Acceleration of Frank-Wolfe Algorithms with Open-Loop Step-Sizes

Elias Wirth
wirth@math.tu-berlin.de
Berlin Institute of Technology

Thomas Kerdreux
thomaskerdreux@gmail.com
Geolabe LLC

Sebastian Pokutta
pokutta@zib.de
Berlin Institute of Technology &
Zuse Institute Berlin

Abstract

Frank-Wolfe algorithms (FW) are popular first-order methods for solving constrained convex optimization problems that rely on a linear minimization oracle instead of potentially expensive projection-like oracles. Many works have identified accelerated convergence rates under various structural assumptions on the optimization problem and for specific FW variants when using line-search or short-step, requiring feedback from the objective function. Little is known about accelerated convergence regimes when utilizing open-loop step-size rules, a.k.a. FW with pre-determined step-sizes, which are algorithmically extremely simple and stable. Not only is FW with open-loop step-size rules not always subject to the same convergence rate lower bounds as FW with line-search or short-step, but in some specific cases, such as kernel herding in infinite dimensions, it has been empirically observed that FW with open-loop step-size rules leads to faster convergence than FW with line-search or short-step. We propose a partial answer to this unexplained phenomenon in kernel herding, characterize a general setting for which FW with open-loop step-size rules converges non-asymptotically faster than with line-search or short-step, and derive several accelerated convergence results for FW with open-loop step-size rules.

1 INTRODUCTION

In this paper, we address the constrained convex optimization problem

$$\min_{x \in \mathcal{C}} f(x), \tag{OPT}$$

where $\mathcal{C} \subseteq \mathbb{R}^d$ is a compact convex set and $f\colon \mathcal{C} \to \mathbb{R}$ is a convex and L-smooth function. Throughout, let $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ be the constrained optimal solution. A classical approach to addressing (OPT) is to apply *projected gradient descent*. When the geometry of $\mathcal{C}$ is too complex, the projection step can become computationally too expensive. In these situations, the *Frank-Wolfe algorithm* (FW) (Frank and Wolfe, 1956), a.k.a. the conditional gradients algorithm (Levitin and Polyak, 1966), described in Algorithm 1, is an efficient alternative, as it only requires first-order access to the objective f and access to a linear minimization oracle (LMO) for the feasible region, that is, given a vector $c \in \mathbb{R}^d$, the LMO outputs $\operatorname{argmin}_{x \in \mathcal{C}} \langle c, x \rangle$. At each iteration, the algorithm calls the LMO, $p_t \in \operatorname{argmin}_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle$, and takes a step in the direction of the vertex p_t to obtain the next iterate $x_{t+1} = (1-\eta_t)x_t + \eta_t p_t$. As a convex combination of elements of $\mathcal{C}$, x_t remains in the feasible region $\mathcal{C}$ throughout the algorithm's execution. Various options exist for the choice of η_t, such as the *open-loop step-size*[1], a.k.a. *agnostic step-size*, rules $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 1}$ (Dunn and Harshbarger, 1978) or line-search $\eta_t \in \operatorname{argmin}_{\eta \in [0,1]} f((1-\eta)x_t + \eta p_t)$. Another classical approach, the *short-step* step-size $\eta_t = \min\{\frac{\langle \nabla f(x_t), x_t - p_t \rangle}{L \|x_t - p_t\|_2^2}, 1\}$, henceforth referred to as short-step, is determined by minimizing a quadratic upper bound on the L-smooth objective function. There also exist variants that adaptively estimate local L-smoothness parameters (Pedregosa et al., 2018).

Algorithm 1 Frank-Wolfe algorithm (FW)

Input: $x_0 \in \mathcal{C}$, step-size $\eta_t \in [0,1]$ for $t \in \{0, \dots, T-1\}$.
for $t = 0, \dots, T-1$ **do**
  $p_t \in \operatorname{argmin}_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle$
  $x_{t+1} \leftarrow (1-\eta_t)x_t + \eta_t p_t$
end for

[1] Open-loop is a term from control theory and here implies that there is no feedback from the objective function to the step-size.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

Table 1: Comparison of convergence rates for the Frank-Wolfe algorithm under different assumptions. We denote the optimal solution by $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$. For all results, convexity of $\mathcal{C}$ and convexity and smoothness of f are assumed and thus not explicitly mentioned. The big-O notation $\mathcal{O}(\cdot)^*$ indicates that a result only holds asymptotically and "str. convex" is an abbreviation for strongly convex. For step-size rule, "any" refers to line-search, short-step, and open-loop step-size $\eta_t = \frac{2}{t+2}$. Shading is used to separate results belonging to similar settings and our results are denoted in bold.

References	Region C	Objective f	Location of x	Rate	Step-size rule
(Jaggi, 2013)	-	-	unrestricted	$O(1/t)$	any
(Garber and Hazan, 2015)	-	str. convex	interior	$O(e^{-t})$	line-search, short-step
Theorem 3.6	-	str. convex	interior	$O(1/t^2)$	open-loop $\eta_t = \frac{4}{t+4}$
(Levitin and Polyak, 1966) (Demianov and Rubinov, 1970) (Dunn, 1979)	str. convex	$\|\nabla f(x)\|_2 > 0$ for all $x \in C$	unrestricted	$O(e^{-t})$	line-search, short-step
Theorem D.2	str. convex	$\|\nabla f(x)\|_2 > 0$ for all $x \in C$	unrestricted	$O(1/t^2)$	open-loop $\eta_t = \frac{4}{t+4}$
Remark D.3	str. convex	$\|\nabla f(x)\|_2 > 0$ for all $x \in C$	unrestricted	$O(1/t^{\ell/2})$	open loop $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 4}$
Remark D.3	str. convex	$\|\nabla f(x)\|_2 > 0$ for all $x \in C$	unrestricted	$O(e^{-t})$	constant
(Garber and Hazan, 2015)	str. convex	str. convex	unrestricted	$O(1/t^2)$	line-search, short-step
Theorem E.1	str. convex	str. convex	unrestricted	$O(1/t^2)$	open-loop $\eta_t = \frac{4}{t+4}$
(Wolfe, 1970)	polytope	str. convex	interior of face	$(1/t^{1+})$	line-search, short-step
(Bach, 2021)	polytope	str. convex	interior of face	$O(1/t^2)$	open-loop $\eta_t = \frac{2}{t+2}$
Theorem 4.3	polytope	str. convex	interior of face	$O(1/t^2)$	open-loop $\eta_t = \frac{4}{t+4}$

1.1 Related Work

Frank-Wolfe algorithms (FW) enjoy various appealing properties (Jaggi, 2013). They are first-order methods, easy to implement, projection-free, affine-invariant (Lacoste-Julien and Jaggi, 2013; Lan, 2013; Kerdreux et al., 2021c; Pena, 2021), and iterates are sparse convex combinations of extreme points of the feasible region. FW is thus an attractive algorithm for practitioners that work at scale and appears in a variety of scenarios in machine learning, for example, deep learning, optimal transport, structured prediction, and video co-localization (Ravi et al., 2018; Courty et al., 2016; Giesen et al., 2012; Joulin et al., 2014). The drawback of FW is its slow convergence rate in primal gap of $h_t = f(x_t) - f(x^*) = \mathcal{O}(1/t)$, where $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$, in comparison to proximal methods. Under mild assumptions, Wolfe (1970) proved that when the feasible region is a polytope and the optimum lies in the relative interior of an at least one-dimensional face, for any $\epsilon > 0$, FW with line-search or short-step converges at a rate of $\Omega\left(1/t^{1+\epsilon}\right)$, see also Canon and Cullum (1968). To circumnavigate the lower bound and achieve linear convergence rates, algorithmic modifications of FW are necessary, see, for example, Lacoste-Julien and Jaggi (2015); Garber and Hazan (2016); Braun et al. (2019); Combettes and Pokutta (2020); Garber (2020). For FW with open-loop step-size rules, the lower bound of Wolfe (1970) does not hold and Bach (2021) proved an asymptotic convergence rate of $\mathcal{O}(1/t^2)$ in the setting of Wolfe (1970). Proving that the latter result holds non-asymptotically remains an open problem. Other drawbacks of line-search and short-step are that the former can be difficult to compute and the latter requires knowledge of the smoothness constant of the objective f. Open-loop step-size rules, on the other hand, are problem-agnostic and, thus, easy to compute. Finally, FW with open-loop step-size $\eta_t = \frac{1}{t+1}$ is known to be equivalent to the kernel-herding algorithm (Bach et al., 2012). Thus, FW has also been studied in kernel herding and in the infinite-dimensional kernel-herding setting in Bach et al. (2012, Figure 3, right), it is empirically observed that FW with open-loop step-size rules converges at the optimal rate of $\mathcal{O}(1/t^2)$, whereas FW with line-search or short-step converges at a rate of essentially $\Omega(1/t)$.

1.2 Contributions

Despite the recent research interest in FW and its variants, FW with open-loop step-size rules is still not fully understood. Especially the practically relevant kernel-herding problem in Bach et al. (2012) where FW with open-loop step-size rules converges faster than FW with line-search or short-step warrants further investigation. The goal of this paper is to address the current gaps in our understanding of FW with open-loop step-size rules and characterize settings in which FW with open-loop step-size rules converges at accelerated rates. Our contributions are four-fold:

First, we prove accelerated convergence rates of FW with open-loop step-size rules in several settings where FW with line-search or short-step enjoys accelerated convergence rates. These results are summarized in Table 1. Since FW with open-loop step-size rules is not a descent method, we require a different proof technique than for proving accelerated convergence results for FW with line-search or short-

step.

Second, we characterize a setting for which FW with open-loop step-size rules is non-asymptotically faster than FW with line-search or short-step.

Third, we provide a theoretical analysis of the accelerated convergence rate of FW with open-loop step-size rules in kernel herding that was left unexplained in Bach et al. (2012).

Finally, we provide numerical experiments that illustrate our results and lead to several open questions.

1.3 Outline

Preliminaries are introduced in Section 2. In Section 3, we present a proof blueprint for obtaining accelerated convergence rates for FW with open-loop step-sizes. In Section 4, we characterize a problem setting where FW with open-loop step-size rules converges faster than FW with line-search or short-step. In Section 5, we prove accelerated convergence rates for FW with open-loop step-size rules in the infinite-dimensional kernel-herding setting of Bach et al. (2012). The numerical experiments are found in Section 6. Finally, we discuss our results in Section 7.

2 PRELIMINARIES

Throughout, let $d \in \mathbb{N}$. Let $\mathbb{0} \in \mathbb{R}^d$ denote the all-zeros vector, let $\mathbb{1} \in \mathbb{R}^d$ denote the all-ones vector, and let $\bar{\mathbb{1}} \in \mathbb{R}^d$ be a vector such that $\bar{\mathbb{1}}_i = 0$ for all $i \in \{1, \dots, \lceil d/2 \rceil\}$ and $\bar{\mathbb{1}}_i = 1$ for all $i \in \{\lceil d/2 \rceil + 1, \dots, d\}$. For $i \in \{1, \dots, d\}$, let $e^{(i)} \in \mathbb{R}^d$ be the ith unit vector such that $e^{(i)}_i = 1$ and $e^{(i)}_j = 0$ for all $j \in \{1, \dots, d\} \setminus \{i\}$. Given a vector $x \in \mathbb{R}^d$, define its support as $\operatorname{supp}(x) = \{i \in \{1, \dots, d\} \mid x_i \neq 0\}$. Let $I \in \mathbb{R}^{d \times d}$ denote the identity matrix. Given a set $\mathcal{C} \subseteq \mathbb{R}^d$, let $\operatorname{aff}(\mathcal{C})$, $\operatorname{conv}(\mathcal{C})$, $\operatorname{span}(\mathcal{C})$, and $\operatorname{vert}(\mathcal{C})$ denote the affine hull, the convex hull, the span, and the set of vertices of $\mathcal{C}$, respectively. For $z \in \mathbb{R}^d$ and $\beta > 0$, the *ball* of radius β around z is defined as $B_\beta(z) := \{x \in \mathbb{R}^d \mid \|x - z\|_2 \leq \beta\}$. For the iterates of Algorithm 1, we denote the *primal gap* at iteration $t \in \{0, \dots, T\}$ by $h_t = f(x_t) - f(x^*)$, where $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$. Finally, for $x \in \mathbb{R}$, let $[x] := x - \lfloor x \rfloor$. We introduce several definitions.

Definition 2.1 (Uniformly convex set)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set, $\alpha > 0$, and $q > 0$. We say that $\mathcal{C}$ is (α, q)-*uniformly convex* with respect to $\|\cdot\|_2$ if for all $x, y \in \mathcal{C}$, $\gamma \in [0, 1]$, and $z \in \mathbb{R}^d$ such that $\|z\|_2 = 1$, it holds that

$$\gamma x + (1 - \gamma) y + \gamma(1 - \gamma) \frac{\alpha}{2} \|x - y\|_2^q z \in \mathcal{C}.$$

We refer to $(\alpha, 2)$-uniformly convex sets as α-*strongly convex sets*.

Definition 2.2 (Smooth function)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set, let $f \colon \mathcal{C} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{C}$, and let $L > 0$. We say that f is L-*smooth* over $\mathcal{C}$ with respect to $\|\cdot\|_2$ if for all $x, y \in \mathcal{C}$, it holds that

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|x - y\|_2^2.$$

Definition 2.3 (Hölderian error bound)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set, let $f \colon \mathcal{C} \to \mathbb{R}$ be convex and differentiable in an open set containing $\mathcal{C}$, let $\mu > 0$, and let $\theta \in [0, 1/2]$. We say that f satisfies a (μ, θ)-*Hölderian error bound* if for all $x \in \mathcal{C}$ and $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$, it holds that

$$\mu (f(x) - f(x^*))^\theta \geq \min_{y \in \operatorname{argmin}_{z \in \mathcal{C}} f(z)} \|x - y\|_2. \tag{1}$$

Note that $\theta \leq 1/2$ is necessary because we only consider smooth functions in this work. Throughout, for ease of notation, we assume that $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ is unique. This follows, for example, from the assumption that f is strictly convex. When $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ is unique, (1) becomes

$$\mu (f(x) - f(x^*))^\theta \geq \|x - x^*\|_2. \tag{HEB}$$

However, with the appropriate modifications, all of our results also extend to functions that are not strictly convex. An important family of functions satisfying (HEB) is the family of uniformly convex functions, which interpolate between convex functions ($\theta = 0$) and strongly convex functions ($\theta = 1/2$).

Definition 2.4 (Uniformly convex function)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set, let $f \colon \mathcal{C} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{C}$, let $\alpha_f > 0$, and let $r \geq 2$. We say that f is (α_f, r)-*uniformly convex* over $\mathcal{C}$ with respect to $\|\cdot\|_2$ if for all $x, y \in \mathcal{C}$, it holds that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha_f}{2} \|x - y\|_2^r.$$

We refer to $(\alpha_f, 2)$-uniformly convex functions as α_f-*strongly convex*.

