Combinatorial semi-bandit with known covariance

Authored by:

R?my Degenne
Vianney Perchet

Abstract

The combinatorial stochastic semi-bandit problem is an extension of the classical multi-armed bandit problem in which an algorithm pulls more than one arm at each stage and the rewards of all pulled arms are revealed. One difference with the single arm variant is that the dependency structure of the arms is crucial. Previous works on this setting either used a worst-case approach or imposed independence of the arms. We introduce a way to quantify the dependency structure of the problem and design an algorithm that adapts to it. The algorithm is based on linear regression and the analysis uses techniques from the linear bandit literature. By comparing its performance to a new lower bound, we prove that it is optimal, up to a poly-logarithmic factor in the number of arms pulled.

1 Paper Body

and setting

The multi-armed bandit problem (MAB) is a sequential learning task in which an algorithm takes at each stage a decision (or, ?pulls an arm?). It then gets a reward from this choice, with the goal of maximizing the cumulative reward [Robbins, 1985]. We consider here its stochastic combinatorial extension, in which the algorithm chooses at each stage a subset of arms [Audibert et al., 2013, Cesa-Bianchi and Lugosi, 2012, Chen et al., 2013, Gai et al., 2012]. These arms could form, for example, the path from an origin to a destination in a network. In the combinatorial setting, contrary to the the classical MAB, the inter-dependencies between the arms can play a role (we consider that the distribution of rewards is invariant with time). We investigate here how the covariance structure of the arms affects the difficulty of the learning task and whether it is possible to design a unique algorithm capable of performing optimally in all cases from the simple scenario with independent rewards to the more challenging scenario of general correlated rewards. Formally, at each stage t ? N, t ? 1, an algorithm pulls m ? 1 arms among d ? m. Such a set of m

arms is called an ?action? and will be denoted by At ? {0, 1}d , a vector with exactly m non-zero entries. The possible actions are restricted to an arbitrary fixed subset A ? {0, 1}d . After choosing d action At , the algorithm receives the reward A¿ t Xt , where Xt ? R is the vector encapsulating the reward of the d arms at stage t. The successive reward vectors (Xt)t?1 are i.i.d with unknown mean ? ? Rd . We consider a semi-bandit feedback system: after choosing the action At , the algorithm observes the reward of each of the arms in that action, but not the other rewards. Other possible feedbacks previously studied include bandit (only A¿ t Xt is revealed) and full information (Xt is revealed). The goal of the algorithm is to maximize the cumulated reward up to stage T ? 1 or equivalently to minimize the expected regret, which is the difference of the reward that would have been gained by choosing the best action in hindsight A? and what was actually gained: ERT = E

T X (A?¿ ? ? A¿ t ?) . t=1

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

For an action A ? A, the difference ?A = (A?¿ ? ? A¿ ?) is called gap of A. We denote by ?t the PT gap of At , so that regret rewrites as ERT = E t=1 ?t . We also define the minimal gap of an arm, ?i,min = min{A?A:i?A} ?A . This setting was already studied Cesa-Bianchi and Lugosi [2012], most recently in Combes et al. [2015], Kveton et al. [2015], where two different algorithms are used to tackle on one hand the case where the arms have independent rewards and on the other hand the general bounded case. The regret guaranties of the two algorithms are different and reflect that the independent case is easier. Another algorithm for the independent arms case based on Thompson Sampling was introduced in Komiyama et al. [2015]. One of the main objectives of this paper is to design a unique algorithm that can adapt to the covariance structure of the problem when prior information is available. The following notations will be used throughout the paper: given a matrix M (resp. vector v), its (i, j)th (resp. ith) coefficient is denoted by M (ij) (resp. v (i)). For a matrix M , the diagonal matrix with same diagonal as M is denoted by ?M . We denote by ?t the noise in the reward, i.e. ?t := Xt ? ?. We consider a subgaussian setting, in which we suppose that there is a positive semi-definite matrix C such that for all t ? 1, ?u ? Rd , E[eu

¿
?t
1
¿
] ? e2u
Cu
.

