Frank-Wolfe Algorithms for Saddle Point Problems

Gauthier Gidel
INRIA - Sierra Project-team
École normale supérieure, Paris

Tony Jebara
Department of Computer Science
Columbia U. & Netflix Inc., NYC

Simon Lacoste-Julien
Department of CS & OR (DIRO)
Université de Montréal, Montréal

Abstract

We extend the Frank-Wolfe (FW) optimization algorithm to solve constrained smooth convex-concave saddle point (SP) problems. Remarkably, the method only requires access to linear minimization oracles. Leveraging recent advances in FW optimization, we provide the first proof of convergence of a FW-type saddle point solver over polytopes, thereby partially answering a 30 year-old conjecture. We also survey other convergence results and highlight gaps in the theoretical underpinnings of FW-style algorithms. Motivating applications without known efficient alternatives are explored through structured prediction with combinatorial penalties as well as games over matching polytopes involving an exponential number of constraints.

1 Introduction

The Frank-Wolfe (FW) optimization algorithm (Frank and Wolfe, 1956), also known as the conditional gradient method (Demyanov and Rubinov, 1970), is a first-order method for smooth constrained optimization over a compact set. It has recently enjoyed a surge in popularity thanks to its ability to cheaply exploit the structured constraint sets appearing in machine learning applications (Jaggi, 2013; Lacoste-Julien and Jaggi, 2015). A known forte of FW is that it only requires access to a *linear minimization oracle* (LMO) over the constraint set, i.e., the ability to minimize linear functions over the set, in contrast to projected gradient methods which require the minimization of *quadratic* functions or other nonlinear functions. In this paper, we extend the applicability of the FW algorithm to solve the following convex-concave saddle point problems:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

$$\text{with only access to } \mathrm{LMO}(\mathbf{r}) \in \operatorname*{arg\,min}_{\boldsymbol{s} \in \mathcal{X} \times \mathcal{Y}} \langle \boldsymbol{s}, \mathbf{r} \rangle ,$$

where $\mathcal{L}$ is a smooth (with L-Lipschitz continuous gradient) *convex-concave function*, i.e., $\mathcal{L}(\cdot, \boldsymbol{y})$ is convex for all $\boldsymbol{y} \in \mathcal{Y}$ and $\mathcal{L}(\boldsymbol{x}, \cdot)$ is concave for all $\boldsymbol{x} \in \mathcal{X}$. We also assume that $\mathcal{X} \times \mathcal{Y}$ is a convex compact set such that its LMO is cheap to compute. A *saddle point solution* to (1) is a pair $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathcal{X} \times \mathcal{Y}$ (Hiriart-Urruty and Lemaréchal, 1993, VII.4) such that: $\forall \boldsymbol{x} \in \mathcal{X}$, $\forall \boldsymbol{y} \in \mathcal{Y}$,

$$\mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}) \leq \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^*) \leq \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2}$$

Examples of saddle point problems. Taskar et al. (2006) cast the maximum-margin estimation of structured output models as a bilinear saddle point problem $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\top M \boldsymbol{y}$, where $\mathcal{X}$ is the regularized set of parameters and $\mathcal{Y}$ is an encoding of the set of possible structured outputs. They considered settings where the projection on $\mathcal{X}$ and $\mathcal{Y}$ was efficient, but one can imagine many situations where only LMO's are efficient. For example, we could use a structured sparsity inducing norm (Martins et al., 2011) for the parameter $\boldsymbol{x}$, such as the overlapping group lasso for which the projection is expensive (Bach et al., 2012), while $\mathcal{Y}$ could be a combinatorial object such as a the ground state of a planar Ising model (without external field) which admits an efficient oracle (Barahona, 1982) but has potentially intractable projection.

Similarly, two-player games (Von Neumann and Morgenstern, 1944) can often be solved as bilinear minimax problems. When a strategy space involves a polynomial number of constraints, the equilibria of such games can be solved efficiently (Koller et al., 1994). However, in situations such as the Colonel Blotto game or the Matching Duel (Ahmadinejad et al., 2016), the strategy space is intractably large and defined by an exponential number of linear constraints. Fortunately, despite this apparent prohibitive structure, some linear minimization oracles such as the blossom algorithm (Edmonds, 1965) can efficiently optimize over the matching polytopes.

Proceedings of the 20^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, Florida, USA. JMLR: W&CP volume 54. Copyright 2017 by the author(s).

Robust learning is also often cast as a saddle point minimax problem (Kim et al., 2005). Once again, a FW implementation could leverage fast linear oracles while projection methods would be plagued by slower or intractable sub-problems. For instance, if the LMO is max-flow, it could have almost linear runtime while the corresponding projection would require cubic runtime quadratic programming (Kelner et al., 2014). Finally, note that the popular generative adversarial networks (Goodfellow et al., 2014) are formulated as a (non-convex) saddle point optimization problem.

Related work. The standard approaches to solve smooth constrained saddle point problems are projection-type methods (surveyed in Xiu and Zhang (2003)), with in particular variations of Korpelevich's extragradient method (Korpelevich, 1976), such as (Nesterov, 2007) which was used to solve the structured prediction problem (Taskar et al., 2006) mentioned above. There is surprisingly little work on FW-type methods for saddle point problems, although they were briefly considered for the more general *variational inequality* problem (VIP):

$$\text{find} \quad \boldsymbol{z}^* \in \mathcal{Z} \quad \text{s.t.} \; \langle \boldsymbol{r}(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^* \rangle \geq 0, \quad \forall \boldsymbol{z} \in \mathcal{Z}, \quad (3)$$

where $\boldsymbol{r}$ is a Lipschitz mapping from $\mathbb{R}^p$ to itself and $\mathcal{Z} \subseteq \mathbb{R}^p$. By using $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and $\boldsymbol{r}(\boldsymbol{z}) = (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$, the VIP (3) reduces to the equivalent optimality conditions for the saddle point problem (1). Hammond (1984) showed that a FW algorithm with a step size of $O(1/t)$ converges for the VIP (3) when the set $\mathcal{Z}$ is strongly convex, while FW with a generalized line-search on a saddle point problem is sometimes non-convergent when $\mathcal{Z}$ is a polytope (see also (Patriksson, 1999, § 3.1.1)). She conjectured though that using a step size of $O(1/t)$ was also convergent when $\mathcal{Z}$ is a polytope – a problem left open up to this point. More recently, Juditsky and Nemirovski (2016) (see also Cox et al. (2015)) proposed a method to transform a VIP on $\mathcal{Z}$ where one has only access to a LMO, to a "dual" VIP on which they can use a projection-type method. Lan (2013) proposes to solve the saddle point problem (1) by running FW on $\mathcal{X}$ on the *smoothed* version of the problem $\max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$, thus requiring a projection oracle on $\mathcal{Y}$. In contrast, in this paper we study simple approaches that do not require any transformations of the problem (1) nor any projection oracle on $\mathcal{X}$ or $\mathcal{Y}$. Finally, He and Harchaoui (2015) introduced an interesting extragradient-type method to solve (3) by approximating the projections using linear oracles. In contrast to our proposal, their work does not cover the geometric convergence for the strongly convex case.

Contributions. In § 2, we extend several variants of the FW algorithm to solve the saddle point problem (1) that we think could be of interest to the machine learning community. In § 3, we give a first proof of (geometric) convergence for these methods over polytope domains under the assumptions of sufficient strong convex-concavity of $\mathcal{L}$, giving a partial answer to the conjecture from Hammond (1984). In § 4, we extend and refine the previous convergence results when $\mathcal{X}$ and $\mathcal{Y}$ are strongly convex sets and the gradient of $\mathcal{L}$ is non-zero over $\mathcal{X} \times \mathcal{Y}$, while we survey the pure bilinear case in § 5. We finally present illustrative experiments for our theory in § 6, noticing that the convergence theory is still incomplete for these methods.

