Thompson Sampling Guided Stochastic Searching on the Line for Adversarial Learning

Sondre Glimsdal, Ole-Christoffer Granmo

Thompson Sampling Guided Stochastic Searching on the Line for Adversarial Learning

Sondre Glimsdal and Ole-Christoffer Granmo

Dept. of ICT, University of Agder, Grimstad, Norway

Abstract. The multi-armed bandit problem has been studied for decades. In brief, a gambler repeatedly pulls one out of N slot machine arms, randomly receiving a reward or a penalty from each pull. The aim of the gambler is to maximize the expected number of rewards received, when the probabilities of receiving rewards are unknown. Thus, the gambler must, as quickly as possible, identify the arm with the largest probability of producing rewards, compactly capturing the exploration-exploitation dilemma in reinforcement learning. In this paper we introduce a particular challenging variant of the multi-armed bandit problem, inspired by the so-called N-Door Puzzle. In this variant, the gambler is only told whether the optimal arm lies to the "left" or to the "right" of the one pulled, with the feedback being erroneous with probability $1 - p$. Our novel scheme for this problem is based on a Bayesian representation of the solution space, and combines this representation with Thompson sampling to balance exploration against exploitation. Furthermore, we introduce the possibility of traitorous environments that lie about the direction of the optimal arm (adversarial learning problem). Empirical results show that our scheme deals with both traitorous and non-traitorous environments, significantly outperforming competing algorithms.
Keywords: *N-Door Puzzle, Multi-armed Bandit Problem, Adversarial Learning, Bayesian Learning, Thompson Sampling*

1 Introduction

A reinforcement learning (RL) problem, at the most basic level, consists of a *learner* that interacts with a *teacher*, often also called the environment. The learner performs some action and the teacher either rewards or penalizes the learner to reinforce a particular behavior. In most cases the teacher is modeled with a stochastic component that represents noise or imperfection in the environment. Traditionally, the teacher is restricted to be *on-average* informative. That is, given that the learner performs the intended behavior, the teacher will more often than not reward the learner. This allows the learner to, over time, identify the correct behavior. However, in this paper we also consider the highly interesting scenario where the teacher may also be of a deceptive nature, giving a penalty on correct behavior instead of a reward. This leads to a novel variant of the well known exploration versus exploitation dilemma. Should the learner trust the teacher and follow the teachers advice (exploit) or should the learner try to explore the truthfulness of the teacher.

A puzzle found in the book "To Mock a Mockingbird" [**?**] illustrates this point: *"Someone was sentenced to death, but since the king loves riddles, he threw this guy into a room with two doors. One leading to death, one leading to freedom. There are two guards, each one guarding one door. One of the guards is a perfect liar, the other one will always tell the truth. The man is allowed to ask one guard a single yes-no question and then has to decide, which door to take. What single question can he ask to guarantee his freedom?"*

To avoid spoiling the puzzle for any interested reader, we omit the solution here and note that asking a double negative question will often be the correct course of action in these types of puzzles. Extending this puzzle into a general setting, lead us to the *N-door* puzzle: Here instead of simply 2 doors, the prisoner is faced with N doors, with a guard posted between each door. Now, a sole door leads to freedom and the remaining doors leads to death. At sunrise each day the prisoner is allowed to ask a single guard if the door is to the guards left or to the guards right. However, only a fixed proportion of the guards answers truthfully, the rest are compulsive liars, and since the guards are randomly assigned a position each sunrise, knowing who lies and who tells the truth is impossible. As an additional complication, depending on the mood of the king, the prisoner may be ordered to walk through one door of his choosing at an arbitrary day. Therefore, to save his life, it is imperative that the prisoner as quickly as possible zero in on the door leading to his freedom.

