(O'.')-o t('.'t)

Mike Lam


DO NOT TOUCH

A General Criterion and an Algorithmic Framework For Learning In Multi-Agent Systems (2006)

R. Powers Y. Shoham T. Vu

Overview

- Brief Introduction
- Some Previous Criteria
- New Criteria
- Algorithmic Framework
  - Stationary
  - Adaptive
- Conclusions
- Future work

Assumptions

- Each player knows all the actions and their reward after each round
- Each player interested in maximizing its reward
- All players know the structure and payoffs of the game at all times

Learning Focus

- Agent-centric: How to maximize rewards in an environment with other agents (that might be learning as well)
- Environment: Repeated stage game
- Learn while protecting against arbitrarily complex/strange and wrong guesses

Dictionary.com: “Learning”

- 1. knowledge acquired by systematic study in any field of scholarly application.
- 2. the act or process of acquiring knowledge or skill.
- 3. Psychology. the modification of behavior through practice, training, or experience.

Dictionary.com: “Learning”

- 1. knowledge acquired by systematic study in any field of scholarly application.
- 2. the act or process of acquiring knowledge or skill.
- 3. AI-ology . the modification of behavior through practice, training, or experience.

Learning Focus

- Agent-centric: How to maximize rewards in an environment with other agents (that might be learning as well)
- Environment: Repeated stage game
- Learn while protecting against arbitrarily complex/strange and wrong guesses

Learning Focus

- Best response against A: Strategy to use to maximize rewards when all other players use joint strategy A.
- Security value: Maximum reward guaranteed regardless of what policies the opponents use.

Previous Criteria

- Bowling and Veloso (2002)
- Rationality
  - If the other players' policies converge to stationary policies then the learning algorithm will converge to a stationary policy that is a best-response (in the stage game) to the other players' policies.
- Convergence
  - The learner will necessarily converge to a stationary policy.
- Self-play

Previous Criteria

- Fudenberg and Levine (1995)
- Safety
  - The learning rule must guarantee at least the minimax payoff of the game.
- Consistency
  - The learning rule must guarantee that it does at least as well as the best response to the empirical distribution of play when playing against an opponent whose play is governed by independent draws from any fixed distribution

Previous Criteria

- No regret
- Ex. GIGA-WoLF
- No focus on the agent's effect on opponent's future play – Tit-for-tat
- Richer notion of regret?

New Criteria

- Goal Oriented – High payoff vs target opponents while maintaining security guarantee vs other opponents
- Given:
  - n-player repeated game G
  - A history H
  - X designed players, Y target opponents, Z unconstrained
  - C a specification of playing strategy for Y players

New Criteria

- X-Enforceable Payoff Profiles
- ε-Pareto-efficient enforceable
- (ε-δ)-Guardedly optimal

- Formal criterion: Guarded Optimality
  - Given a class S of possible opponent strategies, an algorithm is guardedly optimal if for any choice of $\varepsilon > 0$, $\delta > 0$, and any n-player repeated game G, the algorithm is (ε, δ)-guardedly optimal for G given S.

New Criteria

- 2-player case:
- Targeted optimality
- Auto-compatibility
- Safety

Algorithmic Framework

- Modular design – we can design a cupcake to best both the hamster and Bart.

