Thresholding Bandits with Augmented UCB

Subhojyoti Mukherjee
Roll No: CS15S300
Guide: Dr. Balaraman Ravindran
Co-Guide: Dr. Nandan Sudarsanam

IIT Madras

July 14, 2017

Overview

Stochastic Multi-Armed Bandit Problem (SMAB)

- The thresholding bandit problem falls under the broad area of stochastic multi-armed bandit problem.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The thresholding bandit problem falls under the broad area of stochastic multi-armed bandit problem.
- A finite set of actions or arms belonging to set A such that $|A| = K$.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The thresholding bandit problem falls under the broad area of stochastic multi-armed bandit problem.
- A finite set of actions or arms belonging to set A such that $|A| = K$.
- The rewards for each of the arms are identical and independent random variables drawn from distribution specific to the arm.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The thresholding bandit problem falls under the broad area of stochastic multi-armed bandit problem.
- A finite set of actions or arms belonging to set A such that $|A| = K$.
- The rewards for each of the arms are identical and independent random variables drawn from distribution specific to the arm.
- The learner does not know the mean $r_i, \forall i \in A$ of the distribution or the variance σ_i^2.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The distributions for each of the arms are fixed throughout the time horizon denoted by T.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The distributions for each of the arms are fixed throughout the time horizon denoted by T.
- The estimated reward $\hat{r}_i = \frac{1}{n_i}\sum_{z=1}^{n_i} X_{i,z}$.

Stochastic Multi-Armed Bandit Problem (SMAB)

- The distributions for each of the arms are fixed throughout the time horizon denoted by T.
- The estimated reward $\hat{r}_i = \frac{1}{n_i}\sum_{z=1}^{n_i} X_{i,z}$.
- Due to the uncertainty in $\hat{r}_i$ we have carefully conduct exploration.

Problem Definition of TBP

- Primary aim: Identify the arms whose mean of the reward distribution (r_i) is above a particular threshold τ given as input.

Problem Definition of TBP

- Primary aim: Identify the arms whose mean of the reward distribution (r_i) is above a particular threshold τ given as input.
- Condition: This has to be achieved within T timesteps of exploration and this is termed as a fixed-budget problem.

Problem Definition of TBP

- Primary aim: Identify the arms whose mean of the reward distribution (r_i) is above a particular threshold τ given as input.
- Condition: This has to be achieved within T timesteps of exploration and this is termed as a fixed-budget problem.
- At the end of the given T timesteps the learner must recommend a set of arms which (according to it) are the arms having reward mean above τ.

Problem Definition of TBP

- We define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$.

Problem Definition of TBP

- We define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$.
- S_τ^c denote the complement of S_τ, i.e., $S_\tau^c = \{i \in \mathcal{A} : r_i < \tau\}$.

Problem Definition of TBP

- We define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$.
- S_τ^c denote the complement of S_τ, i.e., $S_\tau^c = \{i \in \mathcal{A} : r_i < \tau\}$.
- Let $\hat{S}_\tau$ denote the recommendation of a learning algorithm after T time units of exploration, while $\hat{S}_\tau^c$ denotes its complement.

Problem Definition of TBP

- We define the set $S_\tau = \{i \in \mathcal{A} : r_i \geq \tau\}$.
- S_τ^c denote the complement of S_τ, i.e., $S_\tau^c = \{i \in \mathcal{A} : r_i < \tau\}$.
- Let $\hat{S}_\tau$ denote the recommendation of a learning algorithm after T time units of exploration, while $\hat{S}_\tau^c$ denotes its complement.
- The goal of the learning agent is to minimize the expected loss:

$$\mathbb{E}[\mathcal{L}(T)] = \mathbb{P}\Big(\underbrace{\{S_\tau \cap \hat{S}_\tau^c \neq \emptyset\}}_{\textbf{Rejected good arms}} \cup \underbrace{\{\hat{S}_\tau \cap S_\tau^c \neq \emptyset\}}_{\textbf{Accepted bad arms}} \Big)$$

Challenges in the TBP Settings

- Closer arms' mean to the threshold $\Rightarrow$ harder is to discriminate.

Challenges in the TBP Settings

- Closer arms' mean to the threshold $\Rightarrow$ harder is to discriminate.
- Lesser the budget $\Rightarrow$ harder the problem.

Challenges in the TBP Settings

- Closer arms' mean to the threshold $\Rightarrow$ harder is to discriminate.
- Lesser the budget $\Rightarrow$ harder the problem.
- Higher variance of the arms' $\Rightarrow$ harder the problem.