2 Saddle point Frank-Wolfe (SP-FW)

The algorithms. This article will explore three SP extensions of the classical *Frank-Wolfe* (FW) algorithm (Alg. 1) which are summarized in Alg. 2, 3 and 4.[1] We denote by $\boldsymbol{z}^{(t)} := (\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ the iterate computed after t steps. We first obtain the *saddle point FW* (SP-FW) algorithm (Alg. 2) by simultaneously doing a FW update on both convex functions $\mathcal{L}(\cdot, \boldsymbol{y}^{(t)})$ and $-\mathcal{L}(\boldsymbol{x}^{(t)}, \cdot)$ with a properly chosen step size. As in standard FW, the point $\boldsymbol{z}^{(t)}$ has a sparse representation as a convex combination of the points previously given by the FW oracle, that is,

$$\boldsymbol{x}^{(t)} = \sum_{\boldsymbol{v}_x \in \mathcal{S}_x^{(t)}} \alpha_{\boldsymbol{v}_x} \boldsymbol{v}_x \quad \text{and} \quad \boldsymbol{y}^{(t)} = \sum_{\boldsymbol{v}_y \in \mathcal{S}_y^{(t)}} \alpha_{\boldsymbol{v}_y} \boldsymbol{v}_y. \quad (4)$$

These two sets $\mathcal{S}_x^{(t)}$, $\mathcal{S}_y^{(t)}$ of points are called the *active sets*, and we can maintain them separately (thanks to the product structure of $\mathcal{X} \times \mathcal{Y}$) to run the other two FW variants that we describe below (see L13 of Alg. 3).

If we assume that $\mathcal{X}$ and $\mathcal{Y}$ are the convex hulls of two finite sets of points $\mathcal{A}$ and $\mathcal{B}$, we can also extend the *away-step Frank-Wolfe* (AFW) algorithm (Guélat and Marcotte, 1986; Lacoste-Julien and Jaggi, 2015) to saddle point problems. As for AFW, this new algorithm can choose an *away* direction $\boldsymbol{d}_{\mathcal{A}}$ to remove mass from "bad" atoms in the active set, i.e. to reduce $\alpha_{\boldsymbol{v}}$ for some $\boldsymbol{v}$ (see L9 of Alg. 3), thereby avoiding the zig-zagging problem that slows down standard FW (Lacoste-Julien and Jaggi, 2015). Note that because of the special product structure of the domain, we consider more away directions than proposed in (Lacoste-Julien and Jaggi, 2015) for AFW (see Appendix A for more details). Finally, a straightforward saddle point generalization for the *pairwise Frank-Wolfe* (PFW) algorithm (Lacoste-Julien and Jaggi, 2015) is given in Alg. 4. The proposed algorithms all preserve several nice properties of previous FW methods (in addition to only requiring LMO's): simplicity of implementation, affine invariance (Jaggi, 2013), gap certificates computed for free, sparse representation of the iterates and the possibility to have

[1] Alg. 2 was already proposed by Hammond (1984) for VIPs, while our step sizes and Alg. 3 & 4 are novel.

Algorithm 1 Frank-Wolfe algorithm

```
1: Let $\boldsymbol{x}^{(0)} \in \mathcal{X}$
2: for $t = 0 \dots T$ do
3:     Compute $\boldsymbol{r}^{(t)} = \nabla f(\boldsymbol{x}^{(t)})$
4:     Compute $\boldsymbol{s}^{(t)} := \operatorname*{argmin}_{\boldsymbol{s} \in \mathcal{X}} \langle \boldsymbol{s}, \boldsymbol{r}^{(t)} \rangle$
5:     Compute $g_t := \langle \boldsymbol{x}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{r}^{(t)} \rangle$
6:     if $g_t \leq \epsilon$ then return $\boldsymbol{x}^{(t)}$
7:     Let $\gamma = \frac{2}{2+t}$ (or do line-search)
8:     Update $\boldsymbol{x}^{(t+1)} := (1-\gamma)\boldsymbol{x}^{(t)} + \gamma \boldsymbol{s}^{(t)}$
9: end for
```

Algorithm 2 Saddle point Frank-Wolfe algorithm: **SP-FW**

```
1: Let $\boldsymbol{z}^{(0)} = (\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}) \in \mathcal{X} \times \mathcal{Y}$
2: for $t = 0 \dots T$ do
3:     Compute $\boldsymbol{r}^{(t)} := \begin{pmatrix} \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \\ -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \end{pmatrix}$
4:     Compute $\boldsymbol{s}^{(t)} := \operatorname*{argmin}_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \langle \boldsymbol{z}, \boldsymbol{r}^{(t)} \rangle$
5:     Compute $g_t := \langle \boldsymbol{z}^{(t)} - \boldsymbol{s}^{(t)}, \boldsymbol{r}^{(t)} \rangle$
6:     if $g_t \leq \epsilon$ then return $\boldsymbol{z}^{(t)}$
7:     Let $\gamma = \min\left(1, \frac{\nu}{2C} g_t\right)$ or $\gamma = \frac{2}{2+t}$    (ν and C set as case (I) in Thm. 1)
8:     Update $\boldsymbol{z}^{(t+1)} := (1-\gamma)\boldsymbol{z}^{(t)} + \gamma \boldsymbol{s}^{(t)}$
9: end for
```

Algorithm 3 Saddle point away-step Frank-Wolfe algorithm: **SP-AFW**$(\boldsymbol{z}^{(0)}, \mathcal{A} \times \mathcal{B}, \epsilon)$

```
1: Let $\boldsymbol{z}^{(0)} = (\boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}) \in \mathcal{A} \times \mathcal{B}$, $\mathcal{S}_x^{(0)} := \{\boldsymbol{x}^{(0)}\}$ and $\mathcal{S}_y^{(0)} := \{\boldsymbol{y}^{(0)}\}$
2: for $t = 0 \dots T$ do
3:     Let $\boldsymbol{s}^{(t)} := \mathrm{LMO}_{\mathcal{A} \times \mathcal{B}}(\boldsymbol{r}^{(t)})$ and $\boldsymbol{d}_{\mathrm{FW}}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}$    ($\boldsymbol{r}^{(t)}$ as defined in L3 in Algorithm 2)
4:     Let $\boldsymbol{v}^{(t)} \in \operatorname*{arg\,max}_{\boldsymbol{v} \in \mathcal{S}_x^{(t)} \times \mathcal{S}_y^{(t)}} \langle \boldsymbol{r}^{(t)}, \boldsymbol{v} \rangle$ and $\boldsymbol{d}_{\mathrm{A}}^{(t)} := \boldsymbol{z}^{(t)} - \boldsymbol{v}^{(t)}$    (the away direction)
5:     if $g_t^{\mathrm{FW}} := \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} \right\rangle \leq \epsilon$ then return $\boldsymbol{z}^{(t)}$    (FW gap is small enough, so return)
6:     if $\left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} \right\rangle \geq \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{A}}^{(t)} \right\rangle$ then
7:         $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{FW}}^{(t)}$, and $\gamma_{\max} := 1$    (choose the FW direction)
8:     else
9:         $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{A}}^{(t)}$, and $\gamma_{\max} := \min\left\{ \frac{\alpha_{\boldsymbol{v}_x^{(t)}}}{1-\alpha_{\boldsymbol{v}_x^{(t)}}}, \frac{\alpha_{\boldsymbol{v}_y^{(t)}}}{1-\alpha_{\boldsymbol{v}_y^{(t)}}} \right\}$    (maximum feasible step size; a drop step is when $\gamma_t = \gamma_{\max}$)
10:    end if
11:    Let $g_t^{\mathrm{PFW}} = \left\langle -\boldsymbol{r}^{(t)}, \boldsymbol{d}_{\mathrm{FW}}^{(t)} + \boldsymbol{d}_{\mathrm{A}}^{(t)} \right\rangle$ and $\gamma_t = \min\left\{ \gamma_{\max}, \frac{\nu^{\mathrm{PFW}}}{2C} g_t^{\mathrm{PFW}} \right\}$    (ν and C set as case (P) in Thm. 1)
12:    Update $\boldsymbol{z}^{(t+1)} := \boldsymbol{z}^{(t)} + \gamma_t \boldsymbol{d}^{(t)}$    (and accordingly for the weights $\boldsymbol{\alpha}^{(t+1)}$, see Lacoste-Julien and Jaggi (2015))
13:    Update $\mathcal{S}_x^{(t+1)} := \{\boldsymbol{v}_x \in \mathcal{A}$ s.t. $\alpha_{\boldsymbol{v}_x}^{(t+1)} > 0\}$ and $\mathcal{S}_y^{(t+1)} := \{\boldsymbol{v}_y \in \mathcal{B}$ s.t. $\alpha_{\boldsymbol{v}_y}^{(t+1)} > 0\}$
14: end for
```

