On the convergence of learning dynamics in games

Julien Grand-Clément (HEC Paris)

.

With: Y. Cai (Yale), G. Farina (MIT), C. Kroer (Columbia), C.-W. Lee (Meta), H. Luo (USC), W. Zheng (Yale).

European TOM Seminar Series - March 2024

Recent successes for learning in games

AlphaGo beats top Go player Lee Sedol in 2016 [SHM+16]:

The New York Times

THE GLOBAL PROFILE

Defeated by A.I., a Legend in the Board Game Go Warns: Get Ready for What's Next

Lee Saedol was one of the world's top Go players, and his shocking loss to an A.I. opponent was a harbinger of a new, unsettling era. "It may not be a happy ending," he says.

Recent successes for learning in games

AlphaGo beats top Go player Lee Sedol in 2016 [SHM+16]:


The New York Times

THE GLOBAL PROFILE

Defeated by A.I., a Legend in the Board Game Go Warns: Get Ready for What's Next

Lee Saedol was one of the world's top Go players, and his shocking loss to an A.I. opponent was a harbinger of a new, unsettling era. "It may not be a happy ending," he says.

Abstract of the *Nature* paper:
"[Our algorithms] are trained by a novel combination of supervised learning from human expert games, and reinforcement learning from games of self-play".

Recent successes for learning in games

AIs beating top poker players: Libratus [BS18], Pluribus [BS19]

Poker Bot Pluribus First AI to Beat Humans in Multiplayer No-Limit Hold'em

Valerie Cross

Jul 19, 2019 • 8 min read

The New York Times

ence › | Google's Anthropic Investment | What Is Vibecoding? | OpenAI and Musk | A Look at OpenAI'

Hold 'Em or Fold 'Em? This A.I. Bluffs With the Best

Pluribus, a poker-playing algorithm, can beat the world's top human players, proving that machines, too, can master our mind games.

Recent successes for learning in games

AIs beating top poker players: Libratus [BS18], Pluribus [BS19]

Poker Bot Pluribus First AI to Beat Humans in Multiplayer No-Limit Hold'em

Valerie Cross

Jul 19, 2019 • 8 min read

The New York Times

ence › | Google's Anthropic Investment | What Is Vibecoding? | OpenAI and Musk | A Look at OpenAI'

Hold 'Em or Fold 'Em? This A.I. Bluffs With the Best

Pluribus, a poker-playing algorithm, can beat the world's top human players, proving that machines, too, can master our mind games.

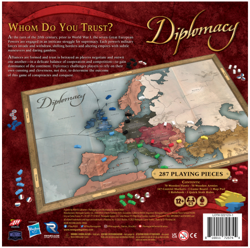
Description of Pluribus in the *Science* paper:

Description of Pluribus

The core of Pluribus's strategy was computed through self-play, in which the AI plays against copies of itself, without any data of human or prior AI play used as input. The AI starts from scratch by playing randomly and gradually improves as it determines which actions, and which probability distribution over those actions, lead to better outcomes against earlier versions of its strategy. Forms of self-play have previously

Other recent achievements

AIs for Stratego [PDVH$^+$22] and Diplomacy [FBB$^+$22]:



Other areas of applications of self-play: boosting [FS96], training generative adversarial networks [DISZ18], fine-tuning large language models [MVC$^+$23], protein folding [WTH$^+$23]

Why should people in Operations Management care?

- Powerful tools developed for solving multi-agent decision problems

- At the core: regret minimization and online learning...
  ... already used in several areas in OM
  - online resource allocation [BLM22]
  - pricing in auctions [CBGM14]
  - online market equilibrium [GPK21]
  - network revenue management [MW21]

This Talk In One Slide

Main objective:
Understanding the convergence of learning algorithms (OMWU)

Why it's interesting?
See previous slides

Main results:

1. The most popular algorithms converge arbitrarily slow ...
   ... because they don't forget the past quickly enough!
2. Last- vs. best- vs. random- vs. average-iterate convergence
3. Uniform vs. universal convergence

Matrix games

Think of rock-paper-scissors:

Setup: two players with d_1 and d_2 actions, zero-sum payoff A_{ij}

Strategies: $x \in \Delta^{d_1}, y \in \Delta^{d_2}$

Payoff: the first player pays $x^\top Ay$ to the second player

Goal: Compute a Nash equilibrium $(x^\star, y^\star)$

Matrix games

Think of rock-paper-scissors:

Setup: two players with d_1 and d_2 actions, zero-sum payoff A_{ij}

Strategies: $x \in \Delta^{d_1}, y \in \Delta^{d_2}$

Payoff: the first player pays $x^\top Ay$ to the second player

Goal: Compute a Nash equilibrium $(x^\star, y^\star)$

Approach: $(x^\star, y^\star)$ N.E. $\iff$ $\mathrm{DualGap}(x^\star, y^\star) = 0$ with

$$\mathrm{DualGap}(x^\star, y^\star) := \max_{y\in\Delta^{d_2}} (x^\star)^\top Ay - \min_{x\in\Delta^{d_1}} x^\top Ay^\star.$$

Self-play

Self-play: the machine learns by playing against itself.

