Parametric Bandits: The Generalized Linear Case

Authored by:

Csaba Szepesv?ri
Sarah Filippi
Olivier Cappe
Aur?lien Garivier

Abstract

We consider structured multi-armed bandit tasks in which the agent is guided by prior structural knowledge that can be exploited to efficiently select the optimal arm(s) in situations where the number of arms is large, or even infinite. We pro- pose a new optimistic, UCB-like, algorithm for non-linearly parameterized bandit problems using the Generalized Linear Model (GLM) framework. We analyze the regret of the proposed algorithm, termed GLM-UCB, obtaining results similar to those recently proved in the literature for the linear regression case. The analysis also highlights a key difficulty of the non-linear case which is solved in GLM-UCB by focusing on the reward space rather than on the parameter space. Moreover, as the actual efficiency of current parameterized bandit algorithms is often deceiving in practice, we provide an asymptotic argument leading to significantly faster convergence. Simulation studies on real data sets illustrate the performance and the robustness of the proposed GLM-UCB approach.

1 Paper Body

In the classical K-armed bandit problem, an agent selects at each time step one of the K arms and receives a reward that depends on the chosen action. The aim of the agent is to choose the sequence of arms to be played so as to maximize the cumulated reward. There is a fundamental trade-off between gathering experimental data about the reward distribution (exploration) and exploiting the arm which seems to be the most promising. In the basic multi-armed bandit problem, also called the independent bandits problem, the rewards are assumed to be random and distributed independently according to a probability distribution that is specific to each arm ?see [1, 2, 3, 4] and references therein. Recently, structured bandit problems in which the distributions of the rewards pertaining to each arm are connected by a common unknown parameter have received much attention [5, 6, 7, 8, 9]. This model is motivated by the many practical

applications where the number of arms is large, but the payoffs are interrelated. Up to know, two different models were studied in the literature along these lines. In one model, in each times step, a side-information, or context, is given to the agent first. The payoffs of the arms depend both on this side information and the index of the arm. Thus the optimal arm changes with the context [5, 6, 9]. In the second, simpler model, that we are also interested in here, there is no side-information, but the agent is given a model that describes the possible relations 1

between the arms? payoffs. In particular, in ?linear bandits? [10, 8, 11, 12], each arm a ? A is associated with some d-dimensional vector ma ? Rd known to the agent. The expected payoffs of the arms are given by the inner product of their associated vector and some fixed, but initially unknown parameter vector ?? . Thus, the expected payoff of arm a is m?a ?? , which is linear in ?? .1 In this article, we study a richer generalized linear model (GLM) in which the expectation of the reward conditionally to the action a is given by ?(m?a ??), where ? is a real-valued, non-linear function called the (inverse) link function. This generalization allows to consider a wider class of problems, and in particular cases where the rewards are counts or binary variables using, respectively, Poisson or logistic regression. Obviously, this situation is very common in the fields of marketing, social networking, web-mining (see example of Section 5.2 below) or clinical studies. Our first contribution is an ?optimistic? algorithm, termed GLM-UCB, inspired by the Upper Confidence Bound (UCB) approach [2]. GLM-UCB generalizes the algorithms studied by [10, 8, 12]. Our next contribution are finite-time bounds on the statistical performance of this algorithm. In particular, we show that the performance depends on the dimension of the parameter but not on the number of arms, a result that was previously known in the linear case. Interestingly, the GLM-UCB approach takes advantage of the particular structure of the parameter estimate of generalized linear models and operates only in the reward space. In contrast, the parameter-space confidence region approach adopted by [8, 12] appears to be harder to generalize to non-linear regression models. Our second contribution is a tuning method based on asymptotic arguments. This contribution addresses the poor empirical performance of the current algorithms that we have observed for small or moderate sample-sizes when these algorithms are tuned based on finite-sample bounds. The paper is organized as follows. The generalized linear bandit model is presented in Section 2, together with a brief survey of needed statistical results. Section 3 is devoted to the description of the GLM-UCB algorithm, which is compared to related approaches. Section 4 presents our regret bounds, as well as a discussion, based on asymptotic arguments, on the optimal tuning of the method. Section 5 reports the results of two experiments on real data sets. 2

