Adaptively Perturbed Mirror Descent for Learning in Games

Atsushi Iwasaki
(University of Electro-Communications)
joint with
Kenshi Abe, Mitsuki Sakamoto, Kaito Ariu
(CyberAgent, Inc.)

International Workshop on Learning in Misspecified Models and Beyond
The University of Tokyo Market Design Center
February 2024

Summary

- This paper proposes a payoff perturbation technique for the Mirror Descent (MD) algorithm
- Existing algorithms typically find an equilibrium in an average sense (*average-iterate convergence*)
- Perturbing payoffs leads us to approximate an equilibrium (a stationary point)
  - The magnitude depends on the distance between current strategy and an anchoring or *slingshot* strategy
- Our Adaptively Perturbed MD updates the slingshot at an interval
  - Stationary points gradually get close to an exact equilibrium (*last-iterate convergence*)

Two-Person Zero-Sum Games

- Biased Rock-Paper-Scissor Game

$A =$

	R	P	S
R	0,0	-1,1	3,-3
P	1,-1	0,0	-1,1
S	-3,3	1,-1	0,0

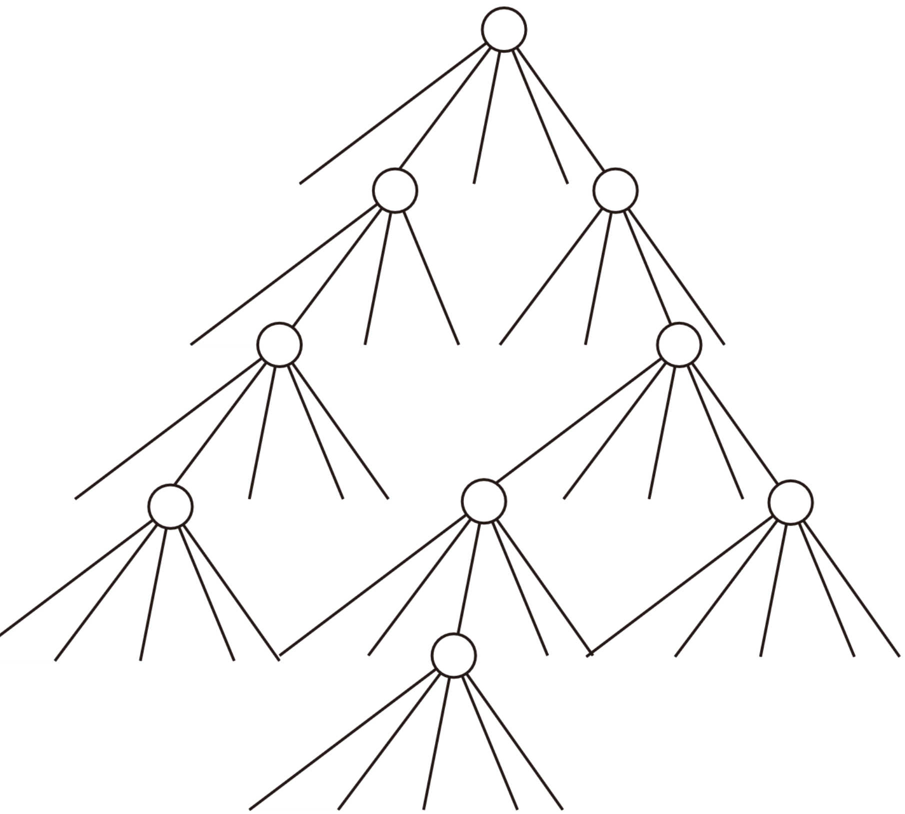
- Our work covers *N-player monotone games*, including Cournot competition

You (may) think nothing left

- Linear programming (LP) can solve all
- Player 1's strategy is obtained by solving
  - $\max_{\pi \in \Delta(X)} v$
  - $s.t. \ \sum_i \pi_i A_{ij} \geq v$ for each action j of Player 2
  - $\sum_i \pi_i = 1$
  - $\pi_i \geq 0$ for each action i of Player 1

Players doesn’t know everything

Large Setting



Can’t reason by the end

Online Setting

	1	2	…	10	…	100
1	0,0	1,-1	?,?	-1,1	?,?	?,?
2	?,?	?,?	0,0	?,?	1,-1	-1,1
…						
10	0,0	1,-1	?,?	-1,1	?,?	?,?
…						
100	?,?	?,?	0,0	?,?	1,-1	-1,1

