Fenchel Game No Regret Dynamics

Jacob Abernethy

November 6, 2025

Papers and Co-Authors 1

On Frank-Wolfe and Equilibrium Computation

Jacob Abernethy
Georgia Institute of Technology
prof@gatech.edu

Jun-Kun Wang
Georgia Institute of Technology
jimwang@gatech.edu

Acceleration through Optimistic No-Regret Dynamics

Jun-Kun Wang
College of Computing
Georgia Institute of Technology
Atlanta, GA 30313
jimwang@gatech.edu

Jacob Abernethy
College of Computing
Georgia Institute of Technology
Atlanta, GA 30313
prof@gatech.edu

Faster Rates for Convex-Concave Games

Jacob Abernethy PROF@GATECH.EDU
Georgia Tech

Kevin A. Lai KEVINLAI@GATECH.EDU
Georgia Tech

Kfir Y. Levy YEHUDA.LEVY@INF.ETHZ.CH
ETH Zurich

Jun-Kun Wang JIMWANG@GATECH.EDU
Georgia Tech

Papers and Co-Authors 2

Linear Separation via Optimism

Rafael Hanashiro RHANASHIRO3@GATECH.EDU
College of Computing, Georgia Institute of Technology

Jacob Abernethy PROF@GATECH.EDU
College of Computing, Georgia Institute of Technology

ON ACCELERATED PERCEPTRONS AND BEYOND

Guanghui Wang[1], Rafael Hanashiro[2], Etash Guha[1], Jacob Abernethy[1,3]
[1]College of Computing, Georgia Tech, Atlanta, GA, USA
[2]Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA
[3]Google Research, Atlanta, GA 30309
{gwang369,etash}@gatech.edu, rafah@mit.edu, abernethyj@google.com

Faster Margin Maximization Rates for Generic Optimization Methods

Guanghui Wang[1], Zihao Hu[1], Vidya Muthukumar[2,3], Jacob Abernethy[1,4]
[1]College of Computing, Georgia Institute of Technology
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology
[3]School of Industrial and Systems Engineering, Georgia Institute of Technology
[4]Google Research, Atalanta
{gwang369,zihaohu,vmuthukumar8}@gatech.edu, abernethyj@google.com

Papers and Co-Authors 3

Mathematical Programming
https://doi.org/10.1007/s10107-023-01976-y

FULL LENGTH PAPER

Series A

Check for updates

No-regret dynamics in the Fenchel game: a unified framework for algorithmic convex optimization

Jun-Kun Wang[1] · Jacob Abernethy[2] · Kfir Y. Levy[3]

Mathematical Programming
https://doi.org/10.1007/s10107-025-02283-4

FULL LENGTH PAPER

Series B

Check for updates

Faster margin maximization rates for generic and adversarially robust optimization methods

Guanghui Wang [1] · Zihao Hu [4] · Claudio Gentile[3] · Vidya Muthukumar [1] · Jacob Abernethy[1,2]

The Min-Max Theorem (Linear Version)

- Let M be an $n \times m$ payoff matrix.
- We have a p-player choosing a distribution $p \in \Delta_n$ to minimize $p^\top Mq$.
- We have a q-player choosing a distribution $q \in \Delta_m$ to maximize $p^\top Mq$.
- The Min-Max Theorem states that:

$$\min_{p \in \Delta_n} \max_{q \in \Delta_m} p^\top Mq = \max_{q \in \Delta_m} \min_{p \in \Delta_n} p^\top Mq$$

A Repeated Game Protocol

- In each round $t = 1, ..., T$:
  1. p-player chooses $p_t \in \Delta_n$ using Mult. Weights (MW), i.e.

$$p_{t+1,i} \propto p_{t,i} \exp(-\eta(Mq_t)_i)$$

  2. q-player chooses $j_t = \arg\max_{j \in \{1,\ldots,m\}} p_t^\top M e_j$ (i.e. $q_t = e_{j_t}$)

The proof is simply two trivial inequalities:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} p_t^\top M q_t &:= \min_p p^\top M\left(\tfrac{1}{T}\textstyle\sum_{t=1}^T q_t\right) + \frac{\mathrm{Reg}_T}{T} \\
&\leq \max_q \min_p p^\top M q + \tfrac{1}{T}\mathrm{Reg}_T \text{ (which } \to 0)
\end{aligned}$$

