Linear Bayes Policy for Learning in Contextual-Bandits

José Antonio Martín H.[a,*], Ana M. Vargas[b]

[a] *Computer Architecture and Automation, Universidad Complutense de Madrid, Spain.*
[b] *Industrial Engineering, Business Administration and Statistics, Universidad Politécnica de Madrid, Spain.*

Abstract

Machine and Statistical Learning techniques are used in almost all online advertisement systems. The problem of discovering which content is more demanded (*e.g.* receive more clicks) can be modeled as a multi-armed bandit problem. Contextual bandits (*i.e.* bandits with covariates, side information or associative reinforcement learning) associate, to each specific content, several features that define the "context" in which it appears (*e.g.* user, web page, time, region). This problem can be studied in the stochastic/statistical setting by means of the conditional probability paradigm using the Bayes' theorem. However, for very large contextual information and/or real-time constraints, the exact calculation of the Bayes' rule is computationally infeasible. In this article, we present a method that is able to handle large contextual information for learning in contextual-bandits problems. This method was tested in the Challenge on Yahoo! dataset at ICML2012's Workshop "new Challenges for Exploration & Exploitation 3", obtaining the second place. Its basic exploration policy is deterministic in the sense that for the same input data (as a time-series) the same results are obtained. We address the deterministic exploration vs. exploitation issue, explaining the way in which the proposed method deterministically finds an effective dynamic trade-off based solely in the input-data, in contrast to other methods that use a random number generator.

Keywords: Contextual bandits, Online advertising, Recommender systems, One-to-one Marketing, Empirical Bayes

1. Introduction

In statistical decision theory, conditional probability through the Bayes' theorem provides an optimal decision rule. However, in sequential decision problems under stochastic conditions, when informed decisions must be computed sequentially with larger previously unknown side information, or the time to take decisions is critical, the exact calculation of the Bayes' optimal rule is computationally intractable or –even worse– non-computable. Nowadays, this is an increasingly common picture and so empirical approximations to Bayesian methods (Robbins, 1964) are finding great applicability in the Machine and Statistical Learning fields.

*address: Facultad de Informática, Universidad Complutense de Madrid, C. Prof. José García Santesmases, s/n., 28040, Madrid, (Spain). Tel.: +34 91.394.764, Fax: +34 91.394.7510

Email addresses: jamartinh@fdi.ucm.es (José Antonio Martín H.), ana.vargas@upm.es (Ana M. Vargas)

One of the most general and theoretically sound approaches in Machine Learning is the Inductive Inference Theory, where the learner tries to *predict* future events from the history (sequence) of past events (see Solomonoff, 1964 and Vovk et al., 2005), that is, "learning from observations". However, in general, there is no way to predict a sequence completely. Thus, one can divide a sequence in two parts: all that can be predicted, and all that cannot. Moreover, even for the predictable part some kind of search or trial-error process is often required. This effort is what is called *exploration* and deciding when one should explore or just use the current knowledge to make an educated guess (*i.e. exploit* current knowledge) is the so called *exploration vs. exploitation problem* (see Holland, 1975, March, 1991 or Kaelbling et al., 1996). The balance, ratio or proportion between these two opposed alternatives is called *the exploration / exploitation trade-off* and is performed following an arbitrary (ad hoc) algorithm referred to as the *exploration policy*.

In this article, we present a (empirical) Bayes-like method for learning in contextual-bandits problems. This method is able to take effective advantage of large contextual information in an efficient manner. In addition, one key-feature is the autonomous exploration / exploitation trade-off that it achieves deterministically, based solely in the input-data, in contrast to other methods that use a random number generator.

1.1. Bandit problems and the exploration vs. exploitation problem

The *multi-armed bandit problem* (MAB) is the sequential decision task where an agent (gambler or player) must decide or choose (pull) a set of actions (arms) to take at each time step, by following some informed strategy (a policy). For each chosen action, the agent receives a corresponding numerical *payoff* (reward) following an unknown probability distribution that may evolve in time for each different arm. Then, the agent can use such payoffs to improve its selection strategy to decide the choice of actions in the future to maximize the cumulative payoff in the long-run. Therefore, this becomes the problem of estimating the payoff-probability of each arm (over time).

A standard way of analyzing this maximization problem is to define it in terms of the minimization of the loss or *regret* with respect to the *optimal-policy* that always plays the best arm $a^*(t)$ at trial t (Auer et al., 2002). Hence, a natural measure of optimality in terms of the regret $R_{\mathsf{A}}(T)$ can be expressed in the following way:

$$R_{\mathsf{A}}(T) \quad \stackrel{\text{def}}{=} \quad \underbrace{\mathbf{E}\left[\sum_{t=1}^{T} r(t, a^*)\right]}_{\text{optimal}} - \underbrace{\mathbf{E}\left[\sum_{t=1}^{T} r(t, a)\right]}_{\text{current}}, \tag{1}$$

where A is the current algorithm and $r(t, a)$ the payoff obtained by playing arm a at trial t.

