Developing and analyzing algorithms for the Multi-armed Bandit

Sanit Gupta

December 4, 2019

Preliminaries

- What is the multi-armed bandit (MAB) problem?
- In the MAB problem, an agent must choose to pull one of n available arms. Each arm has a reward distribution associated with it. These distributions are fixed but unknown.

Preliminaries

- What is the multi-armed bandit (MAB) problem?
- In the MAB problem, an agent must choose to pull one of n available arms. Each arm has a reward distribution associated with it. These distributions are fixed but unknown.
- We study the regret minimization setting for Bernoulli bandits.
- Expected regret is defined as:

$$E[R(T)] = \mu^* T - E[\sum_{t=1}^{T} r(t)]$$

  where $R(T)$ is the cumulative regret in T time steps and $r(t)$ is the reward received in the t^{th} time step
- We wish to design algorithms aimed at minimizing this quantity.

Previous Work (1)

- Lai and Robbins (1985) gave lower bounds on regret for all bandit algorithms:

$$E[R(T)] \geq [\Sigma_{i:\mu_i<\mu^*} \frac{\Delta_i}{D(\mu_i||\mu^*)} + o(1)] lnT$$

where D is KL divergence

Previous Work (1)

- Lai and Robbins (1985) gave lower bounds on regret for all bandit algorithms:

$$E[R(T)] \geq [\Sigma_{i:\mu_i<\mu^*} \frac{\Delta_i}{D(\mu_i||\mu^*)} + o(1)] lnT$$

  where D is KL divergence
- Some popular algorithms that match the lower bound are UCB1 and Thompson Sampling.

Previous Work (2)

- Upper bound on expected regret for UCB1 from Auer et al. (2002):

$$E[R(T)] \leq 8[\Sigma_{i:\mu_i<\mu^*} \frac{lnT}{\Delta_i}] + (1 + \frac{\pi^2}{3})(\Sigma_{j=1}^{K} \Delta_j)$$

Previous Work (2)

- Upper bound on expected regret for UCB1 from Auer et al. (2002):

$$E[R(T)] \leq 8[\Sigma_{i:\mu_i<\mu^*} \frac{lnT}{\Delta_i}] + (1 + \frac{\pi^2}{3})(\Sigma_{j=1}^{K} \Delta_j)$$

- Upper bound on expected regret of Thompson Sampling from Kaufmann et al. (2012):

$$E[R(T)] \leq (1+\epsilon)\Sigma_{i:\mu_i<\mu^*} \frac{\Delta_i(ln(T) + ln(ln(T)))}{D(\mu_i||\mu^*)} + C(\epsilon, \mu_1, .., \mu_n)$$

where ϵ and C are problem dependent constants

Key Idea - Persistence

- A bandit algorithm is a map from the whole history of arms pulled and rewards observed to an arm choice or a probability distribution over arms
- Given a history, a bandit algorithm will return an arm or a distribution over arms.

Key Idea - Persistence

- A bandit algorithm is a map from the whole history of arms pulled and rewards observed to an arm choice or a probability distribution over arms
- Given a history, a bandit algorithm will return an arm or a distribution over arms.
- We can add 'Persistence' to any bandit algorithm.
- In the persistence variant of any bandit algorithm, whenever one gets a 1 reward, they stick with their choice for the next time instance ignoring the rest of the history.

Persistence - Pseudocode

Algorithm 1 Persistence Variant of Bandit_Algorithm

```
n ← Number of Arms
T ← Time Horizon
for i = 1 to n do
    true_reward_distribution[i] ← Bernoulli(μ_i)
reward_history ← [ ]
action_history ← [ ]

r ← 0
for i = 1 to T do
    if r == 0 then  //remove this condition −> regular bandit algorithm
        action_choice ← bandit_algorithm(action_history, reward_history)
    r ← sample(true_reward_distribution[action_choice])
    reward_history ← reward_history.append(r)
    action_history ← action_history.append(action_choice)
```

Persistence - Intuition

- Why should persistence work?

Persistence - Intuition

- Why should persistence work?
- The intuition behind persistence is that it will lead to arms with higher means being pulled more often and hence result in lower regret.

Persistence - Intuition

- Why should persistence work?
- The intuition behind persistence is that it will lead to arms with higher means being pulled more often and hence result in lower regret.
- When an arm with mean μ_i is picked, in expectation, with persistence, it will be picked $\frac{1}{1-\mu_i}$ times before the next decision needs to be made about which arm to pick.

