Tutorial on Bandit

Subhojyoti Mukherjee

IIT Madras

February 21, 2017

Overview

Major Objections raised by AISTATS

- Gap-Dependent and Gap-Independent bounds take two different sets of hyper-parameters
- Difference between empirical performance and theoretical guarantees. Mainly taking ρ_a, ρ_s as constant in theoretical proof and reducing it in experiments.
- Why didn't you test against OCUCB (which reaches the optimal bounds) or Optimistic Gittins indices (NIPS) against your algorithm?
- What is the optimal choice of number of clusters p?
- What is the optimal choice of ρ_a, ρ_s and ψ?
- The algorithm is not anytime

Gap-Dependent and Gap-Independent bounds

- In both the cases we now choose $\rho_a = \frac{1}{2}$, $\rho_s = \frac{1}{2}$ and $\psi = \frac{T}{\log K}$.
- We achieve a gap-dependent regret of $O\left(\frac{K \log\left(\frac{T\Delta^2}{\sqrt{\log(K)}}\right)}{\Delta}\right)$, better than UCB1, MOSS and UCB-Improved. Not better than OCUCB.
- We achieve a gap-independent regret bound of $O\left(\sqrt{KT \log K}\right)$, which is not better than MOSS, OCUCB's $O\left(\sqrt{KT}\right)$.
- But empirically we outperform both in different test cases for large horizon T and large K.

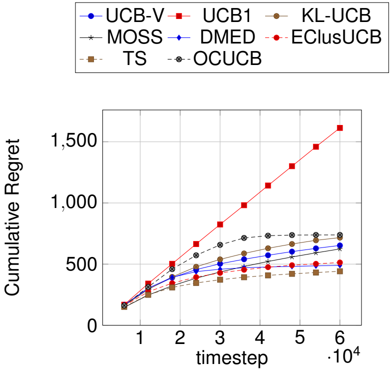
Choosing ρ_a, ρ_s, p and ψ

- We achieve a gap-independent regret bound of $O\left(\sqrt{KT\log K}\right)$ with $\rho_a = \frac{1}{2}$, $\rho_s = \frac{1}{2}$, $\psi = \frac{T}{\log K}$ and $p = \left\lceil \frac{K}{\log K} \right\rceil$.
- We prove that for $p = \left\lceil \frac{K}{\log K} \right\rceil$, we achieve a error elimination regret bound of $O(\frac{K}{K+\log K}\sqrt{KT\log K})$ for ClusUCB(employing arm elimination and cluster elimination simultaneously) while we achieve a worse elimination regret bound for ClusUCB-AE (only arm elimination) of order $O(\sqrt{KT\log K})$.
- This is corroborated in the experiments as well.

Handling theoretical and empirical performance mismatch

- Our approach of expecting ClusUCB will beat other algorithms is wrong, because it is a round-based algorithm which is pulling all the arms equally in each round.
- Simple solution is to pull in each timestep the arm that maximizes the UCB and divide each round into $|B_m|n_m$ timesteps, so that we can achieve the same theoretical guarantees as ClusUCB.
- This is the Efficient ClusUCB (EClusUCB) which outperforms most other algorithms. All we do is to prove in Theorem 2 that EClusUCB has a regret upper bound same as ClusUCB. Rest all corollaries will follow.

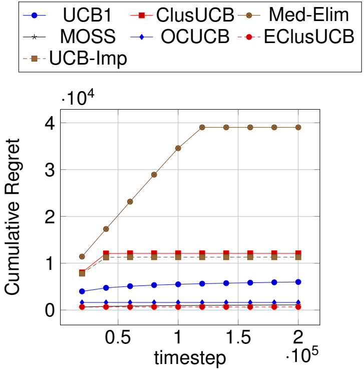
Experiments



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

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

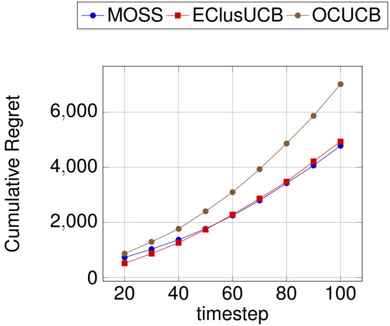
Experiments



(a) 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.