Applications

- Selecting the best channels (out of several existing channels) for mobile communications in a very short duration whose qualities are above an acceptable threshold (see [Audibert and Bubeck(2010)]).

Applications

- Selecting the best channels (out of several existing channels) for mobile communications in a very short duration whose qualities are above an acceptable threshold (see [Audibert and Bubeck(2010)]).
- Selecting a small set of best workers (out of a very large pool of workers) whose productivity is above a threshold.

Applications

- Selecting the best channels (out of several existing channels) for mobile communications in a very short duration whose qualities are above an acceptable threshold (see [Audibert and Bubeck(2010)]).
- Selecting a small set of best workers (out of a very large pool of workers) whose productivity is above a threshold.
- In anomaly detection and classification (see Locatelli et al. (2016)).

Contributions

- We propose the Augmented UCB (AugUCB) [Mukherjee et al. (2017)] algorithm for the fixed-budget TBP setting.

Contributions

- We propose the Augmented UCB (AugUCB) [Mukherjee et al. (2017)] algorithm for the fixed-budget TBP setting.
- AugUCB takes into account the empirical variances of the arms along with mean estimates.

Contributions

- We propose the Augmented UCB (AugUCB) [Mukherjee et al. (2017)] algorithm for the fixed-budget TBP setting.
- AugUCB takes into account the empirical variances of the arms along with mean estimates.
- It is the first variance-based arm elimination algorithm for the considered TBP settings.

Contributions

- We propose the Augmented UCB (AugUCB) [Mukherjee et al. (2017)] algorithm for the fixed-budget TBP setting.
- AugUCB takes into account the empirical variances of the arms along with mean estimates.
- It is the first variance-based arm elimination algorithm for the considered TBP settings.
- It addresses an open problem discussed in [Auer and Ortner(2010)] of designing an algorithm that can eliminate arms based on variance estimates.

Contributions

- We propose the Augmented UCB (AugUCB) [Mukherjee et al. (2017)] algorithm for the fixed-budget TBP setting.
- AugUCB takes into account the empirical variances of the arms along with mean estimates.
- It is the first variance-based arm elimination algorithm for the considered TBP settings.
- It addresses an open problem discussed in [Auer and Ortner(2010)] of designing an algorithm that can eliminate arms based on variance estimates.
- We also define a new problem complexity which uses empirical variance estimates along with arm's mean for giving the theoretical bound.

The Upper Confidence Bound (UCB) Approach

- Since there is an initial uncertainty in the estimated mean ($\hat{r}_i$) introduce a confidence interval term c_i.

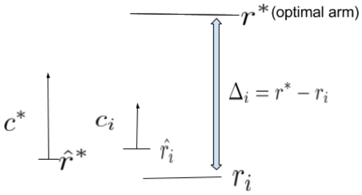
The Upper Confidence Bound (UCB) Approach

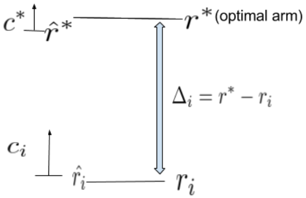
- Since there is an initial uncertainty in the estimated mean ($\hat{r}_i$) introduce a confidence interval term c_i.
- c_i ensures that the arm i is properly explored and is gradually reduced with time as one pulls the arm i more.

The Upper Confidence Bound (UCB) Approach

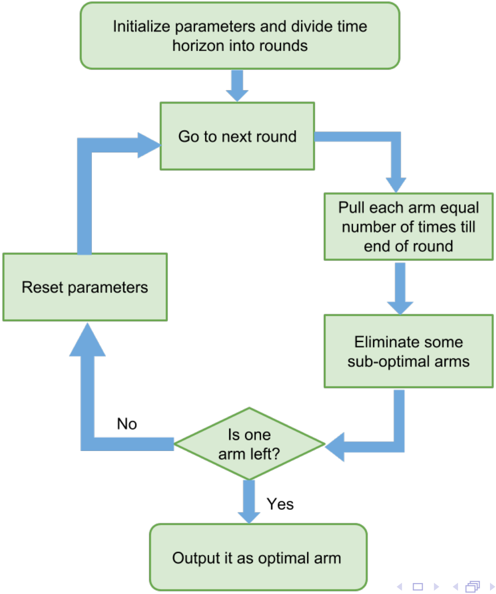
- Since there is an initial uncertainty in the estimated mean ($\hat{r}_i$) introduce a confidence interval term c_i.
- c_i ensures that the arm i is properly explored and is gradually reduced with time as one pulls the arm i more.
- At every timestep pull arm that has the maximum value of $\hat{r}_i + c_i$ and this will ensure that proper exploration is done.

