Convergence of the Frank-Wolfe Algorithm for Monotone Variational Inequalities

Matthew Hough[1*]

[1*]Department of Combinatorics and Optimization, University of Waterloo, 200 University Ave. W., Waterloo, N2L 3G1, Ontario, Canada.

Corresponding author(s). E-mail(s): mhough@uwaterloo.ca;

Abstract

We consider the Frank-Wolfe algorithm for solving variational inequalities over compact, convex sets under a monotone $\boldsymbol{C^1}$ operator and vanishing, non-summable step sizes. We introduce a continuous-time interpolation of the discrete iteration and use tools from dynamical systems theory to analyze its asymptotic behavior. This allows us to derive convergence results for the original discrete algorithm. Consequently, every cluster point of the iterates is a solution of the underlying variational inequality, the distance from the iterates to the solution set converges to zero, and the Frank-Wolfe gap vanishes asymptotically. In the strongly monotone case, the solution is unique and the iterates converge to it. In particular, this proves Hammond's conjecture on the convergence of generalized fictitious play. We also discuss rates of convergence and under what assumptions rates can be shown.

Keywords: Frank-Wolfe, conditional gradient method, projection-free, variational inequality problem

1 Introduction

We consider the variational inequality problem

$$\text{Find } x^* \in \mathcal{C} \text{ such that } \langle F(x^*), z - x^* \rangle \geq 0, \quad \forall z \in \mathcal{C}, \tag{VIP}$$

and study the Frank-Wolfe iteration for solving (VIP)

$$x_{k+1} = x_k + \gamma_{k+1}(s_k - x_k), \quad s_k \in \beta(F(x_k)), \quad x_0 \in \mathcal{C}, \tag{FW}$$

where

$$\beta(u) := \operatorname*{argmin}_{s \in \mathcal{C}} \langle u, s \rangle$$

denotes the linear minimization oracle (LMO). In general, we assume that $\mathcal{C} \subseteq \mathbb{R}^n$ is nonempty, compact, and convex, that $F : \mathcal{C} \to \mathbb{R}^n$ is monotone and C^1, and that the stepsizes satisfy $\gamma_k \in (0, 1]$, $\gamma_k \to 0$, and $\sum_{k=1}^{\infty} \gamma_k = \infty$. This algorithm may be viewed as a generalization of the Frank-Wolfe algorithm [1] to the setting of monotone variational inequalities. It also contains generalized fictitious play (GFP) [2, Section 4.3.1] as the special case where $\gamma_k = 1/k$. Brown's classical fictitious play algorithm (FP) [3] is obtained as a further specialization when $\gamma_k = 1/k$, $\mathcal{C} = \Delta_n \times \Delta_m$ is a product of simplices, and

$$F(x, y) = (-Ay, A^\top x),$$

with $A \in \mathbb{R}^{n \times m}$ the payoff matrix. We will denote the set of solutions to (VIP) by $\mathcal{S}$, i.e.

$$\mathcal{S} := \{x \in \mathcal{C} : \langle F(x), z - x \rangle \geq 0, \quad \forall z \in \mathcal{C}\}.$$

To measure convergence to a solution of (VIP), we introduce $V : \mathcal{C} \to \mathbb{R}_+$ the *Frank-Wolfe gap*, defined by

$$V(x) := \max_{s \in \mathcal{C}} \langle F(x), x - s \rangle.$$

We will see that V is nonnegative on $\mathcal{C}$, and vanishes only on $\mathcal{S}$. Moreover, for the special case of (FW) corresponding to FP, V is equivalent to the gap function used to measure convergence to a Nash equilibrium (x^*, y^*) in the analysis of FP. Indeed, V is nonnegative for all $(x, y) \in \Delta_n \times \Delta_m$ and zero iff (x, y) is a Nash equilibrium.

Related work.

In the special case of FP, convergence was proven in [4], which was later extended in [5] to the convergence rate $\mathcal{O}(k^{-1/(m+n-2)})$. Karlin later conjectured that the faster rate $\mathcal{O}(k^{-1/2})$ could be obtained, but this was refuted recently in [6] by the construction of an adversarial tie-breaking strategy achieving a $\Omega(k^{-1/n})$ rate of convergence where A is the $n \times n$ identity matrix. After the discovery of this counterexample, the question of whether FP could still converge at Karlin's conjectured rate under a lexicographic tie-breaking rule remained open. In 2021, it was proven in [7] that for diagonal A, the convergence rate is indeed $\mathcal{O}(k^{-1/2})$. Notably, this class of A includes the identity matrix used in the counterexample [6]. However, the very recent result of Wang [8] showed that the weaker form of Karlin's conjecture, where ties are assumed to be broken lexicographically, was indeed false. Wang constructed a 10×10 matrix for which FP converges at $\Omega(k^{-1/3})$ and no ties occur except at the first step.

Despite the rich literature for FP, no convergence results are known for (FW) without additional assumptions on $\mathcal{C}$, even if F is taken to be strongly monotone instead of just monotone. Hammond conjectured the following for GFP in [2, Section 4.3.1]:

> *If F is strongly monotone and $\mathcal{C}$ is a polytope, then generalized fictitious play will solve* (VIP).

To the best of our knowledge, prior to this work Hammond's conjecture remained open.

Note that when F is only assumed to be monotone and not strongly monotone, (FW) can converge no faster than FP, in general, for the step size $\gamma_k = 1/k$. It is unknown whether this statement can be extended to other choices of γ_k.

Variants of (FW) have drawn interest recently in the fields of optimization and computer science due to the projection-free nature of the iterations. In particular, [9] introduce SP-FW, which considers the case of $\mathcal{C} = \mathcal{X} \times \mathcal{Y}$ and $F(x, y) = (\nabla_x \mathcal{L}(x, y), -\nabla_y \mathcal{L}(x, y))$, where $\mathcal{L}$ is assumed to be smooth and convex-concave. Since the subgradient of a convex function is monotone, it follows that $F(x, y)$ in this case is monotone. The authors are able to obtain some convergence results for specific step sizes, but under strong assumptions such as strong convexity of $\mathcal{X}$ and $\mathcal{Y}$, or uniform strong convex-concavity of $\mathcal{L}$ in addition to $\mathcal{X}$ and $\mathcal{Y}$ being polytopes with a very restrictive pyramidal width[1] bound. Note that the uniform strong convex-concavity of $\mathcal{L}$ implies F is strongly monotone. Another related work is [11], in which the authors consider using iterations similar to (FW) to solve linear programs. They call their algorithm FWLP, which on each iteration performs the following

$$\begin{cases} r_k \in \operatorname{argmin}_{r \in \Delta} \langle c - A^T y_k, r \rangle, \\ x_{k+1} = x_k + \gamma_{k+1}(r_k - x_k), \\ s_k \in \operatorname{argmin}_{s \in \Gamma} \langle b - A x_{k+1}, s \rangle, \\ y_{k+1} = y_k + \gamma_{k+1}(s_k - y_k), \end{cases}$$

where Δ, Γ are polytopes and $\gamma_k = 1/k$. Interestingly, the s_k update of FWLP relies on x_{k+1} instead of x_k. If in the s_k update, x_{k+1} was replaced by x_k, FWLP could be rewritten completely in the framework of (FW) with monotone operator $F(x, y) = (c - A^T y, Ax - b)$. The authors in [11] were unable to prove convergence of FWLP.

Contributions.

We prove that (FW) converges asymptotically to a solution of (VIP) in the sense that the Frank-Wolfe gap $V(x_k) \to 0$. Moreover, if F is additionally assumed to be strongly monotone, we show that $x_k \to x^* \in \mathcal{S}$. Hammond's conjecture applies to the special case where $\gamma_k = 1/k$, hence our result proves Hammond's conjecture. We also provide convergence rates in cases where $\mathcal{C}$ is assumed to be strongly convex.

1.1 Preliminaries

Since $\mathcal{C}$ is compact, its diameter is well-defined. We denote it by $\operatorname{diam}(\mathcal{C}) := \max_{x,y \in \mathcal{C}} \|x - y\|$. Throughout, $\|\cdot\|$ denotes the Euclidean norm, and $B(c, r)$ denotes the closed ball centered at $c \in \mathbb{R}^n$ with radius $r > 0$.

The following are standard definitions that we will use throughout.