The N-door puzzle is a reinforcement learning puzzle where the agent (the prisoner) at each iteration chooses freely between $N-1$ guards, each stationed between two doors. However, what separate N-door from traditional RL problems is the possible existence of a deceptive teacher. Let $\pi = \frac{\#\text{truthful guards}}{\#\text{guards}}$ be the ratio of truthful guards. As the guards are randomly assigned each day the probability of a truthful answer is governed by π. If $\pi < 0.5$ then the majority of the guards are compulsive liars and the guards, as an entity are said to be *deceptive*. Conversely, if $\pi > 0$ then the majority of the guards are truthful and the the guards can be seen as *informative*. For completeness we mention that the puzzle is unsolvable for the case where π is exactly equal to $\frac{1}{2}$ as each answer is then turned into a coin flip with a fair coin and consequentially voiding any possible information gain.

Since the guards or teacher may be deceptive the traditional RL approaches is rendered ineffective, as they inherently trust the teacher to *on average* be informative. To handle the case of a *possibly* deceptive teacher any solution scheme must not only use the feedback provided by the guards to locate the door that leads to freedom but also learn the nature of the guards. Are they informative or deceptive?

This paper introduces Thompson Sampling guided Stochastic Point Search (TS-SPL) – a novel scheme for solving the N-door puzzle. A probabilistic model of all possible N-door instances is simultaneously explored and exploited as determined by Thompson Sampling (TS), gradually filtering away unlikely instances until we have identified the correct one, thus revealing the optimal decision. This means that instead of directly solving the N-door puzzle, we – in a Bayesian line

of thought – start with all possible N-door puzzles and progressively try to determine which puzzle has the highest probability of giving us the current feedback.

As we shall see, not only does this scheme handle an arbitrary level of noise, but it also exhibits better performance characteristics than the current state of art. TS-SPL also fulfill the *anytime* component of the N-door puzzle, namely that the prisoner should be able, at any time, to provide a *best* guess on what door leads to his freedom.

Stochastic Point Location

To see the applications of the N-door puzzle it is profitable to review the Stochastic Point Location (SPL) problem [**?**]. Consider an agent (algorithm or learning mechanism) that moves on a line attempting to locate a particular location λ^*. The agent can communicate with a teacher (or oracle) that informs the agent its current location λ is greater or lower than λ^*. However, the teacher is of a stochastic nature and with probability $1 - \pi$ feeds the agent erroneous feedback.

The SPL problem, due to its simplicity has been shown to have fascinating applications in meta-optimizing such as improving the performance of Multidimensional Scaling (MDS) [**?**] and estimation of non-stationary binomial distributions [**?**] as well as within routing optimization [**?**].

2 Related Work

Stochastic Searching on a Line. Baeza-Yates *et* al. [**?**] introduced the precursor to the SPL problem as *Searching for a Point on a Line* (PL). The PL problem is near identical to SPL with the obvious difference that as opposed to SPL, PL operate solely with deterministic feedback.

In his pioneering work Oommen et al. [**?**] introduced SPL as a meta-optimization algorithm. Following this paper, the general theme for investigation has been to apply various Learning Automata (LA) schemes [**? ? ?**]. And to the best of our knowledge the general state-of-art solver is Adaptive Step Searching (ASS) [**?**], although it is outperformed by the Hierarchical Stochastic Searching on the Line (HSSL)[**?**] for highly volatile non-stationary environments [**?**]. However, all the above schemes are restricted to the case where the teacher is informative ($\pi > 0.5$).

The SPL problem with a possible deceptive teacher is a magnitude harder problem than the original SPL problem and the only known solution, CPL-AdS, can be found in [**?**]. The CPL-AdS scheme is a two phase algorithm, first phase consist of using a sequence of intelligently selected questions to, with arbitrary high probability, classify the teacher as either informative or deceptive. In the second phase a regular SPL solving scheme is used where all the feedback from the teacher is negated if the first phase decided that the teacher was deceptive.