Generalized Linear Bandits, Generalized Linear Models

We consider a structured bandit model with a finite, but possibly very large, number of arms. At each time t, the agent chooses an arm At from the set A (we shall denote the cardinality of A by K). The prior knowledge available to the agent consists of a collection of vectors {ma }a?A of features which are specific to each arm and a so-called (inverse) link function ? : R ? R. The generalized

linear bandit model investigated in this work is based on the assumption that the payoff Rt received at time t is conditionally independent of the past payoffs and choices and it satisfies (1) E [Rt — At] = ?(m?At ??) ,

for some unknown parameter vector ?? ? Rd . This framework generalizes the linear bandit model considered by [10, 8, 12]. Just like the linear bandit model builds on linear regression, our model capitalizes on the well-known statistical framework of Generalized Linear Models (GLMs). The advantage of this framework is that it allows to address various, specific reward structures widely found in applications. For example, when rewards are binary-valued, a suitable choice of ? is ?(x) = exp(x)/(1 + exp(x)), leading to the logistic regression model. For integer valued rewards, the choice ?(x) = exp(x) leads to the Poisson regression model. This can be easily extended to the case of multinomial (or polytomic) logistic regression, which is appropriate to model situations in which the rewards are associated with categorical variables. To keep this article self-contained, we briefly review the main properties of GLMs [13]. A univariate probability distribution is said to belong to a canonical exponential family if its density with respect to a reference measure is given by p? (r) = exp (r? ? b(?) + c(r)) ,

(2)

where ? is a real parameter, c(?) is a real function and the function b(?) is assumed to be twice continuously differentiable. This family contains the Gaussian and Gamma distributions when the reference measure is the Lebesgue measure and the Poisson and Bernoulli distributions when the 1

Throughout the paper we use the prime to denote transposition.

2

reference measure is the counting measure on the integers. For a random variable R with density ? defined in (2), E(R) = b(?) and Var(R) = ?b(?), where b? and ?b denote, respectively, the first and second derivatives of b. In addition, ?b(?) can also be shown to be equal to the Fisher information matrix for the parameter ?. The function b is thus strictly convex. Now, assume that, in addition to the response variable R, we have at hand a vector of covariates X ? Rd . The canonical GLM associated to (2) postulates that p? (r—x) = px? ? (r), where ? ? Rd is a vector of parameter. Denote by ? = b? the so-called inverse link function. From the properties of b, we know that ? is continuously differentiable, strictly increasing, and thus one-to-one. The maximum likelihood estimator ??t , based on observations (R1 , X1), . . . (Rt?1 , Xt?1), is defined as the maximizer of the function t?1 X

k=1

log p? (Rk —Xk) =

t?1 X

k=1

Rk Xk? ? ? b(Xk? ?) + c(Rk) ,

a strictly concave function in ?. Upon differentiating, we obtain that ??t is the unique solution of the following estimating equation 2

t?1 X

k=1

(Rk ? ?(Xk? ?)) Xk = 0 ,

(3)

? In practice, the solution of (3) may be found efficiently where we have used the fact that ? = b. using, for instance, Newton?s algorithm. A semi-parametric version of the above model is obtained by assuming only that E? [R—X] = ?(X ? ?) without (much) further assumptions on the conditional distribution of R given X. In this case, the estimator obtained by solving (3) is referred to as the maximum quasi-likelihood estimator. It is a remarkable fact that this estimator is consistent under very general assumptions as long as the Pt?1 design matrix k=1 Xk Xk? tends to infinity [14]. As we will see, this matrix also plays a crucial role in the algorithm that we propose for bandit optimization in the generalized linear bandit model. 3

The GLM-UCB Algorithm

According to (1), the agent receives, upon playing arm a, a random reward whose expected value is ?(m?a ??), where ?? ? ? is the unknown parameter. The parameter set ? is an arbitrary closed subset of Rd . Any arm with largest expected reward is called optimal. The aim of the agent is to quickly find an optimal arm in order to maximize the received rewards. The greedy action argmaxa?A ?(m?a ??t) may lead to an unreliable algorithm which does not sufficiently explore to guarantee the selection of an optimal arm. This issue can be addressed by resorting to an ?optimistic approach?. As described by [8, 12] in the linear case, an optimistic algorithm consists in selecting, at time t, the arm At = argmax max E? [Rt — At = a] s.t. k? ? ??t kMt ? ?(t) , a

?

