Adaptive Multi-armed Bandit Algorithms for Markovian and i.i.d Rewards

Arghyadip Roy (Mehta Family School of Data Science & Artificial Intelligence, IIT Guwahati)
Joint work with
Sanjay Shakkottai (University of Texas at Austin)
&
R. Srikant (University of Illinois at Urbana-Champaign)

Outline

- Introduction
- Related Work
- TV-KL-UCB Algorithm
- Regret Upper Bound
- Conclusions & Future Work

Introduction

- Exploration vs. Exploitation
  - Fundamental trade-off in decision making
    - Exploitation: Choose best action given current knowledge
    - Exploration: Gather more knowledge
  - Example: Go to favorite restaurant vs try a new one, online advertisement

- Multi-armed bandit problem
  - Choose arms sequentially from a set of arms
    - Each arm produces reward: statistics of reward distribution unknown
    - Maximize total reward (minimize total regret)

Introduction

Stochastic MAB

I.I.D

Markovian

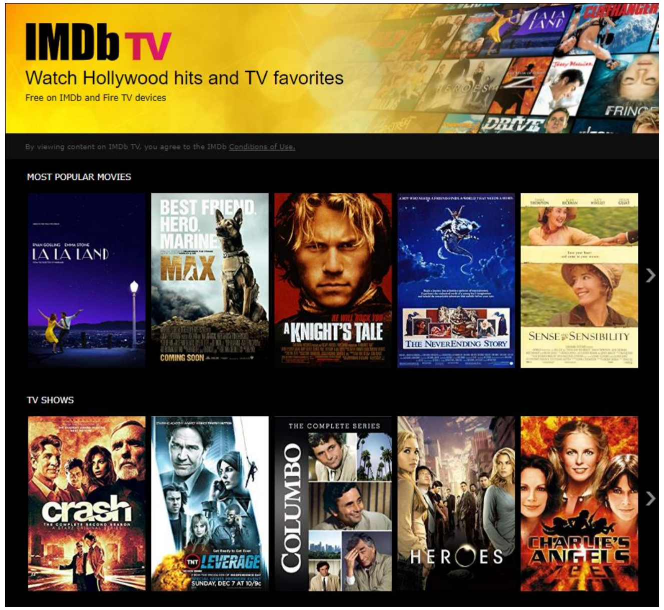
[Lai et al, 1985]
[Thompson, 1933]
[Anantharam et al, 1987]
UCB [Auer et al, 2002]
KL-UCB [Garivier et al, 2011]

Rested MAB

Restless MAB

State of arms not played : Frozen

Introduction



High reward in current play-> low reward in next play (high probability)

Future movie selection depends on customer's past response

Outcome of an intervention depends on those of past interventions

Source: Google

Related Work

- [Anantharam et al 1987]
  - Index policy: matches lower bound
- [Moulos 2020]
  - Multiple play: Extension of KL-UCB using sample mean

Single-parameter family of transition matrix

- [Tekin et al. 2010]
  - UCB based policy: sample mean reward
  - Logarithmic regret: constants not optimal