The multi-armed bandit problem was observed and studied by Robbins (1952), as the problem of sequential design of experiments (Wald, 1947), and extensively studied in statistics by Berry & Fristedt (1986). It was formalized and solved optimally by Gittins et al. (1989) for the special case in which the payoffs of all the arms are independent and that only one arm may evolve at each play.

The exploration vs. exploitation dilemma is what makes the MAB especially useful in several disciplines (Berry & Fristedt, 1986; Jun, 2004; Audibert et al., 2009; Bubeck & Cesa-Bianchi, 2012). That is, finding the right proportion between these two opposed "intentions":

Table 1: Some examples of equivalent terms to exploration and exploitation that are used in different fields.

Area or Discipline	Exploration	vs.	Exploitation
Sequential decision making	exploration	vs.	exploitation
Compressed sensing	sensor-reading	vs.	signal-reconstruction
Statistics and Machine Learning	memorizing data	vs.	generalizing
Curve-fitting	acquire-points	vs.	interpolation
Economics	risk-taking	vs.	risk-avoiding
Finance	investing	vs.	saving
Marketing	diversification/proliferation	vs.	concentration strategy
Medicine	experimental treatments	vs.	safety and efficacy
Data-compression	store-data	vs.	space-savings

1. Exploit the current knowledge to guess the best choice.
2. Explore an unknown or suboptimal choice to improve the knowledge about the problem (when possible).

Intuitively, your first impulse is to minimize exploration in order to increase the chance of getting a higher payoff, or –conversely– experience a low regret. However, which is the minimum exploration rate to minimize the regret in the long-run? Table 1 show us a subtle clue! The last row gives the relation between the exploration vs. exploitation problem with the data-compression one, for which we know that it is non-computable, *i.e.* , no general lossless compression method may exist.

This follows directly from the non-computability of Kolmogorov's complexity $K(s)$ of a string s. In the general case, we can't encode any sequence S_φ of length $\ell(S_\varphi)$ in a shorter sequence S_ϵ of length $\ell(S_\epsilon) < \ell(S_\varphi)$.

Now, since any *optimal* run of a sequential decision problem φ defines (obviously) a sequence of decisions S^*_φ of length $\ell(S_\varphi)^*$, then we cannot, in general, find a shorter sequence S_ϵ of length $\ell(S_\epsilon) < \ell(S^*_\varphi)$ that would predict S^*_φ (for any non-trivial S_φ).

Therefore, in general, a shorter sequence that specifies an optimal exploration / exploitation trade-off that serves to predict the optimal sequence of actions does not exists. Otherwise, we would be able to create a universal lossless compression program by encoding particular playing sequences as strings.

This tell us that there are tasks in which the optimal solution is a pure exploration approach since there will be problems in which learning (and so prediction) is impossible at all. However, despite this bad news, in a sense, this is a full employment theorem for bandits, and so it is possible to find suboptimal exploration policies that significantly improve learning.

1.2. Contextual Bandits and Online advertising/recommender systems

Nowadays, Machine Learning and statistical techniques are used in almost all online advertisement and recommendation systems (Konstan & Ried, 2012; Agarwal et al., 2013). The problem of discovering which content is more demanded (*e.g.* receive more clicks), or which product is more likely to be consumed if displayed in an online advertisement system, can be modeled mathematically as a multi-armed bandit problem.

In online advertising, the *click-through rate* (CTR) is an index used to measure purchase propensity. This index is calculated as the proportion obtained by dividing the number of clicks

received by an advertising-banner by the number of its impressions or displays (Agarwal et al., 2009; Wang et al., 2011). From here, a common approach is to model online advertising as a multi-armed bandit problem for maximizing the CTR of the repeated interaction cycle whereby the system selects an article (arm) from a pool, recommends it by displaying the article to a particular user (pull the arm) and then observes whether the user clicked or not the recommended article (get the payoff or reward).

The Contextual bandits model, also known as bandits with covariates, side information, associative bandits or associative reinforcement learning (Langford & Zhang, 2007; Li et al., 2010), or simply the reinforcement learning case when there are multiple states but reinforcement is immediate (Kaelbling et al., 1996), is a natural extension of the multi-armed bandit problem. Contextual bandits incorporates additional information (context) to the decision making process. The assumption is that the payoff obtained by playing an arm, is –up to some degree if not totally– dependent on such contextual information (*i.e.* a covariate). This kind of problem appears to have wider applicability in practice, since problems that can be solved optimally without considering contextual information are not so common (Langford & Zhang, 2007). For example, feature-based recommender systems in general (Weng & Liu, 2004), and particularly news recommendation systems (Li et al., 2010; Liu et al., 2010) can be modeled naturally as contextual bandits.

