CSC 2541: Neural Net Training Dynamics
Lecture 10 - Differentiable Games

Guodong Zhang

University of Toronto, Winter 2021
Differentiable Games

- So far, we have been exclusively discussing minimization problems:

  \[ z^* \in \arg \min_z f(z) \]

  (minimizing a single objective)

- What if we have multiple players and each of them optimizes their own objective?

  \[ z_i^* \in \arg \min_{z_i} f_i(z_i, z_{-i}^*) \]

  (now, we’re trying to find local/global Nash equilibrium)

- Examples: Generative Adversarial Networks, multi-agent RL, PCA, off-policy evaluation, robust optimization, ...
Generative Adversarial Networks

$$\min_G \max_D f(D, G) = \mathbb{E}_{x \sim p_{\text{data}}} \left[ \log(D(x)) \right] + \mathbb{E}_{z \sim p_{z}} \left[ \log(1 - D(G(z))) \right]$$
Nash Equilibrium

\[ z_i^* \in \arg\min_{z_i} f_i(z_i, z_{-i}^*) \]

---

**THE PRISONER’S DILEMMA**

<table>
<thead>
<tr>
<th>( A ) stays silent (cooperates)</th>
<th>( B ) stays silent (cooperates)</th>
<th>( B ) stays silent (cooperates)</th>
<th>( B ) betrays ( A ) (defects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both serve 1 year</td>
<td>A goes free, B serves 3 years</td>
<td>A serves 3 years, B goes free</td>
<td>Both serve 2 years</td>
</tr>
<tr>
<td>A stays silent (cooperates)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B betrays A (defects)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Differentiable Games

- Differentiable games are much harder to solve (even only two-player)!
  \[
  \min_x \max_y f(x, y)
  \]
  (It’s called minimax optimization, saddle-point problem)

- Why are they more challenging?
  - In the nonconvex-nonconcave case, local Nash equilibria might not exist. Even when they exist, finding a local Nash equilibrium is PPAD-complete.
  - In the convex-concave setting, standard gradient descent can diverge with any positive step size or enter limit cycles.
  - Even when gradient descent converges, the rate of convergence may be too slow in practice (our focus today).

Left: bilinear game with \( f(x, y) = 10xy \)
Right: \( f(x, y) = 0.5x^2 + 10xy - 0.5y^2 \)
Today

- We are going to focus on two-player, strongly-convex strongly-concave, zero-sum games.

\[
\min_x \max_y f(x, y)
\]

(many insights carry over to more general settings)

- Strong duality (minimax theorem) holds, i.e.,

\[
\min_x \max_y f(x, y) = \max_y \min_x f(x, y)
\]

- Local Nash equilibrium is global and it is unique.

- Even for this simple setting, convergence can be slow because the “rotational force” necessitate extremely small learning rates.
A Closer Look of Linear Case

- Consider the general dynamics:
  \[ z^{(k+1)} = z^{(k)} - \eta F(z^{(k)}) \]
  (where \( F \) is a vector field)

- Linear case: \( F(z) = Hz \)
  - Minimization: \( H \) is symmetric and all eigenvalues are real
  - Differentiable Games: \( H \) is non-symmetric and can have complex eigenvalues (with large imaginary parts)

\[
\begin{align*}
\min f(x, y) &= 0.5x^2 + 0.5y^2 \\
\min_x \max_y f(x, y) &= 0.5x^2 + 10xy - 0.5y^2
\end{align*}
\]
Simultaneous Gradient Descent-Ascent
Simultaneous Gradient Descent-Ascent

- **Sim-GDA** is a naïve extension to gradient descent to the game setting

  \[
  \begin{align*}
  x^{(k+1)} &\leftarrow x^{(k)} - \eta \nabla_x f(x^{(k)}, y^{(k)}) \\
y^{(k+1)} &\leftarrow y^{(k)} + \eta \nabla_y f(x^{(k)}, y^{(k)})
  \end{align*}
  \]

- We can compactly write it as \( z^{(k+1)} \leftarrow z^{(k)} - \eta F(z^{(k)}) \) where \( z = [x^\top, y^\top]^\top \) and \( F(z) = [\nabla_x f(x, y)^\top, -\nabla_y f(x, y)^\top]^\top \).

