Distributed Algorithms for Learning and Cognitive Medium Access with Logarithmic Regret

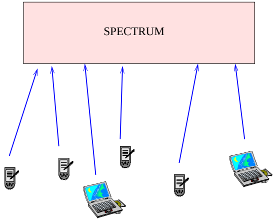
Anima Anandkumar

Electrical Engineering and Computer Science
University of California Irvine

Joint work with Nithin Michael and Ao Tang, Cornell University

USC EE Systems Seminar

Introduction: Distributed Medium Access



Constraints of users

- Sensing constraints: Sense only part of spectrum at any time
- Local information: No centralized control
- Lack of coordination: Collisions among users
- Unknown channel conditions Lost opportunities

Cooperation and Competition among the Users

Distributed and Adaptive Medium Access Control

Introduction: Cognitive Radio Networks

Two types of users

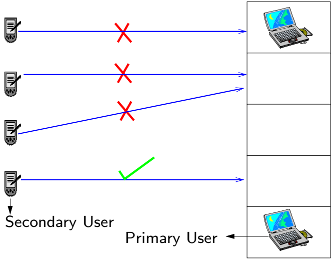
- Primary Users

  Priority for channel access

- Secondary or Cognitive Users

  Opportunistic access

  Channel sensing abilities



Key Difference

Availability of Sensing Samples: Leverage for learning channel conditions

Performance Measures for Distributed Mechanisms

Questions

- Can sensing samples be used to learn channels with good availability?
- Can we design distributed medium access to avoid collisions among the users?
- What is the cost of learning and distributed access?

Learning Criteria

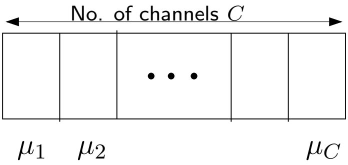
- Consistency: Learn the good channels with large number of sensing samples
- Low Regret: Minimize access of bad channels

Channel Access Criteria

- Orthogonalization: Minimize collisions among users

Maximize total secondary throughput under distributed learning and access

Setup: Distributed Learning and Access

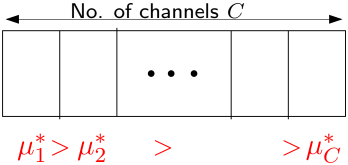


- Slotted tx. with U cognitive users and $C > U$ channels
- Channel Availability for Cognitive Users: Mean availability μ_i for channel i and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_C]$.
- $\boldsymbol{\mu}$ unknown to secondary users: learning through sensing samples
- No explicit communication/cooperation among cognitive users

Objectives for secondary users

- Users ultimately access orthogonal channels with best availabilities $\boldsymbol{\mu}$
- Max. Total Cognitive System Throughput $\equiv$ Min. Regret

Setup: Distributed Learning and Access

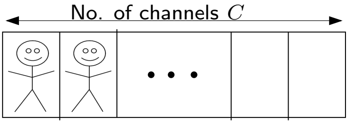


- Slotted tx. with U cognitive users and $C > U$ channels
- Channel Availability for Cognitive Users: Mean availability μ_i for channel i and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_C]$.
- $\boldsymbol{\mu}$ unknown to secondary users: learning through sensing samples
- No explicit communication/cooperation among cognitive users

Objectives for secondary users

- Users ultimately access orthogonal channels with best availabilities $\boldsymbol{\mu}$
- Max. Total Cognitive System Throughput $\equiv$ Min. Regret

Setup: Distributed Learning and Access



$\mu_1^* > \mu_2^* \quad > \quad > \mu_C^*$

- Slotted tx. with U cognitive users and $C > U$ channels
- Channel Availability for Cognitive Users: Mean availability μ_i for channel i and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_C]$.
- $\boldsymbol{\mu}$ unknown to secondary users: learning through sensing samples
- No explicit communication/cooperation among cognitive users