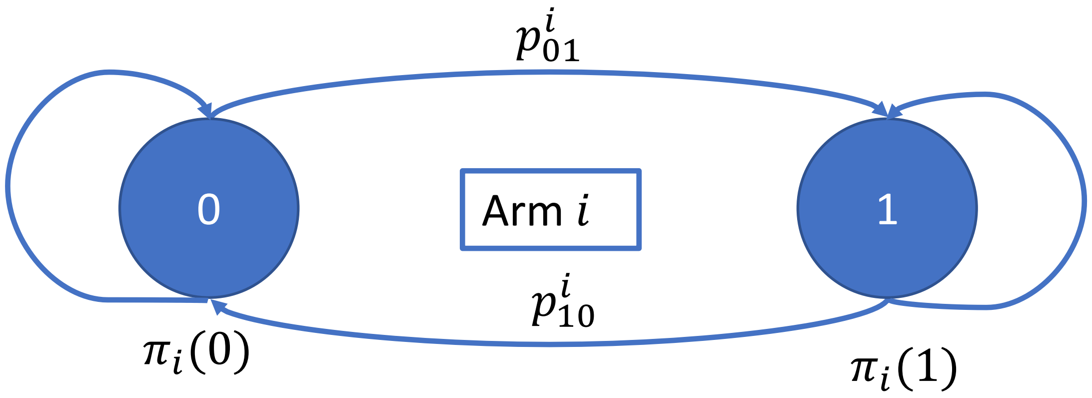
Our Contributions

- Extension of KL-UCB for Markovian bandits
  - Sample transition probability based KL-UCB
  - Outperforms [Tekin et al, 2010] for Markovian rewards
  - Bad for i.i.d rewards (special case of Markovian bandits)
- Identify rewards: Markovian/i.i.d
  - TV distance based test using estimates of transition probability
  - Switch from sample transition probability to sample mean based KL-UCB
- Upper bound on regret
  - Invertibility of KL divergence: Does not hold in multi-parameter setting
  - Collection of single parameter problems
  - Appropriate condition on TV distance satisfied infinitely often

Our Contributions

- No parameterization on transition probability matrix
- Only assumption: Irreducibility of Markov chain
- Lower regret than [Tekin et al 2010]
  - Sample mean : Not a unique representation for truly Markovian arms
  - KL-UCB: Tighter confidence bound than UCB

Problem Formulation & Preliminaries



- Reward from arm i in state $s = r(s,i) = s$
- Mean reward from arm i is $\mu_i = \sum_{s=0}^{1} s\pi_i(s) = \pi_i(1) = \frac{p_{01}^{i}}{p_{10}^{i}+p_{01}^{i}}$
- $\mu^* = \max_i \mu_i = \mu_1, \; \Delta_i = \mu_1 - \mu_i$ (suboptimality gap)

Problem Formulation & Preliminaries

- Regret of policy α till time $n = R_\alpha(n)$

$$= n\mu_1 - E_\alpha\left[\sum_{t=1}^{n} r\big(s(\alpha(t)), \alpha(t)\big)\right]$$

i.e., regret of policy α till time $n =$ Difference of mean rewards under optimal policy and policy α

Bandit Algorithms

- Never explore (Greedy)
  - Choose arm with greatest mean estimate
  - May lock into sub-optimal arm: Linear regret
- Forever explore (ϵ-greedy)
  - Explore with probability ϵ, exploit with remaining probability
  - Linear regret
- ϵ_t-greedy
  - Exploration probability decays with time
  - Sublinear (logarithmic) regret: Requires knowledge of mean reward
- Design mechanism with sublinear regret without reward knowledge

Bandit Algorithms

- [Lai et al 1985] Lower bound on regret is logarithmic in time asymptotically: at least $O(\log n)$ suboptimal pulls
- Algorithm is order-optimal if regret= $O(\log n)$
- ϵ-greedy: Exploration w/o any preference for nearly greedy/arm with uncertain estimate
- Add upper confidence to estimated mean: overestimate true mean with high probability
- Large $T_i(t)$: small upper confidence
- Select arm which maximizes upper confidence bound
  - Explore uncertain arms, exploit arms with high estimates
  - As $t \to \infty$, select optimal arm

KL-UCB-SM Algorithm [Garivier et al, 2011]

- Rewards from arm i: Bernoulli(μ_i)
- Usage of KL distance as upper confidence bound
- $D(a||b) = a\log\frac{a}{b} + (1-a)\log\frac{1-a}{1-b}$

Assume $W_1, W_2, \ldots, W_n$ is a sequence of Bernoulli R.V. with mean μ. Then,

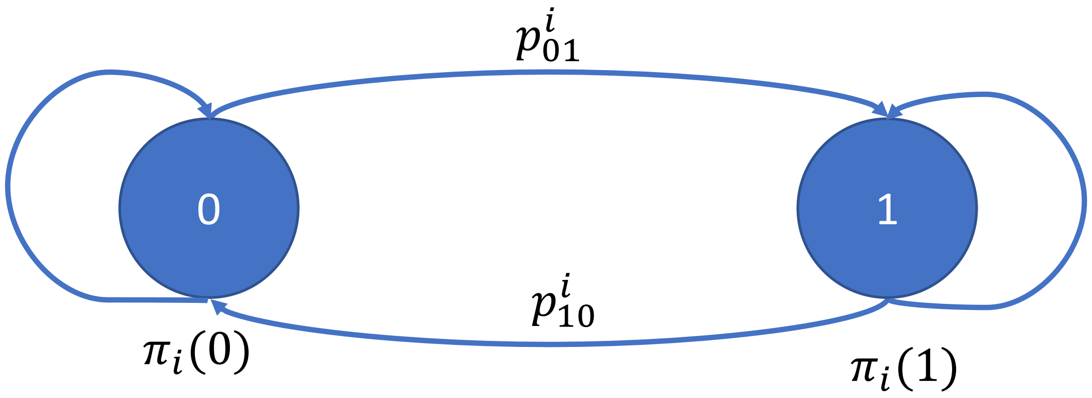
$$P(\hat{\mu} \le \mu - \epsilon) \le \exp(-nD(\mu - \epsilon||\mu))$$