Following the terminology of Li et al. (with some minor variations): a contextual-bandit algorithm A proceeds in discrete trials $t = 1, 2, 3, \ldots T$. At each trial t:

1. The algorithm observes a set $A(t)$ of arms (e.g. articles, options, choices) and a features vector $\mathbf{x}(t)$ (the *context*).
2. Based on observed payoffs from previous trials, A chooses an arm $a(t) \in \mathcal{A}(t)$, and receives payoff $r(t, a)$ whose expectation may depend on both; the context $\mathbf{x}(t)$ and the arm $a(t)$.
3. The algorithm then improves its selection rule (policy) from the tuple: context, arm and reward; $\mathbf{x}(t)$, $a(t)$ and $r(t, a)$ respectively.

2. Materials and Methods

The current algorithm was developed and tested in the "Exploration and Exploitation 3 Challenge"[1]. Hence, a good opportunity to describe the proposed algorithm in an applied form is to take advantage of this challenge, i.e., serving (recommending) news articles on a web site. Here, the quantity to be optimized is the overall CTR[2].

The dataset is a collection of user visits to the Yahoo! Front Page Today Module obtained from the "User Click Log Dataset" (Yahoo! Academic Relations, 2012). Each row represents a user visit, which is composed by the recommended article, a value of 0 or 1 indicating if the user clicked the recommended article, a series of features and a series of available articles to recommend. Each user visit contains a 136-dimensional binary features vector. Feature IDs take integer values in $\{1, 2, ..., 136\}$. Feature #1 is the constant (always 1) feature, and features #2-136 corresponds to other user information such as age, gender, behavior-targeting features, etc. Some user features are not present, since not all users are logged into Yahoo! when they visited the front page.

[1] New Challenges for Exploration and Exploitation workshop at the International Conference on Machine Learning - June 26–July 1, 2012 - Edinburgh, Scotland. `http://icml.cc/2012/`

[2] `https://explochallenge.inria.fr/the-evaluation-process/`

A unique property of this data set is that the displayed article was chosen uniformly at random from the candidate article pool. Therefore, one can use an unbiased off-line evaluation method (Li et al., 2011) to compare bandit algorithms in a reliable way. An example row (user-visit) from this dataset is shown next:

"1317513291 id-560620 0 |user 1 9 11 13 23 16 18 17 19 15 43 14 39 30 66 50 27 104 20 |*id*-552077 |*id*-555224 |*id*-555528 ... |*id*-565822", where the following information is present:

- time-stamp: 1317513291.
- displayed article id: id-560620.
- user-action (0 for no-click and 1 for click): 0.
- string "user" indicates the start of user features.
- user features: the IDs of the nonzero features are listed.
- The set of available articles for recommendation for each user visit is the set of articles that appear in that line of data.

The proposed method handles contextual information (user/web-page features) in the form of a vector of binary features $\mathbf{x}(t) \in \{0, 1\}^*$, provided at each trial t, when the algorithm is confronted to the selection of one article from a set $\mathcal{A}$ of possible items to recommend.

The system operation has three main parts:

1. Item recommendation: the algorithm, at trial t, has to select one article $a(t)$ to recommend from a list $\mathcal{A}(t)$ of available items, taking advantage (if possible) of the available features $\mathbf{x}(t)$ (the context).
2. "Preference-value" computation (p_value): this part computes a preference-value for an article a given the context $\mathbf{x}(t)$ at trial t, i.e., $v = p_v(a, \mathbf{x}, t)$. The computed value is used to produce a ranking over the set $\mathcal{A}(t)$ in order to select the item with a higher "preference-value".
3. Policy update: once the selected article is recommended (and if it coincides with the data, see Li et al., 2010 for evaluation methodology), a feedback in the form of a binary number $r \in \{0, 1\}$ is received and it is used to update the necessary statistical information required by the "Preference-values" computation.

2.1. Non-contextual bandits

The most simple algorithm is to forget the context and just select always at trial t the article $a(t)$ which historically received more clicks $r = 1$. This algorithm can be implemented just by having a counter ($clicks[a]$) for each article a.

$$a(t) = \underset{a}{\operatorname{argmax}} \Big(clicks[a] \Big). \tag{2}$$

However, if different articles are recommended (and selected) in different proportions, this criterion is unfair since an article (a_1) recommended (and selected), say, on 1000 trials having received only 10 clicks will be preferred to an article (a_2) that have been recommended (and selected) only on 10 trials but received 9 clicks. It is of common sense to feel a preference for the a_2 article.

2.1.1. Naïve approach II

A second approach is to maintain also a counter ($selections[a]$) of the quantity of trials in which each article have been recommended (and selected), so that a proportion $clicks[a]/selections[a]$ could be calculated and used as the preference criterion. A key-advantage of this second (proportion-based) approach is that it allows to "learn" preferences even with a very different selection rate for each article, and so it creates room for balancing exploration and exploitation.

$$P_a = \frac{clicks[a]}{selections[a]}, \tag{3}$$

$$a(t) = \underset{a}{\operatorname{argmax}} \left(P_a \right). \tag{4}$$

This approach, although extremely basic, will perform well under the assumptions that: (1) the preferences for the recommended news are universal across all the users of the web site and (2) that a well enough exploration / exploitation trade-off is used.

