Thompson Sampling for Contextual Bandits with Linear Payoffs

Shipra Agrawal SHIPRA@MICROSOFT.COM
Microsoft Research India

Navin Goyal NAVINGO@MICROSOFT.COM
Microsoft Research India

Abstract

Thompson Sampling is one of the oldest heuristics for multi-armed bandit problems. It is a randomized algorithm based on Bayesian ideas, and has recently generated significant interest after several studies demonstrated it to have better empirical performance compared to the state-of-the-art methods. However, many questions regarding its theoretical performance remained open. In this paper, we design and analyze a generalization of Thompson Sampling algorithm for the stochastic contextual multi-armed bandit problem with linear payoff functions, when the contexts are provided by an adaptive adversary. This is among the most important and widely studied version of the contextual bandits problem. We prove a high probability regret bound of $\tilde{O}(\frac{d^2}{\epsilon}\sqrt{T^{1+\epsilon}})$ in time T for any $0 < \epsilon < 1$, where d is the dimension of each context vector and ϵ is a parameter used by the algorithm. Our results provide the first theoretical guarantees for the contextual version of Thompson Sampling, and are close to the lower bound of $\Omega(d\sqrt{T})$ for this problem. This essentially solves a COLT open problem of Chapelle and Li [COLT 2012].

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

1. Introduction

Multi-armed bandit (MAB) problems model the exploration/exploitation trade-off inherent in many sequential decision problems. There are many versions of multi-armed bandit problems; a particularly useful version is the contextual multi-armed bandit problem. In this problem, in each of T rounds, a learner is presented with the choice of taking one out of N actions, referred to as N arms. Before making the choice of which arm to play, the learner sees d-dimensional feature vectors b_i, referred to as "context", associated with each arm i. The learner uses these feature vectors along with the feature vectors and rewards of the arms played by her in the past to make the choice of the arm to play in the current round. Over time, the learner's aim is to gather enough information about how the feature vectors and rewards relate to each other, so that she can predict, with some certainty, which arm is likely to give the best reward by looking at the feature vectors. The learner competes with a class of predictors, in which each predictor takes in the feature vectors and predicts which arm will give the best reward. If the learner can guarantee to do nearly as well as the predictions of the best predictor in hindsight (i.e., have low regret), then the learner is said to successfully compete with that class.

In the contextual bandits setting with *linear payoff functions*, the learner competes with the class of all "linear" predictors on the feature vectors. That is, a predictor is defined by a d-dimensional parameter $\overline{\mu} \in \mathbb{R}^d$, and the predictor ranks the arms according to $b_i^T\overline{\mu}$. We consider stochastic contextual bandit problem under linear realizability assumption, that is, we assume that there is an unknown underlying parameter $\mu \in \mathbb{R}^d$ such that the expected reward for each arm i, given context b_i, is $b_i^T\mu$. Under this realizability assumption, the linear predictor corresponding to μ is in fact the best predictor and the learner's aim is to learn this underlying parameter. This realizability assumption is standard in the existing literature on contextual

multi-armed bandits, e.g. (Auer, 2002; Filippi et al., 2010; Chu et al., 2011; Abbasi-Yadkori et al., 2011).

Thompson Sampling (TS) is one of the earliest heuristics for multi-armed bandit problems. The first version of this Bayesian heuristic is around 80 years old, dating to Thompson (1933). Since then, it has been rediscovered numerous times independently in the context of reinforcement learning, e.g., in Wyatt (1997); Ortega & Braun (2010); Strens (2000). It is a member of the family of *randomized probability matching* algorithms. The basic idea is to assume a simple prior distribution on the underlying parameters of the reward distribution of every arm, and at every time step, play an arm according to its posterior probability of being the best arm. The general structure of TS for the contextual bandits problem involves the following elements:

1. a set Θ of parameters $\tilde{\mu}$;
2. a prior distribution $P(\tilde{\mu})$ on these parameters;
3. past observations $\mathcal{D}$ consisting of (context b, reward r) for the past time steps;
4. a likelihood function $P(r|b, \tilde{\mu})$, which gives the probability of reward given a context b and a parameter $\tilde{\mu}$;
5. a posterior distribution $P(\tilde{\mu}|\mathcal{D}) \propto P(\mathcal{D}|\tilde{\mu})P(\tilde{\mu})$, where $P(\mathcal{D}|\tilde{\mu})$ is the likelihood function.

In each round, TS plays an arm according to its posterior probability of having the best parameter. A simple way to achieve this is to produce a sample of parameter for each arm, using the posterior distributions, and play the arm that produces the best sample. In this paper, we design and analyze a natural generalization of Thompson Sampling (TS) for contextual bandits; this generalization fits the above general structure, and uses Gaussian prior and Gaussian likelihood function. We emphasize that although TS is a Bayesian approach, the description of the algorithm and our analysis apply to the prior-free stochastic MAB model, and our regret bounds will hold irrespective of whether or not the actual reward distribution matches the Gaussian likelihood function used to derive this Bayesian heuristic. Thus, our bounds for TS algorithm are directly comparable to the UCB family of algorithms which form a frequentist approach to the same problem. One could interpret the priors used by TS as a way of capturing the current knowledge about the arms.

Recently, TS has attracted considerable attention. Several studies (e.g., Granmo (2010); Scott (2010); Graepel et al. (2010); Chapelle & Li (2011); May & Leslie (2011); Kaufmann et al. (2012)) have empirically demonstrated the efficacy of TS: Scott (2010) provides a detailed discussion of probability matching techniques in many general settings along with favorable empirical comparisons with other techniques. Chapelle & Li (2011) demonstrate that for the basic stochastic MAB problem, empirically TS achieves regret comparable to the lower bound of Lai & Robbins (1985); and in applications like display advertising and news article recommendation modeled by the contextual bandits problem, it is competitive to or better than the other methods such as UCB. In their experiments, TS is also more robust to delayed or batched feedback than the other methods. TS has been used in an industrial-scale application for CTR prediction of search ads on search engines (Graepel et al., 2010). Kaufmann et al. (2012) do a thorough comparison of TS with the best known versions of UCB and show that TS has the lowest regret in the long run.