The UCB Approach



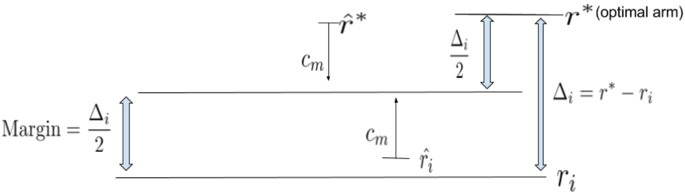


Approach of UCB-Improved (UCB-Imp)



Intuition of UCB-Improved (UCB-Imp)

Arm Elimination: $\hat{r}_i + c_m < \hat{r}_{max} - c_m$



APT Approach

- The Anytime Parameter Free (APT) [Locatelli et al. (2016)] algorithm was proposed for TBP setting in ICML 2016.

APT Approach

- The Anytime Parameter Free (APT) [Locatelli et al. (2016)] algorithm was proposed for TBP setting in ICML 2016.
- This algorithm uses only mean estimation to find the S_τ.

APT Approach

- The Anytime Parameter Free (APT) [Locatelli et al. (2016)] algorithm was proposed for TBP setting in ICML 2016.
- This algorithm uses only mean estimation to find the S_τ.
- Theoretically they proved this algorithm to be almost optimal when only mean estimation is used as a metric of comparison.

APT Approach

- The Anytime Parameter Free (APT) [Locatelli et al. (2016)] algorithm was proposed for TBP setting in ICML 2016.
- This algorithm uses only mean estimation to find the S_τ.
- Theoretically they proved this algorithm to be almost optimal when only mean estimation is used as a metric of comparison.
- Empirically it outperformed other state-of-the-art algorithms which were modified to perform in the TBP setting.

APT Algorithm

Algorithm 1 APT

Input: Time horizon T, threshold τ, tolerance factor $\epsilon \geq 0$

Pull each arm once

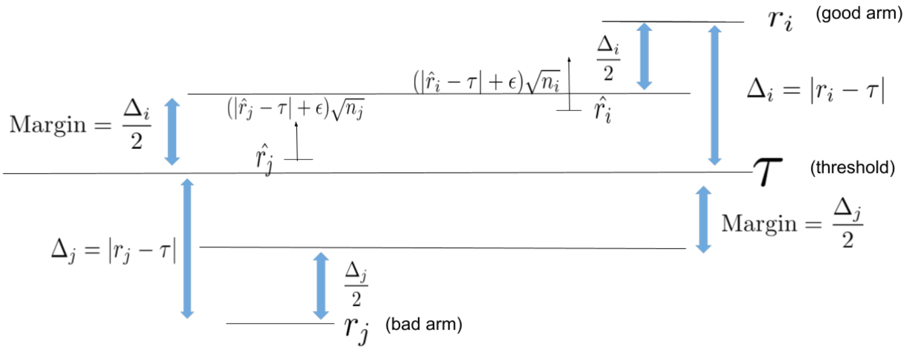
for $t = K + 1, .., T$ **do**

Pull arm $j \in \arg\min_{i \in A} \left\{ (|\hat{r}_i - \tau| + \epsilon) \sqrt{n_i} \right\}$ and observe the reward for arm j.

end for

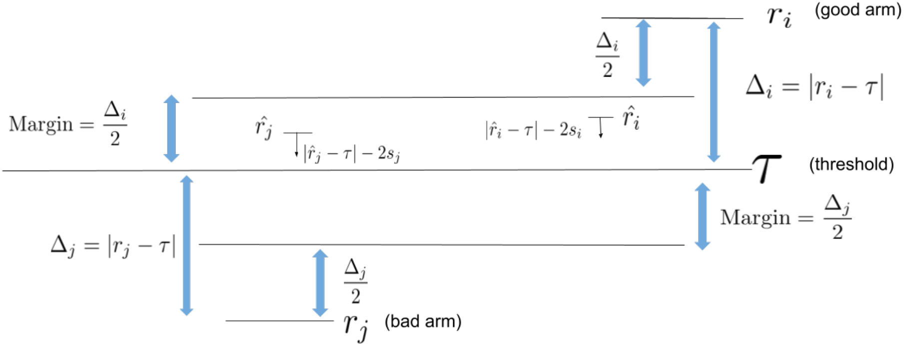
Output: $\hat{S}_\tau = \{i : \hat{r}_i \geq \tau\}$.

