Optimal Resource Allocation with Semi-Bandit Feedback

Tor Lattimore
Dept. of Computing Science
University of Alberta, Canada

Koby Crammer
Dept. of Electrical Engineering
The Technion, Israel

Csaba Szepesvári*
Microsoft Research
Redmond, USA

Abstract

We study a sequential resource allocation problem involving a fixed number of recurring jobs. At each time-step the manager should distribute available resources among the jobs in order to maximise the expected number of completed jobs. Allocating more resources to a given job increases the probability that it completes, but with a cut-off. Specifically, we assume a linear model where the probability increases linearly until it equals one, after which allocating additional resources is wasteful. We assume the difficulty of each job is unknown and present the first algorithm for this problem and prove upper and lower bounds on its regret. Despite its apparent simplicity, the problem has a rich structure: we show that an appropriate optimistic algorithm can improve its learning speed dramatically beyond the results one normally expects for similar problems as the problem becomes resource-laden.

1 INTRODUCTION

Assume that there are K jobs and at each time-step t a learner must distribute the available resources with $M_{k,t} \geq 0$ going to job k, subject to a budget constraint,

$$\sum_{k=1}^{K} M_{k,t} \leq 1.$$

The probability that the kth job completes in time-step t is $\min\{1, M_{k,t}/\nu_k\}$, where the unknown cut-off parameter $\nu_k \in (0, \infty]$ determines the difficulty of job k. After every time-step the resources are replenished and all jobs are restarted regardless of whether or not they completed successfully in the previous time-step. The goal of the learner is to maximise the expected number of jobs that successfully complete up to some known time horizon n.

Despite the simple model, the problem is surprisingly rich. Given its information structure, the problem belongs to the class of stochastic partial monitoring problems, which was first studied by Agrawal et al. [1989][1], where in each time step the learner receives noisy information about a hidden "parameter" while trying to maximise the sum of rewards and both the information received and the rewards depend in a known fashion on the actions and the hidden parameter. While partial monitoring by now is relatively well understood, either in the stochastic or the adversarial framework when the action set is finite [Bartók et al., 2011, Foster and Rakhlin, 2012, Bartók, 2013], the case of continuous action sets has received only limited attention [Broder and Rusmevichientong, 2012, and references therein]. To illustrate the difficulty of the problem, notice that over-assigning resources to a given job means that the job completes with certainty and provides little information about the job's difficulty. On the other hand, if resources are under-assigned, then the information received allows one to learn about the payoff associated with all possible arms, which is reminiscent of bandit problems where the arms have "correlated payoffs" (e.g., Filippi et al. 2010, Russo and Roy 2013 and the references therein). Finally, allocating less resources yields high-variance estimates.

Our motivation to study this particular framework comes from the problem of cache allocation. In particular, data collected offline from existing and experimental allocation strategies showed a relatively good fit to the above parametric model. In this problem each job is a computer process, which is successful in a given time-step if there were no cache misses (cache misses are very expensive). Besides this specific resource allocation problem, we also envision other applications, such as load balancing in networked environments, or any other computing applications where some precious resource (bandwidth, radio spectrum, CPU, etc.) is to be subdivided amongst competing processes. In fact, we anticipate numerous extensions and adaptations for

*On sabbatical leave from the Department of Computing Science, University of Alberta, Canada

[1]The name was invented later by (perhaps) [Rustichini, 1999].

specific applications, such as in the case of bandits (see, Bubeck and Cesa-Bianchi [2012] for an overview of this rich literature). Finally, let us point out that although our problem is superficially similar to the so-called budgeted bandit problems (or, budget limited bandit problems), there are some major differences: in budgeted bandits, the information structure is still that of bandit problems and the resources are not replenished. Either learning stops when the budget is exhausted (e.g., Tran-Thanh et al. 2012, Ding et al. 2013, Badanidiyuru et al. 2013)[2], or performance is measured against the total resources consumed in an ongoing fashion (e.g., György et al. 2007).

The main contribution besides the introduction of a new problem is a new optimistic algorithm for this problem that is shown to suffer poly-logarithmic regret with respect to optimal omniscient algorithm that knows the parameters $(\nu_k)_k$ in advance. The structure of the bound depends significantly on the problem dynamics, ranging from a (relatively) easy full-information-like setting, corresponding to a resource-laden regime, to a bandit-like setting, corresponding to the resource-scant setting. Again, to contrast this work to previous works, note that the results we obtain for the full-information-like setting are distinct from those possible in the finite action case, where the full-information setting allows one to learn with finite regret [Agrawal et al., 1989]. On the technical side, we believe that our study and use of weighted estimators in situations where some samples are more informative than others might be of independent interest, too.

Problems of allocating resources to jobs were studied in the community of architecture and operating systems. Liu et al. [2004] build static profile-based allocation of L2-cache banks to different processes using their current miss rate data. Suh et al. [2002] proposed a hit-rate optimisation using hardware counters which used a model-based estimation of hit-rate vs allocated cache. However, they all assume the model is fully known and no learning is required. Bitirgen et al. [2008] used ANNs to predict individual program performance as a function of resources. Finally, Ipek et al. [2008] used reinforcement learning to allocate DRAM to multi-processors.

2 PRELIMINARIES

In each time-step t the learner chooses $M_{k,t} \geq 0$ subject to the constraint, $\sum_{k=1}^{K} M_{k,t} \leq 1$. Then all jobs are executed and $X_{k,t} \in \{0,1\}$ indicates the success or failure of job k in time-step t and is sampled from a Bernoulli distribution with parameter $\beta(M_{k,t}/\nu_k) := \min\{1, M_{k,t}/\nu_k\}$. The goal is to maximise the expected number of jobs that successfully complete, $\sum_{k=1}^{K} \beta(M_{k,t}/\nu_k)$. We define the gaps $\Delta_{j,k} = \nu_j^{-1} - \nu_k^{-1}$. We assume throughout for convenience, and without loss of generality, that $\nu_1 < \nu_2 < \cdots < \nu_K$. It can be shown that the optimal allocation distributes the resources to jobs in increasing order of difficulty.

$$M_k^* = \min\left\{1 - \sum_{i=1}^{k-1} M_i^*,\ \nu_k\right\}.$$

We let ℓ be the number of jobs that are fully allocated under the optimal policy: $\ell = \max\{i : M_i^* = \nu_i\}$. The overflow is denoted by $S^* = M_{\ell+1}^*$, which we assume to vanish if $\ell = K$. The expected reward (number of completed jobs) when following the optimal allocation is

$$\sum_{k=1}^{K} \frac{M_k^*}{\nu_k} = \ell + \frac{S^*}{\nu_{\ell+1}},$$

where we define $\nu_{K+1} = \infty$ in the case that $\ell = K$. The (expected n-step cumulative) regret of a given allocation algorithm is the difference between the expected number of jobs that complete under the optimal policy and those that complete given the algorithm,

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} r_t\right], \quad r_t = \sum_{k=1}^{K} \beta(M_k^*/\nu_k) - \sum_{k=1}^{K} \beta(M_{k,t}/\nu_k)$$
$$= \left(\ell + \frac{S^*}{\nu_{\ell+1}}\right) - \sum_{k=1}^{K} \beta(M_{k,t}/\nu_k).$$

3 OVERVIEW OF ALGORITHM

We take inspiration from the optimal policy for known ν_k, which is to fully allocate the jobs with the smallest ν_k (easiest jobs) and allocate the remainder/overflow to the next easiest job. At each time-step t we replace the unknown ν_k by a high-probability lower bound $\underline{\nu}_{k,t-1} \leq \nu_k$. This corresponds to the optimistic strategy, which assumes that each job is as easy as reasonably possible. The construction of a confidence interval about ν_k is surprisingly delicate. There are two main challenges. First, the function $\beta(M_{k,t}/\nu_k)$ is non-differentiable at $M_{k,t} = \nu_k$, and for $M_{k,t} \geq \nu_k$ the job will always complete and little information is gained. This is addressed by always using a lower estimate of ν_k in the algorithm. The second challenge is that $M_{k,t}$ will vary with time, so the samples $X_{k,t}$ are not identically distributed. This would normally be unproblematic, since martingale inequalities can be applied, but the specific structure of this problem means that a standard sample average estimator is a little weak in the sense that its estimation accuracy can be dramatically improved. In particular, we will propose an estimator that is able to take advantage of the fact that the variance of $X_{k,t}$ decreases to zero as $M_{k,t}$ approaches ν_k from below.

As far as the estimates are concerned, rather than estimate the parameters ν_k, it turns out that learning the reciprocal

[2]Besides Badanidiyuru et al. [2013], all works consider finite action spaces and unstructured reward functions.