However, the theoretical understanding of TS is limited. Granmo (2010) and May et al. (2011) provided weak guarantees, namely, a bound of $o(T)$ on the expected regret in time T. For the the basic (i.e. without contexts) version of the stochastic MAB problem, some significant progress was made by Agrawal & Goyal (2012), Kaufmann et al. (2012) and, more recently, by Agrawal & Goyal (2013), who provided optimal regret bounds on the expected regret. But, many questions regarding theoretical analysis of TS remained open, including high probability regret bounds, and regret bounds for the more general contextual bandits setting. In particular, the contextual MAB problem does not seem easily amenable to the techniques used so far for analyzing TS for the basic MAB problem. In Section 3.1, we describe some of these challenges. Some of these questions and difficulties were also formally raised as a COLT 2012 open problem (Chapelle & Li, 2012).

In this paper, we use novel martingale-based analysis techniques to demonstrate that TS (i.e., our Gaussian prior based generalization of TS for contextual bandits) achieves high probability, near-optimal regret bounds for stochastic contextual bandits with linear payoff functions. To our knowledge, ours are the first non-trivial regret bounds for TS for the contextual bandits problem. Additionally, our results are the first high probability regret bounds for TS, even in the case of basic MAB problem. This essentially solves the COLT 2012 open problem by(Chapelle & Li, 2012) for contextual bandits with linear payoffs.

Our version of Thompson Sampling algorithm for the contextual MAB problem, described formally in Section 2.2, uses Gaussian prior and Gaussian likelihood functions. Our techniques can be extended to the use of other prior distributions, satisfying certain condi-

tions, as discussed in Section 4.

2. Problem setting and algorithm description

2.1. Problem setting

There are N arms. At time $t = 1, 2, \ldots$, a context vector $b_i(t) \in \mathbb{R}^d$, is revealed for every arm i. These context vectors are chosen by an adversary in an adaptive manner after observing the arms played and their rewards up to time $t-1$, i.e. history $\mathcal{H}_{t-1}$,

$$\mathcal{H}_{t-1} = \{a(\tau), r_{a(\tau)}(\tau), b_i(\tau), i = 1, \ldots, N, \tau = 1, \ldots, t-1\},$$

where $a(\tau)$ denotes the arm played at time τ. Given $b_i(t)$, the reward for arm i at time t is generated from an (unknown) distribution with mean $b_i(t)^T\mu$, where $\mu \in \mathbb{R}^d$ is a fixed but unknown parameter.

$$\mathbb{E}\left[r_i(t) \,\middle|\, \{b_i(t)\}_{i=1}^N, \mathcal{H}_{t-1}\right] = \mathbb{E}\left[r_i(t) \mid b_i(t)\right] = b_i(t)^T\mu.$$

An algorithm for the *contextual bandit problem* needs to choose, at every time t, an arm $a(t)$ to play, using history $\mathcal{H}_{t-1}$ and current contexts $b_i(t), i = 1, \ldots, N$. Let $a^*(t)$ denote the optimal arm at time t, i.e. $a^*(t) = \arg\max_i b_i(t)^T\mu$. And let $\Delta_i(t)$ be the difference between the mean rewards of the optimal arm and of arm i at time t, i.e.,

$$\Delta_i(t) = b_{a^*(t)}(t)^T\mu - b_i(t)^T\mu.$$

Then, the regret at time t is defined as

$$\text{regret}(t) = \Delta_{a(t)}(t).$$

The objective is to minimize the total regret $\mathcal{R}(T) = \sum_{t=1}^T \text{regret}(t)$ in time T. The time horizon T is finite but possibly unknown.

We assume that $\eta_{i,t} = r_i(t) - b_i(t)^T\mu$ is conditionally R-sub-Gaussian for a constant $R \geq 0$, i.e.,

$$\forall\lambda \in \mathbb{R}, \mathbb{E}[e^{\lambda\eta_{i,t}} | \{b_i(t)\}_{i=1}^N, \mathcal{H}_{t-1}] \leq \exp\left(\frac{\lambda^2R^2}{2}\right).$$

This assumption is satisfied whenever $r_i(t) \in [b_i(t)^T\mu - R, b_i(t)^T\mu + R]$ (see Remark 1 in Appendix A.1 of Filippi et al. (2010)). We will also assume that $||b_i(t)|| \leq 1$, $||\mu|| \leq 1$, and $\Delta_i(t) \leq 1$ for all i, t (the norms, unless otherwise indicated, are ℓ_2-norms). These assumptions are required to make the regret bounds scale-free, and are standard in the literature on this problem. If $||\mu|| \leq c, ||b_i(t)|| \leq c, \Delta_i(t) \leq c$ instead, then our regret bounds would increase by a factor of c.

Remark 1. *An alternative definition of regret that appears in the literature is*

$$\textit{regret}(t) = r_{a^*(t)}(t) - r_{a(t)}(t).$$

We can obtain the same regret bounds for this alternative definition of regret. The details are provided in the supplementary material in Appendix A.5.

2.2. Thompson Sampling algorithm

We use Gaussian likelihood function and Gaussian prior to design our version of Thompson Sampling algorithm. More precisely, suppose that the **likelihood** of reward $r_i(t)$ at time t, given context $b_i(t)$ and parameter μ, were given by the pdf of Gaussian distribution $\mathcal{N}(b_i(t)^T\mu, v^2)$. Here, $v = R\sqrt{\frac{24}{\epsilon}d\ln(\frac{1}{\delta})}$, with $\epsilon \in (0,1)$ which parametrizes our algorithm. Let

$$B(t) = I_d + \textstyle\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)b_{a(\tau)}(\tau)^T$$

$$\hat{\mu}(t) = B(t)^{-1}\left(\textstyle\sum_{\tau=1}^{t-1} b_{a(\tau)}(\tau)r_{a(\tau)}(\tau)\right).$$

Then, if the **prior** for μ at time t is given by $\mathcal{N}(\hat{\mu}(t), v^2B(t)^{-1})$, it is easy to compute the **posterior** distribution at time $t+1$,

$$\Pr(\tilde{\mu}|r_i(t)) \propto \Pr(r_i(t)|\tilde{\mu})\Pr(\tilde{\mu})$$

as $\mathcal{N}(\hat{\mu}(t+1), v^2B(t+1)^{-1})$ (details of this computation are in Appendix A.1). In our Thompson Sampling algorithm, at every time step t, we will simply generate a sample $\tilde{\mu}(t)$ from the distribution $\mathcal{N}(\hat{\mu}(t), v^2B(t)^{-1})$, and play the arm i that maximizes $b_i(t)^T\tilde{\mu}(t)$.

