Progress Report MUL

Ronald Ortner

Montanuniversität Leoben

Delta Meeting Lille
4 Oct 2018

Outline

1. Introduction
2. Sliding Window UCRL
3. Tracking the Best Arm in Switching Bandit Problem
4. Variational Regret Bounds
5. Open Questions / Future Work

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? (EWRL)

Outline

1 Introduction

2 Sliding Window UCRL

3 Tracking the Best Arm in Switching Bandit Problem

4 Variational Regret Bounds

5 Open Questions / Future Work

Setting

Setting for RL with changes:

- Horizon T
- MDP is allowed to change ℓ times up to step T.
- All MDPs the learner has to deal with have diameter bounded by D.

The regret in this setting can be defined as

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

where ρ_t^* is the optimal average reward of the MDP the learner acts on at step t.

UCRL Summary

- Optimistic UCRL algorithm (Jaksch et al., 2010)
- Regret bounds of $O(DS\sqrt{AT})$ for UCRL in MDPs with S states, A actions and diameter D

UCRL in Changing Environments

Idea to deal with up to ℓ (possibly abrupt) changes:

Restart UCRL every $\tau := \left(\frac{T}{\ell}\right)^{2/3}$ steps.

UCRL in Changing Environments

Idea to deal with up to ℓ (possibly abrupt) changes:

Restart UCRL every $\tau := \left(\frac{T}{\ell}\right)^{2/3}$ steps.

Why it works / Regret Bound:

- In ℓ periods in which MDP changes the regret is at most $\ell \cdot \left(\frac{T}{\ell}\right)^{2/3} = \ell^{1/3} T^{2/3}$.
- In the other $\ell^{2/3} T^{1/3}$ periods the regret is bounded by $\ell^{2/3} T^{1/3} \cdot \left(\frac{T}{\ell}\right)^{1/3} = \ell^{1/3} T^{2/3}$.

From Standard UCRL ...

Now, instead of restarts, we want to use a sliding window.

UCRL (Auer, Jaksch, Ortner 2008 & 2010)

For episodes $k = 1, 2, \ldots$ do:

1. Maintain UCB-like confidence intervals for rewards and transition probabilities to define set of plausible MDPs $\mathbb{M}$.
2. Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi),$$

   where $\rho(\mathcal{M}, \pi)$ is the average reward of policy π in MDP $\mathcal{M}$.
3. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled.

Sliding Window UCRL

Input: Window size W

For episodes $k = 1, 2, \ldots$ do:

1. Maintain UCB-like confidence intervals for rewards and transition probabilities to define set of plausible MDPs $\mathbb{M}$ computed from the previous W steps.
2. Calculate optimal policy $\tilde{\pi}$ in optimistic model $\tilde{\mathcal{M}} \in \mathbb{M}$, i.e.

$$\rho(\tilde{\mathcal{M}}, \tilde{\pi}) = \max_{\pi, \mathcal{M} \in \mathbb{M}} \rho(\mathcal{M}, \pi),$$

   where $\rho(\mathcal{M}, \pi)$ is the average reward of policy π in MDP $\mathcal{M}$.
3. Execute $\tilde{\pi}$ until the visits in some state-action pair have doubled (when compared to the number visits in the previous W steps from the episode start).

Regret Analysis:

- In episodes with a change (either in the estimation window or the episode itself), we lose at most the episode length, which is $\leq W$.
  $\rightsquigarrow$ Respective total regret: $O(\ell W)$.

Regret Analysis for Sliding Window UCRL

Regret Analysis:

- In episodes with a change (either in the estimation window or the episode itself), we lose at most the episode length, which is $\leq W$.
  $\rightsquigarrow$ Respective total regret: $O(\ell W)$.
- In any other episode of length τ we obtain by UCRL bound that regret is bounded by $\tilde{O}(DS\sqrt{A\tau})$.

Regret Analysis:

- In episodes with a change (either in the estimation window or the episode itself), we lose at most the episode length, which is $\leq W$.
  $\rightsquigarrow$ Respective total regret: $O(\ell W)$.
- In any other episode of length τ we obtain by UCRL bound that regret is bounded by $\tilde{O}(DS\sqrt{A\tau})$.
- As there are $\tilde{O}(\frac{SAT}{W})$ episodes and the lengths sum up to T, one gets by Jensen inequality that the respective total regret is

$$\tilde{O}\left(DS\sqrt{A}\cdot\sqrt{T\cdot\frac{SAT}{W}}\right)=\tilde{O}\left(DS^{3/2}T\sqrt{\frac{A}{W}}\right)$$

Regret Bound for Sliding Window UCRL

A bit more sophisticated analysis gets rid of factor $\sqrt{S}$:

Theorem

In an MDP with S states, A actions, diameter D and ℓ changes, with probability of at least $1-\delta$ the regret of SW-UCRL with window size W after T steps is bounded by

$$\tilde{O}\left(\ell W + DST\sqrt{\frac{A}{W}}\right).$$

Optimizing the window size W gives:

Choosing $W = \left(\frac{T}{\ell}\right)^{2/3}$ one obtains regret

$$\tilde{O}\left(\ell^{1/3} T^{2/3}\right)$$

just as for UCRL with restarts.