At iteration t:

1. Players choose strategies x^t, y^t
2. x-player receives loss $Ay^t \in \mathbb{R}^{d_1}$
3. y-player receives loss $-A^\top x^t \in \mathbb{R}^{d_2}$

Stop at iteration T, return average iterates: $\frac{1}{T}\sum_{t=1}^{T}(x^t, y^t)$

Next question:
How should player choose their strategies next, given the losses?

Multiplicative Weight Update (MWU)[1]

Example for the x-player, with loss $\ell^t = Ay^t \in \mathbb{R}^{d_1}$:
For each action i,

$$x_i^t \;\propto\; \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau\right)\right) \qquad \text{(MWU)}$$

[1] Also called *Hedge, online mirror descent, dual averaging, FTRL, etc.*

Multiplicative Weight Update (MWU)[1]

Example for the x-player, with loss $\ell^t = Ay^t \in \mathbb{R}^{d_1}$:
For each action i,

$$x_i^t \propto \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau\right)\right) \qquad \text{(MWU)}$$

- MWU decreases the proba. to play action i with large loss ℓ_i

[1] Also called *Hedge, online mirror descent, dual averaging, FTRL, etc.*

Multiplicative Weight Update (MWU)[1]

Example for the x-player, with loss $\ell^t = Ay^t \in \mathbb{R}^{d_1}$:
For each action i,

$$x_i^t \;\propto\; \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau\right)\right) \qquad \text{(MWU)}$$

- MWU decreases the proba. to play action i with large loss ℓ_i
- At iteration t, MWU uses all past losses $\ell^1, .., \ell^{t-1}$

[1] Also called *Hedge, online mirror descent, dual averaging, FTRL, etc.*

Multiplicative Weight Update (MWU)[1]

Example for the x-player, with loss $\ell^t = Ay^t \in \mathbb{R}^{d_1}$:
For each action i,

$$x_i^t \propto \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau\right)\right) \qquad \text{(MWU)}$$

- MWU decreases the proba. to play action i with large loss ℓ_i
- At iteration t, MWU uses all past losses $\ell^1, .., \ell^{t-1}$

Optimistic MWU: count the last loss twice

$$x_i^t \propto \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau + \ell_i^{t-1}\right)\right) \qquad \text{(OMWU)}$$

[1] Also called *Hedge, online mirror descent, dual averaging, FTRL, etc.*

Advantages of OMWU:

1. Closed-form updates
2. Regret bounds logarithmic in the size of payoff matrix
3. Regret bound in $\tilde{O}(1)$ in n-player games [DFG21]
4. $\tilde{O}(1/T)$ average convergence to (coarse) correlated equilibrium in general-sum games [ADF$^+$22]

Advantages of OMWU:

1. Closed-form updates
2. Regret bounds logarithmic in the size of payoff matrix
3. Regret bound in $\tilde{O}(1)$ in n-player games [DFG21]
4. $\tilde{O}(1/T)$ average convergence to (coarse) correlated equilibrium in general-sum games [ADF$^+$22]

Best-in class guarantees!

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Average convergence:

$$\lim_{T \to +\infty} \mathrm{DualGap}\left(\frac{1}{T} \sum_{t=1}^{T} (x^t, y^t) \right) = 0.$$

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Average convergence:

$$\lim_{T \to +\infty} \text{DualGap}\left(\frac{1}{T}\sum_{t=1}^{T}(x^t, y^t)\right) = 0.$$

Seminal results [RS13, SALS15]:
OMWU gives $O(1/T)$ average convergence in matrix games.

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Average convergence:

$$\lim_{T\to+\infty} \mathrm{DualGap}\left(\frac{1}{T}\sum_{t=1}^{T}(x^t, y^t)\right) = 0.$$

Seminal results [RS13, SALS15]:
OMWU gives $O(1/T)$ average convergence in matrix games.

What if computing running averages is too cumbersome?