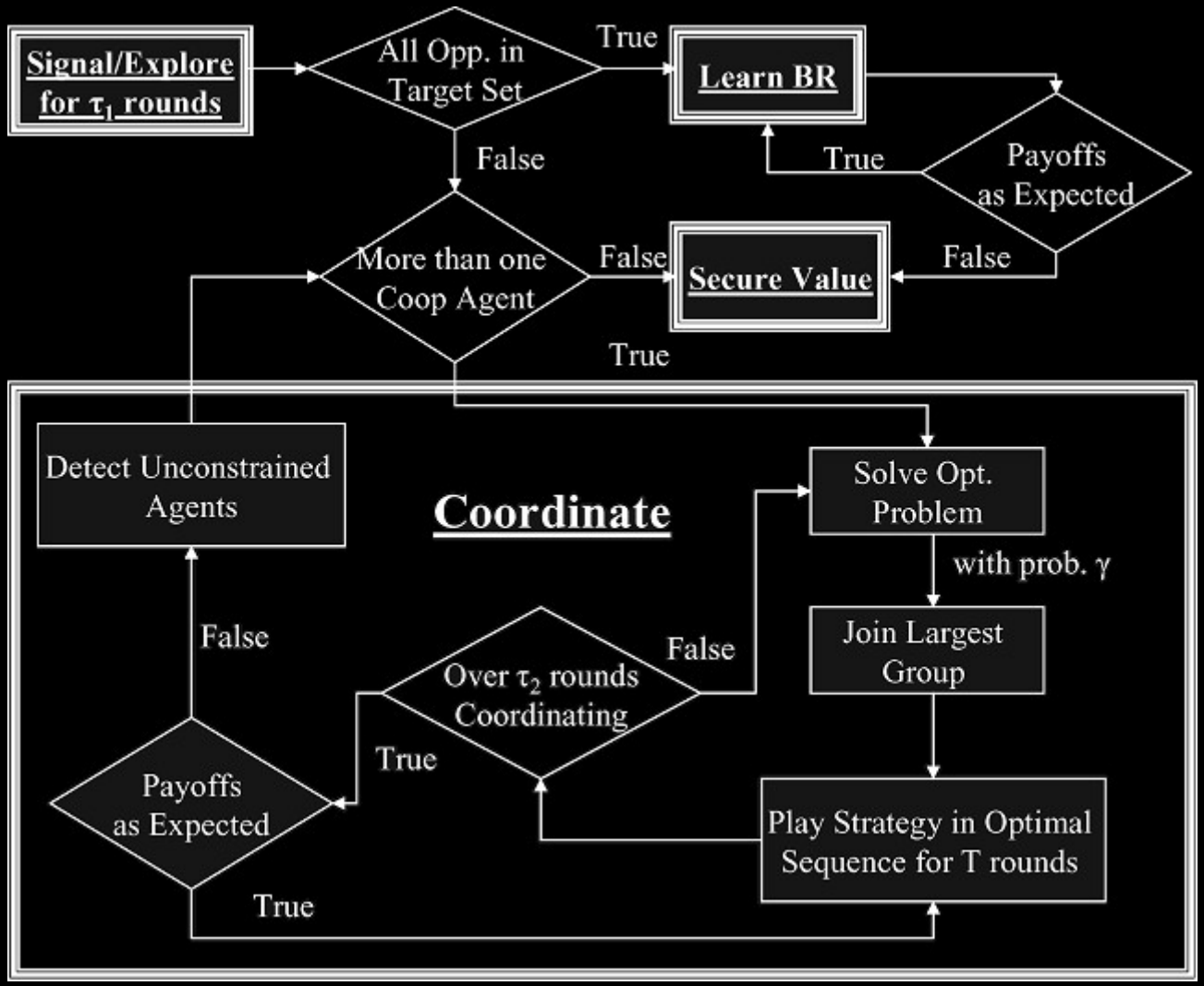
- Learn Best Response
- Coordinate
- Secure Value
- Signal
- Teach

