Stochastic Bandits with Changing Reward Distributions

Peter Auer[1], Ronald Ortner[1] and Pratik Gajane[1,2]

[1]Montanuniversität Leoben

[2]chist-era project DELTA
Austrian Science Fund (FWF): I 3437

Multi Armed Bandit Workshop
London 26 Sep 2019

Switching Bandit Setting

Stochastic multi-armed bandit problem with changes

- Set of arms $\{1, \ldots, K\}$.
- Learner chooses arm a_t at steps $t = 1, 2, \ldots, T$.
- Learner receives random reward $r_t \in [0, 1]$ with (unknown) mean $\mathbb{E}[r_t] = \mu_t(a_t)$.
- **Note:** The mean rewards $\mu_t(a)$ depend on time t.

Regret Definition

We define the **regret** in this setting as

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

where $\mu_t^* := \max_a \mu_t(a)$ is the optimal mean reward at step t.

Note: We compete against the policy that keeps track of the best arm!

The regret will depend on how the reward distributions change:

▷ We consider the number of changes L,
i.e., the number of times when $\mu_{t-1}(a) \neq \mu_t(a)$ for some a.

Previous Work

When the number of changes L is known:

- Upper bounds of $\tilde{O}(\sqrt{KLT})$ for algorithms which use number of changes L:
  - EXP3.S (Auer et al., SIAM J. Comput. 2002)
  - Garivier& Moulines, ALT 2011
  - Allesiardo et al, IJDSA 2017
- Lower bound of $\Omega(\sqrt{KLT})$, which holds even when L is known.

Why Knowledge of L Helps

Sampling rate for inferior arms:

- Assume an inferior arm a is Δ-worse than the best arm.
- To detect a change of arm a sample a with probability $p = \sqrt{L/(KT)}/\Delta$:

  When there is no change:

  - Each sample contributes Δ to the regret, which results in sampling costs of $pT\Delta = \sqrt{LT/K}$.
  - Summing over all inferior arms, this contributes $\sqrt{KLT}$ to the regret.

Why Knowledge of L Helps

Sampling rate for inferior arms:

- Assume an inferior arm a is Δ-worse than the best arm.
- To detect a change of arm a sample a with probability $p = \sqrt{L/(KT)}/\Delta$:

  When arm a changes by $\epsilon > \Delta$:

  - In this case, $\approx 1/\epsilon^2$ samples of a are sufficient to detect the change.
  - Hence the change is detected after $1/(p\epsilon^2)$ time steps, and the respective regret is at most

$$\epsilon/(p\epsilon^2) = \Delta\sqrt{KT/L}/\epsilon < \sqrt{KT/L}.$$

  - Summing over the changes gives a regret contribution of $\sqrt{KLT}$.

Algorithm sketch for two arms

Idea: Try to detect changes and use the respective current estimate for L to set the sample probability for bad arm.

ADSWITCH for two arms (Sketch)

For episodes (≈ estimated changes) $\ell = 1, 2, \ldots$ 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.
  - W. prob. $\frac{\sqrt{(\ell+1)/T}}{\Delta}$ sample bad arm to check for change of size* Δ.
    If a change is detected then start a new episode.

Algorithm sketch for two arms

AdSwitch for two arms (Sketch)

For episodes ($\approx$ estimated changes) $\ell = 1, 2, \ldots$ 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.
  - W. prob. $\frac{\sqrt{(\ell+1)/T}}{\Delta}$ sample bad arm to check for change of size* Δ.
    If a change is detected then start a new episode.

* Since we do not know the size of the change Δ,
we have to check for a change of different values of Δ !

Algorithm AdSwitch for two arms

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

- **Estimation phase**:
  Sample both arms alternatingly in rounds $n = 1, 2, 3, \ldots$ until $\left|\hat{\mu}_1 - \hat{\mu}_2\right| > \sqrt{\frac{C_1 \log T}{n}}$. Set $\hat{\Delta} := \hat{\mu}_1 - \hat{\mu}_2$.
- **Checking and exploitation phase**:
  - Let $d_i = 2^{-i}$ and $I_\ell = \max\{i : d_i \geq \hat{\Delta}\}$.
  - Randomly choose i from $\{1, 2, \ldots, I_\ell\}$ with probabilities $d_i \sqrt{\frac{\ell+1}{T}}$.
  - 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.
  - With remaining probability choose empirically best arm and repeat phase.