We emphasize that the Gaussian priors and the Gaussian likelihood model for rewards are only used above to design the Thompson Sampling algorithm for contextual bandits. Our analysis of the algorithm allows these models to be completely unrelated to the *actual* reward distribution. The assumptions on the actual reward distribution are only those mentioned in Section 2.1, i.e., the R-sub-Gaussian assumption.

Algorithm 1 Thompson Sampling for Contextual bandits

Set $B = I_d, \hat{\mu} = 0_d$, $f = 0_d$.
for all $t = 1, 2, \ldots,$ **do**
  Sample $\tilde{\mu}(t)$ from distribution $\mathcal{N}(\hat{\mu}, v^2B^{-1})$.
  Play arm $a(t) := \arg\max_i b_i(t)^T\tilde{\mu}(t)$, and observe reward r_t.
  Update $B = B + b_{a(t)}(t)b_{a(t)}(t)^T$, $f = f + b_{a(t)}(t)r_t$, $\hat{\mu} = B^{-1}f$.
end for

Every step t of Algorithm 1 consists of generating a d-dimensional sample $\tilde{\mu}(t)$ from a multivariate Gaussian distribution, and solving the problem $\arg\max_i b_i(t)^T\tilde{\mu}(t)$. Therefore, even if the number of arms N is large (or infinite), the above algorithm is efficient as long as the problem $\arg\max_i b_i(t)^T\tilde{\mu}(t)$ is

efficiently solvable. This is the case, for example, when the set of arms at time t is given by a d-dimensional convex set (every vector in the convex set is a context vector, and thus corresponds to an arm).

2.3. Our Results

Theorem 1. *For the stochastic contextual bandit problem with linear payoff functions, with probability $1-\delta$, the total regret in time T for Thompson Sampling (Algorithm 1) is bounded by $O\left(\frac{d^2}{\epsilon}\sqrt{T^{1+\epsilon}}\left(\ln(Td)\ln\frac{1}{\delta}\right)\right)$, for any $0<\epsilon<1$, $0<\delta<1$. Here, ϵ is a parameter used by the Thompson Sampling algorithm.*

Remark 2. *The parameter ϵ can be chosen to be any constant in $(0,1)$. If T is known, one could choose $\epsilon=\frac{1}{\ln T}$, to get $\tilde{O}(d^2\sqrt{T})$ regret bound.*

Remark 3. *Our regret bound in Theorem 1 does not depend on N, and is applicable to the case of infinite arms, with only notational changes required in the analysis.*

In the main body of this paper, we will discuss the proof of the above result. Below, we state two additional results; their proofs require small changes to the proof of Theorem 1 and are provided in the supplementary material.

The first result is for the setting where each of the N arms is associated with a different d-dimensional parameter $\mu_i\in\mathbb{R}^d$, so that the mean reward for arm i at time t is $b_i(t)^T\mu_i$. This setting is a direct generalization of the basic MAB problem to d-dimensions. Thompson Sampling for this setting will maintain a separate posterior distribution for each arm i which would be updated only at the time instances when i is played. And, at every time step t, instead of a single sample $\tilde{\mu}(t)$, N independent samples will have to be generated: $\tilde{\mu}_i(t)$ for each arm i. We prove the following regret bound for this setting.

Theorem 2. *For the setting with N different parameters, with probability $1-\delta$, the total regret in time T for Thompson Sampling is bounded by $O\left(d\sqrt{\frac{NT^{1+\epsilon}\ln N}{\epsilon}}\left(\ln T\ln\frac{1}{\delta}\right)\right)$, for any $0<\epsilon<1$, $0<\delta<1$.*

The details of the algorithm for N-parameter setting and the proof of Theorem 2 appear in the supplementary material in Appendix C.

Note that unlike Theorem 1, the regret bound in Theorem 2 has a dependence on N, which is expected because Theorem 2 deals with a setting where there are N different parameters to learn. However, the bound in Theorem 2 has a better dependence on d. This improvement results from the independence of $\theta_i(t)=b_i(t)^T\tilde{\mu}_i(t)$ in the algorithm for this setting. On the other hand in Algorithm 1, used for the single parameter setting of Theorem 1, a single $\tilde{\mu}(t)$ is generated, and so $\theta_i(t)=b_i(t)^T\tilde{\mu}(t)$ are not independent.

This motivates us to consider a modification of Algorithm 1 for the single parameter setting, in which the $\theta_i(t)$'s are independently generated, each with marginal distribution $b_i(t)^T\tilde{\mu}(t)$. The arm with the highest value of $\theta_i(t)$ is played at time t. Although, this modified algorithm could be inefficient compared to Algorithm 1 if N is large (say exponential) compared to d, the better dependence on d in regret bounds could be useful if d is large.

Theorem 3. *For the modified algorithm in single parameter setting, with probability $1-\delta$, the total regret in time T is bounded by $O\left(d\sqrt{\frac{T^{1+\epsilon}\ln N}{\epsilon}}\left(\ln T\ln\frac{1}{\delta}\right)\right)$, for any $0<\epsilon<1$, $0<\delta<1$.*

The details of the modified algorithm and the proof of the above theorem appears in the supplementary material in Appendix B.

2.4. Related Work

The contextual bandit problem with linear payoffs is a widely studied problem in statistics and machine learning often under different names as mentioned by Chu et al. (2011): bandit problems with co-variates (Woodroofe, 1979; Sarkar, 1991), associative reinforcement learning (Kaelbling, 1994), associative bandit problems (Auer, 2002; Strehl et al., 2006), bandit problems with expert advice (Auer et al., 2002), and linear bandits (Dani et al., 2008; Abbasi-Yadkori et al., 2011; Bubeck et al., 2012). The name *contextual bandits* was coined in Langford & Zhang (2007).