Algorithm 4 Saddle point pairwise Frank-Wolfe algorithm: **SP-PFW**$(\boldsymbol{z}^{(0)}, \mathcal{A} \times \mathcal{B}, \epsilon)$

```
1: In Alg. 3, replace L6 to 10 by:  $\boldsymbol{d}^{(t)} := \boldsymbol{d}_{\mathrm{PFW}}^{(t)} := \boldsymbol{s}^{(t)} - \boldsymbol{v}^{(t)}$, and $\gamma_{\max} := \min\left\{ \alpha_{\boldsymbol{v}_x^{(t)}}, \alpha_{\boldsymbol{v}_x^{(t)}} \right\}$.
```

adaptive step sizes using the gap computation. We next analyze the convergence of these algorithms.

The suboptimality error and the gap. To establish convergence, we first define several quantities of interest. In classical convex optimization, the suboptimality error h_t is well defined as $h_t := f(\boldsymbol{x}^{(t)}) - \min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$. This quantity is clearly non-negative and proving that h_t goes to 0 is enough to establish convergence. Unfortunately, in the saddle point setting the quantity $\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \mathcal{L}^*$ is no longer non-negative and can be equal to zero for an infinite number of points $(\boldsymbol{x}, \boldsymbol{y})$ while $(\boldsymbol{x}, \boldsymbol{y}) \notin (\mathcal{X}^*, \mathcal{Y}^*)$. For instance, if $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ with $\mathcal{X} = \mathcal{Y} = [-1, 1]$, then $\mathcal{L}^* = 0$ and $(\mathcal{X}^*, \mathcal{Y}^*) = \{(0, 0)\}$. But for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{y} \in \mathcal{Y}$, $\boldsymbol{x} \cdot 0 = 0 \cdot \boldsymbol{y} = \mathcal{L}^*$. The saddle point literature thus considers a non-negative gap function (also known as a merit function (Larsson and Patriksson, 1994; Zhu and Marcotte, 1998) and (Patriksson, 1999, Sec 4.4.1)) which is zero only for optimal points, in order to quantify progress towards the saddle point. We can define the following *suboptimality error* h_t for our saddle point problem:

$$
\begin{aligned}
& h_t := \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}), \\
\text{where} \quad & \widehat{\boldsymbol{x}}^{(t)} := \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}^{(t)}), \\
\text{and} \quad & \widehat{\boldsymbol{y}}^{(t)} := \operatorname*{arg\,max}_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}).
\end{aligned}
\tag{5}
$$

This is an example of *primal-dual* gap function by noticing that

$$
\begin{aligned}
h_t &= \mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}^* + \mathcal{L}^* - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}) \\
&= p(\boldsymbol{x}^{(t)}) - p(\boldsymbol{x}^*) + g(\boldsymbol{y}^*) - g(\boldsymbol{y}^{(t)}),
\end{aligned}
\tag{6}
$$

where $p(\boldsymbol{x}) := \max_{\boldsymbol{y} \in \mathcal{Y}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ is the convex primal function and $g(\boldsymbol{y}) := \min_{\boldsymbol{x} \in \mathcal{X}} \mathcal{L}(\boldsymbol{x}, \boldsymbol{y})$ is the concave dual function. By convex-concavity, h_t can be upper-bounded by the following FW linearization gap (Jaggi, 2011, 2013; Larsson and Patriksson, 1994; Zhu and Marcotte, 1998):

$$g_t^{\mathrm{FW}} := \underbrace{\max_{\boldsymbol{s}_x \in \mathcal{X}} \left\langle \boldsymbol{x}^{(t)} - \boldsymbol{s}_x, \nabla_x \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle}_{:= g_t^{(x)}} + \underbrace{\max_{\boldsymbol{s}_y \in \mathcal{Y}} \left\langle \boldsymbol{y}^{(t)} - \boldsymbol{s}_y, -\nabla_y \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) \right\rangle}_{:= g_t^{(y)}} . \tag{7}$$

This gap is easy to compute and gives a stopping criterion since $g_t^{\mathrm{FW}} \geq h_t$.

Compensation phenomenon and difficulty for SP. Even when equipped with a suboptimality error and a gap function (as in the convex case), we still cannot apply the standard FW convergence analysis. The usual FW proof sketch uses the fact that the gradient of f is Lipschitz continuous to get

$$h_{t+1} \leq h_t - \gamma_t g_t^{\mathrm{FW}} + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2} \tag{8}$$

which then provides a rate of convergence. Roughly, since $g_t \geq h_t$ by convexity, if γ_t is small enough then (h_t) will decrease and converge. For simplicity, in the main paper, $\|\cdot\|$ will refer to the ℓ_2 norm of $\mathbb{R}^d$. The partial Lipschitz constants and the diameters of the sets are defined with respect to this norm (see (40) in Appendix B.1 for more general norms).

Using the L-Lipschitz continuity of $\mathcal{L}$ and letting $\mathcal{L}_t := \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$ as a shorthand, we get

$$\mathcal{L}_{t+1} \leq \mathcal{L}_t + \gamma_t \left\langle \boldsymbol{d}_x^{(t)}, \nabla_x \mathcal{L}_t \right\rangle + \gamma_t \left\langle \boldsymbol{d}_y^{(t)}, \nabla_y \mathcal{L}_t \right\rangle + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2} \tag{9}$$

where $\boldsymbol{d}_x^{(t)} = \boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}$ and $\boldsymbol{d}_y^{(t)} = \boldsymbol{s}_y^{(t)} - \boldsymbol{y}^{(t)}$. Then

$$\mathcal{L}_{t+1} - \mathcal{L}^* \leq \mathcal{L}_t - \mathcal{L}^* - \gamma_t \left(g_t^{(x)} - g_t^{(y)} \right) + \gamma_t^2 \frac{L\|\boldsymbol{d}^{(t)}\|^2}{2}. \tag{10}$$

Unfortunately, the quantity g_t^{FW} does *not* appear above and we therefore cannot control the oscillation of the sequence (the quantity $g_t^{(x)} - g_t^{(y)}$ can make the sequence increase or decrease). Instead, we must focus on more specific SP optimization settings and introduce other quantities of interest in order to establish convergence.

The asymmetry of the SP. Hammond (1984, p. 165) showed the divergence of the SP-FW algorithm with an extended line-search step-size on some bilinear objectives. She mentioned that the difficulty for SP optimization is contained in this bilinear coupling between $\boldsymbol{x}$ and $\boldsymbol{y}$. More generally, most of the examples of SP functions cited in the introduction can be written in the form:

$$\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{x}^\top M \boldsymbol{y} - g(\boldsymbol{y}), \ f \text{ and } g \text{ convex.} \tag{11}$$

In this setting, the bilinear part M is the only term preventing us to apply theorems on standard FW. Hammond (1984, p. 175) also conjectured that the SP-FW algorithm with $\gamma_t = {}^1/_{(t+1)}$ performed on a uniformly strongly convex-concave objective function (see (12)) over a polytope should converge. We give a partial answer to this conjecture in the following section.

