HAL open science

A New Theoretical Framework for Fast and Accurate Online Decision-Making

Nicolò Cesa-Bianchi, Tommaso R Cesari, Yishay Mansour, Vianney Perchet

To cite this version:

Nicolò Cesa-Bianchi, Tommaso R Cesari, Yishay Mansour, Vianney Perchet. A New Theoretical Framework for Fast and Accurate Online Decision-Making. 2021. hal-02976864v3

HAL Id: hal-02976864

https://hal.science/hal-02976864v3

Preprint submitted on 30 Jun 2021 (v3), last revised 22 Dec 2021 (v4)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A New Theoretical Framework for Fast and Accurate Online Decision-Making

Nicolò Cesa-Bianchi
Università degli Studi di Milano & DSRC

Tommaso Cesari
Toulouse School of Economics (TSE) & Artificial and Natural Intelligence Toulouse Institute (ANITI)

Yishay Mansour
Tel Aviv University & Google research

Vianney Perchet
CREST, ENSAE & Criteo AI Lab, Paris

June 30, 2021

Abstract

We introduce a novel theoretical framework for Return On Investment (ROI) maximization in repeated decision-making. Our setting is motivated by the use case of companies that regularly receive proposals for technological innovations and want to quickly decide whether they are worth implementing. We design an algorithm for learning ROI-maximizing decision-making policies over a sequence of innovation proposals. Our algorithm provably converges to an optimal policy in class Π at a rate of order $\min\left\{1/(N\Delta^2), N^{-1/3}\right\}$, where N is the number of innovations and Δ is the suboptimality gap in Π. A significant hurdle of our formulation, which sets it aside from other online learning problems such as bandits, is that running a policy does not provide an unbiased estimate of its performance.

1 Introduction

Often, companies have to make yes/no decisions, such as whether to adopt a new technology or retire an old product. However, finding out the best option in all circumstances could mean spending too much time or money in the evaluation process. If the decisions to make are many, one could be better off making more of them quickly and inexpensively, provided that these decisions have an overall positive effect. In this paper, we investigate the problem of determining a decision policy to balance the reward over cost ratio optimally (i.e., to maximize the return on investment).

A motivating example. Consider a technology company that keeps testing innovations to increase some chosen metric (e.g., benefits, gross revenue, revenue excluding the traffic acquisition cost). Before deploying an innovation, the company wants to figure out whether it is profitable. As long as each innovation can be tested on i.i.d. samples of users, the company can perform randomized tests and make statistically sound decisions. However, there is an incentive to make these tests run as quickly as possible because, for example, the testing process is expensive. Another reason could be that keeping a team on a project that has negative, neutral, or even borderline positive potential prevents it from testing other ideas that might lead to a significantly better improvement. In other words, it is crucial to learn when to drop barely positive innovations in favor of highly positive ones, so to increase the overall flow of improvement over time (i.e., the ROI of the tests).

More generally, our framework describes problems where an agent faces a sequence of decision tasks consisting of either accepting or rejecting an innovation. Before making each decision, the agent can invest resources into reducing the uncertainty on the value brought by the innovation. The global objective is to maximize the total ROI. Namely, the ratio between the total value accumulated by accepting innovations and the total cost. For an in-depth discussion on alternative goals, we refer the reader to the Appendix (Section D).

The model. Each task n in the sequence is associated with a pair $(\mu_n, \mathcal{D}_n)$ that the learner can *never* directly observe.

- μ_n is a random variable representing the (possibly negative) true value of the n-th innovation.
- $\mathcal{D}_n$ is a probability distribution over the real numbers with expectation μ_n, modeling the feedback on the n-th innovation that the learner can gather from testing (see below).

During the n-th task, the learner can draw arbitrarily many i.i.d. samples $X_{n,1}, X_{n,2}, \ldots$ from $\mathcal{D}_n$, accumulating information on the unknown value μ_n of the innovation currently being tested. After stopping drawing samples, the learner can decide to either accept the innovation, earning μ_n as a reward, or reject it and gain nothing instead. We measure the agent performance during N tasks as the (expected) total amount of value accumulated by accepting innovations μ_n divided by the (expected) total number of samples requested throughout all tasks. In Section 3 we present this setting in more detail and introduce the relevant notation.

I.I.D. assumption. We assume that the value μ_n of the n-th innovation is drawn i.i.d. from an unknown and fixed distribution. This assumption is meaningful if past decisions do not influence future innovations whose global quality remains stable over time. In particular, it applies whenever innovations can progress in many orthogonal directions, each yielding a similar added value (e.g., when different teams within the same company test improvements relative to individual aspects of the company). If both the state of the agent and that of the environment evolve, but the ratio of good versus bad innovations remains essentially the same, then this i.i.d. assumption is still justified. In other words, it is not necessarily the absolute quality of the innovations that has to remain stationary, but rather the relative added value of the innovations given the current state of the system. This case is frequent in practice, especially when a system is close to its technological limit. Last but not least, algorithms designed under stochastic assumptions often perform surprisingly well in practice, even if i.i.d. assumptions are not fully satisfied or simply hard to check.

A baseline strategy and policy classes. A natural, yet suboptimal, approach for deciding if an innovation is worth accepting is to gather samples sequentially, stopping as soon as the absolute value of their running average surpasses a threshold, and then accepting the innovation if and only if the average is positive. The major drawback of this approach is that the value μ_n of an innovation n could be arbitrarily close to zero. In this case, the number of samples needed to reliably determine its sign (which is of order $1/\mu_n^2$) would become prohibitively large. This would result is a massive time investment for an innovation whose return is negligible at best. In hindsight, it would have been better to reject the innovation early and move on to the next task. For this reason, testing processes in practice needs hard termination rules of the form: *if after drawing a certain number of samples no confident decision can be taken, then terminate the testing process rejecting the current innovation.* Denote by τ this capped early stopping rule and by accept the accept/reject decision rule that comes with it. We say that the pair $\pi = (\tau, \text{accept})$ is a *policy* because it fully characterize the decision-making process for an innovation. Policies defined by capped early stopping rules (see (4) for a concrete example) are of great practical importance [20, 22]. However, policies can be defined more generally by any reasonable stopping rule and decision function. Given a (possibly infinite) set of policies, and assuming that $\mu_1, \mu_2, \ldots$ are drawn i.i.d. from some unknown but fixed distribution, the goal is to learn efficiently, at the lowest cost, the best policy $\pi_\star$ in the set with respect to a sensible metric. Competing against fixed policy classes is a common modeling choice that allows to express the intrinsic constraints that are imposed by the nature of the decision-making problem. For example, even if some policies outside of the class could theoretically yield better performance, they might not be implementable because of time, budget, fairness, or technology constraints.

Challenges. One of the biggest challenges arising in our framework is that running a decision-making policy generates a collection of samples that —in general— cannot be used to form an unbiased estimate of the policy reward (see the impossibility result in Section E of the Appendix). The presence of this bias is a significant departure from settings like multiarmed and firing bandits [3, 18], where the learner observes an unbiased sample of the target quantity at the end of every round (see the next section for additional details). Moreover, contrary to standard online learning problems, the relevant performance measure is neither additive in the number of innovations nor in the number of samples per innovation. Therefore, algorithms have to be analyzed globally, and bandit-like techniques —in which the regret is additive over rounds— cannot be directly applied. We argue that these technical difficulties are a worthy price to pay in order to define a plausible setting, applicable to real-life scenarios.

Main contributions. The first contribution of this paper is providing a mathematical formalization of our ROI maximization setting for repeated decision making (Section 3). We then design an algorithm called Capped Policy Elimination (Algorithm 1, CAPE) that applies to finite policy classes (Section 4). We prove that CAPE converges to the optimal policy at rate $1/(\Delta^2 N)$, where N is the number of tasks and Δ is the unknown gap between the performance of the two best policies, and at rate $N^{-1/3}$ when Δ is small (Theorem 1) . In Section 5 we tackle the challenging problem of infinitely large policy classes. For this setting, we design a preprocessing step (Algorithm 2, ESC) that leads to the ESC-CAPE algorithm. We prove that this algorithm converges to the optimal policy in an infinite set at a rate of $N^{-1/3}$ (Theorem 4).

Limitations. Although we do not investigate lower bounds in this paper, we conjecture that our $N^{-1/3}$ convergence rate it is optimal due to similarities with bandits with weakly observable feedback graphs (see Section 4, "Divided we fall"). Another limitation of our theory is that it only applies to i.i.d. sequences of values μ_n. It would be interesting to extend our analysis to distributions of μ_n that evolve over time. These two intriguing problems are left open for future research.

2 Related Work

Return on Investment (ROI) was developed and popularized by Donaldson Brown in the early Nineties [12] and it is still considered an extremely valuable metric by the overwhelming majority of marketing managers [11]. Beyond economics, mathematics, and computer science, ROI finds applications in other fields, such as cognitive science and psychology [7]. Despite this, to the best of our knowledge, no theoretical online learning framework has been developed specifically for ROI maximization. However, our novel formalization of this sequential decision problem does share some similarities with other known online learning settings. In this section, we review the relevant literature regarding these settings and stress the differences with ours.

Prophet inequalities and Pandora's box. In prophet inequalities [24, 9, 1], an agent observes sequentially (usually non-negative) random variables $Z_1, \dots, Z_n$ and decides to stop at some time τ; the reward is then Z_τ. Variants include the possibility of choosing more than one random variable (in which case the reward is some function of the selected random variables), and the possibility to go back in time (to some extent). The Pandora's box problem is slightly different [32, 21, 10]; in its original formulation, the agent can pay a cost $c_n \geq 0$ to observe any Z_n. After stopping exploring, the agent's final utility is the maximum of the observed Z_n's minus the cumulative cost (or, in other variants, some function of these). Similarly to the (general) prophet inequality, the agent in our sequential problem faces random variables ($Z_n = \mu_n$ in our notation) and sequentially selects any number of them (possibly with negative values) without the possibility to go back in time and change past decisions. The significant difference is that the agent in our setting never observes the value of μ_n. In Pandora's box, the agent can see this value by paying some price (that approximately scales as $1/\varepsilon^2$ where ε is the required precision). Finally, the global reward is the cumulative sum (as in prophets) and not the maximum (as in Pandora's box) of the selected variables, normalized by the total cost (as in Pandora's box, but our normalization is multiplicative instead of additive, as it represents a ROI).

Multi-armed bandits. If we think of the set of all policies used by the agent to determine whether or not to accept innovations as arms, our setting becomes somewhat reminiscent of multi-armed bandits [29, 6, 27]. However, there are several notable differences between these two problems. In stochastic bandits, the agent observes an unbiased estimate of the expected reward of each pulled arm. In our setting, the agent not only does not see it directly, but it is mathematically impossible to define such an estimator solely with the feedback received (see the impossibility result in Section E of the Appendix). Hence, off-the-shelf bandit algorithms cannot be run to solve our problem. In addition, the objective in bandits is to maximize the cumulative reward, which is additive over time, while the ROI is not. Thus, it is unclear how formal guarantees for bandit algorithms would translate to our problem.