Note that (α_f, r)-uniformly convex functions satisfy a $((2/\alpha_f)^{1/r}, 1/r)$-(HEB):

$$\begin{aligned} f(x) - f(x^*) &\geq \langle \nabla f(x^*), x - x^* \rangle + \frac{\alpha_f}{2} \|x - x^*\|_2^r \\ &\geq \frac{\alpha_f}{2} \|x - x^*\|_2^r. \end{aligned}$$

3 ACCELERATED CONVERGENCE RESULTS

FW with open-loop step-size rules was already studied by Dunn and Harshbarger (1978) and currently, two open-loop step-sizes are prevalent, $\eta_t = \frac{1}{t+1}$, for which the best known convergence rate is $\mathcal{O}(\log(t)/t)$, and $\eta_t = \frac{2}{t+2}$, for

which a faster convergence rate of $\mathcal{O}(1/t)$ holds, see, for example, Dunn and Harshbarger (1978) and Jaggi (2013), respectively. In this section, we present accelerated convergence results for FW with the open-loop step-size $\eta_t = \frac{4}{t+4}$. Note that the convergence-rate results presented in Table 1 proved in this paper for FW with $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 1}$ can always be generalized (up to a constant) to $\eta_t = \frac{j}{t+j}$ for $j \in \mathbb{N}_{\geq \ell}$.

3.1 Convergence Rate Of $\mathcal{O}(1/t)$

We begin the analysis of FW with open-loop step-size rules by first recalling the, to the best of our knowledge, best general convergence rate of the algorithm. Consider the setting when $\mathcal{C} \subseteq \mathbb{R}^d$ is a compact convex set and $f\colon \mathcal{C} \to \mathbb{R}$ is a convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Then, the iterates of Algorithm 1 with any step-size $\eta_t \in [0,1]$ satisfy

$$h_{t+1} \leq h_t - \eta_t \langle \nabla f(x_t), x_t - p_t\rangle + \frac{\eta_t^2 L \|x_t - p_t\|_2^2}{2}, \tag{Progress-Bound}$$

which follows from the smoothness of f. With (Progress-Bound), it is possible to derive a baseline convergence rate for FW with open-loop step-size $\eta_t = \frac{4}{t+4}$ similar to Jaggi (2013, Theorem 1) for $\eta_t = \frac{2}{t+2}$.

Proposition 3.1 ($\mathcal{O}(1/t)$ convergence rate)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, it holds that $h_t \leq \frac{8L\delta^2}{t+3} = \eta_{t-1} 2L\delta^2$ for all $t \in \{1,\dots,T\}$.*

To prove accelerated convergence rates for FW with open-loop step-size rules, we require bounds on the *Frank-Wolfe gap* (FW gap) $\max_{p\in\mathcal{C}} \langle \nabla f(x_t), x_t - p\rangle$, which appears in the middle term in (Progress-Bound).

3.2 Optimal Solution In The Interior Of $\mathcal{C}$, A Blueprint For Acceleration

Traditionally, to prove accelerated convergence rates for FW with line-search or short-step, the geometry of the feasible region, curvature assumptions on the objective function, and information on the location of the optimum are exploited, see, for example, Levitin and Polyak (1966); Demianov and Rubinov (1970); Guélat and Marcotte (1986); Garber and Hazan (2015). We demonstrate that a similar approach leads to acceleration results for FW with open-loop step-size rules, however, requiring a different proof technique as FW with open-loop step-size rules is not monotonous in primal gap. We present a blueprint of the technique used to derive most of the accelerated rates via the setting when the optimum of f is in the relative interior of the feasible region $\mathcal{C}$ and the objective function f satisfies (HEB).

Our approach for proving accelerated convergence rates is based on bounding the FW gap to counteract the error accumulated from the right-hand term in (Progress-Bound). More formally, we prove the existence of $\phi > 0$, such that there exists an iteration $S \in \mathbb{N}$ such that for all iterations $t \geq S$ of FW, it holds that

$$\frac{\langle \nabla f(x_t), x_t - p_t\rangle}{\|x_t - p_t\|_2} \geq \phi \frac{\langle \nabla f(x_t), x_t - x^*\rangle}{\|x_t - x^*\|_2}. \tag{Scaling}$$

Inequalities that bound (Scaling) from either side are referred to as *scaling inequalities*. Intuitively speaking, scaling inequalities relate the *FW direction* $\frac{p_t - x_t}{\|p_t - x_t\|_2}$ with the *optimal descent direction* $\frac{x^* - x_t}{\|x^* - x_t\|_2}$. Scaling inequalities stem from the geometry of the feasible region, properties of the objective function, or information on the location of the optimum. The scaling inequality below exploits the latter property.

Lemma 3.2 (Guélat and Marcotte, 1986)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and suppose that there exists $\beta > 0$ such that $\operatorname{aff}(\mathcal{C}) \cap B_\beta(x^*) \subseteq \mathcal{C}$. Then, for all $x \in \mathcal{C} \cap B_\beta(x^*)$,*

$$\frac{\langle \nabla f(x), x - p\rangle}{\|x - p\|_2} \geq \frac{\beta}{\delta} \|\nabla f(x)\|_2, \tag{Scaling-INT}$$

where $p \in \operatorname{argmin}_{v\in\mathcal{C}} \langle \nabla f(x), v\rangle$.

We prove in the lemma below that there exists an iteration $S \in \mathbb{N}$, such that for all $t \geq S$, it holds that $x_t \in B_\beta(x^*)$ and (Scaling-INT) is satisfied.

Lemma 3.3 (Distance to optimum)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function satisfying a (μ,θ)-(HEB) for some $\mu > 0$ and $\theta \in]0, 1/2]$ with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and let $\beta > 0$. Let*

$$S = \left\lceil \left(8L\delta^2 (\mu/\beta)^{1/\theta}\right)^{\theta} \right\rceil \geq 1, \tag{2}$$

$T \in \mathbb{N}$, and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, it holds that $\|x_t - x^*\|_2 \leq \beta$ for all $t \in \{S,\dots,T\}$.

We require an additional scaling inequality based on the objective satisfying (HEB).

Lemma 3.4. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set and let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex function satisfying a (μ,θ)-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer*

$x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$. *Then, for all* $x \in \mathcal{C}$,

$$\|\nabla f(x)\|_2 \geq \frac{\langle \nabla f(x), x - x^* \rangle}{\|x - x^*\|_2} \geq \frac{1}{\mu}(f(x) - f(x^*))^{1-\theta}. \quad \text{(Scaling-HEB)}$$

For $t \geq S$, where $S = \left\lceil 8L\delta^2 (2\mu/\beta)^{1/\theta} \right\rceil$, we can chain (Scaling-INT) and (Scaling-HEB) together and plug the resulting inequality into (Progress-Bound) yielding

$$h_{t+1} \leq h_t - \eta_t \frac{\beta^2}{2\mu\delta} h_t^{1-\theta} + \frac{\eta_t^2 L \delta^2}{2}$$

for all $t \in \{S, \dots, T-1\}$. Combined with (9), we can then bound the primal-gap progress via

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t \frac{\beta^2}{4\mu\delta} h_t^{1-\theta} + \frac{\eta_t^2 L \delta^2}{2} \quad (3)$$

for all $t \in \{S, \dots, T-1\}$. For sequences satisfying this type of inequality, the lemma below implies accelerated convergence rates.

Lemma 3.5. *Let* $\psi \in [0, 1/2]$, $S, T \in \mathbb{N}_{\geq 1}$, *and* $\eta_t = \frac{4}{t+4}$ *for all* $t \in \mathbb{Z}$. *Suppose that there exist constants* $A, B, C > 0$, *a nonnegative sequence* $\{C_t\}_{t=S}^{T-1}$ *such that* $C \geq C_t \geq 0$ *for all* $t \in \{S, \dots, T-1\}$, *and a nonnegative sequence* $\{h_t\}_{t=S}^{T}$ *such that*

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t A C_t h_t^{1-\psi} + \eta_t^2 B C_t \quad (4)$$

for all $t \in \{S, \dots, T-1\}$. *Then,*

$$h_t \leq \max\left(\left(\frac{t}{S}\right)^{2/(1-\psi)} h_S, \left(\frac{t^2 B}{A}\right)^{1/(1-\psi)} + \frac{2}{t} BC \right) : \quad (5)$$

for all $t \in \{S, \dots, T\}$.

We conclude the presentation of our proof blueprint by stating the first accelerated convergence rate for FW with open-loop step-size $\eta_t = \frac{4}{t+4}$ when the optimum lies in the relative interior of $\mathcal{C}$ and the objective function satisfies (HEB), a setting for which multiple accelerated convergence results are known: FW with line-search or short-step converges linearly if the objective function is strongly convex, see, for example, Guélat and Marcotte (1986); Garber and Hazan (2015). Further, FW with open-loop step-size $\eta_t = \frac{1}{t+1}$ converges at a rate of $\mathcal{O}(1/t^2)$ when the optimum lies in the relative interior of the feasible region and the objective function has the form $f(x) = \frac{1}{2}\|x - b\|_2^2$ for some $b \in \mathcal{C}$ (Chen et al., 2012).

Theorem 3.6 (Optimal solution in the interior of $\mathcal{C}$)**.** *Let* $\mathcal{C} \subseteq \mathbb{R}^d$ *be a compact convex set of diameter* $\delta > 0$, *let* $f : \mathcal{C} \to \mathbb{R}$ *be a convex and* L*-smooth function satisfying a* (μ, θ)*-(HEB) for some* $\mu > 0$ *and* $\theta \in]0, 1/2]$ *with unique minimizer* $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$, *and suppose that there exists* $\beta > 0$ *such that* $\operatorname{aff}(\mathcal{C}) \cap B_\beta(x^*) \subseteq \mathcal{C}$. *Let*

$$S = \left\lceil 8L\delta^2 (2\mu/\beta)^{1/\theta} \right\rceil, \quad (6)$$

$T \in \mathbb{N}$, *and* $\eta_t = \frac{4}{t+4}$ *for all* $t \in \mathbb{Z}$. *Then, for the iterates of Algorithm 1 with open-loop step-size* η_t, *it holds that*

$$h_t \leq \max\left(\left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^{1/(1-\theta)} h_S, \left(\frac{\eta_{t-2} 2\mu L \delta^3}{\beta^2}\right)^{1/(1-\theta)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right)$$

for all $t \in \{S, \dots, T\}$.

We complement Theorem 3.6 with a discussion on the lower bound of the convergence rate of FW when the optimum is in the interior of the probability simplex by Jaggi (2013). We recall the result below.

Lemma 3.7 (Jaggi, 2013)**.** *Let* $\mathcal{C} \subseteq \mathbb{R}^d$ *be the probability simplex,* $f(x) = \|x\|_2^2$, *and* $t \in \{1, \dots, d\}$. *It holds that*

$$\min_{\substack{x \in \mathcal{C} \\ |\operatorname{supp}(x)| \leq t}} f(x) = \frac{1}{t},$$

where $|\operatorname{supp}(x)|$ *denotes the number of non-zero entries of* x.

Remark 3.8 (Compatibility with lower bound from Jaggi (2013))**.** In Lemma 3.7, the optimum $x^* = \frac{1}{d}\mathbb{1} \in \mathbb{R}^d$ lies in the interior of $\mathcal{C}$ and $\min_{x \in \mathcal{C}} f(x) = 1/d$. When $\mathcal{C}$ is the probability simplex, all of its vertices are of the form $e^{(i)} = (0, \dots, 0, 1, 0, \dots, 0)^\intercal \in \mathbb{R}^d$ for $i \in \{1, \dots, d\}$, where the ith entry of $e^{(i)}$ is 1. Thus, any iteration of FW can modify at most one entry of iterate x_t and the primal gap is at best $h_t = 1/t - 1/d$ for $t \in \{1, \dots, d\}$. Applying Theorem 3.6 to the setting of Lemma 3.7, we observe that $\beta = 1/d$ and acceleration starts only after $S = \Omega\left(d^{1/\theta}\right) \geq \Omega(d)$ iterations. Thus, Theorem 3.6 does not contradict the lower bound from Lemma 3.7.

3.3 Other Settings

We also derive accelerated convergence rates for the settings when the norm of the gradient of f is bounded from below by a nonnegative constant and the feasible region $\mathcal{C}$ is uniformly convex in Appendix D and when f satisfies (HEB) and the feasible region $\mathcal{C}$ is uniformly convex in Appendix E.

4 OPTIMAL SOLUTION IN THE INTERIOR OF AN AT LEAST ONE-DIMENSIONAL FACE OF $\mathcal{C}$

In this section, we consider the setting when the feasible region is a polytope, the objective function is strongly convex, and the optimum lies in the relative interior of an at least one-dimensional face $\mathcal{C}^*$ of $\mathcal{C}$. Then, under mild assumptions, FW with line-search or short-step converges at a rate of $\Omega(1/t^{1+\epsilon})$ for any $\epsilon > 0$ (Wolfe, 1970).

We prove that in the same setting, FW with open-loop step-size rules admits a convergence rate of $\mathcal{O}(1/t^2)$ and is thus non-asymptotically faster than FW with line-search or short-step. To prove the result, we require two assumptions, the first of which stems from *active set identification*, that is, identifying the face $\mathcal{C}^*$ containing the optimal solution x^*, an important problem: After having determined the active face, it is possible to apply faster methods and the dimension dependence of the convergence rate can often be reduced to the dimension of the optimal face, see, for example, Bertsekas (1982); Guélat and Marcotte (1986); Birgin and Martínez (2002); Hager and Zhang (2006); Bomze et al. (2019, 2020) for examples with a focus on FW. For our current setting, it is possible to determine the number of iterations necessary for FW with open-loop step-size rules to identify the optimal face when the following regularity assumption, already used in Wolfe (1970); Guélat and Marcotte (1986); Garber (2020), is satisfied.

Assumption 4.1 (Strict complementarity)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope and let $f\colon \mathcal{C} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{C}$. Suppose that $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ is unique and contained in a face $\mathcal{C}^*$ of $\mathcal{C}$ and that there exists $\kappa > 0$ such that if $p \in \operatorname{vert}(\mathcal{C}) \setminus \mathcal{C}^*$, then $\langle \nabla f(x^*), p - x^* \rangle \geq \kappa$; otherwise, if $p \in \operatorname{vert}(\mathcal{C}^*)$, then $\langle \nabla f(x^*), p - x^* \rangle = 0$.

We also assume the optimum to lie in the relative interior of an at least one-dimensional face $\mathcal{C}^*$ of $\mathcal{C}$.

Assumption 4.2 (Optimal solution in the interior of a face of $\mathcal{C}$)**.** Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope and let $f\colon \mathcal{C} \to \mathbb{R}$. Suppose that $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ is unique and contained in the relative interior of an at least one-dimensional face $\mathcal{C}^*$ of $\mathcal{C}$, that is, there exists $\beta > 0$ such that $\emptyset \neq B_\beta(x^*) \cap \operatorname{aff}(\mathcal{C}^*) \subseteq \mathcal{C}$.

Following the proof blueprint presented in Section 3, with Assumptions 4.1 and 4.2, we derive two scaling inequalities in Lemmas B.2 and B.4 to prove the accelerated convergence rate for FW with open-loop step-size rules below, which can be thought of as the non-asymptotic version of Proposition 2.2 in Bach (2021).

