Lipschitz Bandit

Heuna Kim

Contents

- Recap: Bandit and multi-armed Bandit
- Lipschitz Bandit
  - Definition
  - Fixed Discretization and lower bound
- Algorithms
  - Zooming algorithm by Kleinberg
  - Hierarchical Optimistic Optimization by Bubeck
- Applications

Recap: Multi-armed Bandit Example

- Gambling Machine
- Medication Prescription
- Mouse pushing buttons to get cheese

Recap: k-armed Bandit

- For each of the k actions, an expected or mean reward (value)

$$q_*(a) \doteq \mathbb{E}[R_t \mid A_t = a] \,.$$

- Qt (a) = the estimated value of action a at time step t

  = the average rewards so far (Monte-Carlo estimates)

- Exploration: Epsilon-greedy strategy
  - With probability 1- epsilon
  - With probability epsilon

$$A_t \doteq \underset{a}{\arg\max}\, Q_t(a)$$

Recap: Regret

- The optimal value

$$v_* = \max_{a \in \mathcal{A}} q(a) = \max_a \mathbb{E}[R_t \mid A_t = a]$$

- For each step, regret is $v_* - q(A_t)$
- Total regret is the sum of regrets over time

Recap: Upper Confidence Bound

$$A_t \doteq \underset{a}{\arg\max} \left[Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

Theorem (Lai and Robbins)

Asymptotic total regret is at least logarithmic in number of steps

$$\lim_{t\to\infty} L_t \geq \log t \sum_{a|\Delta_a>0} \frac{\Delta_a}{KL(\mathcal{R}^a||\mathcal{R}^{a_*})}$$

Theorem (Auer et al., 2002)

The UCB algorithm achieves logarithmic expected total regret

$$L_t \leq 8 \sum_{a|\Delta_a>0} \frac{\log t}{\Delta_a} + O(\sum_a \Delta_a)$$

for any t

Theorems captured from the lecture note by Hasselt (References: the last page)

Infinitely-many-armed Bandit

- Intractable

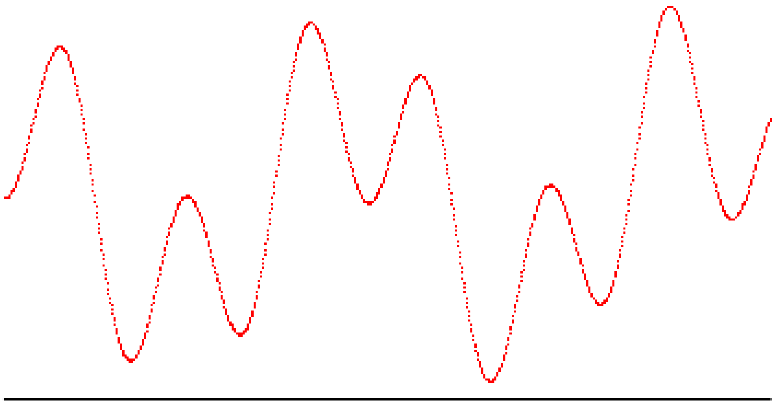
Contents

- Recap: Bandit and multi-armed Bandit
- Lipschitz Bandit
  - Definition
  - Fixed Discretization and lower bound
- Algorithms
  - Zooming algorithm by Kleinberg
  - Hierarchical Optimistic Optimization by Bubeck
- Applications

Lipschitz Bandit: definition

- Each arm x is an IID sample from some fixed distribution with expectation μ(x) with x in X = [0,1]

$$|\mu(x) - \mu(y)| \leq L \cdot |x - y| \quad \text{for any two arms } x, y \in X$$



The picture captured from the paper by Bubeck

Fixed Discretization

- Let S be the discretized set and W(ALG) is the return of the algorithm: the regret of ALG is

$$\begin{aligned} R(T) &= \mu^*(X) - W(ALG) \\ &= (\mu^*(S) - W(ALG)) + (\mu^*(X) - \mu^*(S)) \\ &= R_S(T) + \mathtt{DE}(S), \end{aligned}$$

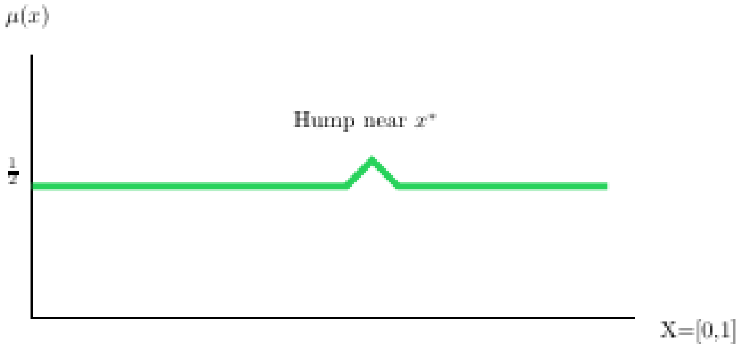
$$\mathbb{E}[R(T)] \le O(\sqrt{|S| T \log T}) + \mathtt{DE}(S) \cdot T.$$