A lower bound of $\Omega(d\sqrt{T})$ for this problem was given by Dani et al. (2008), when the number of arms is allowed to be infinite. In particular, they prove their lower bound using an example where the set of arms correspond to all vectors in the intersection of a d-dimensional sphere and a cube. They also provide an upper bound of $\tilde{O}(d\sqrt{T})$, although their setting is slightly restrictive in the sense that the context vector for every arm is fixed in advanced and is not allowed to change with time. Abbasi-Yadkori et al. (2011) analyze a UCB-style algorithm and provide a regret upper bound of $O(d\log(T)\sqrt{T}+\sqrt{dT\log(T/\delta)})$. Apart from the dependence on ϵ, our bounds are essentially away by a factor of d from these bounds.

For finite N, Chu et al. (2011) show a lower bound

of $\Omega(\sqrt{Td})$ for $d^2 \leq T$. Auer (2002) and Chu et al. (2011) analyze SupLinUCB, a complicated algorithm using UCB as a subroutine, for this problem. Chu et al. (2011) achieve a regret bound of $O(\sqrt{Td\ln^3(NT\ln(T)/\delta)})$ with probability at least $1-\delta$ (Auer (2002) proves similar results). This regret bound is not applicable to the case of infinite arms, and assumes that context vectors are generated by an *oblivious* adversary. Also, this regret bound would give $O(d^2\sqrt{T})$ regret if N is exponential in d. The state-of-the-art bounds for linear bandits problem in case of finite N are given by Bubeck et al. (2012). They provide an algorithm based on exponential weights, with regret of order $\sqrt{dT\log N}$ for any finite set of N actions. However, the exponential weights based algorithms are not efficient if N is large (sampling complexity of $O(N)$ in every step). Also, their setting is slightly different from ours. The set of arms and the associated b_i vectors are *non-adaptive* and fixed in advance. And, they consider a non-stochastic (adversarial) bandit setting where the reward at time t for arm i is $b_i^T\mu_t$ with μ_t chosen by an adversary.

Very recent work Russo & Roy (2013) provides near-optimal bounds on *Bayesian regret* in many general settings. This result is incomparable to ours because of the different notion of regret used.

While the regret bounds provided in this paper do not match or better the best available regret bounds for the extensively studied problem of linear contextual bandits, our results demonstrate that the natural and efficient heuristic of Thompson Sampling can achieve theoretical bounds that are close to the best bounds. The main contribution of this paper is to provide new tools for analysis of Thompson Sampling algorithm for contextual bandits, which despite being popular and empirically attractive, has eluded theoretical analysis. We believe the techniques used in this paper will provide useful insights into the workings of this Bayesian algorithm, and may be useful for further improvements and extensions.

3. Regret Analysis: Proof of Theorem 1

3.1. Challenges and proof outline

The contextual version of the multi-armed bandit problem presents new challenges for the analysis of TS algorithm, and the techniques used so far for analyzing the basic multi-armed bandit problem by Agrawal & Goyal (2012); Kaufmann et al. (2012) do not seem directly applicable. Let us describe some of these difficulties and our novel ideas to resolve them.

In the basic MAB problem there are N arms, with mean reward $\mu_i \in \mathbb{R}$ for arm i, and the regret for playing a suboptimal arm i is $\mu_{a^*} - \mu_i$, where a^* is the arm with the highest mean. Let us compare this to a 1-dimensional contextual MAB problem, where arm i is associated with a parameter $\mu_i \in \mathbb{R}$, but in addition, at every time t, it is associated with a context $b_i(t) \in \mathbb{R}$, so that mean reward is $b_i(t)\mu_i$. The best arm $a^*(t)$ at time t is the arm with the highest mean at time t, and the regret for playing arm i is $b_{a^*(t)}(t)\mu_{a^*(t)} - b_i(t)\mu_i$.

In general, the basis of regret analysis for stochastic MAB is to prove that the variances of empirical estimates for all arms decrease fast enough, so that the regret incurred until the variances become small enough, is small. In the basic MAB, the variance of the empirical mean is inversely proportional to the number of plays $k_i(t)$ of arm i at time t. Thus, every time the suboptimal arm i is played, we know that even though a regret of $\mu_{i^*} - \mu_i \leq 1$ is incurred, there is also an improvement of exactly 1 in the number of plays of that arm, and hence, corresponding decrease in the variance. The techniques for analyzing basic MAB rely on this observation to precisely quantify the exploration-exploitation tradeoff. On the other hand, the variance of the empirical mean for the contextual case is given by inverse of $B_i(t) = \sum_{\tau=1:a(\tau)=i}^{t} b_i(\tau)^2$. When a suboptimal arm i is played, if $b_i(t)$ is small, the regret $b_{a^*(t)}(t)\mu_{a^*(t)} - b_i(t)\mu_i$ could be much higher than the improvement $b_i(t)^2$ in $B_i(t)$.

In our proof, we overcome this difficulty by dividing the arms into two groups at any time: saturated and unsaturated arms, based on whether the standard deviation of the estimates for an arm is smaller or larger compared to the standard deviation for the optimal arm. The optimal arm is included in the group of unsaturated arms. We show that for the unsaturated arms, the regret on playing the arm can be bounded by a factor of the standard deviation, which improves every time the arm is played. This allows us to bound the total regret due to unsaturated arms. For the saturated arms, standard deviation is small, or in other words, the estimates of the means constructed so far are quite accurate in the direction of the current contexts of these arms, so that the algorithm is able to distinguish between them and the optimal arm. We utilize this observation to show that the probability of playing such arms at any step is bounded by a function of the probability of playing the unsaturated arms.

