Adapting Stochastic Online Decision-Making for ROI Maximization: A Concave Cost Function Approach to Bulk Purchasing in Innovation Testing Scenarios

Mateo Pedro
Georgia Institute Of Technology

Tommaso Cesari
Toulouse School of Economics (TSE)

Vianney Perchet
CREST, ENSAE & Criteo AI Lab, Paris

July 2023

Abstract

We present an adaptation of the theoretical framework for Return On Investment (ROI) maximization in repeated decision-making, introducing a concave cost function to account for volume-based discounts in bulk purchasing of equipment for innovation testing. This altered cost structure adds a new dimension to the exploration cost involved in the original setting, where companies regularly review and decide on the potential of technological innovation proposals. We adapt and modify the CAPE and ESC-CAPE algorithms to handle this nonlinearity in costs, allowing for more realistic and efficient decision-making policies over a sequence of innovation proposals. The revised algorithm incorporates decreasing marginal costs, a common characteristic in scenarios of bulk procurement. It continues to offer convergence to an optimal policy in class Π, taking into account the number of innovations N and the suboptimality gap Δ, with adjusted rates to reflect the altered cost structure. This adaptation broadens the scope of the original framework, providing unbiased estimates of performance even in the context of volume-based discounting and concave cost functions.

1 Introduction

This paper builds upon the research presented in the paper "ROI Maximization in Stochastic Online Decision-Making", and a thorough understanding of the original research is necessary to fully comprehend the adaptations introduced here. We focus on the repeated decision-making model presented in the original paper, where each task in a sequence is associated with a pair $(\mu_n, \mathcal{D}_n)$, representing the true value of the n-th innovation and the feedback on the n-th innovation respectively. The learner can never directly observe this pair, but can draw arbitrarily many i.i.d. samples from $\mathcal{D}_n$ to accumulate information on the unknown value μ_n of the current innovation. Post sampling, the learner decides to either accept or reject the innovation, impacting the ROI.

In the original paper, the primary contribution is the mathematical formalization of the ROI maximization model and the introduction of the Capped Policy Elimination (CAPE) algorithm. CAPE applies to finite policy classes and proves convergence to the optimal policy at rates of $1/(\Delta^2 N)$ and $N^{-1/3}$. Furthermore, for infinitely large policy classes, the authors introduced a preprocessing step leading to the development of the ESC-CAPE algorithm, which converges to the optimal policy in an infinite set at a rate of $N^{-1/3}$.

The focus of this paper is to adapt the aforementioned framework and algorithms to scenarios involving bulk procurement, particularly in situations where vendors offer volume discounts. These situations are characterized by decreasing marginal costs, modeled using a concave cost function. We present adaptations to both the CAPE and ESC-CAPE algorithms to handle this non-linearity in cost structures. These revised algorithms provide an even more realistic and efficient decision-making process for businesses facing bulk purchasing decisions, thus further expanding the versatility of the ROI maximization framework.

2 Related Work

It is of utmost importance to highlight the foundational work presented in the parent article "ROI Maximization in Stochastic Online Decision-Making". This seminal piece of research set the stage for our current investigation and has been influential in its innovative approach to maximizing ROI in repeated decision-making scenarios. It lays the groundwork for understanding the relationship between sequential decisions and ROI, creating a mathematical framework and introducing the CAPE and ESC-CAPE algorithms.

This parent paper meticulously reviews and incorporates the relevant preceding work, creating a clear lineage of the research progression leading up to their findings. Given the substantial amount of related research cited and discussed in the original paper, it is recommended for readers to refer directly to it for a comprehensive overview of previous work in this area. This approach will provide the necessary context and depth of understanding required to fully appreciate the developments presented in our adaptation.

Our work carries forward the spirit of the parent article, extending its implications to the realistic and prevalent scenario of bulk procurement in innovation testing. By tailoring the original algorithms to account for concave cost structures, our research contributes a valuable dimension to this evolving field of study. Therefore, the parent article serves as an essential point of reference not only for its related work but also for the methodological foundation upon which we build our own investigations.

3 New Settings and notations

We introduce a new function CCF(x), which we define as the square root function, a concave function (meaning first derivative decreasing and second derivative negative). This function maps the number of observations $\tau(x)$ made by the learner to gather information on the current innovation to a non-negative real number C. Hence, $\tau(x)$, called duration, maps a sequence of observations x = (x1,x2,...) to an integer d (the no. of observations after which the learner stops gathering info on the current innovation), and CCF(d) calculates the cost incurred in making these observations.

