Bandits and Structured Bandits

Abhishek Sinha
Laboratory for Information and Decision Systems

MIT

Talk at: CNRG Meeting

April 25, 2016

Outline

Outline

General Multi-armed Bandit Problem - Setup

- There are K arms with the i^{th} arm having unknown reward expectations $\mu_i, i = 1, 2, \ldots, K$. The distributions are i.i.d. w.r.t. time with support in $[0, 1]$.

General Multi-armed Bandit Problem - Setup

- There are K arms with the i^{th} arm having unknown reward expectations $\mu_i, i = 1, 2, \ldots, K$. The distributions are i.i.d. w.r.t. time with support in $[0, 1]$.
- At time-step t we select one of the arms $I_t \in \{1, 2, \ldots, K\}$ to play, yielding a random reward X_{I_t}. Action (or the policy) I_t may depend on past actions and their outcomes. Hence over a time-horizon of n slots, we gather an expected reward of

$$\mathbb{E} \sum_{t=1}^{n} X_{I_t} = \sum_{t=1}^{n} \mu_{I_t}, \quad \text{(linearity of expectation)}$$

General Multi-armed Bandit Problem - Setup

- There are K arms with the i^{th} arm having unknown reward expectations $\mu_i, i = 1, 2, \ldots, K$. The distributions are i.i.d. w.r.t. time with support in $[0, 1]$.
- At time-step t we select one of the arms $I_t \in \{1, 2, \ldots, K\}$ to play, yielding a random reward X_{I_t}. Action (or the policy) I_t may depend on past actions and their outcomes. Hence over a time-horizon of n slots, we gather an expected reward of

$$\mathbb{E} \sum_{t=1}^{n} X_{I_t} = \sum_{t=1}^{n} \mu_{I_t}, \quad \text{(linearity of expectation)}$$

- If we had known the best arm $i^* \in \arg\max \mu_i$ apriori and played that arm throughout, we would have obtained an expected reward of $n\mu^*$.

General Multi-armed Bandit Problem - Setup

- There are K arms with the i^{th} arm having unknown reward expectations $\mu_i, i = 1, 2, \ldots, K$. The distributions are i.i.d. w.r.t. time with support in $[0, 1]$.
- At time-step t we select one of the arms $I_t \in \{1, 2, \ldots, K\}$ to play, yielding a random reward X_{I_t}. Action (or the policy) I_t may depend on past actions and their outcomes. Hence over a time-horizon of n slots, we gather an expected reward of

$$\mathbb{E} \sum_{t=1}^{n} X_{I_t} = \sum_{t=1}^{n} \mu_{I_t}, \quad \text{(linearity of expectation)}$$

- If we had known the best arm $i^* \in \arg\max \mu_i$ apriori and played that arm throughout, we would have obtained an expected reward of $n\mu^*$.
- The expected regret (or, pseudo-regret) up to time n is defined as their difference, which we want to minimize over admissible policies.

$$\mathbb{E}(\text{regret}(n)) = n\mu^* - \sum_{t=1}^{n} \mu_{I_t}$$

Lower bounds and Achievability

In the special case when the reward distributions are Bernoulli (μ_i), we have

Lower bounds (Lai and Robins (1985))

$$\liminf_n \frac{\mathbb{E}(\text{regret}(n))}{\ln(n)} \geq \sum_{i:\mu^* - \mu_i > 0} \frac{\mu^* - \mu_i}{D(\mu_i, \mu^*)} \stackrel{(\text{def})}{=} C_l$$

In other words, for large enough n, for any admissible policy

$$\mathbb{E}(\text{regret}(n)) \geq C_l \ln(n) + o(\ln(n))$$

Lower bounds and Achievability

In the special case when the reward distributions are Bernoulli (μ_i), we have

Lower bounds (Lai and Robins (1985))

$$\liminf_n \frac{\mathbb{E}(\text{regret}(n))}{\ln(n)} \geq \sum_{i:\mu^*-\mu_i>0} \frac{\mu^*-\mu_i}{D(\mu_i,\mu^*)} \stackrel{\text{(def)}}{=} C_l$$

In other words, for large enough n, for any admissible policy

$$\mathbb{E}(\text{regret}(n)) \geq C_l \ln(n) + o(\ln(n))$$

Fortunately, there exists a simple policy UCB (Upper Confidence Bound, described next) which achieves non-asymptotic logarithmic regret bound

Achievability

$$\mathbb{E}(\text{regret}(n)) \leq C_u \ln(n) + 3K$$

where $C_u = 6\sum_{i:\mu^*-\mu_i} \frac{1}{\mu^*-\mu_i}$.

The UCB Policy: Explore and Exploit

The principle is simple and intuitive : at time $t+1$ play the arm which maximizes sum of an exploit and explore index.