Below is a more technical outline of the proof of Theorem 1. At any time step t, we divide the arms into two groups:

- *saturated arms* defined as those with $g(T)s_{t,i} < \ell(T)s_{t,a^*(t)}$,
- *unsaturated arms* defined as those with $g(T)s_{t,i} \geq \ell(T)s_{t,a^*(t)}$,

where $s_{t,i} = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$ and $g(T)$, $\ell(T)$ ($g(T) > \ell(T)$) are constants (functions of T, d, δ) defined later. Note that $s_{t,i}$ is the standard deviation of the estimate $b_i(t)^T\hat{\mu}(t)$ and $vs_{t,i}$ is the standard deviation of the random variable $b_i(t)^T\tilde{\mu}(t)$.

We use concentration bounds for $\tilde{\mu}(t)$ and $\hat{\mu}(t)$ to bound the regret at any time t by $g(T)(s_{t,a^*(t)}+s_{t,a(t)})$. Now, if an unsaturated arm is played at time t, then using the definition of unsaturated arms, the regret is at most $\frac{2g(T)^2}{\ell(T)}s_{t,a(t)}$. This is useful because of the inequality $\sum_t s_{t,a(t)} = O(\sqrt{Td\ln T})$ (derived along the lines of Auer (2002)), which allows us to bound the total regret due to unsaturated arms.

For saturated arms, we prove that the probability of playing a saturated arm at any time t is within p of the probability of playing an unsaturated arm, where $p = \frac{1}{4e\sqrt{\pi T^\epsilon}}$. More precisely, we define $\mathcal{F}_{t-1}$ as the union of history $\mathcal{H}_{t-1}$ and the contexts $b_i(t), i = 1, \ldots, N$ at time t, and prove that for "most" (in a high probability sense) $\mathcal{F}_{t-1}$,

$$\Pr\left(a(t) \text{ is a saturated arm} \mid \mathcal{F}_{t-1}\right) \leq \frac{1}{p} \cdot \Pr\left(a(t) \text{ is an unsaturated arm} \mid \mathcal{F}_{t-1}\right) + \frac{1}{pT^2},$$

We use these observations to establish that $(X_t; t \geq 0)$, where

$$X_t \simeq \text{regret}(t) - \frac{g(T)}{p} I(a(t) \text{ is unsaturated}) s_{t,a^*(t)} - \frac{2g(T)^2}{\ell(T)} s_{t,a(t)} - \frac{2g(T)}{pT^2},$$

is a super-martingale difference process adapted to filtration $\mathcal{F}_t$. Then, using the Azuma-Hoeffding inequality for super-martingales, along with the inequality $\sum_t s_{t,a(t)} = O(\sqrt{Td\ln T})$, we will obtain the desired high probability regret bound.

3.2. Formal proof

For quick reference, the notations introduced below also appear in a table of notations at the beginning of the supplementary material.

Definition 1. *For all i, define $\theta_i(t) = b_i(t)^T\tilde{\mu}(t)$, and $s_{t,i} = \sqrt{b_i(t)^T B(t)^{-1} b_i(t)}$. By definition of $\tilde{\mu}(t)$, marginal distribution of each $\theta_i(t)$ is Gaussian with mean $b_i(t)^T\hat{\mu}(t)$ and standard deviation $vs_{t,i}$. Also, $s_{t,i}$ is the standard deviation of estimate $b_i(t)^T\hat{\mu}(t)$.*

Definition 2. *Recall that $\Delta_i(t) = b_{a^*(t)}(t)^T\mu - b_i(t)^T\mu$, the difference between the mean reward of optimal arm and arm i at time t. .*

Definition 3. *Define $\ell(T) = R\sqrt{d\ln(T^3)\ln(\frac{1}{\delta})} + 1$, $v = R\sqrt{\frac{24}{\epsilon}d\ln(\frac{1}{\delta})}$, and $g(T) = \sqrt{4d\ln(Td)}\; v + \ell(T)$.*

Definition 4. *Define $E^\mu(t)$ and $E^\theta(t)$ as the events that $b_i(t)^T\hat{\mu}(t)$ and $\theta_i(t)$ are concentrated around their respective means. More precisely, define $E^\mu(t)$ as the event that*

$$\forall i : |b_i(t)^T\hat{\mu}(t) - b_i(t)^T\mu| \leq \ell(T)\; s_{t,i}.$$

Define $E^\theta(t)$ as the event that

$$\forall i : |\theta_i(t) - b_i(t)^T\hat{\mu}(t)| \leq \sqrt{4d\ln(Td)}\; vs_{t,i}.$$

Definition 5. *An arm i is called* saturated *at time t if $g(T)\; s_{t,i} < \ell(T)\; s_{t,a^*(t)}$, and* unsaturated *otherwise. Let $C(t)$ denote the set of saturated arms at time t. Note that the optimal arm is always unsaturated at time t, i.e., $a^*(t) \notin C(t)$. An arm may keep shifting from saturated to unsaturated and vice-versa over time.*

Definition 6. *Define filtration $\mathcal{F}_{t-1}$ as the union of history until time $t-1$, and the contexts at time t, i.e., $\mathcal{F}_{t-1} = \{\mathcal{H}_{t-1}, b_i(t), i = 1, \ldots, N\}$.*

By definition, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \cdots \subseteq \mathcal{F}_{T-1}$. Observe that the following quantities are determined by the history $\mathcal{H}_{t-1}$ and the contexts $b_i(t)$ at time t, and hence are included in $\mathcal{F}_{t-1}$,

- *$\hat{\mu}(t), B(t)$,*
- *$s_{t,i}$, for all i,*
- *the identity of the optimal arm $a^*(t)$ and the set of saturated arms $C(t)$,*
- *whether $E^\mu(t)$ is true or not,*
- *the distribution $\mathcal{N}(\hat{\mu}(t), B(t)^{-1})$ of $\tilde{\mu}(t)$, and hence the joint distribution of $\theta_i(t) = b_i(t)^T\tilde{\mu}(t), i = 1, \ldots, N$.*

Lemma 1. *For all t, $0 < \delta < 1$, $\Pr(E^\mu(t)) \geq 1 - \frac{\delta}{T^2}$. And, for all possible filtrations $\mathcal{F}_{t-1}$, $\Pr(E^\theta(t)|\mathcal{F}_{t-1}) \geq 1 - \frac{1}{T^2}$.*

Proof. The complete proof of this lemma appears in Appendix A.3. The probability bound for $E^\mu(t)$ will be proven using a concentration inequality given by Abbasi-Yadkori et al. (2011), stated as Lemma 7 in Appendix A.2. The R-sub-Gaussian assumption on rewards will be utilized here. The probability bound for $E^\theta(t)$ will be proven using a concentration inequality for Gaussian random variables from Abramowitz & Stegun (1964) stated as Lemma 5 in Appendix A.2 . □

The next lemma lower bounds the probability that $\theta_{a^*(t)}(t) = b_{a^*(t)}(t)^T\tilde{\mu}(t)$ for the optimal arm at time t will exceed its mean reward $b_{a^*(t)}(t)^T\mu$ plus $\ell(T)s_{t,a^*(t)}$.