(4)

where ? is an appropriate, ?slowly increasing? function, Mt =

t?1 X

mAk m?Ak

(5)

k=1

? is the design matrix corresponding to the first t ? 1 timesteps and kvkM = v ? M v denotes the matrix norm induced by the positive semidefinite matrix M . The region k? ? ??t kMt ? ?(t) is a confidence ellipsoid around the estimated parameter ??t . Generalizing this approach beyond the case of linear link functions looks challenging. In particular, in GLMs, the relevant confidence regions may have a more complicated geometry in the parameter space than simple ellipsoids. As a consequence, the benefits of this form of optimistic algorithms appears dubious.3 2

Here, and in what follows log denotes the natural logarithm. Note that maximizing ?(m?a ?) over a convex confidence region is equivalent to maximizing m?a ? over the same region since ? is strictly increasing. Thus, computationally, this approach is not more difficult than it is for the linear case. 3

3

An alternative approach consists in directly determining an upper confidence bound for the expected reward of each arm, thus choosing the action a that maximizes E??t [Rt — At = a] + ?(t)kma kM ?1 . t

4

In the linear case the two approaches lead to the same solution [12]. Interestingly, for non-linear bandits, the second approach looks more appropriate. In the rest of this section, we apply this second approach to the GLM bandit model defined in (1). According to (3), the maximum quasi-likelihood estimator of the parameter in the GLM is the unique solution of the estimating equation t?1

X Rk ? ?(m?Ak ??t) mAk = 0 ,

(6)

k=1

where A1 , . . . , At?1 denote the arms played so far and R1 , . . . , Rt?1 are the corresponding rewards. Pt?1 Let gt (?) = k=1 ?(m?Ak ?)mAk be the invertible function such that the estimated parameter ??t P ? satisfies gt (??t) = t?1 k=1 Rk mAk . Since ?t might be outside of the set of admissible parameters ?, ? we ?project it? to ?, to obtain ?t :

??t = argmin gt (?) ? gt (??t) ???

Mt?1

t?1

X

= argmin gt (?) ? Rk mAk ???

k=1

Mt?1

.

(7)

Note that if ??t ? ? (which is easy to check and which happened to hold always in the examples we dealt with) then we can let ??t = ??t . This is important since computing ??t is non-trivial and we can save this computation by this simple check. The proposed algorithm, GLM-UCB, is as follows: Algorithm 1 GLM-UCB 1: Input: {ma }a?A 2: Play actions a1 , . . . , ad , receive R1 , . . . , Rd . 3: for t ¿ d do 4: Estimate ??t according to (6) ? 5: if ??t ? ? let ??t = ??t else compute o n ?t according to (7) 6: Play the action At = argmaxa ?(m?a ??t) + ?(t)kma kM ?1 , receive Rt t 7: end for At time t, for each arm a, an upper bound ?(m?a ??t) + ?ta is computed, where the ?exploration bonus? ?ta = ?(t)kma kM ?1 is the product of two terms. The quantity ?(t) is a slowly increasing t function; we prove in Section 4 that ?(t) can be set to guarantee high-probability bounds on the expected regret (for the actual form used, see (8)). Note that the leading term of ?ta is kma kM ?1 t which decreases to zero as t increases. As we are mostly interested in the case when the number of arms K is much larger than the dimension d, the algorithm is simply initialized by playing actions a1 , . . . , ad such that the vectors ma1 . . . , mad form a basis of M = span(ma , a ? A). Without loss of generality, here and in what follows we assume that the dimension of M is equal to d. Then, by playing a1 , . . . , ad in the first d steps the agent ensures that Mt is invertible for all t. An alternative strategy would be to initialize M0 = ?0 I, where I is the d ? d identify matrix. 3.1 Discussion The purpose of this section is to discuss some properties of Algorithm 1, and in particular the interpretation of the role played by kma kM ?1 . t

