Tutorial on Bandit

Subhojyoti Mukherjee

IIT Madras

February 13, 2017

Overview

- The bandit problem is a sequential decision making process where at each timestep we have to choose one action or arm from a set of arms.

Introduction

- The bandit problem is a sequential decision making process where at each timestep we have to choose one action or arm from a set of arms.
- After say pulling each arm once we are presented with an *exploration-exploitation* trade-off, that is whether to continue to pull the arm for which we have observed the highest estimated reward till now(exploitation) or to explore a new arm(exploration).

Introduction

- The bandit problem is a sequential decision making process where at each timestep we have to choose one action or arm from a set of arms.
- After say pulling each arm once we are presented with an *exploration-exploitation* trade-off, that is whether to continue to pull the arm for which we have observed the highest estimated reward till now(exploitation) or to explore a new arm(exploration).
- If we become too greedy and always exploit we may miss the chance of actually finding the optimal arm and get stuck with a sub-optimal arm.

Why study bandits at all?

- We all know of ϵ-greedy [Sutton and Barto(1998)] algorithm, we can simply stick to it.

Why study bandits at all?

- We all know of ϵ-greedy [Sutton and Barto(1998)] algorithm, we can simply stick to it.
- But ϵ-greedy only gives us an asymptotic guarantee. There is no guarantee that in a highly regressive environment how ϵ-greedy will behave. Can we be better in our search?

Why study bandits at all?

- We all know of ϵ-greedy [Sutton and Barto(1998)] algorithm, we can simply stick to it.
- But ϵ-greedy only gives us an asymptotic guarantee. There is no guarantee that in a highly regressive environment how ϵ-greedy will behave. Can we be better in our search?
- Bandits allows us to study this behavior in a more formal way giving us strict guarantees regarding the performance of our algorithm.

Why study bandits at all?

- We all know of ϵ-greedy [Sutton and Barto(1998)] algorithm, we can simply stick to it.
- But ϵ-greedy only gives us an asymptotic guarantee. There is no guarantee that in a highly regressive environment how ϵ-greedy will behave. Can we be better in our search?
- Bandits allows us to study this behavior in a more formal way giving us strict guarantees regarding the performance of our algorithm.
- They form the linking pieces of a larger problem.

Why study bandits at all?

- We all know of ϵ-greedy [Sutton and Barto(1998)] algorithm, we can simply stick to it.
- But ϵ-greedy only gives us an asymptotic guarantee. There is no guarantee that in a highly regressive environment how ϵ-greedy will behave. Can we be better in our search?
- Bandits allows us to study this behavior in a more formal way giving us strict guarantees regarding the performance of our algorithm.
- They form the linking pieces of a larger problem.
- They are easy to implement.

Some practical applications

- Selecting the best channel (out of several existing channels) for mobile communications in a very short duration.

Some practical applications

- Selecting the best channel (out of several existing channels) for mobile communications in a very short duration.
- Selecting a small set of best workers (out of a very large pool of workers) whose productivity is above a threshold.

Some practical applications

- Selecting the best channel (out of several existing channels) for mobile communications in a very short duration.
- Selecting a small set of best workers (out of a very large pool of workers) whose productivity is above a threshold.
- Selecting the best possible route for a message to pass through in a peer-to-peer network connection.

Stochastic Multi-Armed Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a finite set of actions or arms.

Stochastic Multi-Armed Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a finite set of actions or arms.
- The rewards for each of the arms drawn from distributions are identical and independent random variables.

Stochastic Multi-Armed Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a finite set of actions or arms.
- The rewards for each of the arms drawn from distributions are identical and independent random variables.
- The learner does not know the mean of the distributions, denoted by r_i.

Stochastic Multi-Armed Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a finite set of actions or arms.
- The rewards for each of the arms drawn from distributions are identical and independent random variables.
- The learner does not know the mean of the distributions, denoted by r_i.
- The learner has to find the optimal arm the mean of whose distribution is denoted by r^* such that $r^* > r_i, \forall i \in A$.

