Making the most of your day: online learning for optimal allocation of time

Etienne Boursier
Centre Borelli, ENS Paris-Saclay, France
etienne.boursier1@gmail.com

Tristan Garrec
Centre Borelli, ENS Paris-Saclay, France
EDF Lab, Palaiseau, France
tristan.garrec@ut-capitole.fr

Vianney Perchet
CREST, ENSAE Paris, France
Criteo AI Lab, Paris, France
vianney@ensae.fr

Marco Scarsini
Department of Economics and Finance, LUISS, Rome
marco.scarsini@luiss.it

Abstract

We study online learning for optimal allocation when the resource to be allocated is time. An agent receives task proposals sequentially according to a Poisson process and can either accept or reject a proposed task. If she accepts the proposal, she is busy for the duration of the task and obtains a reward that depends on the task duration. If she rejects it, she remains on hold until a new task proposal arrives. We study the regret incurred by the agent, first when she knows her reward function but does not know the distribution of the task duration, and then when she does not know her reward function, either. This natural setting bears similarities with contextual (one-armed) bandits, but with the crucial difference that the normalized reward associated to a context depends on the whole distribution of contexts.

1 Introduction

Motivation. A driver filling her shift with rides, a landlord renting an estate short-term, an independent deliveryman, a single server that can make computations online, a communication system receiving a large number of calls, etc. all face the same trade-off. There is a unique resource that can be allocated to some tasks/clients for some duration. The main constraint is that, once it is allocated, the resource becomes unavailable for the whole duration. As a consequence, if a "better" request arrived during this time, it could not be accepted and would be lost. Allocating the resource for some duration has some cost but generates some rewards – possibly both unknown and random. For instance, an estate must be cleaned up after each rental, thus generating some fixed costs; on the other hand, guests might break something, which explains why these costs are both random and unknown beforehand. Similarly, the relevance of a call is unknown beforehand. Concerning duration, the shorter the request the better (if the net reward is the same). Indeed, the resource could be allocated twice in the same amount of time.

The ideal request would therefore be of short duration and large reward; this maximizes the revenue per time. A possible policy could be to wait for this kind of request, declining the other ones (too long and/or less profitable). On the other hand, such a request could be very rare. So it might be more rewarding in the long run to accept any request, at the risk of "missing" the ideal one.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Some clear trade-offs arise. The first one is between a greedy policy that accepts only the highest profitable requests – at the risk of staying idle quite often – and a safe policy that accepts every request – but unfortunately also the non-profitable ones. The second trade-off concerns the learning phase; indeed, at first and because of the randomness, the actual net reward of a request is unknown and must be learned on the fly. The safe policy will gather a lot of information (possibly at a high cost) while the greedy one might lose some possible valuable information for the long run (in trying to optimize the short term revenue).

We adopt the perspective of an agent seeking to optimize her earned income for some large duration. The agent receives task proposals sequentially, following a Poisson process. When a task is proposed, the agent observes its expected duration and can then either accept or reject it. If she accepts it, she cannot receive any new proposals for the whole duration of the task. At the end of the task she observes her reward, which is a function of the duration of the task. If, on the contrary, she rejects the task, she remains on hold until she receives a new task proposal.

The agent's policies are evaluated in terms of their expected regret, which is the difference between the cumulative rewards obtained until T under the optimal policy and under the implemented agent policy (as usual the total length could also be random or unknown (Degenne and Perchet, 2016)). In this setting, the "optimal" policy is within the class of policies that accept – or not – tasks whose length belongs to some given acceptance set (say, larger than some threshold, or in some specific Borel subset, depending on the regularity of the reward function).

Organization and main contributions. In Section 2, we formally introduce the model and the problem faced by an oracle who knows the distribution of task durations as well as the reward function (quite importantly, we emphasize again that the oracle policy must be independent of the realized rewards). Using continuous-time dynamic programming principles, we construct a quasi-optimal policy in terms of accepted and rejected tasks: this translates into a single optimal threshold for the ratio of the reward to the duration, called the profitability function. Any task with a profitability above this threshold is accepted, and the other ones are declined. As a benchmark, we first assume in Section 3 that the agent knows the reward function $r(\cdot)$, but ignores the distribution of task durations. The introduced techniques can be generalized, in the following sections, to further incorporate estimations of $r(\cdot)$. In that case, our base algorithm has a regret scaling as $\mathcal{O}(\sqrt{T})$; obviously, this cannot be improved without additional assumptions, ensuring minimax optimality. In Section 4, the reward function is not known to the agent anymore and the reward realizations are assumed to be noisy. To get non-trivial estimation rates, regularity – i.e., (L, β)-Hölder – of the reward function is assumed. Modifying the basic algorithm to incorporate non-parametric estimation of r yields a regret scaling as $\mathcal{O}(T^{1-\eta}\sqrt{\ln T})$ where $\eta = \beta/(2\beta + 1)$. As this is the standard error rate in classification (Tsybakov, 2006), minimax optimality (up to $\log$-term) is achieved again. Finally, our different algorithms are empirically evaluated on simple toy examples in Section 5. Due to space constraints, all the proofs are deferred to the Appendix.

Related work. As data are gathered sequentially, our problem bears similarities with online learning and multi-armed bandit (Bubeck and Cesa-Bianchi, 2012). The main difference with multi-armed bandit or resource allocation problems (Fontaine et al., 2020) is that the agent's only resource is her time, which has to be spent wisely and, most importantly, saving it actually has some unknown value. The problem of time allocation actually goes way back. In economic theory, Becker's seminal paper (Becker, 1965) evaluates the full costs of non-working activities as the sum of their market prices and the forgone value of the time used to enjoy the activities. Various empirical works have followed (see, e.g., Juster and Stafford, 1991; Chabris et al., 2009).

The problem of online scheduling has been studied in several possible variations (see, e.g., Pruhs et al., 2004, for a survey). Various subsequent papers consider scheduling models that combines online and stochastic aspects (see, e.g., Megow et al., 2006; Chou et al., 2006; Vredeveld, 2012; Marbán et al., 2012; Skutella et al., 2016). Yet these models are different as they aim at minimizing the makespan of all received jobs, while actions here consist in accepting/declining tasks. In online admission control, a controller receives requests and decides on the fly to accept/reject them. However, the problem of admission control is more intricate since a request is served through some path, chosen by the controller, in a graph. Even solving it offline thus remains a great challenge (Wu and Bertsekas, 2001; Leguay et al., 2016).

An online model of reward maximization over time budget has been proposed by Cayci et al. (2019, 2020), where the agent does not observe the duration of a task before accepting it. As a consequence, the optimal policy always chooses the same action, which is maximizing the profitability. Since the duration of a task is observed before taking the decision in our problem, the decision of the optimal policy depends on this duration, used as a covariate. Cayci et al. (2019) thus aim at determining the arm with the largest profitability, while our main goal is here to estimate the profitability threshold from which we start accepting the tasks. The considered problems and proposed solutions thus largely differ in these two settings.

Our problem is strongly related to contextual multi-armed bandits, where each arm produces a noisy reward that depends on an observable context. Indeed, the agent faces a one arm contextual bandit problem (Sarkar, 1991), with the crucial difference that the normalized reward associated to a context depends on the whole distribution of contexts (and not just the current context). The literature on contextual bandits, also known as bandits with covariates, actually goes back to Woodroofe (1979) and has seen extensive contributions in different settings (see, e.g., Yang and Zhu, 2002; Wang et al., 2005; Rigollet and Zeevi, 2010; Goldenshluger and Zeevi, 2009; Perchet and Rigollet, 2013).

This problem is also related to bandits with knapsacks (see Slivkins, 2019, Chapter 10) introduced by Badanidiyuru et al. (2013) and even more specifically to contextual bandits with knapsacks (Badanidiyuru et al., 2014; Agrawal et al., 2016), where pulling an arm consumes a limited resource. Time is the resource of the agent here, while both time and resource are well separated quantities in bandits with knapsacks. Especially, bandits with knapsacks assume the existence of a null arm, which does not consume any resource. This ensures the feasibility of the linear program giving the optimal fixed strategy. Here, the null arm (declining the task) still consumes the waiting time before receiving the next task proposal. Bandits with knapsacks strategies are thus not adapted to our problem, which remains solvable thanks to a particular problem structure. The problem of online knapsacks is also worth mentioning (Noga and Sarbua, 2005; Chakrabarty et al., 2008; Han and Makino, 2009; Böckenhauer et al., 2014), where the resources consumed by each action are observed beforehand. This is less related to our work, as online knapsacks consider a competitive analysis, whereas we here aim at minimizing the regret with stochastic contexts/rewards.

2 Model and benchmark

2.1 The problem: allocation of time

All the different notations used in the paper are summarized at the beginning of the Appendix to help the reader. We consider an agent who sequentially receives task proposals and decides whether to accept or decline them on the fly. The durations of the proposed tasks are assumed to be i.i.d. with an unknown law, and X_i denotes the duration of the i-th task. If this task is accepted, the agent earns some reward Y_i with expectation $r(X_i)$. The profitability is then defined as the function $x \mapsto r(x)/x$. We emphasize here that Y_i is not observed before accepting (or actually completing) the i-th task and that the expected reward function $r(\cdot)$ is unknown to the agent at first. If task i is accepted, the agent cannot accept any new proposals for the whole duration X_i. We assume in the following that the function r is bounded in $[E, D]$ with $E \leq 0 \leq D$, and that the durations X_i are upper bounded by C.

After completing the i-th task, or after declining it, the agent is on hold, waiting for a new proposal. We assume that idling times – denoted by S_i – are also i.i.d., following some exponential law of parameter λ. This parameter is supposed to be known beforehand as it has a small impact on the learning cost (it can be quickly estimated, independently of the agent's policy). An equivalent formulation of the task arrival process is that proposals follow a Poisson process (with intensity λ) and the agent does not observe task proposals while occupied. This is a mere consequence of the memoryless property of Poisson processes.

The agent's objective is to maximize the expected sum of rewards obtained by choosing an appropriate acceptance policy. Given the decisions $(a_i)_{i\geq 1}$ in $\{0, 1\}$ (decline/accept), the total reward accumulated by the agent after the first n proposals is equal to $\sum_{i=1}^{n} Y_i a_i$ and the required amount of time for this is $\mathcal{T}_n \coloneqq \sum_{i=1}^{n} S_i + X_i a_i$. This amount of time is random and strongly depends on the policy. As a consequence, we consider that the agent optimizes the cumulative reward up to time T, so that the number of received tasks proposals is random and equal to $\theta \coloneqq \min\{n \in \mathbb{N} \mid \mathcal{T}_n > T\}$. Mathematically, a policy of the agent is a function π – taking as input a proposed task X, the

time t and the history of observations $\mathcal{H}_t \coloneqq (X_i, Y_i a_i)_{i:\mathcal{T}_i<t}$ – returning the decision $a \in \{0,1\}$ (decline/accept). In the following, the optimal policy refers to the strategy π maximizing the expected reward U_π at time T, given by

$$U_\pi(T) = \mathbb{E}\left[\sum_{n=1}^{\theta} r(X_n)\pi(X_n, t_n, \mathcal{H}_{t_n})\right], \tag{2.1}$$

where $t_n \coloneqq S_n + \sum_{i=1}^{n-1} S_i + X_i a_i$ is the time at which the n-th task is proposed.

We allow a slight boundary effect, i.e., the ending time can slightly exceed T, because the last task may not end exactly at time T. A first alternative would be to compute the cumulative revenue over completed tasks only, and another alternative would be to attribute the relative amount of time before T spent on the last task. Either way, the boundary effect is of the order of a constant and has a negligible impact as T increases.

2.2 The benchmark: description of the optimal policy

The benchmark to which the real agent is compared is an oracle who knows beforehand both the distribution of tasks and the reward function. It is easier to describe the reward of the optimal policy than the policy itself. Indeed, at the very last time T, the agent cannot accumulate any more reward, hence the remaining value is 0. We then compute the reward of the policy backward, using continuous dynamic programming principles. Formally, let $v(t)$ denote the "value" function, i.e., the expected reward that the optimal policy will accumulate on the remaining time interval $[t, T]$ if the agent is on hold at this very specific time $t \in [0, T]$. As mentioned before, we assume the boundary condition $v(T) = 0$. The next proposition gives the dynamic programming equation satisfied by $v(t)$ (with the notation $(z)_+ = \max\{z, 0\}$ for any $z \in \mathbb{R}$), as well as a simple function approximating this value.

Proposition 2.1. *The value function v satisfies the dynamic programming equation*

$$\begin{cases} v'(t) = -\lambda\,\mathbb{E}\left[(r(X) + v(t+X) - v(t))_+\right] & \text{for all } t < T, \\ v(t) = 0 & \text{for all } t \geq T. \end{cases} \tag{2.2}$$

If $X \leq C$, then its solution is bounded by the affine function $w : t \mapsto c^*(T-t)$ for any $t \in [0, T]$ as follows

$$w(t - C) \geq v(t) \geq w(t), \tag{2.3}$$

where c^* is the unique root of the function

$$\Phi : \begin{array}{l} \mathbb{R}_+ \to \mathbb{R} \\ c \mapsto \lambda\,\mathbb{E}\left[(r(X) - cX)_+\right] - c. \end{array} \tag{2.4}$$

The constant c^* represents the optimal reward per time unit, hereafter referred to as the *optimal profitability threshold*. The function w is the value function of the optimal policy when neglecting boundary effects. The proof of Proposition 2.1 largely relies on the memoryless property of the idling times and determining a benchmark policy without this memoryless assumption becomes much more intricate. Based on this, it is now possible to approach the optimal policy by only accepting task proposals with profitability at least c^*. This results in a stationary policy, i.e., its decisions depend neither on the time nor on past observations but only on the duration of the received task.

Theorem 2.2. *The policy π^* which accepts a task with duration x if and only if $r(x) \geq c^* x$ is c^*C-suboptimal, i.e., $U_{\pi^*} \geq \max_\pi U_\pi - c^* C$.*

In the considered model, it is thus possible to compute a quasi-optimal online policy.

2.3 Online learning policies and regret

Below we investigate several contexts where the agent lacks information about the environment, such as the task durations distribution and/or the (expected) reward function. As mentioned before, the objective of the agent is to maximize the cumulative reward gathered until T (up to the same boundary effect as the optimal policy) or equivalently to minimize the policy regret, defined for the policy π as

$$R(T) = c^* T - U_\pi(T). \tag{2.5}$$

Note that $R(T)$ is the difference between the expected rewards of strategies π^* and π, up to some constant term due to the boundary effect.

3 Warm up: known reward, unknown distribution

First, we assume that the reward function r is known to the agent, or equivalently, is observed along with incoming task proposals. However, the agent does not know the distribution F of task durations and we emphasize here that the agent does not observe incoming task proposals while occupied with a task (though the converse assumption should only affect the results by a factor C^{-1}, the inverse of the maximal length of a task). We now define

$$\Phi_n : c \mapsto \lambda \frac{1}{n} \sum_{i=1}^{n} (r(X_i) - cX_i)_+ - c, \tag{3.1}$$

which is the empirical counterpart of Φ. Moreover, let c_n be the unique root of Φ_n. One has $\mathbb{E}[\Phi_n(c)] = \Phi(c)$ for all $c \geq 0$ and all $n \geq 1$.