Persistence - Intuition

- Why should persistence work?
- The intuition behind persistence is that it will lead to arms with higher means being pulled more often and hence result in lower regret.
- When an arm with mean μ_i is picked, in expectation, with persistence, it will be picked $\frac{1}{1-\mu_i}$ times before the next decision needs to be made about which arm to pick.
- Therefore, with persistence we expect the better arms to be picked more often and hence incur lesser regret as compared to the regular variant.

Empirical Results

- Our experiments are done for the two-armed case because, usually, analysis of the two-armed bandit can be generalised to the n-armed bandit.
- We wish to examine how ‘persistence’ influences the performance of various bandit algorithms.

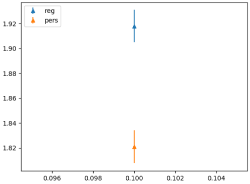
Empirical Results

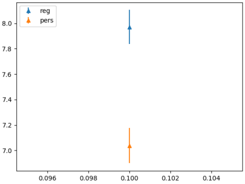
- Our experiments are done for the two-armed case because, usually, analysis of the two-armed bandit can be generalised to the n-armed bandit.
- We wish to examine how ‘persistence’ influences the performance of various bandit algorithms.
- For ϵ-greedy, we use a constant value of ϵ. In our experiments, we keep $\epsilon = 0.05$. It picks the arm with the highest empirical mean with probability $1 - \epsilon$, and a random arm with ϵ.

Empirical Results

- Our experiments are done for the two-armed case because, usually, analysis of the two-armed bandit can be generalised to the n-armed bandit.
- We wish to examine how 'persistence' influences the performance of various bandit algorithms.
- For ϵ-greedy, we use a constant value of ϵ. In our experiments, we keep $\epsilon = 0.05$. It picks the arm with the highest empirical mean with probability $1 - \epsilon$, and a random arm with ϵ.
- The other algorithm we run experiments for is Thompson Sampling.

Graphs for ϵ-greedy (1)





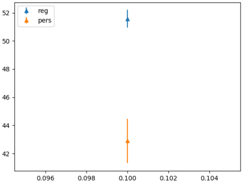
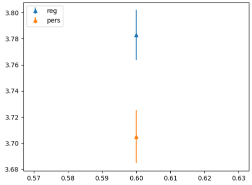
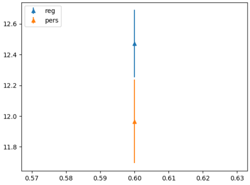


Table: Regret for (0.3, 0.1) Horizons = 100, 1000 and 10000; Orange = Persistence; Blue = Regular

Graphs for ϵ-greedy (2)





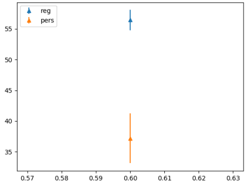


Table: Regret for (0.8, 0.6) Horizons = 100, 1000 and 10000; Orange = Persistence; Blue = Regular

Graphs for ϵ-greedy (3)

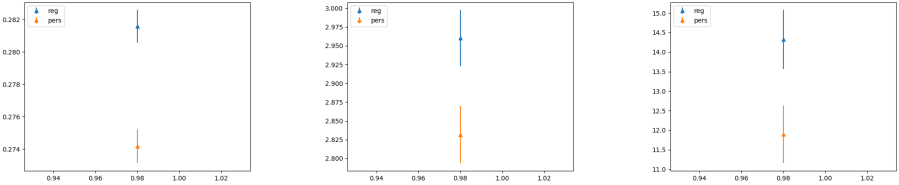


Table: Regret for (0.99, 0.98) Horizons = 100, 1000 and 10000; Orange = Persistence; Blue = Regular

For ϵ-greedy

- We can clearly see that throughout these graphs, persistent ϵ-greedy outperforms regular ϵ-greedy
- At least empirically, it seems clear that persistence improves ϵ-greedy

Graphs for Thompson Sampling (1)

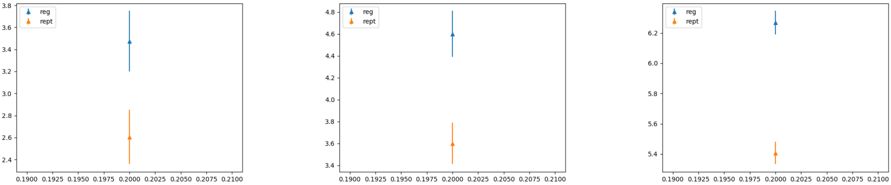


Table: Regret for (0.7, 0.2) Horizons = 100, 1000 and 10000; Orange = Persistence; Blue = Regular

