Bandit Problems

Rashish Tandon, Harsh Pareek

University of Texas at Austin

November 30, 2011

- Best Arm identification in Multi Arm Bandits: When the final payoff does not include the reward from the exploration stage
  - Channel selection for cell phones: Test a number of channels, then use one to transmit
  - Clinical trials: Sequentially test a number of formulae. Choose the best for commercialization
- Best Arm identification in Multi-Bandit Multi-Arm settings
  - Clinical Trials with M subpopulations, and K_m options for treating the m^{th} population
  - Online advertising with M subpopulations, and K_m options for advertising to the m^{th} population

Problem definition

- **Known Parameters**: Number of arms k and number of rounds n
- **Unknown Parameters**: Reward distributions $\nu_1, \ldots, \nu_k$ and unique arm i^* with maximal mean
- For each round:
  - The forecaster chooses an arm $I_t \in 1, \ldots, K$
  - The arm draws a reward from ν_{I_t}

  Finally, the forecaster outputs a recommendation J_n
- **Assumptions** Each arm has a finite first moment
- **Goal** Find the arm with maximal mean
- **Evaluation**
  - Regret: $r_n = \mu^* - \mu_{J_n}$
  - Gap: $\Delta_i = \mu^* - \mu_i$
  - Minimum Gap: $\Delta_{i^*} = \min_{i \neq i^*} \Delta_i$
  - Probability of error: $e_n = P(J_n \neq i^*)$

Lower Bound

Lower Bound

Let $\{\nu_1, \ldots, \nu_K\}$ be Bernoulli distributions with parameters in $[\frac{1}{3}, \frac{2}{3}]$. For any forecaster, $\exists c > 0$ such that, up to a permutation of the arms,

$$e_n \geq \exp\left(-c\frac{n\log(K)}{H}\right)$$

Here, $H = \sum_{i \neq i^*} \left(\frac{1}{\Delta_i}\right)^2$

Informally, we say, the algorithm required $O(H/\log K)$ rounds to find the best arm

Also, let $H' = \max_i \frac{i}{\Delta_{(i)}^2}$, where $\Delta_{(i)}$ is the gap for the i^{th} best arm.

Notice that $H' \leq H \leq \log(2K)H'$.

Uniform Strategy

Select each arm $k \in \{1, \ldots, K\}$ $\lceil \frac{n}{K} \rceil$ times

Output $J_n = \mathrm{argmax}_{i \in K} \hat{X}_i$

Probability of Error Bound

$\exists c > 0$ such that $e_n \leq \exp\left(-c\frac{n \min_i \Delta_i^2}{K}\right)$

Proof: (Union Bound and Hoeffdings inequality)

$$
\begin{aligned}
P(J_n \neq i^*) &\leq \sum_{i \neq i^*} P(\hat{X}_i \geq \hat{X}_{i^*}) \\
&\leq \sum_{i \neq i^*} P(\hat{X}_i - \mu_i + \mu_{i^*} - \hat{X}_{i^*} \geq \Delta_i) \\
&\leq K \exp(-\frac{cn \min \Delta_i^2}{k})
\end{aligned}
$$

Successive Rejects

Let $\overline{\log}(K) = \frac{1}{2} + \sum_{i=2}^{K} \frac{1}{i}$

Let $n_0 = 0, n_k = \lceil \frac{1}{\overline{\log}(K)} \frac{n-K}{k+1-K} \rceil$ for $1 \leq k \leq K-1$

Let $A_1 = 1, \ldots, K$. For $k = 1 \ldots K$:

- $\forall i \in A_k$, select arm i $n_k - n_{k-1}$ times.
- $A_{k+1} = A_k \backslash \text{argmin}_{i \in A_k} \hat{X}_{i,n_k}$, where $\hat{X}_{i,n_k}$ is the emperical value of the mean of arm i

Output the unique element in A_K as J_n

Probability of Error Bound

$\exists c > 0$ such that

$$e_n \leq \exp\left(-c \frac{n}{H' \overline{\log}(K)}\right)$$

Upper Confidence Bound - Exploration

Let $B_{i,t} = \hat{X}_{i,T_i(t)} + \sqrt{\frac{cn/H}{T_i(t)}}$,
where $T_i(t) =$ No. of times arm i is pulled till stage t

For each round t,
Draw $I_t \in \mathrm{argmax}_{i \in 1,\ldots,K} B_{i,t-1}$
Finally, $J_n = \mathrm{argmax}_{i,\ldots,K} \hat{X}_{i,T_i(t)}$

