

# From Bandits to Experts: On the Value of Side-Observations

Shie Mannor  
The Technion

Ohad Shamir\*  
Microsoft Research  
New England

NIPS  
December 2011

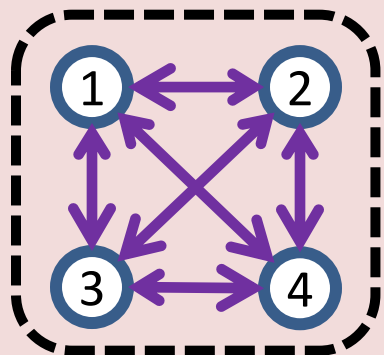
# What We Do

- Sequential decision making: repeatedly choose among  $k$  actions
- **“Experts” Setting:** Learner sees rewards of all actions
- **“Bandits” Setting:** Learner sees only its own reward

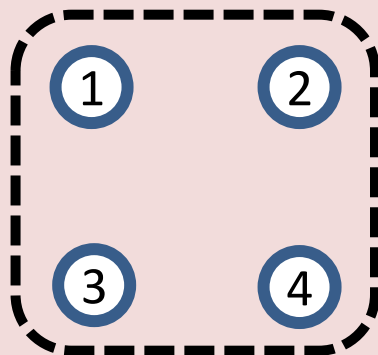
## Our (More General) Model

By picking an action, Learner gets **side-information** on **some subset** of other actions. Captures:

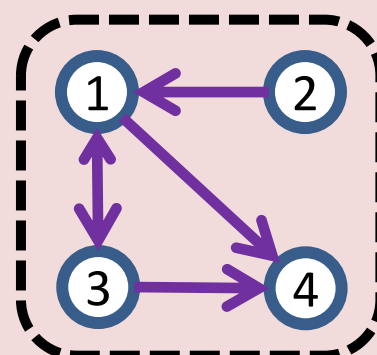
Experts...



...bandits...



...and in between!



# Why

- Captures **structure** between the actions
  - Web Advertising



- Sensor and communication networks

- **Related work:** Focus on affinity (Bandits in Metric Spaces) or stochastic correlations, while we make no such assumptions



# Results

- New Efficient Algorithms
- Regret upper and lower bounds
  - Non-trivial dependence on information feedback structure
- Experiments
- Many open questions!



Come see our poster!

## From Bandits to Experts: On the Value of Side-Observations

Shie Mannor      Ohad Shamir  
The Technion      MSR New England

### Experts and Bandits

- Fixed set of  $k$  actions
- Each round  $t = 1, 2, \dots, T$ :
  - Nature chooses rewards vector  $g^t \in [0, 1]^k$
  - Learner chooses action  $i_t \in \{1, \dots, k\}$
  - Learner gets reward  $g_{i_t}^t$
  - Experts: Learner then sees  $g^t$
  - Bandits: Learner then sees  $g_{i_t}^t$  only
- Goal: minimize regret with respect to best action
 
$$\max_{i^*} \sum_{t=1}^T g_{i^*}^t - \sum_{t=1}^T g_{i_t}^t$$
- Experts: Can get  $O(\sqrt{k \log(k)T})$  regret
- Bandits: Can get  $O(\sqrt{k \log(k)T})$  regret

### Our Model:

- Graph  $G$  over  $\{1, \dots, k\}$
- Each round  $t$ :
  - Learner picks  $i_t \in \{1, \dots, k\}$
  - Learner then sees  $g_{i_t}^t$  as well as  $g_j^t$  for all neighbors  $j$  of  $i_t$  in the graph
- We can also deal with time varying graphs, noisy observations...

Can always get  $O(\sqrt{k \log(k)T})$  regret by ignoring side-observations  
Main question: can we do better?

Experts...

...bandits...

...and in between

### Motivation

Captures structure between the actions

- Web Advertising
- Sensor and communication networks

Related work: Focus on affinity (Bandits in Metric, Sparse) or stochastic correlations, while we make no such assumptions

### Lower Bound

**Theorem**

- Exists adversary strategy such that any Learner has expected regret  $\Omega(\sqrt{\alpha(G)T})$
- $\alpha(G)$  is size of largest independent set

**Proof Idea**

Adversary can make game as hard as standard multi-armed bandits over  $\alpha(G)$  actions. Known lower bound of  $\Omega(\sqrt{\alpha(G)T})$  for multi-armed bandits over  $\alpha$  actions.