ν_k^{-1} is both more approachable and ultimately more useful for proving regret bounds. Fix k and let $M_{k,1}, \ldots, M_{k,t} \leq \nu_k$ be a sequence of allocations with $M_{k,s} \leq \nu_k$ and $X_{k,s} \sim \text{Bernoulli}\left(M_{k,s}/\nu_k\right)$. Then a natural (unbiased) estimator of ν_k^{-1} is given by

$$\frac{1}{\hat{\nu}_{k,t}} := \frac{1}{t}\sum_{s=1}^{t} \frac{X_{k,s}}{M_{k,s}}.$$

The estimator has some interesting properties. First, the random variable $X_{k,s}/M_{k,s} \in [0, 1/M_{k,s}]$ has a large range for small $M_{k,s}$, which makes it difficult to control the error $\hat{\nu}_{k,t}^{-1} - \nu_k^{-1}$ via the usual Azuma/Bernstein inequalities. Secondly, if $M_{k,s}$ is close to ν_k, then the range of $X_{k,s}/M_{k,s}$ is small, which makes estimation easier. Additionally, the variance is greatly decreased for $M_{k,s}$ close to ν_k. This suggests that samples for which $M_{k,s}$ is large are more useful than those where $M_{k,s}$ is small, which motivates the use of the weighted estimator,

$$\frac{1}{\hat{\nu}_{k,t}} := \frac{\sum_{s=1}^{t} w_s X_{k,s}}{\sum_{s=1}^{t} w_s M_{k,s}},$$

where w_s will be chosen in a data-dependent way, but with the important characteristic that w_s is large for $M_{k,s}$ close to ν_k. The pseudo-code of the main algorithm is shown on Algorithm Listing 1. It accepts as input the horizon n, the number of jobs, and a set $\{\underline{\nu}_{k,0}\}_{k=1}^{K}$ for which $0 < \underline{\nu}_{k,0} \leq \nu_k$ for each k. In Section 7 we present a simple (and efficient) algorithm that relaxes the need for the lower bounds $\underline{\nu}_{k,0}$.

Remark 1. Later (in Lemma 6) we will show that with high probability $1 \leq w_{k,s} \leq O(s)$. By definition the confidence bounds $\underline{\nu}_{k,t}$ and $\bar{\nu}_{k,t}$ are non-decreasing/increasing respectively. These results are sufficient to guarantee that the new algorithm is numerically stable. It is also worth noting that the running time of Algorithm 1 is $O(1)$ per time step, since all sums can be computed incrementally.

Algorithm 1 Optimistic Allocation Algorithm

input: $n, K, \{\underline{\nu}_{k,0}\}_{k=1}^{K}$
$\delta \leftarrow (nK)^{-2}$ and $\bar{\nu}_{k,0} = \infty$ for each k
for $t \in 1, \ldots, n$ **do**
  /* Optimistically choose $M_{k,t}$ using $\underline{\nu}_{k,t-1}$ */
  $(\forall k \in 1, \ldots, K)$ initialise $M_{k,t} \leftarrow 0$
  for $i \in 1, \ldots, K$ **do**
    $k \leftarrow \operatorname*{arg\,min}_{k: M_{k,t}=0} \underline{\nu}_{k,t-1}$
    $M_{k,t} \leftarrow \min\left\{\underline{\nu}_{k,t-1},\ 1 - \sum_{j=1}^{K} M_{j,t}\right\}$
  end for
  $(\forall k \in 1, \ldots, K)$ observe $X_{k,t}$
  $(\forall k \in 1, \ldots, K)$ compute weighted estimates:

$$w_{k,t} \leftarrow \frac{1}{1 - \frac{M_{k,t}}{\bar{\nu}_{k,t-1}}} \qquad \frac{1}{\hat{\nu}_{k,t}} \leftarrow \frac{\sum_{s=1}^{t} w_{k,s} X_{k,s}}{\sum_{s=1}^{t} w_{k,s} M_{k,s}}$$

  $(\forall k \in 1, \ldots, K)$ update confidence intervals:

$$R_{k,t} \leftarrow \max_{s \leq t} w_{k,s} \qquad \hat{V}_{k,t}^2 \leftarrow \sum_{s \leq t} \frac{w_{k,s} M_{k,s}}{\underline{\nu}_{k,t-1}}$$

$$\tilde{\varepsilon}_{k,t} \leftarrow \frac{f(R_{k,t}, \hat{V}_{k,t}^2, \delta)}{\sum_{s=1}^{t} w_{k,s} M_{k,s}}$$

$$\frac{1}{\underline{\nu}_{k,t}} \leftarrow \min\left\{\frac{1}{\underline{\nu}_{k,t-1}}, \frac{1}{\hat{\nu}_{k,t}} + \tilde{\varepsilon}_{k,t}\right\}$$

$$\frac{1}{\bar{\nu}_{k,t}} \leftarrow \max\left\{\frac{1}{\bar{\nu}_{k,t-1}}, \frac{1}{\hat{\nu}_{k,t}} - \tilde{\varepsilon}_{k,t}\right\}$$

end for

function $f(R, V^2, \delta)$
  $\delta_0 \leftarrow \frac{\delta}{3(R+1)^2(V^2+1)^2}$
  return $\frac{R+1}{3}\log\frac{2}{\delta_0} + \sqrt{2(V^2+1)\log\frac{2}{\delta_0} + \left(\frac{R+1}{3}\right)^2 \log^2\frac{2}{\delta_0}}$
end function

4 UPPER BOUNDS ON THE REGRET

The regret of Algorithm 1 depends in a subtle way on the parameters ν_k. There are four natural cases, which will appear in our main result.

Case 1: Insufficient budget for any jobs. In this case $\ell = 0$ and the optimal algorithm allocates all available resources to the easiest task, which means $M_1^* = 1$. Knowing that $\ell = 0$, the problem can be reduced to a K-armed Bernoulli bandit by restricting the action space to $M_{k,t} = 1$ for all k. Then a bandit algorithm such as UCB1 [Auer et al., 2002] will achieve logarithmic (problem dependent) regret with some dependence on the gaps $\Delta_{1,k} = \frac{1}{\nu_1} - \frac{1}{\nu_k}$. In particular, the regret looks like $R_n \in O\left(\sum_{k=2}^{K} \frac{\log n}{\Delta_{1,k}}\right)$.

Case 2: Sufficient budget for all jobs. In this case $\ell = K$ and the optimal policy assigns $M_{k,t} = \nu_k$ for all k, which enjoys a reward of K at each time-step. Now Algorithm 1 will choose $M_{k,t} = \underline{\nu}_{k,t-1}$ for all time-steps and by Theorem 4 stated below we will have $\underline{\nu}_{k,t-1}/\nu_k \in O(1 - \frac{1}{t}\log n)$. Consequently, the regret may be bounded by $R_n \in O\left(\log^2 n\right)$ with *no dependence on the gaps*.

Case 3: Sufficient budget for all but one job. Now the algorithm must learn which jobs should be fully allocated. This introduces a weak dependence on the gaps $\Delta_{\ell,k}$ for $k > \ell$, but choosing the overflow job is trivial. Again we expect the regret to be $O(\log^2 n)$, but with an additional modest dependence on the gaps.

Case 4: General case. In the completely general case even

the choice of the overflow job is non-trivial. Ultimately it turns out that in this setting the problem decomposes into two sub-problems. Choosing the jobs to fully allocate, and choosing the overflow job. The first component is fast, since we can make use of the faster learning when fully allocating. Choosing the overflow reduces to the bandit problem as described in case 1.

Our main result is the following theorem bounding the regret of our algorithm.

Theorem 2. *Let δ be as in the algorithm, $\eta_k = \min\{1, \nu_k\}/\underline{\nu}_{k,0}$, $\tilde{\delta}_k = \frac{\delta}{48\eta_k^4 n^6}$, $c_{k,1} = 27\log\frac{2}{\tilde{\delta}_k}$, $c_{k,2} = 6\log\frac{2}{\tilde{\delta}_k}$, $u_{k,j} = \frac{c_{k,1}}{\underline{\nu}_{k,0}\Delta_{j,k}}$. Then Algorithm 1 suffers regret at most*

$$R_n \le 1 + \sum_{k=1}^{\ell} c_{k,1}\eta_k(1+\log n) + \mathbb{1}\{\ell < K\}\left[\sum_{k=\ell+2}^{K} \frac{c_{k,2}}{\underline{\nu}_{k,0}\Delta_{\ell+1,k}} + \sum_{k=1}^{\ell+1} c_{k,1}\eta_k(1+\log n) + \sum_{k=\ell+2}^{K} c_{k,1}\eta_k(1+\log u_{\ell+1,k}) + \sum_{k=\ell+1}^{K} c_{k,1}\eta_k(1+\log u_{\ell,k})\right].$$

If we assume $\eta_k \in O(1)$ for each k (reasonable as discussed in Section 7), then the regret bound looks like

$$R_n \in O\left(\ell\log^2 n + \sum_{k=\ell+1}^{K}\left(\log\frac{1}{\nu_k\Delta_{\ell,k}}\right)\log n + \sum_{k=\ell+2}^{K}\left(\log\frac{1}{\nu_k\Delta_{\ell+1,k}}\right)\log n + \sum_{k=\ell+1}^{K}\frac{\log n}{\Delta_{\ell+1,k}}\right), \tag{1}$$

where the first term is due to the gap between $\underline{\nu}_{k,t}$ and ν_k, the second due to discovering which jobs should be fully allocated, while the third and fourth terms are due to mistakes when choosing the overflow job.

The proof is broken into two components. In the first part we tackle the convergence of $\hat{\nu}_{t,k}$ to ν_k and analyse the width of the confidence intervals, which are data-dependent and shrink substantially faster when $M_{k,t}$ is chosen close to ν_k. In the second component we decompose the regret in terms of the width of the confidence intervals. While we avoided large constants in the algorithm itself, in the proof we focus on legibility. Optimising the constants would complicate an already long result.