Stochastic Multi-Armed Bandit Problem

- In stochastic multi-armed bandit problem we are presented with a finite set of actions or arms.
- The rewards for each of the arms drawn from distributions are identical and independent random variables.
- The learner does not know the mean of the distributions, denoted by r_i.
- The learner has to find the optimal arm the mean of whose distribution is denoted by r^* such that $r^* > r_i, \forall i \in A$.
- The distributions for each of the arms are fixed throughout the time horizon.

Basic Notations

- Goal: To minimize Regret

Basic Notations

- Goal: To minimize Regret
- Average reward of best action is r^* and any other action i as r_i. There are K total actions. $T_i(n)$ is number of times tried action i is executed till n-timesteps.

Basic Notations

- Goal: To minimize Regret
- Average reward of best action is r^* and any other action i as r_i. There are K total actions. $T_i(n)$ is number of times tried action i is executed till n-timesteps.
- Cumulative Regret: The loss we suffer because of not pulling the optimal arm till the total number of timesteps T.

$$R_T = r^* T - \sum_{i \in A} r_i T_i(T),$$

Basic Notations

- Goal: To minimize Regret
- Average reward of best action is r^* and any other action i as r_i. There are K total actions. $T_i(n)$ is number of times tried action i is executed till n-timesteps.
- Cumulative Regret: The loss we suffer because of not pulling the optimal arm till the total number of timesteps T.

$$R_T = r^* T - \sum_{i \in A} r_i T_i(T),$$

- The expected regret of an algorithm after T rounds can be written as

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \mathbb{E}[T_i(T)]\Delta_i,$$

- $\Delta_i = r^* - r_i$ denotes the gap between the means of the optimal arm and of the i-th arm.

Another Notion of Regret

- Goal: To minimize Regret

Another Notion of Regret

- Goal: To minimize Regret
- Can we have a policy which achieves the minimum regret among all the possible environments available?

Another Notion of Regret

- Goal: To minimize Regret
- Can we have a policy which achieves the minimum regret among all the possible environments available?
- This is called the worst case gap-independent regret or sometimes called the minimax regret.

Another Notion of Regret

- Goal: To minimize Regret
- Can we have a policy which achieves the minimum regret among all the possible environments available?
- This is called the worst case gap-independent regret or sometimes called the minimax regret.
- It is generally found by setting all the gaps to equal values of order $O\left(1/\sqrt{T}\right)$.

Literature Survey

- A considerable amount of research has been conducted on SMAB. We can divide the gamut of literature on SMAB into broadly two categories:

Literature Survey

- A considerable amount of research has been conducted on SMAB. We can divide the gamut of literature on SMAB into broadly two categories:
  - **Frequentist approach:** In this approach for each of the arms compute a dynamic allocation index that depends only on the number of draws on the arm and choose the arm with the maximal index. Eg: *UCB* variants

Literature Survey

- A considerable amount of research has been conducted on SMAB. We can divide the gamut of literature on SMAB into broadly two categories:
  - **Frequentist approach:** In this approach for each of the arms compute a dynamic allocation index that depends only on the number of draws on the arm and choose the arm with the maximal index. Eg: *UCB* variants
  - **Bayesian Approach:** In this approach you start with a prior guess over the performance of each of the arms, then you pull an arm and based on the reward received we update our posterior guess on the performance of the arm.

Literature Survey

- A considerable amount of research has been conducted on SMAB. We can divide the gamut of literature on SMAB into broadly two categories:
  - **Frequentist approach:** In this approach for each of the arms compute a dynamic allocation index that depends only on the number of draws on the arm and choose the arm with the maximal index. Eg: *UCB* variants
  - **Bayesian Approach:** In this approach you start with a prior guess over the performance of each of the arms, then you pull an arm and based on the reward received we update our posterior guess on the performance of the arm.
- We will be focusing on UCB based approaches in our work.

