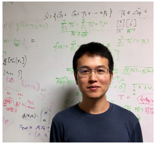
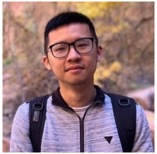
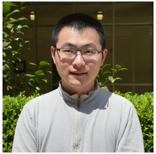
Linear Last-iterate Convergence in Constrained Saddle-point Optimization

Chung-Wei Lee

joint with **Haipeng Luo**, **Chen-Yu Wei** and **Mengxiao Zhang**







USC University of Southern California

USC Viterbi School of Engineering

A One-sentence (Informal) Summary

We prove that the last-iterate of

Optimistic Gradient Descent Ascent (OGDA) and

Optimistic Multiplicative Weights Update (OMWU)

converges to the Nash equilibrium **exponentially** fast,

in various constrained settings including matrix games and

strongly-convex-strongly-concave functions.

Saddle-point Optimization

- Consider constrained saddle-point optimization in the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets, and f is a continuous differentiable function that is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$.

Saddle-point Optimization

- Consider constrained saddle-point optimization in the form

$$\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}),$$

where $\mathcal{X}$ and $\mathcal{Y}$ are closed convex sets, and f is a continuous differentiable function that is convex in $\boldsymbol{x}$ and concave in $\boldsymbol{y}$.

- Goal: find a *Nash equilibrium* $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X}^* \times \mathcal{Y}^*$ satisfying

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \leq f(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq f(\boldsymbol{x}, \boldsymbol{y}^*)$$

for any $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}$.

"No-Regret" Algorithms

- (Projected) Gradient Descent Ascent (GDA):

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), \quad \boldsymbol{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\boldsymbol{y}_t + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right)$$

"No-Regret" Algorithms

- (Projected) Gradient Descent Ascent (GDA):

$$\boldsymbol{x}_{t+1} = \Pi_{\mathcal{X}}\left(\boldsymbol{x}_t - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), \quad \boldsymbol{y}_{t+1} = \Pi_{\mathcal{Y}}\left(\boldsymbol{y}_t + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right)$$

- Multiplicative Weights Update (MWU):

$$\boldsymbol{x}_{t+1} \propto \left(\boldsymbol{x}_t \odot \exp(-\eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right), \quad \boldsymbol{y}_{t+1} \propto \left(\boldsymbol{y}_t \odot \exp(\eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right),$$

when $\mathcal{X}$ and $\mathcal{Y}$ are simplex, and $\odot$ denotes the element-wise product.

Optimistic No-Regret Algorithms

- Optimistic Gradient Descent Ascent (OGDA):

$$
\begin{aligned}
\widehat{\boldsymbol{x}}_{t+1} &= \Pi_{\mathcal{X}}\left(\widehat{\boldsymbol{x}}_t - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), & \widehat{\boldsymbol{y}}_{t+1} &= \Pi_{\mathcal{Y}}\left(\widehat{\boldsymbol{y}}_t + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right) \\
\boldsymbol{x}_{t+1} &= \Pi_{\mathcal{X}}\left(\widehat{\boldsymbol{x}}_{t+1} - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), & \boldsymbol{y}_{t+1} &= \Pi_{\mathcal{Y}}\left(\widehat{\boldsymbol{y}}_{t+1} + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right)
\end{aligned}
$$

Optimistic No-Regret Algorithms

- Optimistic Gradient Descent Ascent (OGDA):

$$
\begin{aligned}
\widehat{\boldsymbol{x}}_{t+1} &= \Pi_{\mathcal{X}}\left(\widehat{\boldsymbol{x}}_t - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), \quad & \widehat{\boldsymbol{y}}_{t+1} &= \Pi_{\mathcal{Y}}\left(\widehat{\boldsymbol{y}}_t + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right) \\
\boldsymbol{x}_{t+1} &= \Pi_{\mathcal{X}}\left(\widehat{\boldsymbol{x}}_{t+1} - \eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right), \quad & \boldsymbol{y}_{t+1} &= \Pi_{\mathcal{Y}}\left(\widehat{\boldsymbol{y}}_{t+1} + \eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t)\right)
\end{aligned}
$$

- Optimistic Multiplicative Weights Update (OMWU):

$$
\begin{aligned}
\widehat{\boldsymbol{x}}_{t+1} &\propto \left(\widehat{\boldsymbol{x}}_t \odot \exp(-\eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right), \quad & \widehat{\boldsymbol{y}}_{t+1} &\propto \left(\widehat{\boldsymbol{y}}_t \odot \exp(\eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right) \\
\boldsymbol{x}_{t+1} &\propto \left(\widehat{\boldsymbol{x}}_{t+1} \odot \exp(-\eta \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right), \quad & \boldsymbol{y}_{t+1} &\propto \left(\widehat{\boldsymbol{y}}_{t+1} \odot \exp(\eta \nabla_{\boldsymbol{y}} f(\boldsymbol{x}_t, \boldsymbol{y}_t))\right)
\end{aligned}
$$

Average-iterate Convergence

- Convergence of average-iterate $(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t)$ is well known in many settings.