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} p_t^\top M q_t &= \frac{1}{T}\sum_{t=1}^{T}\left(\max_q p_t^\top M q\right) \\
&\geq \max_q \left(\tfrac{1}{T}\textstyle\sum_{t=1}^T p_t\right)^\top M q \quad \geq \quad \min_p \max_q p^\top M q
\end{aligned}$$

What's amazing about this proof:

1. Every step is a definition or trivial observation
2. The proof is **actually a construction**!
   - Indeed the pair $(\bar{p}_T, \bar{q}_T)$ is a $\frac{1}{T}\mathrm{Reg}_T$-approx minmax
3. It says nothing about the choice of online learning algorithm (except that $\mathrm{Reg}_T = o(T)$)
4. It does not depend at all on the payoff matrix
   - Indeed can be trivially extended to convex-concave $g(p, q)$
5. It says nothing about the domain of p and q (besides they need be convex)
6. Who goes first? Can player 2 see what player 1 does?
7. It permits the q-player to use any online learning algorithm instead of Best Response
8. The rounds of the game need not be uniformly weighted!

- Imagine the following protocol - note weights $\alpha_1, \ldots, \alpha_T > 0$
  1. Input: set $\mathcal{Z} \subset \mathbb{R}^n$, len T, weights $\alpha_{1:T}$, alg OAlg
  2. for $t = 1, 2, ..., T$ do
     1. Return: $z_t \leftarrow$ OAlg
     2. Receive: $\alpha_t, \ell_t(\cdot) \rightarrow$ OAlg
     3. Evaluate Loss: $\alpha_t \ell_t(z_t)$
  3. end for
- Goal: minimize weighted regret

$$\alpha\text{-REG}^{\mathcal{Z}} := \sum_{t=1}^{T} \alpha_t \ell_t(z_t) - \min_{z \in \mathcal{Z}} \sum_{t=1}^{T} \alpha_t \ell_t(z)$$

$$\overline{\alpha\text{-REG}}^{\mathcal{Z}} := \frac{\alpha\text{-REG}^{\mathcal{Z}}}{A_T} \quad \text{where } A_T = \alpha_1 + ... + \alpha_T$$

- **To whet your appetite: OGD (Online Gradient Descent)**
  - $x_{t+1} = x_t - \eta \nabla \ell_t(x_t)$
  - Proposition: OGD obtains an (average) regret rate of $O(1/\sqrt{T})$.

Master Protocol: dynamics of two-player game

Goal: find approx. min-max pair (x^*, y^*) for $g : \mathcal{X} \times \mathcal{Y}$,

$$\sup_{y \in \mathcal{Y}} g(x^*, y) - \inf_{x \in \mathcal{X}} g(x, y^*) \leq \epsilon$$

1. Input: len T, $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}^d$, $\alpha_{1:T} > 0$, algs $\text{OAlg}^{\mathcal{Y}}, \text{OAlg}^{\mathcal{X}}$
2. for $t = 1, \ldots, T$ do
    1. Return: $y_t \leftarrow \text{OAlg}^{\mathcal{Y}}$
    2. Update: $\alpha_t, h_t(\cdot) \rightarrow \text{OAlg}^{\mathcal{X}}$ where $h_t(\cdot) := -g(\cdot, y_t)$
    3. Return: $x_t \leftarrow \text{OAlg}^{\mathcal{X}}$
    4. Update: $\alpha_t, \ell_t(\cdot) \rightarrow \text{OAlg}^{\mathcal{Y}}$ where $\ell_t(\cdot) := g(x_t, \cdot)$
3. end for
4. Output $(\bar{x}_T, \bar{y}_T) := \left(\frac{\sum \alpha_s x_s}{A_T}, \frac{\sum \alpha_s y_s}{A_T}\right)$