Regret Bound

Theorem (EWRL 2018)

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

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

Facts about the Algorithm

W.h.p. the algorithm

- will identify the better arm in the exploration phase,
- will make no false detections of a change,
  i.e. there are at most L episodes.

For the regret analysis we have to show that the algorithm will detect significant changes in the exploitation phase, while the overhead for additional sampling is not too large,

Regret Analysis

The regret can be decomposed into

1. regret from steps in the exploration phase,
2. regret from exploitation or checking when there are no or just small changes,
3. regret from exploitation or checking when there are large changes.

❶ Regret in the Exploration Phase

Consider τ consecutive steps with no change in the exploration phase of some episode ℓ.

Let Δ be the true gap during these steps.

- $\frac{c \log T}{\Delta^2}$ samples are sufficient to detect a gap of size Δ, i.e.

$$\tau \leq \frac{c \log T}{\Delta^2}.$$

- Regret in these τ steps is $\leq \max\left\{\frac{c \log T}{\Delta}, \tau\Delta\right\} \leq \sqrt{c\tau \log T}$
- Since there are at most $2L + 1$ such intervals of consecutive steps with no change in an episode, summing over these intervals bounds the respective regret by $\sqrt{cT(2L+1) \log T}$.

2 Regret for Sampling with small or no changes

Next, we consider τ_ℓ steps in an episode ℓ when $\left|\hat{\mu}_i - \mu_i\right| \leq \frac{\hat{\Delta}}{4}$.

- Then $|\mu_1 - \mu_2| \leq \frac{3\hat{\Delta}}{2}$.
- The expected regret for sampling is hence bounded by

$$\begin{aligned}
& c' \cdot \frac{3\hat{\Delta}}{2}\tau_\ell \sum_i \left(d_i \sqrt{\tfrac{\ell+1}{T}} \right) \frac{\log T}{{d_i}^2} \\
= \quad & c' \cdot \frac{3\hat{\Delta}}{2}\tau_\ell (\log T) \sqrt{\tfrac{\ell+1}{T}} \sum_i \frac{1}{d_i} \\
\leq \quad & c' \cdot \frac{3\hat{\Delta}}{2}\tau_\ell (\log T) \sqrt{\tfrac{\ell+1}{T}} \cdot \frac{2}{\hat{\Delta}}
\end{aligned}$$

- Summing over all episodes gives a bound of $c''(\log T)\sqrt{(L+1)T}$.

3 Regret for Sampling with large changes

Finally, we consider the remaining steps in the exploitation phase when $\left|\hat{\mu}_i - \mu_i\right| > \frac{\hat{\Delta}}{4}$.

We analyse intervals $[a_j, b_j]$ of τ_j consecutive steps with no change.

Short intervals:

- If $\tau_j \leq c\frac{\log T}{\Delta^2}$, then $\Delta \leq c'\sqrt{\frac{\log T}{\tau_j}}$.
- Hence the regret in $[a_j, b_j]$ is bounded by $\Delta\tau_j \leq c'\sqrt{(\log T)\tau_j}$.
- Summing over all short intervals gives a regret contribution of $c'\sqrt{(\log T)LT}$.

3 Regret for Sampling with large changes

Finally, we consider the remaining steps in the exploitation phase when $\left|\hat{\mu}_i - \mu_i\right| > \frac{\hat{\Delta}}{4}$.
We analyse intervals $[a_j, b_j]$ of τ_j consecutive steps with no change.

Long intervals:

- If $\tau_j > c\frac{\log T}{\Delta^2}$, then a change will be detected w.h.p. as soon as a check for a change of size Δ is done.
- Such a check is done at each step with probability $\Delta\sqrt{\frac{\ell+1}{T}}$.
- In expectation this takes $\frac{1}{\Delta}\sqrt{\frac{T}{\ell+1}}$ steps with resp. regret of $\sqrt{\frac{T}{\ell+1}}$.
- Summing over all long intervals gives regret contribution $c\sqrt{TL}$.

Algorithm AdSwitch for K arms (Sketch)

Main problem for generalization from 2 to K arms:

- Cannot separate exploration from exploitation/checking phase.
- $\rightsquigarrow$ need to interweave these phases:

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 a 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.