Theorem 4.3 (Optimal solution in the interior of a face of $\mathcal{C}$)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be an α_f-strongly convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and suppose that there exist $\beta, \kappa > 0$ such that Assumptions 4.1 and 4.2 are satisfied. Let $M = \max_{x\in\mathcal{C}} \|\nabla f(x)\|_2$,*

$$S = \max\left\{(16L^2)/(\alpha_f^2)\,;\ (64L^3\alpha^4)/(\alpha_f^2)\right\}\,; \tag{7}$$

$T \in \mathbb{N}$, and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, it holds that

$$h_t \leq \eta_{t-2}^2 \max\left\{\frac{h_S}{\eta_{S-1}^2}, \frac{B^2}{A^2} + B, \frac{D}{\eta_S^2} + E\right\}$$

for all $t \in \{S, \dots, T\}$, where

$$A = \frac{\sqrt{\alpha_f}\beta}{2\sqrt{2}}, \qquad B = \frac{L\delta^2}{2} + \frac{\beta\sqrt{\alpha_f\beta M}}{\eta_S 2\sqrt{2}} + \frac{L\beta^2}{\eta_S 2},$$
$$D = \beta M, \qquad E = \frac{L\delta^2}{2}.$$

We make two remarks. First, we discuss the compatibility of Theorem 4.3 with the lower bound due to Jaggi (2013).

Remark 4.4 (Compatibility with lower bound from Jaggi (2013))**.** Let $\mathcal{C} = \operatorname{conv}(\{e^{(1)}, \dots, e^{(d)}, \mathbb{1}\}) \subseteq \mathbb{R}^d$. Note that the probability simplex is a face of $\mathcal{C}$. Thus, Lemma 3.7 implies that the convergence rate of FW for $\mathcal{C}$ and $f(x) = \|x\|_2^2$ is bounded from below by $\frac{1}{t} - \frac{1}{d}$ for the first $t \in \{1, \dots, d\}$ iterations and that $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ lies in the interior of an at least one-dimensional face of $\mathcal{C}$. By similar arguments as in Remark 3.8, Theorem 4.3 does not violate this lower bound, due to the dependence of S on β and δ.

In the second remark for Theorem 4.3, we discuss how to relax strict complementarity.

Remark 4.5 (Relaxation of strict complementarity)**.** In the proof of Theorem 4.3, strict complementarity is only needed to guarantee that after a specific iteration $S \in \{1, \dots, T-1\}$, for all $t \geq S$, it holds that $p_t \in \operatorname{vert}(\mathcal{C}^*)$, that is, only vertices that lie in the optimal face get returned by FW's LMO. However, strict complementarity is only a sufficient but not necessary criterion to guarantee that only vertices in the optimal face are obtained from the LMO for iterations $t \in \{S, \dots, T-1\}$: Consider, for example, the minimization of $f(x) = \frac{1}{2}\|x - b\|_2^2$ for $b = (0, 1/2, 1/2)^\intercal \in \mathbb{R}^3$ over the probability simplex $\mathcal{C} = \operatorname{conv}\{e^{(1)}, e^{(2)}, e^{(3)}\}$. Note that $\mathcal{C}^* = \operatorname{conv}\{e^{(2)}, e^{(3)}\}$. It holds that $x^* = b$ and $\nabla f(x^*) = (0, 0, 0)^\intercal \in \mathbb{R}^3$. Thus, strict complementarity is violated. However, for any $x_t = (u, v, w)^\intercal \in \mathbb{R}^3$ with $u+v+w = 1$ and $u, v, w \geq 0$, it holds, by case distinction, that either $\langle \nabla f(x_t), e^{(1)} - x_t \rangle > \min\{\langle \nabla f(x_t), e^{(2)} - x_t \rangle, \langle \nabla f(x_t), e^{(3)} - x_t \rangle\}$, or $x^* = x_t$. Thus, $p_t \in \mathcal{C}^*$ for all $t \geq 0$ without strict complementarity being satisfied.

5 KERNEL HERDING

In this section, we answer the following unexplained phenomenon observed in Bach et al. (2012):

In the kernel-herding setting of Figure 3 in Section 5.1 of Bach et al. (2012), why does FW with open-loop step-size rules converge at a rate of $\mathcal{O}(1/t^2)$?

5.1 Kernel Herding And The Frank-Wolfe Algorithm

Kernel herding is equivalent to solving a quadratic optimization problem in a *reproducing kernel Hilbert space* (RKHS) with FW. To describe this application of FW, we use the following notation: Let $\mathcal{Y} \subseteq \mathbb{R}$ be an observation space, $\mathcal{H}$ a RKHS with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and $\Phi \colon \mathcal{Y} \to \mathcal{H}$ the feature map associating a real function on $\mathcal{Y}$ to any element of $\mathcal{H}$ via $x(y) = \langle x, \Phi(y) \rangle_{\mathcal{H}}$ for $x \in \mathcal{H}$ and $y \in \mathcal{Y}$. The positive-definite kernel associated with Φ is denoted by $k \colon (y, z) \mapsto k(y, z) = \langle \Phi(y), \Phi(z) \rangle_{\mathcal{H}}$ for $y, z \in \mathcal{Y}$. In kernel herding, the feasible region is usually the *marginal polytope* $\mathcal{C}$, the convex hull of all functions $\Phi(y)$ for $y \in \mathcal{Y}$, that is, $\mathcal{C} = \operatorname{conv}(\{\Phi(y) \mid y \in \mathcal{Y}\}) \subseteq \mathcal{H}$. We consider a fixed probability distribution p over $\mathcal{Y}$ and denote the associated mean element by

$$\mu = \mathbb{E}_{p(y)} \Phi(y) \in \mathcal{C},$$

where $\mu \in \mathcal{C}$ follows from the fact that the support of p is contained in $\mathcal{Y}$. In Bach et al. (2012), kernel herding was shown to be equivalent to solving the following optimization problem with FW and step-size $\eta_t = \frac{1}{t+1}$:

$$\min_{x \in \mathcal{C}} f(x), \tag{OPT-KH}$$

where $f(x) := \frac{1}{2} \|x - \mu\|_{\mathcal{H}}^2$. This equivalence led to the study of FW (variants) with other step-size rules to solve (OPT-KH), see, for example, Bach et al. (2012); Chen et al. (2012); Lacoste-Julien et al. (2015); Tsuji et al. (2022). Under the assumption that $\|\Phi(y)\|_{\mathcal{H}} = R$ for some constant $R > 0$ and all $y \in \mathcal{Y}$, the herding procedure is well-defined and all extreme points of $\mathcal{C}$ are of the form $\Phi(y)$ for $y \in \mathcal{Y}$ (Bach et al., 2012). Thus, the linear minimization oracle (LMO) in FW always returns an element of the form $\Phi(y) \in \mathcal{C}$ for $y \in \mathcal{Y}$. Hence, FW constructs iterates of the form $x_t = \sum_{i=1}^{t} v_i \Phi(y_i)$, where $v = (v_1, \dots, v_t)^\intercal$ is a weight vector, that is, $\sum_{i=1}^{t} v_i = 1$ and $v_i \geq 0$ for all $i \in \{1, \dots, t\}$, and x_t corresponds to an empirical distribution $\tilde{p}_t$ over $\mathcal{Y}$ with empirical mean

$$\tilde{\mu}_t = \mathbb{E}_{\tilde{p}_t(y)} \Phi(y) = \sum_{i=1} v_i \Phi(y_i) = x_t \in \mathcal{C}.$$

Then, according to Bach et al. (2012),

$$\sup_{x \in \mathcal{H}, \|x\|_{\mathcal{H}} = 1} |\mathbb{E}_{p(y)} x(y) - \mathbb{E}_{\tilde{p}_t(y)} x(y)| = \|\mu - \tilde{\mu}_t\|_{\mathcal{H}}.$$

Thus, a bound on $\|\mu - \tilde{\mu}_t\|_{\mathcal{H}}$ implies control on the error in computing the expectation for all $x \in \mathcal{H}$ such that $\|x\|_{\mathcal{H}} = 1$. In kernel herding, since the objective function is a quadratic, line-search and short-step are identical.

5.2 Explaining The Phenomenon In Bach et al. (2012)

We briefly recall the infinite-dimensional kernel-herding setting of Section 5.1 in Bach et al. (2012), see also Section 2.1 in Wahba (1990). Let $\mathcal{Y} = [0, 1]$ and

$$\mathcal{H} = \left\{ x \colon [0;1] \to \mathbb{R} \mid x^{(0)}(y) \in L^2([0;1]); \; x(y) = \sum_{j=1} (a_j \cos(2\pi j y) + b_j \sin(2\pi j y)); a_j; b_j \in \mathbb{R} \right\} : \tag{8}$$

For $w, x \in \mathcal{H}$, $\langle w, x \rangle_{\mathcal{H}} := \int_{[0,1]} w'(y) x'(y) dy$ defines an inner product and $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ is a Hilbert space. Moreover, $\mathcal{H}$ is also a RKHS and for $y, z \in [0, 1]$, $\mathcal{H}$ has the reproducing kernel

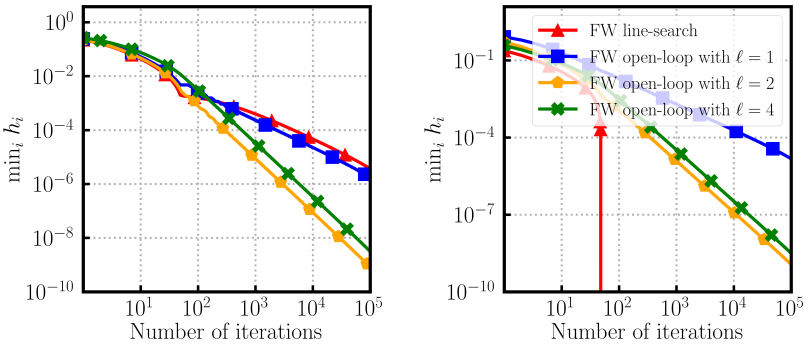
$$\begin{aligned} k(y, z) &= \sum_{j=1} \frac{2}{(2\pi j)^2} \cos(2\pi j (y - z)) \\ &= \frac{1}{2} B_2(y - z - \lfloor y - z \rfloor) \\ &= \frac{1}{2} B_2([y - z]), \qquad \text{(Bernoulli-kernel)} \end{aligned}$$

where for $y \in \mathbb{R}$, $[y] := y - \lfloor y \rfloor$, and $B_2(y) = y^2 - y + \frac{1}{6}$ is a *Bernoulli polynomial*. In the right plot of Figure 3 in Bach et al. (2012), kernel herding on $[0, 1]$ and Hilbert space $\mathcal{H}$ is considered for the uniform density $p(y) := 1$ for all $y \in [0, 1]$. Then, for all $z \in [0, 1]$, we have

$$\begin{aligned} \mu(z) &= \int_{[0,1]} k(z, y) p(y) dy \\ &= \int_{[0,1]} \sum_{j=1} \frac{2}{(2\pi j)^2} \cos(2\pi j (z - y)) \cdot 1 dy \\ &= \sum_{j=1} 0 \\ &= 0, \end{aligned}$$

where the integral and the sum can be interchanged due to the theorem of Fubini, see, for example, Royden and Fitzpatrick (1988). For the remainder of this section, we assume that $p(y) = 1$ and, thus, $\mu(y) = 0$ for all $y \in [0, 1]$. Thus, $f(x) = \frac{1}{2} \|x\|_{\mathcal{H}}^2$. For this setting, Bach et al. (2012) observed empirically that FW with open-loop step-size $\eta_t = \frac{1}{t+1}$ converges at a rate of $\mathcal{O}(1/t^2)$, whereas FW with line-search converges at a rate of $\mathcal{O}(1/t)$, see Figure 2a. The theorem below explains the accelerated convergence rate for FW with open-loop step-size $\eta_t = \frac{1}{t+1}$.

Theorem 5.1 (Kernel herding)**.** *Let $\mathcal{H}$ be the Hilbert space defined in (8), let $k \colon \mathbb{R} \times \mathbb{R} \to \mathcal{H}$ be the kernel defined in (Bernoulli-kernel), let $\Phi \colon [0, 1] \to \mathcal{H}$ be the feature map associated with k restricted to $[0, 1] \times [0, 1]$, let $\mathcal{C} = \operatorname{conv}(\{\Phi(y) \mid y \in [0, 1]\})$ be the marginal polytope, and let $\mu = 0$ such that $f(x) = \frac{1}{2} \|x\|_{\mathcal{H}}^2$. Let $T \in \mathbb{N}$*



(a) Line-search: sublinear convergence.

(b) Line-search: linear convergence.

Figure 1: Solving (OPT) with optimum in the interior of an at least one-dimensional face of the feasible region for $\mathcal{C} \subseteq \mathbb{R}^{100}$ the probability simplex and $f(x) = \frac{1}{2}\|x - \rho\bar{\mathbb{1}}\|_2^2$, where $\rho \in \{\frac{1}{4}, 2\}$, Figure 1a and 1b, respectively. In the setting of the plots, FW with short-step is identical to FW with line-search and, thus, omitted. To avoid the oscillating behavior of the primal gap, the y-axis represents $\min_{i\in\{1,\dots,t\}} h_i$, where t denotes the number of iterations and h_i the primal gap. Figure 1a illustrates that there exist problem settings for which FW with open-loop step-size rules converges faster than FW with line-search or short-step. In Figure 1b, FW with line-search solves the problem exactly after $|\operatorname{supp}(x^*)|$ iterations.

and $\eta_t = \frac{1}{t+1}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t and the LMO satisfying Assumption C.2 (a tie-breaking rule), it holds that $f(x_t) = 1/(24t^2)$ for all $t \in \{1, \dots, T\}$ such that $t = 2^m$ for $m \in \mathbb{N}$.

The proof of Theorem 5.1 implies that the iterates of FW with open-loop step-size $\eta_t = \frac{1}{t+1}$ are identical to the Sobol sequence at any iteration $t = 2^m$, where $m \in \mathbb{N}$. The Sobol sequence is known to converge at the optimal rate of $\mathcal{O}(1/t^2)$ (Bach et al., 2012) in this infinite-dimensional kernel-herding setting. Here, the equivalence of FW with kernel herding leads to the study and discovery of new convergence rates for FW. This is in contrast to other papers (Chen et al., 2012; Bach et al., 2012; Tsuji et al., 2022) in which FW is exploited to improve kernel-herding methods.

6 NUMERICAL EXPERIMENTS

In this section, we present the numerical experiments. All of our numerical experiments are implemented in PYTHON and performed on an NVIDIA GeForce RTX 2060 GPU with 6GB RAM and an Intel Core i7-9750H CPU at 2.60GHz with 16 GB RAM. Our code is publicly available on GitHub.