Generalizing UCB The standard UCB algorithm for K arms [2] can be seen as a special case of GLM-UCB where the vectors of covariates associated with the arms form an orthogonal system and $?(x) = x$. Indeed, take $d = K$, $A = \{1, . . . , K\}$, define the vectors $\{m_a\}_{a?A}$ as the canonical basis $\{e_a\}_{a?A}$ of R_d, and take $? ? R_d$ the vector whose component $?_a$ is the expected reward for arm a. 4

Then, M_t is a diagonal matrix whose a-th diagonal element is the number $N_t(a)$ of times the a-th arm haspbeen played up to time t. Therefore, the exploration bonus in GLM-UCB is given by $?_a = ??_t(a) ?_{ta} = ?(t)/N_t(a)$. Moreover, the maximum quasi-likelihood estimator $??_t$ satisfies $R t P t?1 ?a = 1 I R$ is the empirical mean of the rewards received for all a ? A, where R k $\{A =a\} t t k=1 N_t(a)$ while playing arm a. Algorithm 1 then reduces to the familiar UCB algorithm. In this case, it is known that the expected cumulated regret can be controlled upon setting the slowly varying p function ? to $?(t) = 2 \log(t)$, assuming that the range of the rewards is bounded by one [2].

Generalizing linear bandits Obviously, setting $?(x) = x$, we obtain a linear bandit model. In this case, assuming that $? = R_d$, the algorithm will reduce to those described in the papers [8, 12]. In particular, the maximum quasi-likelihood estimator becomes the least-squares estimator and as noted earlier, the algorithm behaves identically to one which chooses the parameter optimistically within the confidence ellipsoid $\{? : k? ? ??_t kM_t ? ?(t)\}$.

Dependence in the Number of Arms In contrast to an algorithm such as UCB, Algorithm 1 does not need that all arms be played even once.4 To understand this phenomenon, observe that, 2 2 2 2 $(1 + kmA_t kM ?1)$ for as $M_{t+1} = M_t + m_{A_t} m?_{A_t}$, $kma kM ?1 = kma kM ?1 ? m?_a M_{t?1} mA_t t t t+1 a$ any arm a. Thus the exploration bonus $?_{t+1}$ decreases for all arms, except those which are exactly orthogonal to mA_t (in the $M_t?1$ metric). The decrease is most significant for arms that are colinear to mA_t. This explains why the regret bounds obtained in Theorems 1 and 2 below depend on d but not on K.

4

Theoretical analysis

In this section we first give our finite sample regret bounds and then show how the algorithm can be tuned based on asymptotic arguments. 4.1 Regret Bounds To quantify the performance of the GLM-UCB algorithm, we consider the cumulated (pseudo) regret defined as the expected difference between the optimal reward obtained by always playing an optimal arm and the reward received following the algorithm: RegretT =

T X t=1

$?(m?_{a?} ??) ? ?(m?_{A_t} ??)$.

For the sake of the analysis, in this section we shall assume that the following assumptions hold: Assumption 1. The link function $? : R ? R$ is continuously differentiable, Lipschitz with constant $k?$ and such that $c? = \inf ???,_{a?A} ?(m ? ?_a ?) ¿ 0$. For the logistic function $k? = 1/4$, while the value of $c?$ depends on $\sup???,_{a?A} —m?_a ?—$.