Proposition 3.1. *For all $\delta \in (0, 1]$ and $n \geq 1$, $\mathbb{P}\left(c_n - c^* > \lambda(D - E)\sqrt{\frac{\ln(1/\delta)}{2n}}\right) \leq \delta$.*

Remark 3.2. Proposition 3.1 states that the error in the estimation of c^* scales with $\lambda(D-E)$. Notice that if the reward function is multiplied by some factor (and λ is fixed), then c^* is multiplied by the same factor; as a consequence, a linear dependency in $D - E$ is expected.
Similarly, since $c^* = \lambda \mathbb{E}\left[(r(X) - c^*X)_+\right]$, a small variation of λ induces a variation of c^* of the same order. As a consequence, a linear dependency in λ (for small values of λ) is expected. And as both effects are multiplied, the dependency in $\lambda(D - E)$ is correct.
On the other hand, as λ goes to infinity, c^* converges to $\max_x r(x)/x$, so our deviation result seems irrelevant (for a small number of samples). This is due to the fact that for large λ, c^* is not really the expectation of a random variable, but its essential supremum. In the proof, this appears when $\Phi(c^* + \varepsilon)$ is bounded by $-\varepsilon$. It is not difficult to see that a tighter upper-bound is

$$-\varepsilon(1 + \lambda q_\varepsilon), \quad \text{with} \quad q_\varepsilon \coloneqq \mathbb{E}[X \mathbb{1}(r(X) \geq (c^* + \varepsilon)X)]. \tag{3.2}$$

Hence the approximation error scales with $\lambda(D - E)/(1 + \lambda q_\varepsilon)$ as λ goes to infinity. However, this does not give explicit confidence intervals and the regime $\lambda \to \infty$ is not really interesting. As a consequence, the formulation of Proposition 3.1 is sufficient for our purposes.

The used policy is straightforward: upon receiving the n-th proposal, the agent computes c_n and accepts the task X_n if and only if $r(X_n) \geq c_n X_n$. The pseudo-code of this policy is given in Algorithm 1 below. The regret of this policy can be rewritten as

$$R(T) = c^*T - \mathbb{E}\left[\sum_{n=1}^{\theta} \gamma_n \mathbb{E}\left[X_n \mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right]\right], \tag{3.3}$$

$$\text{where} \quad \gamma_n \coloneqq \frac{\lambda \mathbb{E}[r(X_n) \mathbb{1}(r(X_n) \geq c_n X_n)]}{1 + \lambda \mathbb{E}[X_n \mathbb{1}(r(X_n) \geq c_n X_n)]} = \frac{\mathbb{E}[r(X_n) \mathbb{1}(r(X_n) \geq c_n X_n)]}{\mathbb{E}[S_n + X_n \mathbb{1}(r(X_n) \geq c_n X_n)]} \tag{3.4}$$

corresponds to the expected per-time reward obtained for the n-th task under the policy induced by $(c_n)_{n\geq 1}$. In the last expression, the numerator is indeed the expected reward per task while the denominator is the expected time spent per task (including the waiting time before receiving the proposal). Proposition D.1, postponed to the Appendix, gives a control of γ_n entailing the following.

Theorem 3.3. *In the known reward, unknown distribution setting, the regret of Algorithm 1 satisfies*

$$R(T) \leq \lambda(D - E)C\sqrt{\frac{\pi}{2}}\sqrt{\lambda T + 1}. \tag{3.5}$$

Remark 3.4. Once again, one might question the dependency on the different parameters. As mentioned before, $\lambda(D - E)$ represents the scale at which c^* grows and C is the scale of X, so that $\lambda(D - E)C$ is the total global scaling of rewards. On the other hand, the parameter λ also appears in the square-root term. This is because λT is approximately the order of magnitudes of the observed tasks.

Remark 3.5. The estimated profitability threshold c_n can be efficiently approximated using binary search, since the function Φ_n is decreasing. The approximation error is ignored in the regret, as approximating c_n up to, e.g., n^{-2} only leads to an additional constant term in the regret.
The computation and memory complexities of Algorithm 1 are then respectively of order $n \log(n)$

ALGORITHM 1: Known reward algorithm

input : r, λ
$n = 0$
while $\sum_{i=1}^{n} S_i + X_i \mathbb{1}(r(X_i) \geq c_i X_i) < T$ **do**
  $n = n + 1$
  Wait S_n and observe X_n
  Compute c_n as the unique root of Φ_n defined in Equation (3.1)
  if $r(X_n) \geq c_n X_n$ **then** accept task and receive reward $r(X_n)$
  else reject task
end

and n when receiving the n-th proposal, because the complete history of tasks is used to compute c_n, the root of Φ_n. This can be improved by noting that only the tasks with profitability in

$$\left[c_n - \lambda(D-E)\sqrt{\frac{\ln(1/\delta)}{2n}},\ c_n + \lambda(D-E)\sqrt{\frac{\ln(1/\delta)}{2n}} \right]$$

need to be exactly stored, thanks to Proposition 3.1. Indeed, with high probability, tasks with smaller profitability do not contribute to $\Phi(c^*)$, while tasks with larger profitability fully contribute to $\Phi(c^*)$, meaning that only their sum has to be stored. As this interval is of length $1/\sqrt{n}$, the complexities of the algorithm become sublinear in n under regularity assumptions on X and the profitability function. The computational complexity can even be further improved to $\log(n)$ using binary search trees, as explained in Appendix B.

4 Bandit Feedback

We now assume that rewards are noisy and not known upfront. When the agent accepts a task X_i, the reward earned and observed over the period of time X_i is $Y_i = r(X_i) + \varepsilon_i$, where $(\varepsilon_i)_{i \geq 1}$ is a sequence of σ^2-subgaussian independent random variables with $\mathbb{E}[\varepsilon_i] = 0$. If the task X_i is rejected, then the associated reward is not observed; on the other hand, the agent is not occupied and thus may be proposed another task, possibly within the time window of length X_i.

To get non-trivial estimation rates, in the whole section we assume the reward function r to be (β, L)-Hölder, whose definition is recalled below.

Definition 4.1. Let $\beta \in (0, 1]$ and $L > 0$. The function r is (β, L)-Hölder if

$$|r(x) - r(x')| \leq L|x - x'|^{\beta}, \qquad \forall x, x' \in [0, C]\ .$$

The reward function $r(\cdot)$ is estimated by $\hat{r}_n(\cdot)$, constructed as a variant of a regressogram. The state space $[0, C]$ is partitioned regularly into M bins $B_1, \ldots, B_M$ of size $h = C/M$. For every bin B, let $x^B := \min\{x \in B\}$. Similarly to Algorithm 1, c^* is estimated by $\hat{c}_n$. To define the learning algorithm, we also construct an upper-estimate of $r(\cdot)$, denoted by $\hat{r}_n^+(\cdot)$, and a lower-estimate of c^*, denoted by $\hat{c}_n^-$.

The learning algorithm is actually quite simple. A task X_n is accepted if and only if its best-case reward $\hat{r}_{n-1}^+(X_n)$ is bigger than the worst-case per-time value of its bin $\hat{c}_{n-1}^- x^{B(X_n)}$, where X_n is in the bin $B(X_n)$. Notice that, if $\hat{r}_n^+(\cdot)$ is bigger than $r(\cdot)$ and $\hat{c}_n^-$ is always smaller than c^*, then any task accepted by the optimal algorithm (i.e., such that $r(X) \geq c^* X$) is accepted by the learning algorithm. Hence regret is incurred solely by accepting sub-optimal tasks. The question is therefore how fast tasks can be detected as sub-optimal, and whether declining them affects, or not, the estimation of c^*. The pseudo-code is given in Algorithm 2 at the end of this section.

We now describe the construction of $\hat{r}_n$ and $\hat{c}_n$ (as well as their upper and lower-estimates). Let $(X_1, Y_1, \ldots, X_n, Y_n)$ be the vector of observations of pairs (task, reward). We define for all bins B,

$$\begin{gathered} \hat{r}_n(x) = \hat{r}_n^B = \frac{1}{N_B} \sum_{i=1}^{n} Y_i \mathbb{1}(X_i \in B \text{ and } a_i = 1), \quad \forall x \in B \\ \text{with } N_B = \textstyle\sum_{i=1}^{n} \mathbb{1}(X_i \in B \text{ and } a_i = 1). \end{gathered} \tag{4.1}$$

We emphasize here that declined tasks are not used in the estimation of $r(\cdot)$ (as we shall see later, they are not used in the estimation of c^*, either). The upper estimate of $r(\cdot)$ is now defined as

$$\hat{r}_n^+(x) = \hat{r}_n(x) + \underbrace{\sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}}\sqrt{\frac{\ln(M/\delta)}{2N_B}} + L\left(\frac{C}{M}\right)^{\beta}}_{:=\eta_{n-1}(x)}, \quad \forall x \in B. \tag{4.2}$$

The following Lemma 4.2 states that $\hat{r}_n^+(\cdot)$ is indeed an upper-estimate of $r(\cdot)$.

Lemma 4.2. *For every $n \in \mathbb{N}$, we have $\mathbb{P}\left(\forall x \in [0, C],\ \hat{r}_n^+(x) \geq r(x)\right) \geq 1 - \delta$.*

It remains to construct $\hat{c}_n$. We define iteratively for every bin B

$$\tilde{r}_n^B = \begin{cases} 0 & \text{if a task in } B \text{ has ever been rejected,} \\ \hat{r}_n^B & \text{otherwise.} \end{cases} \tag{4.3}$$

So $\tilde{r}_n$ is equal to the reward estimate $\hat{r}_n$, except on the eliminated bins that are known to be suboptimal, for which it is instead 0. We then introduce the empirical counterpart of Φ,

$$\hat{\Phi}_n : c \mapsto \lambda \sum_{j=1}^{M} \frac{N_{B_j}}{n}\left(\tilde{r}_n^{B_j} - c x^{B_j}\right)_+ - c. \tag{4.4}$$

Let $\hat{c}_n$ be the unique root of $\hat{\Phi}_n$; the lower estimate of c^* is then

$$\hat{c}_n^- = \hat{c}_n - \underbrace{\left(2\lambda\sqrt{\sigma^2 + \frac{(D-E)^2}{4}}\sqrt{\frac{\ln(1/\delta)}{n}} + \kappa\lambda \max\left(\sigma, \frac{D-E}{2}\right)\sqrt{\frac{\log(n)+1}{hn}} + \sqrt{8}\lambda\frac{Lh^\beta}{2^\beta} + \lambda^2 Dh\right)}_{:=\xi_{n-1}}, \tag{4.5}$$

where $\kappa \leq 150$, is the square root of the universal constant introduced in (Györfi et al., 2002, Theorem 11.3). Although the following lemma looks trivial, it is actually a keystone of our main result. It states that the optimal profitability threshold may be computed solely based on the accepted tasks – no matter what the rewards associated to declined tasks are. This is of crucial relevance, since it implies that a learning algorithm may actually decline tasks that are clearly non-profitable, without degrading the quality and the speed of profitability threshold estimation.

Lemma 4.3. *Let c^* be the optimal profitability threshold associated to r, i.e., c^* is the unique root of $c \mapsto \Phi(c) = \lambda\,\mathbb{E}\left[(r(X) - cX)_+\right] - c$ and let $\mathcal{E} \subset \{x \in [0, C] \,|\, r(x) \leq c^* x\}$ be any subset of sub-profitable tasks.*
Then the profitability threshold $\tilde{c}$ associated to the modified reward function $r_{\mathcal{E}}(x) \coloneqq r(x)\,\mathbb{1}(x \notin \mathcal{E})$ is equal to c^*, or, stated otherwise,

$$c^* = \lambda\,\mathbb{E}\left[(r_{\mathcal{E}}(X) - c^* X)_+\right], \quad \forall \mathcal{E} \subset \{x \in [0, C] \,|\, r(x) \leq c^* x\}.$$

This implies that $\hat{c}_n^-$ is indeed an under-estimate.

Proposition 4.4. *For all $N \in \mathbb{N}$, we have $\mathbb{P}(\forall n \leq N, \hat{c}_n^- \leq c^*) \geq 1 - 2N\delta$.*

We can now state our main result.

Theorem 4.5. *If r is (L, β)-Hölder, then the regret of Algorithm 2, taking $\delta = 1/T^2$ and $M = \left\lceil CL^{\frac{2}{2\beta+1}}(\lambda T + 1)^{\frac{1}{2\beta+1}}\right\rceil$, satisfies*

$$R(T) \leq \kappa_1 \lambda C \max(\sigma, D - E) L^{\frac{1}{2\beta+1}}(\lambda T + 1)^{\frac{\beta+1}{2\beta+1}}\sqrt{\ln(\lambda T + 1)},$$

where κ_1 is some universal constant independent of any problem parameter.

A similar lower bound is proved in Appendix A.4, which implies that Algorithm 2 optimally scales with T, up to $\log$ terms. Even faster rates of convergence hold under additional assumptions on the task distribution or the profitability function and are provided in Appendix A for the cases of finite support of X, margin condition or monotone profitability function.

Remark 4.6. The complexity of Algorithm 2 when receiving a single task scales with the number of bins M, because the function $\hat{\Phi}_n$ uses the discretization over the bins. In particular, it uses the representatives x^B instead of the exact observations X_i in $\hat{\Phi}_n$, which is the reason for the additional $\lambda^2 Dh$ term in ξ_n. Similarly to Algorithm 1, it can be improved to a $\log(M)$ computational complexity with binary search trees as detailed in Appendix B.

ALGORITHM 2: Bandit algorithm

input : δ, λ, C, D, E, T, σ, κ
$\hat{c}_0^- = 0$; $\hat{r}_0^+ = \infty$; $n = 0$
while $\sum_{i=1}^{n} S_i + X_i \mathbb{1}(\hat{r}_i^+(X_i) \geq \hat{c}_i^- X_i) < T$ **do**
  $n = n + 1$
  Wait S_n and observe X_n
  if $\hat{r}_{n-1}^+(X_n) \geq \hat{c}_{n-1}^- x^{B(X_n)}$ **then** accept task and receive reward Y_n
  else reject task
  For $B = B(X_n)$ compute $\hat{r}_n^B, \hat{r}_n^{B+}, \tilde{r}_n^B$ as described by Equations (4.1) to (4.3)
  Compute $\hat{c}_n$ as the unique root of $\hat{\Phi}_n$ defined in Equation (4.4)
  $\hat{c}_n^- = \hat{c}_n - \xi_{n-1}$ as described by Equation (4.5)
end

5 Simulations

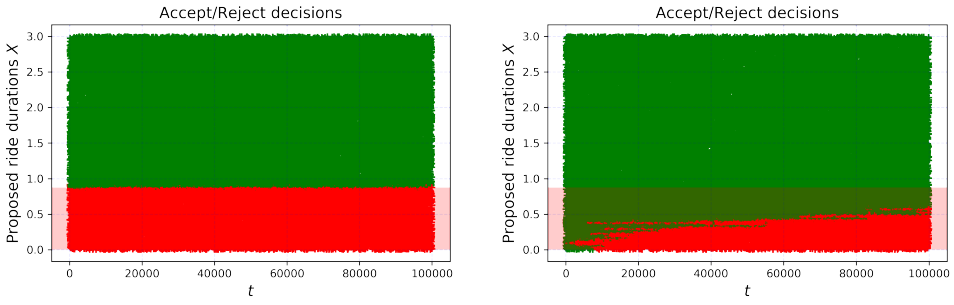
Algorithms 1 and 2 are computationally efficient when using binary search trees as described in Appendix B. This section compares empirically these different algorithms on toy examples, considering affine and concave reward functions. The code used in this section is available at `github.com/eboursier/making_most_of_your_time`.