Literature Survey

- The literature under UCB can be broadly classified into three categories:

Literature Survey

- The literature under UCB can be broadly classified into three categories:
  - **Mean-based estimation:** In this approach at every timestep we choose an arm based on $\hat{r}_i$ and its confidence interval c_i. Eg: UCB1 [Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer], MOSS [Audibert and Bubeck(2009)], UCB-Improved [Auer and Ortner(2010)]

Literature Survey

- The literature under UCB can be broadly classified into three categories:
  - **Mean-based estimation:** In this approach at every timestep we choose an arm based on $\hat{r}_i$ and its confidence interval c_i. Eg: UCB1 [Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer], MOSS [Audibert and Bubeck(2009)], UCB-Improved [Auer and Ortner(2010)]
  - **Mean and Variance based Estimation:** Here, at every timestep we choose an arm based on $\hat{r}_i$, $\hat{V}_i$ and its confidence interval c_i. Eg: UCB-Norma [Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer]l, UCB-V [Audibert et al.(2009)Audibert, Munos, and Szepesvári].

Literature Survey

- The literature under UCB can be broadly classified into three categories:
  - **Mean-based estimation:** In this approach at every timestep we choose an arm based on $\hat{r}_i$ and its confidence interval c_i. Eg: UCB1 [Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer], MOSS [Audibert and Bubeck(2009)], UCB-Improved [Auer and Ortner(2010)]
  - **Mean and Variance based Estimation:** Here, at every timestep we choose an arm based on $\hat{r}_i$, $\hat{V}_i$ and its confidence interval c_i. Eg: UCB-Norma [Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer]l, UCB-V [Audibert et al.(2009)Audibert, Munos, and Szepesvári].
  - **Divergence based methods:** Eg: KL-UCB [Garivier and Cappé(2011)], DMED [Honda and Takemura(2010)].

UCB1 Algorithm ([Auer et al.(2002a)Auer, Cesa-Bianchi, and Fischer])

Algorithm 1 UCB1

1: Pull each arm once
2: **for** $t = K + 1, ..., T$ **do**
3: Pull the arm such that $\max_{i \in A} \left\{ \hat{r}_i + \sqrt{\frac{2 \log t}{s_i}} \right\}$
4: $t := t + 1$
5: **end for**

- Maintain an upper confidence bound (c_i) for each of the arms
- This c_i will help in sufficiently exploring sub-optimal arms and then exploiting the optimal arm.
- The gap-independent regret bound of $O\left(\sqrt{KT \log T}\right)$ and gap-dependent bound of $O\left(\frac{K \log(T)}{\Delta}\right)$.

Minimax Optimal Strategy in the Stochastic Case ([Audibert and Bubeck(2009)])

Algorithm 2 MOSS

1: Pull each arm once
2: **for** $t = K+1, ..., T$ **do**
3: Pull the arm such that $\max_{i \in A} \left\{ \hat{r}_i + \sqrt{\frac{\max\{0, \log(\frac{T}{Ks_i})\}}{s_i}} \right\}$
4: $t := t + 1$
5: **end for**

- UCB1 suffers from a worst case regret of $O\left(\sqrt{KT \log T}\right)$.
- MOSS corrects this and gives us a gap-independent regret bound of $O\left(\sqrt{KT}\right)$ and gap-dependent bound of $O\left(\frac{K^2 \log(\frac{T\Delta^2}{K})}{\Delta}\right)$.

Approach of UCB-Improved

- The basic idea of UCB-Improved is to divide the horizon into phases or rounds and initialize parameters.

Approach of UCB-Improved

- The basic idea of UCB-Improved is to divide the horizon into phases or rounds and initialize parameters.
- Pull all surviving arms equal number of times during a round.

Approach of UCB-Improved

- The basic idea of UCB-Improved is to divide the horizon into phases or rounds and initialize parameters.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate some arms based on some criteria.

