# Frank-Wolfe for White Box Adversarial Attacks

Department of Mathematics "Tullio Levi-Civita"
Master's Degree in Data Science

Eleonora Brasola
Alberto Cocco
Greta Farnea

Università degli Studi di Padova

# Introduction to Adversarial attacks



- What are in general
- Untargeted and targeted attacks
- White and Black box attack

**Algorithm 1** PGM

1: **for** $k = 1 \dots$ **do**
2: Set $\bar{x}_k = \rho_C(x_k + s_k \nabla f(x_k))$ ▷ if untargeted attack, with $s_k > 0$
3: Set $\bar{x}_k = \rho_C(x_k - s_k \nabla f(x_k))$ ▷ if targeted attack, with $s_k > 0$
4: If $\bar{x}_k$ satisfies some specific condition, then STOP
5: Set $x_{k+1} = x_k + \gamma_k(\bar{x}_k - x_k)$ ▷ with $\gamma_k \in (0, 1]$
6: **end for**

**Algorithm 2** MI-FGSM

1: Fix $g_0 = 0$ and $x_0^*$
2: **for** $t = 0$ to $T - 1$ **do**
3: Input $x_t$ and obtain the gradient $\nabla_x f(x_t)$
4: Accumulate velocity

$$g_{t+1} = \beta \cdot g_t + \frac{\nabla_x f(x_t)}{\|\nabla_x f(x_t)\|_1}$$

5: Update

$$\begin{aligned} x_{t+1} &= x_t + \gamma \cdot \mathrm{sign}(g_{t+1}) && \text{if untargeted attack} \\ x_{t+1} &= x_t - \gamma \cdot \mathrm{sign}(g_{t+1}) && \text{if targeted attack} \end{aligned}$$

6: **end for**

# FW-white



**Algorithm 3** FW-White

1: Set $x_0 = x_{\text{ori}}$, $m_{-1} = -\nabla_x f(x_0)$ if untargeted attack, $m_{-1} = \nabla_x f(x_0)$ if targeted attack
2: **for** $t = 0$ to $T - 1$ **do**
3: $\quad m_t = \beta \cdot m_{t-1} - (1 - \beta) \cdot \nabla f(x_t)$ ▷ if untargeted
4: $\quad m_t = \beta \cdot m_{t-1} + (1 - \beta) \cdot \nabla f(x_t)$ ▷ if targeted
5: $\quad v_t = \text{argmin}_{x \in C} \langle x, m_t \rangle = -\epsilon \cdot \text{sign}(m_t) + x_{\text{ori}}$
6: $\quad d_t = v_t - x_t$
7: $\quad x_{t+1} = x_t + \gamma d_t$
8: **end for**

# Demo.py



Now we hand over the word to Alberto so that we can see a demo of our project.

# Indice

Università degli Studi di Padova

Ogni giocatore ha lo scopo di massimizzare il proprio payoff.
Si trovano $N$ problemi di Controllo Ottimo interdipendenti.

**Definition (Equilibrio di Nash)**

Un *equilibrio di Nash* è una combinazione di strategie $(\hat{\varphi}_1, \dots, \hat{\varphi}_N) \in \times_{i=1}^N U_i$, tale per cui si abbia

$$J_i(\hat{\varphi}_1, \dots, \hat{\varphi}_N) \geq J_i([\varphi_i, \hat{\varphi}_{-i}])$$

per ogni $\varphi_i \in U_i, i = 1, \dots, N$.

L'equilibrio di un gioco è determinato anche dalle informazioni disponibili a ciascun giocatore. Si può caratterizzare la definizione a seconda della *struttura informativa*.

---

Si definisce $[\varphi_i, \hat{\varphi}_{-i}] = (\hat{\varphi}_1, \dots, \hat{\varphi}_{i-1}, \varphi_i, \hat{\varphi}_{i+1}, \dots, \hat{\varphi}_N)$.

# Due tipi di equilibri



## Definition (Equilibrio di Nash markoviano)

L'$N$-upla di controlli $(\hat{\varphi}_1, \ldots, \hat{\varphi}_N) \in \times_{i=1}^N U_i$ si dice *equilibrio di Nash markoviano* se vale $J_i(\hat{\varphi}_1, \ldots, \hat{\varphi}_N) \geq J_i([\varphi_i, \hat{\varphi}_{-i}])$ per ogni giocatore $i \in \mathcal{I}$ e

per ogni sua strategia $\varphi_i(t, x(t)) \in U_i$.

## Definition (Equilibrio di Nash open-loop)

L'$N$-upla di controlli $(\hat{u}_1, \ldots, \hat{u}_N) \in \times_{i=1}^N U_i$ si dice *equilibrio di Nash open-loop* se vale $J_i(\hat{u}_1, \ldots, \hat{u}_N) \geq J_i([u_i, \hat{u}_{-i}])$ per ogni giocatore $i \in \mathcal{I}$ e

per ogni sua strategia $u_i(t) \in U_i$.