Intuition of APT



AugUCB algorithm (Intuition, Arm pulling)

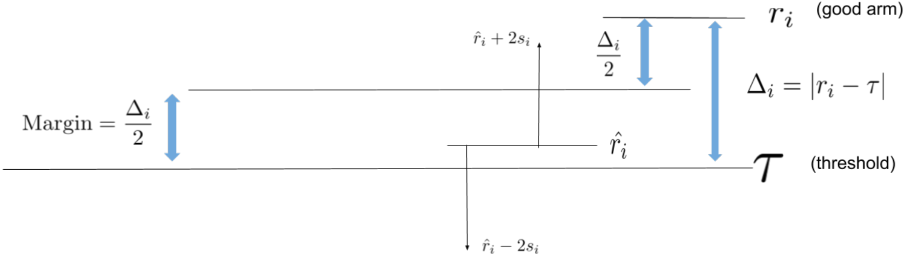
- Like UCB-Imp, AugUCB also divides the time budget T into rounds.
- At every timestep we pull arm j s.t. $j \in \arg\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2s_i \right\}$ (like APT).



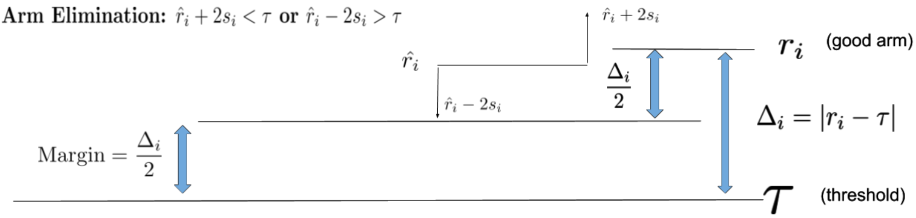
AugUCB algorithm (Intuition, Arm Elimination)

- It is risky to eliminate arm i while $\hat{r}_i$ is inside *Margin*.
- Confidence interval s_i will make sure arm i is not eliminated while inside Margin with a high probability.

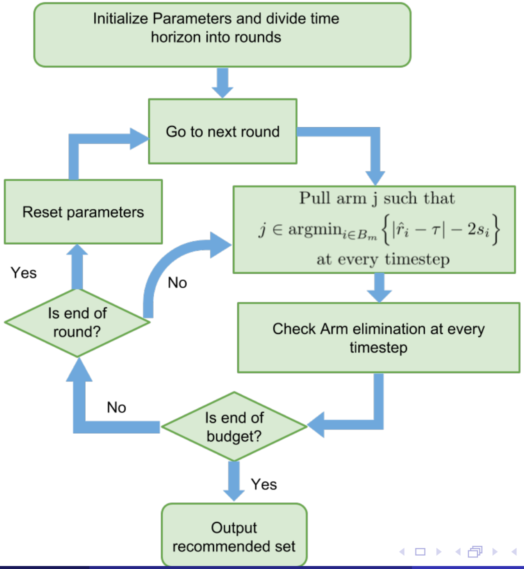
Arm Elimination: $\hat{r}_i + 2s_i < \tau$ **or** $\hat{r}_i - 2s_i > \tau$



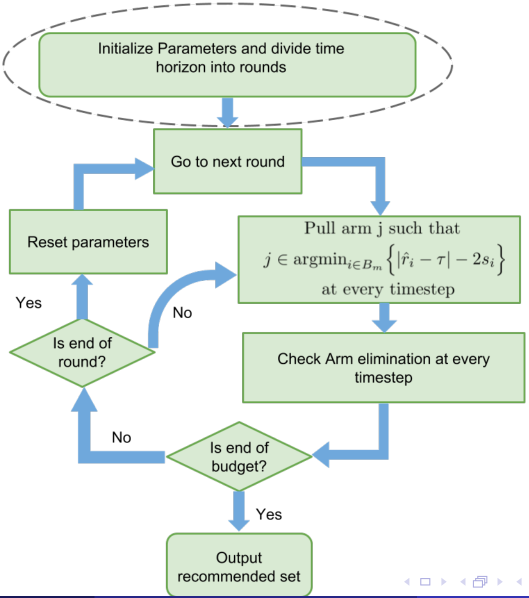
AugUCB algorithm (Intuition, Arm Elimination)



AugUCB algorithm



AugUCB parameter initialization



Parameter initialization

- We define $\ell_0 = \left\lceil \frac{2\psi_0 \log(T\epsilon_0)}{\epsilon_0} \right\rceil$ as the budget allocated to each arm in a round.