Definition 1 An operator $F : \mathcal{C} \to \mathbb{R}^n$ is called monotone if for all $x, y \in \mathcal{C}$,

$$\langle F(x) - F(y), x - y \rangle \geq 0.$$

[1] Pyramidal width was first introduced in [10]

Definition 2 An operator $F : \mathcal{C} \to \mathbb{R}^n$ is called μ-strongly monotone for some $\mu > 0$ if for all $x, y \in \mathcal{C}$,

$$\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2.$$

Definition 3 An operator $F : \mathcal{C} \to \mathbb{R}^n$ is called β-cocoercive for some $\beta > 0$ if for all $x, y \in \mathcal{C}$,

$$\langle F(x) - F(y), x - y \rangle \geq \beta \|F(x) - F(y)\|^2.$$

Definition 4 A nonempty set $C \subseteq \mathbb{R}^n$ is α-strongly convex if for every $x, y \in C$ and every $t \in [0, 1]$,

$$B\left((1-t)x + ty, \frac{\alpha}{2} t(1-t)\|x - y\|^2\right) \subseteq C.$$

Definition 5 Let $I \subseteq \mathbb{R}$ be an interval and let $G : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$. A map $x : I \to \mathbb{R}^n$ is called a solution of the differential inclusion

$$\dot{x}(t) \in G(x(t)),$$

if x is absolutely continuous and

$$\dot{x}(t) \in G(x(t)), \qquad \text{for a.e. } t \in I.$$

All solutions in this paper are understood in this sense.

2 Asymptotic convergence to solutions of the VIP

Throughout this section, we assume that $F : \mathcal{C} \to \mathbb{R}^n$ is monotone and C^1, that $\mathcal{C} \subseteq \mathbb{R}^n$ is nonempty, compact, and convex, and that $\{x_k\}$ is the sequence generated by (FW).

Lemma 1 *The solution set $\mathcal{S}$ is nonempty.*

Proof This follows from the Hartman-Stampacchia theorem [12, Lemma 3.1] and the assumption that F is C^1 and $\mathcal{C}$ is nonempty, convex, and compact. □

Define the following set-valued map on $\mathcal{C}$:

$$\mathcal{F}(x) := \beta(F(x)) - x, \qquad x \in \mathcal{C}.$$

To analyze (FW), we introduce the differential inclusion

$$\dot{x}(t) \in \mathcal{F}(x(t)) = \beta(F(x(t))) - x(t), \qquad x(t) \in \mathcal{C}. \tag{DI-$\mathcal{C}$}$$

This is the continuous-time counterpart of the Frank-Wolfe iteration, since

$$\frac{x_{k+1} - x_k}{\gamma_{k+1}} = s_k - x_k \in \beta(F(x_k)) - x_k = \mathcal{F}(x_k).$$

Thus, if the step size γ_{k+1} is viewed as a small time increment, the discrete update formally approaches the inclusion above. For each $x \in \mathcal{C}$, the set $\mathcal{F}(x) = \beta(F(x)) - x$ consists of all vectors of the form $s - x$ with $s \in \beta(F(x))$. Thus, the differential inclusion (DI-$\mathcal{C}$) means that, at each time t, the rate of change of the trajectory $x(t)$ is given by the vector from the current point $x(t)$ to some current solution $s(t) \in \beta(F(x(t)))$ of the linear minimization oracle.

However, $\mathcal{F}$ is defined only for points in $\mathcal{C}$, while the results of [13] we appeal to are stated for inclusions on all of $\mathbb{R}^n$. We therefore introduce the projected extension (cf. [13, Remark 1.2])

$$\widetilde{\mathcal{F}}(x) := \beta(F(P(x))) - x = \big\{s - x : s \in \beta(F(P(x)))\big\},$$

where $P : \mathbb{R}^n \to \mathcal{C}$ is the Euclidean projection. Since $P(x) = x$ for every $x \in \mathcal{C}$, the extension agrees with $\mathcal{F}$ on $\mathcal{C}$:

$$\widetilde{\mathcal{F}}(x) = \mathcal{F}(x), \quad x \in \mathcal{C}.$$

Accordingly, we will study the following differential inclusion defined on the whole space:

$$\dot{x}(t) \in \widetilde{\mathcal{F}}(x(t)) = \beta(F(P(x(t)))) - x(t), \qquad t \in \mathbb{R}. \tag{DI}$$

Lemma 2 *Recall the gap function $V : \mathcal{C} \to \mathbb{R}_+$ given by*

$$V(x) := \langle F(x), x\rangle - \min_{s\in\mathcal{C}} \langle F(x), s\rangle = \max_{s\in\mathcal{C}} \langle F(x), x - s\rangle.$$

The following hold.

(a) V is Lipschitz continuous on $\mathcal{C}$, $V(x) \geq 0$ for all $x \in \mathcal{C}$, and $V^{-1}(0) = \mathcal{S}$. Moreover, $\mathcal{S}$ is closed.

(b) Let $x : [0, \infty) \to \mathcal{C}$ be any solution of (DI-$\mathcal{C}$). Then $t \mapsto V(x(t))$ satisfies

$$\frac{d}{dt} V(x(t)) \leq -V(x(t)) \quad \text{for a.e. } t \geq 0.$$

Consequently,

$$V(x(t)) \leq e^{-t} V(x(0)) \quad \forall t \geq 0.$$

In particular, if $x(0) \notin \mathcal{S}$ then $V(x(t)) < V(x(0))$ for every $t > 0$, and if $x(0) \in \mathcal{S}$ then $V(x(t)) = 0$ for all $t \geq 0$.

Proof (a) The fact that V is Lipschitz on $\mathcal{C}$ follows from [14, Theorem 1] with the assumption that F is C^1. Nonnegativity is clear from the definition of V and the fact that $x \in \mathcal{C}$. $V(x) = 0$ iff $x \in \mathcal{S}$, because for any $x \in \mathcal{C}$

$$\max_{s\in\mathcal{C}} \langle F(x), x - s\rangle = 0 \iff \langle F(x), x - z\rangle \leq 0 \quad \forall z \in \mathcal{C} \iff \langle F(x), z - x\rangle \geq 0 \quad \forall z \in \mathcal{C}$$

To show $\mathcal{S}$ is closed, we note that since $\mathcal{S} = V^{-1}(\{0\})$ and V is Lipschitz continuous, $\mathcal{S}$ is the preimage of a closed set under a continuous map and therefore must be closed in $\mathcal{C}$. Because $\mathcal{C}$ is closed in $\mathbb{R}^n$, $\mathcal{S}$ must be closed in $\mathbb{R}^n$.

(b) The differential inequality comes from the proof of [14, Theorem 1]. Let x be a solution of (DI-$\mathcal{C}$). To obtain the bound on $V(x(t))$ for all $t \geq 0$, fix some $T > 0$. Recall from (a) that V is Lipschitz continuous, and $t \mapsto x(t)$ is absolutely continuous by Definition 5. Hence, their composition $V(x(t))$ is absolutely continuous on $[0, T]$. We proceed by showing a Grönwall-type inequality inspired by the proof of [15, Theorem 1.12]. Let $v(t) := V(x(t))$ and $u(t) := v(t)e^t$, then $u(t)$ is absolutely continuous on $[0, T]$, since $v(t)$ is. Moreover, since $v'(t) \leq -v(t)$ for a.e. $t \geq 0$,

$$u'(t) = v'(t)e^t + v(t)e^t \leq -v(t)e^t + v(t)e^t = 0,$$

for a.e. $t \in [0, T]$, which along with the absolute continuity of $u(t)$, implies that for any $0 \leq r \leq t \leq T$,

$$u(t) - u(r) = \int_r^t u'(s)\, ds \leq 0.$$

Hence, $u(t)$ is nonincreasing on $[0, T]$, and in particular $u(t) \leq u(0)$ for all $t \in [0, T]$. This can be rewritten as

$$v(t)e^t \leq v(0)e^0 = v(0), \qquad \forall t \in [0, T],$$

which after rearranging gives

$$V(x(t)) \leq e^{-t}V(x(0)) \quad \forall t \in [0, T].$$

Since $T > 0$ was arbitrary, the above inequality holds for all $t \geq 0$. □

We now construct an interpolating curve for the algorithm (FW). Set $\tau_0 := 0$ and

$$\tau_{k+1} := \tau_k + \gamma_{k+1} = \sum_{i=1}^{k+1} \gamma_i.$$

Since $\sum_{k=1}^{\infty} \gamma_k = \infty$, it follows that $\tau_k \to \infty$ as $k \to \infty$. Suppose $\{x_k\}$ is the sequence of iterates generated by (FW) under the assumptions of Section 1. Define the interpolating curve $w : [0, \infty) \to \mathcal{C}$ so that $w(\tau_k) = x_k$ for each k and it is linear on the time segments $t \in [\tau_k, \tau_{k+1}]$:

$$w(t) := (1 - \theta(t))x_k + \theta(t)x_{k+1}, \quad t \in [\tau_k, \tau_{k+1}],$$

where

$$\theta(t) = \frac{t - \tau_k}{\gamma_{k+1}} = \frac{t - \tau_k}{\tau_{k+1} - \tau_k}, \quad t \in [\tau_k, \tau_{k+1}].$$