5.1 Affine reward function

We first study the simple case of affine reward function $r(x) = x - 0.5$. The profitability function is increasing and Algorithm 3, given in Appendix A.3, can be used in this specific case. We consider a uniform distribution on $[0, 3]$ for task distribution, a Poisson rate $\lambda = 1$ and a uniform distribution on $[-1, 1]$ for the noise ε. As often in bandit algorithms, we tune the constants scaling the uncertainty terms η and ξ for better empirical results.

Figure 1 shows the typical decisions of Algorithms 1 and 2 on a single run of this example. The former especially seems to take almost no suboptimal decision. This is because in simpler settings with good margin conditions, the regret of Algorithm 1 can be shown to be only of order $\log(T)$. On the other hand, it can be seen how Algorithm 2 progressively eliminates bins, represented by red traces.

Figure 2 gives the upper estimate of the reward function by representing $\hat{r}_n^{B+}/x^B$ for each bin and the lower estimate of $\hat{c}_n^-$ after a time $t = 10^5$. It illustrates how the algorithm takes decisions at this stage. Especially, the bins with estimated profitability above $\hat{c}_n^-$, while their actual profitability is below c^*, still have to be eliminated by the algorithm. Some bins are badly estimated, since the algorithm stops accepting and estimating them as soon as they are detected suboptimal. The algorithm thus adapts nicely to the problem difficulty, by eliminating very bad bins way faster than slightly suboptimal ones.



(a) Algorithm 1: known reward.

(b) Algorithm 2: bandit.

Figure 1: Accept/reject decisions on a single run. A single point corresponds to a proposed task, with the time of appearance on x axis and task durations on y axis. It is marked in green (resp. red) if it is accepted (resp. rejected) by the algorithm, while the light red area highlights the suboptimal region. Each task in this colored region thus has a profitability below c^*.

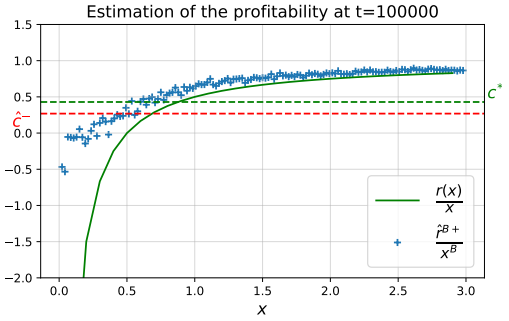


Figure 2: Estimations of $r(\cdot)$ and c^* by Algorithm 2 at $t = 10^5$.

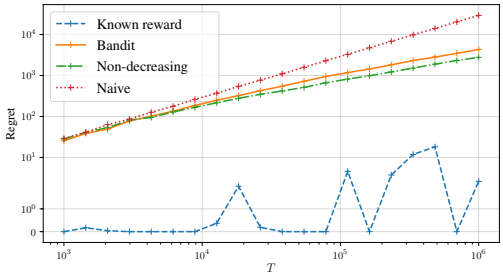


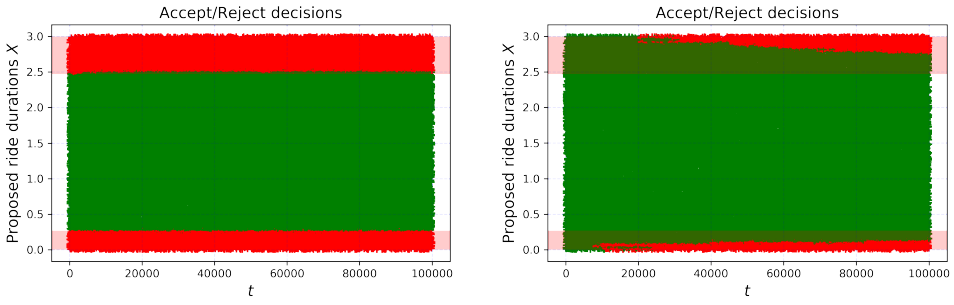
Figure 3: Evolution of regret with T for different algorithms.

Figure 3 shows in logarithmic scale the evolution of regret with the horizon T for the different algorithms. The regret is averaged over 50 runs (and 500 runs for Known reward). Algorithm 1 performs way better than its $\sqrt{T}$ bound, due to a better performance with margin conditions as mentioned above. Its regret is non-monotonic in T here and is sometimes negative, in which case we round it to 0 for clarity. This is only due to the variance in the Monte-Carlo estimates of the expected regret. For such a small regret, much more than 500 runs are required to observe accurately the expected value. We unfortunately could not afford to simulate more runs for computational reasons.

Non-decreasing corresponds to Algorithm 3 and performs slightly better than Algorithm 2 here. This remains a small improvement since Algorithm 2 benefits from margin conditions as shown in Appendix A.2. Moreover, this improvement comes with additional computations. Clearly, all these algorithms perform way better than the naive algorithm, which simply accepts every task. We indeed observe that learning occurs and the regret scales sublinearly with T for all non-naive algorithms.

5.2 Concave reward function

This section now considers the case of the concave reward function $r(x) = -0.3x^2 + x - 0.2$. The profitability is not monotone here, making Algorithm 3 unadapted. We consider the same task distribution as in Section 5.1 and a Gaussian noise of variance 0.1 for ε. Similarly to the affine case, Figures 4 to 6 show the behaviors of the different algorithms, as well as the regret evolution. Here as well, the regret is averaged over 50 runs (and 500 runs for Known reward).



(a) Algorithm 1: known reward.

(b) Algorithm 2: bandit setting.

Figure 4: Accept/Reject decisions on a single run, illustrated as in Figure 1.

The problem is now harder than in the previous affine case, since there are multiple suboptimal regions. Algorithm 1 still performs very well for the same reasons, while Algorithm 2 requires more time to detect suboptimal tasks, but still performs well in the end. Similarly to Figure 3, the regret of Known reward is non-monotonic at first here. However it is increasing for larger values of T, as the variance becomes negligible with respect to the regret in this setting.

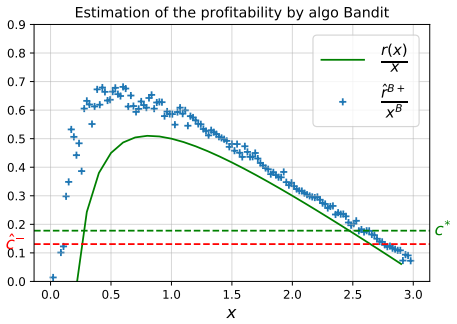


Figure 5: Estimation of profitability function by Algorithm 2 after a time $t = 10^5$.

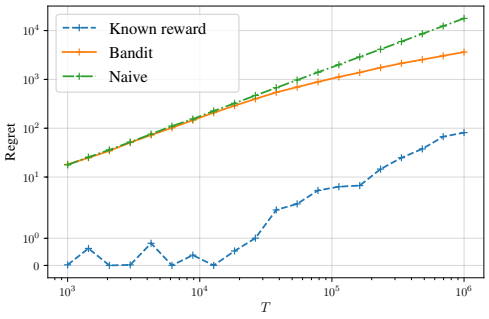


Figure 6: Evolution of regret with time T.

6 Conclusion

We introduced the problem of online allocation of time where an agent observes the duration of a task before accepting/declining it. Its main novelty and difficulty resides in the estimation and computation of a threshold, which depends both on the whole duration distribution and the reward function. After characterizing a baseline policy, we proposed computationally efficient learning algorithms with optimal regret guarantees, when the reward function is either known or unknown. Even faster learning rates are shown in Appendix A under stronger assumptions on the duration distribution or the reward function.

The model proposed in this work is rather simple and aims at laying the foundations for possible extensions, that might be better adapted to specific applications. We believe that Algorithm 2 can be easily adapted to many natural extensions. For instance, we here assume that the reward function only depends on the duration of the task, while it could also depend on additional covariates; using contextual bandits techniques with our algorithm should directly lead to an optimal solution in this case. Similarly, time is the only limited resource here, while extending this work to multiple limited resources seems possible using knapsack bandits techniques. It is also possible to extend it to several actions (instead of a single accept/decline action) by simultaneously estimating multiple rewards functions.

On the other hand, some extensions deserve careful considerations, as they lead to complex problems. In call centers for example, the agent might be able to complete several tasks simultaneously instead of a single one and characterizing the benchmark becomes much more intricate in this setting. Also, we consider stochastic durations/rewards here. Extending this work to the adversarial case is far from obvious. We yet believe this might be related to online knapsack problems and the use of a competitive analysis might be necessary in this case. For ride-hailing applications, the spatial aspect is also crucial as a driver moves from a starting point to his destination during each ride. Adding a state variable to the model might thus be interesting here.

Finally, we believe this work only aims at proposing a first, general and simple model for online learning problems of time allocation, which can be extended to many settings, depending on the considered application.

Acknowledgements

E. Boursier was supported by an AMX scholarship. V. Perchet acknowledges support from the French National Research Agency (ANR) under grant number #ANR-19-CE23-0026 as well as the support grant, as well as from the grant "Investissements d'Avenir" (LabEx Ecodec/ANR-11-LABX-0047). This research project received partial support from the COST action GAMENET. M. Scarsini is a member of INdAM-GNAMPA. His work was partially supported by the INdAM-GNAMPA 2020 Project "Random walks on random games" and the Italian MIUR PRIN 2017 Project ALGADIMAR "Algorithms, Games, and Digital Markets." He gratefully acknowledges the kind hospitality of Ecole Normale Supérieure Paris-Saclay, where this project started.

References

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pages 4–18, 2016.

Jean-Yves Audibert and Alexandre B. Tsybakov. Fast learning rates for plug-in classifiers. *Ann. Statist.*, 35(2):608–633, 2007.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134, 2014.

Gary S. Becker. A theory of the allocation of time. *The Economic Journal*, 75(299):493–517, 1965.

Hans-Joachim Böckenhauer, Dennis Komm, Richard Královič, and Peter Rossmanith. The online knapsack problem: Advice and randomization. *Theoretical Computer Science*, 527:61–72, 2014.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Clément L Canonne. A short note on Poisson tail bounds, 2017. URL `http://www.cs.columbia.edu/~ccanonne`.

Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. Learning to control renewal processes with bandit feedback. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 3 (2):1–32, 2019.

Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. Budget-constrained bandits over general cost and reward distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4388–4398. PMLR, 2020.

Christopher F. Chabris, Carrie L. Morris, Dmitry Taubinsky, David Laibson, and Jonathon P. Schuldt. The allocation of time in decision-making. *Journal of the European Economic Association*, 7(2-3): 628–637, 05 2009.

Deeparnab Chakrabarty, Yunhong Zhou, and Rajan Lukose. Online knapsack problems. In *Workshop on internet and network economics (WINE)*, 2008.

Mabel C. Chou, Hui Liu, Maurice Queyranne, and David Simchi-Levi. On the asymptotic optimality of a simple on-line algorithm for the stochastic single-machine weighted completion time problem and its extensions. *Oper. Res.*, 54(3):464–474, 2006. ISSN 0030-364X. doi: 10.1287/opre.1060.0270. URL `https://doi.org/10.1287/opre.1060.0270`.

Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, third edition, 2009.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1587–1595, New York, New York, USA, 20–22 Jun 2016. PMLR.

Xavier Fontaine, Shie Mannor, and Vianney Perchet. An adaptive stochastic optimization algorithm for resource allocation. In Aryeh Kontorovich and Gergely Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 319–363, San Diego, California, USA, 08 Feb–11 Feb 2020. PMLR.

Alexander Goldenshluger and Assaf Zeevi. Woodroofe's one-armed bandit problem revisited. *Ann. Appl. Probab.*, 19(4):1603–1633, 2009.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag, New York, 2002.

Xin Han and Kazuhisa Makino. Online minimization knapsack problem. In *International Workshop on Approximation and Online Algorithms*, pages 182–193. Springer, 2009.

F. Thomas Juster and Frank P. Stafford. The allocation of time: Empirical findings, behavioral models, and problems of measurement. *J. Econ. Literature*, 29(2):471–522, 1991.

Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36, 2014.

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

Jérémie Leguay, Lorenzo Maggi, Moez Draief, Stefano Paris, and Symeon Chouvardas. Admission control with online algorithms in sdn. In *NOMS 2016-2016 IEEE/IFIP Network Operations and Management Symposium*, pages 718–721. IEEE, 2016.

Sebastián Marbán, Cyriel Rutten, and Tjark Vredeveld. Learning in stochastic machine scheduling. In *Approximation and online algorithms*, volume 7164 of *Lecture Notes in Comput. Sci.*, pages 21–34. Springer, Heidelberg, 2012. doi: 10.1007/978-3-642-29116-6_3. URL `https://doi.org/10.1007/978-3-642-29116-6_3`.

Nicole Megow, Marc Uetz, and Tjark Vredeveld. Models and algorithms for stochastic online scheduling. *Math. Oper. Res.*, 31(3):513–525, 2006. ISSN 0364-765X. doi: 10.1287/moor.1060.0201. URL `https://doi.org/10.1287/moor.1060.0201`.

John Noga and Veerawan Sarbua. An online partially fractional knapsack problem. In *8th International Symposium on Parallel Architectures, Algorithms and Networks (ISPAN'05)*, pages 5–pp. IEEE, 2005.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *Ann. Statist.*, 41(2):693–721, 2013.

Kirk Pruhs, Jiri Sgall, and Eric Torng. Online scheduling. In Joseph Y-T. Leung, editor, *Handbook of Scheduling: Algorithms, Models, and Performance Analysis*. Chapman and Hall/CRC, 2004.

Philippe Rigollet and Assaf Zeevi. Nonparametric bandits with covariates. *COLT 2010*, page 54, 2010.

Jyotirmoy Sarkar. One-armed bandit problems with covariates. *Ann. Statist.*, 19(4):1978–2002, 1991.

Martin Skutella, Maxim Sviridenko, and Marc Uetz. Unrelated machine scheduling with stochastic processing times. *Math. Oper. Res.*, 41(3):851–864, 2016. ISSN 0364-765X. doi: 10.1287/moor.2015.0757. URL `https://doi.org/10.1287/moor.2015.0757`.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*, 2019.

Alexandre Tsybakov. Statistique appliquée, 2006. Lecture Notes.

Tjark Vredeveld. Stochastic online scheduling. *Computer Science - Research and Development*, 27(3):181–187, 2012. doi: 10.1007/s00450-011-0153-5. URL `https://doi.org/10.1007/s00450-011-0153-5`.

Chih-Chun Wang, Sanjeev R. Kulkarni, and H. Vincent Poor. Bandit problems with side observations. *IEEE Trans. Automat. Control*, 50(3):338–355, 2005.