Lemma 2. *For any filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true,*

$$\Pr\left(\theta_{a^*(t)}(t) > b_{a^*(t)}(t)^T\mu + \ell(T)s_{t,a^*(t)} \,\middle|\, \mathcal{F}_{t-1}\right) \geq \frac{1}{4e\sqrt{\pi}T^\epsilon}.$$

Proof. The proof uses anti-concentration of Gaussian random variable $\theta_{a^*(t)}(t) = b_{a^*(t)}(t)^T\tilde{\mu}(t)$, which has mean $b_{a^*(t)}(t)^T\hat{\mu}(t)$ and standard deviation $vs_{t,a^*(t)}$, provided by Lemma 5 in Appendix A.2, and the concentration of $b_{a^*(t)}(t)^T\hat{\mu}(t)$ around $b_{a^*(t)}(t)^T\mu$ provided by the event $E^\mu(t)$. The details of the proof are in Appendix A.4. □

The following lemma bounds the probability of playing saturated arms in terms of the probability of playing unsaturated arms.

Lemma 3. *Given any filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true,*

$$\Pr(a(t) \in C(t) \mid \mathcal{F}_{t-1}) \leq \frac{1}{p}\Pr(a(t) \notin C(t) \mid \mathcal{F}_{t-1}) + \frac{1}{pT^2},$$

where $p = \frac{1}{4e\sqrt{\pi}T^\epsilon}$.

Proof. The algorithm chooses the arm with the highest value of $\theta_i(t) = b_i(t)^T\tilde{\mu}(t)$ to be played at time t. Therefore, if $\theta_{a^*(t)}(t)$ is greater than $\theta_j(t)$ for all saturated arms, i.e., $\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t)$, then one of the unsaturated arms (which include the optimal arm and other suboptimal unsaturated arms) must be played. Therefore,

$$\begin{aligned}&\Pr(a(t) \notin C(t) \mid \mathcal{F}_{t-1})\\ &\geq \Pr\left(\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right). \quad (1)\end{aligned}$$

By definition, for all saturated arms, i.e. for all $j \in C(t)$, $g(T)s_{t,j} < \ell(T)s_{t,a^*(t)}$. Also, if both the events $E^\mu(t)$ and $E^\theta(t)$ are true then, by the definitions of these events, for all $j \in C(t)$, $\theta_j(t) \leq b_j(t)^T\mu + g(T)s_{t,j}$. Therefore, given an $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true, either $E^\theta(t)$ is false, or else for all $j \in C(t)$,

$$\theta_j(t) \leq b_j(t)^T\mu + g(T)s_{t,j} \leq b_{a^*(t)}(t)^T\mu + \ell(T)s_{t,a^*(t)}.$$

Hence, for any $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true,

$$\begin{aligned}&\Pr\left(\theta_{a^*(t)}(t) > \theta_j(t), \forall j \in C(t) \,\middle|\, \mathcal{F}_{t-1}\right)\\ \geq\ &\Pr\left(\theta_{a^*(t)}(t) > b_{a^*(t)}(t)^T\mu + \ell(T)s_{t,a^*(t)} \,\middle|\, \mathcal{F}_{t-1}\right)\\ &- \Pr\left(\overline{E^\theta(t)} \,\middle|\, \mathcal{F}_{t-1}\right)\\ \geq\ &p - \frac{1}{T^2}.\end{aligned}$$

The last inequality uses Lemma 2 and Lemma 1. Substituting in Equation (1), this gives,

$$\Pr(a(t) \notin C(t) \mid \mathcal{F}_{t-1}) + \frac{1}{T^2} \geq p,$$

which implies

$$\frac{\Pr(a(t) \in C(t) \mid \mathcal{F}_{t-1})}{\Pr(a(t) \notin C(t) \mid \mathcal{F}_{t-1}) + \frac{1}{T^2}} \leq \frac{1}{p}.$$

□

Definition 7. *Recall that regret(t) was defined as, regret(t) $= \Delta_{a(t)}(t) = b_{a^*(t)}(t)^T\mu - b_{a(t)}(t)^T\mu$. Define regret′(t) $=$ regret(t) $\cdot\, I(E^\mu(t))$.*

Next, we establish a super-martingale process that will form the basis of our proof of the high-probability regret bound.

Definition 8. *Let*

$$\begin{aligned}X_t &= \text{regret}'(t) - \frac{g(T)}{p}I(a(t) \notin C(t))s_{t,a^*(t)}\\ &\quad - \frac{2g(T)^2}{\ell(T)}s_{t,a(t)} - \frac{2g(T)}{pT^2},\\ Y_t &= \sum_{w=1}^{t} X_w,\end{aligned}$$

where $p = \frac{1}{4e\sqrt{\pi}T^\epsilon}$.

