L11 – Week 6
Min-max Optimization: Local Nash and Last iterate convergence

CS 295 Optimization for Machine Learning
Ioannis Panageas
Min-max in bilinear

- Previously we motivated the Last iterate convergence.
- We show that Gradient Descent Ascent (GDA) diverges even for $x^T Ay$.

**Intuition:** Given the bilinear problem below let’s run the continuous GDA.

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} x^T Ay.$$ 

Consider continuous GDA that is the system of odes:

Recall GDA:

$$
x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t),
$$

$$
y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t).
$$
Min-max in bilinear

- Previously we motivated the Last iterate convergence.
- We show that Gradient Descent Ascent (GDA) diverges even for $x^T Ay$.

**Intuition:** Given the bilinear problem below let's run the continuous GDA.

$$\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} x^T Ay.$$ 

Consider **continuous GDA** that is the system of odes:

$$\frac{dx}{dt} = -\eta Ay,$$
$$\frac{dy}{dt} = \eta A^T x.$$ 

**Recall GDA:**

$$x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t),$$
$$y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t).$$
Min-max in bilinear

- Previously we motivated the Last iterate convergence.
- We show that Gradient Descent Ascent (GDA) diverges even for \( x^T Ay \).

**Intuition:** Given the bilinear problem below let’s run the continuous GDA.

\[
\min_{x \in \mathbb{R}^n} \max_{y \in \mathbb{R}^m} x^T Ay.
\]

Consider continuous GDA that is the system of odes:

\[
\frac{dx}{dt} = -\eta Ay, \\
\frac{dy}{dt} = \eta A^T x.
\]

**Recall GDA:**

\[
x_{t+1} = x_t - \eta \nabla_x f(x_t, y_t), \\
y_{t+1} = y_t + \eta \nabla_y f(x_t, y_t).
\]

**Lemma (Cycles).** It holds that \( \|x\|_2^2 + \|y\|_2^2 \) is constant w.r.t \( t \).
Min-max in bilinear

Proof. It suffices to prove

$$\frac{d}{dt} \left\{ \|x\|_2^2 + \|y\|_2^2 \right\} = 0.$$
Min-max in bilinear

Proof. It suffices to prove

$$\frac{d}{dt} \{ \|x\|_2^2 + \|y\|_2^2 \} = 0.$$ 

Observe that

$$\frac{dx_i^2}{dt} = 2x_i \frac{dx}{dt} = -\eta 2x_i (Ay)_i,$$

$$\frac{dy_j^2}{dt} = 2y_j \frac{dy}{dt} = \eta 2y_j (A^T x)_j.$$
Min-max in bilinear

**Proof.** It suffices to prove

\[
\frac{d}{dt} \left\{ \|x\|_2^2 + \|y\|_2^2 \right\} = 0.
\]

Observe that

\[
\frac{dx_i^2}{dt} = 2x_i \frac{dx}{dt} = -\eta 2x_i (Ay)_i, \quad \frac{dy_j^2}{dt} = 2y_j \frac{dy}{dt} = \eta 2y_j (A^T x)_j.
\]

Hence

\[
\frac{d}{dt} \left\{ \|x\|_2^2 + \|y\|_2^2 \right\} = -2\eta x^T Ay + 2\eta x^T Ay = 0.
\]
Min-max in bilinear

• Question: Can we fix this behavior? We can use “optimism” (negative momentum).

\[
\begin{align*}
x_{t+1} &= x_t - \eta \cdot \nabla_x f(x_t, y_t) \\
&\quad + \frac{\eta}{2} \cdot \nabla_x f(x_{t-1}, y_{t-1}) \\
y_{t+1} &= y_t + \eta \cdot \nabla_y f(x_t, y_t) \\
&\quad - \frac{\eta}{2} \cdot \nabla_y f(x_{t-1}, y_{t-1})
\end{align*}
\]
Min-max in bilinear (OGDA)

\[ x_{t+1} = x_t - \eta \cdot \nabla_x f(x_t, y_t) + \frac{\eta}{2} \cdot \nabla_x f(x_{t-1}, y_{t-1}) \]

\[ y_{t+1} = y_t + \eta \cdot \nabla_y f(x_t, y_t) - \frac{\eta}{2} \cdot \nabla y f(x_{t-1}, y_{t-1}) \]
Min-max in bilinear (OGDA)