Michael Woodroofe. A one-armed bandit problem with a concomitant variable. *J. Amer. Statist. Assoc.*, 74(368):799–806, 1979.

Cynara C Wu and Dimitri P Bertsekas. Admission control for wireless networks. *IEEE Trans. Veh. Technol*, 50:504–514, 2001.

Yuhong Yang and Dan Zhu. Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates. *Ann. Statist.*, 30(1):100–121, 2002.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No] The experiments can be run on any personal laptop.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

List of symbols

a_i	decision concerning i-th task
B_j	j-th bin
c^*	unique root of the function Φ
c_n	unique root of Φ_n
$\hat{c}_n$	estimate of c^*
$\hat{c}_n^-$	lower estimate of c^*
$\tilde{c}$	profitability threshold associated to $r_{\mathcal{E}}$
C	upper bound for X_i
D	upper bound for r
E	lower bound for r
$\mathcal{E}$	subset of sub-profitable tasks, introduced in Lemma 4.3
h	size of bins
$\mathcal{H}_t$	history at time t
K	$\mathrm{card}(\mathrm{supp}(X_1))$
K_n	$\mathrm{card}\,\{X_i \mid 1 \leq i \leq n\}$
L	constant of the Hölder class, defined in Definition 4.1
M	number of bins
n_B	number of received task proposals in the bin B
N_B	$\sum_{i=1}^{n} \mathbb{1}(X_i \in B \text{ and } a_i = 1)$
$N_n(x)$	$\sum_{i=1} \mathbb{1}(X_i = x \text{ and } a_i = 1)$, defined in **??**
p	reward per time unit function, defined in Equation (A.1)
p_n	empirical counterpart of p, defined in Equation (A.2)
q_ε	$\mathbb{E}[X\,\mathbb{1}(r(X) \geq (c^* + \varepsilon)X)]$, defined in Equation (3.2)
r	reward function
$\hat{r}_n$	estimate of r
$r(X_i)$	$\mathbb{E}[Y_i]$
$\hat{r}_n^B$	$\hat{r}_n(x)\ \forall x \in B$, defined in Equation (4.1)
$\hat{r}_n^+$	upper estimate of r
$\tilde{r}_n^B$	defined in Equation (4.3)
$r_{\mathcal{E}}(x)$	$r(x)\,\mathbb{1}(x \notin \mathcal{E})$, introduced in Lemma 4.3
R	regret, defined in Equation (2.5)
s	threshold
s^*	threshold such that $r(s^*)/s^* = c^*$
s_n	$\min \mathcal{S}_n$
S_i	idling time after i-th task
$\mathcal{S}_n$	set of all potentially optimal thresholds, defined in Equation (A.4)
T	problem horizon
t_n	time at which the n-th task is proposed
$\mathcal{T}_n$	amount of time after n proposals
U_π	expected reward, defined in Equation (2.1)
v	value function
w	bound on the value function
x^B	$\min\{x \in B\}$
X_i	duration of the i-th task
Y_i	reward of the i-th task
α	parameter of the margin condition
β	exponent of the Hölder class, defined in Definition 4.1
γ_n	$\frac{\lambda\,\mathbb{E}[r(X_n)\,\mathbb{1}(r(X_n) \geq c_n X_n)]}{1+\lambda\,\mathbb{E}[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n)]}$, defined in Equation (3.4)
δ	probability bound, introduced in Proposition 3.1
$\Delta_{\min}$	$\inf\{c^* x - r(x) \,\vert\, r(x) < c^* x\}$, introduced in Theorem A.1
ε_i	subgaussian noise
ζ_n	defined in Equation (A.3)
η_{n-1}	defined in Equation (4.2)
θ	total number of accepted task proposals
κ	universal constant
λ	arrival rate of Poisson process

ξ_{n-1}	defined in Equation (4.5)
π	policy
π^*	threshold policy
σ^2	proxy of the subgaussian noise
Φ	function $c \mapsto \lambda \mathbb{E}\left[(r(X) - cX)_+\right] - c$, defined in Equation (2.4)
Φ_n	empirical counterpart of Φ, defined in Equation (3.1)

Appendix

A Fast rates

In this section, we investigate several cases where the general slow rate convergence can be improved. We use two assumptions about the distribution of task durations X_i and one about the reward function r. For the sake of clarity, the proofs of all the results from this section are deferred to Appendix F.

A.1 Finite support

The first assumption that can be used to obtain fast rates of convergence is the so-called "finite support assumption". Under this assumption, the learning algorithm is quite simple and it consists in keeping an under-estimation $\hat{c}_n^-$ of c^* and an upper-estimation $\hat{r}_n^+(x)$ of $r(x)$ for all different values of x, whose number is denoted by $K = \operatorname{card}(\operatorname{supp}(X_1))$.

The finite algorithm is then based on Algorithm 2, with the difference that each bin corresponds to a single value of the support of the distribution: $B_j = \{x_j\}$, where x_j is the j-th element in $\operatorname{supp}(X_1)$. The estimates $\hat{c}_n$ and $\hat{r}_n$ are then computed similarly to Algorithm 2, but their uncertainty bounds are tighter. This yields the following optimistic estimates:

$$\hat{r}_n^+(x) = \hat{r}_n(x) + \sigma\sqrt{\frac{\ln(K/\delta)}{2N_B}} \qquad \text{for } B = \{x\} \text{ and}$$

$$\hat{c}_n^- = \hat{c}_n - 2\lambda\sqrt{\sigma^2 + \frac{(D-E)^2}{4}}\sqrt{\frac{\ln(1/\delta)}{n}} - \lambda\sigma\sqrt{\frac{K}{2n}} - 8\lambda\frac{K(D-E)}{n}.$$

The task X_{n+1} is then accepted if and only if $\hat{r}_n^+(X_{n+1}) \geq \hat{c}_n^- X_{n+1}$.

Theorem A.1. *Assume that the support of X has cardinality K. Then, taking $\delta = 1/T$, the finite algorithm's regret scales as*

$$R(T) \leq \kappa_2(1 + \lambda C)\max(\sigma, D-E)\sqrt{KT\ln(T)},$$

where κ_2 is some universal constant independent of any problem parameter.

Moreover, if $\Delta_{\min} \coloneqq \inf\{c^*x - r(x) \,|\, r(x) < c^*x\}$, then

$$R(T) \leq \kappa_3(1 + \lambda^2 C^2)\max(\sigma, D-E)^2 K\frac{\ln T}{\Delta_{\min}},$$

for another universal constant κ_3.

Remark A.2. The algorithm complexity scales with K in this particular setting. The size K of the support is assumed to be known for simplicity. For an unknown K, this algorithm can be adapted by using the observed value $K_n \coloneqq \operatorname{card}\{X_i \mid 1 \leq i \leq n\}$ instead of K and restarting every time K_n increases. Such an algorithm yields a similar performance, if not better in practice.

A.2 Margin condition

The finite support assumption gives very fast learning rates but is quite strong. It is possible to study some "intermediate" regime between the general Hölder case and the finite support assumption, thanks to the margin condition recalled below. Intuitively, this condition states that slightly suboptimal tasks have a small probability of being proposed, thus limiting the regret from declining them.

Definition A.3. The distribution of task durations X_i satisfies the *margin condition* with parameter $\alpha \geq 0$ if there exists $\kappa_0 > 0$ such that for every $\varepsilon > 0$,

$$\mathbb{P}\left(c^* X - \varepsilon \leq r(X) < c^* X\right) \leq \kappa_0 \varepsilon^\alpha.$$

If the margin condition α is small enough (i.e., the problem is not too easy and $\alpha < 1$), then we can apply the very same algorithm as in the general case when r is (L, β)-Hölder. We also assume that the distribution of task durations has a positive density lower-bounded by $\underline{\kappa} > 0$ and upper bounded by $\overline{\kappa} > 0$; with a slight abuse of notation, we call this being Lebesgue-equivalent.

Theorem A.4. *If r is (L, β)-Hölder, the distribution of X satisfies the margin condition with (unknown) parameter $\alpha \in [0, 1)$, and is Lebesgue-equivalent, then the regret of Algorithm 2 (with same δ and M as in Theorem 4.5) scales as*

$$R(T) \leq \kappa_4 \sigma^{1+\alpha} \left(\lambda C\right)^{1+\alpha} L^{\frac{1+\alpha}{2\beta+1}} (\lambda T + 1)^{1 - \frac{\beta}{2\beta+1}(1+\alpha)} \log(\lambda T + 1)^{\frac{1+\alpha}{2}},$$

where κ_4 is a constant depending only on $\underline{\kappa}$ and κ_0 defined in the margin condition assumption.

Remark A.5. The proof follows the lines of Theorem 4.1 in Perchet and Rigollet (2013). The term $\log(\lambda T + 1)$ can be removed with a refined analysis using decreasing confidence levels $\delta_n = 1/n^2$ instead of a fixed δ.
For larger margin conditions $\alpha \geq 1$, improved rates are still possible using an adaptive binning instead. The analysis should also follow the lines of Perchet and Rigollet (2013).

A.3 Monotone profitability function

We consider in this section an additional assumption on the profitability function $x \mapsto r(x)/x$, motivated by practical considerations. We assume that this function is non-decreasing (as the impact of some fixed cost declines with the task duration) – the analysis would follow the same line for a non-increasing profitability function (to model diminishing returns for instance). A sufficient condition is convexity of the reward function r with $r(0) \leq 0$. It is not difficult to see that a non-decreasing profitability function ensures the existence of a threshold[1] $s^* \in [0, C]$, such that $r(s^*)/s^* = c^*$. Moreover, by monotonicity, it is optimal to accept a task with duration x if and only if $x \geq s^*$. Hence the problem reduces to learning this optimal threshold based on the reward gathered in the first tasks. In particular, we consider plug-in strategies that accept any task whose duration is large enough (and decline the other ones). The reward per time unit of these strategies is then a function of the threshold s, defined by

$$p(s) = \frac{\lambda\, \mathbb{E}[r(X)\, \mathbb{1}(X \geq s)]}{1 + \lambda\, \mathbb{E}[X\, \mathbb{1}(X \geq s)]}. \tag{A.1}$$

Our policy uses and updates a threshold s_n for all rounds. Let $S = 2(\lambda T + 1)$ be an upper bound on the total number of received tasks with high probability. As the objective is to find the optimum of the function $p(\cdot)$, we introduce its empirical counterparts defined, at the end of stage n, by

$$p_n(s) = \frac{\frac{\lambda}{n} \sum_{i=1}^n Y_i\, \mathbb{1}(X_i \geq s)}{1 + \frac{\lambda}{n} \sum_{i=1}^n X_i\, \mathbb{1}(X_i \geq s)}. \tag{A.2}$$

Given some confidence level $\delta \in (0, 1]$ – to be chosen later –, we also introduce the error term

$$\zeta_n = \left(\sqrt{\sigma^2 + \frac{(D-E)^2}{4}} + \frac{D-E}{\sqrt{2}}(\lambda C + 2)\right) \sqrt{\frac{\ln\left(\frac{2(S+1)}{\delta}\right)}{n-1}} + \lambda \frac{D-E}{n}, \tag{A.3}$$

in the estimation of $p(\cdot)$ by $p_n(\cdot)$, see Lemma F.1 in the Appendix.

The policy starts with a first threshold $s_1 = 0$. After observing a task, the first estimate $p_1(\cdot)$ is computed as in Equation (A.2) and we proceed inductively. After $n-1$ tasks, given the estimate $p_{n-1}(\cdot)$ on $[s_{n-1}, C]$, we compute a lower estimate of $p(s^*)$ by $\max_s p_n(s) - \zeta_n$ (for notational convenience, we also introduce $s^*_{n-1} = \arg\max_{s \in [s_{n-1}, C]} p_{n-1}(s)$). A threshold is then potentially

[1] For non-continuous reward, we only have that $\lim_{\substack{x \to s^* \\ x < s^*}} \frac{r(x)}{x} \leq c^* \leq \lim_{\substack{x \to s^* \\ x > s^*}} \frac{r(x)}{x}$, but the following remains valid.

optimal if its associated reward per time unit (plus the error term) $p_n(s) + \zeta_n$ exceeds this lower estimate. Mathematically, the set of all potentially optimal thresholds is therefore:

$$\mathcal{S}_n = \left\{ s \in [s_{n-1}, C] \,\middle|\, (p_{n-1}(s) - p_{n-1}(s^*_{n-1}))\left(\frac{1}{\lambda} + \frac{1}{n-1}\sum_{i=1}^{n-1} X_i \mathbb{1}(X_i \geq s)\right) + 2\zeta_{n-1} \geq 0 \right\}. \tag{A.4}$$

We then define the new threshold to be used in the next stage as $s_n = \min \mathcal{S}_n$. The rationale behind this choice is that it provides information for all the (potentially optimal) thresholds in $\mathcal{S}_n$. The pseudo-code is given in Algorithm 3 below.

We can finally state the main result when profitability is monotone. Regret scales as $\sqrt{T}$ independently of the regularity of the reward function r.

Theorem A.6. *Assume the profitability function $x \mapsto r(x)/x$ is non-decreasing. Then, the regret of Algorithm 3 (taking $\delta = 1/T^2$) scales as*

$$R(T) \leq \kappa_5(\sigma + D - E)(1 + \lambda C)\sqrt{(\lambda T + 1)\ln(\lambda T)}, \tag{A.5}$$

where κ_5 is a constant independent from any problem parameter.

ALGORITHM 3: Non-decreasing profitability algorithm

input : δ, λ, C, D, σ, E, T
$s_1 = 0$; $S = 2\lambda T + 1$; $n = 0$
while $\sum_{i=1}^{n} S_i + X_i \mathbb{1}(X_i \geq s_i) < T$ **do**
  $n = n + 1$
  Wait S_n and observe X_n
  if $X_n \geq s_n$ **then** accept task and receive reward Y_n
  else reject task
  Compute for $s \in [s_n, C]$, $\quad p_n(s) = \frac{\lambda \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}(X_i \geq s)}{1 + \lambda \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}(X_i \geq s)}$
  Compute $s^*_n = \arg\max_{s \in [s_n, C]} p_n(s)$
  Let $\mathcal{S}_{n+1} = \left\{ s \in [s_n, C] \,|\, (p_n(s) - p_n(s^*_n))\left(\frac{1}{\lambda} + \frac{1}{n}\sum_{i=1}^{n} X_i \mathbb{1}(X_i \geq s)\right) + 2\zeta_n \geq 0 \right\}$
  Set $s_{n+1} = \min \mathcal{S}_{n+1}$
end

Remark A.7. The function $p_n(s)$ is piecewise constant in Algorithm 3 and thus only requires to be computed at all the points X_i. The algorithm then requires to store the complete history of tasks and has a complexity of order n when receiving the n-th proposition.
However, a trick similar to Remark 3.5 leads to improved complexity. The algorithm actually requires to compute $p_n(s)$ only for all $s \in \mathcal{S}_n$. Only tasks of length in $\mathcal{S}_n$ then have to be individually stored and the computation of all "interesting" values of p_n scales with the number of X_i in $\mathcal{S}_n$. This yields a sublinear complexity in n under regularity conditions on the distribution of X.