Objectives for secondary users

- Users ultimately access orthogonal channels with best availabilities $\boldsymbol{\mu}$
- Max. Total Cognitive System Throughput $\equiv$ Min. Regret

Summary of Results: Three Algorithms

- Distributed algorithms based on local information
- Performance guarantees under self play

- ρ^{PRE}: under pre-allocated ranks among cognitive users
  - Learning of channels corresponding to assigned ranks

- ρ^{RAND}: no assigned ranks but number of secondary users known
  - Learning channel ranks and adapting to collisions

- ρ^{EST}: no assigned ranks and unknown number of secondary users
  - Learning channel ranks, adapting to collisions and estimating number of users based on number of collisions

Summary of Results (Contd.)

- Provable guarantees on sum regret under the three policies
  - Convergence to optimal configuration (orthogonal occupancy in the best channels)
  - Regret grows in no. of access slots as $R(n) \sim O(\log n)$ for ρ^{PRE} and ρ^{RAND}
  - Regret grows in no. of access slots as $R(n) \sim O(f(n) \log n)$ for any $f(n) \to \infty$ for ρ^{EST}
- Lower bound for any uniformly-good policy: also logarithmic in no. of access slots $R(n) \sim \Omega(\log n)$

We propose order-optimal distributed learning and allocation policies

---

A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition" in Proc. of INFOCOM, (San Deigo, USA), Mar. 2010.

A. Anandkumar, N. Michael, A.K. Tang, and A. Swami, "Distributed Learning and Allocation of Cognitive Users with Logarithmic Regret", to appear, IEEE JSAC on Cognitive Radio.

Related Work

Multi-armed Bandits

- Single cognitive user (Lai & Robbins 85)
- Multiple users with centralized allocation (Ananthram et. al 87)
  Key Result: Regret $R(n) \sim O(\log n)$ and optimal as $n \to \infty$
- Auer et. al. 02: order optimality for sample mean policies

Cognitive Medium Access & Learning

- Liu et. al. 08: Explicit communication among users
- Li 08: Q-learning, Sensing all channels simultaneously
- Liu & Zhao 10: Learning under time division access: users do not orthogonalize to different channels
- Gai et. al. 10: Combinatorial bandits, centralized learning under heterogeneous channel availability

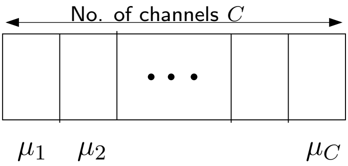
Outline

System Model

Primary and Cognitive Networks

- Slotted tx. with U cognitive users and C channels
- Primary Users: IID tx. in each slot and channel

  Channel Availability for Cognitive Users: In each slot, IID with prob. μ_i for channel i and $\boldsymbol{\mu} = [\mu_1, \ldots, \mu_C]$.
- Perfect Sensing: Primary user always detected
- Collision Channel: tx. successful only if sole user
- Equal rate among secondary users:
  Throughput $\equiv$ total no. of successful tx.



Problem Formulation

Distributed Learning Through Sensing Samples

- No information exchange/coordination among secondary users
- All secondary users employ same policy

Throughput under perfect knowledge of $\boldsymbol{\mu}$ and coordination

$$S^*(n; \boldsymbol{\mu}, U) := n \sum_{j=1}^{U} \mu(j^*)$$

where j^* is j^{th} largest entry in $\boldsymbol{\mu}$ and n: no. of access slots

Regret under learning and distributed access policy ρ

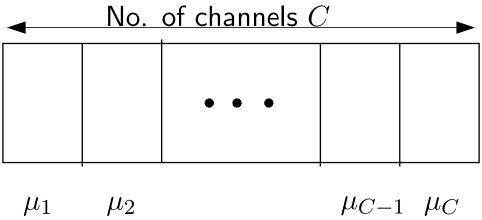
Loss in throughput due to learning and collisions

$$R(n; \boldsymbol{\mu}, U, \rho) := S^*(n; \boldsymbol{\mu}, U) - S(n; \boldsymbol{\mu}, U, \rho)$$