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Last-iterate convergence:

$$\lim_{T \to +\infty} \text{DualGap}(x^T, y^T) = 0.$$

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Last-iterate convergence:

$$\lim_{T \to +\infty} \text{DualGap}(x^T, y^T) = 0.$$

Best-iterate convergence:

$$\lim_{T \to +\infty} \min_{t \leq T} \text{DualGap}(x^t, y^t) = 0.$$

Notions of convergence

Consider the sequence $\{x^t, y^t\}_t$ computed by self-play.

Last-iterate convergence:

$$\lim_{T \to +\infty} \mathrm{DualGap}(x^T, y^T) = 0.$$

Best-iterate convergence:

$$\lim_{T \to +\infty} \min_{t \leq T} \mathrm{DualGap}(x^t, y^t) = 0.$$

Random-iterate convergence:

$$\lim_{T \to +\infty} \frac{1}{T} \sum_{t=1}^{T} \mathrm{DualGap}(x^t, y^t) = 0.$$

Why do we care about convergence in iterates?

- Computationally cheaper than averaging (think of LLMs)
- Eventually the players sample actions from an equilibrium
- W/o convergence, undesirable recurrence/chaotic behavior
- Practical performance may be better than averaging

Last-iterate dynamics of OMWU:

- Convergence result without rates [DP19, MLZ+19, HAM21]
- Unique N.E.: linear rate with large constant $C > 0$ [WLZL21]:

$$\text{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

Last-iterate dynamics of OMWU:

- Convergence result without rates [DP19, MLZ$^+$19, HAM21]
- Unique N.E.: linear rate with large constant $C > 0$ [WLZL21]:

$$\mathrm{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

Problem: C may be arbitrarily large, even for $A \in [0,1]^{2x2}$!

Last-iterate dynamics of OMWU:

- Convergence result without rates [DP19, MLZ+19, HAM21]
- Unique N.E.: linear rate with large constant $C > 0$ [WLZL21]:

$$\mathrm{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

Problem: C may be arbitrarily large, even for $A \in [0,1]^{2x2}$!
For $\delta =$ min. non-zero probability in N.E., $C = \Omega\left(\exp\left(\frac{1}{\delta}\right)\right)$

Last-iterate dynamics of OMWU:

- Convergence result without rates [DP19, MLZ+19, HAM21]
- Unique N.E.: linear rate with large constant $C > 0$ [WLZL21]:

$$\mathrm{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

Problem: C may be arbitrarily large, even for $A \in [0,1]^{2x2}$!
For $\delta =$ min. non-zero probability in N.E., $C = \Omega\left(\exp\left(\frac{1}{\delta}\right)\right)$

Open research questions before our work:
Better rates for last-iterate convergence of OMWU?
What about best-/random-iterate convergence?

Uniform vs. universal rates:

- *Universal* rate: depends on T, d_1, d_2 and payoff matrix A

Uniform vs. universal rates:

- *Universal* rate: depends on T, d_1, d_2 and payoff matrix A

  Example [WLZL21]:

$$\forall A, \exists C > 0, \mathrm{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

Uniform vs. universal rates:

- *Universal* rate: depends on T, d_1, d_2 and payoff matrix A

  Example [WLZL21]:

$$\forall A, \exists C > 0, \mathrm{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

- *Uniform* rate: depends only on T, d_1, d_2.

Uniform vs. universal rates:

- *Universal* rate: depends on T, d_1, d_2 and payoff matrix A

  Example [WLZL21]:

$$\forall A, \exists C > 0, \text{DualGap}\left(x^T, y^T\right) = O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)$$

- *Uniform* rate: depends only on T, d_1, d_2.

  Research question: can we find a function f such that

$$\exists C > 0, \forall A, \text{DualGap}\left(x^T, y^T\right) = O\left(f(C, T)\right)$$

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate		

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. ‡: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	universal	uniform
Last iterate	$O\Big(C \cdot \exp\big(-\frac{T}{C}\big)\Big)^{\dagger}$	$\Omega\Big(1\Big)$

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	$\Omega\left(1\right)$
Random iterate		

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. ‡: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	$\Omega\Big(1\Big)$
Random iterate	$O\left(\log(C)^{\frac{1}{2}} \cdot T^{-\frac{1}{4}}\right)$	

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	$\Omega\left(1\right)$
Random iterate	$O\left(\log(C)^{\frac{1}{2}} \cdot T^{-\frac{1}{4}}\right)$	$\Omega\left(\frac{1}{\log T}\right)$

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells) [CFGC+24, CFGC+25]:

Convergence	**universal**	**uniform**
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	$\Omega\left(1\right)$
Random iterate	$O\left(\log(C)^{\frac{1}{2}} \cdot T^{-\frac{1}{4}}\right)$	$\Omega\left(\frac{1}{\log T}\right)$
Best iterate	$O(T^{-\frac{1}{6}})^{\ddagger}$	

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

Our results for OMWU (in blue cells):

Convergence	universal	uniform
Last iterate	$O\left(C \cdot \exp\left(-\frac{T}{C}\right)\right)^{\dagger}$	$\Omega\left(1\right)$
Random iterate	$O\left(\log(C)^{\frac{1}{2}} \cdot T^{-\frac{1}{4}}\right)$	$\Omega\left(\frac{1}{\log T}\right)$
Best iterate	$O(T^{-\frac{1}{6}})^{\ddagger}$	

Table: †: $C := \Omega(\exp(\frac{1}{\delta}))$ with $\delta > 0$ is the min. proba. in N.E.. $\ddagger$: This upper bound only holds for 2×2 games.

1. Uniform last-iterate rates are *impossible*
2. Uniform random-iterate rates are no faster than $1/\log(T)$
3. Uniform best-iterate rates are polynomial in $1/T$

Example of impossibility result

Theorem (Informal)

Consider two-player zero-sum games with matrix entries in $[0, 1]$*, and* d_1 *and* d_2 *are the number of actions.*

For OMWU with constant step size, no function f *can satisfy*

1. $\mathrm{DualGap}(x^T, y^T) \leq f(d_1, d_2, T)$ *for all* T*.*
2. $\lim_{T\to\infty} f(d_1, d_2, T) \to 0$.

Example of impossibility result

Theorem (Informal)

Consider two-player zero-sum games with matrix entries in $[0,1]$*, and* d_1 *and* d_2 *are the number of actions.*

For OMWU with constant step size, no function f *can satisfy*

1. $\mathrm{DualGap}(x^T, y^T) \leq f(d_1, d_2, T)$ *for all* T*.*
2. $\lim_{T\to\infty} f(d_1, d_2, T) \to 0$.

Note: for each instance A, $\mathrm{DualGap}(x^T, y^T) \to 0$ ...
... but we can make this convergence arbitrarily slow!

A difficult matrix game for OMWU

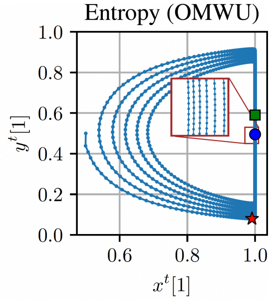
Consider the matrix game A_δ with $0 < \delta < 1/2$:

$$A_\delta := \begin{bmatrix} \frac{1}{2} + \delta & \frac{1}{2} \\ 0 & 1 \end{bmatrix}$$

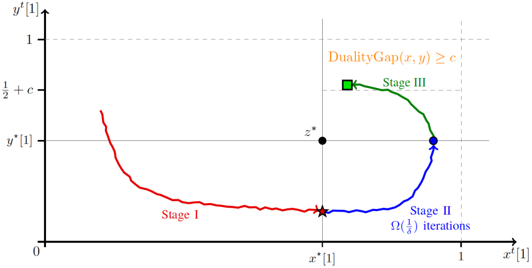
A_δ has a unique N.E., δ-close to the simplex boundary.

Only 2 actions $\Rightarrow$ dynamics fully described by $x^t[1], y^t[1] \in (0, 1)$

Running OMWU on A_δ

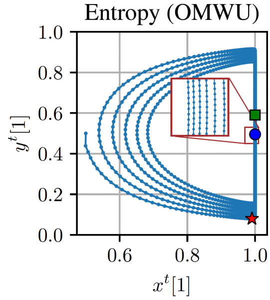


(a) Dynamics of OMWU

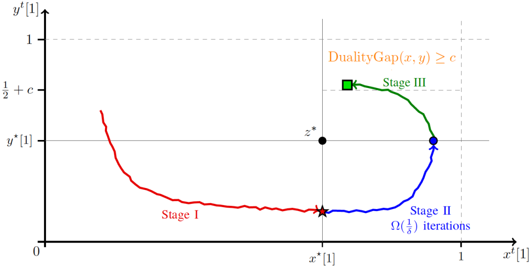


(b) Dynamics of OMWU (analysis)