6.1 Optimum In The Relative Interior Of An At Least One-dimensional Face Of A Polytope

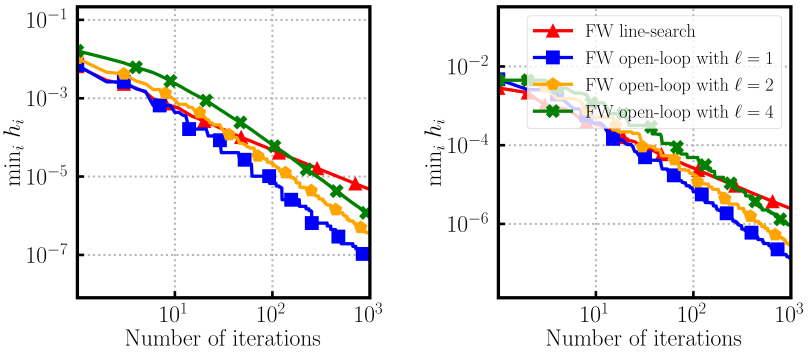
In this section, we compare the convergence rates of FW with open-loop step-size rules and line-search when the optimum lies in the relative interior of an at least one-dimensional face of a polytope and the objective is strongly convex. These experiments correspond to the setting of Section 4.

6.1.1 Setup

For $d = 100$, we address (OPT) with FW with different step-sizes for $\mathcal{C} \subseteq \mathbb{R}^d$ the probability simplex and $f(x) = \frac{1}{2}\|x - \rho\bar{\mathbb{1}}\|_2^2$, where $\rho \geq \frac{2}{d}$ and $\bar{\mathbb{1}}$ is the vector with zeros for the first $\lceil d/2 \rceil$ entries and ones for the remaining entries. Then, $\frac{2}{d}\bar{\mathbb{1}} = x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$ is the unique minimizer of f. For $\rho \in \{\frac{1}{4}, 2\}$, we compare FW with line-search and open-loop step-sizes $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \{1, 2, 4\}$ starting with $x_0 = e^{(1)}$. The algorithms run for 100,000 iterations. In this setting, short-step is identical to line-search and, thus, omitted. We plot the results of the experiments in log-log plots in Figure 1. To avoid the oscillating behavior of the primal gap, the y-axis represents $\min_{i\in\{1,\dots,t\}} h_i$, where t denotes the number of iterations and h_i the primal gap.

6.1.2 Results

For $\rho \in \{\frac{1}{4}, 2\}$, that is, in Figures 1a and 1b, FW with open-loop step-size rules converges at a rate of $\mathcal{O}(1/t^2)$ whereas FW with open-loop step-size $\eta_t = \frac{1}{t+1}$ converges at a rate of $\mathcal{O}(1/t)$. For $\rho \in \{\frac{1}{4}, 2\}$, that is, in Figures 1a and 1b, FW with line-search converges at a rate of $\Omega(1/t)$ and linearly, respectively. In Figure 1b, FW with line-search solves the problem exactly after $|\operatorname{supp}(x^*)|$ iterations.



(a) Uniform density.

(b) Non-uniform density.

Figure 2: Solving (OPT-KH) for the setting presented in Section 5.2 with uniform and non-uniform densities, Figures 2a and 2b, respectively. In kernel herding, FW with short-step is identical to FW with line-search and, thus, omitted. To avoid the oscillating behavior of the primal gap, the y-axis represents $\min_{i \in \{1,\dots,t\}} h_i$, where t denotes the number of iterations and h_i the primal gap. Both for the uniform and non-uniform density, FW with open-loop step-size rules converges at a rate of $\mathcal{O}(1/t^2)$ whereas FW with line-search converges at a rate of $\mathcal{O}(1/t)$.

6.2 Kernel Herding

In this section, we compare the convergence rates of FW with open-loop step-size rules and line-search for various kernel-herding problems. These experiments correspond to the setting of Section 5.

6.2.1 Setup

Consider the kernel-herding setting of Section 5.2 over $\mathcal{Y} = [0, 1]$. Given either the uniform density or a random non-uniform density of the form $p(y) \propto \left(\sum_{i=1}^{n}(a_i \cos(2\pi i y) + b_i \sin(2\pi i y))\right)^2$ with $a_i, b_i \in \mathbb{R}$ and $n \leq 5$ such that $\int_{[0,1]} p(y)dy = 1$, we address (OPT-KH) with FW with line-search and open-loop step-sizes $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \{1, 2, 4\}$. The LMO is implemented as an exhaustive search over $[0, 1]$ and run for 1,000 iterations and the algorithms run for 1,000 iterations. We plot the results of the experiments in log-log plots in Figure 2. To avoid the oscillating behavior of the primal gap, the y-axis represents $\min_{i \in \{1,\dots,t\}} h_i$, where t denotes the number of iterations and h_i the primal gap.

6.2.2 Results

For both settings, FW with open open-loop step-size rules converges at a rate of $\mathcal{O}(1/t^2)$, whereas FW with line-search converges at a rate of $\mathcal{O}(1/t)$.

7 DISCUSSION

The central motivation for studying FW with open-loop step-size rules was the unexplained phenomenon in Bach et al. (2012), a problem in the foundational area of kernel herding. To study the mechanisms at large, we revisited with open-loop step-size rules the regimes where FW algorithms are known to enjoy accelerated convergence rates with line-search or short-step, even those that do not correspond to kernel herding directly. For these settings, we derived accelerated convergence rates for FW with open-loop step-size rules, that, in combination with our kernel-herding analysis characterize the acceleration of FW with open-loop step-size rules. The analysis of non-kernel-herding settings also culminated in the characterization of a setting in which FW with open-loop step-size rules converges faster than FW with line-search or short-step, see Theorem 4.3 and the lower bound due to Wolfe (1970). Despite closing gaps in our understanding of FW, open questions remain:

1. FW is a famously affine-invariant algorithm. Our results, however, rely on norms, which are not affine-invariant. It remains an open problem to restate the accelerated convergence rates for FW with open-loop step-sizes in affine-invariant form.
2. Theorem 5.1, our kernel-herding result in Section 5, is limited to the uniform density whereas numerical experiments in Section 6.2 suggest that a result similar to Theorem 5.1 could hold for non-uniform densities. Future research might be able to address this discrepancy between theory and practice.

Acknowledgements

This research was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany´s Excellence Strategy – The Berlin Mathematics Research Center MATH$^+$ (EXC-2046/1, project ID 390685689, BMS Stipend).

References

Bach, F. (2021). On the effectiveness of richardson extrapolation in data science. *SIAM Journal on Mathematics of Data Science*, 3(4):1251–1277.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the International Conference on Machine Learning*, pages 1355–1362. PMLR.

Bertsekas, D. P. (1982). Projected newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20(2):221–246.

Birgin, E. G. and Martínez, J. M. (2002). Large-scale active-set box-constrained optimization method with spectral projected gradients. *Computational Optimization and Applications*, 23(1):101–125.

Bomze, I. M., Rinaldi, F., and Bulo, S. R. (2019). First-order methods for the impatient: Support identification in finite time with convergent frank–wolfe variants. *SIAM Journal on Optimization*, 29(3):2211–2226.

Bomze, I. M., Rinaldi, F., and Zeffiro, D. (2020). Active set complexity of the away-step frank–wolfe algorithm. *SIAM Journal on Optimization*, 30(3):2470–2500.

Braun, G., Pokutta, S., Tu, D., and Wright, S. (2019). Blended conditional gradients. In *Proceedings of the International Conference on Machine Learning*, pages 735–743. PMLR.

Canon, M. D. and Cullum, C. D. (1968). A tight upper bound on the rate of convergence of frank-wolfe algorithm. *SIAM Journal on Control*, 6(4):509–516.

Chen, Y., Welling, M., and Smola, A. (2012). Super-samples from kernel herding. *arXiv preprint arXiv:1203.3472*.

Combettes, C. and Pokutta, S. (2020). Boosting frank-wolfe by chasing gradients. In *Proceedings of the International Conference on Machine Learning*, pages 2111–2121. PMLR.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9):1853–1865.

Demianov, V. F. and Rubinov, A. M. (1970). *Approximate methods in optimization problems*. Number 32. Elsevier Publishing Company.

Dunn, J. C. (1979). Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. *SIAM Journal on Control and Optimization*, 17(2):187–211.

Dunn, J. C. and Harshbarger, S. (1978). Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110.

Garber, D. (2020). Revisiting frank-wolfe for polytopes: Strict complementarity and sparsity. volume 33, pages 18883–18893.

Garber, D. and Hazan, E. (2015). Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Proceedings of the International Conference on Machine Learning*. PMLR.

Garber, D. and Hazan, E. (2016). A linearly convergent variant of the conditional gradient algorithm under strong convexity, with applications to online and stochastic optimization. *SIAM Journal on Optimization*, 26(3):1493–1528.

Giesen, J., Jaggi, M., and Laue, S. (2012). Optimizing over the growing spectrahedron. In *European Symposium on Algorithms*, pages 503–514. Springer.

Guélat, J. and Marcotte, P. (1986). Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119.

Hager, W. W. and Zhang, H. (2006). A new active set algorithm for box constrained optimization. *SIAM Journal on Optimization*, 17(2):526–557.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the International Conference on Machine Learning*, pages 427–435. PMLR.

Joulin, A., Tang, K., and Fei-Fei, L. (2014). Efficient image and video co-localization with frank-wolfe algorithm. In *Proceedings of the European Conference on Computer Vision*, pages 253–268, Cham. Springer International Publishing.

Kerdreux, T., d'Aspremont, A., and Pokutta, S. (2021a). Local and global uniform convexity conditions. *arXiv preprint arXiv:2102.05134*.

Kerdreux, T., d'Aspremont, A., and Pokutta, S. (2021b). Projection-free optimization on uniformly convex sets. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 19–27. PMLR.

Kerdreux, T., Liu, L., Lacoste-Julien, S., and Scieur, D. (2021c). Affine invariant analysis of frank-wolfe on

strongly convex sets. In *Proceedings of the International Conference on Machine Learning*, pages 5398–5408. PMLR.

Lacoste-Julien, S. and Jaggi, M. (2013). An affine invariant linear convergence analysis for Frank-Wolfe algorithms. *arXiv preprint arXiv:1312.7864*.

Lacoste-Julien, S. and Jaggi, M. (2015). On the global linear convergence of Frank-Wolfe optimization variants. In *Proceedings of Advances in Neural Information Processing Systems*, pages 496–504.

Lacoste-Julien, S., Lindsten, F., and Bach, F. (2015). Sequential kernel herding: Frank-wolfe optimization for particle filtering. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 544–552. PMLR.

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

Levitin, E. S. and Polyak, B. T. (1966). Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50.

Li, B., Coutino, M., Giannakis, G. B., and Leus, G. (2021). A momentum-guided frank-wolfe algorithm. *IEEE Transactions on Signal Processing*, 69:3597–3611.

Pedregosa, F., Askari, A., Negiar, G., and Jaggi, M. (2018). Step-size adaptivity in projection-free optimization. *arXiv preprint arXiv:1806.05123*.

Pena, J. (2021). Affine invariant convergence rates of the conditional gradient method. *arXiv preprint arXiv:2112.06727*.

Ravi, S. N., Dinh, T., Lokhande, V., and Singh, V. (2018). Constrained deep learning using conditional gradient and applications in computer vision. *arXiv preprint arXiv:1803.06453*.

Royden, H. L. and Fitzpatrick, P. (1988). *Real Analysis*, volume 32. Macmillan New York.

Tsuji, K. K., Tanaka, K., and Pokutta, S. (2022). Pairwise conditional gradients without swap steps and sparser kernel herding. In *Proceedings of the International Conference on Machine Learning*, pages 21864–21883. PMLR.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wolfe, P. (1970). Convergence theory in nonlinear programming. *Integer and Nonlinear Programming*, pages 1–36.

A MISSING PROOFS FROM SECTION 3

Proof of Proposition 3.1. In the literature, the proof is usually done by induction (Jaggi, 2013). Here, for convenience and as a brief introduction for things to come, we proceed with a direct approach. Since $\eta_0 = 1$, by L-smoothness, we have $h_1 \leq \frac{L\delta^2}{2}$. Let $t \in \{1, \dots, T-1\}$. By optimality of p_t and convexity of f, $\langle \nabla f(x_t), x_t - p_t \rangle \geq \langle \nabla f(x_t), x_t - x^* \rangle \geq h_t$. Plugging this bound into (Progress-Bound) and with $\|x_t - p_t\|_2 \leq \delta$, it holds that

$$\begin{aligned} h_{t+1} &\leq (1-\eta_t)h_t + \eta_t^2 \frac{L\|x_t - p_t\|_2^2}{2} \\ &\leq (1-\eta_t) \quad (1-\eta_{t-1})h_{t-1} + \eta_{t-1}^2 \frac{L\delta^2}{2} \quad + \eta_t^2 \frac{L\delta^2}{2} \\ &\leq \prod_{i=1}^{t} (1-\eta_i)h_1 + \frac{L\delta^2}{2} \sum_{i=1}^{t} \eta_i^2 \prod_{j=i+1}^{t} (1-\eta_j) \qquad (9) \\ &\leq \frac{L\delta^2}{2} \quad \frac{4!}{(t+1)\cdots(t+4)} + \sum_{i=1}^{t} \frac{4^2}{(i+4)^2} \frac{(i+1)\cdots(i+4)}{(t+1)\cdots(t+4)} \Bigg) \qquad (10) \\ &\leq 8L\delta^2 \quad \frac{1}{(t+4-1)(t+4)} + \frac{t}{(t+4-1)(t+4)} \\ &\leq \frac{8L\delta^2}{t+4}, \end{aligned}$$

where for the third inequality, we use that

$$\prod_{j=i+1}^{t} (1-\eta_j) = \prod_{j=i+1}^{t} \frac{j}{j+4} = \frac{(i+1)(i+2)\cdots t}{(i+5)(i+6)\cdots(t+4)} = \frac{(i+1)(i+2)(i+3)(i+4)}{(t+1)(t+2)(t+3)(t+4)}.$$

□

Proof of Lemma 3.3. Let $t \in \{S, \dots, T\}$, where S is as in (2). Then, by (HEB) and Proposition 3.1, it holds that

$$\|x_t - x^*\|_2 \leq \mu h_t^\theta \leq \mu \left(\frac{8L\delta^2}{t+3} \right)^\theta \leq \mu \left(\frac{8L\delta^2}{8L\delta^2 \frac{\mu}{\beta}} \right)^{1/\theta} \leq \beta.$$

□

Proof of Lemma 3.4. The statement holds for $x = x^*$. For $x \in \mathcal{C} \setminus \{x^*\}$, by convexity and (HEB),

$$f(x) - f(x^*) \leq \frac{\langle \nabla f(x), x - x^* \rangle}{\|x - x^*\|_2} \|x - x^*\|_2 \leq \frac{\langle \nabla f(x), x - x^* \rangle}{\|x - x^*\|_2} \mu (f(x) - f(x^*))^\theta.$$

Dividing by $\mu(f(x) - f(x^*))^\theta$, which cannot be equal to zero by the assumption that x is not optimal, yields (Scaling-HEB). □

Proof of Lemma 3.5. For all $t \in \{S, \dots, T\}$, we first prove that

$$h_t \leq \max \left(\left(\frac{\eta_{t-2}\eta_{t-1}}{\eta_{S-2}\eta_{S-1}} \right)^{1/(2(1-\psi))} h_S, \left(\frac{\eta_{t-2}\eta_{t-1}B^2}{A^2} \right)^{1/(2(1-\psi))} + \eta_{t-2}\eta_{t-1}BC \right), \qquad (11)$$

which then implies (5). The proof is a straightforward modification of Footnote 3 in the proof of Proposition 2.2 in Bach (2021) and is by induction. The base case of (11) with $t = S$ is immediate, even if $S = 1$, as $\eta_{-1} \geq \eta_0 = 1$. Suppose that (11) is correct for a specific iteration $t \in \{S, \dots, T-1\}$. We distinguish between two cases.

