# Linear Multi-Resource Allocation with Semi-Bandit Feedback

**Tor Lattimore**
Department of Computing Science
University of Alberta, Canada
tor.lattimore@gmail.com

**Koby Crammer**
Department of Electrical Engineering
The Technion, Israel
koby@ee.technion.ac.il

**Csaba Szepesvári**
Department of Computing Science
University of Alberta, Canada
szepesva@ualberta.ca

## Abstract

We study an idealised sequential resource allocation problem. In each time step the learner chooses an allocation of several resource types between a number of tasks. Assigning more resources to a task increases the probability that it is completed. The problem is challenging because the alignment of the tasks to the resource types is unknown and the feedback is noisy. Our main contribution is the new setting and an algorithm with nearly-optimal regret analysis. Along the way we draw connections to the problem of minimising regret for stochastic linear bandits with heteroscedastic noise. We also present some new results for stochastic linear bandits on the hypercube that significantly improve on existing work, especially in the sparse case.

## 1 Introduction

Economist Thomas Sowell remarked that "The first lesson of economics is scarcity: There is never enough of anything to fully satisfy all those who want it."[1] The optimal allocation of resources is an enduring problem in economics, operations research and daily life. The problem is challenging not only because you are compelled to make difficult trade-offs, but also because the (expected) outcome of a particular allocation may be unknown and the feedback noisy.

We focus on an idealised resource allocation problem where the economist plays a repeated resource allocation game with multiple resource types and multiple tasks to which these resources can be assigned. Specifically, we consider a (nearly) linear model with $D$ resources and $K$ tasks. In each time step $t$ the economist chooses an allocation of resources $M_t \in \mathbb{R}^{D \times K}$ where $M_{tk} \in \mathbb{R}^D$ is the $k$th column and represents the amount of each resource type assigned to the $k$th task. We assume that the $k$th task is completed successfully with probability $\min\{1, \langle M_{tk}, \nu_k \rangle\}$ and $\nu_k \in \mathbb{R}^D$ is an unknown non-negative vector that determines how the success rate of a given task depends on the quantity and type of resources assigned to it. Naturally we will limit the availability of resources by demanding that $M_t$ satisfies $\sum_{k=1}^{K} M_{tdk} \leq 1$ for all resource types $d$. At the end of each time step the economist observes which tasks were successful. The objective is to maximise the number of successful tasks up to some time horizon $n$ that is known in advance. This model is a natural generalisation of the one used by Lattimore et al. [2014a], where it was assumed that there was a single resource type only.

[1] He went on to add that "The first lesson of politics is to disregard the first lesson of economics." Sowell [1993]

An example application might be the problem of allocating computing resources on a server between a number of Virtual Private Servers (VPS). In each time step (some fixed interval) the controller chooses how much memory/cpu/bandwidth to allocate to each VPS. A VPS is said to fail in a given round if it fails to respond to requests in a timely fashion. The requirements of each VPS are unknown in advance, but do not change greatly with time. The controller should learn which VPS benefit the most from which resource types and allocate accordingly.

The main contribution of this paper besides the new setting is an algorithm designed for this problem along with theoretical guarantees on its performance in terms of the regret. Along the way we present some additional results for the related problem of minimising regret for stochastic linear bandits on the hypercube. We also prove new concentration results for weighted least squares estimation, which may be independently interesting.

The generalisation of the work of Lattimore et al. [2014a] to multiple resources turns out to be fairly non-trivial. Those with knowledge of the theory of stochastic linear bandits will recognise some similarity. In particular, once the nonlinearity of the objective is removed, the problem is equivalent to playing $K$ linear bandits in parallel, but where the limited resources constrain the actions of the learner and correspondingly the returns for each task. Stochastic linear bandits have recently been generating a significant body of research (e.g., Auer [2003], Dani et al. [2008], Rusmevichientong and Tsitsiklis [2010], Abbasi-Yadkori et al. [2011, 2012], Agrawal and Goyal [2012] and many others). A related problem is that of online combinatorial optimisation. This has an extensive literature, but most results are only applicable for discrete action sets, are in the adversarial setting, and cannot exploit the additional structure of our problem. Nevertheless, we refer the interested reader to (say) the recent work by Kveton et al. [2014] and references there-in. Also worth mentioning is that the resource allocation problem at hand is quite different to the "linear semi-bandit" proposed and analysed by Krishnamurthy et al. [2015] where the action set is also finite (the setting is different in many other ways besides).

Given its similarity, it is tempting to apply the techniques of linear bandits to our problem. When doing so, two main difficulties arise. The first is that our payoffs are non-linear: the expected reward is a linear function only up to a point after which it is clipped. In the resource allocation problem this has a natural interpretation, which is that over-allocating resources beyond a certain point is fruitless. Fortunately, one can avoid this difficulty rather easily by ensuring that with high probability resources are never over-allocated. The second problem concerns achieving good regret regardless of the task specifics. In particular, when the number of tasks $K$ is large and resources are at a premium the allocation problem behaves more like a $K$-armed bandit where the economist must choose the few tasks that can be completed successfully. For this kind of problem regret should scale in the worst case with $\sqrt{K}$ only [Auer et al., 2002, Bubeck and Cesa-Bianchi, 2012]. The standard linear bandits approach, on the other hand, would lead to a bound on the regret that depends linearly on $K$. To remedy this situation, we will exploit that if $K$ is large and resources are scarce, then many tasks will necessarily be under-resourced and will fail with high probability. Since the noise model is Bernoulli, the variance of the noise for these tasks is extremely low. By using weighted least-squares estimators we are able to exploit this and thereby obtain an improved regret. An added benefit is that when resources are plentiful, then all tasks will succeed with high probability under the optimal allocation, and in this case the variance is also low. This leads to a poly-logarithmic regret for the resource-laden case where the optimal allocation fully allocates every task.

## 2 Preliminaries

If $F$ is some event, then $\neg F$ is its complement (i.e., it is the event that $F$ does not occur). If $A$ is positive definite and $x$ is a vector, then $\|x\|_A^2 = x^\top A x$ stands for the weighted 2-norm. We write $|x|$ to be the vector of element-wise absolute values of $x$. We let $\nu \in \mathbb{R}^{D \times K}$ be a matrix with columns $\nu_1, \ldots \nu_K$. All entries in $\nu$ are non-negative, but otherwise we make no global assumptions on $\nu$. At each time step $t$ the learner chooses an allocation matrix $M_t \in \mathcal{M}$ where

$$\mathcal{M} = \left\{ M \in [0,1]^{D \times K} : \sum_{k=1}^{K} M_{dk} \leq 1 \text{ for all } d \right\} .$$

The assumption that each resource type has a bound of 1 is non-restrictive, since the units of any resource can be changed to accommodate this assumption. We write $M_{tk} \in [0,1]^D$ for the $k$th

column of $M_t$. The reward at time step $t$ is $\|Y_t\|_1$ where $Y_{tk} \in \{0, 1\}$ is sampled from a Bernoulli distribution with parameter $\psi(\langle M_{tk}, \nu_k \rangle) = \min\{1, \langle M_{tk}, \nu_k \rangle\}$. The economist observes all $Y_{tk}$, however, not just the sum. The optimal allocation is denoted by $M^*$ and defined by

$$M^* = \operatorname*{arg\,max}_{M \in \mathcal{M}} \sum_{k=1}^{K} \psi(\langle M_k, \nu_k \rangle) \,.$$

We are primarily concerned with designing an allocation algorithm that minimises the expected (pseudo) regret of this problem, which is defined by

$$R_n = n \sum_{k=1}^{K} \psi(\langle M_k^*, \nu_k \rangle) - \mathbb{E}\left[ \sum_{t=1}^{n} \sum_{k=1}^{K} \psi(\langle M_{tk}, \nu_k \rangle) \right] ,$$

where the expectation is taken over both the actions of the algorithm and the observed reward.

**Optimal Allocations**

If $\nu$ is known, then the optimal allocation can be computed by constructing an appropriate linear program. Somewhat surprisingly it may also be computed exactly in $O(K \log K + D \log D)$ time using Algorithm 1 below. The optimal allocation is not so straight-forward as, e.g., simply allocating resources to the incomplete task for which the corresponding $\nu$ is largest in some dimension. For example, for $K = 2$ tasks and $d = 2$ resource types:

$$\nu = \begin{pmatrix} \nu_1 & \nu_2 \end{pmatrix} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 1 \end{pmatrix} \quad \Longrightarrow \quad M^* = \begin{pmatrix} M_1^* & M_2^* \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1/2 & 1/2 \end{pmatrix} .$$

We see that even though $\nu_{22}$ is the largest parameter, the optimal allocation assigns only half of the second resource ($d = 2$) to this task. The right approach is to allocate resources to incomplete tasks using the ratios as prescribed by Algorithm 1. The intuition for allocating in this way is that resources should be allocated as efficiently as possible, and efficiency is determined by the ratio of the expected success due to the allocation of a resource and the amount of resources allocated.