A.4 Lower bounds

All the algorithms we propose are actually "optimal" for their respective class of problems. Here, optimality must be understood in the minimax sense, as a function of the horizon T and up to $\operatorname{poly}\log(T)$ terms. More formally, a minimax lower-bound for a given class of problems (say, (L, β)-Hölder reward functions with bandit noisy feedback) states that, for any given algorithm, there always exists a specific problem instance such that $R(T) \geq \kappa' T^{\gamma}$ for some number $\gamma \in [0, 1]$ and κ' is independent of T. An algorithm whose regret never increases faster than $\kappa'' T^{\gamma} \log^{\gamma'}(T)$, for some constants κ'' and γ' independent of T is then *minimax optimal*.

Proposition A.8. *Our algorithms are all minimax optimal for their respective class of problems.*

The proof is rather immediate and postponed to Appendix F.3. Indeed, the agent is facing a problem more complicated than a contextual (one-armed) bandit problem because c^* is unknown at first. On the other hand, if we assume that it is given to the agent, the problem is actually a one-armed contextual bandit problem. As a consequence, lower bounds for this problem are also lower-bounds for the agent's problem; as lower and upper bounds match (up to $\operatorname{poly}\log$ terms), the devised algorithms are optimal.

B Fast implementation using search trees

First recall that the complexity of computing the n-th decision of Algorithm 1 is of order n in worst cases, while Algorithm 2 scales with M, which is of order $T^{\frac{1}{2\beta+1}}$. The computational complexities of these algorithms at the n-th round can actually be respectively improved to $\log(n)$ and $\log(M)$, while keeping the same space complexity, by the use of augmented balanced binary search trees data structures, e.g., augmented red black trees (Cormen et al., 2009, Chapters 12-14). Such kind of trees stores nodes with keys in an increasing order. The search, insert and deletion function can thus all be computed in a time $\log(n)$, where n is the number of nodes in the tree. An augmented structure allows to store additional information at each node, which is here used to compute the root of Φ_n.

Fast Algorithm 1. For Algorithm 1, each task i is stored as a node with the key $r(X_i)/X_i$, with the additional piece of information $(r(X_i), X_i)$. Using augmented trees, it is possible to also store in each node i the sum of $r(X_j)$ and X_j for all the nodes j in its right subtree, and similarly for its left subtree with the same complexity.

The main step of Algorithm 1 is the computation of c_n, which can now be computed in a time $\log(n)$, thanks to the described structure. Indeed, note that

$$\Phi_n\left(\frac{r(X_i)}{X_i}\right) = \frac{\lambda}{n}\left(\sum_{j:\ \frac{r(X_j)}{X_j} > \frac{r(X_i)}{X_i}} r(X_j) - \frac{r(X_i)}{X_i} \sum_{j:\ \frac{r(X_j)}{X_j} > \frac{r(X_i)}{X_i}} X_j\right) - \frac{r(X_i)}{X_i}.$$

As the right subtree of the root exactly corresponds to the tasks j with a larger profitability than the root, $\Phi_n(c)$ can be computed in time 1 where c is the profitability of the root node, using the additional stored information. As Φ_n is decreasing, the algorithm continues in the left subtree if $\Phi_n(c) < 0$ and in the right subtree otherwise. Evaluating Φ_n at the key of the following node also takes a time 1 using the different stored information. The algorithm then searches the tree in a time $\log(n)$ to find the task i^* with the largest profitability such that $\Phi_n\left(\frac{r(X_{i^*})}{X_{i^*}}\right) > 0$.

Note that Φ_n is piecewise linear with a partition given by the profitability of the received tasks. In particular, it is linear on $[\frac{r(X_{i^*})}{X_{i^*}}, c_n]$ and c_n is given by

$$c_n = \frac{\lambda \sum_{j:\ \frac{r(X_j)}{X_j} > \frac{r(X_{i^*})}{X_{i^*}}} r(X_j)}{n + \lambda \sum_{j:\ \frac{r(X_j)}{X_j} > \frac{r(X_{i^*})}{X_{i^*}}} X_j}.$$

Recall that it was suggested in Remark 3.5 to remove all tasks with profitability outside

$$\left[c_n - \lambda(D-E)\sqrt{\frac{\ln(1/\delta)}{2n}},\ c_n + \lambda(D-E)\sqrt{\frac{\ln(1/\delta)}{2n}}\right]$$

and to only store the sum of rewards and durations for those above $c_n + \lambda(D-E)\sqrt{\frac{\ln(1/\delta)}{2n}}$. Removing all nodes above or below some threshold in a search tree is called the split operation and can be done in time $\log(n)$ with the structures used here. Thanks to this, keeping a reduced space complexity is possible without additional computational cost.

Fast Algorithm 2. For Algorithm 2, each node represents a bin with the key $(\frac{\tilde{r}_n^B}{x^B}, B)$ and the additional information $(n_B\, \tilde{r}_n^B, n_B\, x^B)$. B is also required in the key to handle the case of bins with same profitabilities.

Here again at each node, we also store the sum of the additional information of the left and right subtrees. Note that at each time step, the algorithm updates $\tilde{r}_n^B$ for the bin of the received task. This is done by deleting the node with the old value[2] of $\tilde{r}_n^B/x^B$ and insert a node with the new value.

The computation of $\hat{c}_n$ then follows the lines of the computation of c_n in Algorithm 1 above.

[2]The algorithm also requires to store a list to find $\tilde{r}_n^B/x^B$ in time 1 for any bin B.

Unfortunately, search trees do not seem adapted to Algorithm 3 since maximizing the function $p_n(s)$ requires to browse the whole tree here.

C Proofs of Section 2

Proof of Proposition 2.1. For $h > 0$, let $B(t,h)$ denote the event "the agent is proposed exactly one task in the interval $[t, t+h]$." Since the probability of receiving more than one task in this interval is $o(h)$, one has for $t < T$

$$\begin{aligned} v(t) &= \mathbb{P}(B(t,h))\,\mathbb{E}[\max(r(X)+v(t+X), v(t+h)) \,|\, B(t,h)] + (1-\mathbb{P}(B(t,h)))v(t+h) + o(h) \\ &= (\lambda h + o(h))\,\mathbb{E}[\max(r(X)+v(t+X), v(t+h)) \,|\, B(t,h)] + (1-\lambda h + o(h))v(t+h) + o(h). \end{aligned}$$

By definition of the value function, the optimal strategy accepts a task X at time t if and only if $r(X) + v(t+X) \geq v(t)$. Hence

$$\frac{v(t+h)-v(t)}{h} = -(\lambda + o(1))\,\mathbb{E}[(r(X)+v(t+X)-v(t+h))_+ \,|\, B(t,h)] + o(1). \tag{C.1}$$

Letting $h \to 0$ one has $v'(t) = -\lambda\,\mathbb{E}[(r(X)+v(t+X)-v(t))_+]$.

Since $(t,v) \mapsto -\lambda\,\mathbb{E}[(r(X)+v(t+X)-v(t))_+]$ is uniformly Lipschitz-continuous in v, uniqueness of the solution follows the same lines as the proof of the Picard-Lindelöf (a.k.a. Cauchy-Lipschitz) theorem. Hence, v is the unique solution of the dynamic programming Equation (2.2).

Note that w is actually the unique solution of the same equation without the boundary effect

$$\begin{cases} w'(t) = -\lambda\,\mathbb{E}\left[(r(X)+w(t+X)-w(t))_+\right] & \text{for all } t \in \mathbb{R}, \\ w(T) = 0. \end{cases} \tag{C.2}$$

Similarly to the argument above, w is the value function of the optimal strategy of an alternative program, defined as the original program, with the difference that, if the agent ends at time $t > T$, then she suffers an additional loss $(t-T)c^*$ in her reward.

The optimal strategy does not only maximize the value function at time $t = 0$, but actually maximizes it for any time $t \in [0,T]$. By definition, any strategy earns less in the alternative program than in the original program. By optimality of the strategy giving the value v, this yields $v(t) \geq w(t)$ for any time t.

By translation, $w(\cdot - C)$ is also the value of the optimal strategy in the alternative program, but with horizon $T + C$. As the length of any task is at most C, following the optimal strategy in the original program of horizon T and rejecting all tasks proposed between T and $T + C$ yields exactly the value v in this *delayed* alternative program; thus $w(t-C) \geq v(t)$ by optimality. □

Proof of Theorem 2.2. Recall that the strategies considered in the proof of Proposition 2.1 accept a task X at time t if and only if $r(X) + v(t+X) \geq v(t)$ where v is its value function. Thus, the optimal strategy in the alternative program described in the proof of Proposition 2.1 accepts a task X if and only if $r(X) \geq c^* X$. Moreover, the cumulative reward of this strategy in the original program is larger than $w(0)$. The relation between v and w given by Proposition 2.1 then yields the result. □

D Proofs of Section 3

Proof of Proposition 3.1. Let $\varepsilon > 0$. One has,

$$\begin{aligned} \mathbb{P}(c_n - c^* > \varepsilon) &= \mathbb{P}(\Phi_n(c_n) < \Phi_n(c^*+\varepsilon)) \text{ since } \Phi_n \text{ is decreasing} \\ &= \mathbb{P}(0 < \Phi_n(c^*+\varepsilon)) \text{ since } c_n \text{ is the root of } \Phi_n \\ &\leq \mathbb{P}(\varepsilon < \Phi_n(c^*+\varepsilon) - \Phi(c^*+\varepsilon)) \text{ since } \Phi(c^*+\varepsilon) \leq -\varepsilon \\ &\leq \exp\left(\frac{-2n\varepsilon^2}{\lambda^2(D-E)^2}\right), \end{aligned}$$

where the last inequality follows from Hoeffding's inequality. □

Proposition D.1. *For all $n \geq 1$, the following bound holds*

$$(c^* - \gamma_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right] \;\leq \frac{\lambda(D-E)C}{2}\sqrt{\frac{\pi}{2n}}.$$

Proof of Proposition D.1. First note that the definition of c^* implies

$$c^* = \frac{\lambda\,\mathbb{E}[r(X)\,\mathbb{1}(r(X) \geq c^* X)]}{1 + \lambda\,\mathbb{E}[X\,\mathbb{1}(r(X) \geq c^* X)]}.$$

Let us denote ν the numerator and μ the denominator of the above expression. First decompose

$$\mathbb{1}(r(X) \geq c_n X) = \mathbb{1}(r(X) \geq c^* X) + \mathbb{1}(c^* X > r(X) \geq c_n X) - \mathbb{1}(c_n X > r(X) \geq c^* X).$$

Denote

$$Q = \lambda\,\mathbb{E}\left[X\Big(\mathbb{1}(c^* X > r(X) \geq c_n X) - \mathbb{1}(c_n X > r(X) \geq c^* X)\Big)\right].$$

One has

$$\begin{aligned}
\gamma_n &= \frac{\nu + \lambda\,\mathbb{E}[r(X)(\mathbb{1}(c^* X > r(X) \geq c_n X) - \mathbb{1}(c_n X > r(X) \geq c^* X))]}{\mu + \lambda\,\mathbb{E}[X(\mathbb{1}(c^* X > r(X) \geq c_n X) - \mathbb{1}(c_n X > r(X) \geq c^* X))]} \\
&\geq \frac{\nu + \lambda\,\mathbb{E}[c_n X(\mathbb{1}(c^* X > r(X) \geq c_n X) - \mathbb{1}(c_n X > r(X) \geq c^* X))]}{\mu + Q} \\
&= c^* + \frac{\mathbb{E}\left[(c_n - c^*)Q\right]}{\mu + Q} \\
&\geq c^* - \frac{\lambda C\,\mathbb{E}[(c_n - c^*)_+]}{1 + \lambda\,\mathbb{E}\left[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n)\right]} \\
&\geq c^* - \frac{\lambda(D-E)C}{2}\sqrt{\frac{\pi}{2n}}\frac{1}{\mathbb{E}\left[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right]},
\end{aligned}$$

where the last inequality follows from Proposition 3.1. □

Proof of Theorem 3.3. From Equation (3.3) one has

$$\begin{aligned}
R(T) &= \mathbb{E}\left[\sum_{n=1}^{\theta}(c^* - \gamma_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right]\right] + c^*\left(T - \mathbb{E}\left[\sum_{n=1}^{\theta} X_n\,\mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right]\right) \\
&\leq \mathbb{E}\left[\sum_{n=1}^{\theta}(c^* - \gamma_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(r(X_n) \geq c_n X_n) + S_n\right]\right] \\
&\leq \frac{\lambda(D-E)C}{2}\sqrt{\frac{\pi}{2}}\,\mathbb{E}\left[\sum_{n=1}^{\theta}\sqrt{\frac{1}{n}}\right] \text{ by Proposition D.1} \\
&\leq \lambda(D-E)C\sqrt{\frac{\pi}{2}}\,\mathbb{E}\left[\sqrt{\theta}\right] \text{ by Wald's formula.}
\end{aligned}$$

It remains to control the expected number of tasks observed $\mathbb{E}\,\theta$. First remark that

$$\theta - 1 \leq \min\{n \geq 1 \,|\, \sum_{i=1}^{n} S_i > T\} - 1 = \sup\{n \geq 1 \,|\, \sum_{i=1}^{n} S_i \leq T\} \tag{D.1}$$

and that the law of the right-hand side of Equation (D.1) is Poisson of parameter λT. So we get that $\mathbb{E}\,\theta \leq \lambda T + 1$ and thus $\mathbb{E}\sqrt{\theta} \leq \sqrt{\lambda T + 1}$ by Jensen's inequality. □

E Proofs of Section 4

Proof of Lemma 4.2. It follows from Hoeffding's inequality, and the fact that r is Hölder continuous. □

Proof of Lemma 4.3. The proof of this crucial lemma is actually straightforward, and follows from the fact that $c^* \geq 0$ and thus $(r_{\mathcal{E}}(X) - c^* X)_+ = (r(X) - c^* X)_+$. □

Proof of Proposition 4.4. Let us, in the first step, assume that all the first N tasks have been accepted, so that (X_i, Y_i) are i.i.d., and $\mathbb{E}[Y_i | X_i] = r(X_i)$. We define

$$\bar{\Phi}_n : c \mapsto \lambda \, \mathbb{E}[\Big(\hat{r}_n(X) - c x^{B(X)}\Big)_+] - c,$$

where $B(X)$ is the bin corresponding to X. Hence for all $c \geq 0$, $\mathbb{E}[\hat{\Phi}_n(c)] = \bar{\Phi}_n(c)$.

Second, remark that for all $c \geq 0$

$$\begin{aligned} |\bar{\Phi}_n(c) - \Phi(c)| &= \lambda |\, \mathbb{E}[(\hat{r}_n(X) - c x^{B(X)})_+ - (r(X) - cX)_+]| && \text{(E.1)} \\ &\leq \lambda \, \mathbb{E}[|\hat{r}_n(X) - r(X)|] + \lambda c h && \text{(E.2)} \\ &\leq \lambda \, \mathbb{E}[|\hat{r}_n(X) - r(X)|] + \lambda^2 D h. && \text{(E.3)} \end{aligned}$$

The last inequality is obtained by noting that $c^* \leq \lambda D$, and thus we only consider $c \leq \lambda D$.

