Adapting to Easy Data in Prediction with Limited Advice

Tobias Sommer Thune & Yevgeny Seldin

Department of Computer Science, University of Copenhagen

contact: tobias.thune@di.ku.dk

Motivation

In Online Learning we consider the problem of choosing between a number of possible actions, for instance medical treatments, trading strategies or scientific models. The learner faces a stream of problem instances, for example patients, with the goal of using the feedback from each instance to improve the decision strategy in the future.

To be able to make good decisions, the possibilities need to be explored. This however has a cost since the exploration is done by trying out different actions, even suboptimal ones. A good strategy balances this *exploration/exploitation tradeoff*.

Commonly strategies are characterised by their worst case performance. Here we consider how the learner can exploit easier learning scenarios and thereby improve their performance.

Abstract

We consider a scenario where the learner gets feedback from the chosen action and one additional action. We construct a novel algorithm that maintains the canonical worst case performance and simultaneously enjoys improved performance for two kinds of easy settings:

- Stochastic outcomes, where the outcome of each action has a constant expectation
- Arbitrary outcomes with small effective range, where the differences between the actions' outcomes are small for each problem instance

With this result we bypass the impossibility result of Gerchinovitz and Lattimore [2016] and improve on a similar result by Cesa-Bianchi and Shamir [2017] by relaxing the assumptions.

Setting

Prediction with limited advice models sequential decision processes as a repeated game, where a learner in each round chooses an action A_t out of K possible actions. The associated loss $\ell_t^{A_t}$ is revealed to and suffered by the learner. The learner then chooses a second action B_t and observes its loss, but this is not suffered.

$$\begin{array}{cccccc} \ell_1^1, & \ell_2^1, & \cdots & \ell_t^1, & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \\ \ell_1^a, & \ell_2^a, & \cdots & \ell_t^a, & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \\ \ell_1^K, & \ell_2^K, & \cdots & \ell_t^K, & \cdots \end{array}$$

$$\longrightarrow \text{time}$$

Fig. 1: Each round a decision $a = 1, \ldots, K$ is made and the loss of the decision is suffered.

The performance of the learner is measured by the *expected regret* of the decisions made in the first T rounds, compared to the best action in hindsight:

$$\mathcal{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^a\right]. \tag{1}$$

Easy data

We consider two models for the losses of each action:

Stochastic losses The first type of easiness the learner can exploit is the restriction that the losses are generated I.I.D with expectation $\mathbb{E}[\ell_t^a] = \mu_a$ for all t. Denoting the best action by $a^\star$, the setting can be described in terms of *suboptimality gaps* $\Delta_a := \mu_a - \mu_{a^\star}$.

Effective loss range The second type of easiness is that the losses have a small range within each round. We define the *effective loss range*, denoted ε, such that for every round t and actions a, a' we have $|\ell_t^a - \ell_t^{a'}| \leq \varepsilon$.

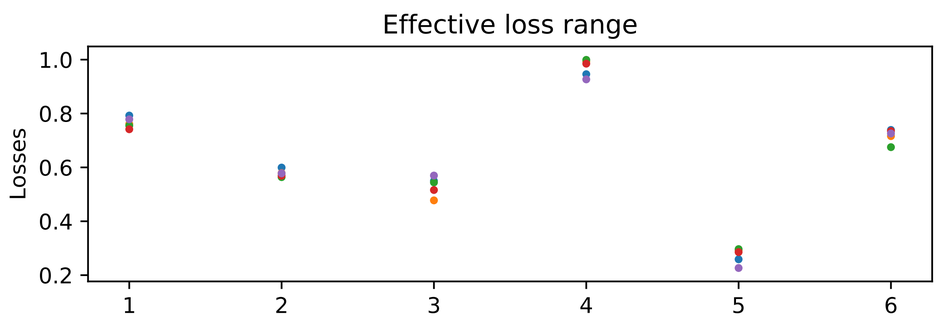

Fig. 2: Illustration of the effective loss range. The losses span the entire unit interval, but are clustered in each round.