First, suppose that $h_t \leq (\frac{\eta_t B}{A})^{1/(1-\psi)}$. Plugging this bound into (4), we obtain (11) at iteration $t+1$:

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - 0 + \eta_t^2 BC_t \leq \left(\frac{\eta_t B}{A}\right)^{1/(1-\psi)} + \eta_t^2 BC \leq \left(\frac{\eta_{t-1}\eta_t B^2}{A^2}\right)^{1/(2(1-\psi))} + \eta_{t-1}\eta_t BC.$$

Next, suppose that $h_t \geq (\frac{\eta_t B}{A})^{1/(1-\psi)}$. Plugging this bound on h_t into (4) and using the induction assumption (11) at iteration t yields

$$\begin{aligned}
h_{t+1} &\leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t A C_t \frac{\eta_t B}{A} + \eta_t^2 BC_t \\
&= \frac{t+2}{t+4} h_t \\
&= \frac{\eta_t}{\eta_{t-2}} h_t \\
&\leq \frac{\eta_t}{\eta_{t-2}} \max\left(\left(\frac{\eta_{t-2}\eta_{t-1}}{\eta_{S-2}\eta_{S-1}}\right)^{1/(2(1-\psi))} h_S, \left(\frac{\eta_{t-2}\eta_{t-1}B^2}{A^2}\right)^{1/(2(1-\psi))} + \eta_{t-2}\eta_{t-1}BC \right) \\
&\leq \max\left(\left(\frac{\eta_{t-1}\eta_t}{\eta_{S-2}\eta_{S-1}}\right)^{1/(2(1-\psi))} h_S, \left(\frac{\eta_{t-1}\eta_t B^2}{A^2}\right)^{1/(2(1-\psi))} + \eta_{t-1}\eta_t BC \right),
\end{aligned}$$

where the last inequality holds due to $\frac{\eta_t}{\eta_{t-2}^{\ 2}}(\eta_{t-2}\eta_{t-1})^{1/(2(1-\psi))} \leq (\eta_{t-1}\eta_t)^{1/(2(1-\psi))}$ for $\frac{\eta_t}{\eta_{t-2}^{\ 2}} \in [0,1]$ and $1/(2(1-\psi)) \in [1/2, 1]$. In either case, (11) is satisfied for $t+1$, proving the lemma. □

Proof of Theorem 3.6. Let $t \in \{S, \dots, T-1\}$, where S is as in (6). By Lemma 3.3, $\|x_t - x^*\|_2 \leq \beta/2$ and, by triangle inequality, we have $\|x_t - p_t\|_2 \geq \beta/2$. Thus, for all $t \in \{S, \dots, T\}$, it follows that (3) holds. This inequality allows us to apply Lemma 3.5 with $A = \frac{\beta^2}{4\mu\delta}$, $B = \frac{L\delta^2}{2}$, $C = 1$, $C_t = 1$ for all $t \in \{S, \dots, T-1\}$, and $\psi = \theta$, resulting in

$$h_t \leq \max\left(\left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^{1/(1-\theta)} h_S, \left(\frac{\eta_{t-2} 2\mu L \delta^3}{\beta^2}\right)^{1/(1-\theta)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right)$$

for all $t \in \{S, \dots, T\}$. □

B MISSING PROOFS FROM SECTION 4

The proof of Theorem 4.3 follows the proof blueprint of Section 3, that is, is built on two scaling inequalities, which hold when Assumptions 4.1 and 4.2 are satisfied. In the proof of Theorem 5 in Garber (2020), the authors showed that there exists an iterate $S \in \mathbb{N}$ such that for all $t \geq S$, the FW vertices p_t lie in the optimal face, assuming that the objective function is strongly convex. Below, we generalize their result to convex functions satisfying (HEB).

Lemma B.1 (Active set identification)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function satisfying a (μ, θ)-(HEB) for some $\mu > 0$ and $\theta \in]0, 1/2]$ with unique minimizer $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$, and suppose that there exists $\kappa > 0$ such that Assumption 4.1 is satisfied. Let*

$$S = \left\lceil 8L\delta^2 \left(\frac{2\mu L \delta}{\kappa}\right)^{1/\theta} \right\rceil, \tag{12}$$

$T \in \mathbb{N}$, and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with step-size η_t, it holds that $p_t \in \operatorname{vert}(\mathcal{C}^*)$ for all $t \in \{S, \dots, T-1\}$.

Proof. Let $t \in \{S, \dots, T-1\}$, where S is as in (12). Note that in Line 3 of Algorithm 1, $p_t \in \operatorname{argmin}_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle$ can always be chosen such that $p_t \in \operatorname{argmin}_{p \in \operatorname{vert}(\mathcal{C})} \langle \nabla f(x_t), p - x_t \rangle$. For $p \in \operatorname{vert}(\mathcal{C})$, it holds that

$$\begin{aligned}
\langle \nabla f(x_t), p - x_t \rangle &= \langle \nabla f(x_t) - \nabla f(x^*) + \nabla f(x^*), p - x^* + x^* - x_t \rangle \\
&= \langle \nabla f(x_t) - \nabla f(x^*), p - x_t \rangle + \langle \nabla f(x^*), p - x^* \rangle + \langle \nabla f(x^*), x^* - x_t \rangle.
\end{aligned} \tag{13}$$

We distinguish between vertices $p \in \operatorname{vert}(\mathcal{C}) \setminus \mathcal{C}^*$ and vertices $p \in \operatorname{vert}(\mathcal{C}^*)$. First, suppose that $p \in \operatorname{vert}(\mathcal{C}) \setminus \mathcal{C}^*$. Using strict complementarity, Cauchy-Schwarz, L-smoothness, and (HEB) to bound (13) yields

$$\begin{aligned}\langle \nabla f(x_t), p - x_t \rangle &\geq -\|\nabla f(x_t) - \nabla f(x^*)\|_2 \|p - x_t\|_2 + \kappa + \langle \nabla f(x^*), x^* - x_t \rangle \\ &\geq \kappa - L\delta \|x_t - x^*\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle \\ &\geq \kappa - \mu L \delta h_t^\theta + \langle \nabla f(x^*), x^* - x_t \rangle.\end{aligned}$$

Next, suppose that $p \in \operatorname{vert}(\mathcal{C}^*)$. Using strict complementarity, Cauchy-Schwarz, L-smoothness, and (HEB) to bound (13) yields

$$\begin{aligned}\langle \nabla f(x_t), p - x_t \rangle &\leq \|\nabla f(x_t) - \nabla f(x^*)\|_2 \|p - x_t\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle \\ &\leq L\delta \|x_t - x^*\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle \\ &\leq \mu L \delta h_t^\theta + \langle \nabla f(x^*), x^* - x_t \rangle.\end{aligned}$$

By Proposition 3.1, it holds that

$$\mu L \delta h_t^\theta \leq \mu L \delta h_S^\theta \leq \mu L \delta \left(\frac{8L\delta^2}{8L\delta^2 \left(\frac{2\mu L \delta}{\kappa}\right)^{1/\theta} + 3} \right)^\theta < \frac{\kappa}{2}.$$

Hence, for $t \in \{S, \ldots, T-1\}$,

$$\langle \nabla f(x_t), p - x_t \rangle = \begin{cases} > \frac{\kappa}{2} + \langle \nabla f(x^*), x^* - x_t \rangle, & \text{if } p \in \operatorname{vert}(\mathcal{C}) \setminus \mathcal{C}^* \\ < \frac{\kappa}{2} + \langle \nabla f(x^*), x^* - x_t \rangle, & \text{if } p \in \operatorname{vert}(\mathcal{C}^*). \end{cases}$$

Then, by optimality of p_t, for all iterations $t \in \{S, \ldots, T-1\}$ of Algorithm 1, it holds that $p_t \in \operatorname{vert}(\mathcal{C}^*)$. □

Using Assumption 4.2, Bach (2021) derived the following scaling inequality, a variation of (Scaling-INT).

Lemma B.2 (Bach, 2021)**.** *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope, let $f \colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$, and suppose that there exists $\beta > 0$ such that Assumption 4.2 is satisfied. Then, for all $x \in \mathcal{C}$ such that $p \in \operatorname{argmin}_{v \in \mathcal{C}} \langle \nabla f(x), v \rangle \subseteq \mathcal{C}^*$, it holds that*

$$\langle \nabla f(x), x - p \rangle \geq \beta \|\Pi \nabla f(x)\|_2, \tag{Scaling-BOR}$$

where Πx denotes the orthogonal projection of $x \in \mathbb{R}^d$ onto the span of $\{x^* - p \mid p \in \mathcal{C}^*\}$.

Proof. Suppose that $x \in \mathcal{C}$ such that $p \in \operatorname{argmin}_{v \in \mathcal{C}} \langle \nabla f(x), v \rangle \subseteq \mathcal{C}^*$. Then,

$$\begin{aligned}\langle \nabla f(x), x - p \rangle &= \max_{v \in \mathcal{C}} \langle \nabla f(x), x - v \rangle \\ &\geq \langle \nabla f(x), x - x^* \rangle + \quad \nabla f(x), \beta \frac{\Pi \nabla f(x)}{\|\Pi \nabla f(x)\|_2} \\ &= \langle \nabla f(x), x - x^* \rangle + \quad \Pi \nabla f(x) + (\mathrm{I} - \Pi) \nabla f(x), \beta \frac{\Pi \nabla f(x)}{\|\Pi \nabla f(x)\|_2} \\ &= \langle \nabla f(x), x - x^* \rangle + \beta \|\Pi \nabla f(x)\|_2 \\ &\geq \beta \|\Pi \nabla f(x)\|_2,\end{aligned}$$

where the first equality follows from the construction of $p \in \operatorname{argmin}_{v \in \mathcal{C}} \langle \nabla f(x), v \rangle$, the first inequality follows from the fact that the maximum is at least as large as the maximum attained on $B_\beta(x^*) \cap \mathcal{C}^*$, the second equality follows from the definition of the orthogonal projection, the third equality follows from the fact that Πx and $(\mathrm{I} - \Pi)x$ are orthogonal for any $x \in \mathbb{R}^d$, and the second inequality follows from the convexity of f. □

We next bound the distance between x_t and the optimal face $\mathcal{C}^*$.

Lemma B.3 (Distance to optimal face). *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function satisfying a (μ, θ)-(HEB) for some $\mu > 0$ and $\theta \in]0, 1/2]$ with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and suppose that there exist $\beta, \kappa > 0$ such that Assumptions 4.1 and 4.2 are satisfied. Let*

$$S = \max\left(8L\delta^2 \left(\frac{\mu}{\beta}\right)^{1/\theta}, 8L\delta^2 \left(\frac{2\mu L\delta}{\kappa}\right)^{1/\theta}\right), \tag{14}$$

$T \in \mathbb{N}$, and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, it holds that

$$\|(I - \Pi)(x_t - x^*)\|_2 \leq \frac{\eta_t^4}{\eta_S^4}\beta \tag{15}$$

for all $t \in \{S, \dots, T-1\}$, where Πx denotes the orthogonal projection of $x \in \mathbb{R}^d$ onto the span of $\{x^* - p \mid p \in \mathcal{C}^*\}$.

Proof. Let $t \in \{S, \dots, T-1\}$, where S is as in (14). By Lemma B.1, $p_t \in \operatorname{vert}(\mathcal{C}^*)$. Thus, $(\mathrm{I} - \Pi)(p_t - x^*) = \mathbb{0}$ and

$$\begin{aligned}
(\mathrm{I} - \Pi)(x_{t+1} - x^*) &= (1 - \eta_t)(\mathrm{I} - \Pi)(x_t - x^*) + \eta_t(\mathrm{I} - \Pi)(p_t - x^*) \\
&= (1 - \eta_t)(\mathrm{I} - \Pi)(x_t - x^*) \\
&= \prod_{i=S}^{t}(1 - \eta_i)(\mathrm{I} - \Pi)(x_S - x^*) \\
&= \frac{S(S+1)\cdots t}{(S+4)(S+5)\cdots(t+4)}(\mathrm{I} - \Pi)(x_S - x^*) \\
&= \frac{S(S+1)(S+2)(S+3)}{(t+1)(t+2)(t+3)(t+4)}(\mathrm{I} - \Pi)(x_S - x^*).
\end{aligned}$$

Hence,

$$\begin{aligned}
\|(I - \Pi)(x_{t+1} - x^*)\|_2 &\leq \frac{S(S+1)(S+2)(S+3)}{(t+1)(t+2)(t+3)(t+4)}\|(I - \Pi)(x_S - x^*)\|_2 \\
&\leq \frac{(S+1)(S+2)(S+3)(S+4)}{(t+2)(t+3)(t+4)(t+5)}\|(I - \Pi)(x_S - x^*)\|_2 \\
&\leq \frac{\eta_{t+1}^4}{\eta_S^4}\|(I - \Pi)(x_S - x^*)\|_2 \\
&\leq \frac{\eta_{t+1}^4}{\eta_S^4}\beta,
\end{aligned}$$

where the last inequality follows from Lemma 3.3. □

We require a second scaling inequality, relying on Assumptions 4.1 and 4.2.

Lemma B.4. *Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a polytope of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be an α_f-strongly convex and L-smooth function with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$, and suppose that there exist $\beta, \kappa > 0$ such that Assumptions 4.1 and 4.2 are satisfied. Let $M = \max_{x\in\mathcal{C}} \|\nabla f(x)\|_2$,*

$$S = \max\left(\frac{16L\delta^2}{\alpha_f\beta^2}, \frac{64L^3\delta^4}{\alpha_f\kappa^2}\right), \tag{16}$$

$T \in \mathbb{N}$, and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t and $t \in \{S, \dots, T-1\}$, it holds that

$$\|\Pi\nabla f(x_t)\|_2 \geq \left(\sqrt{\frac{\alpha_f}{2}}\sqrt{h_t} - \frac{\eta_t^2}{\eta_S^2}\sqrt{\frac{\alpha_f\beta M}{2}}\right) - \frac{\eta_t^4}{\eta_S^4}L\beta \tag{Scaling-CVX}$$

or $h_t \leq \frac{\eta_t^4}{\eta_S^4}\beta M$, where Πx denotes the orthogonal projection of $x \in \mathbb{R}^d$ onto the span of $\{x^* - p \mid p \in \mathcal{C}^*\}$.