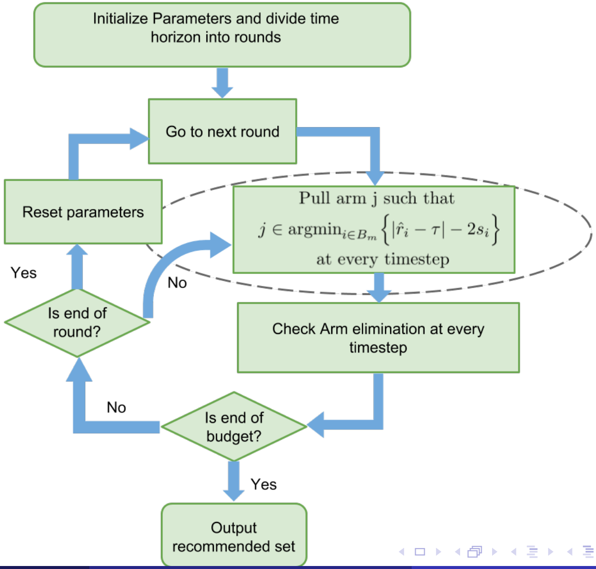
Parameter initialization

- We define $\ell_0 = \left\lceil \frac{2\psi_0 \log(T\epsilon_0)}{\epsilon_0} \right\rceil$ as the budget allocated to each arm in a round.
- The first round gets divided into $N_0 = K\ell_0$ timesteps.

Parameter initialization

- We define $\ell_0 = \left\lceil \frac{2\psi_0 \log(T\epsilon_0)}{\epsilon_0} \right\rceil$ as the budget allocated to each arm in a round.
- The first round gets divided into $N_0 = K\ell_0$ timesteps.
- We define a large exploration regulatory factor $\psi_0 = \frac{T\epsilon_0}{128\left(\log(\frac{3}{16}K\log K)\right)^2}$ which controls exploration.

AugUCB arm pull



Arm pull

- We pull the arm that minimizes $j \in \arg\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2s_i \right\}$

Arm pull

- We pull the arm that minimizes $j \in \arg\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2s_i \right\}$
- We define the confidence interval $s_i = \sqrt{\frac{\rho \psi_m (\hat{v}_i + 1) \log(T \epsilon_m)}{4 n_i}}$.

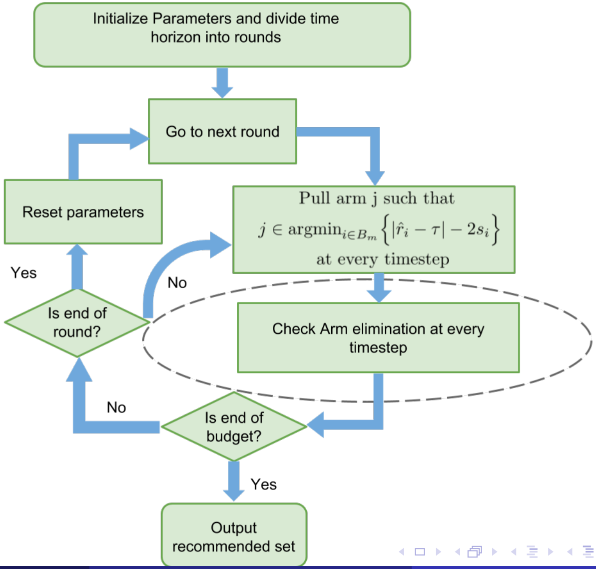
Arm pull

- We pull the arm that minimizes $j \in \arg\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2s_i \right\}$
- We define the confidence interval $s_i = \sqrt{\frac{\rho\psi_m(\hat{v}_i+1)\log(T\epsilon_m)}{4n_i}}$.
- s_i decreases with more n_i and ψ_m and ρ ensures that it decreases at a correct rate.

Arm pull

- We pull the arm that minimizes $j \in \arg\min_{i \in B_m} \left\{ |\hat{r}_i - \tau| - 2s_i \right\}$
- We define the confidence interval $s_i = \sqrt{\frac{\rho\psi_m(\hat{v}_i+1)\log(T\epsilon_m)}{4n_i}}$.
- s_i decreases with more n_i and ψ_m and ρ ensures that it decreases at a correct rate.
- Note that $\hat{v}_i$ estimated variance in s_i makes the algorithm pull the arm which shows more variance.

AugUCB arm elimination



Arm elimination

- Arm elimination condition is checked at every timestep.