Master Protocol MinMax Theorem

$$
\begin{aligned}
\frac{1}{A_T}\sum_{t=1}^{T}\alpha_t g(x_t, y_t) &= \frac{1}{A_T}\sum_{t=1}^{T} -\alpha_t \ell_t(y_t) \\
&\geq -\frac{1}{A_T}\inf_{y\in\mathcal{Y}}\left(\sum \alpha_t \ell_t(y)\right) - \overline{\alpha\text{-REG}}^{\mathcal{Y}} \\
&= \sup_{y\in\mathcal{Y}}\left(\frac{1}{A_T}\sum \alpha_t g(x_t, y)\right) - \overline{\alpha\text{-REG}}^{\mathcal{Y}} \geq \sup_{y\in\mathcal{Y}} g(\bar{x}_T, y) - \overline{\alpha\text{-REG}}^{\mathcal{Y}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{1}{A_T}\sum_{t=1}^{T}\alpha_t g(x_t, y_t) &= \text{ (... same steps ...)} \\
&\leq \inf_{x\in\mathcal{X}} g(x, \bar{y}_T) + \overline{\alpha\text{-REG}}^{\mathcal{X}}
\end{aligned}
$$

Therefore, $\quad \sup_{y\in\mathcal{Y}} g(\bar{x}_T, y) - \inf_{x\in\mathcal{X}} g(x, \bar{y}_T) \leq \overline{\alpha\text{-REG}}^{\mathcal{X}} + \overline{\alpha\text{-REG}}^{\mathcal{Y}}$

OCO: There's (sort of) only one algorithm

Follow the Regularized Leader (FTRL)

- Select cvx regularizer $R : \mathcal{Z} \to \mathbb{R}$, learning rate $\eta > 0$
- Receive loss fns $\ell_1(\cdot), \ell_2(\cdot), \ldots$, weights $\alpha_1, \alpha_2, \ldots$
- At time t set

$$z_t \leftarrow \mathrm{argmin}_{z \in \mathcal{Z}} \left(\frac{1}{\eta} R(z) + \sum_{s=1}^{t-1} \alpha_s \ell_s(z) \right)$$

- For a β-strongly convex regularizer $R(\cdot)$ and μ-strongly convex losses $\ell_t(\cdot)$, the regret is bounded by:

$$\mathrm{Reg}_T \leq \frac{1}{\eta}(R(z^*) - R(z_1)) + \sum_{t=1}^{T} \frac{2\alpha_t^2}{(\sum_{s=1}^{t} \alpha_s \mu) + \beta} \|\nabla \ell_t(z_t)\|_*^2$$

- With $\alpha_t = 1$ and $\mu > 0$: $\mathrm{Reg}_T \leq \frac{G \log(T+1)}{2\mu}$

Optimization $\to$ MinMax: The Fenchel Game

- I have a convex optimization problem $\min_{x \in \mathcal{X}} f(x)$.
- We can convert it into the **Fenchel Game**:

$$g(x, y) := \langle x, y \rangle - f^*(y)$$

where $f^*(y)$ is the Fenchel conjugate of f;

$$f^*(y) := \sup_x \langle x, y \rangle - f(x).$$

- **Key Lemma:** if $(\hat{x}, \hat{y})$ an ϵ-approx. minmax of $g(\cdot, \cdot)$, then
  1. $f(\hat{x}) - \min_{x^*} f(x^*) \leq \epsilon$, and
  2. $\hat{y} \approx \mathbf{0}$

Example: The Frank-Wolfe Algorithm (v1)

The Frank-Wolfe goes back to 1956; an excellent method for convex opt when you have a linear minimization oracle!

- **Goal:** minimize $f(w)$ for $w \in \mathcal{K}$.
- For $t = 1, \ldots, T$:

$$\begin{aligned} \gamma_t &\leftarrow \frac{2}{t+1} \\ v_t &\leftarrow \operatorname{argmin}_{v \in \mathcal{K}} \langle v, \nabla f(w_{t-1}) \rangle \\ w_t &\leftarrow (1 - \gamma_t) w_{t-1} + \gamma_t v_t \end{aligned}$$

- **Theorem 4 (orig):** Assuming $\|\mathcal{K}\| \leq D$ and $\|\nabla f\| \leq L$:

$$f(w_T) - \min_{w \in \mathcal{K}} f(w) \leq \frac{8LD}{T+1}$$

The (almost) Frank-Wolfe Algorithm (v2)

