Advanced Topics in Machine Learning 2017-2018

Yevgeny Seldin Christian Igel Tobias Sommer Thune Julian Zimmert

Home Assignment 5

Deadline: Tuesday, 10 October 2017, 23:59

The assignments must be answered individually - each student must write and submit his/her own solution. We encourage you to work on the assignments on your own, but we do not prevent you from discussing the questions in small groups. If you do so, you are requested to list the group partners in your individual submission.

Submission format: *Please, upload your answers in a single* `.pdf` *file and additional* `.zip` *file with all the code that you used to solve the assignment. (The* `.pdf` *should* ***not*** *be part of the* `.zip` *file.)*

IMPORTANT: *We are interested in how you solve the problems, not in the final answers. Please, write down the calculations and comment your solutions.*

1 Introduction of New Products (33 points)

Imagine that we have an established product on the market, which sells with probability 0.5. We have received a new product which sells with an unknown probability μ. Assume that at every sales round we can offer only one product, so we have the choice between offering the established or the new product. Propose a strategy for maximizing the number of sales and analyze its pseudo-regret. Write your answer in terms of the gap $\Delta = 0.5 - \mu$. (Separate between positive and negative gap.)

2 Tighter analysis of the Hedge algorithm (33 points)

Apply Hoeffding's lemma (Lemma 1.5 in Yevgeny's lecture notes) in order to derive a better parametrization and a tighter bound on the expected regret of the Hedge algorithm. Guidance:

1. Traverse the analysis of the Hedge algorithm that we did in class. There will be a place where you will have to bound expectation of an exponent of a function ($\sum_a p_t(a)e^{-\eta X_t^a}$). Instead of going the way we did, apply Hoeffding's inequality.

2. Find the value of η that minimizes the new bound. (You should get $\eta = \sqrt{\frac{8 \ln N}{T}}$ - please, prove this formally.)

3. At the end you should obtain $\mathbb{E}[R_T] \leq \sqrt{\frac{1}{2} T \ln N}$. (I.e., you will get an improvement by a factor of 2 compared to what we did in class.)

Remark: Note that the regret upper bound matches the lower bound up to a constant. This is an extremely rare case.

3 Empirical comparison of FTL and Hedge (34 points)

Assume that you have to predict a binary sequence $X_1, X_2, \ldots$ and that you know that X_i-s are i.i.d. Bernoulli variables with unknown bias μ. (You know that X_i-s are i.i.d., but you do not know the value of μ.) On every round you can predict "0" or "1" (i.e., you have two actions - "predict 0" or "predict 1") and your loss is the zero-one loss depending on whether your prediction matches the outcome. The regret is computed with respect to the performance of the two possible actions.

1. Write a simulation that compares numerically the performance of Follow The Leader (FTL) algorithm with performance of the Hedge algorithm that we presented in class (with $\eta = \sqrt{\frac{2 \ln N}{T}}$) and the performance of reparametrized Hedge algorithm from Question 2 (with $\eta = \sqrt{\frac{8 \ln N}{T}}$). The Hedge algorithm should operate with the aforementioned two actions. To make things more interesting we will add an "anytime" version of Hedge to the comparison. "Anytime" algorithm is an algorithm that does not depend on the time horizon. Let t be a running time index ($t = 1, \ldots, T$). Anytime Hedge corresponding to the simple analysis uses $\eta_t = \sqrt{\frac{\ln N}{t}}$ and anytime Hedge corresponding to the tighter analysis in Question 2 uses $\eta_t = 2\sqrt{\frac{\ln N}{t}}$ (the learning rate η_t of anytime Hedge changes with time and does not depend on the time horizon). Some instructions for the simulation:

   - Take time horizon $T = 1000$. (In general, time horizon should be large in comparison to $\frac{1}{(\mu - \frac{1}{2})^2}$.)
   - Test several values of μ. We suggest $\mu = \frac{1}{2} - \frac{1}{4}$, $\mu = \frac{1}{2} - \frac{1}{8}$, $\mu = \frac{1}{2} - \frac{1}{16}$.
   - Plot the pseudo regret of the five algorithms with respect to the best out of "0" and "1" actions as a function of t for the different values of μ (make a separate plot for each μ). Make 10 runs of each algorithm and report the average pseudo regret over the 10 runs and the average pseudo regret + one standard deviation over the 10 runs. Do not forget to add a legend to your plot.

2. Which values of μ lead to a higher regret? How the relative performance of the algorithms evolves with time and does it depend on μ?

3. Design an adversarial (non-i.i.d.) sequence on which you expect the FTL algorithm to perform poorly. Explain the design of your adversarial sequence and report a plot with a simulation, where you compare the performance of FTL with the different versions of Hedge. As before, make 10 repetitions of the experiment and report the average regret (in this case you should use regret and not pseudo regret) and the average + one standard deviation. Comment on your observations.

4 [Optional, not for submission] Decoupling exploration and exploitation in i.i.d. multiarmed bandits

Assume an i.i.d. multiarmed bandit game, where the observations are not coupled to the actions. Specifically, we assume that on each round of the game the player is allowed to observe the reward of a single arm, but it does not have to be the same arm that was played on that round (and if it's not the same arm, the player does not observe his own reward, but he observes the reward of an alternative option).

Derive a playing strategy and a regret bound for this game. (You should solve this problem for a general K and you should get that the regret does not grow with time.)

Remark: note that in this setting the exploration is "free", because we do not have to play suboptimal actions in order to test their quality. And if we contrast this with the standard multiarmed bandit setting we observe that the regret stops growing with time instead of growing logarithmically with time. Actually, the result that you should get is much closer to the regret bound for FTL with full information than to the regret bound for multiarmed bandits. Thus, it is not the fact that we have just a single observation that makes i.i.d. multiarmed bandits harder than full information games, but the fact that this single observation is linked to the action. (In adversarial multiarmed bandits the effect of decoupling is more involved (Avner et al., 2012, Seldin et al., 2014).)

Good luck!
Yevgeny, Christian, Tobias & Julian

References

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

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.