This is equivalent to the usual setting for bandits where we suppose ? that the individual arms are (i) subgaussian. Indeed if we have such a matrix C then each ?t is C (ii) -subgaussian. And under a subgaussian arms assumption, such a matrix always exists. This setting encompasses the case of bounded rewards. We call C a subgaussian covariance matrix of the noise (see appendix A of the

supplementary material). A good knowledge of C can simplify the problem greatly, as we will show. In the case of 1-subgaussian independent rewards, in which C can be chosen diagonal, a known lower bound d on the regret appearing in Combes et al. [2015] is ? log T , while Kveton et al. [2015] proves a dm log T lower bound in general. Our goal here is to investigate the spectrum of intermediate cases ? between these two settings, from the uninformed general case to the independent case in which one has much information on the relations between the arm rewards. We characterize the difficulty of the problem as a function of the subgaussian covariance matrix C. We suppose that we know a positive semi-definite matrix ? such that for all vectors v with positive coordinates, v ¿ Cv ? v ¿ ?v, property that we denote by C + ?. ? reflects the prior information available about the possible degree of independence of the arms. We will study algorithms that enjoy regret bounds as functions of ?. The matrix ? ? can be chosen such that all its coefficients are non-negative and verify for all i, j, ?(ij) ? ?(ii) ?(jj) . From now on, we suppose that it is the case. In the following, we will use t such that ?t = C 1/2 t and write for the reward: Xt = ? + C 1/2 t .

2

Lower bound

We first prove a lower bound on the regret of any algorithm, demonstrating the link between the subgaussian covariance matrix and the difficulty of the problem. It depends on the maximal off-diagonal (ij) correlation coefficient of the covariance matrix. This coefficient is ? = max{(i,j)?[d],i6=j} ? C . C (ii) C (jj) The bound is valid for consistent algorithms [Lai and Robbins, 1985], for which the regret on any problem verifies ERt = o(ta) as t ? +? for all a ¿ 0. Theorem 1. Suppose to simplify that d is a multiple of m. Then, for any ? ¿ 0, for any consistent algorithm, there is a problem with gaps ?, ?-subgaussian arms and correlation coefficients smaller than ? ? [0, 1] on which the regret is such that lim inf t?+?

ERt 2? 2 (d ? m) ? (1 + ?(m ? 1)) log t ?

This bound is a consequence of the classical result of Lai and Robbins [1985] for multi-armed bandits, applied to the problem of choosing one among d/m paths, each of which has m different successive edges (Figure 1). The rewards in the same path are correlated but the paths are independent. A complete proof can be found in appendix B.1 of the supplementary material. 2

Figure 1: Left: parallel paths problem. Right: regret of OLS-UCB as a function of m and ? in the parallel paths problem with 5 paths (average over 1000 runs).

3

OLS-UCB Algorithm and analysis

Faced with the combinatorial semi-bandit at stage t ? 1, the observations from t ? 1 stages form as many linear equations and the goal of an algorithm is to choose the best action. To find the action with the highest mean, we estimate the mean of all arms. This can be viewed as a regression problem. The design of our algorithm stems from this observation and is inspired by linear regression in the fixed design setting, similarly to what was done in the stochastic linear

bandit literature [Rusmevichientong and Tsitsiklis, 2010, Filippi et al., 2010]. There are many estimators for linear regression and we focus on the one that is simple enough and adaptive: Ordinary Least Squares (OLS). 3.1

Fixed design linear regression and OLS-UCB algorithm

For an action A ? A, let IA be the diagonal matrix with a 1 at line i if A(i) = 1 and 0 otherwise. For a matrix M , we also denote by MA the matrix IA M IA . At stage t, if all actions A1 , . . . , At were independent of the rewards, we would have observed a set of linear equations IA1 X1 = IA1 ? + IA1 ?1 .. . IAt?1 Xt?1 = IAt?1 ? + IAt?1 ?t?1 and we could use the OLS estimator to estimate ?, which is unbiased and has a known subgaussian constant controlling its variance. This is however not true in our online setting since the successive actions are not independent. At stage t, we define (i)

nt =

t?1 X

(ij)

I{i ? As }, nt

s=1

=

t?1 X

I{i ? As }I{j ? As } and Dt =

s=1

t?1 X

IAs ,

s=1

(i)

where nt is the number of times arm i has been pulled before stage t and Dt is a diagonal matrix of these numbers. The OLS estimator is, for an arm i ? [d], (i)