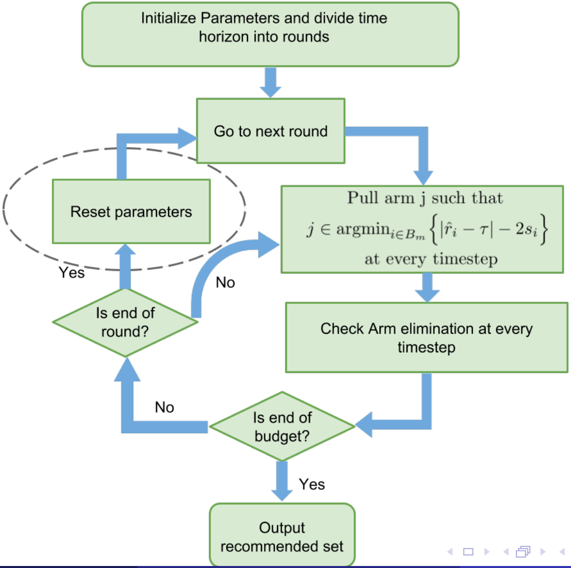
Arm elimination

- Arm elimination condition is checked at every timestep.
- It eliminated arm which are above or below margin.

Arm elimination

- Arm elimination condition is checked at every timestep.
- It eliminated arm which are above or below margin.
- Re-allocates the remaining budget for surviving arms.

AugUCB parameter reset



AugUCB parameter reset

- Reduce the exploration factor as time progresses.

AugUCB parameter reset

- Reduce the exploration factor as time progresses.
- Recalculate the budget for each surviving arms.

AugUCB parameter reset

- Reduce the exploration factor as time progresses.
- Recalculate the budget for each surviving arms.
- Recalculate the length of each round on the number of surviving arms.

Problem Complexity

- We must delve into the notion of hardness which come from the general pure exploration bandit literature.

Problem Complexity

- We must delve into the notion of hardness which come from the general pure exploration bandit literature.
- We define $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$ and $H_2 = \min_{i \in \mathcal{A}} \frac{i}{\Delta_{(i)}^2}$ where $\Delta_{(i)}$ is an increasing ordering of Δ_i.

Problem Complexity

- We must delve into the notion of hardness which come from the general pure exploration bandit literature.
- We define $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$ and $H_2 = \min_{i \in \mathcal{A}} \frac{i}{\Delta_{(i)}^2}$ where $\Delta_{(i)}$ is an increasing ordering of Δ_i.
- The relationship between H_1 and H_2 can be derived as,

$$H_2 \leq H_1 \leq \log(2K) H_2$$

Problem Complexity

- For a variance aware algorithm we define $H_{\sigma,1}$ (as in Gabillon et al. (2011)) that incorporates reward variances into its expression as:

$$H_{\sigma,1} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}.$$

Problem Complexity

- For a variance aware algorithm we define $H_{\sigma,1}$ (as in Gabillon et al. (2011)) that incorporates reward variances into its expression as:

$$H_{\sigma,1} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}.$$

- Finally, analogous to H_2, we introduce $H_{\sigma,2}$, such that $H_{\sigma,2} = \max_{i \in \mathcal{A}} \frac{i}{\tilde{\Delta}_{(i)}^2}$, where $\tilde{\Delta}_i^2 = \frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}$, $(\tilde{\Delta}_{(i)})$ is an increasing ordering of $(\tilde{\Delta}_i)$.

Problem Complexity

- For a variance aware algorithm we define $H_{\sigma,1}$ (as in Gabillon et al. (2011)) that incorporates reward variances into its expression as:

$$H_{\sigma,1} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}.$$

- Finally, analogous to H_2, we introduce $H_{\sigma,2}$, such that $H_{\sigma,2} = \max_{i \in \mathcal{A}} \frac{i}{\tilde{\Delta}_{(i)}^2}$, where $\tilde{\Delta}_i^2 = \frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}$, $(\tilde{\Delta}_{(i)})$ is an increasing ordering of $(\tilde{\Delta}_i)$.
- From [Audibert and Bubeck(2010)], we can show that

$$H_{\sigma,2} \leq H_{\sigma,1} \leq \log(2K) H_{\sigma,2}.$$

Problem Complexity

- For a variance aware algorithm we define $H_{\sigma,1}$ (as in Gabillon et al. (2011)) that incorporates reward variances into its expression as:

$$H_{\sigma,1} = \sum_{i=1}^{K} \frac{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}{\Delta_i^2}.$$

- Finally, analogous to H_2, we introduce $H_{\sigma,2}$, such that $H_{\sigma,2} = \max_{i \in \mathcal{A}} \frac{i}{\tilde{\Delta}_{(i)}^2}$, where $\tilde{\Delta}_i^2 = \frac{\Delta_i^2}{\sigma_i + \sqrt{\sigma_i^2 + (16/3)\Delta_i}}$, $(\tilde{\Delta}_{(i)})$ is an increasing ordering of $(\tilde{\Delta}_i)$.
- From [Audibert and Bubeck(2010)], we can show that

$$H_{\sigma,2} \le H_{\sigma,1} \le \log(2K) H_{\sigma,2}.$$

- Note that H_1, H_2 and $H_{\sigma,1}, H_{\sigma,2}$ are not directly comparable to each other except in a special case when variances and gaps (Δ_i) are very low we can say that $H_{\sigma,1} < H_1$.