3 SP-FW for strongly convex functions

Uniform strong convex-concavity. In this section, we will assume that $\mathcal{L}$ is uniformly $(\mu_{\mathcal{X}}, \mu_{\mathcal{Y}})$-strongly convex-concave, which means that the following function is convex-concave:

$$(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) - \frac{\mu_{\mathcal{X}}}{2}\|\boldsymbol{x}\|^2 + \frac{\mu_{\mathcal{Y}}}{2}\|\boldsymbol{y}\|^2. \tag{12}$$

A new merit function. To prove our theorem, we use a different quantity w_t which is smaller than h_t but still a valid merit function in the case of *strongly convex-concave* SPs (where $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is thus unique); see (14) below. For $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ a solution of (1), we define the non-negative quantity w_t:

$$w_t := \underbrace{\mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}^*}_{:= w_t^{(x)}} + \underbrace{\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)})}_{:= w_t^{(y)}} . \tag{13}$$

Notice that $w_t^{(x)}$ and $w_t^{(y)}$ are non-negative, and that $w_t \leq h_t$ since:

$$\mathcal{L}(\boldsymbol{x}^{(t)}, \widehat{\boldsymbol{y}}^{(t)}) - \mathcal{L}(\widehat{\boldsymbol{x}}^{(t)}, \boldsymbol{y}^{(t)}) \geq \mathcal{L}(\boldsymbol{x}^{(t)}, \boldsymbol{y}^*) - \mathcal{L}(\boldsymbol{x}^*, \boldsymbol{y}^{(t)}).$$

In general, w_t can be zero even if we have not reached a solution. For example, with $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ and $\mathcal{X} = \mathcal{Y} = [-1, 1]$, then $\boldsymbol{x}^* = \boldsymbol{y}^* = \mathbf{0}$, implying $w_t = 0$ for any $(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)})$. But for a uniformly strongly convex-concave $\mathcal{L}$, this cannot happen and we can prove that w_t has the following nice property (akin to $\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \sqrt{{}^2/_\mu (f(\boldsymbol{x}) - f(\boldsymbol{x}^*))}$ for a μ-strongly convex function f; see Proposition 15 in Appendix B.6):

$$h_t \leq \sqrt{2} P_{\mathcal{L}} \sqrt{w_t}\,, \tag{14}$$

where

$$P_{\mathcal{L}} \leq \sqrt{2} \sup_{\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}} \left\{ \frac{\|\nabla_x \mathcal{L}(\boldsymbol{z})\|_{\mathcal{X}^*}}{\sqrt{\mu_{\mathcal{X}}}}, \frac{\|\nabla_y \mathcal{L}(\boldsymbol{z})\|_{\mathcal{Y}^*}}{\sqrt{\mu_{\mathcal{Y}}}} \right\}. \tag{15}$$

Pyramidal width and distance to the border. We now provide a theorem that establishes convergence in two situations: (I) when the SP belongs to the interior of $\mathcal{X} \times \mathcal{Y}$; (P) when the set is a polytope, i.e. when there exist two finite sets such that $\mathcal{X} = \mathrm{conv}(\mathcal{A})$ and $\mathcal{Y} = \mathrm{conv}(\mathcal{B})$. Our convergence result holds when (roughly) the strong convex-concavity of $\mathcal{L}$ is big enough in comparison to the cross Lipschitz constants L_{XY}, L_{YX} of $\nabla\mathcal{L}$ (defined in (20) below) multiplied by geometric "condition numbers" of each set. The condition number of $\mathcal{X}$ (and similarly for $\mathcal{Y}$) is defined as the ratio of its *diameter* $D_{\mathcal{X}} := \sup_{\boldsymbol{x},\boldsymbol{x}'\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{x}'\|$ over the following appropriate notions of "width":

$$\text{border distance: } \delta_{\mathcal{X}} := \min_{\boldsymbol{s}\in\partial\mathcal{X}} \|\boldsymbol{x}^* - \boldsymbol{s}\| \quad \text{for (I)}, \tag{16}$$

$$\text{pyramidal width: } \delta_{\mathcal{A}} := PWidth(\mathcal{A}) \qquad \text{for (P)}. \tag{17}$$

The pyramidal width (17) is formally defined in Eq. 9 of Lacoste-Julien and Jaggi (2015) and in Appendix B.3. Given the above constants, we can state below a non-affine invariant version of our convergence theorem (for simplicity). The affine invariant versions of this theorem are given in Thm. 24 and 25 in Appendix D.2 (with proofs).

Theorem 1. *Let $\mathcal{L}$ be a convex-concave function and $\mathcal{X}\times\mathcal{Y}$ a convex and compact set. Assume that the gradient of $\mathcal{L}$ is L-Lipschitz continuous, that $\mathcal{L}$ is $(\mu_{\mathcal{X}},\mu_{\mathcal{Y}})$-strongly convex-concave, and that we are in one of the two following situations:*

(I) *The SP belongs to the interior of $\mathcal{X}\times\mathcal{Y}$. In this case, set $g_t = g_t^{\text{FW}}$ (as in L5 of Alg. 3), $\delta_\mu := \sqrt{\min(\mu_{\mathcal{X}}\delta_{\mathcal{X}}^2, \mu_{\mathcal{Y}}\delta_{\mathcal{Y}}^2)}$ and $a := 1$. "Algorithm" then refers to SP-FW.*

(P) *The sets $\mathcal{X}$ and $\mathcal{Y}$ are polytopes. In this case, set $g_t = g_t^{\text{PFW}}$ (as in L11 of Alg. 3), $\delta_\mu := \sqrt{\min(\mu_{\mathcal{X}}\delta_{\mathcal{A}}^2, \mu_{\mathcal{Y}}\delta_{\mathcal{B}}^2)}$ and $a := \frac{1}{2}$. "Algorithm" then refers to SP-AFW. Here δ_μ needs to use the Euclidean norm for its defining constants.*

In both cases, if $\nu := a - \frac{\sqrt{2}}{\delta_\mu}\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\sqrt{\mu_{\mathcal{X}}}}\right\}$ is positive, then the errors h_t (5) of the iterates of the algorithm with step size $\gamma_t = \min\{\gamma_{\max}, \frac{\nu}{2C}g_t\}$ decrease geometrically as

$$h_t = O\left((1-\rho)^{\frac{k(t)}{2}}\right) \ \text{and} \ \min_{s\le t} g_s^{\text{FW}} = O\left((1-\rho)^{\frac{k(t)}{2}}\right)$$

where $\rho := \nu^2\frac{\delta_\mu^2}{2C}$, $C := \frac{LD_{\mathcal{X}}^2+LD_{\mathcal{Y}}^2}{2}$ and $k(t)$ is the number of non-drop step after t steps (see L9 in Alg. 3). In case (I) we have $k(t) = t$ and in case (P) we have $k(t) \ge t/3$. For both algorithms, if $\delta_\mu > 2\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\mu_{\mathcal{X}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\mu_{\mathcal{Y}}}\right\}$, we also obtain a sublinear rate with the universal choice $\gamma_t = \min\{\gamma_{\max}, \frac{2}{2+k(t)}\}$. This yields the rates:

$$\min_{s\le t} h_s \le \min_{s\le t} g_s^{\text{FW}} = O\left(\frac{1}{t}\right). \tag{18}$$

Clearly, the sublinear rate seems less interesting than the linear one but has the added convenience that the step size can be set without knowledge of various constants that characterize $\mathcal{L}$. Moreover, it provides a partial answer to the conjecture from Hammond (1984).