We could also see firing bandits [18] as a variant of our problem, where μ_n belongs to $[0, 1]$, $\mathcal{D}_n$ are Bernoulli distribution with parameter μ_n, and policies have a specific form that allows to easily define unbiased estimates of their rewards (which, we reiterate, is not possible in our setting in general). Furthermore, in firing bandits, it is possible to go back and forth in time, sampling from any of the past distributions $\mathcal{D}_n$ and gathering any number of samples from it. This is a reasonable assumption for the original motivations of firing bandits because the authors thought of μ_n as the value of a project in a crowdfunding platform, and, in their setting, drawing samples from $\mathcal{D}_n$ corresponds to displaying projects on web pages. However, in our setting, μ_n represents the theoretical increment (or decrement) of a company's profit through a given innovation, and it is unlikely that a company would show new interest in investing in a technology that has been tested before and did not prove to be useful (a killed project is seldom re-launched). Hence, when the sampling of $\mathcal{D}_n$ stops, an irrevocable decision is made. After that, the learner cannot draw any more samples in the future. Finally, as in multi-armed bandits, the performance criterion in firing bandits is the cumulative reward and not the global ROI.

Another online problem that shares some similarities with ours is bandits with knapsacks [5]. In this problem, playing an arm consumes one unit of time together with some other resources, and the learner receives an unbiased estimate of its reward as feedback. The process ends as soon as time or any one of the other resources is exhausted. As usual, the goal is to maximize the cumulative regret. As it turns out, we can also think of our problem as a budgeted problem. In this restatement, there is a budget of T samples. The repeated decision-making process proceeds as before, but it stops as soon as the learner has drawn a total of T samples across all decision tasks. The goal is again to maximize the total expected reward of accepted innovations divided by T (see Section D of the Appendix for more details on the reduction). As per the other bandit problems, there are two crucial differences. First, running a policy does not reveal an unbiased estimate of its reward. Second, our objective is different, and regret bounds do not directly imply convergence to optimal ROI.

Repeated A/B testing. We can view our problem as a framework for repeated A/B testing [30, 14, 13, 17, 19, 4, 23, 28], in which assessing the value of an innovation corresponds to performing an A/B test, and the goal is maximizing the ROI. A popular metric to optimize sequential A/B tests is the so-called *false discovery rate* (FDR) —see [25, 33] and references therein. Roughly speaking, the FDR is the ratio of accepted μ_n that are negative over the total number of accepted μ_n (or more generally, the number of incorrectly accepted tests over the total number if the metric used at each test changes with time). This, unfortunately, disregards the relative values of tests μ_n that must be taken into account when optimizing a single metric [8, 26]. Indeed, the effect of many even slightly negative accepted tests could be overcome by a few largely positive ones. For instance, assume that the samples $X_{n,i}$ of any distribution $\mathcal{D}_n$ belong to $\{-1, 1\}$, and that their expected value μ_n is uniformly distributed on $\{-\varepsilon, \varepsilon\}$. To control the FDR, each A/B test should be run for approximately $1/\varepsilon^2$ times, yielding a ratio of the average value of an accepted test to the number of samples of order ε^3. A better strategy, using just one sample from each A/B test, is simply to accept μ_n if and only if the first sample is positive. Direct computations show that this policy, which fits our setting, achieves a significantly better performance of order ε.

Some other A/B testing settings are more closely related to ours, but make stronger additional assumptions or suppose preliminary knowledge: for example, smoothness assumptions can be made on both $\mathcal{D}_n$ and the

distributions of μ_n [4], or the distribution of μ_n is known, and the distribution of samples belongs to a single parameter exponential family, also known beforehand [28].

Rational metareasoning. Our setting is loosely related to the AI field of meta-reasoning [15, 16]. In a metalevel decision problem, determining the utility (or reward) of a given action is computationally intractable. Instead, the learner can run a simulation, investing a computational cost to gather information about this hidden value. The high-level idea is then to learn *which* actions to simulate. After running some simulations, the learner picks an action to play, gains the corresponding (hidden) reward, and the state of the system changes. In rational meta-reasoning, the performance measure is the value of computation (VOC): the difference between the increment in expected utility gained by executing a simulation and the cost incurred by doing so. This setting is not directly comparable to ours for two reasons. First, the performance measure is different, and the additive nature of the difference that defines the VOC gives no guarantees on our multiplicative notion of ROI. Second, in this problem, one can pick which actions to simulate, while in our settings, innovations come independently of the learner, who has to evaluate them in that order.

3 Setting and Notation

In this section, we formally introduce the repeated decision-making protocol for an agent whose goal is to maximize the total return on investment in a sequence of decision tasks.

The only two choices that an agent makes in a decision task are when to stop gathering information on the current innovation and whether or not to accept the innovation based on this information. In other words, the behavior of the agent during each task is fully characterized by the choice of a pair $\pi = (\tau, \text{accept})$ that we call a *(decision-making) policy* (for the interested reader, Section A of the Appendix contains a short mathematical discussion on policies), where:

- $\tau(\boldsymbol{x})$, called *duration*, maps a sequence of observations $\boldsymbol{x} = (x_1, x_2, \ldots)$ to an integer d (the no. of observations after which the learner stops gathering info on the current innovation);
- $\text{accept}(d, \boldsymbol{x})$, called *decision*, maps the firs d observations of a sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ to a boolean value in $\{0, 1\}$ (where 1 represents accepting the current innovation).

An instance of our repeated decision-making problem is therefore determined by a set of admissible policies $\Pi = \{\pi_k\}_{k\in\mathcal{K}} = \{(\tau_k, \text{accept})\}_{k\in\mathcal{K}}$ (with $\mathcal{K}$ finite or countable) and a distribution μ on $[-1, 1]$, modelling the value of innovations.[1] Naturally, the former is known beforehand but the latter is unknown and should be learned.

For a fixed choice of Π and μ, the protocol is formally described below. In each decision task n:

1. the *value* μ_n of the current innovation is drawn i.i.d. according to μ;
2. $\boldsymbol{X}_n$ is a sequence of i.i.d. (given μ_n) *observations* with $X_{n,i} = \pm 1$ and $\mathbb{E}[X_{n,i} \mid \mu_n] = \mu_n$;
3. the agent picks $k_n \in \mathcal{K}$ or, equivalently, a policy $\pi_{k_n} = (\tau_{k_n}, \text{accept}) \in \Pi$;
4. the agent draws the first $d_n = \tau_{k_n}(\boldsymbol{X}_n)$ *samples*[2] of the sequence of observations $\boldsymbol{X}_n$;
5. on the basis of these sequential observations, the agent makes the decision $\text{accept}\big(d_n, \boldsymbol{X}_n\big)$.

Crucially, μ_n is *never* revealed to the learner. We say that the agent *runs a policy* $\pi_k = (\tau_k, \text{accept})$ (on a value μ_n) when steps 4–5 occur (with $k_n \leftarrow k$). We also say that they accept (resp., rejects) μ_n if their decision at step 5 is equal to 1 (resp., 0). Moreover, we say that the *reward* obtained and the *cost* payed by

[1] We assume that the values of the innovations and the learner's observations belong to $[-1, 1]$ and $\{-1, 1\}$ respectively. We do this merely for the sake of readability (to avoid carrying over awkward constants or distributions $\mathcal{D}_n$). With a standard argument, both $[-1, 1]$ and $\{-1, 1\}$ can be extended to arbitrary codomains straightforwardly under a mild assumption of subgaussianity.

[2] Given μ_n, the random variable d_n is a stopping time w.r.t. the natural filtration associated to $\boldsymbol{X}_n$.

running a policy $\pi_k = (\tau_k, \text{accept})$ on a value μ_n are, respectively,

$$\text{reward}(\pi_k, \mu_n) = \mu_n \, \text{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \in \{\mu_n, 0\} \qquad \text{cost}(\pi_k, \mu_n) = \tau_k(\boldsymbol{X}_n) \in \mathbb{N} \tag{1}$$

The objective of the agent is to converge to the highest ROI of a policy in Π, i.e., to guarantee that

$$R_N = \sup_{k \in \mathcal{K}} \frac{\sum_{n=1}^N \mathbb{E}\big[\text{reward}(\pi_k, \mu_n)\big]}{\sum_{m=1}^N \mathbb{E}\big[\text{cost}(\pi_k, \mu_m)\big]} - \frac{\sum_{n=1}^N \mathbb{E}\big[\text{reward}(\pi_{k_n}, \mu_n)\big]}{\sum_{m=1}^N \mathbb{E}\big[\text{cost}(\pi_{k_m}, \mu_m)\big]} \to 0 \quad \text{as } N \to \infty \tag{2}$$

where the expectations are taken with respect to μ_n, $\boldsymbol{X}_n$, and (possibly) the random choices of k_n.

To further lighten notations, we denote the expected reward, cost, and ROI of a policy π by

$$\text{reward}(\pi) = \mathbb{E}\big[\text{reward}(\pi, \mu_n)\big], \, \text{cost}(\pi) = \mathbb{E}\big[\text{cost}(\pi, \mu_n)\big], \, \text{ROI}(\pi) = \text{reward}(\pi)/\text{cost}(\pi) \tag{3}$$

respectively and we say that $\pi_{k^\star}$ is an *optimal policy* if $k^\star \in \text{argmax}_{k \in \mathcal{K}} \text{ROI}(\pi_k)$. Note that $\text{reward}(\pi)$ and $\text{cost}(\pi)$ do not depend on n because μ_n is drawn i.i.d. according to μ.

For each policy $(\tau, \text{accept}) \in \Pi$ and all tasks n, we allow the agent to reject the value μ_n regardless of the outcome of the sampling. Formally, the agent can always run the policy $(\tau, 0)$, where the second component of the pair is the decision identically equal to zero (i.e., the rule "always reject").

We also allow the agent to draw arbitrarily many extra samples in addition to the number $\tau(\boldsymbol{X}_n)$ that they would otherwise draw when running a policy $(\tau, \text{accept}) \in \Pi$ on a value μ_n, provided that these additional samples are not taken into account in the decision to either accept or reject μ_n. Formally, the agent can always draw $\tau(\boldsymbol{X}_n) + k$ many samples (for any $k \in \mathbb{N}$) before making the decision $\text{accept}\big(\tau(\boldsymbol{X}_n), \boldsymbol{X}_n\big)$, where we stress that the first argument of the decision function accept is $\tau(\boldsymbol{X}_n)$ and not $\tau(\boldsymbol{X}_n) + k$. Oversampling this way worsens the objective and might seem utterly counterproductive, but it will be crucial for recovering unbiased estimates of μ_n.

4 Competing Against K policies (CAPE)

As we mentioned in the introduction, in practice the duration of a decision task is defined by a capped early-stopping rule —e.g., drawing samples until 0 falls outside of a confidence interval around the empirical average, or a maximum number of draws has been reached. More precisely, if N tasks have to be performed, one could consider the natural policy class $\{(\tau_k, \text{accept})\}_{k \in \{1, \dots, K\}}$ given by

$$\tau_k(\boldsymbol{x}) = \min\big(k, \ \inf\{d \in \mathbb{N} : |\overline{x}_d| \geq \alpha_d\}\big) \qquad \text{and} \qquad \text{accept}(d, \boldsymbol{x}) = \mathbb{I}\{\overline{x}_d \geq \alpha_d\} \tag{4}$$

where $\overline{x}_d = (1/d) \sum_{i=1}^d x_i$ is the average of the first d elements of the sequence $\boldsymbol{x} = (x_1, x_2, \dots)$ and $\alpha_d = c\sqrt{(1/d)\ln(KN/\delta)}$, for some $c > 0$ and $\delta \in (0, 1)$. While in this example policies are based on an Hoeffding concentration rule, in principle the learner is free to follow any scheme. Thus, we now generalize this notion and present an algorithm with provable guarantees against these finite families of policies.

Finite sets of policies. In this section, we focus on finite sets of K policies $\Pi = \{\pi_k\}_{k \in \{1, \dots, K\}} = \{(\tau_k, \text{accept})\}_{k \in \{1, \dots, K\}}$ where accept is an arbitrary decision and $\tau_1, \dots, \tau_K$ is any sequence of bounded durations (say, $\tau_k \leq k$ for all k).[3] For the sake of convenience, we assume the durations are sorted by index ($\tau_k \leq \tau_h$ if $k \leq h$), so that τ_1 is the shortest and τ_K is the longest.