Can’t know payoffs at the beginning

Dynamics for Learning in Games

- LP and minimax theorem frontiered learning dynamics
  - Players choose their actions with a simple procedure
  - They observe the outcomes and learn the next actions
- Possibility of online learning techniques
  - (Un)Constrained optimization
  - Robustness to adversarial environments
  - Convergence at faster rate
- *No-regret learning* has been emerged
  - Associates the consequences with equilibrium concepts

No-regret Learning

- Compared to LP, the advantage lies in the simplicity
  - Follow-The-Regularized-Leader (FTRL)
  - Mirror Descent (MD)
- MD is quite different from FTRL, but sometimes equivalent
  - If the regularizer is entropy, both becomes Multiplicative Weights Update (MWU)
- This talk concentrates on MD, but the same holds on FTRL

Mirror Descent

[Nemirovskij & Yudin, 1983; Beck & Teboulle, 2003]

- A class of algorithms for online convex optimization

Make strategies with higher expected values more likely

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

Next strategy

Don't move too far away from current strategy

- $D_\psi(\pi_i, \pi_i')$: Bregman divergence with strongly convex function ψ

Multiplicative Weights Update (MWU)

- MD with entropy regularizer
  - Bregman divergence: $D_\psi(x, \pi^t) = \sum_i^N D_\psi(x_i, \pi_i^t)$
  - Let $\psi(\pi_i') = \sum_j \pi_{ij}' \ln \pi_{ij}'$ where $\pi_i' = x_i$ or π_i^t

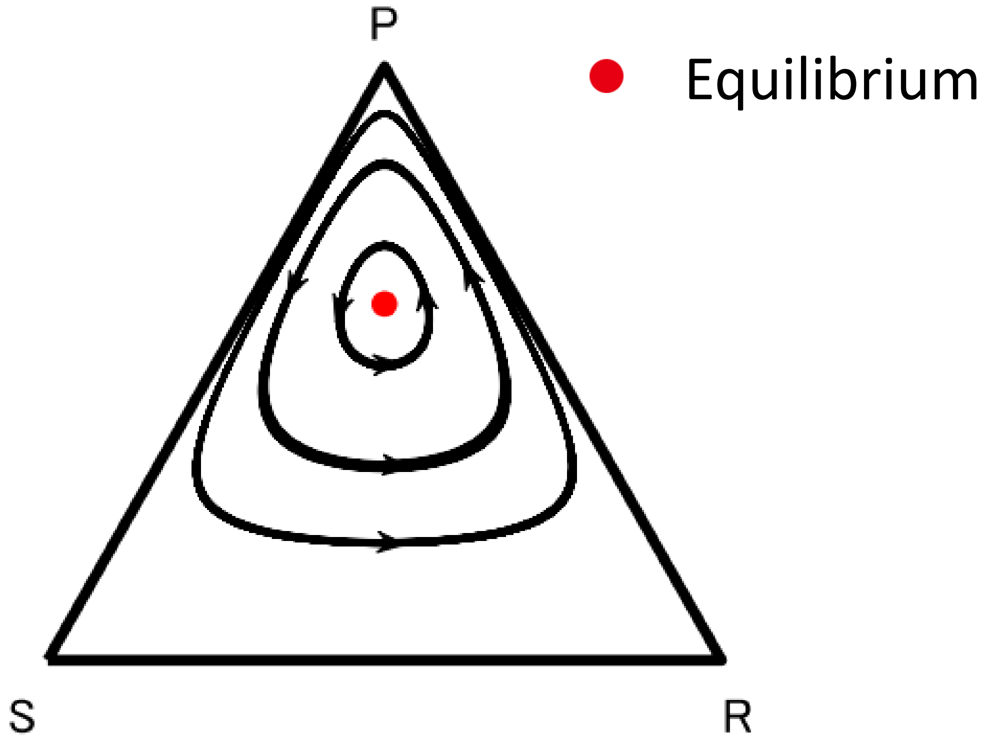
$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t), x \right\rangle - \sum_j (x_{ij} \ln \frac{x_{ij}}{\pi_{ij}^t}) \right\}$$