Expected Loss of AugUCB

Theorem

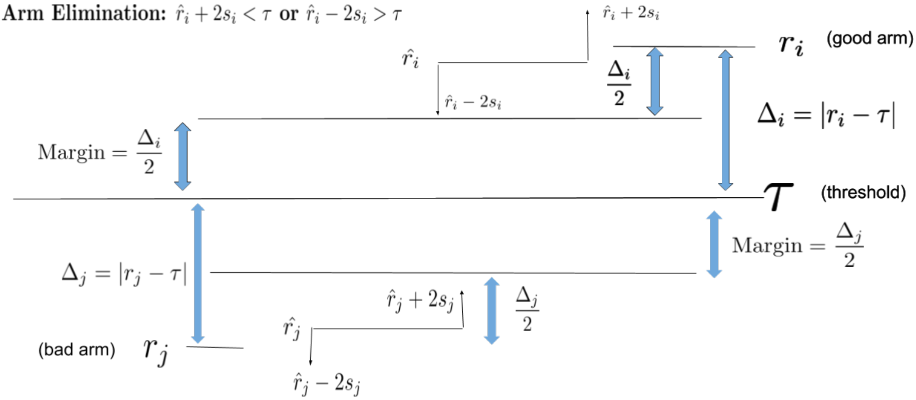
For $K \geq 4$ and $\rho = 1/3$, the expected loss of the AugUCB algorithm is given by,

$$\mathbb{E}[\mathcal{L}(T)] \leq 2KT \exp\left(- \frac{T}{4096 \log(K \log K) H_{\sigma,2}} \right).$$

Table: AugUCB vs. State of the art

Algorithm	Upper Bound on Expected Loss
AugUCB	$\exp\left(- \frac{T}{4096 \log(K \log K) H_{\sigma,2}} + \log(2KT) \right)$
UCBEV	$\exp\left(- \frac{1}{512} \frac{T - 2K}{H_{\sigma,1}} + \log(6KT) \right)$
APT	$\exp\left(- \frac{T}{64 H_1} + 2 \log((\log(T) + 1) K) \right)$
CSAR	$\exp\left(- \frac{T - K}{72 \log(K) H_{CSAR,2}} + 2 \log(K) \right)$

Sketch of the proof



Finally, experiment!!!

- We compare with APT, AugUCB, UCBE, UCBEV, CSAR, UA.

Finally, experiment!!!

- We compare with APT, AugUCB, UCBE, UCBEV, CSAR, UA.
- Note that UCBE and UCBEV require access to H_1 and $H_{\sigma,1}$ as input and hence not implementable in real life.
- By access we mean that an oracle supplies them the H_1 or $H_{\sigma,1}$. They do not have access to individual means and variances.

Finally, experiment!!!

- We compare with APT, AugUCB, UCBE, UCBEV, CSAR, UA.
- Note that UCBE and UCBEV require access to H_1 and $H_{\sigma,1}$ as input and hence not implementable in real life.
- By access we mean that an oracle supplies them the H_1 or $H_{\sigma,1}$. They do not have access to individual means and variances.
- APT, AugUCB, CSAR, UA do not require access to H_1 or $H_{\sigma,1}$.

Finally, experiment!!!

- We compare with APT, AugUCB, UCBE, UCBEV, CSAR, UA.
- Note that UCBE and UCBEV require access to H_1 and $H_{\sigma,1}$ as input and hence not implementable in real life.
- By access we mean that an oracle supplies them the H_1 or $H_{\sigma,1}$. They do not have access to individual means and variances.
- APT, AugUCB, CSAR, UA do not require access to H_1 or $H_{\sigma,1}$.
- UCBE, UCBEV, CSAR and UA come from the pure exploration lineage and are modified suitably to perform in TBP setting.

Experimental Setup

- This setup involves Gaussian reward distributions with $K = 100$, $T = 10000$ and $\tau = 0.5$ with the reward means set in two groups.

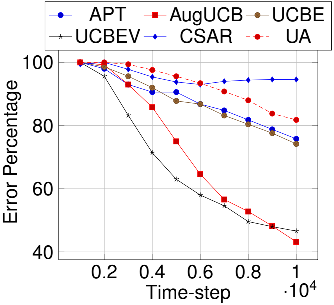
Experimental Setup

- This setup involves Gaussian reward distributions with $K = 100$, $T = 10000$ and $\tau = 0.5$ with the reward means set in two groups.
- The first 10 arms partitioned into two groups; the respective means are $r_{1:5} = 0.45$, $r_{6:10} = 0.55$.