Chernoff's Bound

- Using Chernoff's bound and appropriate confidence interval, we get
  - Choose $A_t = \underset{i}{arg\,max}\, sup\left\{\tilde{\mu} \in [\hat{\mu}^i(t-1), 1]: D\left(\hat{\mu}^i(t-1), \tilde{\mu}\right) \le \frac{log\, f(t)}{T_i(t-1)}\right\}$

    where $f(t) = 1 + t\, log^2(t)$
- Failure of confidence interval goes to zero slightly faster than $\frac{1}{t}$: Logarithmic regret
- Asymptotically optimal for i.i.d rewards

KL-UCB-MC Algorithm



$$\mu_i = \frac{p_{01}^i}{p_{10}^i + p_{01}^i}$$

- Two parameters instead of one parameter as in i.i.d. arms
- Simultaneous confidence bounds on $\hat{p}_{01}^i$ and $\hat{p}_{10}^i$: Analysis difficult
- Use confidence bound on estimate of one parameter at a time
- Use raw estimate of the other parameter

KL-UCB-MC Algorithm

- Index of arm i
  - Use upper confidence bound on $\hat{p}_{01}^{i}$ and $\hat{p}_{10}^{i}$ in state 0

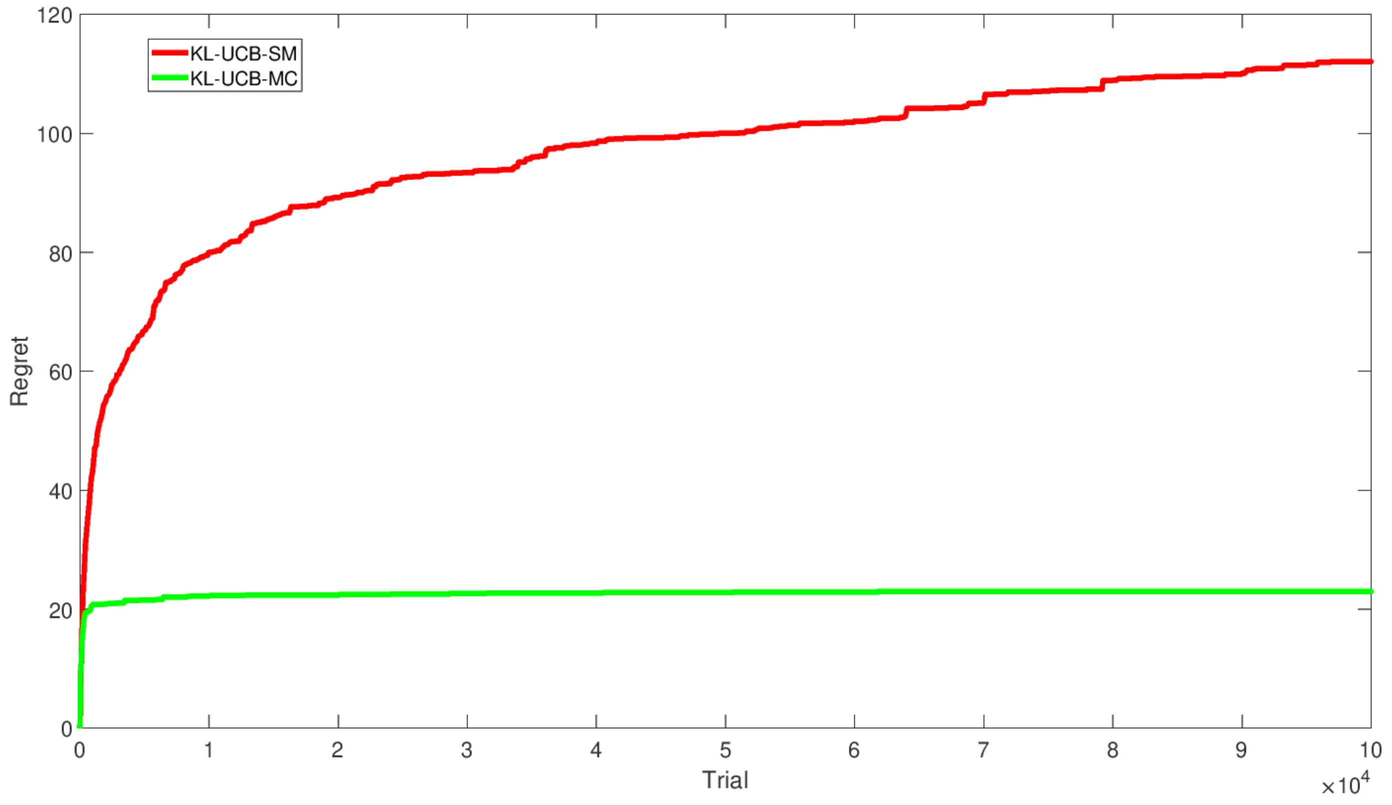
$$U_i = sup\left\{\frac{\tilde{p}}{\tilde{p} + \hat{p}_{10}^{i}(t-1)} : D\left(\hat{p}_{01}^{i}(t-1), \tilde{p}\right) \leq \frac{log\, f(t)}{T_i(t-1)}\right\}$$