**Theorem 1.** *Algorithm 1 returns $M^*$.*

**Algorithm 1**

**Input:** $\nu$
$M = 0 \in \mathbb{R}^{D \times K}$ and $B = \mathbb{1} \in \mathbb{R}^D$
**while** $\exists\, k, d$ s.t $\langle M_k, \nu_k \rangle < 1$ and $B_d > 0$ **do**

$$\mathcal{A} = \{k : \langle M_k, \nu_k \rangle < 1\} \text{ and } \mathcal{B} = \{d : B_d > 0\}$$

$$k, d = \operatorname*{arg\,max}_{(k,d) \in \mathcal{A} \times \mathcal{B}} \min_{i \in \mathcal{A} \setminus \{k\}} \left( \frac{\nu_{dk}}{\nu_{di}} \right)$$

$$M_{dk} = \min\left\{ B_d, \frac{1 - \langle M_k, \nu_k \rangle}{\nu_{dk}} \right\}$$

**end while**
**return** $M$

The proof of Theorem 1 can be found in Appendix D.

We are interested primarily in the case when $\nu$ is unknown, so Algorithm 1 will not be directly applicable. Nevertheless, the algorithm is useful as a module in the implementation of a subsequent algorithm that estimates $\nu$ from data.

## 3 Optimistic Allocation Algorithm

We follow the optimism in the face of uncertainty principle. In each time step $t$, the algorithm constructs an estimator $\hat{\nu}_{kt}$ for each $\nu_k$ and a corresponding confidence set $C_{tk}$ for which $\nu_k \in C_{tk}$ holds with high probability. The algorithm then takes the optimistic action subject to the assumption that $\nu_k$ does indeed lie in $C_{tk}$ for all $k$. The main difficulty is the construction of the confidence sets. Like other authors [Dani et al., 2008, Rusmevichientong and Tsitsiklis, 2010, Abbasi-Yadkori et al., 2011] we define our confidence sets to be ellipses, but the use of a weighted least-squares estimator means that our ellipses may be significantly smaller than the sets that would be available by using these previous works in a straightforward way. The algorithm accepts as input the number of tasks and resource types, the horizon and constants $\alpha > 0$ and $\beta$ where constant $\beta$ is defined by

$$\delta = \frac{1}{nK}, \qquad N = \left(4n^4 D^2\right)^D , \qquad B \geq \max_k \|\nu_k\|_2^2 , \qquad \text{so that}$$

$$\beta = \left( 1 + \sqrt{\alpha B} + 2\sqrt{\log\left( \frac{6nN}{\delta} \log\left( \frac{3nN}{\delta} \right) \right)} \right)^2 . \tag{1}$$

Note that $B$ must be a known bound on $\max_k \|\nu_k\|_2^2$, which might seem like a serious restriction, until one realizes that it is easy to add an initialisation phase where estimates are quickly made while incurring minimal additional regret, as was also done by Lattimore et al. [2014a]. The value of $\alpha$ determines the level of regularisation in the least squares estimation and will be tuned later to optimise the regret.

**Algorithm 2** Optimistic Allocation Algorithm

1: **Input** $K$, $D$, $n$, $\alpha$, $\beta$
2: **for** $t \in 1, \dots, n$ **do**
3: **// Compute confidence sets for all tasks $k$:**
4: $G_{tk} = \alpha I + \sum_{\tau<t} \gamma_{\tau k} M_{\tau k} M_{\tau k}^\top$
5: $\hat{\nu}_{tk} = G_{tk}^{-1} \sum_{\tau<t} \gamma_{\tau k} M_\tau Y_{\tau k}$
6: $C_{tk} = \left\{\tilde{\nu}_k : \|\tilde{\nu}_k - \hat{\nu}_{tk}\|_{G_{tk}}^2 \le \beta\right\}$ and $C'_{tk} = \left\{\tilde{\nu}_k : \|\tilde{\nu}_k - \hat{\nu}_{tk}\|_{G_{tk}}^2 \le 4\beta\right\}$
7: **// Compute optimistic allocation:**
8: $M_t = \arg\max_{M_t \in \mathcal{M}} \max_{\tilde{\nu}_k \in C_{tk}} \psi(\langle M_{tk}, \tilde{\nu}_k \rangle)$
9: **// Observe success indicators $Y_{tk}$ for all tasks $k$:**
10: $Y_{tk} \sim \text{Bernoulli}(\psi(\langle M_{tk}, \nu_k \rangle))$
11: **// Compute weights for all tasks $k$:**
12: $\gamma_{tk}^{-1} = \arg\max_{\tilde{\nu}_k \in C'_{tk}} \langle M_{tk}, \tilde{\nu}_k \rangle \left(1 - \langle M_{tk}, \tilde{\nu}_k \rangle\right)$
13: **end for**

### Computational Efficiency

We could not find an efficient implementation of Algorithm 2 because solving the bilinear optimisation problem in Line 8 is likely to be NP-hard (Bennett and Mangasarian [1993] and also Petrik and Zilberstein [2011]). In our experiments we used a simple algorithm based on optimising for $M$ and $\nu$ in alternative steps combined with random restarts, but for large $D$ and $K$ this would likely not be efficient. In Appendix E we present an alternative algorithm that is efficient, but relies on the assumption that $\|\nu_k\|_1 \le 1$ for all $k$. In this regime it is impossible to over-allocate resources and this fact can be exploited to obtain an efficient and practical algorithm with strong guarantees. Along the way, we are able to construct an elegant algorithm for linear bandits on the hypercube that enjoys optimal regret and adapts to sparsity.

Computing the weights $\gamma_{tk}$ (Line 12) is (somewhat surprisingly) straight-forward. Define

$$\bar{p}_{tk} = \langle M_{tk}, \hat{\nu}_{tk} \rangle + 2\sqrt{\beta}\, \|M_{tk}\|_{G_{tk}^{-1}} \quad \text{and} \quad \underline{p}_{tk} = \langle M_{tk}, \hat{\nu}_{tk} \rangle - 2\sqrt{\beta}\, \|M_{tk}\|_{G_{tk}^{-1}} \,.$$

Then the weights can be computed by

$$\gamma_{tk}^{-1} = \begin{cases} \bar{p}_{tk}(1 - \bar{p}_{tk}) & \text{if } \bar{p}_{tk} \le \frac{1}{2} \\ \underline{p}_{tk}(1 - \underline{p}_{tk}) & \text{if } \underline{p}_{tk} \ge \frac{1}{2} \\ \frac{1}{4} & \text{otherwise}\,. \end{cases} \tag{2}$$

A curious reader might wonder why the weights are computed by optimising within confidence set $C'_{tk}$, which has double the radius of $C_{tk}$. The reason is rather technical, but essentially if the true parameter $\nu_k$ were to lie on the boundary of the confidence set, then the corresponding weight could become infinite. For the analysis to work we rely on controlling the size of the weights. It is not clear whether or not this trick is really necessary.

## 4 Worst-case Regret for Algorithm 2

We now analyse the regret of Algorithm 2. First we offer a worst-case bound on the regret that depends on the time-horizon like $O(\sqrt{n})$. We then turn our attention to the resource-laden case where the optimal allocation satisfies $\langle M_k^*, \nu_k \rangle = 1$ for all $k$. In this instance we show that the dependence on the horizon is only poly-logarithmic, which would normally be unexpected when the

action-space is continuous. The improvement comes from the weighted estimation that exploits the fact that the variance of the noise under the optimal allocation vanishes.

**Theorem 2.** *Suppose Algorithm 2 is run with bound $B \geq \max_k \|\nu_k\|_2^2$. Then*

$$R_n \leq 1 + 4D\sqrt{2\beta nK\left(\max_k \|\nu_k\|_\infty + 4\sqrt{\beta/\alpha}\right)\log(1+4n^2)}\,.$$

Choosing $\alpha = B^{-1}\log\left(\frac{6nN}{\delta}\log\left(\frac{3nN}{\delta}\right)\right)$ and assuming that $B \in O(\max_k \|\nu_k\|_2^2)$, then

$$R_n \in O\left(D^{3/2}\sqrt{nK\max_k \|\nu_k\|_2 \log n}\right)\,.$$

The proof of Theorem 2 will follow by carefully analysing the width of the confidence sets as the algorithm makes allocations. We start by proving the validity of the confidence sets, and then prove the theorem.

### Weighted Least Squares Estimation

For this sub-section we focus on the problem of estimating a single unknown $\nu = \nu_k$. Let $M_1, \ldots, M_n$ be a sequence of allocations to task $k$ with $M_t \in \mathbb{R}^D$. Let $\{\mathcal{F}_t\}_{t=0}^n$ be a filtration with $\mathcal{F}_t$ containing information available at the end of round $t$, which means that $M_t$ is $\mathcal{F}_{t-1}$-measurable. Let $\gamma_1, \ldots, \gamma_n$ be the sequence of weights chosen by Algorithm 2. The sequence of outcomes is $Y_1, \ldots, Y_n \in \{0,1\}$ for which $\mathbb{E}[Y_t|\mathcal{F}_{t-1}] = \psi(\langle M_t, \nu\rangle)$. The weighted regularised gram matrix is $G_t = \alpha I + \sum_{\tau<t}\gamma_\tau M_\tau M_\tau^\top$ and the corresponding weighted least squares estimator is

$$\hat{\nu}_t = G_t^{-1}\sum_{\tau<t}\gamma_t M_\tau Y_\tau\,.$$

**Theorem 3.** *If $\|\nu\|_2^2 \leq B$ and $\beta$ is chosen as in Eq. (1), then $\|\hat{\nu}_t - \nu\|_{G_t}^2 \leq \beta$ for all $t \leq n$ with probability at least $1 - \delta = 1/(nK)$.*

Similar results exist in the literature for unweighted least-squares estimators (for example, Dani et al. [2008], Rusmevichientong and Tsitsiklis [2010], Abbasi-Yadkori et al. [2011]). In our case, however, $G_t$ is the weighted gram matrix, which may be significantly larger than an unweighted version when the weights become large. The proof of Theorem 3 is presented in Appendix C.

