Regret Bounds for Switching Bandits

Ronald Ortner

Montanuniversität Leoben

Delta Meeting Liège
30 April 2019

Outline

Overview WP 3 (Exploration)

- **Task 3.1:**
  RL algorithms for changing environments (M1–M12)
- **Task 3.2:**
  Open-ended exploration in changing environments (M11–M24)
- **Task 3.3:**
  Incorporating state space partitions into exploration (M18–32)

Overview WP 3 (Exploration)

- **Task 3.1:**
  RL algorithms for changing environments (M1–M12)
- **Task 3.2:**
  Open-ended exploration in changing environments (M11–M24)
- **Task 3.3:**
  Incorporating state space partitions into exploration (M18–32)

Plans from Kickoff-Meeting

Task 3.1:
RL algorithms for changing environments (M1–M12) :

Plans for gradually changing environments:

- Give more weight to more recent experience
  (instead of complete restart):
  - Sliding window
  - Discounted averages
- Attainable bounds will depend on changes.
- What are suitable models for gradual changes?
- When are $\sqrt{T}$ bounds possible?

Plans from Kickoff-Meeting

Task 3.1:
RL algorithms for changing environments (M1–M12) :

Plans for gradually changing environments:

- Give more weight to more recent experience (instead of complete restart):
  - Sliding window (LLARLA Workshop Best Paper)
  - Discounted averages
- Attainable bounds will depend on changes.
- What are suitable models for gradual changes?
- When are $\sqrt{T}$ bounds possible?
- What if number of changes is not known? (COLT 2019)

Outline

Setting

Setting for multi-armed bandit problem with changes:

- Horizon T
- Reward distributions may change change L times up to step T.

The regret in this setting can be defined as

$$\sum_{t=1}^{T} \left(\mu_t^* - r_t\right),$$

where μ_t^* is the optimal mean reward at step t.

Previous Work

- Upper bounds of $\tilde{O}(\sqrt{LT})$ for algorithms which use number of changes L:
  - Garivier& Moulines, ALT 2011
  - Allesiardo et al, IJDSA 2017
- Lower bound of $\Omega(\sqrt{LT})$, which holds even when L is known.
- Results for two arms (EWRL 2018)

AdSwitch for K arms (Sketch)

For episodes ($\approx$ estimated changes) $\ell = 1, 2, \ldots$ do:

- Let the set *GOOD* contain all arms.
- Select all arms in *GOOD* alternatingly.
- Remove bad arms from *GOOD*.
- Sometimes sample discarded arms not in *GOOD* (to be able to check for changes).
- Check for changes (of all arms). If a change is detected, start a new episode.

Regret Bound for AdSwitch

W.h.p. the algorithm

- will identify the bad arms,
- will detect significant changes, while the overhead for additional sampling is not too large,
- will make no false detections of a change.

Regret Bound for AdSwitch

W.h.p. the algorithm

- will identify the bad arms,
- will detect significant changes, while the overhead for additional sampling is not too large,
- will make no false detections of a change.

Theorem

The regret of AdSwitch in a switching bandit problem with K arms and L changes is at most

$$O\big(\sqrt{(L+1)KT(\log T)}\big).$$

Outline

Variational Bounds

- Regret Bound depends on the number of changes L.
- For gradual changes this is a bad model, as one can have in principle changes at every time step.
- An alternative measure for gradual changes could be the variation of the changes:

$$V := \sum_t \max_{a \in A} \left| \mu_{t+1}(a) - \mu_t(a) \right|$$

  would be the variation of a bandit problem with arm set A and mean $\mu_t(a)$ of arm a at step t.

Variational Bounds: Previous Work

Besbes et al. (NIPS 2014) consider variational bounds for bandit problems with changes:

- They show lower bound on regret of
$$\Omega\left((KV)^{1/3}T^{2/3}\right).$$
- They propose an algorithm based on EXP3 with restarts and show regret bound of
$$\tilde{O}\left((KV)^{1/3}T^{2/3}\right).$$
- **Note:** Algorithm knows and uses V to set restart times.

Variational Bounds from L-dependent Bounds

Slightly adapting AdSwitch one can guarantee that a new episode $\ell + 1$ starts only when there is a significant change in variation V_ℓ of current episode ℓ, that is, w.h.p.

$$V_\ell \geq \sqrt{\frac{\ell K \log T}{T}}\,. \tag{1}$$

Variational Bounds from L-dependent Bounds

Slightly adapting AdSwitch one can guarantee that a new episode $\ell + 1$ starts only when there is a significant change in variation V_ℓ of current episode ℓ, that is, w.h.p.

$$V_\ell \geq \sqrt{\frac{\ell K \log T}{T}}\,. \tag{1}$$

Rewriting (1) gives

$$\sqrt{\ell} \leq V_\ell \sqrt{\frac{T}{K \log T}},$$

Variational Bounds from L-dependent Bounds

Slightly adapting AdSwitch one can guarantee that a new episode $\ell + 1$ starts only when there is a significant change in variation V_ℓ of current episode ℓ, that is, w.h.p.

$$V_\ell \geq \sqrt{\frac{\ell K \log T}{T}} . \tag{1}$$

Rewriting (1) gives

$$\sqrt{\ell} \leq V_\ell \sqrt{\frac{T}{K \log T}},$$

and summing up over episodes we get

$$\sum_{\ell=1}^{L} \sqrt{\ell} \leq V \sqrt{\frac{T}{K \log T}}.$$

Variational Bounds from L-dependent Bounds

Slightly adapting AdSwitch one can guarantee that a new episode $\ell + 1$ starts only when there is a significant change in variation V_ℓ of current episode ℓ, that is, w.h.p.

$$V_\ell \geq \sqrt{\frac{\ell K \log T}{T}}\,. \tag{1}$$

Rewriting (1) gives

$$\sqrt{\ell} \leq V_\ell \sqrt{\frac{T}{K \log T}},$$

and summing up over episodes we get

$$L^{3/2} \approx \sum_{\ell=1}^{L} \sqrt{\ell} \leq V \sqrt{\frac{T}{K \log T}}.$$

Variational Bounds from L-dependent Bounds

Now from

$$L^{3/2} \leq V \sqrt{\frac{T}{K \log T}}.$$

we have

$$\sqrt{L} \leq V^{1/3} \left(\frac{T}{K \log T} \right)^{1/6}.$$

Variational Bounds from L-dependent Bounds

Now from

$$L^{3/2} \leq V \sqrt{\frac{T}{K \log T}}.$$

we have

$$\sqrt{L} \leq V^{1/3} \left(\frac{T}{K \log T} \right)^{1/6}.$$

Plugging this into our regret bound we finally get a regret bound of

$$\begin{aligned} \sqrt{LKT \log T} &\leq V^{1/3} \left(\frac{T}{K \log T} \right)^{1/6} \sqrt{KT \log T} \\ &= V^{1/3} T^{2/3} (K \log T)^{1/3} \end{aligned}$$

Variational Bounds from L-dependent Bounds

- Thus, we obtain a regret bound of $V^{1/3}T^{2/3}$.
- This is best possible (Besbes et al, NIPS 2014).
- Unlike in (Besbes at al, NIPS 2014), this has been achieved without knowing the variation V in advance.
- Another COLT 2019 paper of Y. Chen, C. Lee, H. Luo, and C. Wei that is based on our EWRL paper for the two-arms-case considers contextual bandits and subsumes our results.