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

Elias Wirth  
Institute of Mathematics  
Berlin Institute of Technology  
Strasse des 17. Juni 135, Berlin, Germany

Thomas Kerdreux  
Geolabe LLC  
1615 Central Avenue, Los Alamos, New Mexico, USA

Sebastian Pokutta  
Institute of Mathematics & AI in Society, Science, and Technology  
Berlin Institute of Technology & Zuse Institute Berlin  
Strasse des 17. Juni 135, Berlin, Germany

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 enjoys faster convergence rates 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-sizes. Finally, we demonstrate that FW with open-loop step-sizes can compete with momentum-based open-loop FW variants.

Keywords: Frank-Wolfe algorithm, open-loop step-sizes, acceleration, kernel herding, convex optimization

1. Introduction

In this paper, we address the constrained convex optimization problem

$$\min_{x \in \mathcal{C}} f(x),$$  \hspace{1cm} (OPT)

where $\mathcal{C} \subseteq \mathbb{R}^d$ is a compact convex set and $f : \mathcal{C} \to \mathbb{R}$ is a convex and $L$-smooth function. Let $x^* \in \arg\min_{x \in \mathcal{C}} f(x)$ 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 $\arg\min_{x \in \mathcal{C}} \langle c, x \rangle$. At each iteration, the algorithm calls the LMO, $p_t \in \arg\min_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle$, 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
They are easy to implement, projection-free, affine invariant (Lacoste-Julien and Jaggi, 2013; Lan, 2013; 2014). See Braun et al. (2022), for a survey. For several settings, FW with line-search or short-step admits projection-free, affine invariant (Lacoste-Julien and Jaggi, 2013; Lan, 2013; 2014). See Braun et al. (2022), for a survey. For several settings, FW with line-search or short-step admits accelerated convergence rates in primal gap $h_t = f(x_t) - f(x^*)$, where $x^* = \arg\min_{x \in C} f(x)$ is the minimizer of $f$. Specifically, when the objective is strongly convex and the optimal solution lies in the relative interior of the feasible region, FW with line-search or short-step converges linearly (Guélat and Marcotte, 1986). Moreover, when the feasible region is strongly convex and the norm of the gradient of the objective is bounded from below by a nonnegative constant, FW with line-search or short-step converges linearly (Levitin and Polyak, 1966; Demianov and Rubinov, 1970; Dunn, 1979). Finally, when the feasible region and objective are strongly convex, FW with line-search or short-step converges at a rate of order $O(1/t^2)$, see also Table 1. However, the drawback of FW is its slow convergence rate when the feasible region $C$ is a polytope and the optimal solution lies in the relative interior of an at least one-dimensional face $C^*$ of $C$. In this setting, for any $\epsilon > 0$, FW with line-search or short-step converges at a rate of order $O(1/t^{1+\epsilon})$ (Wolfe, 1970; Canon and Cullum, 1968). To achieve linear convergence rates in this setting, algorithmic modifications of FW are necessary (Lacoste-Julien and Jaggi, 2015; Garber and Meshi, 2016; Braun et al., 2019; Combettes and Pokutta, 2020; Garber, 2020).

FW with open-loop step-size rules, on the other hand, has a convergence rate that is not governed by the lower bound of Wolfe (1970). Indeed, Bach (2021) proved an asymptotic convergence rate of order $O(1/t^2)$ for FW with open-loop step-sizes in the setting of Wolfe (1970). However, proving that the latter result holds non-asymptotically remains an open problem. Other disadvantages 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 $f$. On the other hand, open-loop step-size rules are problem-agnostic and, thus, easy to compute. Nevertheless, little is known about the settings in which FW with open-loop step-size rules admits acceleration, except for two momentum-exploiting variants that achieve convergence rates of order up to $O(1/t^2)$: The primal-averaging Frank-Wolfe algorithm (PAFW), presented in Algorithm 2, was first proposed by Lan (2013) and later analyzed by Kerdreux et al. (2021a). PAFW employs the open-loop step-size $\eta_t = \frac{1}{t^2}$ and momentum to achieve convergence rates of order up to $O(1/t^2)$ when the feasible region is uniformly convex and the gradient norm of the objective is bounded from below by a nonnegative constant. For the same setting, the momentum-guided Frank-Wolfe algorithm (MFW) (Li et al., 2021), presented in

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.
We demonstrate that the Algorithm 3, employs the open-loop step-size and the MFW with open-loop step-sizes admit convergence rates of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$ in the setting of the lower bound due to Wolfe (1970), that is, we prove the non-asymptotic version of the result due to Bach (2021). We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We prove the non-asymptotic version of the result due to Bach (2021). We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes admits a convergence rate of order $\mathcal{O}(1/t^2)$, assuming strict complementarity. We thus characterize a setting for which FW with open-loop step-sizes adm
and Jaggi, 2015) with open-loop step-sizes converge at rates of order $O(1/t^2)$ without the assumption of strict complementarity.

Fourth, we compare FW with open-loop step-sizes to PAFW and MFW for the problems of logistic regression and collaborative filtering. The results indicate that FW with open-loop step-sizes converges at comparable rates as or better rates than PAFW and MFW. This implies that faster convergence rates can not only be achieved by studying algorithmic variants of FW but can also be obtained via deeper understanding of vanilla FW and its various step-size rules.

Finally, we provide a theoretical analysis of the accelerated convergence rate of FW with open-loop step-sizes in the kernel herding setting of Bach et al. (2012, Figure 3, right).