[3] We chose $\tau_k \leq k$ for the sake of concreteness. All our results can be straightforwardly extended to arbitrary $\tau_k \leq D_k$ by simply assuming without loss of generality that $k \mapsto D_k$ is monotone and replacing k with D_k.

Divided we fall. A common strategy in online learning problems with limited feedback is explore-then-commit (ETC). ETC consists of two phases. In the first phase (explore), each action is played for the same amount of rounds, collecting this way i.i.d. samples of all rewards. In the subsequent commit phase, the arm with the best empirical observations is played consistently. Being very easy to execute, this strategy is popular in practice, but unfortunately, it is theoretically suboptimal in some applications. A better approach is performing action elimination. In a typical implementation of this strategy, all actions in a set are played with a round-robin schedule, collecting i.i.d. samples of their rewards. At the end of each cycle, all actions that are deemed suboptimal are removed from the set, and a new cycle begins. Neither one of these strategies can be applied directly because running a policy in our setting does not return an unbiased estimate of its reward (for a quick proof of this simple result, see Section E in the Appendix). However, it turns out that we can get an i.i.d. estimate of a policy π by playing a *different* policy π'. Namely, one that draws *more* samples than π. This is reminiscent of bandits with a weakly observable feedback graph, a related problem for which the time-averaged regret over T rounds vanishes at a $T^{-1/3}$ rate [2]. Albeit none of these three techniques works on its own, suitably interweaving all of them does.

United we stand. With this in mind, we now present our simple and efficient algorithm (Algorithm 1, CAPE) whose ROI converges (with high probability) to the best one in a finite family of policies. We will later discuss how to extend the analysis even further, including countable families of policies. Our algorithm performs policy elimination (lines 1–5) for a certain number of tasks (line 1) or until a single policy is left (line 6). After that, it runs the best policy left in the set (line 7) for all remaining tasks. During each policy elimination step, the algorithm oversamples (line 2) by drawing twice as many samples as it would suffice to take its decision $\mathrm{accept}\big(\tau_{\max(C_n)}(\boldsymbol{X}_n), \boldsymbol{X}_n\big)$ (at line 3). These extra samples are used to compute rough estimates of rewards and costs of all potentially optimal policies and more specifically to build *unbiased* estimates of these rewards. The test at line 4 has the only purpose of ensuring that the denominators $\widehat{c}_n^-(k)$ at line 5 are bounded away from zero so that all quantities are well-defined.

Algorithm 1: Capped Policy Elimination (CAPE)

Input: finite policy set Π, number of tasks N, confidence parameter δ, exploration cap N_{ex}
Initialization: let $C_1 \leftarrow \{1, \dots, K\}$ be the set of indices of all currently optimal candidates

1 **for** *task* $n = 1, \dots, N_{\mathrm{ex}}$ **do**
2 draw the first $2\max(C_n)$ samples $X_{n,1}, \dots, X_{n,2\max(C_n)}$ of $\boldsymbol{X}_n$
3 make the decision $\mathrm{accept}\big(\tau_{\max(C_n)}(\boldsymbol{X}_n), \boldsymbol{X}_n\big)$
4 **if** $n \geq 2K^2 \ln(4KN_{\mathrm{ex}}/\delta)$ **then** let $C_{n+1} \leftarrow C_n \setminus C'_n$, where
5 $$C'_n = \big\{k \in C_n \,:\, \big(\widehat{r}_n^+(k) \geq 0 \text{ and } \widehat{r}_n^+(k)/\widehat{c}_n^-(k) < \widehat{r}_n^-(j)/\widehat{c}_n^+(j), \text{ for some } j \in C_n\big) \text{ or } \big(\widehat{r}_n^+(k) < 0 \text{ and } \widehat{r}_n^+(k)/\widehat{c}_n^+(k) < \widehat{r}_n^-(j)/\widehat{c}_n^-(j), \text{ for some } j \in C_n\big)\big\}$$

$$\widehat{r}_n^{\pm}(k) = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{\max(C_m)} \frac{X_{m,\max(C_m)+i}}{\max(C_m)}\,\mathrm{accept}\big(\tau_k(\boldsymbol{X}_m), \boldsymbol{X}_m\big) \pm \sqrt{\frac{2}{n}\ln\frac{4KN_{\mathrm{ex}}}{\delta}} \tag{5}$$

$$\widehat{c}_n^{\pm}(k) = \frac{1}{n}\sum_{m=1}^{n}\tau_k(\boldsymbol{X}_m) \pm (k-1)\sqrt{\frac{1}{2n}\ln\frac{4KN_{\mathrm{ex}}}{\delta}} \tag{6}$$

6 **if** $|C_{n+1}| = 1$ **then** let $\widehat{r}_{N_{\mathrm{ex}}}^{\pm}(k) \leftarrow \widehat{r}_n^{\pm}(k)$, $\widehat{c}_{N_{\mathrm{ex}}}^{\pm}(k) \leftarrow \widehat{c}_n^{\pm}(k)$, $C_{N_{\mathrm{ex}}+1} \leftarrow C_{n+1}$, **break**
7 run policy $\pi_{k'}$ for all remaining tasks, where

$$k' \in \begin{cases} \operatorname*{argmax}_{k \in C_{N_{\mathrm{ex}}+1}} \big(\widehat{r}_{N_{\mathrm{ex}}}^+(k)/\widehat{c}_{N_{\mathrm{ex}}}^-(k)\big) & \text{if } \widehat{r}_{N_{\mathrm{ex}}}^+(k) \geq 0 \text{ for some } k \in C_{N_{\mathrm{ex}}+1} \\ \operatorname*{argmax}_{k \in C_{N_{\mathrm{ex}}+1}} \big(\widehat{r}_{N_{\mathrm{ex}}}^+(k)/\widehat{c}_{N_{\mathrm{ex}}}^+(k)\big) & \text{if } \widehat{r}_{N_{\mathrm{ex}}}^+(k) < 0 \text{ for all } k \in C_{N_{\mathrm{ex}}+1} \end{cases} \tag{7}$$

As usual in online learning, the *gap* in performance between optimal and sub-optimal policies is a complexity parameter. We define it as $\Delta = \min_{k \neq k^\star} \big(\mathrm{ROI}(\pi_{k^\star}) - \mathrm{ROI}(\pi_k)\big)$, where we recall that $k^\star \in \operatorname{argmax}_k \mathrm{ROI}(\pi_k)$ is the index of an optimal policy. Conventionally, we set $1/0 = \infty$.

Theorem 1. *If Π is a finite set of K policies, then the ROI of Algorithm 1 run for N tasks with exploration cap $N_{\mathrm{ex}} = \lceil N^{2/3} \rceil$ and confidence parameter $\delta \in (0,1)$ converges to the optimal $\mathrm{ROI}(\pi_{k^\star})$, with probability at least $1-\delta$, at a rate*

$$R_N = \widetilde{\mathcal{O}}\left(\min\left(\frac{K^3}{\Delta^2 N}, \frac{K}{N^{1/3}}\right)\right)$$

as soon as $N \geq K^3$ (where the $\widetilde{\mathcal{O}}$ notation hides only logarithmic terms, including a $\log(1/\delta)$ term).

Proof sketch. This theorem relies on four technical lemmas (Lemmas 5-8) whose proofs are deferred to Section B of the Appendix.

With a concentration argument (Lemma 5), we leverage the definitions of $\widehat{r}_n^{\pm}(k), \widehat{c}_n^{\pm}(k)$ and the i.i.d. assumptions on the samples $X_{n,i}$ to show that, with probability at least $1-\delta$, the event

$$\widehat{r}_n^-(k) \leq \mathrm{reward}(\pi_k) \leq \widehat{r}_n^+(k) \qquad \text{and} \qquad \widehat{c}_n^-(k) \leq \mathrm{cost}(\pi_k) \leq \widehat{c}_n^+(k) \tag{8}$$

occurs simultaneously for all $n \leq N_{\mathrm{ex}}$ and all $k \leq \max(C_n)$. For the rewards, the key is oversampling, because $\mathrm{accept}\big(\tau_k(\boldsymbol{X}_m), \boldsymbol{X}_m\big)$ in eq. (5) depends only on the first $k \leq \max(C_m)$ samples of $\boldsymbol{X}_m$ and is therefore independent of $X_{m,\max(C_m)+i}$ for all i. Assume now that (8) holds.

If $\Delta > 0$ (i.e., if there is a unique optimal policy), we then obtain (Lemma 6) that suboptimal policies are eliminated after at most N'_{ex} tasks, where $N'_{\mathrm{ex}} \leq 288\, K^2 \ln(4KN_{\mathrm{ex}}/\delta)/\Delta^2 + 1$. To prove it we show that a confidence interval for $\mathrm{ROI}(\pi_k) = \mathrm{reward}(\pi_k)/\mathrm{cost}(\pi_k)$ is given by

$$\left[\frac{\widehat{r}_n^-(k)}{\widehat{c}_n^+(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) \geq 0\big\} + \frac{\widehat{r}_n^-(k)}{\widehat{c}_n^-(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) < 0\big\},\ \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^-(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) \geq 0\big\} + \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^+(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) < 0\big\}\right]$$

we upper bound its length, and we compute an N'_{ex} such that this upper bound is smaller than $\Delta/2$.

Afterwards, we analyze separately the case in which the test at line 6 is true for some task $N'_{\mathrm{ex}} \leq N_{\mathrm{ex}}$ and its complement (i.e., when the test is always false).

In the first case, by (8) there exists a unique optimal policy, i.e., we have that $\Delta > 0$. This is where the policy-elimination analysis comes into play. We can apply the bound above on N'_{ex}, obtaining a deterministic upper bound N''_{ex} on the number N'_{ex} of tasks needed to identify the optimal policy. Using this upper bound, writing the definition of R_N, and further upper bounding (Lemma 7) yields

$$R_N \leq \min\left(\frac{(2K+1)N_{\mathrm{ex}}}{N},\ \frac{(2K+1)\big(288\,(K/\Delta)^2 \ln(4KN_{\mathrm{ex}}/\delta) + 1\big)}{N}\right) \tag{9}$$

Finally, we consider the case in which the test at line 6 is false for all tasks $n \leq N_{\mathrm{ex}}$, and line 7 is executed with $C_{N_{\mathrm{ex}}+1}$ containing two or more policies. This is covered by a worst case explore-then-commit analysis. The key idea here is to use the definition of k' in Equation (7) to lower-bound $\mathrm{reward}(\pi_{k'})$ in terms of $\mathrm{reward}(\pi_{k^\star})/\mathrm{cost}(\pi_{k^\star})$. This, together with some additional technical estimations (Lemma 8) leads to the result. □

5 Competing Against Infinitely Many Policies (ESC-CAPE)

Theorem 1 provides theoretical guarantees on the convergence rate R_N of CAPE to the best ROI of a finite set of policies. Unfortunately, the bound becomes vacuous when the cardinality K of the policy set is large compared to the number of tasks N. It is therefore natural to investigate whether the problem becomes impossible in this scenario.

Infinite sets of policies. With this goal in mind, we now focus on policy sets $\Pi = \{\pi_k\}_{k\in\mathcal{K}} = \big\{(\tau_k, \text{accept})\big\}_{k\in\mathcal{K}}$ as in the previous section, with $\mathcal{K} = \mathbb{N}$ rather than $\mathcal{K} = \{1, \ldots, K\}$.

