Advanced Topics in Machine Learning 2015-2016

Yevgeny Seldin Christian Igel Brian Brost

Home Assignment 6

Deadline: Sunday, 25 October, 2015, 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 your 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 all your calculations.*

Question 1 (Regret bound for i.i.d. full information games - 40 points)**.** Follow the leader (FTL) is a playing strategy that on round t plays the action that was most successful up to round t ("the leader"). (This is the strategy that you used in Question 2 in Home Assignment 4. Remember that this strategy does not work in adversarial environments.) Derive a regret bound for FTL in i.i.d. full information games with K possible actions and outcomes bounded in $[0,1]$ interval (you can work with rewards or losses, as you like). You can use the following guidelines:

1. You are allowed to solve the problem for $K = 2$. (The guidelines are not limited to $K = 2$.)
2. Let $\mu(a)$ be expected reward of action a and let $\hat{\mu}_t(a)$ be empirical estimate of the reward of action a on round t (the average of rewards observed so far). Let a^* be an optimal action (there may be more than one optimal action, but then things only get better [convince yourself that this is true], so we can assume that there is a single a^*). Let $\Delta(a) = \mu(a^*) - \mu(a)$. FTL plays $a \neq a^*$ on rounds t for which $\hat{\mu}_{t-1}(a) \geq \hat{\mu}_{t-1}(a^*)$. So let us analyze how often this may happen.
3. Construct Hoeffding's confidence bounds on the deviation of $\hat{\mu}_t(a)$ from $\mu(a)$. Use the following form of the inequality $\mathbb{P}\left\{\hat{\mu}_t(a) - \mu(a) \geq \sqrt{\frac{\ln\frac{1}{\delta_t}}{2t}}\right\} \leq \delta_t$ (check yourself which side you need for which action).
4. Note that when the width of confidence intervals for a and a^* is smaller than $\frac{1}{2}\Delta(a)$ we can distinguish between a and a^* unless one of the confidence intervals fails, which happens with probability at most δ_t.
5. Note that δ_t is a free parameter that we can set any way we like. So lets set it in such a way that the width of the confidence interval will be $\frac{1}{2}\Delta(a)$.
6. Note that δ_t cannot be larger than 1, so you should get an initial period in the game, where confidence intervals are not under control. What is the length of this period?
7. Now if we put everything together - we have an initial period during which we have no control over the estimates and after the initial period we can get a bound on the expected number of rounds when confidence intervals fail. Combine the two results to get a bound on the expected number of times $\mathbb{E}[N_T(a)]$ that a is played and use it to get a bound on $\mathbb{E}[R_T]$. Overall, you should get a bound of a form $\mathbb{E}[R_T] \leq \sum_{a:\Delta(a)>0}\left(\frac{c_1}{\Delta(a)} + \frac{c_2}{1-\exp(\Delta(a)^2/2)}\Delta(a)\right)$, where c_1 and c_2 are constants. *Note that in the full information i.i.d. setting the regret does not grow with time!!!* (Since the bound is independent of T.)

Question 2 (Decoupling exploration and exploitation in i.i.d. multiarmed bandits - 20 points)**.** 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 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 in Question 1 than to the regret bound for multiarmed bandits. Thus, it is not the fact that we have just 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 involved (Avner et al., 2012, Seldin et al., 2014).)

Question 3 (The value of additional observations in i.i.d. multiarmed bandits - 5 points)**.** Assume an i.i.d. multiarmed bandit game, where on each round of the game the player is allowed to observe the reward of two arms - the arm that was played and any single additional arm. Derive a playing strategy and a regret bound for this game. (Hint: If you reduce this problem to the previous one, the solution is immediate.)

Question 4 (Empirical comparison of different algorithms for i.i.d. multiarmed bandits - 35 points)**.** Implement and compare the performance of UCB1 and EXP3 in i.i.d. multiarmed bandit setting. You can use the following settings:

- Time horizon $T = 10000$.
- Take a single best arm with bias $\mu^* = 0.5$.
- Take $K - 1$ suboptimal arms for $K = 2, 4, 8, 16$.
- For suboptimal arms take $\mu = \mu^* - \frac{1}{4}$, $\mu = \mu^* - \frac{1}{8}$, $\mu = \mu^* - \frac{1}{16}$ (3 different experiments with all suboptimal arms being equally bad).
- [Optional] Repeat the experiments with $\mu^* = 0.9$.

Make 10 repetitions of each experiment and for each experiment plot the average regret (over the 10 repetitions) as a function of time and the average regret + one standard deviation (over the 10 repetitions). Remember that in i.i.d. setting the regret is defined as $\mathbb{E}[R_t] = \sum_{s=1}^{t} \mathbb{E}[N_t(a)]\,\Delta(a)$, so each time an algorithm plays a suboptimal action a it accumulates $\Delta(a)$ regret.

You are welcome to try other settings - if you modify the experiments explain clearly what you do.

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 *JMLR Workshop and Conference Proceedings*, volume 32 (ICML), 2014.

Good luck!
Yevgeny, Christian, & Brian