Graphs for Thompson Sampling (2)

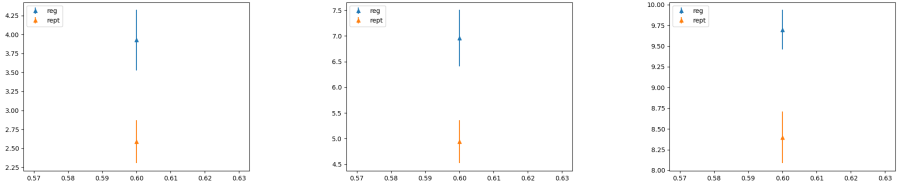
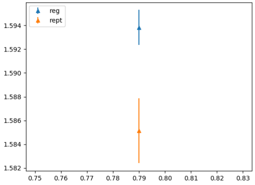
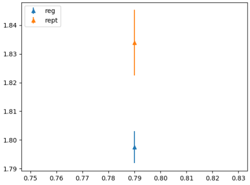


Table: Regret for (0.8, 0.6) Horizons = 100, 1000 and 10000; Orange = Persistence; Blue = Regular

Graphs for Thompson Sampling (3)





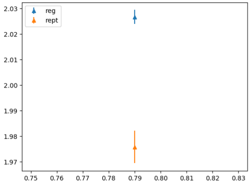


Table: Regret for (0.99, 0.79) Horizons = 50, 100 and 500; Orange = Persistence; Blue = Regular

Graphs for Thompson Sampling (4)

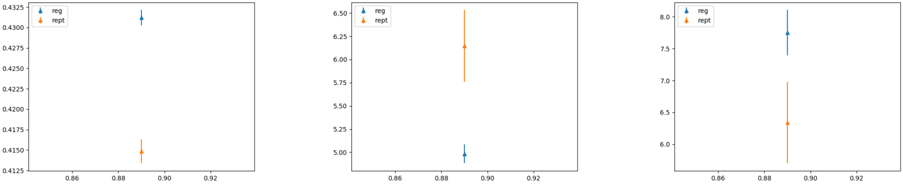


Table: Regret for (0.99, 0.89) Horizons = 10, 10^6 and 10^8; Orange = Persistence; Blue = Regular

For Thompson Sampling

- Here, we observe that for instances where both means aren't very high, the persistence version outperforms the regular version.
- On the other hand, for those instances, we observe a 'slump' in Graphs (3) and (4).

For Thompson Sampling

- Here, we observe that for instances where both means aren't very high, the persistence version outperforms the regular version.
- On the other hand, for those instances, we observe a 'slump' in Graphs (3) and (4).
- We notice that the higher the two arms' means are, the longer is the slump.

For Thompson Sampling

- Here, we observe that for instances where both means aren't very high, the persistence version outperforms the regular version.
- On the other hand, for those instances, we observe a 'slump' in Graphs (3) and (4).
- We notice that the higher the two arms' means are, the longer is the slump.
- But, for large enough time horizon, the persistence version again starts to outperform the regular version.

Theoretical Guarantees (1)

- For ϵ-greedy, we prove that we are in something called a 'bad' state at most for a constant number of time steps.

Theoretical Guarantees (1)

- For ϵ-greedy, we prove that we are in something called a 'bad' state at most for a constant number of time steps.
- We define a 'good' state as a state in which the arm with the highest empirical value is indeed the one with the highest true mean.
- If this is not the case, we are in a 'bad' state.

Theoretical Guarantees (1)

- For ϵ-greedy, we prove that we are in something called a 'bad' state at most for a constant number of time steps.
- We define a 'good' state as a state in which the arm with the highest empirical value is indeed the one with the highest true mean.
- If this is not the case, we are in a 'bad' state.
- We prove that, whenever we are in a good state, in expectation, the regret incurred is lower for the persistence variant.

Theoretical Guarantees (1)

- For ϵ-greedy, we prove that we are in something called a 'bad' state at most for a constant number of time steps.
- We define a 'good' state as a state in which the arm with the highest empirical value is indeed the one with the highest true mean.
- If this is not the case, we are in a 'bad' state.
- We prove that, whenever we are in a good state, in expectation, the regret incurred is lower for the persistence variant.
- Together, these two statements are enough to say that, in expectation, the persistence variant does better than the regular variant beyond a certain horizon.
- We'll do all our analysis for the two-armed case because, usually, analysis of the two-armed bandit problem can be generalised to the n-armed bandit.