We will show how such a countable set of policies can be reduced to a finite one containing all optimal policies with high probability (Algorithm 2, ESC). After this is done, we can run CAPE on the smaller policy set, obtaining theoretical guarantees for the resulting algorithm.

Estimating rewards and costs. Similarly to eqs. (5) and (6), we first introduce estimators for our target quantities. If at least $2k$ samples are drawn during each of n_2 consecutive tasks $n_1+1, \ldots, n_1+n_2$, we can define, for all $\varepsilon > 0$, the following lower confidence bound on reward(π_k):

$$\widehat{r}_k^-(n_1, n_2, \varepsilon) = \frac{1}{n_2} \sum_{n=n_1+1}^{n_1+n_2} \sum_{i=1}^{k} \frac{X_{n,k+i}}{k} \,\text{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) - 2\varepsilon \tag{10}$$

If at least $\tau_k(\boldsymbol{X}_n)$ samples are drawn during each of m_0 consecutive tasks $n_0+1, \ldots, n_0+m_0$, we can define the following empirical average of cost(π_k):

$$\overline{c}_k(n_0, m_0) = \big(\tau_k(\boldsymbol{X}_{n_0+1}) + \ldots + \tau_k(\boldsymbol{X}_{n_0+m_0})\big)/m_0 \tag{11}$$

A key observation. The key idea behind Algorithm 2 (ESC) is simple. Since all optimal policies $\pi_{k^\star}$ have to satisfy the relationships reward(π_k)/cost$(\pi_k) \le$ reward$(\pi_{k^\star})$/cost$(\pi_{k^\star}) \le 1$/cost$(\pi_{k^\star})$, then, for all policies π_k with positive reward(π_k), the cost of any optimal policy $\pi_{k^\star}$ must satisfy the relationship cost$(\pi_{k^\star}) \le$ cost(π_k)/reward(π_k). In other words, *optimal policies cannot draw too many samples* and their cost can be controlled by estimating the reward and cost of *any* policy with positive reward.

We recall that running a policy $(\tau, 0)$ during a task n means drawing the first $\tau(\boldsymbol{X}_n)$ samples of $\boldsymbol{X}_n = (X_{n,1}, X_{n,2}, \ldots)$ and always rejecting μ_n, regardless of the observations.

Algorithm 2: Extension to Countable (ESC)

Input: countable policy set Π, number of tasks N, confidence parameter δ, accuracy levels $(\varepsilon_n)_n$
Initialization: for all j, let $m_j \leftarrow \big\lceil \ln\big(j(j+1)/\delta\big)/2\varepsilon_j^2 \big\rceil$ and $M_j = m_1 + \ldots + m_j$

1 **for** $j = 1, 2, \ldots$ **do**
2 run policy $\big(2 \cdot 2^j, 0\big)$ for m_j tasks and compute $\widehat{r}_{2^j}^- \leftarrow \widehat{r}_{2^j}^-(M_{j-1}, m_j, \varepsilon_j)$ as in (10)
3 **if** $\widehat{r}_{2^j}^- > 0$ **then** let $j_0 \leftarrow j$ and $k_0 \leftarrow 2^{j_0}$
4 **for** $l = j_0+1, j_0+2, \ldots$ **do**
5 run policy $\big(\tau_{2^l}, 0\big)$ for m_l tasks and compute $\overline{c}_{2^l} \leftarrow \overline{c}_{2^l}(M_{l-1}, m_l)$ as in (11)
6 **if** $\overline{c}_{2^l} > 2^l\,\varepsilon_l + k_0/\widehat{r}_{k_0}^-$ **then** let $j_1 \leftarrow l$ and **return** $K \leftarrow 2^{j_1}$

Thus, Algorithm 2 (ESC) first finds a policy π_{k_0} with reward$(\pi_{k_0}) > 0$ (lines 1–3), memorizing an upper estimate $k_0/\widehat{r}_{k_0}^-$ of the ratio cost(π_{k_0})/reward$(\pi_{k_0}) = 1$/ROI(π_{k_0}). By the argument above, this estimate upper bounds the expected number of samples cost$(\pi_{k^\star})$ drawn by *all* optimal policies $\pi_{k^\star}$. Then ESC simply proceeds to finding the smallest (up to a factor of 2) K such that cost$(\pi_K) \ge k_0/\widehat{r}_{k_0}^-$ (lines 4–6). Being $k_0/\widehat{r}_{k_0}^- \ge$ cost(π_{k_0})/reward$(\pi_{k_0}) \ge$ cost$(\pi_{k^\star})$ by construction, the index K determined this way upper bounds $k^\star$ for all optimal policies $\pi_{k^\star}$. (All the previous statements are intended to hold with high probability.) This is formalized in the following key lemma, whose full proof we defer to Section C of the Appendix.

Lemma 2. *Let Π be a countable set of policies. If ESC is run with $\delta \in (0,1)$, $\varepsilon_1, \varepsilon_2, \ldots \in (0,1]$, and halts returning K, then $k^\star \le K$ for all optimal policies $\pi_{k^\star}$ with probability at least $1-\delta$.*

Before proceeding with the main result of this section, we need a final lemma upper bounding the expected cost of our ESC algorithm. This step is crucial to control the total ROI because in this setting with

arbitrarily long durations, picking the wrong policy even once is, in general, enough to drop the performance of an algorithm down to essentially zero, compromising the convergence to an optimal policy. This is another striking difference with other common online learning settings like stochastic bandits, where a single round has a negligible influence on the overall performance of an algorithm. To circumvent this issue, we designed ESC so that it tests shorter durations first, stopping as soon as the previous lemma applies, and a finite upper bound K on $k^\star$ is determined.

Lemma 3. *Let Π be a countable set of policies. If ESC is run with $\delta \in (0,1)$, $\varepsilon_1, \varepsilon_2, \ldots \in (0,1]$, and halts returning K, then the total number of samples it draws before stopping (i.e., its cost) is upper bounded by $\widetilde{\mathcal{O}}\big((K/\varepsilon^2)\log(1/\delta)\big)$, where $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{\log_2 K}\}$.*

Proof. Note that, by definition, $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{j_1}\} > 0$. Algorithm 2 (ESC) draw samples only when lines 2 or 5 are executed. Whenever line 2 is executed ($j = 1, \ldots, j_0$) the algorithm performs m_j tasks drawing $2 \cdot 2^j$ samples each time. Similarly, whenever line 5 is executed ($l = j_0 + 1, \ldots, j_1$) the algorithm draws at most 2^l samples during each of the m_l tasks. Therefore, recalling that $j_1 = \log_2 K$, the total number of samples drawn by ESC before stopping is at most

$$\sum_{j=1}^{j_0} 2 \cdot 2^j m_j + \sum_{l=j_0+1}^{j_1} 2^l m_l \le 2 \sum_{j=1}^{j_1} 2^j m_j \le 2 j_1 2^{j_1} \left\lceil \frac{1}{2\varepsilon^2} \ln \frac{j_1(j_1+1)}{\delta} \right\rceil \qquad \square$$

The ESC-CAPE algorithm. We can now join together our two algorithms obtaining a new one, that we call ESC-CAPE, which takes as input a countable policy set Π, the number of tasks N, a confidence parameter δ, some accuracy levels $\varepsilon_1, \varepsilon_2, \ldots$, and an exploration cap N_{ex}. The joint algorithm runs ESC first with parameters $\Pi, N, \delta, \varepsilon_1, \varepsilon_2, \ldots$. Then, if ESC halts returning K, it runs CAPE with parameters $\{(\tau_k, \text{accept})\}_{k \in \{1,\ldots,K\}}, N, \delta, N_{\text{ex}}$.

Analysis of ESC-CAPE. Since ESC rejects all values μ_n, the sum of the rewards accumulated during its run is zero. Thus, the only effect that ESC has on the convergence rate R_N of ESC-CAPE is an increment on the total cost in the denominator of its ROI. We control this cost by minimizing its upper bound in Lemma 3. This is not a simple matter of taking all ε_j's as large as possible. Indeed, if all the ε_j's are large, the **if** clause at line 3 might never be verified. In other words, the returned index K depends on ε and grows unbounded in general as ε approaches $1/2$. This follows directly from the definition of our lower estimate on the rewards (10). Thus, there is a trade-off between having a small K (which requires small ε_j's) and a small $1/\varepsilon^2$ to control the cost of ESC (for which we need large ε_j's). A direct computation shows that picking constant accuracy levels $\varepsilon_j = N^{-1/3}$ for all j achieves the best of both worlds and immediately gives our final result.

Theorem 4. *If Π is a countable set of policies, then the ROI of ESC-CAPE run for N tasks with confidence parameter $\delta \in (0,1)$, constant accuracy levels $\varepsilon_j = N^{-1/3}$, and exploration cap $N_{\text{ex}} = \lceil N^{2/3} \rceil$ converges to the optimal $\text{ROI}(\pi_{k^\star})$, with probability at least $1 - \delta$, at a rate*

$$R_N = \widetilde{\mathcal{O}}\left(\frac{1 + K \mathbb{I}\{\textit{ESC halts returning } K\}}{N^{1/3}} \right)$$

where the $\widetilde{\mathcal{O}}$ notation hides only logarithmic terms, including a $\log(1/\delta)$ term.

6 Conclusions and Open Problems

After formalizing the problem of ROI maximization in repeated decision making, we presented an algorithm (ESC-CAPE) that is competitive against infinitely large policy sets (Theorem 4). For this algorithm, we prove a convergence rate of order $1/N^{1/3}$ with high probability. To analyze it, we first proved a convergence

result for its finite counterpart CAPE (Theorem 1), which is of independent interest. Notably, this finite analysis guarantees a significantly faster convergence of order $1/N$ on easier instances in which there is a positive gap in performance between the two best policies.

Acknowledgements

An earlier version of this work was done during Tommaso Cesari's Ph.D. at the University of Milan. Nicolò Cesa-Bianchi and Tommaso R. Cesari gratefully acknowledge partial support by Criteo AI Lab through a Faculty Research Award and by the MIUR PRIN grant Algorithms, Games, and Digital Markets (ALGADIMAR). This work has also benefited from the AI Interdisciplinary Institute ANITI. ANITI is funded by the French "Investing for the Future – PIA3" program under the Grant agreement n. ANR-19-PI3A-0004. Yishay Mansour was supported in part by a grant from the Israel Science Foundation (ISF). Vianney Perchet was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH, in a joint call with Gaspard Monge Program for optimization, operations research and their interactions with data sciences. Vianney Perchet also acknowledges the support of the ANR under the grant ANR-19-CE23-0026.

A Decision-Making Policies

In this section, we give a formal functional definition of the decision-making policies introduced in Section 3. During each task, the agent sequentially observes samples $x_i \in [-1,1]$ representing realizations of stochastic observations of the current innovation value. A map $\tau\colon [-1,1]^{\mathbb{N}} \to \mathbb{N}$ is a *duration* (of a decision task) if for all $\boldsymbol{x} \in [-1,1]^{\mathbb{N}}$, its value $d = \tau(\boldsymbol{x}) \in \mathbb{N}$ at $\boldsymbol{x}$ depends only on the first d components $x_1, x_2, \ldots, x_d$ of $\boldsymbol{x} = (x_1, x_2, \ldots)$; mathematically speaking, if $\boldsymbol{X}$ is a discrete stochastic process (i.e., a random sequence), then $\tau(\boldsymbol{X})$ is a stopping time with respect to the filtration generated by $\boldsymbol{X}$. This definition reflects the fact that the components $x_1, x_2, \ldots$ of the sequence $\boldsymbol{x} = (x_1, x_2, \ldots)$ are generated sequentially, and the decision to stop testing an innovation depends only on what occurred so far. A concrete example of a duration function is the one, mentioned in the introduction and formalized in (4), that keeps drawing samples until the empirical average of the observed values x_i surpasses/falls below a certain threshold, or a maximum number of samples have been drawn.