### Analysing the Regret

We start with some technical lemmas. Let $F$ be the failure event that $\|\hat{\nu}_{tk} - \nu_k\|_{G_{tk}}^2 > \beta$ for some $t \leq n$ and $1 \leq k \leq K$.

**Lemma 4** (Abbasi-Yadkori et al. [2012])**.** *Let $x_1, \ldots, x_n$ be an arbitrary sequence of vectors with $\|x_t\|_2^2 \leq c$ and let $G_t = I + \sum_{s=1}^{t-1} x_s x_s^\top$. Then $\sum_{t=1}^n \min\left\{1, \|x_t\|_{G_t^{-1}}^2\right\} \leq 2D\log\left(1 + \frac{c\cdot n}{D}\right)$.*

**Corollary 5.** *If $F$ does not hold, then* $\displaystyle\sum_{t=1}^n \gamma_{tk}\min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}}^2\right\} \leq 8D\log(1+4n^2)$.

*Proof.* Since $F$ does not hold we can apply Lemma 15 in the appendix to obtain

$$\gamma_{tk}\min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}}^2\right\} \leq 4\min\left\{1, \gamma_{tk}\|M_{tk}\|_{G_{tk}^{-1}}^2\right\}\,.$$

By Lemma 13 in the appendix, $\gamma_{tk}\|M_{tk}\|_2^2 \leq 4tD \leq 4nD$. Then simply apply Lemma 4. □

**Lemma 6.** *Suppose $F$ does not hold, then* $\displaystyle\sum_{k=1}^K \gamma_{tk}^{-1} \leq D\left(\max_k \|\nu_k\|_\infty + 4\sqrt{\beta/\alpha}\right)$.

*Proof.* We exploit the fact that $\gamma_{tk}^{-1}$ is an estimate of the variance, which is small whenever $\|M_{tk}\|_1$ is small:

$$\begin{aligned}\gamma_{tk}^{-1} &= \arg\max_{\tilde{\nu}_k \in C'_{tk}} \langle M_{tk}, \tilde{\nu}_k \rangle \left(1 - \langle M_{tk}, \tilde{\nu}_k \rangle\right) \leq \arg\max_{\tilde{\nu}_k \in C'_{tk}} \langle M_{tk}, \tilde{\nu}_k \rangle \\ &= \langle M_{tk}, \nu \rangle + \arg\max_{\tilde{\nu}_k \in C_{tk'}} \langle M_{tk}, \tilde{\nu}_k - \nu \rangle \overset{(a)}{\leq} \|M_{tk}\|_1 \|\nu_k\|_\infty + 4\sqrt{\beta} \|M_{tk}\|_{G_{tk}^{-1}} \\ &\overset{(b)}{\leq} \|M_{tk}\|_1 \|\nu_k\|_\infty + 4\sqrt{\beta} \|M_{tk}\|_{I/\alpha} \overset{(c)}{\leq} \|M_{tk}\|_1 \left(\|\nu_k\|_\infty + 4\sqrt{\beta/\alpha}\right),\end{aligned}$$

where (a) follows from Cauchy-Schwartz and the fact that $\nu_k \in C'_{tk}$, (b) since $G_{tk}^{-1} \leq I/\alpha$ and Proposition 8, (c) since $\|M_{tk}\|_{I/\alpha} = \sqrt{1/\alpha}\,\|M_{tk}\|_2 \leq \sqrt{1/\alpha}\,\|M_{tk}\|_1$. The result is completed since the resource constraints implies that $\sum_{k=1}^K \|M_{tk}\|_1 \leq D$. □

*Proof of Theorem 2.* By Theorem 3 we have that $F$ holds with probability at most $\delta = 1/(nK)$. If $F$ does not hold, then by the definition of the confidence set we have $\nu_k \in C_{tk}$ for all $t$ and $k$. Therefore

$$R_n = \mathbb{E}\sum_{t=1}^n \sum_{k=1}^K \left(\langle M_k^*, \nu_k \rangle - \psi(\langle M_{tk}, \nu_k \rangle)\right) \leq 1 + \mathbb{E}\left[\mathbb{1}\left\{\neg F\right\} \sum_{t=1}^n \sum_{k=1}^K \langle M_k^* - M_{tk}, \nu_k \rangle\right].$$

Note that we were able to replace $\psi(\langle M_{tk}, \nu_k \rangle) = \langle M_{tk}, \nu_k \rangle$, since if $F$ does not hold, then $M_{tk}$ will never be chosen in such a way that resources are over-allocated. We will now assume that $F$ does not hold and bound the argument in the expectation. By the optimism principle we have:

$$\begin{aligned}\sum_{t=1}^n \sum_{k=1}^K \langle M_k^* - M_{tk}, \nu_k \rangle &\overset{(a)}{\leq} \sum_{t=1}^n \sum_{k=1}^K \min\left\{1, \langle M_{tk}, \tilde{\nu}_{tk} - \nu_k \rangle\right\} \\ &\overset{(b)}{\leq} \sum_{t=1}^n \sum_{k=1}^K \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}} \|\tilde{\nu}_{tk} - \nu_k\|_{G_{tk}}\right\} \\ &\overset{(c)}{\leq} 2\sum_{t=1}^n \sum_{k=1}^K \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}} \sqrt{\beta}\right\} \\ &\overset{(d)}{\leq} 2\sqrt{n\sum_{t=1}^n \beta \left(\sum_{k=1}^K \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}}\right\}\right)^2} \\ &\overset{(e)}{\leq} 2\sqrt{n\sum_{t=1}^n \beta \left(\sum_{k=1}^K \gamma_{tk}^{-1}\right)\left(\sum_{k=1}^K \gamma_{tk} \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}}^2\right\}\right)} \\ &\overset{(f)}{\leq} 2\sqrt{nD\left(\max_k \|\nu_k\|_\infty + 4\sqrt{\frac{\beta}{\alpha}}\right)\sum_{t=1}^n \beta \left(\sum_{k=1}^K \gamma_{tk} \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}}^2\right\}\right)} \\ &\overset{(g)}{\leq} 4D\sqrt{2\beta nK\left(\max_k \|\nu_k\|_\infty + 4\sqrt{\frac{\beta}{\alpha}}\right)\log(1+4n^2)}.\end{aligned}$$

where (a) follows from the assumption that $\nu_k \in C_{tk}$ for all $t$ and $k$ and since $M_t$ is chosen optimistically, (b) by the Cauchy-Schwarz inequality, (c) by the definition of $\tilde{\nu}_{kt}$, which lies inside $C_{tk}$, (d) by Jensen's inequality, (e) by Cauchy-Schwarz again, (f) follows from Lemma 6. Finally (g) follows from Corollary 5. □

## 5 Regret in Resource-Laden Case

We now show that if there are enough resources such that the optimal strategy can complete every task with certainty, then the regret of Algorithm 2 is poly-logarithmic (in contrast to $O(\sqrt{n})$ otherwise). As before we exploit the low variance, but now the variance is small because $\langle M_{tk}, \nu_k \rangle$ is

close to 1, while in the previous section we argued that this could not happen too often (there is no contradiction as the quantity $\max_k \|\nu_k\|$ appeared in the previous bound).

**Theorem 7.** *If* $\sum_{k=1}^{K} \langle M_k^*, \nu_k \rangle = K$, *then* $R_n \leq 1 + 8\beta KD \log(1 + 4n^2)$.

*Proof.* We start by showing that the weights are large:

$$\gamma_{tk}^{-1} = \max_{\underline{\nu} \in C'_{tk}} \langle M_{tk}, \underline{\nu} \rangle (1 - \langle M_{tk}, \underline{\nu} \rangle) \leq \max_{\underline{\nu} \in C'_{tk}} (1 - \langle M_{tk}, \underline{\nu} \rangle)$$
$$\leq \max_{\bar{\nu}, \underline{\nu} \in C'_{tk}} \langle M_{tk}, \bar{\nu} - \underline{\nu} \rangle \leq \|M_{tk}\|_{G_{tk}^{-1}} \max_{\bar{\nu}, \underline{\nu} \in C'_{tk}} \|\bar{\nu} - \underline{\nu}\|_{G_{tk}} \leq \|M_{tk}\|_{G_{tk}^{-1}} 4\sqrt{\beta} \,.$$

