}^{(H)}$, let $\mathrm{Alt}_{\mathcal{M}}(\mu) := \{\mu' \in \mathcal{M} \cap \mathcal{S}^{(H)} : \pi^*(\mu') \neq \pi^*(\mu)\}$ and $c_w(\mu, \mu') := \sum_h \sum_{a \in A_h} w_{h,a}(\mu_h(a) - \mu'_h(a))^2/(2\sigma^2)$. The two times are

$$T^*_{\mathrm{epi},\mathcal{M}}(\mu)^{-1} := \sup_{w \in \Omega_{\mathrm{epi}}} \inf_{\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)} c_w(\mu, \mu'), \qquad T^*_{\mathrm{gen},\mathcal{M}}(\mu)^{-1} := \sup_{w \in \Omega_{\mathrm{gen}}} \inf_{\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)} c_w(\mu, \mu').$$

$T^*_{\mathrm{epi,G}}$, $T^*_{\mathrm{gen,blind}}$ and so on denote these times for $\mathcal{M}^{(H)}_{\mathrm{G}}$ and $\mathcal{M}^{(H)}_{\mathrm{blind}}$.

Theorem 2 (Composition law and lower bounds)**.** *Under Definitions 8 to 9 and Assumption 3, let $\mathcal{M} \in \{\mathcal{M}^{(H)}_{\mathrm{G}}, \mathcal{M}^{(H)}_{\mathrm{blind}}\}$ and $\mu \in \mathcal{M} \cap \mathcal{S}^{(H)}$. Then (i) $T^*_{\mathrm{epi},\mathcal{M}}(\mu) = \max_{h \leq H} T^*_{\mathcal{M},h}$ and (ii) $T^*_{\mathrm{gen},\mathcal{M}}(\mu) = \sum_{h \leq H} T^*_{\mathcal{M},h}$. For every $\delta \in (0, 1/2)$, every strategy that is δ-correct relative to $\mathcal{M}$ satisfies (iii) $\mathbb{E}_\mu[N] \geq \big(\max_h T^*_{\mathcal{M},h}\big)\,\mathrm{kl}(\delta, 1-\delta)$ under the episodic protocol and (iv) $\mathbb{E}_\mu[\tau] \geq \big(\sum_h T^*_{\mathcal{M},h}\big)\,\mathrm{kl}(\delta, 1-\delta)$ under the generative protocol.*

Proof sketch in Appendix C, full proof in Appendix E.6. In the two-stage store one day is one episode. The episodic times are the larger stage times, 55.06 for the blind learner against 32 for the graph-aware one, and the generative times are the sums, 110.12 against 64.

5 Algorithm and guarantees

Intervention-Aware Track-and-Stop (IATS, Algorithm 1) runs the Track-and-Stop strategy of Garivier & Kaufmann (2016) at every stage h, on the ancestor bandit $\nu_h := (\mu_h(a))_{a \in A^{\mathrm{anc}}_h}$ with $K_h := |A^{\mathrm{anc}}_h|$ arms. Each stage stops by the Chernoff rule with threshold

$$\beta_h(t, \delta') := \beta_{\mathrm{an}}(t, \delta'; K_h) := \log\Big(\frac{(K_h - 1)\,\zeta(2)\,t^3}{\delta'}\Big), \qquad \zeta(2) = \sum_{s \geq 1} s^{-2}, \tag{4}$$

at the per-stage confidence $\delta' := \delta/H$. Track-and-Stop recomputes the optimal allocation w^* of Lemma 2 at the empirical means and pulls the arm whose count lags most behind its target share $t\,w^*_a$ (D-tracking). If some arm has fewer than about $\sqrt{t}$ pulls, the rule pulls it instead (forced exploration). Here $\hat{\mu}_h(t) = (\hat{\mu}_{h,a}(t))_{a \in A^{\mathrm{anc}}_h}$ collects the empirical means of the stage-h ancestor arms after t pulls of that stage, and $N_{h,a}(t)$ is the number of pulls of arm a. The empirical best arm is $\hat{a}_h(t) := \arg\max_a \hat{\mu}_{h,a}(t)$. The Chernoff statistic $Z_h(t) := \min_{b \neq \hat{a}_h(t)} \big(\hat{\mu}_{h,\hat{a}_h(t)}(t) - \hat{\mu}_{h,b}(t)\big)^2 \big/ \big[2\sigma^2\big(1/N_{h,\hat{a}_h(t)}(t) + 1/N_{h,b}(t)\big)\big]$, with b ranging over A^{anc}_h, is the generalised likelihood ratio between the empirical best arm and its closest rival. In words, it is their squared gap in units of their combined noise. Each stage first pulls its ancestor arms in turn for K_0 pulls, then follows the D-tracking rule. It stops at $\tau_h := \inf\{t \geq K_0 :\ Z_h(t) \geq \beta_h(t, \delta/H)\}$ and recommends $\hat{a}_h := \hat{a}_h(\tau_h)$, and $\hat{\pi} := (\hat{a}_h)_{h \leq H}$. The generative variant runs the stages one after another with $K_0 = K_h$, so $\tau = \sum_h \tau_h$. The episodic variant runs all stages in parallel on the episode counter, with initialisation length $K_0 = \bar{K} := \max_h K_h$. Once a stage stops, IATS freezes it on its recommendation and keeps pulling that arm, so every stage is pulled once per episode as the protocol requires, and $N = \max_h \tau_h$. The blind version uses A_h and $K_h = |A_h|$. In the running example, IATS ignores the k notification arms at each stage and runs Track-and-Stop on the two price arms, while the blind version runs it on all $k + 2$ arms.

The analysis rests on two single-stage tools stated in Appendix C. Lemma 10 gives δ-correctness of the Chernoff rule under β_{an} for any sampling rule that pulls every arm once. Lemma 9 gives almost-sure stopping at the rate of the characteristic time with an exponential tail.

Theorem 3 (Achievability by IATS)**.** *Let Assumption 3 hold and let $\mu \in \mathcal{M}^{(H)}_{\mathrm{G}} \cap \mathcal{S}^{(H)}$. For every $\delta \in (0, 1)$, run IATS (Definition 10) with the threshold* (4) *at the per-stage confidence δ/H, so $\beta_h(t, \delta/H) = \log(H(K_h - 1)\zeta(2)t^3/\delta)$. Then IATS (i) is δ-correct relative to $\mathcal{M}^{(H)}_{\mathrm{G}}$,*

Algorithm 1: IATS, episodic variant, with threshold (4) at confidence δ/H per stage. With $r = 1$ it is the strategy of Definition 10. A_h^{anc} is an ordered list and arm j its j-th element. $w^*(\hat{\mu}_h(t))$ maximises the program of Lemma 2 at $\hat{\mu}_h(t)$, recomputed every r episodes and cached in between.

Input: confidence δ, graph $G^{(H)}$, arm sets $A_1, \dots, A_H$, variance σ^2, reallocation period $r \geq 1$
Output: recommended policy $\hat{\pi} = (\hat{a}_1, \dots, \hat{a}_H)$ and episode count N
foreach *stage* h **do** $A_h^{\mathrm{anc}} \leftarrow A_h \setminus B_h$ with B_h from $G^{(H)}$ (Definition 8), $K_h \leftarrow |A_h^{\mathrm{anc}}|$;
$\bar{K} \leftarrow \max_h K_h$, $n \leftarrow 0$, all stages active;
while *some stage is active, with* $n \leftarrow n+1$ *at the start of each episode* **do**
  foreach *stage* h **do**
    if h *is frozen* **then** pull $\hat{a}_h$ at stage h and discard the observation;
    else if $n \leq \bar{K}$ **then** pull arm $((n-1) \bmod K_h) + 1$ of A_h^{anc} (initialisation);
    else with $t = n-1$ and $\mathcal{L}_t = \{a : N_{h,a}(t) < \sqrt{t} - K_h/2\}$, pull $\arg\min_{a \in \mathcal{L}_t} N_{h,a}(t)$ if $\mathcal{L}_t \neq \emptyset$ (forced exploration), else $\arg\max_a \left(t\, w_a^*(\hat{\mu}_h(t)) - N_{h,a}(t)\right)$ (D-tracking);
  foreach *active stage* h *with* $n \geq \bar{K}$ **do**
    update $\hat{\mu}_h(n)$, $\hat{a}_h(n)$ and $Z_h(n)$;
    if $Z_h(n) \geq \beta_h(n, \delta/H)$ **then** $\tau_h \leftarrow n$, $\hat{a}_h \leftarrow \hat{a}_h(n)$, mark h frozen;
return $\hat{\pi} = (\hat{a}_h)_{h \leq H}$ and $N = n = \max_h \tau_h$;

under both protocols, (ii) under the generative protocol satisfies $\limsup_{\delta \to 0} \mathbb{E}_\mu[\tau]/\log(1/\delta) \leq \sum_{h \leq H} T^*_{\mathrm{anc}}(\mu_h)$*, and (iii) under the episodic protocol satisfies* $\limsup_{\delta \to 0} \mathbb{E}_\mu[N]/\log(1/\delta) \leq \max_{h \leq H} T^*_{\mathrm{anc}}(\mu_h)$*. If* (H) *holds at every stage, then* $T^*_{\mathrm{anc}}(\mu_h) = T^*_{\mathrm{G},h}$ *(Lemma 3), so the rates in (ii) and (iii) are those of Theorem 2. The same statements hold for the blind version of the strategy with* $K_h = |A_h|$*,* $\mathcal{M}^{(H)}_{\mathrm{blind}}$ *in place of* $\mathcal{M}^{(H)}_{\mathrm{G}}$ *and* $T^*_{\mathrm{blind},h}$ *in place of* $T^*_{\mathrm{anc}}(\mu_h)$*.*

Proof sketch in Appendix C, full proof in Appendix E.9. Without (H) the rate $T^*_{\mathrm{anc}}(\mu_h)$ can exceed the lower bound $T^*_{\mathrm{G},h}$ (Remark 3), and closing this gap is open. The simulations of Section 6 use the analysed threshold β_{an}, which is δ-correct by Lemma 10. The calibrated threshold β_{cal} of Garivier & Kaufmann (2016, Section 6) serves for comparison only and is not proved δ-correct.

6 Experiments

Both learners identified the best policy in every run at the target confidence, with zero errors in every cell of the correctness and rate check. With two or more notification arms, the graph-aware learner needed 1.42 to 6.77 times fewer samples, a speed-up 15 to 43 percent larger than the leading-order prediction. No public benchmark exists for fixed-confidence identification with a known graph, so the released code generates every instance. The confirmatory family F-IATS has four experiments with 300 seeds per cell and meets 32 of its 33 pre-specified checks, all but the rate-stability criterion of the correctness and rate check. The run protocol and the result files label the four confirmatory experiments E1 (exact times), E2 (correctness and rate), E3 (realized speed-up) and E4 (horizon), and the illustrative coupled-case study E5. Appendix C reports E1 to E4 in full, Appendix F lists the criteria (Table 5) and Appendix D.3 holds E5.

Exact-time computation. The store of Section 1 with $k \in \{0, 1, 2, 4, 8, 16, 32, 64\}$ notification arms gives $T^*_{\mathrm{G}} = 32$ for all k, as Theorem 1(ii) predicts, while T^*_{blind} rises to 264.16. The ratio rises from 1.12 at $k = 1$ to 8.26 at $k = 64$ (Figure 2).

Correctness and rate check. On the store with $k = 8$ notifications, both learners with the threshold β_{an} of (4) made no error at any $\delta \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The per-cell Clopper-Pearson bound 9.9×10^{-3} then certifies the level of Theorem 3 at $\delta \geq 10^{-2}$. Theorem 3 says that the mean stopping time $\bar{\tau}$ relative to $T^* \log(1/\delta)$ tends to 1, with T^* the characteristic time. At the tested levels this ratio instead tracks $\beta_{\mathrm{an}}(\bar{\tau}, \delta; K)/\log(1/\delta)$ (Appendix C). The check that $\bar{\tau}/(T^* \log(1/\delta))$ is stable in δ therefore fails by construction under β_{an} and holds under β_{cal}.

Realized speed-up experiment. The store with $k \in \{0, 2, 4, 8, 16, 32\}$ notifications at $\delta = 0.01$ gives blind-to-IATS ratios $\bar{\tau}_{\mathrm{blind}}/\bar{\tau}_{\mathrm{IA}}$ of 1.00, 1.42, 1.75, 2.46, 3.84, 6.77, and the 95% bootstrap

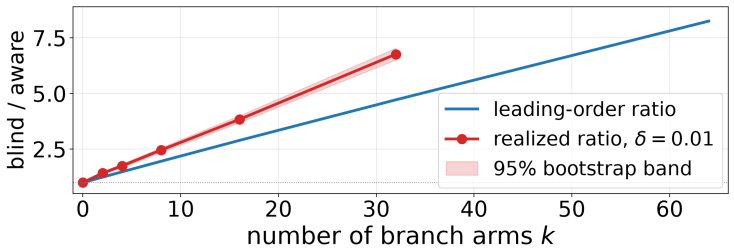


Figure 2: Separation in the running example against the branch count $k \leq 64$. The blue line is the exact leading-order ratio $T^*_{\text{blind}}/T^*_{\text{G}}$ of Definition 6. The red line is the realized ratio $\bar{\tau}_{\text{blind}}/\bar{\tau}_{\text{IA}}$ from the realized speed-up experiment, at $\delta = 0.01$ over 300 seeds with an unpaired bootstrap 95% band.

intervals exclude 1 for $k \geq 2$. They exceed the leading-order ratio by a factor 1.15 to 1.43, inside the pre-specified envelope $[0.7, 1.6]$ (Appendix C explains the excess).

Horizon experiment. A day has H selling periods, each a store with 6 notifications. The episodic ratio $\max_h T^*_{\text{blind},h} / \max_h T^*_{\text{G},h}$ is $55.06/32 = 1.72$ for all $H \in \{1, 2, 4, 8, 16\}$, and the generative difference $\sum_h (T^*_{\text{blind},h} - T^*_{\text{G},h})$ is $23.06\,H$ (exact linear fit, $R^2 = 1.0$). Both match Theorem 2.

7 Related work

Fixed-confidence pure exploration runs from PAC best-arm identification (Mannor & Tsitsiklis, 2004; Even-Dar et al., 2006; Kalyanakrishnan et al., 2012; Jamieson et al., 2014) through the characteristic-time program and Track-and-Stop (Garivier & Kaufmann, 2016; Kaufmann et al., 2016) and extensions (Degenne et al., 2019; Degenne & Koolen, 2019; Kaufmann & Koolen, 2021; Jourdan et al., 2022). It reaches tabular MDPs (Al Marjani & Proutiere, 2021; Al Marjani et al., 2021) and PAC reinforcement learning (Fiechter, 1994; Wagenmaker et al., 2022; Domingues et al., 2021). A known graph reduces single-step exploration (Lattimore et al., 2016; Sen et al., 2017; Nair et al., 2021; Xiong & Chen, 2023). The closest bounds are one-sided (Xiong & Chen, 2023), upper bounds only (Sen et al., 2017; Nair et al., 2021) or fixed-budget simple regret (Madhavan et al., 2021). None proves a two-sided separation from an optimal blind learner (Appendix B, Table 3).

8 Conclusion

A known causal graph enters fixed-confidence best-policy identification through the alternative set. Branches are point-identified (Lemma 1), so the graph-aware characteristic time is free of their number while the graph-blind one grows linearly in it (Theorem 1). Under decoupling the separation lifts to layered causal MDPs, where the characteristic time composes as a maximum or a sum according to the protocol (Theorem 2). IATS attains these times with a threshold valid for any sampling rule, whenever no reward-ancestral intervention is worse than the null one (Theorem 3). The experiments match them up to the finite-δ terms of the threshold. Under the episodic protocol the bottleneck stage sets the speed-up, which fell to $818.90/800 = 1.02$ in the heterogeneous-horizon variant of Appendix C because that stage gains little from the graph. The coupled case remains open. Appendix D shows that the blind lower bound survives coupling and conjectures achievability.

The theory assumes a known and causally sufficient graph (Definition 1), Gaussian rewards with known variance (Assumption 1) and decoupling in the MDP part (Assumption 3). It also assumes a known observational baseline, which Remark 2 shows how to remove. Unknown graphs and unobserved confounders are outside its scope. Appendix B relates our setting to that literature (Lee & Bareinboim, 2018; Shahverdikondori et al., 2026).

AI use statement

In this work, we used a large language model assistant for tasks with required disclosure. It assisted in formulating intermediate lemmas and checking proofs and derivations for the theoretical results.

We did not use generative AI tools to generate synthetic data sets or to formulate the research questions. The remaining tasks with required disclosure are not applicable to this work. We also used the assistant to edit the text of the paper, a task with recommended disclosure. We have reviewed all AI-assisted work. Every proof was checked by the authors. We take full responsibility for the final content of this work.

Ethics statement

This paper is theoretical, and its experiments run on synthetic instances generated by the released code. No human subjects, personal data or proprietary data are involved.

Reproducibility statement

Every assumption is stated in Sections 2 and 4. Every result stated in the main text is proved in full, in Appendix E or in Appendix C. Appendix C also holds the deferred remarks and proof sketches. Appendix A lists the complete notation. Appendix F describes the released code, the exact configurations, the pre-specified acceptance gates and the committed result files behind every number and figure. The code is available at `https://anonymous.4open.science/r/iats-crl-AC81`, and one command reproduces all experiments.

References

Aymen Al Marjani and Alexandre Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *PMLR*, pp. 7459–7468, 2021.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in markov decision processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 25852–25864, 2021.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, pp. 41–53, 2010.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sébastien Racanière, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *International Conference on Learning Representations (ICLR)*, 2019.

Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 379–387, 2014.

Arnoud de Kroon, Joris Mooij, and Danielle Belgrave. Causal bandits without prior knowledge using separating sets. In *Proceedings of the 1st Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *PMLR*, pp. 407–427, 2022.

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

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

Omar Darwiche Domingues, Pierre Ménard, Emilie Kaufmann, and Michal Valko. Episodic reinforcement learning in finite MDPs: Minimax lower bounds revisited. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, volume 132 of *PMLR*, pp. 578–598, 2021.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7:1079–1105, 2006.

Shi Feng and Wei Chen. Combinatorial causal bandits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7550–7558, 2023.

Claude-Nicolas Fiechter. Efficient reinforcement learning. In *Proceedings of the 7th Annual ACM Conference on Computational Learning Theory (COLT)*, pp. 88–97, 1994.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory (COLT)*, volume 49 of *PMLR*, pp. 998–1027, 2016.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pp. 423–439, 2014.

Yassir Jedra and Alexandre Proutiere. Optimal best-arm identification in linear bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Marc Jourdan, Rémy Degenne, Dorian Baudry, Rianne de Heide, and Emilie Kaufmann. Top two algorithms revisited. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. PAC subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 655–662, 2012.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 1238–1246, 2013.

Emilie Kaufmann and Wouter M. Koolen. Mixture martingales revisited with applications to sequential tests and confidence intervals. *Journal of Machine Learning Research*, 22(246):1–44, 2021.

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

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 2568–2578, 2018.

Yangyi Lu, Amirhossein Meisami, Ambuj Tewari, and William Yan. Regret analysis of bandit problems with causal background knowledge. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *PMLR*, pp. 141–150, 2020.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. Efficient reinforcement learning with prior causal knowledge. In *Proceedings of the 1st Conference on Causal Learning and Reasoning (CLeaR)*, volume 177 of *PMLR*, pp. 526–541, 2022.

Rahul Madhavan, Aurghya Maiti, Gaurav Sinha, and Siddharth Barman. Intervention efficient algorithm for two-stage causal MDPs. *arXiv preprint arXiv:2111.00886*, 2021.

Aurghya Maiti, Vineet Nair, and Gaurav Sinha. A causal bandit approach to learning good atomic interventions in presence of unobserved confounders. In *Proceedings of the 38th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 180 of *PMLR*, pp. 1328–1338, 2022.

Shie Mannor and John N. Tsitsiklis. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5:623–648, 2004.

Vineet Nair, Vishakha Patil, and Gaurav Sinha. Budgeted and non-budgeted causal bandits. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 130 of *PMLR*, 2021.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, UK, 2nd edition, 2009. ISBN 978-0-521-89560-6.

Rajat Sen, Karthikeyan Shanmugam, Alexandros G. Dimakis, and Sanjay Shakkottai. Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, 2017.

Mohammad Shahverdikondori, Jalal Etesami, and Negar Kiyavash. Graph learning is suboptimal in causal bandits. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Burak Varıcı, Karthikeyan Shanmugam, Prasanna Sattigeri, and Ali Tajer. Causal bandits for linear structural equation models. *Journal of Machine Learning Research*, 24(297):1–59, 2023.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent PAC reinforcement learning. In *Proceedings of the 35th Conference on Learning Theory (COLT)*, volume 178 of *PMLR*, pp. 358–418, 2022.

Po-An Wang, Ruo-Chun Tzeng, and Alexandre Proutiere. Fast pure exploration via frank-wolfe. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Nuoya Xiong and Wei Chen. Combinatorial pure exploration of causal bandits. In *International Conference on Learning Representations (ICLR)*, 2023.

Akihiro Yabe, Daisuke Hatano, Hanna Sumita, Shinji Ito, Naonori Kakimura, Takuro Fukunaga, and Ken-ichi Kawarabayashi. Causal bandits with propagating inference. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 5512–5520, 2018.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 13401–13411, 2019.

The appendix is organised as follows. Appendix A lists the notation. Appendix B is the extended related work with the comparison table. Appendix C collects the remarks, auxiliary statements and proofs deferred from the main text, in the order of the main text. Appendix D treats the coupled program beyond decoupling, with partial results and conjectures. Appendix E contains the full proofs of the results of Sections 2 to 5. Appendix F documents the numerical witnesses, the pre-specified acceptance criteria and the reproduction protocol.

A Notation

Table 2 collects the notation used in the paper.

Table 2: Notation.

Symbol	Meaning
$G=(V,E)$, Y, $\mathrm{An}(Y)$	causal DAG, reward variable, ancestors of Y in G (Defs. 1 and 3)
P, $P_{\mathrm{do}(X=x)}$, $P_{\mathrm{do}(a)}$	observational distribution, distribution under the hard intervention $\mathrm{do}(X{=}x)$, law of Y under arm a (Defs. 1 and 2)
X, x, X_a, x_a, V_i, pa_i	intervened variable and its value, the variable and value of arm a, a generic variable of V and the values of its parents in G (Defs. 1 and 2)
A, K, $a=\mathrm{do}(X_a{=}x_a)$, a_0	arm set, $K=\lvert A\rvert$ (also the arm count of a generic bandit $\nu\in\mathbb{R}^K$ in Lemmas 9 and 10, and the largest action count of a decision state in the coupled-case study E5), an arm, the null intervention (Remark 2)
B, k, A^{anc}, $\mu\lvert_{A^{\mathrm{anc}}}$	branches ($X_a\notin\mathrm{An}(Y)$), $k=\lvert B\rvert$ (also $\lvert B_s\rvert$ in Prop. 2), ancestor arms $A\setminus B$, restriction of μ to A^{anc} (Def. 3)
D, $\bar{D}$	$\mathrm{An}(Y)\cup\{Y\}$ and its complement in V (proof of Lemma 1), and in the proof of Lemma 6 the unrolled versions $V=\bigcup_h V_h$ and $D=\bigcup_h D_h$ with $D_h=\mathrm{An}(Y_h)\cup\{Y_h\}$
μ, μ_a, μ_0, σ^2	mean vector $(\mu_a)_{a\in A}$ (Assumption 1), mean of Y under $\mathrm{do}(a)$, observational mean, known variance
δ, δ'	error-probability parameter of a δ-correct strategy (Def. 5), per-stage confidence δ/H of IATS (Def. 10)

