IFT 6756 - Lecture N
(LECTURE TITLE)

This version of the notes has not yet been thoroughly checked. Please report any bugs to the scribes or instructor.

Scribes **Instructor:** Gauthier Gidel
Winter 2021: [Arnold(Zicong) Mo, Pascal Jutras-Dubé, and Mojtaba Faramarzi.]

1 Summary

1.1 Reference

Today's lecture is based on these three papers:

- Gidel, Gauthier, et al. "A variational inequality perspective on generative adversarial networks." ICLR 2019 [2]
- Mokhtari, Aryan, Asuman Ozdaglar, and Sarath Pattathil. "A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach." International Conference on Artificial Intelligence and Statistics. PMLR, 2020. [3]
  - Provides an analysis on extra gradient by considering it as an approximation of the sub-optimal point method.
- Azizian, Waïss, et al. "A tight and unified analysis of gradient-based methods for a whole spectrum of differentiable games." International Conference on Artificial Intelligence and Statistics. PMLR, 2020.[1]
  - Analyze in depth about extra-gradient. Showed that doing multiple step is useless and only gets marginal improvement.

1.2 Last Time

Our goal is to solve this minmax optimization problem, where the payoff is convex-concave:

$$\min_{\theta} \max_{\phi} L(\theta, \phi)$$

Example 1. *Bi-linear minmax:*

$$\min_{\theta} \max_{\phi} (\theta - \theta^*)^T A(\phi - \phi^*)$$

Linear in θ and ϕ, and we have the bi-linear product with a matrix in them.

We learned 3 methods last time:

Definition 2 (Simultaneous Gradient Descent-Ascent(Sim-GDA))**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t \end{cases} \tag{1}$$

Definition 3 (Alternated Gradient Descent-Ascent(Alt-GDA))**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases} \tag{2}$$

Definition 4 (Proximal Point Method)**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\nabla_\theta L(\theta_{t+1}, \phi_{t+1}) \\ \phi_{t+1} = \phi_t + \eta\nabla_\theta L(\theta_{t+1}, \phi_{t+1}) \end{cases} \tag{3}$$

Although it converges, this is an implicit update and it is not practical.

1.3 Today

Today's goal is to learn a method that is practical, has similar properties as the proximal point method, but revert the inconvenient of being an implicit method.

2 Variational Inequality Perspective

Proposing a way to see this update on a more compact way, so that we won't have lines of equations. At the end, we only care about the gradient-based updates:

$$F(\theta_t, \phi_t) := \begin{pmatrix} \nabla_\theta L(\theta_t, \phi_t) \\ -\nabla_\phi L(\theta_t, \phi_t) \end{pmatrix}$$

We see that everything depends on the pair of (θ_t, ϕ_t):

$$w_t := (\theta_t, \phi_t)$$

2.1 Examples of the VIP

Example 5 (Sim-GDA)**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_t \end{cases} \tag{4}$$

VIP form:

$$w_{t+1} = w_t - \eta F(w_t)$$

Example 6 (Prox-Point)**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases} \tag{5}$$

VIP form:

$$w_{t+1} = w_t - \eta F(w_{t+1})$$

Example 7 (Alt-GDA)**.**

$$\begin{cases} \theta_{t+1} = \theta_t - \eta\phi_t \\ \phi_{t+1} = \phi_t + \eta\theta_{t+1} \end{cases} \tag{6}$$

No applicable to VIP

2.2 Goal

Our goal is to find a stationary point of the vector fields:

$$F(w^*) = 0$$

In zero sum game, this is equivalent to find a nash in the game where the pair is convex concave. We have reduce the problem to finding a stationary point. We want to follow the vector field until finding it.

3 Extragradient

Proximal Point method:

$$w_{t+1} = w_t - \eta F(w_{t+1})$$

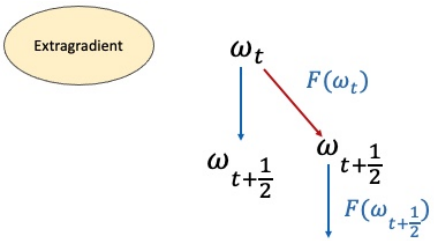
Idea is to approximate w_{t+1} with a gradient step, by doing a simple update step:

$$w_{t+1/2} = w_t - \eta F(w_t)$$

Replace w_{t+1} by the above approximation, we get:

$$w_{t+1} = w_t - \eta F(w_{t+1/2})$$

This is the extragradient, and the method is explicit.