$\mathtt{DE}(S) \le L\epsilon$. Picking $\epsilon = (TL^2 / \log T)^{-1/3}$ $\quad \mathbb{E}[R(T)] \le O\left(L^{1/3} \cdot T^{2/3} \cdot \log^{1/3}(T)\right).$

ALG for CAB instance $\mathcal{I}$	ALG' for K-armed bandits instance $\mathcal{J}$
chooses arm $x \in \left[\frac{a}{K} - \epsilon, \frac{a}{K} + \epsilon\right), a \in [K]$	chooses arm a
receives reward $r_x \in \{0,1\}$ with mean $\mu(x)$	receives reward $r \in \{0,1\}$ with mean $\mu_{\mathcal{J}}(a)$

Lower bound

$$\mu(x) = \begin{cases} \frac{1}{2}, & |x - x^*| \geq \epsilon/L \\ \frac{1}{2} + \epsilon - L \cdot |x - x^*|, & \text{otherwise.} \end{cases}$$



$$\mathbb{E}[R(T) \mid \mathcal{I}] \geq \mathbb{E}[R'(T) \mid \mathcal{J}]$$

Picture captured from the lecture note by Silvkins

$$\mathbb{E}[R(T) \mid \mathcal{J}] \geq \Omega(\epsilon T).$$

$$K = (T/4c)^{1/3}$$

$$\epsilon \leq \sqrt{cK/T}$$

$$\mathbb{E}[R'(T) \mid \mathcal{J}] \geq \Omega(\sqrt{\epsilon T}) = \Omega(T^{2/3})$$

Contents

Zooming Algorithm (UCB inspired)

Activation rule. The activation rule is very simple:

If some arm y becomes uncovered by confidence balls of the active arms, activate y.

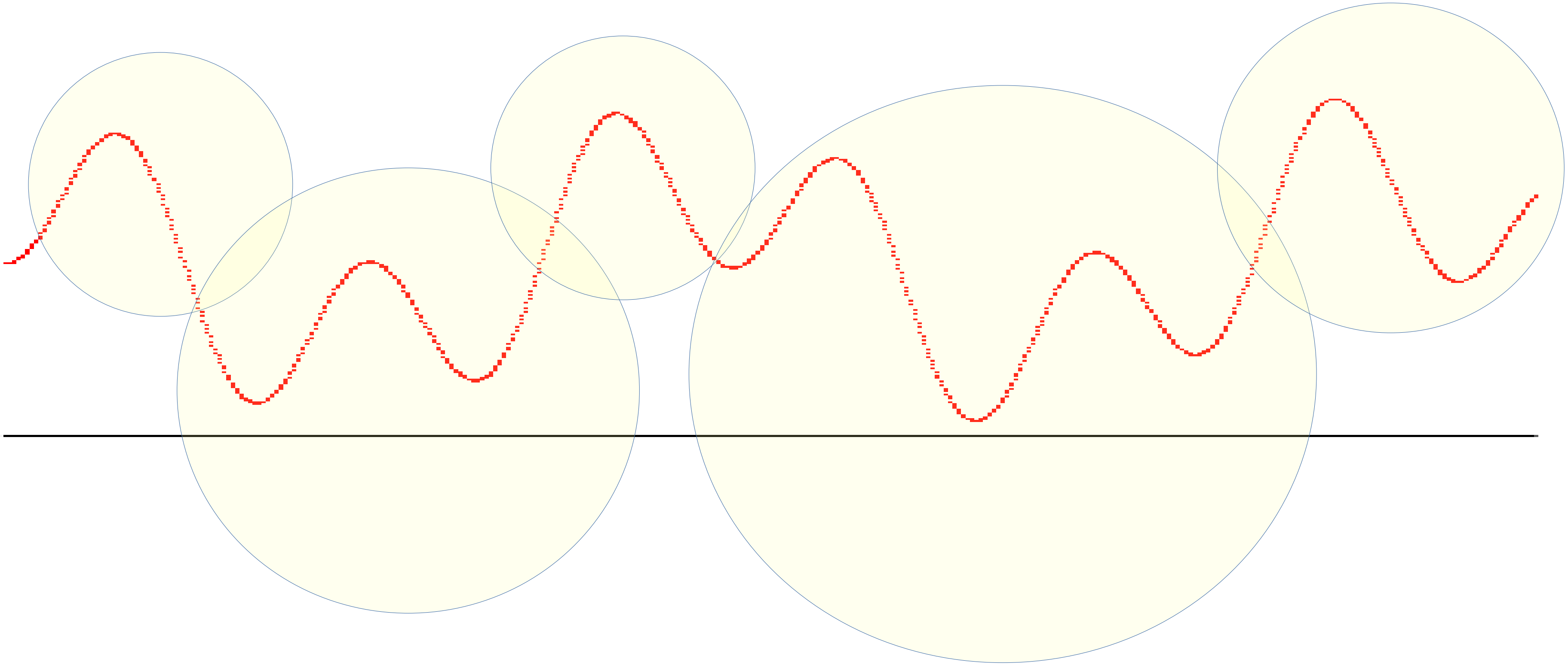
$$\mathbf{B}_t(x) = \{y \in X : \mathcal{D}(x, y) \leq r_t(x)\}.$$