- Fast convergence
- (Coarse) correlated equilibrium in general-sum games

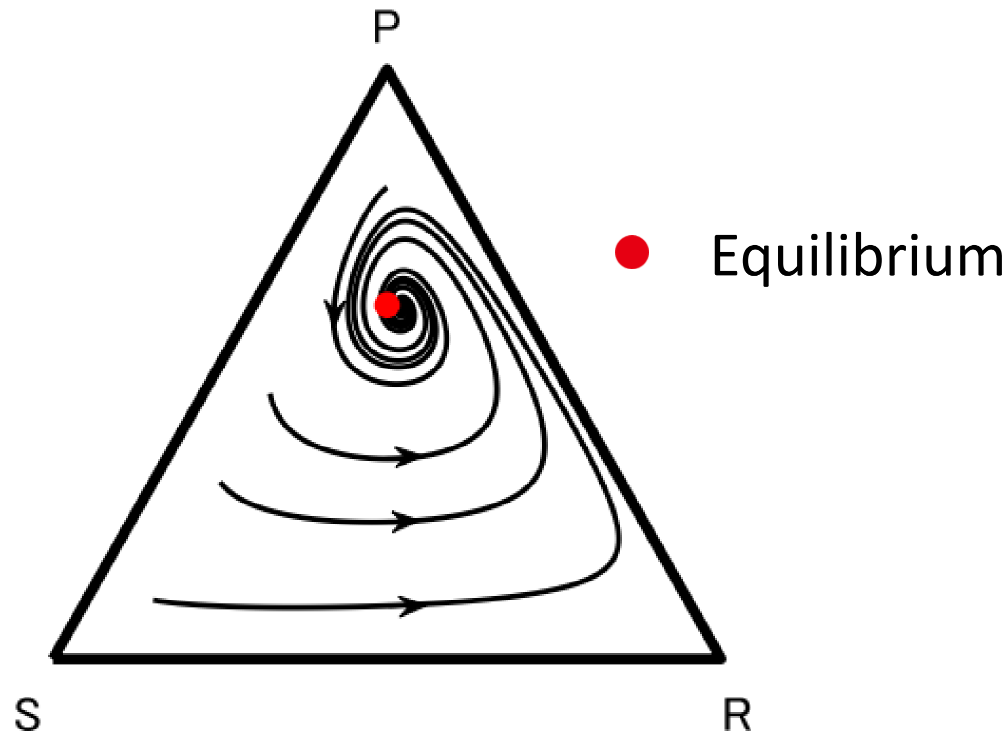
MWU enters a limit cycle

- Average-Iterate $\frac{1}{t}\sum_t \pi_i^t$ converges to an equilibrium as $t \to \infty$



Aim of this work is

- Let last-iterate π_i^t converge to an equilibrium

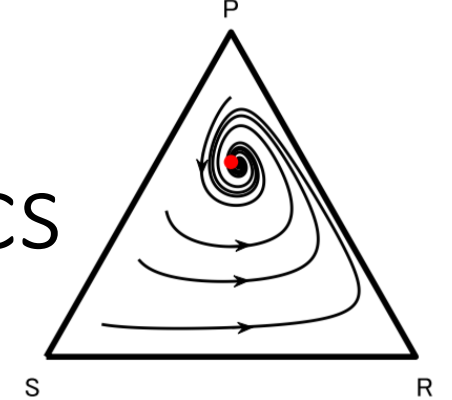


- Optimistic family is the central of the recent success [Daskalakis et al., 2018; Daskalakis & Panageas, 2019; Mertikopoulos et al., 2019]
  - Recency bias: the outcome of the second last-iterate is outweighed

Perturbation approach

- Instead of recency bias, we perturb the expected payoff vector [Perolat et al. 2021, Liu et al. 2023, Abe et al. 2022, 2023]
- This idea is analogue to mutate actions
  - Players may mistakenly choose a different action from the one they intended
- MWU is equivalent to replicator dynamics, assuming continuous time
  - $\dot{x}_j = x_j \left(f_j(x) - \phi(x) \right)$
- Introducing mutation makes dynamics likely to converge to a stationary point

Replicator-Mutator Dynamics



- Mutation stabilizes learning dynamics [Bauer et al. 2019]

$$\dot{x}_j = x_j\left(f_j(x) - \phi(x)\right) - \mu x_j + \frac{1}{n}(\mu x_1 + \cdots + \mu x_n)$$
$$= x_j\left(f_j(x) - \phi(x)\right) + \mu\left(\frac{1}{n} - x_j\right)$$

  - n: Number of strategies
- After producing strategy j, with probability μ, it mutates to others with equal probability
- Special case of *Mutant MWU* [Abe et al. AISTATS 2023]
  - Guaranteed to last-iterately converge to a 2μ-Nash equilibrium