? ?t =

1

X

(i) nt s¡t:i?As

Xs(i) = ?(i) + (Dt?1

t?1 X

IAs C

1/2

s)(i) .

s=1

Then for all A ? A, A¿ (? ?t ? ?) in the fixed design setting has a subgaussian matrix equal to Pt?1 Dt?1 (s=1 CAs)Dt?1 . We get confidence intervals for the estimates and can use an upper confidence bound strategy [Lai and Robbins, 1985, Auer et al., 2002]. In the online learning setting the actions are not independent but we will show that using this estimator still leads to estimates that are well concentrated around ?, with confidence intervals given by the same subgaussian matrix. The algorithm OLS-UCB (Algorithm 1) results from an application of an upper confidence bound strategy with this estimator. We

now turn to an analysis of the regret of OLS-UCB. At any stage t ? 1 of the algorithm, let ?(ij) ?t = max{(i,j)?At ,i6=j} ? (ii) be the maximal off-diagonal correlation coefficient of ?At and ? ?(jj) let ? = max{t?[T]} ?t be the maximum up to stage T . 3

Algorithm 1 OLS-UCB. Require: Positive semi-definite matrix ?, real parameter ? ¿ 0. 1: Choose actions such that each arm is pulled at least one time. 2: loop: at stage t, 3: At = arg maxA A¿ ? ?t +qEt (A) p Pt?1 with Et (A) = 2f (t) A¿ Dt?1 (??? Dt + s=1 ?As)Dt?1 A. 4: Choose action At , observe IAt Xt . 5: Update ? ?t , Dt . 6: end loop

Theorem 2. The OLS-UCB algorithm with parameter ? ¿ 0 and f (t) = log t + (m + 2) log log t + m e 2 log(1 + ?) enjoys for all times T ? 1 the regret bound !

2 X ?(ii) log m E[RT] ?16f (T) 5(? + 1 ? ?) + 45?m ?i,min 1.6 i?[d]

8dm2 maxi {C (ii) }?max + + 4?max , ?2min where dxe stands for the smallest positive integer bigger than or equal to x. In particular, d0e = 1. This bound shows the transition between a general case with a 2

dm log T ?

regime and an independent

d log m log T ?

T case with a upper bound (we recall that the lower bound is of the order of d log ?). The weight of each case is given by the maximum correlation parameter ?. The parameter ? seems to be an artefact of the analysis and can in practice be taken very small or even equal to 0.

Figure 1 illustrates the regret of OLS-UCB on the parallel paths problem used to derive the lower bound. It shows a linear dependency in ? and supports the hypothesis that the true upper bound matches the lower bound with a dependency in m and ? of the form (1 + ?(m ? 1)). Corollary 1. The OLS-UCB algorithm with matrix ? and parameter ? ¿ 0 has a regret bounded as v ! u 2

u log m t (ii) + 45?m) . E[RT] ? O(dT log T max{? } 5(? + 1 ? ?) 1.6 i?[d] Proof. We write that the regret up to stage T is bounded by ?T for actions with gap smaller than some ? and bounded using theorem 2 for other actions (with ?min ? ?). Maximizing over ? then gives the result. 3.2

Comparison with other algorithms

Previous works supposed that the rewards of the individual arms are in [0, 1], which gives them a property. Hence we suppose (?i ? [d], C (ii) = 1/2) for our comparison.

1/2-subgaussian

In the independent case, our algorithm is the same as ESCB-2 from Combes et al. [2015], up to the ? d m log T parameter ?. That paper shows that ESCB-2 enjoys an O( ) upper bound but our analysis ? tighten it to O(d log

2

m log T ?

).

log T In the general (worst) case, Kveton et al. [2015] prove an O(dm ?) upper bound (which is tight) using CombUCB1, a UCB based algorithm

introduced in Chen et al. [2013] which at stage t uses q ? P (i) the exploration term 1.5 log t i?A 1/ nt . Our exploration term always verifies Et (A) ? q p P (i) f (t) i?A 1/ nt with f (t) ? log t (see section 3.6). Their exploration term is a worst-case confidence interval for the means. Their broader confidence intervals however have the desirable property that one can find the action that realizes the maximum index by solving a linear optimization problem, making their algorithm computationally efficient, quality that both ESCB and OLS-UCB are lacking.