**Theorem (Convergence).** Consider the bilinear game $x^TAy$ where $A$ is full rank. Optimistic GDA converges pointwise and reaches an $\epsilon$ neighborhood in

$$T := \Theta \left( \frac{\lambda_{\text{max}}(AA^T)}{\lambda_{\text{min}}(AA^T)} \log \frac{1}{\epsilon} \right)$$

choosing learning rate $\eta = \frac{1}{4\sqrt{\lambda_{\text{max}}(AA^T)}}$.
Min-max in bilinear (OGDA)

**Theorem (Convergence).** Consider the bilinear game $x^T Ay$ where $A$ is full rank. Optimistic GDA converges pointwise and reaches an $\epsilon$ neighborhood in

$$T := \Theta \left( \frac{\lambda_{\text{max}}(AA^T)}{\lambda_{\text{min}}(AA^T)} \log \frac{1}{\epsilon} \right)$$

choosing learning rate $\eta = \frac{1}{4 \sqrt{\lambda_{\text{max}}(AA^T)}}$.

The idea behind the proof is to analyze the following dynamical system

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \left( I - \begin{pmatrix} 0 & 2\eta A \\ -2\eta A^T & 0 \end{pmatrix} \right) \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \eta \begin{pmatrix} 0 & 2\eta A \\ -2\eta A^T & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}$$
Min-max in bilinear (OGDA)

**Theorem (Convergence).** Consider the bilinear game $x^T Ay$ where $A$ is full rank. Optimistic GDA converges pointwise and reaches an $\epsilon$ neighborhood in

$$T := \Theta \left( \frac{\lambda_{\text{max}}(AA^T)}{\lambda_{\text{min}}(AA^T)} \log \frac{1}{\epsilon} \right)$$

choosing learning rate $\eta = \frac{1}{4\sqrt{\lambda_{\text{max}}(AA^T)}}$.

The idea behind the proof is to analyze the following dynamical system

$$\begin{pmatrix} x_{t+1} \\ y_{t+1} \end{pmatrix} = \left( I - \begin{pmatrix} 0 & 2\eta A \\ -2\eta A^T & 0 \end{pmatrix} \right) \begin{pmatrix} x_t \\ y_t \end{pmatrix} + \eta \begin{pmatrix} 0 & 2\eta A \\ -2\eta A^T & 0 \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \end{pmatrix}$$

Let’s make it linear system!
Min-max in bilinear (OGDA)

Consider the linear dynamical system

\[
\begin{pmatrix}
  x_{t+1} \\
  y_{t+1} \\
  z_{t+1} \\
  w_{t+1}
\end{pmatrix} =
\begin{pmatrix}
  I & -2\eta A & 0 & \eta A \\
  2\eta A^T & I & -\eta A^T & 0 \\
  I & 0 & 0 & 0 \\
  0 & I & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  y_t \\
  z_t \\
  w_t
\end{pmatrix}
\]
Min-max in bilinear (OGDA)

Consider the linear dynamical system

\[
\begin{pmatrix}
  x_{t+1} \\
  y_{t+1} \\
  z_{t+1} \\
  w_{t+1}
\end{pmatrix}
= \begin{pmatrix}
  I & -2\eta A & 0 & \eta A \\
  2\eta A^T & I & -\eta A^T & 0 \\
  I & 0 & 0 & 0 \\
  0 & I & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  y_t \\
  z_t \\
  w_t
\end{pmatrix}
\]

Observe that

\[
\begin{pmatrix}
  x_{t+1} \\
  y_{t+1} \\
  x_t \\
  y_t
\end{pmatrix}
= \begin{pmatrix}
  I & -2\eta A & 0 & \eta A \\
  2\eta A^T & I & -\eta A^T & 0 \\
  I & 0 & 0 & 0 \\
  0 & I & 0 & 0
\end{pmatrix}
\begin{pmatrix}
  x_t \\
  y_t \\
  x_{t-1} \\
  y_{t-1}
\end{pmatrix}
\]

**Lemma (Eigenvalues).** The matrix above has eigenvalues that are less than one for the appropriate choice of $\eta$. 

Optimization for Machine Learning
Min-max in bilinear (constrained)

Consider the problem

\[
\min_{x \in \Delta_n} \max_{y \in \Delta_m} x^T A y.
\]

- Projected Optimistic GDA not clear if works... Let’s do Optimistic MWU!

