Frank-Wolfe Algorithm for Saddle Point Problems

Gauthier Gidel[1], **Tony Jebara**[2] and **Simon Lacoste-Julien**[3]

[1]Sierra Project-team - INRIA, [2]Columbia University, [3]Université de Montréal

Overview

Summary

Frank-Wolfe algorithm (FW) **gained in popularity** in the last couple of years because of some ***key properties*** (only requires LMO).

- We tried to ***extend*** FW nice properties to solve saddle point problem.
- **Straightforward** extension but **Non trivial** analysis.

Contributions

- We **extend** several variants of the FW algorithm to solve the saddle point problem.
- We **prove convergence** results for these methods over **polytope** domains giving a partial answer to **Hammond's conjecture** [1].

Saddle Point Problem

Let $\mathcal{L} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{X}$ and $\mathcal{Y}$ are convex and compact.

Saddle point problem solve $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} \mathcal{L}(\boldsymbol{x},\boldsymbol{y})$

A solution $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is called a ***Saddle Point***.

Stationary conditions

$$\langle \boldsymbol{x} - \boldsymbol{x}^*, \quad \nabla_x \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)\rangle \geq 0$$

$$\langle \boldsymbol{y} - \boldsymbol{y}^*, -\nabla_y \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*)\rangle \geq 0$$

Variational inequality

$$\forall \boldsymbol{z} \in \mathcal{X}\times\mathcal{Y} \ \langle \boldsymbol{z} - \boldsymbol{z}^*, g(\boldsymbol{z}^*)\rangle \geq 0$$

where $g(\boldsymbol{z}) = (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$

and $(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{z}^*$

Global solution if $\mathcal{L}$ *convex-concave*: $\forall (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{X}\times\mathcal{Y}$

$\boldsymbol{x}' \mapsto \mathcal{L}(\boldsymbol{x}', \boldsymbol{y})$ is convex and $\boldsymbol{y}' \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}')$ is concave.

Simple method to solve SP

projected gradient

Simple algorithm to solve Saddle point optimization:

$$\boldsymbol{x}^+ = P_{\mathcal{X}}(\boldsymbol{x} - \eta \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}))$$
$$\boldsymbol{y}^+ = P_{\mathcal{Y}}(\boldsymbol{y} + \eta \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}))$$

Non-smooth function:

$$\frac{1}{T}\sum_{t=1}^{T} \left(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}\right) \underset{T\to\infty}{\longrightarrow} (\boldsymbol{x}^*, \boldsymbol{y}^*)$$

projected extra-gradient

$$\bar{\boldsymbol{x}} = P_{\mathcal{X}}(\boldsymbol{x} - \eta \nabla_x \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}))$$
$$\bar{\boldsymbol{y}} = P_{\mathcal{Y}}(\boldsymbol{y} + \eta \nabla_y \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}))$$
$$\boldsymbol{x}^+ = P_{\mathcal{X}}(\boldsymbol{x} - \eta \nabla_x \mathcal{L}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}))$$
$$\boldsymbol{y}^+ = P_{\mathcal{Y}}(\boldsymbol{y} + \eta \nabla_y \mathcal{L}(\bar{\boldsymbol{x}}, \bar{\boldsymbol{y}}))$$

Faster for **Smooth** function:

$$(\boldsymbol{x}^{(T)}, \boldsymbol{y}^{(T)}) \underset{T\to\infty}{\longrightarrow} (\boldsymbol{x}^*, \boldsymbol{y}^*)$$

SP-FW

Update rule

Update

$$\boldsymbol{r}(\boldsymbol{z}) := \begin{pmatrix} \nabla_x \mathcal{L}(\boldsymbol{z}) \\ -\nabla_y \mathcal{L}(\boldsymbol{z}) \end{pmatrix}$$
$$\boldsymbol{s} \in \underset{\boldsymbol{z}\in\mathcal{X}\times\mathcal{Y}}{\text{argmin}} \ \langle \boldsymbol{z}, \boldsymbol{r}(\boldsymbol{z})\rangle$$
$$\boldsymbol{z}^+ := (1-\gamma)\boldsymbol{z} + \gamma \boldsymbol{s}$$

Stopping criterion:

$$g_t := \langle \boldsymbol{r}(\boldsymbol{z}), \boldsymbol{z} - \boldsymbol{s}\rangle \leq \varepsilon$$