Symbol	Meaning
t, a_t, Y_t, n	round (pull) index within a stage, arm pulled and reward observed at round t (Def. 5), episode index (Def. 9)
a^*, Δ_a, Δ_0, $\Delta_{\min}$, $\underline{\Delta}$	best arm and the gaps $\mu_{a^*} - \mu_a$, $\mu_{a^*} - \mu_0$, $\min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} \Delta_a$, $\min_{a \neq a^*} \Delta_a$
$\mathcal{S}_A$, $a^*(\mu)$	mean vectors with a unique best arm, and that arm
$\mathcal{M}$, $\mathcal{M}_{\mathrm{G}}$, $\mathcal{M}_{\mathrm{blind}}$, $\mathcal{M}'_{\mathrm{blind}}$	a model class $\subseteq \mathbb{R}^A$ (Def. 5), model classes of the two learners (Def. 4), and the blind class restricted by Assumption 1(c) (Lemma 3)
(U_t), $\mathcal{F}_t$, τ, τ_δ, $\hat{a}$	internal randomisation, filtration, stopping time, stopping time of the strategy run at confidence δ (Fact 1, Lemma 9), recommendation (Def. 5)
$\mathbb{P}_\mu$, $\mathbb{E}_\mu$, $\mathbb{P}_\nu$, $\mathbb{E}_\nu$, $\mathbb{P}_M$, $\mathbb{E}_M$	law and expectation of the observations under the mean vector μ (Def. 5), under a Gaussian bandit ν (Lemma 9), under an MDP model M (Def. 11)
$N_a(t)$, $\hat{\mu}_a(t)$	pull count and empirical mean of a after t rounds
$\mathrm{Alt}_{\mathcal{M}}(\mu)$, $\mathrm{Alt}_{\mathrm{G}}$, $\mathrm{Alt}_{\mathrm{blind}}$, c_w	alternative set and its shorthands for the two classes, weighted transportation cost (Def. 6), both overloaded for layered vectors (Def. 9) and, as $\mathrm{Alt}_{\mathcal{M}}(M)$ and $c_w(M, M')$, for MDP models (Def. 11)
λ, μ', M'	alternative mean vector (Def. 6), alternative layered vector (Def. 9), alternative MDP model (Def. 11)
Σ_A, $\Sigma_{A^{\mathrm{anc}}}$, Σ_H	probability simplex on A, on the ancestor arms and on $\{1, \dots, H\}$ (Def. 6 and Section 4)
w, w_a, w_h, $w_{h,a}$, u, u_h	sampling allocation in Σ_A and its coordinates (Def. 6), stage-h block and coordinates of a multi-stage allocation (Def. 9), an allocation in $\Sigma_{A^{\mathrm{anc}}}$ or Σ_{A_h}
$T^*_{\mathcal{M}}$, T^*_{G}, T^*_{blind}, T^*_{anc}, T^*	characteristic times (Def. 6, Lemma 3), bare $T^*(\mu)$ meaning $T^*_{\mathbb{R}^A}(\mu)$ (Def. 6) and, in Sections 5–6, the time of the learner under discussion, and the coupled program $T^*_{\mathcal{M}}(M)$ of an MDP (Def. 11)
$h_a(w)$, w^*, $F(w)$, $F_{\mathrm{anc}}(w)$	rival term (3), maximiser of the blind program, blind inner infimum $\min_{a \neq a^*} h_a(w)$ (Lemma 2), its ancestor-only version (App. E.2)
z^*	common value $(w_{a^*}\mu_{a^*} + w_a\mu_a)/(w_{a^*} + w_a)$ of the two-arm boundary alternative (Remark 3, App. E.1)
$\mu^{(k)}$, $B^{(k)}$, $\bar{T}_{\mathrm{G}}$, c_1, c_2	separation family with k branches at μ_0, its branch set, its common graph-aware time $T^*_{\mathrm{G}}(\mu^{(1)})$, separation constants (Def. 7, Theorem 1)
KL, $\mathrm{kl}(p, q)$	Kullback–Leibler divergence, binary relative entropy $p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q}$
H, h, $G^{(H)}$, V_h	horizon, stage index, unrolled DAG, stage-h variables (Def. 8)
A_h, B_h, k_h, A_h^{anc}, K_h	stage-h arms, branches, $\lvert B_h \rvert$, ancestor arms, arm count tracked at stage h
a_h, a^*_h, $\mathrm{do}(a_1, \dots, a_H)$, μ_h	action at stage h, best arm of stage h, joint intervention of an episode, stage-h mean vector $(\mu_h(a))_{a \in A_h}$ (Def. 8, Assumption 3)
$P_{\mathrm{do}(a)}$, $\mu_h(a)$, $\mu_{h,0}$, $\Delta_{0,h}$, Y_h	law under the single intervention $a \in A_h$ in $G^{(H)}$, stage-h mean under a, stage-h observational mean, stage-h branch gap $\mu_h(a^*_h) - \mu_{h,0}$, stage-h reward
π, $\pi^*(\mu)$, $V_\mu(\pi)$	policy $(a_1, \dots, a_H)$, optimal policy, value $\sum_h \mu_h(a_h)$
$\mathcal{M}^{(H)}_{\mathrm{G}}$, $\mathcal{M}^{(H)}_{\mathrm{blind}}$, $\mathcal{M}_h$, $\mathcal{M}_{\mathrm{G},h}$, $\mathcal{S}^{(H)}$	product model classes, stage-h factor of a product class, stage-h graph-aware class, vectors with a unique optimal policy (Section 4)
N, $N_{h,a}$, Ω_{epi}, Ω_{gen}	episode count, stage-h pull counts, feasible allocations (Def. 9)
$T^*_{\mathrm{epi},\mathcal{M}}$, $T^*_{\mathrm{gen},\mathcal{M}}$, $T^*_{\mathcal{M},h}$, T^*_h, $T^*_{\mathrm{G},h}$, $T^*_{\mathrm{blind},h}$, $T^*_{\mathrm{epi,G}}$, $T^*_{\mathrm{gen,blind}}$, G_h, θ_h	episodic and generative times, stage-h time and its shorthand, the stage-h, episodic and generative times relative to $\mathcal{M}^{(H)}_{\mathrm{G}}$ and $\mathcal{M}^{(H)}_{\mathrm{blind}}$ (Section 4), stage-h inner infimum (5), stage mass $\sum_a w_{h,a}$ (proof of Theorem 2)
ν, ν_h, $T^*(\nu)$, $\Delta_{\min,h}$	a Gaussian bandit, the stage-h ancestor bandit, its characteristic time relative to $\mathbb{R}^K$, smallest gap of ν_h (Def. 10, Lemma 9, Section 5)
$\hat{\mu}_h(t)$, $\hat{\mu}_{h,a}(t)$, $N_{h,\cdot}(t)$, $\hat{a}_h(t)$	stage-h empirical mean vector, its coordinates, vector of stage-h pull counts, empirical best arm after t pulls (Def. 10), with $Z(t)$, $\hat{a}(t)$ the single-stage versions of $Z_h(t)$, $\hat{a}_h(t)$ (Lemma 10)
$\mathcal{L}_t$, $Z_h(t)$, $\beta_h(t, \delta')$, K_0, $\bar{K}$	under-sampled arms (forced exploration), Chernoff statistic, stage-h threshold $\log((K_h - 1)\zeta(2)t^3/\delta')$, initialisation length, $\max_h K_h$ (Def. 10)
β_{an}, β_{cal}, $\zeta(2)$, $\zeta(3)$	analysed threshold $\log((K-1)\zeta(2)t^3/\delta)$ (Def. 10, Lemma 10), calibrated comparison threshold $\log((1 + \log t)(K-1)/\delta)$ (Sec. 5), $\zeta(2) = \sum_{s \geq 1} s^{-2} = \pi^2/6$ and $\zeta(3) = \sum_{s \geq 1} s^{-3}$ (error bound of Lemma 10)
$X_{a,s}$, $\bar{\mu}_{a,m}$	s-th draw of arm a in the stack-of-rewards representation, mean of its first m draws (App. E.8)

Symbol	Meaning
τ_h, $\hat{a}_h$, $\hat{\pi}$	stage-h stopping time and recommendation, and the recommended policy
α, $r(t)$, C_r	exponent, rate function and constant in the hypothesis $r(t) \le C_r t^{\alpha}$ of Lemma 9 ($r(t) = (K-1)\zeta(2)t^3$, $\alpha = 3$, $C_r = (K-1)\zeta(2)$ for β_{an})
ε, $\varepsilon_0(\nu)$, $\mathcal{I}_\varepsilon$, $\mathcal{E}_T$, b_ε, c_ε	neighbourhood radius of Lemma 9 and its largest admissible value, neighbourhood, event and constants of Lemma 19 of Garivier & Kaufmann (2016) (App. E.7)
$C^*_\varepsilon(\nu)$, T_ε, T_1, ψ, $T_0(\delta)$, $T_\varepsilon(\delta)$, D_ε, $\kappa(\varepsilon)$	rate constant and deterministic times, auxiliary function $\psi(T)$ of one variable, tail constant and rate factor of Lemma 9 (App. E.7), where the two-variable function $\chi(x,y) = xy/(x+y)$ of App. E.1 is a different function
$\chi(x,y)$, v^*, $\mathcal{G}_\delta$, KL_a, Π_h, Π^0_h, Σ_H, ℓ_δ, n_*, n_b	local symbols of the proofs, the function $xy/(x+y)$ and the optimal value of the blind program (App. E.1), the success event and the per-arm divergence $(\mu_a - \lambda_a)^2/(2\sigma^2)$ (App. E.3), the stage factors of the truncated factorisation (App. E.5), the simplex on the stage indices (App. E.6), $\log(1/\delta)$ (App. E.7), and the counts $N_{a^*}(\tau)$, $N_b(\tau)$ (App. E.8)
M, S, S_h, S_{dec}, s_{root}, $p(s' \mid s,a)$, $p_0(\cdot \mid s)$	finite-horizon causal MDP, states, layer h, decision states, root state, transition kernel, observational kernel (Def. 11)
A_s, $\mathcal{A}$, $\mu(s,a)$, B_s, $\mu_0(s)$, $\mu(L)$, $\mu(R)$	action set at state s, all state-action pairs $\bigsqcup_s A_s$, mean of action a at s, branch actions and observational mean at s (Def. 11), means of the single leaf actions of Example 1
Q_M, V_M, π^*_M, π°_M	state-action value, state value, optimal policy, local-argmax policy (Def. 11)
$\Omega(M)$, $w(s,a)$, ρ_π, q	occupancy polytope, occupancy, visit probabilities (Def. 11), and root occupancy $w(s_{\mathrm{root}}, 0)$ in App. E.13
$\mathcal{M}_{\mathrm{blind}}(M)$, $\mathcal{M}_{\mathrm{G}}(M)$, $\mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M)$, $\mathcal{M}^{\mathrm{al}}_{\mathrm{blind}}(M)$, $\mathcal{M}^{\mathrm{al}}$	model classes of the coupled case, their greedy-aligned members, and the aligned class corresponding to the class at hand (Def. 11, Prop. 3)
$g_{s,a}$, $R_{\mathcal{M}}$, $T^*_{\mathrm{sf},\mathcal{M}}$, $T^*_{\mathrm{sf,G}}$, $T^*_{\mathrm{sf,blind}}$, T^*_{sf}	single-flip rival term, rival set, single-flip program, its versions for the two classes, and the common value when both agree (Def. 11)
$Q_0(s)$, $\Delta_Q(s)$, Δ_{root}	Q-value of a branch at s, Q-margin of the optimal action over the branches (Prop. 2), and the root Q-gap $\lvert Q_M(s_{\mathrm{root}},0) - Q_M(s_{\mathrm{root}},1)\rvert$ of Example 1
Γ, Φ, Ψ	quantile map of the kernel (App. E.10), and model and occupancy maps (App. E.11)
$\bar{\tau}$, $\bar{\tau}_{\mathrm{IA}}$, $\bar{\tau}_{\mathrm{blind}}$, $\bar{N}_H$	mean stopping time over seeds, its values for IATS and the blind learner, mean episode count of the episodic run with horizon H (Sections 5–6)
ptp, CP	peak-to-peak range $\max - \min$ of a sequence, Clopper–Pearson upper bound on the error rate (Section 6, Table 5)

B Extended related work

Fixed-confidence characteristic time. The change-of-measure program that defines T^* and the Track-and-Stop strategy that attains it are due to Garivier & Kaufmann (2016), building on the transportation inequality of Kaufmann et al. (2016). Al Marjani & Proutiere (2021) and Al Marjani et al. (2021) instantiate the program for best-policy identification in tabular MDPs. Their upper bounds match the characteristic time up to a factor 4 (Al Marjani & Proutiere, 2021, Thm. 3) and 2 (Al Marjani et al., 2021, Thm. 6). Fixed-confidence pure exploration has a long lineage. It includes PAC best-arm and subset identification via elimination and confidence-based sampling rules (Even-Dar et al., 2006; Kalyanakrishnan et al., 2012; Jamieson et al., 2014), and fixed-budget and exponential-gap algorithms (Audibert et al., 2010; Karnin et al., 2013). It also includes combinatorial pure exploration (Chen et al., 2014) and the sample-complexity lower bound of Mannor & Tsitsiklis (2004). Later work extends the Track-and-Stop template in several directions. These include a game-theoretic non-asymptotic analysis (Degenne et al., 2019), multiple correct answers (Degenne & Koolen, 2019) and mixture-martingale stopping thresholds (Kaufmann & Koolen, 2021). They also include linear structure (Jedra & Proutiere, 2020), Frank–Wolfe allocation solvers (Wang et al., 2021) and Top-Two sampling (Jourdan et al., 2022). The PAC-RL best-policy-identification formulation dates to Fiechter (1994). Instance-dependent PAC-RL (Wagenmaker et al., 2022) and minimax episodic lower bounds (Domingues et al., 2021) bracket the tabular-MDP rate. None of these works has causal or graph content. Our lower-bound object is the program of Garivier & Kaufmann (2016) applied to two model classes (Definition 6). Our algorithm is their algorithm run on a graph-restricted arm set. This paper adds three things. It defines the alternative set through a causal

graph (Lemma 1, Corollary 1). It compares the graph-aware learner with an otherwise optimal blind learner (Theorem 1). It proves the composition law (Theorem 2).

Relation to Track-and-Stop for MDPs. IATS can be read as the Track-and-Stop variant of Al Marjani & Proutiere (2021) for best-policy identification in MDPs, restricted by the graph. This reading leaves four points open, and the analysis has to settle them. First, Al Marjani & Proutiere (2021) allocate samples over the state-action pairs of the MDP. IATS instead restricts the stage-h allocation to the reward-ancestral arms A_h^{anc} and never pulls a branch arm. This restriction is admissible for three reasons. Lemma 1 pins every branch mean at the observational mean. Assumption 3(c) excludes a branch from being optimal. Assumption 3 decouples the stages (Lemma 6). Second, the analysed threshold (4) has the exponent $\alpha = 3$, which lies outside the range $[1, e/2]$ of Theorem 14 of Garivier & Kaufmann (2016). The rate T^* rather than αT^* therefore requires the direct tail bound of Lemma 9 (Remark 8). Third, the δ-correctness of that threshold (Lemma 10) is a union bound over the pairs of pull counts, with the explicit Gaussian constant $1 - \zeta(3)/\zeta(2)$. The bound holds for every sampling rule. This is what lets the released code cache the allocation without touching the error guarantee. Fourth, under the episodic protocol the stages share one clock. A stopped stage is frozen on its recommendation and keeps being pulled, and the episode count is $N = \max_h \tau_h$. The leading constant $\max_h T^*_{\mathrm{G},h}$ of Theorem 3(iii) therefore follows from the tail bound of Lemma 9, not from its consequence for the expected stopping time alone. The upper bound of Al Marjani & Proutiere (2021) matches its characteristic time up to a factor 4, whereas Theorem 3 attains the constant of Theorem 2 under (H).

Causal bandits ($H = 1$). Prior work establishes that a known graph reduces exploration in single-step best-intervention identification. The causal-bandit literature spans cumulative-regret methods with causal background knowledge (Lu et al., 2020), propagating-inference simple regret (Yabe et al., 2018) and atomic interventions under unobserved confounders (Maiti et al., 2022). It also covers linear structural-equation models (Varıcı et al., 2023) and combinatorial intervention spaces (Feng & Chen, 2023). The bounds closest to ours are of a different type. Lattimore et al. (2016) give fixed-budget simple-regret bounds with a minimax guarantee. Nair et al. (2021) prove instance-dependent upper bounds for simple regret (Thms. 1–2) and cumulative regret (Thms. 3–4) but no lower bound of their own. Sen et al. (2017) give instance-dependent upper bounds (Thms. 1–2) and leave the matching lower bound open (their Section 5). Xiong & Chen (2023) give the first gap-dependent adaptive pure-exploration bounds for causal bandits. Their lower bound (Thm. 3) is one-sided and nearly matches their upper bounds (Thms. 1–2) on a parallel-graph instance class. It involves no inf-over-alternatives characteristic time and no lower bound for a graph-blind learner, since they compare informally with LUCB1. Their mechanism is observation-probability reweighting rather than contraction of the alternative set. de Kroon et al. (2022) treat an unknown graph. Shahverdikondori et al. (2026) prove that learning the graph is suboptimal for regret under an unknown graph, a complementary result. None of these works states an instance-optimal characteristic time or a lower bound for a graph-blind learner (Table 3).

Unknown graphs and unobserved confounders. Two standing assumptions of this paper, a perfectly known graph and causal sufficiency, are relaxed in recent work. Lee & Bareinboim (2018) characterise, from the causal structure alone, which intervention sets can be optimal, so the remaining arms need not be played. This reduction of the arm set is of the same kind as the restriction to A^{anc} here. Shahverdikondori et al. (2026) treat an unknown graph and show that learning the graph first is suboptimal for regret. Both questions come before the one studied here. Our question starts from a fixed arm set and asks how fast a δ-correct learner can certify the best policy. The bounds of this paper are the known-graph, causally sufficient base case. Any extension to an unknown graph or to unobserved confounders must recover T^*_{G} and the separation of Theorem 1 when the graph is known and no confounder is present.

Causal MDPs. Causal structure has been used in several parts of sequential decision making. Examples are near-optimal reinforcement learning in confounded dynamic treatment regimes (Zhang & Bareinboim, 2019), counterfactually guided policy search (Buesing et al., 2019) and sample-efficient reinforcement learning from prior causal knowledge of the transition structure (Lu et al., 2022). None of these targets a fixed-confidence characteristic time. The closest known-graph causal-MDP result is Madhavan et al. (2021) on two-stage causal MDPs. It gives fixed-budget simple regret

Table 3: Comparison with the closest prior results. The column *Char.-time obj.* records whether the bound is expressed through an inf-over-alternatives program, schematically $\sup_w \inf_{M'} \sum w \,\mathrm{KL}$ (Definition 6). Here w is a sampling allocation, M' an alternative model and KL the Kullback–Leibler divergence. The column *Blind LB* records whether a lower bound is proved for a learner denied the causal graph. In the Setting column, regret stands for the fixed-budget or regret setting and PE for fixed-confidence pure exploration or best-policy identification.

Work	Setting	Bound type	Two-sided	Char.-time obj.	Blind LB
GK / KCG (Garivier & Kaufmann, 2016; Kaufmann et al., 2016)	PE (BAI)	instance-dep.	yes (Thm. 1)	yes	n/a
AMP (Al Marjani & Proutiere, 2021), AMGP (Al Marjani et al., 2021)	PE (BPI, MDP)	instance-dep.	to $4\times/2\times$ (Thm. 3/6)	yes	no
Xiong–Chen (Xiong & Chen, 2023)	PE (causal)	instance-dep.	one-sided (Thm. 1–3)	no	no (informal)
Madhavan (Madhavan et al., 2021)	simple regret (causal MDP)	instance-dep.	up to logs (Thm. 1–2)	no	no
Sen (Sen et al., 2017)	regret (causal)	instance-dep.	UB only (Thm. 1–2)	no	no
Nair (Nair et al., 2021)	regret (causal)	instance-dep.	UB only (Thm. 1–4)	no	no
This paper	PE (causal, BPI)	instance-dep.	yes (Thm. 1, 2, 3)	yes	yes (Fact 1, Thm. 1)

with an instance-dependent upper bound (Thm. 1) and a minimax lower bound up to logarithmic factors (Thm. 2). It is two-sided, with an upper and a lower bound, but it is not a characteristic-time object. It has no graph-blind baseline and proves no separation from a graph-blind learner. Their cross-stage maximum is a maximum of two regret constants inside one bound. The episodic $\max_h$ of Theorem 2 is instead a composition of per-stage characteristic times for δ-correct identification. Their use of convex optimisation for the exploration design is consistent with Proposition 1.

Comparison. To our knowledge, no prior work proves a two-sided characteristic-time separation (Definition 7) for known-graph causal best-policy identification. Such a separation compares a graph-aware learner with an otherwise optimal graph-blind learner. Table 3 records three properties of each of the closest works. These are whether the bound is two-sided, whether it is expressed through an inf-over-alternatives program, and whether a lower bound for a graph-blind learner is proved. The claim rests on this conjunction only. The separation mechanism is standard (Section 3), and the change-of-measure program and the Track-and-Stop template are those of Garivier & Kaufmann (2016). The single-step advantage of a known graph is prior art (Lattimore et al., 2016).

C Statements and proofs deferred from the main text

This appendix collects the remarks, corollaries, full proofs and auxiliary statements that the main text states in outline or refers to by number. The subsections follow the order of the main text.

C.1 Asymptotic conventions (Section 1)

Throughout, $o(1)$ and $(1 + o(1))$ refer to $\delta \to 0$ with the instance μ fixed. Statements "of order k" refer to $k \to \infty$ along the family $\{\mu^{(k)}\}$ defined in Theorem 1, in which everything except the branch count is fixed. No other asymptotic regime is used, and the finite-δ ratios reported in Section 6 are not asymptotic statements. Three remarks on the results follow. The lower bound of Garivier & Kaufmann (2016) (Fact 1) on each class turns the contraction of the alternative set into a difference of characteristic times. Through Fact 1, the separation of Theorem 1 is therefore a separation in expected sample complexity against every δ-correct blind strategy (Corollary 2, Appendix C). The episodic rate of Theorem 3 rests on a tail estimate for Track-and-Stop (Lemma 9). Section 7 places these results among prior work, and Appendix D treats the coupled case as a conjecture with partial results.

C.2 The two learners and the alternative set (Section 1)

In fixed-confidence identification, δ is the allowed probability of recommending a wrong arm. Structural knowledge enters through the alternative set. The graph-aware learner knows the vertices and edges of a directed acyclic graph G, but neither the structural equations nor any reward mean. Arms are hard interventions $\mathrm{do}(X{=}x)$ on variables X of G, as in causal bandits (Lattimore et al., 2016; Sen et al., 2017), and the reward is a variable Y of G. A branch is an arm that intervenes on a non-ancestor of Y. It leaves the law of Y unchanged (Pearl, 2009) (Lemma 1), so its mean is μ_0. Once some ancestral arm beats μ_0 (Assumption 1), the blind learner must sample every branch to exclude it, and the graph-aware learner need not.

Table 4: Three models, what the learner knows, what it must learn, and where each appears.

Model	Learner knows	Learner must learn	Where
MDP	state and action sets, sampled transitions and rewards	best action per state	background
Causal MDP	also the graph of each stage	best intervention per stage	Appendix D
Decoupled causal MDP	also that stages do not interact	best intervention per stage	Sections 2 to 5

C.3 Three models (Section 1)

Table 4 sets the three models of Section 1 side by side.

C.4 Causal bandit with a known graph (Section 2)

Section 2 defines the objects used in the rest of the paper. These are the causal bandit with a known graph, the two learners and their model classes, δ-correct strategies, alternative sets and characteristic times. Table 2 (Appendix A) collects the notation. The truncated factorisation (1) is that of Pearl (2009, Section 1.3.1, eq. (1.37)). Assumption 1(a) is the standard Gaussian causal-bandit assumption, and it restricts the interventional laws of Y only.

Proof of Lemma 1

Proof. Let $D := \mathrm{An}(Y) \cup \{Y\}$ and $\bar{D} := V \setminus D$. Since $X \notin \mathrm{An}(Y)$ and $X \neq Y$, $X \in \bar{D}$. If $V_i \in \bar{D}$ has a child $V_j \in D$, then $V_i \to V_j \to \cdots \to Y$ (or $V_j = Y$) is a directed path, so $V_i \in \mathrm{An}(Y)$, a contradiction. So every child of a member of $\bar{D}$ lies in $\bar{D}$, and every parent of a member of D lies in D. The factors $P(v_j \mid \mathrm{pa}_j)$ with $V_j \in D$ therefore involve only variables in D, and none involves X. Marginalising (1) over all variables other than Y (sums for discrete variables, integrals otherwise),

$$P_{\mathrm{do}(X=x)}(y) = \sum_{(v_i)_{V_i \in D \setminus \{Y\}}} \prod_{V_j \in D} P(v_j \mid \mathrm{pa}_j) \cdot \left[\sum_{(v_i)_{V_i \in \bar{D} \setminus \{X\}}} \prod_{V_i \in \bar{D} \setminus \{X\}} P(v_i \mid \mathrm{pa}_i) \right].$$

In the bracket, take $V_i \in \bar{D} \setminus \{X\}$ maximal in a topological order of G. Its children lie in $\bar{D}$. Each child either equals X, whose factor is absent from (1), or comes later in the order. So V_i occurs only in its own factor, which sums to one. Iterating over $\bar{D} \setminus \{X\}$ in decreasing topological order shows that the bracket equals one. Hence $P_{\mathrm{do}(X=x)}(y) = \sum_{(v_i)_{V_i \in D \setminus \{Y\}}} \prod_{V_j \in D} P(v_j \mid \mathrm{pa}_j)$, which does not depend on x. The same computation applied to P itself gives the same expression for $P(y)$. There the factor of X is present, and it is summed out in the bracket in the same way because the children of X lie in $\bar{D}$. Thus $P_{\mathrm{do}(X=x)}(Y \in \cdot) = P(Y \in \cdot)$ and $\mu_a = \mu_0$. The last claim follows from Assumption 1(a) applied to any $a \in B$. □

Remark 1 (What the graph does and does not give). Lemma 1 can also be read off rule 3 of the do-calculus (Pearl, 2009, Theorem 3.4.1). Consider the graph obtained by deleting the edges into X. Every path from X to Y in this graph leaves X through a child. Since X is not an ancestor of Y, the path must contain a collider, so X and Y are d-separated. The graph gives only the implication "non-ancestor $\Rightarrow$ no effect". An ancestor of Y may have zero or negative effect, so ancestor arms must be sampled. The graph gives the *equality* $\mu_a = \mu_0$ for $a \in B$, not the value of μ_0. This is the reason for Assumption 2 below. Under Assumption 1(c) a branch is never the best arm, so $a^* \in A^{\mathrm{anc}}$.

C.5 Learners, model classes and characteristic times (Section 2)

This paragraph expands on the model classes of Definition 4. The true mean vector belongs to both classes, by Lemma 1 and Assumption 1(c). The class $\mathcal{M}_{\mathrm{G}}$ encodes what the graph-aware learner can infer from G, μ_0 and Assumption 1(c). The class $\mathcal{M}_{\mathrm{blind}}$ is all of $\mathbb{R}^A$, for two reasons. First, a learner that does not know G cannot exclude any mean vector on graphical grounds. Every mean vector is consistent with some graph satisfying Definition 2, for instance one in which every

intervened variable is a parent of Y. Second, we also require the blind learner to be correct on the vectors with $\max_a \mu_a \leq \mu_0$, which Assumption 1(c) excludes for the true instance. Lemma 5 below shows that, under condition (H), restricting $\mathcal{M}_{\text{blind}}$ to the vectors with $\max_a \mu_a > \mu_0$ does not change the blind characteristic time. So this convention costs the blind learner nothing at the leading order in the cases we study. The two learners face the same instance, the same feedback and the same δ, and differ only in the class on which correctness is required. Fact 1 below gives the change-of-measure lower bound for each class. Section 4 shows when it is attained, unconditionally for the blind class and under condition (H) for the graph-aware class.

Remark 2 (Discharging Assumption 2). Assumption 2 enters only through the definition of $\mathcal{M}_{\text{G}}$. The sampling and stopping rules of Section 5 do not use the value of μ_0. The assumption can be removed by adding the null intervention a_0 (no intervention) to the arm set and pooling the branches with a_0. This pooling is justified because Lemma 1 makes the branches copies of a_0. The graph-aware learner then faces the bandit on $A^{\text{anc}} \cup \{a_0\}$, whose characteristic time again does not depend on k, at the price of one extra arm. We do not pursue this variant, and the released code follows the convention of Definition 4.

Two observations accompany Assumption 2 and Definition 4. Assumption 2 is discharged by adding a null-intervention arm a_0 with mean μ_0 to the arm set. This costs one extra arm and leaves the graph-aware time independent of k (Remark 2, Appendix C). The graph yields exactly the constraint $\mu_j = \mu_0$ for $j \in B$ and no constraint on the ancestor arms. So $\mathcal{M}_{\text{G}}$ encodes all the non-parametric information in G (Remark 1, Appendix C).

One point about Definition 5 concerns internal randomisation. The change-of-measure inequality of Kaufmann et al. (2016, Lemma 1), used in Fact 1 below, is stated there for strategies without internal randomisation. It holds verbatim for the strategies of Definition 5. The likelihood ratio between two mean vectors on $\mathcal{F}_\tau$ depends on the rewards only, because the law of (U_t) is the same under both (Appendix E.3).

For $\mathcal{M} = \mathbb{R}^A$, (2) is equation (1) of Garivier & Kaufmann (2016). Corollary 1 of Section 2, proved next, is the consequence of Lemma 1 that contracts the alternative set.

PROOF OF COROLLARY 1

Proof. (i) $\mathcal{M}_{\text{G}} \subseteq \mathbb{R}^A$ gives the inclusion. If $k \geq 1$, pick $j \in B$ and let $\lambda_j := \mu_{a^*} + 1$, $\lambda_a := \mu_a$ for $a \neq j$. Then $\lambda \in \text{Alt}_{\text{blind}}(\mu)$ but $\lambda \notin \mathcal{M}_{\text{G}}$. (ii) An infimum over a smaller set is larger. (iii) For $\lambda \in \mathcal{M}_{\text{G}}$ and $j \in B$, $\lambda_j = \mu_0 = \mu_j$ by Lemma 1, so $c_w(\mu, \lambda) = \sum_{a \in A^{\text{anc}}} w_a(\mu_a - \lambda_a)^2/(2\sigma^2)$. A vector $\lambda \in \mathcal{M}_{\text{G}}$ has a unique best arm if and only if the maximum of $(\lambda_a)_{a \in A^{\text{anc}}}$ is attained at a single ancestor arm. The constraint $\max_{a \in A^{\text{anc}}} \lambda_a > \mu_0$ excludes the branches from the maximum. Hence, identified with its ancestor coordinates, $\text{Alt}_{\text{G}}(\mu) = \{\lambda \in \mathbb{R}^{A^{\text{anc}}} : \max_a \lambda_a > \mu_0,\ \arg\max_a \lambda_a \text{ is a single arm } \neq a^*\}$, which does not depend on k. Finally, $c_w(\mu, \lambda)$ is non-decreasing in each w_a, so moving allocation mass from B to a^* does not decrease the inner infimum. The supremum over Σ_A therefore equals the supremum over allocations supported on A^{anc}. □

C.6 STRUCTURE OF THE BLIND PROGRAM AND THE BLIND CLASS UNDER (H) (SECTION 2)

Lemma 2 of Section 2 computes the blind inner infimum. It is Lemma 3 of Garivier & Kaufmann (2016) specialised to Gaussian arms. The maximiser of the blind program has the positivity and equalisation properties stated next. They are labelled (b) and (c) to match their proof, which appears together with that of Lemma 2 in Appendix E.1.