Algorithmic Framework

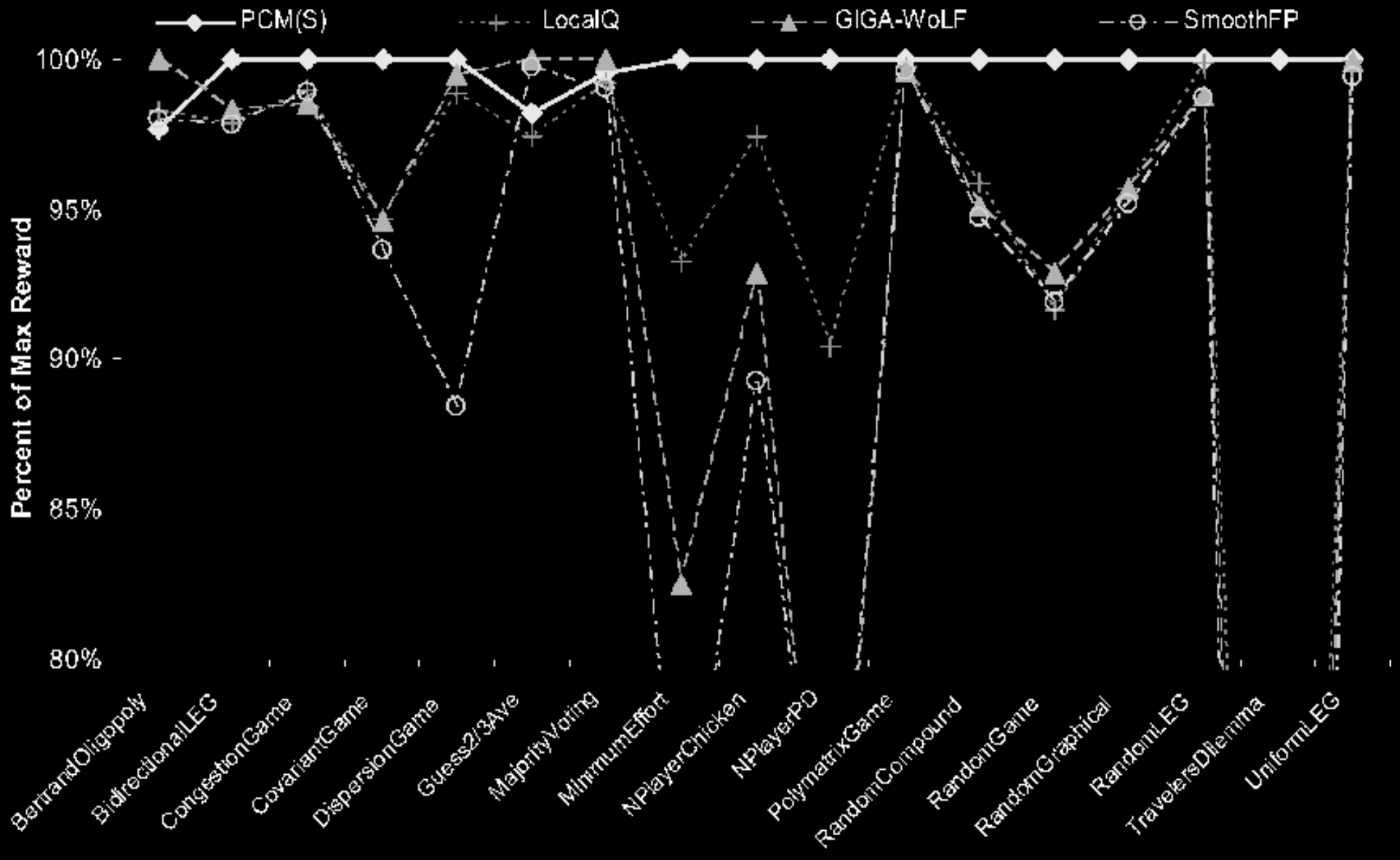
- PCMs – Partition, Coordinate, Monitor

- Stationary Opponents: PCM(S)
- Adaptive Opponents: PCM(A)

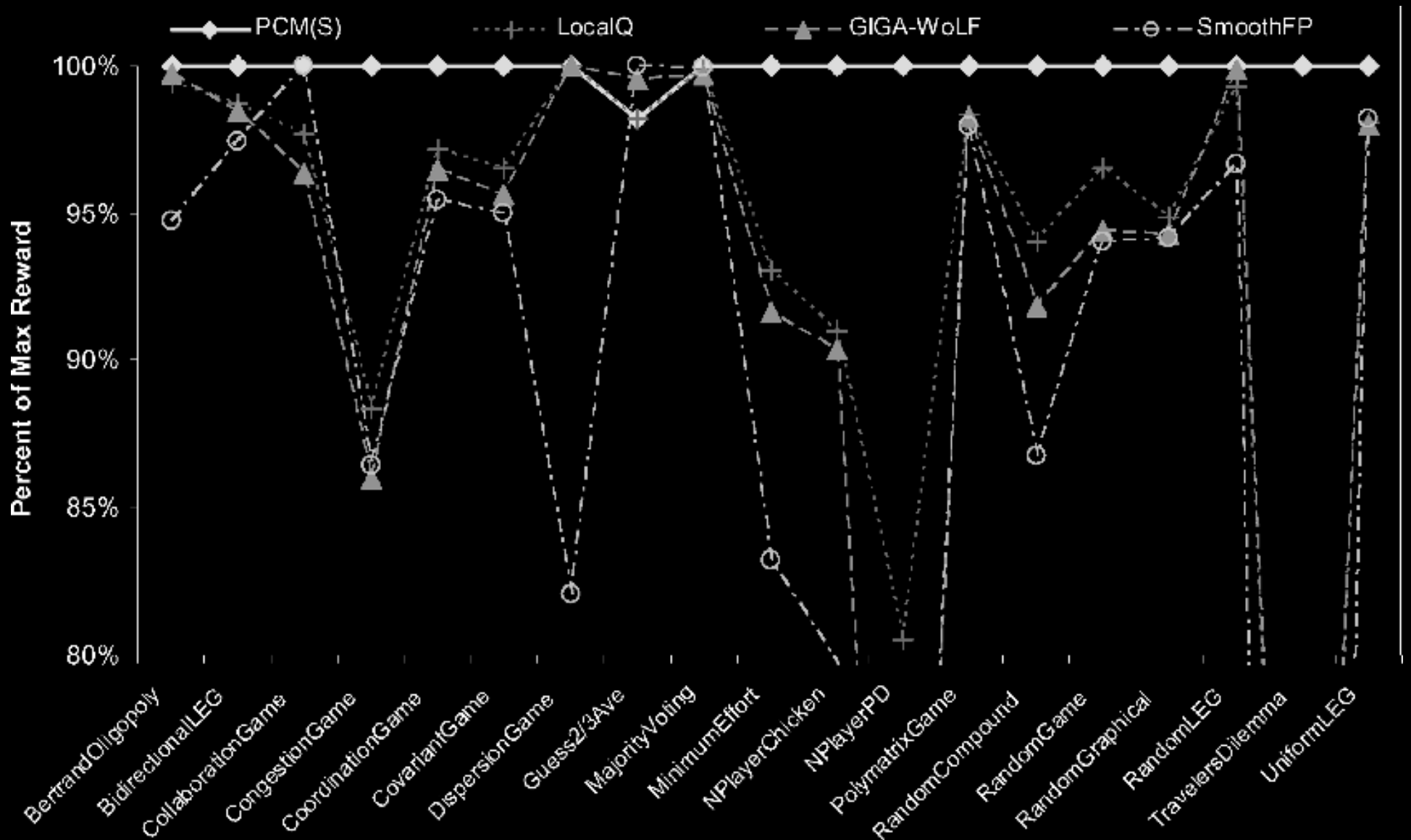
PCM(S)