Proof sketch. Strong convexity is an essential assumption in our proof; it allows us to relate w_t to how close we are to the optimum. Actually, by $\mu_{\mathcal{Y}}$-strong concavity of $\mathcal{L}(\boldsymbol{x}^*,\cdot)$, we have

$$\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\| \le \sqrt{\frac{2}{\mu_{\mathcal{Y}}}\left(\mathcal{L}^* - \mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right)} = \sqrt{\frac{2}{\mu_{\mathcal{Y}}}w_t^{(y)}}. \tag{19}$$

Now, recall that we assumed that $\nabla\mathcal{L}$ is Lipschitz continuous. In the following, we will call L the *Lipschitz continuity constant* of $\nabla\mathcal{L}$ and L_{XY} and L_{YX} its (cross) *partial Lipschitz constants.* For all $\boldsymbol{x}$, $\boldsymbol{x}'\in\mathcal{X}$, $\boldsymbol{y}$, $\boldsymbol{y}'\in\mathcal{Y}$, these constants satisfy

$$\begin{aligned} \|\nabla_x\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_x\mathcal{L}(\boldsymbol{x},\boldsymbol{y}')\|_{\mathcal{X}^*} &\le L_{XY}\|\boldsymbol{y}-\boldsymbol{y}'\|_{\mathcal{Y}}, \\ \|\nabla_y\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) - \nabla_y\mathcal{L}(\boldsymbol{x}',\boldsymbol{y})\|_{\mathcal{Y}^*} &\le L_{YX}\|\boldsymbol{x}-\boldsymbol{x}'\|_{\mathcal{X}}. \end{aligned} \tag{20}$$

Note that $L_{XY}, L_{YX} \le L$ if $\|(\boldsymbol{x},\boldsymbol{y})\| := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$. Then, using Lipschitz continuity of the gradient,

$$\begin{aligned} \mathcal{L}(\boldsymbol{x}^{(t+1)},\boldsymbol{y}^*) \le \mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^*) &+ \gamma\langle \boldsymbol{d}_x^{(t)}, \nabla_x\mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^*)\rangle \\ &+ \gamma^2\frac{L\|\boldsymbol{d}_x^{(t)}\|^2}{2}. \end{aligned} \tag{21}$$

Furthermore, setting $(\boldsymbol{x},\boldsymbol{y}) = (\boldsymbol{x}^{(t)},\boldsymbol{y}^*)$ and $\boldsymbol{y}' = \boldsymbol{y}^{(t)}$ in Equation (20), we have

$$\begin{aligned} w_{t+1}^{(x)} \le w_t^{(x)} - \gamma g_t^{(x)} &+ \gamma D_{\mathcal{X}}L_{XY}\|\boldsymbol{y}^{(t)}-\boldsymbol{y}^*\| \\ &+ \gamma^2\frac{LD_{\mathcal{X}}^2}{2}. \end{aligned} \tag{22}$$

Finally, combining (22) and (19), we get

$$\begin{aligned} w_{t+1}^{(x)} \le w_t^{(x)} - \gamma g_t^{(x)} &+ \gamma D_{\mathcal{X}}L_{XY}\sqrt{\frac{2}{\mu_{\mathcal{Y}}}}\sqrt{w_t^{(y)}} \\ &+ \gamma^2\frac{LD_{\mathcal{X}}^2}{2}. \end{aligned} \tag{23}$$

A similar argument on $-\mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t+1)})$ gives a bound on $w_t^{(y)}$ much like (23). Summing both yields:

$$\begin{aligned} w_{t+1} \le w_t - \gamma g_t &+ 2\gamma\max\left\{\frac{D_{\mathcal{X}}L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \frac{D_{\mathcal{Y}}L_{YX}}{\sqrt{\mu_{\mathcal{X}}}}\right\}\sqrt{w_t} \\ &+ \gamma^2\frac{LD_{\mathcal{X}}^2+LD_{\mathcal{Y}}^2}{2}. \end{aligned} \tag{24}$$

We now apply recent developments in the convergence theory of FW methods for strongly convex objectives. Lacoste-Julien and Jaggi (2015) crucially upper bound the square root of the suboptimality error on a convex function with the FW gap if the optimum is in the interior, or with the PFW gap if the set is a polytope (Lemma 18 in Appendix C.2). We continue our proof sketch for case (I) only:[2]

$$\begin{aligned} 2\mu_{\mathcal{X}}\delta_{\mathcal{X}}^2\left(\mathcal{L}(\boldsymbol{x}^{(t)},\boldsymbol{y}^{(t)}) - \mathcal{L}(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right) &\le \left(g_t^{(x)}\right)^2 \\ \text{where} \quad \delta_{\mathcal{X}} := \min_{\boldsymbol{s}\in\partial\mathcal{X}}\|\boldsymbol{x}^*-\boldsymbol{s}\|. \end{aligned} \tag{25}$$

[2] The idea is similar for case (P), but with the additional complication of possible drop steps.

We can also get the respective equation on $\boldsymbol{y}$ with $\delta_{\mathcal{Y}} := \min_{\boldsymbol{y} \in \partial \mathcal{Y}} \|\boldsymbol{y}^* - \boldsymbol{y}\|$ and sum it with the previous one (25) to get:

$$\delta_\mu \sqrt{2w_t} \leq g_t \text{ where } \delta_\mu := \sqrt{\min(\mu_{\mathcal{X}} \delta_{\mathcal{X}}^2, \mu_{\mathcal{Y}} \delta_{\mathcal{Y}}^2)}. \quad (26)$$

Plugging this last equation into (23) gives us

$$\begin{aligned} w_{t+1} \leq w_t - \nu\gamma g_t + \gamma^2 C \text{ where } C := \tfrac{LD_{\mathcal{X}}^2 + LD_{\mathcal{Y}}^2}{2} \\ \text{and } \nu := 1 - \tfrac{\sqrt{2}}{\delta_\mu} \max\left\{ \tfrac{D_{\mathcal{X}} L_{XY}}{\sqrt{\mu_{\mathcal{Y}}}}, \tfrac{D_{\mathcal{Y}} L_{YX}}{\sqrt{\mu_{\mathcal{X}}}} \right\}. \end{aligned} \quad (27)$$

The recurrence (27) is typical in the FW literature. We can re-apply standard techniques on the sequence w_t to get a sublinear rate with $\gamma_t = \frac{2}{2+t}$, or a linear rate with $\gamma_t = \min\left\{\gamma_{\max}, \frac{\nu g_t}{2C}\right\}$ (which minimizes the RHS of (27) and actually guarantees that w_t will be *decreasing*). Finally, thanks to strong convexity, a rate on w_t gives us a rate on h_t (by (14)). □

4 SP-FW with strongly convex sets

Strongly convex set. One can (roughly) define strongly convex sets as sublevel sets of strongly convex functions (Vial, 1983, Prop. 4.14). In this section, we replace the strong convex-concavity assumption on $\mathcal{L}$ with the assumption that $\mathcal{X}$ and $\mathcal{Y}$ are β-strongly convex *sets*.

Definition 2 (Vial (1983); Polyak (1966))**.** *A convex set $\mathcal{X}$ is said to be β-strongly convex with respect to $\|.\|$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$ and any $\gamma \in [0,1]$, $B_\beta(\gamma, \boldsymbol{x}, \boldsymbol{y}) \subset \mathcal{X}$ where $B_\beta(\gamma, \boldsymbol{x}, \boldsymbol{y})$ is the $\|.\|$-ball of radius $\gamma(1-\gamma)\frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2$ centered at $\gamma\boldsymbol{x} + (1-\gamma)\boldsymbol{y}$.*

Frank-Wolfe for convex optimization over strongly convex sets has been studied by Levitin and Polyak (1966); Demyanov and Rubinov (1970) and Dunn (1979), amongst others. They all obtained a linear rate for the FW algorithm if the norm of the gradient is lower bounded by a constant. More recently, Garber and Hazan (2015) proved a sublinear rate $O(1/t^2)$ by replacing the lower bound on the gradient by a strong convexity assumption on the function. In the VIP setting (3), the linear convergence has been proved if the optimization is done under a strongly convex set but this assumption does *not* extend to $\mathcal{X} \times \mathcal{Y}$ which *cannot* be strongly convex if $\mathcal{X}$ or $\mathcal{Y}$ is not reduced to a single element. In order to prove the convergence, we first prove the Lipschitz continuity of the *FW-corner* function $\boldsymbol{s}(\cdot)$ defined below. A proof of this theorem is given in Appendix E.