Approach of UCB-Improved

- The basic idea of UCB-Improved is to divide the horizon into phases or rounds and initialize parameters.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate some arms based on some criteria.
- Reset parameters and proceed to next round.

Approach of UCB-Improved

- The basic idea of UCB-Improved is to divide the horizon into phases or rounds and initialize parameters.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate some arms based on some criteria.
- Reset parameters and proceed to next round.
- UCB-Imp achieves a gap-independent regret bound of $O\left(\sqrt{KT \log K}\right)$ and gap-dependent bound of $O\left(\frac{K \log(T\Delta^2)}{\Delta}\right)$.

UCB-Improved ([Auer and Ortner(2010)])

Algorithm 3 UCB-Improved

Input: Time horizon T

Initialization: Set $B_0 := A$ and $\tilde{\Delta}_0 := 1$.

for $m = 0, 1, .. \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$ **do**

Pull each arm in B_m, $n_m = \left\lceil \frac{2 \log (T \tilde{\Delta}_m^2)}{\tilde{\Delta}_m} \right\rceil$ number of times.

Arm Elimination

For each $i \in B_m$, delete arm i from B_m if,

$$\bar{X}_i + \sqrt{\frac{\log (T \tilde{\Delta}_m^2)}{2 n_m}} < \max_{j \in B_m} \left\{ \bar{X}_j - \sqrt{\frac{\log (T \tilde{\Delta}_m^2)}{2 n_m}} \right\}$$

Set $\tilde{\Delta}_{m+1} := \frac{\tilde{\Delta}_m}{2}$, Set $B_{m+1} := B_m$

Stop if $|B_m| = 1$ and pull $i \in B_m$ till n is reached.

end for

Some technical details of UCB-Improved

- We do not know the true means $\mu_i, \forall i \in A$ of the distributions so we estimate it by the $\tilde{\Delta}$ by initializing it from 1.

Some technical details of UCB-Improved

- We do not know the true means $\mu_i, \forall i \in A$ of the distributions so we estimate it by the $\tilde{\Delta}$ by initializing it from 1.
- All rewards are assume to be bounded between $[0, 1]$ and so $\Delta_i \in [0, 1], \forall i \in A$ as well.

Some technical details of UCB-Improved

- We do not know the true means $\mu_i, \forall i \in A$ of the distributions so we estimate it by the $\tilde{\Delta}$ by initializing it from 1.
- All rewards are assume to be bounded between $[0, 1]$ and so $\Delta_i \in [0, 1], \forall i \in A$ as well.
- As opposed to UCB1, MOSS and OCUCB, UCB-Improved has fixed confidence interval $c_m = \sqrt{\dfrac{\log(T\tilde{\Delta}_m^2)}{2n_m}}$ for all arms in a particular phase.

Some technical details of UCB-Improved

- We do not know the true means $\mu_i, \forall i \in A$ of the distributions so we estimate it by the $\tilde{\Delta}$ by initializing it from 1.
- All rewards are assume to be bounded between $[0, 1]$ and so $\Delta_i \in [0, 1], \forall i \in A$ as well.
- As opposed to UCB1, MOSS and OCUCB, UCB-Improved has fixed confidence interval $c_m = \sqrt{\dfrac{\log(T\tilde{\Delta}_m^2)}{2n_m}}$ for all arms in a particular phase.
- c_m ensures that whenever $\tilde{\Delta}_m < \dfrac{\Delta_i}{2}$ in the m-th round, the arm i gets eliminated.

Optimally confident UCB ([Lattimore(2015)])

Algorithm 4 MOSS

Input: K,T, α, ψ

Pull each arm once

for $t = K + 1, ..., T$ **do**

  Pull the arm such that $\max_{i \in A} \left\{ \hat{r}_i + \sqrt{\alpha \frac{\max\{0, \log(\frac{\psi T}{s_i})\}}{s_i}} \right\}$

  $t := t + 1$

end for

- UCB1 is too conservative in exploiting, MOSS is not conservative enough and tends to explore more often than required.
- OCUCB correctly balances this and achieves a gap-independent regret bound $O\left(\sqrt{KT}\right)$ and gap-dependent bound $O\left(\frac{K \log(T/H)}{\Delta}\right)$.