To conclude a task, the agent has to make a decision: either accepting or rejecting the current innovation. Formally, we say that a function $\mathrm{accept}\colon \mathbb{N} \times [-1,1]^{\mathbb{N}} \to \{0,1\}$ is a *decision* (to accept) if for all $d \in \mathbb{N}$ and $\boldsymbol{x} \in [-1,1]^{\mathbb{N}}$, its value $\mathrm{accept}(d, \boldsymbol{x}) \in \{0,1\}$ at $(d, \boldsymbol{x})$ depends only on the first d components $x_1, \ldots, x_d$ of $\boldsymbol{x} = (x_1, x_2, \ldots)$. Again, this definition reflects the fact that the decision $\mathrm{accept}(d, \boldsymbol{x})$ to either accept ($\mathrm{accept}(d, \boldsymbol{x}) = 1$) or reject ($\mathrm{accept}(d, \boldsymbol{x}) = 0$) the current innovation after observing the first d values $x_1, \ldots, x_d$ of $\boldsymbol{x} = (x_1, x_2, \ldots)$ is oblivious to all future observations $x_{d+1}, x_{d+2}, \ldots$. Following up on the concrete example above, the decision function is accepting the current innovation if and only if the the empirical average of the observed values x_i surpasses a certain threshold.[4]

Since the only two choices that an agent makes in a decision task are when to stop drawing new samples and whether or not to accept the current innovation, the behavior of the agent during each task is fully characterized by the choice of a pair $\pi = (\tau, \mathrm{accept})$ that we call a (*decision-making*) *policy*, where τ is a duration and accept is a decision.

B Technical Lemmas for Theorem 1

In this section, we give formal proofs of all results needed to prove Theorem 1.

[4]Note that, even for decision functions that only look at the mean of the first d values, our definition is more general than simple threshold functions of the form $\mathbb{I}\{\text{mean} \geq \varepsilon_d\}$, as it also includes all decisions of the form $\mathbb{I}\{\text{mean} \in A_d\}$, for all measurable $A_d \subset \mathbb{R}$.

Lemma 5. *Under the assumptions of Theorem 1, the event*

$$\widehat{r}_n^-(k) \le \mathrm{reward}(\pi_k) \le \widehat{r}_n^+(k) \qquad \textit{and} \qquad \widehat{c}_n^-(k) \le \mathrm{cost}(\pi_k) \le \widehat{c}_n^+(k) \tag{12}$$

occurs simultaneously for all $n = 1, \ldots, N_{\mathrm{ex}}$ and all $k = 1, \ldots, \max(C_n)$ with probability at least $1 - \delta$.

Proof. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\mathrm{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^+(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^+(k) - (k-1)\varepsilon_n \tag{13}$$

Note that $\overline{c}_n(k)$ is the empirical average of n i.i.d. samples of $\mathrm{cost}(\pi_k)$ for all n, k by definitions (13), (6), (1), (3), and point 4 in the formal definition of our protocol (Section 3). We show now that $\overline{r}_n(k)$ is the empirical average of n i.i.d. samples of $\mathrm{reward}(\pi_k)$ for all n, k; then claim (8) follows by Hoeffding's inequality. Indeed, by the conditional independence of the samples and being $\mathrm{accept}(k, \boldsymbol{x})$ independent of the variables $(x_{k+1}, x_{k+2}, \ldots)$ by definition, for all tasks n, all policies $k \in C_n$, and all $i > \max(C_n)$ ($\ge k$ by monotonicity of $k \mapsto k$),

$$\begin{aligned}
\mathbb{E}\left[X_{n,i}\,\mathrm{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\middle|\, \mu_n\right] &= \mathbb{E}\left[X_{n,i} \mid \mu_n\right] \mathbb{E}\Big[\mathrm{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big] \\
&= \mu_n\, \mathbb{E}\Big[\mathrm{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big] \\
&= \mathbb{E}\Big[\mu_n\, \mathrm{accept}\big(\tau_k(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big]
\end{aligned}$$

Taking expectations with respect to μ_n on both sides of the above, and recalling definitions (13), (5), (1), (3), (4) proves the claim. Thus, Hoeffding's inequality implies, for all fixed n, k,

$$\begin{aligned}
\mathbb{P}\big(\widehat{r}_n^-(k) \le \mathrm{reward}(\pi_k) \le \widehat{r}_n^+(k)\big) &= \mathbb{P}\Big(\big|\overline{r}_n(k) - \mathrm{reward}(\pi_k)\big| \le 2\varepsilon_n\Big) \ge 1 - \frac{\delta}{2KN_{\mathrm{ex}}} \\
\mathbb{P}\big(\widehat{c}_n^-(k) \le \mathrm{cost}(\pi_k) \le \widehat{c}_n^+(k)\big) &= \mathbb{P}\Big(\big|\overline{c}_n(k) - \mathrm{cost}(\pi_k)\big| \le (K-1)\varepsilon_n\Big) \ge 1 - \frac{\delta}{2KN_{\mathrm{ex}}}
\end{aligned}$$

Applying a union bound shows that event (8) occurs simultaneously for all $n \in \{1, \ldots, N_{\mathrm{ex}}\}$ and $k \in \{1, \ldots, \max(C_n)\}$ with probability at least $1 - \delta$. □

Lemma 6. *Under the assumptions of Theorem 1, if the event* (12) *occurs simultaneously for all $n = 1, \ldots, N_{\mathrm{ex}}$ and all $k = 1, \ldots, \max(C_n)$, and $\Delta > 0$, (i.e., if there is a unique optimal policy), then all suboptimal policies are eliminated after at most N'_{ex} tasks, where*

$$N'_{\mathrm{ex}} \le \frac{288\, K^2 \ln(4KN_{\mathrm{ex}}/\delta)}{\Delta^2} + 1 \tag{14}$$

Proof. Note first that (12) implies, for all $n \ge 2K^2 \ln(4KN_{\mathrm{ex}}/\delta)$ (guaranteed by line 5) and all $k \in C_n$

$$\begin{aligned}
\frac{\widehat{r}_n^-(k)}{\widehat{c}_n^+(k)} \le \frac{\mathrm{reward}(\pi_k)}{\mathrm{cost}(\pi_k)} \le \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^-(k)} \qquad &\text{if } \widehat{r}_n^+(k) \ge 0 \\
\frac{\widehat{r}_n^-(k)}{\widehat{c}_n^-(k)} \le \frac{\mathrm{reward}(\pi_k)}{\mathrm{cost}(\pi_k)} \le \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^+(k)} \qquad &\text{if } \widehat{r}_n^+(k) < 0
\end{aligned}$$

In other words, the interval

$$\left[\frac{\widehat{r}_n^-(k)}{\widehat{c}_n^+(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) \ge 0\big\} + \frac{\widehat{r}_n^-(k)}{\widehat{c}_n^-(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) < 0\big\},\ \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^-(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) \ge 0\big\} + \frac{\widehat{r}_n^+(k)}{\widehat{c}_n^+(k)}\mathbb{I}\big\{\widehat{r}_n^+(k) < 0\big\}\right]$$

is a confidence interval for the value reward(π_k)/cost(π_k) that measures the performance of π_k. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\text{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^+(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^+(k) - (k-1)\varepsilon_n \tag{15}$$

If $\widehat{r}_n^+(k) \geq 0$, by the definitions in (15), the length of this confidence interval is

$$\frac{\overline{r}_n(k) + 2\varepsilon_n}{\overline{c}_n(k) - (k-1)\varepsilon_n} - \frac{\overline{r}_n(k) - 2\varepsilon_n}{\overline{c}_n(k) + (k-1)\varepsilon_n} = \frac{2\varepsilon_n\big(2\,\overline{c}_n(k) + (k-1)\,\overline{r}_n(k)\big)}{\overline{c}_n(k)^2 - (k-1)^2\,\varepsilon_n^2} \leq 12\,K\varepsilon_n$$

where for the numerator we used the fact that $\overline{c}_n(k)$ (resp., $\overline{r}_n(k)$) is an average of random variables all upper bounded by k (resp., 1) and the denominator is lower bounded by $1/2$ because $\overline{c}_n(k)^2 \geq 1$, $(k^2-1)\,\varepsilon_n^2 \leq 1/2$ by $n \geq 2K^2\ln(4KN_{\text{ex}}/\delta)$ (line 4), and $k/K \leq 1$ (by monotonicity of $k \mapsto k$). Similarly, if $\widehat{r}_n^+(k) < 0$, the length of the confidence interval is

$$\frac{\overline{r}_n(k) + 2\varepsilon_n}{\overline{c}_n(k) + (k-1)\varepsilon_n} - \frac{\overline{r}_n(k) - 2\varepsilon_n}{\overline{c}_n(k) - (k-1)\varepsilon_n} = \frac{2\varepsilon_n\big(2\,\overline{c}_n(k) - (k-1)\,\overline{r}_n(k)\big)}{\overline{c}_n(k)^2 - (k-1)^2\,\varepsilon_n^2} \leq 12\,K\varepsilon_n$$

where, in addition to the considerations above, we used $0 < -\widehat{r}_n^+(k) < -\overline{r}_n(k) \leq 1$. Hence, as soon as the upper bound $12\,K\varepsilon_n$ on the length of each of the confidence interval above falls below $\Delta/2$, all such intervals are guaranteed to be disjoint and by definition of C_n (line 5), all suboptimal policies are guaranteed to have left C_{n+1}. In formulas, this happens at the latest during task n, where $n \geq 2K^2\ln(4KN_{\text{ex}}/\delta)$ satisfies

$$12\,K\varepsilon_n < \frac{\Delta}{2} \iff n > 288\,(K/\Delta)^2\ln(4KN_{\text{ex}}/\delta)$$

This proves the result. □

Lemma 7. *Under the assumptions of Theorem 1, if the event* (12) *occurs simultaneously for all* $n = 1, \ldots, N_{\text{ex}}$ *and all* $k = 1, \ldots, \max(C_n)$*, and the test at line 6 is true for some* $N'_{\text{ex}} \leq N_{\text{ex}}$*, then*

$$R_N \leq \min\left(\frac{(2K+1)N_{\text{ex}}}{N},\ \frac{(2K+1)\big(288\,(K/\Delta)^2\ln(4KN_{\text{ex}}/\delta) + 1\big)}{N}\right) \tag{16}$$

Proof. Note that if the test at line 6 is true, than by (12) there exists a unique optimal policy, i.e., we have $\Delta > 0$. We can therefore apply Lemma 6, obtaining a deterministic upper bound N''_{ex} on the number N'_{ex} of tasks needed to identify the optimal policy, where

$$N''_{\text{ex}} = \min\left(N_{\text{ex}},\ \frac{128\,K^2\ln(4KN_{\text{ex}}/\delta)}{\Delta^2} + 1\right)$$

The total expected reward of Algorithm 1 divided by its total expected cost is lower bounded by

$$\xi = \frac{\mathbb{E}\left[-N'_{\text{ex}} + \sum_{n=N'_{\text{ex}}+1}^{N}\text{reward}(\pi_{k^\star}, \mu_n)\right]}{\mathbb{E}\left[2\sum_{m=1}^{N'_{\text{ex}}}\max(C_m) + \sum_{n=N'_{\text{ex}}+1}^{N}\text{cost}(\pi_{k^\star}, \mu_n)\right]}$$