Algorithm AdSwitch for K arms (Sketch)

- Cannot separate exploration from exploitation/checking phase.
- $\rightsquigarrow$ need to interweave these phases:

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 a 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.

Algorithm ADSWITCH (Sketch with more details)

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

- Let the set GOOD contain all arms.
- Select all arms in $\text{GOOD} \cup \mathcal{S}$ alternatingly.
- Remove bad arms a from GOOD.
  Keep in mind empirical eviction gaps $\hat{\Delta}(a)$.
- Sometimes sample discarded arms not in GOOD:
  - Define set $\mathcal{S}$ of arms $a \notin \text{GOOD}$ to be sampled.
  - At each step t, each $a \notin \text{GOOD}$, for $d_i \approx \hat{\Delta}(a), 2\hat{\Delta}(a), 4\hat{\Delta}(a), \ldots$, with probability $d_i\sqrt{\ell/(KT)}$ add a to $\mathcal{S}$.
  - Keep a in $\mathcal{S}$ until it has been sampled $1/d_i^2$ times.
- Check for changes (of all arms).
  If a change is detected, start a new episode.

Regret Bound for AdSwitch

Theorem (COLT 2019)

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

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

Facts about the Algorithm

By standard confidence intervals, w.h.p. the algorithm

- will only remove suboptimal arms from GOOD,
- will make no false detections of a change,
  i.e. there are at most L episodes.

Regret decomposition

"Horizontal" regret decomposition:

The regret at each step t can be decomposed as:

$$\mu_t^* - \mu_t(a_t) = \mu_t^* - \max_{a \in \mathrm{GOOD}_t} \mu_t(a) + \max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t)$$

Note: At steps where optimal arm is in GOOD the first term is 0.

Regret w.r.t. best arm in GOOD

"Vertical" regret decomposition:

We decompose all time steps t in an episode ℓ into the following categories:

1. Time steps t when a_t is in GOOD:

- Considering intervals $[a_i, b_i]$ with no changes, in each interval the regret is bounded by the sum over the confidence intervals in each step, which gives regret of $\tilde{O}(\sqrt{b_i - a_i})$.
- Summing over all intervals and episodes gives a regret contribution of $\tilde{O}(\sqrt{KLT})$.

Regret w.r.t. best arm in GOOD

“Vertical” regret decomposition:

We decompose all time steps t in an episode ℓ into the following categories:

1. Time steps t when a_t is in GOOD. $\surd$
2. Time steps t when a_t is not in GOOD, and $\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t) \lesssim \hat{\Delta}$:

- An arm like a_t is only sampled when checking for changes.
- The regret analysis is similar to the two arms case for sampling with no or small changes and gives a contribution of $\tilde{O}(\sqrt{KLT})$.

Regret w.r.t. best arm in GOOD

"Vertical" regret decomposition:

We decompose all time steps t in an episode ℓ into the following categories:

1. Time steps t when a_t is in GOOD. $\surd$
2. Time steps t when a_t is not in GOOD, and $\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t) \lesssim \hat{\Delta}$. $\surd$
3. Time steps t when a_t is not in GOOD, and $\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t) > \hat{\Delta}$:

- If the reward for a_t has decreased significantly since its eviction from GOOD, it cannot be played often before detecting the change.
- Otherwise, the best arm in GOOD has been significantly improved. The regret until this is detected is controlled by the confidence intervals for checking changes.

Regret w.r.t. best arm in GOOD

"Vertical" regret decomposition:

We decompose all time steps t in an episode ℓ into the following categories:

1. Time steps t when a_t is in GOOD. $\surd$
2. Time steps t when a_t is not in GOOD, and $\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t) \lesssim \hat{\Delta}$. $\surd$
3. Time steps t when a_t is not in GOOD, and $\max_{a \in \mathrm{GOOD}_t} \mu_t(a) - \mu_t(a_t) > \hat{\Delta}$:

The respective regret is bounded again by $\tilde{O}(\sqrt{KLT})$. $\surd$

Regret when optimal arm is not in GOOD

Finally, we consider the distance $\mu_t^* - \max_{a \in \mathrm{GOOD}_t} \mu_t(a)$.