Perturbed Mirror Descent with Uniform Distribution

- Let us perturb MD, along with RMD

Make strategies with higher expected values more likely

Perturbation Strength

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G\left(\pi_i^t, \frac{1}{n}\right), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

Next strategy

Perturbation Function

Don’t move too far away from current strategy

Perturbed MD with slingshot σ_i

- Let σ_i be a slingshot strategy, generalizing the uniform strategy $\frac{1}{n}$

Perturbation term

current strategy π_i^t gets close to σ_i

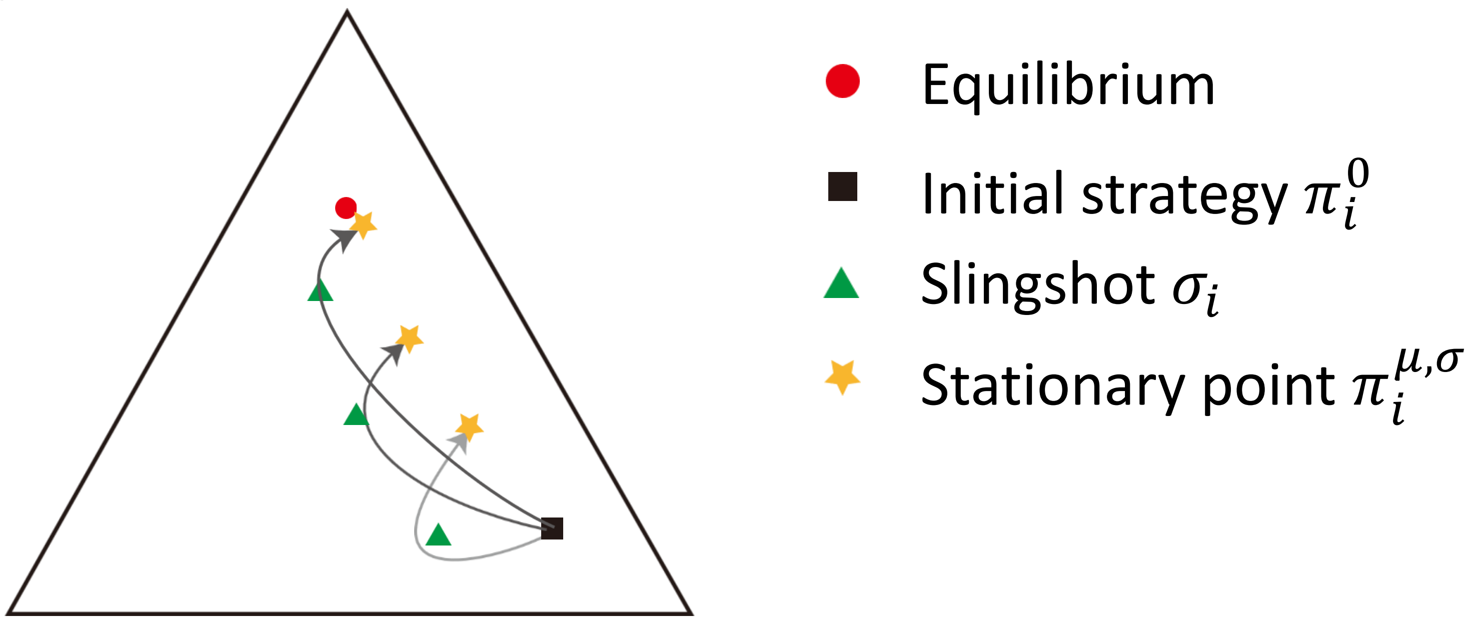
$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

- Current strategy converges to a stationary point that balances the payoff gradient with the perturbation term

$$\widehat{\nabla}_{\pi_i} v_i(\pi^t) \approx \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i)$$