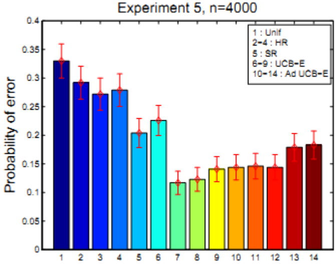
Probability of Error Bound

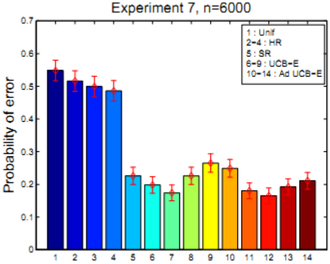
For c small enough, $\exists c' > 0$ such that

$$e_n \leq \exp(-c'n/H)$$

Here, c is an exploration parameter.
Note that H is unknown, thus effectively a parameter, which must be tuned.

Experiments





Multi-Bandit Multi-Arm Problem

Multi-Bandit Best Arm Identification - Gabillon, Ghavamzadeh, Lazaric, Bubeck (NIPS 2011)

Setup similar to before.
M bandits. The m^{th} bandit has K_m arms

Evaluation

Regret

$$r(n) = \frac{1}{M}\sum_{m=1}^{M} r_m(n) = \frac{1}{M}\sum_{m=1}^{M}\left(\mu_m^* - \mu_{J_m(n)}\right)$$

Average probability of error

$$e(n) = \frac{1}{M}\sum_{m=1}^{M} r_m(n) = \frac{1}{M}\sum_{m=1}^{M}\left(P(J_m(n) \neq k_{m^*})\right)$$

Global max error

$$l(n) = \max_m l_m(n) = \max_m P(J_m(n) \neq k_{m^*})$$

Gap Exploration

Similar to UCB-E

Parameters: Number of rounds n, exploration parameter a, maximum range b

Let $T_{mk}(0) = 0, \hat{\Delta}_{mk}(0) = 0, \forall m, k$

For $t = 1, \ldots, n$:

- $B_{mk}(t) = -\hat{\Delta}_{mk}(t-1) + b\sqrt{\frac{a}{T_{mk}(t-1)}}$
- Draw $I_t \in \mathrm{argmax}_{m,k}\ B_m k(t)$
- Pull arm I_t and update the selected bandit and arm

Probability of Error Bound

$\exists c_1, c_2 > 0$ such that

$$l(n) \leq P(\exists m : J_m(n) \neq k_{m^*}) \leq c_1 MKn \exp\left(-c_2 \frac{n - MK}{H}\right)$$

Here, $H = \sum_{m,k} \frac{b^2}{\Delta_{mk}^2}$

Further remarks and Summary

- For the algorithms which have parameters, corresponding adaptive versions also exist, and perform well in practice
- $H = 1/\Delta_i^2$ measures the complexity of the problem
- We need at least $H/\log(K)$ rounds to find the best arm
- SR is parameterless and find the best arms in $H\log(K)$ rounds
- UCB - E require H rounds, but needs the value of H
- Gap-E generalizes UCB-E to the multi bandit case

Linear Optimization for Bandit Problems

The Price of Bandit Information for Online Optimization - Dani, Hayes, Kakade (NIPS 2008)

We are given a set $D \subset \mathbb{R}^n$ - the decision space

D could be possibly infinite

Each vector in D is a possible decision or action

For $t = 1, 2, \ldots, T$

- Adversary chooses $L_t \in \mathbb{R}^n$
- Learner chooses $x_t \in D$
- A loss of $l_t = L_t^T x_t$ is incurred - Learner only sees this

Possible Applications

- Path planning
- Network Routing

Assumptions

- Learner has a randomized strategy. Chooses x_t probabilistically
- L_t must be admissible i.e. $0 \leq L_t^T x \leq 1, \; \forall x \in D$
- D contains a barycentric spanner i.e. there are n l.i. vectors $\{v_1, \ldots, v_n\}$ in D and any vector $x \in D$ is s.t. $x = \sum_i \lambda_i v_i$ with $\lambda_i \in [-1, 1]$.
- Such a barycentric spanner can be obtained for any n-dimensional compact set in $\mathbb{R}^n$ (*Awerbuch and Kleinberg*, STOC 2004)
- WLOG, $\{e_1, e_2, \ldots, e_n\} \subset D$ and $D \subset [-1, 1]^n$

Define regret for a particular choice of actions $\{x_1, \ldots, x_T\}$ as,

$$R = \sum_t L_t^T x_t - \min_{x \in D} \sum_t L_t^T x$$

AIM : Minimize the regret

Since we have a randomized strategy, we want to bound $E[R]$

Previous Bounds

- Use general K-arm bandit strategy by *Auer et. al*. Gives, $E[R] = O(\sqrt{TK \log K})$. $K = |D|$ in this setting. Does not exploit linearity of loss. Moreover, $|D|$ can be very high.
- Approaches prior to this - Bound of $O(poly(n)T^{\frac{2}{3}})$ (*Dani and Hayes*, SODA 2006) or worse.