Assumption 2. The norm of covariates in $\{m_a : a ? A\}$ is bounded: there exists $c_m ¡ ?$ such that for all a ? A, $kma k2 ? c_m$. Finally, we make the

following assumption on the rewards: Assumption 3. There exists Rmax ¿ 0 such that for any t ? 1, 0 ? Rt ? Rmax holds a.s. Let ?t = Rt ? ?(m?At ??). For all t ? 1, it holds that E [?t —mAt , ?t?1 , . . . , mA2 , ?1 , mA1] = 0 a.s. As for the standard UCB algorithm, the regret can be analyzed in terms of the difference between the expected reward received playing an optimal arm and that of the best sub-optimal arm: ?(??) =

min

a:?(m?a ??)¡?(m?a? ??)

?(m?a? ??) ? ?(m?a ??) .

Theorem 1 establishes a high probability bound on the regret underlying using GLM-UCB with 2k? ?Rmax p ?(t) = 2d log(t) log(2 d T /?) , (8) c? 4

Of course, the linear bandit algorithms also share this property with our algorithm.

5

p where T is the fixed time horizon, ? = 3 + 2 log(1 + 2c2m /?0) and ?0 denotes the smallest Pd ? eigenvalue of i=1 mai mai , which by our previous assumption is positive. Theorem 1 (Problem Dependent Upper Bound). Let s = max(1, c2m /?0). Then, under Assumptions 1?3, for all T ? 1, the regret satisfies:

2 32?2 Rmax k?2 C d2 2d T 2 P RegretT ? (d + 1)Rmax + ? 1 ? ? with C = log [s T] log . ?(??) ? c2? Note that the above regret bound depends on the true value of ?? through ?(??). The following theorem provides an upper-bound of the regret independently of the ?? . Theorem 2 (Problem Independent Upper Bound). Let s = max(1, c2m /?0). Then, under Assumptions 1?3, for all T ? 1, the regret satisfies s

! 2dT 8Rmax k? ? P RegretT ? (d + 1)Rmax + Cd log [s T] T log ? 1 ? ? with C = . ? c? The proofs of Theorems 1?2 can be found in the supplementary material. The main idea is to use the explicit form of the estimator given by (6) to show that t?1

X

k

? mAk ?k ?1 . kmAt kM ?1 ?(m?At ??) ? ?(m?At ??t) ? t c? Mt k=1

Bounding the last term on the right-hand side is then carried out following the lines of [12]. 4.2 Asymptotic Upper Confidence Bound

Preliminary experiments carried out using the value of ?(t) defined equation (8), including the case where ? is the identity function ?i.e., using the algorithm described by [8, 12], revealed poor performance for moderate sample sizes. A look into the proof of the regret bound easily explains this observation as the mathematical involvement of the arguments is such that some approximations seem unavoidable, in particular several applications of the Cauchy-Schwarz inequality, leading to pessimistic confidence bounds. We provide here some asymptotic arguments that suggest to choose significantly smaller exploration bonuses, which will in turn be validated by the numerical experiments presented in Section 5. Consider the canonical GLM associated with an inverse link function ? and assume that the vectors of covariates X are drawn independently under a fixed distribution. This random design model would for instance

7

describe the situation when the arms are drawn randomly from a fixed distribution. Standard statistical arguments show that the Fisher information matrix pertaining to this model is given by $J = E[?(X ? ? ??)XX ?]$ and that the maximum likelihood estimate $??t$ is such D D that $t?1/2$ $(??t ? ??)$?? N $(0, J ?1)$, where ?? stands for convergence in distribution. Moreover, a.s. ? ?1 t Mt ?? ? where ? $= E[XX]$. Hence, using the delta-method and Slutsky?s lemma D kma k?1?1 $(?(m? ??t) ? ?(m? ??))$?? N $(0, ?(m ? ? ??)km? k2?1 km? k2 ?1)$. Mt

a
a
a
a ?
a J

The right-hand variance is smaller than $k? /c?$ as $J c? ?$. Hence, for any sampling distribution such that J and ? are positive definite and sufficiently large t and small ?,

q ?1 ? ? ? P kma kM ?1 $(?(ma ?t) ? ?(ma ??))$ ¿ $2k? /c? \log(1/?)$ t

is asymptotically bounded by ?. Based on the above asymptotic argument, we ppostulate that using p $?(t) = 2k? /c? \log(t)$, i.e., inflating the exploration bonus by a factor of $k? /c?$ compared to the usual UCB setting, is sufficient. This is the setting used in the simulations below. 5