\[
\begin{align*}
    x_{i}^{t+1} &= x_{i}^{t} + \frac{1+2\eta(A y^{t})_{i} - \eta(A y^{t-1})_{i}}{\sum_{j} x_{j}^{t}(1+2\eta(A y^{t})_{j} - \eta(A y^{t-1})_{j})}, \\
    y_{i}^{t+1} &= y_{i}^{t} + \frac{1-2\eta(A^{\top} x^{t})_{i} + \eta(A^{\top} x^{t-1})_{i}}{\sum_{j} y_{j}^{t}(1-2\eta(A^{\top} x^{t})_{j} + \eta(A^{\top} x^{t-1})_{j})}.
\end{align*}
\]
Min-max in bilinear (constrained)

Consider the problem

\[
\min_{x \in \Delta_n} \max_{y \in \Delta_m} x^T Ay.
\]

- Projected Optimistic GDA not clear if works... Let’s do Optimistic MWU!

\[
\begin{align*}
    x_i^{t+1} &= x_i^t \frac{1+2\eta(Ay^t)_i - \eta(Ay^{t-1})_i}{\sum_j x_j^t (1+2\eta(Ay^t)_j - \eta(Ay^{t-1})_j)}, \\
    y_i^{t+1} &= y_i^t \frac{1-2\eta(A^T x^t)_i + \eta(A^T x^{t-1})_i}{\sum_j y_j^t (1-2\eta(A^T x^t)_j + \eta(A^T x^{t-1})_j)}.
\end{align*}
\]

Theorem (Convergence). Let \( A \) be the payoff matrix of a zero sum game and the game has a unique Nash equilibrium. It holds that for \( \eta \) sufficiently small (depends on \( n, m, A, \eta \) can be exponentially small in \( n, m \)), starting from uniform distribution \( \lim_{t \to \infty} (x^t, y^t) = (x^*, y^*) \) under OMWU dynamics.
Min-max in general settings

- Min-max theorem is not applicable. How can we solve such a problem?
Min-max in general settings

- Min-max theorem is not applicable. How can we solve such a problem?

Relax the solution concept...
Min-max in general settings

- Min-max theorem is not applicable. How can we solve such a problem?

Relax the solution concept...

**Definition (Local Nash).** A critical point \((x^*, y^*)\) is a local Nash if there exists a neighborhood \(U\) around \((x^*, y^*)\) so that for all \((x, y) \in U\) we have that

\[
f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*).
\]

- Does there always exist a local Nash? Is it a good solution concept?
Min-max in general settings

- Min-max theorem is not applicable. How can we solve such a problem?

**Relax the solution concept...**

**Definition (Local Nash).** A critical point \((x^*, y^*)\) is a local Nash if there exists a neighborhood \(U\) around \((x^*, y^*)\) so that for all \((x, y) \in U\) we have that

\[
f(x^*, y) \leq f(x^*, y^*) \leq f(x, y^*).
\]

- Does there always exist a local Nash? Is it a good solution concept?

**No! Not sure...**
Min-max in general settings

**Theorem (Local Convergence).** Under some mild assumptions on $f(x, y)$ and step-size we have

\[
\text{Local Nash} \subset \text{GDA-stable} \subset \text{OGDA-stable}
\]

**Remarks**

- This is a **local** result!
- Unfortunately the inclusions can be **strict**!
Min-max in general settings

**Theorem (Local Convergence).** Under some mild assumptions on \( f(x, y) \) and step-size we have

\[
\text{Local Nash} \subset \text{GDA-stable} \subset \text{OGDA-stable}
\]

**Remarks**
- This is a local result!
- Unfortunately the inclusions can be strict!

**Lemma (Inclusion strict).** There are functions with critical points that are GDA-stable but not local Nash. An example is \( f(x, y) = -\frac{1}{8}x^2 - \frac{1}{2}y^2 + \frac{6}{10}xy \).
Min-max in general settings

Proof. Let \( f(x, y) = -\frac{1}{8}x^2 - \frac{1}{2}y^2 + \frac{6}{10}xy \).
Computing the Jacobian of the update rule of OGDA at \((0, 0)\) we get

\[
J_{\text{GDA}} = \begin{pmatrix}
1 + \frac{1}{4}\eta & -\frac{6}{10}\eta \\
\frac{6}{10}\eta & 1 - \eta
\end{pmatrix}
\]

Both eigenvalues of \( J_{\text{GDA}} \) have magnitude less than 1 (for any \( 0 < \alpha < 1.34 \)). GDA is contracting around \((0, 0)\).

However it is clear that \((0, 0)\) is not a local Nash. Why?
Min-max in general settings
Conclusion

• Introduction to min-max optimization.
  – Negative momentum for last iterate convergence.
  – Bilinear unconstrained and constrained
  – Local Nash

• Next lecture we will talk about Multi-Armed Bandits.