This type of scheme can be seen as a semi off-line estimator scheme where the parameters is first collected from a long sequence of observations, then given a sufficient high number of observations, estimate the best parameters. The main

drawback with this approach is the fact that the observations in phase one is not given any consideration until the parameters are ready to be estimated. And as we shall see, leads to significantly worse performance compared with schemes that operate on the data in an on-line fashion. Another major drawback is the fact that once the scheme has decided on the nature of the teacher, it is stuck with that decision. Thus the first phase must sufficiently guarantee a correct decision. Consequentially CPL-AdS scheme is not anytime due to the simple fact that it cannot guarantee us any results until it has converged.

Thompson Sampling. The Thompson Sampling (TS) principle was introduced by Thompson in 1933 [**?**], and is currently primarily employed to solve the Multi-Armed Bandit (MAB) problem.

The iconic MAB is concerned by maximizing the profit of a gambler playing slots on one armed bandits machines. The gambler knows that one of the arms pays out more than the others and wants to identify this optimal arm. However, he also does not want to waste his tokens by playing sub-optimal arms. Solving a MAB problem thus involves determining a strategy that identify the optimal arm while minimizing the loss inflicted by not playing the optimal arm. For this paper we will not be concerned with the traditional MAB problem where each arm is independent of each other, instead we will establish a global-informative MAB (GI-MAB) [**?**] where each selected arm rewards all other arms.

TS has proven to be one of the top contenders for traditional MAB problems [**? ?**], contextual MAB problems [**?**], and has also been utilized for Gaussian Process optimization [**?**] as well as a foundation for solving the Maximum a Posteriori Estimation (MAP) problem [**?**].

2.1 Contribution of the Paper

The contributions of this paper can be summarized as follows:

1. We demonstrate the efficiency of a Bayesian line of thought to progressively explore and exploit (using Thompson Sampling) a probabilistic model over all possible N-door instances instead of focusing directly on solving a single specific problem.
2. We present a novel scheme, TS-SPL for solving the N-door puzzle (and SPL) that outperforms the current state-of-art in the case of an informative teacher. If the teacher is deceptive TS-SPL outperform the state-of-art by an order of magnitude.
3. The performance of TS-SPL is also highly stable compared to the current state of art, i.e. it exhibits a low variance.
4. A secondary benefit of TS-SPL is that it not only outperform state-of-art, but in addition also estimates the degree of truthfulness of the teacher.

2.2 Paper outline

In Section 3, we present our scheme for Thompson Sampling guided Stochastic Point Search (TS-SPL). Here we start with a brief outline of the TS-SPL scheme

and a intuitive explanation on how TS-SPL works. Thereafter, we derive the probabilistic model that tracks the probabilities for which door is the correct choice, as well as the truthfulness of the guards. Using this probabilistic model as a foundation, we formulate a MAB problem that we solve using TS in such a manner that by solving the MAB problem we also solve the N-door puzzle.

Then, in Section 4, we define an ensemble of empirical experiments that we employ to evaluate the performance of TS-SPL compared to state-of-the-art. We conclude in Section 5 and present interesting venues for further work.

3 Thompson Sampling Guided Stochastic Search on the Line (TS-SPL)

The Thompson Sampling guided Stochastic Searching on the Line (TS-SPL) scheme introduced in this paper utilizes a Bayesian approach to capture the N-door puzzle. A N-door puzzle can be succinctly represented as a 2-tuple (λ^*, π) where $\lambda* \in D$ is the door leading to freedom and $\pi \in T$ is the truthfulness of the guards. Let Ω be a set containing all possible N-door puzzles. Then, assuming no a-priori information, we formulate a uniform prior over all N-door puzzles in Ω. The essence of TS-SPL is then to gradually refine the posterior over Ω by letting TS guide our questioning and then update our posterior with the resulting new information. As this process is repeated the likelihood of any other puzzle than (λ^*, π) is lowered and thus the probability of (λ^*, π) is increased. Consequently, given enough iterations TS-SPL will correctly identify λ^* as the door leading to freedom as the posterior probability of (λ^*, π) approaches unity.