If $\xi < 0$, we can further lower bound it by

$$\frac{(N - N''_{\text{ex}})\,\text{reward}(\pi_{k^\star}) - N''_{\text{ex}}}{(N - N''_{\text{ex}})\,\text{cost}(\pi_{k^\star}) + 2N''_{\text{ex}}} \geq \frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - \frac{3N''_{\text{ex}}}{N}$$

where the inequality follows by $(a-b)/(c+d) \geq a/c - (d+b)/(c+d)$ for all $a, b, c, d \in \mathbb{R}$ with $0 \neq c > -d$ and $a/c \leq 1$, and then using $c + d \geq N$ which holds because $\text{cost}(\pi_{k^\star}) \geq 1$. Similarly, if $\xi \geq 0$, we can further lower bound it by

$$\frac{(N - N''_{\text{ex}})\,\text{reward}(\pi_{k^\star}) - N''_{\text{ex}}}{(N - N''_{\text{ex}})\,\text{cost}(\pi_{k^\star}) + 2KN''_{\text{ex}}} \geq \frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - \frac{(2K+1)N''_{\text{ex}}}{N}$$

Thus, the result follows by $K \geq 1$ and the definition of N''_{ex}. □

Lemma 8. *Under the assumptions of Theorem 1, if the event* (12) *occurs simultaneously for all* $n = 1, \dots, N_{\text{ex}}$ *and all* $k = 1, \dots, \max(C_n)$*, and the test at line 6 is false for all tasks* $n \le N_{\text{ex}}$ *(i.e., if line 7 is executed with* $C_{N_{\text{ex}}+1}$ *containing two or more policies), then*

$$R_T \le (K+1)\sqrt{\frac{8\ln(4KN_{\text{ex}}/\delta)}{N_{\text{ex}}}} + \frac{(2K+1)N_{\text{ex}}}{N}$$

Proof. Note first that by (12) and the definition of C_n (line 5), all optimal policies belong to $C_{N_{\text{ex}}+1}$. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\text{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^{+}(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^{+}(k) - (k-1)\varepsilon_n \tag{17}$$

By (12) and the definitions of k', $\widehat{r}_n^{\pm}(k)$, and ε_n (line 7, (5), (5), and (17) respectively), for all optimal policies $\pi_{k^\star}$, if $\widehat{r}_{N_{\text{ex}}}^{+}(k^\star) \ge 0$, then also $\widehat{r}_{N_{\text{ex}}}^{+}(k') \ge 0$[5] and

$$\frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} \le \frac{\widehat{r}_{N_{\text{ex}}}^{+}(k^\star)}{\widehat{c}_{N_{\text{ex}}}^{-}(k^\star)} \le \frac{\widehat{r}_{N_{\text{ex}}}^{+}(k')}{\widehat{c}_{N_{\text{ex}}}^{-}(k')} \le \frac{\text{reward}(\pi_{k'}) + 4\varepsilon_n}{\text{cost}(\pi_{k'}) - 2(k'-1)\varepsilon_n}$$
$$\le \frac{\text{reward}(\pi_{k'})}{\text{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\text{cost}(\pi_{k'}) - 2(k'-1)\varepsilon_n}$$

where all the denominators are positive because $N_{\text{ex}} \ge 8(K-1)^2\ln(4KN_{\text{ex}}/\delta)$ and the last inequality follows by $(a+b)/(c-d) \le a/c + (d+b)/(c-d)$ for all $a \le 1$, $b \in \mathbb{R}$, $c \ge 1$, and $d < c$; next, if $\widehat{r}_{N_{\text{ex}}}^{+}(k^\star) < 0$ but $\widehat{r}_{N_{\text{ex}}}^{+}(k') \ge 0$ the exact same chain of inequalities hold; finally, if both $\widehat{r}_{N_{\text{ex}}}^{+}(k^\star) < 0$ and $\widehat{r}_{N_{\text{ex}}}^{+}(k') < 0$, then $\widehat{r}_{N_{\text{ex}}}^{+}(k) < 0$ for all $k \in C_{N_{\text{ex}}+1}$[6], hence, by definition of k' and the same arguments used above

$$\frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} \le \frac{\widehat{r}_{N_{\text{ex}}}^{+}(k^\star)}{\widehat{c}_{N_{\text{ex}}}^{+}(k^\star)} \le \frac{\widehat{r}_{N_{\text{ex}}}^{+}(k')}{\widehat{c}_{N_{\text{ex}}}^{+}(k')} \le \frac{\text{reward}(\pi_{k'}) + 4\varepsilon_n}{\text{cost}(\pi_{k'}) + 2(k'-1)\varepsilon_n}$$
$$\le \frac{\text{reward}(\pi_{k'})}{\text{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\text{cost}(\pi_{k'}) + 2(k'-1)\varepsilon_n} \le \frac{\text{reward}(\pi_{k'})}{\text{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\text{cost}(\pi_{k'}) - 2(k'-1)\varepsilon_n}$$

That is, for all optimal policies $\pi_{k^\star}$, the policy $\pi_{k'}$ run at line 7 satisfies

$$\text{reward}(\pi_{k'}) \ge \text{cost}(\pi_{k'})\left(\frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - \frac{2(k'+1)\varepsilon_n}{\text{cost}(\pi_{k'}) - 2(k'-1)\varepsilon_n}\right)$$
$$\ge \text{cost}(\pi_{k'})\left(\frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - 4(K+1)\varepsilon_n\right)$$

where in the last inequality we lower bounded the denominator by $1/2$ using $\text{cost}(\pi_{k'}) \ge 1$ and $\varepsilon_n \le \varepsilon_{N_{\text{ex}}} \le 1/2$ which follows by $n \ge N_{\text{ex}} \ge 8K^2\ln(4KN_{\text{ex}}/\delta)$ and the monotonicity of $k \mapsto k$. Therefore, for all optimal policies $\pi_{k^\star}$, the total expected reward of Algorithm 1 divided by its total expected cost (i.e., the negative

[5] Indeed, $k' \in \operatorname{argmax}_{k \in C_{N_{\text{ex}}+1}}\big(\widehat{r}_{N_{\text{ex}}}^{+}(k)/\widehat{c}_{N_{\text{ex}}}^{-}(k)\big)$ in this case, and $\widehat{r}_{N_{\text{ex}}}^{+}(k') \ge 0$ follows by the two inequalities $\widehat{r}_{N_{\text{ex}}}^{+}(k')/\widehat{c}_{N_{\text{ex}}}^{-}(k') \ge \widehat{r}_{N_{\text{ex}}}^{+}(k^\star)/\widehat{c}_{N_{\text{ex}}}^{-}(k^\star) \ge 0$.

[6] Otherwise k' would belong to the set $\operatorname{argmax}_{k \in C_{N_{\text{ex}}+1}}\big(\widehat{r}_{N_{\text{ex}}}^{+}(k)/\widehat{c}_{N_{\text{ex}}}^{-}(k)\big)$ which in turn would be included in the set $\big\{k \in C_{N_{\text{ex}}+1} : \widehat{r}_{N_{\text{ex}}}^{+}(k) \ge 0\big\}$ and this would contradict the fact that $\widehat{r}_{N_{\text{ex}}}^{+}(k') < 0$.

addend in (2)) is at least

$$
\begin{aligned}
&\frac{\mathbb{E}\big[-N_{\text{ex}} + (N - N_{\text{ex}})\,\text{reward}(\pi_{k'})\big]}{\mathbb{E}\big[2\sum_{n=1}^{N_{\text{ex}}} \max(C_n) + (N - N_{\text{ex}})\,\text{cost}(\pi_{k'})\big]} \\
&\qquad \geq \frac{-N_{\text{ex}}}{2\sum_{n=1}^{N_{\text{ex}}} \mathbb{E}\big[\max(C_n)\big] + (N - N_{\text{ex}})\,\mathbb{E}\big[\text{cost}(\pi_{k'})\big]} \\
&\qquad + \frac{(N - N_{\text{ex}})\,\mathbb{E}\big[\text{cost}(\pi_{k'})\big]}{2\sum_{n=1}^{N_{\text{ex}}} \mathbb{E}\big[\max(C_n)\big] + (N - N_{\text{ex}})\,\mathbb{E}\big[\text{cost}(\pi_{k'})\big]} \left(\frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - 4(K+1)\varepsilon_n\right) \\
&\qquad \geq \frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - 4(K+1)\varepsilon_n - \frac{N_{\text{ex}} + 2\sum_{n=1}^{N_{\text{ex}}} \mathbb{E}\big[\max(C_n)\big]}{2\sum_{n=1}^{N_{\text{ex}}} \mathbb{E}\big[\max(C_n)\big] + (N - N_{\text{ex}})\,\mathbb{E}\big[\text{cost}(\pi_{k'})\big]} \\
&\qquad \geq \frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} - 4(K+1)\varepsilon_n - \frac{(2K+1)N_{\text{ex}}}{N}
\end{aligned}
$$

where we used $\frac{a}{b+a}(x-y) \geq x - y - \frac{b}{b+a}$ for all $a, b, y > 0$ and all $x \leq 1$ to lower bound the third line, then the monotonicity of $k \mapsto k$ and $2\mathbb{E}\big[\max(C_n)\big] \geq \mathbb{E}\big[\text{cost}(\pi_{k'})\big] \geq 1$ for the last inequality. Rearranging the terms of the first and last hand side in the previous display, using the monotonicity of $k \mapsto k$, and plugging in the value of ε_n, gives

$$
R_T \leq 4(K+1)\varepsilon_n + \frac{(2K+1)N_{\text{ex}}}{N} = (K+1)\sqrt{\frac{8\ln(4KN_{\text{ex}}/\delta)}{N_{\text{ex}}}} + \frac{(2K+1)N_{\text{ex}}}{N}
$$

□

C A Technical Lemma for Theorem 4

In this section, we give a formal proof for a result needed to prove Theorem 4.

Lemma 2. *Let Π be a countable set of policies. If ESC is run with $\delta \in (0,1)$, $\varepsilon_1, \varepsilon_2, \ldots \in (0,1]$, and halts returning K, then $k^\star \leq K$ for all optimal policies $\pi_{k^\star}$ with probability at least $1-\delta$.*

Proof. Note fist that $\widehat{r}_{2^j}^- + 2\varepsilon_j$ (line 2) is an empirical average of m_j i.i.d. unbiased estimators of $\text{reward}(\pi_{2^j})$. Indeed, being $\text{accept}(k, \boldsymbol{x})$ independent of the variables $(x_{k+1}, x_{k+2}, \ldots)$ by definition of duration and the conditional independence of the samples (recall the properties of samples in step 4 of our online protocol, Section 3), for all tasks n performed at line 2 during iteration j and all $i > 2^j$,

$$
\begin{aligned}
\mathbb{E}\left[X_{n,i}\,\text{accept}\big(\tau_{2^j}(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\middle|\, \mu_n\right] &= \mathbb{E}\left[X_{n,i} \mid \mu_n\right] \mathbb{E}\Big[\text{accept}\big(\tau_{2^j}(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big] \\
&= \mu_n\,\mathbb{E}\Big[\text{accept}\big(\tau_{2^j}(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big] = \mathbb{E}\Big[\mu_n\,\text{accept}\big(\tau_{2^j}(\boldsymbol{X}_n), \boldsymbol{X}_n\big) \,\Big|\, \mu_n\Big]
\end{aligned}
$$