Consequently, the reward and cost obtained by running a policy $\pi_k = (\tau_k, accept)$ on a value μ_n are, respectively,

$$reward(\pi_k, \mu_n) = \mu_n accept(\tau_k(X_n), X_n) \in \{\mu_n, 0\} \; and \; cost(\pi_k, \mu_n) = CCF(\tau_k(X_n) \in R \tag{1}$$

To simplify our calculations and discussions, we will set our concave cost function CCF to be the square root function.

4 CAPE adaptation

Algorithm 1: Capped Policy Elimination (CAPE)

```
Input: finite policy set $\Pi$, number of tasks $N$, confidence parameter $\delta$, exploration cap $N_{\mathrm{ex}}$
Initialization: let $C_1 \leftarrow \{1, ..., K\}$ 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 $X_{n,1}, ..., X_{n,2\max(C_n)}$ 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         $C'_n = \big\{k \in C_n \,:\, \big(\widehat{r}^{\,+}_n(k) \geq 0$ and $\widehat{r}^{\,+}_n(k)/\widehat{c}^{\,-}_n(k) < \widehat{r}^{\,-}_n(j)/\widehat{c}^{\,+}_n(j)$, for some $j \in C_n\big)$
              or $\big(\widehat{r}^{\,+}_n(k) < 0$ and $\widehat{r}^{\,+}_n(k)/\widehat{c}^{\,+}_n(k) < \widehat{r}^{\,-}_n(j)/\widehat{c}^{\,-}_n(j)$, for some $j \in C_n\big)\big\}$

          $$\widehat{r}^{\,\pm}_n(k) = \frac{1}{n}\sum_{m=1}^{n}\sum_{i=1}^{\max(C_m)} \frac{X_{m,\max(C_m)+i}}{\max(C_m)} \,\mathrm{accept}\big(\tau_k(\boldsymbol{X}_m), \boldsymbol{X}_m\big) \pm \sqrt{\frac{2}{n}\ln\frac{4KN_{\mathrm{ex}}}{\delta}} \qquad (2)$$
          $$\widehat{c}^{\,\pm}_n(k) = \frac{1}{n}\sum_{m=1}^{n} \tau_k(\boldsymbol{X}_m) \pm (\sqrt{k}-1)\sqrt{\frac{1}{2n}\ln\frac{4KN_{\mathrm{ex}}}{\delta}} \qquad (3)$$

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

          $$k' \in \begin{cases} \operatorname*{argmax}_{k \in C_{N_{\mathrm{ex}}+1}} \big(\widehat{r}^{\,+}_{N_{\mathrm{ex}}}(k)/\widehat{c}^{\,-}_{N_{\mathrm{ex}}}(k)\big) & \text{if } \widehat{r}^{\,+}_{N_{\mathrm{ex}}}(k) \geq 0 \text{ for some } k \in C_{N_{\mathrm{ex}}+1} \\ \operatorname*{argmax}_{k \in C_{N_{\mathrm{ex}}+1}} \big(\widehat{r}^{\,+}_{N_{\mathrm{ex}}}(k)/\widehat{c}^{\,+}_{N_{\mathrm{ex}}}(k)\big) & \text{if } \widehat{r}^{\,+}_{N_{\mathrm{ex}}}(k) < 0 \text{ for all } k \in C_{N_{\mathrm{ex}}+1} \end{cases} \qquad (4)$$
```

Above is our enhanced adaptation of the CAPE algorithm. For an in-depth understanding and to observe the modifications implemented, we strongly recommend cross-referencing this version with the original one presented in the seminal research paper.

The overall functioning of the CAPE algorithm is kept unchanged. For comprehension purposes the explanation, as written by its author, is copied below:

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 estimates of rewards and costs of all potentially optimal policies and more specifically to build unbiased elimination step, the algorithm oversamples (line 2) by drawing twice as many samples as it would suffice to take its decision accept$(\tau_{max(C_n)}(X_n), X_n)$ (at line 3). These extra samples are used to compute rough 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.