  - Use lower confidence bound on $\hat{p}_{10}^{i}$ and $\hat{p}_{01}^{i}$ in state 1

$$U_i = sup\left\{\frac{\hat{p}_{01}^{i}(t-1)}{\hat{p}_{01}^{i}(t-1) + \tilde{q}} : D\left(\hat{p}_{10}^{i}(t-1), \tilde{q}\right) \leq \frac{log\, f(t)}{T_i(t-1)}\right\}$$

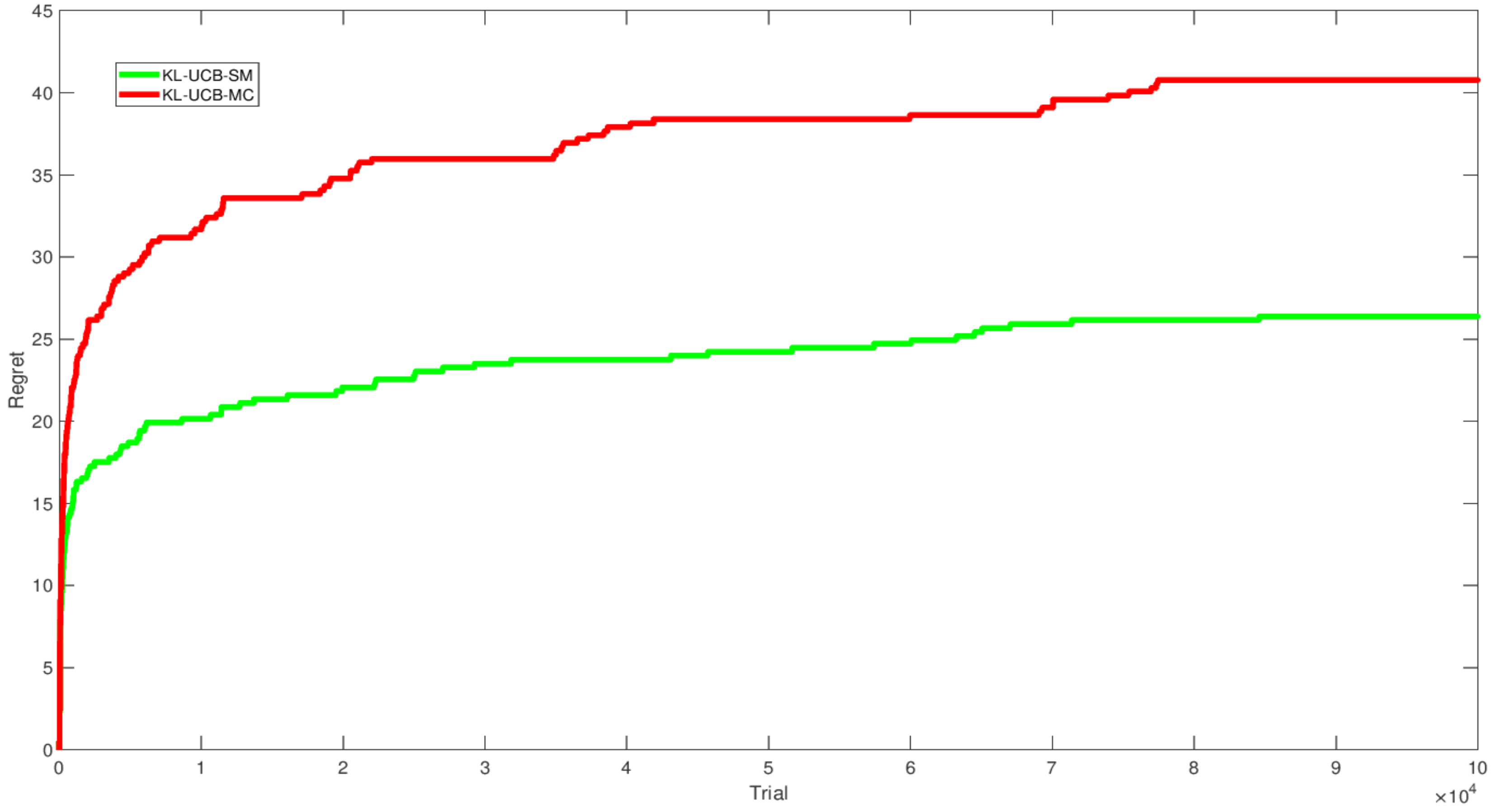
- Choose $A_t = arg\,\max_i U_i$

KL-UCB-MC Algorithm



Truly Markovian rewards

KL-UCB-MC Algorithm



i.i.d rewards

TV-KL-UCB Algorithm

- i.i.d. rewards: Special case of Markovian rewards ($p_{01}^{i} + p_{10}^{i} = 1$)
- KL-UCB-SM [Garivier et al, 2011] known to be optimal for i.i.d rewards
- KL-UCB-MC performs bad for i.i.d rewards
- Design of test for detecting i.i.d/ Markovian arm online
  - Switch from KL-UCB-MC to KL-UCB-SM if i.i.d. reward
  - Truly Markovian arm: Can be described uniquely by p_{01}^{i} and p_{10}^{i}
  - i.i.d arm: can be described uniquely by μ_i
  - Appropriate condition satisfied infinitely often
  - Regret due to incorrect variant vanishes asymptotically

TV-KL-UCB Algorithm

- TV distance: Depicts similarity between two probability distributions (Similar to KL distance)
- Two discrete prob. dist. $A = (a_1, \ldots, a_k)$ and $\mathrm{B} = (b_1, \ldots, b_k)$

$$TV(A||B) = \frac{1}{2}\sum_{i=1}^{k} |a_i - b_i|$$

- TV distance chosen for analytical convenience
- Test for detecting i.i.d/ Markovian arm:
  - i.i.d arms: $p_{01}^{i} + p_{10}^{i} = 1$
  - TV distance between p_{01}^{i} and $1 - p_{10}^{i}$
  - Condition for testing: $TV(\hat{p}_{01}(t)||1 - \hat{p}_{10}(t)) < \frac{1}{t^{\frac{1}{4}}}$

TV-KL-UCB Algorithm

Algorithm 1 Total Variation KL-UCB Algorithm (TV-KL-UCB)