- Assuming a quadratic problem \( f(x, y) = \frac{1}{2} x^\top A x + x^\top B y - \frac{1}{2} y^\top C y \)
  - We have the dynamics:

    \[
    z^{(k+1)} \leftarrow (I - \eta H)z^{(k)}
    \]

    where \( z = [x^\top, y^\top]^\top \) and \( H = \begin{bmatrix} A & B \\ -B^\top & C \end{bmatrix} \)
Convergence Analysis of Sim-GDA

- **Setting:** Smooth and strongly-monotone games
  - Define the gradient vector field $F(z) = [\nabla_x f(x, y)^\top, -\nabla_y f(x, y)^\top]^\top$
  - **Lipschitz Smooth:** a vector field $F$ is Lipschitz if for any $z_1, z_2$ and a parameter $L$:
    \[
    \|F(z_1) - F(z_2)\| \leq L\|z_1 - z_2\|
    \]
  - **Strongly Monotone:** a vector field $F$ is strongly monotone if for any $z_1, z_2$ and a parameter $\mu$:
    \[
    (F(z_1) - F(z_2))^\top (z_1 - z_2) \geq \mu\|z_1 - z_2\|^2
    \]
  - **Condition number:** $\kappa \triangleq \frac{L}{\mu}$
  - **Quadratic case:** $F(z) = Hz$ where $H \succeq \mu I$ and $\|H\| \leq L$
Convergence Analysis of Sim-GDA

- Recall that the dynamics of Sim-GDA: $z^{(k+1)} \leftarrow (I - \eta H)z^{(k)}$
- Its convergence rate is $\min_\eta \rho(I - \eta H) = \min_\eta \max_{\lambda \in \text{Sp}(H)} \|1 - \eta \lambda\|$

The best convergence rate is limited by the eigenvalue $\lambda = \mu + \sqrt{L^2 - \mu^2}i$.

The optimal convergence rate is $1 - \frac{1}{\kappa^2}$, which implies that Sim-GDA takes roughly $\mathcal{O}(\kappa^2)$ steps to converge. Recall that gradient descent only takes $\mathcal{O}(\kappa)$ steps to converge in minimizing a strongly-convex function!
Could we accelerate the convergence of Sim-GDA?
# Alternating Gradient Descent-Ascent

- **Alt-GDA** updates multiple players sequentially:
  \[
  \begin{align*}
  x^{(k+1)} &\leftarrow x^{(k)} - \eta \nabla_x f(x^{(k)}, y^{(k)}) \\
  y^{(k+1)} &\leftarrow y^{(k)} + \eta \nabla_y f(x^{(k+1)}, y^{(k)})
  \end{align*}
  \]

- Alt-GDA converges with $O(\kappa)$ steps (which matches the coarse lower-bound).
- The result could be extended to n-player setting (ongoing work).
- In the bilinear case, Alt-GDA is a symplectic integrator applied on the continuous dynamic.
- The discussion of simultaneous and alternating updates dates back to the Jacobi and Gauss-Seidel methods in numerical linear algebra, see the celebrated Stein-Rosenberg theorem.

---

see more details in “Don’t fix what ain’t broke: near-optimal local convergence of alternating gradient descent-ascent for minimax optimization”
Alternating Gradient Descent-Ascent

- Consider the quadratic problem $f(x, y) = \frac{1}{2} x^\top A x + x^\top B y - \frac{1}{2} y^\top C y$.
- We have Alt-GDA as the following form:

$$
\begin{bmatrix}
    x^{(k+1)} \\
    y^{(k+1)}
\end{bmatrix}
\leftarrow
\begin{bmatrix}
    I - \eta A \\
    \eta B^\top (I - \eta A)
\end{bmatrix}
\begin{bmatrix}
    I - \eta C - \eta^2 B^\top B
\end{bmatrix}
\begin{bmatrix}
    x^{(k)} \\
    y^{(k)}
\end{bmatrix}
$$