5 ESTIMATION

We consider a single job with parameter ν and analyse the estimator and confidence intervals used by Algorithm 1. We start by showing that the confidence intervals contain the truth with high-probability and then analyse the rate at which the intervals shrink as more more data is observed. Somewhat surprisingly the rate has a strong dependence on the data with larger allocations leading to faster convergence.

Let $\{\mathcal{F}_t\}_{t=0}^{\infty}$ be a filtration and let $M_1, \dots, M_n$ be a sequence of positive random variables such that M_t is $\mathcal{F}_{t-1}$-measurable. Define X_t to be sampled from a Bernoulli distribution with parameter $\beta(M_t/\nu)$ for some $\nu \in [\underline{\nu}_0, \infty]$ and assume that X_t is $\mathcal{F}_t$-measurable. Our goal is to construct a sequence of confidence intervals $\{[\underline{\nu}_t, \bar{\nu}_t]\}_{t=1}^{n}$ such that $\nu \in [\underline{\nu}_t, \bar{\nu}_t]$ with high probability and $\bar{\nu}_t - \underline{\nu}_t \to 0$ as fast as possible. We assume a known lower bound $\underline{\nu}_0 \le \nu$ and define $\bar{\nu}_0 = \infty$. Recall that the estimator used by Algorithm 1 is defined by

$$w_s = \frac{1}{1 - \frac{M_t}{\bar{\nu}_{t-1}}}, \qquad \frac{1}{\hat{\nu}_t} = \frac{\sum_{s=1}^{t} w_s X_s}{\sum_{s=1}^{t} w_s M_s}.$$

Fix a number $0 < \delta < 1$ and define $\tilde{\varepsilon}_t = f(R_t, \hat{V}_t^2, \delta)/\sum_{s=1}^{t} w_s M_s$, where the function f is defined in Algorithm 1, $R_t = \max_{s\le t} w_s$ and $\hat{V}_t^2 = \sum_{s=1}^{t}\frac{w_s M_s}{\underline{\nu}_{t-1}}$. The lower and upper confidence bounds on ν^{-1} are defined by,

$$\frac{1}{\underline{\nu}_t} = \min\left\{\frac{1}{\underline{\nu}_{t-1}}, \frac{1}{\hat{\nu}_t} + \tilde{\varepsilon}_t\right\}, \quad \frac{1}{\bar{\nu}_t} = \max\left\{\frac{1}{\bar{\nu}_{t-1}}, \frac{1}{\hat{\nu}_t} - \tilde{\varepsilon}_t\right\}.$$

We define $\varepsilon_t = \underline{\nu}_t^{-1} - \bar{\nu}_t^{-1}$ to be the (decreasing) width of the confidence interval. Note that both $\underline{\nu}_t$ and $\bar{\nu}_t$ depend on δ, although this dependence is not shown to minimise clutter.

Theorem 3. *If M_s is chosen such that $M_s \le \underline{\nu}_{s-1}$ for all s then $\mathbb{P}\{\exists s \le t \text{ s.t. } \nu \notin [\underline{\nu}_s, \bar{\nu}_s]\} \le t\delta$ holds for any $0 < \delta < 1$.*

Proof of Theorem 3. Let F_t be the event $F_t = \{\nu \in [\underline{\nu}_t, \bar{\nu}_t]\}$. Note that since $[\underline{\nu}_t, \bar{\nu}_t] \subset [\underline{\nu}_{t-1}, \bar{\nu}_{t-1}] \subset \cdots \subset [\underline{\nu}_0, \bar{\nu}_0]$, $F_t \subset F_{t-1} \subset \cdots \subset F_0$. Hence, $F_t = \cap_{s\le t} F_s$ and it suffices to prove that $\mathbb{P}\{F_t^c\} \le t\delta$.[3]

Define $Y_s = w_s X_s - \frac{w_s M_s}{\nu}$ and $S_t = \sum_{s=1}^{t} Y_s$ and $V_t^2 = \sum_{s=1}^{t}\mathrm{Var}[Y_s|\mathcal{F}_{s-1}]$. We proceed by induction. Assume $\mathbb{P}\{F_{t-1}^c\} \le (t-1)\delta$, which is trivial for $t = 1$. Now, on F_{t-1},

$$V_t^2 \overset{(a)}{=} \sum_{s=1}^{t}\mathrm{Var}[Y_s|\mathcal{F}_{s-1}] \overset{(b)}{=} \sum_{s=1}^{t}\frac{w_s^2 M_s}{\nu}\left(1 - \frac{M_s}{\nu}\right) \overset{(c)}{=} \sum_{s=1}^{t}\frac{w_s M_s}{\nu}\left(\frac{1 - \frac{M_s}{\nu}}{1 - \frac{M_s}{\bar{\nu}_{s-1}}}\right) \overset{(d)}{\le} \sum_{s=1}^{t}\frac{w_s M_s}{\nu} \overset{(e)}{\le} \hat{V}_t^2,$$

where (a) is the definition of V_t^2, (b) follows since w_s is $\mathcal{F}_{s-1}$-measurable, (c) follows by substituting the definition of w_s, (d) and (e) are true since given F_{t-1} we

[3] For an event E, we use E^c to denote its complement.

know that $\underline{\nu}_{s-1} \le \nu \le \bar{\nu}_{s-1}$. Therefore $f(R_t, V_t^2, \delta) \le f(R_t, \hat{V}_t^2, \delta)$, which follows since f is monotone increasing in its second argument. Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\left|\frac{1}{\hat{\nu}_t} - \frac{1}{\nu}\right| \ge \tilde{\varepsilon}_t \wedge F_{t-1}\right\} \\
&= \mathbb{P}\left\{\left|\frac{\sum_{s=1}^t w_s X_s}{\sum_{s=1}^t w_s M_s} - \frac{1}{\nu}\right| \ge \frac{f(R_t, \hat{V}_t^2, \delta)}{\sum_{s=1}^t w_s M_s} \wedge F_{t-1}\right\} \\
&\le \mathbb{P}\left\{\left|\sum_{s=1}^t w_s X_s - \sum_{s=1}^t \frac{w_s M_s}{\nu}\right| \ge f(R_t, V_t^2, \delta) \wedge F_{t-1}\right\} \\
&= \mathbb{P}\left\{|S_t| \ge f(R_t, V_t^2, \delta) \wedge F_{t-1}\right\}. \qquad (2)
\end{aligned}$$

By the union bound we have

$$\begin{aligned}
&\mathbb{P}\left\{|S_t| \ge f(R_t, \hat{V}_t^2, \delta) \vee F_{t-1}^c\right\} \\
&\le \mathbb{P}\left\{|S_t| \ge f(R_t, V_t^2, \delta) \wedge F_{t-1}\right\} + \mathbb{P}\left\{F_{t-1}^c\right\} \\
&\overset{(a)}{\le} \delta + \mathbb{P}\left\{F_{t-1}^c\right\} \le \delta + (t-1)\delta = t\delta\,,
\end{aligned}$$

where (a) follows from Theorem 13 in the Appendix. Therefore $\mathbb{P}\left\{|S_t| \le f(R_t, V_t^2, \delta) \wedge F_{t-1}\right\} \ge 1 - t\delta$ and so with probability at least $1 - t\delta$ we have that F_{t-1} and

$$\left|\frac{1}{\hat{\nu}_t} - \frac{1}{\nu}\right| \le \frac{f(R_t, \hat{V}_t^2, \delta)}{\sum_{s=1}^t w_s M_s} = \tilde{\varepsilon}_t,$$

in which case

$$\frac{1}{\underline{\nu}_t} = \min\left\{\frac{1}{\underline{\nu}_{t-1}}, \frac{1}{\hat{\nu}_t} + \tilde{\varepsilon}_t\right\} \ge \frac{1}{\nu}\,,$$

and similarly $\frac{1}{\bar{\nu}_t} \le \frac{1}{\nu}$, which implies F_t. Therefore $\mathbb{P}\{F_t^c\} \le t\delta$ as required. □

We now analyse the width $\varepsilon_t \equiv \underline{\nu}_t^{-1} - \bar{\nu}_t^{-1}$ of the confidence interval obtained after t samples are observed. We say that a job is *fully allocated* at time-step s if $M_s = \underline{\nu}_{s-1}$. The first theorem shows that the width ε_t drops with order $O(1/T(t))$, where $T(t) = \sum_{s=1}^t \mathbb{1}\{M_s = \underline{\nu}_{s-1}\}$ is the number of fully allocated time-steps. The second theorem shows that for any $\alpha > 0$, the width ε_t drops with order $O(\sqrt{1/(\alpha U_\alpha(t))})$, where $U_\alpha(t) = \sum_{s=1}^t \mathbb{1}\{M_s \ge \alpha\}$. The dramatic difference in speeds is due to the low variance $\mathrm{Var}[X_t|\mathcal{F}_{t-1}]$ when M_t is chosen close to ν. For the next results define $\eta = \min\{1, \nu\}/\underline{\nu}_0$ and $\tilde{\delta} = \frac{\delta}{48\eta^4 n^6}$.