3.1 Exercise

3.1.1 Exercise 1 - Update Rules for EG

Write the updates rules for EG for the following case

$$\min_{\theta} \max_{\phi} \theta, \phi$$

Answer:

$$\begin{cases} \theta_{t+1} = \theta_t - \eta(\phi_t + \eta\phi_t) = \theta_t - \eta(\phi_{t+1/2}) \\ \phi_{t+1} = \phi_t + \eta(\theta_t - \eta\phi_t) = \phi_t + \eta(\theta_{t+1/2}) \end{cases} \tag{7}$$

3.1.2 Exercise 2 - For a small step size

Show that for a small enough step-size:

$$\theta_t^2 + \phi_t^2 \leq p^t(\theta_0^2 + \phi_0^2) \quad where\, 0 < p < 1$$

Proof:

$$\begin{aligned} \theta_t^2 + \phi_t^2 &= ((\theta_t - \eta(\phi_t + \eta\theta_t))^2 + ((\phi_t + \eta(\theta_t - \eta\phi_t))^2 \\ &= (\theta^2 + \phi^2)(1 - \eta^2 + \eta^4) \end{aligned}$$

Therefore EG converges iff $\eta < 1$.

3.2 Standard Assumption

Definition 8 (Monotone operator)**.**

$$< F(w) - F(w'), w - w' > \geq 0, \forall w, w'$$

3.2.1 Intuition

Monotonicity implies:

$$< F(w), w^* - w_t > \geq 0, \forall w, w'$$

It is a generalization of convexity.

3.2.2 Exercice 1: prove F is monotone

For $\min_{\theta} \max_{\phi} \theta^T A \phi$, we have:

$$F(\theta_t, \phi_t) = \begin{pmatrix} A\phi \\ -A^T\phi \end{pmatrix}$$

Show that F is monotone.
Proof:

$$< F(\theta_t, \phi_t) - F(\theta'_t, \phi'_t), (\theta_t, \phi_t) - (\theta'_t, \phi'_t) >$$
$$= (\phi - \phi')A(\theta - \theta') - (\phi - \phi')^T A(\theta - \theta') = 0 \quad \forall \theta, \theta', \phi, \phi'$$

3.2.3 Examples

Example 9. *The vector field*

$$F(x, y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$

is monotone.

Example 10. *The vector field*

$$F(x, y) = \begin{pmatrix} (y - 0.5)(y + 0.5) \\ -x \end{pmatrix}$$

is not monotone.

A monotone vector cannot have two connected optimal points.

Example 11. *The vector field*

$$F(x, y) = \begin{pmatrix} -y - x \\ x - y \end{pmatrix}$$

is monotone.

3.3 Convergence of Extra Gradient (General case)

Recall from the lecture on gradient descent, we had

Lemma 12 (Convergence of Gradient Descent)**.**

$$\|\theta_{t+1} - \theta^*\|_2^2 = \|\theta_t - \theta^*\|_2^2 - 2\eta g(\theta_t)^T(\theta_t - \theta^*) + \|\theta_{t+1} - \theta_t\|_2^2$$

The second term in the right hand-side is the local progress due to monotonicity and the last term is the error due to discretization. Roughly, what the lemma says is that we can decrease the distance to the optimum if the inner product in the second term is positive, meaning we are progressing in the right direction, and if η is small enough, meaning the error due to discretization is also small. For Extra Gradient, we can show the following similar lemma [1].

