An Adversarial Analysis of Thompson Sampling for Full-information Online Learning: from Finite to Infinite Action Spaces

Alexander Terenin (Cornell University) and Jeffrey Negrea (University of Waterloo and the Vector Institute)

Abstract

We develop a form Thompson sampling for online learning under full feedback—also known as prediction with expert advice—where the learner's prior is defined over the space of an adversary's future actions, rather than the space of experts. We show regret decomposes into regret the learner expected a priori, plus a prior-robustness-type term we call excess regret. In the classical finite-expert setting, this recovers optimal rates. As an initial step towards practical online learning in settings with a potentially-uncountably-infinite number of experts, we show that Thompson sampling over the d-dimensional unit cube, using a certain Gaussian process prior widely-used in the Bayesian optimization literature, has a $\mathcal{O}\Big(\beta\sqrt{Td\log(1+\sqrt{d}\frac{\lambda}{\beta})}\Big)$ rate against a β-bounded λ-Lipschitz adversary.

The Online Learning Game

At each time $t = 1, .., T$:

1. Learner picks a random *action* $x_t \sim p_t \in \mathcal{M}_1(X)$.
2. Adversary responds with a *reward function* $y_t : X \to \mathbb{R}$, chosen adaptively from a given function class.

Regret:

$$R(p, y) = \mathop{\mathbb{E}}_{x_t \sim p_t} \underbrace{\sup_{x \in X} \sum_{t=1}^{T} y_t(x)}_{\substack{\text{single best action} \\ \text{in hindsight}}} - \underbrace{\sum_{t=1}^{T} y_t(x_t)}_{\substack{\text{total reward of} \\ \text{learner's actions}}}.$$

Adversarial problem: seems to have nothing to do with Bayes' Rule?

Importance of no-regret algorithms:

1. Many complicated decision problems can be reduced to online learning, including certain forms of reinforcement learning.
2. Imply existence of generalization bounds.
3. Basic building block of equilibrium-computation algorithms.

Follow-the-Regularized-Leader and Online Mirror Descent

Standard approach: FTRL or OMD, the former of which chooses

$$p_t = \mathop{\arg\max}_{p \in \mathcal{M}_1(X)} \mathop{\mathbb{E}}_{x \sim p} \sum_{t=1}^{T} y_t(x) - \Gamma(p)$$

where $\Gamma : \mathcal{M}_1(X) \to \overline{\mathbb{R}}$ is a convex regularizer.

- If $X = [N]$ is a finite set, entropy is a typical choice.
- Not obvious how to choose Γ in infinite-dimensional setting.
- Similar difficulties with follow-the-perturbed-leader variants.

Thompson Sampling

Our approach: think of this game in a Bayesian way.

1. Place a prior $q^{(\gamma)}$ on the *adversary's future reward functions* $\gamma_1, .., \gamma_T \in \mathcal{M}_1(Y)$.
2. Condition on observed rewards $\gamma_1 = y_1, .., \gamma_{t-1} = y_{t-1}$.
3. Draw a sample from the posterior, and play

$$x_t = \mathop{\arg\max}_{x \in X} \sum_{\tau=1}^{t-1} y_\tau(x) + \sum_{\tau=t}^{T} \gamma_\tau(x).$$

This is a *very specific* form of follow-the-perturbed-leader, with an atypical learning rate schedule. Why should it be a good idea?

> ***Result*** *(Gravin, Peres, and Sivan, 2016). For $X = [3]$, the (infinite-horizon discounted) online learning game's Nash equilibrium takes the form of Thompson sampling, where the prior is taken to be the optimal adversary from the online learning game's maximin dual.*

Conjecture. Thompson sampling algorithms, with priors given by strong adversaries, are minimax-strong.

This work: prove the finite and simplest infinite-dimensional case.

A Bayesian-type Regret Decomposition

Approach: analyze a Bayesian algorithm in a Bayesian way

Proposition. We have

$$\underbrace{R(p, y)}_{\text{regret}} = \underbrace{R(p, q^{(\gamma)})}_{\text{prior regret}} + \underbrace{E_{q^{(\gamma)}}(p, y)}_{\text{excess regret}}.$$

Excess regret:

$$E_{q^{(\gamma)}}(p, y) = \sum_{t=1}^{T} \Gamma^*_{t+1}(y_{1:t}) - \Gamma^*_t(y_{1:t-1}) - \langle y_t \mid p_t \rangle + \mathbb{E}\langle \gamma_t \mid p_t^{(\gamma)} \rangle$$

where $\Gamma^*_t(f) = \sup_{x \in X} f(x) + \gamma_{t:T}(x)$.

- Quantifies how much adversary can exploit the learner's strategy, compared to what they expected based on the prior.

Strong Priors for Adversarial Feedback: from theory to practice

What kind of virtual adversaries make for good priors?

Equalizing adversary: same expected reward for all actions.

Strong adversary: *equalizing* and certifies sharp regret lower bound

$$R(\cdot, q^{(\gamma)}) \leq \min_p \max_q R(p, q).$$

Practical approximation: Gaussian process with matching kernel.

- Similar to priors used today in Bayesian optimization.

Numerical implementation. Requires two oracles:

1. *Gaussian process sampling oracle*: easy to implement via random Fourier features, with approximation guarantees.
2. *Optimization oracle*: implement via multi-start gradient-based methods such as LBFGS which perform well in practice.

A Bregman Divergence Bound on Excess Regret

Proposition. For an equalizing adversary, which is independent across time, and has almost surely no ties, we have

$$E_{q^{(\gamma)}}(p, y) \leq \sum_{t=1}^{T} D_{\Gamma^*_t}(y_{1:t} \,||\, y_{1:t-1})$$

where $D_{\Gamma^*_t}$ is the Bregman divergence induced by Γ^*_t.

- Connects probabilistic perspective with well-developed tools from convex analysis.
- Completely general: works essentially anywhere the algorithm and Bregman divergences are well-defined.
- Compared to prior work involving similar bounds, does not require any learning-rate-type restrictions, which usually preclude Thompson sampling.

Regret Guarantees

Finite X with ℓ^∞ adversary $R(p, q) \leq 2\sqrt{T \log N}$.
$[0, 1]^d$ with BL adversary: $R(p, q) \leq \mathcal{O}\Big(\beta\sqrt{Td\log(1+\sqrt{d}\frac{\lambda}{\beta})}\Big)$.

References

Nick Gravin, Yuval Peres, and Balasubramanian Sivan. Towards Optimal Algorithms for Prediction with Expert Advice. Symposium on Discrete Algorithms, 2016.