Third, for a fixed $c \geq 0$, $\hat{\Phi}_n(c)$ as a function of $(Y_1, \ldots, Y_n)$, is $\frac{\lambda}{n}$-Lipschitz with respect to each variable. Hence,

$$\begin{aligned} \mathbb{P}(\hat{c}_n - c^* > \varepsilon) &= \mathbb{P}(\hat{\Phi}_n(\hat{c}_n) < \hat{\Phi}_n(c^* + \varepsilon)) \text{ since } \hat{\Phi}_n \text{ is decreasing} \\ &= \mathbb{P}(0 < \hat{\Phi}_n(c^* + \varepsilon)) \text{ since } \hat{c}_i \text{ is the root of } \hat{\Phi}_i \\ &= \mathbb{P}(-\bar{\Phi}_n(c^* + \varepsilon) < \hat{\Phi}_n(c^* + \varepsilon) - \bar{\Phi}_n(c^* + \varepsilon)) \\ &\leq \mathbb{P}(-\Phi(c^* + \varepsilon) - \lambda \, \mathbb{E}[|\hat{r}_n(X) - r(X)|] - \lambda^2 D h < \hat{\Phi}_n(c^* + \varepsilon) - \bar{\Phi}_n(c^* + \varepsilon)) \text{ by } \textit{Equation} \text{ (E.3)} \\ &\leq \mathbb{P}(\varepsilon - \lambda \, \mathbb{E}[|\hat{r}_n(X) - r(X)|] - \lambda^2 D h < \hat{\Phi}_n(c^* + \varepsilon) - \bar{\Phi}_n(c^* + \varepsilon)) \text{ since } \Phi(c^* + \varepsilon) \leq -\varepsilon \\ &\leq \exp\left(\frac{-n(\varepsilon - \lambda \, \mathbb{E}[|\hat{r}_n(X) - r(X)|] - \lambda^2 D h)^2}{4\lambda^2(\sigma^2 + \frac{(D-E)^2}{4})} \right) \text{ by McDiarmid's inequality.} \end{aligned}$$

The last inequality uses McDiarmid's inequality for subgaussian variables (Kontorovich, 2014, Theorem 1). We conclude using Györfi et al. (2002, Corollary 11.2), which yields that $\mathbb{E}[|\hat{r}_n(X) - r(X)|] \leq \kappa \max(\sigma, \frac{D-E}{2}) \sqrt{\frac{\log(n)+1}{hn}} + \sqrt{8} \frac{L h^\beta}{2^\beta}$.

Now, we focus on the case where some of the first N tasks have been declined by the agent. We are going to prove the result by induction. Consider the event where, on the first $n-1$ tasks, it always happened that $\hat{r}_i^+(\cdot) \geq r(\cdot)$ and $\hat{c}_i^- \leq c^*$. As a consequence, on this event, only sub-optimal tasks are declined. Using Lemma 4.3, this yields that the optimal value per time step associated to $r(\cdot)$ and to $r(\cdot) \, \mathbb{1}(\tilde{r}_n(\cdot) > 0)$ are equal.

As a consequence, the precedent arguments can be applied to the following function

$$\tilde{\Phi}_n : c \mapsto \lambda \, \mathbb{E}\left[(\tilde{r}_n(X) - c x^{B(X)}) \right]$$

instead of $\bar{\Phi}_n$. Although Györfi et al. (2002, Corollary 11.2) requires the function r to be (β, L)-Hölder, it also holds here as $r(\cdot) \, \mathbb{1}(\tilde{r}_n(\cdot) > 0)$ is (β, L)-Hölder on every interval of the subdivision $B_1, \ldots, B_M$.

□

Note that in the proof above, we always consider $\varepsilon > \sqrt{8} \lambda \frac{L h^\beta}{2^\beta} + \lambda^2 D h$. Thus, by removing the term $\sqrt{8} \lambda \frac{L h^\beta}{2^\beta} + \lambda^2 D h$ in ξ_{n-1}, all tasks with profitability above $c^* + \varepsilon$ remain observed. With a slight modification of the last argument, it can still be shown that $\hat{\Phi}_n(\hat{c}_n^-) < 0$ with high probability. While this has no influence on the theoretical order of the regret, it is used in the experiments as it yields a significant practical improvement.

Proof of Theorem 4.5. Recall that upon the n-th call, the agent computes $\hat{r}_{n-1}$ and $\hat{c}_{n-1}$ and accepts the task $X_n \in B$ if and only if $\hat{r}^+_{n-1}(X_n) \geq \hat{c}^-_{n-1} x^{B(X_n)}$. Let us denote by $A(n)$ this event so that the regret can be decomposed into

$$R(T) = c^* T - \mathbb{E}\left[\sum_{n=1}^{\theta} r(X_n)\,\mathbb{1}(A(n))\right]$$

where

$$\theta \coloneqq \min\{N \in \mathbb{N} \,|\, \sum_{n=1}^{N} S_n + X_n\,\mathbb{1}\left(A(n)\right) > T\}.$$

The regret can then be rewritten as

$$\begin{aligned} R(T) &= c^* T - \mathbb{E}\left[\sum_{n=1}^{\theta} \mathbb{E}\left[r(X_n)\,\mathbb{1}(A(n))\right]\right] \text{ by Wald's equation} \\ &= c^* T - \mathbb{E}\left[\sum_{n=1}^{\theta} \hat{\gamma}_n\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right]\right], \end{aligned}$$

where the reward per time unit of task n is

$$\hat{\gamma}_n \coloneqq \frac{\lambda\,\mathbb{E}[r(X_n)\,\mathbb{1}(A(n))]}{1 + \lambda\,\mathbb{E}[X_n\,\mathbb{1}(A(n)]}.$$

We further decompose the regret into

$$\begin{aligned} R(T) &= \mathbb{E}\left[\sum_{n=1}^{\theta} (c^* - \hat{\gamma}_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right]\right] + c^*\left(T - \mathbb{E}\left[\sum_{n=1}^{\theta} X_n\,\mathbb{1}(A(n)) + S_n\right]\right) \\ &\leq \mathbb{E}\left[\sum_{n=1}^{\theta} (c^* - \hat{\gamma}_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right]\right] \\ &\leq \mathbb{E}\left[\sum_{n=1}^{\theta} 2C\xi_{n-1} + 4nD\delta + c^* h + 2\,\mathbb{E}[\eta_{n-1}(X_n)]\right], \end{aligned}$$

where the last inequality is a consequence of the following Proposition E.1. As a consequence, it only remains to bound the last term

$$\begin{aligned} &\mathbb{E}\left[\sum_{n=1}^{\theta} \mathbb{E}[\eta_{n-1}(X_n)]\right] \\ &= \mathbb{E}\left[\sum_{n=1}^{\theta} \mathbb{E}\left[\sum_{j=1}^{M} \eta_{n-1}(B_j)\,\mathbb{1}(X_n \in B_j)\right]\right] \\ &= \mathbb{E}\left[\sum_{n=1}^{\theta} \mathbb{E}\left[\sum_{j=1}^{M} \sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}} \sqrt{\frac{\ln(M/\delta)}{2N_{B_j}(n-1)}}\,\mathbb{1}(X_n \in B_j)\right] + \theta L\left(\frac{C}{M}\right)^{\beta}\right] \\ &= \sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}} \sqrt{\ln(M/\delta)}\,\mathbb{E}\left[\sum_{j=1}^{M}\sum_{n=1}^{\theta} \frac{\mathbb{1}(X_n \in B_j)}{\sqrt{2N_{B_j}(n-1)}}\right] + \theta L\left(\frac{C}{M}\right)^{\beta} \\ &\leq \sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}} \sqrt{\ln(M/\delta)}\,\mathbb{E}\left[\sum_{j=1}^{M} \sqrt{N_{B_j}(\theta - 1)}\right] + \theta L\left(\frac{C}{M}\right)^{\beta} \\ &\leq \sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}} \sqrt{\ln(M/\delta)}\,\mathbb{E}\left[\sqrt{M\theta}\right] + \theta L\left(\frac{C}{M}\right)^{\beta} \end{aligned}$$

$$\leq \sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}}\sqrt{\ln(M/\delta)}\sqrt{M(\lambda T+1)} + (\lambda T+1)L\left(\frac{C}{M}\right)^{\beta}$$

Hence, putting all things together, we get

$$\begin{aligned}
R(T) \leq 2\lambda C\Big(&4\sqrt{\sigma^2 + \frac{(D-E)^2}{4}}\sqrt{\ln(1/\delta)}\sqrt{\lambda T+1} + \kappa\max(\sigma, \frac{D-E}{2})\sqrt{\frac{M}{C}\log(e(\lambda T+1))}\sqrt{\lambda T+1} \\
&+ 2\sqrt{2}\frac{LC^{\beta}}{(2M)^{\beta}}(\lambda T+1) + \lambda D\frac{C}{M}\Big) \\
&+ 4D(\lambda T+1)\delta + c^*(\lambda T+1)\frac{C}{M} \\
&+ 2\sqrt{\sigma^2 + \frac{L^2}{4}\left(\frac{C}{M}\right)^{2\beta}}\sqrt{\ln(M/\delta)}\sqrt{M(\lambda T+1)} + 2(\lambda T+1)L\left(\frac{C}{M}\right)^{\beta}.
\end{aligned}$$

The result follows from the specific choices $\delta = \frac{1}{T^2}$ and $M = \left\lceil CL^{\frac{2}{2\beta+1}}(\lambda T+1)^{\frac{1}{2\beta+1}}\right\rceil$. □

Proposition E.1. *For all $n \geq 1$,*

$$0 \leq (c^* - \hat{\gamma}_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right] \leq 2C\xi_{n-1} + 4nD\delta + 2\,\mathbb{E}[\eta_{n-1}(X_n)] + c^*h.$$

Proof. Recall that

$$c^* = \frac{\lambda\,\mathbb{E}[r(X)\,\mathbb{1}(r(X) \geq c^*X)]}{1 + \lambda\,\mathbb{E}[X\,\mathbb{1}(r(X) \geq c^*X)]} = \frac{\nu}{\mu}.$$

We first decompose the indicator function in the numerator of $\hat{\gamma}_n$. Recall that the task n is accepted (the event $A(n)$ occurs) if $\hat{r}^+_{n-1} \geq \hat{c}^-_{n-1}x^{B(X_n)}$. The following holds

$$\begin{aligned}
&\mathbb{1}(\overbrace{\hat{r}^+_{n-1} \geq \hat{c}^-_{n-1}x^{B(X_n)}}^{A(n)}) + \mathbb{1}(\overbrace{c^*X_n < \hat{r}^+_{n-1}(X_n) < \hat{c}^-_{n-1}x^{B(X_n)}}^{\mathcal{E}_2}) + \mathbb{1}(\overbrace{\hat{r}^+_{n-1}(X_n) \leq c^*X_n \leq r(X_n)}^{\mathcal{E}_4}) \\
&= \mathbb{1}(\underbrace{r(X_n) \geq c^*X_n}_{\mathcal{E}_0}) + \mathbb{1}(\underbrace{c^*X_n \geq \hat{r}^+_{n-1}(X_n) \geq \hat{c}^-_{n-1}x^{B(X_n)}}_{\mathcal{E}_1}) + \mathbb{1}(\underbrace{\hat{r}^+_{n-1}(X_n) > c^*X_n > r(X_n)}_{\mathcal{E}_3}).
\end{aligned}$$

To prove it, just notice the followings:

- $A(n) \cap \mathcal{E}_4 = \mathcal{E}_0 \cap \mathcal{E}_1$,
- $\mathcal{E}_2$ is disjoint with both $A(n)$ and $\mathcal{E}_4$,
- $\mathcal{E}_3$ is disjoint with both $\mathcal{E}_0$ and $\mathcal{E}_1$,
- $A(n) \cup \mathcal{E}_2 \cup \mathcal{E}_4 = \mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_3$.

It then comes

$$\begin{aligned}
\mathbb{1}(A(n)) + \mathbb{1}(\mathcal{E}_2) + \mathbb{1}(\mathcal{E}_4) &= \mathbb{1}(A(n) \cup \mathcal{E}_2 \cup \mathcal{E}_4) + \mathbb{1}(A(n) \cap \mathcal{E}_4) \\
&= \mathbb{1}(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_3) + \mathbb{1}(\mathcal{E}_0 \cap \mathcal{E}_1) \\
&= \mathbb{1}(\mathcal{E}_0) + \mathbb{1}(\mathcal{E}_1) + \mathbb{1}(\mathcal{E}_3).
\end{aligned}$$

The first equality comes from the second point; the second from the first and last point; while the last equality comes from the third point. This gives the following

$$\mathbb{1}(A(n)) = \mathbb{1}(\mathcal{E}_0) + \mathbb{1}(\mathcal{E}_1) - \mathbb{1}(\mathcal{E}_2) + \mathbb{1}(\mathcal{E}_3) - \mathbb{1}(\mathcal{E}_4).$$

The quantity of interest is then rewritten as:

$$\begin{aligned}
(c^* - \hat{\gamma}_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right] &= c^*\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right] - \mathbb{E}\left[r(X_n)\,\mathbb{1}(A(n))\right] \\
&= c^*\,\mathbb{E}\left[X_n\big(\,\mathbb{1}(\mathcal{E}_1) - \mathbb{1}(\mathcal{E}_2) + \mathbb{1}(\mathcal{E}_3) - \mathbb{1}(\mathcal{E}_4)\big)\right]
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}\left[r(X_n)\big(\mathbb{1}(\mathcal{E}_1) - \mathbb{1}(\mathcal{E}_2) + \mathbb{1}(\mathcal{E}_3) - \mathbb{1}(\mathcal{E}_4)\big)\right] \\
&\leq c^* \, \mathbb{E}\left[X_n\big(\mathbb{1}(\mathcal{E}_1) + \mathbb{1}(\mathcal{E}_3)\big)\right] - \mathbb{E}\left[r(X_n)\big(\mathbb{1}(\mathcal{E}_1) - \mathbb{1}(\mathcal{E}_2) + \mathbb{1}(\mathcal{E}_3) - \mathbb{1}(\mathcal{E}_4)\big)\right]
\end{aligned}
$$

Let us now bound the last four terms.

1. Recall that $\mathcal{E}_1 = \{c^* X_n \geq \hat{r}^+_{n-1}(X_n) \geq \hat{c}^-_{n-1} x^{B(X_n)}\}$, so that

$$
\begin{aligned}
\mathbb{E}[r(X_n)\,\mathbb{1}(\mathcal{E}_1)] =& \mathbb{E}[\hat{r}^+_{n-1}(X_n)\,\mathbb{1}(\mathcal{E}_1)] + \mathbb{E}[(r(X_n) - \hat{r}^+_{n-1}(X_n))\,\mathbb{1}(\mathcal{E}_1)] \\
\geq& \mathbb{E}[\hat{c}^-_{n-1} x^{B(X_n)}\,\mathbb{1}(\mathcal{E}_1)] - \mathbb{E}[|r(X_n) - \hat{r}^+_{n-1}(X_n)|\,\mathbb{1}(\mathcal{E}_1)] \\
\geq& c^* \, \mathbb{E}[x^{B(X_n)}\,\mathbb{1}(\mathcal{E}_1)] - C\,\mathbb{E}[|\hat{c}^-_{n-1} - c^*|\,\mathbb{1}(\mathcal{E}_1)] - \mathbb{E}[|r(X_n) - \hat{r}^+_{n-1}(X_n)|\,\mathbb{1}(\mathcal{E}_1)] \\
\geq& c^* \, \mathbb{E}[X_n\,\mathbb{1}(\mathcal{E}_1)] - C\,\mathbb{E}[|\hat{c}^-_{n-1} - c^*|\,\mathbb{1}(\mathcal{E}_1)] - \mathbb{E}[|r(X_n) - \hat{r}^+_{n-1}(X_n)|\,\mathbb{1}(\mathcal{E}_1)] + c^* h \mathbb{P}(\mathcal{E}_1).
\end{aligned}
$$

2. $\mathcal{E}_2$ happens with probability at most $n\delta$ since $x^{B(X_n)} \leq X_n$ and using Proposition 4.4, which upper bounds the second term by $Dn\delta$.