Theorem 4. $\varepsilon_t \le \dfrac{c_1}{\underline{\nu}_0(T(t)+1)}$ *where* $c_1 = 27\log\frac{2}{\tilde{\delta}}$.

Theorem 5. $\varepsilon_t \le \sqrt{\dfrac{c_2}{\alpha\underline{\nu}_0 U_\alpha(t)}}$ *where* $c_2 = 6\log\frac{2}{\tilde{\delta}}$.

The proofs are based on the following lemma that collects some simple observations:

Lemma 6. *The following hold for any* $t \ge 1$*:*

1. $w_t M_t \le \frac{1}{\varepsilon_{t-1}}$, *with equality if* $M_t = \underline{\nu}_{t-1}$.
2. $1 \le R_t \le \frac{1}{\underline{\nu}_0 \varepsilon_{t-1}}$.
3. $\varepsilon_t \ge \frac{1}{t\min\{1,\nu\}}$.
4. $1 - \frac{\underline{\nu}_t}{\nu} \le \underline{\nu}_t \varepsilon_t$.

Proof. Using the definition of w_s and the fact that M_s is always chosen to be smaller or equal to $\underline{\nu}_{s-1}$, we get

$$w_s \equiv \left(1 - \frac{M_s}{\bar{\nu}_{s-1}}\right)^{-1} \overset{(a)}{\le} \left(1 - \frac{\underline{\nu}_{s-1}}{\bar{\nu}_{s-1}}\right)^{-1} = \frac{1}{\varepsilon_{s-1}\underline{\nu}_{s-1}}\,.$$

The first claim follows since the inequality (a) can be replaced by equality if $M_s = \underline{\nu}_{s-1}$. The second follows from the definition of R_t and the facts that $(\varepsilon_s)_s$ is non-increasing and $(\underline{\nu}_s)_s$ is non-decreasing. For the third claim we recall that $R_t = \max_{s\le t} w_s$ and $M_s \le \nu$. Therefore,

$$\begin{aligned}
\varepsilon_t &\overset{(a)}{\ge} \min\left\{\varepsilon_{t-1}, \frac{R_t}{\sum_{s=1}^t w_s M_s}\right\} \\
&\overset{(b)}{\ge} \min\left\{\varepsilon_{t-1}, \frac{1}{t\min\{1,\nu\}}\right\},
\end{aligned}$$

where (a) follows from the definition of ε_t and naive bounding of the function f, (b) follows since $R_t \ge w_s$ for all $s \le t$ and because $M_s \le \min\{1,\nu\}$ for all s. Trivial induction and the fact that $\varepsilon_0 = \underline{\nu}_0^{-1} \ge \nu^{-1}$ completes the proof of the third claim. For the final claim we use the facts that $\underline{\nu}_t^{-1} \le \nu^{-1} + \varepsilon_t$. Therefore, $1 - \frac{\underline{\nu}_t}{\nu_t} = \underline{\nu}_t\left(\frac{1}{\underline{\nu}_t} - \frac{1}{\nu}\right) \le \underline{\nu}_t\varepsilon_t$. □

Lemma 7. $\varepsilon_t \le \dfrac{6R_t\log\frac{2}{\tilde{\delta}}}{\sum_{s=1}^t w_s M_s} + \sqrt{\dfrac{8\log\frac{2}{\tilde{\delta}}}{\underline{\nu}_0\sum_{s=1}^t w_s M_s}}$.

Proof. Let $\delta_t = \delta/(3(R_t+1)^2(\hat{V}_t^2+1)^2) < 1$. By the definition of ε_t,

$$\begin{aligned}
\varepsilon_t &\le \frac{2f(R_t, \hat{V}_t^2, \delta)}{\sum_{s=1}^t w_s M_s} \\
&\overset{(a)}{\le} \frac{\frac{4(R_t+1)}{3}\log\frac{2}{\delta_t} + 2\sqrt{2(\hat{V}_t^2+1)\log\frac{2}{\delta_t}}}{\sum_{s=1}^t w_s M_s} \\
&\overset{(b)}{\le} \frac{6R_t\log\frac{2}{\delta_t} + \sqrt{\frac{8}{\underline{\nu}_0}\sum_{s=1}^t w_s M_s\log\frac{2}{\delta_t}}}{\sum_{s=1}^t w_s M_s} \\
&= \frac{6R_t\log\frac{2}{\delta_t}}{\sum_{s=1}^t w_s M_s} + \sqrt{\frac{8\log\frac{2}{\delta_t}}{\underline{\nu}_0\sum_{s=1}^t w_s M_s}},
\end{aligned}$$

where in (a) we used the definition of f, in (b) we substituted the definition of $\hat{V}_t^2$ and used the facts that $R_t \ge 1$ and $\underline{\nu}_0 \le \underline{\nu}_{t-1}$ and we also used a naive bound. The proof is completed by proving $2/\delta_t \le 2/\tilde{\delta}$. Indeed, by Lemma 6,

$1 \le R_t \le \frac{1}{\varepsilon_{t-1}\underline{\nu}_0} \le \frac{1}{\varepsilon_t \underline{\nu}_0}$. We also have $\hat{V}_t^2 \le t R_t^2$. Thus,

$$\frac{2}{\delta_t} = \frac{6(R_t+1)^2(\hat{V}_t^2+1)^2}{\delta} \le \frac{6}{\delta}\left(\frac{16t^2}{(\varepsilon_t \underline{\nu}_0)^4}\right) \overset{(a)}{\le} \frac{2}{\tilde{\delta}},$$

where in (a) we used Lemma 6(3). □

Proof of Theorem 4. By Lemma 7,

$$\varepsilon_t \le \frac{6R_t \log\frac{2}{\tilde{\delta}}}{\sum_{s=1}^t w_s M_s} + \sqrt{\frac{8}{\underline{\nu}_0 \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}}}. \tag{3}$$

We proceed by induction. Assume that $\varepsilon_{s-1} \le \frac{c_1}{\underline{\nu}_0(T(s-1)+1)}$, which is trivial for $s=1$. By Lemma 6(1),

$$\sum_{s=1}^t w_s M_s \ge \sum_{s=1}^{T(t)} \frac{s\underline{\nu}_0}{c_1} = \frac{\underline{\nu}_0 T(t)(T(t)+1)}{2c_1}. \tag{4}$$

Therefore,

$$\sqrt{\frac{8}{\underline{\nu}_0 \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}}} \overset{(a)}{\le} \frac{1}{\underline{\nu}_0 T(t)} \sqrt{4c_1 \log\frac{2}{\tilde{\delta}}}. \tag{5}$$

Now we work on the first term in (3). If $\varepsilon_{t-1} \le \frac{c_1}{\underline{\nu}_0(T(t)+1)}$, then we are done, since ε_s is non-increasing. Otherwise, we use Lemma 6(2) to obtain,

$$\begin{aligned}\frac{6R_t}{\sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}} &\le \frac{6}{\underline{\nu}_0 \varepsilon_{t-1} \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}} \\ &\overset{(a)}{\le} \frac{3}{\underline{\nu}_0 T(t)} \log\frac{2}{\tilde{\delta}},\end{aligned} \tag{6}$$

where in (a) we used (4) and the lower bound on ε_{t-1}. Substituting (5) and (6) into (3) we have

$$\varepsilon_t \le \frac{1}{\underline{\nu}_0 T(t)} \sqrt{4c_1 \log\frac{2}{\tilde{\delta}}} + \frac{3}{\underline{\nu}_0 T(t)} \log\frac{2}{\tilde{\delta}}.$$

Choosing $c_1 = 27 \log\frac{2}{\tilde{\delta}}$ leads to

$$\begin{aligned}\varepsilon_t &\le \frac{1}{\underline{\nu}_0 T(t)} \sqrt{4 \cdot 27 \log^2\frac{2}{\tilde{\delta}}} + \frac{3}{\underline{\nu}_0 T(t)} \log\frac{2}{\tilde{\delta}} \\ &\le \frac{27}{\underline{\nu}_0 (T(t)+1)} \log\frac{2}{\tilde{\delta}} = \frac{c_1}{\underline{\nu}_0 (T(t)+1)},\end{aligned}$$

which completes the induction and proof. □

Proof of Theorem 5. Firstly, by Lemma 7,

$$\varepsilon_t \le \frac{6R_t}{\sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}} + \sqrt{\frac{8}{\underline{\nu}_0 \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}}}.$$

The second term is easily bounded by using the fact that $w_s \ge 1$ and the definition of $U_\alpha(t)$,

$$\sqrt{\frac{8}{\underline{\nu}_0 \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}}} \le \sqrt{\frac{8}{\underline{\nu}_0 U_\alpha(t) \alpha} \log\frac{2}{\tilde{\delta}}}.$$

For the first term we assume $\varepsilon_{t-1} \ge \sqrt{\frac{c_2}{\underline{\nu}_0 U_\alpha(t)\alpha}}$, since otherwise we can apply monotonicity of ε_t. Therefore

$$\begin{aligned}\frac{6R_t}{\sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}} &\le \frac{6}{\underline{\nu}_0 \varepsilon_{t-1} \sum_{s=1}^t w_s M_s} \log\frac{2}{\tilde{\delta}} \\ &\le \sqrt{\frac{U_\alpha(t)\alpha\underline{\nu}_0}{c_2}} \cdot \frac{6\log\frac{2}{\tilde{\delta}}}{\underline{\nu}_0 U_\alpha(t)\alpha} \le 6\sqrt{\frac{1}{c_2 \alpha \underline{\nu}_0 U_\alpha(t)}} \log\frac{2}{\tilde{\delta}}.\end{aligned}$$

Now choose $c_2 = 6\log\frac{2}{\tilde{\delta}}$ to complete the result. □

6 PROOF OF THEOREM 2

We are now ready to use the results of Section 5 to bound the regret of Algorithm 1. The first step is to decompose the regret into two cases depending on whether or not the confidence intervals contain the truth. The probability that they do not is low, so this contributes negligibly to the regret. When the confidence intervals are valid we break the problem into two components. The first is the selection of the processes to fully allocation, which leads to the $O(\log^2 n)$ part of the bound. The second component involves analysing the selection of the overflow process, where the approach is reminiscent of the analysis for the UCB algorithm for stochastic bandits [Auer et al., 2002].

