ICML | 2022
Thirty-ninth International Conference on Machine Learning

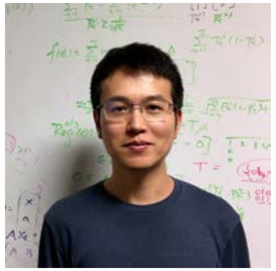
南京大學 NANJING UNIVERSITY

USC University of Southern California

No-Regret Learning in Time-Varying Zero-Sum Games

Peng Zhao[1]

joint work with


Mengxiao Zhang[2] Haipeng Luo[2] Zhi-Hua Zhou[1]

1. Nanjing University; 2. University of Southern California

Introduction

- Uncoupled learning dynamics for *a fixed game* is well studied.

- What if the game is ***changing***?
  - in some cases, changes are due to the other players' decisions
  - in other cases, changes may come from the environmental factors

morning rush-hour traffic

modern air combat

Our Contributions

- **Focus**: uncoupled learning over a sequence of ***time-varying*** zero-sum games decided exogenously by the environments.

First part: how to *measure the performance*?

- review an existing measure (and argue why it is problematic)
- consider/propose three natural measures (one is new)

Second part: propose *a single algorithm* that

- is parameter-free (i.e., no need prior info. on environments)
- achieves strong guarantees under all three measures
- recovers best known results when the game is fixed

Time-Varying Zero-Sum Games

For each round $t = 1, \ldots, T$:

- environment decides a payoff matrix $A_t \in [-1, 1]^{m \times n}$;
- without knowing A_t, x-player decides a mixed strategy $x_t \in \Delta_m$ and y-player decides a mixed strategy $y_t \in \Delta_n$;
- x-player suffers loss $x_t^\top A_t y_t$ and observes $A_t y_t$, while y-player receives reward $x_t^\top A_t y_t$ and observes $x_t^\top A_t$ (mixture feedback).

More applications:

- online linear programming (Agrawal-Wang-Ye'14)
- adversarial bandits w. knapsacks (Immorlica-Sankararaman-Schapire-Slivkins'18)

Our Result: Performance Measures

We investigate/propose the following three measures:

- ***individual regret***

$$\mathrm{Reg}_T^x = \sum_{t=1}^{T} x_t^\top A_t y_t - \min_{x \in \Delta_m} \sum_{t=1}^{T} x^\top A_t y_t$$

- ***dynamic NE-regret***

$$\text{DynNE-Reg}_T = \left| \sum_{t=1}^{T} x_t^\top A_t y_t - \sum_{t=1}^{T} \min_{x \in \Delta_m} \max_{y \in \Delta_n} x^\top A_t y \right|$$

new, proposed by this paper

- ***duality gap***

$$\text{Dual-Gap}_T = \sum_{t=1}^{T} \left(\max_{y \in \Delta_n} x_t^\top A_t y - \min_{x \in \Delta_m} x^\top A_t y_t \right)$$

Our Result: Algorithm and Theory

- We propose a parameter-free algorithm that obtains the following simultaneously (when deployed by both players):

Measure	Time-Varying Game	Fixed Game
Individual Regret	$\widetilde{\mathcal{O}}(\sqrt{1+Q_T})$	$\widetilde{\mathcal{O}}(1)$ recovers [HAM'21]
Dynamic NE-Reg	$\widetilde{\mathcal{O}}(\min\{\sqrt{(1+V_T)(1+P_T)}+P_T, 1+W_T\})$	$\widetilde{\mathcal{O}}(1)$ recovers [HAM'21]
Duality Gap	$\widetilde{\mathcal{O}}(\min\{T^{\frac{3}{4}}(1+Q_T)^{\frac{1}{4}}, T^{\frac{1}{2}}(1+Q_T^{\frac{3}{2}}+P_T Q_T)^{\frac{1}{2}}\})$	$\widetilde{\mathcal{O}}(\sqrt{T})$ recovers [WLZL'21]