3. Recall that $\mathcal{E}_3 = \{\hat{r}^+_{n-1}(X_n) > c^* X_n > r(X_n)\}$ so that the third term is bounded as

$$
\begin{aligned}
\mathbb{E}[r(X_n)\,\mathbb{1}(\mathcal{E}_3)] &= \mathbb{E}[\hat{r}^+_{n-1}(X_n)\,\mathbb{1}(\mathcal{E}_3)] + \mathbb{E}[(r(X_n) - \hat{r}^+_{n-1}(X_n))\,\mathbb{1}(\mathcal{E}_3)] \\
&\geq c^* \, \mathbb{E}[X_n\,\mathbb{1}(\mathcal{E}_3)] - \mathbb{E}[|r(X_n) - \hat{r}^+_{n-1}(X_n)|\,\mathbb{1}(\mathcal{E}_3)].
\end{aligned}
$$

4. $\mathcal{E}_4$ happens with probability at most $n\delta$ thanks to Lemma 4.2, which upper bounds the fourth term by $Dn\delta$.

Putting everything together we get

$$
\begin{aligned}
(c^* - \hat{\gamma}_n)\,\mathbb{E}\left[X_n\,\mathbb{1}(A(n)) + S_n\right] \leq \mathbb{E}[|r(X_n) - \hat{r}^+_{n-1}(X_n)|\,\mathbb{1}(\mathcal{E}_1 \cup \mathcal{E}_3)] + C\,\mathbb{E}[|\hat{c}^-_{n-1} - c^*|\,\mathbb{1}(\mathcal{E}_1)] \\
+ 2Dn\delta + c^* h.
\end{aligned}
$$

The result follows by noting that on $\mathcal{E}_1$, with probability at least $1 - n\delta$, then $c^* - 2\xi_{n-1} \leq \hat{c}^-_{n-1}$, and similarly for $\hat{r}^+_{n-1} - r$ on $\mathcal{E}_1 \cup \mathcal{E}_3$. □

F Proofs of Appendix A

F.1 Proofs of Appendices A.1 and A.2

Proof of Theorem A.1. The proof of this theorem is almost a direct consequence of the proofs of Proposition E.1 and Theorem 4.5, it only requires a few tweaks.

Similarly to Lemma 4.2, it can be shown that $|\hat{r}_n(x) - r(x)| \leq \sigma\sqrt{\frac{\ln(K/\delta)}{2N\{x\}}}$ with probability at least $1 - 2N\delta$ for all $n \leq N$ and x. The uncertainty on $\hat{c}_n$ is shown similarly, except for the term $\mathbb{E}[|\hat{r}_n(X_n) - r(X_n)|]$ which is bounded as follows:

$$\begin{aligned}
\mathbb{E}[|\hat{r}_n(X_n) - r(X_n)|] &= \mathbb{E}\left[\sum_x p(x)\,\mathbb{E}[|\hat{r}_n(X_n) - r(X_n)| \mid N_{\{x\}}]\right] \\
&\leq \mathbb{E}\left[\sum_x p(x)\min\left(\frac{\sigma}{2\sqrt{N_{\{x\}}}}, D - E\right)\right] \quad \text{using Hoeffding's inequality} \\
&\leq \sum_x p(x)\left(\frac{\sigma}{\sqrt{2p(x)n}} + (D-E)e^{-\frac{p(x)n}{8}}\right) \\
&\leq \sigma\sqrt{\frac{K}{2n}} + \frac{8K(D-E)}{n}.
\end{aligned}$$

The Chernoff's bound $\mathbb{P}(N_{\{x\}} \leq \frac{p(x)}{2n}) \leq e^{-\frac{p(x)n}{8}}$ yields the second inequality; while the third one comes from Cauchy-Schwarz inequality and uses that $pe^{-\frac{pn}{8}} \leq \frac{8}{n}$.

Following the same arguments and as standard in multi-armed bandit, we basically need to compute the number of times tasks x are incorrectly accepted. Consider the event where, for all tasks, it holds $\hat{r}_n^+(x) \leq r(x) + 2(D-E)\sqrt{\frac{\ln(K/\delta)}{N(x)}}$ (with the standard convention that $1/0 = +\infty$) and that $\hat{c}_n^- \geq c^* - 4\lambda(D-E)\sqrt{K\frac{\ln(1/\delta)}{n}}$. Then, for n such that $r(x) + 2\eta_n(x) < (c^* - 2\xi_n)x$, x stops being accepted. In particular, this yields for some constant κ' that the total number of accepted tasks of duration x (on this event) is smaller than

$$N_{\{x\}} \leq \kappa' \frac{K\lambda^2(C^2 + (D-E)^2) + (\sigma^2 + (D-E)^2)(1 + \lambda^2 C^2)\ln(K/\delta)}{(c^*x - r(x))^2} + \sqrt{K},$$

which gives a contribution to the regret of the order of (up to a multiplicative constant)

$$\frac{K\lambda^2(C^2 + (D-E)^2) + (\sigma^2 + (D-E)^2)(1 + \lambda^2 C^2)\ln(K/\delta)}{c^*x - r(x)} + \sqrt{K}(c^*x - r(x)),$$

We conclude as usual thanks to the choice of δ that ensures that the contribution to the regret on the complimentary event is negligible. □

Proof of Theorem A.4. Similarly to the proof of Theorem 4.5, denoting $\Delta(x) = c^*x - r(x)$, the regret can be decomposed as

$$\begin{aligned}
R(T) &\leq \mathbb{E}\sum_{n=1}^{\theta}\sum_{j=1}^{M} \mathbb{1}(x \in B_j)\Delta(x)\,\mathbb{1}(2\eta_{n-1}(x) + 2\xi_{n-1}C \geq \Delta(x) \geq 0) + 4nD\delta \\
&\leq \mathbb{E}\sum_{n=1}^{\theta}\sum_{j=1}^{M} \mathbb{1}(x \in B_j)\Delta(x)\,\mathbb{1}(4\eta_{n-1}(x)C \geq \Delta(x) \geq 0) + \mathbb{1}(x \in B_j)\Delta(x)\,\mathbb{1}(4\xi_{n-1}C \geq \Delta(x) \geq 0) + 4nD\delta
\end{aligned} \tag{F.1}$$

The contribution of the third term can be bounded similarly to Theorem 4.5 and is $\mathcal{O}(1)$.

Using the margin condition, the second term scales with $\sum_{n=1}^{\lambda T+1}(C\xi_{n-1})^{1+\alpha}$, which is of order

$$(\lambda C L^{1-\frac{2}{2\beta+1}})^{1+\alpha}(\lambda T + 1)^{1-\frac{\beta}{2\beta+1}(1+\alpha)}$$

.

It now remains to bound the first term. It can be done using the analysis of Perchet and Rigollet (2013), which we sketch here.

The idea is to divide the bins into two categories for some constant c_1 scaling with $\sigma C^{\frac{2\beta+1}{2}} L\sqrt{\log(\lambda T+1)}$:

- *well behaved* bins, for which $\exists x \in B, \Delta(x) \geq c_1 M^{-\beta}$,
- *ill behaved* bins, for which $\forall x \in B, \Delta(x) < c_1 M^{-\beta}$.

The first term in Equation (F.1) for ill behaved bins is directly bounded, using the margin condition, by a term scaling with

$$c_1^{1+\alpha} M^{-\beta(1+\alpha)}(\lambda T+1) \approx \sigma^{1+\alpha} C^{\frac{1+\alpha}{2}} L^{\frac{1+\alpha}{2\beta+1}} \log(\lambda T+1)^{\frac{1+\alpha}{2}} (\lambda T+1)^{1-\frac{\beta}{2\beta+1}(1+\alpha)}.$$

Now denote $\mathcal{J} \subset \{1,\dots,M\}$ the set of well behaved bins and for any $j \in \mathcal{J}$, $\Delta_j = \max_{x \in B_j} \Delta(x)$. Using classical bandit arguments, the event $\{x \in B_j\} \cap \{4\eta_{n-1}(x)C \geq \Delta(x) \geq 0\}$ only holds at most $\kappa''(\sigma^2 + L^2\left(\frac{C}{M}\right)^2)\frac{\log(\lambda T)}{\Delta_j^2}$ times.

Assuming w.l.o.g. that $\mathcal{J} = \{1,\dots,j_1\}$ and $\Delta_1 \leq \dots \leq \Delta_{j_1}$, the margin condition can be leveraged to show that $\Delta_j \geq \left(\frac{\kappa j}{\kappa_0 M}\right)^{\frac{1}{\alpha}}$. The contribution of well behaved bins then scales as

$$(\sigma^2 + L^2\left(\frac{C}{M}\right)^2)\sum_{j=1}^{j_1} \frac{\log(\lambda T+1)}{\Delta_j} \leq \sigma^2 \log(\lambda T+1)\left(\frac{j_2 M^\beta}{c_1} + \sum_{j=j_2+1}^{j_1}\left(\frac{j}{M}\right)^{-\frac{1}{\alpha}}\right), \quad \text{(F.2)}$$

where j_2 is some integer of order $c_1^\alpha M^{1-\alpha\beta}$ so that for any $j \leq j_2$, $c_1 M^{-\beta} \geq \left(\frac{\kappa j}{\kappa_0 M}\right)^{\frac{1}{\alpha}}$.

The first term of Equation (F.2) can then be bounded by

$$(\sigma^2 + L^2\left(\frac{C}{M}\right)^2) c_1^{\alpha-1} M^{1+\beta(1-\alpha)} \log(\lambda T+1) \lesssim \sigma^{1+\alpha} C^{\frac{1+\alpha}{2}} L^{\frac{1+\alpha}{2\beta+1}} (\lambda T+1)^{1-\frac{\beta}{2\beta+1}(1+\alpha)} \log(\lambda T+1)^{\frac{1+\alpha}{2}},$$

where $\lesssim$ means that the inequality holds up to some universal constant κ'.

The sum in Equation (F.2) can be bounded as follows:

$$\begin{aligned}\sum_{j=j_2+1}^{j_1}\left(\frac{j}{M}\right)^{-\frac{1}{\alpha}} &\leq \int_{\frac{j_2}{M}}^1 x^{\frac{1}{\alpha}} \mathrm{d}x \\ &\lesssim \frac{c_1^{\alpha-1} M^{\beta(1-\alpha)}}{1-\alpha}.\end{aligned}$$

Finally, the first term of Equation (F.1) scales as

$$\sigma^{1+\alpha} C^{\frac{1+\alpha}{2}} L^{\frac{1+\alpha}{2\beta+1}} (\lambda T+1)^{1-\frac{\beta}{2\beta+1}(1+\alpha)} \log(\lambda T+1)^{\frac{1+\alpha}{2}},$$

which finally yields Theorem A.4 when gathering everything. □

F.2 Proofs of Appendix A.3

The following Lemma indicates that with arbitrarily high probability, our upper/lower estimations are correct

Lemma F.1. *With a constant κ_6 independent from n, the events*

$$|p_n(s) - p(s)|\,\mathbb{E}[X\,\mathbb{1}(X \geq s) + \frac{1}{\lambda}] \leq \zeta_n + \frac{\kappa_6}{n}.$$

hold with probability at least $1-\delta$, simultaneously for all $n \in \{1,\dots,S\}$ and all $s \in [s_n, C]$.

Proof. Let

$$p(s) = \frac{\lambda\,\mathbb{E}[r(X)\,\mathbb{1}(X \geq s)]}{1 + \lambda\,\mathbb{E}[X\,\mathbb{1}(X \geq s)]} =: \frac{\nu(s)}{\mu(s)}. \tag{F.3}$$

and

$$p_n(s) = \frac{\lambda \frac{1}{n}\sum_{i=1}^n Y_i\,\mathbb{1}(X_i \geq s)}{1 + \lambda \frac{1}{n}\sum_{i=1}^n X_i\,\mathbb{1}(X_i \geq s)} =: \frac{\nu_n(s)}{\mu_n(s)}. \tag{F.4}$$

Let us rewrite

$$\frac{\mu(s)}{\lambda} = \frac{1}{\lambda} + s(1-F(s)) + \int_s^C 1 - F(x)dx \text{ and}$$
$$\frac{\mu_n(s)}{\lambda} = \frac{1}{\lambda} + s(1-F_n(s)) + \int_s^C 1 - F_n(x)dx,$$

where F_n is the empirical distribution function $F_n(x) \coloneqq \frac{1}{n}\sum_{i=1}^n \mathbb{1}(X_i \leq x)$. Hence

$$\begin{aligned}\frac{|\mu(s) - \mu_n(s)|}{\lambda} &\leq s|F(s) - F_n(s)| + \int_s^C |F(x) - F_n(x)|dx \\ &\leq s|F(s) - F_n(s)| + (C-s)\max_{x\in[s,C]}|F(x) - F_n(x)| \\ &\leq C \max_{x\in[s,C]}|F(x) - F_n(x)|.\end{aligned}$$

The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality ensures that the event $|\mu(s) - \mu_n(s)| \leq \lambda C\sqrt{\frac{1}{2n}\ln\left(\frac{2S}{\delta}\right)},\ \forall s \in [0, C]$, holds with probability at least $1 - \frac{\delta}{S}$.

Denote by $s^1, s^2, \dots, s^n$ a permutation of observed task durations $X_1, \dots, X_n$ such that $s^1 \leq s^2 \leq \dots \leq s^n$. Also define $s^0 = s_n$ for completeness. Let $s \in [s^k, s^{k+1})$, since $E \leq 0$, it comes

$$\begin{aligned}\nu(s) &= \lambda\,\mathbb{E}[r(X)\,\mathbb{1}(X \geq s)] \\ &\leq \nu(s^k) - \lambda E(F(s) - F(s^k)) \\ &\leq \nu(s^k) - \lambda E(F(s^{k+1}) - F(s^k)) \\ &= \nu(s^k) - \lambda E\Big(F(s^{k+1}) - F_n(s^{k+1}) + F_n(s^{k+1}) - F_n(s^k) + F_n(s^k) - F(s^k)\Big)\end{aligned}$$

Hence,

$$\nu(s) \leq \nu(s^k) - 2\lambda E\sqrt{\frac{1}{2n}\ln\left(\frac{S}{\delta}\right)} - \frac{\lambda E}{n} \quad \forall k \in \{1,\dots,n\},$$

holds as soon as the DKW inequality above holds. We prove analogously for the event

$$\nu(s) \geq \nu(s^k) - 2\lambda D\sqrt{\frac{1}{2n}\ln\left(\frac{S}{\delta}\right)} - \frac{\lambda D}{n}\ \forall k \in \{1,\dots,n\}.$$

Recall that $s^k = X_{i_k}$ for some i_k for $k \in \{1,\dots,n\}$. By Hoeffding's inequality with $(X_i)_{i\neq i_k}$ as samples and a union bound on $\{0,\dots,n\}$, the event

$$|\nu(s^k) - \nu_n(s^k)| \leq \lambda\sqrt{\sigma^2 + \frac{(D-E)^2}{4}}\sqrt{\frac{1}{2(n-1)}\ln\left(\frac{2S(n+1)}{\delta}\right)} \quad \forall k \in \{0,\dots,n\}$$

holds with probability at least $1 - \frac{\delta}{S}$.