Let $F_{k,t}$ denote the event when none of the confidence intervals underlying job k fail up to time t:

$$F_{k,t} = \{\forall s \le t : \nu \in [\underline{\nu}_{k,s}, \bar{\nu}_{k,s}]\}\ .$$

The algorithm uses $\delta = (nK)^{-2}$, which is sufficient by a union bound and Theorem 3 to ensure that,

$$\mathbb{P}\{G^c\} \le \frac{1}{nK}\,, \quad \text{where } G = \bigcap_{k=1}^K F_{k,n}\,. \tag{7}$$

The regret can be decomposed into two cases depending on whether G holds:

$$\begin{aligned}R_n &= \mathbb{E}\sum_{t=1}^n r_t \overset{(a)}{=} \mathbb{E}\,\mathbb{1}\{G^c\}\sum_{t=1}^n r_t + \mathbb{E}\,\mathbb{1}\{G\}\sum_{t=1}^n r_t \\ &\overset{(b)}{\le} \mathbb{E}\,\mathbb{1}\{G^c\}\,nK + \mathbb{E}\,\mathbb{1}\{G\}\sum_{t=1}^n r_t \overset{(c)}{\le} 1 + \mathbb{E}\,\mathbb{1}\{G\}\sum_{t=1}^n r_t,\end{aligned} \tag{8}$$

where (a) follows from the definition of expectation, (b) is true by bounding $r_t \le K$ for all t, and (c) follows from (7). For the remainder we assume G holds and use Theorems 4 and 5 combined with the definition of the algorithm to control the second term in (8). The first step is to decompose the regret in round t:

$$r_t = \ell^* + \frac{S^*}{\nu_{\ell+1}} - \sum_{k=1}^K \beta\left(\frac{M_{k,t}}{\nu_k}\right).$$

By the assumption that G holds we know for all $t \le n$ and k that $\bar{\nu}_{k,t}^{-1} \le \nu_k^{-1} \le \underline{\nu}_{k,t}^{-1}$. Therefore $M_{k,t} \le \underline{\nu}_{k,t-1} \le \nu_k$, which means that $\beta(M_{k,t}/\nu_k) = M_{k,t}/\nu_k$. Define $\pi_t(i) \in \{1, \dots, K\}$ such that $\underline{\nu}_{\pi_t(i),t-1} \le \underline{\nu}_{\pi_t(i+1),t-1}$. Also let

$$A_t = \{k : M_{k,t} = \underline{\nu}_{k,t-1}\},$$
$$A_t^{\le j} = A_t \cap \{\pi_i(t) : 1 \le i \le j\},$$
$$T_k(t) = \sum_{s=1}^{t} \mathbb{1}\{k \in A_t\} \quad \text{and} \quad B_t = \pi_t(\ell+1).$$

Informally, A_t is the set of jobs that are fully allocated at time-step t, $A_t^{\le j}$ is a subset of A_t containing the j jobs believed to be easiest, $T_k(t)$ is the number of times job k has been fully allocated at time-step t, and B_t is the $(\ell+1)$th easiest job at time-step t (this is only defined if $\ell < K$ and will only be used in that case).

Lemma 8. *For all t, $|A_t| \ge \ell$ and if $|A_t| = \ell$, then $M_{B_t,t} \ge S^*$.*

Proof. $|A_t| = \max\left\{j : \sum_{i=1}^{j} \underline{\nu}_{\pi_t(i),t-1} \le 1\right\}$. But $\underline{\nu}_{k,t-1} \le \nu_k$ for all k and t, so $\sum_{i=1}^{\ell} \underline{\nu}_{\pi_t(i),t-1} \le \sum_{k=1}^{\ell} \underline{\nu}_{k,t-1} \le \sum_{k=1}^{\ell} \nu_k \le 1$. Therefore $|A_t| \ge \ell$. If $|A_t| = \ell$, then $B_t \notin A_t$ is the overflow job and so $M_{B_t,t} = 1 - \sum_{k \in A_t} \underline{\nu}_{k,t-1} \ge 1 - \sum_{k \in A^*} \underline{\nu}_{k,t-1} \ge 1 - \sum_{k \in A^*} \nu_k \equiv S^*$ □

We now decompose the regret, while still assuming that G holds:

$$\sum_{t=1}^{n} r_t = \sum_{t=1}^{n} \left(\ell + \frac{S^*}{\nu_{\ell+1}} - \sum_{k=1}^{K} \frac{M_{k,t}}{\nu_k}\right)$$
$$\le \sum_{t=1}^{n} \sum_{k \in A_t^{\le \ell}} \left(1 - \frac{M_{k,t}}{\nu_k}\right) \tag{9}$$
$$+ \mathbb{1}\{\ell < K\} \sum_{t=1}^{n} \left(\frac{S^*}{\nu_{\ell+1}} - \frac{M_{B_t,t}}{\nu_{B_t}}\right). \tag{10}$$

Let us bound the first sum:

$$\sum_{t=1}^{n} \sum_{k \in A_t^{\le \ell}} \left(1 - \frac{M_{k,t}}{\nu_k}\right)$$
$$= \sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \left(1 - \frac{\underline{\nu}_{k,t-1}}{\nu_k}\right)$$
$$\overset{(a)}{\le} \sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \underline{\nu}_{k,t-1} \varepsilon_{k,t-1}$$
$$\overset{(b)}{\le} \sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \frac{c_{k,1} \underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0} T_k(t)}, \tag{11}$$

where (a) follows by Lemma 6 and (b) by Theorem 4.

Lemma 9. *If $k > j$, then*

$$\sum_{t=1}^{n} \mathbb{1}\left\{k \in A_t^{\le j}\right\} \le \frac{c_{k,1}}{\underline{\nu}_{k,0} \Delta_{j,k}} =: u_{j,k}.$$

Proof. Assume $k \in A_t^{\le j}$. Therefore $\underline{\nu}_{k,t-1} \le \nu_j$. But if $u_{j,k} < \sum_{s=1}^{t} \mathbb{1}\left\{k \in A_s^{\le j}\right\} \le T_k(t-1) + 1$, then

$$\frac{1}{\underline{\nu}_{k,t-1}} \le \frac{1}{\nu_k} + \varepsilon_{k,t-1} = \frac{1}{\nu_j} + \varepsilon_{k,t-1} - \Delta_{j,k}$$
$$\overset{(a)}{\le} \frac{1}{\nu_j} + \frac{c_{k,1}}{\underline{\nu}_{k,0}(T_k(t-1)+1)} - \Delta_{j,k} < \frac{1}{\nu_j},$$

where (a) follows from Theorem 4. Therefore $k \in A_t^{\le j}$ implies that $\sum_{s=1}^{t} \mathbb{1}\left\{k \in A_s^{\le j}\right\} \le u_{j,k}$ and so $\sum_{t=1}^{n} \mathbb{1}\left\{k \in A_t^{\le j}\right\} \le u_{j,k}$ as required. □

Continuing (11) by applying Lemma 9 with $j = \ell$:

$$\sum_{t=1}^{n} \sum_{k=1}^{K} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \frac{c_{k,1} \underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0} T_k(t)}$$
$$= \sum_{t=1}^{n} \sum_{k \in A^*} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \frac{c_{k,1} \underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0} T_k(t)}$$
$$+ \sum_{t=1}^{n} \sum_{k \notin A^*} \mathbb{1}\left\{k \in A_t^{\le \ell}\right\} \frac{c_{k,1} \underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0} T_k(t)} \tag{12}$$
$$\overset{(a)}{\le} \sum_{k \in A^*} \sum_{t=1}^{n} \frac{c_{k,1} \eta_k}{t} + \sum_{k \notin A^*} \sum_{t=1}^{u_{\ell,k}} \frac{c_{k,1} \eta_k}{t}$$
$$\le \sum_{k=1}^{\ell} c_{k,1} \eta_k (1 + \log n) + \sum_{k=\ell+1}^{K} c_{k,1} \eta_k (1 + \log u_{\ell,k}),$$

where (a) follows by Lemma 9 and the fact that $k \in A_t^{\le \ell}$ implies that $\frac{\underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0}} \le \eta_k$. Now if $\ell = K$, then the second term in (9) is zero and the proof is completed by substituting the above result into (9) and then into (8). So now we assume $\ell > K$ and bound the second term in (9) as follows:

$$\sum_{t=1}^{n} \left(\frac{S^*}{\nu_{\ell+1}} - \frac{M_{B_t,t}}{\nu_{B_t}}\right) \le \sum_{t=1}^{n} \mathbb{1}\{B_t \in A_t\} \left(1 - \frac{\underline{\nu}_{B_t,t-1}}{\nu_{B_t}}\right)$$
$$+ \sum_{t=1}^{n} \mathbb{1}\{B_t \notin A_t\} \left(\frac{S^*}{\nu_{\ell+1}} - \frac{S^*}{\nu_{B_t}}\right), \tag{13}$$

where we used Lemma 8 and $S^* \le 1$ and that if $B_t \in A_t$,

then $M_{B_t,t} = \underline{\nu}_{B_t,t-1}$. Bounding each term separately:

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{1}\{B_t \in A_t\} \left(1 - \frac{\underline{\nu}_{B_t,t-1}}{\nu_{B_t}}\right) \\
&\overset{(a)}{\leq} \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbb{1}\left\{k \in A_t^{\leq \ell+1}\right\} \left(1 - \frac{\underline{\nu}_{k,t-1}}{\nu_k}\right) \\
&\overset{(b)}{\leq} \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbb{1}\left\{k \in A_t^{\leq \ell+1}\right\} \underline{\nu}_{k,t-1}\varepsilon_{k,t-1} \qquad (14) \\
&\overset{(c)}{\leq} \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbb{1}\left\{k \in A_t^{\leq \ell+1}\right\} \frac{c_{k,1}\underline{\nu}_{k,t-1}}{\underline{\nu}_{k,0} T_k(t)} \\
&\overset{(d)}{\leq} \sum_{k=1}^{\ell+1} c_{k,1}\eta_k(1+\log n) + \sum_{k=\ell+2}^{K} c_{k,1}\eta_k(1+\log u_{\ell+1,k}),
\end{aligned}$$

where (a) follows since $B_t \in A_t$ implies that $B_t \in A_t^{\leq \ell+1}$, (b) follows from Lemma 6(4), (c) by Theorem 4, and (d) follows from Lemma 9 and the same analysis as (12). For the second term we need the following lemma, which uses Theorem 5 and a reasoning analogues to that of Auer et al. [2002] to bound the regret of the UCB algorithm for stochastic bandits:

Lemma 10. *Let* $U_k(t) = \sum_{s=1}^{t} \mathbb{1}\{M_{k,s} \geq S^*\}$ *and* $k > \ell + 1$. *If* $U_k(t) \geq \frac{c_{k,2}}{S^* \underline{\nu}_{k,0} \Delta_{\ell+1,k}^2} =: v_k$, *then* $k \neq B_t$.

Proof. If $\underline{\nu}_{k,t-1} > \nu_{\ell+1}$, then $k \neq B_t$. Furthermore, if $U_k(t) > v_k$, then

$$\begin{aligned}
\frac{1}{\underline{\nu}_{k,t-1}} &\leq \frac{1}{\nu_k} + \varepsilon_{k,t-1} = \frac{1}{\nu_{\ell+1}} - \Delta_{\ell+1,k} + \varepsilon_{k,t-1} \\
&\overset{(a)}{\leq} \frac{1}{\nu_{\ell+1}} - \Delta_{\ell+1,k} + \sqrt{\frac{c_{k,2}}{\underline{\nu}_{k,0} S^* U_k(t)}} < \frac{1}{\nu_{\ell+1}},
\end{aligned}$$

where (a) follows from Theorem 5. □

Therefore

$$\begin{aligned}
&\sum_{t=1}^{n} \mathbb{1}\{B_t \notin A_t\} \left(\frac{S^*}{\nu_{\ell+1}} - \frac{S^*}{\nu_{B_t}}\right) \\
&\overset{(a)}{\leq} S^* \sum_{k=1}^{K}\sum_{t=1}^{n} \mathbb{1}\{k = B_t \notin A_t\} \Delta_{\ell+1,k} \\
&\overset{(b)}{\leq} S^* \sum_{k=\ell+2}^{K}\sum_{t=1}^{n} \mathbb{1}\{k = B_t \notin A_t\} \Delta_{\ell+1,k} \\
&\overset{(c)}{\leq} S^* \sum_{k=\ell+2}^{K}\sum_{t=1}^{n} \mathbb{1}\{k = B_t \wedge M_{k,t} \geq S^*\} \Delta_{\ell+1,k} \\
&\overset{(d)}{\leq} \sum_{k=\ell+2}^{K} S^* \Delta_{\ell+1,k} v_k \overset{(e)}{=} \sum_{k=\ell+2}^{K} \frac{c_{k,2}}{\underline{\nu}_{k,0}\Delta_{\ell+1,k}}, \qquad (15)
\end{aligned}$$

where (a) follows from the definition of $\Delta_{\ell+1,k}$ and the fact that if $B_t \notin A_t$, then $|A_t| = \ell$, (b) follows since $\Delta_{\ell+1,k}$ is negative for $k \leq \ell + 1$, (c) by Lemma 8, (d) by Lemma 10, and (e) by the definition of v_k. Substituting (14) and (15) into (13) we have

$$\begin{aligned}
&\sum_{t=1}^{n} \left(\frac{S^*}{\nu_{\ell+1}} - \frac{M_{B_t,t}}{\nu_{B_t}}\right) \leq \sum_{k=1}^{\ell+1} c_{k,1}\eta_k(1+\log n) \\
&+ \sum_{k=\ell+2}^{K} c_{k,1}\eta_k(1+\log u_{\ell+1,k}) + \sum_{k=\ell+2}^{K} \frac{c_{k,2}}{\underline{\nu}_{k,0}\Delta_{\ell+1,k}}.
\end{aligned}$$

We then substitute this along with (12) into (9) and then (8) to obtain

$$\begin{aligned}
R_n &\leq 1 + \sum_{k=1}^{\ell} c_{k,1}\eta_k(1+\log n) \\
&+ \mathbb{1}\{\ell < K\}\Bigg[\sum_{k=\ell+2}^{K} \frac{c_{k,2}}{\underline{\nu}_{k,0}\Delta_{\ell+1,k}} + \sum_{k=1}^{\ell+1} c_{k,1}\eta_k(1+\log n) \\
&+ \sum_{k=\ell+2}^{K} c_{k,1}\eta_k(1+\log u_{\ell+1,k}) + \sum_{k=\ell+1}^{K} c_{k,1}\eta_k(1+\log u_{\ell,k})\Bigg].
\end{aligned}$$

7 INITIALISATION

Previously we assumed a known lower bound $\underline{\nu}_{k,0} \leq \nu_k$ for each k. In this section we show that these bounds are easily obtained using a halving trick. In particular, the following algorithm computes a lower bound $\underline{\nu}_0 \leq \nu$ for a single job with unknown parameter ν.

Algorithm 2 Initialisation of $\underline{\nu}_0$

```
for t ∈ 1, ..., ∞ do
    Allocate M_t = 2^{-t} and observe X_t
    if X_t = 0 then return ν_0 ← 2^{-t}.
end for
```

A naive way to eliminate the need for the lower bounds $(\underline{\nu}_{k,0})_k$ is simply to run Algorithm 2 for each job sequentially. Then the following proposition shows that $\eta \in O(1)$ is reasonable, which justifies the claim made in (1) that the η_k terms appearing in Theorem 2 are $O(1)$.

Proposition 11. *If* $\eta = \frac{\min\{1,\nu\}}{\underline{\nu}_0}$, *then* $\mathbb{E}\eta \leq 4$.

Proof. Let p_t be the probability that the algorithm ends after time-step t, which is

$$p_t = (1 - \beta(2^{-t}/\nu)) \prod_{s=1}^{t-1} \beta(2^{-s}/\nu).$$

Therefore

$$\begin{aligned}\mathbb{E}\eta &= \mathbb{E}\left[\frac{\min\{1,\nu\}}{\nu_0}\right] = \sum_{t=1}^{\infty} p_t \cdot \frac{\min\{1,\nu\}}{M_t} \\ &= \min\{1,\nu\}\sum_{t=1}^{\infty} 2^t(1-\beta(2^{-t}/\nu))\prod_{s=1}^{t-1}\beta(2^{-s}/\nu) \\ &\leq 4,\end{aligned}$$

where the final inequality follows from an arduous, but straight-forward, computation. □

The problem with the naive method is that the expected running time of Algorithm 2 is $O(\log \frac{1}{\nu})$, which may be arbitrary large for small ν and lead to a high regret *during the initialisation period*. Fortunately, the situation when ν is small is easy to handle, since the amount of resources required to complete such a job is also small. The trick is to run K offset instances of Algorithm 2 alongside a modified version of Algorithm 1. First we describe the parallel implementations of Algorithm 2. For job k, start Algorithm 2 in time-step k, which means that the total amount of resources used by the parallel copies of Algorithm 2 in time-step t is bounded by

$$\sum_{k=1}^{K} \mathbb{1}\{t \geq k\}\, 2^{k-t-1} \leq \min\left\{1, 2^{K-t}\right\}. \quad (16)$$

Job	$M_{k,1}$	$M_{k,2}$	$M_{k,3}$	$M_{k,4}$
1	1/2	1/4	1/8	1/16
2	0	1/2	1/4	1/8
3	0	0	1/2	1/4
$\sum_{k=1}^{K} M_{k,t}$	1/2	3/4	7/8	7/16

Algorithm 1 is implemented starting from time-step 1, but only allocates resources to jobs for which the initialisation process has completed. Estimates are computed using only the samples for which Algorithm 1 chose the allocation, which ensures that they are based on allocations with $M_{k,t} \leq \nu_k$. Note that the analysis of the modified algorithm does not depend on the order in which the parallel processes are initialised. The regret incurred by the modified algorithm is given in order notation in (1). The proof is omitted, but relies on two observations. First, that the expected number of time-steps that a job is not (at least) fully allocated while it is being initialised is 2. The second is that the resources available to Algorithm 1 at time-step t converges exponentially fast to 1 by (16).