Input K (number of arms).
Choose each arm once.
while TRUE **do**
  if $(|\hat{p}^i_{01}(t-1) + \hat{p}^i_{10}(t-1) - 1|) > \frac{1}{(t-1)^{1/4}}$ **(procedure** STP_PHASE) **then**
    if (state of arm $i = 0$) **then**

$$U_i = \sup\{\frac{\tilde{p}}{\tilde{p} + \hat{p}^i_{10}(t-1)} : D(\hat{p}^i_{01}(t-1)||\tilde{p}) \leq \frac{\log f(t)}{T_i(t-1)}\}. \quad (1)$$

    else

$$U_i = \sup\{\frac{\hat{p}^i_{01}(t-1)}{\hat{p}^i_{01}(t-1) + \tilde{q}} : D(\hat{p}^i_{10}(t-1)||\tilde{q}) \leq \frac{\log f(t)}{T_i(t-1)}\}. \quad (2)$$

    end if
  else (procedure SM_PHASE)

$$U_i = \sup\{\tilde{\mu} \in [\hat{\mu}^i(t-1), 1] : D(\hat{\mu}^i(t-1)||\tilde{\mu}) \leq \frac{\log f(t)}{T_i(t-1)}\}. \quad (3)$$

  end if
  Choose $A_t = \arg\max_i U_i$.
end while

KL-UCB-MC:
Markovian arms

KL-UCB-SM:
i.i.d. arms

Regret Upper Bound

Fractions of visits to states close to stationary probabilities

+

Estimates of Transition probabilities close to true values

+

TV distance condition satisfied i.o. after a large time τ

+

Confidence bounds for estimates of transition probabilities of optimal arm close to true values

$\tau = \mathrm{C}_1 \log \mathrm{n} + \mathrm{C}_2$,
Where $\mathrm{C}_1 = f(p_{01}^i, p_{10}^i, p_{01}^1, p_{10}^1)$,
$\mathrm{C}_2 = g(p_{01}^i, p_{10}^i, p_{01}^1, p_{10}^1)$

$\Downarrow$

Index of sub-optimal arm is not often greater than index of optimal arm

Regret Upper Bound

Theorem 1: *Asymptotic regret is bounded by*

(Truly Markovian optimal and sub-optimal arms):

$$\limsup_{n\to\infty} \frac{R_n}{\log n} \leq \sum_{i\neq 1}\left[\frac{2}{D\left(p_{01}^i \| \frac{p_{01}^1 p_{10}^i}{p_{10}^1}\right)} 1\{p_{01}^1 p_{10}^i < p_{10}^1\} + \frac{2}{D\left(p_{10}^i \| \frac{p_{10}^1 p_{01}^i}{p_{01}^1}\right)}\right]$$

- Similarly, upper bound for other three combinations can be derived.
- Upper bound matches the lower bound when all arms are i.i.d.

$$\liminf_{n\to\infty} \frac{R_n}{\log n} \geq \sum_{i\neq 1} \frac{1}{\pi_i(0) D(p_{01}^i \| p_{01}^1) + \pi_i(1) D(p_{10}^i \| p_{10}^1)}$$

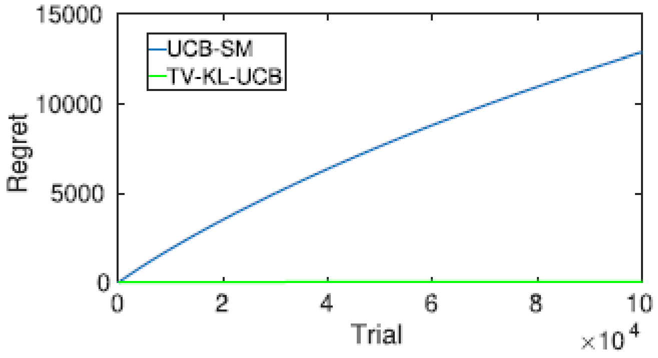
Lower bound

Regret Upper Bound

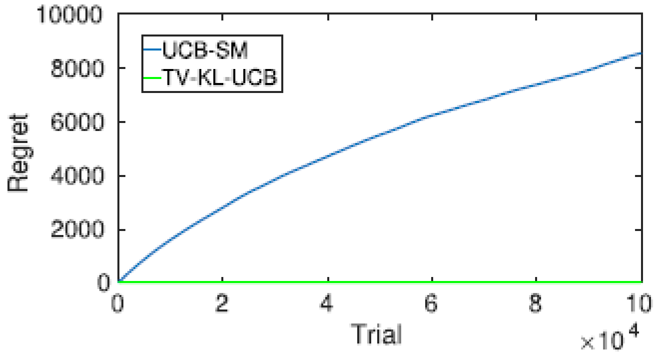
Theorem 2: *Let the eigenvalue gap of arm i be σ_i. Asymptotic upper bound smaller than UCB-SM [Tekin et al, 2010]*