Average-iterate Convergence

- Convergence of average-iterate $(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t)$ is well known in many settings.
- GDA and MWU are known to enjoy a converging duality gap of $\mathcal{O}(1/\sqrt{T})$. [FS99]

Average-iterate Convergence

- Convergence of average-iterate $(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t)$ is well known in many settings.
- GDA and MWU are known to enjoy a converging duality gap of $\mathcal{O}(1/\sqrt{T})$. [FS99]
- Optimistic algortihms such as OGDA and OMWU improve the converging rate to $\mathcal{O}(1/T)$. [RS13,DDK15,SALS15]

Average-iterate Convergence

- Convergence of average-iterate $(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{y}_t)$ is well known in many settings.
- GDA and MWU are known to enjoy a converging duality gap of $\mathcal{O}(1/\sqrt{T})$. [FS99]
- Optimistic algortihms such as OGDA and OMWU improve the converging rate to $\mathcal{O}(1/T)$. [RS13,DDK15,SALS15]
- However, averaging large neural networks is usually prohibited.

Average-iterate Convergence

- Convergence of average-iterate $(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t, \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t)$ is well known in many settings.
- GDA and MWU are known to enjoy a converging duality gap of $\mathcal{O}(1/\sqrt{T})$. [FS99]
- Optimistic algortihms such as OGDA and OMWU improve the converging rate to $\mathcal{O}(1/T)$. [RS13,DDK15,SALS15]
- However, averaging large neural networks is usually prohibited.
- This motivates us to consider the last-iterate $(\boldsymbol{x}_T, \boldsymbol{y}_T)$ convergence.

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.
  - Learning rate is exponentially small, which is inconsistent with practice.

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.
  - Learning rate is exponentially small, which is inconsistent with practice.
- OGDA achieves last-iterate convergence

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.
  - Learning rate is exponentially small, which is inconsistent with practice.
- OGDA achieves last-iterate convergence
  - when $\mathcal{X}$ and $\mathcal{Y}$ are unconstrained. [DISZ18,DP18,LS19,MOP19]

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.
  - Learning rate is exponentially small, which is inconsistent with practice.
- OGDA achieves last-iterate convergence
  - when $\mathcal{X}$ and $\mathcal{Y}$ are unconstrained. [DISZ18,DP18,LS19,MOP19]

Last-iterate Convergence

- For MWU and GDA, last-iterate diverges. [BP18,CP19]
- On the contrary, for *Extra-Gradient*, a standard algortithm for saddle-point optimization, last-iterate convergence has been shown in various settings. [T95,LS19,MOP20]
- OMWU achieves last-iterate convergence when f is bilinear on simplex (i.e. matrix game) when Nash Equilibrium is unique. [DP19]
  - No concrete convergence rate.
  - Learning rate is exponentially small, which is inconsistent with practice.
- OGDA achieves last-iterate convergence
  - when $\mathcal{X}$ and $\mathcal{Y}$ are unconstrained. [DISZ18,DP18,LS19,MOP19]

Question: Whether OGDA and OMWU can achieve last-iterate convergence in **constrained** saddle-point optimization with **concrete** convergence rate?

Our Contributions

- Under uniqueness assumption made by Daskalakis and Panageas (2019), we show that OMWU with constant learning rate has exponential convergence rate.

Our Contributions

- Under uniqueness assumption made by Daskalakis and Panageas (2019), we show that OMWU with constant learning rate has exponential convergence rate.

- For OGDA, we get more general results: under a sufficient condition called SP-MS, OGDA with constant learning rate converges exponentially fast.

Our Contributions

- Under uniqueness assumption made by Daskalakis and Panageas (2019), we show that OMWU with constant learning rate has exponential convergence rate.
- For OGDA, we get more general results: under a sufficient condition called SP-MS, OGDA with constant learning rate converges exponentially fast.
- The SP-MS condition includes many settings such as bilinear games over any polytope and strongly-convex-strongly-concave functions without uniqueness assumption.

Experiments

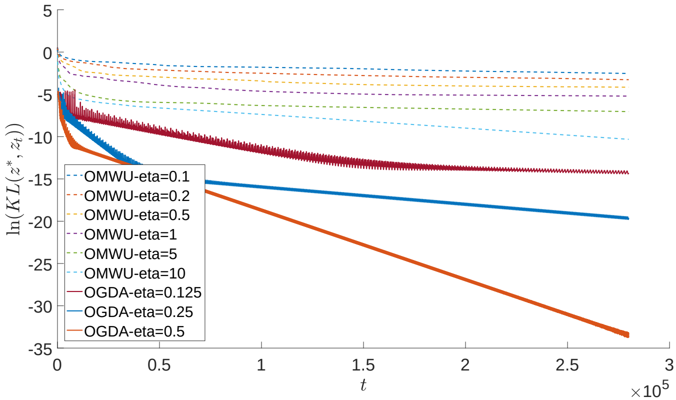


Figure: Experiments of OGDA and OMWU with different learning rates for a matrix game.

Future directions

- One future direction is to get rid of the uniqueness assumption for OMWU.

Future directions

- One future direction is to get rid of the uniqueness assumption for OMWU.
- It is also interesting to generalize the results to Markov/Stochastic Games.

Future directions

- One future direction is to get rid of the uniqueness assumption for OMWU.
- It is also interesting to generalize the results to Markov/Stochastic Games.
- For this direction, see our new paper *Last-iterate Convergence of Decentralized Optimistic Gradient Descent/Ascent in Infinite-horizon Competitive Markov Games* on arXiv.