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, for the setting of the lower bound of Wolfe (1970) and assuming strict complementarity, we prove that FW with open-loop step-sizes converges faster than FW with line-search or short-step. In Section 5, we introduce two algorithmic variants of FW with open-loop step-sizes that admit accelerated convergence rates in the problem setting of the lower bound of Wolfe (1970) without relying on strict complementarity. In Section 6, we prove accelerated convergence rates for FW with open-loop step-sizes in the infinite-dimensional kernel-herding setting of Bach et al. (2012, Figure 3, right). Section 7 contains the numerical experiments. Finally, we discuss our results in Section 8.

2. Preliminaries

Throughout, let $d \in \mathbb{N}$. Let $0 \in \mathbb{R}^d$ denote the all-zeros vector, let $\mathbb{1} \in \mathbb{R}^d$ denote the all-ones vector, and let $\mathbb{1}_i \in \mathbb{R}^d$ be a vector such that $\mathbb{1}_i = 0$ for all $i \in \{1, \ldots, [d/2]\}$ and $\mathbb{1}_i = 1$ for all $i \in \{[d/2] + 1, \ldots, d\}$. For $i \in \{1, \ldots, d\}$, let $e^{(i)} \in \mathbb{R}^d$ be the $i$th unit vector such that $e^{(i)}_i = 1$ and $e^{(i)}_j = 0$ for all $j \in \{1, \ldots, d\} \setminus \{i\}$. Given a vector $x \in \mathbb{R}^d$, define its support as $\text{supp}(x) = \{i \in \{1, \ldots, d\} \mid x_i \neq 0\}$. Let $I \in \mathbb{R}^{d \times d}$ denote the identity matrix. Given a set $C \subseteq \mathbb{R}^d$, let $\text{aff}(C)$, $\text{conv}(C)$, $\text{span}(C)$, and $\text{vert}(C)$ denote the affine hull, the convex hull, the span, and the set of vertices of $C$, respectively. For $z \in \mathbb{R}^d$ and $\beta > 0$, the ball of radius $\beta$ 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 $t \in \{0, \ldots, T\}$ by $h_t := f(x_t) - f(x^*)$, where $x^* \in \text{argmin}_{x \in C} f(x)$. Finally, for $x \in \mathbb{R}$, let $[x] := x - \lfloor x \rfloor$. We introduce several definitions.

Definition 2.1 (Uniformly convex set). Let $C \subseteq \mathbb{R}^d$ be a compact convex set, $\alpha_C > 0$, and $q > 0$. We say that $C$ is $(\alpha_C, q)$-uniformly convex with respect to $\|\cdot\|_2$ if for all $x, y \in C$, $\gamma \in [0, 1]$, and $z \in \mathbb{R}^d$ such that $\|z\|_2 = 1$, it holds that $\gamma x + (1 - \gamma) y + \gamma(1 - \gamma) \alpha_C \|x - y\|_2^2 z \in C$. We refer to $(\alpha_C, 2)$-uniformly convex sets as $\alpha_C$-strongly convex sets.

Definition 2.2 (Smooth function). Let $C \subseteq \mathbb{R}^d$ be a compact convex set, let $f : C \rightarrow \mathbb{R}$ be differentiable in an open set containing $C$, and let $L > 0$. We say that $f$ is $L$-smooth over $C$ with respect to $\|\cdot\|_2$ if for all $x, y \in C$, 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 $C \subseteq \mathbb{R}^d$ be a compact convex set, let $f : C \rightarrow \mathbb{R}$ be convex, let $\mu > 0$, and let $\theta \in [0, 1/2]$. We say that $f$ satisfies a $(\mu, \theta)$-Hölderian error bound if for all $x \in C$ and $x^* \in \text{argmin}_{x \in C} f(x)$, it holds that

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

(2.1)

Throughout, for ease of notation, we assume that $x^* \in \text{argmin}_{x \in C} f(x)$ is unique. This follows, for example, from the assumption that $f$ is strictly convex. When $x^* \in \text{argmin}_{x \in C} f(x)$ is unique, (2.1) becomes

$$
\mu (f(x) - f(x^*))^\theta \geq \|x - x^*\|_2.
$$

(HEB)
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 \(C \subseteq \mathbb{R}^d\) be a compact convex set, let \(f: C \to \mathbb{R}\) be differentiable in an open set containing \(C\), let \(\alpha_f > 0\), and let \(r \geq 2\). We say that \(f\) is \((\alpha_f, r)\)-uniformly convex over \(C\) with respect to \(\| \cdot \|_2\) if for all \(x, y \in C\), it holds that \(f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha_f}{r^2} \|x-y\|^2\). We refer to \((\alpha_f, 2)\)-uniformly convex functions as \(\alpha_f\)-strongly convex.

Note that \((\alpha_f, r)\)-uniformly convex functions satisfy a \(((r/\alpha_f)^{1/r}, 1/r)\)-(HEB): \(f(x) - f(x^*) \geq \langle \nabla f(x^*), x-x^* \rangle + \frac{\alpha_f}{r^2} \|x-x^*\|^2\).