Clearly, if $t = \tau_k$ $w(t) = x_k$ and if $t = \tau_{k+1}$, $w(t) = x_{k+1}$. Also,

$$0 = \frac{\tau_k - \tau_k}{\tau_{k+1} - \tau_k} \leq \frac{t - \tau_k}{\tau_{k+1} - \tau_k} \leq \frac{\tau_{k+1} - \tau_k}{\tau_{k+1} - \tau_k} = 1,$$

so $\theta(t) \in [0, 1]$. Therefore, for any t, $w(t)$ is a convex combination of points in $\mathcal{C}$, hence $w(t) \in \mathcal{C}$ for all t. On each open interval (τ_k, τ_{k+1}), w is linear and is written in full as

$$w(t) = t\frac{x_{k+1} - x_k}{\gamma_{k+1}} + \left(1 + \frac{\tau_k}{\gamma_{k+1}}\right) x_k - \frac{\tau_k}{\gamma_{k+1}} x_{k+1},$$

with derivative

$$\dot{w}(t) = \frac{x_{k+1} - x_k}{\gamma_{k+1}} = \frac{x_{k+1} - x_k}{\tau_{k+1} - \tau_k}.$$

Since $x_{k+1} - x_k = \gamma_{k+1}(s_k - x_k)$, we have for $t \in (\tau_k, \tau_{k+1})$

$$\dot{w}(t) = s_k - x_k \in \beta(F(x_k)) - x_k = \mathcal{F}(x_k),$$

where the derivative exists for all $t \geq 0$, except the breakpoints $t = \tau_k$ for all k. Hence, $\dot{w}(t) \in \mathcal{F}(x_k)$ a.e. on $t \geq 0$. The next few results show that this fits into the framework of [13].

Lemma 3 *The projected extension $\widetilde{\mathcal{F}}$ satisfies the following:*

(a) $\widetilde{\mathcal{F}}(x)$ is nonempty, compact, and convex for all $x \in \mathbb{R}^n$.

(b) $\widetilde{\mathcal{F}}$ has closed graph in $\mathbb{R}^n \times \mathbb{R}^n$.

(c) There exists $c > 0$ such that $\sup_{z \in \widetilde{\mathcal{F}}(x)} \|z\| \leq c(1 + \|x\|)$ for all $x \in \mathbb{R}^n$.

Proof Since $\mathcal{C}$ is compact and $s \mapsto \langle u, s \rangle$ is continuous and affine, $\beta(u)$ is nonempty and compact for every u. It is also convex because it is the set of minimizers of a linear functional over a convex set. Also, since $\mathcal{C}$ is compact and convex, P is single-valued and nonexpansive, so continuous.

(a) Fix $x \in \mathbb{R}^n$. Then $\beta(F(P(x)))$ is nonempty, compact, convex, and contained in $\mathcal{C}$. Hence

$$\widetilde{\mathcal{F}}(x) = \beta(F(P(x))) - x$$

is a translation of a nonempty, compact, and convex set, so it is itself nonempty, compact, and convex.

(b) Let $x_k \to x$ in $\mathbb{R}^n$ and $y_k \to y$ in $\mathbb{R}^n$ with $y_k \in \widetilde{\mathcal{F}}(x_k)$. Then there exist $b_k \in \beta(F(P(x_k)))$ such that

$$y_k = b_k - x_k.$$

Since $b_k \in \mathcal{C}$ and $\mathcal{C}$ is compact, we may pass to a subsequence such that $b_k \to b \in \mathcal{C}$. Because P and F are continuous, $F(P(x_k)) \to F(P(x))$.

Now fix any $s \in \mathcal{C}$. Optimality of b_k gives

$$\langle F(P(x_k)), b_k \rangle \leq \langle F(P(x_k)), s \rangle, \quad \forall k.$$

Taking $k \to \infty$ yields

$$\langle F(P(x)), b \rangle \leq \langle F(P(x)), s \rangle, \quad \forall s \in \mathcal{C},$$

so $b \in \beta(F(P(x)))$. Finally, $y_k = b_k - x_k \to b - x =: y$, hence $y \in \beta(F(P(x))) - x = \widetilde{\mathcal{F}}(x)$. Therefore $\mathrm{gra}(\widetilde{\mathcal{F}})$ is closed.

(c) Let $M := \max_{u \in \mathcal{C}} \|u\|$ and note $M < \infty$ because $\mathcal{C}$ is compact. For any $z \in \widetilde{\mathcal{F}}(x)$ we can write $z = b - x$ with $b \in \mathcal{C}$, hence $\|z\| \leq \|b\| + \|x\| \leq M + \|x\|$. Thus $\sup_{z \in \widetilde{\mathcal{F}}(x)} \|z\| \leq M + \|x\| \leq c(1 + \|x\|)$ with $c := \max(M, 1)$. □

Remark 1 By Lemma 3, the differential inclusion (DI) admits at least one solution $x : \mathbb{R} \to \mathbb{R}^n$ through every initial condition $x(0) = x_0 \in \mathbb{R}^n$ (see [13, Section 1.2]).

Lemma 4 *The interpolated curve w is Lipschitz continuous.*

Proof Recall, on $[\tau_k, \tau_{k+1}]$

$$w(t) = (1 - \theta(t))x_k + \theta(t)x_{k+1},$$

so for any $t, t' \in [\tau_k, \tau_{k+1}]$,

$$w(t) - w(t') = \theta(t')x_k - \theta(t')x_{k+1} - \theta(t)x_k + \theta(t)x_{k+1} = \big(\theta(t) - \theta(t')\big)\,(x_{k+1} - x_k).$$

Hence,

$$\|w(t) - w(t')\| = \|\big(\theta(t) - \theta(t')\big)\,(x_{k+1} - x_k)\| \leq |\theta(t) - \theta(t')|\|x_{k+1} - x_k\|.$$

But $x_{k+1} - x_k = \gamma_{k+1}(s_k - x_k)$ and $x_k, s_k \in \mathcal{C}$, so $\|s_k - x_k\| \leq \operatorname{diam}(\mathcal{C})$, and $\operatorname{diam}(\mathcal{C}) < \infty$ by compactness. Also,

$$\theta(t) - \theta(t') = \frac{t - \tau_k}{\gamma_{k+1}} - \frac{t' - \tau_k}{\gamma_{k+1}} = \frac{t - t'}{\gamma_{k+1}}.$$

Thus,

$$\|w(t) - w(t')\| \leq \operatorname{diam}(\mathcal{C})|t - t'|.$$

Now consider the case of arbitrary $a, b \in \mathbb{R}_+$. We may assume wlog that $a < b$. If a, b lie in the same interval, we are in the case of the above, so suppose $a \in [\tau_m, \tau_{m+1}]$ and $b \in [\tau_n, \tau_{n+1}]$ where $m \leq n$. We may write $w(b) - w(a)$ using a telescoping sum:

$$w(b) - w(a) = (w(b) - w(\tau_n)) + (w(\tau_{m+1}) - w(a)) + \sum_{k=m+1}^{n-1} (w(\tau_{k+1}) - w(\tau_k))\,.$$

By the triangle inequality and the above case when the two breakpoints are in the same interval, we have

$$\begin{aligned}
\|w(b) - w(a)\| &\leq \|w(b) - w(\tau_n)\| + \|w(\tau_{m+1}) - w(a)\| + \sum_{k=m+1}^{n-1} \|w(\tau_{k+1}) - w(\tau_k)\| \\
&\leq \operatorname{diam}(\mathcal{C})(b - \tau_n) + \operatorname{diam}(\mathcal{C})(\tau_{m+1} - a) + \operatorname{diam}(\mathcal{C}) \sum_{k=m+1}^{n-1} (\tau_{k+1} - \tau_k) \\
&= \operatorname{diam}(\mathcal{C})(b - \tau_n) + \operatorname{diam}(\mathcal{C})(\tau_{m+1} - a) + \operatorname{diam}(\mathcal{C})(\tau_n - \tau_{m+1}) \\
&= \operatorname{diam}(\mathcal{C})(b - a).
\end{aligned}$$

It follows that w is Lipschitz continuous. □

The interpolated curve w is not, in general, an exact solution of the differential inclusion (DI). Nevertheless, because it is obtained by linearly interpolating the Frank-Wolfe iterates, it provides a natural approximation of the continuous-time dynamics, up to a discretization error that vanishes asymptotically as $\gamma_k \to 0$. To formalize this idea, we use the notion of a *perturbed solution* [13, Definition II]. Roughly speaking, a perturbed solution is an absolutely continuous curve that satisfies the differential inclusion up to errors that become negligible in the long run. This notion is useful because, even though a perturbed solution is only approximate, we will see that its asymptotic behavior can still be analyzed through the corresponding differential inclusion.