Proof. Given a vector $x \in \mathbb{R}^d$, let $\Pi_{\mathrm{aff}(\mathcal{C})}x$ denote the projection of x onto $\mathrm{aff}(\mathcal{C}^*)$, that is, $\Pi_{\mathrm{aff}(\mathcal{C})}x \in \operatorname{argmin}_{y\in\mathrm{aff}(\mathcal{C})}\|y-x\|_2$. Since $\mathrm{aff}(\mathcal{C}^*) = x^* + \mathrm{span}(\{x^* - p \mid p \in \mathcal{C}^*\})$, there has to exist some $y \in \mathbb{R}^d$ such that $\Pi_{\mathrm{aff}(\mathcal{C})}x = (I-\Pi)x^* + \Pi x + \Pi y$. By orthogonality of Π, we have

$$\|\Pi_{\mathrm{aff}(\mathcal{C})}x - x\|_2 = \|(I-\Pi)x^* - (I-\Pi)x + \Pi y\|_2 = \|(I-\Pi)x^* - (I-\Pi)x\|_2 + \|\Pi y\|_2.$$

The right-hand side is minimized when $\Pi y = \mathbb{0}$. Thus,

$$\Pi_{\mathrm{aff}(\mathcal{C})}x = (I-\Pi)x^* + \Pi x \in \operatorname*{argmin}_{y\in\mathrm{aff}(\mathcal{C})}\|y-x\|_2.$$

Let $t \in \{S, \dots, T-1\}$, where S is as defined in (16). By Lemma 3.3, $\|x_t - x^*\|_2 \leq \beta$ and, thus, by Assumption 4.2, $\Pi_{\mathrm{aff}(\mathcal{C})}x_t \in \mathcal{C}^*$. By L-smoothness of f, it holds that

$$\|\nabla f(x_t) - \nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \leq L\|x_t - \Pi_{\mathrm{aff}(\mathcal{C})}x_t\|_2 = L\|(I-\Pi)(x_t - x^*)\|_2.$$

By Lemma B.3, it then holds that

$$\|\nabla f(x_t) - \nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \leq \frac{\eta_t^4}{\eta_S^4}L\beta. \tag{17}$$

Since for any $x \in \mathbb{R}^d$, we have that $\|\Pi x\|_2 \leq \|\Pi x\|_2 + \|(I-\Pi)x\|_2 = \|x\|_2$, Inequality (17) implies that $\|\Pi\nabla f(x_t) - \Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \leq \frac{\eta_t^4}{\eta_S^4}L\beta$. Combined with the triangle inequality,

$$\|\Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \leq \|\Pi\nabla f(x_t)\|_2 + \|\Pi\nabla f(x_t) - \Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \leq \|\Pi\nabla f(x_t)\|_2 + \frac{\eta_t^4}{\eta_S^4}L\beta,$$

which we rearrange to

$$\|\Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 - \frac{\eta_t^4}{\eta_S^4}L\beta \leq \|\Pi\nabla f(x_t)\|_2. \tag{18}$$

For the remainder of the proof, we bound $\|\Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2$ from below. To do so, define the function $g\colon \mathcal{C}\cap B_\beta(x^*) \to \mathbb{R}$ via $g(x) := f(\Pi_{\mathrm{aff}(\mathcal{C})}x) = f((I-\Pi)x^* + \Pi x)$. The gradient of g at $x \in \mathcal{C}\cap B_\beta(x^*)$ is $\nabla g(x) = \Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x) = \Pi\nabla f((I-\Pi)x^* + \Pi x)$. Since f is α_f-strongly convex in $\mathcal{C}$ and $g(x) = f(x)$ for all $x \in \mathrm{aff}(\mathcal{C}^*)\cap B_\beta(x^*)$, g is α_f-strongly convex in $\mathrm{aff}(\mathcal{C}^*)\cap B_\beta(x^*)$. Since the projection onto $\mathrm{aff}(\mathcal{C}^*)$ is idempotent, $\Pi_{\mathrm{aff}(\mathcal{C})}x_t \in \mathrm{aff}(\mathcal{C}^*)\cap B_\beta(x^*)$ for all $t \in \{S, \dots, T-1\}$, and g is α_f-strongly convex in $\mathrm{aff}(\mathcal{C}^*)\cap B_\beta(x^*)$, it holds that

$$\begin{aligned}
\|\Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 &= \|\Pi\nabla f(\Pi^2_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \\
&= \|\nabla g(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 \\
&\geq \sqrt{\frac{\alpha_f}{2}}\sqrt{g(\Pi_{\mathrm{aff}(\mathcal{C})}x_t) - g(x^*)} \\
&= \sqrt{\frac{\alpha_f}{2}}\sqrt{f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t) - f(x^*)}.
\end{aligned}$$

Suppose that $h_t \geq \frac{\eta_t^4}{\eta_S^4}\beta M$. By Lemma B.3 and Cauchy-Schwarz, we obtain $h_t - \langle\nabla f(x_t), (I-\Pi)(x_t - x^*)\rangle \geq h_t - \frac{\eta_t^4}{\eta_S^4}\beta M \geq 0$. Combined with convexity of f, we have

$$\begin{aligned}
\|\Pi\nabla f(\Pi_{\mathrm{aff}(\mathcal{C})}x_t)\|_2 &\geq \sqrt{\frac{\alpha_f}{2}}\sqrt{f(x_t) + \langle\nabla f(x_t), \Pi_{\mathrm{aff}(\mathcal{C})}x_t - x_t\rangle - f(x^*)} \\
&= \sqrt{\frac{\alpha_f}{2}}\sqrt{h_t - \langle\nabla f(x_t), (I-\Pi)(x_t - x^*)\rangle} \\
&\geq \sqrt{\frac{\alpha_f}{2}}\sqrt{h_t - \frac{\eta_t^4}{\eta_S^4}\beta M}.
\end{aligned}$$

Since for $a, b \in \mathbb{R}$ with $a \geq b \geq 0$, it holds that $\sqrt{a-b} \geq \sqrt{a} - \sqrt{b}$, we obtain

$$\|\Pi \nabla f(\Pi_{\text{aff}(\mathcal{C})} x_t)\|_2 \geq \left(\sqrt{\frac{\alpha_f}{2}}\sqrt{h_t} - \sqrt{\frac{\eta_t^4}{\eta_S^4}\beta M}\right) = \sqrt{\frac{\alpha_f}{2}}\sqrt{h_t} - \frac{\eta_t^2}{\eta_S^2}\sqrt{\beta M} \quad .$$

Combining this inequality with (18), we obtain

$$\|\Pi \nabla f(x_t)\|_2 \geq \sqrt{\frac{\alpha_f}{2}}\sqrt{h_t} - \frac{\eta_t^2}{\eta_S^2}\sqrt{\frac{\alpha_f \beta M}{2}} - \frac{\eta_t^4}{\eta_S^4} L\beta.$$

□

Finally, we prove Theorem 4.3, that is, we prove that when the feasible region $\mathcal{C}$ is a polytope, the objective function f is strongly convex, and the unique minimizer $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ lies in the relative interior of an at least one-dimensional face $\mathcal{C}^*$ of $\mathcal{C}$, FW with the open-loop step-size $\eta_t = \frac{4}{t+4}$ converges at a rate of $\mathcal{O}(1/t)$ for iterations $t \leq S$ and at a non-asymptotic rate of $\mathcal{O}(1/t^2)$ for iterations $t \geq S$, where S is defined as in (16). Our result can be seen as the non-asymptotic version of Bach (2021, Proposition 2.2). Contrary to the result of Bach et al. (2012), our result is in primal gap, we no longer require bounds on the third-order derivatives, and do not have to invoke affine-invariance of FW to obtain the accelerated convergence rate.

Proof of Theorem 4.3. Let $t \in \{S, \dots, T-1\}$, where S is as in (7). Furthermore, suppose that $h_t \geq \frac{\eta_t^4}{\eta_S^4}\beta M$. Combine (9) and (Progress-Bound) to obtain

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - \frac{\eta_t}{2}\langle \nabla f(x_t), x_t - p_t \rangle + \frac{\eta_t^2 L \|x_t - p_t\|_2^2}{2}.$$

We plug (Scaling-BOR) and (Scaling-CVX) into the inequality above, resulting in

$$\begin{aligned}
h_{t+1} &\leq \left(1 - \frac{\eta_t}{2}\right) h_t - \frac{\eta_t}{2}\langle \nabla f(x_t), x_t - p_t \rangle + \frac{\eta_t^2 L \|x_t - p_t\|_2^2}{2} \\
&\leq \left(1 - \frac{\eta_t}{2}\right) h_t - \frac{\eta_t \beta}{2}\|\Pi \nabla f(x_t)\|_2 + \frac{\eta_t^2 L \delta^2}{2} \\
&\leq \left(1 - \frac{\eta_t}{2}\right) h_t - \frac{\eta_t \beta}{2}\left(\sqrt{\frac{\alpha_f}{2}}\sqrt{h_t} - \frac{\eta_t^2}{\eta_S^2}\sqrt{\frac{\alpha_f \beta M}{2}} - \frac{\eta_t^4}{\eta_S^4} L\beta\right) + \frac{\eta_t^2 L \delta^2}{2} \\
&\leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t \frac{\sqrt{\alpha_f}\beta}{2\sqrt{2}}\sqrt{h_t} + \frac{\eta_t^2 L \delta^2}{2} + \frac{\eta_t^3 \beta}{\eta_S^2}\sqrt{\frac{\alpha_f \beta M}{2}} + \frac{\eta_t^5 L \beta^2}{\eta_S^4 2}.
\end{aligned}$$

Since $\eta_t / \eta_S \leq 1$ for all $t \in \{S, \dots, T-1\}$, it holds that

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t \frac{\sqrt{\alpha_f}\beta}{2\sqrt{2}}\sqrt{h_t} + \eta_t^2 \left(\frac{L\delta^2}{2} + \frac{\beta}{\eta_S}\sqrt{\frac{\alpha_f \beta M}{2}} + \frac{L\beta^2}{\eta_S 2}\right) \quad . \tag{19}$$

Let

$$A = \frac{\sqrt{\alpha_f}\beta}{2\sqrt{2}}, \qquad B = \frac{L\delta^2}{2} + \frac{\beta}{\eta_S}\sqrt{\frac{\alpha_f \beta M}{2}} + \frac{L\beta^2}{\eta_S 2}, \qquad C = C_t = 1$$

for all $t \in \{S, \dots, T-1\}$, and $\psi = 1/2$. Ideally, we could now apply Lemma 3.5. However, Inequality (19) is only guaranteed to hold in case that $h_t \geq \frac{\eta_t^4}{\eta_S^4}\beta M$. Thus, we have to extend the proof of Lemma 3.5 for the case that $h_t \leq \frac{\eta_t^4}{\eta_S^4}\beta M$. In case $h_t \leq \frac{\eta_t^4}{\eta_S^4}\beta M$, (9) implies that

$$h_{t+1} \leq (1 - \eta_t) h_t + \eta_t^2 \frac{L\|x_t - p_t\|_2^2}{2} \leq h_t + \eta_t^2 \frac{L\delta^2}{2} \leq \eta_{t-1}\eta_t \left(\frac{\beta M}{\eta_S^2} + \frac{L\delta^2}{2}\right) = \eta_{t-1}\eta_t \left(\frac{D}{\eta_S^2} + E\right) \quad ,$$

where $D = \beta M$ and $E = \frac{L\delta^2}{2}$. Thus, in the proof of Lemma 3.5, the induction assumption (11) has to be replaced by

$$h_t \leq \max\left(\frac{\eta_{t-2}\eta_{t-1}}{\eta_{S-2}\eta_{S-1}}h_S, \frac{\eta_{t-2}\eta_{t-1}B^2}{A^2} + \eta_{t-2}\eta_{t-1}BC, \eta_{t-2}\eta_{t-1}\left(\frac{D}{\eta_S^2} + E\right)\right).$$

Then, using the same analysis as in Lemma 3.5, extended by the case that $h_t \leq \frac{\eta_t^4}{\eta_S^4}\beta M$, proves that

$$h_t \leq \max\left(\left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^2 h_S, \left(\frac{\eta_{t-2}B}{A}\right)^2 + \eta_{t-2}^2 B, \eta_{t-2}^2\left(\frac{D}{\eta_S^2} + E\right)\right)$$

for all $t \in \{S, \dots, T\}$. □

C MISSING PROOFS FROM SECTION 5

We first present a lemma characterizing the kernel used in Section 5.

Lemma C.1. *Let $\mathcal{H}$ be the Hilbert space defined in (8) and let $k\colon \mathbb{R}\times\mathbb{R} \to \mathcal{H}$ be the kernel defined in (Bernoulli-kernel). For $y, z \in [0,1]$ and $n \in \mathbb{Z}$, it holds that*

$$k(y,z) = k(z,y) = k(|y-z|, 0) = \frac{1}{2}B_2(|y-z|) \qquad \text{and} \qquad k(y,z) = k(y, z+n).$$

Proof. We first prove that for $y, z \in [0,1]$, it holds that $k(y,z) = k(z,y)$. Let $a \in [0,1[$. Then,

$$\begin{aligned} [a] &= a - \lfloor a \rfloor = a - 0 = a, \\ [-a] &= -a - \lfloor -a \rfloor = -a - (-1) = 1 - a, \\ B_2([a]) &= a^2 - a + \frac{1}{6} = (1-a)^2 - (1-a) + \frac{1}{6} = B_2[-a]. \end{aligned} \tag{20}$$

Similarly, note that

$$\begin{aligned} [1] &= 1 - \lfloor 1 \rfloor = 1 - 1 = 0, \\ [-1] &= -1 - \lfloor -1 \rfloor = -1 - (-1) = 0, \\ B_2([1]) &= B_2([-1]). \end{aligned} \tag{21}$$

By (20) and (21), for any $y, z \in [0,1]$, it holds that

$$k(y,z) = \frac{1}{2}B_2([y-z]) = \frac{1}{2}B_2([z-y]) = k(z,y). \tag{22}$$

Next, we prove that for $y, z \in [0,1]$, it holds that $k(y,z) = k(|y-z|, 0) = \frac{1}{2}B_2(|y-z|)$. Let $y, z \in [0,1]$ such that $|y-z| = a \in [0,1[$. Then, by (20),

$$k(y,z) = \frac{1}{2}B_2([y-z]) = \frac{1}{2}B_2([|y-z|]) = \frac{1}{2}B_2(|y-z|). \tag{23}$$

Furthermore, $k(y,z) = \frac{1}{2}B_2([y-z]) = \frac{1}{2}B_2([|y-z|]) = k(|y-z|, 0)$. Next, let $y, z \in [0,1]$ such that $|y-z| = 1$. Then, by (21),

$$k(y,z) = \frac{1}{2}B_2([y-z]) = \frac{1}{2}B_2([|y-z|]) = \frac{1}{2}B_2([1]) = \frac{1}{12} = \frac{1}{2}\left(1^2 - 1 + \frac{1}{6}\right) = \frac{1}{2}B_2(1) = \frac{1}{2}B_2(|y-z|). \tag{24}$$

Furthermore, $k(y,z) = \frac{1}{2}B_2([y-z]) = \frac{1}{2}B_2([|y-z|]) = \frac{1}{2}B_2([1]) = k(|y-z|, 0)$.

Finally, we prove that for $y, z \in [0,1]$ and $n \in \mathbb{Z}$, it holds that $k(y,z) = k(y, z+n)$. Indeed,

$$k(y,z) = \frac{1}{2}B_2(y - z - \lfloor y - z \rfloor) = \frac{1}{2}B_2(y - z - n - \lfloor y - z - n \rfloor) = k(y, z+n).$$

□

Next, we provide a proof sketch for Theorem 5.1.