## Theorem (Condizioni sufficienti per l'equilibrio markoviano)

*Sia* $(\varphi_1, \ldots, \varphi_N)$ *una data N-upla di funzioni* $\varphi_i : [t_0, t_1] \times \mathbb{R}^n \to U_i$ *e valgano:*

- *esista una funzione assolutamente continua* $x : [t_0, t_1] \to \mathbb{R}^n$ *soluzione delle equazioni del moto con le condizioni iniziali,*
- *per ogni* $i \in \mathcal{I}$ *esiste una funzione* $V^i : [t_0, t_1] \times \mathbb{R}^n \to \mathbb{R}$ *differenziabile con continuità tale che sia soddisfatta l'equazione HJB:*

$$\rho_i V^i(t,x) - V^i_t(t,x) = \max_{u_i \in \mathbb{R}^{m_i}} \left\{ V^i_x(t,x) f(x, [u_i, \varphi_{-i}], t) + f_{0i}(x, [u_i, \varphi_{-i}], t) \right\},$$

- $V^i(t_1, x) = S_i(x)$ *per ogni* $i \in \mathcal{I}$ *e per ogni* $x \in \mathbb{R}^n$.

*Sia* $\Phi_i(t,x) = \arg\max_{u_i \in \mathbb{R}^{m_i}} \left\{ V^i_x(t,x) f(x, [u_i, \varphi_{-i}], t) + f_{0i}(x, [u_i, \varphi_{-i}], t) \right\}$.
*Se* $\varphi_i(t,x) \in \Phi_i(t,x)$ *per ogni* $i \in \mathcal{I}$ *e per ogni* $t \in [t_0, t_1]$, *allora* $(\varphi_1, \ldots, \varphi_N)$ *è un equilibrio di Nash markoviano.*

# Teorema di condizioni sufficienti openloop



## Theorem (Condizioni sufficienti per l'equilibrio open-loop)

*Sia* $(\varphi_1, \ldots, \varphi_N)$ *una data* $N$*-upla di funzioni* $\varphi_i : [t_0, t_1] \to U_i$ *e sia* $H^i : \mathbb{R}^n \times \mathbb{R}^{m_i} \times \mathbb{R}^n \times [t_0, t_1] \to \mathbb{R}$ *la funzione hamiltoniana:*

$$H^i(x, u_i, \lambda^i, t) = f_{0i}(x, [u_i, \varphi_{-i}(t)], t) + \lambda^i f(x, [u_i, \varphi_{-i}(t)], t).$$

*Esistano* $N$ *funzioni continue e di classe* $C^1$ *a tratti* $\lambda^i : [t_0, t_1] \to \mathbb{R}^n$ *tali che:*

- $\varphi_i(t) \in \arg\max_{u_i \in \mathbb{R}^{m_i}} H^i(x(t), u_i, \lambda_i(t), t)$ *per ogni* $t \in [t_0, t_1]$,
- *per ogni* $t \in [t_0, t_1]$ *valga l'equazione aggiunta*

$$\dot{\lambda}^i(t) = -\frac{\partial}{\partial x} H^i(x(t), \varphi_i(t), \lambda^i(t), t) + \rho_i \lambda^i(t),$$

- $\lambda^i(t_1) = \frac{\partial}{\partial x} S_i(x(t_1))$ *per ogni* $i \in \mathcal{I}$.

*Allora* $(\varphi_1, \ldots, \varphi_N)$ *è un equilibrio di Nash open-loop.*

Modello di Lanchester per un duopolio, con $i, j \in \{1, 2\}$:

- $x_i(t)$: quota di mercato,
- $a_i(t)$: intensità degli sforzi pubblicitari,
- dinamica delle variabili di stato, con $i \neq j$:

$$\dot{x}_i(t) = \varphi_i a_i(t) x_j(t) - \varphi_j a_j(t) x_i(t).$$

Modello di Lanchester, esteso da Jørgensen e Sigué:

- $p_i(t)$: prezzo del prodotto al dettaglio,
- dinamica delle variabili di stato, con $i \neq j$:

$$\dot{x}_i(t) = a_i(t) \frac{p_j(t)}{p_i(t)} \sqrt{x_j(t)} - a_j(t) \frac{p_i(t)}{p_j(t)} \sqrt{x_i(t)}.$$