Running OMWU on A_δ



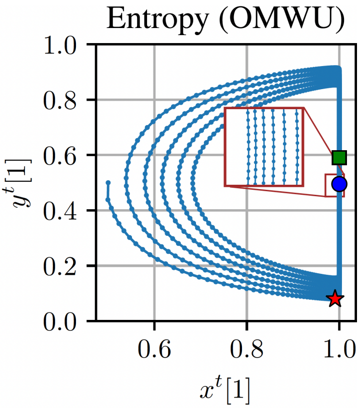
(a) Dynamics of OMWU



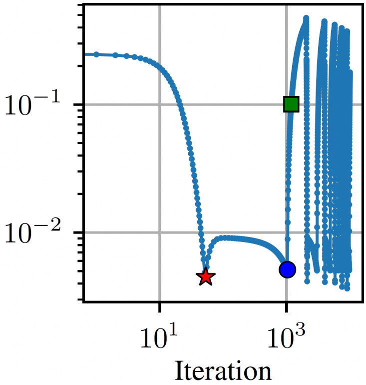
(b) Dynamics of OMWU (analysis)

In A_δ, a duality gap of $c > 0$ is attained after $\Omega\left(\frac{1}{\delta}\right)$ iterations!

Running OMWU on A_δ



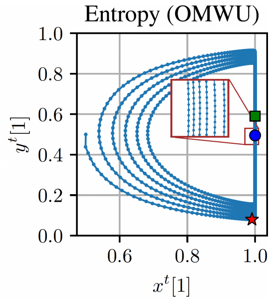
(a) Dynamics of OMWU



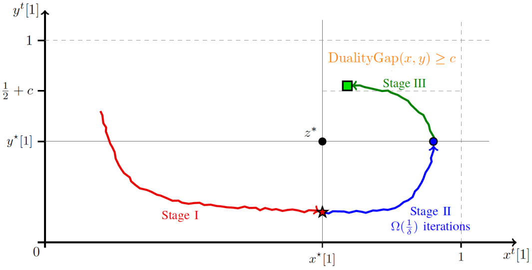
(b) Duality gap of last iterate

In A_δ, a duality gap of $c > 0$ is attained after $\Omega\left(\frac{1}{\delta}\right)$ iterations!

Running OMWU on A_δ

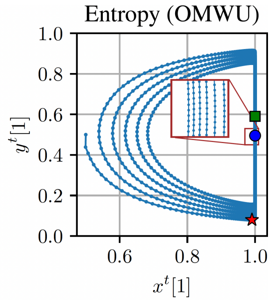


(a) Dynamics of OMWU

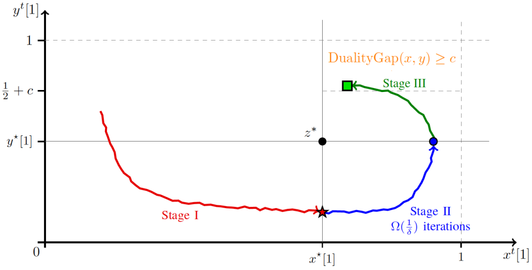


(b) Dynamics of OMWU (analysis)

Running OMWU on A_δ



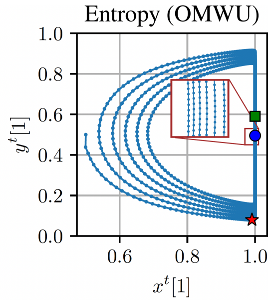
(a) Dynamics of OMWU



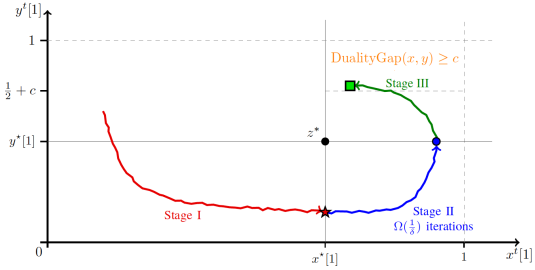
(b) Dynamics of OMWU (analysis)

Main issue: OMWU does not forget the past quickly enough!

Running OMWU on A_δ



(a) Dynamics of OMWU

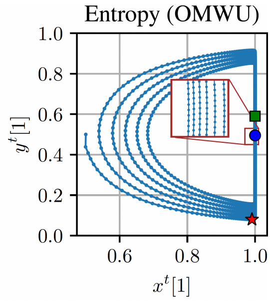


(b) Dynamics of OMWU (analysis)

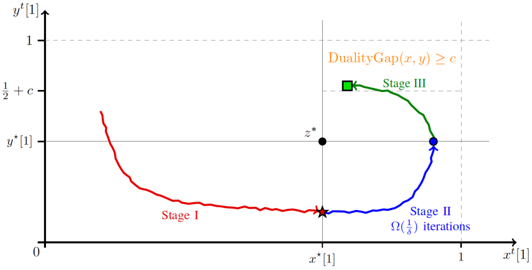
Main issue: OMWU does not forget the past quickly enough!