### 3. Accelerated convergence rates for FW with open-loop step-sizes

FW with open-loop step-size rules was already studied by Dunn and Harshbarger (1978) and Jaggi (2013), respectively, two open-loop step-sizes are prevalent, \(\eta_t = \frac{1}{t+1}\), for which the best known convergence rate is \(O(\log(t)/t)\), and \(\eta_t = \frac{\alpha_f}{r^2}\), for which a faster convergence rate of order \(O(1/t)\) holds, see, for example, Dunn and Harshbarger (1978) and Jaggi (2013), respectively. In this section, we derive convergence rates for FW with open-loop step-size \(\eta_t = \frac{\alpha_f}{r^2}\). Convergence results for FW with \(\eta_t = \frac{1}{t+1}\) for \(t \in \mathbb{N}_{\geq 1}\) presented throughout this paper, except for those in Section 6, can always be generalized (up to a constant) to \(\eta_t = \frac{1}{t+j}\) for \(j \in \mathbb{N}_{\geq 1}\).

This section is structured as follows. First, we derive a baseline convergence rate of order \(O(1/t)\) in Section 3.1. Then, in Section 3.2, we present the proof blueprint used throughout most parts of the paper to derive accelerated convergence rates and directly apply our approach to the setting where the objective satisfies (HEB) and the optimal solution \(x^* \in \text{argmin}_{x \in C} f(x)\) lies in the relative interior of the feasible region. In Section 3.3, we prove accelerated rates when the feasible region is uniformly convex and the norm of the gradient of the objective is bounded from below by a nonnegative constant. Finally, in Section 3.4, we prove accelerated rates when the feasible region is uniformly convex and the objective satisfies (HEB).

#### 3.1 Convergence rate of order \(\mathcal{O}(1/t)\)

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 \(C \subseteq \mathbb{R}^d\) is a compact convex set and \(f: C \to \mathbb{R}\) is a convex and \(L\)-smooth function with unique minimizer \(x^* \in \text{argmin}_{x \in C} f(x)\). Then, the iterates of Algorithm 1 with any step-size \(\eta_t \in [0, 1]\) satisfy

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

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 \(\eta_t = \frac{1}{t+4}\) similar to the one derived by Jaggi (2013) for FW with \(\eta_t = \frac{\alpha_f}{r^2}\).

**Proposition 3.1** (Convergence rate of order \(\mathcal{O}(1/t)\)). Let \(C \subseteq \mathbb{R}^d\) be a compact convex set of diameter \(\delta > 0\), let \(f: C \to \mathbb{R}\) be a convex and \(L\)-smooth function with unique minimizer \(x^* \in \text{argmin}_{x \in C} f(x)\). Let \(T \in \mathbb{N}\) and \(\eta_t = \frac{1}{t+4}\) for all \(t \in \mathbb{Z}\). Then, for the iterates of Algorithm 1 with step-size \(\eta_t\), it holds that

\[
h_t \leq \frac{h_1 L \delta^2}{2} = n_{t-1} 2L \delta^2 \text{ for all } t \in \{1, \ldots, T\}.
\]

**Proof.** 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 \(h_1 \leq \frac{h_1 L \delta^2}{2}\). Let \(t \in \{1, \ldots, T - 1\}\). 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
\[
h_{t+1} \leq (1 - \eta_t) h_t + \frac{L \|x_t - p_t\|^2}{2} \]
\[
\leq \prod_{i=1}^t (1 - \eta_i) h_1 + \frac{L \delta^2}{2} \sum_{i=1}^t \eta_i^2 \prod_{j=i+1}^t (1 - \eta_j)
\leq \frac{L \delta^2}{2} \frac{4!}{(t+1) \cdots (t+4)} + \frac{L \delta^2}{2} \sum_{i=1}^t \frac{(i+1) \cdots (i+4)}{(i+4)^2 (t+1) \cdots (t+4)}
\leq 8L \delta^2 \frac{1}{(t+4-1)(t+4)} + \frac{L \delta^2}{(t+4-1)(t+4)}
\leq \frac{8L \delta^2}{t+4},
\] where we used that \( \prod_{j=i+1}^t (1 - \eta_j) = \frac{(i+1)(i+2)\cdots t}{(t+5)(t+6)\cdots(t+4)} = \frac{(i+1)(i+2)(i+3)(t+4)}{(t+1)(t+2)(t+3)(t+4)}. \]  

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

### 3.2 Optimal solution in the relative interior – 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 optimal solution are exploited (Levitin and Polyak, 1966; Demianov and Rubinov, 1970; Guélat and Marcotte, 1986; Garber and Hazan, 2015). A similar approach leads to acceleration results for FW with open-loop step-sizes, however, requiring a different proof technique as FW with open-loop step-sizes is not monotuous in primal gap. Here, we introduce the proof blueprint used to derive most of the accelerated rates in this paper via the setting when the objective \( f \) satisfies (HEB) and the minimizer of \( f \) is in the relative interior of the feasible region \( C \).

Our goal is to bound 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
\[
\frac{\langle \nabla f(x_t), x_t - p_t \rangle}{\|x_t - p_t\|_2} \geq \frac{\phi}{\|x_t - x^\star\|_2} \frac{\langle \nabla f(x_t), x_t - x^\star \rangle}{\|x_t - x^\star\|_2}.
\]  

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 \( \frac{x^\star - x_t}{\|x^\star - x_t\|_2} \). Scaling inequalities stem from the geometry of the feasible region, properties of the objective function, or information on the location of the optimal solution. The scaling inequality below exploits the latter property.