1. *Truly Markovian suboptimal arms: if* $\min_i \sigma_i \geq \frac{1}{1440}$.
2. *i.i.d. suboptimal arms: Always*

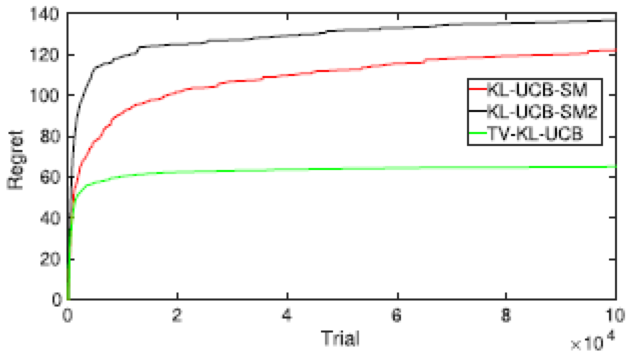
Experimental Evaluation



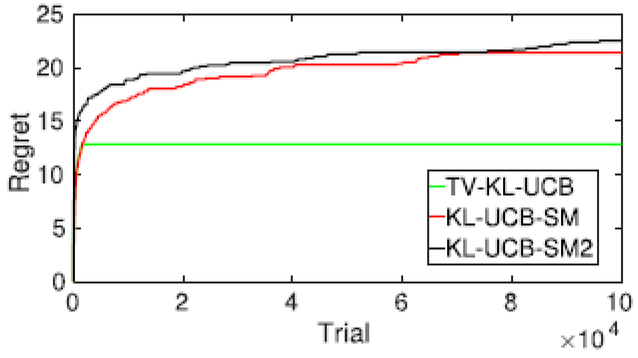
(a)



(b)



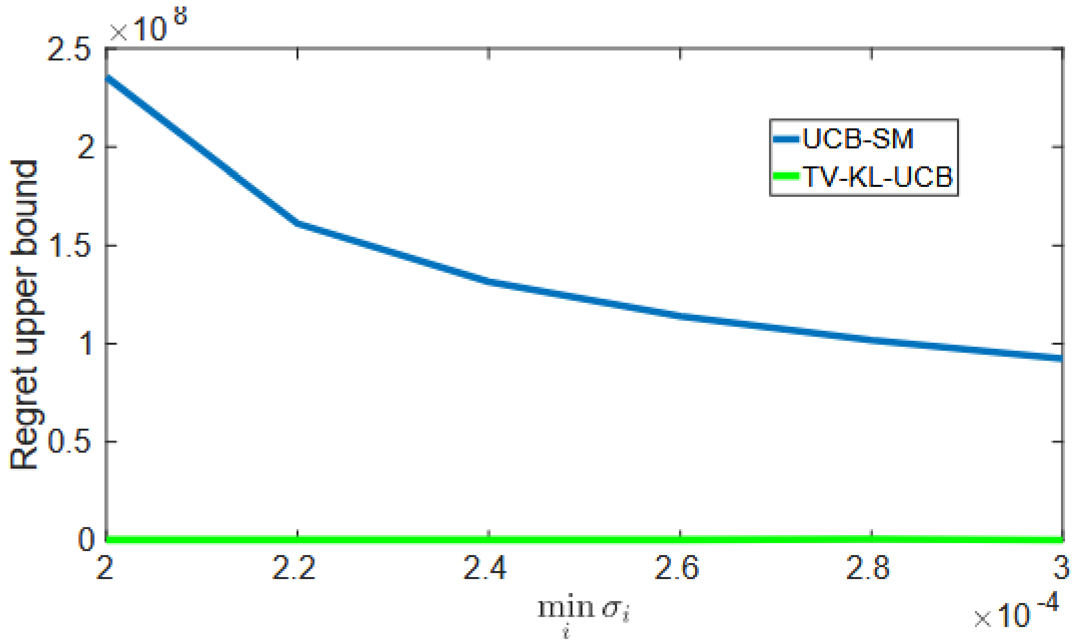
(c)



(d)

- UCB-SM: [Tekin et al, 2010]
- KL-UCB-SM: KL-UCB version of [Tekin et al, 2010]
- KL-UCB-SM2: Single play version of [Moulos, 2020]