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).*

Moving forward, our primary endeavor is to adapt and recalibrate the underlying lemmas associated with the CAPE algorithm. It is important to note that the proof of Theorem 1 remains unaltered in our adaptation, hence it is not discussed here. Nevertheless, this theorem builds upon a number of critical lemmas which necessitate appropriate modifications to fit into our newly designed framework.

4.1 Lemmas for CAPE adaptation

In this particular subsection, our focus will be on customizing Lemmas 5 through 8, originally presented in Appendix B of the base research. For a comprehensive understanding of the modifications and their rationale, we suggest reviewing this portion in alignment with the original research document.

Lemma 5:

Under the assumptions of Theorem 1, the event

$$\widehat{r}_n^-(k) \leq \text{reward}(\pi_k) \leq \widehat{r}_n^+(k) \qquad \text{and} \qquad \widehat{c}_n^-(k) \leq \text{cost}(\pi_k) \leq \widehat{c}_n^+(k) \tag{5}$$

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

Proof. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\text{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^+(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^+(k) - (\sqrt{k} - 1)\varepsilon_n \tag{6}$$

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

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

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

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

Applying a union bound shows that event (5) occurs simultaneously for all $n \in \{1, ..., N_{\text{ex}}\}$ and $k \in \{1, ..., \max(C_n)\}$ with probability at least $1 - \delta$. □

Lemma 6:

Under the assumptions of Theorem 1, if the event (5) occurs simultaneously for all $n = 1, ..., N_{\text{ex}}$ and all $k = 1, ..., \max(C_n)$, 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

$$N'_{\text{ex}} \leq \frac{288\, K^2 \ln(4KN_{\text{ex}}/\delta)}{\Delta^2} + 1 \tag{7}$$

Proof. Note first that (5) implies, for all $n \geq 2K^2 \ln(4KN_{\text{ex}}/\delta)$ and all $k \in C_n$

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

In other words, the interval

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

is a confidence interval for the value $\text{reward}(\pi_k)/\text{cost}(\pi_k)$ that measures the performance of π_k. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\text{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^{\,+}(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^{\,+}(k) - (\sqrt{k}-1)\varepsilon_n \tag{8}$$

If $\widehat{r}_n^{\,+}(k) \ge 0$, by the definitions in (8), the length of this confidence interval is

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

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 $\overline{c}_n(k)^2 \ge 1$, $(\sqrt{k}^2 - 1)\,\varepsilon_n^2 \le 1/2$ by $n \ge 2K^2\ln(4KN_{\text{ex}}/\delta)$, and $k/K \le 1$ (by monotonicity of $k \mapsto k$). Similarly, if $\widehat{r}_n^{\,+}(k) < 0$, the length of the confidence interval is

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

where, in addition to the considerations above, we used $0 < -\widehat{r}_n^{\,+}(k) < -\overline{r}_n(k) \le 1$. 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 , all suboptimal policies are guaranteed to have left C_{n+1}. In formulas, this happens at the latest during task n, where $n \ge 2K^2\ln(4KN_{\text{ex}}/\delta)$ 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 does not require any adaptation following the change to a concave cost function. Therefore, it will not be discussed in this paper and it is advised to refer to the original paper for comprehensive reasons.

Lemma 8: Under the assumptions of Theorem 1, if the event (5) occurs simultaneously for all $n = 1, ..., N_{\text{ex}}$ and all $k = 1, ..., \max(C_n)$, and the test at line 6 of the algorithm 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 (5) and the definition of C_n (line 5), all optimal policies belong to $C_{N_{\text{ex}}+1}$. Let, for all n, k,

$$\varepsilon_n = \sqrt{\frac{\ln(4KN_{\text{ex}}/\delta)}{2n}}, \qquad \overline{r}_n(k) = \widehat{r}_n^{\,+}(k) - 2\varepsilon_n, \qquad \overline{c}_n(k) = \widehat{c}_n^{\,+}(k) - (\sqrt{k}-1)\varepsilon_n \tag{9}$$

By (5) and the definitions of k', $\widehat{r}_n^{\,\pm}(k)$, and ε_n, for all optimal policies $\pi_{k^\star}$, if $\widehat{r}_{N_{\text{ex}}}^{\,+}(k^\star) \ge 0$, then also $\widehat{r}_{N_{\text{ex}}}^{\,+}(k') \ge 0$[1] and