**Lemma 3.2** (Guélat and Marcotte, 1986). Let \( C \subseteq \mathbb{R}^d \) be a compact convex set of diameter \( \delta > 0 \), let \( f: C \to \mathbb{R} \) be a convex and \( L \)-smooth function with unique minimizer \( x^\star \in \text{argmin}_{x \in C} f(x) \), and suppose that there exists \( \beta > 0 \) such that \( \text{aff}(C) \cap B_\beta(x^\star) \subseteq C \). Then, for all \( x \in C \cap B_\beta(x^\star) \), it holds that
\[
\frac{\langle \nabla f(x), x - p \rangle}{\|x - p\|_2} \geq \frac{\beta}{\delta} \|\nabla f(x)\|_2,
\]  

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

Below, we prove that there exists \( S \in \mathbb{N} \) such that for all \( t \geq S \), \( x_t \in B_\beta(x^\star) \) and (Scaling-INT) is satisfied.
Lemma 3.3. Let $C \subseteq \mathbb{R}^d$ be a compact convex set of diameter $\delta > 0$, let $f : C \to \mathbb{R}$ be a convex and $L$-smooth function satisfying a $(\mu, \theta)$-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer $x^* \in \text{argmin}_{x \in C} f(x)$, and let $\beta > 0$. Let $S = [8L^2(\mu/\beta)^{1/\theta}]$, $T \in \mathbb{N}$, and $\eta = \frac{4}{t+4}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with step-size $\eta_t$, it holds that $\|x_t - x^*\|_2 \leq \beta$ for all $t \in \{S, \ldots, T\}$.

Proof. By (HEB) and Proposition 3.1, $\|x_t - x^*\|_2 \leq \mu \eta_t^\theta \leq \mu(\frac{8L^2}{8L^2(\mu/\beta)^{1/\theta}})^\theta \leq \beta$ for all $t \in \{S, \ldots, T\}$.

The second scaling inequality follows from the objective satisfying (HEB).

Lemma 3.4. Let $C \subseteq \mathbb{R}^d$ be a compact convex set and let $f : C \to \mathbb{R}$ be a convex function satisfying a $(\mu, \theta)$-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer $x^* \in \text{argmin}_{x \in C} f(x)$. Then, for all $x \in C$, it holds that

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

Proof. The statement holds for $x = x^*$. For $x \in C \setminus \{x^*\}$, by convexity and (HEB), $f(x) - f(x^*) \leq \langle \nabla f(x^*), x - x^* \rangle \|x - x^*\|_2 \leq \frac{\mu}{2} \|x - x^*\|_2^2 \mu (f(x) - f(x^*))^{\theta}$. Dividing by $\mu (f(x) - f(x^*))^{\theta}$ yields (Scaling-HEB).

For $t \in \{S, \ldots, T - 1\}$, where $S = [8L^2(2\mu/\beta)^{1/\theta}]$, we plug (Scaling-INT) and (Scaling-HEB) into (Progress-Bound) to obtain $h_{t+1} \leq h_t - \eta_t \frac{\beta^2}{2L^2} h_t^{1-\theta} + \eta_t^2 L^2 \beta^2$. Combined with (3.1), we have

$$h_{t+1} \leq (1 - \frac{\eta_t}{2}) h_t - \eta_t \frac{\beta^2}{2L^2} h_t^{1-\theta} + \eta_t^2 L^2 \beta^2.$$  \hspace{1cm} \text{(3.2)}

for all $t \in \{S, \ldots, T - 1\}$. If the primal gaps of FW with open-loop step-sizes satisfy an inequality of this type, 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 $t \in \{S, \ldots, T - 1\}$, and a nonnegative sequence $\{h_t\}_{t=S}^{T}$ such that

$$h_{t+1} \leq (1 - \frac{\eta_t}{2}) h_t - \eta_t AC_t h_t^{1-\psi} + \eta_t^2 BC_t$$  \hspace{1cm} \text{(3.3)}

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

$$h_t \leq \max \left\{ \left( \frac{\eta_{t-2}}{\eta_{S-1}} \right)^{1/(1-\psi)} h_{S}, \left( \frac{\eta_{t-2}B}{A} \right)^{1/(1-\psi)} + \eta_{t-2}^2 BC_t \right\}.$$  \hspace{1cm} \text{(3.4)}

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

Proof. For all $t \in \{S, \ldots, T\}$, we first prove that

$$h_t \leq \max \left\{ \left( \frac{\eta_{t-2}h_{t-1}}{\eta_{S-1}h_{S-1}} \right)^{1/(2(1-\psi))} h_{S}, \left( \frac{\eta_{t-2}h_{t-1}B^2}{A^2} \right)^{1/(2(1-\psi))} \right\}, \hspace{1cm} \text{(3.5)}$$

which then implies (3.4). 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 (3.5) with $t = S$ is immediate, even if $S = 1$, as $\eta_{t-1} \geq \eta_S = 1$. Suppose that (3.5) is correct for a specific iteration $t \in \{S, \ldots, T - 1\}$. We distinguish between two cases. First, suppose that $h_t \leq \left( \frac{\eta_{t}B}{A} \right)^{1/(1-\psi)}$. Plugging this bound into (3.3), we obtain $h_{t+1} \leq (1 - \frac{\eta_t}{2}) h_t - 0 + \eta_t^2 BC_t \leq \left( \frac{\eta_{t}B}{A} \right)^{1/(1-\psi)} + \eta_t^2 BC \leq \left( \frac{\eta_{t-1}B^2}{A^2} \right)^{1/(2(1-\psi))} + \eta_{t-1}h_t BC$. Next, suppose
that $h_t \geq \left(\frac{mB}{A}\right)^{1/(1-\psi)}$ instead. Plugging this bound on $h_t$ into (3.3) and using the induction assumption (3.5) at iteration $t$ yields