Signal/Explore for τ_1 rounds
All Opp. in Target Set
True
Learn BR
False
True
Payoffs as Expected
More than one Coop Agent
False
Secure Value
False
True
Detect Unconstrained Agents
Coordinate
Solve Opt. Problem
with prob. γ
Join Largest Group
False
Over τ_2 rounds Coordinating
False
True
Payoffs as Expected
Play Strategy in Optimal Sequence for T rounds
True

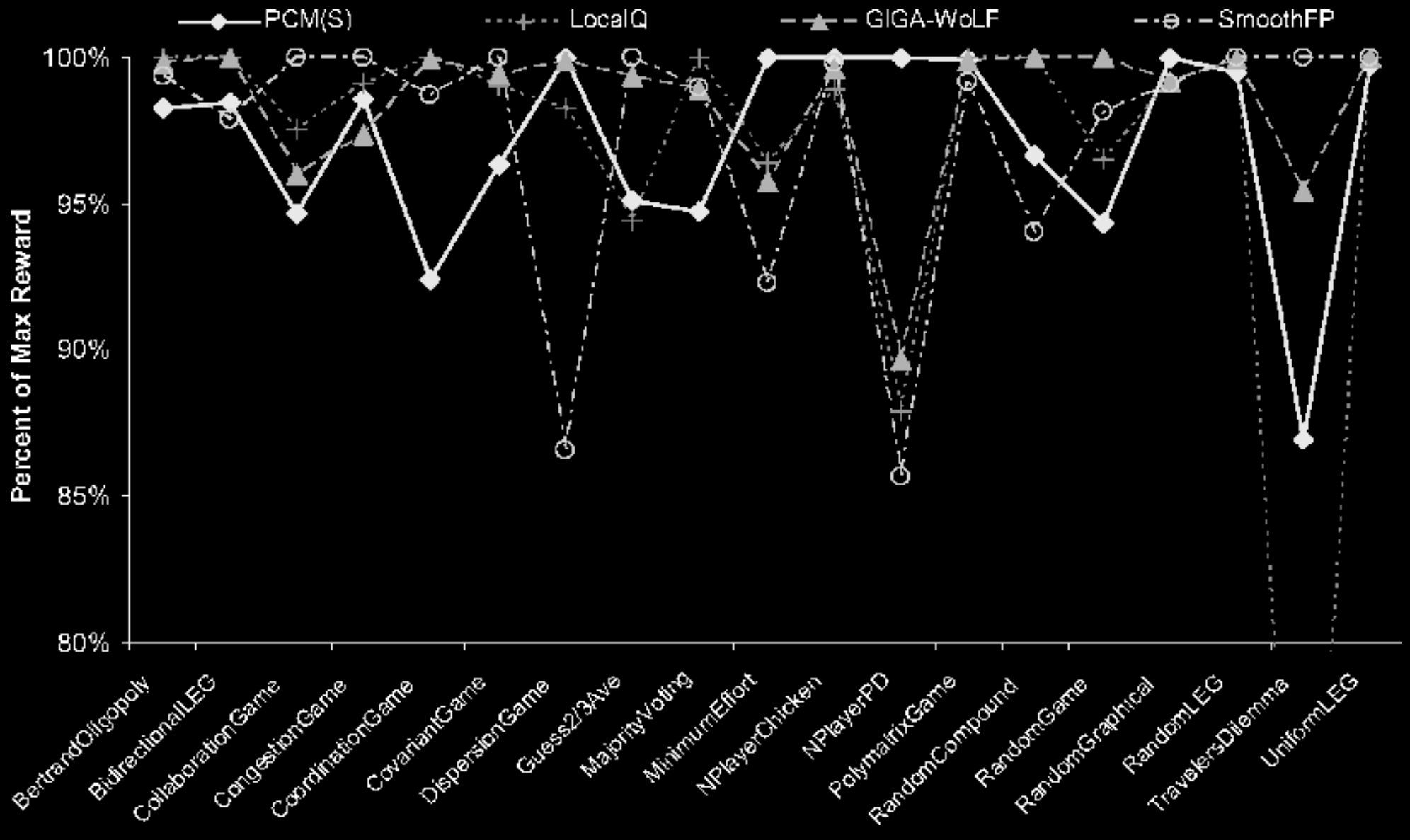
2v2


PCM(S)
LocalQ
GIGA-WoLF
SmoothFP
100%
95%
90%
85%
80%
Percent of Max Reward
BertrandOligopoly
BidirectionalLEG
CongestionGame
CovariantGame
DispersionGame