- $Q_T = V_T + \min\{P_T, W_T\}$
- the last column also holds when A_t changes $\mathcal{O}(1)$ times
- robustness: $\mathrm{Reg}_T^x = \widetilde{\mathcal{O}}(\sqrt{T})$ even if y-player behaves arbitrarily

Technique Highlight

- ***dynamic regret***: a central concept to achieve different adaptivity

For time-varying games, it is important to compete with arbitrary time-varying strategies $u_1, \ldots, u_T \in \Delta_m$. We thus propose Dynamic RVU:

$$\sum_{t=1}^{T}(x_t - u_t)^\top A_t y_t \le \frac{\alpha P_T^u}{\eta} + \eta\beta \sum_{t=2}^{T} \|A_t y_t - A_{t-1} y_{t-1}\|_\infty^2 - \frac{\gamma}{\eta} \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1^2$$

where $P_T^u = 1 + \sum_{t=2}^{T} \|u_t - u_{t-1}\|_1$.

[1] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. ICML 2003.
[2] Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. NeurIPS 2018.
[3] Peng Zhao, Yu-Jie Zhang, Lijun Zhang, and Zhi-Hua Zhou. Adaptivity and non-stationarity: Problem-dependent dynamic regret for online convex optimization. ArXiv preprint:2112.14368, 2021.

Algorithm Overview (for *x*-player)

Online Ensemble

two-layer structure

Input: any base algorithm $\mathcal{A}(\eta)$ satisfying DRVU with learning rate η.

Initialize: a set of $\mathcal{O}(\log T)$ base learners $\mathcal{S}$, each of which is $\mathcal{A}(\eta)$ with a certain η or a dummy learner always selecting a fixed action

Idea 1: make sure the meta-algorithm comparable to Nash

For $t = 1, \ldots, T$:

- receive $x_{t,i} \in \Delta_m$ from each base learner $i \in \mathcal{S}$.
- compute "prediction vector m_t" and update $p_t \in \Delta_{\mathcal{S}}$ as:

$$p_t = \operatorname*{argmin}_{p \in \Delta_{\mathcal{S}}} \epsilon_t \langle p, m_t \rangle + \|p - \widehat{p}_t\|_2^2$$

- play the final action $x_t = \sum_{i \in \mathcal{S}} p_{t,i} x_{t,i}$
- suffer loss $x_t^\top A_t y_t$, observe $A_t y_t$, and send it to each base learner
- compute "loss vector ℓ_t" and update $\widehat{p}_{t+1}$ as:

$$\widehat{p}_{t+1} = \operatorname*{argmin}_{p \in \Delta_{\mathcal{S}}} \epsilon_t \langle p, \ell_t \rangle + \|p - \widehat{p}_t\|_2^2$$

Idea 2: inject correction term to bias towards more stable learners

Algorithm Overview (for x-player)

Online Ensemble

two-layer structure

Input: any base algorithm $\mathcal{A}(\eta)$ satisfying DRVU with learning rate η.

Initialize: a set of $\mathcal{O}(\log T)$ base learners $\mathcal{S}$, each of which is $\mathcal{A}(\eta)$ with a certain η or a dummy learner always selecting a fixed action

For $t = 1, \ldots, T$:

- receive $x_{t,i} \in \Delta_m$ from each base learner $i \in \mathcal{S}$.

Idea 1: make sure the meta-algorithm comparable to Nash

Injecting a correction term into feedback loss and optimism

$$\ell^x_{t,i} = x_{t,i}^\top A_t y_t + \lambda \|x_{t,i} - x_{t-1,i}\|_1^2$$

$$m^x_{t,i} = x_{t,i}^\top A_{t-1} y_{t-1} + \lambda \|x_{t,i} - x_{t-1,i}\|_1^2$$