$$h_{t+1} \leq \left(1 - \frac{\eta_t}{2}\right) h_t - \eta_t AC_t \frac{\eta_t B}{A} + \eta_t^2 BC_t$$

$$= \frac{t + 2}{t + 4} h_t$$

$$= \frac{\eta_t}{\eta_{t-2}} h_t$$

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

$$\leq \max \left\{ \left(\frac{\eta_t - 1\eta_t}{\eta_{t-2}\eta_{t-1}}\right)^{1/(2(1-\psi))} h_S, \left(\frac{\eta_t - 1\eta_tB^2}{A^2}\right)^{1/(2(1-\psi))} + \eta_{t-1}\eta_tBC \right\},$$

where the last inequality holds due to $\frac{\eta_t}{\eta_{t-2}}(\eta_t - 2\eta_{t-1})^{1/(2(1-\psi))} \leq (\eta_t - 1\eta_t)^{1/(2(1-\psi))}$ for $\eta_t/\eta_{t-2} \in [0, 1]$ and $1/(2(1-\psi)) \in [1/2, 1]$. In either case, (3.5) is satisfied for $t+1$. By induction, the lemma follows.

We conclude the presentation of our proof blueprint by stating the first accelerated convergence rate for FW with open-loop step-size $\eta_t = \frac{\eta}{\sqrt{t+1}}$ when the the objective function $f$ satisfies (HEB) and the minimizer lies in the relative interior of the feasible region $\mathcal{C}$. For this setting, FW with line-search or short-step converges linearly if the objective function is strongly convex (Guélat and Marcotte, 1986; Garber and Hazan, 2015). Further, FW with open-loop step-size $\eta_t = \frac{\eta}{\sqrt{t+1}}$ converges at a rate of order $O(1/t^2)$ when the objective is of 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 relative interior of $\mathcal{C}$). Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set of diameter $\delta > 0$, let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a convex and $L$-smooth function satisfying a $\mu$-$(\mu, \theta)$-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer $x^* \in \arg\min_{x \in \mathcal{C}} f(x)$, and suppose that there exists $\beta > 0$ such that $\text{aff}(\mathcal{C}) \cap B_\beta(x^*) \subseteq \mathcal{C}$. Let $S = [8L\delta^2 (2\mu/\beta)^{1/\theta}]$, $T \in \mathbb{N}$, and $\eta_t = \frac{\eta}{\sqrt{t+1}}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with step-size $\eta_t$, it holds that

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

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

**Proof.** Let $t \in \{S, \ldots, T - 1\}$. 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 $t \in \{S, \ldots, T\}$, it follows that (3.2) holds. We apply Lemma 3.5 with $A = \frac{\eta_t^2}{4\delta^2}$, $B = \frac{L\delta^2}{2}$, $C = 1$, $C_t = 1$ for all $t \in \{S, \ldots, T - 1\}$, and $\psi = \theta$, resulting in (3.6) holding for all $t \in \{S, \ldots, T\}$.

We complement Theorem 3.6 with a discussion on the lower bound of the convergence rate of FW when the optimal solution is in the relative interior of the probability simplex.

**Lemma 3.7** (Jaggi, 2013). Let $\mathcal{C} \subseteq \mathbb{R}^d$ be the probability simplex, $f(x) = ||x||_2^2$, and $t \in \{1, \ldots, d\}$. It holds that $\min_{x \in \mathcal{C}} f(x) = \frac{1}{d}$, where $||\text{supp}(x)||$ denotes the number of non-zero entries of $x$.

**Remark 3.8** (Compatibility with lower bound from Jaggi (2013)). In Lemma 3.7, the optimal solution $x^* = \frac{1}{d} \mathbb{1} \in \mathbb{R}^d$ lies in the relative 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 $e^{(i)} = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^d$, $i \in \{1, \ldots, d\}$. 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 $t \in \{1, \ldots, d\}$. Applying Theorem 3.6 to the setting of Lemma 3.7, we observe that $\beta = 1/d$ and acceleration starts only after $S = \Omega(d^{1/\theta}) \geq \Omega(d)$ iterations. Thus, Theorem 3.6 does not contradict Lemma 3.7.
Algorithm 2: Primal-averaging Frank-Wolfe algorithm (PAFW) (Lan, 2013)

```
Input: $x_0 \in C$, step-sizes $\eta_t = \frac{\ell}{t+\ell}$, where $\ell \in \mathbb{N}_{\geq 1}$, for $t \in \{0, \ldots, T-1\}$.

1. $v_0 \leftarrow x_0$
2. For $t = 0, \ldots, T-1$ do
   3. $y_t \leftarrow (1 - \eta_t)x_t + \eta_tv_t$
   4. $w_{t+1} \leftarrow \nabla f(y_t)$
   5. $v_{t+1} \in \text{argmin}_{v \in C}(w_{t+1}, v)$
   6. $x_{t+1} \leftarrow (1 - \eta_t)x_t + \eta_tv_{t+1}$
```

Algorithm 3: Momentum-guided Frank-Wolfe algorithm (MFW) (Li et al., 2021)

```
Input: $x_0 \in C$, step-sizes $\eta_t = \frac{\ell}{t+\ell}$, where $\ell \in \mathbb{N}_{\geq 1}$, for $t \in \{0, \ldots, T-1\}$.

1. $v_0 \leftarrow x_0$; $w_0 \leftarrow 0$
2. For $t = 0, \ldots, T-1$ do
   3. $y_t \leftarrow (1 - \eta_t)x_t + \eta_tv_t$
   4. $w_{t+1} \leftarrow (1 - \eta_t)w_t + \eta_t\nabla f(y_t)$
   5. $v_{t+1} \in \text{argmin}_{v \in C}(w_{t+1}, v)$
   6. $x_{t+1} \leftarrow (1 - \eta_t)x_t + \eta_tv_{t+1}$
```