Condition 1 is extremely restrictive and it may only happen in very specialized contexts in which additional contextual information is redundant, i.e., a non-contextual bandit (a simple k-armed bandit), which is not the current case. Condition 2 can be addressed in many ways since there are many alternatives and combinations between them, e.g., ϵ-greedy, soft-max, UCB (see Kuleshov & Precup 2010 for a comparison of some).

Here we simply use "optimistic initial values" (see Sutton & Barto, 1998, chap. 2.7) to force exploration, which is a simpler approach to the key-idea of "optimism in front of uncertainty" implemented in UCB-like algorithms. This method can be implemented simply by assuming, as a starting point, that every article has been selected and clicked one time, i.e., a proportion of 1:

$$initial\ values \begin{cases} clicks[a] & = 1, \\ selections[a] & = 1, \end{cases} \tag{5}$$

for any non-previously selected articles a.

This exploration strategy works in the following way, at trial 1 it selects the first available article a_1 since all non-yet recommended articles have preference 1 at starting point; and continue to recommend a_1 until a no click event occur in which a case the preference of a_1 will be less than 1. Then at the next trial, the next article with a preference of 1 is recommended. This cycle continues until all articles adapt its preference estimate very close to the true click-rate of every article.

Some optimizations can be made in order to avoid the computation of the proportion (3) at every query. For this purpose, it is just needed to maintain the current proportion $P[a]$ for each article a continuously updated:

$$P_a(t+1) = P_a(t) + \frac{r(t) - P_a(t)}{selections[a] + 1}, \tag{6}$$

where $r(t) = 1$ indicates that article a was clicked at trial t and $r(t) = 0$ if not. Note that equations (3) and (6) compute exactly the same value, and that (6) is equivalent to the pursuit method used in single-step temporal difference equation of reinforcement learning:

$$\mu(t+1) = \mu(t) + \alpha \left[r - \mu(t) \right], \tag{7}$$

where α is known as the learning rate parameter and r is the target to be learned (in this case 0 or 1).

2.2. Contextual bandits: naïve approach III

Let us now see how to incorporate in a useful way the available contextual information $\mathbf{x}(t)$. The first key-idea is quite simple:

Let us assume that each context $\mathbf{x}(t)$ define some characteristic features of a group of users; such as time-zone, country, language, previous navigation history and even direct knowledge about the kind of news they prefer to read. Hence, there would be some contexts (user-groups from now on) $\mathbf{x}(t)$ that are more likely to click on certain articles than on others. Here, the assumption is that each binary feature gives information of a particular fact. Thus, each feature should be treated as positive evidence that a visitor belongs to a certain user-group, that is, the features are independent and non-mutually exclusive. Hence, a user-group is defined by a set of features that individual users may have in common, but not that must have in common. For example, a user-group interested in sports may have interest in baseball OR football OR tennis, instead of being defined as having interest in baseball AND football AND tennis. In this case, there will be preference for a sports related news when there is a preference for baseball OR football and, more importantly, this preference will be maximized if the visitor shows preference for baseball AND football.

Hence, the visitor preference for a particular article can be measured simply by adding up all the individual preferences for each feature $\mathbf{x}_i$: $P_{a,i}$, i.e., a preference (feature) is specified when the i^{th} element of $\mathbf{x}_i = 1$ and so on:

In this case, $clicks[a][i]$ accumulates clicks for article a when $\mathbf{x}_i(t) = 1$ (*i.e.* , the feature is present) and $r(t) = 1$, while $selections[a][i]$ does the same independent of the value of $r(t)$:

$$clicks[a][i] = \sum_t \mathbf{x}_i(t) r(t); \text{ for } a(t) = a \tag{8}$$

$$selections[a][i] = \sum_t \mathbf{x}_i(t); \text{ for } a(t) = a. \tag{9}$$

Hence the incrementally calculated proportion is:

$$P_{a,i}(t+1) = P_{a,i}(t) + \mathbf{x}_i(t) \frac{r(t) - P_{a,i}(t)}{1 + selections[a][i]}. \tag{10}$$

And the recommended article $a(t)$ at trial t is:

$$a(t) = \underset{a}{\text{argmax}} \left(\frac{\sum_{i \neq 0} clicks[a][i]}{\sum_{i \neq 0} selections[a][i]} \right), \tag{11}$$

or alternatively:

$$a(t) = \underset{a}{\text{argmax}} \left(\sum_{i \neq 0} P_{a,i}(t) \right), \tag{12}$$

$$\text{where } P_{a,i} = \left(\frac{clicks[a][i]}{selections[a][i]} \right). \tag{13}$$

Table 2: Which feature should indicate or predict in a mayor degree a user-click on article a?. Consider a feature i whose click-rate for article a is CTR(i, a) and whose overall click-rate is CTR(i)

$CTR(i, a)$	$CTR(i)$	Feature evaluation
High	Low	almost all clicks related to feature i are going to article a, *i.e.* $P(click_a\|i) \to 1$
Low	High	almost no click related to feature i is going to article a, *i.e.* $P(click_a\|i) \to 0$
High	High	ambiguous
Low	Low	ambiguous

2.3. Contextual bandits: a Linear Bayes' Method

Now, an extension to the last preference measure is derived from an intuitive observation. Instead of basing the preference in the simple summation of the proportions $\sum_i P_{a,i}(t)$, between clicks to a specific article a by some user-group and the number of times this article has been recommended to this user-group, what about if we find a way of determining which specific features, and so which specific proportion $P_{a,i}(t)$ should contribute in a mayor degree to the overall sum?