Recall that $g(y, v) := \langle y, v\rangle - f^*(y)$

- For $t = 1, \ldots, T$:

$$\gamma_t \leftarrow \frac{1}{t} \quad \text{(... only change ...)}$$

$$y_t \leftarrow \text{argmax}_y \langle y, w_{t-1}\rangle - f^*(y)$$

$$= \text{argmax}_y \sum_{s=1}^{t-1} g(v_s, y) \;=\; \nabla f(w_{t-1})$$

$$v_t \leftarrow \text{argmin}_{v\in\mathcal{K}} \langle v, y_t\rangle = \text{argmin}_{v\in\mathcal{K}} g(v, y_t)$$

$$w_t \leftarrow (1-\gamma_t)w_{t-1} + \gamma_t v_t = \frac{1}{t}\sum_{s=1}^{t} v_s$$

- **Thm (v2):** $f(w_T) - \min_{w\in\mathcal{K}} f(w) = O\left(\frac{LD\log T}{T}\right)$

The (almost) Frank-Wolfe Algorithm (v2)

Recall that $g(y, v) := \langle y, v\rangle - f^*(y)$

- For $t = 1, \ldots, T$:

$$\gamma_t \leftarrow \frac{1}{t} \quad \text{(... only change ...)}$$

$$\begin{aligned} y_t &\leftarrow \operatorname{argmax}_y \langle y, w_{t-1}\rangle - f^*(y) \\ &= \operatorname{argmax}_y \sum_{s=1}^{t-1} g(v_s, y) \; = \; \nabla f(w_{t-1}) \\ v_t &\leftarrow \operatorname{argmin}_{v \in \mathcal{K}} \langle v, y_t\rangle = \operatorname{argmin}_{v \in \mathcal{K}} g(v, y_t) \\ w_t &\leftarrow (1 - \gamma_t) w_{t-1} + \gamma_t v_t = \frac{1}{t} \sum_{s=1}^{t} v_s \end{aligned}$$

- **Thm (v2):** $f(w_T) - \min_{w \in \mathcal{K}} f(w) = O\left(\frac{LD \log T}{T}\right)$

The (FIXED) Frank-Wolfe Algorithm (v1)

Recall that $g(y,v) := \langle y, v\rangle - f^*(y)$

- For $t = 1, \ldots, T$:

$$\gamma_t \leftarrow \frac{2}{t+1} \quad \Longleftrightarrow \quad \alpha_t \leftarrow t$$

$$y_t \leftarrow \mathrm{argmax}_y \langle w_{t-1}, y\rangle - f^*(y) \quad \text{(FTL)}$$

$$= \mathrm{argmax}_y \sum_{s=1}^{t-1} \alpha_s g(v_s, y) \; = \; \nabla f(w_{t-1})$$

$$v_t \leftarrow \mathrm{argmin}_{v\in\mathcal{K}} \langle v, y_t\rangle = \mathrm{argmin}_{v\in\mathcal{K}} g(v, y_t) \quad \text{(BestResp)}$$

$$w_t \leftarrow (1-\gamma_t) w_{t-1} + \gamma_t v_t = \frac{1}{\alpha_1 + \ldots + \alpha_t} \sum_{s=1}^{t} \alpha_s v_s$$

- **Thm (v1):** $f(w_T) - \min_{w\in\mathcal{K}} f(w) = \leq \frac{8LD}{T+1}$

The (FIXED) Frank-Wolfe Algorithm (v1)

Recall that $g(y, v) := \langle y, v\rangle - f^*(y)$

- **Thm:** $f(w_T) - \min_{w\in\mathcal{K}} f(w) = \leq \frac{8LD}{T+1}$
- **Proof (v2)**:

$$
\begin{aligned}
f(w_T) - \min_{w\in\mathcal{K}} f(w) &\leq \overline{\alpha\text{-REG}}(\text{FTL}) + \overline{\alpha\text{-REG}}(\text{BestResp}) \\
&= \frac{LD\sum_{t=1}^{T} O(\frac{\alpha_s^2}{A_t})}{A_T} + 0 \\
&= \frac{LD\sum_{t=1}^{T} O(\frac{t^2}{t^2})}{T^2} \\
&= O(\frac{LDT}{T^2}) = O(\frac{LD}{T})
\end{aligned}
$$