3.3 Unconstrained minimizer in the exterior – lower-bounded gradient norm

In this section, we apply the proof blueprint from the previous section to the setting when the feasible region $C$ 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 converges linearly when the feasible region is also strongly convex (Levitin and Polyak, 1966; Demianov and Rubinov, 1970; Garber and Hazan, 2015). When the feasible region is only uniformly convex, rates interpolating between $O(1/t)$ and linear convergence are known (Kerdreux et al., 2021b). Two FW variants employ open-loop step-sizes and enjoy accelerated convergence rates of order up to $O(1/t^2)$ when the feasible region $C$ is uniformly convex and the norm of the gradient of $f$ is bounded from below by a nonnegative constant: the primal-averaging Frank-Wolfe algorithm (PAFW) (Lan, 2013; Kerdreux et al., 2021a), presented in Algorithm 2, and the momentum-guided FW algorithm (MFW) (Li et al., 2021), presented in Algorithm 3. Below, for the same setting, we prove that FW with open-loop step-size $\eta_t = \frac{\ell}{t+\ell}$ also admits accelerated convergence rates of order up to $O(1/t^2)$ depending on the uniform convexity of the feasible region. Furthermore, when the feasible region is strongly convex, we prove that FW with open-loop step-size $\eta_t = \frac{\ell}{t+\ell}$, where $\ell \in \mathbb{N}_{\geq 2}$, converges at a rate of order $O(1/t^{\ell/2})$, which is faster than the convergence rates known for PAFW and MFW. To prove these results, we require two new scaling inequalities, the first of which follows directly from the assumption that the norm of the gradient of $f$ is bounded from below by a nonnegative constant. More formally, let $C \subseteq \mathbb{R}^d$ be a compact convex set and let $f: C \to \mathbb{R}$ be a convex and $L$-smooth function such that there exists $\lambda > 0$ such that for all $x \in C$,

$$\|
abla f(x)\|_2 \geq \lambda. \quad \text{(Scaling-EXT)}$$

In case $f$ is well-defined, convex, and differentiable on $\mathbb{R}^d$, (Scaling-EXT) is, for example, implied by the convexity of $f$ and the assumption that the unconstrained minimizer of $f$, that is, $\text{argmin}_{x \in \mathbb{R}^d} f(x)$, lies in the exterior of $C$. 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. The result stated below is then obtained by bounding the FW gap from below with the primal gap.

**Lemma 3.9 (Kerdreux et al., 2021b).** For $\alpha > 0$ and $q \geq 2$, let $C \subseteq \mathbb{R}^d$ be a compact $(\alpha, q)$-uniformly convex set and let $f: C \to \mathbb{R}$ be a convex function that is differentiable in an open set containing $C$ with unique minimizer $x^* \in \text{argmin}_{x \in C} f(x)$. Then, for all $x \in C$, it holds that

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

where $p \in \text{argmin}_{v \in C}(\nabla f(x), v)$.

Combining (Scaling-EXT) and (Scaling-UNIF), we derive the following accelerated convergence result.

**Theorem 3.10 (Norm of the gradient of $f$ is bounded from below by a nonnegative constant).** For $\alpha > 0$ and $q \geq 2$, let $C \subseteq \mathbb{R}^d$ be a compact $(\alpha, q)$-uniformly convex set of diameter $\delta > 0$, let $f: C \to \mathbb{R}$ 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 $x^* = \arg\min_{x \in \mathcal{C}} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{t}{t+1}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with step-size $\eta_t$, when $q \geq 4$, it holds that

$$
\eta_t \leq \max \left\{ \frac{\eta_{t-2}^{1/(1-2/q)}}{2} L_2^2, \left( \frac{\eta_{t-2} L \left( \frac{2}{\alpha \lambda} \right)^{2/q}}{2} \right)^{1/(1-2/q)} + \frac{\eta_{t-2}^2 L_2^2}{2} \right\} \quad (3.7)
$$

for all $t \in \{1, \ldots, T\}$, and letting $S = \lfloor 8L\delta^2 \rfloor$, when $q \in [2, 4]$, it holds that

$$
\eta_t \leq \max \left\{ \left( \frac{\eta_{t-2}}{\eta_{S-1}} \right)^2 h_S, \left( \frac{\eta_{t-2} L}{2} \left( \frac{\eta}{\alpha \lambda} \right)^{2/q} \right)^2 + \frac{\eta_{t-2}^2 L_2^2}{2} \right\} \quad (3.8)
$$

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

**Proof.** Let $t \in \{1, \ldots, T-1\}$. Combining (Scaling-UNIF) and (Scaling-EXT), it holds that $\langle \nabla f(x_t), x_t - p_t \rangle \geq \|x_t - p_t\|_2^2 \eta_t^{1-2/q}$. Then, using (Progress-Bound), we obtain $h_{t+1} \leq \eta_t \|x_t - p_t\|^2_2 (\frac{\alpha \lambda}{2})^{2/q} h_t^{1-2/q} + \eta_{t-2}^2 \|x_t - p_t\|^2_2$. Combined with (3.1), we obtain

$$
h_{t+1} \leq (1 - \frac{\eta_t}{h}) h_t + \eta_t \|x_t - p_t\|^2_2 \left( \eta_t L - \left( \frac{\alpha \lambda}{2} \right)^{2/q} h_t^{1-2/q} \right) \quad (3.9)
$$