Max. Throughput $\equiv$ Min. Sum Regret

Outline

Single Cognitive User: Multi-armed Bandit

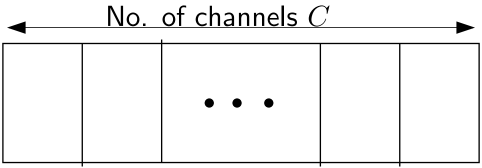


Exploration vs. Exploitation Tradeoff

- Exploration: channels with good availability are not missed
- Exploitation: obtain good throughput

Explore in the beginning and exploit in the long run

Single Cognitive User: Multi-armed Bandit



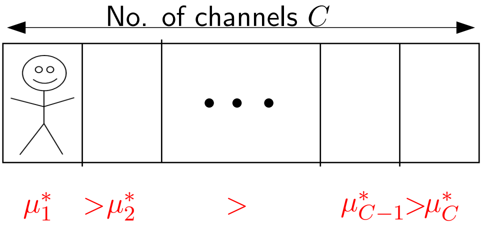
$\mu_1^* > \mu_2^* > \cdots > \mu_{C-1}^* > \mu_C^*$

Exploration vs. Exploitation Tradeoff

- Exploration: channels with good availability are not missed
- Exploitation: obtain good throughput

Explore in the beginning and exploit in the long run

Single Cognitive User: Multi-armed Bandit



Exploration vs. Exploitation Tradeoff

- Exploration: channels with good availability are not missed
- Exploitation: obtain good throughput

Explore in the beginning and exploit in the long run

Single Cognitive User: Multi-armed Bandit (Contd.)

- $T_i(n)$: no. of slots where user j selects channel i
- $\overline{X}_i(T_i(n))$: sample mean availability of channel i acc. to user j
- 1^*: channel with the highest mean availability $\mu(1^*) \geq \mu_j, \forall j$.
- 1-worst channel: channels which have lower availability than 1^*

Two Policies based on Sample Mean (Auer et. al. 02)

- Deterministic Policy: Select channel with highest g-statistic:

$$g(i;n) := \overline{X}_i(T_i(n)) + \sqrt{\frac{2\log n}{T_i(n)}}$$

- Randomized Greedy Policy: Select channel with highest $\overline{X}_i(T_i(n))$ with prob. $1-\epsilon_n$ and with prob. ϵ_n unif. select other channels, where

$$\epsilon_n := \min[\frac{\beta}{n}, 1]$$

Policies with Logarithmic Regret

Simplification of Regret $R(n) := S^*(n) - S(n)$

$$R(n) = \sum_{i \in 1\text{-worst}} \Delta(1^*, i)\mathbb{E}[T_i(n)].$$

where $\Delta(1^*, i) := \mu(1^*) - \mu(i)$.

Theorem (Deterministic Policy)

Time spent in any channel i which is 1-worst under g-statistic policy, where $g(i;n) := \overline{X}_i(T_i(n)) + \sqrt{\frac{2\log n}{T_i(n)}}$ is

$$\mathbb{E}[T_i(n)] \leq \Delta(1^*, i)\left[\frac{8\log n}{\Delta(1^*, i)^2} + 1 + \frac{\pi^2}{3}\right], \quad \forall i = 1, \ldots, C, i \in 1\text{-}\mathit{worst}.$$

Policies with Logarithmic Regret (Contd.)

Theorem (Randomized Policy)

No. of slots a channel $i \neq 1^*$ is accessed under randomized greedy policy satisfies

$$\mathbb{E}[T_i(n)] \leq \frac{\beta}{C} \log n + \delta, \quad \forall i = 1, \ldots, C, i \in 1\text{-}\textit{worst},$$

when

$$\beta > \max[20, \frac{4}{\Delta_{\min}^2}],$$

where $\Delta_{\min} := \min_{i \in 1\text{-}\textit{worst}} \Delta(1^*, i)$ is minimum separation.

