Thompson Sampling in Adversarial Environments

Travis Dunlop

July 1, 2018

Advisors: Gergely Neu, Mihalis Markakis

Abstract

Thompson Sampling is an increasingly popular family of algorithms for decision making in online optimization. And, this is for good reason. If the loss of each action is independent and identically distributed, the external regret is bounded by $\mathcal{O}()$ (where N is the number of actions and T total time steps) []. This rivals the performance of state-of-the art follow the perturbed leader algorithms. However, what is not yet understood is it's performance if the losses are not iid. In this thesis, we provide empirical evidence that Thompson Sampling has a low bound on regret even if the losses are set by an adversary. We hypothesize that it's bounded in the order of $\mathcal{O}()$.

1 INTRODUCTION

In 1933, William R. Thompson came up with Thompson Sampling [] while researching the best way treat patitents with novel medicines. Consider a scenerio where patients with the same ailment come to a doctor over time. The doctor has a number of medicines she could prescribe. She has prescribed some of them many times before and knows how well they work. But she faces a dilema: should she try out a new medicine? It could work even better than the others, or perhaps even worse. She needs to balance the need to *explore* the possibilities with *exploiting* the treatments she knows work well. How should she prescribe medicine such that negative health outcomes are minimized?

Thompson's answer to this question is a Bayesian approach. For each of the medicines, the doctor has a prior belief on the effect. This is expressed as a probability distribution over the health outcomes. As she prescribes the medicines, her beliefs are updated. To make a decision, she samples a value from each of the distributions and chooses the one with minimum sampled loss. The probability distribution over the health outcomes captures her uncertainty about the true distribution as well as the inherent randomness of the problem. Sampling from these distributions allows for each medicine to be chosen in proportion to her belief of it's quality. Exploration is induced by chosing the appropriate prior for medicines whose effects have not yet been seen.

Thompson Sampling is known to perform well in the case where a patient's reaction to the medicines are independent from each other and identically distributed. Unfortunately, in many situations there is no guarantee that this assumption holds. Perhaps the population grows resistent to a

treatment, or environmental factors change it's efficacy. Thus, we would like to know how robust is Thompson Sampling when the data is not iid. What is common in literature, is to analyze the worst case scenerio - when the losses are not just non-iid, but they are set by an adversary who specifically tries induce poor performance. If an algorithm works well not just in easy situations, but in these adversarial cases, then it is one worth investing in.

Now we give an overview of the rest of the document. We will first give a more explicit mathematical description of the problem at hand. Then, we describe the particular flavor of Thompson Sampling we analyze. Next, we discuss other algorithms which have been used to tackle this problem. After, we recast this problem within game theory and leverage some theoretical results to estimate the regret of these algorithms. Finally, we use an evolutionary strategy to maximize the regret on these algorithms and report the results.

2 PROBLEM SETUP

We consider a version of the 'prediction with expert advice' framework from the textbook *Prediction, Learning, and Games* [2]. In this setup, there is a forecaster (analagous to the doctor) that plays a repeated game against their adversary - the environment.

First, the environment choses a loss for each action and time step $\ell_{i,t}$. Where $i \in \{1, 2, ...N\}$ corresponds with the possible actions of the forecaster and $t \in \{1, 2, ..., T\}$ is the timestep. Then, the forecaster plays the game. For each timestep t: the forecaster chooses an action $a_t \in \{1, ..., N\}$, and incurrs loss $\ell_{a_t,t}$. The losses for all actions are then revealed to the forecaster.

The forecaster tries to learn from the losses it's seen to choose good actions and the environment tries to trick the forecaster into incurring high loss. This setup can be thought of as a mulit-armed bandit problem with full information.

Note, we are cheating here compared with the story of the doctor. When the doctor chooses medicines, she only observes the effect of the medicine she chose. By assuming we have full information, we assume we know what would have happened had the doctor chosen another medicine. We do this for mathematical ease. If we are able to prove a regret bound in this easier framework, it will be a much smaller step to prove it in the harder one.