Experiments

To the best of our knowledge, there is currently no public benchmark available to test bandit methods on real world data. On simulated data, the proposed method unsurprisingly outperforms its competitors when the data is indeed simulated from a well-specified generalized linear model. In order to evaluate the potential of the method in more challenging scenarios, we thus carried out two experiments using real world datasets. 6

5.1 Forest Cover Type Data In this first experiment, we test the performance of the proposed method on a toy problem using the ?Forest Cover Type dataset? from the UCI repository. The dataset (centered and normalized with constant covariate added, resulting in 11-dimensional vectors, ignoring all categorical variables) has been partitioned into $K = 32$ clusters using unsupervised k-means. The values of the response variable for the data points assigned to each cluster are viewed as the outcomes of an arm while the centroid of the cluster is taken as the 11-dimensional vector of covariates characteristic of the arm. To cast the problem into the logistic regression framework, each response variable is binarized by associating the first class (?Spruce/Fir?) to a response $R = 1$ and all other six classes to $R = 0$. The proportions of responses equal to 1 in each cluster (or, in other word, the expected reward associated with each arm) ranges from 0.354 to 0.992, while the proportion on the complete set of 581,012 data points is equal to 0.367. In effect, we try to locate as fast as possible the cluster that contains the maximal proportion of trees from a given species. We are faced with a 32-arm problem in a 11-dimensional space with binary rewards. Obviously, the logistic regression model is not satisfied, although we do expect

some regularity with respect to the position of the cluster?s centroid as the logistic regression trained on all data reaches a 0.293 misclassification rate.

2000
Regrett
1500
UCB GLM?UCB ??greedy
1000 500 0 0
1000
2000
3000
4000
5000 t
6000
7000
8000
6000
9000 10000
GLM?UCB UCB
4000 2000 0
2
4
6
8
10 arm a
12
14
16
18

Figure 1: Top: Regret of the UCB, GLM-UCB and the ?-greedy algorithms. Bottom: Frequencies of the 20 best arms draws using the UCB and GLM-UCB. We compare the performance of three algorithms. First, the GLM-UCB algorithm, with parameters tuned as indicated in Section 4.2. Second, the standard UCB algorithm that ignores the covariates. Third, an ?-greedy algorithm that performs logistic regression and plays the best estimated action, At = argmaxa ?(m?a ??t), with probability 1 ? ? (with ? = 0.1). We observe in the top graph of Figure 1 that the GLM-UCB algorithm achieves the smallest average regret by a large margin. When the parameter is well estimated, the greedy algorithm may find the best arm in little time and then leads to small regrets. However, the exploration/exploitation tradeoff is not correctly handled by the ?-greedy approach causing a large variability in the regret. The lower plot of Figure 1 shows the number of times each of the 20 best arms have been played by the UCB and GLM-UCB algorithms. The arms are sorted in decreasing order of expected reward. It can be observed that GML-UCB only plays a small subset of all possible arms, concentrating on the bests. This behavior is made possible by the predictive power of the covariates: by sharing information between arms,

it is possible to obtain sufficiently accurate predictions of the expected rewards of all actions, even for those that have never (or rarely) been played. 7