Fenchel Game No Regret Dynamics (FGNRD)

- We have a recipe now!
  - Take any iterative optimization algorithm
  - Determine if the iterates can be decomposed into primal-dual
  - Determine primal player's algorithm $\text{OAlg}^{\mathcal{X}}$
  - Determine dual player's algorithm $\text{OAlg}^{\mathcal{Y}}$
  - Determine weights $\alpha_1, \alpha_2, \ldots$
- **Bonus:** Convergence rate comes for free.

$$f(w_T) - \min_{w \in \mathcal{K}} f(w) \leq \overline{\alpha\text{-REG}}(\text{OAlg}^{\mathcal{X}}) + \overline{\alpha\text{-REG}}(\text{OAlg}^{\mathcal{Y}})$$

Simple Example: Gradient Descent

- **EXAMPLE: Vanilla grad descent w/ iterate averaging**
  - $\eta \leftarrow \frac{R}{G\sqrt{T}}$
  - $w_t \leftarrow w_{t-1} - \eta \partial f(w_{t-1})$
  - $\bar{w}_t \leftarrow \frac{1}{t}\sum_{s=1}^{t} w_s$
- **FGNRD Equivalence:**
  - $g(x, y) := \langle x, y \rangle - f^*(y)$
  - $\alpha_t := 1$
  - $\text{OAlg}^{\mathcal{X}} := \text{OMD}[\frac{1}{2}\|\cdot\|_2^2, x_0, \eta]$
  - $\text{OAlg}^{\mathcal{Y}} := \text{BESTRESP}^+$

Theorem: Assuming f is convex we have

$$f(\bar{w}_T) - \min_{w \in \mathcal{K}} f(w) \le \overline{\alpha\text{-REG}}(\text{OMD}) + \overline{\alpha\text{-REG}}(\text{BESTRESP}^+)$$

$$= O\left(\eta G T + \frac{R}{\eta}\right) + 0 \; = \; O\left(\frac{GR}{\sqrt{T}}\right)$$

The Heavy Ball Algorithm

$$\begin{aligned} \eta_t &\leftarrow \tfrac{t}{4(t+1)L}, \quad \beta_t \leftarrow \tfrac{t-1}{t+2} \\ v_t &\leftarrow w_{t-1} - w_{t-2} \\ w_t &\leftarrow w_{t-1} - \eta_t \nabla f(w_{t-1}) \\ &\quad + \beta_t v_t \end{aligned}$$

Iterative Description

...

$$\begin{aligned} g(x,y) &:= \langle x, y\rangle - f^*(y) \\ \alpha_t &:= t \text{ for } t = 1, \ldots, T \\ \text{OAlg}^Y &:= \text{FTL}[\nabla f(w_0)] \\ \text{OAlg}^X &:= \text{MD}[\tfrac{1}{2}\|\cdot\|_2^2, \tfrac{1}{8L}] \end{aligned}$$

FGNRD Equivalence

Can we benefit from "coupling" the players?

- We might have oversimplified things by decomposing
- Recall the Fenchel game decomposition:

$$f(w_T) - \min_{w \in \mathcal{K}} f(w) \leq \overline{\alpha\text{-REG}}(\text{OAlg}^{\mathcal{X}}) + \overline{\alpha\text{-REG}}(\text{OAlg}^{\mathcal{Y}})$$

- Must we bound these two regrets independently?
- **Key Idea:** The two regret bounds can sometimes cancel each other out.
- Technique goes back to *Rakhlin & Sridharan 2013*, *Daskalakis et al 2015*, and *Syrgkanis et al 2015*.