Theorem 3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be β-strongly convex sets. If $\min(\|\nabla_x L(\boldsymbol{z})\|_{\mathcal{X}^*}, \|\nabla_y L(\boldsymbol{z})\|_{\mathcal{Y}^*}) \geq \delta > 0$ for all $\boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$, then the oracle function $\boldsymbol{z} \mapsto \boldsymbol{s}(\boldsymbol{z}) := \arg\min_{\boldsymbol{s} \in \mathcal{X} \times \mathcal{Y}} \langle \boldsymbol{s}, \boldsymbol{r}(\boldsymbol{z}) \rangle$ is well defined and is $\frac{4L}{\delta\beta}$-Lipschitz continuous (using the norm $\|(\boldsymbol{x}, \boldsymbol{y})\|_{\mathcal{X} \times \mathcal{Y}} := \|\boldsymbol{x}\|_{\mathcal{X}} + \|\boldsymbol{y}\|_{\mathcal{Y}}$), where $\boldsymbol{r}(\boldsymbol{z}) := (\nabla_x \mathcal{L}(\boldsymbol{z}), -\nabla_y \mathcal{L}(\boldsymbol{z}))$.*

Convergence rate. When the FW-corner function $\boldsymbol{s}(\cdot)$ is Lipschitz continuous (by Theorem 3), we can actually show that the FW gap is decreasing in the FW direction and get a similar inequality as the standard FW one (8), but, in this case, on the *gaps*: $g_{t+1} \leq g_t(1-\gamma_t) + \gamma_t^2 \|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta$. Moreover, one can show that the FW gap on a strongly convex set $\mathcal{X}$ can be lower-bounded by $\|\boldsymbol{s}_x^{(t)} - \boldsymbol{x}^{(t)}\|^2$ (Lemma 27 in Appendix E), by using the fact that $\mathcal{X}$ contains a ball of sufficient radius around the midpoint between $\boldsymbol{s}_x^{(t)}$ and $\boldsymbol{x}^{(t)}$. From these two facts, we can prove the following linear rate of convergence (*not* requiring any *strong* convex-concavity of $\mathcal{L}$).

Theorem 4. *Let $\mathcal{L}$ be a convex-concave function and $\mathcal{X}$ and $\mathcal{Y}$ two compact β-strongly convex sets. Assume that the gradient of $\mathcal{L}$ is L-Lipschitz continuous and that there exists $\delta > 0$ such that $\min(\|\nabla_x L(\boldsymbol{z})\|_*, \|\nabla_y L(\boldsymbol{z})\|_*) \geq \delta \ \forall \boldsymbol{z} \in \mathcal{X} \times \mathcal{Y}$. Set $C_\delta := 2L + \frac{8L^2}{\beta\delta}$. Then the gap g_t^{FW} (7) of the SP-FW algorithm with step size $\gamma_t = \frac{g_t^{\mathrm{FW}}}{\|\boldsymbol{s}^{(t)} - \boldsymbol{z}^{(t)}\|^2 C_\delta}$ converges linearly as $g_t^{\mathrm{FW}} \leq g_0 (1-\rho)^t$, where $\rho := \frac{\beta\delta}{16 C_\delta}$.*

5 SP-FW in the bilinear setting

Fictitious play. In her thesis, Hammond (1984, § 4.3.1) pointed out that for the bilinear setting:

$$\min_{\boldsymbol{x} \in \Delta_p} \max_{\boldsymbol{y} \in \Delta_q} \boldsymbol{x}^\top M \boldsymbol{y} \quad (28)$$

where Δ_p is the probability simplex on p elements, the SP-FW algorithm with step size $\gamma_t = 1/(1+t)$ is equivalent to the fictitious play (FP) algorithm introduced by Brown (1951). The FP algorithm has been widely studied in the game literature. Its convergence has been proved by Robinson (1951), while Shapiro (1958) showed that one can deduce from Robinson's proof a $O(t^{-1/(p+q-2)})$ rate. Around the same time, Karlin (1960) conjectured that the FP algorithm converged at the better rate of $O(t^{-1/2})$, though this conjecture is still open and Shapiro's rate is the only one we are aware of. Interestingly, Daskalakis and Pan (2014) recently showed that Shapiro's rate is also a lower bound if the tie breaking rule gets the worst pick an infinite number of times. Nevertheless, this kind of adversarial tie breaking rule does not seems realistic since this rule is a priori defined by the programmer. In practical cases (by setting a fixed prior order for ties or picking randomly for example), Karlin's Conjecture (Karlin, 1960) is still open. Moreover, we always observed an empirical rate of at least $O(t^{-1/2})$ during our experiments, we thus believe the conjecture to be true for realistic tie breaking rules.

Rate for SP-FW. Via the affine invariance of the FW algorithm and the fact that every polytope with p vertices is the affine transformation of a probability simplex of dimension p, any rate for the fictitious play algorithm implies a rate for SP-FW.

Corollary 5. *For polytopes $\mathcal{X}$ and $\mathcal{Y}$ with p and q vertices respectively and $\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\top M\boldsymbol{y}$, the SP-FW algorithm with step size $\gamma_t = \frac{1}{t+1}$ converges at the rate $h_t = O\left(t^{-\frac{1}{p+q-2}}\right)$.*

This (very slow) convergence rate is mainly of theoretical interest, providing a safety check that the algorithm actually converges. Moreover, if Karlin's strong conjecture is true, we can get a $O(1/\sqrt{t})$ worst case rate which is confirmed by our experiments.

6 Experiments

Toy experiments. First, we test the empirical convergence of our algorithms on a simple saddle point problem over the unit cube in dimension d (whose pyramidal width has the explicit value $1/\sqrt{d}$ by Lemma 4 from Lacoste-Julien and Jaggi (2015)). Thus $\mathcal{X} = \mathcal{Y} := [0,1]^d$ and the linear minimization oracle is simply $\text{LMO}(\cdot) = -0.5 \cdot (\text{sign}(\cdot) - \mathbf{1})$. We consider the following objective function:

$$\frac{\mu}{2}\|\boldsymbol{x}-\boldsymbol{x}^*\|_2^2+(\boldsymbol{x}-\boldsymbol{x}^*)^\top M(\boldsymbol{y}-\boldsymbol{y}^*)-\frac{\mu}{2}\|\boldsymbol{y}-\boldsymbol{y}^*\|_2^2 \quad (29)$$

for which we can control the location of the saddle point $(\boldsymbol{x}^*,\boldsymbol{y}^*) \in \mathcal{X}\times\mathcal{Y}$. We generate a matrix M randomly as $M \sim \mathcal{U}([-0.1, 0.1]^{d\times d})$ and keep it fixed for all experiments. For the interior point setup (I), we set $(\boldsymbol{x}^*,\boldsymbol{y}^*) \sim \mathcal{U}([0.25, 0.75]^{2d})$, while we set $\boldsymbol{x}^*$ and $\boldsymbol{y}^*$ to some fixed random vertex of the unit cube for the setup (P). With all these parameters fixed, the constant ν is a function of μ only. We thus vary the strong convexity parameter μ to test various ν's.