Outline

1 Introduction

2 Sliding Window UCRL

3 Tracking the Best Arm in Switching Bandit Problem

4 Variational Regret Bounds

5 Open Questions / Future Work

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 $\tilde{O}(\sqrt{\ell T})$ for algorithms which use number of changes ℓ:
  - Garivier& Moulines, ALT 2011
  - Allesiardo et al, IJDSA 2017
- Lower bound of $\Omega(\sqrt{\ell T})$, which holds even when ℓ is known.

Our Algorithm (for two arms)

Algorithm for unknown ℓ:

AdSwitch for two arms (Sketch)

For episodes $k = 1, 2, \ldots$ do:

- **Estimation phase**:
  Select both arms are selected alternatingly,
  until better arm has been identified.
- **Exploitation and checking phase**:
  - Mostly exploit the empirical best arm.
  - Sometimes sample both arms to check for change.
    If a change is detected then start a new episode.

Our Algorithm

AdSwitch for two arms

For episodes $k = 1, 2, \ldots$ do:

- **Estimation phase**:
  Sample both arms alternatingly in rounds $n = 1, 2, 3, ...$ until $\left|\hat{\mu}_1 - \hat{\mu}_2\right| > \sqrt{\frac{C_1 \log T}{n}}$. Set $\hat{\Delta} := \hat{\mu}_1 - \hat{\mu}_2$.
- **Exploitation and checking phase**:
  - Let $d_i = 2^{-i}$ and $I_k = \max\{i : d_i \geq \hat{\Delta}\}$.
  - Randomly choose i from $\{1, 2, \ldots, I_k\}$ with probabilities $d_i \sqrt{\frac{k+1}{T}}$.
  - With remaining probability choose empirically best arm and repeat phase.
  - If an i is chosen, sample both arms alternatingly for $2 \left\lceil \frac{C_2 \log T}{d_i{}^2} \right\rceil$ steps to check for changes of size d_i:
    If $\hat{\mu}_1 - \hat{\mu}_2 \notin \left[\hat{\Delta} - \frac{d_i}{4}, \hat{\Delta} + \frac{d_i}{4}\right]$, then start a new episode.

Regret Bound for AdSwitch

W.h.p. the algorithm

- will identify the better arm in the exploration phase,
- will detect significant changes in the exploitation phase, 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 two arms and ℓ changes is at most

$$O\Big((\log T)\sqrt{(\ell+1)T}\Big).$$

AdSwitch for K arms

AdSwitch for K arms (Sketch)

For episodes $k = 1, 2, \ldots$ do:

- Let the set A^+ of active arms contain all arms.
- Select all arms in A^+ alternatingly.
- Remove bad arms from A^+.
- Sometimes sample discarded arms not in A^+ to check for change. If a change is detected, start a new episode.

AdSwitch for K arms

AdSwitch for K arms (Sketch)

For episodes $k = 1, 2, \ldots$ do:

- Let the set A^+ of active arms contain all arms.
- Select all arms in A^+ alternatingly.
- Remove bad arms from A^+.
- Sometimes sample discarded arms not in A^+ to check for change. If a change is detected, start a new episode.

We expect this algorithm to achieve $O\left(K(\log T)\sqrt{(\ell + 1)T}\right)$ regret.

Outline

1. Introduction
2. Sliding Window UCRL
3. Tracking the Best Arm in Switching Bandit Problem
4. Variational Regret Bounds
5. Open Questions / Future Work

Variational Bounds

- Regret Bounds presented so far depend 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

How to obtain variational from ℓ-dependent bounds (two arms case):

- Assume $\Delta := |\mu_t(1) - \mu_t(2)|$ remains the same over all time steps.