Lemma 4. *$(Y_t; t = 0, \ldots, T)$ is a super-martingale process with respect to filtration $\mathcal{F}_t$.*

Proof. See Definition 9 in Appendix A.2 for the definition of super-martingales. We need to prove that for all $t \in [1, T]$, and any $\mathcal{F}_{t-1}$, $\mathbb{E}[Y_t - Y_{t-1}|\mathcal{F}_{t-1}] \leq 0$, i.e.

$$\begin{aligned}&\mathbb{E}\left[\text{regret}'(t) \mid \mathcal{F}_{t-1}\right] \leq\\ &\frac{g(T)}{p}\Pr(a(t) \notin C(t) \mid \mathcal{F}_{t-1})\, s_{t,a^*(t)} +\\ &\frac{2g(T)^2}{\ell(T)}\mathbb{E}\left[s_{t,a(t)} \,\middle|\, \mathcal{F}_{t-1}\right] + \frac{2g(T)}{pT^2}.\end{aligned}$$

If $\mathcal{F}_{t-1}$ is such that $E^\mu(t)$ is not true, then regret′$(t) =$ regret$(t) \cdot I(E^\mu(t)) = 0$, and the above inequality holds trivially. So, we consider $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ holds.

We observe that if the events $E^\mu(t), E^\theta(t)$ are true, then $\Delta_{a(t)}(t) \leq g(T)(s_{t,a(t)} + s_{t,a^*(t)})$. This is because if an arm i is played at time t, then it must be true that $\theta_i(t) \geq \theta_{a^*(t)}(t)$. And, if $E^\theta(t)$ and $E^\mu(t)$ are true, then,

$$\begin{aligned}b_i(t)^T\mu &\geq \theta_i(t) - g(T)s_{t,i}\\ &\geq \theta_{a^*(t)}(t) - g(T)s_{t,i}\\ &\geq b_{a^*(t)}(t)^T\mu - g(T)s_{t,a^*(t)} - g(T)s_{t,i}.\end{aligned}$$

Therefore, given a filtration $\mathcal{F}_{t-1}$ such that $E^\mu(t)$ is true, either $\Delta_{a(t)}(t) \leq g(T)(s_{t,a(t)} + s_{t,a^*(t)})$ or $E^\theta(t)$

is false. And, hence,

$$\begin{aligned}
&\mathbb{E}\left[\text{regret}'(t) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\Delta_{a(t)}(t) \mid \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[g(T)(s_{t,a^*(t)} + s_{t,a(t)}) \mid \mathcal{F}_{t-1}\right] + \Pr\left(\overline{E^\theta(t)}\right) \\
&= g(T)\ \mathbb{E}\left[s_{t,a^*(t)} I(a(t) \in C(t)) \mid \mathcal{F}_{t-1}\right] \\
&\quad + g(T)\ \mathbb{E}\left[s_{t,a^*(t)} I(a(t) \notin C(t)) \mid \mathcal{F}_{t-1}\right] \\
&\quad + g(T)\ \mathbb{E}\left[s_{t,a(t)} \mid \mathcal{F}_{t-1}\right] + \Pr\left(\overline{E^\theta(t)}\right) \\
&\leq g(T) s_{t,a^*(t)} \Pr\left(a(t) \in C(t) \mid \mathcal{F}_{t-1}\right) \\
&\quad + g(T)\mathbb{E}\left[\left(\tfrac{g(T)}{\ell(T)}\right) s_{t,a(t)} I(a(t) \notin C(t)) \,\middle|\, \mathcal{F}_{t-1}\right] \\
&\quad + g(T)\mathbb{E}\left[s_{t,a(t)} \mid \mathcal{F}_{t-1}\right] + \tfrac{1}{T^2} \\
&\leq g(T) s_{t,a^*(t)} \cdot \tfrac{1}{p} \Pr\left(a(t) \notin C(t) \mid \mathcal{F}_{t-1}\right) + g(T)\tfrac{1}{pT^2} \\
&\quad + \left(\tfrac{2g(T)^2}{\ell(T)}\right) \mathbb{E}\left[s_{t,a(t)} \mid \mathcal{F}_{t-1}\right] + \tfrac{1}{T^2} \\
&\leq g(T) s_{t,a^*(t)} \cdot \tfrac{1}{p} \Pr\left(a(t) \notin C(t) \mid \mathcal{F}_{t-1}\right) \\
&\quad + \left(\tfrac{2g(T)^2}{\ell(T)}\right) \mathbb{E}\left[s_{t,a(t)} \mid \mathcal{F}_{t-1}\right] + \tfrac{2g(T)}{pT^2}.
\end{aligned}$$

In the first inequality we used that for all i, $\Delta_i(t) \leq 1$. The second inequality used the definition of unsaturated arms to apply $s_{t,a^*(t)} \leq \frac{g(T)}{\ell(T)} s_{t,a(t)}$, and Lemma 1 to apply $\Pr\left(\overline{E^\theta(t)}\right) \leq \frac{1}{T^2}$. The third inequality used Lemma 3, and also the observation that $0 \leq s_{t,a^*(t)} \leq ||b_{a^*(t)}(t)|| \leq 1$. □

Now, we are ready to prove Theorem 1.

Proof of Theorem 1 We observe that the absolute value of each of the four terms in the definition of X_t is bounded by $\frac{2}{p}\frac{g(T)^2}{\ell(T)}$, therefore the super-martingale Y_t has bounded difference $|Y_t - Y_{t-1}| \leq \frac{8}{p}\frac{g(T)^2}{\ell(T)}$, for all $t \geq 1$. Thus, we can apply Azuma-Hoeffding inequality (see Lemma 6 in Appendix A.2), to obtain that with probability $1 - \frac{\delta}{2}$,

$$\begin{aligned}
&\textstyle\sum_{t=1}^T \text{regret}'(t) \\
&\leq \textstyle\sum_{t=1}^T \left(\frac{g(T)}{p} I(a(t) \notin C(t)) s_{t,a^*(t)}\right) + \frac{2g(T)}{pT} \\
&\quad + \frac{2g(T)^2}{\ell(T)} \textstyle\sum_{t=1}^T s_{t,a(t)} + \frac{8}{p}\frac{g(T)^2}{\ell(T)}\sqrt{2T\ln(\frac{2}{\delta})} \\
&\leq \textstyle\sum_{t=1}^T \left(\frac{g(T)^2}{\ell(T)}\frac{1}{p} I(a(t) \notin C(t)) s_{t,a(t)}\right) + \frac{2g(T)}{pT} \\
&\quad + \frac{2g(T)^2}{\ell(T)} \textstyle\sum_{t=1}^T s_{t,a(t)} + \frac{8}{p}\frac{g(T)^2}{\ell(T)}\sqrt{2T\ln(\frac{2}{\delta})} \\
&\leq \tfrac{g(T)^2}{\ell(T)}\tfrac{3}{p} \textstyle\sum_{t=1}^T s_{t,a(t)} + \frac{2g(T)}{pT} + \frac{8}{p}\frac{g(T)^2}{\ell(T)}\sqrt{2T\ln(\frac{2}{\delta})}.
\end{aligned}$$

The second inequality used the observation that if an unsaturated arm is played, i.e., $a(t) \notin C(t)$, then, $g(T)s_{t,a(t)} \geq \ell(T)s_{t,a^*(t)}$.