4

None of the two former algorithms benefits from guaranties in the other regime. The regret of ESCB in the general possibly correlated case is unknown and the regret bound for CombUCB1 is not improved in the independent case. In contrast, OLS-UCB is adaptive in the sense that its performance gets better when more information is available on the independence of the arms. 3.3

Regret Decomposition (i)

?t Let Hi,t = {—? ?t ? ?(i) — ? 2m } and Ht = ?di=1 Hi,t . Ht is the event that at least one coordinate of ? ?t is far from the true mean. Let Gt = {A?¿ ? ? A?¿ ? ?t + Et (A?)} be the event that the estimate of the optimal action is below its true mean by a big margin. We decompose the regret according to these events:

RT ?
T X
?t I{Gt , Ht } +
t=1
T X
?t I{Gt } +
t=1
T X
?t I{Ht }
t=1

Events Gt and Ht are rare and lead to a finite regret (see below). We first simplify the regret due to Gt ? Ht and show that it is bounded by the "variance" term of the algorithm. Lemma 1. With the algorithm choosing at stage t the action At = arg maxA (A¿ ? ?t + Et (A)), we have ?t I{Gt , Ht } ? 2Et (At)I{?t ? Et (At)}. Proof in appendix B.2 of the supplementary material. Then the regret is cut into three terms, RT ? 2

T X
Et (At)I{?t ? 2Et (At)} +
T X t=1
t=1
?t I{Gt } +
T X
?t I{Ht } .
t=1

The three terms will be bounded as follows: ? The Ht term leads to a finite regret from a simple application of Hoeffding?s inequality. ? The Gt term leads

to a finite regret for a good choice of f (t). This is where we need to show that the exploration term of the algorithm gives a high probability upper confidence bound of the reward. ? The Et (At) term, or variance term, is the main source of the regret and is bounded using ideas similar to the ones used in existing works on semi-bandits. 3.4

Expected regret from Ht

PT Lemma 2. The expected regret due to the event Ht is E[t=1 ?t I{Ht }] ?

8dm2 maxi {C (ii) }?max ?2min

.

The proof uses Hoeffding?s inequality on the arm mean estimates and can be found in appendix B.2 of the supplementary material. 3.5

Expected regret from Gt

We want to bound the probability that the estimated reward for the optimal action is far from its mean. We show that it is sufficient to control a self-normalized sum and do it using arguments from Pe?a et al. [2008], or Abbasi-Yadkori et al. [2011] who applied them to linear bandits. The analysis also involves a peeling argument, as was done in one dimension by Garivier [2013] to bound a similar quantity. e Lemma 3. Let ?t ¿ 0. With f?(?t) = log(1/?t) + m log log t + m 2 log(1 + ?) and an algorithm q q P t?1 given by the exploration term Et (A) = 2f?(?t) A¿ Dt?1 (??? Dt + s=1 ?As)Dt?1 A , then the event Gt = {A?¿ ? ? A?¿ ? ?t + Et (A?)} verifies P{Gt } ? ?t . 1 ? With ?1 = 1 and ?t = t log 2 t for t ? 2, such that f (?t) = f (t), the regret due to Gt is finite in expectation, bounded by 4?max .

5

Proof. We use a peeling argument: let ? ¿ 0 and for a = (a1 , . . . , am) ? Nm , let Da ? [T] be a (i) subset of indices defined by (t ? Da ? ?i ? A? , (1 + ?)ai ? nt ¡ (1 + ?)ai +1). For any Bt ? R,

X ?¿

P A?¿ (? ? ? ?t) ? Bt ? P A (? ? ? ?t) ? Bt —t ? Da . a

The number of possible sets Da for t is bounded by (log t/ log(1 + ?))m , since each number of pulls