We verify the linear convergence expected for the SP-FW algorithm for case (I) in Figure 1a, and for the SP-AFW algorithm for case (P) in Figure 1b. As the adaptive step size (and rate) depends linearly on ν, the linear rate becomes quite slow for small ν. In this regime (in red), the step size $2/(2 + k(t))$ (in orange) can actually perform better, despite its theoretical sublinear rate.

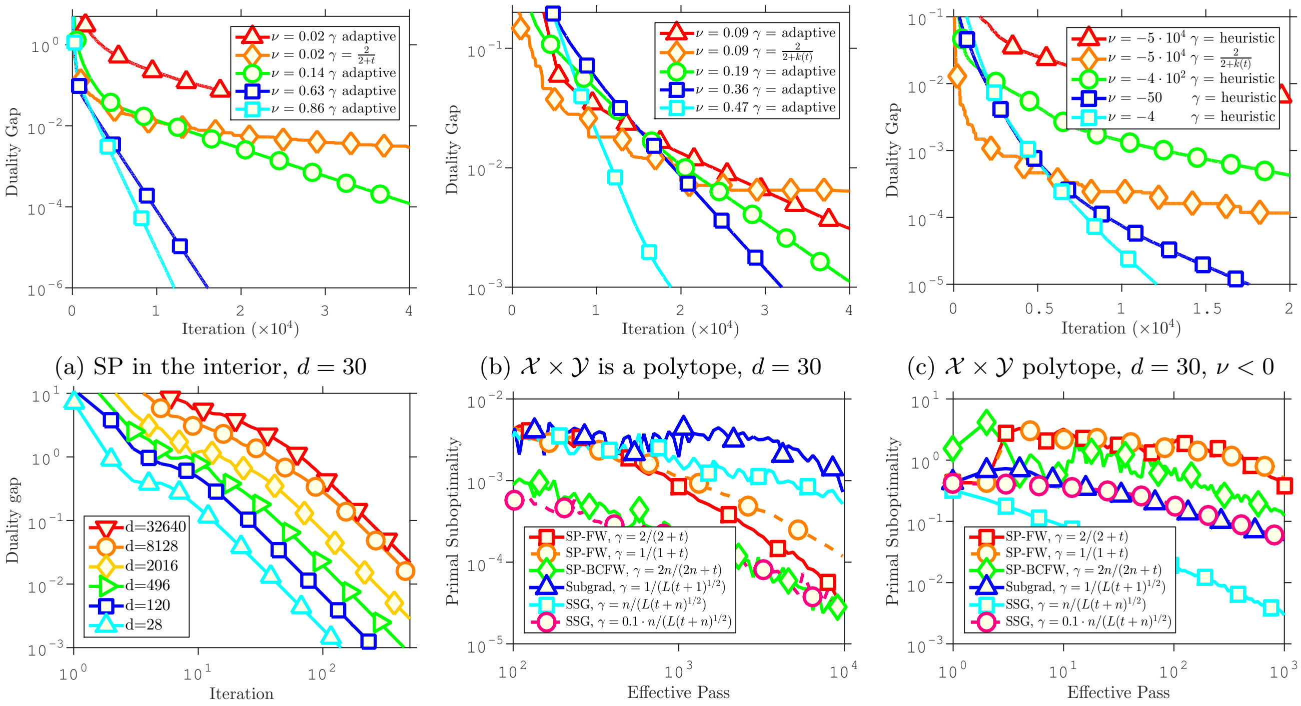
Finally, figure 1c shows that we can observe a linear convergence of SP-AFW even if ν is negative by using a different step size. In this case, we use the heuristic adaptive step size $\gamma_t := g_t/\tilde{C}$ where $\tilde{C} := LD_{\mathcal{X}}^2 + LD_{\mathcal{Y}}^2 + L_{XY}L_{YX}\left(D_{\mathcal{X}}^2/\mu_{\mathcal{X}} + D_{\mathcal{Y}}^2/\mu_{\mathcal{Y}}\right)$. Here $\tilde{C}$ takes into account the coupling between the concave and the convex variable and is motivated from a different proof of convergence that we were not able to complete. The empirical linear convergence in this case is not yet supported by a complete analysis, highlighting the need for more sophisticated arguments.

Graphical games. We now consider a bilinear objective $\mathcal{L}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\top M\boldsymbol{y}$ where exact projections on the sets is intractable, but we have a tractable LMO. The problem is motivated from the following setup. We consider a game between two universities (A and B) that are admitting s students and have to assign pairs of students into dorms. If students are unhappy with their dorm assignments, they will go to the other university. The game has a payoff matrix M belonging to $\mathbb{R}^{(s(s-1)/2)^2}$ where $M_{ij,kl}$ is the expected tuition that B gets (or A gives up) if A pairs student i with j and B pairs student k with l. Here the actions $\boldsymbol{x}$ and $\boldsymbol{y}$ are both in the marginal polytope of all perfect unipartite matchings. Assume that we are given a graph $G = (V, E)$ with vertices V and edges E. For a subset of nodes $S \subseteq V$, let the induced subgraph $G(S) = (S, E(S))$. Edmonds (1965) showed that any subgraph forming a triangle can contain at most one edge of any perfect matching. This forms an exponential set of linear equalities which define the matching polytope $\mathcal{P}(G) \subset \mathbb{R}^E$ as

$$\{\boldsymbol{x}|\boldsymbol{x}_e\geq 0,\sum_{e\in E(S)}\boldsymbol{x}_e\leq k,\forall S\subseteq V,|S|=2k+1,\forall e\in E\}. \quad (30)$$

While this strategy space seems daunting, the LMO can be solved in $\mathcal{O}(s^3)$ time using the blossom algorithm (Edmonds, 1965). We run the SP-FW algorithm with $\gamma_t = {}^2/_{(t+2)}$ on this problem with $s = 2^j$ students for $j = 3, \ldots, 8$ with results given in Figure 1d ($d = s(s-1)/2$ in the legend represents the dimensionality of the $\boldsymbol{x}$ and $\boldsymbol{y}$ variables). The order of the complexity of the LMO is then $O(d^{3/2})$. In Figure 1d, the observed empirical rate of the SP-FW algorithm (using $\gamma_t = {}^2/_{(t+2)}$) is $O(1/t^2)$. Empirically, faster rates seem to arise if the solution is at a corner (a pure equilibrium, to be expected for random payoff matrices in light of (Bárány et al., 2007)).

Sparse structured SVM. We finally consider a challenging optimization problem arising from structured prediction. We consider the saddle point formulation (Taskar et al., 2006) for a ℓ_1-regularized structured SVM objective that minimizes the primal cost function $p(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^n \tilde{H}_i(\boldsymbol{w})$, where $\tilde{H}_i(\boldsymbol{w}) = \max_{\boldsymbol{y}\in\mathcal{Y}_i} L_i(\boldsymbol{y}) - \langle \boldsymbol{w}, \boldsymbol{\psi}_i(\boldsymbol{y})\rangle$ is the structured hinge loss (using the notation from Lacoste-Julien et al. (2013)). We only assume access to the linear oracle computing $\tilde{H}_i(\boldsymbol{w})$. Let M_i have $\left(\boldsymbol{\psi}_i(\boldsymbol{y})\right)_{\boldsymbol{y}\in\mathcal{Y}_i}$ as columns. We can rewrite the minimization problem as a bilinear saddle point problem:



(a) SP in the interior, $d = 30$

(b) $\mathcal{X} \times \mathcal{Y}$ is a polytope, $d = 30$

(c) $\mathcal{X} \times \mathcal{Y}$ polytope, $d = 30$, $\nu < 0$

(d) Graphical games

(e) OCR dataset, $R = 0.01$.

(f) OCR dataset, $R = 5$.

