Combinatorial Pure Exploration with Continuous and Separable Reward Functions and Its Applications

Weiran Huang
Tsinghua University

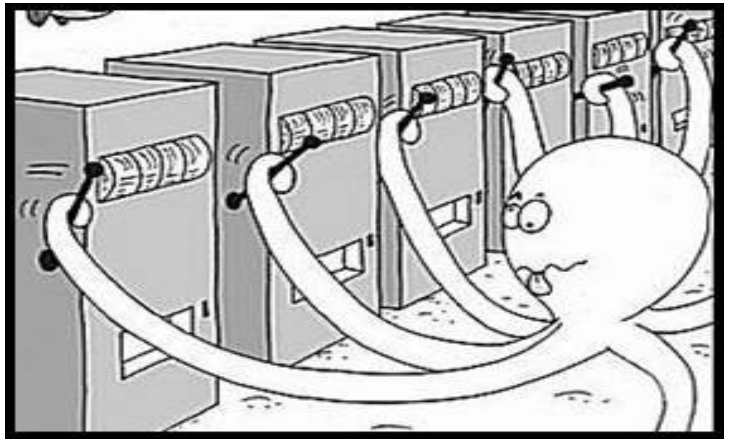
Jungseul Ok
KTH

Liang Li
Ant Financial Group

Wei Chen
Microsoft Research

Multi-Armed Bandit (Explore-Exploit Tradeoff)

- A gambler faces m slot-machines ("armed bandits").
- Each machine will provide a random reward from an unknown distribution specific to that machine, when its arm is pulled.
- In which order should the gambler pull arms based on the feedback collected, to maximize the sum of rewards?



[Lai and Robbins, 1985; Auer et al., 2002b; Auer et al., 2002a; Bubeck and CesaBianchi, 2012]

Pure Exploration Bandit

Another interesting problem is how to adaptively select arms to pull to identify the optimal arm with high probability using as few samples as possible -- best arm identification. [Bubeck et al., 2010; Audibert et al., 2010; Gabillon et al., 2012]

Extensions:

- Top-k best arm identification [Kalyanakrishnan and Stone, 2010; Kalyanakrishnan et al., 2012; Bubeck et al., 2013; Kaufmann and Kalyanakrishnan, 2013; Zhou et al., 2014]
- Multi-bandit best arm identification [Gabillon et al., 2011]
- Combinatorial Pure Exploration with Linear reward functions [Chen et al., 2014; Chen et al., 2016a]

Combinatorial Pure Exploration with Continuous and Separable Rewards (CPE-CS)

- There are m arms and a finite set of decisions $\mathcal{Y} \subseteq \mathbb{R}^m$. Each arm is associated with an unknown distribution $D_i \in [0,1]$ and an unknown parameter θ_i^* of D_i.
- In each round, the player chooses an arm to pull and gets a random outcome.
- Reward function: $r(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_i f_i(\theta_i, y_i)$, continuous in $\boldsymbol{\theta}$ for each $\boldsymbol{y} \in \mathcal{Y}$.
- Task: design an algorithm with the following components:
  - A stopping condition: decide whether the algorithm should stop in the current round.
  - An arm selection component: select the arm to play in the current round when the stopping condition is false.
  - An output component: output a decision when the stopping condition is true.
- Goal: identify the unique optimal decision $\boldsymbol{y}^o = \arg\max_{\boldsymbol{y} \in \mathcal{Y}} r(\boldsymbol{\theta}^*; \boldsymbol{y})$.

Special Cases of CPE-CS

Best arm identification

- $r(\boldsymbol{\theta}^*; \boldsymbol{y}) = \sum_{i=1}^{m} \theta_i^* \cdot y_i$, where θ_i^* is the mean of each arm i
- $\mathcal{Y} = \{\underbrace{(1,0,\dots,0)}_{m \text{ bits}}, \dots, \underbrace{(0,\dots,0,1)}_{m \text{ bits}}\}.$

Special Cases of CPE-CS

Top-k best arms identification

- $r(\boldsymbol{\theta}^*; \boldsymbol{y}) = \sum_{i=1}^{m} \theta_i^* \cdot y_i$, where θ_i^* is the mean of each arm i
- $\mathcal{Y} = \{\boldsymbol{y} \in \{0,1\}^m \mid k$ bits of ones$\}$