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

$$\le \frac{\text{reward}(\pi_{k'})}{\text{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\text{cost}(\pi_{k'}) - 2(\sqrt{k'}-1)\varepsilon_n}$$

where all the denominators are positive because $N_{\text{ex}} \ge 8(\sqrt{K}-1)^2\ln(4KN_{\text{ex}}/\delta)$ 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 $\widehat{r}_{N_{\text{ex}}}^{\,+}(k') \ge 0$

[1] 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 $\widehat{r}_{N_{\text{ex}}}^{\,+}(k')/\widehat{c}_{N_{\text{ex}}}^{\,-}(k') \ge \widehat{r}_{N_{\text{ex}}}^{\,+}(k^\star)/\widehat{c}_{N_{\text{ex}}}^{\,-}(k^\star) \ge 0$.

the exact same chain of inequalities hold; finally, if both $\widehat{r}^{+}_{N_{\mathrm{ex}}}(k^\star) < 0$ and $\widehat{r}^{+}_{N_{\mathrm{ex}}}(k') < 0$, then $\widehat{r}^{+}_{N_{\mathrm{ex}}}(k) < 0$ for all $k \in C_{N_{\mathrm{ex}}+1}$[2], hence, by definition of k' and the same arguments used above

$$\begin{aligned}\frac{\mathrm{reward}(\pi_{k^\star})}{\mathrm{cost}(\pi_{k^\star})} \le \frac{\widehat{r}^{+}_{N_{\mathrm{ex}}}(k^\star)}{\widehat{c}^{+}_{N_{\mathrm{ex}}}(k^\star)} \le \frac{\widehat{r}^{+}_{N_{\mathrm{ex}}}(k')}{\widehat{c}^{+}_{N_{\mathrm{ex}}}(k')} &\le \frac{\mathrm{reward}(\pi_{k'}) + 4\varepsilon_n}{\mathrm{cost}(\pi_{k'}) + 2(\sqrt{k'}-1)\varepsilon_n} \\ &\le \frac{\mathrm{reward}(\pi_{k'})}{\mathrm{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\mathrm{cost}(\pi_{k'}) + 2(\sqrt{k'}-1)\varepsilon_n} \le \frac{\mathrm{reward}(\pi_{k'})}{\mathrm{cost}(\pi_{k'})} + \frac{2(k'+1)\varepsilon_n}{\mathrm{cost}(\pi_{k'}) - 2(\sqrt{k'}-1)\varepsilon_n}\end{aligned}$$

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

$$\begin{aligned}\mathrm{reward}(\pi_{k'}) &\ge \mathrm{cost}(\pi_{k'}) \left(\frac{\mathrm{reward}(\pi_{k^\star})}{\mathrm{cost}(\pi_{k^\star})} - \frac{2(k'+1)\varepsilon_n}{\mathrm{cost}(\pi_{k'}) - 2(\sqrt{k'}-1)\varepsilon_n} \right) \\ &\ge \mathrm{cost}(\pi_{k'}) \left(\frac{\mathrm{reward}(\pi_{k^\star})}{\mathrm{cost}(\pi_{k^\star})} - 4(K+1)\varepsilon_n \right)\end{aligned}$$

where in the last inequality we lower bounded the denominator by $1/2$ using $\mathrm{cost}(\pi_{k'}) \ge 1$ and $\varepsilon_n \le \varepsilon_{N_{\mathrm{ex}}} \le 1/2$ which follows by $n \ge N_{\mathrm{ex}} \ge 8K^2 \ln(4KN_{\mathrm{ex}}/\delta)$ 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 is at least

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

where we used $\frac{a}{b+a}(x-y) \ge x - y - \frac{b}{b+a}$ for all $a, b, y > 0$ and all $x \le 1$ to lower bound the third line, then the monotonicity of $k \mapsto k$ and $2\mathbb{E}\big[\max(C_n)\big] \ge \mathbb{E}\big[\mathrm{cost}(\pi_{k'})\big] \ge 1$ 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

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

□

5 ESC-CAPE Adaptation

The authors originally designed the ESC algorithm as a pre-processing mechanism to address a challenge observed with the CAPE (Capped Policy Elimination) algorithm. When the cardinality of the policy set greatly exceeds the number of tasks, the guarantees provided by Theorem 1 regarding the convergence rate of CAPE to the best ROI become ineffective. The ESC algorithm's essential premise is to exploit the relationship between the reward and cost of optimal policies and those with a positive reward. This insight allows the algorithm to control the cost of optimal policies by estimating the reward and cost of any policy with a positive reward. Thus, the ESC algorithm aims to provide a more refined selection of policies for the CAPE algorithm to operate on, effectively controlling the number of samples drawn and ensuring that the learning process remains effective even when the policy set is substantially large.