Theoretical Guarantees (2)

- Without loss of generality, let us assume that the two arms have means μ_1 and μ_2 with $\mu_1 > \mu_2$. Let $\Delta = \mu_1 - \mu_2$
- **Fact 1** Hoeffding's Inequality: Let $X_1, ..., X_t$ be i.i.d random variable bounded by the interval [0, 1] and such that $\mu = E[X_i]$ and $M(k) = (X_1 + ... + X_k)/k$

$$P(M(k) - \mu \geq c) \leq e^{-2kc^2}$$

where $c \geq 0$

Theoretical Guarantees: For the regular version (1)

- At any given time t, in expectation, at least $\epsilon t/2$ pulls of each arm have been made.

Theoretical Guarantees: For the regular version (1)

- At any given time t, in expectation, at least $\epsilon t/2$ pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon t/2$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon t \Delta^2/4}$$

Theoretical Guarantees: For the regular version (1)

- At any given time t, in expectation, at least $\epsilon t/2$ pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon t/2$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon t \Delta^2/4}$$

$$=> P(M_1(t) < M_2(t)) \leq e^{-\epsilon t \Delta^2/4}$$

Theoretical Guarantees: For the regular version (1)

- At any given time t, in expectation, at least $\epsilon t/2$ pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon t/2$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon t \Delta^2/4}$$

$$=> P(M_1(t) < M_2(t)) \leq e^{-\epsilon t \Delta^2/4}$$

- This is the probability of being in a 'bad' state in ϵ-greedy at time t.

Theoretical Guarantees: For the regular version (2)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon t \Delta^2/4}$$

Theoretical Guarantees: For the regular version (2)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon t \Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\epsilon\Delta^2/4} - e^{-\epsilon(T+1)\Delta^2/4}}{1 - e^{\epsilon\Delta^2/4}} \leq \frac{e^{-\epsilon\Delta^2/4}}{1 - e^{\epsilon\Delta^2/4}}$$

Theoretical Guarantees: For the regular version (2)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon t \Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\epsilon \Delta^2/4} - e^{-\epsilon(T+1)\Delta^2/4}}{1 - e^{\epsilon \Delta^2/4}} \leq \frac{e^{-\epsilon \Delta^2/4}}{1 - e^{\epsilon \Delta^2/4}}$$

- For at least $T - k_2(T)$ decision times, we are in a good state.

Theoretical Guarantees: For the regular version (2)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon t \Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\epsilon\Delta^2/4} - e^{-\epsilon(T+1)\Delta^2/4}}{1 - e^{\epsilon\Delta^2/4}} \leq \frac{e^{-\epsilon\Delta^2/4}}{1 - e^{\epsilon\Delta^2/4}}$$

- For at least $T - k_2(T)$ decision times, we are in a good state.
- In a good state, the expected regret in one time step is:

$$\epsilon\frac{\Delta}{2} \qquad (1)$$

Theoretical Guarantees: For the persistence version (1)

- The analysis is going to be similar but we are going to only look at the times when compound pulls are made i.e. our time scale, instead of looking at each pull, will now look at only the time when a new decision needs to be made.

Theoretical Guarantees: For the persistence version (1)

- The analysis is going to be similar but we are going to only look at the times when compound pulls are made i.e. our time scale, instead of looking at each pull, will now look at only the time when a new decision needs to be made.
- Note that, after arm i is chosen, in expectation it will get pulled $\frac{1}{1-\mu_i}$ times before a new choice needs to be made. Here, time t represents the time when the t^{th} choice is made.

Theoretical Guarantees: For the persistence version (2)

- Now, at any given time t, in expectation, at least $\epsilon t/2$ compound pulls of each arm have been made.

Theoretical Guarantees: For the persistence version (2)

- Now, at any given time t, in expectation, at least $\epsilon t/2$ compound pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon\frac{t}{2(1-\mu_2)}$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon\frac{t}{1-\mu_2}\Delta^2/4}$$

Theoretical Guarantees: For the persistence version (2)

- Now, at any given time t, in expectation, at least $\epsilon t/2$ compound pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon \frac{t}{2(1-\mu_2)}$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

$$=> P(M_1(t) < M_2(t)) \leq e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

Theoretical Guarantees: For the persistence version (2)

- Now, at any given time t, in expectation, at least $\epsilon t/2$ compound pulls of each arm have been made.
- Using Fact 1 for the rewards of arm 2, with $c = \Delta/2$ and $k = \epsilon \frac{t}{2(1-\mu_2)}$,

$$P(M_2(t) - \mu_2 \geq \Delta/2) \leq e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

$$=> P(M_1(t) < M_2(t)) \leq e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

- This is the probability of being in a 'bad' state in persistent ϵ-greedy at time t.