Special Cases of CPE-CS

Multi-bandit best arm identification

- $r(\boldsymbol{\theta}^*; \boldsymbol{y}) = \sum_{i=1}^{mn} \theta_i^* \cdot y_i$, where θ_i^* is the mean of each arm i
- $\mathcal{Y} = \bigotimes_{j=1}^{n} \{\underbrace{(1,0, \dots, 0)}_{m \text{ bits}}, \dots, \underbrace{(0, \dots, 0,1)}_{m \text{ bits}}\}$

n bandits
each has m arms

Special Cases of CPE-CS

Combinatorial Pure Exploration with Linear rewards (CPE-L)

- $r(\boldsymbol{\theta}^*; \boldsymbol{y}) = \sum_{i=1}^{m} \theta_i^* \cdot y_i$, where θ_i^* is the mean of each arm i
- $\mathcal{Y} = \{\mathbf{1}_S |\ S \text{ is an super arm}\}$

Assumptions

- Suppose we have a deterministic oracle ϕ for the offline problem ($\boldsymbol{\theta}^*$ is known):

$$\phi(\boldsymbol{\theta}) = \left(\phi_1(\boldsymbol{\theta}), \dots, \phi_m(\boldsymbol{\theta})\right) \in \arg\max_{\boldsymbol{y} \in \mathcal{Y}} r(\boldsymbol{\theta}; \boldsymbol{y})$$

- Suppose we have an estimator for the statistic θ_i^*.
- Suppose we can get the confidence interval of each estimate.

Estimator

Recall that each θ_i^* is a statistic of distribution D_i.

Suppose some arm i has been observed T_i times and output samples $X_{i,1}, X_{i,2}, \dots, X_{i,T_i}$.

- Mean: $\hat{\theta}_{i,t} = EST_{mean}(X_{i,1}, \dots, X_{i,T_i}) = \sum_{j=1}^{T_i} X_{i,j} / T_i$
- Variance: $\hat{\theta}_{i,t} = EST_{var}(X_{i,1}, \dots, X_{i,T_i}) = \frac{1}{T_i}\left(\sum_{j=1}^{T_i} X_{i,j}^2 - \frac{1}{T_i}\left(\sum_{j=1}^{T_i} X_{i,j}\right)^2\right)$

Confidence Interval

Which confidence interval of $\hat{\theta}_{i,t}$ contains the true value θ_i^* for all $t \geq t_0$ with high probability $1-\delta$?

- Event $\xi = \left\{\forall t \geq t_0, \forall i \in [m], \left|\hat{\theta}_{i,t} - \theta_i^*\right| \leq rad_{i,t}\right\}$ occurs with probability at least $1-\delta$.
- $rad_{i,t} = \sqrt{\frac{1}{2T_i} \ln \frac{4t^3}{\tau\delta}}$

  $\tau = 1$ for mean, $\tau = 2$ for variance

McDiarmid's inequality

$$\Pr\left[\left|\hat{\theta}_i - \theta_i^*\right| \geq \varepsilon\right] \leq 2\exp(-2T_i\varepsilon^2)$$

Problem Solving

Framework of Algorithm

1. (Initialization) Get an initial estimate $\hat{\theta}_{i,t_0}$ and confidence interval $\left[\hat{\theta}_{i,t_0} - rad_{i,t_0}, \hat{\theta}_{i,t_0} + rad_{i,t_0}\right]$ for each θ_i^*.
2. (Loop) For round $t = t_0 + 1, t_0 + 2, \ldots$
   - Calculate the current candidate set C_t
   - If $C_t = \emptyset$, the algorithm stops; if not, picks the arm with the largest confidence radius to pull.
   - Update the estimate $\hat{\theta}_{i,t}$ and confidence radius $rad_{i,t}$

Candidate Set C_t

We use $\widehat{\Theta}_t = \left\{ \boldsymbol{\theta} \in [0,1]^m \mid \left|\theta_i - \hat{\theta}_{i,t}\right| \leq rad_{i,t}, \forall i \in [m] \right\}$

to denote the confidence interval space.

- Let $C_t \leftarrow \emptyset$.
- For each $i \in [m]$, if $\max_{\theta \in \widehat{\Theta}_{t-1}} \phi_i(\boldsymbol{\theta}) \neq \min_{\theta \in \widehat{\Theta}_{t-1}} \phi_i(\boldsymbol{\theta})$, then

$$C_t \leftarrow C_t \cup \{i\}.$$

COCI: Consistently Optimal Confidence Interval Algorithm for CPE-CS

Input: Confidence error bound $\delta \in (0,1)$, maximization oracle ϕ.

Output: $\boldsymbol{y}^o = (y_1, y_2, \ldots, y_m) \in \mathcal{Y}$.