Taking expectations to both sides proves the claim. Thus, Hoeffding's inequality implies

$$
\mathbb{P}\Big(\widehat{r}_{2^j}^- > \text{reward}(\pi_{2^j})\Big) = \mathbb{P}\Big(\big(\widehat{r}_{2^j}^- + 2\varepsilon_j\big) - \text{reward}(\pi_{2^j}) > 2\varepsilon_j\Big) \leq \frac{\delta}{j(j+1)}
$$

for all $j \leq j_0$. Similarly, for all $l > j_0$, $\mathbb{P}\big(\overline{c}_{2^l} - \text{cost}(\pi_{2^l}) > 2^l\,\varepsilon_l\big) \leq \frac{\delta}{l(l+1)}$. Hence, the event

$$
\big\{\widehat{r}_{2^j}^- \leq \text{reward}(\pi_{2^j})\big\} \;\wedge\; \big\{\overline{c}_{2^l} \leq \text{cost}(\pi_{2^l})) + 2^l\,\varepsilon_l\big\} \qquad \forall j \leq j_0, \forall l > j_0 \tag{18}
$$

occurs with probability at least

$$
1 - \sum_{j=1}^{j_0} \frac{\delta}{j(j+1)} - \sum_{l=j_0+1}^{j_1} \frac{\delta}{l(l+1)} \geq 1 - \delta \sum_{j \in \mathbb{N}} \frac{1}{j(j+1)} = 1 - \delta
$$

Note now that for each policy π_k with $\text{reward}(\pi_k) \geq 0$ and each optimal policy $\pi_{k^\star}$,

$$\frac{\text{reward}(\pi_k)}{k} \leq \frac{\text{reward}(\pi_k)}{\text{cost}(\pi_k)} \leq \frac{\text{reward}(\pi_{k^\star})}{\text{cost}(\pi_{k^\star})} \leq \frac{1}{\text{cost}(\pi_{k^\star})} \tag{19}$$

Hence, all optimal policies $\pi_{k^\star}$ satisfy $\text{cost}(\pi_{k^\star}) \leq k/\text{reward}(\pi_k)$ for all policies π_k such that $\text{reward}(\pi_k) > 0$. Being durations sorted by index, for all $k \leq h$

$$\text{cost}(\pi_k) = \mathbb{E}\big[\text{cost}(\pi_k, \mu_n)\big] \leq \mathbb{E}\big[\text{cost}(\pi_h, \mu_n)\big] = \text{cost}(\pi_h) \tag{20}$$

Thus, with probability at least $1 - \delta$, for all $k > K$

$$\text{cost}(\pi_k) \overset{(20)}{\geq} \text{cost}(\pi_K) \overset{(18)}{\geq} \overline{c}_K - K\,\varepsilon_{\log_2 K} \overset{\text{line } 6}{>} \frac{k_0}{\widehat{r}^-_{k_0}} \geq \frac{k_0}{\text{reward}(k_0)}$$

where $\text{reward}(k_0) \geq \widehat{r}^-_{k_0} > 0$ by (18) and line (3); i.e., π_k do not satisfy (19). Therefore, with probability at least $1 - \delta$, all optimal policies $\pi_{k^\star}$ satisfy $k^\star \leq K$. □

D Choice of Performance Measure

In this section, we discuss our choice of measuring the performance of policies π with

$$\frac{\sum_{n=1}^N \mathbb{E}\big[\text{reward}(\pi, \mu_n)\big]}{\sum_{m=1}^N \mathbb{E}\big[\text{cost}(\pi, \mu_m)\big]} = \frac{\text{reward}(\pi)}{\text{cost}(\pi)}$$

We compare several different benchmarks and investigate the differences if the agent had a budget of samples and a variable number of tasks, rather than the other way around. We will show that all "natural" choices essentially go in the same direction, except for one (perhaps the most natural) which turns out to be the worst.

At a high level, an agent constrained by a budget would like to maximize its ROI. This can be done in several different ways. If the constraint is on the number N of tasks, then the agent could aim at maximizing (over $\pi = (\tau, \text{accept}) \in \Pi$) the objective $g_1(\pi, N)$ defined by

$$g_1(\pi, N) = \mathbb{E}\left[\frac{\sum_{n=1}^N \text{reward}(\pi, \mu_n)}{\sum_{m=1}^N \text{cost}(\pi, \mu_m)}\right]$$

This is equivalent to the maximization of the ratio

$$\frac{\text{reward}(\pi)}{\text{cost}(\pi)} = \frac{\mathbb{E}\big[\text{reward}(\pi, \mu_n)\big]}{\mathbb{E}\big[\text{cost}(\pi, \mu_n)\big]}$$

in the sense that, multiplying both the numerator and the denominator in $g_1(\pi, N)$ by $1/N$ and applying Hoeffding's inequality, we get $g_1(\pi, N) = \Theta\big(\text{reward}(\pi)/\text{cost}(\pi)\big)$. Furthermore, by the law of large numbers and Lebesgue's dominated convergence theorem, $g_1(\pi, N) \to \text{reward}(\pi)/\text{cost}(\pi)$ when $N \to \infty$ for any $\pi \in \Pi$.

Assume now that the constraint is on the total number of samples instead. We say that the agent has a *budget of samples* T if as soon as the total number of samples reaches T during task N (which is now a random variable), the agent has to interrupt the run of the current policy, reject the current value μ_N, and end the process. Formally, the random variable N that counts the total number of tasks performed by repeatedly running a policy $\pi = (\tau, \text{accept})$ is defined by

$$N = \min\left\{ m \in \mathbb{N} \,\middle|\, \sum_{n=1}^m \tau(\boldsymbol{X}_n) \geq T \right\}$$

In this case, the agent could aim at maximizing the objective

$$g_2(\pi, T) = \mathbb{E}\left[\frac{\sum_{n=1}^{N-1} \text{reward}(\pi, \mu_n)}{T}\right]$$

where the sum is 0 if $N = 1$ and it stops at $N - 1$ because the the last task is interrupted and no reward is gained. As before, assume that $\tau \leq D$, for some $D \in \mathbb{N}$. Note first that by the independence of μ_n and $\boldsymbol{X}_n$ from past tasks, for all deterministic functions f and all $n \in \mathbb{N}$, the two random variables $f(\mu_n, \boldsymbol{X}_n)$ and $\mathbb{I}\{N \geq n\}$ are independent, because $\mathbb{I}\{N \geq n\} = \mathbb{I}\big\{\sum_{i=1}^{n-1} \tau(\boldsymbol{X}_i) < T\big\}$ depends only on the random variables $\tau(\boldsymbol{X}_1), \ldots, \tau(\boldsymbol{X}_{n-1})$. Hence

$$\mathbb{E}\Big[\text{reward}(\pi, \mu_n)\, \mathbb{I}\{N \geq n\}\Big] = \text{reward}(\pi)\, \mathbb{P}(N \geq n)$$
$$\mathbb{E}\big[\text{cost}(\pi, \mu_n)\, \mathbb{I}\{N \geq n\}\big] = \text{cost}(\pi)\, \mathbb{P}(N \geq n)$$

Moreover, note that during each task at least one sample is drawn, hence $N \leq T$ and

$$\sum_{n=1}^{\infty} \mathbb{E}\Big[\big|\text{reward}(\pi, \mu_n)\big|\, \mathbb{I}\{N \geq n\}\Big] \leq \sum_{n=1}^{T} \mathbb{E}\Big[\big|\text{reward}(\pi, \mu_n)\big|\Big] \leq T < \infty$$
$$\sum_{n=1}^{\infty} \mathbb{E}\big[\text{cost}(\pi, \mu_n)\, \mathbb{I}\{N \geq n\}\big] \leq \sum_{n=1}^{T} \mathbb{E}\big[\text{cost}(\pi, \mu_n)\big] = T\, \text{cost}(\pi) \leq TD < \infty$$

We can therefore apply Wald's identity [31] to deduce

$$\mathbb{E}\left[\sum_{n=1}^{N} \text{reward}(\pi, \mu_n)\right] = \mathbb{E}[N]\, \text{reward}(\pi) \qquad \text{and} \qquad \mathbb{E}\left[\sum_{n=1}^{N} \text{cost}(\pi, \mu_n)\right] = \mathbb{E}[N]\, \text{cost}(\pi)$$

which, together with

$$\mathbb{E}\left[\sum_{n=1}^{N} \text{cost}(\pi, \mu_n)\right] \geq T \geq \mathbb{E}\left[\sum_{n=1}^{N} \text{cost}(\pi, \mu_n)\right] - D$$

and

$$\mathbb{E}\left[\sum_{n=1}^{N} \text{reward}(\pi, \mu_n)\right] - 1 \leq \mathbb{E}\left[\sum_{n=1}^{N-1} \text{reward}(\pi, \mu_n)\right] \leq \mathbb{E}\left[\sum_{n=1}^{N} \text{reward}(\pi, \mu_n)\right] + 1$$

yields

$$\frac{\mathbb{E}[N]\, \text{reward}(\pi) - 1}{\mathbb{E}[N]\, \text{cost}(\pi)} \leq g_2(\pi, T) \leq \frac{\mathbb{E}[N]\, \text{reward}(\pi) + 1}{\mathbb{E}[N]\, \text{cost}(\pi) - D}$$

if the denominator on the right-hand side is positive, which happens as soon as $T > D^2$ by $ND \geq \sum_{n=1}^{N} \tau(\boldsymbol{X}_n) \geq T$ and $\text{cost}(\pi) \geq 1$. I.e., $g_2(\pi, T) = \Theta\big(\text{reward}(\pi)/\text{cost}(\pi)\big)$ and noting that $\mathbb{E}[N] \geq T/D \to \infty$ if $T \to \infty$, we have once more that $g_2(\pi, T) \to \text{reward}(\pi)/\text{cost}(\pi)$ when $T \to \infty$ for any $\pi \in \Pi$.

This proves that having a budget of tasks, samples, or using any of the three natural objectives introduced so far is essentially the same.

Before concluding the section, we go back to the original setting and discuss a very natural definition of objective which should be avoided because, albeit easier to maximize, it is not well-suited for this problem. Consider as objective the average payoff of accepted values per amount of time used to make the decision, i.e.,

$$g_3(\pi) = \mathbb{E}\left[\frac{\text{reward}(\pi, \mu_n)}{\text{cost}(\pi, \mu_n)}\right]$$

We give some intuition on the differences between the ratio of expectations and the expectation of the ratio g_3 using the concrete example (4) and we make a case for the former being better than the latter.

More precisely, if N decision tasks have to be performed by the agent, consider the natural policy class $\{\pi_k\}_{k\in\{1,\dots,K\}} = \big\{(\tau_k, \text{accept})\big\}_{k\in\{1,\dots,K\}}$ given by

$$\tau_k(\boldsymbol{x}) = \min\left(k,\ \inf\left\{n \in \mathbb{N} : |\overline{x}_n| \geq c\sqrt{\frac{\ln \frac{KN}{\delta}}{n}}\right\}\right), \quad \text{accept}(n, \boldsymbol{x}) = \mathbb{I}\left\{\overline{x}_n \geq c\sqrt{\frac{\ln \frac{KN}{\delta}}{n}}\right\}$$

for some $c > 0$ and $\delta \in (0,1)$, where $\overline{x}_n = (1/n)\sum_{i=1}^n x_i$ is the average of the first n elements of the sequence $\boldsymbol{x} = (x_1, x_2, \dots)$.