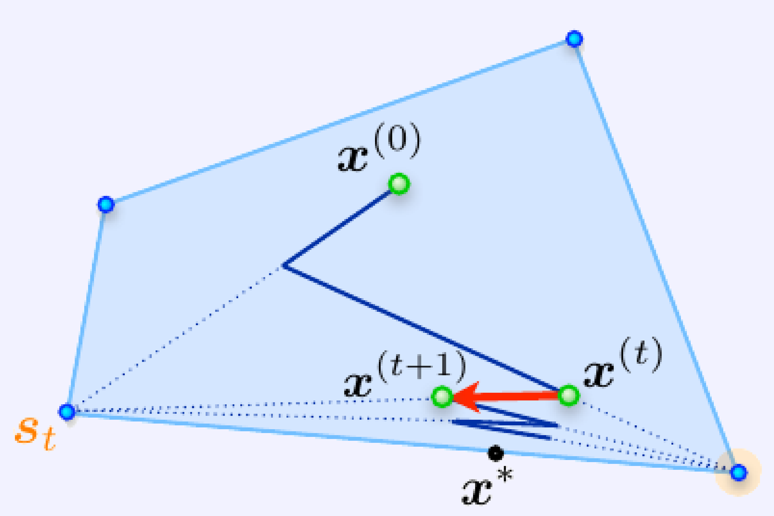
Properties

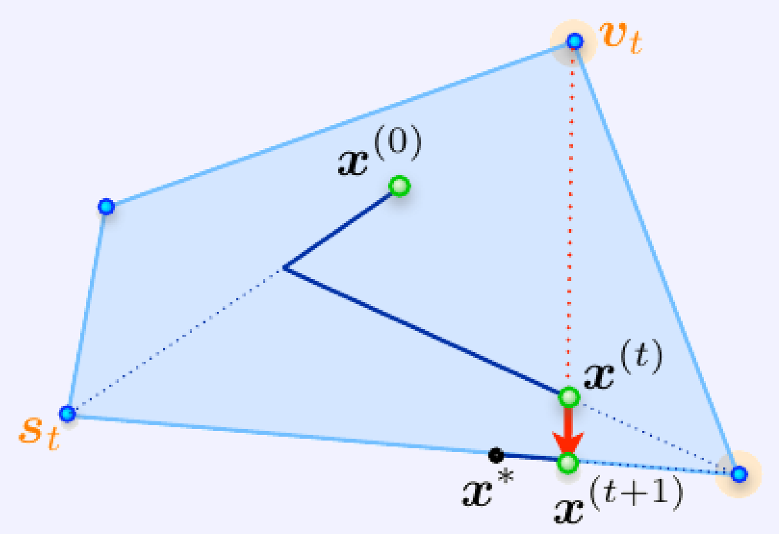
Only LMO.

- Gap certificate for free.
- Algorithm *affine invariant*.
- Universal step size $\gamma_t := \frac{2}{2+t}$ and adaptive step size $\gamma_t \sim g_t$.
- **No** line-search.
- ***Sparsity*** of the iterates.

Away step

$$\boldsymbol{v} \in \underset{\boldsymbol{v}'\in\mathcal{S}_x\times\mathcal{S}_y}{\arg\max} \langle \boldsymbol{r}(\boldsymbol{z}), \boldsymbol{v}'\rangle, \quad \boldsymbol{d}_{\mathrm{A}} := \boldsymbol{z} - \boldsymbol{v} \ \text{ and } \ g_t^{\mathrm{A}} := \langle \boldsymbol{d}_{\mathrm{A}}, -\boldsymbol{r}(\boldsymbol{z})\rangle$$





Theoretical contribution

Convergence

SP extension of FW with **away step**:

- ***Linear*** rate with ***adaptive*** step size $\gamma_t := \frac{\nu}{LD^2} g_t$.

$$\min_{s\leq t} g_s \leq \left(1 - \nu^2 \frac{\delta^2}{D^2}\frac{\mu}{2L}\right)^t$$

- ***Sublinear*** rate with ***universal*** step size $\gamma_t := \frac{2}{2+k(t)}$.

Hypothesis

Similar hypothesis as AFW:

- $\mathcal{L}$ **smooth** and **strongly convex**.
- $\mathcal{X}$ and $\mathcal{Y}$ **polytopes**.

Additional assumption on **bilinearity**:

$$\nu := \frac{1}{2} - \frac{\sqrt{2}\|M\|}{\mu}\frac{D}{\delta} > 0$$

Details on the additional assumption

"*Strong convexity* μ *big enough compared to the bilinear coupling* $\|M\|$"

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{x}^\top M \boldsymbol{y} - g(\boldsymbol{y}).$$

$$D := \max\{\mathrm{diam}(\mathcal{X}), \mathrm{diam}(\mathcal{Y})\}, \ \delta := \min\{PWidth(\mathcal{X}), PWidth(\mathcal{Y})\}$$

Difficulties for SP

FW proof technique

The gradient is L-Lipschitz,

$$h_{t+1} \leq h_t - \gamma_t g_t + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2}$$

Set $\gamma_t = \frac{g_t}{C}$, decreasing scheme:

$$h_{t+1} \leq h_t - \frac{g_t^2}{2C} \leq h_t - \frac{h_t^2}{2C}$$

Same derivation for SP

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t - \gamma_t \underbrace{(g_t^{(x)} - g_t^{(y)})}_{\text{arbitrary sign}} + \gamma_t^2 \frac{LD^2}{2}$$

- Cannot control oscillations of $\mathcal{L}_t$.
- Must introduce other quantities.
- Proof use **recent advances** on AFW [3].