Now, we can use $\sum_{t=1}^T s_{t,a(t)} \leq 5\sqrt{dT\ln T}$, which can be derived along the lines of Lemma 3 of Chu et al. (2011) using Lemma 11 of Auer (2002) (see Appendix A.5 for details). Also, recalling the definitions of $p, \ell(T)$, and $g(T)$ (see the Table of notations in the beginning of the supplementary material), and substituting in above, we get

$$\textstyle\sum_{t=1}^T \text{regret}'(t) = O\left(\frac{d^2}{\epsilon}\sqrt{T^{(1+\epsilon)}}\ln(\frac{1}{\delta})\ln(Td)\right).$$

Also, because $E^\mu(t)$ holds for all t with probability at least $1 - \frac{\delta}{2}$ (see Lemma 1), $\text{regret}'(t) = \text{regret}(t)$ for all t with probability at least $1 - \frac{\delta}{2}$. Hence, with probability $1 - \delta$,

$$\mathcal{R}(T) = \textstyle\sum_{t=1}^T \text{regret}(t) = \sum_{t=1}^T \text{regret}'(t) = O\left(\frac{d^2}{\epsilon}\sqrt{T^{(1+\epsilon)}}\ln(\frac{1}{\delta})\ln(Td)\right).$$

The proof for the alternate definition of regret mentioned in Remark 1 is provided in Appendix A.5.

4. Conclusions

Detailed concluding remarks appear in supplementary materials Sec. D.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved Algorithms for Linear Stochastic Bandits. In *NIPS*, pp. 2312–2320, 2011.

Abramowitz, Milton and Stegun, Irene A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, 1964.

Agrawal, Shipra and Goyal, Navin. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *COLT*, 2012.

Agrawal, Shipra and Goyal, Navin. Further Optimal Regret Bounds for Thompson Sampling. *AISTATS*, 2013.

Auer, Peter. Using Confidence Bounds for Exploitation-Exploration Trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

Auer, Peter, Cesa-Bianchi, Nicolò, Freund, Yoav, and Schapire, Robert E. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

Bubeck, Sébastien, Cesa-Bianchi, Nicolò, and Kakade, Sham M. Towards minimax policies for online linear optimization with bandit feedback. *Proceedings of the 25th Conference on Learning Theory (COLT)*, pp. 1–14, 2012.

Chapelle, Olivier and Li, Lihong. An Empirical Evaluation of Thompson Sampling. In *NIPS*, pp. 2249–2257, 2011.

Chapelle, Olivier and Li, Lihong. Open Problem: Regret Bounds for Thompson Sampling. In *COLT*, 2012.

Chu, Wei, Li, Lihong, Reyzin, Lev, and Schapire, Robert E. Contextual Bandits with Linear Payoff Functions. *Journal of Machine Learning Research - Proceedings Track*, 15:208–214, 2011.

Dani, Varsha, Hayes, Thomas P., and Kakade, Sham M. Stochastic Linear Optimization under Bandit Feedback. In *COLT*, pp. 355–366, 2008.

Filippi, Sarah, Cappé, Olivier, Garivier, Aurélien, and Szepesvári, Csaba. Parametric Bandits: The Generalized Linear Case. In *NIPS*, pp. 586–594, 2010.

Graepel, Thore, Candela, Joaquin Quiñonero, Borchert, Thomas, and Herbrich, Ralf. Web-Scale Bayesian Click-Through rate Prediction for Sponsored Search Advertising in Microsoft's Bing Search Engine. In *ICML*, pp. 13–20, 2010.

Granmo, O.-C. Solving Two-Armed Bernoulli Bandit Problems Using a Bayesian Learning Automaton. *International Journal of Intelligent Computing and Cybernetics (IJICC)*, 3(2):207–234, 2010.

Kaelbling, Leslie Pack. Associative Reinforcement Learning: Functions in k-DNF. *Machine Learning*, 15(3):279–298, 1994.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson Sampling: An Optimal Finite Time Analysis. *ALT*, 2012.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

Langford, John and Zhang, Tong. The Epoch-Greedy Algorithm for Multi-armed Bandits with Side Information. In *NIPS*, 2007.

May, Benedict C. and Leslie, David S. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol, 2011.

May, Benedict C., Korda, Nathan, Lee, Anthony, and Leslie, David S. Optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:01, Statistics Group, Department of Mathematics, University of Bristol, 2011.

Ortega, Pedro A. and Braun, Daniel A. Linearly Parametrized Bandits. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

Russo, Daniel and Roy, Benjamin Van. Learning to optimize via posterior sampling. *CoRR*, abs/1301.2609, 2013.

Sarkar, Jyotirmoy. One-armed badit problem with covariates. *The Annals of Statistics*, 19(4):1978–2002, 1991.

Scott, S. A modern Bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26:639–658, 2010.

Strehl, Alexander L., Mesterharm, Chris, Littman, Michael L., and Hirsh, Haym. Experience-efficient learning in associative bandit problems. In *ICML*, pp. 889–896, 2006.

Strens, Malcolm J. A. A Bayesian Framework for Reinforcement Learning. In *ICML*, pp. 943–950, 2000.

Thompson, William R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.

Woodroofe, Michael. A one-armed bandit problem with a concomitant variable. *Journal of the American Statistics Association*, 74(368):799–806, 1979.

Wyatt, Jeremy. *Exploration and Inference in Learning from Reinforcement*. PhD thesis, Department of Artificial Intelligence, University of Edinburgh, 1997.