The probabilistic model. The main purpose for the probabilistic model is to determine a posterior distribution over the possible N-door puzzle instances, Ω. Let the probability of the doors and the truthfulness of the guards be defined by two random variables (r.v.) denoted D and T, respectively. We here assume that D and T are independent of each other. The information we obtain from questioning the guards are represented as a set of conditionally independent r.v. $Q = \{Q_1, \ldots, Q_n\}$, each conditioned on D and T. For each question Q_k we can compute the probability of the answer (left or right) that we received from the guard as follows:

Data: guard, truthfulness $t \in T$ and door $d \in D$
Result: $P(Q_k|d, t)$ – P(left) and P(right) given the input
if *guard is left of door* **then**
  P(left | guard, door, t) $= t$
  P(right | guard, door, t) $= 1 - t$
else
  P(left | guard, door, t) $= 1 - t$
  P(right | guard, door, t) $= t$
end

Algorithm 1: How to compute the probability of a guards answer.

Applying the algorithm given above to find $P(Q|d,t)$ we derive a simple expression that find the posterior distribution given the answers. Shown here is the derivation of $P(d \in D|Q)$, however the derivation of $P(t \in T|Q)$ is analogous.

$$P(d \in D|Q) = \sum_{t \in T} P(d|Q,t)P(t) \tag{1}$$

$$= \sum_{t \in T} \frac{P(Q|d,t)P(d|t)P(t)}{P(Q|t)} \tag{2}$$

$$\propto \sum_{t \in T} P(Q|d,t)P(d|t)P(t) \tag{3}$$

$$\propto \sum_{t \in T} P(Q|d,t)P(d)P(t) \tag{4}$$

$$\propto \sum_{t \in T} \hat{Q}Q^{+}P(d)P(t) \tag{5}$$

Where $\hat{Q} = \prod_{k=1}^{n-1} P(Q_k|d,t)$ and $Q^{+} = P(Q_n|d,t)$. (2) follows directly from Bayes Theorem. (3) From marginalizing out $Q(Q|t)$. (4) From the independence of D and T, and (5) from the independence between the questions in Q. Applying memoization on Q allow us then to swiftly compute the next probability term as the number of answers in Q grow.

Multi-armed-bandit The above section describes how once we have obtained information from the guards we can turn it into probabilities for the different doors. However, as mentioned in the introduction we have trade-off between exploring different doors and zeroing in on the correct door. To handle this trade-off we model the door selection as a MAB.

To apply MAB to the N-door puzzle we enable a bijective mapping between the MAB arms and the doors found in the N-door puzzle. To decide on what arm to select for each turn we solve the MAB by utilizing the principle of TS. Here, the selection process is simply to select a random door proportional to the probability that that door is the best. Once the door has been selected we need to find what guard to ask that is standing adjacent to that door. We do this by, again, select a guard randomly proportionally to the sum of the probability of the door next to the guard on each side. So for example, if we have three doors $d_k, 0 < k < 4$, and thus three arms with probability $P(d_1) = 0.1, P(d_2) = 0.2, P(d_3) = 0.7$, then, in lieu with the TS principle these are also the probabilities of selecting that particular arm. Since selecting a door affect D, the probability of each door, we have a GI-MAB as opposed to a traditional MAB. The advantage of using TS in the selection process is evident as it provides a effective yet intuitive handling of arm selection in GI-MAB. The resulting scheme allows for more exploration in the start and gradually become increasingly greedy as the probability distribution over the arms converge.

4 Experimental results

In section we will thoroughly demonstrate the superiority of TS-SPL as compared with existing state-of-the-art. Firstly, how it outperforms the state-of-art a setting where the teacher is either informative or deceptive. Secondly, we will demonstrate how it compares with the classical informative SPL schemes. Thirdly, we will consider how the performance of TS-SPL can be enhanced by applying prior information. And finally, we will investigate to what degree TS-SPL is able in a on-line fashion to track the truthfulnesses of the teacher, before concluding with some final remarks on the behavior of TS-SPL.