Let us define the overall click-rate for a feature i as:

$$P_i(t+1) = P_i(t) + \mathbf{x}_i(t)\frac{r(t) - P_i(t)}{1 + \sum_a selections[a][i]}. \tag{14}$$

Now, which feature should indicate or predict in a mayor degree a user-click? Table 2 shows the information given by a feature i. As we can observe, the most informative cases occur when the estimated values of $P_{a,i}$ and P_i are very different, while the situation in which these to values are almost equal yields the most uncertain scenario.

Also, what role plays the overall click-rate $P_a(t)$ of article a in this puzzle? *I.e*, a feature i whose click-rate for article a is high and whose overall click-rate is also high but the overall click-rate $P_a(t)$ of article a is high/low? In the case when a feature does not significantly influence the posterior probability $P(click_a|i)$, then the information is contained exclusively in the estimated prior $P_a(t)$. Therefore, to create a weighted average, the idea is just weighting each $P_{a,i}(t)$ by $P_a(t)/P_i(t)$:

$$a(t) = \underset{a}{\operatorname{argmax}}\left(\sum_{i\neq 0} P_{a,i}(t)\frac{P_a(t)}{P_i(t)}\right), \tag{15}$$

where $P_a(t)$ can go out the summation because it remains constant.

From this, we can make the following equivalences with the standard notation:

$$P(\mathbf{x}_i|a) \approx P_{a,i}(t) \tag{16}$$

$$P(\mathbf{x}_i) \approx P_i(t), \tag{17}$$

$$P(a) \approx P_a(t), \tag{18}$$

$$P(a|\cup \mathbf{x}_i) \approx \sum_{i\neq 0}\frac{P(\mathbf{x}_i|a)}{P(\mathbf{x}_i)}P(a) \tag{19}$$

So, the following three equations are equivalent:

$$a(t) = \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{P_{a,i}(t)}{P_i(t)} P_a(t) \right), \tag{20}$$

$$a(t) = \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{P(\mathbf{x}_i(t)|a)}{P(\mathbf{x}_i(t))} P(a) \right), \tag{21}$$

$$a(t) = \underset{a}{\operatorname{argmax}} \Big(P(a| \cup \mathbf{x}_i(t)) \Big). \tag{22}$$

Therefore the applied preference selection is a linear approximation to the Bayes' rule, i.e., for the union of the informative events. Finally, a slight variation is to include an additional term $P(\cap \mathbf{x}_i|a)$ defined as:

$$P(\cap \mathbf{x}_i|a) = \prod_{i \neq 0} P(\mathbf{x}_i|a), \tag{23}$$

that expresses the joint probability of all posteriors $P(\mathbf{x}_i|a)$, assuming independence. Equation 23 is known as the naïve Bayes method. Hence, the final recommendations of the presented algorithm are done in the following way:

$$a(t) = \underset{a}{\operatorname{argmax}} \Big(P(a| \cup \mathbf{x}_i(t)) \; P(\cap \mathbf{x}_i(t)|a) \Big). \tag{24}$$

2.4. Exploration vs. Exploitation: does exactly what it says on the tin

The most important feature for any algorithm designed to solve a multi-armed bandit problem, being contextual or not, is the exploration policy. In Section 2.1.1 it was described how optimistic initial values induce an implicit exploration that converge to near optimal click-rates under benevolent conditions. However, by applying the selection rule (24) the exploration policy becomes non-trial. In this sense, how does this algorithm explore? How does it balance exploration and exploitation?

It should be emphasized the importance of finding simple and intuitive explanations of effective exploration / exploitation techniques, for many obvious reasons. For instance, Vermorel & Mohri (2005) and later Kuleshov & Precup (2010) have shown that very simple techniques such as ϵ-greedy and softmax perform extremely well when compared to more elaborate and theoretically-regret-guaranteed techniques, which are far from being intuitive for the uninitiated despite their simple and elegant key-ideas, but also that the situation can vary significantly from one domain to another.

Some of the most used policies that has succeeded in practical applications are pure-exploration, pure-exploitation, ϵ-greedy likes, Bolstzmann/softmax likes and interval-estimation (Powell, 2010). These policies try to balance exploration and exploitation focusing on minimizing regret (*i.e.* safer-exploration or risk-averse approach), without paying much attention to the value of the information to be obtained. For instance, Sutton & Barto (1998) comment that the softmax algorithm, per se, tries to implement a softer dichotomy (or remove it at all). However, in strict sense, it will actually end up at each trial selecting between the best choice and suboptimal ones, based on a trade-off defined by the "temperature parameter" of the Boltzmann-Gibs distribution. However, this is another kind of dichotomy after all but of a different nature: the payoff that is gained with a safer-exploration will have, as a price, the information loss of the exploratory trial. One

exception is the interval estimation policy, which increase the priority of choices with higher uncertainty. That is, it assigns a value to information gathering actions.