Problem Framework

Parameters: Number of actions: N, Number of timesteps: T
Environment chooses losses $\ell_{i,t} \in [0, 1]$ for $i \in \{1, 2, .., N\}$ & $t \in \{1, 2, ..., T\}$
For each timestep $t = 1, 2, ..., T$

1. Forecaster chooses action $a_t \in \{1, 2, ..., N\}$
2. Environment reveals losses for each action $\ell_{i,t}$ for $i \in \{1, 2, ..., N\}$
3. Forecaster suffers loss of chosen action $\ell_{a_t,t}$

Now, we need some way of scoring the game between the forecaster and environment. At first glance, a natural choice is the cummulative loss of the forecaster: $\widehat{L} = \sum_{t=1}^{T} \ell_{a_t,t}$. However, this gives too much power to the environment. They could simply maximize loss by choosing $\ell_{i,t} = 1$ for all actions and time steps.

A better choice of metric is regret. We compare what the forecaster did with what would have been a good action in hindsight. While there are several forms of regret, here we consider *external* regret - the difference between the forecasters loss and that of the best fixed action:

$$R_T = \widehat{L}_T - L_T^*$$

Where $L^* = \min_j \sum_{t=1}^{T} \ell_{j,t}$ is the loss of the best fixed action.

The ultimate goal of this work is to find a tight upper bound on the expected regret, maximized with respect to the losses the adversary could choose:

$$\max_{\ell} E[R_T]$$

3 THOMPSON SAMPLING

Since we are considering a framework where the losses are bounded between 0 and 1, we choose to analyze the popular Beta-Bernoulli variant of Thompson Sampling. In this case we assume that the losses, $\ell_{i,t}$, are Bernoulli distributed with probibility they are equal to one is $\theta_{i,t}$.

$$\underset{beta}{\overset{posterior}{P(\theta_{i,t}|\ell_{i,t})}} = \underset{bernoulli}{\overset{liklihood}{P(\ell_{i,t}|\theta_{i,t})}} \underset{beta}{\overset{prior}{P(\theta_{i,t})}}$$

Recall that for a random variable $\theta \sim \text{Beta}(\alpha, \beta)$, the density is $P(\theta = x) \propto x^{\alpha}(1-x)^{\beta-1}$. Here, α and β are shape parameters. The higher the value of α the more the density is shifted towards one, the higher β the more of a shift towards 0. Thus, this leads to a straightforward update rule. If we observe $\ell_{i,t} = 1$, we add one to our current estimate of α_i, if we observe $\ell_{i,t} = 0$ we add one to β_i. Of course, by default this just supports $\ell_{i,t} \in \{0, 1\}$ and not $\ell_{i,t} \in [0, 1]$. In order for the algorithm to support values between 0 and 1, we introduce a secondary loss $\widetilde{\ell}_{i,t}$ which is Bernoulli distributed according to the primary loss. Thus, the probability of updating the parameter is in proportion to the level of the observed loss. This rule is shown in item three of the boxed explanation below.

Thompson Sampling: Beta-Bernoulli

Set parameters $\alpha_i = 1$ and $\beta_i = 1$ for all $i \in \{1, ..., N\}$
For each timestep $t = 1, 2, ..., T$

1. Sample $\theta_{i,t} \sim \text{Beta}(\alpha_i, \beta_i)$ and choose action $a_t = \text{argmin}_i\, \theta_{i,t}$
2. Observe losses $\ell_{i,t}$ for $i \in \{1, 2, ..., N\}$
3. Perform Bernoulli trial for each action: $\widetilde{\ell}_{i,t} \sim \text{Bernoulli}(\ell_{i,t})$
4. Update parameters: $\alpha_i = \alpha_i + \widetilde{\ell}_{i,t}$
   $\beta_i = \beta_i + 1 - \widetilde{\ell}_{i,t}$

For a more thorough treatment of Beta-Bernoulli Thompson Sampling, please refer to []

4 OTHER ALGORITHMS

Of course, Thompson Sampling is not the only algorithm that could be used in this 'prediction with expert advice' framework. In this section we discuss

two of the more popular strategies. Both of them have provable bounds on adversarial regret.

4.1 *Follow the Perturbed Leader*

The first of these algorithms is based of the idea of 'following the leader'. That is, choose the action which has the lowest cummulative loss so far. Unfortunately, it is well known result that in this most naive implementation, one can construct losses such that the regret grows linearly with time.

To see this imagine two actions with losses (1, 0, 1/2, 0, 1/2, 0,) (1, 1/2, 0, 1/2, 0, 1/2...).