$$r_t(x) = \sqrt{\frac{2 \log T}{n_t(x) + 1}}.$$

- Maintain a set of active arms S.
- Initially $S = \emptyset$, activate arms one by one.
- In each round t,
  - Activate uncovered arms according to *Activation Rule.*
  - Play the active arm with the largest index $\texttt{index}_t(x)$.

$$\texttt{index}_t(x) = \bar{\mu}_t(x) + 2r_t(x)$$

Zooming algorithm - Example



Zooming Algorithm - Bound

Theorem 3.5. *Consider Lipschitz MAB problem with time horizon T. For any given problem instance and any $c > 0$, the zooming algorithm attains regret*

$$\mathbb{E}[R(T)] \leq O\left(T^{\frac{d+1}{d+2}} (c \log T)^{\frac{1}{d+2}}\right),$$

where d is the zooming dimension with multiplier c.

d is a constant depending on Δ

1. Most of them are covered by confidence balls

$$\mathcal{E}_x = \{|\mu_t(x) - \mu(x)| \leq r_t(x), \quad \forall t\}$$

By Hoeffding Inequality

$$\Pr[\mathcal{E}] \geq 1 - \frac{1}{T^2}.$$

2. The arms with low rewards cannot be played often.

$$\Delta(x) = \mu^* - \mu(x)$$

$\Delta(x) \leq 3\, r_t(x)$ *for each arm x and each round t.*

Remind that: $\texttt{index}_t(x) = \bar{\mu}_t(x) + 2r_t(x)$

Contents

- Recap: Bandit and multi-armed Bandit
- Lipschitz Bandit
  - Definition
  - Fixed Discretization and lower bound
- Algorithms
  - Zooming algorithm by Kleinberg
  - Hierarchical Optimistic Optimization by Bubeck
- Applications

Hierarchical Optimistic Optimization

- Build a tree corresponding to the regions $(\mathcal{P}_{h,i})_{h \geqslant 0, 1 \leqslant i \leqslant 2^h}$
  - Select the child with the highest B-value
  - Extend tree
    - Add a node with a path and U-value
    - Update B-value

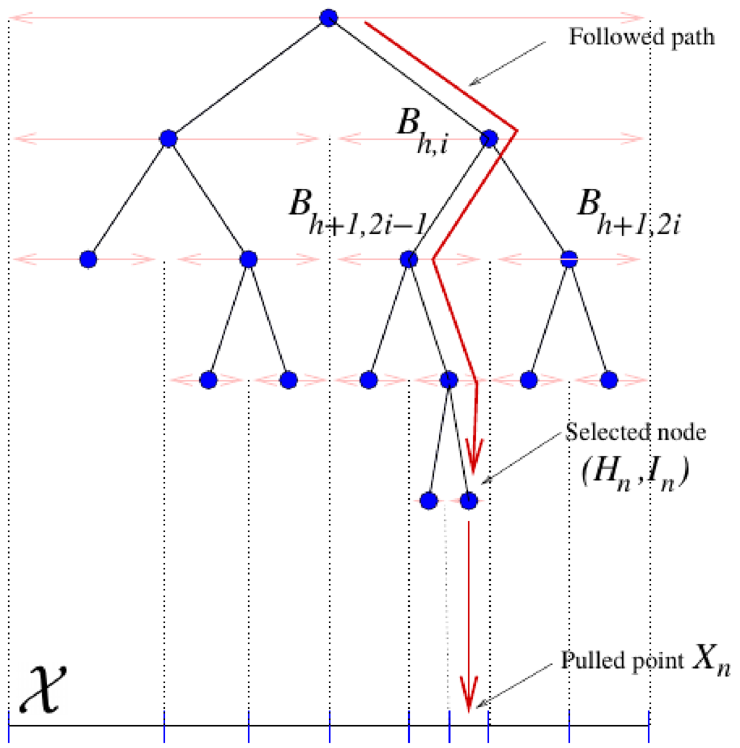
$$B_{h,i}(n) = \begin{cases} \min\Big\{U_{h,i}(n), \max\big\{B_{h+1,2i-1}(n), B_{h+1,2i}(n)\big\}\Big\}, & \text{if } (h,i) \in \mathcal{T}_n; \\ + \quad , & \text{otherwise.} \end{cases}$$