The configuration of TS-SPL remains the same throughout all the different scenarios. We used a lambda set: $\lambda_0, \lambda_1, \ldots, \lambda_{200}$ evenly distributed over 0.0 - 1.0.

Regret is measured as the distance d between the point x selected and the optimal point λ_*. If the distance $|d|$ is less than $\Omega = 2^{-8}$ we clamp d to zero. This is done to allow the algorithms that operate with continues values to *pseudo converge* and facilitates easy comparison between the discrete and continues schemes.

4.1 Deceptive or informative teacher

With the underlying π taking on values in the interval $[0, 1] - \{\frac{1}{2}\}$ we test TS-SPL vs CPL-AdS[**?**] for speed of convergence and how much regret on average one accumulate before converging. However, due to the vast difference in regret between the two method a direct comparison is almost meaningless. The large regret accumulated by CPL-AdS is mainly occurring as it requires a separate phase for detecting if a environment is, informative or deceptive.

From their paper this decision phase needs approximately 200 time steps, and by this time TS-SPL is already close to converging. The other parameters in CPL-AdS the same as in their paper, LRI-$\theta = 0.8$ and LRI-$\epsilon = 0.05$.

To illustrate this point see Table 1 where it is clear that TS-SPL is superior to CPL-AdS by several orders of magnitude.

	$\pi = 0.85$	$\pi = 0.15$
TS-SPL ($\lambda^* = 0.85$)	6.2	6.2
CPL-AdS ($\lambda^* = 0.85$)	501.6 / 354.9	842.8/502.3

Table 1: The accumulative regret for TS-SPL and CPL-AdS. All entries was estimated taking the average of an ensemble of 1000 independent trials and each entry corresponds to the estimated cumulative regret using $N = 1000$ time steps, leading to a negligible variance of the estimates relative to the large difference in performance among the competing schemes. The number after the slash corresponds to the regret applied after CPL-AdS determines if the teacher is informative or deceptive.

4.2 Informative teacher

For the case where we restrict the teacher to be informative $\pi > 0.5$ we can compare TS-SPL against both the Adaptive Step Searching (ASS) [**?**] and the Hierarchical Stochastic Searching on the Line (HSSL)[**?**]. Another highly fascinating scheme is found by modifying TS-SPL to use an informative prior, as such, TS-SPL knows that $0.5 < \pi \leq 1$. We denote this modified scheme as TS-SPL-INF.

For HSSL we used a tree branching factor of $D = 8$, and for ASS we set $N_{\max} = 200$. This give the closest resolution between the tree schemes.

	$\pi = 0.55$	$\pi = 0.65$	$\pi = 0.75$	$\pi = 0.85$
TS-SPL ($\lambda^* = 0.85$)	280.1	34.3	12.5	6.2
TS-SPL-INF ($\lambda^* = 0.85$)	102.6	18.4	6.7	3.6
ASS ($\lambda^* = 0.85$)	180.3	41.3	14.9	8.1
HSSL ($\lambda^* = 0.85$)	311.7	109.8	19.9	7.1

Table 2: The cumulative regret for the competitive schemes for solving a informative SPL problem. All entries were estimated taking the average of an ensemble of 1000 independent trials and each entry corresponds to the estimated cumulative regret using $N = 1000$ time steps.

An important point to make is that in this scenario TS-SPL is configured with a uniform prior over π, so that the TS-SPL does not take advantage of the fact that π is restricted to certain values. Changing this by applying a uniform distribution over the feasible interval of $\pi = [0.5, 1.0]$ and zero elsewhere we obtain an significant increase in performance, as evident by TS-SPL-INF. This is one of the main advantages that TS-SPL present, the ability to, in a straight forward way, encode any prior information, both concerning λ^* and π. This is useful if one believe that for instance π should be close to $\frac{1}{2}$ or that one prefer a more cautious approach and thus instead of applying a uniform distribution one could for instance select a Gaussian distribution with a mean of $\frac{1}{2}$ and a suitable variance.