$y^{t+1}[1]$ can only decrease if $x^t[1] < x^\star[1]$.

Running OMWU on A_δ



(a) Dynamics of OMWU



(b) Dynamics of OMWU (analysis)

Main issue: OMWU does not forget the past quickly enough!

$y^{t+1}[1]$ can only decrease if $x^t[1] < x^\star[1]$.

But the x-player uses all the losses from the past:

$$x_i^t \propto \exp\left(-\eta \cdot \left(\sum_{\tau=1}^{t-1} \ell_i^\tau + \ell_i^{t-1}\right)\right) \qquad \text{(OMWU)}$$

Running OMWU on A_δ

Recall $\delta > 0$ captures how close is N.E. to the boundary

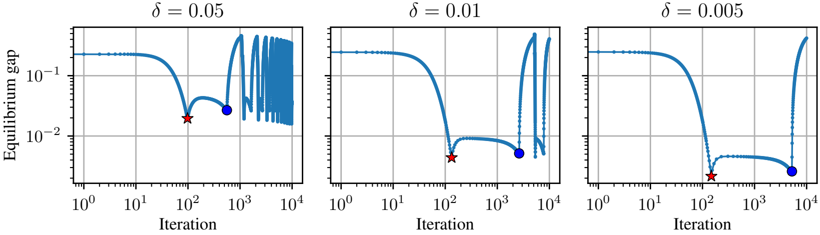
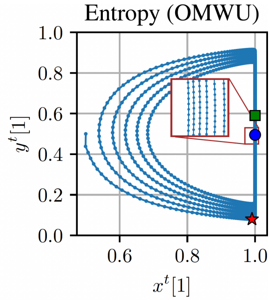


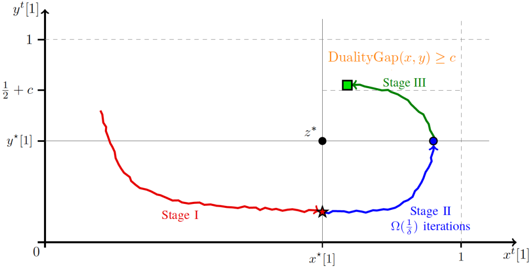
Figure: Duality gap of last-iterates produced by OMWU in the game A_δ for various values of δ.

In A_δ, a duality gap of $c > 0$ is attained after $\Omega\left(\frac{1}{\delta}\right)$ iterations!

Random-iterate performance

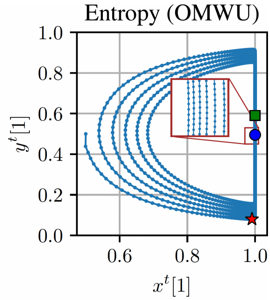


(a) Dynamics of OMWU

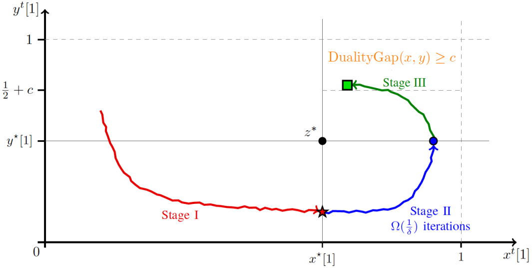


(b) Dynamics of OMWU (analysis)

Random-iterate performance



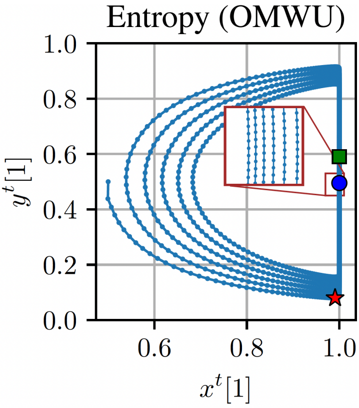
(a) Dynamics of OMWU



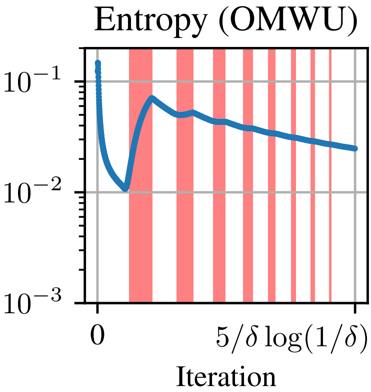
(b) Dynamics of OMWU (analysis)

In A_δ, after $O\left(\frac{1}{\delta \log(\delta)}\right)$ iterations, the duality gap remains larger than $c >$ for $\Omega\left(\frac{1}{\delta}\right)$ iterations!

