# To Teach or not to Teach? Decision Making Under Uncertainty in Ad Hoc Teams


**PETER STONE**
Department of Computer Science
The University of Texas at Austin
pstone@cs.utexas.edu

**SARIT KRAUS**
Department of Computer Science, Bar-Ilan University
Institute for Advanced Computer Science
University of Maryland
sarit@cs.biu.ac.il


## Abstract


- Ad hoc teams: **collaboration without pre-coordination**
- Testbed scenario: **cooperative $k$-armed bandit**
- We characterize potentially optimal actions


## Teamwork


- Typical scenario: pre-coordination
  - People practice together
  - Robots given coordination languages, protocols
  - "Locker room agreement" [Stone & Veloso, '99]

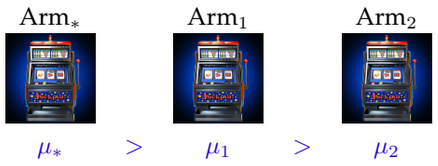
## Ad Hoc Teams

- Ad hoc team player is an individual
  - Unknown teammates (programmed by others)
- May or may not be able to communicate
- Teammates likely sub-optimal: no control


**Goal: Create a good team player**

- Minimal representative scenario
  - One teammate, no communication
  - Fixed and known behavior

## 3-armed bandit




- Agent A: teacher
  - Knows payoff distributions
  - Objective: maximize expected sum of payoffs
  - If alone, always Arm$_*$
- Agent B: learner
  - Can only pull Arm$_1$ or Arm$_2$
  - Selects arm with highest observed sample average

## Assumptions

- Alternate actions (teacher first)
- Results of all actions fully observable (to both)
- Number of rounds remaining finite, known to teacher

**Objective**: maximize expected sum of payoffs

## Formalism

- $\mu_i$: expected payoff of Arm$_i$ ($i \in \{1, 2, *\}$)
  - Assume $\mu_* > \mu_1 > \mu_2$: only interesting case
- $n_i$: number of times Arm$_i$ has been pulled
- $m_i$: cumulative payoff from past pulls of Arm$_i$
- $\bar{x_i} = \frac{m_i}{n_i}$: observed sample average so far
- $r$: number of rounds left

Which arm should the teacher pull, as a function of $r$ and all the $\mu_i$, $n_i$, and $\bar{x_i}$?


## Acknowledgements

Thanks to Yonatan Aumann, Vincent Conitzer, Reshef Meir, Daniel Stronger and members of the UT Austin Learning Agents Research Group (LARG) for helpful comments and suggestions. The research is supported in part by grants from NSF (IIS-0917122, 0705587), ONR (N00014-09-1-0658), DARPA (FA8650-08-C-7812), U.S. Army Lab (W911NF-08-1-0144), ISF (#1357/07), the FHWA (DTFH61-07-H-00030), and the Fulbright and Guggenheim Foundations.


## Teacher should consider Arm$_1$


$r = 1$

| $i$ | $m_i$ | $n_i$ | $\bar{x_i}$ |
|---|---|---|---|
| 1 | 6.0 | 1 | 6.0 |
| 2 | 7.0 | 1 | 7.0 |

- Teacher Arm$_1$ expected value (EV):
  - Define $\eta$: probability Arm$_1$ returns $> 8$
  - Assume: $\eta > \frac{1}{2}$
  - EV: $\mu_1 + \eta\mu_1 + (1-\eta)\mu_2 > 9 + \frac{9}{2} + \frac{5}{2} = \mathbf{16}$
- Teacher Arm$_*$ expected value:
  - EV: $\mu_* + \mu_2 = \mathbf{15}$

## Should teacher consider Arm$_2$?

- $\bar{x_1} < \bar{x_2} \implies$ **no**
  - Sequence of values from Arm$_2$: $u_0, u_1, u_2, \ldots$
  - Optimal from Arm$_2$: $u_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, u_0, \mu_*, a, b, c, d, e, \ldots, w, x$
- $\bar{x_1} > \bar{x_2} \implies$ **?**
  - Subtle, but still **no**
  - Challenge: prove it!

## Never teach when $\bar{x_1} > \bar{x_2}$

- Same proof
  - Sequence of values from Arm$_1$: $v_0, v_1, v_2, \ldots$
  - Optimal from Arm$_1$: $v_0, a, b, c, d, e, \ldots w, x, y, z$
  - Also possible: $\mu_*, v_0, \mu_*, a, b, c, d, e, \ldots, w, x$
- Only need to consider Arm$_1$ when $\bar{x_1} < \bar{x_2}$
  - Depends on distributions
  - Consider binary and normal