Experimental Evaluation



Upper bound on regret smaller than UCB-SM even when condition on Theorem 2 not satisfied

Conclusions

- TV-KL-UCB detects arm reward Markovian/i.i.d using TV distance based test
  - Arm i can be represented uniquely using p_{01}^{i} and p_{10}^{i}
  - If arm i.i.d., unique representation using μ_i
  - Switch from sample transition probability KL-UCB to sample mean KL-UCB
- Regret upper bound matches lower bound when arm reward i.i.d
- Significant improvement over state-of-the-art bandit algorithms

Future Work

- Use of other metric such as KL distance for testing Markovian/i.i.d
  - Easy to obtain upper bound involving additive separability of estimates with TV/Hellinger distance
  - Difficult in case of KL distance

- Design of asymptotically optimal algorithm for truly Markovian arms

Roy, Arghyadip, Sanjay Shakkottai, and R. Srikant. "Adaptive KL-UCB based Bandit Algorithms for Markovian and iid Settings." Vol 69, Issue 4, IEEE Transactions on Automatic Control, pp-2637-2644, 2024.

Other Research Activities

Task Scheduling Policy for IoT based Mobile Edge Computing

Multi-armed Bandit Algorithms for Beam Tracking in mm-wave MIMO

UAV placement in next-generation wireless systems

Federated Learning for IoT systems

M. Moharrami, Y. Murthy, A. Roy and R. Srikant, "*A Policy Gradient Algorithm for the Risk-Sensitive Exponential Cost MDP*," Accepted in Mathematics of Operations Research, 2024
S. Badireddi, R. Banerjee, P. Shah, A. Roy,“ *Exploiting Bias in Reinforcement Learning for Task Allocation in a Mobile Edge Computing System,"* IEEE International Conference on Signal Processing and Communications (SPCOM) ,2024
A. Kumar, A. Roy and R. Bhattacharjee, “ *Actively Adaptive Multi-armed Bandit Based Beam Tracking for mmWave MIMO Systems,"* IEEE Wireless Communications and Networking Conference (WCNC), 2024
A. Roy and N. Biswas,“ *GoPro: A Low Complexity Task Allocation Algorithm for a Mobile Edge Computing System,*” IEEE National Conference on Communications (NCC) 2022

Arghyadip Roy | IIT Guwahati

References

- H. Robbins, “Some aspects of the sequential design of experiments,” Bulletin of the American Mathematical Society, vol. 58, no. 5, pp. 527–535, 1952.
- T. L. Lai and H. Robbins, “Asymptotically efficient adaptive allocation rules,” Advances in applied mathematics, vol. 6, no. 1, pp. 4–22, 1985.
- P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire, “The nonstochastic multiarmed bandit problem,” SIAM journal on computing, vol. 32, no. 1, pp. 48–77, 2002.
- W. R. Thompson, “On the likelihood that one unknown probability exceeds another in view of the evidence of two samples,” Biometrika, vol. 25, no. 3/4, pp. 285–294, 1933.
- V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays part i: IID rewards,” IEEE Transactions on Automatic Control, vol. 32, no. 11, pp. 968–976, 1987.
- P. Auer, N. Cesa-Bianchi, and P. Fischer, “Finite-time analysis of the multiarmed bandit problem,” Machine learning, vol. 47, no. 2-3, pp. 235–256, 2002.
- A. Garivier and O. Cappé, “The KL-UCB algorithm for bounded stochastic bandits and beyond,” in conference on learning theory, 2011,pp. 359–376.

References

- C. Tekin and M. Liu, “Online algorithms for the multi-armed bandit problem with Markovian rewards,” in IEEE Annual Allerton Conference on Communication, Control, and Computing, 2010, pp. 1675–1682.
- V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays part ii: Markovian rewards,” IEEE Transactions on Automatic Control, vol. 32, no. 11, pp. 977–982, 1987.
- V. Moulos, “Finite-time analysis of Kullback-Leibler upper confidence bounds for optimal adaptive allocation with multiple plays and Markovian rewards,” arXiv preprint arXiv:2001.11201, 2020.
- T. Lattimore and C. Szepesvári, “Bandit algorithms,” preprint, 2018.

Thank you