- Recall Sim-GDA dynamics for the quadratic case:

$$
\begin{bmatrix}
    x^{(k+1)} \\
    y^{(k+1)}
\end{bmatrix}
\leftarrow
\begin{bmatrix}
    I - \eta A \\
    \eta B^\top
\end{bmatrix}
\begin{bmatrix}
    I - \eta C
\end{bmatrix}
\begin{bmatrix}
    x^{(k)} \\
    y^{(k)}
\end{bmatrix}
$$

- Alt-GDA allows us to use larger step sizes. The optimal step size for Sim-GDA is $\frac{\mu}{L^2}$ while the optimal one for Alt-GDA is roughly $\frac{1}{L}$.
• We are implicitly using alternating updates in GAN training.

DCGAN on CIFAR-10. **Left**: SGD as base optimizer; **Right**: AMSGrad as base optimizer.
Negative Momentum

- **Negative momentum** is basically Heavy-ball momentum with a negative damping value:

\[ z^{(k+1)} \leftarrow (1 + \beta) z^{(k)} - \beta z^{(k-1)} - \eta F(z^{(k)}) \]

- Intuition: negative momentum reduces the imaginary parts of complex eigenvalues, and hence suppresses the rotational behaviour. (recall the rate of Sim-GDA was limited by the eigenvalue \( \lambda = \mu + \sqrt{L^2 - \mu^2 i} \))

- Negative momentum converges in \( O(\kappa^{1.5}) \) steps, which is slightly faster than Sim-GDA (recall the complexity of \( O(\kappa^2) \)). However, this rate is suboptimal as some other algorithms converge in \( O(\kappa) \) steps.

- Proving this convergence rate is extremely **hard**! Need to leverage the connection between Chebyshev polynomial and Heavy-ball momentum. Check out my paper “On the suboptimality of negative momentum for minimax optimization”.

Negative Momentum

- **Negative momentum** is basically Heavy-ball momentum with a negative damping value:

\[
    z^{(k+1)} \leftarrow (1 + \beta)z^{(k)} - \beta z^{(k-1)} - \eta F(z^{(k)})
\]

- **Fact:** Heavy-ball momentum with an optimally-tuned damping parameter is optimal when all eigenvalues of $H$ fall within an ellipse in the complex plane.

\[
    \left(\Re \lambda - d\right)^2 + \left(\Im \lambda\right)^2 \leq \frac{a^2}{b^2}
\]

- $a > b$: optimal $\beta$ is positive
- $a < b$: optimal $\beta$ is negative
- $a = b$: optimal $\beta$ is zero

- **Another fun fact:** Negative momentum retains the same convergence rate when the function $f$ is not quadratic. (Recall that Heavy-ball momentum only achieves acceleration when $f$ is quadratic)

see more details in “Don’t fix what ain’t broke: near-optimal local convergence of alternating gradient descent-ascent for minimax optimization”
Negative Momentum

Image Credit: Negative Momentum for Improved Game Dynamics
The proximal point method (Rockafeller, 1976) is an implicit method:

$$z^{(k+1)} \leftarrow z^{(k)} - \eta F(z^{(k+1)})$$

Intuition: compute gradient at a future point, but it is not implementable in many cases (chicken and egg situation).

In optimization, the proximal point method is largely regarded as a “conceptual” guiding principle for accelerating optimization algorithms. NAG can be derived from the proximal point method (see “From Proximal Point Method to Nesterov’s Acceleration” paper).

It can be shown that for smooth and strongly monotone games, the proximal point method converges linearly for any $\eta$:

$$\|z^{(k)} - z^*\|^2 \leq \left( \frac{1}{1 + 2\eta \mu} \right)^k \|z^{(0)} - z^*\|^2$$

check out the proof in “A Unified Analysis of First-Order Methods for Smooth Games via Integral Quadratic Constraints”
Could we approximate proximal point method and achieve acceleration?
Extra-gradient method

- The Extra-gradient method computes the gradient with one-step lookahead (extrapolated gradient):

  \[ z^{(k+1/2)} \leftarrow z^{(k)} - \eta F(z^{(k)}) \]

  \[ z^{(k+1)} \leftarrow z^{(k)} - \eta F(z^{(k+1/2)}) \]

- It was first proposed by Korpelevich in 70’s to solve monotone variational inequality problem.