Lemma 4 (Structure of the blind program)**.** *Let $\mu \in \mathcal{S}_A$ with $K \geq 2$ and $\Delta_a > 0$ for all $a \neq a^*$. Then the following hold.*

(b) $F(w) := \min_{a \neq a^*} h_a(w)$ is continuous on Σ_A and attains its maximum. Every maximiser w^* satisfies $w^*_a > 0$ for all $a \in A$.

(c) At every maximiser w^*, $h_a(w^*) = T^*_{\text{blind}}(\mu)^{-1}$ for every $a \neq a^*$.

The blind counterpart of Lemma 3 concerns condition (H). Under that condition, requiring correctness on all of $\mathbb{R}^A$ rather than only on the mean vectors compatible with Assumption 1(c) does not change the blind characteristic time. Its part is labelled (ii) to match its proof, which is given together with that of Lemma 3 in Appendix E.2.

Lemma 5 (Reduction of the blind program under (H)). *Under Definitions 1 to 6 and Assumptions 1 and 2, suppose in addition that* (H) *holds.* (ii) *Let* $\mathcal{M}'_{\mathrm{blind}} := \{\lambda \in \mathbb{R}^A : \max_{a \in A} \lambda_a > \mu_0\}$*, the set of mean vectors compatible with Assumption 1(c). For every* $w \in \Sigma_A$, $\inf_{\lambda \in \mathrm{Alt}_{\mathcal{M}'_{\mathrm{blind}}}(\mu)} c_w(\mu, \lambda) = \min_{a \neq a^*} h_a(w)$*, and consequently* $T^*_{\mathcal{M}'_{\mathrm{blind}}}(\mu) = T^*_{\mathrm{blind}}(\mu)$.

Remark 3. Condition (H) says that no reward-ancestral intervention is worse than doing nothing. It holds in all experiments of Section 6 ($\mu|_{A^{\mathrm{anc}}} = (1.0, 0.5)$, $\mu_0 = 0.2$), so there $T^*_{\mathrm{G}} = 8\sigma^2/0.5^2 = 32$. Without (H), Corollary 1(iii) still gives that $T^*_{\mathrm{G}}(\mu)$ does not depend on $k \geq 1$. Also $T^*_{\mathrm{G}}(\mu) \leq T^*_{\mathrm{anc}}(\mu)$, because every element of $\mathrm{Alt}_{\mathrm{G}}(\mu)$, read in ancestor coordinates, is an alternative of the ancestor bandit. Condition (H) is what makes the lower bound T^*_{G} and the rate T^*_{anc} attained by the algorithm of Section 5 coincide. The inequality can be strict without (H). For $A^{\mathrm{anc}} = \{a^*, a\}$, $\mu = (1, -5)$, $\mu_0 = 0$ and $\sigma = 1$, the alternative λ of Appendix E.2 with $\lambda_a = \lambda_{a^*} = z^* := (w_{a^*}\mu_{a^*} + w_a\mu_a)/(w_{a^*} + w_a)$ has $\max_b \lambda_b \leq \mu_0$ and is not in $\mathcal{M}_{\mathrm{G}}$. The graph-aware inner infimum at $w = (1/2, 1/2)$ is $13/(2\sigma^2)$ (approached as $\lambda \to (0, 0)$), whereas $h_a(w) = 9/(2\sigma^2)$. Lemma 5 plays the same role for the blind learner. Under (H), requiring correctness on all of $\mathbb{R}^A$ rather than on $\mathcal{M}'_{\mathrm{blind}}$ does not change T^*_{blind}. The proof of Lemma 3(i) uses (H) only to ensure that the approximating alternatives have a maximum above μ_0. It suffices that for every rival $a \in A^{\mathrm{anc}} \setminus \{a^*\}$ either $\mu_a \geq \mu_0$ or some $b \in A^{\mathrm{anc}} \setminus \{a, a^*\}$ has $\mu_b > \mu_0$.

The following details accompany Section 2. $T^*_{\mathcal{M}}(\mu)$ is the number of samples per unit of $\log(1/\delta)$ that the best allocation needs to make every alternative implausible. Table 4 lists what is known and learned in each of the three models of Section 1. Ancestor arms may have zero or negative effect and must be sampled (Remark 1). The true mean vector belongs to both classes, by Lemma 1 and Assumption 1(c). A generic $\mu \in \mathbb{R}^A$ need not be realised by any causal model with graph G, for instance one with $\mu_j \neq \mu_0$ for a branch j. Every maximiser of the blind program of Lemma 2 gives positive mass to every arm and equalises all rival terms (Lemma 4). Definition 5 is the δ-PAC property of Garivier & Kaufmann (2016) with model class $\mathcal{M} \cap \mathcal{S}_A$. Lemma 5 shows that requiring blind correctness also when $\max_a \mu_a \leq \mu_0$ does not change its characteristic time under condition (H). That condition holds in all experiments of Section 6 and makes the lower bound T^*_{G} and the rate T^*_{anc} attained in Section 5 coincide (Remark 3). Throughout, $N_a(t) := \sum_{s \leq t} \mathbf{1}\{a_s = a\}$ denotes the number of pulls of a after t rounds, and $\hat{\mu}_a(t)$ denotes their empirical mean.

C.7 Proof sketch, mechanism and magnitude of Theorem 1 (Section 3)

Proof sketch of Theorem 1. For the blind class the inner infimum at an allocation w is $\min_{a \neq a^*} h_a(w)$ (Lemma 2(a)). The k branches share mass at most one, so some branch has $w_j \leq 1/k$ and $h_j(w) \leq \Delta_0^2/(2\sigma^2 k)$. The uniform allocation gives every rival term at least $\underline{\Delta}^2/(4\sigma^2(|A^{\mathrm{anc}}| + k))$, which proves (i). Corollary 1(iii) gives (ii), and dividing (i) by $\bar{T}_{\mathrm{G}}$ gives (iii). The full proof is in Appendix E.4.

Remark 4. The lower bound in Theorem 1(i) is the arm-count scaling of best-arm identification. The k branches have the common gap Δ_0, and each contributes a rival term $h_j(w) \leq \Delta_0^2 w_j/(2\sigma^2)$. The constraint $\sum_{j \in B} w_j \leq 1$ forces $\min_j h_j(w) \leq \Delta_0^2/(2\sigma^2 k)$ for every allocation. The upper bound is certified by the uniform allocation, which is a feasible point of the program and not a maximiser. What is separated is the coefficient of $\log(1/\delta)$, which does not vanish as $\delta \to 0$. The contribution is not the scaling law. It is the identification, from the graph and before any sample is drawn, of the k arms that a blind learner must nevertheless sample (Lemma 4(b)).

Corollary 2 (Sample-complexity separation). *Under the hypotheses of Theorem 1 and Assumptions 1–2, for every* $k \geq 1$ *and* $\delta \in (0, 1/2)$ *the following hold.*

(i) Every strategy which is δ*-correct relative to* $\mathcal{M}_{\mathrm{blind}}$ *satisfies* $\mathbb{E}_{\mu^{(k)}}[\tau] \geq \dfrac{2\sigma^2 k}{\Delta_0^2} \mathrm{kl}(\delta, 1-\delta)$.

(ii) The graph-aware strategy of Definition 10 with $H = 1$ *is* δ*-correct relative to* $\mathcal{M}_{\mathrm{G}}$ *and satisfies* $\limsup_{\delta \to 0} \mathbb{E}_{\mu^{(k)}}[\tau_\delta]/\log(1/\delta) \leq T^*_{\mathrm{anc}}(\mu)$*, a quantity that does not depend on* k.

Proof. (i) is Fact 1 with $\mathcal{M} = \mathcal{M}_{\mathrm{blind}}$ combined with Theorem 1(i). (ii) is Theorem 3 with $H = 1$. □

Remark 5 (Magnitude). Consider the instance of Section 6, with $\mu|_{A^{\mathrm{anc}}} = (1.0, 0.5)$, $\mu_0 = 0.2$ and $\sigma = 1$. There $\Delta_0 = 0.8$, $\Delta_{\min} = \underline{\Delta} = 0.5$ and (H) holds. Lemma 3 gives $T^*_{\mathrm{G}} = 8\sigma^2/\Delta^2_{\min} = 32$ for every k. The constant in the lower bound of Theorem 1(i) is $2\sigma^2/\Delta_0^2 = 3.125$. The exact ratios $T^*_{\mathrm{blind}}/T^*_{\mathrm{G}}$ computed from Definition 6 for $k \le 64$ are summarised in Section 6 and listed in the full report of experiment E1 below. They show that the separation is linear in k with a small constant, and that the ratio is single-digit for $k \le 64$. They also show that the harmonic coupling of the rival terms through w_{a^*} keeps the finite-k ratio above the lower bound of (i).

The following details accompany Section 3. The lower bound of Garivier & Kaufmann (2016) is restated as a fact because it is not a contribution of this paper. By Lemma 1, $\mu^{(k)}$ is the mean vector of every causal bandit satisfying Definition 2 and Assumption 1 with the given ancestor data and k branches. Theorem 1 therefore covers every such instance. The graph identifies, before any sample is drawn, the k arms that a blind learner must nevertheless sample (Lemma 4(b)). The restriction to $\delta < 1/2$ is the only range used here. It is needed because the bound fails for δ close to one, and it cannot be dropped. For $K = 2$, stopping after one pull and recommending uniformly at random is δ-correct for every $\delta \ge 1/2$.

C.8 Remarks, auxiliary lemmas and the IATS strategy in full (Section 4)

This section collects the material that Section 4 states in outline or refers to by number. It contains the remarks, the proofs of the auxiliary lemmas, the formal definition of the strategy and two single-stage tools.

On Definition 8. The branch set B_h of Definition 8 is a set determined by $G^{(H)}$, so the branches of every stage are known before any sample is drawn.

Factorisation under decoupling. The following lemma, used throughout Section 4, is proved in Appendix E.5.

Lemma 6 (Factorisation under decoupling)**.** *Under Definition 8 and Assumption 3(a), for every episode $(a_1, \dots, a_H) \in \prod_h A_h$ the following hold. (i) $P_{\mathrm{do}(a_1,\dots,a_H)}(y_1, \dots, y_H) = \prod_{h=1}^{H} P_{\mathrm{do}(a_h)}(y_h)$, so $Y_1, \dots, Y_H$ are independent under every joint intervention and the law of Y_h depends on a_h only. (ii) For every $j \in B_h$, $P_{\mathrm{do}(j)}(Y_h \in \cdot) = P(Y_h \in \cdot)$, hence $\mu_h(j) = \mu_{h,0}$. (iii) $B_h = \{a \in A_h : X_a \notin \mathrm{An}_{G^{(H)}}(Y_h)\}$.*

Remark on Assumption 3.

Remark 6. Under Assumption 3, Lemma 6 gives the value of a policy as $V_\mu(\pi) = \sum_h \mu_h(a_h)$. The optimal policy $\pi^*(\mu)$ is unique and equals $(a^*_h)_{h \le H}$, and every branch has the stage observational mean. The layered causal MDP is therefore exactly H causal bandits of Definition 2, one per stage. Assumption 3(c) plays the role of Assumptions 1–2 at every stage. What couples the stages is the sampling protocol (Definition 9). Assumption 3(a) is a condition on $G^{(H)}$ alone and is checkable before any sample is drawn. It cannot be replaced by the weaker requirement that the law of Y_h under a joint intervention depend on a_h only, as the following example shows. Let $H = 2$, $V_1 = \{X_1, Y_1\}$ and $V_2 = \{X_2, Z_2, Y_2\}$ with edges $X_1 \to Y_1$, $X_1 \to Y_2$, $X_2 \to Y_2$. Let $Y_2 = f(X_1) + X_2 + \xi$ with $f(0) = 0$, $f(1) = f(2) = c > 0$ and $X_1 = 0$ under P. Let $A_1 = \{\mathrm{do}(X_1{=}1), \mathrm{do}(X_1{=}2)\}$ and $A_2 = \{\mathrm{do}(X_2{=}1), \mathrm{do}(X_2{=}2), \mathrm{do}(Z_2{=}0)\}$. Under every joint intervention the law of Y_2 depends on a_2 only, and Z_2 is a non-ancestor of Y_1 and Y_2. However, $\mathbb{E}[Y_2 \mid \mathrm{do}(a_1, Z_2{=}0)] = c + \mathbb{E}_P[X_2] \neq \mathbb{E}_P[Y_2] = \mathbb{E}_P[X_2]$, so the branch does not have the observational mean. The true instance is then outside $\mathcal{M}^{(H)}_{\mathrm{G}}$, defined below. A strategy that never samples branches is wrong with probability close to one when the branch is the best stage-2 arm. Appendix D treats the case in which (b) fails.

The stage-h inner infimum. Take $\mathcal{S}^{(H)}$, the product classes $\mathcal{M}^{(H)}_{\mathrm{G}}$, $\mathcal{M}^{(H)}_{\mathrm{blind}}$ and the stage-h times $T^*_{\mathcal{M},h}, T^*_{\mathrm{G},h}, T^*_{\mathrm{blind},h}$ as in Section 4. We abbreviate the last three to T^*_h when the class is clear. Under Assumption 3 the true mean vector belongs to $\mathcal{M}^{(H)}_{\mathrm{G}} \cap \mathcal{S}^{(H)}$ by Lemma 6(ii) and Assumption 3(c). For $v \in \mathbb{R}_+^{A_h}$ let

$$G_h(v) := \inf_{\lambda \in \mathrm{Alt}_{\mathcal{M}_h}(\mu_h)} \sum_{a \in A_h} v_a \, \frac{(\mu_h(a) - \lambda_a)^2}{2\sigma^2}, \tag{5}$$

so that G_h restricted to Σ_{A_h} is the inner infimum of (2) and $\sup_{u\in\Sigma_{A_h}} G_h(u) = 1/T^*_{\mathcal{M},h}$. For both classes the stage-h characteristic time $T^*_{\mathcal{M},h}$ is finite and positive. Indeed, $\mathrm{Alt}_{\mathcal{M}_h}(\mu_h) \neq \emptyset$ because $|A_h^{\mathrm{anc}}| \geq 2$, and the uniform allocation gives a positive inner infimum.

Conditional interventions and overloaded symbols (Definition 9). Selecting $a_{h,n}$ after observing $Y_{1,n}, \ldots, Y_{h-1,n}$ would be a conditional intervention, whose semantics Definition 8 does not specify. We do not consider such strategies. The strategy of Definition 10 selects its arms at episode boundaries. By Lemma 6(i) and Assumption 3(c) the true instance is $\mathbb{P}_\mu$ for its mean vector. The symbols $\mathrm{Alt}_{\mathcal{M}}$ and c_w are thus overloaded, as is $T^*_{\mathcal{M}}$ in Appendix D. The argument determines which version is meant. It is a single-stage vector, a layered vector $(\mu_h)_{h\leq H}$, or an MDP model.

Auxiliary lemmas for Theorem 2.

Lemma 7 (Single-stage flips are the cheapest alternatives)**.** *Under Definitions 8–9 and Assumption 3, for $\mathcal{M} \in \{\mathcal{M}_{\mathrm{G}}^{(H)}, \mathcal{M}_{\mathrm{blind}}^{(H)}\}$, $\mu \in \mathcal{M} \cap \mathcal{S}^{(H)}$ and every $w \in \Omega_{\mathrm{epi}} \cup \Omega_{\mathrm{gen}}$,*

$$\inf_{\mu'\in \mathrm{Alt}_{\mathcal{M}}(\mu)} c_w(\mu, \mu') \;=\; \min_{h\leq H} G_h(w_h).$$

Proof. Let $\mathcal{M} = \prod_h \mathcal{M}_h$. By Lemma 6(i) and Remark 6, for every $\mu' \in \mathcal{M} \cap \mathcal{S}^{(H)}$ the optimal policy is $\pi^*(\mu') = (a_h^*(\mu'_h))_h$, and c_w is a sum of per-stage blocks. ($\geq$) Let $\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)$. Since $\pi^*(\mu')$ differs from $\pi^*(\mu)$, there is a stage h with $a_h^*(\mu'_h) \neq a_h^*(\mu_h)$. As $\mu'_h \in \mathcal{M}_h \cap \mathcal{S}_{A_h}$, we have $\mu'_h \in \mathrm{Alt}_{\mathcal{M}_h}(\mu_h)$. So $c_w(\mu, \mu')$ is at least its stage-h block, which is at least $G_h(w_h) \geq \min_h G_h(w_h)$. ($\leq$) Fix h and $\lambda \in \mathrm{Alt}_{\mathcal{M}_h}(\mu_h)$, and let μ' agree with μ at every stage other than h and equal λ at stage h. Then $\mu' \in \mathcal{M} \cap \mathcal{S}^{(H)}$ and $\pi^*(\mu')$ differs from $\pi^*(\mu)$ at stage h, so $\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)$ and the infimum is at most $\sum_a w_{h,a}(\mu_h(a) - \lambda_a)^2/(2\sigma^2)$. Taking the infimum over λ and then the minimum over h gives the claim. Assumption 3(a) enters through Lemma 6. It gives the block structure of c_w, the separable optimal policy, and the membership of the branches in $\mathcal{M}_{\mathrm{G},h}$. Assumption 3(b) enters through the definition of the protocols. □

Lemma 8 (Homogeneity)**.** *For every h, $r \geq 0$ and $v \in \mathbb{R}_+^{A_h}$, $G_h(rv) = r\, G_h(v)$.*

Proof. For each fixed alternative λ the cost $\sum_a v_a(\mu_h(a) - \lambda_a)^2/(2\sigma^2)$ is linear in v. For $r > 0$ the infimum of r times a family of numbers is r times the infimum, and for $r = 0$ both sides are 0 because $\mathrm{Alt}_{\mathcal{M}_h}(\mu_h) \neq \emptyset$. □

Remark 7 (Structure of the episodic optimum)*.* By Lemma 7 the episodic program is $\sup_{w\in\Omega_{\mathrm{epi}}} \min_h G_h(w_h)$. The proof of Theorem 2(i) shows that its value is attained by any w with two properties. Its stage-h component maximises G_h at a bottleneck stage $h \in \arg\max_h T^*_{\mathcal{M},h}$, and it satisfies $G_h(w_h) \geq 1/\max_h T^*_{\mathcal{M},h}$ elsewhere. The equalisation property of Lemma 4(c) therefore holds only within stages. At the episodic optimum the rivals of the bottleneck stage sit at the rate $1/\max_h T^*_{\mathcal{M},h}$, while faster stages may be allocated in many ways. Positivity, Lemma 4(b), holds at every stage for the blind class. A zero-weight rival at any stage makes $\min_h G_h(w_h) = 0$.

The IATS strategy in full.

Definition 10 (The IATS strategy)**.** Intervention-Aware Track-and-Stop (IATS) runs, at every stage h, the Track-and-Stop strategy of Garivier & Kaufmann (2016) on the ancestor bandit $\nu_h := (\mu_h(a))_{a\in A_h^{\mathrm{anc}}}$ with $K_h := |A_h^{\mathrm{anc}}|$ arms. It uses the threshold (4) at the per-stage confidence $\delta' := \delta/H$. In full, the threshold is $\beta_h(t,\delta') := \beta_{\mathrm{an}}(t,\delta'; K_h) := \log\big((K_h - 1)\zeta(2)t^3/\delta'\big)$ with $\zeta(2) = \sum_{s\geq 1} s^{-2} = \pi^2/6$, so that $\beta_h(t, \delta/H) = \log\big(H(K_h - 1)\zeta(2)t^3/\delta\big)$. The error bound of Lemma 10 below also involves $\zeta(3) := \sum_{s\geq 1} s^{-3} \approx 1.202$. The strategy consists of the following rules.

- *Initialisation.* The first $K_0 \geq K_h$ pulls of stage h cycle through the arms of A_h^{anc} in a fixed order, so that every arm is pulled at least once. Pull $t \leq K_0$ takes arm $((t-1) \bmod K_h)+1$. Here $K_0 = K_h$ in the generative variant and $K_0 = \bar{K} := \max_h K_h$ in the episodic variant, as described below.

- *Sampling rule* (D-tracking with forced exploration, see Garivier & Kaufmann, 2016, Section 3). After $t \geq K_0$ pulls at stage h, let $N_{h,a}(t)$ be the counts, collected in the vector $N_{h,\cdot}(t)$, and $\hat{\mu}_{h,a}(t)$ the empirical means, collected in the vector $\hat{\mu}_h(t) := (\hat{\mu}_{h,a}(t))_{a \in A_h^{\mathrm{anc}}}$. Let $\mathcal{L}_t := \{a \in A_h^{\mathrm{anc}} : N_{h,a}(t) < \sqrt{t} - K_h/2\}$ be the set of under-sampled arms, written U_t by Garivier & Kaufmann (2016). If $\mathcal{L}_t \neq \emptyset$, pull $\arg\min_{a \in \mathcal{L}_t} N_{h,a}(t)$. Otherwise pull $\arg\max_{a \in A_h^{\mathrm{anc}}} \big(t\, w_a^*(\hat{\mu}_h(t)) - N_{h,a}(t)\big)$. Here $w^*(\nu)$ is the maximiser of the program of Lemma 2 for the K_h-arm bandit ν. This maximiser is unique and depends continuously on $\nu \in \mathcal{S}_{A_h^{\mathrm{anc}}}$ (Garivier & Kaufmann, 2016, Theorem 5 and Proposition 6). When $\hat{\mu}_h(t)$ has a tie for the best arm, an event of probability zero, a fixed allocation is used. All ties in the $\arg\min$ and $\arg\max$ are broken by a fixed order of the arms. The sampling rule is defined for every $t \geq 1$, and the stopping rule below only fixes the time τ_h at which the recommendation is read off.

- *Stopping rule* (Chernoff, see Garivier & Kaufmann, 2016, eq. (8)). Let $\hat{a}_h(t) := \arg\max_a \hat{\mu}_{h,a}(t)$ and

$$Z_h(t) := \min_{b \in A_h^{\mathrm{anc}} \setminus \{\hat{a}_h(t)\}} \frac{\big(\hat{\mu}_{h,\hat{a}_h(t)}(t) - \hat{\mu}_{h,b}(t)\big)^2}{2\sigma^2\big(1/N_{h,\hat{a}_h(t)}(t) + 1/N_{h,b}(t)\big)}.$$

  Stage h stops at $\tau_h := \inf\{t \geq K_0 :\ Z_h(t) \geq \beta_h(t, \delta/H)\}$.

- *Recommendation.* $\hat{a}_h := \hat{a}_h(\tau_h)$, and the recommended policy is $\hat{\pi} := (\hat{a}_h)_{h \leq H}$.

Generative variant. The stages are run one after another, and the total number of samples is $\tau = \sum_h \tau_h$. *Episodic variant.* All stages run in parallel on the shared episode counter. The protocol pulls every stage once in every episode. The initialisation of the stage with the most arms therefore occupies $\bar{K} = \max_h K_h$ episodes, and every stage receives $K_0 = \bar{K}$ initialisation pulls. When all K_h are equal, as in every run of Section 6, this is the single-stage initialisation at every stage. In episode $n > \bar{K}$, every stage with $\tau_h > n - 1$ makes its n-th pull according to its sampling rule. Every stage with $\tau_h \leq n - 1$ is *frozen*. A frozen stage pulls its recommendation $\hat{a}_h$, and its further observations are not used. The arms of episode n are thus selected from the data of episodes $1, \ldots, n-1$ only. The episode count is $N = \max_h \tau_h$. In both variants the stage-h run uses stage-h observations only, so it is a single-stage Track-and-Stop on the K_h-arm ancestor bandit with initialisation length K_0. The *blind* version of the strategy runs the same rules on A_h instead of A_h^{anc}, with $K_h = |A_h|$.

The threshold (4) is the one analysed in Lemma 10 and Theorem 3 and the one used by the released code (Section 5). Algorithm 1 in Section 5 gives the episodic variant in pseudocode. Any threshold written as $\log(r_h(t)/\delta')$ can be substituted in the rate statement of Lemma 9 below, and the threshold (4) has $r_h(t) = (K_h - 1)\zeta(2)t^3$.

Two single-stage tools.

Lemma 9 (Tail bound for Track-and-Stop, from the proof of Theorem 14 of Garivier & Kaufmann, 2016)**.** *Let $\nu \in \mathbb{R}^K$, $K \geq 2$, be a Gaussian bandit with known variance σ^2 and a unique best arm, and let $T^*(\nu) := T^*_{\mathbb{R}^K}(\nu)$. Let $\mathbb{P}_\nu$ and $\mathbb{E}_\nu$ denote the law and expectation of the observations when the arm means are ν. These are the $\mathbb{P}_\mu$ and $\mathbb{E}_\mu$ of Definition 5 with $\mu = \nu$. Let $\alpha \geq 0$ and $C_r > 0$, and let $r : \mathbb{N} \to (0, \infty)$ satisfy $r(t) \leq C_r t^\alpha$ for all $t \geq 1$. Consider the single-stage Track-and-Stop strategy of Definition 10 on ν with stopping time τ_δ. This strategy runs an initialisation of a fixed length $K_0 \geq K$ cycling through the arms, then D-tracking with forced exploration for all $t \geq 1$. It stops by the Chernoff rule with the threshold $\beta(t, \delta) = \log(r(t)/\delta)$ in place of* (4)*. There is $\varepsilon_0(\nu) > 0$ with the following property. For every $\varepsilon \in (0, \varepsilon_0(\nu))$ there are a constant $D_\varepsilon < \infty$ that does not depend on δ and a number $\kappa(\varepsilon)$ with $\kappa(\varepsilon) \to 1$ as $\varepsilon \to 0$, and for every $\delta \in (0, 1)$ a deterministic integer $T_\varepsilon(\delta)$, such that*

$$\limsup_{\delta \to 0} \frac{T_\varepsilon(\delta)}{\log(1/\delta)} \ \leq\ \kappa(\varepsilon)\, T^*(\nu) \qquad \text{and} \qquad \sup_{\delta \in (0,1)} \mathbb{E}_\nu\big[(\tau_\delta - T_\varepsilon(\delta))^+\big] \ \leq\ D_\varepsilon.$$

In particular $\tau_\delta < \infty$ almost surely for every $\delta \in (0, 1)$, and $\limsup_{\delta \to 0} \mathbb{E}_\nu[\tau_\delta]/\log(1/\delta) \leq T^*(\nu)$.