Guess2/3Ave
MajorityVoting
MinimumEffort
NPlayerChicken
NPlayerPD
PolymatrixGame
RandomCompound
RandomGame
RandomGraphical
RandomLEG
TravelersDilemma
UniformLEG

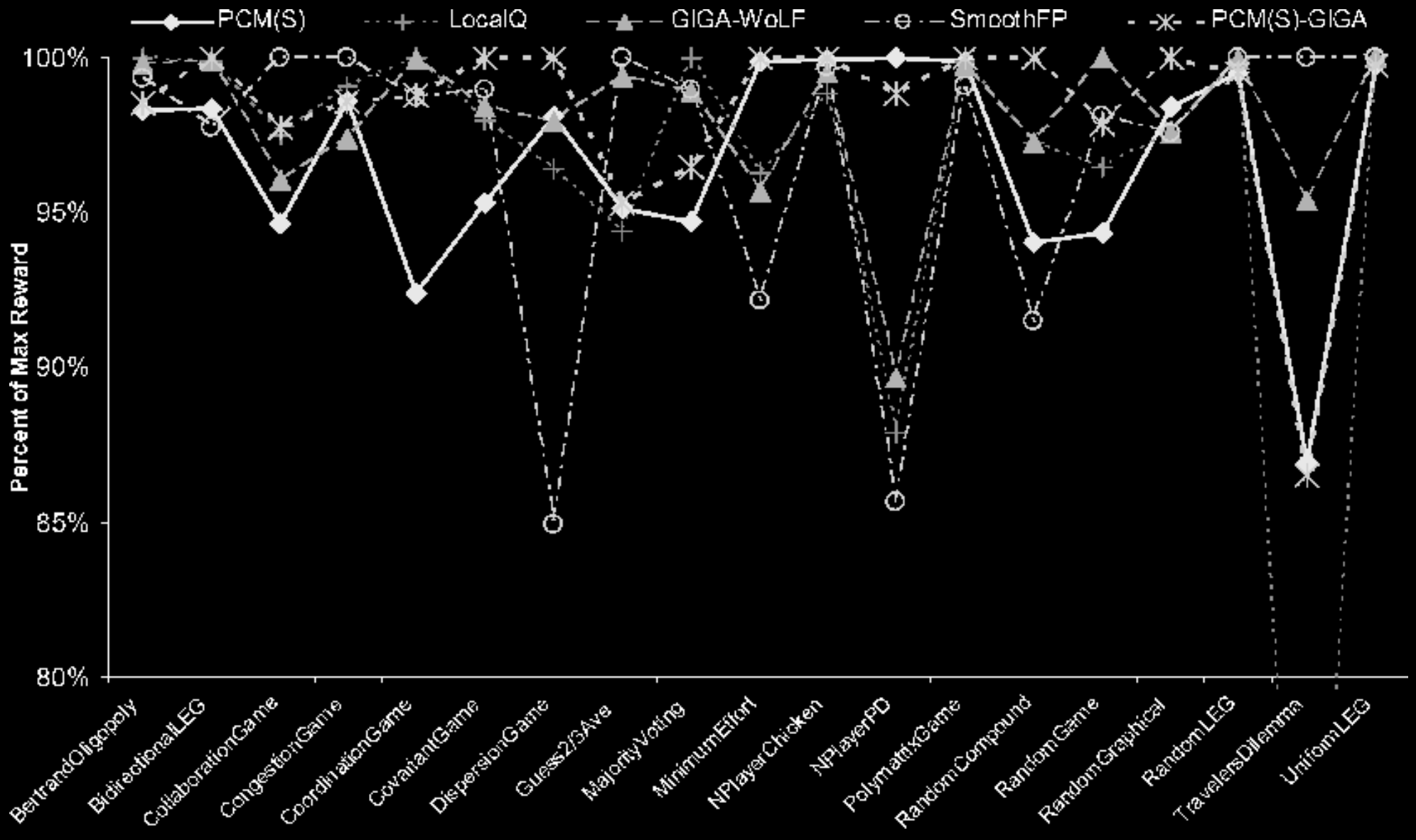
2v1


PCM(S)
LocalQ
GIGA-WoLF
SmoothFP
100%
95%
90%
85%
80%
Percent of Max Reward
BertrandOligopoly
BidirectionalLEG
CollaborationGame
CongestionGame
CoordinationGame
CovariantGame
DispersionGame
Guess2/3Ave
MajorityVoting
MinimumEffort
NPlayerChicken
NPlayerPD
PolymatrixGame
RandomCompound
RandomGame
RandomGraphical
RandomLEG
TravelersDilemma
UniformLEG