A totally different kind of approach is to explicitly choose an option to gain as much information as possible from every exploratory trial. This approach has been studied in depth by Frazier et al. (2008) and Powell (2010) as the knowledge gradient (KG) method. The KG is an *explicit* method to maximize the marginal value of the information obtained by exploring an alternative, that is, it "identifies the measurement which will do the most to identify the best choice" (Powell & Ryzhov, 2012).

The proposed Bayesian policy, however, does not approach explicitly the problem of maximizing information gain. However, the algorithm achieves an implicit dynamic balance between exploration and exploitation that maximizes, in a sense, an information gain criterion.

In order to study such mechanism, let us now return to the news recommendation context. Let us suppose we want to create a special exploration procedure that tries to gain as much information as possible from every exploratory trial.

A simple intuitive but powerful rule could be the next one: recommend an article a such that the following *concurrent* conditions hold:

1. Article has been little selected by the current user-group $\mathbf{x}(t)$.
2. Article has been highly clicked by all the user-groups.
3. Despite its low inner-group selection rate it shows some clicks for the current user-group.
4. The article has a high prior, i.e., $P(a)$ is high.
5. The user-group has a low click-rate.

Thus, any unexplored article or a new one just arriving to the web site (a truly news article) will have maximal priority to be explored by using the above list of desired conditions. However, to understand the complete picture, let us observe the following equations which are equivalent to (21):

$$a(t) = \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{P(\mathbf{x}_i|a)}{P(\mathbf{x}_i)} P(a) \right), \tag{25}$$

$$= \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{C_{ai}/S_{ai}}{P(\mathbf{x}_i)} P(a) \right), \tag{26}$$

$$= \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{C_{ai}/S_{ai}}{C_i/S_i} P(a) \right), \tag{27}$$

$$= \underset{a}{\operatorname{argmax}} \left(\sum_{i \neq 0} \frac{C_{ai}\ S_i\ C_a}{C_i\ S_{ai}\ S_a} \right), \tag{28}$$

where $C_{ai} = clicks[a][i]$, $S_{ai} = selections[a][i]$, $C_i = \sum_a clicks[a][i]$ and $S_i = \sum_a selections[a][i]$.

We can see that, indeed, these equations maximize all the conditions above, e.g. the term S_i as well as C_{ai} and $P(a)$ are directly proportional to the selection preference (conditions 2,3,4), however C_i and S_{ai} are inversely proportional to the selection preference (conditions 1 and 5).

Therefore, in the presented algorithm converge the two opposed "intentions" in just one selection rule; it explores and exploits depending on the particular context.

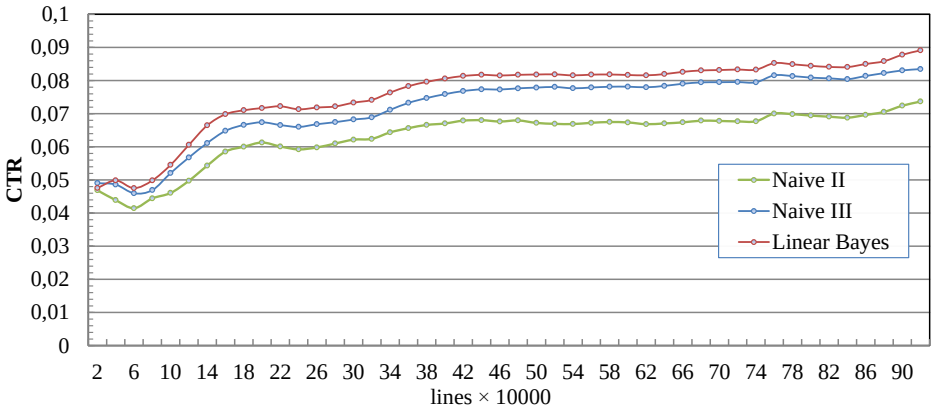

Figure 1: Performance comparison in CTR vs. processed lines between the Naïve II, III and Linear Bayes method over the first 9200000 rows on the R6B - Yahoo! Front Page Today Module User Click Log Dataset, version 2.0 (300 MB).