Suppose that $q \geq 4$. Then, (3.9) allows us to 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 $t \in \{1, \ldots, T-1\}$, and $\psi = 2/q \in [0, 1/2]$, resulting in (3.7) holding for all $t \in \{1, \ldots, T\}$, since $h_1 \leq \frac{L_2^2}{2} \eta_1$ and $\eta_1 - \eta_0 = 1$. Next, suppose that $q \in [2, 4]$ and note that $2/q > 1/2$. Thus, Lemma 3.5 can be applied after a burn-in phase of slower convergence. Let $t \in \{S, \ldots, T-1\}$. By Proposition 3.1, $h_t \leq h_S \leq 1$. Since $1 - 2/q \leq 1/2$, we have $h_{t-2}^{1-2/q} \geq h_t^{1/2} = h_t^{1-1/2}$. Combined with (3.9), it holds that $h_{t+1} \leq (1 - \frac{\eta_t}{h}) h_t + \eta_t \|x_t - p_t\|^2_2 (\eta_t L - \left( \frac{\alpha \lambda}{2} \right)^{2/q} h_t^{1-1/2})$. 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 $t \in \{S, \ldots, T-1\}$, and $\psi = 1/2$, resulting in (3.8) holding for all $t \in \{S, \ldots, T\}$. Note that the lemma holds even if $S = 1$ since $\eta_1 \geq \eta_0 = 1$.

As we discuss below, in the setting of Theorem 3.10, when $q = 2$, FW with open-loop step-sizes $\eta_t = \frac{t}{t+1}$, where $\ell \in \mathbb{N}_{>2}$, converges at a rate of order $O(1/t^{\ell/2})$.

**Remark 3.11** (Acceleration beyond rates of order $O(1/t^{\ell/2})$). Under the assumptions of Theorem 3.10, analogously to Proposition 3.1, one can prove convergence rates of order $O(1/\ell)$ for FW with step-sizes $\eta_t = \frac{t}{t+1}$, where $\ell \in \mathbb{N}_{>2}$, depending on $L, \delta, \ell$ and $\lambda$. Thus, for $q = 2$, there exists $S \in \mathbb{N}$ depending only on $L, \alpha, \delta, \lambda, \ell$, such that for all $t \in \{S, \ldots, T-1\}$, it holds that

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

Thus, (3.9) becomes $h_{t+1} \leq (1 - \frac{\eta_t}{h}) h_t$ for all $t \in \{S, \ldots, T-1\}$. Then, by induction, for even $\ell \in \mathbb{N}_{>2}$, it 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 $t \in \{S, \ldots, T-1\}$, resulting in a convergence rate of order $O(1/\ell^{\ell/2})$. For $\ell \in \mathbb{N}_{>6}$, this convergence rate is better than the convergence rates of order $O(1/t^{\ell/2})$ known for PAFW and MFW. Using similar arguments, one can prove that FW with the constant open-loop step-size $\eta_t = \frac{\alpha \lambda}{2\ell}$ converges linearly, that is, $h_t \leq (1 - \frac{\alpha \lambda}{4\ell^2}) h_0$ for all $t \in \{0, \ldots, T\}$.

The results in Figure 1, see Section 7.1.1 for details, show that in the setting of Theorem 3.10 and Remark 3.11, FW with open-loop step-sizes $\eta_t = \frac{t}{t+1}$, where $\ell \in \mathbb{N}_{>1}$, converges at a rate of order $O(1/t^{\ell/2})$ and FW with constant step-size $\eta_t = \frac{\alpha \lambda}{2\ell}$ converges linearly in Figure 1a. The convergence rates for FW with $\eta_t = \frac{t}{t+1}$ are better than predicted by Remark 3.11 and indicate a gap between theory and practice. Note that we observe acceleration beyond $O(1/t^{\ell/2})$ even when the feasible region is only uniformly convex, a behaviour which our current theory does not explain.
be a compact satisfying a, Theorem 3.12

4. Optimal solution in the relative interior of a face of

Figure 1: Comparison of FW with different step-sizes when the feasible region $C \subseteq \mathbb{R}^{100}$ is an $\ell_p$-ball, the objective $f$ is not strongly convex, and the unconstrained optimal solution $\text{argmin}_{x \in \mathbb{R}^d} f(x)$ lies in the exterior of $C$, implying that $\|\nabla f(x)\|_2 \geq \lambda > 0$ for all $x \in C$ for some $\lambda > 0$. The $y$-axis represents the minimum primal gap. FW with open-loop step-sizes $\eta = \frac{c}{t^\ell}$, where $\ell \in \mathbb{N}_{\geq 1}$, converges at a rate of order $O(1/t^\ell)$ and FW with constant step-size converges linearly.

3.4 No assumptions on the location of the optimal solution

In this section, we address the setting when the feasible region $C$ is uniformly convex, the objective function $f$ satisfies (HEB), and no assumptions are made on the location of the optimal solution $x^* = \text{argmin}_{x \in C} f(x)$.

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 convergence rate of order $O(1/t^2)$ for FW with line-search or short-step. Kerdreux et al. (2021b) relaxed these assumptions and proved convergence rates for FW with line-search or short-step interpolating between $O(1/t)$ and $O(1/t^2)$. Below, for the same setting, we prove that FW with open-loop step-sizes also admits rates interpolating between $O(1/t)$ and $O(1/t^2)$.