Random-iterate performance



(a) Dynamics of OMWU



(b) Average Duality gap

Main issue: OMWU does not forget the past quickly enough!

Time permitting: experiments with other regularizers:

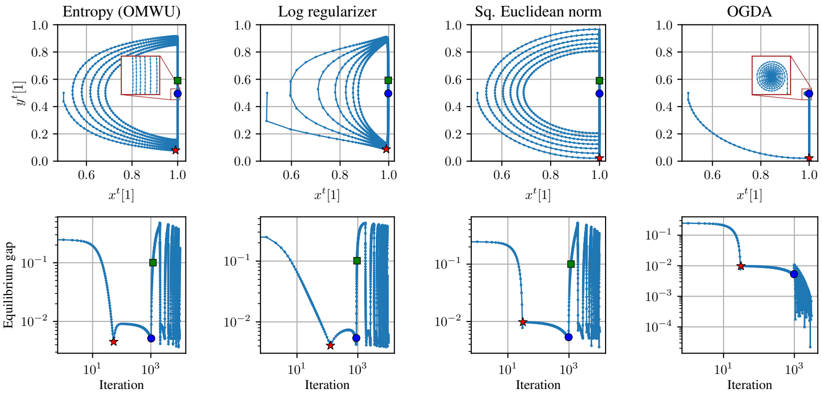
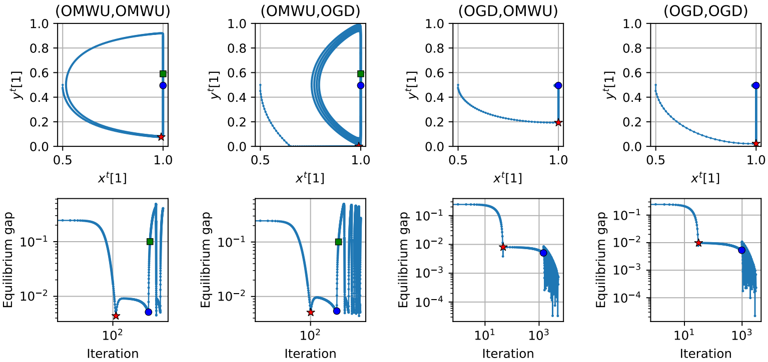


Figure: OFTRL with different regularizers and OGDA in A_δ.

1. OMWU = FTRL with entropy as regularizer.
   Pathological behaviors persist with other regularizers!
2. Optimistic Gradient Descent (OGDA, only uses last loss) fixes the issue

Time permitting: mixing OMWU and OGD



⇒ Only one non-forgetful player seems to lead to pathological behavior!

Conclusion

- **Main result:** separation last-/best-/random-iterate convergence for a widely studied algorithms
- **Next steps:**
  How to alleviate this "pathological behavior"?
  Uniform best-iterate conv. rate beyond 2×2 games?
- **More in the two papers:**
  follow-the-regularized-leader (FTRL) vs. online gradient descent (OGD)
  Papers/slides/code available on my website

Conclusion

- **Main result:** separation last-/best-/random-iterate convergence for a widely studied algorithms
- **Next steps:**
  How to alleviate this "pathological behavior"?
  Uniform best-iterate conv. rate beyond 2×2 games?
- **More in the two papers:**
  follow-the-regularized-leader (FTRL) vs. online gradient descent (OGD)
  Papers/slides/code available on my website

Thank you!

References I

Ioannis Anagnostides, Constantinos Daskalakis, Gabriele Farina, Maxwell Fishelson, Noah Golowich, and Tuomas Sandholm.
Near-optimal no-regret learning for correlated equilibria in multi-player general-sum games.
In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, 2022.

Santiago R Balseiro, Haihao Lu, and Vahab Mirrokni.
The best of many worlds: Dual mirror descent for online allocation problems.
Operations Research, 2022.

Noam Brown and Tuomas Sandholm.
Superhuman ai for heads-up no-limit poker: Libratus beats top professionals.
Science, 359(6374):418–424, 2018.

References II

Noam Brown and Tuomas Sandholm.
Superhuman ai for multiplayer poker.
Science, 365(6456):885–890, 2019.

Nicolo Cesa-Bianchi, Claudio Gentile, and Yishay Mansour.
Regret minimization for reserve prices in second-price auctions.
IEEE Transactions on Information Theory, 61(1):549–564, 2014.