Applying the optimism principle and using the bound above combined with Corollary 5 gives the result:

$$\mathbb{E}R_n \leq 1 + \mathbb{E}\left[\mathbb{1}\left\{\neg F\right\} \sum_{t=1}^{n} \sum_{k=1}^{K} \min\left\{1, \langle M_{tk}, \tilde{\nu}_{kt} - \nu_k \rangle\right\}\right]$$
$$\leq 1 + 2\mathbb{E}\left[\mathbb{1}\left\{\neg F\right\} \sum_{t=1}^{n} \sum_{k=1}^{K} \min\left\{1, \|M_{tk}\|_{G_{tk}^{-1}} \sqrt{\beta}\right\}\right]$$
$$= 1 + 2\mathbb{E}\left[\mathbb{1}\left\{\neg F\right\} \sum_{t=1}^{n} \sum_{k=1}^{K} \min\left\{1, \gamma_{tk}^{-1}\gamma_{tk} \|M_{tk}\|_{G_{tk}^{-1}}\right\} \sqrt{\beta}\right]$$
$$\leq 1 + 8\beta\, \mathbb{E}\left[\mathbb{1}\left\{\neg F\right\} \sum_{t=1}^{n} \sum_{k=1}^{K} \min\left\{1, \gamma_{tk} \|M_{tk}\|_{G_{tk}^{-1}}^2\right\}\right]$$
$$\leq 1 + 8\beta KD \log(1 + 4n^2) \,. \qquad \square$$

## 6 Experiments

We present two experiments to demonstrate the behaviour of Algorithm 2. All code and data is available in the supplementary material. Error bars indicate 95% confidence intervals, but sometimes they are too small to see (the algorithm is quite conservative, so the variance is very low). We used $B = 10$ for all experiments. The first experiment demonstrates the improvements obtained by using a weighted estimator over an unweighted one, and also serves to give some idea of the rate of learning. For this experiment we used $D = K = 2$ and $n = 10^6$ and

$$\nu = \begin{pmatrix} \nu_1 & \nu_2 \end{pmatrix} = \begin{pmatrix} 8/10 & 2/10 \\ 4/10 & 2 \end{pmatrix} \quad \Longrightarrow \quad M^* = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix} \quad \text{and} \quad \sum_{k=1}^{K} \langle M_k^*, \nu_k \rangle = 2 \,,$$

where the $k$th column is the parameter/allocation for the $k$th task. We ran two versions of the algorithm. The first, exactly as given in Algorithm 2 and the second identical except that the weights were fixed to $\gamma_{tk} = 4$ for all $t$ and $k$ (this value is chosen because it corresponds to the minimum inverse variance for a Bernoulli variable). The data was produced by taking the average regret over 8 runs. The results are given in Fig. 1. In Fig. 2 we plot $\gamma_{tk}$. The results show that $\gamma_{tk}$ is increasing linearly with $t$. This is congruent with what we might expect because in this regime the estimation error should drop with $O(1/t)$ and the estimated variance is proportional to the estimation error. Note that the estimation error for the algorithm with $\gamma_{tk} = 4$ will be $O(\sqrt{1/t})$.

For the second experiment we show the algorithm adapting to the environment. We fix $n = 5 \times 10^5$ and $D = K = 2$. For $\alpha \in (0, 1)$ we define

$$\nu_\alpha = \begin{pmatrix} 1/2 & \alpha/2 \\ 1/2 & \alpha/2 \end{pmatrix} \quad \Longrightarrow \quad M^* = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \quad \text{and} \quad \sum_{k=1}^{K} \langle M_k^*, \nu_k \rangle = 1 \,.$$

The unusual profile of the regret as $\alpha$ varies can be attributed to two factors. First, if $\alpha$ is small then the algorithm quickly identifies that resources should be allocated first to the first task. However, in the early stages of learning the algorithm is conservative in allocating to the first task to avoid over-allocation. Since the remaining resources are given to the second task, the regret is larger for small

$\alpha$ because the gain from allocating to the second task is small. On the other hand, if $\alpha$ is close to 1, then the algorithm suffers the opposite problem. Namely, it cannot identify which task the resources should be assigned to. Of course, if $\alpha = 1$, then the algorithm must simply learn that all resources can be allocated safely and so the regret is smallest here. An important point is that the algorithm never allocates all its resources at the start of the process because this risks over-allocation, so even in "easy" problems the regret will not vanish.

**Figure 1:** Weighted vs unweighted estimation

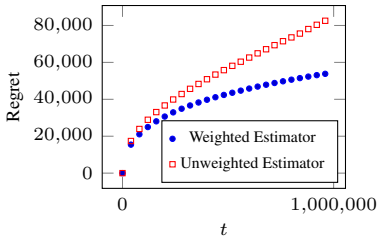



**Figure 2:** Weights

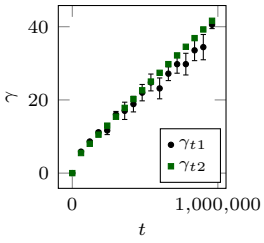
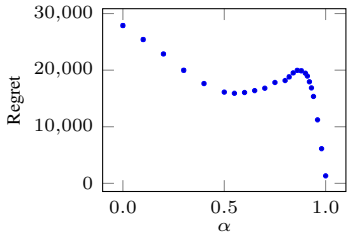



**Figure 3:** "Gap" dependence




## 7 Conclusions and Summary

We introduced the stochastic multi-resource allocation problem and developed a new algorithm that enjoys near-optimal worst-case regret. The main drawback of the new algorithm is that its computation time is exponential in the dimension parameters, which makes practical implementations challenging unless both $K$ and $D$ are relatively small. Despite this challenge we were able to implement that algorithm using a relatively brutish approach to solving the optimisation problem, and this was sufficient to present experimental results on synthetic data showing that the algorithm is behaving as the theory predicts, and that the use of the weighted least-squares estimation is leading to a real improvement.

Despite the computational issues, we think this is a reasonable first step towards a more practical algorithm as well as a solid theoretical understanding of the structure of the problem. As a consolation (and on their own merits) we include some other results:

- An efficient (both in terms of regret and computation) algorithm for the case where over-allocation is impossible.
- An algorithm for linear bandits on the hypercube that enjoys optimal regret bounds *and* adapts to sparsity.
- Theoretical analysis of weighted least-squares estimators, which may have other applications (e.g., linear bandits with heteroscedastic noise).

There are many directions for future research. The most natural is to improve the practicality of the algorithm. We envisage such an algorithm might be obtained by following the program below:

- Generalise the Thompson sampling analysis for linear bandits by Agrawal and Goyal [2012]. This is a highly non-trivial step, since it is no longer straight-forward to show that such an algorithm is optimistic with high probability. Instead it will be necessary to make do with some kind of local optimism for each task.
- The method of estimation depends heavily on the algorithm over-allocating its resources only with extremely low probability, but this significantly slows learning in the initial phases when the confidence sets are large and the algorithm is acting conservatively. Ideally we would use a method of estimation that depended on the real structure of the problem, but existing techniques that might lead to theoretical guarantees (e.g., empirical process theory) do not seem promising if small constants are expected.

It is not hard to think up extensions or modifications to the setting. For example, it would be interesting to look at an adversarial setting (even defining it is not so easy), or move towards a non-parametric model for the likelihood of success given an allocation.

## References


Yasin Abbasi-Yadkori, Csaba Szepesvári, and David Tax. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online-to-confidence-set conversions and application to sparse stochastic bandits. In *AISTATS*, volume 22, pages 1–9, 2012.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. *arXiv preprint arXiv:1209.3352*, 2012.

Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *The Journal of Machine Learning Research*, 3:397–422, 2003.

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

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Kristin P Bennett and Olvi L Mangasarian. Bilinear separation of two sets inn-space. *Computational Optimization and Applications*, 2(3):207–227, 1993.

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

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.

Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, pages 355–366, 2008.

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

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Akshay Krishnamurthy, Alekh Agarwal, and Miroslav Dudik. Efficient contextual semi-bandit learning. *arXiv preprint arXiv:1502.05890*, 2015.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. *arXiv preprint arXiv:1410.0949*, 2014.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2014a.

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. *arXiv preprint arXiv:1406.3840*, 2014b.

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

Marek Petrik and Shlomo Zilberstein. Robust approximate bilinear programming for value function approximation. *The Journal of Machine Learning Research*, 12:3027–3063, 2011.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

Thomas Sowell. *Is Reality Optional?: And Other Essays*. Hoover Institution Press, 1993.


## A Linear Algebra

We collect some well-known results in linear algebra for easy of reference. A square matrix $A$ is said to be positive definite (semi-definite) if it is symmetric and all its eigenvalues are positive (nonnegative). For $A, B$ positive definite, $A \leq B$ means $B - A$ is positive semi-definite.

**Proposition 8.** *Let $A$ and $B$ be positive-definite and $x$ and $y$ be vectors. The following hold:*

1. *If $A \leq B$, then $\|x\|_A \leq \|x\|_B$.*
2. *If $A \leq B$, then $A^{-1} \geq B^{-1}$.*
3. *If $A$ has maximum eigenvalue $\lambda_{\max}$, then $\|Ax\|_2 \leq \lambda_{\max} \|x\|_2$ and $\lambda_{\max} \leq \operatorname{trace}(A)$.*

## B Concentration Bounds

**Theorem 9.** *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 $Z \subseteq \{1, \ldots, n\}$ be such that $\mathbb{1}\{t \in Z\}$ is $\mathcal{F}_{t-1}$-measurable and let $R_t$ be $\mathcal{F}_{t-1}$-measurable such that $|X_t| \leq R_t$ almost surely. Now define*

$$V = \sum_{t \in Z} \mathbb{V}[X_t|\mathcal{F}_{t-1}] + \sum_{t \notin Z} R_t^2/2\,, \qquad R = \max_{t \in Z} R_t\,, \qquad \text{and} \qquad S = \sum_{t=1}^{n} X_t\,.$$

*Then $\mathbb{P}\{S \geq f(R,V)\} \leq \delta$, where*

$$f(r,v) = \frac{2(r+1)}{3}\log\frac{2}{\delta_{r,v}} + \sqrt{2(v+1)\log\frac{2}{\delta_{r,v}}}\,, \qquad \text{and}$$

$$\delta_{r,v} = \frac{\delta}{3(r+1)^2(v+1)^2}\,.$$

The proof follows along precisely the same lines as the proof of Theorem 13 by Lattimore et al. [2014b], which itself is essentially just a modification of the Freedman's version of the Bernstein's inequality [Bernstein, 1946, Freedman, 1975]. The only modification required is to merge the proofs of Theorems 3.14 and 3.15 by McDiarmid [1998] using either Lemma 2.6 or 2.7 in that work depending on whether $t \in Z$ or otherwise. The intuition is that we are locally able to use either Hoeffding's lemma (Lemma 2.6) or Bennet's variance-dependent lemma (Lemma 2.7). See also the classical works by Bennett [1962] and Hoeffding [1963]. Once this is done, a simple peeling argument, identical to that used in the proof of their Theorem 13 by Lattimore et al. [2014b], is used.