Definition 6 Let $y : [0,\infty) \to \mathbb{R}^n$ be continuous. We say that y is a perturbed solution of the differential inclusion (DI) if:

(i) y is absolutely continuous;

(ii) there exist a locally integrable function $t \mapsto U(t)$ and $\delta : [0,\infty) \to \mathbb{R}_+$ with $\delta(t) \to 0$ as $t \to \infty$ such that:

(a) for all $T > 0$,
$$\lim_{t\to\infty} \sup_{0\le v\le T} \left\| \int_t^{t+v} U(s)\, ds \right\| = 0;$$

(b) $\dot{y}(t) - U(t) \in \widetilde{\mathcal{F}}^{\delta(t)}(y(t))$ for almost every $t > 0$, where
$$\widetilde{\mathcal{F}}^{\delta}(x) := \big\{u \in \mathbb{R}^n : \exists z \in \mathbb{R}^n \text{ with } \|z - x\| < \delta \text{ and } \inf_{f\in\widetilde{\mathcal{F}}(z)} \|u - f\| < \delta\big\}.$$

Condition (ii)(a) says that the cumulative effect of the additive perturbation U on every bounded time window becomes negligible as $t \to \infty$, while condition (ii)(b) permits a vanishing approximation error both in the point at which the map is evaluated and in the inclusion itself. In our setting, the interpolated curve w satisfies this definition with $U \equiv 0$. Thus the only discrepancy from being an exact solution of (DI) is that, on each interval (τ_k, τ_{k+1}), one has

$$\dot{w}(t) = s_k - x_k \in \widetilde{\mathcal{F}}(x_k),$$

whereas an exact solution would require

$$\dot{w}(t) \in \widetilde{\mathcal{F}}(w(t)).$$

Since $w(t)$ remains close to x_k on this interval and the distance tends to zero as $k \to \infty$, this discrepancy is asymptotically negligible.

Lemma 5 *The interpolated curve w is a perturbed solution of the differential inclusion (DI).*

Proof By definition of w, it is a continuous function from $[0,\infty)$ to $\mathcal{C} \subseteq \mathbb{R}^n$. Moreover, since w is Lipschitz continuous from Lemma 4, it is absolutely continuous. It remains to satisfy the conditions (ii) in Definition 6. In Definition 6, take $U \equiv 0$, then (ii)(a) must be satisfied. To satisfy (ii)(b), we need to show $\dot{w}(t) \in \widetilde{\mathcal{F}}^{\delta(t)}(w(t))$ for almost every $t > 0$. Define $\delta(t) = \gamma_{k+1}(1 + \mathrm{diam}(\mathcal{C}))$ for each $t \in [\tau_k, \tau_{k+1})$, then clearly $\delta : [0,\infty) \to \mathbb{R}$ and $\delta(t) \to 0$. Recall
$$\widetilde{\mathcal{F}}^{\delta}(x) = \{u \in \mathbb{R}^n : \exists z \text{ with } \|z - x\| < \delta, \inf_{f\in\widetilde{\mathcal{F}}(z)} \|u - f\| < \delta\}.$$
We have already shown that $\dot{w}(t) \in \mathcal{F}(x_k) = \widetilde{\mathcal{F}}(x_k)$, where the equality comes from $x_k \in \mathcal{C}$. Thus, if we take $z = x_k$ in the definition of $\widetilde{\mathcal{F}}^{\delta(t)}(w(t))$, we have
$$\inf_{f\in\widetilde{\mathcal{F}}(x_k)} \|\dot{w}(t) - f\| = 0 < \delta(t).$$

Moreover, for $t \in (\tau_k, \tau_{k+1})$,

$$w(t) - x_k = \theta(t)(x_{k+1} - x_k),$$

and $\theta(t) \in (0,1)$, so

$$\|w(t) - x_k\| = \theta(t)\|x_{k+1} - x_k\| \le \|x_{k+1} - x_k\| = \gamma_{k+1}\|s_k - x_k\| \le \gamma_{k+1}\,\mathrm{diam}(\mathcal{C}) < \delta(t).$$

We have shown that $\dot{w}(t) \in \widetilde{\mathcal{F}}^{\delta(t)}(w(t))$ for all t in the open segments (τ_k, τ_{k+1}). Since the endpoints form a countable set, we have shown that $\dot{w}(t) \in \widetilde{\mathcal{F}}^{\delta(t)}(w(t))$ for almost every $t > 0$, hence we have satisfied condition (ii)(b). □

Recall the definition of the limit set of w,

$$L(w) := \bigcap_{t \ge 0} \overline{\{w(s) : s \ge t\}}.$$

Our goal will be to show that $L(w) \subseteq \mathcal{S}$, and then conclude that every cluster point of $\{x_k\}$ belongs to $\mathcal{S}$. To do so, we use the notion of invariance of a set with respect to the differential inclusion (DI).

Definition 7 A set $A \subseteq \mathbb{R}^n$ is said to be invariant if for all $z \in A$ there exists a solution x to (DI) with $x(0) = z$ where $x(\mathbb{R}) \subseteq A$.

Theorem 6 *$L(w) \subseteq \mathcal{S}$.*

Proof Since $w([0,\infty)) \subseteq \mathcal{C}$ and $\mathcal{C}$ is compact, we have $L(w) \subseteq \mathcal{C}$ and $L(w)$ is nonempty.

By Lemma 5, w is a bounded perturbed solution of the inclusion $\dot{x}(t) \in \widetilde{\mathcal{F}}(x(t))$, hence by [13, Theorem 3.6] the set $L(w)$ is internally chain transitive [13, Definition VI] for the dynamical system generated by the differential inclusion (DI). Therefore, by [13, Lemma 3.5], $L(w)$ is invariant. That is, for every $z \in L(w)$ there exists a solution $x : \mathbb{R} \to \mathbb{R}^n$ of $\dot{x}(t) \in \widetilde{\mathcal{F}}(x(t))$ with $x(0) = z$ and $x(\mathbb{R}) \subseteq L(w) \subseteq \mathcal{C}$. Fix $z \in L(w)$ and let x be such a solution in $L(w)$ with $x(0) = z$. Because $x(\mathbb{R}) \subseteq \mathcal{C}$, we have $\widetilde{\mathcal{F}}(x(t)) = \mathcal{F}(x(t))$ for all $t \in \mathbb{R}$, hence x is also a solution of the differential inclusion (DI-$\mathcal{C}$) for a.e. $t \in \mathbb{R}$.

Let $M := \max_{u \in \mathcal{C}} V(u) < \infty$, where the finiteness comes from the compactness of $\mathcal{C}$ and continuity of V (from V Lipschitz, cf. Lemma 2(a)). For any $T > 0$, define $y : [0,\infty) \to \mathcal{C}$ by $y(t) := x(t - T)$. Since x is a solution of (DI-$\mathcal{C}$) on $\mathbb{R}$, it follows that y is a solution of (DI-$\mathcal{C}$) on $[0,\infty)$. Hence, we may apply Lemma 2(b) to y:

$$V(y(t)) \le e^{-t} V(y(0)), \qquad \forall t \ge 0,$$

which when taking $t = T$ is equivalent to

$$V(x(0)) \le e^{-T} V(x(-T)).$$

Since $x(-T) \in \mathcal{C}$, we must have $V(x(-T)) \le M$, so we may write

$$V(z) = V(x(0)) \le e^{-T} M, \qquad \forall\ T > 0.$$

Letting $T \to \infty$ gives $V(z) = 0$. By Lemma 2(a), $V^{-1}(0) = \mathcal{S}$, so $z \in \mathcal{S}$. Thus $L(w) \subseteq \mathcal{S}$. □

Corollary 7 *Every cluster point of the sequence $\{x_k\}$ belongs to $\mathcal{S}$. In particular, as $k \to \infty$,*

$$\operatorname{dist}(x_k, \mathcal{S}) \to 0$$

Proof Let $\bar{x}$ be a cluster point of $\{x_k\}$. Then there exists a subsequence $\{x_{k_j}\}$ such that $x_{k_j} \to \bar{x}$ as $j \to \infty$. Since $x_k = w(\tau_k)$ for all k and $\tau_k \to \infty$, we also have $\tau_{k_j} \to \infty$ and $w(\tau_{k_j}) = x_{k_j} \to \bar{x}$. We claim that $\bar{x} \in L(w)$. Indeed, fix any $t \geq 0$. Since $\tau_{k_j} \to \infty$, there exists j_0 such that $\tau_{k_j} \geq t$ for all $j \geq j_0$. Hence

$$w(\tau_{k_j}) \in \{w(s) : s \geq t\}, \qquad \forall j \geq j_0.$$

Taking the limit and using $w(\tau_{k_j}) \to \bar{x}$, we obtain

$$\bar{x} \in \overline{\{w(s) : s \geq t\}}.$$

Since $t \geq 0$ was arbitrary, it follows that

$$\bar{x} \in \bigcap_{t \geq 0} \overline{\{w(s) : s \geq t\}} = L(w).$$

By Theorem 6, $L(w) \subseteq \mathcal{S}$, so $\bar{x} \in \mathcal{S}$. This proves the first claim.