Comparison of UCB1, MOSS, OCUCB, UCB-Improved

Table: Cumulative Regret of Algorithms

Algorithm	Upper bound on Cumulative Regret
UCB1	$\min\left\{O\left(\frac{K\log T}{\Delta}\right), O\left(\sqrt{KT\log T}\right)\right\}$
MOSS	$\min\left\{O\left(\frac{K^2\log(T\Delta^2/K)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
OCUCB	$\min\left\{O\left(\frac{K\log(T/H)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
UCB-Improved	$\min\left\{O\left(\frac{K\log(T\Delta^2)}{\Delta}\right), O\left(\sqrt{KT\log K}\right)\right\}$

Problems of UCB-Improved

- UCB-Improved conducts too much early exploration.

Problems of UCB-Improved

- UCB-Improved conducts too much early exploration.
- The arm elimination condition of UCB-Imp is very conservative.

Problems of UCB-Improved

- UCB-Improved conducts too much early exploration.
- The arm elimination condition of UCB-Imp is very conservative.
- When the gaps are small and uniform UCB-Imp performs very badly.

Problems of UCB-Improved

- UCB-Improved conducts too much early exploration.
- The arm elimination condition of UCB-Imp is very conservative.
- When the gaps are small and uniform UCB-Imp performs very badly.
- There is a gap in theoretical guarantee and empirical performance.

Problem Definition

- In our work, we ask the following set of questions:

Problem Definition

- In our work, we ask the following set of questions:
  - Can we achieve a better gap-dependent/gap-independent regret bound than UCB-Improved?

Problem Definition

- In our work, we ask the following set of questions:
  - Can we achieve a better gap-dependent/gap-independent regret bound than UCB-Improved?
  - Can we bridge the theoretical versus empirical performance of UCB-Imp?

Problem Definition

- In our work, we ask the following set of questions:
  - Can we achieve a better gap-dependent/gap-independent regret bound than UCB-Improved?
  - Can we bridge the theoretical versus empirical performance of UCB-Imp?
  - Can we use ideas from Clustering to achieve this?

Problem Definition

- In our work, we ask the following set of questions:
  - Can we achieve a better gap-dependent/gap-independent regret bound than UCB-Improved?
  - Can we bridge the theoretical versus empirical performance of UCB-Imp?
  - Can we use ideas from Clustering to achieve this?
  - Can we study the effect of Clustering in SMAB?

Problem Definition

- In our work, we ask the following set of questions:
  - Can we achieve a better gap-dependent/gap-independent regret bound than UCB-Improved?
  - Can we bridge the theoretical versus empirical performance of UCB-Imp?
  - Can we use ideas from Clustering to achieve this?
  - Can we study the effect of Clustering in SMAB?
- The answer to all of this is ClusUCB.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.
- It then creates p fixed clusters (given as an input) and randomly assign arms into each of them such that each cluster contains equal number of arms.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.
- It then creates p fixed clusters (given as an input) and randomly assign arms into each of them such that each cluster contains equal number of arms.
- Pull all surviving arms equal number of times during a round.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.
- It then creates p fixed clusters (given as an input) and randomly assign arms into each of them such that each cluster contains equal number of arms.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate arms inside each cluster by comparing its performance against the best arm in the cluster.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.
- It then creates p fixed clusters (given as an input) and randomly assign arms into each of them such that each cluster contains equal number of arms.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate arms inside each cluster by comparing its performance against the best arm in the cluster.
- Also eliminate clusters with all of its arms by comparing its performance against the globally best arm.
- Reset parameters and move to the next round.