## C Proof of Theorem 3

Our approach generalises that used by Lattimore et al. [2014a] to the multi-dimensional case. Note that similar results were given by Dani et al. [2008], Rusmevichientong and Tsitsiklis [2010] and Abbasi-Yadkori et al. [2011], but none are able to effectively handle the heteroscedastic noise and so are unsuitable for our needs. Unfortunately we were not able to generalise the beautiful method of Abbasi-Yadkori et al. [2011], but our approach still enjoys relatively small constants and (in our view) is relatively insightful.

We will abbreviate the notation for simplicity. Pick some task $k$ and let $M_1, \ldots, M_t$ be a sequence of allocations chosen for it, $Y_1, \ldots, Y_t$ the corresponding rewards and $\gamma_1, \ldots, \gamma_t$ the weights as chosen by Algorithm 2. Fixing $k$, we omit the $k$-dependence in this section.

Recall that for $t \geq 1$ the gram matrix and weighted least-squares estimator are defined by

$$G_t = \alpha I + \sum_{s<t} \gamma_s M_s M_s^\top\,,$$

$$\hat{\nu}_t = G_t^{-1} \sum_{s<t} \gamma_s M_s Y_s\,.$$

We also set $G_0 = I$. Remember also that $Y_s$ is sampled from a Bernoulli distribution with parameter $\langle M_s, \nu\rangle$. Assuming that $\langle M_s, \nu\rangle \leq 1$, we can separate signal and noise by writing

$$Y_t = \langle M_s, \nu\rangle + \eta_s\,,$$

where $\eta_s \in [-1,1]$, $\mathbb{E}[\eta_s|\mathcal{F}_{s-1}] = 0$ and $\mathbb{V}[\eta_s|\mathcal{F}_{s-1}] = \langle M_s, \nu\rangle(1 - \langle M_s, \nu\rangle)$. Note that the algorithm is crafted in such a way that $\langle M_s, \nu\rangle \leq 1$ unless some confidence interval fails, which only occurs on a low probability failure event $F_t$ to be defined shortly. For this reason we are able to ignore the non-linear part of the pay-off for this section, but the price we pay is that the confidence intervals must be chosen wide enough that the failure probability is very low, while other algorithms (such as UCB) are able to recover from failing confidence intervals. The confidence sets are given by

$$C_s = \left\{\tilde{\nu} : \|\nu - \hat{\nu}_s\|_{G_s}^2 \leq \beta\right\} \qquad \text{and} \qquad C_s' = \left\{\tilde{\nu} : \|\nu - \hat{\nu}_s\|_{G_s}^2 \leq 4\beta\right\}\,.$$

The key in the proof of Theorem 3 will be controlling the size of $S_t$ defined by

$$S_t = \sum_{s<t} \gamma_s \eta_s M_s \,.$$

Let $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n$ be a sequence of failure events defined by

$$F_t = \left\{ \exists s \leq t \text{ such that } \|S_s\|_{G_s^{-1}} + \sqrt{\alpha B} \geq \sqrt{\beta} \right\} .$$

**Lemma 10.** *Let $t \geq 1$. If $F_t$ does not hold, then $\nu \in C_t$.*

*Proof.* Since $F_t$ does not hold, we have that $\|S_t\|_{G_t^{-1}} + \sqrt{\alpha B} \leq \sqrt{\beta}$. Then,

$$\begin{aligned}
\|\hat{\nu}_t - \nu\|_{G_t} &\overset{(a)}{=} \left\| G_t^{-1} S_t + G_t^{-1} \sum_{s<t} \gamma_s M_s M_s^\top \nu - G_t^{-1} G_t \nu \right\|_{G_t} \\
&\overset{(b)}{=} \left\| G_t^{-1} S_t - \alpha G_t^{-1} \nu \right\|_{G_t} \\
&\overset{(c)}{\leq} \|S_t\|_{G_t^{-1}} + \alpha \|\nu\|_{I/\alpha} \\
&\overset{(d)}{\leq} \sqrt{\beta} - \sqrt{\alpha B} + \sqrt{\alpha} \|\nu\|_2 \\
&\overset{(e)}{\leq} \sqrt{\beta} \,,
\end{aligned}$$

where (a), (b) are immediate by substituting the definitions, (c) from the triangle inequality and Proposition 8.§2. Finally, (d) and (e) follow from the assumptions and because $\|\nu\|_2 \leq \sqrt{B}$. Therefore $\nu \in C_t$. □

*Proof of Theorem 3.* Note that by definition $\beta \geq B$, hence $F_0$ holds. Let $t \leq n$ and assume that $F_{t-1}$ holds, which by Lemma 10 implies that $\nu \in C_s$ for all $s < t$. Shortly we will show that

$$\mathbb{P}\left\{ \|S_t\|_{G_t^{-1}} + \sqrt{\alpha B} \geq \sqrt{\beta} \text{ and not } F_{t-1} \right\} \leq \delta / n \,. \tag{3}$$

Then by induction we see that $\mathbb{P}\{\neg F_n\} \geq 1 - \delta$, and so by Lemma 10 it follows that $\nu \in C_t$ for all $t \leq n$ with probability at least $1 - \delta$.

We now work on showing Eq. (3). Let $\lambda \in \mathbb{R}^D$ and define

$$\begin{aligned}
V_{s,\lambda} &= \begin{cases} \mathbb{V}[\eta_s | \mathcal{F}_{s-1}] \gamma_s^2 \langle M_s, \lambda \rangle^2 \,, & \text{if } \gamma_s > 4; \\ \gamma_s \langle M_s, \lambda \rangle^2 \,, & \text{otherwise,} \end{cases} \\
R_\lambda &= \max_{s<t} \{ \gamma_s \langle M_s, \lambda \rangle : \gamma_s > 4 \} \,.
\end{aligned}$$

Then, by Theorem 9 we have with probability at least $1 - \delta/n$ that

$$\begin{aligned}
\langle S_t, \lambda \rangle &= \sum_{s<t} \gamma_s \eta_s \langle M_s, \lambda \rangle \\
&\leq \frac{2(R_\lambda + 1)}{3} \log \frac{1}{\delta_\lambda} + \sqrt{2 \left( 1 + \sum_{s<t} V_{s,\lambda} \right) \log \frac{1}{\delta_\lambda}} \,,
\end{aligned}$$

where

$$\delta_\lambda = \frac{3n}{\delta \left(1 + R_\lambda\right)^2 \left(1 + \sum_{s<t} V_{s,\lambda}\right)^2} \,.$$

Let $\varepsilon > 0$ and $C > 0$ be constants to be chosen later and define the covering set

$$\Lambda = \{-C, -C + \varepsilon, \ldots, 0, \ldots, \varepsilon, \ldots, C - \varepsilon, C\}^D \,,$$

which has size $N = |\Lambda| = (2C/\varepsilon)^D$. Then, by the union bound we have with probability at least $1-\delta$ that

$$\langle S_t, \lambda \rangle \le \frac{2(R_\lambda + 1)}{3} \log \frac{N}{\delta_\lambda} + \sqrt{2\left(1 + \sum_{s<t} V_{s,\lambda}\right) \log \frac{N}{\delta_\lambda}} \quad \text{for all } \lambda \in \Lambda\,. \tag{4}$$

From now on, assume this event occurs. Since $F_{t-1}$ does not hold we can apply Lemma 14 to get

$$\left\|G_t^{-1} S_t\right\|_\infty \le \|S_t\|_1 \le 2t^2 D = C\,.$$

Let $\lambda = G_t^{-1} S_t$ (which is a random quantity) for which $\|\lambda\|_\infty \le C$. Then there exists a $\lambda' \in \Lambda$ such that $\lambda' \le \lambda$ and $\|\lambda' - \lambda\|_\infty \le \varepsilon$.

$$\|S_t\|^2_{G_t^{-1}} = \langle S_t, \lambda \rangle \le \|S_t\|_1 \varepsilon + \langle S_t, \lambda' \rangle\ .$$