Variational Bounds from ℓ-dependent Bounds

How to obtain variational from ℓ-dependent bounds (two arms case):

- Assume $\Delta := |\mu_t(1) - \mu_t(2)|$ remains the same over all time steps.
- Then $V = \ell\Delta$, where ℓ is the number of changes and

$$\ell = \frac{V}{\Delta}. \tag{1}$$

Variational Bounds from ℓ-dependent Bounds

How to obtain variational from ℓ-dependent bounds (two arms case):

- Assume $\Delta := |\mu_t(1) - \mu_t(2)|$ remains the same over all time steps.
- Then $V = \ell\Delta$, where ℓ is the number of changes and

$$\ell = \frac{V}{\Delta}. \tag{1}$$

- We have a regret bound of order $\sqrt{\ell T} \leq T\Delta$, so that

$$\Delta \geq \sqrt{\frac{\ell}{T}}, \text{ or equivalently } \frac{1}{\Delta} \leq \sqrt{\frac{T}{\ell}}$$

Variational Bounds from ℓ-dependent Bounds

How to obtain variational from ℓ-dependent bounds (two arms case):

- Assume $\Delta := |\mu_t(1) - \mu_t(2)|$ remains the same over all time steps.
- Then $V = \ell\Delta$, where ℓ is the number of changes and

$$\ell = \frac{V}{\Delta}. \tag{1}$$

- We have a regret bound of order $\sqrt{\ell T} \leq T\Delta$, so that

$$\Delta \geq \sqrt{\frac{\ell}{T}}, \text{ or equivalently } \frac{1}{\Delta} \leq \sqrt{\frac{T}{\ell}}$$

- With (1) we get

$$\ell = \frac{V}{\Delta} \leq V\sqrt{\frac{T}{\ell}} \text{ and hence } \ell \leq V^{2/3} T^{1/3}.$$

Variational Bounds from ℓ-dependent Bounds

How to obtain variational from ℓ-dependent bounds (two arms case):

- Assume $\Delta := |\mu_t(1) - \mu_t(2)|$ remains the same over all time steps.
- Then $V = \ell\Delta$, where ℓ is the number of changes and

$$\ell = \frac{V}{\Delta}. \tag{1}$$

- We have a regret bound of order $\sqrt{\ell T} \leq T\Delta$, so that

$$\Delta \geq \sqrt{\frac{\ell}{T}}, \text{ or equivalently } \frac{1}{\Delta} \leq \sqrt{\frac{T}{\ell}}$$

- With (1) we get

$$\ell = \frac{V}{\Delta} \leq V\sqrt{\frac{T}{\ell}} \text{ and hence } \ell \leq V^{2/3}T^{1/3}.$$

- It follows that the regret is bounded by

$$\sqrt{\ell T} \leq V^{1/3}T^{2/3}.$$

Variational Bounds from ℓ-dependent Bounds

- Thus, we obtain 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.

Definition of Variation in Changing MDPs

For RL in MDPs one may consider defining the variation

$$V := \sum_t \max_{\pi:S \to A} \left| \rho_{t+1}(\pi) - \rho_t(\pi) \right|$$

via the average rewards $\rho_t(\pi)$ of policies π at step t.

Definition of Variation in Changing MDPs

For RL in MDPs one may consider defining the variation

$$V := \sum_t \max_{\pi : S \to A} \left| \rho_{t+1}(\pi) - \rho_t(\pi) \right|$$

via the average rewards $\rho_t(\pi)$ of policies π at step t.

However, this does not work:
The mean reward of a policy might change little or not at all, while the underlying rewards and transition probabilities may change a lot.
The variation would be small, but the learning effort large.

Definition of Variation in Changing MDPs

Thus, we have to define variation "bottom-up" via rewards and transition probabilities:

$$
\begin{aligned}
V^r &:= \sum_t \max_{s,a \in S \times A} \left| r_{t+1}(s,a) - r_t(s,a) \right| \\
V^p &:= \sum_t \max_{s,a \in S \times A} \left\| p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a) \right\|
\end{aligned}
$$

Definition of Variation in Changing MDPs

Thus, we have to define variation "bottom-up" via rewards and transition probabilities:

$$
\begin{aligned}
V^r &:= \sum_t \max_{s,a \in S \times A} \left| r_{t+1}(s,a) - r_t(s,a) \right| \\
V^p &:= \sum_t \max_{s,a \in S \times A} \left\| p_{t+1}(\cdot|s,a) - p_t(\cdot|s,a) \right\|
\end{aligned}
$$

Perturbation bounds for MDPs show that variations V^r and V^p result in variation $\leq V^r + D \cdot V^p$ with respect to the average reward of any policy.

Variational Bounds for RL in Changing MDPs

- Can again use UCRL with restarts after any $\frac{T^{2/3}}{V^{2/3}}$ steps with $V := V^r + D \cdot V^p$.
- Respective regret is bounded by $V^{1/3} T^{2/3}$.

Outline

1. Introduction
2. Sliding Window UCRL
3. Tracking the Best Arm in Switching Bandit Problem
4. Variational Regret Bounds
5. Open Questions / Future Work

Future Work in Task T3.1

- Meaningful experiments comparing UCRL with restarts to SW-UCRL
- Generalize AdSwitch to K arms
- Generalize variational bounds to K arms and arbitrary gaps
- Variational bounds for SW-UCRL
- Lower bounds