Approach of ClusUCB

- The basic idea of ClusUCB directly follows from UCB-Imp. It starts by dividing the horizon into rounds and initializing parameters.
- It then creates p fixed clusters (given as an input) and randomly assign arms into each of them such that each cluster contains equal number of arms.
- Pull all surviving arms equal number of times during a round.
- At the end of the round eliminate arms inside each cluster by comparing its performance against the best arm in the cluster.
- Also eliminate clusters with all of its arms by comparing its performance against the globally best arm.
- Reset parameters and move to the next round.
- At a higher level ClusUCB behaves like p independently running UCB-Imp with the exploration parameters ρ_a, ρ_s and ψ helping in overcoming early exploration.

Approach of ClusUCB I

Input: Number of clusters p, time horizon T, exploration parameters ρ_a, ρ_s and ψ.

Initialization: Set $B_0 := A$, $S_0 = S$ and $\epsilon_0 := 1$.

Create a partition S_0 of the arms at random into p clusters of size up to $\ell = \left\lceil \frac{K}{p} \right\rceil$ each.

for $m = 0, 1, .. \left\lfloor \frac{1}{2} \log_2 \frac{7T}{K} \right\rfloor$ **do**

  Pull each arm in B_m so that the total number of times it has been pulled is $n_m = \left\lceil \frac{2 \log\left(\psi T \epsilon_m^2\right)}{\epsilon_m} \right\rceil$.

  Arm Elimination

    For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

$$\hat{r}_i + \sqrt{\frac{\rho_a \log\left(\psi T \epsilon_m^2\right)}{2 n_m}} < \max_{j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\rho_a \log\left(\psi T \epsilon_m^2\right)}{2 n_m}} \right\}$$

Approach of ClusUCB II

Cluster Elimination

Delete cluster $s_k \in S_m$ and remove all arms $i \in s_k$ from B_m if

$$\max_{i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2n_m}} \right\} < \max_{j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2n_m}} \right\}.$$

Set $\epsilon_{m+1} := \frac{\epsilon_m}{2}$

Set $B_{m+1} := B_m$

Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.

end for

UCB1, MOSS, OCUCB, UCB-Imp, ClusUCB

Algorithm	Upper bound on Cumulative Regret
UCB1	$\min\left\{O\left(\frac{K\log T}{\Delta}\right), O\left(\sqrt{KT\log T}\right)\right\}$
MOSS	$\min\left\{O\left(\frac{K^2\log(T\Delta^2/K)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
OCUCB	$\min\left\{O\left(\frac{K\log(T/H)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
UCB-Improved	$\min\left\{O\left(\frac{K\log(T\Delta^2)}{\Delta}\right), O\left(\sqrt{KT\log K}\right)\right\}$
ClusUCB	$\min\left\{O\left(\frac{K\log(T\Delta^2/\sqrt{\log K})}{\Delta}\right), O\left(\sqrt{KT\log K}\right)\right\}$

Approach of EClusUCB

- One of the main problems of ClusUCB is that it's still a round based algorithm.

Approach of EClusUCB

- One of the main problems of ClusUCB is that it's still a round based algorithm.
- In every round it pulls all the arms equal number of times, which although is less compared to UCB-Improved but still we can be better.

Approach of EClusUCB

- One of the main problems of ClusUCB is that it's still a round based algorithm.
- In every round it pulls all the arms equal number of times, which although is less compared to UCB-Improved but still we can be better.
- One simple solution is to pull the arm with the highest UCB at every timestep. This is called optimistic greedy sampling for UCB-Imp (see [Liu and Tsuruoka(2016)]).

Approach of EClusUCB

- One of the main problems of ClusUCB is that it's still a round based algorithm.
- In every round it pulls all the arms equal number of times, which although is less compared to UCB-Improved but still we can be better.
- One simple solution is to pull the arm with the highest UCB at every timestep. This is called optimistic greedy sampling for UCB-Imp (see [Liu and Tsuruoka(2016)]).
- We introduce this in Efficient ClusUCB or EClusUCB.