Regret under the two policies is $O(\log n)$ for n no. of access slots

Lower Bound on Regret

Uniformly good policy ρ

A policy which enables the single cognitive user to ultimately settle down in the best channel under any channel availabilities $\boldsymbol{\mu}$ and the user spends most of time in the best channel

$$\mathbb{E}_{\boldsymbol{\mu}}[n - T_i(n)] = o(n^\alpha), \quad \forall \alpha > 0, \boldsymbol{\mu} \in (0,1)^C, i \in \text{1-worst}.$$

Satisfied by the two policies of Auer et. al.

Theorem (Lower Bound, Lai & Robbins 85)

Time spent in a 1-worst channel under any uniformly good policy satisfies

$$\lim_{n \to \infty} \mathbb{P}\left[T_i(n;\rho) \geq \frac{(1-\epsilon)\log n}{D(\mu_i, \mu_{1^*})}; \boldsymbol{\mu}, \rho\right] = 1, \forall i \in \textit{1-worst}$$

Hence, $R(n) = \Omega(\log n)$ *for any uniformly good policy.*

Outline

Outline

Learning Under Pre-Allocation

If user j is assigned rank w_j, select channel with w_j^{th} highest $\overline{X}_{i,j}(T_{i,j}(n))$ with prob. $1-\epsilon_n$ and with prob. ϵ_n unif. select other channels, where

$$\epsilon_n := \min[\frac{\beta}{n}, 1]$$

- Allows for heterogeneous users
- Feedback on collisions not required for learning

Regret: user does not select channel of pre-assigned rank

$$\mathbb{E}[T_{i,j}(n)] \leq \sum_{t=1}^{n-1} \frac{\epsilon_{t+1}}{C} + \sum_{t=1}^{n-1}(1-\epsilon_{t+1})\mathbb{P}[\mathcal{E}_{i,j}(n)], \;\; i \neq w_j^*,$$

where $\mathcal{E}_{i,j}(n)$ is the error event that w_j^{th} highest entry of $\bar{X}_{i,j}(T_{i,j}(n))$ is not same as $\mu_{w_j}^*$

Regret Under Pre-allocation

Theorem (Regret Under ρ^{PRE} Policy)

No. of slots user j accesses channel $i \neq w_j^*$ other than pre-allocated channel under ρ^{PRE} satisfies

$$\mathbb{E}[T_{i,j}(n)] \leq \frac{\beta}{C}\log n + \delta, \quad \forall i = 1, \ldots, C, i \neq w_j^*,$$

when

$$\beta > \max[20, \frac{4}{\Delta_{\min}^2}],$$

where $\Delta_{\min} := \min_{i \in U\text{-worst}} \Delta(U^*, i)$ is minimum separation between a U-worst channel and the U^{th} channel.

Regret $R(n) = O(\log n)$ under ρ^{PRE}

Outline

Distributed Learning and Randomized Allocation ρ^{RAND}

User adaptively chooses rank w_j based on feedback for successful tx.

- If collision in previous slot, draw a new w_j uniformly from 1 to U
- If no collision, retain the current w_j

Select channel with w_j^{th} highest entry:

$$g_j(i;n) := \overline{X}_{i,j}(T_{i,j}(n)) + \sqrt{\frac{2\log n}{T_{i,j}(n)}}$$

Upper Bound on Regret

$$R(n) \leq \mu(1^*)\left[\sum_{j=1}^{U} \sum_{i \in U\text{-worst}} \mathbb{E}[T_{i,j}(n) + M(n)]\right]$$

- $\sum\limits_{i \in U\text{-worst}} T_{i,j}(n)$: Time spent in U-worst channels by user j
- $M(n)$: No. of collisions in U-best channels

Distributed Learning and Randomized Allocation ρ^{RAND}

Theorem

Under ρ^{RAND} Policy, $\mathbb{E}[\sum_{i \in U\text{-worst}} T_{i,j}(n)]$ and $\mathbb{E}[M(n)]$ are $O(\log n)$ and hence, regret is $O(\log n)$ where n is the number of access slots.