Lemma 13 (Convergence of Extra Gradient)**.**

$$\|\omega_{t+1} - \omega^*\|_2^2 = \|\omega_t - \omega^*\|_2^2 - 2\eta F(\omega_{t+1/2})^T(\omega_{t+1/2} - \omega^*) + \eta^2\|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2 - \|\omega_{t+1/2} - \omega_t\|_2^2$$

Again, the second term is the progress due to monotonicity and the term $\eta^2\|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2 - \|\omega_{t+1/2} - \omega_t\|_2^2$ is the error due to discretization. Unlike before, if η is small enough the discretization's error can be made negative and thus induce progress. We can also enforce $\|F(\omega_{t+1/2}) - F(\omega_t)\|_2^2$ not to be too big. The natural assumption to do so is to say that the vector field operator is Lipschitz.

Definition 14 (Lipschitz Operator)**.** *The vector field operator F is Lipschitz if there exists $0 < L < \infty$ such that*

$$\|F(\omega) - F(\omega')\| \leq L\|\omega - \omega'\|$$

for all ω and ω'.

3.3.1 Examples

Example 15. *The vector field*

$$F(x, y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$

is Lipschitz.

Proof. We know that

$$F(\omega) - F(\omega') = \nabla F(\widetilde{\omega})(\omega - \omega')$$

with $\widetilde{\omega} \in [\omega, \omega']$ so

$$\|F(\omega) - F(\omega')\| \leq \|\nabla F(\widetilde{\omega})\|\|(\omega - \omega')\|.$$

Therefore, F is Lipschitz if $\|\nabla F(\widetilde{\omega})\| \leq L$ for all $\widetilde{\omega}$. In our case,

$$\nabla F(x, y) = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$$

and we know that $\|\nabla F(x, y)\|$ is smaller than its largest singular value

$$\|\nabla F(x, y)\| \leq \sigma_{max} \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} < \infty.$$

Thus we can pose

$$L = \sigma_{max} \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$$

to complete the proof. □

Example 16. *The vector field*

$$F(x, y) = \begin{pmatrix} (y - 0.5)(y + 0.5) \\ -x \end{pmatrix}$$

is Lipschitz if x and y are bounded but is not Lipschitz if x and y go to the infinity.

Example 17. *The vector field*

$$F(x, y) = \begin{pmatrix} -sign(y) \\ sign(x) \end{pmatrix}$$

is not Lipschitz.

Proof. Consider a point where the sign of one coordinate changes, for example $(x, y) = (1, 0)$. For $\epsilon > 0$, we have

$$\|F(1, \epsilon) - F(1, -\epsilon)\| = \|(-1, 1) - (1, 1)\| = 2$$

and

$$\|(1,\epsilon)-(1,-\epsilon)\| = 2\epsilon.$$

Thus, for all $\epsilon > 0$, L has to be larger than $\frac{1}{\epsilon}$ because we want

$$2 = \|F(1,\epsilon) - F(1,-\epsilon)\| \leq L\|(1,\epsilon)-(1,-\epsilon)\| = L\,2\epsilon$$

but $\frac{1}{\epsilon} \to \infty$ when $\epsilon \to 0$. The Lipschitz property cannot hold. □

If the Lipschitz assumption holds, we know from the lemma 13 on the convergence of the Extra Gradient that

$$\|\omega_{t+1} - \omega^*\|_2^2 \leq \|\omega_t - \omega^*\|_2^2 - 2\eta F(\omega_{t+1/2})^T(\omega_{t+1/2} - \omega^*) + (\eta^2 L^2 - 1)\|\omega_{t+1/2} - \omega_t\|_2^2 < \|\omega_t - \omega^*\|_2^2,$$

meaning that the distance to the optimum decreases.

3.4 Strongly Monotone Operator

Definition 18. *The vector field operator F is strongly monotone if there exists $\mu > 0$ such that*

$$\langle F(\omega) - F(\omega'),\, \omega - \omega'\rangle \geq \mu\|\omega - \omega'\|_2^2$$

for all ω and ω'.

3.4.1 Examples

Example 19. *The vector field*

$$F(x,y) = \begin{pmatrix} -y \\ x - y \end{pmatrix}$$

is not strongly monotone.