Input: Number of clusters p, time horizon T, exploration parameters ρ_a, ρ_s and ψ.
Initialization: Set $m := 0$, $B_0 := A$, $S_0 = S$, $\epsilon_0 := 1$,
$M = \left\lfloor \frac{1}{2} \log_2 \frac{7T}{K} \right\rfloor$, $n_0 = \left\lceil \frac{2 \log(\psi T \epsilon_0^2)}{\epsilon_0} \right\rceil$ and $N_0 = K n_0$.
Create a partition S_0 of the arms at random into p clusters of size up to $\ell = \left\lceil \frac{K}{p} \right\rceil$ each.
Pull each arm once
for $t = K + 1, .., T$ **do**

Pull arm $i \in B_m$ such that $\text{argmax}_{j \in B_m} \left\{ \hat{r}_j + \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2 z_j}} \right\}$,
where z_j is the number of times arm j has been pulled

$t := t + 1$

Arm Elimination

Approach of ClusUCB II

For each cluster $s_k \in S_m$, delete arm $i \in s_k$ from B_m if

$$\hat{r}_i + \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2z_i}} < \max_{j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\rho_a \log(\psi T \epsilon_m^2)}{2z_j}} \right\}$$

Cluster Elimination

Delete cluster $s_k \in S_m$ and remove all arms $i \in s_k$ from B_m if

$$\max_{i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2z_i}} \right\} < \max_{j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho_s \log(\psi T \epsilon_m^2)}{2z_j}} \right\}.$$

if $t \geq N_m$ and $m \leq M$ **then**

Approach of ClusUCB III

$\epsilon_{m+1} := \frac{\epsilon_m}{2}$

$B_{m+1} := B_m$

$n_{m+1} := \left\lceil \frac{2\log(\psi T \epsilon_{m+1}^2)}{\epsilon_{m+1}} \right\rceil$

$N_{m+1} := t + |B_{m+1}| n_{m+1}$

$m := m + 1$

Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.

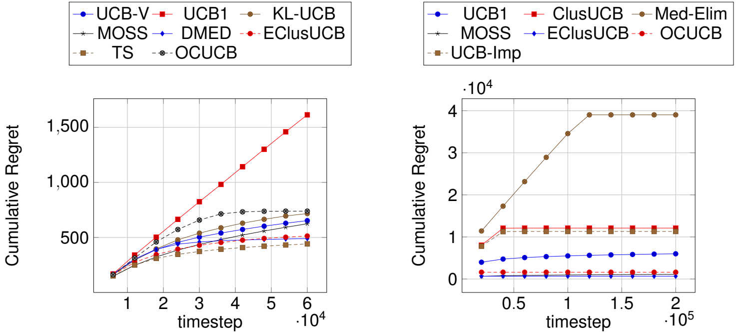
end if

end for

UCB1, MOSS, OCUCB, UCB-Imp, EClusUCB

Algorithm	Upper bound on Cumulative Regret
UCB1	$\min\left\{O\left(\frac{K\log T}{\Delta}\right), O\left(\sqrt{KT\log T}\right)\right\}$
MOSS	$\min\left\{O\left(\frac{K^2\log(T\Delta^2/K)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
OCUCB	$\min\left\{O\left(\frac{K\log(T/H)}{\Delta}\right), O\left(\sqrt{KT}\right)\right\}$
UCB-Improved	$\min\left\{O\left(\frac{K\log(T\Delta^2)}{\Delta}\right), O\left(\sqrt{KT\log K}\right)\right\}$
EClusUCB	$\min\left\{O\left(\frac{K\log(T\Delta^2/\sqrt{\log K})}{\Delta}\right), O\left(\sqrt{KT\log K}\right)\right\}$

Finally, experiment!!!