Observation

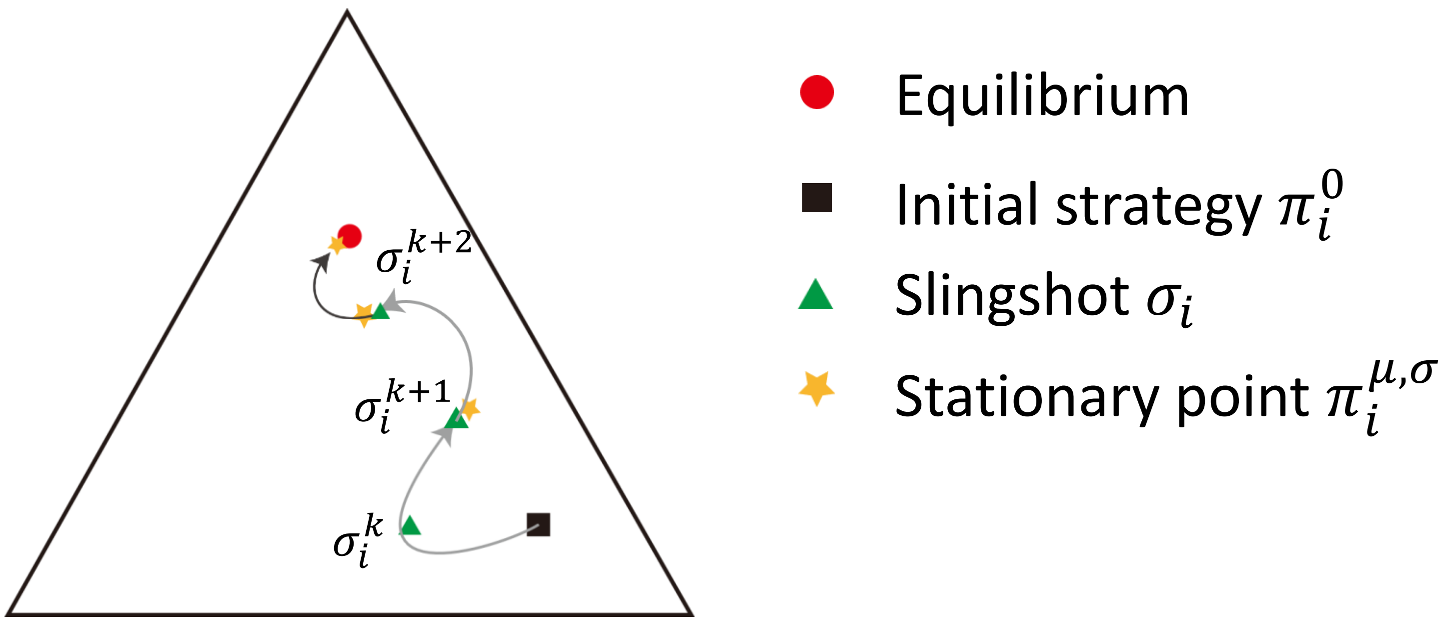
- Different slingshot leads to different stationary point



- As a slingshot gets close to an equilibrium, so does the stationary point.

Intuitive Idea

- Update slingshot at a predefined interval



- Slingshot σ^k is overrode by approximating π^{μ,σ^k}
- The sequence gradually goes to an equilibrium

Adaptively Perturbed MD

- Slingshot is updated at a predefined interval T_σ
  - Let σ_i^k be slingshot updated $k = \left\lfloor \frac{t}{T_\sigma} \right\rfloor$ times for each iteration t.

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \boxed{\sigma_i^k}), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

- π_i^{t+1} approximates the stationary point π^{μ,σ_i^k} during T_σ
- Update the slingshot σ_i^k to $\sigma_i^{k+1} = \pi^{\mu,\sigma^k}$

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \boxed{\sigma_i^{k+1}}), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

- The procedure is repeated T iterations
- We will argue how π_i^T gets close to an equilibrium

Further Notions for APMD

Make strategies
with higher expected
values more likely

Perturbation term

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i^k), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

Next
strategy

- Squared ℓ^2-distance on G and D_ψ
- Feedback types: Full or Noisy
  - Gradient of payoff vector may have noise
- Metric: GAP function

Squared ℓ^2 distance

Perturbation term

Regularization term

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i^k), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

Next strategy

- Perturbation function $G(\pi_i^t, \sigma_i^k) = \frac{1}{2} \| \pi_i^t - \sigma_i^k \|^2$
- Regularizer $D_\psi(\pi_i^t, x)$ where $\psi(\pi_i^t, x) = \frac{1}{2} \| \pi_i^t - x \|^2$
- Note that our results are extend beyond.