The UCB Policy: Explore and Exploit

The principle is simple and intuitive : at time $t+1$ play the arm which maximizes sum of an exploit and explore index.
Suppose the arm i has been played $T_i(t)$ times up to time t, yielding an average reward of $\hat{\mu}_i(T_i(t))$. Then at time $t+1$ the policy UCB plays the arm which maximizes the following index:

UCB policy

$$I_{t+1} = \arg\max_{i=1}^{K} \left(\hat{\mu}_i(T_i(t)) + \sqrt{\frac{3\ln(t)}{2T_i(t)}} \right)$$

The UCB Policy: Explore and Exploit

The principle is simple and intuitive : at time $t+1$ play the arm which maximizes sum of an exploit and explore index.
Suppose the arm i has been played $T_i(t)$ times up to time t, yielding an average reward of $\hat{\mu}_i(T_i(t))$. Then at time $t+1$ the policy UCB plays the arm which maximizes the following index:

UCB policy

$$I_{t+1} = \arg\max_{i=1}^{K} \left(\hat{\mu}_i(T_i(t)) + \sqrt{\frac{3\ln(t)}{2T_i(t)}} \right)$$

Large observed average rewards $\hat{\mu}_i(T_i(t))$ encourages to play that arm: exploit factor!

The UCB Policy: Explore and Exploit

The principle is simple and intuitive : at time $t+1$ play the arm which maximizes sum of an exploit and explore index.
Suppose the arm i has been played $T_i(t)$ times up to time t, yielding an average reward of $\hat{\mu}_i(T_i(t))$. Then at time $t+1$ the policy UCB plays the arm which maximizes the following index:

UCB policy

$$I_{t+1} = \arg\max_{i=1}^{K} \left(\hat{\mu}_i(T_i(t)) + \sqrt{\frac{3\ln(t)}{2T_i(t)}} \right)$$

Large observed average rewards $\hat{\mu}_i(T_i(t))$ encourages to play that arm: exploit factor!

Small number of past plays $T_i(t)$ also encourages to play that arm: explore factor!

Outline

Strutured Bandits

- In the general bandit problem, we did not assume any underlying structure among the distributions of different arms. The resulting constants C_l and C_u are $O(K)$.
- In many interesting combinatorial problems, number of arms K can be very large (e.g., exponential) and hence the general bandit results are not so useful.

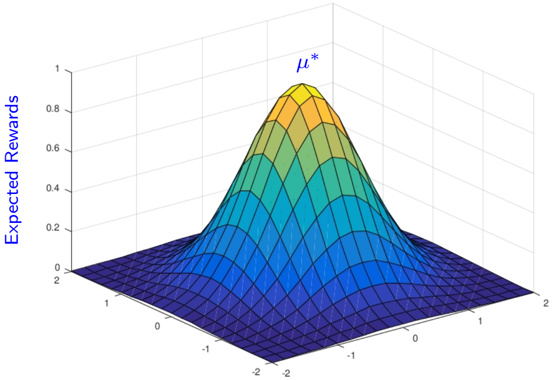
Strutured Bandits

- In the general bandit problem, we did not assume any underlying structure among the distributions of different arms. The resulting constants C_l and C_u are $O(K)$.
- In many interesting combinatorial problems, number of arms K can be very large (e.g., exponential) and hence the general bandit results are not so useful.
- However, many of the interesting combinatorial problem imposes some natural structures among the unknown distributions of the arms. Exploiting these structures significantly improves the constants C_l, C_u.
- Combes and Proutiere (2014) analyzes one such structured bandit problem, called Graphical Unimodal bandits.

Strutured Bandits

- In the general bandit problem, we did not assume any underlying structure among the distributions of different arms. The resulting constants C_l and C_u are $O(K)$.
- In many interesting combinatorial problems, number of arms K can be very large (e.g., exponential) and hence the general bandit results are not so useful.
- However, many of the interesting combinatorial problem imposes some natural structures among the unknown distributions of the arms. Exploiting these structures significantly improves the constants C_l, C_u.
- Combes and Proutiere (2014) analyzes one such structured bandit problem, called Graphical Unimodal bandits.
- Informally, in graphical unimodal bandits, from every arm i, there exist a path to the optimum arm i^* through a sequence of neighbouring arms of non-decreasing expected rewards.

Graphical Unimodal Bandits: In picture



Each intersection corresponds to an arm

Formal Definitions and Lower Bounds

- Consider an undirected graph $\mathcal{G}(V, E)$ whose vertices correspond to arms and incident edges to a vertex $i \in \{1, 2, \ldots, K\}$ denote its neighborhood.
- There is a unique $i^* = \arg\max \mu_i$ and from any arm $i = k_1$ there exists a path $p = (k_1, k_2, \ldots, , k_m = i^*)$ such that $\mu_{k_i} < \mu_{k_{i+1}}, i = 1, 2, \ldots, m - 1$.