Proof Steps for $\mathbb{E}[\sum_{i \in U\text{-worst}} T_{i,j}(n)]$

- Bound time spent in U-worst channels by decay of exploration term
- Chernoff-Hoeffding bounds for concentration of sample mean channel availability
- Techniques similar to Auer et. al.

Idea of Proof for Regret Under ρ^{RAND}

Proof for $\mathbb{E}[M(n)]$: no. of collisions in U-best channels

- Bound collisions under perfect knowledge of $\boldsymbol{\mu}$ as $\Pi(U)$
- Relate $\Pi(U)$ with $\mathbb{E}[M(n)]$ under learning of $\boldsymbol{\mu}$

Analysis of $\Pi(U)$

- Markov chain with state space of all possible user configurations where the absorbing state is the orthogonal configuration
- $\Pi(U)$: mean time to absorption under uniform randomization of colliding users
- $\Pi(U) < \infty$ since finite state Markov chain

Idea of Proof for Regret Under ρ^{RAND} Contd.,

Bound on $\mathbb{E}[M(n)]$ under learning

- Good state: all users estimate order of top-U channels correctly, otherwise bad state
- Bound separately collisions under good and bad states
- Under run of good states: $\Pi(U)$ mean no. of collisions
- Run of bad states is $O(\log n)$

Outline

Learning with Unknown No. of Users

- No. of cognitive users U fixed but unknown to policy
- Horizon length n known to users
- Regret bounds as $n \to \infty$

ρ^{EST}: Joint learning of channel order and no. of users

- Maintain an estimate of no. of users $\widehat{U}$
- Execute $\rho^{\text{RAND}}(n; \widehat{U})$ based on g-statistic assuming $\widehat{U}$ no. of users
- Update $\widehat{U}$ based on feedback (no. of collisions)

Update Rule for $\widehat{U}$ under ρ^{EST}

- Fixed threshold functions $\xi(n; k)$ for $n = 1, 2, \ldots$ and $k = 1, \ldots C$.
- Slow start: Initialize $\widehat{U} \leftarrow 1$.
- If no. of collisions in top-$\widehat{U}$ channels exceeds threshold $\xi(n; \widehat{U})$, then

$$\widehat{U} \leftarrow \widehat{U} + 1$$

Learning with Unknown No. of Users (Contd.)

Theorem

For any class of threshold functions $\xi(n;k)$ *satisfying*

$$\lim_{n\to\infty} \frac{\xi(n;k)}{\log n} = \infty, \quad \forall k > 1,$$

the regret under ρ^{EST} *policy satisfies*

$$\limsup_{n\to\infty} \frac{R(n;\boldsymbol{\mu}, U, \rho^{\mathsf{EST}})}{\xi^*(n;U)} < \infty,$$

where

$$\xi^*(n;U) := \max_{k=1,\ldots,U} \xi(n;k).$$

Regret is asymptotically (slightly more than) logarithmic under ρ^{EST}

Proof Steps

- Overestimation: Probability that $\widehat{U} \leq U$
- Conditional regret: Regret under $\widehat{U} \leq U$

Overestimation Error

- Number of collisions experienced when $\widehat{U} = U$ is $\Theta(\log n)$
- Applying threshold $\xi(n) = \omega(\log n)$ ensures that $\widehat{U}$ is not incremented

Conditional Regret

- Time spent in U-worst channels is $O(\log n)$
- Number of collisions is $O(\xi^*)$