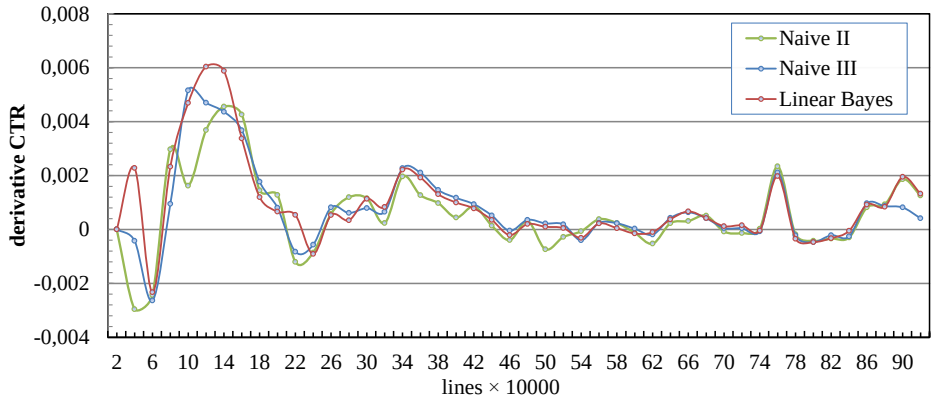

Figure 2: Performance comparison in derivative CTR vs. processed lines between the Naïve II, III and Linear Bayes method over the first 9200000 rows on the R6B - Yahoo! Front Page Today Module User Click Log Dataset, version 2.0 (300 MB).

3. Results

The presented method obtained the second place in the Challenge on Yahoo! dataset at ICML 2012 Workshop new Challenges for Exploration & Exploitation 3. Here the details of the results of that benchmark are presented. Here we present the details of the results obtained by the proposed algorithm on the "R6B - Yahoo! Front Page Today Module User Click Log Dataset, version 2.0 (300 MB)"

Figure 1 shows the different performances obtained for the naïve II, naïve III and Linear Bayes approaches on the first 9200000 rows of the dataset. These results show that the Bayes' method outperforms a simple contextual bandit and that a simple contextual bandit outperforms a simple non-contextual bandit as well. All the three curves present the same shape (*c.f.* Figure 1). This may indicate a strong pattern in user-click behavior that is independent of the contextual information. However, by the same reason, it can be argued that, since this pattern exists in the

Table 3: Leaderboard of the first-phase (training) of the Challenge on Yahoo! dataset at ICML 2012 Workshop "new Challenges for Exploration & Exploitation 3". The entry of the proposed method is indicated in bold-face by the team-name JAMH, Universidad Complutense de Madrid (UCM), with a score of 891.9 (5th place).

NAME	AFFILIATION	SCORE	RANK
Ku-Chun	NTU	905.9	1
tvirot	MIT	903.9	2
edjoesu	MIT	903.4	3
Francis	ULg	895.4	4
JAMH	**UCM**	**891.9**	**5**
exploreit	untitled	891.4	6
EpsilonGreedyRocks	U of A	890.9	7
Exp	LUM	890.9	8
bigwhite	NCU of TW	885.8	9
Allen	NCU	883.8	10

Table 4: Leaderboard of the competition's final phase (testing) of the Challenge on Yahoo! dataset at ICML 2012 Workshop new Challenges for Exploration & Exploitation 3. The entry of the proposed method is indicated in bold-face by the team-name JAMH, Universidad Complutense de Madrid, with a score of 887.6 (2nd place!).

NAME	AFFILIATION	SCORE	RANK
The expert	Montanuniversitaet Leoben	891.9	1
JAMH	**Universidad Complutense de Madrid**	**887.6**	**2**
Allen	NCU	881.1	3
Yildiz	Montanuniversitaet Leoben	874.3	4
lucati	Montanuniversitaet Leoben	873.4	5
Ku-Chun	University of Taiwan	868.3	6
tvirot & edjoesu	MIT	865.9	7
tianhuil	Princeton	863.4	8
EpsilonGreedyRocks	U of A	863.4	9

behavior of all the three experiments, the difference in performance should be strongly attributed to the effective use of contextual information by the Bayes based method. Furthermore, Figure 2 shows the difference in derivative CTR vs. processed lines between the Naïve II, III and Linear Bayes method over the first 9200000 rows. As it can bee seen, the difference is small and focused on certain periods of time.

Tables 3 and 4 show the result of the evaluation of the proposed algorithm by the evaluation server application used by the organizing committee for the challenge. Table 3 shows the leaderboard for the top 10 entries of the first phase (training on the first 9200000). The entry of the proposed method is indicated in bold, by the name JAMH, Universidad Complutense de Madrid (UCM), with a score of 891.9 (5th place of 38 entries).

Table 4 shows the leaderboard for the top 9 entries of the final-stage (evaluation on the complete dataset). The entry of the proposed method is indicated by the name JAMH, Universidad Complutense de Madrid, with a score of 887.6 (2nd place! with a very little gap w.r.t. the first place). This result indicates not only that the algorithm performs very well but that it also generalizes very well. Furthermore, since the algorithm obtained the *fifth-place* in the training phase

and improved up to the *second place* in the final-stage (on the complete dataset), it may indicate that some desired features of the algorithm are working well. For instance, the stability in the long-run, which can be affected by overflows or arithmetic precision when running on big-data; keep a good exploration / exploitation trade-off during all the running time, something that is truly difficult to tune manually; and maintain a good learning trade-off between a long history of estimates and new data, *i.e.* , the plasticity level.