Proof. Appendix E.7. Lemma 9 is not a numbered statement of Garivier & Kaufmann (2016). Appendix E.7 quotes two facts established in the proof of their Theorem 14 in their Appendix D.1. These are their Lemma 19 and the lower bound on the Chernoff statistic that follows their Lemma 20, and neither involves the threshold. It then bounds directly the deterministic time after which the threshold is crossed on the event of their Lemma 19. Their Theorem 14 assumes $\alpha \in [1, e/2]$ and states the rate $\alpha T^*(\nu)$. Both the restriction and the factor come from their technical Lemma 18, which the direct bound replaces (Remark 8). The threshold (4) satisfies the hypothesis with $r(t) = (K-1)\zeta(2)t^3$, $\alpha = 3$ and $C_r = (K-1)\zeta(2)$. □

Lemma 10 (Error bound for the Chernoff stopping rule under β_{an}, any sampling rule)**.** *Let $\nu \in \mathbb{R}^K$, $K \geq 2$, be a Gaussian bandit with known variance σ^2 and a unique best arm a^*, and let $\delta \in (0,1)$. Consider any sampling rule, as defined in Definition 5, under which every arm is pulled at least once during the first K_0 rounds. Let $Z(t)$ be the Chernoff statistic and $\hat{a}(t)$ the empirical best arm, the single-stage versions of $Z_h(t)$ and $\hat{a}_h(t)$ of Definition 10 with the stage index dropped. Let $\tau := \inf\{t \geq K_0 : \ Z(t) \geq \beta_{\mathrm{an}}(t, \delta; K)\}$ be the stopping time, with $\beta_{\mathrm{an}}(t, \delta; K) = \log\big((K-1)\zeta(2)t^3/\delta\big)$, and let $\hat{a} := \arg\max_a \hat{\mu}_a(\tau)$ be the recommendation on $\{\tau < \infty\}$. Then*

$$\mathbb{P}_\nu(\tau < \infty,\ \hat{a} \neq a^*) \ \leq \ \Big(1 - \frac{\zeta(3)}{\zeta(2)}\Big)\,\delta \ < \ \delta.$$

Proof. Appendix E.8. The bound holds for every sampling rule, in particular for D-tracking with a cached allocation (Section 5). The union over the pairs of pull counts is what the factor $\zeta(2)t^3$ in (4) pays for. Proposition 12 of Garivier & Kaufmann (2016) gives, for exponential families, the alternative threshold $\log(C(\alpha, K)t^\alpha/\delta)$ with $\alpha > 1$ and a constant whose value is not made explicit. Lemma 10 is the explicit Gaussian counterpart used here. □

Remarks on Theorem 3.

Remark 8 (On the threshold, the exponent α and Lemma 9). The threshold (4) has $r(t) = (K-1)\zeta(2)t^3$, so Lemma 9 applies to it with $\alpha = 3$ and $C_r = (K-1)\zeta(2)$. The rates in Theorem 3 do not depend on α. Theorem 14 of Garivier & Kaufmann (2016) assumes $\alpha \in [1, e/2]$ and states, for the same strategy, the bound $\alpha T^*(\nu)$. Since $3 > e/2$, that theorem does not cover (4) directly. In their proof, in their Appendix D.1, both the restriction and the factor enter only through their technical Lemma 18. That lemma bounds the solutions x of inequalities $C_1 x \geq \log(C_2 x^\alpha)$ with generic constants $C_1, C_2 > 0$. They use it to bound the deterministic time $T_0(\delta)$ after which the threshold is crossed, step (F3) of Appendix E.7. Their Lemmas 19–20, which Lemma 9 quotes, concern the sampling rule only and do not involve the threshold. Lemma 9 bounds $T_0(\delta)$ directly in step (F3) of Appendix E.7 and therefore obtains $T^*(\nu)$ for every $\alpha \geq 0$. The δ-correctness of (4) is Lemma 10. The threshold of Proposition 12 of Garivier & Kaufmann (2016), $\log(C(\alpha, K)t^\alpha/\delta)$ with $\alpha > 1$, is an alternative. Its constant is not explicit, and we do not use it in this paper.

Under (H) at every stage, Theorems 2 and 3 identify $\max_h T^*_{\mathrm{G},h}$ and $\sum_h T^*_{\mathrm{G},h}$ as the optimal leading-order constants. Every δ-correct family has $\liminf_{\delta\to 0} \mathbb{E}[N]/\log(1/\delta)$ at least the constant, and IATS has $\limsup$ at most the constant. Part (iii) of Theorem 3 is the only place where the tail bound of Lemma 9, rather than its consequence for $\mathbb{E}_\nu[\tau_\delta]$, is used. From $N = \max_h \tau_h \leq \sum_h \tau_h$ alone one obtains only the bound $\sum_h T^*_{\mathrm{anc}}(\mu_h)$.

Remark 9 (The separation under the two protocols). Theorems 2 and 3 hold for both classes, so Theorem 1 lifts. Under the generative protocol the difference $T^*_{\mathrm{gen,blind}} - T^*_{\mathrm{gen,G}} = \sum_h (T^*_{\mathrm{blind},h} - T^*_{\mathrm{G},h})$ is a sum of per-stage differences, each of which is at least $(2\sigma^2/\Delta^2_{0,h})k_h - T^*_{\mathrm{G},h}$ by Theorem 1. Under the episodic protocol the ratio is $\max_h T^*_{\mathrm{blind},h}/\max_h T^*_{\mathrm{G},h}$, a ratio of maxima. It is bounded away from one only if a bottleneck stage of the blind learner carries branches. For identical stages it equals the single-stage ratio for every H, whereas for heterogeneous stages it can be close to one. If the bottleneck stage is a near-tie ancestor stage, its branches add little to an already large $T^*_{\mathrm{blind},h}$. The ratio of maxima is then close to one although every stage carries branches. The horizon experiment E4 gives an example computed from Definition 6. Section 6 summarises it and the full report below states it.

The following details accompany Section 4. The model is H causal bandits of Definition 2 coupled only through the sampling protocol. Under Assumption 3 the true mean vector lies in $\mathcal{M}^{(H)}_{\mathrm{G}} \cap \mathcal{S}^{(H)}$ (Lemma 6(ii) and Assumption 3(c)). Condition (a) of Assumption 3 cannot be weakened (Remark 6), and state-dependent transitions are the subject of Appendix D. The threshold (4) is the one

analysed and the default of the released code, whereas the calibrated threshold of one comparison run is not covered (Section 5). Remark 8 explains why (4), with $\alpha = 3$, needs the direct proof of Lemma 9. Under (H) at every stage the constants of Theorem 3 are optimal, and Theorem 1 lifts to both protocols (Remark 9).

C.9 Proof sketch of Theorem 2

Proof sketch. The cheapest alternative flips a single stage, by Lemma 7 in Appendix C. Both programs therefore read $\sup_w \min_h G_h(w_h)$, where G_h is the stage-h inner infimum (5) at the unnormalised weights w_h. Over Ω_{epi} the stages decouple, which gives (i). Over Ω_{gen}, write $w_h = \theta_h u_h$ with $\theta \in \Sigma_H$ and $u_h \in \Sigma_{A_h}$. Lemma 8 gives $G_h(\theta_h u_h) = \theta_h G_h(u_h)$, so the program becomes $\sup_\theta \min_h \theta_h / T^*_{\mathcal{M},h} = 1/\sum_h T^*_{\mathcal{M},h}$, which is (ii). Parts (iii) and (iv) follow from the change of measure of Kaufmann et al. (2016) for the bandit with arms (h, a). The full proof is in Appendix E.6.

C.10 Released implementation (Section 5)

The released code implements Definition 10 and Algorithm 1. The single-stage routine is `src/iats/bandit/trackandstop.py` and the episodic routine is `src/iats/mdp/layered.py`, at commit `4965fca`, recorded in `results/paper/summary.json`. The threshold β_{an} of (4) is the default (`threshold="analysed"`). The episodic routine uses the per-stage confidence δ/H. It freezes each stage at its own first crossing and keeps pulling the frozen stage's recommendation. It gives every stage the $\bar{K} = \max_h K_h$ initialisation episodes that the protocol imposes, and it recommends the empirical best arm at freezing. Four points remain to be disclosed, together with a note on the allocation solver after item (i).

(i) Allocation cache. The runs of Section 6 recompute the plug-in allocation $w^*(\hat{\mu}_h)$ every 20 rounds in E2 and E3 and every 50 episodes in E4, and track the cached allocation in between. The periods are `REALLOC` and `E4_REALLOC` in `experiments/paper/run.py`, stored under `config.realloc_every` and `config.realloc_every_e4` in the summary. The library default, `realloc_every=1`, recomputes it at every round. This is the D-tracking rule of Definition 10 and Lemma 9, and it would make the full 300-seed run take many hours. The cache affects the guarantees differently. The error bound is unaffected, since Lemma 10 holds for every sampling rule. The probability that a stage stops by certificate with a wrong arm is therefore at most δ/H for the released runs. A run that reached the hard cap of item (ii) would return an uncertified recommendation, an error term not covered by Lemma 10. No run reached the cap. Almost-sure termination is also unaffected, because the cache leaves the forced-exploration branch unchanged. The first claim of Lemma 17 of Garivier & Kaufmann (2016) uses only that branch and gives $N_{h,a}(t) \geq \sqrt{t} - K_h/2 - 1 \to \infty$ for every arm. By the strong law of large numbers, $\hat{\mu}_h(t) \to \nu_h$ almost surely. Eventually $\hat{a}_h(t) = a^*_h$ and $Z_h(t) \geq \Delta^2_{\min,h}(\sqrt{t} - K_h/2 - 1)/(16\sigma^2)$, where $\Delta_{\min,h}$ is the smallest gap of ν_h. The right-hand side eventually exceeds $\beta_h(t, \delta/H) = O(\log t)$, so $\tau_h < \infty$ almost surely. The rate is the statement of Lemma 9 for the per-round rule and is not claimed for the cached rule. For the cached rule the rate is an empirical observation. E2 reports it through the mean stopping time over seeds, $\bar{\tau}$, normalised by the threshold actually crossed, $\bar{\tau}/(T^*\, \beta_{\mathrm{an}}(\bar{\tau}, \delta; K))$, under the key `rate_over_tstar_beta`. Here and in Section 6 a bare T^* denotes the characteristic time of the learner under discussion, so T^*_{G} for IATS and T^*_{blind} for the blind learner. The ratio is 0.978, 0.978, 0.983, 0.983 for IATS and 1.034, 1.029, 1.029, 1.026 for the blind learner at $\delta = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$.

Allocation solver. The solver of the plug-in allocation $w^*(\hat{\mu}_h)$ is a nested line search with cost linear in K_h. It is implemented by `optimal_allocation` in `src/iats/bandit/chartime.py`, which solves the program of Lemma 2(a) for the K_h-arm ancestor bandit without an external optimiser. Let c denote a candidate value of the program and $d_a := \Delta^2_a/(2\sigma^2)$ the rival coefficients. The rival term $h_a(w)$ of (3) reaches c if and only if $w_a \geq 1/(d_a/c - 1/w_{a^*})$. So c is attainable if and only if the convex function $w_{a^*} + \sum_{a \neq a^*} 1/(d_a/c - 1/w_{a^*})$ of $w_{a^*} \in (c/\min_a d_a, 1)$ has minimum at most one. That minimum increases with c. The code runs an outer bisection on c with a fixed budget of 45 steps, which gives a relative precision of about 3×10^{-14} on the value. At each candidate c it runs an inner golden-section minimisation over w_{a^*} with 60 steps. It returns the

weights w_{a^*} and $1/(d_a/c - 1/w_{a^*})$ at the final c, renormalised to the simplex. The two-arm case has the closed form $T^* = 8\sigma^2/\Delta^2$ for the gap Δ between the two means, recorded in the module documentation as a hand check. There is no closed form beyond two arms. The budget of 45 and 60 steps, which locates the minimiser to about 3×10^{-13}, is set at what double precision can use rather than beyond it. Each evaluation of the function is a sum over the $K_h - 1$ rivals, so one solve costs about 45×60 such sums. This cost is linear in K_h and does not depend on t or δ. When the empirical best arm is not unique the solver returns the uniform allocation, and with fewer than two arms it returns the single arm. The solve is the hot spot of Track-and-Stop, with one solve per allocation update. The cache of item (i) keeps it off the per-round path. Both routines recompute the allocation only when the round index is a multiple of `realloc_every` and track the cached vector in between.

(ii) Hard caps. Both routines stop without a certificate at a hard cap and then return the empirical best arms. The cap is 5×10^5 pulls for a single-stage run and 2×10^5 episodes for the episodic run. Capped runs are counted in the summary under `capped_runs`. None occurred in E2 under either threshold, in E3 or in E4. The per-seed stopping times are stored in `results/paper/arrays.npz`. The largest stopping time of any single-stage run is 6961 pulls, for the blind learner in E3 at $k = 32$. The largest in E2 is 3410 pulls, for the blind learner at $\delta = 10^{-4}$. The largest episode count in E4 is 1980, at $H = 8$. All are far below the caps.

(iii) E5. The coupled-MDP strategy of Appendix D.3 (`src/iats/mdp/general.py`) is not the strategy of Definition 10. It uses the calibrated threshold β_{cal} of item (iv) plus $\log|S_{\mathrm{dec}}|$, where S_{dec} is the set of decision states of the MDP of Definition 11. The study is illustrative and non-confirmatory.

(iv) Calibrated threshold, kept for comparison only. Experiments E2, E3 and E4 use β_{an}, whose δ-correctness is Lemma 10. For comparison only, the code retains the threshold $\beta_{\mathrm{cal}}(t, \delta; K) := \log\big((1 + \log t)(K - 1)/\delta\big)$ as `threshold="calibrated"`. This is the experimental threshold of Garivier & Kaufmann (2016, Section 6) up to the factor $K-1$, and its δ-correctness is not proven. In full, that threshold is

$$\beta_{\mathrm{cal}}(t, \delta; K) := \log\Big(\frac{(1 + \log t)(K - 1)}{\delta}\Big).$$

Up to the arm-count factor $K - 1$, this is the threshold $\log((\log t + 1)/\delta)$ that Garivier & Kaufmann (2016, Section 6) use in their experiments and describe as not yet allowed by theory. Its δ-correctness is not proven, since it is covered neither by Lemma 10 nor by Proposition 12 of Garivier & Kaufmann (2016), whose threshold grows like $\alpha \log t$ with $\alpha > 1$. It satisfies the hypothesis of Lemma 9 with $\alpha = 1$ and $C_r = K - 1$, since $1 + \log t \le t$. So the strategy of Definition 10 run with β_{cal} stops almost surely and has the rate T^*. A rate is not a correctness guarantee. In the released run, E2 is repeated with β_{cal} on the same seeds, under the summary key `e2_calibrated`. This repetition is not part of the pre-specified checks. No error occurred in any of the eight cells. The ratio $\bar\tau/\log(1/\delta)$ was 1.48, 1.28, 1.21, 1.17 times T^*_{G} for IATS at $\delta = 10^{-1}, \dots, 10^{-4}$, against 9.54, 5.33, 3.94, 3.23 under the analysed threshold. For the blind learner it was 3.23, 2.11, 1.73, 1.54 times T^*_{blind}, against 12.33, 6.70, 4.85, 3.91 under β_{an}. The normalised quantity $\bar\tau/(T^*\beta_{\mathrm{cal}}(\bar\tau, \delta; K))$ was 0.84, 0.91, 0.95, 0.97 for IATS and 1.15, 1.10, 1.07, 1.05 for the blind learner. Forced exploration was 14.0%, 11.6%, 9.9%, 8.5% of the blind learner's budget under β_{cal}, against 4.4%, 4.0%, 3.7%, 3.5% under β_{an}. It was 0% for IATS under both. At the tested δ, the analysed threshold thus costs a factor between 2.8 and 6.5 in the mean stopping time of IATS, relative to a threshold without a proof. For the blind learner the factor is between 2.5 and 3.8.

The following details accompany Section 5. Algorithm 1 is the episodic variant of the strategy of Definition 10, analysed in Theorem 3. Each stage drops its branch arms B_h and runs the single-stage Track-and-Stop of Garivier & Kaufmann (2016) on the remaining K_h arms with the episode index as its clock. A stopped stage keeps pulling its recommendation so that the one-pull-per-stage rule of the episodic protocol holds. Both variants of Algorithm 1 return the empirical best arms without a certificate at a hard cap of 5×10^5 pulls per single-stage run or 2×10^5 episodes. No released run reached the cap (item (ii)). In the runs of Section 6 the released code recomputes the plug-in allocation $w^*(\hat\mu_h)$ every 20 rounds in E2 and E3 and every 50 episodes in E4. The cache leaves Lemma 10, which holds for every sampling rule, and almost-sure termination unchanged. The rate of Lemma 9 is claimed for the per-round rule $r = 1$ only (item (i)). The calibrated threshold β_{cal} meets the hypothesis of Lemma 9, so it stops almost surely at the rate T^*, which is no correctness

guarantee. It is used for comparison only, in a repetition of E2 on the same seeds outside the pre-specified checks (item (iv)). It is also used in the illustrative study E5 of Appendix D.3, with $\log|S_{\mathrm{dec}}|$ added for the decision states S_{dec} of the MDP.

C.11 Choice of threshold and composition of the stage stopping times (Section 5)

Lemma 10 bounds the error probability of the Chernoff rule under $\beta_{\mathrm{an}}(t,\delta;K)$ by $(1-\zeta(3)/\zeta(2))\delta < \delta$ for any sampling rule that pulls every arm once during initialisation. Here $\zeta(3) := \sum_{s\geq 1} s^{-3} \approx 1.202$. Lemma 9 says that Track-and-Stop with the analysed threshold stops almost surely at the rate of the characteristic time $T^*(\nu) := T^*_{\mathbb{R}^K}(\nu)$. Its stopping time has an exponential tail beyond a deterministic time of that order. The threshold of the experiments of Garivier & Kaufmann (2016) has no proof of δ-correctness, so we analyse the polynomial threshold (4). For this threshold Lemma 9 gives the rate $T^*(\nu)$ without the factor α of their Theorem 14 (Remark 8). Lemma 10 gives δ-correctness under any sampling rule at the per-stage confidence δ/H. Under the episodic protocol with freezing, the stage stopping times compose as their maximum, not their sum. This makes the episodic rate $\max_h T^*_{\mathrm{G},h}$ of Theorem 2(i) attainable under (H). With reallocation period $r = 1$, Algorithm 1 under either protocol is the strategy of Definition 10. Theorem 3 shows that this strategy is δ-correct and attains the rates of Theorem 2. Under the generative protocol each stage run is a function of its own reward streams and internal randomisation (Definition 10, Lemma 6(i)). Consider any interleaving of the stage samplers, adaptive or not, that samples only unstopped stages and eventually samples each of them. Such an interleaving leaves the law of every τ_h and of $\tau = \sum_h \tau_h$ unchanged, so the sequential order is a presentation choice. By Theorem 2(iv) no cross-stage rule improves the leading term $\sum_h T^*_{\mathrm{G},h}$. A rule sharing the confidence unevenly across stages is not covered by Theorem 3, which uses δ/H at every stage.

C.12 Proof sketch of Theorem 3

The proof rests on Lemmas 10 and 9. Both had to be added to Track-and-Stop on MDPs (Al Marjani & Proutiere, 2021) and to the pruning of the branch arms. *Proof sketch.* The stage-h run is the single-stage strategy on ν_h, whose unique best arm is a^*_h, at confidence δ/H on independent draws (Lemma 6(i)). Lemma 10, almost-sure stopping from Lemma 9 and a union bound over the H stages give (i). Lemma 9 and $\log(H/\delta)/\log(1/\delta) \to 1$ give $\limsup_{\delta\to 0} \mathbb{E}_\mu[\tau_h]/\log(1/\delta) \leq T^*_{\mathrm{anc}}(\mu_h)$, and summing gives (ii). For (iii), $N = \max_h \tau_h$ is at most the largest deterministic time of Lemma 9 plus the sum of the stage excesses, whose expectations are bounded by the tail constants. The full proof is in Appendix E.9.

C.13 Finite-δ behaviour (Section 5)

The rate-stability check of the correctness and rate experiment E2 (Section 6) rests on the following observation. Theorems 2 and 3 concern the coefficient of $\log(1/\delta)$ as $\delta \to 0$ and make no statement about finite δ. The threshold (4) is $\beta_{\mathrm{an}}(t,\delta;K) = \log(1/\delta) + 3\log t + \log((K-1)\zeta(2))$. At the tested δ the term $3\log t$ is the larger one, with $3\log t \approx 20.7$ at $t = 10^3$ against $\log(1/\delta)$ between 2.3 and 9.2. Accordingly, the realized $\bar\tau/\log(1/\delta)$ of IATS on the instance of Remark 5 with $k = 8$ was 9.54, 5.33, 3.94 and 3.23 times T^*_{G} at $\delta = 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$ (experiment E2, summarised in Section 6 and reported in full below). That of the blind learner was 12.33, 6.70, 4.85 and 3.91 times T^*_{blind}. Both decrease as δ decreases. In contrast, $\bar\tau/(T^*\beta_{\mathrm{an}}(\bar\tau,\delta;K))$ stays within 4% of one for both learners, from 0.978 to 0.983 and from 1.026 to 1.034. At these δ the realized stopping time is therefore described by $T^*\beta_{\mathrm{an}}(\tau,\delta;K)$ rather than by $T^*\log(1/\delta)$. The approach of $\bar\tau/(T^*\log(1/\delta))$ to one is logarithmically slow in $1/\delta$ under any threshold with a $\log t$ term, and slower under (4) than under β_{cal} (item (iv)). Forced-exploration pulls were 0% of the budget for IATS and between 3.5% and 4.4% for the blind learner. At the tested δ, β_{an} therefore costs a factor between 2.8 and 6.5 in the mean stopping time over seeds $\bar\tau$ for IATS. For the blind learner the factor is between 2.5 and 3.8 (item (iv)).

C.14 Full report of experiments E1 to E4, with Figures 3 and 4

This section reports the confirmatory family F-IATS in full. Section 6 summarises it and contains Figure 2. The experiment codes are those of the run protocol and of the result files. E1 computes exact characteristic times, E2 tests correctness and the rate, E3 measures the realized speed-up, E4 varies the horizon, and E5 is the coupled-case study of Appendix D.3.

Throughout this section, as in Section 5, a bare T^* denotes the characteristic time of the learner under discussion. $\bar{\tau}$ is the mean stopping time over seeds. When both learners appear, $\bar{\tau}_{\mathrm{IA}}$ refers to IATS and $\bar{\tau}_{\mathrm{blind}}$ to the blind learner. $\bar{N}_H$ is the mean episode count over seeds of the episodic run with horizon H. All experiments run on a single CPU. Acceptance criteria, hypotheses and statistics were pre-specified in the run protocol (Appendix F.1, Table 5). The confirmatory family F-IATS consists of E1–E4. It reports significance by Holm–Bonferroni at level 0.05. It compares learners by an unpaired ratio of means with an unpaired bootstrap confidence interval. The two learners are run separately on a common seed sequence, $i = 0, \ldots, 299$, and analysed as unpaired samples. It reports an exact-binomial Clopper–Pearson upper bound on the error rate. With 300 seeds per cell, a $0/300$ cell gives 9.9×10^{-3}. One pre-specified check, the rate-stability check of E2, was not met. It is reported as such below, together with the reason. It fails by construction under the threshold β_{an}. The experiments compare the algorithms with the exact characteristic times and the composition law of Sections 2–4, and the two learners with each other. They do not compare with heuristic baselines. Every number below is produced by the released code (Appendix F), except where stated.

E1 (exact characteristic times). The instance has $\mu|_{A^{\mathrm{anc}}} = (1.0, 0.5)$, $\mu_0 = 0.2$, $\sigma = 1$ and $k \in \{0, 1, 2, 4, 8, 16, 32, 64\}$ branches. We compute T^*_{G} and T^*_{blind} from Definition 6, through Lemmas 2(a) and 3, by a bisection on the value of the program. $T^*_{\mathrm{G}} = 32$ for every k, and $T^*_{\mathrm{blind}} =$ 32.0, 35.88, 39.74, 47.43, 62.65, 92.59, 150.96, 264.16, within the bounds $3.125\,k \le T^*_{\mathrm{blind}} \le 16\,(k+2)$ of Theorem 1(i). The ratio rises from 1.12 at $k = 1$ to 8.26 at $k = 64$ (Figure 2). All four pre-specified checks are met. They are a flat T^*_{G}, the lower bound, a ratio at $k = 64$ more than three times the ratio at $k = 1$, and a positive slope.

E2 (δ-correctness and rate). With $k = 8$ branches for the blind learner, we ran the released implementation of Definition 10 with the threshold β_{an} over 300 seeds at each $\delta \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}\}$. The implementation uses the allocation cache of Section 5, item (i). Theorem 3 bounds the error probability by δ and predicts $\bar{\tau}/(T^* \log(1/\delta)) \to 1$. No error occurred in any of the eight cells. The Clopper–Pearson bound 9.9×10^{-3} certifies the error level at $\delta \ge 10^{-2}$, and for the two smaller δ the pre-specified check is "no error observed". The mean stopping time relative to $T^* \log(1/\delta)$ was 9.54, 5.33, 3.94, 3.23 for IATS and 12.33, 6.70, 4.85, 3.91 for the blind learner. The pre-specified rate-stability check requires the two smallest-δ values for IATS to differ by less than 0.1. They differ by 0.71, so the check is not met. The failure is a property of the threshold, not of the sampling rule. The threshold is $\beta_{\mathrm{an}}(t, \delta; K) = \log(1/\delta) + 3 \log t + \log((K-1)\zeta(2))$. At the tested δ the ratio $\bar{\tau}/(T^* \log(1/\delta))$ therefore tracks $\beta_{\mathrm{an}}(\bar{\tau}, \delta; K)/\log(1/\delta)$. This quantity decreases from about 9.8 to 3.3 over the tested range. It is flat only in the limit $\delta \to 0$ of Theorem 3. The check predates the adoption of β_{an} and was kept unchanged. On the same seeds under the calibrated threshold β_{cal} of Section 5, item (iv), the two values differ by 0.04 and the check is met. This locates the failure in the $3 \log t$ term of β_{an}. The mean stopping time normalised by the threshold actually crossed, $\bar{\tau}/(T^*\, \beta_{\mathrm{an}}(\bar{\tau}, \delta; K))$, is 0.978, 0.978, 0.983, 0.983 for IATS and 1.034, 1.029, 1.029, 1.026 for the blind learner. It is reported as a diagnostic and is not one of the pre-specified checks. Forced-exploration pulls were 4.4%, 4.0%, 3.7% and 3.5% of the blind learner's budget and 0% of IATS's (Figure 3). The same cells under the calibrated threshold β_{cal} are reported in Section 5, item (iv).