These adversarial cases inspired the creation of, follow the *perturbed* leader. That is to choose the action with the lowest cummulative so far, but perturbed by some random noise.

$$a_t = \underset{i}{\operatorname{argmin}}\, L_{i,t-1} + Z_{i,t}$$

There are many choices one could make for the distribution of this perturbation and thus many variants of the algorithm. For example, it could be uniformly or exponentially distributed, or perhaps follow a random walk, or even follow a dropout pattern. The citations and regret bounds for these variants are in the table below.

4.2 *Exponential Weighted Forecaster*

Another popular algorithm for this problem is the exponential weighted forecaster. In this scheme, the probability of choosing an action is in proportion to the exponential of the cummulative loss thus far.

$$P(a_t = i) = \frac{e^{-\eta_t L_{i,t-1}}}{\sum_{j=1}^{N} e^{-\eta_t L_{j,t-1}}}$$

Where η_t is a learning rate parameter. The main ways of varying this algorithm is by choosing different values for this parameter. Some of these rely on having more information about the system. For example the fixed learning rate $\eta = \sqrt{8(\ln N)/T}$ relies on knowing the time horizon T a priori. Ideally we want an algorithm that both has a tight regret bound with as little extra information needed as possible.

4.3 *Comparison of Algorithms*

Algorithm	Type	Adversarial Regret Bound	Citation
Exponential Weighted Average	$\eta_t = \sqrt{8(\ln N)/t}$	$\mathcal{O}(\sqrt{T \log N} + \log N)$	Section 2.3 [2]
	$\eta = \sqrt{8(\ln N)/T}$	$\mathcal{O}(\sqrt{T \log N})$	Section 2.2 [2]
	$\eta_t = \min\{1, C\sqrt{(\ln N)/\mathrm{Var}(\widehat{L}_t)}\}$	$\mathcal{O}(\sqrt{\mathrm{Var}(\widehat{L}_T) \log N} + \log N)$	Equation 13 [1]
	AdaHedge	$\mathcal{O}(\sqrt{L^* \log N} + \log N)$	[3]
Follow the Perturbed Leader	Uniform	$\mathcal{O}(\sqrt{TN})$	Corollary 4.4 [2]
	Random Walk	$\mathcal{O}(\sqrt{T \log N} + \log T)$	[4]
	Exponential	$\mathcal{O}(\sqrt{L^* \log N} + \log N)$	Corollary 4.5 [2]
	Dropout	$\mathcal{O}(\sqrt{L^* \log N} + \log N)$	[6]

5 GAME THEORY

There are deep underlying ties between analyzing regret in 'prediction with expert advice' scenerios and game theory. In this section, we highlight some of those ties as theoretical foundation for our empirical tests.

6 EVOLUTIONARY STRATEGIES

7 RELATED WORK

While there does not yet exist a theoretical bound on the regret for Thompson Sampling, there are a few results in the domain worth mentioning.

Thompson Sampling == FPL with Gaussian

Daniel Russo mention Thompson Sampling in non-stochastic environments

8 CONCLUSION

All of the code used in this thesis can be found at `github.com/TravisDunlop/thompson-sampling-thesis`

9 ACKNOWLEDGEMENT

Thank you to Gergely and Mihalis for all of there advice during this process. Also, thank you to Gabor Lugosi who, along with the other two, helped come up with the topic for this thesis.

REFERENCES

[1] N. Cesa-Bianchi, Y. Mansour, and G. Stoltz. Improved Second-Order Bounds for Prediction with Expert Advice. *ArXiv Mathematics e-prints*, February 2006.

[2] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[3] Steven de Rooij, Tim van Erven, Peter D. Grünwald, and Wouter M. Koolen. Follow the leader if you can, hedge if you must. *CoRR*, abs/1301.0534, 2013.

[4] Luc Devroye, Gábor Lugosi, and Gergely Neu. Prediction by random-walk perturbation. *CoRR*, abs/1302.5797, 2013.

[5] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *CoRR*, abs/1707.02038, 2017.

[6] Tim Van Erven, Wojciech Kotłowski, and Manfred K Warmuth. Follow the leader with dropout perturbations. In *Conference on Learning Theory*, pages 949–974, 2014.