### ExpBan

**Theorem**

- Expected regret at most  $O(\sqrt{c \log(k)T})$  for optimal clique partition,  $c = \beta(G)$ , the clique-partition number of graph  $G$

**Theorem**

- Expected regret at most  $O(\sqrt{c \log(k)T})$  for optimal clique partition,  $c = \beta(G)$ , the clique-partition number of graph  $G$

Simple, but

- Sub-optimal regret
- Finding/approximating optimal clique partition is NP-hard!
- In worst case, no improvement over standard bandit setting

### ELP

(Exponentially-weighted algorithm with Linear Programming)

- Parameters  $\gamma, \beta$ , graph  $G$  with neighborhood sets  $\{N_i\}_{i=1}^k$
- Initialize weights  $w^1 = (\frac{1}{k}, \dots, \frac{1}{k})$
- For  $t = 1, 2, \dots, T$ :
  - Let  $p^t = (1 - \gamma) \frac{w^t}{\sum_{i=1}^k w_i^t} + \gamma w^t$ , where  $w_i = \arg \max_{i \in N_j} \sum_{j=1}^k w_j$  (dominance)
  - Pick action  $i_t$  with probability  $p_{i_t}^t$
  - For all  $j \in \{1, \dots, k\}$ , let  $\beta_j^t = \begin{cases} \frac{1}{\sum_{i \in N_j} p_i^t} & \text{if } i_t \in N_j \\ 0 & \text{if } i_t \notin N_j \end{cases}$
  - For all  $j \in \{1, \dots, k\}$ , let  $w_j^{t+1} = \exp(\beta_j^t)$
- Endfor

### Regret Analysis

Key quantity: upper bound for  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j}$

Empty graph (bandits):  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} = T$   
Regret  $O(\sqrt{k \log(k)T})$

Complete graph (experts):  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} = 1$   
Regret  $O(\sqrt{\log(k)T})$

Combinatorial generalization of cyclic sums: Hazlett's inequality (1970):  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} \leq \frac{1}{2}$

Shapiro's inequality (1974):  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} \leq \frac{1}{2}$

Bastion (1974): upper and lower bounds for  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} = \frac{1}{2}$

For undirected graphs, can show  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} = \alpha(G)$

Results in  $O(\sqrt{\alpha(G) \log(k)T})$  expected regret for ELP algorithm – virtually matching the lower bound

**Proof technique:**

- Start with an arbitrary assignment
- Appropriate weight change only increases value (convexity argument)

End with an independent set. Value equals size of set

For directed graphs, can only show much weaker  $\sum_{t=1}^T \sum_{i=1}^k \frac{p_i}{\sum_{j=1}^k p_j} \leq \beta(G)$

Regret for directed graphs still unclear...

### Simple Examples

**Star Graph**

Single "center-action" reveals all rewards  
Same regret as bandit setting

**Metric Spaces**

Edge between actions  $i, j$  if distance of  $i$  from  $j$  is  $\leq d$   
 $\alpha(G)$  = sphere-packing number  
 $\beta(G)$  = sphere-covering number  
 large difference in high dimensions

**Erdős-Rényi Random Graph**

Actions connected randomly with probability  $p$   
 For constant  $p$ ,  $\alpha(G) = O(\log(k))$ , and  $\beta(G) = \Omega(k / \log(k))$

### Experiments

- Random graph with parameter  $p$  over 300 actions
- Bernoulli rewards

### Extensions and Some Open Questions

- Time-changing graphs; noisy observations
- High probability regret bounds; bounds over policies (EXP4 analogue)...

- What is optimal regret for undirected graphs? What algorithm?
- What is optimal regret for time changing graphs? What algorithm?
- Unknown observation structure?
- Is there a more general information-theoretic (as opposed to combinatorial) characterization of attainable regret?