(i) nt for i ? A? is bounded by t. We now search a bound of the form P A?¿ (? ? ? ?t) ? Bt —t ? Da . Suppose t ? Da and let D be a positive definite diagonal matrix (that depends on a). Pt?1 Pt?1 2 Let St = s=1 IAs ?A? C 1/2 s , Vt = s=1 CAs ?A? and IVt +D () = 12 kSt k(Vt +D)?1 . Lemma 4. Let ?t ¿ 0 and let f?(?t) be a function of ?t . With a choice of D such that IA? D ?IA? ?C Dt for all t in Da ,

q

n o

P A?¿ (??? ?t)? 2f?(?t)A?¿ Dt?1 (??C Dt +Vt)Dt?1 A? t?Da ? P IVt +D ()?f?(?t)—t?Da . Proof in appendix B.2 of the supplementary material. The self-normalized sum IVt () is an interesting quantity for the following reason: exp(12 IVt ()) = Qt?1 maxu?Rd s=1 exp(u¿ IAs ?A? C 1/2 s ? u¿ CAs ?A? u). For a given u, the exponential is smaller that 1 in expectation, from the subgaussian hypothesis. The maximum of the expectation is then smaller

than 1. To control IVt (), we are however interested in the expectation of this maximum and cannot interchange max and E. The method of mixtures circumvents this difficulty: it provides an approximation of the maximum by integrating the exponential against a multivariate normal centered at the point Vt?1 St , where the maximum is attained. The integrals over u and can then be swapped by Fubini?s theorem to get an approximation of the expectation of the maximum using an integral of the expectations. Doing so leads to the following lemma, extracted from the proof of Theorem 1 of Abbasi-Yadkori et al. [2011]. Lemma 5.q Let D be a positive definite matrix that does not depend on t and det D Mt (D) = det(V exp(IVt +D ()). Then E[Mt (D)] ? 1. n t +D) o We rewrite P IVt +D () ? f?(?t) to introduce Mt (D), (n o P IVt +D () ? f?(?t)—t?Da = P Mt (D) ? p

1 det(Id + D?1/2 Vt D?1/2)

exp(f?(?t))t?Da

) .

The peeling lets us bound Vt . Let Da be the diagonal matrix with entry (i, i) equal to (1 + ?)ai for i ? A? and 0 elsewhere. Lemma 6. With D = ??C Da + I[d]A? , det(Id + D?1/2 Vt D?1/2) ? (1 +

1+? m ?)

.

The union bound on the sets Da and Markov?s inequality give

q q P A?¿ (? ? ? ?t) ? 2f?(?t) ?A?¿ ?C Dt?1 A? + A?¿ Dt?1 Vt Dt?1 A?

X 1 + ? ?m/2) exp(f?(?t))—t ? Da ? P Mt (D) ? (1 + ? Da

m log t 1 + ? m/2 ? (1 +) exp(?f?(?t)) log(1 + ?) ? For ? = e ? 1 and f?(?t) as in lemma 3, this is bounded by ?t . The result with ? instead of C is a consequence of C + ?. With ?1 = 1 and ?t = 1/(t log2 t) for t ? 2, the regret due to Gt is E[

T X

?t I{Gt }] ? ?max (1 +

t=1

T X t=2

6

1) ? 4?max . t log2 t

3.6

Bounding the variance term

The goal of this section is to bound Et (At) under ? the event {?t ? Et (At)}. Let ?t ? [0, 1] such that for q all i, j ? At with i 6= j, ?(ij) ? ?t ?(ii) ?(jj) . From the Cauchy-Schwartz inequality, (ij)

nt

(i) (j)

?

nt nt . Using these two inequalities, t?1 X

?1 A¿ t Dt (

?As)Dt?1 At

X n(ij) ?(ij) t

=

s=1
i,j?At
(i) (j) nt nt
? (1 ? ?t)
X ?(ii) i?At
(i) nt
s X
?(ii)
i?At
(i) nt
+ ?t (
)2 .

We recognize here the forms of the indexes used in Combes et al. [2015] for independent arms (left term) and Kveton et al. [2015] for general arms (right term). Using ?t ? Et (At) we get s X ?(ii) X ?(ii) ?2t ? (? + 1 ? ?t) + ?t ( )2 . (1) (i) (i) 8f (t) nt nt i?At
i?At

The strategy from here is to find events that must happen when (1) holds and to show that these events cannot happen very often. For positive integers j and t and for e ? {1, 2}, we define the set of arms (ii)
(i)
j in At that were pulled less than a given threshold: St,e = {i ? At , nt ? ?j,e 8f (t)?
ge (m,?t) }, ?2t j At . (St,e)j?0