If $K \gg 1$, there are numerous policies in the class with a large cap. For concreteness, consider the last one (τ_K, accept) and let $k = \lceil c^2 \ln(KN/\delta)\rceil$. If μ_n is uniformly distributed on $\{-1, 0, 1\}$, then

$$\Big(\tau_K(\boldsymbol{X}_0), \text{accept}\big(\tau_K(\boldsymbol{X}_0), \boldsymbol{X}_0\big)\Big) = \begin{cases} (k,1) & \text{if } \mu_1 = 1 \\ (k,0) & \text{if } \mu_1 = -1 \\ (K,0) & \text{if } \mu_1 = 0 \end{cases}$$

i.e., the agent understands quickly (drawing only k samples) that $\mu_n = \pm 1$, accepting it or rejecting it accordingly, but takes exponentially longer ($K \gg k$ samples) to figure out that the value is nonpositive when $\mu_n = 0$. The fact that for a constant fraction of tasks (1/3 of the total) π invests a long time (K samples) to earn no reward makes it a very poor choice of policy. This is not reflected in the value of $g_3(\pi_K)$ but it is so in $\text{reward}(\pi_K)/\text{cost}(\pi_K)$. Indeed, in this instance

$$\mathbb{E}\left[\frac{\text{reward}(\pi_K, \mu_n)}{\text{cost}(\pi_K, \mu_n)}\right] = \Theta\left(\frac{1}{k}\right) \qquad \gg \qquad \Theta\left(\frac{1}{K}\right) = \frac{\text{reward}(\pi_K)}{\text{cost}(\pi_K)}$$

This is due to the fact that the expectation of the ratio "ignores" outcomes with null (or very small) rewards, even if a large number of samples is needed to learn them. On the other hand, the ratio of expectations weighs the total number of requested samples and it is highly influenced by it, a property we are interested to capture within our model.

E An Impossibility Result

We conclude the paper by showing that, in general, given μ_n it is impossible to define an unbiased estimator of the reward of all policies using only the samples drawn by the policies themselves, unless μ_n is known beforehand.

Take a policy $\pi_1 = (1, \text{accept})$ that draws exactly one sample. Note that such a policy is included in all sets of policies Π so this is by no means a pathological example. As before, assume for the sake of simplicity that samples take values in $\{-1, 1\}$ and consider any decision function accept such that $\text{accept}(1, \boldsymbol{x}) = (1 + x_1)/2$ for all $\boldsymbol{x} = (x_1, x_2, \dots)$. In words, the policy π_1 looks at one single sample $x_1 \in \{-1, 1\}$ and accepts if and only if $x_1 = 1$. As discussed earlier (Section 2, Repeated A/B testing, and Section D, where μ is concentrated around $[-1, 0] \cup \{1\}$), there are settings in which this policy is optimal, so this choice of decision function cannot be dismissed as a mathematical pathology.

The following lemma shows that in the simple, yet meaningful case of the policy π_1 described above, it is impossible to define an unbiased estimator of its expected reward given μ_n

$$\mathbb{E}\big[\mu_n \,\text{accept}(1, \boldsymbol{X}_n) \mid \mu_n\big] = \mu_n\, \mathbb{E}\left[\frac{1 + X_{n,1}}{2} \mid \mu_n\right] = \frac{\mu_n + \mu_n^2}{2}$$

using only $X_{n,1}$, unless μ_n is known beforehand.

Lemma 9. *Let $\tilde{X}$ be a $\{-1,1\}$-valued random variable with $\mathbb{E}[\tilde{X}] = \tilde{\mu}$, for some real number $\tilde{\mu}$. If there exists an unbiased estimator $f(\tilde{X})$ of $\big(\tilde{\mu} + \tilde{\mu}^2\big)/2$, for some $f\colon \{-1,1\} \to \mathbb{R}$, then f satisfies*

$$\begin{cases} f(-1) = 0 & \text{if } \tilde{\mu} = -1 \\ f(1) = \tilde{\mu} - f(-1)\dfrac{1-\tilde{\mu}}{1+\tilde{\mu}} & \text{if } \tilde{\mu} \neq -1 \end{cases}$$

i.e., to define any such f (thus, any unbiased estimator of $\big(\tilde{\mu} + \tilde{\mu}^2\big)/2$) it is necessary to know $\tilde{\mu}$.

Proof. From $\mathbb{E}[\tilde{X}] = 1 \cdot \mathbb{P}(\tilde{X} = 1) + (-1) \cdot \mathbb{P}(\tilde{X} = -1) = -1 + 2\mathbb{P}(\tilde{X} = 1)$ and our assumption $\mathbb{E}[\tilde{X}] = \tilde{\mu}$, we obtain $\mathbb{P}(\tilde{X} = 1) = (1 + \tilde{\mu})/2$.

Let $f\colon \{-1,1\} \to \mathbb{R}$ be any function satisfying $\mathbb{E}\big[f(\tilde{X})\big] = \big(\tilde{\mu} + \tilde{\mu}^2\big)/2$. Then, from the law of the unconscious statistician

$$\mathbb{E}\big[f(\tilde{X})\big] = f(1)\mathbb{P}(\tilde{X} = 1) + f(-1)\mathbb{P}(\tilde{X} = -1) = f(1)\frac{1+\tilde{\mu}}{2} + f(-1)\frac{1-\tilde{\mu}}{2}$$

and our assumption $\mathbb{E}\big[f(\tilde{X})\big] = \big(\tilde{\mu} + \tilde{\mu}^2\big)/2$, we obtain

$$f(1)(1+\tilde{\mu}) + f(-1)(1-\tilde{\mu}) = \tilde{\mu} + \tilde{\mu}^2$$

Thus, if $\tilde{\mu} = -1$, we have $f(-1) = 0$. Otherwise, solving for $f(1)$ gives the result. □

References

[1] Saeed Alaei, MohammadTaghi Hajiaghayi, and Vahid Liaghat. 2012. Online prophet-inequality matching with applications to ad allocation. In *Proceedings of the 13th ACM Conference on Electronic Commerce*. Association for Computing Machinery, New York, NY, USA, 18–35.

[2] Noga Alon, Nicolò Cesa-Bianchi, Ofer Dekel, and Tomer Koren. 2015. Online Learning with Feedback Graphs: Beyond Bandits. In *Proceedings of The 28th Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 40)*, Peter Grünwald, Elad Hazan, and Satyen Kale (Eds.). PMLR, Paris, France, 23–35.

[3] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

[4] Eduardo M. Azevedo, Alex Deng, Jose Luis Montiel Olea, Justin Rao, and E. Glen Weyl. 2018. The A/B Testing Problem. In *Proceedings of the 2018 ACM Conference on Economics and Computation* (Ithaca, NY, USA) *(EC '18)*. Association for Computing Machinery, New York, NY, USA, 461–462. `https://doi.org/10.1145/3219166.3219204`

[5] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2018. Bandits with knapsacks. *Journal of the ACM (JACM)* 65, 3 (2018), 1–55.

[6] Sébastien Bubeck and Nicolo Cesa-Bianchi. 2012. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning* 5, 1 (2012), 1–122.

[7] Christopher F Chabris, Carrie L Morris, Dmitry Taubinsky, David Laibson, and Jonathon P Schuldt. 2009. The allocation of time in decision-making. *Journal of the European Economic Association* 7, 2-3 (2009), 628–637.

[8] Shiyun Chen and Shiva Kasiviswanathan. 2020. Contextual Online False Discovery Rate Control. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, Online, 952–961.

[9] Jose Correa, Patricio Foncea, Ruben Hoeksma, Tim Oosterwijk, and Tjark Vredeveld. 2019. Recent developments in prophet inequalities. *ACM SIGecom Exchanges* 17, 1 (2019), 61–70.

[10] Hossein Esfandiari, MohammadTaghi HajiAghayi, Brendan Lucier, and Michael Mitzenmacher. 2019. Online pandora's boxes and bandits. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (2019), 1885–1892.

[11] P.W. Farris, N. Bendle, P.E. Pfeifer, and D. Reibstein. 2010. *Marketing Metrics: The Definitive Guide to Measuring Marketing Performance*. Pearson Education, New York.

[12] Dale L. Flesher and Gary John Previts. 2013. Donaldson Brown (1885-1965): The power of an individual and his ideas over time. *The Accounting Historians Journal* 40, 1 (2013), 79–101.

[13] Dean P Foster and Robert A Stine. 2008. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, 2 (2008), 429–444.

[14] Christopher R Genovese, Kathryn Roeder, and Larry Wasserman. 2006. False discovery control with p-value weighting. *Biometrika* 93, 3 (2006), 509–524.

[15] Thomas L Griffiths, Frederick Callaway, Michael B Chang, Erin Grant, Paul M Krueger, and Falk Lieder. 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. *Current Opinion in Behavioral Sciences* 29 (2019), 24–30.

[16] Nicholas Hay, Stuart Russell, David Tolpin, and Solomon Eyal Shimony. 2012. Selecting Computations: Theory and Applications. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence* (Catalina Island, CA) *(UAI'12)*. AUAI Press, Arlington, Virginia, USA, 346–355.

[17] Philipp Heesen and Arnold Janssen. 2016. Dynamic adaptive multiple tests with finite sample FDR control. *Journal of Statistical Planning and Inference* 168 (2016), 38–51.

[18] Lalit Jain and Kevin Jamieson. 2018. Firing Bandits: Optimizing Crowdfunding. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 2206–2214.

[19] Adel Javanmard, Andrea Montanari, et al. 2018. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics* 46, 2 (2018), 526–554.

[20] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. 2017. Peeking at A/B Tests: Why It Matters, and What to Do about It. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 1517–1525.

[21] Robert Kleinberg, Bo Waggoner, and E. Glen Weyl. 2016. Descending Price Optimally Coordinates Search. In *Proceedings of the 2016 ACM Conference on Economics and Computation* (Maastricht, The Netherlands) *(EC '16)*. Association for Computing Machinery, New York, NY, USA, 23–24.

[22] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Illinois, USA) *(KDD '13)*. Association for Computing Machinery, New York, NY, USA, 1168–1176.

[23] Ang Li and Rina Foygel Barber. 2019. Multiple testing with the structure-adaptive Benjamini–Hochberg algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81, 1 (2019), 45–74.

[24] Brendan Lucier. 2017. An economic view of prophet inequalities. *ACM SIGecom Exchanges* 16, 1 (2017), 24–47.

[25] Aaditya Ramdas, Fanny Yang, Martin J. Wainwright, and Michael I. Jordan. 2017. Online Control of the False Discovery Rate with Decaying Memory. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 5655–5664.

[26] David S. Robertson and James M. S. Wason. 2018. Online control of the false discovery rate in biomedical research. arXiv:1809.07292 [stat.ME]

[27] Dinah Rosenberg, Eilon Solan, and Nicolas Vieille. 2007. Social learning in one-arm bandit problems. *Econometrica* 75, 6 (2007), 1591–1611.

[28] Sven Schmit, Virag Shah, and Ramesh Johari. 2019. Optimal Testing in the Experiment-rich Regime. In *Proceedings of Machine Learning Research (Proceedings of Machine Learning Research, Vol. 89)*, Kamalika Chaudhuri and Masashi Sugiyama (Eds.). PMLR, Naha, Okinawa, Japan, 626–633.

[29] Aleksandrs Slivkins. 2019. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning* 12, 1-2 (2019), 1–286.

[30] John Wilder Tukey. 1953. The Problem of Multiple Comparisons.

[31] Abraham Wald. 1944. On cumulative sums of random variables. *The Annals of Mathematical Statistics* 15, 3 (1944), 283–296.

[32] Martin L. Weitzman. 1979. Optimal Search for the Best Alternative. *Econometrica* 47, 3 (1979), 641–654.

[33] Fanny Yang, Aaditya Ramdas, Kevin G Jamieson, and Martin J Wainwright. 2017. A framework for Multi-A(rmed)/B(andit) Testing with Online FDR Control. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc., Long Beach, CA, USA, 5957–5966.