Full and Noisy Feedback

Make strategies with higher expected values more likely

$$\pi_i^{t+1} = \arg\max_{x \in \mathcal{X}_i} \left\{ \eta_t \left\langle \widehat{\nabla}_{\pi_i} v_i(\pi^t) - \mu \nabla_{\pi_i} G(\pi_i^t, \sigma_i^k), x \right\rangle - D_\psi(x, \pi_i^t) \right\}$$

Next strategy

- Full feedback: $\widehat{\nabla}_{\pi_i} v_i(\pi_i^t, \pi_{-i}^t) = \nabla_{\pi_i} v_i(\pi_i^t, \pi_{-i}^t)$
- Noisy feedback: $\widehat{\nabla}_{\pi_i} v_i(\pi_i^t, \pi_{-i}^t) = \nabla_{\pi_i} v_i(\pi_i^t, \pi_{-i}^t) + \xi_i^t$
- $\xi_i^t \in \mathbb{R}^{d_i}$ has zero-mean and its variance is bounded

Gap Function

- A strategy profile π^* is a Nash equilibrium iff
  - $\forall i \in [N], \forall \pi_i \in \mathrm{X}_i, v_i(\pi_i^*, \pi_{-i}^*) \geq v_i(\pi_i, \pi_{-i}^*)$
- A metric of the distance current strategy π and a Nash equilibrium
- Given π,

$$\mathrm{GAP}(\pi) := \max_{\tilde{\pi} \in \mathcal{X}} \sum_{i=1}^{N} \langle \nabla_{\pi_i} v_i(\pi_i, \pi_{-i}), \tilde{\pi}_i - \pi_i \rangle$$

- How much π is improvable by unilateral deviation

Convergence Rate under Full Feedback

- Given last iteration T and update interval T_σ,

Theorem 4.1. *If we use the constant learning rate* $\eta_t = \eta \in (0, \frac{2\mu\rho^2}{3\mu^2\rho^2+8L^2})$, *and set* D_ψ *and* G *as the squared* ℓ^2-*distance* $D_\psi(\pi_i, \pi_i') = G(\pi_i, \pi_i') = \frac{1}{2}\|\pi_i - \pi_i'\|^2$, *and set* $T_\sigma = \Theta(\ln T)$, *then the strategy* π^T *updated by APMD satisfies:*

$$\mathrm{GAP}(\pi^T) = \mathcal{O}\left(\frac{\ln T}{\sqrt{T}}\right).$$

- Last-iterate π^T has the bounded GAP on T

Convergence Rate under Noisy Feedback

- Given last iteration T and update interval T_σ,

Theorem 4.5. *Let* $\theta = \frac{3\mu^2\rho^2+8L^2}{2\mu\rho^2}$ *and* $\kappa = \frac{\mu}{2}$. *Assume that* D_ψ *and* G *are set as the squared* ℓ^2*-distance* $D_\psi(\pi_i, \pi_i') = G(\pi_i, \pi_i') = \frac{1}{2}\|\pi_i - \pi_i'\|^2$, *and* $T_\sigma = \Theta(T^{4/5})$. *If we choose the learning rate sequence of the form* $\eta_t = 1/(\kappa(t - T_\sigma \cdot \lfloor t/T_\sigma \rfloor) + 2\theta)$, *then the strategy* π^T *updated by APMD satisfies:*

$$\mathbb{E}\left[\mathrm{GAP}(\pi^T)\right] = \mathcal{O}\left(\frac{\ln T}{T^{\frac{1}{10}}}\right).$$

- Learning rate depends on iteration t to prevent noise from leading dynamics to a wrong stationary point

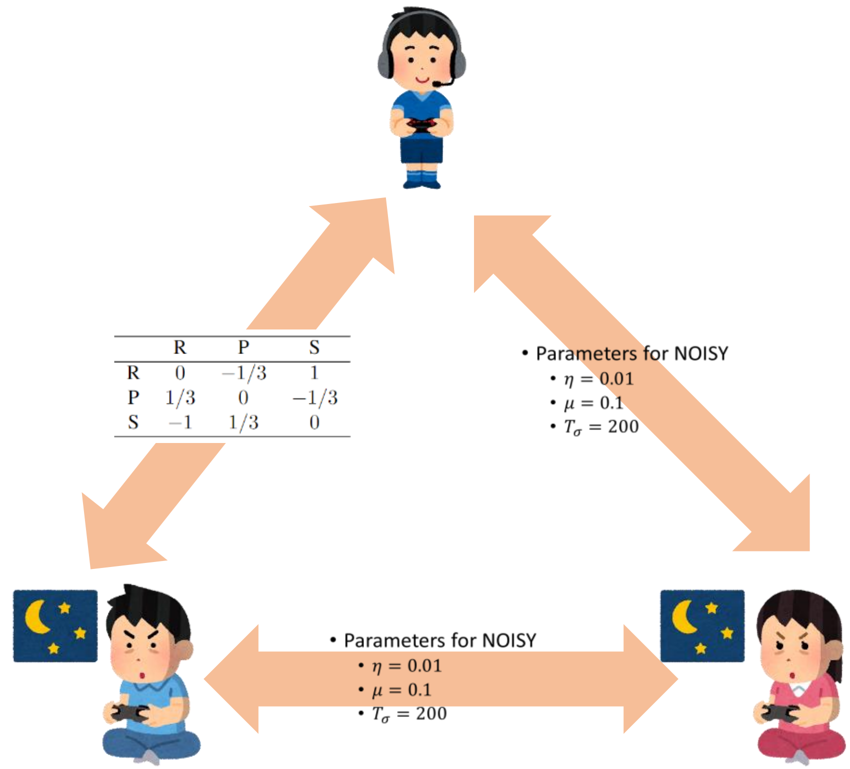
$$\mathrm{GAP}(\pi) := \max_{\tilde{\pi} \in \mathcal{X}} \sum_{i=1}^{N} \langle \nabla_{\pi_i} v_i(\pi_i, \pi_{-i}), \tilde{\pi}_i - \pi_i \rangle$$