Sketch of proof for Theorem 5.1. The main idea behind the proof is that FW with $\eta_t = \frac{1}{t+1}$ leads to iterates $x_t = \frac{1}{t}\sum_{i=1}^{t}\Phi(y_i)$ with $\{y_1, \dots, y_t\} = \left\{\frac{i-1}{t} \mid i = 1, \dots, t\right\}$ for all $t = 2^m$, where $m \in \mathbb{N}$. Then, the proof follows by a series of calculations. We make several introductory observations. Note that Line 3 of Algorithm 1 becomes $p_t \in \operatorname{argmin}_{p\in\mathcal{C}} Df(x_t)(p - x_t) = \operatorname{argmin}_{p\in\mathcal{C}} Df(x_t)(p)$, where, for $w, x \in \mathcal{H}$, $Df(w)(x) = \langle w, x\rangle_{\mathcal{H}}$ denotes the first derivative of f at w. For $x \in \mathcal{C}$ and $x_t \in \mathcal{C}$ of the form $x_t = \frac{1}{t}\sum_{i=1}^{t}\Phi(y_i)$ for $y_1, \dots, y_t \in [0, 1]$, it holds that $Df(x_t)(x) = \left\langle \frac{1}{t}\sum_{i=1}^{t}\Phi(y_i), x\right\rangle_{\mathcal{H}}$. Then, for $y \in [0, 1]$, let

$$g_t(y) := \left\langle \frac{1}{t}\sum_{i=1}^{t}\Phi(y_i), \Phi(y)\right\rangle_{\mathcal{H}} = \frac{1}{t}\sum_{i=1}^{t} k(y_i, y)\,. \tag{25}$$

Since the LMO of FW always returns a vertex of $\mathcal{C}$ of the form $\Phi(y)$ for $y \in [0, 1]$ (Bach et al., 2012), it holds that $\min_{p\in\mathcal{C}} Df(x_t)(p) = \min_{y\in[0,1]} g_t(y)$ and the vertex returned by the LMO during iteration t is contained in the set $\{\Phi(z) \mid z \in \operatorname{argmin}_{y\in[0,1]} g_t(y)\}$. Thus, instead of considering the LMO directly over $\mathcal{C}$, we can perform the computations over $[0, 1]$. To simplify the proof, we make the following assumption on the argmin operation in the LMO of FW, a tie-breaking rule in case $|\operatorname{argmin}_{p\in\mathcal{C}} Df(x_t)(p)| \geq 2$.

Assumption C.2. The LMO of FW always returns $p_t \in \operatorname{argmin}_{p\in\mathcal{C}} Df(x_t)(p)$ such that $p_t = \Phi(z)$ for $z = \min(\operatorname{argmin}_{y\in[0,1]} g_t(y))$.

Recall that FW starts at iterate x_0, but since $\eta_0 = 1$, it holds that $x_1 = \Phi(y_1)$. As we will prove in Lemma C.4, without loss of generality, we can assume that FW starts at iterate $x_1 = \Phi(y_1)$, where $y_1 = 0$. □

We now detail three technical lemmas.

Lemma C.3. *Let $\mathcal{H}$ be the Hilbert space defined in (8), let $k\colon \mathbb{R}\times\mathbb{R} \to \mathcal{H}$ be the kernel defined in (Bernoulli-kernel), let $\Phi\colon [0, 1] \to \mathcal{H}$ be the feature map associated with k restricted to $[0, 1] \times [0, 1]$, let $t \in \mathbb{N}$, let $\{y_1, \dots, y_t\} = \left\{\frac{i-1}{t} \mid i \in \{1, \dots, t\}\right\}$, and let g_t be defined as in (25), that is, $g_t(y) = \frac{1}{t}\sum_{i=1}^{t} k(y_i, y)$. Then, it holds that $\operatorname{argmin}_{y\in[0,1]} g_t(y) = \left\{y_i + \frac{1}{2t} \mid i \in \{1, \dots, t\}\right\}$.*

Proof. Let $t \in \mathbb{N}$ and $\{y_1, \dots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \dots, t\}\}$. We stress that this does not imply that for all $i \in \{1, \dots, t\}$, $y_i = \frac{i-1}{t}$. By Lemma C.1, for all $y \in [0, 1]$, it holds that

$$g_t(y) = \left\langle \frac{1}{t}\sum_{i=1}^{t}\Phi(y_i), \Phi(y)\right\rangle_{\mathcal{H}} = \frac{1}{t}\sum_{i=1}^{t} k(y_i, y) = \frac{1}{2t}\sum_{i=1}^{t} |y_i - y|^2 - |y_i - y| + \frac{1}{6}\,.$$

Then, for $y \in [0, 1] \setminus \{y_1, \dots, y_t\}$, it holds that $g_t'(y) = \frac{1}{2t}\sum_{i=1}^{t} 2(y - y_i) - \frac{y-y_i}{|y-y_i|}$ and since $\sum_{i=1}^{t} y_i = (t-1)/2$,

$$g_t'(y) = \frac{1}{2}\left(2y - \frac{t-1}{t} - \frac{1}{t}\left|\{y_i < y \mid i \in \{1, \dots, t\}\}\right| + \frac{1}{t}\left|\{y_i > y \mid i \in \{1, \dots, t\}\}\right|\right)\,.$$

For $y \in \left(\frac{i-1}{t}, \frac{i}{t}\right)$, where $i \in \{1, \dots, t\}$, it holds that

$$g_t'(y) = \frac{1}{2}\left(2y - \frac{t-1}{t} - \frac{i}{t} + \frac{t-i}{t}\right) = \frac{1}{2}\left(2y + \frac{1}{t} - \frac{2i}{t}\right)$$

and $g_t'(y) = 0$ if and only if $y = \frac{i-\frac{1}{2}}{t}$. Since g_t is strongly convex on $\left(\frac{i-1}{t}, \frac{i}{t}\right)$ for $i \in \{1, \dots, t\}$ and continuous on $[0, 1]$, it holds that $y_i = \frac{i-1}{t}$ cannot be a minimum of g_t on $[0, 1]$ for any $i \in \{1, \dots, t\}$. Since $g_t(0) = g_t(1)$ by Lemma C.1, 1 cannot be a minimum either. Thus, only elements in $\left\{y_i + \frac{1}{2t} \mid i \in \{1, \dots, t\}\right\}$ can be minima of g_t on $[0, 1]$.

We next prove that $g_t(\frac{i-1}{t} + \frac{1}{2t}) = g_t(\frac{j-1}{t} + \frac{1}{2t})$ for all $i, j \in \{1, \dots, t\}$, which concludes the proof of the lemma. To see this, we show that $g_t(\frac{j-1}{t} + \frac{1}{2t}) = g_t(\frac{j}{t} + \frac{1}{2t})$ for all $j \in \{1, \dots, t-1\}$. Using that, by Lemma C.1, $k(y, z) = \frac{1}{2}B_2(|y-z|)$

and $k(1, y) = k(0, y)$ for $y, z \in [0, 1]$, we have that

$$\begin{aligned}\sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j-1}{t} + \frac{1}{2t}\right) - \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j}{t} + \frac{1}{2t}\right) &= \sum_{i=1}^{t} k\left(\frac{i}{t}, \frac{j}{t} + \frac{1}{2t}\right) - \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j}{t} + \frac{1}{2t}\right) \\ &= k\left(\frac{t}{t}, \frac{j}{t} + \frac{1}{2t}\right) - k\left(\frac{0}{t}, \frac{j}{t} + \frac{1}{2t}\right) \\ &= 0\end{aligned}$$

for all $j \in \{1, \ldots, t-1\}$. Thus, $g_t(\frac{j-1}{t} + \frac{1}{2t}) = g_t(\frac{j}{t} + \frac{1}{2t})$ for all $j \in \{1, \ldots, t-1\}$. □

Lemma C.4. *Let $\mathcal{H}$ be the Hilbert space defined in (8), let $k\colon \mathbb{R} \times \mathbb{R} \to \mathcal{H}$ be the kernel defined in (Bernoulli-kernel), let $\Phi\colon [0, 1] \to \mathcal{H}$ be the feature map associated with k restricted to $[0, 1] \times [0, 1]$, let $t \in \mathbb{N}$, let $y_1, \ldots, y_t \in [0, 1]$, and let g_t be defined as in (25), that is, $g_t(y) = \frac{1}{t}\sum_{i=1}^{t} k(y_i, y)$. Suppose that $\operatorname{argmin}_{y\in[0,1]} g_t(y) = \{z_1, \ldots, z_k\} \subseteq [0, 1]$ for some $k \in \mathbb{N}$. Let $c \in \mathbb{R}$, let $\tilde{y}_i = [y_i + c]$ for all $i \in \{1, \ldots, t\}$, and let $\tilde{g}_t(y) = \frac{1}{t}\sum_{i=1}^{t} k(\tilde{y}_i, y)$. Then, $\operatorname{argmin}_{z\in[0,1]} \tilde{g}_t(z) = \{[z_1 + c], \ldots, [z_k + c]\}$.*

Proof. It holds that

$$\begin{aligned}\operatorname*{argmin}_{z\in[0,1]} \tilde{g}_t(z) &= \operatorname*{argmin}_{z=[y+c], y\in\mathbb{R}} \tilde{g}_t(z) \\ &= \operatorname*{argmin}_{z=[y+c], y\in\mathbb{R}} \frac{1}{2t}\sum_{i=1}^{t} B_2([[y_i + c] - [y + c]]) \\ &= \operatorname*{argmin}_{z=[y+c], y\in\mathbb{R}} \frac{1}{2t}\sum_{i=1}^{t} B_2([y_i + c - \lfloor y_i + c\rfloor - (y + c) - (-\lfloor y + c\rfloor)]) \\ &= \operatorname*{argmin}_{z=[y+c], y\in\mathbb{R}} \frac{1}{2t}\sum_{i=1}^{t} B_2([y_i - y - \lfloor y_i + c\rfloor + \lfloor y + c\rfloor]) \\ &= \operatorname*{argmin}_{z=[y+c], y\in\mathbb{R}} \frac{1}{2t}\sum_{i=1}^{t} B_2([y_i - y]) \\ &= \{[z_1 + c], \ldots, [z_k + c]\},\end{aligned}$$

where the second-to-last equality is due to Lemma C.1. □

Lemma C.5. *Let $\mathcal{H}$ be the Hilbert space defined in (8), let $k\colon \mathbb{R} \times \mathbb{R} \to \mathcal{H}$ be the kernel defined in (Bernoulli-kernel), let $\Phi\colon [0, 1] \to \mathcal{H}$ be the feature map associated with k restricted to $[0, 1] \times [0, 1]$, let $\mathcal{C} = \operatorname{conv}(\{\Phi(y) \mid y \in [0, 1]\})$ be the marginal polytope, and let $\mu = 0$ such that $f(x) = \frac{1}{2}\|x\|_{\mathcal{H}}^2$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{1}{t+1}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t and the LMO satisfying Assumption C.2 it holds that $x_t = \frac{1}{t}\sum_{i=1}^{t} \Phi(y_i)$ with $\{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \ldots, t\}\}$ for all $t \in \{1, \ldots, T\}$ such that $t = 2^m$ for $m \in \mathbb{N}$,.*

Proof. Since $\eta_0 = 1$, it holds that $x_1 = \Phi(y_1)$. By Lemma C.4, without loss of generality, we can assume that FW starts with iterate $x_1 = \Phi(y_1)$, where $y_1 = 0$. Let $t \in \{1, \ldots, T\}$. Since we use the step-size $\eta_t = \frac{1}{t+1}$, we obtain uniform weights, that is, $x_t = \frac{1}{t}\sum_{i=1}^{t} \Phi(y_i)$, where $y_i \in [0, 1]$ for all $i \in \{1, \ldots, t\}$. Suppose that $t = 2^m$ for some $m \in \mathbb{N}$. The proof that it holds that $\{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \ldots, t\}\}$ is by induction on $m \in \mathbb{N}$. The base case, $m = 0$, follows from $x_1 = \Phi(y_1)$, where $y_1 = 0$. Suppose that for $t = 2^m$ for some $m \in \mathbb{N}$, it holds that $\{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \ldots, t\}\}$. If we show that

$$\{y_1, \ldots, y_{2t}\} = \left\{\frac{i-1}{2t} \mid i \in \{1, \ldots, 2t\}\right\}, \tag{26}$$

the statement of the lemma follows from induction. (26) is subsumed by the stronger statement that $y_{t+j} = y_j + \frac{1}{2t}$ for all $j \in \{1, \ldots, t\}$, and we prove the latter for the remainder of this proof.

By Lemma C.3 and Assumption C.2, it holds that $y_{t+1} = \frac{1}{2t}$. Suppose that for some $\ell \in \{1, \ldots, t-1\}$, it holds that $y_{t+j} = y_j + \frac{1}{2t}$ for all $j \in \{1, \ldots, \ell\}$. We decompose the function $g_{t+\ell}(y)$ into $g_t(y)$ and $\tilde{g}_\ell(y) =$

$\langle \frac{1}{\ell} \sum_{i=1}^{\ell} \Phi(y_i + \frac{1}{2t}), \Phi(y) \rangle_{\mathcal{H}}$, that is, we consider the decomposition $g_{t+\ell}(y) = \frac{t}{t+\ell} g_t(y) + \frac{\ell}{t+\ell} \tilde{g}_\ell(y)$. By Lemma C.3, $\operatorname{argmin}_{y\in[0,1]} g_t(y) = \{ y_i + \frac{1}{2t} \mid i \in \{1, \dots, t\} \} \subseteq [0,1]$ and by Assumption C.2, $y_{\ell+1} = \min(\operatorname{argmin}_{y\in[0,1]} g_\ell(y))$. Thus, by Lemma C.4, it holds that

$$\min \left(\operatorname*{argmin}_{y\in[0,1]} \tilde{g}_\ell(y) \right) = \min \left(\operatorname*{argmin}_{y\in[0,1]} g_\ell(y) + \frac{1}{2t} \right) = y_{\ell+1} + \frac{1}{2t} \in \left\{ y_i + \frac{1}{2t} \mid i \in \{1, \dots, t\} \right\}.$$

Thus, $\min(\operatorname{argmin}_{y\in[0,1]} \tilde{g}_\ell(y)) \in \operatorname{argmin}_{y\in[0,1]} g_t(y)$ and

$$y_{t+\ell+1} = \min \left(\operatorname*{argmin}_{y\in[0,1]} g_{t+\ell}(y) \right) = \min \left(\operatorname*{argmin}_{y\in[0,1]} \tilde{g}_\ell(y) \right) = y_{\ell+1} + \frac{1}{2t}.$$

By induction, $y_{t+j} = y_j + \frac{1}{2t}$ for all $j \in \{1, \dots, t\}$, as required to conclude the proof. □

Finally, we prove Theorem 5.1.