4.3 Tracking the truthfulness of the teacher

A interesting property of TS-SPL is its inherit ability to not only give a probability distribution over doors, but it can also give a distribution over truthfulness π. This is a great advantage as it present the end-user with a better view into the underlying environment, which can in particular be leveraged upon in the case of repeated trials, where previous trials can be us as a prior on subsequent trials, hence greatly increasing the speed of convergence. Figure 1 shows the probability for each level of noise as the TS-SPL progresses with a deceptive $\pi = 0.2$

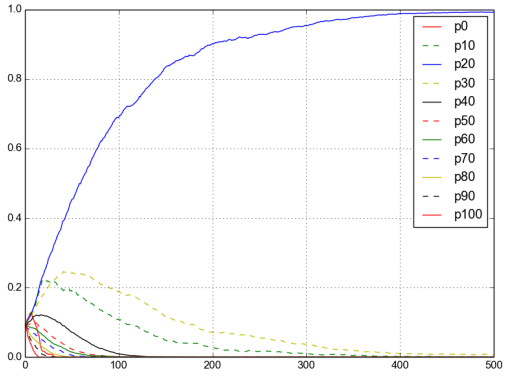

Fig. 1: TS-SPL tracks the probability π of correct feedback (y-axis). Here the true value of π is 0.2, and it is clear from the figure that TS-SPL is close to the optimal choice after only 100 iterations. The figure is based on an average of 1000 independent trials.

4.4 Stability of solution

To investigate the stability of TS-SPL, we study the variance of the regret for 1000 independent trial runs, leaving sta. Low variance means more stable performance in each trial. This property is important as it gives a higher belief that the algorithm will perform well on every occasion. For ease of comparison, the standard deviation is shown instead of the variance. Table 3 shows the results of this stability evaluation.

	$\pi = 0.55$	$\pi = 0.65$	$\pi = 0.75$	$\pi = 0.85$
TS-SPL ($\lambda^* = 0.85$)	104	20	7	2
TS-SPL-INF ($\lambda^* = 0.85$)	104	12	4	2
ASS ($\lambda^* = 0.85$)	211	70	16	4
HSSL ($\lambda^* = 0.85$)	119	80	17	7

Table 3: The standard deviation of the result of 1000 independent trials measured in regret, each trial was done using $N = 1000$ time steps.

5 Conclusions and Further Work

In this paper we investigated a novel reinforcement learning problem, inspired by the N-door puzzle. This puzzle has the fascinating property that it allows the reinforcement of the incorrect behavior, also known as a *deceptive teacher*. This renders traditional RL approaches ineffective due to their dependency on *"on average"* correct feedback, that is, an *informative teacher*.

Thus, to formulate a solution we first mapped the N-door puzzle into a Multi-Armed-Bandit (MAB) Problem, that we applied TS to solve. However, due to the non-trivial properties of the N-door puzzle, a probabilistic model was used to calculate the arm probabilities in the MAB, thus abstracting the possibility of a deceptive environment away from TS.

The resulting scheme, TS-SPL, was applied to the Stochastic Point Location (SPL) problem and demonstrated superior performance when compared to existing state-of-art solutions for a significant set of problem configurations. In the case of an informative teacher, TS-SPL slightly outperforms state-of-the-art, and for a deceptive teacher TS-SPL is an order of magnitude better. For both cases, TS-SPL demonstrates excellent stability, having remarkable low variance compared to state-of-the-art solutions.

However, the TS-SPL scheme is limited in the sense that it does not handle a non-stationary solution that some of the the informative schemes support. Further work therefore includes to investigate how we can incorporate support for non-stationary environments, such as switching environments where the teacher switches from informative to deceptive or vice versa.

Bibliography