Therefore

$$\|S_t\|^2_{G_t^{-1}} \le \|S_t\|_1 \varepsilon + \frac{2(R_\lambda + 1)}{3} \log \frac{N}{\delta_\lambda} + \sqrt{2\left(1 + \sum_{s<t} V_{s,\lambda}\right) \log \frac{N}{\delta_\lambda}}\,, \tag{5}$$

where we used the fact that $R_{\lambda_1} \le R_{\lambda_2}$ and $V_{s,\lambda_1} \le V_{s,\lambda_2}$ and $1/\delta_{\lambda_1} \le 1/\delta_{\lambda_2}$ for $\lambda_1 \le \lambda_2$. We now bound the sum term:

$$\begin{aligned}
\sum_{s<t} V_{s,\lambda} &\overset{(a)}{\le} \sum_{s<t} \gamma_s \langle M_s, \lambda \rangle^2 \\
&\overset{(b)}{=} \sum_{s<t} \gamma_s (G_t^{-1} S_t)^\top M_s M_s^\top G_t^{-1} S_t \\
&\overset{(c)}{=} (G_t^{-1} S_t)^\top \sum_{s<t} \gamma_s M_s M_s^\top G_t^{-1} S_t \\
&\overset{(d)}{\le} S_t^\top G_t^{-1} S_t \overset{(e)}{=} \|S_t\|^2_{G_t^{-1}}\ ,
\end{aligned} \tag{6}$$

where (a) follows from the definition of $V_{s,\lambda}$ and since if $\nu \in C_s$, then $\gamma_s \le \mathbb{V}[\eta_s | \mathcal{F}_{s-1}]$, (b) by substituting the definition of $\lambda$, (c) by calculation, (d) follows since $\sum_{s<t} \gamma_s M_s M_s^\top < G_t$ and (e) is just the definition. We now set $\varepsilon = 1/(2t^2 D)$ to obtain

$$\begin{aligned}
\|S_t\|^2_{G_t^{-1}} &\overset{(a)}{\le} \|S_t\|_1 \varepsilon + \frac{2(R_\lambda + 1)}{3} \log \frac{N}{\delta_\lambda} + \sqrt{2\left(1 + \|S_t\|^2_{G_t^{-1}}\right) \log \frac{N}{\delta_\lambda}} \\
&\overset{(b)}{\le} 1 + \frac{2(R_\lambda + 1)}{3} \log \frac{N}{\delta_\lambda} + \sqrt{2\left(1 + \|S_t\|^2_{G_t^{-1}}\right) \log \frac{N}{\delta_\lambda}} \\
&\overset{(c)}{\le} 1 + \frac{2\left(\frac{2\|S_t\|_{G_t^{-1}}}{\sqrt{\beta}} + 1\right)}{3} \log \frac{N}{\delta_\lambda} + \sqrt{2\left(1 + \|S_t\|^2_{G_t^{-1}}\right) \log \frac{N}{\delta_\lambda}}\,,
\end{aligned} \tag{7}$$

where (a) follows by substituting the previous computation into Eq. (5), (b) since $\|S_t\|_1 \varepsilon \le 1$ by Lemma 14 and the assumption that $F_{t-1}$ does not hold, (c) by Lemma 12 and the assumption that $F_{t-1}$ does not hold. From Eq. (3) and Lemma 12 and the definition of $\beta$ we also obtain

$$\delta_\lambda \le \frac{3n}{\delta\left(1 + \|S_t\|^2_{G_t^{-1}}\right)^2}\,.$$

By rearranging and naively simplifying Eq. (7), it can be shown that

$$\begin{aligned}
\|S_t\|_{G_t^{-1}} + \sqrt{B} &\le 1 + \sqrt{\alpha B} + 2\sqrt{\log \frac{N}{\delta_\lambda}} \\
&= 1 + \sqrt{\alpha B} + 2\sqrt{\log\left(\frac{3nN\left(1 + \|S_t\|^2_{G_t^{-1}}\right)^2}{\delta}\right)}\,.
\end{aligned}$$

The result is finally completed by solving the equation above and choosing

$$\beta = \left(1 + \sqrt{\alpha B} + 2\sqrt{\log\left(\frac{6nN}{\delta}\log\left(\frac{3nN}{\delta}\right)\right)}\right)^2 . \qquad \square$$

It remains to prove the lemmas that were used in this proof.

**Lemma 11.** *For any $s < t$, it holds that*

$$\gamma_s \left\|M_s\right\|_{G_t^{-1}} \le \gamma_s \left\|M_s\right\|_{G_s^{-1}} .$$

*Further, if $F_{t-1}$ does not hold, then for all $s < t$ such that $\gamma_s > 4$,*

$$\gamma_s \left\|M_s\right\|_{G_t^{-1}} \le \gamma_s \left\|M_s\right\|_{G_s^{-1}} \le \frac{2}{\sqrt{\beta}} \le 1 .$$

*Proof.* The first inequality follows because $G_t \ge G_s$ and an application of Proposition 8.§2. Since $F_{t-1}$ does not hold, we have $\nu \in C_s$ and since $\gamma_s > 4$ we have from Eq. (2) that one of the following is true:

$$\gamma_s^{-1} \ge \frac{1}{2}\left(\langle M_s, \hat{\nu}_s\rangle + 2\sqrt{\beta}\left\|M_s\right\|_{G_s^{-1}}\right) \ge \sqrt{\beta}\left\|M_s\right\|_{G_s^{-1}}/2 ;$$
$$\gamma_s^{-1} \ge \frac{1}{2}\left(1 - \langle M_s, \hat{\nu}_s\rangle + 2\sqrt{\beta}\left\|M_s\right\|_{G_s^{-1}}\right) \ge \sqrt{\beta}\left\|M_s\right\|_{G_s^{-1}}/2 . \qquad \square$$

**Lemma 12.** *If $F_{t-1}$ does not hold and $\lambda = G_t^{-1} S_t$, then $R_\lambda \le \dfrac{2\left\|S_t\right\|_{G_t^{-1}}}{\sqrt{\beta}}$.*

*Proof.* We apply Lemma 11 to get

$$\gamma_s \langle M_s, \lambda\rangle \overset{(a)}{\le} \frac{2\langle M_s, \lambda\rangle}{\left\|M_s\right\|_{G_t^{-1}}\sqrt{\beta}} \overset{(b)}{=} \frac{2\langle M_s, G_t^{-1}S_t\rangle}{\left\|M_s\right\|_{G_t^{-1}}\sqrt{\beta}} \overset{(c)}{\le} \frac{2\left\|S_t\right\|_{G_t^{-1}}}{\sqrt{\beta}} ,$$

where (a) follows from Lemma 11, (b) is just the definition of $\lambda$ and (c) is follows from Cauchy-Schwarz. $\square$

**Lemma 13.** *If $F_t$ does not hold, then $\gamma_t \left\|M_t\right\|_1 \le 4tD$ and $\gamma_t \left\|M_t\right\|_2^2 \le 4tD$.*

*Proof.* The result holds trivially if $\gamma_t = 4$. Suppose $\gamma_t > 4$ and let $\lambda_{\max}$ be the maximum eigenvalue of $G_t$. Then, by Lemma 11, we have

$$\begin{aligned}
\gamma_t \left\|M_t\right\|_2^2 &\overset{(a)}{\le} \frac{2}{\sqrt{\beta}}\left\|M_t\right\|_2^2 \left\|M_t\right\|_{G_t^{-1}}^{-1} \\
&\overset{(b)}{\le} 2\sqrt{D/\beta}\left\|M_t\right\|_2 \left\|M_t\right\|_{G_t^{-1}}^{-1} \\
&\overset{(c)}{\le} \left\|G_t^{1/2} G_t^{-1/2} M_t\right\|_2 \left\|M_t\right\|_{G_t^{-1}}^{-1} \\
&\overset{(d)}{\le} \sqrt{\lambda_{\max}} ,
\end{aligned}$$

where (a) follows from Lemma 11, (b) by bounding $\left\|M_t\right\|_2 \le \sqrt{D}$, (c) holds by $4D \le \beta$, and (d) follows from Proposition 8.§3. Similarly,

$$\begin{aligned}
\gamma_t \left\|M_t\right\|_1 &\le \frac{2}{\sqrt{\beta}}\left\|M_t\right\|_1 \left\|M_t\right\|_{G_t^{-1}}^{-1} \\
&\le \frac{2\sqrt{D}}{\sqrt{\beta}}\left\|M_t\right\|_2 \left\|M_t\right\|_{G_t^{-1}}^{-1} \\
&\le \sqrt{\lambda_{\max}} .
\end{aligned}$$

Now assume that $\gamma_s \|M_s\|_2^2 \leq 4sD$ for all $s < t$, which is immediate if $t = 1$. Then,

$$\begin{aligned}\gamma_t \|M_t\|_2^2 &\leq \sqrt{\lambda_{\max}} \overset{(a)}{\leq} \sqrt{\operatorname{trace}(G_t)} \\ &= \sqrt{\left(D + \sum_{s=1}^{t-1} \gamma_t \|M_t\|_2^2\right)} \leq \sqrt{\left(D + 4D \sum_{s=1}^{t-1} s\right)} \\ &= \sqrt{D\left(1 + 2t(t-1)\right)} \leq 4tD\,,\end{aligned}$$

where (a) again follows from Proposition 8.§3 and the remaining steps are immediate. Therefore by induction, we have $\gamma_t \|M_t\|_2^2 \leq 4tD$ for all $t$. □