Putting everything together with the fact that $\nu_n(s) = \nu_n(s^k)$ and $n \leq S$, one gets that the event

$$|\nu_n(s) - \nu(s)| \leq \lambda \left(\sqrt{\sigma^2 + \frac{(D-E)^2}{4}} + \sqrt{2}(D-E) \right) \sqrt{\frac{\ln(2S/\delta)}{n-1}} + \frac{\lambda(D-E)}{n}$$

holds with probability at least $1 - \frac{\delta}{S}$ for all $s \in [s_n, C]$ if the DKW inequality also holds.

Furthermore,

$$\left| \frac{\nu_n(s)}{\mu_n(s)} - \frac{\nu(s)}{\mu(s)} \right| \leq |\nu_n(s)| \left| \frac{\mu(s) - \mu_n(s)}{\mu(s)\mu_n(s)} \right| + \frac{1}{\mu(s)} |\nu_n(s) - \nu(s)| \tag{F.5}$$

$$\leq \frac{\left(\lambda(D-E) + |\nu_n(s) - \nu(s)|\right) |\mu(s) - \mu_n(s)| + |\nu_n(s) - \nu(s)|}{1 + \lambda \, \mathbb{E}[X \, \mathbb{1}(X \geq s)]}. \tag{F.6}$$

Hence with probability at least $1 - \frac{2\delta}{S}$ one has for all $s \in [s_n, C]$,

$$\begin{aligned} |p_n(s) - p(s)| \, \mathbb{E}[X \, \mathbb{1}(X \geq s) + \frac{1}{\lambda}] \leq & \lambda \left(\sqrt{\sigma^2 + \frac{(D-E)^2}{4}} + \frac{D-E}{\sqrt{2}}(\lambda C + 2) \right) \sqrt{\frac{\ln(2S/\delta)}{n-1}} \\ & + \lambda \frac{D-E}{n} \\ & + \lambda C \left(\sqrt{\frac{\sigma^2}{2} + \frac{(D-E)^2}{8}} + D - E \right) \frac{\ln\left(\frac{2S}{\delta}\right)}{n-1} \\ & + \lambda C \frac{D-E}{\sqrt{2} n^{3/2}} \sqrt{\ln\left(\frac{2S}{\delta}\right)}. \end{aligned}$$

We conclude using a union bound over $\{1, \ldots, S\}$. □

The goal of the next lemma is to ensure that the sequence $(s_n)_{1 \leq n \leq S}$ is indeed bellow the the optimal threshold s^* with high probability.

Lemma F.2. *With probability at least $1 - \delta$, the events*

$$s_n \leq s^* \quad \textit{and} \quad 0 \leq (p_n(s_n^*) - p_n(s^*)) \left(\frac{1}{\lambda} + \frac{1}{n} \sum_{i=1}^{n} X_i \, \mathbb{1}(X_i \geq s) \right) \leq 2\zeta_n,$$

hold simultaneously in all stages $n \in \{1, \ldots, S\}$.

Proof. The proof is by induction on $n \geq 1$. For $n = 1$, $s_1 = 0 \leq s^*$. We prove in a similar manner to Lemma F.1, that the events

$$(p(s) - p_n(s)) \left(\frac{1}{\lambda} + \frac{1}{n} \sum_{i=1}^{n} X_i \, \mathbb{1}(X_i \geq s) \right) \leq \zeta_n \quad \forall s \in [s_n, C] \tag{F.7}$$

simultaneously hold for all stages $n \in \{1, \ldots, S\}$, with probability at least $1 - \delta$. The only difference with Lemma F.1 is that the term $|\nu_n(s) - \nu(s)|$ in Equation (F.6) does not appear here, removing the $\frac{\kappa_6}{n}$ term in Lemma F.1. In particular, this implies the events

$$(p(s^*) - p_1(s^*)) \left(\frac{1}{\lambda} + X_i \, \mathbb{1}(X_1 \geq s) \right) \leq \zeta_1$$

$$\text{and } (p_1(s_1^*) - p(s_1^*)) \left(\frac{1}{\lambda} + X_1 \, \mathbb{1}(X_1 \geq s) \right) \leq \zeta_1$$

Since $p(s_1^*) \leq p(s^*)$, one has that

$$(p_1(s_1^*) - p_1(s^*)) \left(\frac{1}{\lambda} + X_i \, \mathbb{1}(X_1 \geq s) \right) \leq 2\zeta_1$$

holds when Equation (F.7) holds.

Now for any $n > 1$, the induction assumption implies that $s^* \in \mathcal{S}_n$, hence $s^* \geq s_n$. The rest of the induction follows the steps of the base case $n = 1$. Finally, one has $0 \leq p_n(s_n^*) - p_n(s^*)$ by definition of s_n^*. □

The following proposition allows to control the regret in stage n.

Proposition F.3. *For some constant κ_7 independent from n, with probability at least $1-\delta$, the events*

$$(p(s^*) - p(s_n))\,\mathbb{E}[X\,\mathbb{1}(X \geq s_n) + \frac{1}{\lambda}] \leq 4\zeta_{n-1} + \frac{\kappa_7}{n}.$$

hold simultaneously in all stages $n \in \{2, \dots, S\}$.

Proof. With probability at least $1-\delta$, the following inequality hold uniformly for all stages $n \geq 2$ by Lemma F.1:

$$(p(s^*) - p(s_n))\,\mathbb{E}[X\,\mathbb{1}(X \geq s_n) + \frac{1}{\lambda}] \leq (p(s^*) - p_{n-1}(s_n))\,\mathbb{E}[X\,\mathbb{1}(X \geq s_n) + \frac{1}{\lambda}] + \zeta_{n-1} + \frac{\kappa_6}{n}$$

Now remark, using the notation of Lemma F.1 that, thanks to the DKW inequality and the fact that $s_n \in \mathcal{S}_n$, under the same probability event,

$$\begin{aligned}(p_{n-1}(s_{n-1}^*) - p_{n-1}(s_n))\frac{\mu(s_n)}{\lambda} &= (p_{n-1}(s_{n-1}^*) - p_{n-1}(s_n))\frac{\mu_n(s_n)}{\lambda} + (p_{n-1}(s_{n-1}^*) - p_{n-1}(s_n))\frac{\mu(s_n) - \mu_n(s_n)}{\lambda} \\ &\leq 2\zeta_{n-1} + 2\zeta_{n-1}C\sqrt{\frac{\ln(\frac{4S}{\delta})}{2(n-1)}}.\end{aligned}$$

The definition of $\mathcal{S}_n$ directly bounds the first term. For the second term, note that the definition of $\mathcal{S}_n$ also bounds $p_{n-1}(s_{n-1}^*) - p_{n-1}(s_n)$. Combined with the concentration on $\mu_n - \mu$, this bounds the second term. It then follows

$$\begin{aligned}(p(s^*) - p(s_n))\frac{\mu(s_n)}{\lambda} &\leq (p(s^*) - p_{n-1}(s_{n-1}^*))\,\mathbb{E}[X\,\mathbb{1}(X \geq s_n) + \frac{1}{\lambda}] + 3\zeta_{n-1} + \frac{\kappa_8}{n} \\ &\leq (p(s^*) - p_{n-1}(s^*))\,\mathbb{E}[X\,\mathbb{1}(X \geq s_{n-1}) + \frac{1}{\lambda}] + 3\zeta_{n-1} + \frac{\kappa_8}{n} \\ &\leq 4\zeta_{n-1} + \frac{\kappa_8 + \kappa_6}{n} \text{ by Lemma F.1 and since } s^* \geq s_{n-1} \text{ by Lemma F.2.}\end{aligned}$$

□

Lemma F.4. *For $S = 2\lambda T + 1$, we have*

$$\mathbb{E}\left[(\theta - S)_+\right] \leq 4.$$

Proof. Similarly to the proof of Theorem 3.3, note that $\theta - 1$ is dominated by a random variable following a Poisson distribution of parameter λT. It now just remains to show that for any random variable Z following a Poisson distribution of parameter λ, it holds that $\mathbb{E}\left[(Z - 2\lambda)_+\right] \leq 4$.

Thanks to Canonne (2017), the cdf of Z can be bounded as follows for any positive x:

$$\mathbb{P}(Z \geq \lambda + x) \leq e^{-\frac{x^2}{2(\lambda + x)}},$$

which implies for $x = \lambda + t$ with $t > 0$:

$$\begin{aligned}\mathbb{P}(Z - 2\lambda \geq t) &\leq e^{-\frac{(t+\lambda)^2}{2(2\lambda + t)}} \\ &\leq e^{-\frac{t+\lambda}{4}}.\end{aligned}$$

From there, the expectation of $(Z - 2\lambda)_+$ can be directly bounded:

$$\begin{aligned}\mathbb{E}[(Z - 2\lambda)_+] &\leq \int_0^\infty e^{-\frac{t+\lambda}{4}}\,\mathrm{d}t \\ &\leq 4.\end{aligned}$$

This concludes the proof. □

Proof of Theorem A.6. Here we have

$$\theta = \min\{n \in \mathbb{N} \,|\, \sum_{i=1}^{n} S_i + X_i \mathbb{1}\left(X_i \geq s_i\right) > T\}.$$

Algorithm 3 yields the regret

$$\begin{aligned}
R(T) &= c^* T - \mathbb{E}\left[\sum_{i=1}^{\theta} r(X_i)\, \mathbb{1}(X_i \geq s_i)\right] \\
&= c^* T - \mathbb{E}\left[\sum_{i=1}^{\theta} \mathbb{E}\left[r(X_i)\, \mathbb{1}(X_i \geq s_i)\right]\right] \text{ by Wald's equation} \\
&= c^* T - \mathbb{E}\left[\sum_{i=1}^{\theta} p(s_i)\, \mathbb{E}\left[X_i\, \mathbb{1}(X_i \geq s_i) + S_i\right]\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^{\theta} (c^* - p(s_i))\, \mathbb{E}\left[X_i\, \mathbb{1}(X_i \geq s_i) + S_i\right]\right].
\end{aligned}$$

Note that conversely to the previous sections, when using Wald's equation the expectation here is taken conditionally to threshold s_i. Bounding separately the first two terms and the whole sum for small probability events, Proposition F.3 yields

$$\begin{aligned}
R(T) &\leq \sum_{n=3}^{S} \left(4\zeta_{n-1} + \frac{\kappa_7}{n}\right) + 2(D-E) + (D-E)(S\delta + \mathbb{E}\left[(\theta - S)_+\right]) \\
&\leq 8\left(\sqrt{\sigma^2 + \frac{(D-E)^2}{4}} + \frac{D-E}{\sqrt{2}}(\lambda C + 2)\right)\sqrt{S \ln\left(\frac{2(S+1)}{\delta}\right)} \\
&\quad + (4\lambda(D-E) + \kappa_7)\ln(eS) + (D-E)(2 + S\delta + \mathbb{E}\left[(\theta - S)_+\right])
\end{aligned}$$

Using the given values for S, δ and Lemma F.4 finally yields Theorem A.6. □

F.3 Proofs of Appendix A.4

Proof of Proposition A.8.
1) We first prove for all algorithms in Section 4 and Appendix A, when the reward function is unknown and noisy observations are observed.

Following the same lines of the proof of Theorem 4.5, we first write the regret in a bandit form:

$$R(T) \geq \mathbb{E}\left[\sum_{n=1}^{\theta} \mathbb{E}\left[\big(r(X_n) - c^* X_n\big)\big(\mathbb{1}(r(X_n) \geq c^* X_n) - \mathbb{1}(A(n))\big)\right]\right] - c^* C. \tag{F.8}$$

Now consider the one arm contextual bandit where, given a context X, the normalized reward is $c^* X$ and the arm returns a noised reward of mean $r(X)$. The sum in Equation (F.8) exactly is the regret incurred by the strategy A in this one armed contextual bandit problem, with horizon θ.

Although θ is random and depends on the strategy of the agent, it is larger than $\frac{\lambda T}{1+C}$ with probability at least $\alpha > 0$, constant in T. Moreover, the one armed contextual bandit problem is easier than the agent problem, as c^* is known only in the former. Thanks to these two points, $R(T)$ is larger than

$$R(T) \geq \alpha \tilde{R}\left(\frac{\lambda T}{1+C}\right) - c^* C$$

where $\tilde{R}$ is the regret in this one armed contextual bandit problem. Proposition A.8 then follows from classical results in contextual bandits (see, e.g., Audibert and Tsybakov, 2007; Rigollet and Zeevi, 2010).

Note here that we only consider a subclass of all the one armed contextual bandit problems. Indeed, we fixed the normalized reward to $c^* X$ and the arm reward must satisfy $c^* = \lambda\, \mathbb{E}[(r(X) - c^* X)_+]$. Yet this subclass is large enough and the existing proofs (Audibert and Tsybakov, 2007; Rigollet and Zeevi, 2010) can be easily adapted to this setup.

2) It now remains to prove a $\Omega(\sqrt{T})$ bound when the reward function r is known. Consider the following setting

$$X = \begin{cases} 1 \text{ with proba } \frac{1}{2} + \varepsilon \\ 2 \text{ with proba } \frac{1}{2} - \varepsilon \end{cases}$$
$$\text{and} \quad \begin{cases} r(1) = \frac{1}{2} \\ r(2) = 2. \end{cases}$$

Now consider two worlds where $\varepsilon = \Delta$ in the first one, and $\varepsilon = -\Delta$ in the second one for some $\Delta > 0$. Basic calculations then give the following

$$c_1^* = \frac{1}{2} - \frac{2\Delta}{5} + o(\Delta) \quad \text{in the first world,}$$
$$c_2^* = \frac{1}{2} + \frac{\Delta}{2} + o(\Delta) \quad \text{in the second world.}$$

The optimal strategy then accepts all tasks in the first world, while it only accepts tasks of duration 2 in the second one. Classical lower bound techniques (see, e.g., Lattimore and Szepesvári, 2020, Theorem 14.2) then show that for $\Delta = \frac{1}{\sqrt{\lambda T}}$, with some constant probability and positive constant α both independent from T, any strategy

- either rejects $\alpha\lambda T$ tasks $X_t = 1$ in the first world,
- or accepts $\alpha\lambda T$ tasks $X_t = 1$ in the second world after receiving λT task propositions.

In any case, this means that any strategy has a cumulative regret of order $\sqrt{T}$ in at least one world. $\square$