Theoretical Guarantees: For the persistence version (3)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

Theoretical Guarantees: For the persistence version (3)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4} - e^{-\epsilon \frac{T+1}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}} \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}$$

Theoretical Guarantees: For the persistence version (3)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon \frac{t}{1-\mu_2}\Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4} - e^{-\epsilon \frac{T+1}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}} \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}$$

- For at least $T - k_2(T)$ decision times, we are in a good state.

Theoretical Guarantees: For the persistence version (3)

- Let the expected number of times this event happens till time T be $k_2(T)$.

$$k_2(T) \leq \Sigma_{t=1}^{T} e^{-\epsilon \frac{t}{1-\mu_2} \Delta^2/4}$$

$$k_2(T) \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4} - e^{-\epsilon \frac{T+1}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}} \leq \frac{e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}{1 - e^{-\frac{\epsilon}{1-\mu_2}\Delta^2/4}}$$

- For at least $T - k_2(T)$ decision times, we are in a good state.
- In a good state, the expected regret per time step (the actual time, not the 'compound' time) is:

$$\frac{\frac{\epsilon\Delta}{2(1-\mu_2)}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}} \qquad (2)$$

Theoretical Guarantees (3)

- Comparing (2) and (1):

Theoretical Guarantees (3)

- Comparing (2) and (1):

$$(1) - (2) = \epsilon\frac{\Delta}{2} - \frac{\frac{\epsilon\Delta}{2(1-\mu_2)}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}$$

Theoretical Guarantees (3)

- Comparing (2) and (1):

$$(1) - (2) = \epsilon\frac{\Delta}{2} - \frac{\frac{\epsilon\Delta}{2(1-\mu_2)}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}$$

$$= \frac{\epsilon\Delta}{2}[1 - \frac{\frac{1}{1-\mu_2}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}]$$

Theoretical Guarantees (3)

- Comparing (2) and (1):

$$(1) - (2) = \epsilon\frac{\Delta}{2} - \frac{\frac{\epsilon\Delta}{2(1-\mu_2)}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}$$

$$= \frac{\epsilon\Delta}{2}[1 - \frac{\frac{1}{1-\mu_2}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}]$$

$$= \frac{\epsilon\Delta}{2}[\frac{\frac{1-\epsilon/2}{1-\mu_1} - \frac{1-\epsilon/2}{1-\mu_2}}{\frac{\epsilon}{2(1-\mu_2)} + \frac{1-\epsilon/2}{1-\mu_1}}] \geq 0$$

Theoretical Guarantees (4)

- Clearly, (2) is less than (1).

Theoretical Guarantees (4)

- Clearly, (2) is less than (1).
- Therefore, except at 'bad' states which occur only finitely many times, the average regret incurred is lesser in persistent ϵ-greedy.

Theoretical Guarantees (4)

- Clearly, (2) is less than (1).
- Therefore, except at 'bad' states which occur only finitely many times, the average regret incurred is lesser in persistent ϵ-greedy.
- Therefore, beyond a certain horizon, persistent ϵ-greedy is going to be better than regular ϵ-greedy.

Discussion: Thompson Sampling (1)

- For Thompson Sampling the picture is a little more complicated.

Discussion: Thompson Sampling (1)

- For Thompson Sampling the picture is a little more complicated.
- For most problem instances, the persistence variant outperforms the regular version consistently. (Graphs (1) and (2)).

Discussion: Thompson Sampling (1)

- For Thompson Sampling the picture is a little more complicated.
- For most problem instances, the persistence variant outperforms the regular version consistently. (Graphs (1) and (2)).
- There is a small set of problematic instances, though, where both μ_1 and μ_2 are high.
- For such instances, the persistence variant has a 'slump' between some t_1 and t_2 compared to the regular version (Graphs (3) and (4)).