Proof Sketch

- Convergence to a stationary point is straightforward
- Derive the upper bound of $GAP(\sigma^{k+1})$ for an arbitrary k
  - Cannot directly be bounded between current and the next strategy
- We decompose the gap using stationary point into three terms
  - One term is bounded by Cai's lemma [Cai et al. 2022]
  - The other two is done by Cauchy-Schwarz inequality

Experiments 1

- Three Player Biased RPS game
- Each player simultaneously joins two BRPS with two other players
- Parameters for FULL
    - $\eta = 0.1$
    - $\mu = 0.1$
    - $T_\sigma = 20$

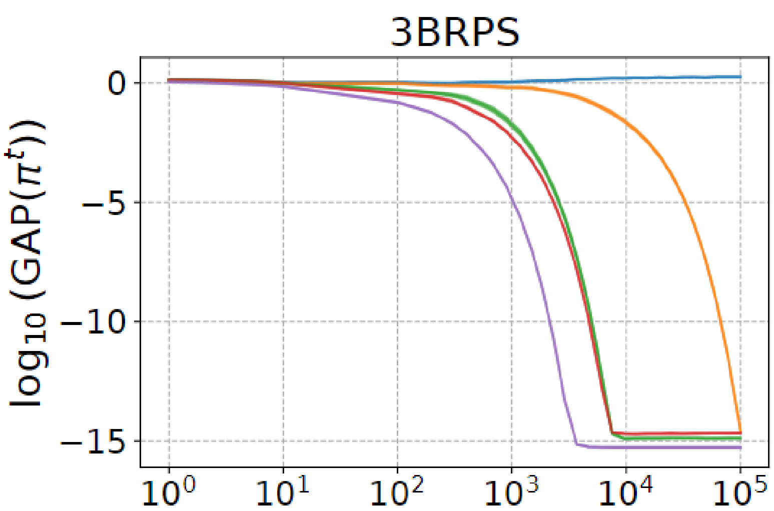


- Parameters for NOISY
    - $\eta = 0.01$
    - $\mu = 0.1$
    - $T_\sigma = 200$