For the second claim, suppose for contradiction that $\operatorname{dist}(x_k, \mathcal{S}) \not\to 0$. Then there exist $\epsilon > 0$ and a subsequence $\{x_{k_j}\}$ such that

$$\operatorname{dist}(x_{k_j}, \mathcal{S}) \geq \epsilon, \qquad \forall j.$$

Since $\{x_k\} \subseteq \mathcal{C}$ and $\mathcal{C}$ is compact, by passing to a further subsequence if necessary, we have that $x_{k_j} \to \bar{x}$ for some $\bar{x} \in \mathcal{C}$. By the above, $\bar{x} \in \mathcal{S}$. Since $\mathcal{S}$ is closed by Lemma 2(a), $x \mapsto \operatorname{dist}(x, \mathcal{S})$ is continuous, and therefore

$$\operatorname{dist}(x_{k_j}, \mathcal{S}) \to \operatorname{dist}(\bar{x}, \mathcal{S}) = 0,$$

which contradicts $\operatorname{dist}(x_{k_j}, \mathcal{S}) \geq \epsilon$ for all j. Hence $\operatorname{dist}(x_k, \mathcal{S}) \to 0$ as $k \to \infty$. □

Corollary 8 *The Frank-Wolfe gap along the iterates $\{x_k\}$ converges to zero:*

$$V(x_k) \to 0 \text{ as } k \to \infty.$$

Proof By Lemma 2(a), V is Lipschitz continuous on $\mathcal{C}$, so there exists some $L > 0$ such that

$$|V(x) - V(y)| \leq L\|x - y\|, \qquad \forall x, y \in \mathcal{C}.$$

Fix $x \in \mathcal{C}$. For any $z \in \mathcal{S}$, Lemma 2(a) gives $V(z) = 0$, so

$$V(x) = |V(x) - V(z)| \leq L\|x - z\|.$$

Taking the infimum of the above over $z \in \mathcal{S}$ yields

$$V(x) \leq L \operatorname{dist}(x, \mathcal{S}).$$

Applying this with $x = x_k$ and using $\operatorname{dist}(x_k, \mathcal{S}) \to 0$ gives $V(x_k) \to 0$. □

2.1 The strongly monotone case

Suppose in addition that F is μ-strongly monotone for some $\mu > 0$, i.e. for all $x, y \in \mathcal{C}$

$$\langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2.$$

Under this additional assumption, the solution set is a singleton.

Lemma 9 *The solution set $\mathcal{S}$ is a singleton.*

Proof Suppose $x, y \in \mathcal{S}$ are distinct. Then,

$$\langle F(x), y - x \rangle \geq 0, \quad \langle F(y), x - y \rangle \geq 0.$$

Adding the two gives

$$\langle F(x) - F(y), x - y \rangle \leq 0.$$

We may combine this with the definition of strong monotonicity to obtain

$$0 \geq \langle F(x) - F(y), x - y \rangle \geq \mu \|x - y\|^2,$$

which implies $x = y$. So $\mathcal{S}$ must contain only one element. □

Theorem 10 *Let x^* be the unique element in $\mathcal{S}$. Then $x_k \to x^*$.*

Proof By Corollary 7, $\operatorname{dist}(x_k, \mathcal{S}) \to 0$. Since $\mathcal{S} = \{x^*\}$, by Lemma 9, we have

$$\|x_k - x^*\| = \operatorname{dist}(x_k, \{x^*\}) = \operatorname{dist}(x_k, \mathcal{S}) \to 0.$$

Hence $x_k \to x^*$. □

3 Rates of convergence

3.1 On obtaining rates from the continuous-time analysis

By Lemma 2(b), every solution x of (DI-$\mathcal{C}$) satisfies

$$V(x(t)) \leq e^{-t} V(x(0)), \qquad t \geq 0.$$

Thus the Frank-Wolfe gap converges to zero at an exponential rate along trajectories of the continuous-time dynamics. It is not clear, however, how to transfer this rate to the discrete iterates. The difficulty is that the interpolated curve w associated with the sequence $\{x_k\}$ is not, in general, a solution of the differential inclusion, but only a perturbed solution. On each interval (τ_k, τ_{k+1}),

$$\dot{w}(t) = s_k - x_k \in \widetilde{\mathcal{F}}(x_k),$$

whereas an exact solution would satisfy

$$\dot{w}(t) \in \widetilde{\mathcal{F}}(w(t)).$$

To deduce a discrete convergence rate from the continuous-time analysis, one would need to control the error introduced by replacing $\widetilde{\mathcal{F}}(w(t))$ with $\widetilde{\mathcal{F}}(x_k)$.

This is the main obstruction in the setting where $\mathcal{C}$ is a general compact, convex set. Without additional geometric assumptions on $\mathcal{C}$, the LMO need not depend continuously on its argument, so small changes in $F(x)$ may produce large changes in the selected minimizer. Hence, even though $\|w(t) - x_k\| = \mathcal{O}(\gamma_{k+1})$ for $t \in [\tau_k, \tau_{k+1}]$, this does not imply that one can choose minimizers in $\beta(F(w(t)))$ and $\beta(F(x_k))$ that are close. For this reason, the continuous-time rate does not directly yield a rate for the discrete algorithm.

In the next subsection, we show that strong convexity of $\mathcal{C}$ gives sufficient regularity of the LMO to control this error and derive rates for the discrete iterates.

3.2 Rates over strongly convex sets

When $\mathcal{C}$ is assumed to be strongly convex in addition to being nonempty and compact, the linear minimization oracle becomes Lipschitz over the unit sphere.

Lemma 11 *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be nonempty, compact, and α-strongly convex for some $\alpha > 0$. Recall the definition of the linear minimization oracle*

$$\beta(u) := \operatorname*{argmin}_{s \in \mathcal{C}} \langle u, s \rangle, \quad u \in \mathbb{R}^n.$$

Then,

(i) For every $u \in \mathbb{R}^n \setminus \{0\}$, the minimizer $\beta(u)$ is unique.

(ii) The function $\beta(u)$ is $(1/\alpha)$-Lipschitz on the unit sphere; that is, for all unit vectors $u, v \in \mathbb{R}^n$

$$\|\beta(u) - \beta(v)\| \leq \frac{1}{\alpha}\|u - v\|. \tag{1}$$

Proof First, since $\mathcal{C}$ is nonempty and compact, $\beta(u)$ must exist. Now suppose $u \neq 0$ and there exist distinct $s_1, s_2 \in \mathcal{C}$ with $\langle u, s_1 \rangle = \langle u, s_2 \rangle = k \in \mathbb{R}$. Define $m = (s_1 + s_2)/2$. By α-strong convexity of $\mathcal{C}$, for $t = 1/2$ we have

$$B(m, \rho) \subseteq \mathcal{C},$$

where $\rho = \frac{\alpha}{8}\|s_1 - s_2\|^2$, which is positive by the assumption that $s_1 \neq s_2$. Take $w = u/\|u\|$ and $z = m - \rho w$. We must have $z \in \mathcal{C}$, while

$$\begin{aligned} \langle u, z \rangle &= \langle u, m \rangle - \rho \langle u, w \rangle \\ &= k - \rho \|u\| \\ &< k. \end{aligned}$$

But we have now found a $z \in \mathcal{C}$ which contradicts the minimality of the claimed minimizers. It can only be that $\rho = 0$, which implies $s_1 = s_2$, completing the proof of (i).