8 MINIMAX LOWER BOUNDS

Despite the continuous action space, the techniques used when proving minimax lower bounds for standard stochastic bandits [Auer et al., 1995] can be adapted to our setting.

Theorem 12. *Given fixed n and $8n \geq K \geq 2$ and an arbitrary algorithm, there exists an allocation problem for which the expected regret satisfies $R_n \geq \frac{\sqrt{nK}}{16\sqrt{2}}$.*

Proof. Let $1 \geq \varepsilon > 0$ be a constant to be chosen later. We consider a set of K allocation problems where in problem k, $\nu_j = 2$ for all $j \neq k$ and $\nu_k = \frac{2}{1+\varepsilon}$. The optimal action in problem k is to assign all available resources to job k when the expected reward is $\frac{1+\varepsilon}{2}$. The interaction between the algorithm and a problem k defines a measure $\mathbb{P}_k$ on the set of outcomes (successes, allocations). We write $\mathbb{E}_k$ for expectations with respect to measure $\mathbb{P}_k$. We have

$$\mathbb{E}_k\left[\sum_{t=1}^{n} M_{k,t}\right] - \mathbb{E}_0\left[\sum_{t=1}^{n} M_{k,t}\right] \leq n\sqrt{\frac{1}{2}\,\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_k)}\,, \quad (17)$$

where $\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_k)$ is the Kullback-Leibler divergence (or relative entropy) between $\mathbb{P}_0$ and $\mathbb{P}_k$. The divergence is bounded by

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_0, \mathbb{P}_k) &\overset{(a)}{\leq} \mathbb{E}_0\left[\sum_{t=1}^{n} \frac{\varepsilon^2 M_{k,t}^2}{4}\left(\frac{1}{\frac{M_{k,t}}{2}} + \frac{1}{1-\frac{M_{k,t}}{2}}\right)\right] \\ &\overset{(b)}{=} \mathbb{E}_0\left[\sum_{t=1}^{n} \frac{\varepsilon^2 M_{k,t}}{2 - M_{k,t}}\right] \overset{(c)}{\leq} \varepsilon^2 \mathbb{E}_0\left[\sum_{t=1}^{n} M_{k,t}\right], \quad (18)\end{aligned}$$

where (a) follows from the telescoping property of the KL divergence and by bounding the KL divergence by the χ-squared distance, (b) is trivial and (c) follows since $M_{k,t} \leq 1$. The n-step expected regret given environment k is

$$\begin{aligned}R_n(k) &= \frac{n(1+\varepsilon)}{2} - \mathbb{E}_k \sum_{t=1}^{n}\sum_{j=1}^{K} \frac{M_{j,t}}{\nu_j} \\ &\overset{(b)}{\geq} \frac{\varepsilon}{2}\left(n - \mathbb{E}_k \sum_{t=1}^{n} M_{k,t}\right) \quad (19)\end{aligned}$$

where (b) follows by recalling that $\nu_j = 2$ unless $j = k$, when $\nu_j = 2/(1+\varepsilon)$. Therefore,

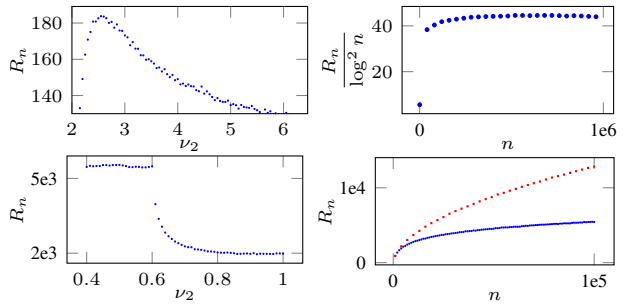
$$\begin{aligned}\sup_k R_n(k) &\overset{(a)}{\geq} \frac{1}{K}\sum_{k=1}^{K} R_n(k) \\ &\overset{(b)}{\geq} \frac{1}{K}\sum_{k=1}^{K}\frac{\varepsilon}{2}\left(n - \mathbb{E}_k\sum_{t=1}^{n} M_{k,t}\right) \\ &\overset{(c)}{\geq} \frac{1}{K}\sum_{k=1}^{K}\frac{\varepsilon}{2}\left(n - \mathbb{E}_0\sum_{t=1}^{n} M_{k,t} - n\varepsilon\sqrt{\frac{1}{2}\mathbb{E}_0\sum_{t=1}^{n} M_{k,t}}\right) \\ &\overset{(d)}{\geq} \frac{\varepsilon}{2K}\left(nK - n - n\varepsilon\sum_{k=1}^{K}\sqrt{\frac{1}{2}\mathbb{E}_0\sum_{t=1}^{n} M_{k,t}}\right) \\ &\overset{(e)}{\geq} \frac{\varepsilon}{2K}\left(nK - n - n\varepsilon\sqrt{\frac{K}{2}\sum_{k=1}^{K}\mathbb{E}_0\sum_{t=1}^{n} M_{k,t}}\right) \\ &\overset{(f)}{\geq} \frac{\varepsilon}{2K}\left(nK - n - n\varepsilon\sqrt{\frac{nK}{2}}\right) \overset{(g)}{\geq} \frac{\varepsilon n}{4} - \frac{\varepsilon^2 n^{\frac{3}{2}}}{2\sqrt{2}K^{\frac{1}{2}}},\end{aligned}$$

where (a) follows since the max is greater than the average, (b) follows from (19), (c) is obtained by combining (17)

and (18), (d) follows from the fact that $\sum_{k=1}^{K} M_{k,t} \leq 1$, (e) is true by Jensen's inequality and (f/g) are straightforward. Choosing $\varepsilon = \sqrt{K/(8n)}$ leads to $\sup_k R_n(k) \geq \sqrt{nK}/(16\sqrt{2})$ as required. □

9 EXPERIMENTS

Data points were generated using the modified algorithm described in Section 7 and by taking the mean of 300 samples. With this many samples the standard error is relatively low (and omitted for readability). We should note that the variance in the regret of the modified algorithm is reasonably large, because the regret depends linearly on the random η_k. For known lower bounds the variance is extremely low. To illustrate the behaviour of the algorithm we performed four experiments on synthetic data with $K = 2$, which are plotted below as TL (top left), TR, BL, BR (bottom right) respectively. In TL we fixed $n = 10^4$, $\nu_1 = 2$ and plotted the regret as a function of $\nu_2 \in [2, 10]$. The experiment shows the usual bandit-like dependence on the gap $1/\Delta_{1,2}$. In TR we fixed $\nu_1 = {}^4\!/_{10}$, $\nu_2 = {}^6\!/_{10}$ and plotted $R_n/\log^2 n$ as a function of n. The experiment lies within case 2 described in Section 4 and shows that the algorithm suffers regret $R_n \approx 45 \log^2 n$ as predicted by Theorem 2. In BL we fixed $n = 10^5$, $\nu_1 = {}^4\!/_{10}$ and plotted the regret as a function of $\nu_2 \in [{}^4\!/_{10}, 1]$. The results show the algorithm suffering $O(\log^2 n)$ regret for both processes until the critical point when $\nu_2 > {}^6\!/_{10}$ when the second process can no longer be fully allocated, which is quickly learned and the algorithm suffers $O(\log^2 n)$ regret for only one process. In BR we fixed $\nu_1 = {}^4\!/_{10}$ and $\nu_2 = {}^6\!/_{10}$ and plotted the regret as a function of n for two algorithms. The first algorithm (solid blue) is the modified version of Algorithm 1 as described in Section 7. The second (dotted red) is the same, but uses the unweighted estimator $w_{k,t} = 1$ for all k and t. The result shows that both algorithms suffer sub-linear regret, but that the weighted estimator is a significant improvement over the unweighted one.



10 CONCLUSIONS

We introduced the linear stochastic resource allocation problem and a new optimistic algorithm for this setting. Our main result shows that the new algorithm enjoys a (squared) logarithmic problem-dependent regret. We also presented a minimax lower bound of $\Omega(\sqrt{nK})$, which is consistent with the problem-dependent upper bound. The simulations confirm the theory and highlight the practical behaviour of the new algorithm. There are many open questions and possibilities for future research. Most important is whether the $\log^2 n$ can be reduced to $\log n$. Problem-dependent lower bounds would be interesting. The algorithm is not anytime (although a doubling trick presumably works in theory). Developing and analysing algorithms when the horizon it not known, and have high-probability bounds are both of interest. We also wonder if Thompson sampling can be efficiently implemented for some reasonable prior, and if it enjoys the same practical and theoretical guarantees in this domain as it does for bandits. Other interesting extensions are when resources are not replenished, or the state of the jobs follow a Markov process. Finally, we want to emphasise that we have made just the first steps towards developing this new and interesting setting. We hope to see significant activity extending and modifying the model/algorithm for specific problems.