Il gioco differenziale si formalizza in questo modo:

$$
\begin{aligned}
\text{massimizza} \quad & J_1(p_1, a_1) = \int_0^T \left[ p_1(t) x_1(t) - \frac{c_1}{2} a_1^2(t) \right] dt + \sigma_1 x_1(T) \\
& J_2(p_2, a_2) = \int_0^T \left[ p_2(t) x_2(t) - \frac{c_2}{2} a_2^2(t) \right] dt + \sigma_2 x_2(T) \\
\text{soggetto a} \quad & \dot{x}_1(t) = a_1(t) \frac{p_2(t)}{p_1(t)} \sqrt{x_2(t)} - a_2(t) \frac{p_1(t)}{p_2(t)} \sqrt{x_1(t)} \\
& \dot{x}_2(t) = a_2(t) \frac{p_1(t)}{p_2(t)} \sqrt{x_1(t)} - a_1(t) \frac{p_2(t)}{p_1(t)} \sqrt{x_2(t)} \\
& x_1(0) = x_1^0, \quad x_2(0) = x_2^0 = 1 - x_1^0 \\
& x_1(t) + x_2(t) = 1 \quad \forall t \in [0, T] \\
& x_1(t), x_2(t) \in [0, 1] \quad \forall t \in [0, T] \\
& a_1(t), a_2(t) \geq 0, \quad p_1(t), p_2(t) > 0 \quad \forall t \in [0, T]
\end{aligned}
$$

---

S. Jørgensen e S. Sigué, «A Lanchester-Type Dynamic Game of Advertising and Pricing» (2020)

Condizioni sufficienti per l'equilibrio di Nash markoviano:

- equazioni di *HJB*

$$-\frac{\partial V_1}{\partial t} = \max_{a_1 \geq 0, p_1 > 0} \left\{ p_1 x_1 - \frac{c_1}{2} a_1^2 + \frac{\partial V_1}{\partial x_1} \dot{x}_1 + \frac{\partial V_1}{\partial x_2} \dot{x}_2 \right\},$$
$$-\frac{\partial V_2}{\partial t} = \max_{a_2 \geq 0, p_2 > 0} \left\{ p_2 x_2 - \frac{c_2}{2} a_2^2 + \frac{\partial V_2}{\partial x_2} \dot{x}_2 + \frac{\partial V_2}{\partial x_1} \dot{x}_1 \right\},$$

- condizioni all'istante finale

$$V_1(x_1, x_2, T) = \sigma_1 x_1(T), \quad V_2(x_1, x_2, T) = \sigma_2 x_2(T).$$

Si ipotizza che le funzioni valore $V_1$ e $V_2$ abbiano una forma lineare:

$$V_1 = \gamma_1(t)\, x_1 + \eta_1(t)\, x_2, \quad V_2 = \gamma_2(t)\, x_2 + \eta_2(t)\, x_1.$$

Si trovano le strategie ottime per l'equilibrio di Nash markoviano $(\hat{a}_1(x(t), t), \hat{p}_1(x(t), t), \hat{a}_2(x(t), t), \hat{p}_2(x(t), t))$.

# Equilibrio di Nash open-loop

Condizioni sufficienti per l'equilibrio di Nash open-loop:

- Funzioni hamiltoniane

$$H^1(x_1, x_2, a_1, p_1, \lambda_1, t), \quad H^2(x_1, x_2, a_2, p_2, \lambda_2, t),$$

- Equazioni aggiunte

$$\dot{\lambda}_1(t) = -p_1(t) + \lambda_1 a_2(t) \frac{p_1(t)}{p_2(t)} \frac{1}{2\sqrt{x_1(t)}},$$
$$\dot{\lambda}_2(t) = -p_2(t) + \lambda_2 a_1(t) \frac{p_2(t)}{p_1(t)} \frac{1}{2\sqrt{x_2(t)}},$$

- Condizioni di trasversalità

$$\lambda_1(T) = \sigma_1, \qquad \lambda_2(T) = \sigma_2.$$

Si massimizzano le funzioni hamiltoniane e si trovano le variabili di controllo candidate. Infine si determinano le strategie ottime per l'equilibrio di Nash open-loop $(a_1^*(t), p_1^*(t), a_2^*(t), p_2^*(t))$.

# Confronto degli equilibri



## Proposition

*Gli equilibri di Nash markoviano e di Nash open-loop del problema presentato non coincidono.*

## Dimostrazione.

Si confrontano le strategie pubblicitarie ottime per gli equilibri all'istante iniziale $t = 0$, dati i valori $x_1^0, x_2^0$:

$$\hat{a}_i(x_1^0, x_2^0, 0) = \frac{2\sigma_j}{2c_i - 3T\sigma_j}\frac{x_i^0}{\sqrt{x_j^0}} \neq a_i^*(x_1^0, x_2^0, 0) = \frac{\sigma_j}{c_i + 3T\sigma_j}\frac{x_i^0}{\sqrt{x_j^0}}$$

per $i, j = 1, 2$ e $i \neq j$.
É sufficiente per concludere che i controlli ottimi hanno traiettorie differenti, e dunque gli equilibri non coincidono.

$\square$

# Conclusioni



Per il problema considerato, l'equivalenza tra equilibrio di Nash markoviano e di Nash open-loop non è soddisfatta.

Si può concludere che l'equilibrio di Nash open-loop non è perfetto nei sottogiochi: per piccole perturbazioni lungo il percorso di equilibrio, le strategie originali possono non essere più ottime per uno o entrambi i giocatori.

Grazie dell'attenzione.