## Arms with Binary Distributions

$\longrightarrow \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases}$

$\mu_i = p_i \qquad m_i =$ number of 1's so far

## One round left: $r = 1$



$p_*$ > $p_1$ > $p_2$

- Consider teaching if:
  1. $\bar{x_1} < \bar{x_2} \quad \equiv \quad \frac{m_1}{n_1} < \frac{m_2}{n_2}$
  2. It could help: $\frac{m_1+1}{n_1+1} > \frac{m_2}{n_2}$
- Teacher Arm$_*$ expected value: $p_* + p_2$
- Teacher Arm$_1$ expected value: $p_1 + p_1 * p_1 + (1-p_1)p_2$

**Teach iff conditions 1, 2, and** $p_* - p_1 < p_1(p_1 - p_2)$

## Multiple Rounds Left: $r \geq 2$

- Polynomial algorithm finds optimal teacher action
  - Takes starting values $M_1, N_1, M_2, N_2$ and $R$
- Dynamic programming
  - Works backwards from $r = 1$
  - Considers all reachable values of $m_1, n_1, m_2, n_2$
- $O(r^5)$ in both memory and runtime
- Generalizes to arm payoffs with any discrete distribution

## Testing an Ad-hoc Team Player [AAAI'10]

**Evaluate**($a_0$, $a_1$, $A$, $D$)

- Initialize performance (reward) counters $r_0$ and $r_1$ for agents $a_0$ and $a_1$ respectively to $r_0 = r_1 = 0$.
- Repeat:
  - Sample a task $d$ from $D$.
  - Randomly draw a subset of agents $B$, $|B| \geq 2$, from $A$ such that $E[s(B, d)] \geq s_{min}$.
  - Randomly select one agent $b \in B$ to remove from the team to create the team $B^-$.
  - Increment $r_0$ by $s(\{a_0\} \cup B^-, d)$.
  - Increment $r_1$ by $s(\{a_1\} \cup B^-, d)$.
- If $r_0 > r_1$ then we conclude that $a_0$ is a better ad-hoc team player than $a_1$ in domain $D$ over the set of possible teammates $A$.

## Arms with Normal Distributions

$\longrightarrow$ N($\mu, \sigma$)

## One round left: $r = 1$



($\mu_*, \sigma_*$) > ($\mu_1, \sigma_1$) > ($\mu_2, \sigma_2$)

- Cost of teaching: $\mu_* - \mu_1$
- Benefit of teaching if successful: $\mu_1 - \mu_2$ ($\bar{x_1} < \bar{x_2}$)
- Probability it's successful: $1 - \Phi_{\mu_1,\sigma_1}(\bar{x_2}(n_1+1) - \bar{x_1}n_1)$
  - Cumulative prob. that pulling Arm$_1$ causes $\bar{x_1} > \bar{x_2}$

**Teach iff** $1 - \Phi_{\mu_1,\sigma_1}(\bar{x_2}(n_1+1) - \bar{x_1}n_1) > \frac{\mu_* - \mu_1}{\mu 1 - \mu_2}$

## Multiple Rounds Left: $r \geq 2$

- Can solve computationally — nested integral
- Not exactly, nor efficiently
- Is there an efficient algorithm?

## Experiments

- Evaluated teacher heuristics
  1. Never teach
  2. Teach iff $\bar{x_1} < \bar{x_2}$
  3. Teach iff it would be optimal to teach if $r = 1$
  - None dominates
- Looked for patterns in optimal action as a function of $r$
  - Conjecture: teach when $r = 1 \implies$ teach when $r = 2$
  - **False!** (binary and normal)

## More than 3 arms


- Additional arms for teacher make no difference
  - Ignore all but the best
- Additional learner arms: most results generalize naturally
  - Never teach with Arm$_z$ (Arm$_1$–Arm$_{z-1}$ possible)
  - Never teach with Arm$_i$ when $\bar{x}_i > \bar{x}_j, \forall j \neq i$
  - **Surprising:** May be best to teach with Arm$_j$ for $j > i$ (teach with Arm$_2$, even though $\bar{x}_1 > \bar{x}_2 > \bar{x}_3$)

## Open Questions

- What if the teacher doesn't know the distributions?
  - Exploration vs. exploitation vs. teaching
- What if the learner isn't greedy: explores on its own?
- How does this extend to the infinite (discounted) case?
- What if there are multiple learners?

## Related Work

(Full references are in the paper)

- Ad hoc human teams [Just, Kildare]
- Softare agents to support human teams [Tambe, Sycara]
- Teacher not embedded [Brafman, Tennenholtz]
- Environment design [Zhang]
- $k$-armed bandits [Lin, Kleinberg]
- Multiagent reinforcement learning [Claus, Schaerf]
- Iterated game theory [with Kaminka, Rosenschein]

| $M1$ | $b_0$ | $b_1$ | $b_2$ |
|---|---|---|---|
| $a_0$ | 25 | 1 | 0 |
| $a_1$ | 10 | 30 | 10 |
| $a_2$ | 0 | 33 | 40 |