The surprising impossibility result of Gerchinovitz and Lattimore [2016] shows that it is impossible to achieve a regret which is linear in ε if we only observe the played action.

Approach

The key ingredient in our approach is the use of importance weighted *difference estimators*:

$$\widetilde{\Delta\ell}_t^a = (K-1)\mathbb{1}(a = B_t)\left(\ell_t^{B_t} - \ell_t^{A_t}\right), \tag{2}$$

and their cumulative first and second moments:

$$D_t(a) := \sum_{s=1}^{t} \widetilde{\Delta\ell}_s^a, \quad S_t(a) := \sum_{s=1}^{t} \left(\widetilde{\Delta\ell}_s^a\right)^2. \tag{3}$$

We use an *exponential weights* algorithm based on these difference estimators instead of the losses themselves. These allow us to consider not just the loss of an action itself, but the relative loss of the action within the round, "anchoring" the estimated losses in each round based on the second chosen action. Intuitively this increases the "resolution" at which the losses are compared. The main played action A_t is chosen randomly with the probability of $A_t = a$ being

$$p_t^a = \frac{\exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)}{\sum_{a=1}^{K} \exp\left(-\eta_t D_{t-1}(a) - \eta_t^2 S_{t-1}(a)\right)}. \tag{4}$$

Algorithm 1: Second Order Difference Adjustments (SODA)

input: Learning rate scheme η_t with $\eta_t \leq (2\varepsilon(K-1))^{-1}$

Set $\boldsymbol{p_1}$ uniform over the arms, $\boldsymbol{p_1} = (1/K, \ldots, 1/K)$.

for $t = 1, 2, \ldots$ **do**

- Draw A_t according to $\boldsymbol{p_t}$;
- Draw B_t uniformly at random from the remaining actions $[K] \setminus \{A_t\}$;
- Observe $\ell_t^{A_t}, \ell_t^{B_t}$ and suffer $\ell_t^{A_t}$;
- Construct $\widetilde{\Delta\ell}_t^a$ by equation (2);
- Update $D_t(a), S_t(a)$ by equation (3);
- Define $\boldsymbol{p_{t+1}}$ by equation (4);

end

The learning rate used in the results below is $\eta_t = \sqrt{\frac{\ln K}{\max_a S_{t-1}(a) + (K-1)^2}}$.

Results

The following two theorems show that our algorithm can adapt to both kinds of easiness, while maintaining the worst case performance.

Theorem 1 *For arbitrary loss sequences with effective loss range ε, the expected regret of* SODA *satisfies*

$$\mathcal{R}_T \leq \mathcal{O}\left(\varepsilon\sqrt{(K-1)T\ln K}\right).$$

Note that a lower bound of $\inf\sup \mathcal{R}_T \geq \mathcal{O}\left(\varepsilon\sqrt{KT}\right)$ holds, which is an extension of the lower bound in Seldin et al. [2014].

Theorem 2 *For stochastic loss sequences with gaps $\Delta_a \leq \varepsilon$, the expected regret of* SODA *satisfies*

$$\mathcal{R}_T \leq \sum_{a:\Delta_a>0} \mathcal{O}\left(\frac{K\varepsilon^2}{\Delta_a}\right).$$

An important point is that the two theorems hold *simultaneously*.

Conclusion

We have introduced a novel algorithm that adapts to two kinds of easiness simultaneously, while being robust to worst case data. The improved performance on easy data means that the algorithm is more suited for real life applications, where the data rarely represents the worst case. This adaptivity comes only at the expense of a single additional observation in each round.

References

Nicolò Cesa-Bianchi and Ohad Shamir. Bandit regret scaling with the effective loss range. Technical report, https://arxiv.org/abs/1705.05091, 2017.

Sébastien Gerchinovitz and Tor Lattimore. Refined lower bounds for adversarial bandits. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Yevgeny Seldin, Peter L. Bartlett, Koby Crammer, and Yasin Abbasi-Yadkori. Prediction with limited advice and multiarmed bandits with paid observations. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

We thank Julian Zimmert for valuable discussion and feedback during this project.