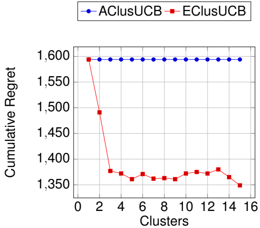
Experiments



(a) Experiment 3: Experiment 3: 20 to 100 Bernoulli-distributed arms with $r_{i_{i \neq *}} = 0.05$ and $r^* = 0.1$.

Figure: Cumulative regret and Error Percentage for ClusUCB variants

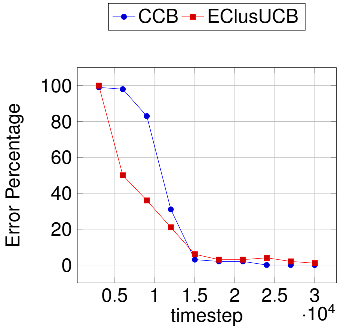
Experiments



(a) Experiment 4: Cumulative regret for ClusUCB for various clusters

Over various number of clusters EClusUCB is found to reach close to the lowest cumulative regret at $p = \left\lceil \frac{K}{\log K} \right\rceil$. For $p = K/2$ it reaches the lowest but it increases the elimination error. So $p = \left\lceil \frac{K}{\log K} \right\rceil$ is a balance between theoretical and empirical performance. ClusUCB-AE ($p = 1$) performs the worst. So clustering has some benefits.

Experiments



(b) Experiment 5: Error Percentage of EClusUCB and CCB

Figure: Cumulative regret and Error Percentage for ClusUCB variants

Logic behind EClusUCB

- At every timestep pull the arm that has the maximum UCB.
- Check arm and cluster elimination conditions at every timestep.
- Divide each round into $|B_m|n_m$ tiesteps where $n_m = \left\lceil \frac{2\log(\psi T \epsilon_m^2)}{\epsilon_m} \right\rceil$.
- The algorithm is not anytime, but it is not as bad as ClusUCB or UCB-Improved. You can stop the algorithm in between a round as you are not pulling all the arms equal number of times in a round (from Liu and Tsuruoka [2016]) .

EClusUCB Algorithm I

Input: Number of clusters p, time horizon T, exploration parameters ρ_a, ρ_s and ψ.
Initialization: Set $m := 0$, $B_0 := A$, $S_0 = S$, $\epsilon_0 := 1$,
$M = \left\lfloor \frac{1}{2} \log_2 \frac{14T}{K} \right\rfloor$, $n_0 = \left\lceil \frac{2 \log (\psi T \epsilon_0^2)}{\epsilon_0} \right\rceil$ and $N_0 = Kn_0$.
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

EClusUCB Algorithm 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**

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

$B_{m+1} := B_m$

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

$N_{m+1} := t + |B_{m+1}|n_{m+1}$
$m := m + 1$
Stop if $|B_m| = 1$ and pull $i \in B_m$ till T is reached.
end if
end for

Regret bound of EClusUCB I

- We have already proved the regret bound of ClusUCB. We just have to show EClusUCB has the same regret upper bound.
- The approach is similar. z_i is the number of times as an arm i is pulled. m_i is the minimum round arm i gets eliminated.
- n_m is the number of pulls allocated for each surviving arms in B_m.
- Case:1 When $z_i = n_{m_i}$ for any sub-optimal arm or cluster arm, we prove that the probability of an arm *not getting eliminated* is exponentially low (as like Theorem 1).
- Case:2 Or $z_i < n_{m_i}$ we upper bound the number of pulls by n_{m_i} (as like Theorem 1).
- Case:3 Or the sub-optimal arm or cluster arm eliminates the optimal arm which is the opposite case of Case 1. We proceed as like Theorem 1.

Regret bound of EClusUCB II

- Combining all we get that EClusUCB has an regret upper bound as like ClusUCB.
- This is also intuitively clear because in any scenario ClusUCB (being round-based) will always pull sub-optimal arms more than EClusUCB.

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.

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.

References II

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.

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 III

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.

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

References IV

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.

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.

References V

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

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