Outline

Lower Bound on Regret

Uniformly good policy ρ

A policy which enables users to ultimately settle down in orthogonal best channels under any channel availabilities $\boldsymbol{\mu}$: user j spends most of time in U-best channels and time spent in $i \in U$-worst channel satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[n - T_{i,j}(n)] = o(n^{\alpha}), \quad \forall \alpha > 0, \boldsymbol{\mu} \in (0,1)^C, i \in U\text{-worst}.$$

Satisfied by ρ^{PRE} and ρ^{RAND} policies

Theorem (Lower Bound for Uniformly Good Policy (Liu & Zhao))

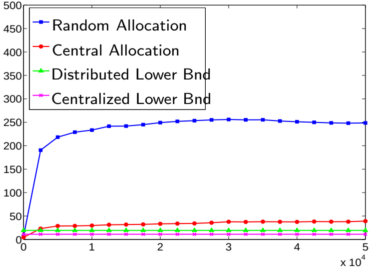
The sum regret satisfies

$$\liminf_{n \to \infty} \frac{R(n; \boldsymbol{\mu}, U, \rho)}{\log n} \geq \sum_{i \in U\text{-worst}} \sum_{j=1}^{U} \frac{\Delta(U^*, i)}{D(\mu_i, \mu_{j^*})}.$$

Order optimal regret under ρ^{PRE} and ρ^{RAND} policies

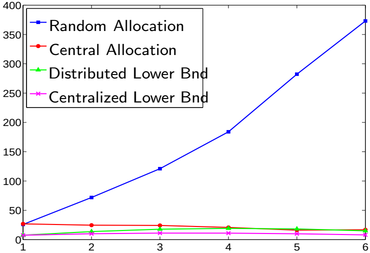
Outline

Simulation Results



Normalized regret $\frac{R(n)}{\log n}$ vs. n slots.

$U = 4$ users, $C = 9$ channels.

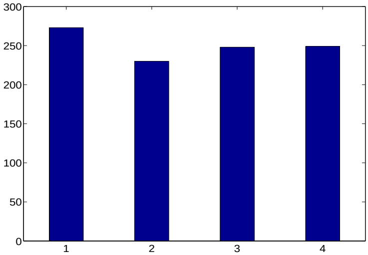


Normalized regret $\frac{R(n)}{\log n}$ vs. U users.

$C = 9$ channels, $n = 2500$ slots.

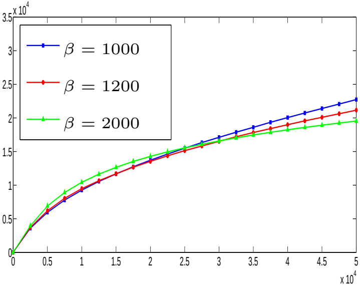
Probability of Availability $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$.

Simulation Results



No. of runs with top rank vs. user.

$U = 4$, $C = 9$, $n = 2500$ slots, ρ^{RAND}.



Regret $R(n)$ vs. n slots (varying β).

$U = 4$ users, $C = 9$ channels, ρ^{PRE}.

Probability of Availability $\boldsymbol{\mu} = [0.1, 0.2, \ldots, 0.9]$.

Outline

Conclusion

Summary

- Considered maximizing total throughput of cognitive users under unknown channel availabilities and no coordination
- Proposed two algorithms which achieve order optimality

  ρ^{PRE} policy works under pre-allocated ranks
  ρ^{RAND} policy does not require prior information
- Proposed ρ^{EST} policy when no. of users is unknown

  ρ^{EST} has asymptotically (slightly more than) logarithmic regret

---

A. Anandkumar, N. Michael, and A.K. Tang, "Opportunistic Spectrum Access with Multiple Users: Learning under Competition" in Proc. of IEEE INFOCOM, (San Deigo, USA), Mar. 2010.

A. Anandkumar, N. Michael, A.K. Tang, and A. Swami, "Distributed Learning and Allocation of Cognitive Users with Logarithmic Regret", available on website.