Discussion: Thompson Sampling (1)

- For Thompson Sampling the picture is a little more complicated.
- For most problem instances, the persistence variant outperforms the regular version consistently. (Graphs (1) and (2)).
- There is a small set of problematic instances, though, where both μ_1 and μ_2 are high.
- For such instances, the persistence variant has a 'slump' between some t_1 and t_2 compared to the regular version (Graphs (3) and (4)).
- Our observations suggest that t_1 keeps decreasing and t_2 keeps increasing as μ_1 and μ_2 become even higher.
- We hypothesize that, for any bandit instance (μ_1, μ_2), there exists, a time that is a function of μ_1 and μ_2, beyond which the persistence outperforms the regular version even for the harder instances.

Discussion: Thompson Sampling (2)

- Why does this 'slump' occur?

Discussion: Thompson Sampling (2)

- Why does this 'slump' occur?
- The algorithm has $1/2$ probability of picking the μ_2 arm in the 1^{st} time step.

Discussion: Thompson Sampling (2)

- Why does this 'slump' occur?
- The algorithm has $1/2$ probability of picking the μ_2 arm in the 1^{st} time step.
- If μ_2 is high and the persistence version has been deployed, this arm will end up being pulled a lot in the beginning ($\frac{1}{1-\mu_2}$ times in expectation) updating its beta parameters to indicate a high mean for arm 2 with high certainty.

Discussion: Thompson Sampling (2)

- Why does this 'slump' occur?
- The algorithm has $1/2$ probability of picking the μ_2 arm in the 1^{st} time step.
- If μ_2 is high and the persistence version has been deployed, this arm will end up being pulled a lot in the beginning ($\frac{1}{1-\mu_2}$ times in expectation) updating its beta parameters to indicate a high mean for arm 2 with high certainty.
- This lowers the probability of μ_1 arm being pulled and ensures that the higher mean of the μ_1 arm is discovered at a much later time.

Discussion: Thompson Sampling (2)

- Why does this 'slump' occur?
- The algorithm has $1/2$ probability of picking the μ_2 arm in the 1^{st} time step.
- If μ_2 is high and the persistence version has been deployed, this arm will end up being pulled a lot in the beginning ($\frac{1}{1-\mu_2}$ times in expectation) updating its beta parameters to indicate a high mean for arm 2 with high certainty.
- This lowers the probability of μ_1 arm being pulled and ensures that the higher mean of the μ_1 arm is discovered at a much later time.
- This means that with $1/2$ probability the persistence algorithm can reach a local minima and remain stuck there for a while.

Discussion: Thompson Sampling (3)

- Things aren't that bad though.

Discussion: Thompson Sampling (3)

- Things aren't that bad though.
- This only happens when the suboptimal arm has a high mean too.

Discussion: Thompson Sampling (3)

- Things aren't that bad though.
- This only happens when the suboptimal arm has a high mean too.
- Therefore, the regret incurred, is only slightly more than that of the regular version.

Future Work

- There are many potential directions for future research.

Future Work

- There are many potential directions for future research.
- First, theoretical guarantees need to be proved for Thompson Sampling.
- We wish to prove in the future that, in expectation, Persistent Thompson Sampling has lesser regret than Regular Thompson Sampling for problem instance dependent time horizons.

Future Work

- There are many potential directions for future research.
- First, theoretical guarantees need to be proved for Thompson Sampling.
- We wish to prove in the future that, in expectation, Persistent Thompson Sampling has lesser regret than Regular Thompson Sampling for problem instance dependent time horizons.
- The analysis needs to be extended to the n armed bandit.

Future Work

- There are many potential directions for future research.
- First, theoretical guarantees need to be proved for Thompson Sampling.
- We wish to prove in the future that, in expectation, Persistent Thompson Sampling has lesser regret than Regular Thompson Sampling for problem instance dependent time horizons.
- The analysis needs to be extended to the n armed bandit.
- A family of persistence algorithms can be looked at, where the maximum persistence, i.e. the maximum time one can go without making a new decision, is a parameter. This parameter can be constant or a variable.

Future Work

- There are many potential directions for future research.
- First, theoretical guarantees need to be proved for Thompson Sampling.
- We wish to prove in the future that, in expectation, Persistent Thompson Sampling has lesser regret than Regular Thompson Sampling for problem instance dependent time horizons.
- The analysis needs to be extended to the n armed bandit.
- A family of persistence algorithms can be looked at, where the maximum persistence, i.e. the maximum time one can go without making a new decision, is a parameter. This parameter can be constant or a variable.
- Making persistence robust to all problem instances.

Miscellaneous

Other things we worked on:

- Tighter bounds on Thompson Sampling
- The Batch Bandit problem
- Discrete Support Thompson Sampling
- Binary Bandits
- Open Loop Algorithm