5.2 Internet Advertisement Data In this experiment, we used a large record of the activity of internet users provided by a major ISP. The original dataset logs the visits to a set of 1222 pages over a six days period corresponding to about 5.108 page visits. The dataset also contains a record of the users clicks on the ads that were presented on these pages. We worked with a subset of 208 ads and 3.105 users. The pages (ads) were partitioned in 10 (respectively, 8) categories using Latent Dirichlet Allocation [15] applied to their respective textual content (in the case of ads, the textual content was that of the page pointed to by the ad?s link). This second experiment is much more challenging, as the predictive power of the sole textual information turns out to be quite limited (for instance, Poisson regression trained on the entire data does not even correctly identify the best arm). The action space is composed of the 80 pairs of pages and ads categories: when a pair is chosen, it is presented to a group of 50 users, randomly selected from the database, and the reward is the number of recorded clicks. As the average reward is typically equal to 0.15, we use a logarithmic link function corresponding to Poisson regression. The vector of covariates for each pair is of dimension 19: it is composed of an intercept followed by the concatenation of two vectors of dimension 10 and 8 representing, respectively, the categories of the pages and the ads. In this problem, the covariate vectors do not span the entire space; to address this issue, it is sufficient to consider the pseudo-inverse of Mt instead of the inverse. On this data, we compared the GLM-UCB algorithm with the two alternatives described in Section 5.1. Figure 2 shows that GLM-UCB once again outperforms its competitors, even though the margin over UCB is now less remarkable. Given the rather limited predictive power of the covariates in this example, this is an encouraging illustration of the potential of techniques which use vectors of covariates in real-life applications.

Regret
3000 2000
UCB GLM?UCB ??greedy
1000 0 0
1000
2000
3000
4000
5000
t

Figure 2: Comparison of the regret of the UCB, GLM-UCB and the ?-greedy (? = 0.1) algorithm on the advertisement dataset. 6

Conclusions

We have introduced an approach that generalizes the linear regression model studied by [10, 8, 12]. As in the original UCB algorithm, the proposed GLM-UCB method operates directly in the reward space. We discussed how to tune the parameters of the algorithm to avoid exaggerated optimism, which would slow down learning. In the numerical simulations, the proposed algorithm was

shown to be competitive and sufficiently robust to tackle real-world problems. An interesting open problem (already challenging in the linear case) consists in tightening the theoretical results obtained so far in order to bridge the gap between the existing (pessimistic) confidence bounds and those suggested by the asymptotic arguments presented in Section 4.2, which have been shown to perform satisfactorily in practice. 8

2 References

[1] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics, 6(1):4?22, 1985. [2] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. Machine Learning, 47(2):235?256, 2002. [3] N. Cesa-Bianchi and G. Lugosi. Prediction, learning, and games. Cambridge Univ Pr, 2006. [4] J. Audibert, R. Munos, and Cs. Szepesv?ari. Tuning bandit algorithms in stochastic environments. Lecture Notes in Computer Science, 4754:150, 2007. [5] C.C. Wang, S.R. Kulkarni, and H.V. Poor. Bandit problems with side observations. IEEE Transactions on Automatic Control, 50(3):338?355, 2005. [6] J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. Advances in Neural Information Processing Systems, pages 817?824, 2008. [7] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. International Conference on Machine learning, pages 721?728, 2007. [8] V. Dani, T.P. Hayes, and S.M. Kakade. Stochastic linear optimization under bandit feedback. Conference on Learning Theory, 2008. [9] S.M. Kakade, S. Shalev-Shwartz, and A. Tewari. Efficient bandit algorithms for online multiclass prediction. In Proceedings of the 25th International Conference on Machine learning, pages 440?447. ACM, 2008. [10] P. Auer. Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3:397?422, 2002. [11] Y. Abbasi-Yadkori, A. Antos, and Cs. Szepesv?ari. Forced-exploration based algorithms for playing in stochastic linear bandits. In COLT Workshop on On-line Learning with Limited Feedback, 2009. [12] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. Mathematics of Operations Research, 35(2):395?411, 2010. [13] P. McCullagh and J. A. Nelder. Generalized Linear Models. Chapman and Hall, 1989. [14] K. Chen, I. Hu, and Z. Ying. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. Annals of Statistics, 27(4):1155?1163, 1999. [15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. Advances in Neural Information Processing Systems, 14:601?608, 2002. [16] V.H. De La Pena, M.J. Klass, and T.L. Lai. Self-normalized processes: exponential inequalities, moment bounds and iterated logarithm laws. Annals of Probability, 32(3):1902?1933, 2004. [17] P. Rusmevichientong and J.N. Tsitsiklis. Linearly parameterized bandits. Arxiv

preprint arXiv:0812.3465v2, 2008.
9