To prove (ii), fix unit vectors $w_1, w_2 \in \mathbb{R}^n$ and let $s_1 := \beta(w_1), s_2 := \beta(w_2), d := \|s_1 - s_2\|$. If $d = 0$ the proof is trivial, so suppose $d > 0$. For any $t \in [0, 1]$, strong convexity of $\mathcal{C}$ implies $B(m_t, r_t) \subseteq \mathcal{C}$, where $m_t = (1-t)s_1 + ts_2$ and $r_t = \frac{\alpha}{2}t(1-t)d^2$. Take $z_t := m_t - r_t w_1$, which must lie in $\mathcal{C}$. By optimality of s_1 for $\min_{s \in \mathcal{C}} \langle w_1, s \rangle$,

$$\langle w_1, s_1 \rangle \leq \langle w_1, z_t \rangle = (1-t)\langle w_1, s_1 \rangle + t\langle w_1, s_2 \rangle - r_t,$$

which after rearranging gives

$$t\langle w_1, s_2 - s_1\rangle \geq r_t = \frac{\alpha}{2}t(1-t)d^2.$$

So for all $t \in (0,1)$,

$$\langle w_1, s_2 - s_1\rangle \geq \frac{\alpha}{2}(1-t)d^2.$$

Taking the limit as $t \downarrow 0$,

$$\langle w_1, s_2 - s_1\rangle \geq \frac{\alpha}{2}d^2. \tag{2}$$

Swapping w_1 and w_2 and s_1 and s_2 in the above working gives the symmetric form

$$\langle w_2, s_1 - s_2\rangle \geq \frac{\alpha}{2}d^2. \tag{3}$$

Adding (2) and (3) we get

$$\langle w_1 - w_2, s_2 - s_1\rangle \geq \alpha d^2,$$

and by Cauchy-Schwarz

$$\alpha d^2 \leq \|w_1 - w_2\| d,$$

which gives the desired inequality

$$\|\beta(w_1) - \beta(w_2)\| \leq \frac{1}{\alpha}\|w_1 - w_2\|.$$

□

Corollary 12 *Suppose that $\mathcal{C}$ is nonempty, compact, and α-strongly convex for some $\alpha > 0$. Then for every $u, v \in \mathbb{R}^n \setminus \{0\}$, and every $s(u) \in \beta(u)$, $s(v) \in \beta(v)$, one has*

$$\|s(u) - s(v)\| \leq \frac{2}{\alpha \min(\|u\|, \|v\|)}\|u - v\|.$$

Proof Let $u, v \neq 0$. Since minimizing $\langle u, \cdot\rangle$ is the same as minimizing $\langle u/\|u\|, \cdot\rangle$, we may write

$$s(u) = s\left(\frac{u}{\|u\|}\right), \quad s(v) = s\left(\frac{v}{\|v\|}\right).$$

Hence

$$\|s(u) - s(v)\| \leq \frac{1}{\alpha}\left\|\frac{u}{\|u\|} - \frac{v}{\|v\|}\right\|.$$

Using the bound

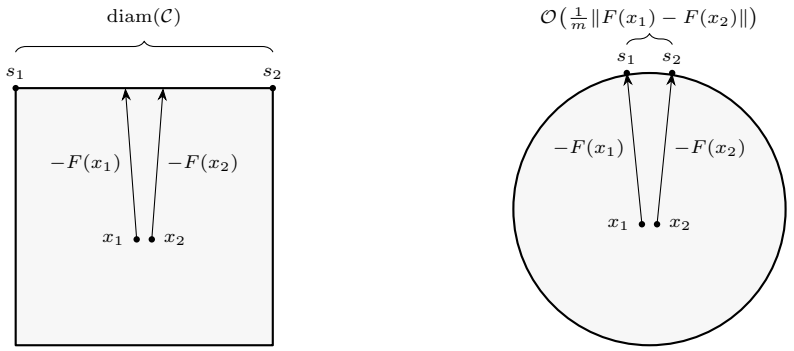
$$\left\|\frac{u}{\|u\|} - \frac{v}{\|v\|}\right\| \leq \frac{2}{\min(\|u\|, \|v\|)}\|u - v\|,$$

the result follows. □

In their PhD thesis [2], Hammond proved convergence of generalized fictitious play to a solution of (VIP) under the assumption that F is C^1 and monotone, $\mathcal{C}$ is compact and strongly convex, and additionally no point $x \in \mathcal{C}$ satisfies $F(x) = 0$. No rate of convergence is proven in [2] for this case. Hammond notes that the assumptions of strong convexity and $F(x) \neq 0$ for all $x \in \mathcal{C}$ are too restrictive. In Section 2 we showed that neither of these assumptions are necessary for convergence. It remains unclear how to prove a rate of convergence without the assumption that $\mathcal{C}$ is strongly convex. However, the assumption that F does not vanish on $\mathcal{C}$ is common in the literature when proving convergence rates over strongly convex sets. For example, in [9], the authors

assume $\min\left(\|\nabla_x \mathcal{L}(z)\|_{\mathcal{X}^*}, \|\nabla_y \mathcal{L}(z)\|_{\mathcal{Y}^*}\right) \geq \delta > 0$ for all $z \in \mathcal{X} \times \mathcal{Y}$ in order to obtain global Lipschitz continuity of the LMO over $\mathcal{X} \times \mathcal{Y}$. The resulting linear convergence guarantee of [9, Theorem 4] depends on δ: as δ decreases, the Lipschitz modulus worsens and the contraction factor in the rate deteriorates. Hence, although the rate remains linear in form, the associated complexity bound can become arbitrarily poor when δ is small.

In the next two theorems, we prove rates of convergence for (FW) over strongly convex sets without assuming $F(x)$ does not vanish over $\mathcal{C}$. The analysis hinges on the observation that when the function value is smaller than some quantity, we can obtain a bound on the Frank-Wolfe gap in terms of this quantity. This analysis relies on showing $\|s_{k+1} - s_k\|$ is decreasing at the rate $\mathcal{O}(\gamma_{k+1})$, which we get from the Lipschitz continuity of the LMO and the operator F. In settings where $\mathcal{C}$ is not uniformly smooth, for example when $\mathcal{C}$ is a polytope, $\|s_{k+1} - s_k\|$ is not necessarily decreasing and thus another proof technique is necessary.



(a) $\mathcal{C}$ is a box, where the distance between any two vertices is $\operatorname{diam}(\mathcal{C})$.

$\mathcal{C}$ is a ball, which is an example of a strongly-convex set.

Fig. 1 Geometry of the linear minimization oracle. On a polytope, crossing the outward normal direction of a face can cause an $\mathcal{O}(\operatorname{diam}(\mathcal{C}))$ jump in the LMO. For a strongly convex set, the LMO is Lipschitz on the unit sphere. With F Lipschitz, this allows a bound in terms of the difference between x_1 and x_2. Here $m := \min\{\|F(x_1)\|, \|F(x_2)\|\}$.

Theorem 13 *In addition to the standing assumptions of Section 2, suppose that $\mathcal{C}$ is α-strongly convex. Then F is Lipschitz on $\mathcal{C}$ with constant $L > 0$, and for all $k \geq 1$,*

$$V(x_{k+1}) \leq \max\left\{(1 - \gamma_{k+1})V(x_k) + B\gamma_{k+1}^{3/2}, C\sqrt{\gamma_{k+1}}\right\},$$

where $B = \frac{2L^2 \operatorname{diam}(\mathcal{C})^2}{\alpha}$ and $C = \operatorname{diam}(\mathcal{C})(1 + L \operatorname{diam}(\mathcal{C}))$. In particular, if $\gamma_k = 1/k$, then

$$V(x_k) \leq \frac{A}{\sqrt{k}},$$

where $A = \max\{V(x_1), 2B, C\}$.

Proof Since F is C^1 on the compact set $\mathcal{C}$, it is Lipschitz on $\mathcal{C}$. We denote its Lipschitz constant by $L > 0$. Observe that $V(x_k) = \langle F(x_k), x_k - s_k \rangle$.