1v2


PCM(S)
LocalQ
GIGA-WoLF
SmoothFP
100%
95%
90%
85%
80%
Percent of Max Reward
BertrandOligopoly
BidirectionalLEG
CollaborationGame
CongestionGame
CoordinationGame
CovariantGame
DispersionGame
Guess2/3Ave
MajorityVoting
MinimumEffort
NPlayerChicken
NPlayerPD
PolymatrixGame
RandomCompound
RandomGame
RandomGraphical
RandomLEG
TravelersDilemma
UniformLEG

1v2 w/GIGA security



Self-play & Average Payoff

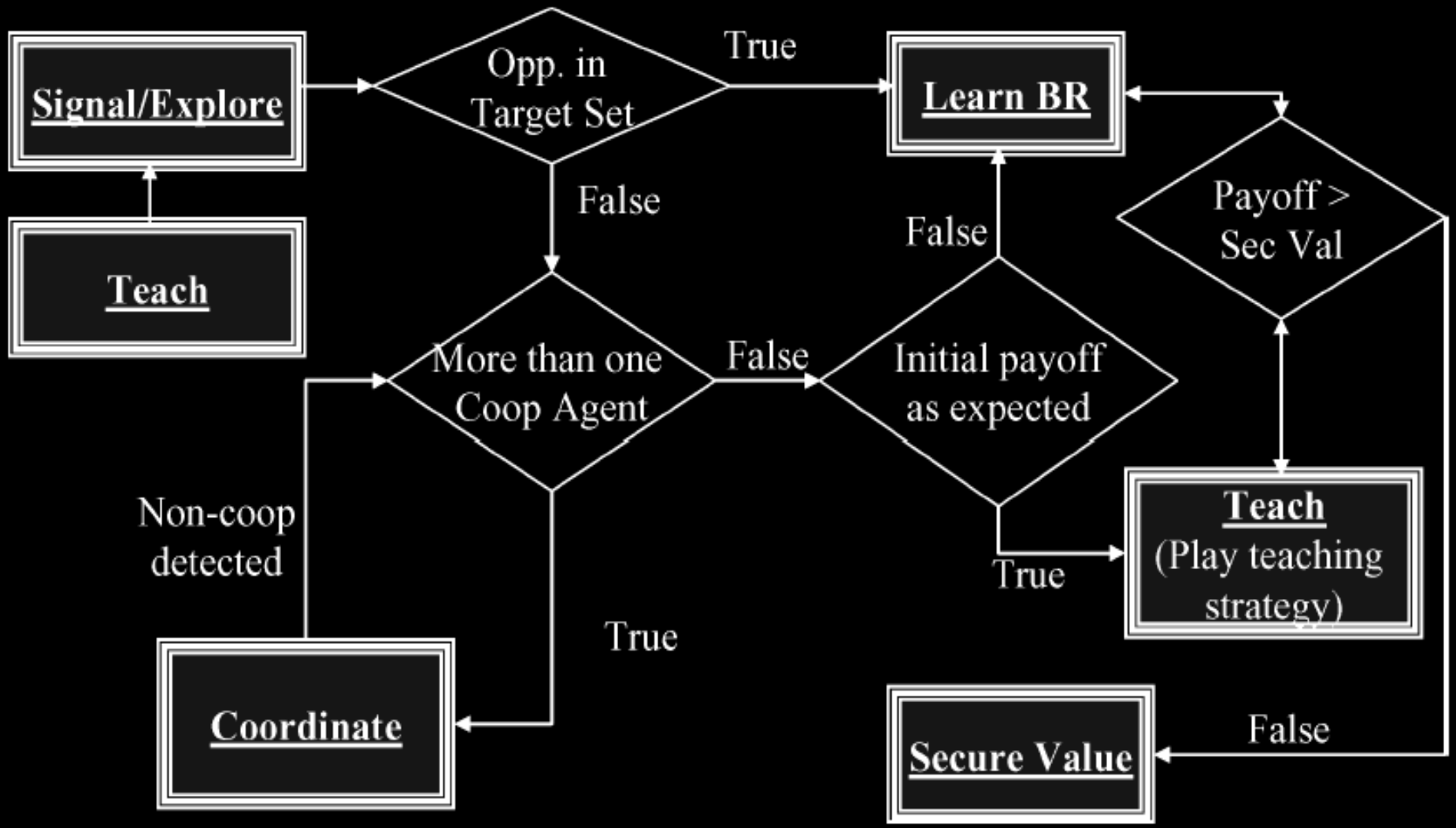
	$N=2$	$N=3$	$N=4$
PCM(S)	0.496	0.675	0.559
LocalQ	0.400	0.550	0.340
WoLF-PHC	0.389	0.449	0.292
StochIGA	0.385	0.422	0.257
GIGA-WoLF	0.374	0.411	0.255
SmoothFP	0.118	0.254	0.027
MiniMax	0.103	0.111	0.023