Furthermore, all the attempts that we have made (by now) to combine this algorithm with other complimentary exploration / exploitation procedures, such as ϵ-greedy or soft-max, have failed to beat the selection rule (24).

4. Conclusion

We have presented a feasible computationally efficient empirical Bayes-like method for learning in contextual-bandit problems. It is able to take effective advantage of large contextual information in an efficient manner. The method obtained the second place in the Challenge on Yahoo! dataset at ICML 2012 Workshop "new Challenges for Exploration & Exploitation 3", among more than 30 teams from several universities.

As shown in Section 3, the method has shown to possess several features that are often a hard requirement for real applications:

1. It presents good stability in the long-run, which can be affected by overflows or arithmetic precision when running on big-data.
2. It keeps a good exploration / exploitation trade-off during all the running time, something that is truly difficult to tune manually.
3. It maintains a good learning trade-off between a long history of estimates and new data, that is, the plasticity level.
4. Its computation is quite simple, fast and efficient.

As a final remark, it is very important to mention that the algorithm scales linear in the number of binary features and scales linear in the number of arms. These are definitely the mayor advantages (together with its predictive performance) of the presented approach to contextual bandits with large binary features context.

Acknowledgments

This work is partially supported by Spanish CICYT Project: DPI2009-14552-C02-01 "Surveillance, search and rescue at sea through collaboration of autonomous marine and aerial vehicles".

We want also thank Yahoo!'s WEBSCOPE program for sharing with us the "R6B - Yahoo! Front Page Today Module User Click Log Dataset, version 2.0 (300 MB)" for academic purposes.

References

Agarwal, D., Chen, B.-C., & Elango, P. (2009). Spatio-temporal models for estimating click-through rate. In *Proceedings of the 18th international conference on World wide web* (pp. 21–30). ACM.

Agarwal, D., Chen, B.-C., Elango, P., & Ramakrishnan, R. (2013). Content recommendation on web portals. *Commun. ACM*, *56*, 92–101. doi:10.1145/2461256.2461277.

Audibert, J.-Y., Munos, R., & Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, *410*, 1876–1902.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, *47*, 235–256.

Berry, D., & Fristedt, B. (1986). Bandit problems. *Journal of Applied Statistics*, *13*.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, *5*, 1–122.

Frazier, P. I., Powell, W. B., & Dayanik, S. (2008). A knowledge-gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, *47*, 2410–2439.

Gittins, J., Weber, R., & Glazebrook, K. (1989). *Multi-armed bandit allocation indices* volume 25. Wiley Online Library.

Holland, J. (1975). *Adaptation in natural and artificial systems*. 53. University of Michigan press.

Jun, T. (2004). A survey on the bandit problem with switching costs. *De Economist*, *152*, 513–541.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Konstan, J. A., & Ried, J. (2012). Deconstructing recommender systems: Recommended for you. *IEEE Spectrum*, (pp. 49–56).

Kuleshov, V., & Precup, D. (2010). Algorithms for the multi-armed bandit problem. *Journal of Machine Learning*, .

Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in Neural Information Processing Systems*, *20*, 1096–1103.

Li, L., Chu, W., Langford, J., & Schapire, R. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web* (pp. 661–670). ACM.

Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining* WSDM '11 (pp. 297–306). New York, NY, USA: ACM. doi:10.1145/1935826.1935878.

Liu, J., Dolan, P., & Pedersen, E. R. (2010). Personalized news recommendation based on click behavior. In *Proceedings of the 15th international conference on Intelligent user interfaces* (pp. 31–40). ACM.

March, J. (1991). Exploration and exploitation in organizational learning. *Organization science*, (pp. 71–87).

Powell, W. B. (2010). The knowledge gradient for optimal learning. *Wiley Encyclopedia of Operations Research and Management Science*, .

Powell, W. B., & Ryzhov, I. O. (2012). *Optimal learning* volume 841. John Wiley & Sons.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.

Robbins, H. (1964). The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, *35*, 1–20.

Solomonoff, R. J. (1964). A formal theory of inductive inference. part i. *Information and control*, *7*, 1–22.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction* volume 1. Cambridge Univ Press.

Vermorel, J., & Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. *Machine Learning: ECML 2005*, (pp. 437–448).

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Wald, A. (1947). Sequential analysis, .

Wang, X., Li, W., Cui, Y., Zhang, R. B., & Mao, J. (2011). Click-through rate estimation for rare events in online advertising. *Online Multimedia Advertising: Techniques and Technologies*, (p. 1).

Weng, S.-S., & Liu, M.-J. (2004). Feature-based recommendations for one-to-one marketing. *Expert Systems with Applications*, *26*, 493 – 508. doi:10.1016/j.eswa.2003.10.008.

Yahoo! Academic Relations (2012). R6A - Yahoo! Front Page Today Module User Click Log Dataset, Version 1.0. URL: http://webscope.sandbox.yahoo.com.