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng.
Fast last-iterate convergence of learning in games requires forgetful algorithms.
arXiv preprint arXiv:2406.10631, 2024.

References III

Yang Cai, Gabriele Farina, Julien Grand-Clément, Christian Kroer, Chung-Wei Lee, Haipeng Luo, and Weiqiang Zheng.
On separation between best-iterate, random-iterate, and last-iterate convergence of learning in games.
arXiv preprint arXiv:2503.02825, 2025.

Constantinos Daskalakis, Maxwell Fishelson, and Noah Golowich.
Near-optimal no-regret learning in general games.
Advances in Neural Information Processing Systems (NeurIPS), 2021.

John Duchi, Elad Hazan, and Yoram Singer.
Adaptive subgradient methods for online learning and stochastic optimization.
Journal of machine learning research, 12(7), 2011.

References IV

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng.
Training gans with optimism.
In *International Conference on Learning Representations (ICLR)*, 2018.

Constantinos Daskalakis and Ioannis Panageas.
Last-iterate convergence: Zero-sum games and constrained min-max optimization.
In *10th Innovations in Theoretical Computer Science Conference (ITCS)*, 2019.

References V

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al.
Human-level play in the game of diplomacy by combining language models with strategic reasoning.
Science, 378(6624):1067–1074, 2022.

Yoav Freund and Robert E Schapire.
Game theory, on-line prediction and boosting.
In *Proceedings of the ninth annual conference on Computational learning theory*, pages 325–332, 1996.

Yuan Gao, Alex Peysakhovich, and Christian Kroer.
Online market equilibrium with application to fair division.
Advances in Neural Information Processing Systems, 34:27305–27318, 2021.

References VI

Yu-Guan Hsieh, Kimon Antonakopoulos, and Panayotis Mertikopoulos.
Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium.
In *Conference on Learning Theory*, pages 2388–2422. PMLR, 2021.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras.
Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile.
In *International Conference on Learning Representations (ICLR)*, 2019.

References VII

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al.
Nash learning from human feedback.
arXiv preprint arXiv:2312.00886, 18, 2023.

Sentao Miao and Yining Wang.
Network revenue management with nonparametric demand learning:\sqrt {T}-regret and polynomial dimension dependency.
Available at SSRN 3948140, 2021.

References VIII

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al.
Mastering the game of stratego with model-free multiagent reinforcement learning.
Science, 378(6623):990–996, 2022.

Sasha Rakhlin and Karthik Sridharan.
Optimization, learning, and games with predictable sequences.
Advances in Neural Information Processing Systems, 2013.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire.
Fast convergence of regularized learning in games.
Advances in Neural Information Processing Systems (NeurIPS), 2015.

References IX

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al.
Mastering the game of go with deep neural networks and tree search.
nature, 529(7587):484–489, 2016.

Chen-Yu Wei, Chung-Wei Lee, Mengxiao Zhang, and Haipeng Luo.
Linear last-iterate convergence in constrained saddle-point optimization.
In *International Conference on Learning Representations (ICLR)*, 2021.

References X

Yi Wang, Hui Tang, Lichao Huang, Lulu Pan, Lixiang Yang, Huanming Yang, Feng Mu, and Meng Yang.
Self-play reinforcement learning guides protein engineering.
Nature Machine Intelligence, 5(8):845–860, 2023.

Adaptive stepsizes

Adaptive stepsize [DHS11]: $\eta_t = 1/\sqrt{\epsilon + \sum_{k=1}^{t-1} \|\boldsymbol{\ell}_k\|_k^2}$

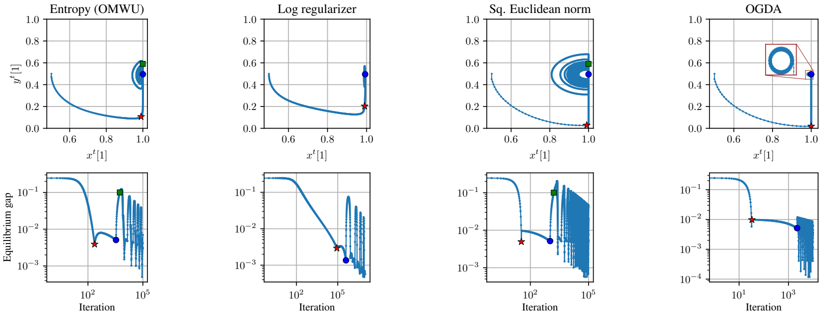


Figure: Here $\delta := 10^{-2}$ and adaptive step size with $\epsilon = 0.1$.