- It was recently re-introduced by Gidel, et.al (2019) and Mokhtari, et.al (2019) in the context of differentiable games and minimax optimization.

- Over the last three years, more than 10 papers discussed the extra-gradient method in different settings.
Extra-gradient method

- The extra-gradient method computes the gradient with one-step lookahead:

\[ z^{(k+1/2)} \leftarrow z^{(k)} - \eta F(z^{(k)}) \]
\[ z^{(k+1)} \leftarrow z^{(k)} - \eta F(z^{(k+1/2)}) \]

- Intuition: approximate \( F(z^{(k+1)}) \) with \( F(z^{(k+1/2)}) \), hoping to inherit the convergence properties of proximal point method.

- Formally, it can shown that starting with the same \( z^{(k)} \), the solution of extra-gradient \( z_{eg}^{(k+1)} \) after one step is relatively close to the solution of proximal point method \( z_{ppm}^{(k+1)} \):

\[ \|z_{eg}^{(k+1)} - z_{ppm}^{(k+1)}\| \leq o(\eta^2) \]

- Under the same set of assumptions, the extra-gradient method converges linearly

\[ \|z^{(k)} - z^*\|^2 \leq \left(1 - \frac{1}{2\kappa}\right)^k \|z^{(0)} - z^*\|^2 \]

see more details in “A Unified Analysis of Extra-gradient and Optimistic Gradient Methods for Saddle Point Problems: Proximal Point Approach”
Optimistic Gradient Method

- Optimistic Gradient update rule:

\[ \mathbf{z}^{(k+1)} \leftarrow \mathbf{z}^{(k)} - \eta F(2\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}) \]

- You could understand it as replacing the first step of extra-gradient with the following:

\[ \mathbf{z}^{(k+1/2)} \leftarrow \mathbf{z}^{(k)} + \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)} \]

- It has pretty much the same convergence properties as extra-gradient but only compute the gradient once in every iteration!

- Under the same set of assumptions, optimistic gradient converges linearly

\[ \|\mathbf{z}^{(k)} - \mathbf{z}^*\|^2 \leq \left(1 - \frac{1}{4\kappa}\right)^k \|\mathbf{z}^{(0)} - \mathbf{z}^*\|^2 \]
In which case should we use optimistic gradient method?

- In the situation that you are only allowed to query the gradient once every iteration.
- In (no-regret) online learning with an arbitrary adversary, extra-gradient is not no-regret.
Comparison between different algorithms

Distances to the optimum as a function of iterations on a quadratic minimax problem.
Important directions that I didn’t cover

- General convex-concave setting (without strong convexity/concavity). In this setting, one can only achieve sublinear convergence (see e.g., [1,2]).
- Stochastic settings (see e.g., [3, 4]).
- Second-order methods (see e.g., [5, 6]).
- Sequential games when $f$ is nonconvex-nonconcave (see e.g., [7, 8]). In this case, Nash equilibrium might not exist and other equilibrium concepts were proposed. Moreover, the order of different players matters since $\min \max \neq \max \min$

[1] Convergence rate of $o(1/k)$ for optimistic gradient and extragradient methods in smooth convex-concave saddle point problems.
[2] Last iterate is slower than averaged iterate in smooth convex-concave saddle point problems.
Summary

- Differentiable game dynamics is more complex.
- In the nonconvex setting, Nash equilibrium might not exist. Even when it exists, finding local solution is much harder than finding local minima in minimization.
- Even for convex-concave two-player setting, standard algorithms could either diverge or cycle around the equilibrium.
- When converges, rotational component (caused by complex eigenvalues) would slow down convergence.
- When it comes to algorithm choice, alternating updates significantly outperform simultaneous updates and negative momentum is preferred in many cases.
- Extra-gradient and optimistic gradient method approximate proximal point method, which accelerate the convergence.