Since $x_{k+1} = x_k + \gamma_{k+1}(s_k - x_k)$, we have

$$x_{k+1} - s_k = (1 - \gamma_{k+1})(x_k - s_k).$$

Therefore,

$$\begin{aligned} V(x_{k+1}) &= \langle F(x_{k+1}), x_{k+1} - s_{k+1} \rangle \\ &= \langle F(x_{k+1}), x_{k+1} - s_k \rangle + \langle F(x_{k+1}), s_k - s_{k+1} \rangle \\ &= (1 - \gamma_{k+1}) \langle F(x_{k+1}), x_k - s_k \rangle + \langle F(x_{k+1}), s_k - s_{k+1} \rangle \\ &= (1 - \gamma_{k+1}) V(x_k) + (1 - \gamma_{k+1}) \langle F(x_{k+1}) - F(x_k), x_k - s_k \rangle + \langle F(x_{k+1}), s_k - s_{k+1} \rangle. \end{aligned}$$

Since $x_{k+1} - x_k = \gamma_{k+1}(s_k - x_k)$ and F is monotone,

$$\langle F(x_{k+1}) - F(x_k), s_k - x_k \rangle = \frac{1}{\gamma_{k+1}} \langle F(x_{k+1}) - F(x_k), x_{k+1} - x_k \rangle \geq 0.$$

Hence

$$(1 - \gamma_{k+1}) \langle F(x_{k+1}) - F(x_k), x_k - s_k \rangle \leq 0.$$

Also, by optimality of $s_k \in \beta(F(x_k))$,

$$\langle F(x_k), s_k - s_{k+1} \rangle \leq 0,$$

so

$$\begin{aligned} \langle F(x_{k+1}), s_k - s_{k+1} \rangle &= \langle F(x_{k+1}) - F(x_k), s_k - s_{k+1} \rangle + \langle F(x_k), s_k - s_{k+1} \rangle \\ &\leq \langle F(x_{k+1}) - F(x_k), s_k - s_{k+1} \rangle. \end{aligned}$$

We conclude that

$$V(x_{k+1}) \leq (1 - \gamma_{k+1}) V(x_k) + \langle F(x_{k+1}) - F(x_k), s_k - s_{k+1} \rangle.$$

Therefore,

$$V(x_{k+1}) \leq (1 - \gamma_{k+1}) V(x_k) + \|F(x_{k+1}) - F(x_k)\| \cdot \|s_k - s_{k+1}\|. \tag{4}$$

Next, define

$$m_k := \min\{\|F(x_k)\|, \|F(x_{k+1})\|\}.$$

If $m_k > 0$, Corollary 12 gives

$$\|s_k - s_{k+1}\| \leq \frac{2}{\alpha\, m_k} \|F(x_{k+1}) - F(x_k)\|.$$

Also, since F is L-Lipschitz on $\mathcal{C}$,

$$\|F(x_{k+1}) - F(x_k)\| \leq L\|x_{k+1} - x_k\| \leq L\gamma_{k+1}\|s_k - x_k\| \leq L \operatorname{diam}(\mathcal{C})\, \gamma_{k+1}.$$

Substituting into (4), we obtain

$$V(x_{k+1}) \leq (1 - \gamma_{k+1}) V(x_k) + \frac{2L^2 \operatorname{diam}(\mathcal{C})^2}{\alpha\, m_k} \gamma_{k+1}^2. \tag{5}$$

Set $\theta_k := \sqrt{\gamma_{k+1}}$. We now split into two cases. In the first case, suppose $m_k > \theta_k$. Then (5) yields

$$V(x_{k+1}) \leq (1 - \gamma_{k+1}) V(x_k) + \frac{2L^2 \operatorname{diam}(\mathcal{C})^2}{\alpha} \gamma_{k+1}^{3/2}.$$

Set

$$B := \frac{2L^2 \operatorname{diam}(\mathcal{C})^2}{\alpha}.$$

Then

$$V(x_{k+1}) \le (1-\gamma_{k+1})V(x_k) + B\,\gamma_{k+1}^{3/2}. \tag{6}$$

In the second case, suppose $m_k \le \theta_k$. If $\|F(x_{k+1})\| \le \theta_k$, then trivially

$$\|F(x_{k+1})\| \le \theta_k.$$

If instead $\|F(x_k)\| \le \theta_k$, then

$$\|F(x_{k+1})\| \le \|F(x_{k+1}) - F(x_k)\| + \|F(x_k)\| \le L\,\mathrm{diam}(\mathcal{C})\,\gamma_{k+1} + \theta_k.$$

Thus in either subcase,

$$\|F(x_{k+1})\| \le L\,\mathrm{diam}(\mathcal{C})\,\gamma_{k+1} + \theta_k.$$

Since

$$V(x_{k+1}) = \langle F(x_{k+1}), x_{k+1} - s_{k+1}\rangle \le \mathrm{diam}(\mathcal{C})\|F(x_{k+1})\|,$$

we get

$$V(x_{k+1}) \le \mathrm{diam}(\mathcal{C})\big(L\,\mathrm{diam}(\mathcal{C})\gamma_{k+1} + \sqrt{\gamma_{k+1}}\big).$$

Because $\gamma_{k+1} \le \sqrt{\gamma_{k+1}}$, it follows that

$$V(x_{k+1}) \le \mathrm{diam}(\mathcal{C})(1 + L\,\mathrm{diam}(\mathcal{C}))\sqrt{\gamma_{k+1}}. \tag{7}$$

Set

$$C := \mathrm{diam}(\mathcal{C})(1 + L\,\mathrm{diam}(\mathcal{C})).$$

We now prove by induction that

$$V(x_k) \le \frac{A}{\sqrt{k}}, \qquad \forall k \ge 1,$$

where

$$A := \max\{V(x_1), 2B, C\}.$$

The case $k = 1$ is immediate, since $V(x_1) = V(x_1)/\sqrt{1} \le A/\sqrt{k}$. Assume now that $V(x_k) \le A/\sqrt{k}$ for some $k \ge 1$.

If the case where $m_k > \theta_k$ holds, then using $\gamma_{k+1} = 1/(k+1)$ and (6),

$$V(x_{k+1}) \le \frac{k}{k+1}\cdot\frac{A}{\sqrt{k}} + \frac{B}{(k+1)^{3/2}}.$$

Also, $\sqrt{k+1}/(\sqrt{k+1}+\sqrt{k}) \ge 1/2$ and $A \ge 2B$,

$$\begin{aligned}\frac{A}{\sqrt{k+1}} - \frac{k}{k+1}\cdot\frac{A}{\sqrt{k}} &= \frac{A}{(k+1)^{3/2}}\left(\frac{\sqrt{k+1}}{\sqrt{k+1}+\sqrt{k}}\right)\\ &\ge \frac{A}{2(k+1)^{3/2}}\\ &\ge \frac{B}{(k+1)^{3/2}},\end{aligned}$$

Hence, in this case

$$V(x_{k+1}) \le \frac{A}{\sqrt{k+1}}.$$

If the case where $m_k \le \theta_k$ holds, then by (7) and the fact that $A \ge C$,

$$V(x_{k+1}) \le C\sqrt{\gamma_{k+1}} \le \frac{C}{\sqrt{k+1}} \le \frac{A}{\sqrt{k+1}}.$$

This completes the induction and proves the theorem. □

When we make the stronger assumption that F is cocoercive, we are able to obtain faster convergence rates. Note that a β cocoercive operator for some $\beta > 0$ is $\frac{1}{\beta}$-Lipschitz continuous.

Theorem 14 *In addition to the standing assumptions of Section 2, suppose that F is β-cocoercive, with $\beta > 0$ and $\mathcal{C}$ is α-strongly convex. Assume there exists some $\bar{\gamma} > 0$ such that $1 - \gamma_{k+1} \geq \bar{\gamma}$ for all $k \geq 1$. Then for all $k \geq 1$,*

$$V(x_{k+1}) \leq \max\left\{(1-\gamma_{k+1})V(x_k), B\gamma_{k+1}\right\},$$

where

$$B = \operatorname{diam}(\mathcal{C})\left(\frac{2}{\alpha\beta\bar{\gamma}} + \frac{\operatorname{diam}(\mathcal{C})}{\beta}\right),$$

In particular, if $\gamma_k = 1/k$, then $\bar{\gamma} = \frac{1}{2}$, and

$$V(x_k) \leq \frac{A}{k},$$

where $A = \max\{V(x_1), B\}$.