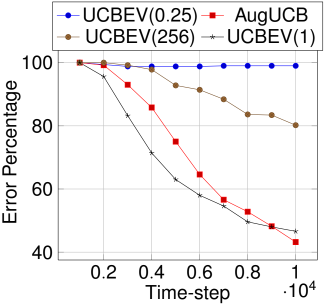
Experimental Setup

- This setup involves Gaussian reward distributions with $K = 100$, $T = 10000$ and $\tau = 0.5$ with the reward means set in two groups.
- The first 10 arms partitioned into two groups; the respective means are $r_{1:5} = 0.45$, $r_{6:10} = 0.55$.
- The means of arms $i = 11 : 100$ are chosen same as $r_{11:100} = 0.4$.
- Variances are set as $\sigma^2_{1:5} = 0.3$ and $\sigma^2_{6:10} = 0.8$; $\sigma^2_{11:100}$ are independently and uniformly chosen in the interval $[0.2, 0.3]$.

Experimental Results



(a) Expt-1: Two Group Setting (Advance)



(b) Expt-2: Two Group Setting (Advance)

Conclusion

- We proposed the AugUCB algorithm for the fixed budget TBP which uses variance estimation and arm elimination to give an improved theoretical and experimental guarantees than APT.
- This work has been accepted in the proceedings of IJCAI 2017.

Conclusion

- We proposed the AugUCB algorithm for the fixed budget TBP which uses variance estimation and arm elimination to give an improved theoretical and experimental guarantees than APT.
- This work has been accepted in the proceedings of IJCAI 2017.
- Further studies are required to establish a lower bound on the expected loss of AugUCB.

Conclusion

- We proposed the AugUCB algorithm for the fixed budget TBP which uses variance estimation and arm elimination to give an improved theoretical and experimental guarantees than APT.
- This work has been accepted in the proceedings of IJCAI 2017.
- Further studies are required to establish a lower bound on the expected loss of AugUCB.
- A more detailed analysis of the non-uniform arm selection and parameter selection is also required.

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.

Jacob D Abernethy, Kareem Amin, and Ruihao Zhu.
Threshold bandits, with and without censored feedback.
In *Advances In Neural Information Processing Systems*, pages 4889–4897, 2016.

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

References II

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 III

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.

References IV

Dimitri P Bertsekas and John N Tsitsiklis.
Neuro-dynamic programming (optimization and neural computation series, 3).
Athena Scientific, 7:15–23, 1996.

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

Sébastien Bubeck, Rémi Munos, and Gilles Stoltz.
Pure exploration in finitely-armed and continuous-armed bandits.
Theoretical Computer Science, 412(19):1832–1852, 2011.

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

- Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Bounded regret in stochastic multi-armed bandits. *arXiv preprint arXiv:1302.1611*, 2013a.
- Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *ICML (1)*, pages 258–265, 2013b.
- Olivier Cappe, Aurelien Garivier, and Emilie Kaufmann. pymabandits, 2012. `http://mloss.org/software/view/415/`.
- Shouyuan Chen, Tian Lin, Irwin King, Michael R Lyu, and Wei Chen. Combinatorial pure exploration of multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 379–387, 2014.

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.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck.
Multi-bandit best arm identification.
In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2011.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric.
Best arm identification: A unified approach to fixed budget and fixed confidence.
In *Advances in Neural Information Processing Systems*, pages 3212–3220, 2012.

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

Mohammad Ghavamzadeh, Shie Mannor, Joelle Pineau, Aviv Tamar, et al.
Bayesian reinforcement learning: a survey.
World Scientific, 2015.

References VIII

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

Kevin Jamieson and Robert Nowak.
Best-arm identification algorithms for multi-armed bandits in the fixed confidence setting.
In *Information Sciences and Systems (CISS), 2014 48th Annual Conference on*, pages 1–6. IEEE, 2014.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone.
Pac subset selection in stochastic multi-armed bandits.
In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 655–662, 2012.

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.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier.
An optimal algorithm for the thresholding bandit problem.
arXiv preprint arXiv:1605.08671, 2016.

References X

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.

Subhojyoti Mukherjee, K. P. Naveen, Nandan Sudarsanam, and Balaraman Ravindran.
Thresholding bandits with augmented UCB.
CoRR, abs/1704.02281, 2017.
URL `http://arxiv.org/abs/1704.02281`.

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

References XI

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