(a) Experiment 1: 20 Bernoulli-distributed arms with $r_{i_{i \neq *}} = 0.07$ and $r^* = 0.1$.

(b) Experiment 2: 100 Gaussian-distributed arms with $r_{i_{i \neq *:1-33}} = 0.1$, $r_{i_{i \neq *:34-99}} = 0.6$ and $r^*_{i=100} = 0.9$.

Figure: Cumulative regret for various bandit algorithms on two stochastic K-armed bandit environments.

References I

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári.
Improved algorithms for linear stochastic bandits.
In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal.
Analysis of thompson sampling for the multi-armed bandit problem.
arXiv preprint arXiv:1111.1797, 2011.

Jean-Yves Audibert and Sébastien Bubeck.
Minimax policies for adversarial and stochastic bandits.
In *COLT*, pages 217–226, 2009.

References II

- Jean-Yves Audibert and Sébastien Bubeck.
  Best arm identification in multi-armed bandits.
  In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

- Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári.
  Exploration–exploitation tradeoff using variance estimates in multi-armed bandits.
  Theoretical Computer Science, 410(19):1876–1902, 2009.

- Peter Auer and Ronald Ortner.
  Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem.
  Periodica Mathematica Hungarica, 61(1-2):55–65, 2010.

References III

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer.
Finite-time analysis of the multiarmed bandit problem.
Machine learning, 47(2-3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire.
The nonstochastic multiarmed bandit problem.
SIAM Journal on Computing, 32(1):48–77, 2002b.

Sébastien Bubeck and Nicolo Cesa-Bianchi.
Regret analysis of stochastic and nonstochastic multi-armed bandit problems.
arXiv preprint arXiv:1204.5721, 2012.

Sébastien Bubeck, Nicolo Cesa-Bianchi, and Gábor Lugosi.
Bandits with heavy tail.
arXiv preprint arXiv:1209.1727, 2012.

References IV

Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet.
Bounded regret in stochastic multi-armed bandits.
arXiv preprint arXiv:1302.1611, 2013.

Olivier Cappe, Aurelien Garivier, and Emilie Kaufmann.
pymabandits, 2012.
`http://mloss.org/software/view/415/`.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour.
Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems.
The Journal of Machine Learning Research, 7:1079–1105, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani.
The elements of statistical learning, volume 1.
Springer series in statistics Springer, Berlin, 2001.

References V

Aurélien Garivier and Olivier Cappé.
The kl-ucb algorithm for bounded stochastic bandits and beyond.
arXiv preprint arXiv:1102.2490, 2011.

Junya Honda and Akimichi Takemura.
An asymptotically optimal bandit algorithm for bounded support models.
In *COLT*, pages 67–79. Citeseer, 2010.

Tze Leung Lai and Herbert Robbins.
Asymptotically efficient adaptive allocation rules.
Advances in applied mathematics, 6(1):4–22, 1985.

Tor Lattimore.
Optimally confident ucb: Improved regret for finite-armed bandits.
arXiv preprint arXiv:1507.07880, 2015.

References VI

- Yun-Ching Liu and Yoshimasa Tsuruoka.
  Modification of improved upper confidence bounds for regulating exploration in monte-carlo tree search.
  Theoretical Computer Science, 2016.

- Shie Mannor and John N Tsitsiklis.
  The sample complexity of exploration in the multi-armed bandit problem.
  Journal of Machine Learning Research, 5(Jun):623–648, 2004.

- Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg.
  Batched bandit problems.
  arXiv preprint arXiv:1505.00369, 2015.

References VII

- Herbert Robbins.
  Some aspects of the sequential design of experiments.
  In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1952.
- Richard S Sutton and Andrew G Barto.
  Reinforcement learning: An introduction.
  MIT press, 1998.
- William R Thompson.
  On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.
  Biometrika, pages 285–294, 1933.
- David Tolpin and Solomon Eyal Shimony.
  Mcts based on simple regret.
  In *AAAI*, 2012.

Thank You