	5 K	10 K	25 K	50 K	100 K	200 K
PCM(S)	0.259	0.266	0.266	0.268	0.269	0.272
GIGA-WoLF	0.223	0.227	0.228	0.227	0.229	0.230

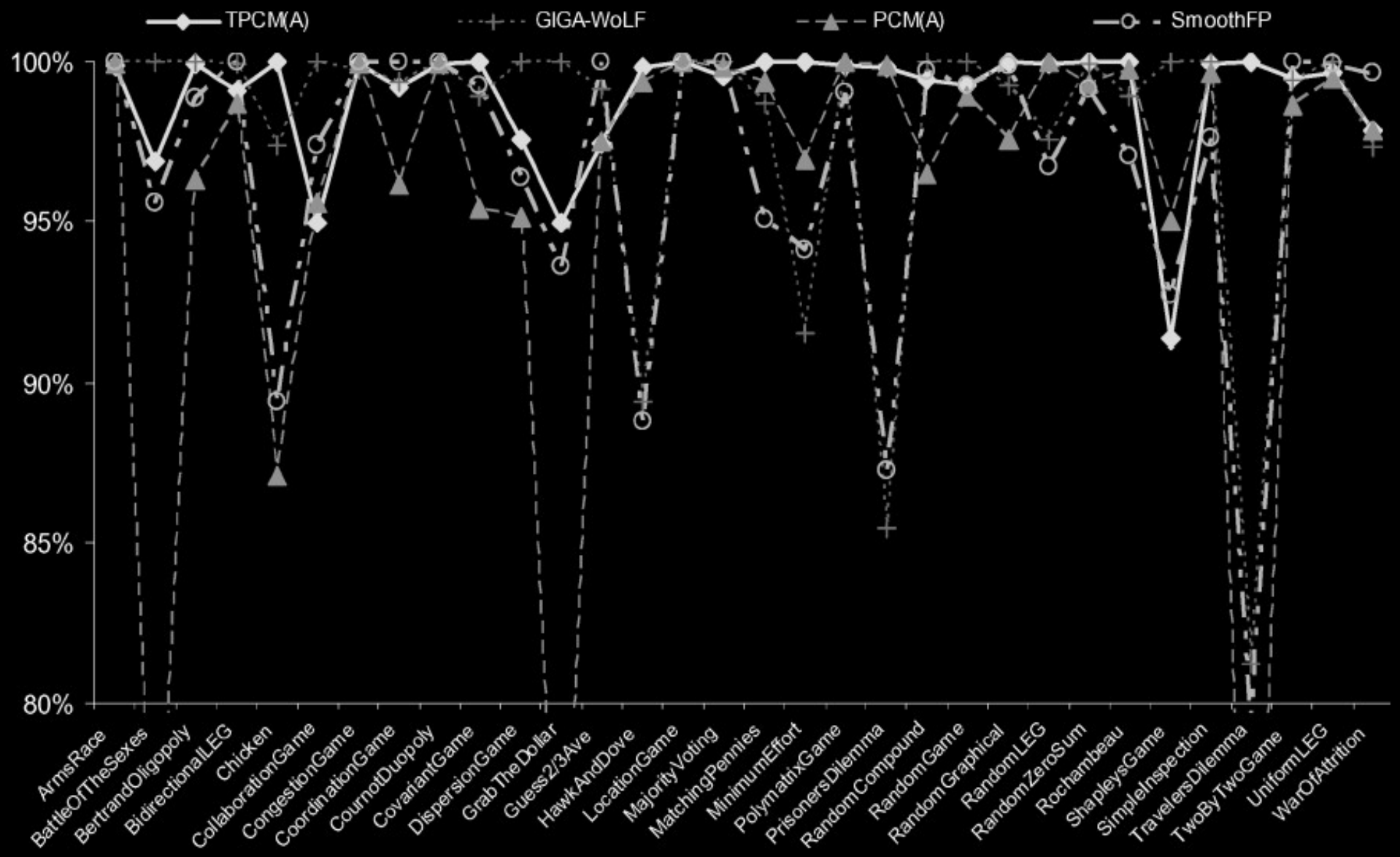
Adaptive Opponents

- Assume: Bounded recall (k turns)
- Same framework
- For 2 players, TPCM(A) – Teach, Partition, Coordinate, Monitor