Figure 1: On Figures 1a, 1b and 1c, we plot on a semilog scale the best gap observed $\min_{s \le t} g_s^{\text{FW}}$ as a function of t. For experiments 1d, 1e and 1f, the objective function is bilinear and the convergence is sublinear. An effective pass is one iteration for SP-FW or the subgradient method and n iterations for SP-BCFW or SSG. We give more details about these experiments in Appendix F.

$$\begin{aligned} &\min_{\|\boldsymbol{w}\|_1 \le R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{y}_i \in \mathcal{Y}_i} \boldsymbol{L}_i^\top \boldsymbol{y}_i - \boldsymbol{w}^\top M_i \boldsymbol{y}_i \Big) \\ = &\min_{\|\boldsymbol{w}\|_1 \le R} \frac{1}{n} \sum_i \Big(\max_{\boldsymbol{\alpha}_i \in \Delta(|\mathcal{Y}_i|)} \boldsymbol{L}_i^\top \boldsymbol{\alpha}_i - \boldsymbol{w}^\top M_i \boldsymbol{\alpha}_i \Big). \end{aligned} \tag{31}$$

Projecting onto $\Delta(|\mathcal{Y}_i|)$ is normally intractable as the size of $|\mathcal{Y}_i|$ is exponential, but the linear oracle is tractable by assumption. We performed experiments with 100 examples from the OCR dataset ($d_\omega = 4028$) (Taskar et al., 2003). We encoded the structure $\mathcal{Y}_i$ of the i^{th} word with a Markov model: its k^{th} character $\mathcal{Y}_i^{(k)}$ only depends on $\mathcal{Y}_i^{k-1}$ and $\mathcal{Y}_i^{k+1}$. In this case, the oracle function is simply the Viterbi algorithm Viterbi (1967). The average length of a word is approximately 8, hence the dimension of $\mathcal{Y}_i$ is $d_{\mathcal{Y}_i} \approx 26^2 \cdot 8 = 5408$ leading to a large dimension for $\mathcal{Y}$, $d_\mathcal{Y} := \sum_{i=1}^n d_{\mathcal{Y}_i} \approx 5 \cdot 10^5$. We run the SP-FW algorithm with step size $\gamma_t = 1/(1+t)$ for which we have a convergence proof (Corollary 5), and with $\gamma_t = 2/(2+t)$, which normally gives better results for FW optimization. We compare with the projected subgradient method (projecting on the ℓ_1-ball is tractable here) with step size $O(1/\sqrt{t})$ (the subgradient of $\tilde{H}_i(\boldsymbol{w})$ is $-\boldsymbol{\psi}_i(\boldsymbol{y}_i^*)$). Following Lacoste-Julien et al. (2013), we also implement a block-coordinate (SP-BCFW) version of SP-FW and compare it with the stochastic projected subgradient method (SSG). As some of the algorithms only work on the primal and to make our result comparable to Lacoste-Julien et al. (2013), we choose to plot the primal suboptimality error $p(\boldsymbol{w}_t) - p^*$ for the different algorithms in Figure 1e and 1f (the $\boldsymbol{\alpha}_t$ iterates for the SP approaches are thus ignored in this error). The performance of SP-BCFW is similar to SSG when we regularize the learning problem heavily (Figure 1e). However, under lower regularization (Figure 1f), SSG (with the correct step size scaling) is faster. This is consistent with the fact that $\boldsymbol{\alpha_t} \neq \boldsymbol{\alpha}^*$ implies larger errors on the primal suboptimality for the SP methods, but we note that an advantage of the SP-FW approach is that the scale of the step size is automatically chosen.

Conclusion. We proposed FW-style algorithms for saddle-point optimization with the same attractive properties as FW, in particular only requiring access to a LMO. We gave the first convergence result for a FW-style algorithm towards a saddle point over polytopes by building on the recent developments on the linear convergence analysis of AFW. However, our experiments let us believe that the condition $\nu > 0$ is not required for the convergence of FW-style algorithms. We thus conjecture that a refined analysis could yield a linear rate for the general uniformly strongly convex-concave functions in both cases (I) and (P), paving the way for further theoretical work.

Acknowledgments

Thanks to N. Ruozzi and A. Benchaouine for helpful discussions. Work supported in part by DARPA N66001-15-2-4026, N66001-15-C-4032 and NSF III-1526914, IIS-1451500, CCF-1302269.

References

A. Ahmadinejad, S. Dehghani, Hajiaghayi, B. Lucier, H. Mahini, and S. Seddighin. From duels to battlefields: Computing equilibria of Blotto and other games. In *AAAI*, 2016.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 2012.

F. Barahona. On the computational complexity of Ising spin glass models. *J. Phys. A: Math. Gen.*, 1982.

I. Bárány, S. Vempala, and A. Vetta. Nash equilibria in random games. *Random Structures & Algorithms*, 31(4):391–405, 2007.

G. Brown. Iterative solution of games by fictitious play. *Activity analysis of production & allocation*, 1951.

B. Cox, A. Juditsky, and A. Nemirovski. Decomposition techniques for bilinear saddle point problems and variational inequalities with affine monotone operators on domains given by linear minimization oracles. *arXiv preprint arXiv:1506.02444*, 2015.

C. Daskalakis and Q. Pan. A counter-example to Karlin's strong conjecture for fictitious play. In *FOCS*, 2014.

V. F. Demyanov and A. M. Rubinov. *Approximate methods in optimization problems*. Elsevier, 1970.

J. C. Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 1979.

J. Edmonds. Paths, trees and flowers. *Canadian Journal of Mathematics*, 1965.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 1956.

D. Garber and E. Hazan. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *ICML*, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014.

J. Guélat and P. Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 1986.

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

N. He and Z. Harchaoui. Semi-proximal mirror-prox for nonsmooth composite minimization. In *NIPS*, 2015.

J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*. Springer, 1993.

M. Jaggi. *Sparse convex optimization methods for machine learning*. PhD thesis, ETH Zürich, 2011.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, 2013.

A. Juditsky and A. Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 2016.

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

J. Kelner, Y. Lee, L. Orrechia, and A. Sidford. An almost-linear-time algorithm for approximate max flow in undirected graphs, and its multicommodity generalizations. In *SODA*, 2014.

S.-J. Kim, A. Magnani, and S. Boyd. Robust Fisher discriminant analysis. In *NIPS*, 2005.

D. Koller, N. Megiddo, and B. Von Stengel. Fast algorithms for finding randomized strategies in game trees. In *STOC*, 1994.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 1976.

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

S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *ICML*, 2013.

G. Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

T. Larsson and M. Patriksson. A class of gap functions for variational inequalities. *Math. Prog.*, 1994.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational mathematics and mathematical physics*, 1966.

A. F. Martins, N. A. Smith, P. M. Aguiar, and M. A. Figueiredo. Structured sparsity in structured prediction. In *EMNLP*, 2011.

Y. Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 2007.

M. Patriksson. *Nonlinear Programming and Variational Inequality Problems: A Unified Approach*. Springer, 1999.

B. T. Polyak. Existence theorems and convergence of minimizing sequences in extremum problems with restrictions. *Soviet Math. Dokl*, 1966.

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

H. N. Shapiro. Note on a computation method in the theory of games. *Communications on Pure and Applied Mathematics*, 1958.

B. Taskar, C. Guestrin, and D. Koller. Max-margin markov networks. In *NIPS*, 2003.

B. Taskar, S. Lacoste-Julien, and M. I. Jordan. Structured prediction, dual extragradient and Bregman projections. *Journal of Machine Learning Research*, 2006.

J.-P. Vial. Strong and weak convexity of sets and functions. *Mathematics of Operations Research*, 1983.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

J. Von Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton press, 1944.

N. Xiu and J. Zhang. Some recent advances in projection-type methods for variational inequalities. *Journal of Computational and Applied Mathematics*, 2003.

D. L. Zhu and P. Marcotte. Convergence properties of feasible descent methods for solving variational inequalities in banach spaces. *Computational Optimization and Applications*, 1998.