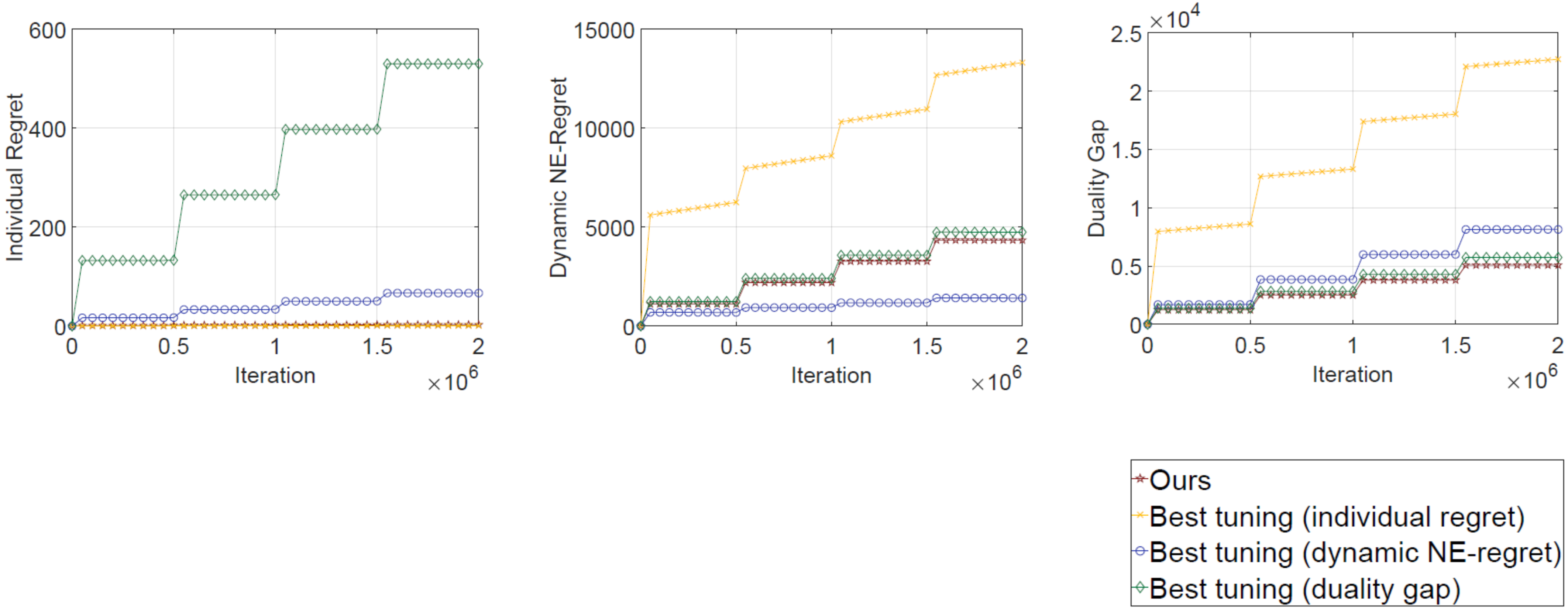
Purpose: bias towards more stable base-learners to make the cancelation in the dynamic regret feasible

$$\widehat{p}_{t+1} = \operatorname*{argmin}_{p \in \Delta_{\mathcal{S}}} \epsilon_t \langle p, \ell_t \rangle + \|p - \widehat{p}_t\|_2^2$$

Idea 2: inject correction term to bias towards more stable learners

Experiments

- A synthetic environment s.t. $P_T = \Theta(\sqrt{T}),\ W_T = \Theta(T^{\frac{3}{4}}),\ V_T = \Theta(\sqrt{T}).$



Summary

- A systematic study for time-varying zero-sum games.
- Rethink existing *performance measures* and propose new one.
- Design *a single parameter-free algorithm* that can simultaneously optimize all three measures (individual regret, dynamic NE-regret, duality gap); can recover best known results when the game is fixed.
- The results build upon *online ensemble* framework, but nevertheless require several new components (correction terms & dummy learners) and exploit the specific structure of games.

Thanks!