0 with ge (m, ?t) to be stated later and (?i,e)i?1 a decreasing sequence. Let also St,e = is decreasing for the inclusion of sets and we impose limj?+? ?j,e = 0, such that there is an index j? j? with St,e = ?. We introduce another positive sequence (?j,e)j?0 and consider the events that j at least m?j,e arms in At are in the set St,e and that the same is false for k ¡ j, i.e. for t ? 1, j j k At,e = {—St,e — ? m?j,e ; ?k ¡ j, —St,e — ¡ m?k,e }. To avoid having some of these events being 0 impossible we choose (?j,e)j?0 decreasing. We also impose ?0,e = 1, such that —St,e — = m?0,e . j Let then At,e = ?+? j=1 At,e and At = At,1 ? At,2 . We will show that At must happen for (1) to be true. First, remark that under a condition on (?j,e)j?0 , At is a finite union of events, 0 Lemma 7. For e ? {1, 2}, if there exists j0,e such that ?j0,e ,e ? 1/m, then At,e = ?jj=1 Ajt,e . We now show that At is impossible by proving a contradiction in (1).

Lemma 8. Under the event At,1 , if there exists j0 such that ?j0 ,1 ? 1/m, then ? ? j0 X ?(ii) X m?2t ? ? ? ? j?1,1 j,1 j ,1 ? ¡ + 0 ?. (i) 8f (t)g1 (m, ?t) j=1 ?j,1 ?j0 ,1 nt i?At

P+? ? Under the event At,2 , if limj?+? ?j,2 / ?j,2 = 0 and j=1 s X
?(ii)
i?At
nt
(i)
?p

?j?1,2 ??j,2 ? ?j,2
exists, then
+? X ?j?1,2 ? ?j,2 . ? ?j,2 8f (t)g2 (m, ?t) j=1
m?t

A proof can be found in appendix B.2 of the supplementary material. To ensure that the conditions of these lemmas are fulfilled, we impose that (?i,1)i?0 and (?i,2)i?0 have limit 0 and ? that limj?+? ?j,2 / ?j,2 = 0. Let j0,1 be the smallest integer such that ?j0,1 ,1 ? 1/m. Let Pj0,1 ?j?1,1 ??j,1 P+? ? ?j ,1 ??j,2 ? l1 = ?j0,1 ,1 + j=1 and l2 = j=1 j?1,2 . Using the two last lemmas with (1), ?j,1 ?j,2 0,1
we get that if At is true, ?2t ?2t ¡ 8f (t) 8f (t)
(? + 1 ? ?t)
ml1 m2 l22 + ?t g1 (m, ?t) g2 (m, ?t)
.

Taking g1 (m, ?t) = 2(? + 1 ? ?t)ml1 and g2 (m, ?t) = 2?t m2 l22 , we get a contradiction. Hence with these choices At must happen. The regret bound will be obtained by a union bound on the events that form At . First suppose that all gaps are equal to the same ?. 7

Lemma 9. Let ? (ii)
d?j,e 8f (T) maxi {? m?j,e ?2
=
maxt?1 ?t .
}ge (m,?)
For j
N? , the event Ajt,e happens at most
?
times. (i)

Proof. Each time that Ajt,e happens, the counter of plays nt of at least m?je arms is incremented. (ii)
After
?j,e 8f (T) maxi {? ?2
}ge (m,?)
(i)
increments, an arm cannot verify the condition on nt any more.
There are d arms, so the event can happen at most d m?1j
e
?j,e 8f (T) maxi {?(ii) }ge (m,?) ?2
times.

If all gaps are equal to ?, an union bound for At gives ? ? j0,1 +? T X X X ?j,1 ?j,2 ? f (T) ? . E[d (? + 1 ? ?)l1 + ?ml22 ?I{Ht ? Gt }] ? 16 max{?(ii) } ? ? ? i?[d] j,1 t=1 j=1 j=1 j,2 The general case requires more involved manipulations but the result is similar and no new important idea is used. The following lemma is proved in appendix B.2 of the supplementary material: Lemma 10. Let ? (i) = max{t,i?At } ?t . The regret from the event Ht ? Gt is such that ? ? j0 T +? X X ?(ii) X X ?j,1 ?j,2 ? ?(? + 1 ? ?)l1 ?t I{Ht ? Gt }] ? 16f (T) E[. + ?ml22 ? ? ? i,min j,1 t=1 j=1 j=1 j,2 i?[d]

Finally we can find sequences (?j,1)j?1 , (?j,2)j?1 , (?j,1)j?0 and (?j,2)j?0 such that !