OptimisticFTRL: Ok there's one more algorithm of note

Optimistic Follow the Regularized Leader (OptFTRL)

- Select cvx regularizer $R : \mathcal{Z} \to \mathbb{R}$, learning rate $\eta > 0$
- Receive loss fns $\ell_1(\cdot), \ell_2(\cdot), \ldots$, weights $\alpha_1, \alpha_2, \ldots$
- Select a "guess" $m_t : \mathcal{Z} \to \mathbb{R}$ of ℓ_t.
- At time t set

$$z_t \leftarrow \text{argmin}_{z \in \mathcal{Z}} \left(\frac{1}{\eta} R(z) + \alpha_t m_t(z) + \sum_{s=1}^{t-1} \alpha_s \ell_s(z) \right)$$

Regret Bound for Optimistic FTRL

Given $\eta > 0$ and a β-strcvx R, weights and losses $\{\alpha_t \ell_t(\cdot)\}_{t=1}^T$ where ℓ_t is μ_t-strcvx. Let $m_1, \ldots, m_T : \mathcal{Z} \to \mathbb{R}$ be the $\hat{\mu}_t$-strcvx *hints* given to OptimFTRL. Then OptimFTRL$[R(\cdot), \eta]$ satisfies

$$
\begin{aligned}
\alpha - \mathrm{Reg}(z^*) \leq \sum_{t=1}^T & \alpha_t \left(\ell_t(z_t) - \ell_t(w_{t+1}) - m_t(z_t) + m_t(w_{t+1})\right) \\
& + \frac{1}{\eta}\left(R(z^*) - R(w_1)\right) \\
& - \frac{1}{2}\sum_{t=1}^T \left(\frac{\beta}{\eta} + \sum_{s=1}^{t-1} \alpha_s \mu_s\right) \|z_t - w_t\|^2 \\
& - \frac{1}{2}\sum_{t=1}^T \left(\frac{\beta}{\eta} + \alpha_t \hat{\mu}_t + \sum_{s=1}^{t-1} \alpha_s \mu_s\right) \|z_t - w_{t+1}\|^2
\end{aligned}
$$

where $w_t := \mathrm{argmin}_{z \in \mathcal{Z}} \sum_{s=1}^{t-1} \alpha_s \ell_s(z) + \frac{1}{\eta} R(z)$; $z^* \in \mathcal{Z}$ arbitrary.

Nesterov Accelerated Gradient Descent

$$\begin{aligned} \beta_t &\leftarrow \tfrac{2}{t+1}, \gamma_t \leftarrow \tfrac{t}{4L} \\ z_t &\leftarrow (1-\beta_t)w_{t-1} + \beta_t v_{t-1} \\ v_t &\leftarrow \operatorname*{argmin}_{x \in \mathcal{K}} \left\{ \gamma_t \langle \nabla f(z_t), x \rangle + D^{\phi}_{v_{t-1}}(x) \right\} \\ w_t &\leftarrow (1-\beta_t)w_{t-1} + \beta_t v_t \end{aligned}$$

Iterative Description

$$\begin{aligned} g(x,y) &:= \langle x, y \rangle - f^*(y) \\ \alpha_t &:= t \text{ for } t = 1, \ldots, T \\ \mathrm{OAlg}^Y &:= \mathrm{OptimFTL}[\nabla f(v_0)] \\ \mathrm{OAlg}^X &:= \mathrm{MD}[\phi(\cdot), \tfrac{1}{4L}] \end{aligned}$$

FGNRD Equivalence

Nesterov's ∞-memory method

$$
\begin{aligned}
\beta_t &\leftarrow \tfrac{2}{t+1}, \gamma_t \leftarrow \tfrac{t}{4L} \\
z_t &\leftarrow (1-\beta_t) w_{t-1} + \beta_t v_{t-1} \\
v_t &\leftarrow \operatorname{argmin}_{x \in \mathcal{K}} \left\{ \sum_{s=1}^{t} \gamma_s \langle \nabla f(z_s), x \rangle + R(x) \right\} \\
w_t &\leftarrow (1-\beta_t) w_{t-1} + \beta_t v_t
\end{aligned}
$$

Iterative Description

$$
\begin{aligned}
g(x,y) &:= \langle x, y \rangle - f^*(y) \\
\alpha_t &:= t \text{ for } t = 1, \ldots, T \\
\text{OAlg}^Y &:= \text{OptimFTL}[\nabla f(v_0)] \\
\text{OAlg}^X &:= \text{FTRL+}[R(\cdot), \tfrac{1}{4L}]
\end{aligned}
$$

FGNRD Equivalence