Formal Definitions and Lower Bounds

- Consider an undirected graph $\mathcal{G}(V, E)$ whose vertices correspond to arms and incident edges to a vertex $i \in \{1, 2, \ldots, K\}$ denote its neighborhood.
- There is a unique $i^* = \arg\max \mu_i$ and from any arm $i = k_1$ there exists a path $p = (k_1, k_2, \ldots, , k_m = i^*)$ such that $\mu_{k_i} < \mu_{k_{i+1}}, i = 1, 2, \ldots, m-1$.

Note that lower-bounds in General bandit does not necessarily imply a corresponding lower-bound for the structured bandit.

Formal Definitions and Lower Bounds

- Consider an undirected graph $\mathcal{G}(V, E)$ whose vertices correspond to arms and incident edges to a vertex $i \in \{1, 2, \ldots, K\}$ denote its neighborhood.
- There is a unique $i^* = \arg\max \mu_i$ and from any arm $i = k_1$ there exists a path $p = (k_1, k_2, \ldots, , k_m = i^*)$ such that $\mu_{k_i} < \mu_{k_{i+1}}, i = 1, 2, \ldots, m-1$.

Note that lower-bounds in General bandit does not necessarily imply a corresponding lower-bound for the structured bandit.

Lower-bound for Unimodal Bandits (Combes and Proutiere (2014))

$$\lim_{n \to \infty} \inf \frac{\mathbb{E}(\text{regret}(n))}{\ln(n)} \geq \sum_{i \in \text{Nbr}(i^*)} \frac{\mu^* - \mu_i}{D(\mu_i, \mu^*)} \stackrel{\text{def}}{=} C_I^{\text{unimodal}}$$

Formal Definitions and Lower Bounds

- Consider an undirected graph $\mathcal{G}(V, E)$ whose vertices correspond to arms and incident edges to a vertex $i \in \{1, 2, \ldots, K\}$ denote its neighborhood.
- There is a unique $i^* = \arg\max \mu_i$ and from any arm $i = k_1$ there exists a path $p = (k_1, k_2, \ldots, , k_m = i^*)$ such that $\mu_{k_i} < \mu_{k_{i+1}}, i = 1, 2, \ldots, m-1$.

Note that lower-bounds in General bandit does not necessarily imply a corresponding lower-bound for the structured bandit.

Lower-bound for Unimodal Bandits (Combes and Proutiere (2014))

$$\lim\inf_{n\to\infty} \frac{\mathbb{E}(\text{regret}(n))}{\ln(n)} \geq \sum_{i \in \text{Nbr}(i^*)} \frac{\mu^* - \mu_i}{D(\mu_i, \mu^*)} \stackrel{\text{def}}{=} C_l^{\text{unimodal}}$$

Comparing this with the lower-bound for general bandit, we immediately see that in general, $C_l^{\text{unimodal}} << C_l$.
C_l^{unimodal} is independent of number of arms (K) and is a function of only local neighbourhood of the optimal arm.

Achievability

The basic strategy for achieving low regret bound is intuitive: explore local neighbourhoods and exploit the currently perceived best neighbouring arm (**Hill Climbing**).

More formally, at time t an index $b_k(t)$ (similar to the general bandit) is computed for all arms in the neighbourhood of the current arm being played. The algorithm simply chooses the neighbouring arm which maximizes this index.

Achievability

The basic strategy for achieving low regret bound is intuitive: explore local neighbourhoods and exploit the currently perceived best neighbouring arm (**Hill Climbing**).

More formally, at time t an index $b_k(t)$ (similar to the general bandit) is computed for all arms in the neighbourhood of the current arm being played. The algorithm simply chooses the neighbouring arm which maximizes this index.

Achievability

The local-search algorithm above is asymptotically optimal for Bermoulli rewards, i.e.

$$\limsup_{n\to\infty} \frac{\mathbb{E}(\text{regret}(n))}{\ln(n)} \leq C_l^{\text{unimodal}}$$

Achievability

The basic strategy for achieving low regret bound is intuitive: explore local neighbourhoods and exploit the currently perceived best neighbouring arm (**Hill Climbing**).

More formally, at time t an index $b_k(t)$ (similar to the general bandit) is computed for all arms in the neighbourhood of the current arm being played. The algorithm simply chooses the neighbouring arm which maximizes this index.

Achievability

The local-search algorithm above is asymptotically optimal for Bermoulli rewards, i.e.

$$\limsup_{n \to \infty} \frac{\mathbb{E}(\mathrm{regret}(n))}{\ln(n)} \leq C_l^{\mathrm{unimodal}}$$

Note that unlike the general bandit case, here the upper and lower-bound constants coincide.

Outline

Application

Combes and Proutiere applied the result of structured bandits to rate adaptation problem in 802.11 systems.

Here the pair (rate, mode) consists of an action (or arm of a bandit) which exhibits graphical unimodal property in a stochastic radio environment .

Using a local neighborhood search method (called G-ORS) they designed an asymptotically optimal rate adaptation policy.

They also extended this result to non-stationary radio environments.

In what directions can these results be extended further ?