$t \leftarrow \tau m$; *// $\tau = 1$ for the mean estimator and $\tau = 2$ for the variance estimator*

for $i = 1, 2, \ldots, m$ **do**

  observe the i-th arm τ times $X_{i,1}, \ldots, X_{i,\tau}$;

  $T_{i,t} \leftarrow \tau$;

  estimate $\hat{\theta}_{i,t} \leftarrow \textsc{Est}_i(X_{i,1}, \ldots, X_{i,T_{i,t}})$;

  $\mathrm{rad}_{i,t} \leftarrow \sqrt{\frac{1}{2T_{i,t}} \ln \frac{4t^3}{\tau\delta}}$; *// confidence radius*

  $\hat{\Theta}_t \leftarrow \{\boldsymbol{\theta} \in [0,1]^m : |\theta_i - \hat{\theta}_{i,t}| \le \mathrm{rad}_{i,t}, \forall i \in [m]\}$;

Initialization

for $t = \tau m + 1, \tau m + 2, \tau m + 3, \ldots$ **do**

  $C_t \leftarrow \emptyset$;

  for $i = 1, 2, \ldots, m$ **do**

    if $\max_{\boldsymbol{\theta} \in \hat{\Theta}_{t-1}} \phi_i(\boldsymbol{\theta}) \neq \min_{\boldsymbol{\theta} \in \hat{\Theta}_{t-1}} \phi_i(\boldsymbol{\theta})$ **then**

      $C_t \leftarrow C_t \cup \{i\}$;

  if $C_t = \emptyset$ **then**

    return $\boldsymbol{y}^o = \phi(\boldsymbol{\theta})$ for an arbitrary $\boldsymbol{\theta} \in \hat{\Theta}_{t-1}$;

  $j \leftarrow \arg\max_{i \in C_t} \mathrm{rad}_{i,t-1}$;

  $T_{j,t} \leftarrow T_{j,t-1} + 1$; $T_{i,t} \leftarrow T_{i,t-1}$ for all $i \neq j$;

  play the j-th arm and observe the outcome $X_{j,T_{j,t}}$;

  update $\hat{\theta}_{j,t} \leftarrow \textsc{Est}_j(X_{j,1}, \ldots, X_{j,T_{j,t}})$;

  update $\hat{\theta}_{i,t} \leftarrow \hat{\theta}_{i,t-1}$ for all $i \neq j$;

  update $\mathrm{rad}_{i,t} \leftarrow \sqrt{\frac{1}{2T_{i,t}} \ln \frac{4t^3}{\tau\delta}}$ for all $i \in [m]$;

  $\hat{\Theta}_t \leftarrow \{\boldsymbol{\theta} \in [0,1]^m : |\theta_i - \hat{\theta}_{i,t}| \le \mathrm{rad}_{i,t}, \forall i \in [m]\}$;

Candidate set

Update parameters

Performance of COCI

We define the consistent optimality radius Λ_i for each arm i as:

$$\Lambda_i = \inf_{\boldsymbol{\theta}:\, \phi_i(\boldsymbol{\theta}) \neq \phi_i(\boldsymbol{\theta}^*)} \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_\infty = \inf_{\boldsymbol{\theta}:\, \phi_i(\boldsymbol{\theta}) \neq \phi_i(\boldsymbol{\theta}^*)} \max_{j \in [m]} \left|\theta_j - \theta_j^*\right|.$$

Thus, $\forall i \in [m]$, if $\left|\theta_j - \theta_j^*\right| < \Lambda_i$ holds for all $j \in [m]$, then

$$\phi_i(\boldsymbol{\theta}) = \phi_i(\boldsymbol{\theta}^*).$$

Define hardness: $H_\Lambda = \sum_{i=1}^m 1/\Lambda_i^2$.

Performance of COCI

Thm 1: With probability at least $1-\delta$, COCI returns the unique true optimal solution and the number of rounds

$$T \leq 2m + 12H_\Lambda \ln 24H_\Lambda + 4H_\Lambda \ln \frac{4}{\tau\delta} = O\left(H_\Lambda \log \frac{H_\Lambda}{\delta}\right)$$

$$\boxed{H_\Lambda = \sum_{i=1}^{m} 1/\Lambda_i^2}$$

Performance of COCI

Thm 2: Given m arms and $\delta \in (0,0.1)$, there exists an instance such that every algorithm for CPE-L which outputs the optimal solution with probability at least $1-\delta$, takes at least

$$\Omega(H_\Lambda + H_\Lambda m^{-1} \log \delta^{-1})$$

samples in expectation.

(Borrowing a lower bound analysis in [Chen *et al.*, 2017])

Applications

- CPE-L: match the bound in Chen et al. [2014]
- Water resource planning
- Other urban planning problems: air pollution control, criminal control, etc.

Thank you!