Proof of Theorem 5.1. By Lemma C.5, we have $x_t = \frac{1}{t} \sum_{i=1}^{t} \Phi\left(\frac{i-1}{t}\right)$ and, since $\mu = 0$,

$$\begin{aligned} f(x_t) &= \frac{1}{2} \|x_t\|_{\mathcal{H}}^2 \\ &= \frac{1}{2t^2} \left\langle \sum_{i=1}^{t} \Phi\left(\frac{i-1}{t}\right), \sum_{j=1}^{t} \Phi\left(\frac{j-1}{t}\right) \right\rangle_{\mathcal{H}} \\ &= \frac{1}{2t^2} \sum_{j=1}^{t} \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j-1}{t}\right) \\ &= \frac{1}{2t} \sum_{i=1}^{t} k\left(\frac{i-1}{t}, 1\right), \end{aligned}$$

where the second-to-last equality follows from the definition of k and the last equality follows from repeatedly applying

$$\sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j-1}{t}\right) = \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j}{t}\right), \tag{27}$$

where $j \in \{1, \dots, t\}$. To see that (27) holds, recall that by Lemma C.1, it holds that

$$\begin{aligned} \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j-1}{t}\right) - \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j}{t}\right) &= \sum_{i=1}^{t} k\left(\frac{i}{t}, \frac{j}{t}\right) - \sum_{i=1}^{t} k\left(\frac{i-1}{t}, \frac{j}{t}\right) \\ &= k(1, \tfrac{j}{t}) - k(0, \tfrac{j}{t}) \\ &= 0 \end{aligned}$$

for all $j \in \{1, \dots, t\}$. Thus,

$$f(x_t) = \frac{1}{2t} \sum_{i=1}^{t} k\left(\frac{i-1}{t}, 1\right) = \frac{1}{2t} \sum_{i=1}^{t} k\left(\frac{i-1}{t}, 0\right) = \frac{1}{2t} \sum_{i=1}^{t} k\left(\frac{i}{t}, 0\right) = \frac{1}{4t} \sum_{i=1}^{t} \left(\frac{i}{t}\right)^2 - \frac{i}{t} + \frac{1}{6},$$

where the second, third, and fourth equalities are due to Lemma C.1. Since $\sum_{i=1}^{t} i = \frac{t(t+1)}{2}$ and $\sum_{i=1}^{t} i^2 = \frac{2t^3+3t^2+t}{6}$, it holds that $f(x_t) = \frac{1}{4t}\left(\frac{2t+3+\frac{1}{t}}{6} - \frac{t+1}{2} + \frac{t}{6}\right) = \frac{1}{24t^2}$. □

D NORM OF THE GRADIENT OF f IS BOUNDED FROM BELOW BY A NONNEGATIVE CONSTANT

In this section, we address the setting when feasible region $\mathcal{C}$ is uniformly convex and the norm of the gradient of f is bounded from below by a nonnegative constant.

For this setting, FW with line-search or short-step admits linear convergence rates when the feasible region is also strongly convex (Levitin and Polyak, 1966; Demianov and Rubinov, 1970; Garber and Hazan, 2015). In Theorem 2.2, Kerdreux et al. (2021b) interpolated between $\mathcal{O}(1/t)$ and the linear convergence rates by relaxing strong convexity of the feasible region to uniform convexity. Two FW variants employ open-loop step-sizes and enjoy accelerated convergence rates when the norm of the gradient of f is bounded from below by a nonnegative constant and the feasible region $\mathcal{C}$ is uniformly convex: the primal averaging conditional gradients algorithm (PACG) (Lan, 2013; Kerdreux et al., 2021a) and a momentum-guided FW variant (Li et al., 2021). Below, we derive Theorem D.2 for FW with open-loop step-size rules, which interpolates between the known convergence rate of $\mathcal{O}(1/t)$ (Jaggi, 2013), and $\mathcal{O}(1/t^2)$ depending on the uniform convexity of the feasible region. To prove the result, we require two new scaling inequalities. Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set and let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function such that there exists $\lambda > 0$ such that for all $x \in \mathcal{C}$,

$$\|\nabla f(x)\|_2 \geq \lambda. \tag{Scaling-EXT}$$

In case f is well-defined, convex, and differentiable on the entirety of $\mathbb{R}^d$, (Scaling-EXT) is, for example, implied by the convexity of f and the assumption that the unconstrained optimum of f, that is, $\operatorname{argmin}_{x\in\mathcal{C}} f(x)$, does not lie in $\mathcal{C}$. The second scaling inequality follows from the uniform convexity of the feasible region and is proved in the proof of Kerdreux et al. (2021b, Theorem 2.2) in FW gap, using Kerdreux et al. (2021b, Lemma 2.1). The result stated below is then obtained by bounding the FW gap from below with the primal gap.

Lemma D.1 (Kerdreux et al., 2021b)**.** *For $\alpha > 0$ and $q \geq 2$, let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact (α, q)-uniformly convex set and let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex function that is differentiable in an open set containing $\mathcal{C}$ with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Then, for all $x \in \mathcal{C}$,*

$$\frac{\langle \nabla f(x), x - p\rangle}{\|x - p\|_2^2} \geq \left(\frac{\alpha}{2}\|\nabla f(x)\|_2\right)^{2/q} (f(x) - f(x^*))^{1-2/q}, \tag{Scaling-UNIF}$$

where $p \in \operatorname{argmin}_{v\in\mathcal{C}} \langle \nabla f(x), v\rangle$.

Combining (Scaling-EXT) and (Scaling-UNIF), we prove the following result.

Theorem D.2 (Norm of the gradient of f is bounded from below by a nonnegative constant)**.** *For $\alpha > 0$ and $q \geq 2$, let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact (α, q)-uniformly convex set of diameter $\delta > 0$, let $f\colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function with lower-bounded gradients, that is, $\|\nabla f(x)\|_2 \geq \lambda$ for all $x \in \mathcal{C}$ for some $\lambda > 0$, with unique minimizer $x^* \in \operatorname{argmin}_{x\in\mathcal{C}} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, when $q \geq 4$, it holds that*

$$h_t \leq \max\left\{\eta_{t-2}^{1/(1-2/q)}\frac{L\delta^2}{2},\ \eta_{t-2}L\left(\frac{2}{\alpha\lambda}\right)^{2/q}\right\}^{1/(1-2/q)} + \eta_{t-2}^2\frac{L\delta^2}{2}\Bigg\},$$

for all $t \in \{1, \dots, T\}$, and when $q \in [2, 4[$, with $S = \left\lceil 8L\delta^2 \right\rceil$, it holds that

$$h_t \leq \max\left\{\left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^2 h_S,\ \eta_{t-2}L\left(\frac{2}{\alpha\lambda}\right)^{2/q}\right\}^2 + \eta_{t-2}^2\frac{L\delta^2}{2}\Bigg\},$$

for all $t \in \{S, \dots, T\}$.

Proof. Let $t \in \{1, \dots, T-1\}$. Combining (Scaling-UNIF) and (Scaling-EXT), it holds that

$$\langle \nabla f(x_t), x_t - p_t\rangle \geq \|x_t - p_t\|_2^2 \left(\frac{\alpha\lambda}{2}\right)^{2/q} h_t^{1-2/q}.$$

Then, using (Progress-Bound), we obtain $h_{t+1} \leq h_t - \eta_t \|x_t - p_t\|_2^2 (\frac{\alpha\lambda}{2})^{2/q} h_t^{1-2/q} + \frac{\eta_t^2 L \|x_t - p_t\|_2^2}{2}$. Combined with (9), we have

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t + \frac{\eta_t \|x_t - p_t\|_2^2}{2} \left(\eta_t L - \left(\frac{\alpha\lambda}{2}\right)^{2/q} h_t^{1-2/q}\right). \tag{28}$$

Suppose that $q \geq 4$. Then, $2/q \in [0, 1/2]$ and we can apply Lemma 3.5 with $A = (\frac{\alpha\lambda}{2})^{2/q}$, $B = L$, $C = \frac{\delta^2}{2}$, $C_t = \frac{\|x_t - p_t\|_2^2}{2}$ for all $t \in \{1, \dots, T-1\}$, and $\psi = 2/q$, resulting in

$$h_t \leq \max\left\{ \left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^{1/(1-2/q)} h_S, \left(\eta_{t-2} L \left(\frac{2}{\alpha\lambda}\right)^{2/q}\right)^{1/(1-2/q)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right\},$$

which, with $S = 1$, $h_1 \leq \frac{L\delta^2}{2}$, and $\eta_{-1} \geq \eta_0 = 1$ proves the first statement of the lemma.

Next, suppose that $q \in [2, 4[$. Note that $2/q > 1/2$. Thus, we require a burn-in phase after which Lemma 3.5 can be applied. Let $S = \lceil 8L\delta^2 \rceil \geq 8L\delta^2$ and $t \in \{S, \dots, T-1\}$. By Proposition 3.1, $h_t \leq \frac{8L\delta^2}{S+3} \leq 1$. Since $1 - 2/q \leq 1/2$, we have $h_t^{1-2/q} \geq h_t^{1/2} = h_t^{1-1/2}$. Combined with (28), it holds that

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t + \frac{\eta_t \|x_t - p_t\|_2^2}{2} \left(\eta_t L - \left(\frac{\alpha\lambda}{2}\right)^{2/q} h_t^{1-1/2}\right).$$

We then apply Lemma 3.5 with $A = (\frac{\alpha\lambda}{2})^{2/q}$, $B = L$, $C = \frac{\delta^2}{2}$, $C_t = \frac{\|x_t - p_t\|_2^2}{2}$ for all $t \in \{S, \dots, T-1\}$, and $\psi = 1/2$, resulting in

$$h_t \leq \max\left\{ \left(\frac{\eta_{t-2}}{\eta_{S-1}}\right)^2 h_S, \left(\eta_{t-2} L \left(\frac{2}{\alpha\lambda}\right)^{2/q}\right)^2 + \eta_{t-2}^2 \frac{L\delta^2}{2} \right\},$$

for all $t \in \{S, \dots, T\}$. Note that the lemma holds even if $S = 1$ since $\eta_{-1} \geq \eta_0 = 1$. □

As we show below, in the setting of Theorem D.2, in case the feasible region is strongly convex, FW with open-loop step-sizes $\eta_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 4}$ an even number converges at rates faster than $\mathcal{O}(1/t^2)$.

Remark D.3 (Open-loop with linear convergence rate)**.** Under the assumptions of Theorem D.2, suppose that $\mathcal{C}$ is α-strongly convex. Since $q = 2$, (28) simplifies to

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t + \frac{\eta_t \|x_t - p_t\|_2^2}{2} \left(\eta_t L - \frac{\alpha\lambda}{2}\right).$$

Analogously to Proposition 3.1, one can prove a convergence rate of $\mathcal{O}(1/t)$ for FW with any step-size $\eta_t = \frac{\ell}{t+\ell}$ for even $\ell \in \mathbb{N}_{\geq 4}$ depending on L, δ, and ℓ. Thus, there exists $S \in \mathbb{N}$ depending only on $L, \alpha, \delta, \lambda$, and ℓ, such that for all $t \in \{S, \dots, T-1\}$, it holds that

$$\frac{\eta_t \|x_t - p_t\|_2^2}{2} \left(\eta_t L - \frac{\alpha\lambda}{2}\right) \leq 0.$$

By induction, for even $\ell \in \mathbb{N}_{\geq 4}$, it then holds that

$$h_t \leq \frac{h_S (S + \ell/2)(S + \ell/2 + 1) \cdots (S + \ell - 1)}{(t + \ell/2)(t + \ell/2 + 1) \cdots (t + \ell - 1)}$$

for all $t \in \{S, \dots, T-1\}$, yielding a convergence rate of $\mathcal{O}(1/t^{\ell/2})$ after an initial burn-in phase with convergence rate $\mathcal{O}(1/t)$ for the first S iterations. Using a similar line of arguments, one can prove that the constant open-loop step-size rule $\eta_t = \frac{\alpha\lambda}{2L}$ admits a linear convergence rate of $h_t \leq (1 - \frac{\alpha\lambda}{4L})^t h_0$. for all $t \in \{0, \dots, T\}$.

E NO ASSUMPTIONS ON THE LOCATION OF THE OPTIMUM

In this section, we address the setting when there are no assumptions on the location of the optimum, the feasible region $\mathcal{C}$ is uniformly convex, and the objective function f satisfies (HEB).

Garber and Hazan (2015) showed that strong convexity of the feasible region and the objective function are enough to modify (Progress-Bound) to prove a $\mathcal{O}(1/t^2)$ convergence rate of FW with line-search or short-step. These assumptions were relaxed in Kerdreux et al. (2021b, Theorem 2.10) and convergence rates for FW with line-search or short-step interpolating between $\mathcal{O}(1/t)$ and $\mathcal{O}(1/t^2)$ were provided. Below, we show that accelerated convergence rates not only hold for line-search or short-step but also open-loop step-size rules, characterizing a problem setting for which FW with open-loop step-size rules converges at the same rate as FW with line-search or short-step, up to a constant.

Theorem E.1 (No assumptions on the location of the optimum)**.** *For $\alpha > 0$ and $q \geq 2$, let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact (α, q)-uniformly convex set of diameter $\delta > 0$, let $f \colon \mathcal{C} \to \mathbb{R}$ be a convex and L-smooth function satisfying a (μ, θ)-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer $x^* \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with open-loop step-size η_t, it holds that*

$$h_t \leq \max \left\{ \eta_{t-2}^{1/(1-2\theta/q)} \frac{L\delta^2}{2}, \quad \eta_{t-2} L \left(\frac{2\mu}{\alpha} \right)^{2/q} \left(\frac{L\delta^2}{2} \right)^{1/(1-2\theta/q)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right\},$$

for all $t \in \{1, \dots, T\}$.

Proof. Let $t \in \{1, \dots, T-1\}$. Combining (Scaling-UNIF) and (Scaling-HEB), we obtain $\langle \nabla f(x_t), x_t - p_t \rangle \geq \|x_t - p_t\|_2^2 (\frac{\alpha}{2\mu})^{2/q} h_t^{1-2\theta/q}$. Then, using (Progress-Bound), we obtain $h_{t+1} \leq h_t - \eta_t \|x_t - p_t\|_2^2 (\frac{\alpha}{2\mu})^{2/q} h_t^{1-2\theta/q} + \frac{\eta_t^2 L \|x_t - p_t\|_2^2}{2}$. Combined with (9), we have

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2} \right) h_t + \frac{\eta_t \|x_t - p_t\|_2^2}{2} \left(\eta_t L - \left(\frac{\alpha}{2\mu} \right)^{2/q} h_t^{1-2\theta/q} \right).$$

This inequality allows us to apply Lemma 3.5 with $A = (\frac{\alpha}{2\mu})^{2/q}$, $B = L$, $C = \frac{\delta^2}{2}$, $C_t = \frac{\|x_t - p_t\|_2^2}{2}$ for all $t \in \{S, \dots, T-1\}$, and $\psi = 2\theta/q \leq 1/2$, resulting in

$$h_t \leq \max \left\{ \left(\frac{\eta_{t-2}}{\eta_{S-1}} \right)^{1/(1-2\theta/q)} h_S, \quad \eta_{t-2} L \left(\frac{2\mu}{\alpha} \right)^{2/q} \left(\frac{L\delta^2}{2} \right)^{1/(1-2\theta/q)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right\},$$

for all $t \in \{S, \dots, T\}$, which, with $S = 1$, $h_1 \leq \frac{L\delta^2}{2}$, and $\eta_{-1} \geq \eta_0 = 1$ proves the theorem. □