Making D finite

Theorem

For any $D \subset [-1,1]^n$, $\exists \tilde{D}$ *s.t.* $|\tilde{D}| \leq (nT)^{\frac{n}{2}}$ *and*

$$|OPT\ REGRET(D) - OPT\ REGRET(\tilde{D})| \leq \sqrt{nT}$$

Also, $\tilde{D}$ *forms a* $1/\sqrt{T}$*-net for* D*.*

- How ? - For every $x \in D$, truncate each coordinate to the first $\frac{1}{2}\log(nT)$ bits. Keep this vector in $\tilde{D}$.

Thus, only concerned with finite decision sets for obtaining sharp regret bounds

Input : $\gamma,\ \eta$

$\forall\, x \in D,\ p_1(x) \leftarrow \frac{1}{|D|}$

for $t \leftarrow 1$ to T **do**

$\forall\, x \in D,\ \hat{p}_t(x) = (1-\gamma)p_t(x) + \frac{\gamma}{n}\mathbf{1}\{x \in \text{ spanner }\}$

Sample x_t from $\hat{p}_t$

Observe the loss $l_t = L_t^T x_t$

Compute the covariance for $\hat{p}_t$, $C_t := E[xx^T]$

Estimate the loss vector as $\hat{L}_t := l_t C_t^{-1} x_t$

$\forall\, x \in D,\ p_{t+1}(x) \propto p_t(x) e^{-\eta \hat{L}_t^T x}$

end for

- Inspired from the multiplicative weight approach of *Auer et. al.*
- A probability distribution, p_t is defined over the decision space D at any stage t
- This is mixed with a uniform distribution on the spanner set to get $\hat{p}_t$
- Action x_t for stage t is chosen from this distribution
- $p_t(x)$ is modified based on the *estimated* loss
- Obtain a bound of $O(n \log n\sqrt{nT})$

Theorem (Main Result)

Choose $\gamma = \frac{n^{3/2}}{\sqrt{T}}$ *and* $\eta = \frac{1}{\sqrt{nT}}$. *For any sequence of loss vectors* $L_1, \ldots, L_T$, *the algorithm satisfies :*

$$E[R] \leq \log|D|\sqrt{nT} + 2n^{3/2}\sqrt{T}$$

Since, $|D| \leq (nT)^{\frac{n}{2}}$, we have the bound.

Proof Roadmap for Bound on Regret

- Observe that $\hat{L}_t$ is a meaningful estimate i.e. $E[\hat{L}_t] = L_t$
- The *estimated* loss vectors are also bounded

$$|\hat{L}_t^T x| \leq \frac{n^2}{\gamma}, \forall x \in D$$

- Use a result similar to *Auer et. al.*

  Given vectors $\hat{L}_1, \ldots, \hat{L}_T$ and probability distributions $p_1, \ldots, p_T$ that undergo an exponential reweighting based on the given vectors, then for any $x^* \in D$,

$$\sum_t \sum_x p_t(x) \left(\hat{L}_t^T x\right) \leq \sum_t \hat{L}_t^T x^* + \frac{\log|D|}{\eta} + \frac{\phi_M(\eta)}{\eta} \sum_t \sum_x p_t(x) \left(\hat{L}_t^T x\right)^2$$

- Here, $\phi_M(\eta) := \frac{e^{M\eta}-1-M\eta}{M^2}$ and M is the upper bound on $|\hat{L}_t^T x|$ i.e. $M = \frac{n^2}{\gamma}$
- Some technical lemmas

$$E[(\hat{L}_t^T x)^2] \leq x^T C_t^{-1} x$$

$$\sum_x \hat{p}_t(x) x^T C_t^{-1} x = n$$

- Use those to give a bound for $\sum_{t,x} \hat{p}_t(x) \left(\hat{L}_t(x)^T x_t\right)$
- The expectation of above is same as $E[\sum_t L_t^T x_t]$

Lower Bound for the Problem $= \Omega\left(n\sqrt{T}\right)$