**Lemma 14.** *If $F_{t-1}$ does not hold, then $\|S_t\|_1 \leq 2t^2 D$.*

*Proof.* We use $|\eta_s| \leq 1$ and the previous lemma to get

$$\|S_t\|_1 = \left\|\sum_{s<t} \gamma_s \eta_s M_s\right\|_1 \leq \sum_{s<t} \gamma_s \|M_s\|_1 \leq \sum_{s<t} 4sD \leq 2t^2 D\,. \qquad \square$$

**Lemma 15.** *If $F_t$ does not hold, then $\gamma_t \min\left\{1, \|M_t\|_{G_t^{-1}}^2\right\} \leq 4 \min\left\{1, \gamma_t \|M_t\|_{G_t^{-1}}^2\right\}$.*

*Proof.* If $\gamma_t = 4$, then the result is trivial. For $\gamma_t > 4$, by Lemma 11, $\gamma_t \|M_t\|_{G_t^{-1}}^2 \leq 1$. Hence, we need to prove $\gamma_t \min\left\{1, \|M_t\|_{G_t^{-1}}^2\right\} \leq 4\gamma_t \|M_t\|_{G_t^{-1}}^2$, which is obvious. □

## D Proof of Theorem 1

Define $e_{di} \in \mathbb{R}^{D \times K}$ to be the matrix with $(e_{di})_{ck} = \mathbb{1}\{c = d \text{ and } k = i\}$. For $M \in \mathcal{M}$ we write $\mu(M) = \sum_{k=1}^K \psi(\langle M_k, \nu_k\rangle)$ to be the reward for allocation $M$. Given an allocation $M \in \mathcal{M}$ we define the conditional optimal allocation function $M^* : \mathcal{M} \to \mathbb{R}^{D \times K}$ by

$$\begin{aligned}M^*(M) &= \operatorname*{arg\,max}_{M' \in \mathcal{M}} \left\{\sum_{k=1}^K \psi(\langle M'_k, \nu\rangle) : M'_{dk} \geq M_{dk} \text{ for all } d \text{ and } k\right\}\,, \\ \mu^*(M) &= \sum_{k=1}^K \psi(\langle M^*(M)_k, \nu_k\rangle)\,, \\ \mu^* &= \mu^*(0)\,.\end{aligned}$$

Note that $M^*(0) = M^*$ is the optimal allocation while $M^*(M)$ is the optimal allocation given that one has committed to allocating at least $M$ already. Let $M_t \in [0,1]^{K \times D}$ be the allocation of Algorithm 1 after $t$ iterations. Assume that $\mu^*(M_{t-1}) = \mu^*$, which is trivial for $t = 1$. Let $(i, d)$ be the task/resource pair selected in the $t$th iteration of Algorithm 1. Suppose that at this point it is sub-optimal to allocate resource $d$ to task $i$. Then

$$\nabla_{e_{di}} \mu^*(M_{t-1}) < 0\,.$$

This implies that under the optimal allocation, resource $d$ should not be allocated to task $i$ and instead to some other task $j \neq i$. Therefore $\psi(\langle M^*(M_{t-1})_i, \nu_i\rangle) = 1$, since otherwise

$$\nabla_{e_{di} - e_{dj}} \mu^*(M_{t-1}) = \nu_{di} - \nu_{dj} \geq 0\,,$$

which is a contradiction. Therefore there exists some other resource $1 \leq c \leq D$ that is assigned to task $i$ under the optimal allocation. We choose

$$\alpha = \frac{\nu_{dj}\nu_{ci}}{\nu_{di}\nu_{cj}} \leq 1$$

and compute the derivative, to get

$$\begin{aligned}
&\nabla_{e_{di}-e_{ci}\nu_{di}/\nu_{ci}-e_{dj}+e_{cj}\alpha\nu_{di}/\nu_{ci}}\mu^*(M_{t-1}) \\
&= \nu_{cj}\alpha\frac{\nu_{di}}{\nu_{ci}} - \nu_{dj} \\
&= 0\,,
\end{aligned}$$

which again implies that allocating resource $d$ to task $i$ is not sub-optimal, which is a contradiction. Therefore $\nabla_{e_{di}}\mu^*(M_{t-1}) = 0$ and so Algorithm 1 is optimal by induction.

## E Resource Allocation when $\|\nu_k\|_1 \leq 1$

Throughout this subsection we assume that $\|\nu_k\|_1 \leq 1$ for all $k$. Therefore $\psi(\langle M_k, \nu_k\rangle) = \langle M_k, \nu_k\rangle$ for all $M \in \mathcal{M}$ and $k$. Therefore the optimal strategy assigns all of resource $d$ to the task $k$ for which $\nu_{kd}$ is the greatest:

$$M^*_{kd} = \mathbb{1}\left\{k = \arg\max_i \nu_{id}\right\}\,,$$

where ties are broken arbitrarily. The algorithm operates by maintaining a set of tasks for each resource that are plausibly still optimal. Each resource type is then allocated to a single task in this set uniformly at random, with tasks being removed from this set in phases as the algorithm proves that allocating a particular resource to this task is sub-optimal with high probability. The structure of the problem then allows us to simultaneously estimate all parameters of $\nu$ using importance sampling, which ultimately leads to an optimal rate. The algorithm is easily implemented to run in $O(KD)$ per iteration.

**Algorithm 3** Unconstrained Allocation Algorithm

**Input:** $K$, $D$, $n$, $\delta$
$\mathcal{A}_d \coloneqq [K]$ and $\Delta_d \coloneqq 1$ and $\tau_d \coloneqq n_d \coloneqq 0$
**for** $t \in 1, \ldots, n$ **do**
  **for** $d \in 1, \ldots, D$ **do**
    **if** $t = \tau_d + 1$ **then**
      $\hat{\nu}_{kd} \coloneqq \frac{1}{n_d}\sum_{s=\tau_d - n_d + 1}^{\tau_d} Z_{skd} Y_{sk}$
      $\mathcal{A}_d \coloneqq \mathcal{A}_d \cap \{k : \hat{\mu}_{kd} + 2\Delta_d \geq \max_j \hat{\mu}_{jd}\}$
      $\Delta_d \coloneqq \Delta_d / 2$
      $n_d \coloneqq n(|\mathcal{A}_d|, \Delta_d)$ and $\tau_d \coloneqq \tau_d + n_d$
    **end if**
    $I_{td} \sim \text{Uniform}(\mathcal{A}_d)$
    Choose $M_{tkd} \coloneqq \mathbb{1}\{k = I_{td}\}$
    $Z_{tkd} \coloneqq |\mathcal{A}_d| M_{tkd} - \frac{|\mathcal{A}_d|}{|\mathcal{A}_d| - 1}(1 - M_{tkd})$
  **end for**
  Observe reward $Y_{tk} \sim \text{Bernoulli}(\langle M_{tk}, \nu_k\rangle)$
**end for**
**function** $n(m, \Delta)$
  **Return** $\left\lceil \frac{2(6 + 3m + \Delta)}{3\Delta^2} \log\frac{2}{\delta} \right\rceil$
**end function**

## F Regret of Algorithm 3

**Theorem 16.** *Define $\nu^*_d = \max_k \nu_{kd}$ and $\Delta_{kd} = \nu^*_d - \nu_{kd} \geq 0$. The regret of Algorithm 3 when run with $\delta = (DKn)^{-2}$ is at most*

$$R_n \in O\left(\sum_{d=1}^{D}\sum_{k:\Delta_{kd}>0} \frac{\log nKD}{\Delta_{kd}}\right)\,. \tag{8}$$

**Corollary 17.** *The regret of Algorithm 3 satisfies $R_n \in \tilde{O}(D\sqrt{Kn})$ in the worst-case.*

The proof of the corollary is omitted, but follows from standard arguments for converting from problem-dependent to problem-independent regret bounds (Bubeck and Cesa-Bianchi [2012] and others).

Before presenting the analysis we compare the regret bound of Theorem 16 to the well-known problem dependent bounds for finite-armed bandits, which look the same as Eq. (8), but with $D = 1$. An incautious reader might believe that a bound similar to Eq. (8) could be derived by simultaneously running $D$ copies of some optimal bandit algorithm. But this is not the case because the algorithm observes a reward for each task and not for each resource. Alternatively one could ignore the semi-bandit feedback and apply an algorithm designed for stochastic linear bandits. This approach also leads to sub-optimal bounds because the $K$ will appear outside of the square root. The optimal regret can only be obtained be exploiting the special structure of the problem. Notably, that if only a small amount of resources are allocated to a particular task, then the probability that it is completed is close to zero and hence the variance of the outcome is significantly reduced. This low variance can then be exploited to accelerate the rate of estimation of the parameters beyond what is normally possible.