$$U_{h,i}(n) = \begin{cases} \widehat{\ }_{h,i}(n) + \sqrt{\dfrac{2\ln n}{T_{h,i}(n)}} + \nu_1 \rho^h, & \text{if } T_{h,i}(n) > 0; \\ + \quad , & \text{otherwise.} \end{cases}$$

$\widehat{\ }_{h,i}(n)$ = MC reward estimates

$\nu_1 > 0$ and $\rho \in (0,1)$ = regularizing tree

HOO - Example



Followed path
$B_{h,i}$
$B_{h+1,2i-1}$
$B_{h+1,2i}$
Selected node
(H_n, I_n)
Pulled point X_n
$\mathcal{X}$

HOO - Bound

Theorem 6 (Regret bound for HOO) *Consider HOO tuned with parameters such that Assumptions A1 and A2 hold for some dissimilarity ℓ. Let d be the $4\nu_1/\nu_2$–near-optimality dimension of the mean-payoff function f w.r.t. ℓ. Then, for all $d' > d$, there exists a constant γ such that for all* $n \geqslant 1$,

$$\mathbb{E}\big[R_n\big] \leqslant \gamma n^{(d'+1)/(d'+2)} \left(\ln n\right)^{1/(d'+2)}.$$

The mean-payoff function = the value function

d here is also a constant depending on how much the reward function is changing over the arms (related to the covering)

In short, it is exactly a comparable result to the zooming algorithm but with tree structure

Contents

- Recap: Bandit and multi-armed Bandit
- Lipschitz Bandit
  - Definition
  - Fixed Discretization and lower bound
- Algorithms
  - Zooming algorithm by Kleinberg
  - Hierarchical Optimistic Optimization by Bubeck
- Applications

Single Kelly Bet

- A simple bet with two outcomes:
  - Losing the entire amount bet
  - Winning the bet * the payoff odds “b”
  - With the probability of winning “p”

  - The expected reward for the “a” bet
    - = - a * (1-p) + b * p
    - If p is a continuous function, finding the best “a” is a lipschitz bandit
    - If p is a constant, finding the best “a” is linear optimization

Lipschitz Contextual MAB

Definition 2. *A* Lipschitz contextual multi-armed bandit problem *(Lipschitz contextual MAB) is a pair of metric spaces—a metric space of* queries (X, L_X) *of and a metric space of* ads (Y, L_Y). *An* instance *of the problem is a* payoff *function* $\mu : X \times Y \to [0,1]$ *which is Lipschitz in each coordinate, that is,* $\forall x, x' \in X,\ \forall y, y' \in Y$,

$$|\mu(x,y) - \mu(x',y')| \leq L_X(x,x') + L_Y(y,y'). \qquad (1)$$

$$\forall x, x' \in X,\ \forall y \in Y, \quad |\mu(x,y) - \mu(x',y)| \leq L_X(x,x'),$$
$$\forall x \in X,\ \forall y, y' \in Y, \quad |\mu(x,y) - \mu(x,y')| \leq L_Y(y,y').$$

Episodic Kelly Bet and LC CAB

- The reward function in a single event is in (X, L_X)
- The current budget in each time step is in (Y, L_Y)

- The algorithm by Lu el al: a similar algorithm to the zooming algorithm with a larger constant in front of the radius for choosing the next index

- We may want to add some trend stability measure to the reward function to follow a variant of UCB policy.

References

- MAB
  - Reinforcement Learning: An Introduction, Chapter 2, Sutton and Barto
  - Lecture Note 2: Exploration and Exploitation, Hasselt
- CAB
  - CMSC 858G Lecture Note 6, Slivkins, 2016
  - Bandits and experts in metric spaces, Kleinberg at al. 2013 - 2018
  - X -Armed Bandits, Bubeck et al. 2011
- Contextual CAB
  - Contextual Multi-Armed Bandits, Lu at al. 2010