E3 (realized speed-up). We use $\delta = 0.01$, 300 seeds and $k \in \{0, 2, 4, 8, 16, 32\}$ branches for the blind learner. The realized speed-up $\bar{\tau}_{\mathrm{blind}}/\bar{\tau}_{\mathrm{IA}}$, a ratio of means, is 1.00, 1.42, 1.75, 2.46, 3.84 and 6.77. The bootstrap 95% intervals are $[0.95, 1.06]$, $[1.36, 1.49]$, $[1.67, 1.83]$, $[2.36, 2.57]$, $[3.69, 4.00]$ and $[6.51, 7.04]$. The exact leading-order ratios are 1.00, 1.24, 1.48, 1.96, 2.89 and 4.72 (Figure 2). The interval excludes 1 for every $k \ge 2$. Holm–Bonferroni rejects the five pre-specified hypotheses H1–H5, with all p-values numerically 0. These are one-sided Mann–Whitney tests of a speed-up at $k = 2, 8, 16, 32$ and of growth of the blind stopping time from $k = 2$ to $k = 32$. A permuted-arm control, with the two ancestor arms in reversed order, gives a ratio of 1.00 and a p-value of 0.996. The pre-specified envelope check requires the realized speed-up to lie within $[0.7, 1.6]$ times the leading-order ratio at $k = 2, 8, 16, 32$. It is met, since realized/theory is 1.15,

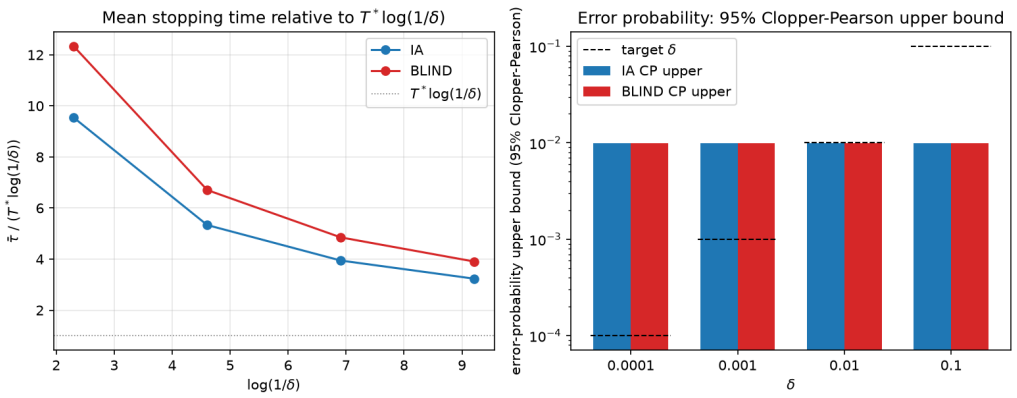


Figure 3: E2. The left panel shows the mean stopping time relative to $T^* \log(1/\delta)$ as a function of $\log(1/\delta)$ under the threshold β_{an}. For IATS it falls from 9.5 at $\delta = 0.1$ to 3.2 at $\delta = 10^{-4}$. The dotted line is the leading-order value 1. The decrease is due to the term $3 \log t$ of β_{an}. Normalised by the threshold actually crossed, the mean stopping time is within 4% of T^* at every δ (E2). The right panel shows the Clopper–Pearson 95% upper bound on the error probability against the target δ (dashed). The bound certifies the error level where the seed count allows ($\delta \geq 10^{-2}$).

1.26, 1.33 and 1.43, and 1.18 at $k = 4$ (Table 5). The realized ratio exceeds the leading-order one because the comparison is between a finite-δ quantity and a leading-order one. At $\delta = 0.01$ the blind learner's $\bar{\tau}/(T^* \log(1/\delta))$ is 6.70 while IATS's is 5.33 (E2). Three effects hit the blind learner harder. They are the term $3 \log t$ of the threshold, evaluated at its larger stopping time, the arm-count term $\log(K-1)$, and the forced exploration of a wider arm set. All three vanish at the leading order but not at $\delta = 0.01$.

E4 (horizon). The stages are identical, with ancestor means $(1.0, 0.5)$ plus 6 branches each. The exact episodic ratio $\max_h T^*_{\text{blind},h} / \max_h T^*_{\text{G},h} = 55.06/32 = 1.72$ is the same for $H = 1, 2, 4, 8, 16$. The exact generative difference $\sum_h (T^*_{\text{blind},h} - T^*_{\text{G},h}) = 23.06\, H$ is linear in H with $R^2 = 1.0$. Both agree with Theorem 2. With heterogeneous stages the episodic ratio need not be bounded away from one (Remark 9). Take 6 branches at $\mu_0 = 0.2$ per stage. A stage with ancestor means $(1.0, 0.5)$ has $T^*_{\text{G}} = 32$ and $T^*_{\text{blind}} = 55.06$, and a stage with ancestor means $(1.0, 0.9)$ has $T^*_{\text{G}} = 800$ and $T^*_{\text{blind}} = 818.90$. Two $(1.0, 0.5)$ stages give the episodic ratio $55.06/32 = 1.72$. One stage of each kind gives $818.90/800 = 1.02$. These values are computed from Definition 6 with the released implementation (Appendix F). The simulation runs the released episodic IATS (Algorithm 1) on $H \in \{1, 2, 4, 8\}$ identical two-arm ancestor stages at $\delta = 0.01$ over 40 seeds. Its episode count grows sub-linearly in H. In these runs $\bar{N}_H / \log(1/\delta)$ is 175.6, 207.0, 239.9 and 276.4. The growth $\bar{N}_8 / \bar{N}_1 = 1.57$, with bootstrap 95% interval $[1.39, 1.80]$, lies below both the pre-specified line $0.5H = 4$ and the additive line $H = 8$. This sub-linear growth is consistent with $N = \max_h \tau_h$, since a maximum of H stopping times of identical stages grows slowly with H. It rules out an additive composition for the released algorithm. It does not test the value $\max_h T^*_h$ pointwise (Figure 4).

The following details accompany Section 6. E1 meets all four checks. E2 meets nine checks and fails the rate-stability check. E3 meets all fifteen checks and E4 meets all four. The blue line of Figure 2 is computed for the instance $\mu|_{A^{\text{anc}}} = (1.0, 0.5)$, $\mu_0 = 0.2$, $\sigma = 1$ of E1.

D THE COUPLED PROGRAM BEYOND DECOUPLING, WITH PARTIAL RESULTS AND CONJECTURES

Assumption 3(a) makes Section 4 a statement about H bandits (Remark 6). Assumption 3(b) makes the feasible allocations a product of simplices. This section defines the objects needed when (b) fails. It proves what the tools of this paper can prove, reports the illustrative study E5, and states the rest as conjectures.

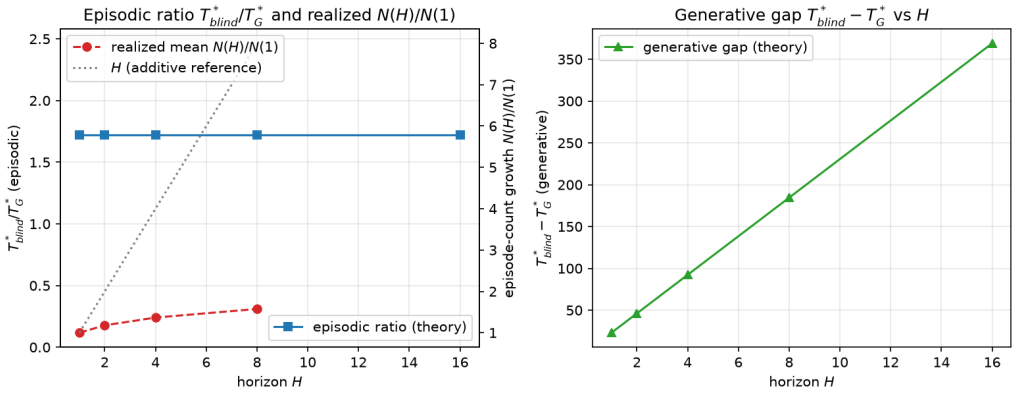


Figure 4: E4. In the left panel the exact episodic ratio is constant in H (blue). The realized episode count $\bar{N}_H/\bar{N}_1$ of the released episodic IATS (red, right-hand axis) grows sub-linearly and stays below the additive line H (grey, right-hand axis). The right panel shows that the exact generative difference is linear in H.

D.1 The coupled program

Definition 11 (Finite-horizon causal MDP, occupancy polytope, coupled and single-flip programs)**.** A finite-horizon causal MDP M consists of four ingredients. The first is a finite state space S partitioned into layers $S_1 = \{s_{\text{root}}\}, S_2, \dots, S_H$. The second is, for each $s \in S$, a finite action set A_s of hard interventions with means $\mu(s,a)$ and a transition kernel $p(\cdot \mid s,a)$ supported on the next layer (terminal at layer H). The kernel is known to both learners. Let $\mathcal{A} := \bigsqcup_{s\in S} A_s$ denote the set of all state-action pairs. The third is a branch set $B_s \subseteq A_s$, part of the model, whose actions all have the observational mean $\mu_0(s)$ and the observational kernel $p_0(\cdot \mid s)$. As motivation, apply the argument of Lemma 1 to the vector of rewards and state variables in place of Y. It gives this property for interventions on variables that are non-ancestors of every reward and of every state variable in an unrolled graph. State variables are not formalised here. The fourth is Gaussian rewards $Y \sim N(\mu(s,a), \sigma^2)$ at every visited (s,a), independent across visits. For a model M, $\mathbb{P}_M$ and $\mathbb{E}_M$ denote the law and expectation of the episodic observations when the means and kernel are those of M. M is *gated* at a state s if the kernels $p(\cdot \mid s,a)$, $a \in A_s$, are not all equal. In words, the action at s influences which states are visited next. An episode starts at s_{root} and follows a policy π (a map from states to distributions over actions) for H steps. Values are defined by backward induction, with $V_M \equiv 0$ after layer H, $Q_M(s,a) := \mu(s,a) + \sum_{s'} p(s' \mid s,a) V_M(s')$ and $V_M(s) := \max_a Q_M(s,a)$. The decision states are $S_{\text{dec}} := \{s : |A_s| \geq 2\}$. M has a *unique optimal policy* π^*_M if $\arg\max_a Q_M(s,a)$ is a single action at every $s \in S_{\text{dec}}$. The *local-argmax policy* is $\pi^\circ_M(s) := \arg\max_a \mu(s,a)$, defined when this is a single action at every $s \in S_{\text{dec}}$, and M is *greedy-aligned* if π°_M is defined and $\pi^\circ_M = \pi^*_M$. For a policy π let $\rho_\pi(s)$ be the probability of visiting s and $w_\pi(s,a) := \rho_\pi(s)\pi(a \mid s)$ its occupancy. The *occupancy polytope* is

$$\Omega(M) := \Big\{ w \geq 0 : \sum_a w(s_{\text{root}}, a) = 1, \ \sum_a w(s',a) = \sum_{s,a} w(s,a)\, p(s' \mid s,a) \ \ \forall s' \notin S_1 \Big\}.$$

There are two model classes. $\mathcal{M}_{\text{blind}}(M)$ is the set of all models with the same S, (A_s), (B_s) and p and arbitrary means. The second is $\mathcal{M}_{\text{G}}(M) := \{M' \in \mathcal{M}_{\text{blind}}(M) : \ \mu'(s,j) = \mu_0(s) \ \forall j \in B_s$, and $\pi^*_{M'}(s) \notin B_s \ \forall s \in S_{\text{dec}}$ whenever $\pi^*_{M'}$ is unique$\}$. The second constraint (no branch is optimal) is Assumption 4 below, imposed on the members of the class. For a class $\mathcal{M}$ and $M \in \mathcal{M}$ with a unique optimal policy, let $\text{Alt}_{\mathcal{M}}(M) := \{M' \in \mathcal{M} : \ M'$ has a unique optimal policy and $\pi^*_{M'} \neq \pi^*_M\}$ and $c_w(M,M') := \sum_{s,a} w(s,a)(\mu(s,a) - \mu'(s,a))^2/(2\sigma^2)$. The *coupled program* is

$$T^*_{\mathcal{M}}(M)^{-1} := \sup_{w \in \Omega(M)} \ \inf_{M' \in \text{Alt}_{\mathcal{M}}(M)} c_w(M,M').$$

For every M for which π°_M is defined, the *single-flip program* is $T^*_{\mathrm{sf},\mathcal{M}}(M)^{-1} := \sup_{w\in\Omega(M)} \min_{(s,a)\in R_{\mathcal{M}}} g_{s,a}(w)$, where

$$g_{s,a}(w) := \frac{\big(\mu(s,\pi^\circ_M(s)) - \mu(s,a)\big)^2}{2\sigma^2} \cdot \frac{w(s,\pi^\circ_M(s))\, w(s,a)}{w(s,\pi^\circ_M(s)) + w(s,a)}$$

with value 0 when the denominator vanishes. The rival sets are $R_{\mathrm{blind}} := \{(s,a) : s \in S_{\mathrm{dec}}, a \neq \pi^\circ_M(s)\}$ and $R_{\mathrm{G}} := R_{\mathrm{blind}} \setminus \{(s,j) : j \in B_s\}$, and $R_{\mathcal{M}}$ is R_{G} or R_{blind} according to the class. Let $T^*_{\mathrm{sf,G}}(M) := T^*_{\mathrm{sf},\mathcal{M}_{\mathrm{G}}(M)}(M)$ and $T^*_{\mathrm{sf,blind}}(M) := T^*_{\mathrm{sf},\mathcal{M}_{\mathrm{blind}}(M)}(M)$. We drop the argument M when it is clear. When the two classes give the same value, as when M has no branches, we drop the class as well and write $T^*_{\mathrm{sf}}(M)$.

Every $w \in \Omega(M)$ is the occupancy w_π of the policy $\pi(a \mid s) := w(s,a)/\sum_b w(s,b)$ (arbitrary where the denominator vanishes), and conversely every w_π lies in $\Omega(M)$. Both claims hold by induction over layers. For a fixed policy, $\rho_\pi(s') = \sum_{s,a} \rho_\pi(s)\pi(a \mid s)p(s' \mid s,a) = \sum_{s,a} w_\pi(s,a)p(s' \mid s,a)$ and $\sum_a w_\pi(s',a) = \rho_\pi(s')$, which are the flow constraints. Conversely, the flow constraints determine $\sum_a w(s,a)$ layer by layer as the visit probability of the policy π just defined. Under Assumption 3(b) with one decision state per layer, $\Omega(M) = \prod_h \Sigma_{A_h}$ and the coupled program is the episodic program of Definition 9. The single-flip program, the object computed by the Frank–Wolfe solver of E5, ranks actions by their local means and treats every local flip as an alternative.

Assumption 4 (No optimal branch)**.** $\pi^*_M(s) \notin B_s$ for every $s \in S_{\mathrm{dec}}$, or equivalently $M \in \mathcal{M}_{\mathrm{G}}(M)$.

Assumption 4 is the coupled analogue of Assumption 1(c). Unlike the equality of the branch means (Lemma 1), it is not implied by the graph. A branch keeps the observational kernel $p_0(\cdot \mid s)$, which may lead to better successor states than any non-branch action. So $Q_M(s,j) = \mu_0(s) + \sum_{s'} p_0(s' \mid s)V_M(s')$ can exceed $\max_{a\notin B_s} Q_M(s,a)$ even when $\mu_0(s) < \max_{a\notin B_s} \mu(s,a)$. It is a property of the instance, not of the graph, and it is a hypothesis of Propositions 2–3 and of Conjectures 1–2. The lower bound of Fact 1 transfers to the coupled program.

Lemma 11 (Lower bound for the coupled program)**.** *Let M be as in Definition 11 with a unique optimal policy. Let $\mathcal{M}$ be any class of models with the same S, (A_s), (B_s) and p that contains M, and let $\delta \in (0, 1/2)$. Consider a strategy that chooses its actions within an episode as a function of the past, of the current state and of internal randomisation, stops at an episode boundary N, and recommends a policy $\hat\pi$. Suppose that $\mathbb{P}_{M'}(N < \infty) = 1$ and $\mathbb{P}_{M'}(\hat\pi \neq \pi^*_{M'}) \leq \delta$ for every $M' \in \mathcal{M}$ with a unique optimal policy. Then $\mathbb{E}_M[N] \geq T^*_{\mathcal{M}}(M)\,\mathrm{kl}(\delta, 1-\delta)$.*

Proof. Appendix E.10. □

Proposition 1 (Convexity)**.** *Let M be as in Definition 11. Then* (i) *$\Omega(M)$ is a convex polytope.* (ii) *For every class $\mathcal{M}$, the map $w \mapsto \inf_{M'\in\mathrm{Alt}_{\mathcal{M}}(M)} c_w(M,M')$ is concave on $\mathbb{R}^{\mathcal{A}}_+$, so the coupled program is the maximisation of a concave function over a polytope.* (iii) *If π°_M is defined, each $g_{s,a}$ is concave on $\mathbb{R}^{\mathcal{A}}_+$, so the single-flip program is also the maximisation of a concave function over a polytope.*

Proof. (i) $\Omega(M)$ is defined by finitely many linear equalities and the inequalities $w \geq 0$. (ii) For each fixed M', $w \mapsto c_w(M,M')$ is linear, and an infimum of linear functions is concave. (iii) On $\{(x,y) \in \mathbb{R}^2_+ : x + y > 0\}$ one has $xy/(x+y) = (x+y)/4 - (x-y)^2/(4(x+y))$. The map $(x,y) \mapsto (x-y)^2/(x+y)$ is convex there, because it is the quadratic-over-linear function u^2/v with $u = x - y$ and $v = x + y > 0$. The extension by 0 at $(0,0)$ is continuous on the closed quadrant, since $0 \leq xy/(x+y) \leq \min(x,y)$, and concave on the interior. So the concavity inequality passes to the limit and the extension is concave on the closed quadrant. Hence each $g_{s,a}$, a positive constant times this function of two coordinates of w, is concave, and so is the minimum of finitely many of them. □

Proposition 2 (The blind lower bound survives coupling)**.** *Let M be as in Definition 11 with a unique optimal policy and satisfying Assumption 4. Let $s \in S_{\mathrm{dec}}$ have $k := |B_s| \geq 1$ branch actions. All $j \in B_s$ share $Q_M(s,j) = Q_0(s) := \mu_0(s) + \sum_{s'} p_0(s' \mid s)V_M(s')$. Let $\Delta_Q(s) := Q_M(s,\pi^*_M(s)) - Q_0(s)$, which is positive by Assumption 4. Then:*

(i) $T^*_{\mathcal{M}_{\mathrm{blind}}(M)}(M) \;\geq\; \dfrac{2\sigma^2}{\Delta_Q(s)^2}\,k$, *and for every* $\delta \in (0, 1/2)$ *every strategy as in Lemma 11 with* $\mathcal{M} = \mathcal{M}_{\mathrm{blind}}(M)$ *satisfies* $\mathbb{E}_M[N] \geq \left(2\sigma^2 k/\Delta_Q(s)^2\right) \mathrm{kl}(\delta, 1-\delta)$.

(ii) $T^*_{\mathcal{M}_{\mathrm{G}}(M)}(M)$ *is the same for every* $k \geq 1$, *because it depends only on* S, p, *the non-branch actions and their means, and* $(\mu_0(s), p_0(\cdot \mid s))$.

Proof. Appendix E.11. □

Proposition 3 (The single-flip program relative to the greedy-aligned class)**.** *Let* M *be greedy-aligned with a unique optimal policy, satisfying Assumption 4. Let* $\mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M) := \{M' \in \mathcal{M}_{\mathrm{G}}(M) : M'$ *greedy-aligned*$\}$ *and* $\mathcal{M}^{\mathrm{al}}_{\mathrm{blind}}(M) := \{M' \in \mathcal{M}_{\mathrm{blind}}(M) : M'$ *greedy-aligned*$\}$. *Then, for* $\mathcal{M} \in \{\mathcal{M}_{\mathrm{G}}(M), \mathcal{M}_{\mathrm{blind}}(M)\}$ *and the corresponding* $\mathcal{M}^{\mathrm{al}}$,

$$T^*_{\mathcal{M}^{\mathrm{al}}}(M) \;\leq\; T^*_{\mathrm{sf},\mathcal{M}}(M) \qquad \textit{and} \qquad T^*_{\mathcal{M}^{\mathrm{al}}}(M) \;\leq\; T^*_{\mathcal{M}}(M).$$

Proof. Appendix E.12. □

Proposition 3 says that the single-flip program is an *upper* bound on the characteristic time relative to the greedy-aligned class, as is the coupled program of the full class. A smaller class has fewer alternatives, so its inner infimum is larger and its characteristic time smaller. The single-flip program therefore yields no lower bound on the sample complexity of any learner. The lower bounds relative to the aligned and full classes are $T^*_{\mathcal{M}^{\mathrm{al}}}(M)\,\mathrm{kl}(\delta, 1-\delta)$ and $T^*_{\mathcal{M}}(M)\,\mathrm{kl}(\delta, 1-\delta)$ (Lemma 11). Example 1 shows that both inequalities of Proposition 3 can be strict and that $T^*_{\mathrm{sf},\mathcal{M}}(M)$ and $T^*_{\mathcal{M}}(M)$ can be in either order.

D.2 Value coupling

Example 1 (Value coupling on a gated MDP)**.** Let $H = 2$ and $S_2 = \{L, R\}$. The root has actions $A_{s_{\mathrm{root}}} = \{0, 1\}$, with action 0 leading to L and action 1 to R deterministically. This is a gated MDP (Definition 11), since the root action determines which leaf is visited. Each leaf has one action, with means abbreviated $\mu(L)$ and $\mu(R)$. Take $\sigma = 1$ and no branches, so $\mathcal{M}_{\mathrm{G}}(M) = \mathcal{M}_{\mathrm{blind}}(M)$, written $\mathcal{M}$, and $\mathcal{M}^{\mathrm{al}} := \mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M)$. The only decision state is the root and a policy is a root action. In all three instances $\mu(s_{\mathrm{root}}, \cdot) = (1.0, 0.5)$, so $\pi^\circ_M = 0$ and the single-flip program depends on the local gap 0.5 only. Appendix E.13 shows that

$$T^*_{\mathcal{M}}(M) = \frac{16\sigma^2}{\Delta^2_{\mathrm{root}}}, \qquad \Delta_{\mathrm{root}} := |Q_M(s_{\mathrm{root}}, 0) - Q_M(s_{\mathrm{root}}, 1)|, \qquad T^*_{\mathrm{sf},\mathcal{M}}(M) = \frac{8\sigma^2}{0.5^2} = 32,$$

and computes $T^*_{\mathcal{M}^{\mathrm{al}}}(M)$ for the two aligned instances. *(a) Misaligned near-tie.* Let $\mu(L) = 0.0$ and $\mu(R) = 0.5645$. Then $Q_M(s_{\mathrm{root}}, \cdot) = (1.0, 1.0645)$, so $\pi^*_M = 1 \neq \pi^\circ_M = 0$, $\Delta_{\mathrm{root}} = 0.0645$ and $T^*_{\mathcal{M}}(M) = 16/0.0645^2 \approx 3846$, about 120 times the single-flip value 32. A stopping rule that certifies the local argmax at the root certifies the wrong policy. *(b) Aligned, small Q-margin.* Let $\mu(L) = 0.5$ and $\mu(R) = 0.9$. Then $Q_M(s_{\mathrm{root}}, \cdot) = (1.5, 1.4)$, so M is greedy-aligned with $\Delta_{\mathrm{root}} = 0.1$, $T^*_{\mathcal{M}}(M) = 1600$, and $T^*_{\mathcal{M}^{\mathrm{al}}}(M) = T^*_{\mathrm{sf},\mathcal{M}}(M) = 32$. *(c) Aligned, large Q-margin.* Let $\mu(L) = 0.9$ and $\mu(R) = 0.0$. Then $Q_M(s_{\mathrm{root}}, \cdot) = (1.9, 0.5)$, so M is greedy-aligned with $\Delta_{\mathrm{root}} = 1.4$, and $T^*_{\mathcal{M}^{\mathrm{al}}}(M) = T^*_{\mathcal{M}}(M) = 16/1.4^2 \approx 8.16 < 32 = T^*_{\mathrm{sf},\mathcal{M}}(M)$. The values $T^*_{\mathcal{M}}(M)$ and $T^*_{\mathrm{sf},\mathcal{M}}(M)$ of all three instances are reproduced by the released code (Appendix F). The value $T^*_{\mathcal{M}^{\mathrm{al}}}(M)$ is computed in Appendix E.13.

Instance (a) shows the mechanism that Assumption 3(b) removes. When the action at a state determines which states are visited, the optimal action is the Q-argmax rather than the local reward argmax. The alternative set then contains models that change the optimal policy through downstream values only. The per-state single-flip infimum of Lemma 7 does not enumerate them. Instance (b) shows that the coupled program can exceed the single-flip program by a large factor even for a greedy-aligned instance. Instance (c) shows that the aligned-class characteristic time can be strictly smaller than the single-flip value. There the cheapest alternative that changes the optimal policy while remaining greedy-aligned moves the leaf means as well as the root means. It is cheaper than any alternative that flips the root's local argmax alone.

D.3 Illustrative study beyond decoupling (E5)

E5 is an illustrative, non-confirmatory study of small MDPs in which several states share a stage or in which Assumption 3(b) fails. It computes the single-flip program of Definition 11 by a Frank–Wolfe method over the occupancy polytope. The method is deterministic and its linear-maximisation oracle is one backward-induction pass. It also simulates a *static-allocation local-certification strategy*. In this strategy each episode follows the policy $\pi(a \mid s) = w^*(s,a)/\sum_b w^*(s,b)$. Here w^* is the Frank–Wolfe maximiser of the single-flip program computed from the *true* means. This is an oracle allocation, without tracking or forced exploration. The run stops at the first episode n at which the local Chernoff statistic of every decision state s exceeds $\beta_{\mathrm{cal}}(n,\delta;K) + \log|S_{\mathrm{dec}}|$. The statistic is computed from the visits to s, over the non-branch rivals for the graph-aware learner. The threshold is the calibrated one of Section 5, item (iv), and K is the largest action count of a decision state. The strategy recommends the local empirical argmax at every decision state (over the non-branch actions for the graph-aware learner). Both instance families of E5 are greedy-aligned (Definition 11). The code reports this (`is_greedy_aligned`) but does not assert it. The first family is a rectangular MDP. Its single-action root leads with probability $1/2$ each to two leaves with ancestor means $(1.0, 0.5)$ and k branches each. The ratio $T^*_{\mathrm{sf,blind}}/T^*_{\mathrm{sf,G}}$ is 1.00, 1.24, 1.48, 1.72 at $k = 0, 2, 4, 6$. On this instance every decision state's actions are terminal, so the optimal policy of every model is its local-argmax policy. The argument of Lemmas 7 and 3(i) then shows that the single-flip program equals the coupled program of Definition 11 for both classes. The second lemma applies because (H) holds at both leaves, $0.5 \geq 0.2$. The second family is a gated $H = 2$ MDP. A root decision with means $(1.0, 0.0)$ determines which of two identical leaves is visited, each with means $(1.0, 0.5)$ plus k branches. The single-flip ratio is 1.00, 1.24, 1.48, 1.72 at $k = 0, 2, 4, 6$. Whether the single-flip program equals the coupled program or the aligned-class program on this instance is not established. On the $k = 6$ gated instance, the static-allocation strategy made no error in 60 runs for either learner at $\delta = 0.01$. These runs use the calibrated threshold, for which this paper proves no error bound. Its realized $N/\log(1/\delta)$ was $2.42\,T^*_{\mathrm{sf,G}}$ for the graph-aware learner, with $T^*_{\mathrm{sf,G}} = 64.2$, and $3.69\,T^*_{\mathrm{sf,blind}}$ for the blind learner, with $T^*_{\mathrm{sf,blind}} = 110.3$. The realized speed-up was 2.62 with bootstrap 95% interval $[2.37, 2.90]$, above the single-flip ratio 1.72. We also checked the concavity of the single-flip objective (Proposition 1) numerically. There were 0 violations of the concavity inequality over 30,000 random pairs of occupancies on the gated instance.