Proof As in the proof of Theorem 13,

$$V(x_{k+1}) = (1-\gamma_{k+1})V(x_k) + (1-\gamma_{k+1})\langle F(x_{k+1}) - F(x_k), x_k - s_k\rangle + \langle F(x_{k+1}), s_k - s_{k+1}\rangle.$$

The definition of β-cocoercivity with $(x, y) = (x_{k+1}, x_k)$ gives

$$\begin{aligned}\langle F(x_{k+1}) - F(x_k), s_k - x_k\rangle &= \frac{1}{\gamma_{k+1}}\langle F(x_{k+1}) - F(x_k), x_{k+1} - x_k\rangle \\ &\geq \frac{\beta}{\gamma_{k+1}}\|F(x_{k+1}) - F(x_k)\|^2. \end{aligned} \tag{8}$$

As in the proof of Theorem 13, we control the $\langle F(x_{k+1}), s_k - s_{k+1}\rangle$ term by observing

$$\langle F(x_{k+1}), s_k - s_{k+1}\rangle \leq \langle F(x_{k+1}) - F(x_k), s_k - s_{k+1}\rangle.$$

This time, β-cocoercivity of F gives us that F is $(1/\beta)$-Lipschitz:

$$\|F(x_{k+1}) - F(x_k)\| \leq \frac{1}{\beta}\|x_{k+1} - x_k\| = \frac{\gamma_{k+1}}{\beta}\|s_k - x_k\| \leq \frac{\gamma_{k+1}}{\beta}\operatorname{diam}(\mathcal{C}). \tag{9}$$

Now, define

$$m_k := \min\{\|F(x_k)\|, \|F(x_{k+1})\|\}.$$

If $m_k > 0$, Corollary 12 gives

$$\|s_k - s_{k+1}\| \leq \frac{2}{\alpha\, m_k}\,\|F(x_{k+1}) - F(x_k)\|.$$

So, in this case we can combine (8) and (9) to obtain the bound

$$V(x_{k+1}) \leq (1-\gamma_{k+1})V(x_k) + \left(\frac{2}{\alpha\, m_k} - \frac{\beta(1-\gamma_{k+1})}{\gamma_{k+1}}\right)\|F(x_{k+1}) - F(x_k)\|^2. \tag{10}$$

Set $\theta_k := \frac{2\gamma_{k+1}}{\alpha\beta(1-\gamma_{k+1})}$. Like before, we split into two cases. First, suppose we are in the case where $m_k > \theta_k$, then

$$\frac{2}{\alpha\, m_k} - \frac{\beta(1-\gamma_{k+1})}{\gamma_{k+1}} < 0,$$

and

$$V(x_{k+1}) \le (1-\gamma_{k+1})V(x_k). \tag{11}$$

In the second case, suppose $m_k \le \theta_k$. If $\|F(x_{k+1})\| \le \theta_k$, then trivially

$$\|F(x_{k+1})\| \le \theta_k.$$

If instead $\|F(x_k)\| \le \theta_k$, we have

$$\|F(x_{k+1})\| \le \frac{\gamma_{k+1}}{\beta}\operatorname{diam}(\mathcal{C}) + \theta_k,$$

as in the proof of Theorem 13. In both subcases, we can say that $\|F(x_{k+1})\| \le \frac{\gamma_{k+1}}{\beta}\operatorname{diam}(\mathcal{C}) + \theta_k$. It follows that

$$\begin{aligned} V(x_{k+1}) &= \langle F(x_{k+1}), x_{k+1} - s_{k+1}\rangle \\ &\le \operatorname{diam}(\mathcal{C})\|F(x_{k+1})\| \\ &\le \operatorname{diam}(\mathcal{C})\left(\theta_k + \frac{\gamma_{k+1}\operatorname{diam}(\mathcal{C})}{\beta}\right) \\ &= \gamma_{k+1}\operatorname{diam}(\mathcal{C})\left(\frac{2}{\alpha\beta(1-\gamma_{k+1})} + \frac{\operatorname{diam}(\mathcal{C})}{\beta}\right) \\ &\le \gamma_{k+1}\operatorname{diam}(\mathcal{C})\left(\frac{2}{\alpha\beta\bar{\gamma}} + \frac{\operatorname{diam}(\mathcal{C})}{\beta}\right) \\ &= B\gamma_{k+1}, \end{aligned} \tag{12}$$

where we set

$$B := \operatorname{diam}(\mathcal{C})\left(\frac{2}{\alpha\beta\bar{\gamma}} + \frac{\operatorname{diam}(\mathcal{C})}{\beta}\right),$$

where we recall $1-\gamma_{k+1} \ge \bar{\gamma} > 0$ for all $k \ge 1$. Therefore, when $m_k \le \theta_k$,

$$V(x_{k+1}) \le B\gamma_{k+1}.$$

We now prove by induction that

$$V(x_k) \le \frac{A}{k}, \qquad \forall k \ge 1,$$

where

$$A := \max\{V(x_1), B\}.$$

The case $k = 1$ is clear, since $V(x_k) = V(x_1)/1 \le A/k$. Assume now that $V(x_k) \le A/k$ for some $k \ge 1$. If the case where $m_k > \theta_k$ holds, then using $\gamma_{k+1} = 1/(k+1)$ and (11),

$$V(x_{k+1}) \le \frac{k}{k+1}V(x_k) = \frac{k}{k+1}\cdot\frac{A}{k} = \frac{A}{k+1}.$$

If the case where $m_k \le \theta_k$ holds, then by (12) and the fact that $A \ge B$,

$$V(x_{k+1}) \le B\gamma_{k+1} = \frac{B}{k+1} \le \frac{A}{k+1}.$$

This completes the induction and proves the claim. □

4 Conclusion

We have shown that the Frank-Wolfe algorithm for solving variational inequalities over compact, convex sets under a monotone C^1 operator and vanishing, nonsummable step sizes converges. We also show iterate convergence to the unique solution in the special case where F is instead assumed to be strongly monotone. The strongly monotone setting generalizes Hammond's generalized fictitious play, which was conjectured by Hammond to converge to a solution of the corresponding variational inequality problem. Thus, this result proves Hammond's longstanding conjecture.

For strongly convex sets, we establish rates of convergence with no assumption that F vanishes over the set $\mathcal{C}$. The convergence rate of Frank-Wolfe for monotone variational inequality problems over sets that aren't uniformly smooth remains an open problem. An important subcase of the nonsmooth setting is where $\mathcal{C}$ is a polytope.

References

[1] Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval Research Logistics Quarterly **3**(1-2), 95–110 (1956) https://doi.org/10.1002/nav.3800030109 https://onlinelibrary.wiley.com/doi/pdf/10.1002/nav.3800030109

[2] Hammond, J.H.: Solving Asymmetric Variational Inequality Problems and Systems of Equations with Generalized Nonlinear Programming Algorithms. PhD thesis, Massachusetts Institute of Technology (1984)

[3] Brown, G.W.: Iterative Solution of Games by Fictitious Play. In: Koopmans, T.C. (ed.) Activity Analysis of Production and Allocation. Wiley, New York (1951)

[4] Robinson, J.: An Iterative Method of Solving a Game. Annals of Mathematics **54**(2), 296–301 (1951)

[5] Shapiro, H.N.: Note on a Computation Method in the Theory of Games. Communications on Pure and Applied Mathematics **11**(4), 587–593 (1958) https://doi.org/10.1002/cpa.3160110408

[6] Daskalakis, C., Pan, Q.: A Counter-example to Karlin's Strong Conjecture for Fictitious Play. In: 55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18–21, 2014, pp. 11–20. IEEE Computer Society, Washington DC, USA (2014). https://doi.org/10.1109/FOCS.2014.10

[7] Abernethy, J., Lai, K.A., Wibisono, A.: Fast Convergence of Fictitious Play for Diagonal Payoff Matrices. In: Proceedings of the 2021 ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1387–1404 (2021). https://doi.org/10.1137/1.9781611976465.84 . https://epubs.siam.org/doi/abs/10.1137/1.9781611976465.84

[8] Wang, Y.: Tie-breaking Agnostic Lower Bound for Fictitious Play (2025). https://arxiv.org/abs/2507.09902

[9] Gidel, G., Jebara, T., Lacoste-Julien, S.: Frank-Wolfe Algorithms for Saddle Point Problems. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (2017)

[10] Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of Frank-Wolfe optimization variants. In: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1. NIPS'15, pp. 496–504. MIT Press, Cambridge, MA, USA (2015)

[11] Hough, M., Vavasis, S.A.: A Primal-Dual Frank-Wolfe Algorithm for Linear Programming (2024). https://arxiv.org/abs/2402.18514

[12] Hartman, P., Stampacchia, G.: On some non-linear elliptic differential-functional equations. Acta Mathematica **115**(1), 271–310 (1966) https://doi.org/10.1007/BF02392210

[13] Benaïm, M., Hofbauer, J., Sorin, S.: Stochastic Approximations and Differential Inclusions. SIAM Journal on Control and Optimization **44**(1), 328–348 (2005) https://doi.org/10.1137/S0363012904439301

[14] Chen, Y.-W., Kizilkale, C., Arcak, M.: Solving Monotone Variational Inequalities with Best Response Dynamics. In: 2024 IEEE 63rd Conference on Decision and Control (CDC), pp. 1751–1756 (2024). https://doi.org/10.1109/CDC56724.2024.10886164

[15] Tao, T.: Nonlinear Dispersive Equations. CBMS Regional Conference Series in Mathematics. American Mathematical Society, Providence, RI (2006)