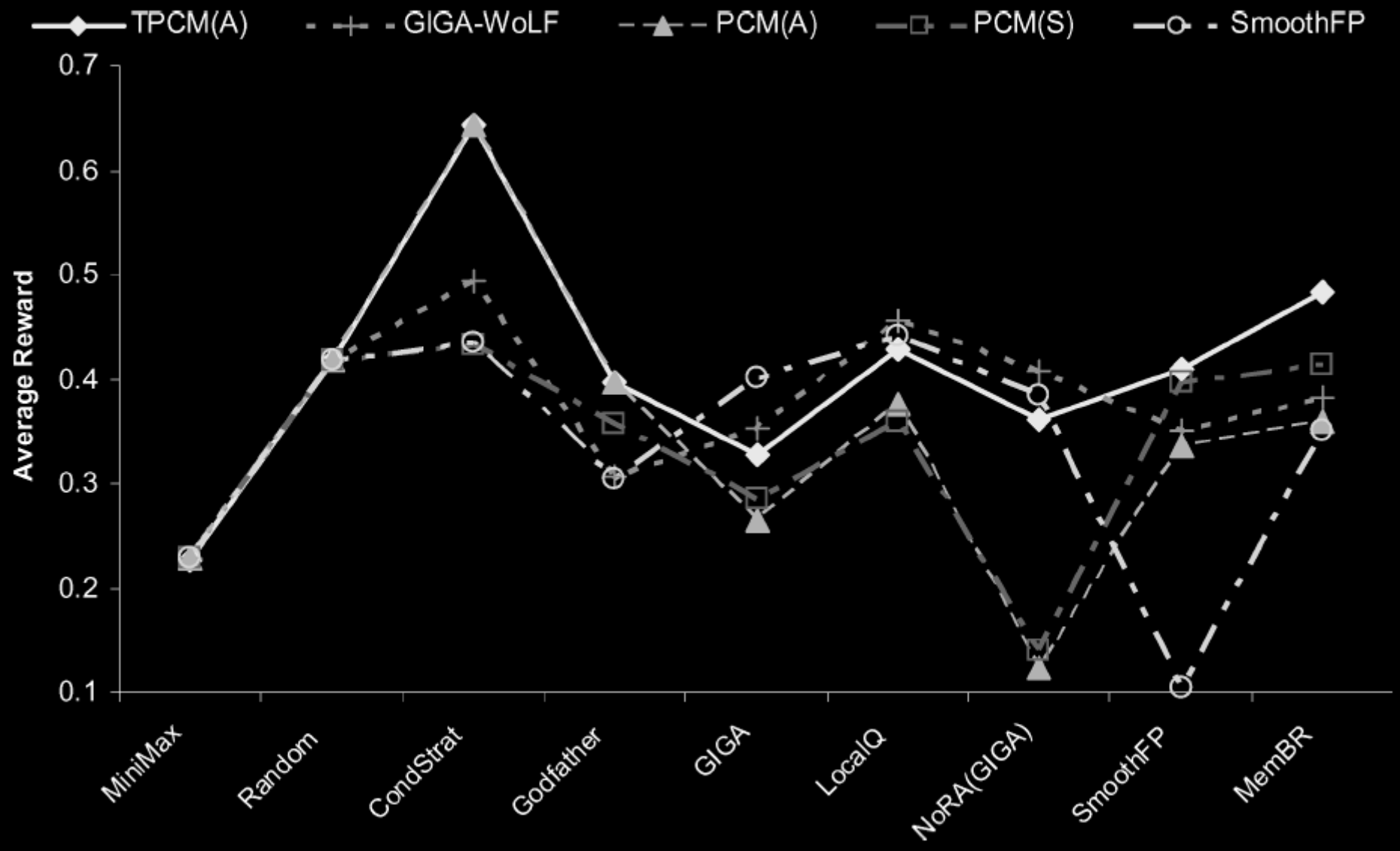
TPCM(A)


Signal/Explore
Opp. in Target Set
True
Learn BR
Teach
False
False
Payoff > Sec Val
More than one Coop Agent
False
Initial payoff as expected
Non-coop detected
Teach (Play teaching strategy)
True
True
Coordinate
Secure Value
False

2 Player Games + Teaching



Target Set = CondStrat



Summary

- New formal criterion for successful multi-agent learning, mainly Guarded Optimality
- Module based framework for learning algorithms
- Empirical results for PCM(S), PCM(A) and TPCM(A)

Future Work

- Different adaptive opponents
- Teaching multiple opponents
- Weaken constraints (full payoff knowledge)
- Stochastic games
- Discounted sum instead of average