Proof. For points of the form $(x,y) = (x,0)$, we have that

$$\langle F(x,y) - F(x',y'),\, (x,y) - (x',y')\rangle = \langle \begin{pmatrix} 0 \\ x - x' \end{pmatrix},\, \begin{pmatrix} x - x' \\ 0 \end{pmatrix}\rangle = 0$$

□

Example 20. *The vector field*

$$F(x,y) = \begin{pmatrix} -y \\ x \end{pmatrix}$$

is not strongly monotone.

Proof. We have that

$$\begin{aligned}\langle F(x,y) - F(x',y'),\, (x,y) - (x',y')\rangle &= \langle \begin{pmatrix} y' - y \\ x - x' \end{pmatrix},\, \begin{pmatrix} x - x' \\ y - y' \end{pmatrix}\rangle \\ &= (y' - y)(x - x') - (y' - y)(x - x') \\ &= 0.\end{aligned}$$

□

Example 21. *The vector field*

$$F(x,y) = \begin{pmatrix} -y - x \\ x - y \end{pmatrix}$$

is strongly monotone.

Proof. We have that

$$\langle F(x,y) - F(x',y'),\, (x,y) - (x',y')\rangle = \frac{1}{2}(x - x')^2 + \frac{1}{2}(y - y')^2 = \frac{1}{2}\|(x,y) - (x',y')\|^2.$$

We can choose $\mu = 1/2$. □

3.5 Convergence Result

3.5.1 Theorem:L-Lipchitz operator

If The operator is strongly monotone: (for $\eta = 1/4$ L)

$$\|\omega_t - \omega^*\|_2^2 \leq \left(1 - \frac{\mu}{4L}\right)^t \|\omega_0 - \omega^*\|_2^2 \tag{8}$$

3.5.2 Proof

Lemma:

$$\|\omega_{t+1} - \omega^*\|^2 \leq \|\omega_t - \omega^*\|^2 - 2\eta F(\omega_{t+\frac{1}{2}})^T(\omega_{t+\frac{1}{2}} - \omega^*) + (\eta^2 L^2 - 1)\|\omega_{t+\frac{1}{2}} - \omega_t\|^2 \tag{9}$$

$$\because \text{Strong monotonicity and} \quad F(\omega^*) = 0 \tag{10}$$

$$\leq \|\omega_t - \omega^*\|^2 - 2\eta\mu\|\omega_{t+\frac{1}{2}} - \omega^*\|^2 + (\eta^2 L^2 - 1)\|\omega_{t+\frac{1}{2}} - \omega_t\|^2 \tag{11}$$

Since we know that:

$$\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \implies \|a\|^2 \leq \|a\|^2 - \frac{1}{2}\|a + b\|^2 \tag{12}$$

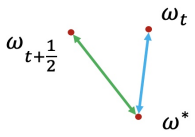
Therefore, we can consider a and b as follows:

$$a = \omega_{t+\frac{1}{2}} - \omega^* \quad \text{and} \quad b = \omega_{t+\frac{1}{2}} - \omega_t \tag{13}$$

According to Equ. 11 and 12 we have:

$$-\|\omega_{t+\frac{1}{2}} - \omega^*\|^2 \leq \|\omega_{t+\frac{1}{2}} - \omega_t\|^2 - \frac{1}{2}\|\omega_t - \omega^*\|^2$$

So geometrically we can consider the distances as follow: by substituting these into Equ. 11 we can get:



$$\|\omega_{t+1} - \omega^*\|^2 \leq \|\omega_t - \omega^*\|^2(1 - \mu\eta) + \underbrace{2\mu\eta\|\omega_{t+\frac{1}{2}} - \omega_t\|^2 + (\eta^2 L^2 - 1)\|\omega_{t+\frac{1}{2}} - \omega_t\|^2}_{\text{good if} \quad \mu\eta + 2\eta^2 L^2 - 1 < 0}$$

$$\eta = \frac{1}{4L} \implies \eta\mu + \eta^2 L^2 - 1 = \frac{\mu}{2L} + \frac{1}{16} - 1 < 0 \quad (\because \mu < L)$$

Overall we get:

$$\|\omega_{t+1} - \omega^*\|^2 \leq (1 - \frac{\mu}{2L})\|\omega_t - \omega^*\|^2$$