D.4 Conjectures and open problems

Conjecture 1 (Coupled achievability)**.** *For every M as in Definition 11 with a unique optimal policy and satisfying Assumption 4, there is a family of strategies, indexed by $\delta \in (0, 1/2)$, each δ-correct relative to $\mathcal{M}_{\mathrm{G}}(M)$ as in Lemma 11, with $\limsup_{\delta\to 0} \mathbb{E}_M[N]/\log(1/\delta) \leq T^*_{\mathcal{M}_{\mathrm{G}}(M)}(M)$, and likewise for $\mathcal{M}_{\mathrm{blind}}(M)$.*

The natural candidate is Track-and-Stop over $\Omega(M)$ with the value-coupled alternative set. Its stopping statistic would be computed by backward induction over the alternatives that change π^*, as Al Marjani et al. (2021) do for tabular MDPs without graph structure. Proposition 1 shows that convexity is not the obstacle, and Example 1(a) shows that the alternative set must be the coupled one. The lower-bound half of the statement is Lemma 11.

Conjecture 2 (Local certification relative to the greedy-aligned class)**.** *Let M be greedy-aligned with a unique optimal policy and satisfying Assumption 4. Consider the strategy that tracks a maximiser of the single-flip program $T^*_{\mathrm{sf,G}}(M)$ over $\Omega(M)$, with forced exploration of every (s,a), $s \in S_{\mathrm{dec}}$, $a \notin B_s$. It stops when the local Chernoff statistic of every decision state exceeds a threshold as in Theorem 3 at level $\delta/|S_{\mathrm{dec}}|$, and recommends the local empirical argmax at every decision state. This strategy is δ-correct relative to $\mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M)$ and satisfies $\limsup_{\delta\to 0} \mathbb{E}_M[N]/\log(1/\delta) \leq T^*_{\mathrm{sf,G}}(M)$. The same holds for the blind class with $T^*_{\mathrm{sf,blind}}(M)$.*

Conjecture 2, if true, gives an achievable rate relative to the aligned class but not an optimal one. By Proposition 3 the lower bound relative to $\mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M)$ is $T^*_{\mathcal{M}^{\mathrm{al}}_{\mathrm{G}}(M)}(M) \leq T^*_{\mathrm{sf,G}}(M)$, with equality in Example 1(b) and strict inequality in Example 1(c). E5 (Appendix D.3) gives numerical evidence for the δ-correctness part only, and for a static oracle-allocation variant of the strategy rather than the tracking strategy stated. The evidence consists of no error in 60 runs per learner at $\delta = 0.01$ on

two aligned instances, with realized episode counts of 2.42 and 3.69 times the single-flip values at this δ. Relative to the full class $\mathcal{M}_{\mathrm{G}}(M)$, local certification is not δ-correct (Example 1(a)).

Open problems. (1) Conjecture 1, and whether $T^*_{\mathcal{M}_{\mathrm{G}}(M)}(M)$ admits a closed form in per-state quantities. Example 1 shows that no form based on local gaps alone can be exact. (2) A δ-correctness proof for the calibrated threshold $\log((1+\log t)(K-1)/\delta)$ for Gaussian arms, or an explicit constant $C(\alpha, K)$ in Proposition 12 of Garivier & Kaufmann (2016). Mixture-martingale thresholds (Kaufmann & Koolen, 2021) are the natural tool, but we have not verified their applicability to the present setting. (3) The finite-δ overhead of the analysed threshold. At the tested δ the mean stopping time of IATS is 3.2 to 9.5 times $T^*_{\mathrm{G}} \log(1/\delta)$, against 1.17 to 1.48 under the calibrated threshold (Section 5). A threshold that is both proven and tight at finite δ is open. (4) Conjecture 2, and a characterisation of the instances on which $T^*_{\mathcal{M}^{\mathrm{al}}}(M) = T^*_{\mathrm{sf},\mathcal{M}}(M)$. (5) Robustness of Lemma 1 to approximate non-ancestry (small rather than zero effects) and to an unknown μ_0 (Remark 2). (6) Rewards that are not stage-additive, and stages with several states visited with policy-dependent probabilities, which is the coupled case of this section.

E Proofs

Throughout, c_w, h_a, Alt, Σ_A and $\mathcal{S}_A$ are as in Section 2, and $\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$.

E.1 Proof of Lemmas 2 and 4

Hypotheses. $\mu \in \mathcal{S}_A$, $K = |A| \geq 2$, $\Delta_a > 0$ for all $a \neq a^*$, with c_w as in Definition 6 and h_a as in (3).

Proof. (a) Fix $w \in \Sigma_A$. For $a \neq a^*$ let $U_a := \{\lambda \in \mathbb{R}^A : \lambda_a > \lambda_{a^*}\}$, an open set. A vector $\lambda \in \mathcal{S}_A$ satisfies $a^*(\lambda) \neq a^*$ if and only if λ_{a^*} is not its maximum. Equivalently, $\lambda \in \bigcup_{a\neq a^*} U_a$. Hence $\mathrm{Alt}_{\mathrm{blind}}(\mu) = \mathcal{S}_A \cap \bigcup_{a\neq a^*} U_a$ and $\inf_{\mathrm{Alt}_{\mathrm{blind}}(\mu)} c_w = \min_{a\neq a^*} \inf_{\mathcal{S}_A \cap U_a} c_w$. The set $\mathcal{S}_A$ is the complement of the finitely many hyperplanes $\{\lambda_a = \lambda_b\}$, $a \neq b$, so it is dense in $\mathbb{R}^A$. Therefore $\mathcal{S}_A \cap U_a$ is dense in the open set U_a. Since $\lambda \mapsto c_w(\mu,\lambda)$ is continuous, $\inf_{\mathcal{S}_A\cap U_a} c_w = \inf_{U_a} c_w = \inf_{\bar{U}_a} c_w$ with $\bar{U}_a = \{\lambda_a \geq \lambda_{a^*}\}$. It remains to compute $\inf_{\bar{U}_a} c_w$. For $\lambda \in \bar{U}_a$, $c_w(\mu,\lambda) \geq \left[w_{a^*}(\mu_{a^*} - \lambda_{a^*})^2 + w_a(\mu_a - \lambda_a)^2\right]/(2\sigma^2)$, with equality when $\lambda_b = \mu_b$ for $b \notin \{a, a^*\}$. If $w_a = 0$, the choice $\lambda_a = \mu_{a^*} + 1$, $\lambda_b = \mu_b$ otherwise, lies in $\bar{U}_a$ and has cost $0 = h_a(w)$. If $w_{a^*} = 0$, the choice $\lambda_{a^*} = \mu_a - 1$, $\lambda_b = \mu_b$ otherwise, does the same. If $w_a > 0$ and $w_{a^*} > 0$, the function $\varphi(x,y) := w_{a^*}(\mu_{a^*} - x)^2 + w_a(\mu_a - y)^2$ is strictly convex with unconstrained minimiser (μ_{a^*}, μ_a), which violates $y \geq x$ because $\mu_a < \mu_{a^*}$. A convex function whose unconstrained minimiser lies outside a closed convex set attains its minimum over that set on the boundary. An interior minimiser would be a local, and therefore global, unconstrained minimiser. On the boundary $x = y = z$, $\varphi(z,z) = w_{a^*}(\mu_{a^*} - z)^2 + w_a(\mu_a - z)^2$ is minimised at $z^* = (w_{a^*}\mu_{a^*} + w_a\mu_a)/(w_{a^*} + w_a)$ with value $w_{a^*}w_a\Delta_a^2/(w_{a^*}+w_a)$. Dividing by $2\sigma^2$ gives $\inf_{\bar{U}_a} c_w = h_a(w)$, and therefore $\inf_{\mathrm{Alt}_{\mathrm{blind}}(\mu)} c_w = \min_{a\neq a^*} h_a(w)$. Taking the supremum over w gives the expression for $T^*_{\mathrm{blind}}(\mu)^{-1}$.

(b) The function $\chi(x,y) := xy/(x+y)$ on $[0,1]^2 \setminus \{(0,0)\}$, extended by $\chi(0,0) := 0$, is continuous because $0 \leq \chi(x,y) \leq \min(x,y)$. Hence each h_a is continuous on Σ_A, and so is $F = \min_{a\neq a^*} h_a$. Since Σ_A is compact, F attains its maximum. For the uniform allocation $w \equiv 1/K$, $h_a(w) = \Delta_a^2/(4\sigma^2 K) > 0$ for every $a \neq a^*$, so $\max F > 0$. If $w_a = 0$ for some a, then $F(w) = 0$, because for $a = a^*$ all rival terms vanish, and for $a \neq a^*$ the term $h_a(w)$ vanishes. So every maximiser has all coordinates positive.

(c) Let w^* maximise F, let $v^* := F(w^*) > 0$, and suppose that $h_{b'}(w^*) > v^*$ for some rival b'. For $0 < \varepsilon < w^*_{b'}$ define w^ε by $w^\varepsilon_{b'} := w^*_{b'} - \varepsilon$, $w^\varepsilon_{a^*} := w^*_{a^*} + \varepsilon$, and $w^\varepsilon_b := w^*_b$ otherwise. Then $w^\varepsilon \in \Sigma_A$. For every rival $a \neq b'$, the coordinate $w_a = w^*_a > 0$ is unchanged. For fixed $y > 0$ the map $x \mapsto xy/(x+y) = y - y^2/(x+y)$ is strictly increasing on $(0,\infty)$, so $h_a(w^\varepsilon) > h_a(w^*) \geq v^*$. By continuity of $h_{b'}$, $h_{b'}(w^\varepsilon) > v^*$ for ε small enough. Hence $F(w^\varepsilon) > v^* = \max F$, a contradiction. Therefore $h_a(w^*) = v^* = T^*_{\mathrm{blind}}(\mu)^{-1}$ for every rival a. □

E.2 Proof of Lemmas 3 and 5

Hypotheses. Definitions 1–6, Assumptions 1–2, and (H), which states $\mu_a \geq \mu_0$ for all $a \in A^{\mathrm{anc}}$.

Proof. (i) By Corollary 1(iii) and its proof, for $\lambda \in \mathcal{M}_{\mathrm{G}}$ the cost $c_w(\mu, \lambda)$ involves only the ancestor coordinates of λ and of w. In ancestor coordinates, $\mathrm{Alt}_{\mathrm{G}}(\mu) = \mathcal{S}' \cap \bigcup_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} U'_a$. Here $R_> := \{\lambda \in \mathbb{R}^{A^{\mathrm{anc}}} : \max_b \lambda_b > \mu_0\}$ is open and $U'_a := \{\lambda \in R_> : \lambda_a > \lambda_{a^*}\}$ is open. The set $\mathcal{S}' := \{\lambda \in R_> : \arg\max_b \lambda_b \text{ is a single arm}\}$ is dense in $R_>$, since it is the complement in $R_>$ of finitely many hyperplanes. When $k = 0$ the same description holds with $B = \emptyset$. By density and continuity, $\inf_{\mathrm{Alt}_{\mathrm{G}}(\mu)} c_w = \min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} \inf_{U'_a} c_w$.

($\geq$) $U'_a \subseteq \{\lambda_a \geq \lambda_{a^*}\}$, so $\inf_{U'_a} c_w \geq h_a(w)$ by the computation in the proof of Lemma 2(a). This step does not use (H).

($\leq$) If $w_a = 0$, take the vector with $\lambda_a = \mu_{a^*} + 1$ and $\lambda_b = \mu_b$ otherwise. It lies in U'_a, because its maximum exceeds μ_0 and $\lambda_a > \lambda_{a^*}$, and it has cost $0 = h_a(w)$. If $w_{a^*} = 0$, take the vector with $\lambda_{a^*} = \mu_a - 1$, $\lambda_a = \mu_a + \eta$ and $\lambda_b = \mu_b$ otherwise. It lies in U'_a for every $\eta > 0$, because $\mu_a + \eta > \mu_0$ by (H). Its cost $w_a \eta^2/(2\sigma^2)$ tends to $0 = h_a(w)$. If $w_a > 0$ and $w_{a^*} > 0$, let $z^* := (w_{a^*}\mu_{a^*} + w_a \mu_a)/(w_{a^*} + w_a) \in [\mu_a, \mu_{a^*}]$ and define λ^η by $\lambda^\eta_a := z^* + \eta$, $\lambda^\eta_{a^*} := z^* - \eta$, $\lambda^\eta_b := \mu_b$ otherwise. By (H), $z^* \geq \mu_a \geq \mu_0$, so $\max_b \lambda^\eta_b \geq z^* + \eta > \mu_0$ and $\lambda^\eta \in U'_a$. Moreover $c_w(\mu, \lambda^\eta) \to \big[w_{a^*}(\mu_{a^*} - z^*)^2 + w_a(\mu_a - z^*)^2\big]/(2\sigma^2) = h_a(w)$ as $\eta \to 0$.

Hence $\inf_{\mathrm{Alt}_{\mathrm{G}}(\mu)} c_w = \min_{a \in A^{\mathrm{anc}} \setminus \{a^*\}} h_a(w) =: F_{\mathrm{anc}}(w)$. This function depends only on the ancestor coordinates of w and is non-decreasing in each of them. Therefore the supremum over Σ_A equals the supremum over $\Sigma_{A^{\mathrm{anc}}}$, since the branch mass can be moved to a^*. The latter supremum is attained by Lemma 4(b) applied to the ancestor bandit, and $T^*_{\mathrm{G}}(\mu) = T^*_{\mathrm{anc}}(\mu)$ for every $k \geq 0$. If $A^{\mathrm{anc}} = \{a^*, a\}$, then $F_{\mathrm{anc}}(u) = \Delta^2_{\min} u_{a^*} u_a/(2\sigma^2(u_{a^*} + u_a)) = \Delta^2_{\min} u_{a^*} u_a/(2\sigma^2)$ on $\Sigma_{A^{\mathrm{anc}}}$, maximised at $u_{a^*} = u_a = 1/2$ with value $\Delta^2_{\min}/(8\sigma^2)$, so $T^*_{\mathrm{anc}}(\mu) = 8\sigma^2/\Delta^2_{\min}$.

(ii) Let $R^A_> := \{\lambda \in \mathbb{R}^A : \max_b \lambda_b > \mu_0\}$, an open set, so that $\mathcal{M}'_{\mathrm{blind}} = R^A_>$ and $\mathrm{Alt}_{\mathcal{M}'_{\mathrm{blind}}}(\mu) = \mathcal{S}_A \cap R^A_> \cap \bigcup_{a \neq a^*} U_a$ with U_a as in Appendix E.1. Since $\mathcal{S}_A$ is dense in $\mathbb{R}^A$ and $R^A_> \cap U_a$ is open, density and continuity give $\inf_{\mathrm{Alt}_{\mathcal{M}'_{\mathrm{blind}}}(\mu)} c_w = \min_{a \neq a^*} \inf_{R^A_> \cap U_a} c_w$. ($\geq$) $R^A_> \cap U_a \subseteq \bar{U}_a$, so the infimum is at least $h_a(w)$ by Appendix E.1. ($\leq$) Fix a rival a. Reuse the vectors of the ($\leq$) step of part (i), with μ_a now the mean of any rival, ancestor or branch. They lie in U_a and have costs tending to $h_a(w)$. Each of them lies in $R^A_>$. In the first case its coordinate a equals $\mu_{a^*} + 1 > \mu_0$. In the second and third cases its coordinate a equals $\mu_a + \eta$ or $z^* + \eta$. Both exceed μ_0, because $z^* \geq \mu_a$ and $\mu_a \geq \mu_0$. The last inequality is (H) if $a \in A^{\mathrm{anc}}$ and Lemma 1 if $a \in B$. Hence $\inf_{R^A_> \cap U_a} c_w \leq h_a(w)$, and $\inf_{\mathrm{Alt}_{\mathcal{M}'_{\mathrm{blind}}}(\mu)} c_w = \min_{a \neq a^*} h_a(w)$, which is the blind inner infimum of Lemma 2(a). Taking the supremum over w gives $T^*_{\mathcal{M}'_{\mathrm{blind}}}(\mu) = T^*_{\mathrm{blind}}(\mu)$. □

E.3 Proof of Fact 1

Hypotheses. $\mathcal{M} \subseteq \mathbb{R}^A$, $\mu \in \mathcal{M} \cap \mathcal{S}_A$, $\delta \in (0, 1/2)$, and a strategy that is δ-correct relative to $\mathcal{M}$ (Definition 5).

Proof. If $\mathrm{Alt}_{\mathcal{M}}(\mu) = \emptyset$ then $T^*_{\mathcal{M}}(\mu) = 0$ and there is nothing to prove. If $\mathbb{E}_\mu[\tau] = \infty$ the bound holds trivially. Assume otherwise and fix $\lambda \in \mathrm{Alt}_{\mathcal{M}}(\mu)$. Let $\mathcal{G}_\delta := \{\hat{a} = a^*(\mu)\} \in \mathcal{F}_\tau$ be the success event. The letter E is reserved for the edge set. Since $\mu, \lambda \in \mathcal{M} \cap \mathcal{S}_A$, $\tau < \infty$ almost surely under both $\mathbb{P}_\mu$ and $\mathbb{P}_\lambda$, and $\mathbb{P}_\mu(\mathcal{G}_\delta) \geq 1 - \delta$. Moreover $\mathbb{P}_\lambda(\mathcal{G}_\delta) \leq \mathbb{P}_\lambda(\hat{a} \neq a^*(\lambda)) \leq \delta$ because $a^*(\mu) \neq a^*(\lambda)$. Lemma 1 of Kaufmann et al. (2016) concerns two bandit models with mutually absolutely continuous arm distributions, here $(N(\mu_a, \sigma^2))_a$ and $(N(\lambda_a, \sigma^2))_a$. It states that for any sampling rule and any almost surely finite stopping time τ,

$$\sum_{a \in A} \mathbb{E}_\mu[N_a(\tau)]\, \mathrm{KL}\big(N(\mu_a, \sigma^2), N(\lambda_a, \sigma^2)\big) \;\geq\; \mathrm{kl}\big(\mathbb{P}_\mu(\mathcal{G}), \mathbb{P}_\lambda(\mathcal{G})\big)$$

for every event $\mathcal{G} \in \mathcal{F}_\tau$. Here $\mathrm{KL}_a := \mathrm{KL}(N(\mu_a, \sigma^2), N(\lambda_a, \sigma^2)) = (\mu_a - \lambda_a)^2/(2\sigma^2)$ is the per-arm divergence. Their statement concerns strategies without internal randomisation. Its proof uses only two facts. The log-likelihood ratio of the observations up to τ has $\mathbb{P}_\mu$-expectation

$\sum_a \mathbb{E}_\mu[N_a(\tau)]\mathrm{KL}_a$, and kl satisfies the data-processing inequality. With the randomisation (U_t) of Definition 5, the likelihood ratio $\mathrm{d}\mathbb{P}_\mu/\mathrm{d}\mathbb{P}_\lambda$ on $\mathcal{F}_\tau$ is the same function of the rewards. The reason is that the conditional law of each U_t given the past is the uniform law under both models. The two steps of their proof therefore apply unchanged. Take $\mathcal{G} = \mathcal{G}_\delta$. Then $\mathbb{P}_\mu(\mathcal{G}_\delta) \geq 1-\delta > \delta \geq \mathbb{P}_\lambda(\mathcal{G}_\delta)$, which is where $\delta < 1/2$ is used. Since $\mathrm{kl}(p,q)$ is non-decreasing in p on $[q,1]$ and non-increasing in q on $[0,p]$, the right side is at least $\mathrm{kl}(1-\delta,\delta) = \mathrm{kl}(\delta,1-\delta)$. Set $w_a := \mathbb{E}_\mu[N_a(\tau)]/\mathbb{E}_\mu[\tau]$, which lies in Σ_A because $\sum_a N_a(\tau) = \tau$. With the value of KL_a above, the inequality reads $\mathbb{E}_\mu[\tau]\, c_w(\mu,\lambda) \geq \mathrm{kl}(\delta,1-\delta)$. It holds for every $\lambda \in \mathrm{Alt}_{\mathcal{M}}(\mu)$, so $\mathbb{E}_\mu[\tau]\, \inf_{\lambda\in\mathrm{Alt}_{\mathcal{M}}(\mu)} c_w(\mu,\lambda) \geq \mathrm{kl}(\delta,1-\delta)$. The infimum is at most $\sup_{w'} \inf_\lambda c_{w'} = T^*_{\mathcal{M}}(\mu)^{-1}$, so $\mathbb{E}_\mu[\tau] \geq T^*_{\mathcal{M}}(\mu)\,\mathrm{kl}(\delta,1-\delta)$. Finally $\mathrm{kl}(\delta,1-\delta) = (1-2\delta)\log\frac{1-\delta}{\delta} = (1-2\delta)\big(\log(1/\delta)+\log(1-\delta)\big)$, so $\mathrm{kl}(\delta,1-\delta)/\log(1/\delta) \to 1$ as $\delta \to 0$, which gives the $\liminf$ statement. □

E.4 Proof of Theorem 1

Hypotheses. $|A^{\mathrm{anc}}| \geq 2$, $(\mu_a)_{a\in A^{\mathrm{anc}}}$ with unique maximiser a^*, $\mu_0 < \mu_{a^*}$, $\mu^{(k)}$ as in the statement, and $k \geq 1$.

Proof. (i), lower bound. Let $w \in \Sigma_A$ with $A = A^{\mathrm{anc}} \sqcup B^{(k)}$. Since $\sum_{j\in B^{(k)}} w_j \leq 1$, some $j \in B^{(k)}$ has $w_j \leq 1/k$. For this j, $h_j(w) = \frac{\Delta_0^2}{2\sigma^2}\cdot\frac{w_{a^*}}{w_{a^*}+w_j}\, w_j \leq \frac{\Delta_0^2}{2\sigma^2}\, w_j \leq \frac{\Delta_0^2}{2\sigma^2 k}$. Hence $\min_{a\neq a^*} h_a(w) \leq \Delta_0^2/(2\sigma^2 k)$ for every w, and by Lemma 2(a), $T^*_{\mathrm{blind}}(\mu^{(k)})^{-1} \leq \Delta_0^2/(2\sigma^2 k)$.

(i), upper bound. Let $K := |A^{\mathrm{anc}}| + k$ and $w \equiv 1/K$. For every rival a, $h_a(w) = \Delta_a^2/(2\sigma^2(K+K)) = \Delta_a^2/(4\sigma^2 K) \geq \underline{\Delta}^2/(4\sigma^2 K)$. This uses that every rival gap is at least $\underline{\Delta} = \min(\Delta_{\min},\Delta_0)$, since the branches have gap Δ_0 by Lemma 1. By Lemma 2(a), $T^*_{\mathrm{blind}}(\mu^{(k)})^{-1} \geq \underline{\Delta}^2/(4\sigma^2 K)$.

(ii). Corollary 1(iii) gives $T^*_{\mathrm{G}}(\mu^{(k)}) = T^*_{\mathrm{G}}(\mu^{(1)})$ for all $k \geq 1$. Moreover $T^*_{\mathrm{G}}(\mu^{(k)}) < \infty$, because the inner infimum is positive for w uniform on A^{anc} and zero on $B^{(k)}$. Indeed, the ($\geq$) step of Appendix E.2 does not use (H) and bounds it below by $\min_{a\in A^{\mathrm{anc}}\setminus\{a^*\}} h_a(w) > 0$. Finally $T^*_{\mathrm{G}}(\mu^{(k)}) > 0$ because $\mathrm{Alt}_{\mathrm{G}}(\mu^{(k)})$ is non-empty, so the inner infimum is finite for every w. The vector with $\lambda_a = \mu_{a^*}+1$ for one $a \in A^{\mathrm{anc}}\setminus\{a^*\}$ and $\lambda_b = \mu_b^{(k)}$ otherwise belongs to it.

(iii). Divide the bounds in (i) by $\bar{T}_{\mathrm{G}}$, and use $|A^{\mathrm{anc}}| + k \leq (1+|A^{\mathrm{anc}}|)k$ for $k \geq 1$. The last statement is Lemma 3(i). □

E.5 Proof of Lemma 6

Hypotheses. Definition 8, Assumption 3(a) ($\mathrm{An}_{G^{(H)}}(Y_h) \subseteq V_h$ for every h), and an episode $(a_1,\dots,a_H)$ with $X_{a_h} \in V_h \setminus \{Y_h\}$.

Proof. With a slight abuse of notation, in this appendix let $V := \bigcup_h V_h$, $D_h := \mathrm{An}(Y_h)\cup\{Y_h\} \subseteq V_h$ and $D := \bigcup_h D_h$. The last union is disjoint, since the V_h are disjoint. Let $\bar{D} := V \setminus D$ and $I := \{X_{a_1},\dots,X_{a_H}\}$. As in the proof of Lemma 1, every parent of a member of D_h lies in D_h, and every child of a member of $\bar{D}$ lies in $\bar{D}$. Indeed, a member of $\bar{D}$ with a child in D would be an ancestor of some $Y_{h'}$. For each h, $X_{a_h} \in V_h$ lies either in D_h or in $\bar{D}$, because if $X_{a_h} \in D_{h'}$ then $X_{a_h} \in V_{h'}$, so $h' = h$. In particular $X_{a_h} \notin \mathrm{An}(Y_h)$ if and only if $X_{a_h} \notin \mathrm{An}(Y_{h'})$ for every h', which is (iii).

(i) The truncated factorisation for the joint intervention is $P_{\mathrm{do}(a_1,\dots,a_H)}(v) = \prod_{V_i\in V\setminus I} P(v_i \mid \mathrm{pa}_i)$ for v consistent with the intervention. Marginalising over all variables other than $Y_1,\dots,Y_H$,

$$P_{\mathrm{do}(a_1,\dots,a_H)}(y_1,\dots,y_H) = \sum_{(v_i)_{V_i\in D\setminus(I\cup\{Y_1,\dots,Y_H\})}} \prod_{V_j\in D\setminus I} P(v_j \mid \mathrm{pa}_j)\cdot\Bigg[\sum_{(v_i)_{V_i\in\bar{D}\setminus I}} \prod_{V_i\in\bar{D}\setminus I} P(v_i \mid \mathrm{pa}_i)\Bigg],$$

where the factors of $D\setminus I$ involve only variables of D. This holds because parents of members of D lie in D and the intervened variables are fixed at their values. For every fixed value of the D-variables the bracket equals one. To see this, take $V_i \in \bar{D}\setminus I$ maximal in a topological order of $G^{(H)}$. Its children lie in $\bar{D}$. Each child is either intervened, so its factor is absent, or later in the

order, so it is already summed out. Therefore V_i occurs only in its own factor, which sums to one. Iterating in decreasing topological order gives the claim. Since $D = \bigsqcup_h D_h$ and the factors of D_h involve only D_h-variables, the remaining sum splits as

$$P_{\mathrm{do}(a_1,\dots,a_H)}(y_1,\dots,y_H) = \prod_{h=1}^{H} \Pi_h(y_h; a_h), \qquad \Pi_h(y_h; a_h) := \sum_{(v_i)_{V_i \in D_h \setminus (I \cup \{Y_h\})}} \prod_{V_j \in D_h \setminus I} P(v_j \mid \mathrm{pa}_j),$$

where Π_h depends on the episode only through a_h, via whether $X_{a_h} \in D_h$ and via its value. The same computation with $I = \{X_{a_h}\}$ gives $P_{\mathrm{do}(a_h)}(y_1,\dots,y_H) = \Pi_h(y_h; a_h) \prod_{h' \neq h} \Pi^0_{h'}(y_{h'})$ with $\Pi^0_{h'}(y_{h'}) := \sum_{(v_i)_{V_i \in D_{h'} \setminus \{Y_{h'}\}}} \prod_{V_j \in D_{h'}} P(v_j \mid \mathrm{pa}_j)$. Each $\Pi^0_{h'}$ sums to one over $y_{h'}$, so $P_{\mathrm{do}(a_h)}(y_h) = \Pi_h(y_h; a_h)$. Indeed, the product over $D_{h'}$ is a joint distribution of the $D_{h'}$-variables by the same topological argument, since parents of members of $D_{h'}$ lie in $D_{h'}$. Hence $P_{\mathrm{do}(a_1,\dots,a_H)}(y_1,\dots,y_H) = \prod_h P_{\mathrm{do}(a_h)}(y_h)$, which gives the independence of $Y_1,\dots,Y_H$ and the dependence of the law of Y_h on a_h only.