Conjectures related

Karlin's conjecture

SP-FW is equivalent to the **fictitious play algorithm** [4] when

$\gamma_t = \frac{1}{1+t}$ and $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top M \boldsymbol{y}$

Karlin [2] conjectured that:

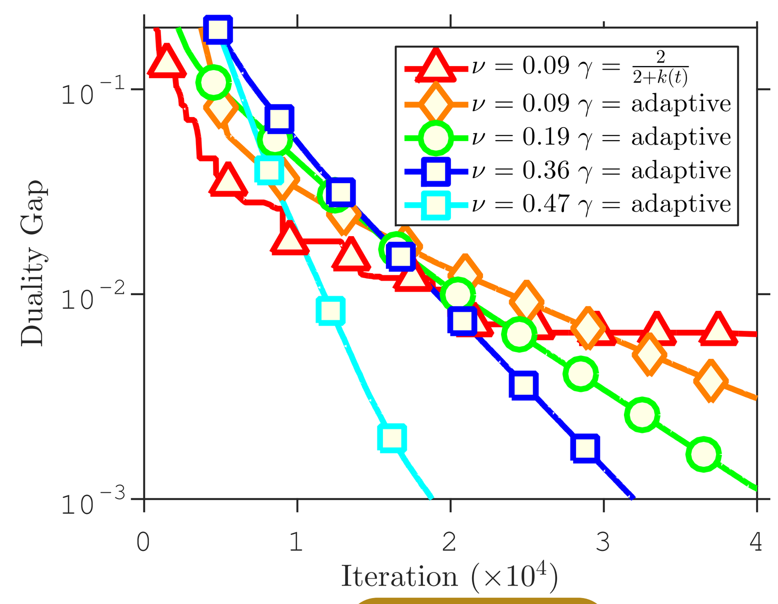
$$g_t \leq O\left(\frac{1}{\sqrt{t}}\right)$$

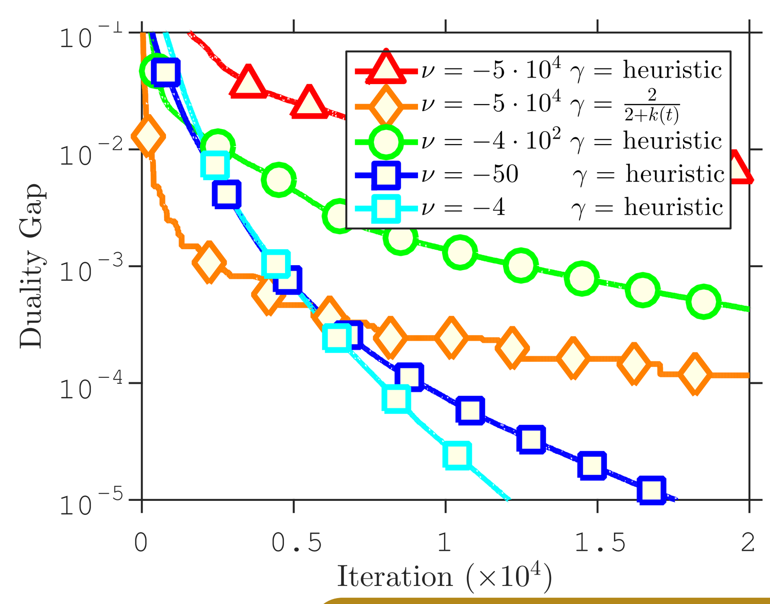
Hammond's conjecture

Hammond [1] conjectured that for Variational inequalities:

If g is **uniformly** monotone and the constraints is a **bounded polyhedron**, then the **fictitious play algorithm** will solve the variational inequality problem.

Toy experiments





Theoretical: $\gamma_t = \frac{\nu}{C} g_t$

Heuristic: $\gamma_t = \dfrac{g_t}{C + 2\frac{\|M\|^2 D^2}{\mu}}$

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) := \frac{\mu}{2}\|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 + (\boldsymbol{x} - \boldsymbol{x}^*)^\top M (\boldsymbol{y} - \boldsymbol{y}^*) - \frac{\mu}{2}\|\boldsymbol{y} - \boldsymbol{y}^*\|_2^2$$

- $\mathcal{X} = \mathcal{Y} := [0,1]^d$
- $d = 30$
- $C := 2LD^2$
- $L = \mu$

References

[1] J. H. Hammond. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms*. PhD thesis, Massachusetts Institute of Technology, 1984.

[2] S. Karlin. Mathematical methods and theory in games, programming and economics, 1960.

[3] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *NIPS*, 2015.

[4] J. Robinson. An iterative method of solving a game. *Annals of mathematics*, 1951.