3.6 Optimistic Method

In optimistic method, we have Extragradient that has the following update:

$$\omega_{t+1/2} = \omega_t - \eta F\left(\omega_t\right)$$
$$\omega_{t+1} = \omega_t - \eta F\left(\omega_{t+1/2}\right)$$

We compute the extrapolated point and then we do update for ω_t using the gradient at extrapolated point. Instead of computing two gradients we can use optimistic method using the past gradient. So we can substitute $\eta F(\omega_t)$ with $\eta F(\omega_{t-1/2})$. Therefore, in optimistic method we have:

$$\omega_{t+1/2} = \omega_t - \eta F\left(\omega_{t-1/2}\right)$$
$$\omega_{t+1} = \omega_t - \eta F\left(\omega_{t+1/2}\right)$$

Now, we can summarize the extrapolation form past as illustrated in figure 3.6 for the Extragradient Method and figure 3.6 for the Optimistic Method.

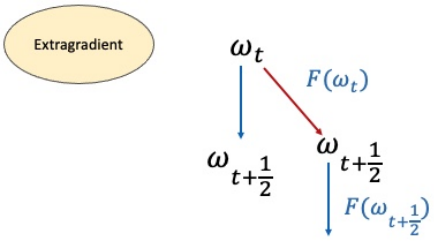


Figure 1: The Extragradient method.

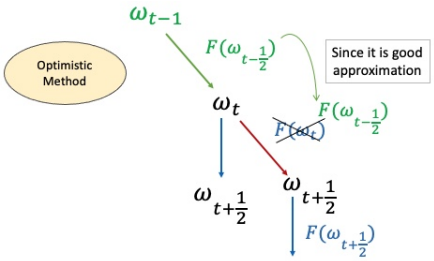


Figure 2: The Optimistic method.

Better understating of Optimistic method:

$$\begin{aligned}\omega_{t+1} \leftarrow & \omega_t - \eta F\left(\omega_t - \eta f\left(\omega_t\right)\right) \approx \omega_t \\ & \approx \omega_t - \eta f\left(\omega_t\right) + \underbrace{\eta^2 \nabla F\left(\omega_\mu\right) f\left(\omega_t\right)}_{\text{corrective term}}\end{aligned}$$

$\nabla F\left(\omega_\mu\right)$ contains the curvature. And the corrective term is approximating the curvature and push towards the optimum point.
Figure 3.6 compares the Extragradient and Optimistic Methods.

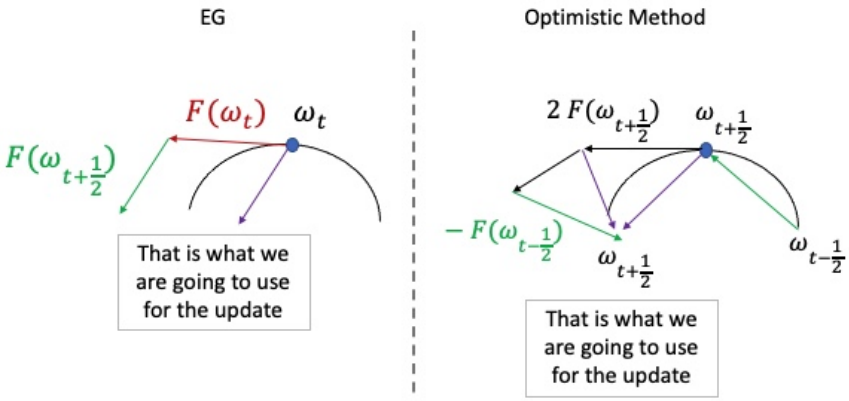


Figure 3: The Optimistic method vs. Extragradient Method.

References

[1] W. Azizian, I. Mitliagkas, S. Lacoste-Julien, and G. Gidel. A tight and unified analysis of gradient-based methods for a whole spectrum of games, 2020.

[2] G. Gidel, H. Berard, G. Vignoud, P. Vincent, and S. Lacoste-Julien. A variational inequality perspective on generative adversarial networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=r1laEnA5Ym`.

[3] A. Mokhtari, A. Ozdaglar, and S. Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach, 2019.