(ii) If $j \in B_h$ then $X_j \in \bar{D}$, so $\Pi_h(y_h; j) = \Pi^0_h(y_h)$. The computation with $I = \emptyset$ gives $P(y_h) = \Pi^0_h(y_h)$. Hence $P_{\mathrm{do}(j)}(Y_h \in \cdot) = P(Y_h \in \cdot)$ and $\mu_h(j) = \mu_{h,0}$. □

E.6 Proof of Theorem 2

Hypotheses. Definitions 8–9, Assumption 3, $\mathcal{M} = \prod_h \mathcal{M}_h \in \{\mathcal{M}^{(H)}_{\mathrm{G}}, \mathcal{M}^{(H)}_{\mathrm{blind}}\}$, $\mu \in \mathcal{M} \cap \mathcal{S}^{(H)}$. In (iii) and (iv), $\delta \in (0, 1/2)$. Let $T^*_h := T^*_{\mathcal{M},h} \in (0, \infty)$.

Proof. (i) By Lemma 7, $T^*_{\mathrm{epi},\mathcal{M}}(\mu)^{-1} = \sup_{w \in \Omega_{\mathrm{epi}}} \min_h G_h(w_h)$. For every $w \in \Omega_{\mathrm{epi}}$ and every h, $\min_{h'} G_{h'}(w_{h'}) \le G_h(w_h) \le \sup_{u \in \Sigma_{A_h}} G_h(u) = 1/T^*_h$, so the supremum is at most $\min_h 1/T^*_h$. Conversely, given $\varepsilon > 0$ choose $u_h \in \Sigma_{A_h}$ with $G_h(u_h) \ge 1/T^*_h - \varepsilon$ for each h. Then $w := (u_h)_h \in \Omega_{\mathrm{epi}}$ and $\min_h G_h(w_h) \ge \min_h 1/T^*_h - \varepsilon$. Hence $T^*_{\mathrm{epi},\mathcal{M}}(\mu)^{-1} = \min_h 1/T^*_h$, which means $T^*_{\mathrm{epi},\mathcal{M}}(\mu) = \max_h T^*_h$.

(ii) By Lemma 7, $T^*_{\mathrm{gen},\mathcal{M}}(\mu)^{-1} = \sup_{w \in \Omega_{\mathrm{gen}}} \min_h G_h(w_h)$. For $w \in \Omega_{\mathrm{gen}}$ let $\theta_h := \sum_a w_{h,a}$, so $\theta \in \Sigma_H := \Sigma_{\{1,\dots,H\}}$, the simplex on the stage index set. When $\theta_h > 0$ write $w_h = \theta_h u_h$ with $u_h \in \Sigma_{A_h}$. When $\theta_h = 0$, $G_h(w_h) = G_h(0) = 0$. By Lemma 8, $\min_h G_h(w_h) = \min_h \theta_h G_h(u_h) \le \min_h \theta_h / T^*_h$. We claim that $\sup_{\theta \in \Sigma_H} \min_h \theta_h / T^*_h = 1/\sum_h T^*_h$. Indeed, if $m := \min_h \theta_h / T^*_h$ then $\theta_h \ge m T^*_h$ for all h and $1 = \sum_h \theta_h \ge m \sum_h T^*_h$. Moreover $\theta_h := T^*_h / \sum_{h'} T^*_{h'}$ attains the value $1/\sum_h T^*_h$. So $T^*_{\mathrm{gen},\mathcal{M}}(\mu)^{-1} \le 1/\sum_h T^*_h$. Conversely, given $\varepsilon > 0$ choose u_h with $G_h(u_h) \ge 1/T^*_h - \varepsilon$ and set $w_h := \theta_h u_h$ with $\theta_h = T^*_h / \sum_{h'} T^*_{h'}$. Then $w \in \Omega_{\mathrm{gen}}$ and, by Lemma 8, $\min_h G_h(w_h) = \min_h \theta_h G_h(u_h) \ge \min_h \theta_h (1/T^*_h - \varepsilon) \ge 1/\sum_h T^*_h - \varepsilon$. Hence $T^*_{\mathrm{gen},\mathcal{M}}(\mu) = \sum_h T^*_h$.

(iii) Consider the bandit whose arms are the pairs (h, a), $h \le H$, $a \in A_h$, with observation $N(\mu_h(a), \sigma^2)$ for arm (h, a). An episodic strategy pulls, in episode n, the arms $(1, a_{1,n}), \dots, (H, a_{H,n})$ in this order. It chooses them at the start of the episode as a function of the data of the previous episodes and of internal randomisation (Definition 9). By Lemma 6(i) the rewards are independent draws from the arm distributions. It is therefore a strategy for this bandit as in Definition 5. The time HN is a stopping time for it, because the decision to stop after episode n depends on the data of the first n episodes. It is almost surely finite under every $\mu' \in \mathcal{M} \cap \mathcal{S}^{(H)}$. Let $\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)$ and $\mathcal{G}_\delta := \{\hat{\pi} = \pi^*(\mu)\}$. As in Appendix E.3, δ-correctness under μ and μ' gives $\mathbb{P}_\mu(\mathcal{G}_\delta) \ge 1 - \delta$ and $\mathbb{P}_{\mu'}(\mathcal{G}_\delta) \le \delta$. Lemma 1 of Kaufmann et al. (2016), extended to randomised rules as in Appendix E.3, together with the monotonicity of kl gives

$$\sum_h \sum_{a \in A_h} \mathbb{E}_\mu[N_{h,a}(N)] \, \frac{(\mu_h(a) - \mu'_h(a))^2}{2\sigma^2} \;\ge\; \mathrm{kl}(\delta, 1 - \delta),$$

where $N_{h,a}(N)$ is the number of pulls of (h, a) in the first N episodes. Assume $\mathbb{E}_\mu[N] < \infty$, since otherwise there is nothing to prove, and set $w_{h,a} := \mathbb{E}_\mu[N_{h,a}(N)] / \mathbb{E}_\mu[N]$. Every stage is visited exactly once per episode by Assumption 3(b), so $\sum_a N_{h,a}(N) = N$ for every h. Hence $w_h \in \Sigma_{A_h}$ and $w \in \Omega_{\mathrm{epi}}$. Thus $\mathbb{E}_\mu[N] \, c_w(\mu, \mu') \ge \mathrm{kl}(\delta, 1 - \delta)$ for every $\mu' \in \mathrm{Alt}_{\mathcal{M}}(\mu)$. Taking the infimum

over μ' and bounding it by the supremum over Ω_{epi} gives $\mathbb{E}_\mu[N] \geq T^*_{\mathrm{epi},\mathcal{M}}(\mu)\,\mathrm{kl}(\delta, 1-\delta)$, and (i) concludes.

(iv) Under the generative protocol the strategy is directly a strategy for the same pair-arm bandit with stopping time τ. The same argument gives $\sum_{h,a} \mathbb{E}_\mu[N_{h,a}(\tau)]\,(\mu_h(a) - \mu'_h(a))^2/(2\sigma^2) \geq \mathrm{kl}(\delta, 1-\delta)$. With $w_{h,a} := \mathbb{E}_\mu[N_{h,a}(\tau)]/\mathbb{E}_\mu[\tau] \in \Omega_{\mathrm{gen}}$ this reads $\mathbb{E}_\mu[\tau]\, c_w(\mu, \mu') \geq \mathrm{kl}(\delta, 1-\delta)$, and (ii) concludes. □

E.7 Proof of Lemma 9

Hypotheses. Gaussian bandit $\nu \in \mathbb{R}^K$ with a unique best arm. Constants $\alpha \geq 0$ and $C_r > 0$ with $r(t) \leq C_r t^\alpha$ for all $t \geq 1$. The single-stage strategy of Definition 10 with the Chernoff stopping rule and the threshold $\beta(t, \delta) = \log(r(t)/\delta)$. The strategy runs an initialisation of fixed length $K_0 \geq K$ cycling through the arms, then D-tracking with forced exploration, for all $t \geq 1$.

Proof. We use two facts established in the proof of Theorem 14 of Garivier & Kaufmann (2016), their Appendix D.1, for exponential-family bandits. Gaussian arms with known variance are a special case. Both concern the sampling rule, run for all $t \geq 1$, and neither involves the threshold. They are stated there for the D-tracking rule run from $t = 1$ without a separate initialisation phase. Their proofs use the sampling rule in two ways only. First, through the lower bound $N_a(t) \geq (\sqrt{t} - K/2)_+ - 1$ of their Lemma 17. Its argument uses the initialisation only through the validity of this bound during the initialisation rounds, and the forced-exploration rule at every later round. A cycling initialisation of length K_0 satisfies the bound at every $t \leq K_0$, since $N_a(t) \geq \lfloor t/K \rfloor \geq \sqrt{t} - K/2 - 1$ because $t/K + K/2 \geq \sqrt{2t}$. It leaves the rule unchanged afterwards, so the bound holds for all t. Second, through the D-tracking rule at rounds $t \geq T^{1/4}$. The counts accumulated before round $T^{1/4}$ are at most $T^{1/4}$ in total, whatever the rule. For a fixed K_0 the two facts therefore hold for every T with $T^{1/4} \geq K_0$. This at most enlarges the constant T_ε below to $\max(T_\varepsilon, K_0^4)$. Every run of Section 6 has $K_0 = K$. Fix $\varepsilon \in (0, \varepsilon_0(\nu))$, where $\varepsilon_0(\nu) > 0$ is such that the constant $C^*_\varepsilon(\nu)$ of (F2) below is positive for $\varepsilon < \varepsilon_0(\nu)$. Such a threshold exists because $C^*_\varepsilon(\nu) \to T^*(\nu)^{-1} > 0$ as $\varepsilon \to 0$. Their proof uses the continuity of w^* (their Proposition 6) to choose a product neighbourhood $\mathcal{I}_\varepsilon$ of ν on which $|w^*_a(\nu') - w^*_a(\nu)| \leq \varepsilon$ for every arm a and the best arm is that of ν. For $T \in \mathbb{N}$ it defines the event $\mathcal{E}_T := \bigcap_{T^{1/4} \leq t \leq T} \{\hat{\nu}(t) \in \mathcal{I}_\varepsilon\}$, where $\hat{\nu}(t)$ is the vector of empirical means of the un-stopped process after t pulls.

(F1) By their Lemma 19, there are constants $b_\varepsilon, c_\varepsilon > 0$, depending on ν and ε only, such that $\mathbb{P}_\nu(\mathcal{E}_T^c) \leq b_\varepsilon T \exp(-c_\varepsilon T^{1/8})$ for every T.

(F2) By their Lemma 20 and the display that follows it, there are an integer T_ε, depending on ν, ε and K only, and a constant $C^*_\varepsilon(\nu) > 0$ with $\lim_{\varepsilon \to 0} C^*_\varepsilon(\nu) = T^*(\nu)^{-1}$, with the following property. For every $T \geq T_\varepsilon$, on $\mathcal{E}_T$, the Chernoff statistic of the un-stopped process satisfies $Z(t) \geq t\, C^*_\varepsilon(\nu)$ for every $t \in [\sqrt{T}, T]$. In particular $Z(T) \geq T\, C^*_\varepsilon(\nu)$. Therefore, for every $\delta \in (0,1)$ and every integer $T \geq T_\varepsilon$ with $T\, C^*_\varepsilon(\nu) \geq \beta(T, \delta)$,

$$\mathcal{E}_T \subseteq \{\tau_\delta \leq T\}.$$

Indeed, on $\mathcal{E}_T$ the stopping condition $Z(T) \geq \beta(T, \delta)$ holds at time T, so the stopping rule has stopped by time T. Garivier & Kaufmann (2016) derive the containment from $\min(\tau_\delta, T) \leq \sqrt{T} + \beta(T, \delta)/C^*_\varepsilon(\nu)$ for every $T \geq \max(T_0(\delta), T_\varepsilon)$. Here $T_0(\delta)$ is the smallest T satisfying $\sqrt{T} + \beta(T, \delta)/C^*_\varepsilon(\nu) \leq T$. We use the direct argument above instead, at every T satisfying the weaker inequality $T\, C^*_\varepsilon(\nu) \geq \beta(T, \delta)$. This inequality follows from $\sqrt{T} + \beta(T, \delta)/C^*_\varepsilon(\nu) \leq T$.

Let $C^* := C^*_\varepsilon(\nu)$ and $\ell_\delta := \log(1/\delta)$. The letter L is reserved for the leaf of Example 1.

(F3) A deterministic time after which the inequality of (F2) holds. For $T \geq 1$, $\beta(T, \delta) = \log r(T) + \ell_\delta \leq \alpha \log T + \log C_r + \ell_\delta$. Let $\psi(T) := T - \sqrt{T} - (\alpha \log T + \log C_r)/C^*$ for real $T \geq 1$. Then $\psi'(T) = 1 - 1/(2\sqrt{T}) - \alpha/(C^* T) \geq 1 - 1/4 - 1/2 > 0$ for $T \geq T_1 := \max(4, 2\alpha/C^*)$, so ψ is increasing on $[T_1, \infty)$ and $\psi(T) \to \infty$. Define

$$T_0(\delta) := \min\{T \in \mathbb{N} :\ T \geq T_1,\ \psi(T) \geq \ell_\delta/C^*\},$$

which is finite. For every integer $T \geq T_0(\delta)$, monotonicity gives $\psi(T) \geq \psi(T_0(\delta)) \geq \ell_\delta/C^*$. Written out, this reads $\sqrt{T}+(\alpha \log T+\log C_r+\ell_\delta)/C^* \leq T$, and therefore $\sqrt{T}+\beta(T,\delta)/C^* \leq T$. Moreover, for every $\eta > 0$, with $T_\delta := \lceil T_1 + (1+\eta)\ell_\delta/C^* \rceil$,

$$\psi(T_\delta) - \frac{\ell_\delta}{C^*} \;\geq\; \frac{\eta \ell_\delta}{C^*} + T_1 - \sqrt{T_\delta} - \frac{\alpha \log T_\delta + \log C_r}{C^*} \;\longrightarrow\; +\infty \qquad (\delta \to 0),$$

because $T_\delta = O(\ell_\delta)$ makes $\sqrt{T_\delta} + \alpha \log T_\delta = o(\ell_\delta)$. Hence $T_0(\delta) \leq T_\delta$ for all δ small enough, and $\limsup_{\delta\to 0} T_0(\delta)/\ell_\delta \leq (1+\eta)/C^*$.

Conclusion. Define $T_\varepsilon(\delta) := \max(T_0(\delta), T_\varepsilon)$, $D_\varepsilon := \sum_{T\geq 1} b_\varepsilon T \exp(-c_\varepsilon T^{1/8}) < \infty$ and $\kappa(\varepsilon) := (1+\varepsilon)/(C^*_\varepsilon(\nu)T^*(\nu))$, so that $\kappa(\varepsilon) \to 1$ as $\varepsilon \to 0$ by (F2). By (F3) with $\eta = \varepsilon$, $\limsup_{\delta\to 0} T_\varepsilon(\delta)/\ell_\delta \leq (1+\varepsilon)/C^* = \kappa(\varepsilon)T^*(\nu)$. For every integer $s \geq T_\varepsilon(\delta)$ we have $s \geq T_\varepsilon$ and, by (F3), $\sqrt{s} + \beta(s,\delta)/C^* \leq s$. Thus $s\,C^* \geq \beta(s,\delta)$, and (F2) gives $\{\tau_\delta > s\} \subseteq \mathcal{E}_s^c$. For integer-valued τ_δ and T, $(\tau_\delta - T)^+ = \sum_{s\geq T} \mathbf{1}\{\tau_\delta > s\}$, so by (F1),

$$\mathbb{E}_\nu\big[(\tau_\delta - T_\varepsilon(\delta))^+\big] = \sum_{s\geq T_\varepsilon(\delta)} \mathbb{P}_\nu(\tau_\delta > s) \leq \sum_{s\geq T_\varepsilon(\delta)} \mathbb{P}_\nu(\mathcal{E}_s^c) \leq \sum_{s\geq 1} b_\varepsilon s \exp(-c_\varepsilon s^{1/8}) = D_\varepsilon,$$

uniformly in $\delta \in (0,1)$. Finally $\mathbb{E}_\nu[\tau_\delta] \leq T_\varepsilon(\delta) + D_\varepsilon < \infty$, so $\tau_\delta < \infty$ almost surely. Dividing by ℓ_δ and letting first $\delta \to 0$ and then $\varepsilon \to 0$ gives $\limsup_{\delta\to 0} \mathbb{E}_\nu[\tau_\delta]/\ell_\delta \leq T^*(\nu)$. □

E.8 Proof of Lemma 10

Hypotheses. Gaussian bandit $\nu \in \mathbb{R}^K$, $K \geq 2$, with known variance σ^2 and a unique best arm a^*. A confidence parameter $\delta \in (0,1)$. Any sampling rule of Definition 5 under which every arm is pulled at least once in the first K_0 rounds. The stopping time $\tau = \inf\{t \geq K_0 :\; Z(t) \geq \beta_{\mathrm{an}}(t,\delta;K)\}$ with $\beta_{\mathrm{an}}(t,\delta;K) = \log((K-1)\zeta(2)t^3/\delta)$, and $\hat{a} = \arg\max_a \hat{\mu}_a(\tau)$ on $\{\tau < \infty\}$.

Proof. Stack-of-rewards representation. Let $(X_{a,s})_{a\in A,\, s\geq 1}$ be independent random variables with $X_{a,s} \sim N(\nu_a, \sigma^2)$, independent of the internal randomisation (U_t). Let the s-th pull of arm a return $X_{a,s}$, so $Y_t := X_{a_t, N_{a_t}(t)}$. Under any sampling rule of Definition 5 the process $(a_t, Y_t, U_t)_{t\geq 1}$ so constructed has the law $\mathbb{P}_\nu$. The reason is that conditionally on $\mathcal{F}_{t-1}$ and on $a_t = a$, the variable $X_{a,N_a(t-1)+1}$ has not been observed, is independent of $\mathcal{F}_{t-1}$ and is distributed as $N(\nu_a,\sigma^2)$. To see this, fix $F \in \mathcal{F}_{t-1}$ and $j \geq 0$. The event $F \cap \{a_t = a,\, N_a(t-1) = j\}$ is measurable with respect to $(U_s)_{s\leq t}$, the stacks of the other arms and $X_{a,1},\dots,X_{a,j}$, so it is independent of $X_{a,j+1}$. Summing over j gives the claim. It therefore suffices to prove the bound in this representation. Let $\bar{\mu}_{a,m} := m^{-1}\sum_{s\leq m} X_{a,s}$ denote the mean of the first m draws of arm a, so that $\hat{\mu}_a(t) = \bar{\mu}_{a,N_a(t)}$. The function $t \mapsto \beta_{\mathrm{an}}(t,\delta;K)$ is increasing, and it is positive on $\{t \geq 1\}$ because $(K-1)\zeta(2) \geq \zeta(2) > 1 > \delta$.

Containment. Fix a rival $b \neq a^*$. On $\{\tau < \infty,\, \hat{a} = b\}$ let $t := \tau$, $n_b := N_b(t) \geq 1$ and $n_* := N_{a^*}(t) \geq 1$, so that $n_* + n_b \leq t$. Since $b = \hat{a}(t)$ is the empirical best arm, $\hat{\mu}_b(t) \geq \hat{\mu}_{a^*}(t)$, and the stopping condition gives

$$\frac{\big(\hat{\mu}_b(t) - \hat{\mu}_{a^*}(t)\big)^2}{2\sigma^2(1/n_b + 1/n_*)} \;\geq\; Z(t) \;\geq\; \beta_{\mathrm{an}}(t,\delta;K) \;\geq\; \beta_{\mathrm{an}}(n_*+n_b,\delta;K) \;>\; 0,$$

so $\bar{\mu}_{b,n_b} - \bar{\mu}_{a^*,n_*} \geq \sigma\sqrt{2\beta_{\mathrm{an}}(n_*+n_b,\delta;K)\,(1/n_* + 1/n_b)}$. Therefore

$$\{\tau < \infty,\, \hat{a} = b\} \;\subseteq\; \bigcup_{n_*\geq 1}\bigcup_{n_b\geq 1} E_{b,n_*,n_b},$$

$$E_{b,n_*,n_b} := \Big\{\bar{\mu}_{b,n_b} - \bar{\mu}_{a^*,n_*} \;\geq\; \sigma\sqrt{2\beta_{\mathrm{an}}(n_*+n_b,\delta;K)\,(1/n_* + 1/n_b)}\Big\},$$

where the events E_{b,n_*,n_b} are defined through the stack of rewards alone and do not depend on the sampling rule.

Tail bound. Under $\mathbb{P}_\nu$, $\bar{\mu}_{b,n_b} - \bar{\mu}_{a^*,n_*} \sim N\big(\nu_b - \nu_{a^*},\, \sigma^2(1/n_* + 1/n_b)\big)$ with $\nu_b - \nu_{a^*} < 0$. Hence, with $\beta := \beta_{\mathrm{an}}(n_*+n_b,\delta;K) > 0$,

$$\mathbb{P}_\nu(E_{b,n_*,n_b}) \;\leq\; \mathbb{P}\big(N(0,1) \geq \sqrt{2\beta}\big) \;\leq\; e^{-\beta} \;=\; \frac{\delta}{(K-1)\,\zeta(2)\,(n_*+n_b)^3},$$

by the Gaussian tail bound $\mathbb{P}(N(0,1) \geq x) \leq e^{-x^2/2}$ for $x \geq 0$.

Union. There are $s-1$ pairs (n_*, n_b) of positive integers with $n_* + n_b = s$, so

$$\mathbb{P}_\nu(\tau < \infty,\ \hat{a} \neq a^*) \leq \sum_{b \neq a^*} \sum_{n_* \geq 1} \sum_{n_b \geq 1} \frac{\delta}{(K-1)\zeta(2)(n_* + n_b)^3} = \frac{\delta}{\zeta(2)} \sum_{s \geq 2} \frac{s-1}{s^3}$$
$$= \frac{\delta}{\zeta(2)} \Big(\sum_{s \geq 2} \frac{1}{s^2} - \sum_{s \geq 2} \frac{1}{s^3} \Big) = \frac{\zeta(2) - \zeta(3)}{\zeta(2)}\, \delta,$$

which is less than δ because $\zeta(3) > 0$. Numerically, $(\zeta(2) - \zeta(3))/\zeta(2) \approx 0.27$. □

E.9 Proof of Theorem 3

Hypotheses. Assumption 3 and $\mu \in \mathcal{M}_{\mathrm{G}}^{(H)} \cap \mathcal{S}^{(H)}$. The threshold (4) with $r_h(t) = (K_h - 1)\zeta(2)t^3$, at the per-stage confidence δ/H, so $\beta_h(t, \delta/H) = \log(H(K_h - 1)\zeta(2)t^3/\delta)$. Initialisation length K_0 as in Definition 10, which is K_h in the generative variant and K in the episodic one.

Proof. Fix h. Since $\mu_h \in \mathcal{M}_{\mathrm{G},h} \cap \mathcal{S}_{A_h}$, the best arm a_h^* of μ_h is an ancestor arm, because branches have mean $\mu_{h,0} < \max_{a \in A_h^{\mathrm{anc}}} \mu_h(a)$. So the ancestor bandit $\nu_h := (\mu_h(a))_{a \in A_h^{\mathrm{anc}}}$ has the unique best arm a_h^*, and $T^*_{\mathbb{R}^{K_h}}(\nu_h) = T^*_{\mathrm{anc}}(\mu_h)$ by definition. In both variants the stage-h run is exactly the single-stage strategy of Definition 10 on ν_h, on its own clock, with initialisation length K_0 and confidence parameter $\delta' := \delta/H \in (0,1)$. Its threshold is therefore $\beta_h(t, \delta') = \beta_{\mathrm{an}}(t, \delta'; K_h) = \log(H(K_h - 1)\zeta(2)t^3/\delta)$. By Lemma 6(i) its observations are independent draws from the arm distributions of ν_h. Gaussian arms with known variance form a one-parameter exponential family, so the results of Garivier & Kaufmann (2016) quoted in Appendix E.7 apply. Lemma 9 applies to this run with $r(t) = (K_h - 1)\zeta(2)t^3$, $\alpha = 3$, $C_r = (K_h - 1)\zeta(2)$ and confidence parameter δ'. Let $T_\varepsilon^h(\delta')$, D_ε^h and $\kappa_h(\varepsilon)$ denote the quantities it provides.

(i) Lemma 10, applied to ν_h with $K = K_h$ and confidence parameter $\delta' = \delta/H$, gives $\mathbb{P}_\mu(\tau_h < \infty, \hat{a}_h \neq a_h^*) \leq \delta/H$ for any sampling rule, and Lemma 9 gives $\mathbb{P}_\mu(\tau_h < \infty) = 1$. Hence $\mathbb{P}_\mu(\hat{a}_h \neq a_h^*) \leq \delta/H$ for every h, and by a union bound $\mathbb{P}_\mu(\hat{\pi} \neq \pi^*(\mu)) \leq \sum_h \mathbb{P}_\mu(\hat{a}_h \neq a_h^*) \leq \delta$. The stopping time, $\sum_h \tau_h$ or $\max_h \tau_h$, is almost surely finite. The argument holds for every $\mu \in \mathcal{M}_{\mathrm{G}}^{(H)} \cap \mathcal{S}^{(H)}$, which is δ-correctness relative to $\mathcal{M}_{\mathrm{G}}^{(H)}$.

(ii) $\mathbb{E}_\mu[\tau] = \sum_h \mathbb{E}_\mu[\tau_h]$, and by Lemma 9, $\mathbb{E}_\mu[\tau_h] \leq T_\varepsilon^h(\delta/H) + D_\varepsilon^h$. Since $\log(H/\delta)/\log(1/\delta) \to 1$, $\limsup_{\delta \to 0} T_\varepsilon^h(\delta/H)/\log(1/\delta) \leq \kappa_h(\varepsilon) T^*_{\mathrm{anc}}(\mu_h)$. So $\limsup_{\delta \to 0} \mathbb{E}_\mu[\tau_h]/\log(1/\delta) \leq \kappa_h(\varepsilon) T^*_{\mathrm{anc}}(\mu_h)$ for every $\varepsilon \in (0, \min_h \varepsilon_0(\nu_h))$. Letting $\varepsilon \to 0$ and summing over h gives (ii). Theorem 14 of Garivier & Kaufmann (2016) requires $\alpha \in [1, e/2]$ and does not cover $\alpha = 3$. Lemma 9 is used instead, as explained in Remark 8.