Acknowledgements This work was supported by the Alberta Innovates Technology Futures, NSERC, by EU Framework 7 Project No. 248828 (ADVANCE), and by Israeli Science Foundation grant ISF- 1567/10. Part of this work was done while Csaba Szepesvári was visiting Technion.

A TECHNICAL INEQUALITIES

The proof of the following theorem is given in the supplementary material.

Theorem 13. *Let $\delta \in (0,1)$ and $X_1, \ldots, X_n$ be a sequence of random variables adapted to filtration $\{\mathcal{F}_t\}$ with $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Let R_t be $\mathcal{F}_{t-1}$-measurable such that $|X_t| \leq R_t$ almost surely, $R = \max_{t \leq n} R_t$. Define $S = \sum_{t=1}^{n} X_t$, $V^2 = \sum_{t=1}^{n} \mathrm{Var}[X_t|\mathcal{F}_{t-1}]$, and*

$$\delta_{r,v} = \frac{\delta}{3(r+1)^2(v+1)^2}\,,$$
$$f(r,v) = \frac{r+1}{3}\log\frac{2}{\delta_{r,v}} + \sqrt{2(v+1)\log\frac{2}{\delta_{r,v}} + \left(\frac{r+1}{3}\right)^2 \log^2\frac{2}{\delta_{r,v}}}\,.$$

Then $\mathbb{P}\left\{|S| \geq f(R, V^2)\right\} \leq \delta$.

Proof of Theorem 13. Note that $f(r,v)$ is strictly monotone increasing in both r and v. We now use a peeling

argument. We have,

$$
\begin{aligned}
&\mathbb{P}\left\{|S_n| \geq f(R, V^2)\right\} \\
&\overset{(a)}{=} \sum_{r=1}^{\infty}\sum_{v=1}^{\infty} \mathbb{P}\left\{|S_n| \geq f(R, V^2), \lceil V^2 \rceil = v, \lceil R \rceil = r\right\} \\
&\overset{(b)}{\leq} \sum_{r=1}^{\infty}\sum_{v=1}^{\infty} \mathbb{P}\left\{|S_n| \geq f(r-1, v-1), \lceil V^2 \rceil = v, \lceil R \rceil = r\right\} \\
&\overset{(c)}{\leq} \sum_{r=1}^{\infty}\sum_{v=1}^{\infty} 2\exp\left(-\frac{f(r-1,v-1)^2}{2v + \frac{2rf(r-1,v-1)}{3}}\right) \\
&\overset{(d)}{\leq} \sum_{r=1}^{\infty}\sum_{v=1}^{\infty} \delta_{r-1,v-1} \overset{(e)}{=} \frac{\delta}{3}\sum_{r=1}^{\infty}\sum_{v=1}^{\infty}\frac{1}{v^2 r^2} \overset{(f)}{\leq} \delta\,,
\end{aligned}
$$

where (a) follows from the positivity of R and V, (b) by the monotonicity of f, (c) by Theorem 14 stated below (a martingale version Bernstein's inequality), (d) by Lemma 15, (e) by the definition of $\delta_{r,v}$, (f) is trivial. □

Theorem 14 (Theorem 3.15 of McDiarmid 1998, see also Freedman 1975 and Bernstein 1946)**.** *Let $X_1, \ldots, X_n$ be a sequence of random variables adapted to the filtration $\{\mathcal{F}_t\}$ with $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. Further, let R_t be $\mathcal{F}_{t-1}$-measurable such that $X_t \leq R_t$ almost surely, $R = \max_{t\leq n} R_t$ and $V^2 = \sum_{t=1}^{n} \mathrm{Var}[X_t|\mathcal{F}_{t-1}]$. Then for any $\varepsilon, r, v > 0$,*

$$
\mathbb{P}\left\{\sum_{t=1}^{n} X_t \geq \varepsilon, V^2 \leq v, R \leq r\right\} \leq \exp\left(-\frac{\varepsilon^2}{2v + \frac{2\varepsilon r}{3}}\right).
$$

We note that although this inequality is usually stated for deterministic R_t, the extension is trivial: Just define $Y_t = X_t\,\mathbb{1}\{R_t \leq r\}$ and apply the standard inequality to Y_t. The result then follows since on $R \leq r$, $Y_t = X_t$ for all t and thus $\sum_{t=1}^{n} X_t = \sum_{t=1}^{n} Y_t$.

Lemma 15. *If $\varepsilon \geq \frac{r}{3}\log\frac{2}{\delta} + \sqrt{2v\log\frac{2}{\delta} + \frac{r^2}{9}\log^2\frac{2}{\delta}}$, then $2\exp\left(-\frac{\varepsilon^2}{2v+\frac{2\varepsilon r}{3}}\right) \leq \delta$.*

B TABLE OF NOTATION

K	number of jobs
n	time horizon
ν_k	parameter characterising difficulty of job k
$\beta(p)$	function $\beta(p) := \min\{1, p\}$
$M_{k,t}$	resources assigned to job k in time-step t
$X_{k,t}$	outcome of job k in time-step t
$\underline{\nu}_{k,t}$	lower bound on ν_k at time-step t
$\bar{\nu}_{k,t}$	upper bound on ν_k at time-step t
δ	bound on probability that some confidence intervals fails
$\pi_t(i)$	ith easiest job at time-step t sorted by $\underline{\nu}_{k,t-1}$
ℓ	number of fully allocated jobs under optimal allocation
S^*	optimal amount of resources assigned to overflow process
A^*	contains the ℓ easiest jobs (sorted by ν_k)
A_t	set of jobs with $M_{k,t} = \underline{\nu}_{k,t-1}$ at time-step t
B_t	equal to $\pi_t(\ell+1)$
η_k	$\frac{\min\{1,\nu_k\}}{\underline{\nu}_{k,0}}$

References

Rajeev Agrawal, Demosthenis Teneketzis, and Venkatachalam Anantharam. Asymptotically efficient adaptive allocation schemes for controlled i.i.d. processes: Finite parameter space. *IEEE Transaction on Automatic Control*, 34:258–267, 1989.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Foundations of Computer Science, 1995. Proceedings., 36th Annual Symposium on*, pages 322–331. IEEE, 1995.

Peter Auer, Nicoló Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *FOCS*, pages 207–216, 2013.

Gábor Bartók. A near-optimal algorithm for finite partial-monitoring games against adversarial opponents. In *COLT*, pages 696–710, 2013.

Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In *COLT 2011*, pages 133–154, 2011.

Sergei Bernstein. *The Theory of Probabilities (Russian)*. Moscow, 1946.

Ramazan Bitirgen, Engin Ipek, and Jose F Martinez. Coordinated management of multiple interacting resources in

chip multiprocessors: A machine learning approach. In *Proceedings of the 41st annual IEEE/ACM International Symposium on Microarchitecture*, pages 318–329. IEEE Computer Society, 2008.

Josef Broder and Paat Rusmevichientong. Dynamic pricing under a general parametric choice model. *Operations Research*, 60(4):965–980, 2012.

Sébastien Bubeck and Nicolò Cesa-Bianchi. *Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems*. Foundations and Trends in Machine Learning. Now Publishers Incorporated, 2012. ISBN 9781601986269.

Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. Multi-armed bandit with budget constraint and variable costs. In *AAAI*, 2013.

Sarah Filippi, Olivier Cappé, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In *NIPS*, pages 586–594, December 2010.

Dean P. Foster and Alexander Rakhlin. No internal regret via neighborhood watch. *Journal of Machine Learning Research - Proceedings Track (AISTATS)*, 22:382–390, 2012.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 02 1975.

András György, Levente Kocsis, Ivett Szabó, and Csaba Szepesvári. Continuous time associative bandit problems. In *IJCAI-07*, pages 830–835, 2007.

Engin Ipek, Onur Mutlu, José F. Martínez, and Rich Caruana. Self-optimizing memory controllers: A reinforcement learning approach. *SIGARCH Comput. Archit. News*, 36(3):39–50, June 2008. ISSN 0163-5964.

Chun Liu, Anand Sivasubramaniam, and Mahmut Kandemir. Organizing the last line of defense before hitting the memory wall for cmps. In *Software, IEE Proceedings-*, pages 176–185. IEEE, 2004.

Colin McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In *NIPS*, pages 2256–2264, 2013.

Aldo Rustichini. Minimizing regret: The general case. *Games and Economic Behavior*, 29(1–2):224–243, 1999.

G Edward Suh, Srinivas Devadas, and Larry Rudolph. A new memory monitoring scheme for memory-aware scheduling and partitioning. In *High-Performance Computer Architecture, 2002. Proceedings. Eighth International Symposium on*, pages 117–128. IEEE, 2002.

Long Tran-Thanh, Archie C. Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI*, 2012.