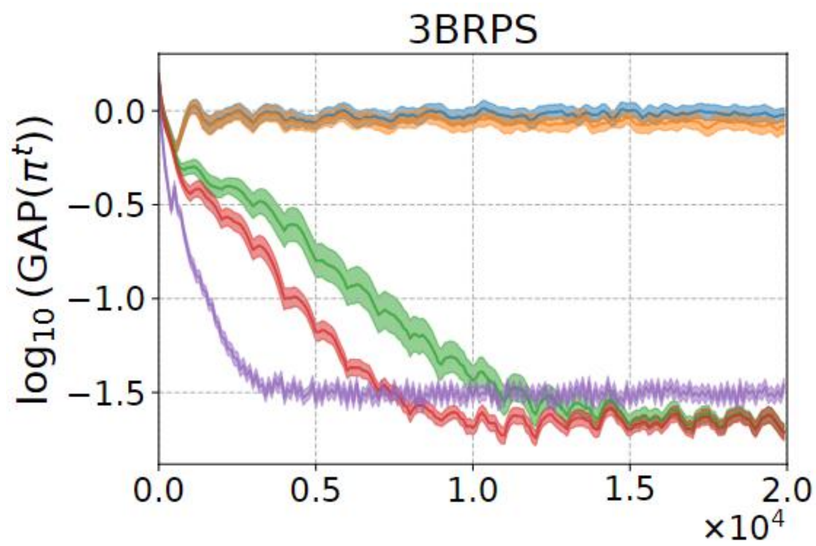
GAP values

APMD with $\mu = 1.0$ and $G = D_\psi = \ell^2$ is sufficiently competitive

Full Feedback



Noisy Feedback



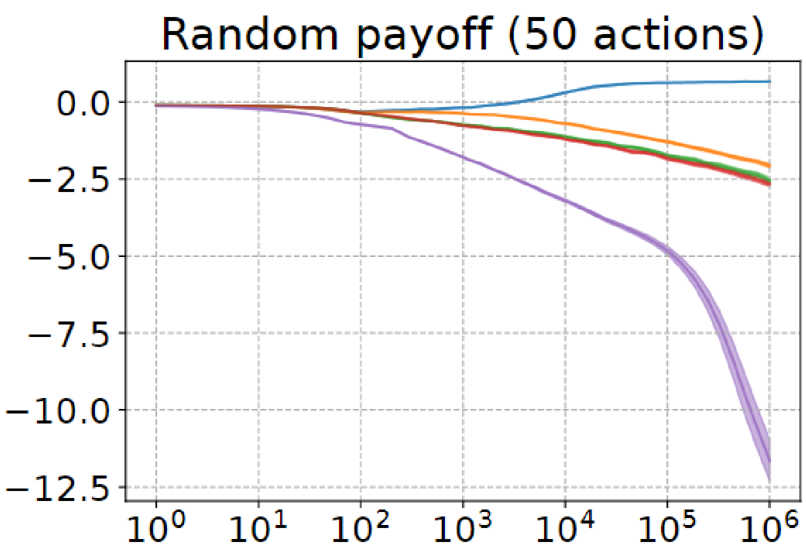
Experiments 2

- Three-Player random payoff games with 50 actions
- Each player i participates in two instances of the game with two other players j simultaneously
- The payoff matrix of each instance is drawn from the uniform distribution
  - Each payoff has the interval of $[-1,1]$
- Full feedback: $\eta = 0.01, \mu = 1.0, T_\sigma = 200$
- Noisy feedback: $\eta = 0.001, \mu = 1.0, T_\sigma = 2000$

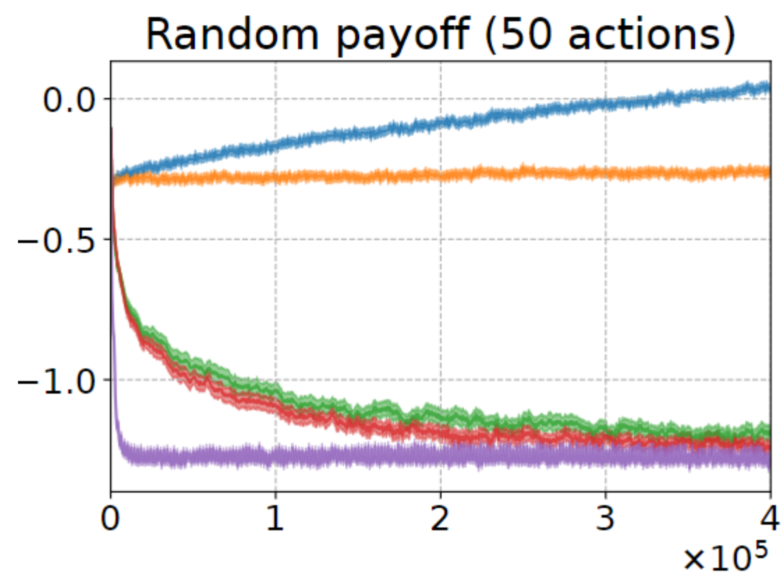
GAP values

APMD with $\mu = 1.0$ and $G = D_\psi = \ell^2$ is sufficiently competitive

Full Feedback



Noisy Feedback



MWU
OMWU
APMD $\mu = 0.1$ D_ψ=KL G=KL
APMD $\mu = 0.1$ D_ψ=KL G=RKL
APMD $\mu = 1.0$ D_ψ=L2 G=L2

Conclusions

- This paper proposes a novel variant of mirror descent (APMD) that achieves last-iterate convergence even when the noise is present
- The adaptive adjust of the perturbation magnitude enables us to bound the gap of values in each iteration
- APMD outperforms optimistic MWU and is competitive against the existing state-of-the-art algorithms
    - E.g., Perolat et al. 2021
- Future works
    - Extensive-form games, Markov games, Mean field games and so on
    - Asymmetric learning