(iii) Fix $\varepsilon \in (0, \min_h \varepsilon_0(\nu_h))$. For every h, $\tau_h \leq T_\varepsilon^h(\delta/H) + (\tau_h - T_\varepsilon^h(\delta/H))^+$, so

$$N = \max_h \tau_h \leq \max_h T_\varepsilon^h(\delta/H) + \sum_h \big(\tau_h - T_\varepsilon^h(\delta/H)\big)^+ \quad \text{and} \quad \mathbb{E}_\mu[N] \leq \max_h T_\varepsilon^h(\delta/H) + \sum_h D_\varepsilon^h.$$

Dividing by $\log(1/\delta)$ and using the first property in Lemma 9 as in (ii), $\limsup_{\delta \to 0} \mathbb{E}_\mu[N]/\log(1/\delta) \leq \max_h \kappa_h(\varepsilon) T^*_{\mathrm{anc}}(\mu_h)$. Letting $\varepsilon \to 0$ gives (iii).

The blind version is the same argument with $\nu_h = \mu_h$ on A_h, $K_h = |A_h|$, and $\mathcal{M}_{\mathrm{blind}}^{(H)}$. The last claim of the theorem is Lemma 3(i). □

E.10 Proof of Lemma 11

Hypotheses. M as in Definition 11 with a unique optimal policy, a class $\mathcal{M}$ of models with the same S, (A_s), (B_s), p, and a strategy as in the statement. A confidence parameter $\delta \in (0, 1/2)$.

Proof. If $\mathrm{Alt}_{\mathcal{M}}(M) = \emptyset$ or $\mathbb{E}_M[N] = \infty$ there is nothing to prove. Let $M' \in \mathrm{Alt}_{\mathcal{M}}(M)$ and $\mathcal{G}_\delta := \{\hat{\pi} = \pi_M^*\}$. Then δ-correctness under M and M' gives $N < \infty$ almost surely under both, $\mathbb{P}_M(\mathcal{G}_\delta) \geq 1 - \delta$ and $\mathbb{P}_{M'}(\mathcal{G}_\delta) \leq \delta$. Consider the bandit whose arms are the pairs (s, a), $s \in S$,

$a \in A_s$, with Gaussian observations. Within an episode the strategy pulls, at each step, an arm (s_t, a_t) whose state component is the current state. The current state is a function of the previous state-action pair and of an independent draw from $p(\cdot \mid s_{t-1}, a_{t-1})$. The law of this draw is the same under M and M' because p is common to the class. Concretely, generate $s_t = \Gamma(s_{t-1}, a_{t-1}, U_{t-1})$, where $\Gamma(s, a, \cdot)$ is the quantile map of $p(\cdot \mid s, a)$ and (U_t) are the i.i.d. uniforms of Definition 5. The strategy is then a randomised sampling rule for this bandit as in Definition 5. Since Γ does not depend on the means, the likelihood ratio between M and M' on $\mathcal{F}_{HN}$ depends on the rewards only, as in Appendix E.3. Moreover HN is an almost surely finite stopping time. Lemma 1 of Kaufmann et al. (2016), extended to randomised rules as in Appendix E.3, and the monotonicity of kl give

$$\sum_{s,a} \mathbb{E}_M[N_{s,a}(N)] \, \frac{(\mu(s,a) - \mu'(s,a))^2}{2\sigma^2} \;\geq\; \mathrm{kl}(\delta, 1-\delta),$$

where $N_{s,a}(N)$ is the number of visits to (s, a) in the first N episodes. Set $w(s,a) := \mathbb{E}_M[N_{s,a}(N)]/\mathbb{E}_M[N]$. We check $w \in \Omega(M)$. Every episode visits s_{root} once, so $\sum_a w(s_{\mathrm{root}}, a) = 1$. For a state s' in layer $h+1 \geq 2$, the number of visits to s' in episode n is $\sum_{(s,a) \text{ in layer } h} \mathbf{1}\{(s,a) \text{ visited in episode } n\}\mathbf{1}\{\text{next state} = s'\}$. The conditional expectation of each summand given the history up to and including the visit to (s,a) is $p(s' \mid s, a)\mathbf{1}\{(s,a) \text{ visited in episode } n\}$, because the transition is drawn from $p(\cdot \mid s, a)$ independently of the past. The event $\{n \leq N\}$ is determined by the data of the first $n-1$ episodes. Summing over $n \leq N$ and taking expectations gives $\mathbb{E}_M[\sum_a N_{s',a}(N)] = \sum_{s,a} p(s' \mid s, a)\mathbb{E}_M[N_{s,a}(N)]$, which is the flow constraint at s'. The sums are finite because $\mathbb{E}_M[N] < \infty$. Hence $w \in \Omega(M)$ and $\mathbb{E}_M[N]\, c_w(M, M') \geq \mathrm{kl}(\delta, 1-\delta)$ for every $M' \in \mathrm{Alt}_{\mathcal{M}}(M)$. Taking the infimum over M' and bounding by the supremum over $\Omega(M)$ gives the claim. □

E.11 Proof of Proposition 2

Hypotheses. M as in Definition 11 with a unique optimal policy and Assumption 4 ($\pi^*_M(s') \notin B_{s'}$ at every decision state s'). The state $s \in S_{\mathrm{dec}}$ has $k = |B_s| \geq 1$ branch actions sharing $(\mu_0(s), p_0(\cdot \mid s))$, and therefore sharing $Q_M(s,j) = Q_0(s)$. In (i), $\delta \in (0, 1/2)$.

Proof. Since $\pi^*_M(s)$ is the unique maximiser of $Q_M(s,\cdot)$ and is not a branch, $\Delta_Q(s) = Q_M(s, \pi^*_M(s)) - Q_0(s) > 0$.

(i) Fix $w \in \Omega(M)$, $j \in B_s$ and $\eta > 0$. Let $M'_{j,\eta}$ be the model obtained from M by replacing $\mu(s,j)$ with $\mu(s,j) + \Delta_Q(s) + \eta$, with all other means unchanged. Then $M'_{j,\eta} \in \mathcal{M}_{\mathrm{blind}}(M)$. Values at states in layers after that of s are unchanged, so $Q_{M'}(s,a) = Q_M(s,a)$ for $a \neq j$ and $Q_{M'}(s,j) = Q_0(s) + \Delta_Q(s) + \eta > Q_M(s, \pi^*_M(s)) \geq Q_M(s,a)$ for all $a \neq j$. Thus j is the unique optimal action at s in $M'_{j,\eta}$, and $V_{M'}(s) = V_M(s) + \eta$. At states in earlier layers the Q-values of $M'_{j,\eta}$ are continuous functions of $\eta \geq 0$ that coincide with those of M at $\eta = 0$. Since M has a unique optimal policy, every argmax at those states is unchanged for all sufficiently small $\eta > 0$. At states in the same or later layers other than s nothing changes. Hence for all sufficiently small $\eta > 0$, $M'_{j,\eta}$ has a unique optimal policy with $\pi^*_{M'}(s) = j \neq \pi^*_M(s)$, so $M'_{j,\eta} \in \mathrm{Alt}_{\mathcal{M}_{\mathrm{blind}}(M)}(M)$ and $c_w(M, M'_{j,\eta}) = w(s,j)(\Delta_Q(s) + \eta)^2/(2\sigma^2)$. The total occupancy of a state is its visit probability, so $\sum_{j \in B_s} w(s,j) \leq \sum_a w(s,a) \leq 1$, and some $j \in B_s$ has $w(s,j) \leq 1/k$. Letting $\eta \downarrow 0$, $\inf_{M' \in \mathrm{Alt}_{\mathcal{M}_{\mathrm{blind}}(M)}(M)} c_w(M, M') \leq \Delta_Q(s)^2/(2\sigma^2 k)$. As w was arbitrary, $T^*_{\mathcal{M}_{\mathrm{blind}}(M)}(M)^{-1} \leq \Delta_Q(s)^2/(2\sigma^2 k)$, and Lemma 11 gives the sample-complexity bound.

(ii) Let M_k denote the model of the statement, with branch set B_s of size k at s, and fix $j_1 \in B_s$. Let M_1 be the model obtained from M_k by deleting the actions $B_s \setminus \{j_1\}$, with all other data unchanged. For a model M' with the state-action structure of M_k let $\Phi(M')$ be the model with the structure of M_1 obtained by deleting the same actions. *Alternatives.* For $M' \in \mathcal{M}_{\mathrm{G}}(M_k)$ all $j \in B_s$ have $Q_{M'}(s,j) = \mu_0(s) + \sum_{s'} p_0(s' \mid s) V_{M'}(s')$, because values at later layers do not involve the actions at s. So $V_{M'}(s) = V_{\Phi(M')}(s)$, and M' and $\Phi(M')$ have the same values and Q-values at every state. If M' has a unique optimal policy then $\pi^*_{M'}(s) \notin B_s$ by the class constraint, so $\Phi(M')$ has the unique optimal policy $\pi^*_{M'}$. Conversely, if $\Phi(M') \in \mathcal{M}_{\mathrm{G}}(M_1)$ has a unique optimal policy then $\pi^*_{\Phi(M')}(s) \neq j_1$ by the class constraint of $\mathcal{M}_{\mathrm{G}}(M_1)$, so M' has the same unique optimal policy. Therefore Φ restricts to a bijection from $\mathrm{Alt}_{\mathcal{M}_{\mathrm{G}}(M_k)}(M_k)$ onto $\mathrm{Alt}_{\mathcal{M}_{\mathrm{G}}(M_1)}(M_1)$. Its inverse

assigns to the deleted actions the mean $\mu_0(s)$ and the kernel $p_0(\cdot \mid s)$. *Costs.* For $M' \in \mathcal{M}_{\mathrm{G}}(M_k)$, $\mu'(s,j) = \mu_0(s) = \mu(s,j)$ for $j \in B_s$, so $c_w(M_k, M') = \sum_{(s',a)\text{ non-branch}} w(s',a)(\mu(s',a) - \mu'(s',a))^2/(2\sigma^2) = c_{\bar{w}}(M_1, \Phi(M'))$ for every $\bar{w}$ that agrees with w on the non-branch coordinates. *Feasible sets.* Define $\Psi : \Omega(M_k) \to \mathbb{R}_+^{\mathcal{A}}$ by $\Psi(w)(s,j_1) := \sum_{j\in B_s} w(s,j)$ and $\Psi(w)(s',a) := w(s',a)$ for all other coordinates. The flow constraints of $\Omega(M_k)$ involve the branch coordinates at s only through two sums. The constraint at s involves $\sum_{j\in B_s} w(s,j)$, and the constraints at the successors s' involve $\sum_{j\in B_s} w(s,j)\, p_0(s' \mid s)$, because all branches share $p_0(\cdot \mid s)$. Hence $\Psi(w) \in \Omega(M_1)$. The map Ψ is onto $\Omega(M_1)$, since given $\bar{w} \in \Omega(M_1)$ one can split $\bar{w}(s,j_1)$ equally among B_s. Combining the three items,

$$\sup_{w\in\Omega(M_k)} \inf_{M'\in \mathrm{Alt}_{\mathcal{M}_{\mathrm{G}}(M_k)}(M_k)} c_w(M_k, M') = \sup_{\bar{w}\in\Omega(M_1)} \inf_{\bar{M}\in \mathrm{Alt}_{\mathcal{M}_{\mathrm{G}}(M_1)}(M_1)} c_{\bar{w}}(M_1, \bar{M}),$$

so $T^*_{\mathcal{M}_{\mathrm{G}}(M_k)}(M_k) = T^*_{\mathcal{M}_{\mathrm{G}}(M_1)}(M_1)$ for every $k \geq 1$, and the right side depends only on the data listed in the statement. □

E.12 Proof of Proposition 3

Hypotheses. M greedy-aligned with a unique optimal policy and satisfying Assumption 4, with $\mathcal{M} \in \{\mathcal{M}_{\mathrm{G}}(M), \mathcal{M}_{\mathrm{blind}}(M)\}$ and $\mathcal{M}^{\mathrm{al}}$ the greedy-aligned members of $\mathcal{M}$.

Proof. Second inequality. $\mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M) \subseteq \mathrm{Alt}_{\mathcal{M}}(M)$, so for every w the infimum over $\mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)$ is at least the infimum over $\mathrm{Alt}_{\mathcal{M}}(M)$. Taking suprema, $T^*_{\mathcal{M}^{\mathrm{al}}}(M)^{-1} \geq T^*_{\mathcal{M}}(M)^{-1}$.

First inequality. Fix $w \in \Omega(M)$ and $M' \in \mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)$. Both M and M' are greedy-aligned, so $\pi^\circ_{M'} = \pi^*_{M'} \neq \pi^*_M = \pi^\circ_M$, and there is a decision state s with $a := \pi^\circ_{M'}(s) \neq \pi^\circ_M(s) =: a^\circ$. Then $\mu'(s,a) > \mu'(s,a^\circ)$. If $\mathcal{M} = \mathcal{M}_{\mathrm{G}}(M)$, then $a \notin B_s$ because $a = \pi^*_{M'}(s)$ and the class excludes optimal branches, so $(s,a) \in R_{\mathrm{G}}$. In the blind case $(s,a) \in R_{\mathrm{blind}}$. Now $c_w(M, M') \geq \big[w(s,a^\circ)(\mu(s,a^\circ) - \mu'(s,a^\circ))^2 + w(s,a)(\mu(s,a) - \mu'(s,a))^2\big]/(2\sigma^2)$. The minimum of the right side over $\{\mu'(s,a) \geq \mu'(s,a^\circ)\}$ is $g_{s,a}(w)$ by the two-coordinate computation in Appendix E.1, with $\mu(s,a^\circ) > \mu(s,a)$ in the role of $\mu_{a^*} > \mu_a$. The value is $0 = g_{s,a}(w)$ if $w(s,a^\circ)w(s,a) = 0$. Hence $c_w(M, M') \geq \min_{(s,a)\in R_{\mathcal{M}}} g_{s,a}(w)$ for every $M' \in \mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)$, so $\inf_{M'\in\mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)} c_w(M, M') \geq \min_{(s,a)\in R_{\mathcal{M}}} g_{s,a}(w)$ for every w. Taking suprema over $w \in \Omega(M)$ gives $T^*_{\mathcal{M}^{\mathrm{al}}}(M)^{-1} \geq T^*_{\mathrm{sf},\mathcal{M}}(M)^{-1}$, which means $T^*_{\mathcal{M}^{\mathrm{al}}}(M) \leq T^*_{\mathrm{sf},\mathcal{M}}(M)$. □

E.13 Computations for Example 1

Let $q := w(s_{\mathrm{root}}, 0)$, so $w(s_{\mathrm{root}}, 1) = 1 - q$, $w(L, \cdot) = q$ and $w(R, \cdot) = 1 - q$ for every $w \in \Omega(M)$. This parametrises $\Omega(M)$ by $q \in [0,1]$. Let $\mu = (\mu_1, \mu_2, \mu_3, \mu_4) := (\mu(s_{\mathrm{root}}, 0), \mu(L), \mu(s_{\mathrm{root}}, 1), \mu(R))$ with weights $(q, q, 1-q, 1-q)$, and let $d_i := \mu'_i - \mu_i$ for an alternative M'. In all three instances $\mu_1 - \mu_3 = 0.5$, so $\pi^\circ_M = 0$ and the single-flip program has the single rival $(s_{\mathrm{root}}, 1)$ with $g(w) = 0.5^2 q(1-q)/(2\sigma^2)$. This is maximised at $q = 1/2$ with value $1/32$, so $T^*_{\mathrm{sf}}(M) = 32$.

Coupled program. Suppose $\pi^*_M = 1$, so that $\Delta_{\mathrm{root}} = \mu_3 + \mu_4 - \mu_1 - \mu_2 > 0$. The case $\pi^*_M = 0$ is symmetric. A model M' has a unique optimal policy $\neq \pi^*_M$ if and only if $\mu'_1 + \mu'_2 > \mu'_3 + \mu'_4$. The set of such M' is dense in its closure $\{d_1 + d_2 - d_3 - d_4 \geq \Delta_{\mathrm{root}}\}$, and c_w is continuous, so the inner infimum is the minimum of $\sum_i w_i d_i^2/(2\sigma^2)$ subject to $\sum_i c_i d_i \geq \Delta_{\mathrm{root}}$ with $c = (1, 1, -1, -1)$. If some $w_i = 0$ the minimum is 0. Otherwise, by the Cauchy–Schwarz inequality, $\min \sum_i w_i d_i^2$ subject to $\sum_i c_i d_i = \Delta_{\mathrm{root}}$ equals $\Delta^2_{\mathrm{root}}/\sum_i (c_i^2/w_i) = \Delta^2_{\mathrm{root}}/\sum_i (1/w_i)$, with equality at $d_i = \Delta_{\mathrm{root}}\, c_i/(w_i \sum_j (1/w_j))$. The constraint is active at the minimum. With the weights above, $\sum_i 1/w_i = 2/q + 2/(1-q)$, so the inner infimum is $\Delta^2_{\mathrm{root}}\, q(1-q)/(4\sigma^2)$, maximised at $q = 1/2$ with value $\Delta^2_{\mathrm{root}}/(16\sigma^2)$. Hence $T^*_{\mathcal{M}}(M) = 16\sigma^2/\Delta^2_{\mathrm{root}}$. In instance (a), $\Delta_{\mathrm{root}} = 1.0645 - 1.0 = 0.0645$ and $16/0.0645^2 = 3845.9$. In instance (b), $\Delta_{\mathrm{root}} = 1.5 - 1.4 = 0.1$ and $16/0.1^2 = 1600$. In instance (c), $\Delta_{\mathrm{root}} = 1.9 - 0.5 = 1.4$ and $16/1.4^2 = 8.163$.

Aligned class, instances (b) and (c). Here $\pi^*_M = \pi^\circ_M = 0$. A model M' belongs to $\mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)$ if and only if it is greedy-aligned with unique optimal policy 1. This means $\mu'_3 > \mu'_1$, so the local argmax is 1, and $\mu'_3 + \mu'_4 > \mu'_1 + \mu'_2$, so policy 1 is optimal. The leaves have one action each. In

the coordinates d the two conditions read $d_3 - d_1 > 0.5$ and $d_3 + d_4 - d_1 - d_2 > \Delta_{\mathrm{root}}$. They define an open convex set whose closure is the set described by the weak inequalities, so the inner infimum is the minimum of $\sum_i w_i d_i^2/(2\sigma^2)$ over the closed set. We use two candidate minimisers. The *local-flip point* $d^{\mathrm{loc}} := (-(1-q)/2,\, 0,\, q/2,\, 0)$ minimises $\sum_i w_i d_i^2$ subject to $d_3 - d_1 \geq 0.5$ alone, with value $q(1-q)/4$ and cost $g(w) = q(1-q)/(8\sigma^2)$. It is the two-coordinate minimiser of Appendix E.1. The *value-flip point* $d^{\mathrm{val}} := \Delta_{\mathrm{root}}\, q(1-q)\,(-1/(2q), -1/(2q), 1/(2(1-q)), 1/(2(1-q)))$ is the Cauchy–Schwarz minimiser above. It minimises $\sum_i w_i d_i^2$ subject to $d_3 + d_4 - d_1 - d_2 \geq \Delta_{\mathrm{root}}$ alone, with cost $\Delta_{\mathrm{root}}^2 q(1-q)/(4\sigma^2)$, and satisfies $d_3^{\mathrm{val}} - d_1^{\mathrm{val}} = \Delta_{\mathrm{root}}/2$ for every $q \in (0,1)$.

(b) $\Delta_{\mathrm{root}} = 0.1$. At d^{loc}, $d_3 + d_4 - d_1 - d_2 = q/2 + (1-q)/2 = 1/2 \geq 0.1$, so d^{loc} satisfies both constraints and the constrained minimum is at most $g(w)$. The proof of Proposition 3 shows that the inner infimum over $\mathrm{Alt}_{\mathcal{M}^{\mathrm{al}}}(M)$ is at least $g(w)$ for every w, so it equals $g(w)$. Hence $T^*_{\mathcal{M}^{\mathrm{al}}}(M) = T^*_{\mathrm{sf}}(M) = 32$.

(c) $\Delta_{\mathrm{root}} = 1.4$. At d^{val}, $d_3 - d_1 = \Delta_{\mathrm{root}}/2 = 0.7 \geq 0.5$, so d^{val} satisfies both constraints. Since it minimises the cost under the value constraint alone, it minimises it under both, and the aligned inner infimum equals the coupled inner infimum $\Delta_{\mathrm{root}}^2 q(1-q)/(4\sigma^2)$ for every q. Hence $T^*_{\mathcal{M}^{\mathrm{al}}}(M) = T^*_{\mathcal{M}}(M) = 16/1.4^2 = 8.163$, whereas $T^*_{\mathrm{sf}}(M) = 32$.

F Numerical witnesses and reproducibility

All code is available at `https://anonymous.4open.science/r/iats-crl-AC81`. The experiment codes are those of the run protocol and of the result files. E1 computes exact characteristic times, E2 tests correctness and the rate, E3 measures the realized speed-up, E4 varies the horizon, and E5 is the coupled-case study. One command, `experiments/paper/run.py`, reruns the confirmatory family F-IATS (E1–E4) with the pre-specified checks from fixed seeds. It writes the aggregate arrays `results/paper/arrays.npz` and the summary `results/paper/summary.json`. A second command (`experiments/paper/figures.py`) renders Figures 2, 3 and 4 from that cache. A third (`experiments/sweeps/general_mdp.py`) produces the E5 witnesses (`results/paper/e5_witnesses.json`). The summary records the threshold (`analysed`), the allocation-cache periods (20 for E2–E3, 50 for E4), the package version and the commit (`4965fca`). The file `arrays.npz` stores the per-seed stopping times of every E2 and E3 cell, under both thresholds for E2. It also stores the per-seed episode counts and per-stage stop episodes of E4. The maxima quoted in Section 5 are read from it. The script `figures.py` refuses a results cache produced before the threshold change. The released run of E1–E4 with 300 seeds took about one hour on one CPU core. The wall-clock time is not recorded in the artefact files. The figure titles are generated by `figures.py` and name the plotted quantities.

Witness instance Layered-Chain. Layered-Chain is the $H = 2$ instance of Definition 8 with one decision per stage, ancestor means $(1.0, 0.5)$ at stage 1 and $(1.0, 0.7)$ at stage 2. The released implementation of Definition 6 (bisection on the value of the program of Lemma 2(a)) gives $T^*((1.0, 0.5)) = 32.0$ and $T^*((1.0, 0.7)) = 88.89$. The episodic composition is $88.89 = \max_h T_h^*$ and the generative composition is $120.89 = \sum_h T_h^*$ (Theorem 2). For the blind stage $(1.0, 0.5, 0.5, 0.5, 0.5)$ the maximiser of Lemma 2 is $w^* = (0.333, 0.167, 0.167, 0.167, 0.167)$, $T^* = 72$, and every rival term equals $h_a(w^*) = 1/72$ (Lemma 4(c)). The heterogeneous example of Remark 9 uses the same routine.

E5 witnesses. The Frank–Wolfe solver over the occupancy polytope is deterministic. Its linear-maximisation oracle is one backward-induction pass and its step is chosen by a golden-section line search on a smoothed objective. The committed artefact records four things. The first is the rectangular and gated ratios of Appendix D.3. The second is the static-allocation run, with 60 seeds, $\delta = 0.01$, no errors and realized speed-up 2.62 with interval $[2.37, 2.90]$. The third is the concavity check, with 0 violations in 30,000 random pairs. The fourth is the misaligned instance of Example 1(a), with $T^*_{\mathrm{sf}} = 32.0$, $T^*_{\mathcal{M}}(M) = 3845.9$ and factor 120.2. The aligned instances of Example 1(b) and (c) are obtained with the same routines, `char_time_polytope` and `value_gap_char_time`. They give $T^*_{\mathrm{sf}} = 32.0$ for both, $T^*_{\mathcal{M}}(M) = 1600.0$ for (b) and

Table 5: Pre-specified acceptance criteria (F-IATS) and outcomes on the released run (threshold β_{an}, with 300 seeds for E2–E3 and 40 seeds for the E4 simulation). E1 computes exact characteristic times, E2 tests correctness and the rate, E3 measures the realized speed-up and E4 varies the horizon. All criteria are met except the E2 rate-stability check, which fails by construction under β_{an} (Section 6, correctness and rate check). The observed value of each check is given in parentheses. Here T^*, $\bar{\tau}$ and $\bar{N}_H$ are as in Section 6, and $\mathrm{ptp}(x) := \max x - \min x$ is the peak-to-peak range of a finite sequence x.

Exp.	Check	What it verifies	Threshold	Outcome
E1	Graph-aware time flat	T^*_{G} is constant in the branch count k	$\mathrm{ptp}(T^*_{\mathrm{G}}) < 10^{-9}$	met
E1	Blind-time floor	blind time satisfies the lower bound of Theorem 1(i)	$T^*_{\mathrm{blind}} \geq \frac{2\sigma^2}{\Delta_0^2} k$	met
E1	Ratio grows	separation ratio grows with k	$\mathrm{ratio}_{64} > 3\,\mathrm{ratio}_1$	met (7.36)
E1	Positive ratio slope	ratio increases in k	slope > 0	met (0.11)
E2	Empirical δ-correctness (8 cells)	Clopper–Pearson error bound, each learner and δ	CP $\leq \delta$ (else 0 errors, n-limited)	met (0 errors in every cell, CP 9.9×10^{-3})
E2	Rate stability	$\bar{\tau}/(T^* \log(1/\delta))$ at the two smallest δ close	$\lvert\Delta\,\mathrm{rate}\rvert < 0.1$	not met (0.71)
E2	Small forced exploration	forced-exploration pulls a small fraction	fraction < 0.2	met (0.040, blind learner, $\delta = 10^{-2}$)
E3	Speed-up excludes 1 (4 cells)	realized speed-up significant (k=2, 8, 16, 32)	bootstrap $\mathrm{CI}_{\mathrm{lo}} > 1$	met (1.36, 2.36, 3.69, 6.51)
E3	Realized/theory envelope (4 cells)	realized speed-up close to the leading-order ratio (k=2, 8, 16, 32)	realized/theory $\in [0.7, 1.6]$	met (1.15, 1.26, 1.33, 1.43)
E3	Holm rejections (5 hypotheses)	hypotheses H1–H5 rejected	reject at level 0.05	met
E3	Control near 1	permuted-arm control shows no effect	ratio $\in [0.9, 1.1]$	met (1.00)
E3	Control not significant	control speed-up not significant	p-value > 0.05	met (0.996)
E4	Episodic ratio flat	episodic separation independent of horizon H	$\mathrm{ptp} < 10^{-9}$	met
E4	Generative gap linear	additive composition in H	$R^2 > 0.99$	met (1.0)
E4	Sub-linear growth	episodic count sub-linear in H	$\bar{N}_H/\bar{N}_1 < 0.5H$	met ($1.57 < 4$)
E4	Growth CI sub-linear	growth CI below the additive line	$\mathrm{CI}_{\mathrm{hi}} < H$	met ($1.80 < 8$)

8.163 for (c). These values, like the heterogeneous-stage value 818.90 of E4, are computed with the released routines and are not stored in the artefact files. The aligned-class values of Appendix E.13 are not computed by the released code.

F.1 Acceptance criteria pre-specified in the run protocol

Table 5 lists the acceptance criteria of the confirmatory family F-IATS (E1–E4). They were pre-specified in the run protocol, so the confirmatory conclusions cannot be selected after seeing the results. E5 is an illustrative study and is not in this family. All 33 criteria are evaluated by the released code, and 32 are met on the released run. The one exception is the E2 rate-stability check. It compares $\bar{\tau}/(T^* \log(1/\delta))$ at $\delta = 10^{-3}$ and 10^{-4}, and under the threshold β_{an} that ratio is not flat at finite δ (Section 6, correctness and rate check). The per-cell table of all checks, with observed values, is written by the released code to `results/paper/summary.json`.