While our proposed modifications to the cost function seamlessly integrate with the original framework of the

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

CAPE and ESC algorithms, as well as their associated lemmas and theorems, we believe it's essential to present the ESC algorithm and its corresponding theorem in their original forms, as articulated by the authors. This approach not only maintains the integrity of the source material but also provides a comprehensive understanding of the algorithm's nuances. We reiterate that these foundational constructs hold true even with the incorporated cost adjustments, emphasizing the robustness of these algorithms and their theoretical underpinnings.

Algorithm 2: Extension to Countable (ESC)

```
  Input: countable policy set $\Pi$, number of tasks $N$, confidence parameter $\delta$, accuracy levels $(\varepsilon_n)_n$
  Initialization: for all $j$, let $m_j \leftarrow \lceil \ln\big(j(j+1)/\delta\big)/2\varepsilon_j^2 \rceil$ and $M_j = m_1 + ... + m_j$
1 for $j = 1, 2, ...$ do
2     run policy $\big(2 \cdot 2^j, 0\big)$ for $m_j$ tasks and compute $\widehat{r}^-_{2^j} \leftarrow \widehat{r}^-_{2^j}(M_{j-1}, m_j, \varepsilon_j)$
3     if $\widehat{r}^-_{2^j} > 0$ then let $j_0 \leftarrow j$ and $k_0 \leftarrow 2^{j_0}$
4         for $l = j_0 + 1, j_0 + 2, ...$ do
5             run policy $\big(\tau_{2^l}, 0\big)$ for $m_l$ tasks and compute $\overline{c}_{2^l} \leftarrow \overline{c}_{2^l}(M_{l-1}, m_l)$
6             if $\overline{c}_{2^l} > 2^l\, \varepsilon_l + k_0/\widehat{r}^-_{k_0}$ then let $j_1 \leftarrow l$ and return $K \leftarrow 2^{j_1}$
```

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 $\varepsilon_1, \varepsilon_2, l$, and an exploration cap N_{ex}. The joint algorithm runs ESC first with parameters $\Pi, N, \delta, \varepsilon_1, \varepsilon_2, l$. Then, if ESC halts returning K, it runs CAPE with parameters $\big\{(\tau_k, \text{accept})\big\}_{k\in\{1,\ldots,K\}}, N, \delta, N_{\text{ex}}$.

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. 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 2. *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.*

Please refer to the original paper for proofs and lemmas.

6 Conclusion and Openings

In conclusion, the smooth integration of a revised cost function into the existing ROI Maximization in Stochastic Online Decision-Making framework stands testament to the robustness and versatility of the CAPE and ESC CAPE algorithms. This adjustment manifests not only the capacity for these algorithms to accommodate variations, but also underlines their applicability in a diverse array of real-world scenarios.

Our exploration has primarily focused on the implementation of a concave cost function, reflecting situations where economies of scale or bulk discounts may apply. This context is particularly relevant in settings where the marginal cost of additional tests diminishes as the number of tests increases, perhaps due to improved efficiency or the exploitation of bulk buying opportunities.

However, the potential to explore the impact of other cost structures remains. One could consider investigating the effect of a convex cost function, which could simulate the dynamics of Increasing Marginal Costs. This might represent a scenario where, as more tests are conducted, resources such as personnel time or lab availability become increasingly scarce, hence making each additional test more expensive. Such a scenario could be prevalent in many industries, for instance in high-demand lab environments.

Another avenue for future exploration is to consider the impact of random cost shocks. These sudden variations could affect the exploration costs and thus the ultimate decisions made by the learner. The incorporation of such randomness in the cost function could result in algorithms that are more resilient to unexpected changes in the environment, thus making them even more practical and robust.

Finally, I would like to express my sincere gratitude to the original authors of the paper "ROI Maximization in Stochastic Online Decision-Making", whose significant work laid a strong foundation for this project. Their guidance and insights have been instrumental in shaping this research, offering new perspectives on how to approach and solve complex problems in the field of stochastic online decision-making. Their contributions to this area of study are truly invaluable.