Let $\tilde{\mu}(a_t^*)$ be the estimate for a_t^* at the time of eviction.

"Vertical" regret decomposition:

We decompose all time steps t in an episode ℓ into the following two categories:

1. Time steps t when $a_t^* \notin \mathrm{GOOD}_t$ and $\mu_t^* \lesssim \tilde{\mu}(a_t^*) + \hat{\Delta}(a_t^*)$:

- This can only happen when the mean of the best arm has dropped significantly.
- The regret till this change is noticed can be bounded by the employed confidence intervals and is bounded by $\tilde{O}(\sqrt{KLT})$.

Regret when optimal arm is not in GOOD

Finally, we consider the distance $\mu_t^* - \max_{a \in \mathrm{GOOD}_t} \mu_t(a)$.

Let $\tilde{\mu}(a_t^*)$ be the estimate for a_t^* at the time of eviction.

"Vertical" regret decomposition:

We decompose all time steps t in an episode ℓ into the following two categories:

1. Time steps t when $a_t^* \notin \mathrm{GOOD}_t$ and $\mu_t^* \lesssim \tilde{\mu}(a_t^*) + \hat{\Delta}(a_t^*)$. $\surd$
2. Time steps t when $a_t^* \notin \mathrm{GOOD}_t$ and $\mu_t^* > \tilde{\mu}(a_t^*) + \hat{\Delta}(a_t^*)$:

- In this case, the mean of a_t^* has significantly increased.
- One has to bound the regret until this change is noticed.
- The analysis is similar to the case in the two arms setting when large changes have occurred.

Regret when optimal arm is not in GOOD

Finally, we consider the distance $\mu_t^* - \max_{a \in \text{GOOD}_t} \mu_t(a)$.

Let $\tilde{\mu}(a_t^*)$ be the estimate for a_t^* at the time of eviction.

“Vertical” regret decomposition:

We decompose all time steps t in an episode ℓ into the following two categories:

1. Time steps t when $a_t^* \notin \text{GOOD}_t$ and $\mu_t^* \lesssim \tilde{\mu}(a_t^*) + \hat{\Delta}(a_t^*)$. $\surd$
2. Time steps t when $a_t^* \notin \text{GOOD}_t$ and $\mu_t^* > \tilde{\mu}(a_t^*) + \hat{\Delta}(a_t^*)$:

The regret in this case is bounded by $\tilde{O}(\sqrt{KLT})$ as well. $\surd$

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

Assume you have an episodic algorithm with $\tilde{O}(\sqrt{KLT})$ regret that starts a new episode $\ell + 1$ only when there is a significant change in variation V_ℓ of current episode ℓ, that is, w.h.p.

$$V_\ell \geq \sqrt{\frac{\ell K \log T}{T}} . \qquad (1)$$

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

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

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.
- A COLT 2019 paper of Y. Chen, C. Lee, H. Luo, and C. Wei based on our EWRL paper for the two-arms-case considers contextual bandits and subsumes our results.

Extensions to the Adversarial Case: Setting

- In adversarial case one usually competes against the best fixed arm in hindsight.
- (Auer et al., SIAM J. Comput. 2002) consider regret against the best strategy that changes arm at most S times.
- Algorithm EXP3.S (a variant of EXP3) gives
  - regret $\tilde{O}(S\sqrt{KT})$,
  - regret $\tilde{O}(\sqrt{SKT})$ if algorithm is tuned w.r.t. S.

Can $\tilde{O}(\sqrt{SKT})$ regret be obtained w.r.t. any S for untuned algorithm?

Extensions to the Adversarial Case: Setting

Can $\tilde{O}(\sqrt{SKT})$ regret be obtained w.r.t. any S for untuned algorithm?

Note:
There is an optimal S maximizing

$$R_S^* - c\sqrt{SKT \log T},$$

where R_S^* is the reward of best T-step strategy with S arm changes.

Extensions to the Adversarial Case: Algorithm

Can $\tilde{O}(\sqrt{SKT})$ regret be obtained w.r.t. any S for untuned algorithm?

What might an algorithm look like?

- We need to count changes (i.e., check when it pays off to switch).
- Sampling itself could be done as by ADSWITCH.
- However, detecting a change is hard.
- Maybe one can use something like EXP3.P ?