2 T X X ?(ii) log m (i) (i) 5(? + 1 ? ?) + 45? m ?I{Ht ? Gt }] ? 16f (T) E[?i,min 1.6 t=1 i?[d]

See appendix C of the supplementary material. In Combes et al. [2015], ?i,1 and ?i,1 were such ? that the log2 m term was replaced by m. Our choice is also applicable to their ESCB algorithm. Our use of geometric sequences is only optimal among sequences such that ?i,1 = ?i,1 for all i ? 1. It is unknown to us if one can do better. With this control of the variance term, we finally proved Theorem 2.

4

Conclusion

We defined a continuum of settings from the general to the independent arms cases which is suitable for the analysis of semi-bandit algorithms. We exhibited a lower bound scaling with a parameter that quantifies the particular setting in this continuum and proposed an algorithm inspired from linear regression with an upper bound that matches the lower bound up to a log2 m term. Finally we showed how to use tools from the linear bandits literature to analyse algorithms for the combinatorial bandit case that are based on linear regression. It would be interesting to estimate the subgaussian covariance matrix online to attain good regret bounds without prior knowledge. Also, our algorithm is not computationally efficient since it requires the computation of an argmax over the actions at each stage. It may be possible to compute this argmax less often and still keep the regret guaranty, as was done in Abbasi-Yadkori et al. [2011] and Combes et al. [2015]. On a broader scope, the inspiration from linear regression could lead to algorithms using different estimators, adapted to the structure of the problem. For example, the weighted least-square estimator is also unbiased and has smaller variance than OLS. Or one could take advantage of a sparse covariance matrix by using sparse estimators, as was done in the linear bandit case in Carpentier and Munos [2012].

Acknowledgements The authors would like to acknowledge funding from the ANR under grant number ANR-13-JS010004 as well as the Fondation Math?matiques Jacques Hadamard and EDF through the Program Gaspard Monge for Optimization and the Irsdi project Tecolere. 8

2 References

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Improved Algorithms for Linear Stochastic Bandits. Neural Information Processing Systems, pages 1?19, 2011. Jean-Yves Audibert, S?bastien Bubeck, and G?bor Lugosi. Regret in online combinatorial optimization. Mathematics of Operations Research, 39(1):31?45, 2013. Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47(2-3):235?256, 2002. Alexandra Carpentier and R?mi Munos. Bandit Theory meets Compressed Sensing for high dimensional Stochastic Linear Bandit.

Advances in Neural Information Processing Systems (NIPS), pages 251?259, 2012. Nicolo Cesa-Bianchi and G?bor Lugosi. Combinatorial bandits. Journal of Computer and System Sciences, 78 (5):1404?1422, 2012. Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. Proceedings of the 30th International Conference on Machine Learning (ICML), pages 151?159, 2013. Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial Bandits Revisited. Neural Information Processing Systems, pages 1?9, 2015. Sarah Filippi, Olivier Capp?, Aur?lien Garivier, and Csaba Szepesv?ri. Parametric Bandits: The Generalized Linear Case. Neural Information Processing Systems, pages 1?9, 2010. Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. IEEE/ACM Transactions on Networking, 20(5):1466?1478, 2012. Aur?lien Garivier. Informational confidence bounds for self-normalized averages and applications. 2013 IEEE Information Theory Workshop, ITW 2013, 2013. Junpei Komiyama, Junya Honda, and Hiroshi Nakagawa. Optimal Regret Analysis of Thompson Sampling in Stochastic Multi-armed Bandit Problem with Multiple Plays. Proceedings of the 32nd International Conference on Machine Learning, 2015. Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, 2015. Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. Advances in applied mathematics, 6(1):4?22, 1985. Victor H Pe?a, Tze Leung Lai, and Qi-Man Shao. Self-normalized processes: Limit theory and Statistical Applications. Springer Science & Business Media, 2008. Herbert Robbins. Some aspects of the sequential design of experiments. In Herbert Robbins Selected Papers, pages 169?177. Springer, 1985. Paat Rusmevichientong and John N. Tsitsiklis. Linearly Parameterized Bandits. Mathematics of Operations Research, (1985):1?40, 2010.

9