*Proof of Theorem 16.* We will analyse the regret using the following decomposition

$$R_n = \sum_{t=1}^{n}\sum_{d=1}^{D}\nu_d^* - \mathbb{E}\left[\sum_{t=1}^{n}\sum_{k=1}^{K}\langle M_{tk}, \nu_k\rangle\right] = \sum_{d=1}^{D}\mathbb{E}\left[\sum_{t=1}^{n}\Delta_{dI_{td}}\right].$$

Now we fix $d$ and analyse the expectation inside the sum. Let $\tau_{d1}, \tau_{d2}, \ldots$ be the sequence of values of $\tau_d$ as it is updated in Line 9 of the algorithm, and let $n_{d1}, n_{d2}, \ldots$ be the corresponding sequence of the values of $n_d$. Similarly, let $\mathcal{A}_{d1}, \mathcal{A}_{d2}, \ldots$ be the sequence of sets of active tasks.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{n}\Delta_{dI_{td}}\right] &= \mathbb{E}\left[\sum_{\ell=1}^{\infty}\sum_{t=\tau_{d\ell}-n_{d\ell}+1}^{\tau_{d\ell}}\Delta_{dI_{td}}\right] = \sum_{\ell=1}^{\infty}\mathbb{E}\left[\sum_{t=\tau_{d\ell}-n_{d\ell}+1}^{\tau_{d\ell}}\Delta_{dI_{td}}\right] \\
&= \sum_{\ell=1}^{\infty}\mathbb{E}\left[n_{d\ell}\sum_{k\in\mathcal{A}_{d\ell}}\frac{\Delta_{kd}}{|\mathcal{A}_{d\ell}|}\right] \le \sum_{\ell=1}^{\infty}\mathbb{E}\left[\frac{20}{3(2^{-\ell})^2}\left(\log\frac{2}{\delta}\right)\sum_{k\in\mathcal{A}_{d\ell}}\Delta_{kd}\right]. \qquad (9)
\end{aligned}$$

Shortly we are going to show with sufficiently high probability that for $\ell \ge \lceil -\log_2(\Delta_{kd}/4)\rceil$ we have $k \notin \mathcal{A}_{d\ell}$ and that $k^*d = \arg\max_k \nu_{kd}$ is in $\mathcal{A}_{d\ell}$. Therefore

$$(9) \le \sum_{k=1}^{K}\sum_{\ell=1}^{\lceil -\log_2(\Delta_{kd}/4)\rceil}\frac{20\Delta_{kd}}{3(2^{-\ell})^2}\log\frac{2}{\delta} \le \sum_{k=1}^{K}\frac{16\cdot 20\cdot 4}{3\Delta_{kd}}\log\frac{2}{\delta}.$$

Let $\{\mathcal{F}_t\}_{t=1}^n$ be the filtration of information available up to each time step. Let $\tau_{d\ell} - n_{d\ell} + 1 \le t \le \tau_{d\ell}$. A straightforward computation shows the following results:

1. $\mathbb{E}[Z_{tkd}Y_{tk}|\mathcal{F}_{t-1}] = \nu_{kd}$.
2. $Z_{tkd}Y_{tk} \in \{0, |\mathcal{A}_d|\}$.
3. $\mathbb{V}[Z_{tkd}Y_{tk}|\mathcal{F}_{t-1}] \le |\mathcal{A}_d| + 2$.

Let $\hat{\nu}_{kd}$ be the estimate of $\nu_{kd}$ made at time step $\tau_\ell$ in Line 6 of Algorithm 3. We apply the martingale version of Bernstein's inequality (Theorem 3.15 by McDiarmid [1998]) to obtain

$$\begin{aligned}
\mathbb{P}\left\{|\hat{\nu}_{kd} - \nu_{kd}| \ge 2^{-\ell}\right\} &= \mathbb{P}\left\{\left|\sum_{s=\tau_d-n_d+1}^{\tau_d} Z_{skd}R_{sk} - n_d\nu_{kd}\right| \ge n_d 2^{-\ell}\right\} \\
&\le 2\exp\left(-\frac{n_d(2^{-\ell})^2}{2\left(|\mathcal{A}_d| + 2 + \frac{|\mathcal{A}_d|2^{-\ell}}{3}\right)}\right) \le \delta.
\end{aligned}$$

But if $|\hat{\nu}_{kd} - \nu_{kd}| \le 2^{-\ell}$ for all $k$, then (a) $k_d^*$ is not removed from $|\mathcal{A}_d|$ and (b) $\ell \ge \lceil -\log_2(\Delta_{kd}/4)\rceil$ implies that $k$ is removed from $\mathcal{A}_d$. The probability that there exists a resource $d$, phase $\ell$ and $k$ such that $|\hat{\nu}_{kd} - \nu_{kd}| \ge 2^{-\ell}$ in the $\ell$th phase is bounded using the union bound by $DK\ell\delta \le DKn\delta$.

Therefore if Algorithm 3 is run with $\delta = (DKn)^{-2}$, then the contribution of the regret to the failure of any confidence set is at most 1, which leads to a regret bound

$$R_n \leq 1 + \frac{4 \cdot 16 \cdot 20}{3} \sum_{d=1}^{D} \sum_{k=1}^{K} \frac{1}{\Delta_{kd}} \log \frac{2}{\delta}$$

as required. □

## G Linear Bandits on the Hypercube

The importance sampling approach used in the previous section can be applied to linear stochastic bandits on the hypercube. In this case the algorithm chooses $M_t \in [0,1]^d$ at each time step and receives reward $Y_t = \langle M_t, \nu \rangle + \eta_t$ where $\nu \in \mathbb{R}^d$ satisfies $\|\nu\|_1 \leq 1$ and $\eta_t \in [-1,1]$ has zero mean. Note that $\nu$ may be negative in some dimensions, so the optimal strategy is not knowable in advance.

**Algorithm 4**

**for** $t \in 1, \ldots, n$ **do**
  **for** $d \in 1, \ldots, D$ **do**
    $\hat{\nu}_{td} = \dfrac{\sum_{\tau=1}^{t-1} \psi_{td} M_{td} Y_t}{\sum_{\tau=1}^{t-1} \psi_{td}}$
    $c_{td} = \sqrt{\dfrac{2 \log(2n^2)}{\sum_{\tau=1}^{t-1} \psi_{td}}}$
    Sample $X_{td} \in \{-1, 1\}$ with $\mathbb{P}\{X_{td} = 1\} = 1/2$
    $\psi_{td} = \begin{cases} 1 & \text{if } \hat{\nu}_{td} \in (-c_{td}, c_{td}) \\ 0 & \text{otherwise} \end{cases}$
    $M_{td} = \begin{cases} 1 & \text{if } \hat{\nu}_{td} - c_{td} > 0 \\ -1 & \text{if } \hat{\nu}_{td} + c_{td} < 0 \\ X_{td} & \text{otherwise} \end{cases}$
  **end for**
**end for**

**Theorem 18.** *The regret of Algorithm 4 is at most* $R_n \leq 3 \|\nu\|_1 + \sum_{d:\nu_d \neq 0} \dfrac{2 \log(2n^2)}{|\nu_d|}$.

This result is especially nice because (a) the algorithm is efficient, (b) the bound scales optimally with the dimension, (c) the problem-dependent bound is essentially correct, and finally (d), the bound is adaptive to sparsity in $\nu$ with no dependence on $D$ if $\nu$ is sparse. The algorithm and proof are significantly more straight-forward than above as the Bernstein's inequality and phases can be replaced by straight-forward union bounds in combination with Azuma's inequality. Details may be found in Appendix G.

*Proof of Theorem 18.* Let $\{\mathcal{F}_t\}$ be the filtration with $\mathcal{F}_t$ containing information up to time step $t$. Then $\hat{\nu}_{td}$, $c_{td}$ and $\psi_{td}$ are all $\mathcal{F}_{t-1}$-measurable. First we note that $Y_t \in [-2, 2]$ and that if $\psi_{td} = 1$, then $\mathbb{E}[\psi_{td} M_{td} Y_t | \mathcal{F}_{t-1}] = \nu_d$ and $\psi_{td} M_{td} Y_t \in [-2, 2]$. Let $F_d$ be the event that there exists a $t \leq n$ for which $|\hat{\nu}_{td} - \nu_d| > c_{td}$. By Azuma's inequality and the union bound we have $\mathbb{P}\{F_d\} \leq 1/n$.

We now decompose the regret

$$\begin{aligned}
R_n &= \mathbb{E}\left[\sum_{t=1}^{n}\sum_{d=1}^{D}(|\nu_d| - M_{td}\nu_d)\right] \\
&= \sum_{d=1}^{D}\mathbb{E}\left[\mathbb{1}\left\{F_d\right\}\cdot 2n|\nu_d| + \mathbb{1}\left\{\neg F_d\right\}\sum_{t=1}^{n}(|\nu_d| - M_{td}\nu_d)\right] \\
&\leq 2\left\|\nu\right\|_1 + \sum_{d=1}^{D}\mathbb{E}\left[\mathbb{1}\left\{\neg F_d\right\}\sum_{t=1}^{n}(|\nu_d| - M_{td}\nu_d)\right] \\
&= 2\left\|\nu\right\|_1 + \sum_{d:\nu_d\neq 0}|\nu_d|\left\lceil\frac{2\log(2n^2)}{|\nu_d|^2}\right\rceil \\
&\leq 3\left\|\nu\right\|_1 + \sum_{d:\nu_d\neq 0}\frac{2\log(2n^2)}{|\nu_d|}\,.
\end{aligned}$$

The second last line follows from two facts. First, if $\psi_{td} = 0$ and $\neg F$, then $|\nu_d| - M_{td}\nu_d = 0$. Second, if $\neg F$ and

$$\sum_{\tau=1}^{t-1}\psi_{\tau d} > \left\lceil\frac{2\log(2n^2)}{|\nu_d|^2}\right\rceil\,,$$

then $\psi_{td} = 0$. □

**Remark 19.** With only a little effort this algorithm could be made anytime. It may also be possible to make the (already quite small) constants smaller.