**Theorem 3.12** (No assumptions on the location of the optimal solution). For $\alpha > 0$ and $q \geq 2$, let $C \subseteq \mathbb{R}^d$ be a compact $(\alpha, q)$-uniformly convex set of diameter $\delta > 0$, let $f : C \to \mathbb{R}$ be a convex and $L$-smooth function satisfying a $(\mu, \theta)$-(HEB) for some $\mu > 0$ and $\theta \in [0, 1/2]$ with unique minimizer $x^* = \text{argmin}_{x \in C} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{4}{t^\ell}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 1 with step-size $\eta_t$, it holds that

$$h_t \leq \max \left\{ \eta_{t-2}^{1/(1-2\theta/q)} \frac{L\delta^2}{2}, \left( \eta_{t-2}L \left( \frac{2\mu}{\alpha} \right)^{2/q} \right)^{1/(1-2\theta/q)} + \eta_{t-2}^2 \frac{L\delta^2}{2} \right\} \tag{3.10}$$

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

**Proof.** Let $t \in \{1, \ldots, T-1\}$. Combining (Scaling-UNIF) and (Scaling-HEB), we obtain $\langle \nabla f(x_t), x_t - p_t \rangle \geq \|x_t - p_t\|^2 \left( \frac{\alpha}{2\mu} \right)^{2/q} h_t^{1-2\theta/q}$. Then, using (Progress-Bound), we obtain $h_{t+1} \leq h_t - \eta_t \|x_t - p_t\|^2 \left( \frac{\alpha}{2\mu} \right)^{2/q} h_t^{1-2\theta/q} + \eta_t^2 \frac{L\delta^2}{2}$. Combined with (3.1), we have $h_{t+1} \leq (1 - \frac{\eta_t}{2})h_t + \eta_t \frac{\|x_t - p_t\|^2}{2} (\eta_t L - (\frac{\alpha}{2\mu})^2 \frac{L\delta^2}{2})$. We apply Lemma 3.5 with $A = (\frac{\mu}{2\alpha})^{2/q}$, $B = L$, $C = \frac{L\delta^2}{2}$, $C_t = \frac{\|x_t - p_t\|^2}{2}$ for all $t \in \{S, \ldots, T-1\}$, and $\psi = 2\theta/q \leq 1/2$, resulting in (3.10) holding for all $t \in \{S, \ldots, T\}$, since $h_1 \leq \frac{L\delta^2}{2}$, and $\eta_{-1} \geq \eta_0 = 1$.

4. Optimal solution in the relative interior of a face of $C$

In this section, we consider the setting when the feasible region is a polytope, the objective function is strongly convex, and the optimal solution lies in the relative interior of an at least one-dimensional face $C^*$ of
\(\mathcal{C}\). Then, under mild assumptions, FW with line-search or short-step converges at a rate of order \(\Omega(1/t^{1+\varepsilon})\) for any \(\varepsilon > 0\) (Wolfe, 1970). Due to this lower bound, several FW variants with line-search or short-step were developed that converge linearly in the described setting, see Section 1.1.

For this setting, following our earlier blueprint from Section 3.2, we prove that FW with open-loop step-sizes converges at a rate of order \(O(1/t^2)\), which is non-asymptotically faster than FW with line-search or short-step. Our result can be thought of as the non-asymptotic version of Proposition 2.2 in Bach (2021). Contrary to the result of Bach et al. (2012), our result is in primal gap, we do not require bounds on the third-order derivatives of the objective, and we do not invoke affine invariance of FW to obtain acceleration. To prove our result, we require two assumptions. The first assumption stems from active set identification, that is, the concept of identifying the face \(\mathcal{C}^* \subseteq \mathcal{C}\) containing the optimal solution \(x^* \in \arg\min_{x \in \mathcal{C}} f(x)\) to then apply faster methods whose convergence rates then often only depend on the dimension of the optimal face (Hager and Zhang, 2006; Bomze et al., 2019; 2020). Here, it is possible to determine the number of iterations necessary for FW with open-loop step-sizes to identify the optimal face when the following regularity assumption, already used in, for example, Garber (2020); Li et al. (2021), is satisfied.

**Assumption 1** (Strict complementarity). Let \(\mathcal{C} \subseteq \mathbb{R}^d\) be a polytope and let \(f: \mathcal{C} \to \mathbb{R}\) be differentiable in an open set containing \(\mathcal{C}\). Suppose that \(x^* \in \arg\min_{x \in \mathcal{C}} f(x)\) is unique and contained in an at least one-dimensional face \(\mathcal{C}^*\) of \(\mathcal{C}\) and that there exists \(\kappa > 0\) such that if \(p \in \text{vert}(\mathcal{C}) \setminus \mathcal{C}^*\), then \(\langle \nabla f(x^*), p - x^* \rangle \geq \kappa\); otherwise, if \(p \in \text{vert}(\mathcal{C}^*)\), then \(\langle \nabla f(x^*), p - x^* \rangle = 0\).

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 4.1** (Active set identification). Let \(\mathcal{C} \subseteq \mathbb{R}^d\) be a polytope of diameter \(\delta > 0\), let \(f: \mathcal{C} \to \mathbb{R}\) be a convex and \(L\)-smooth function satisfying a \((\mu, \theta)\)-(HEB) for some \(\mu > 0\) and \(\theta \in [0, 1/2]\) with unique minimizer \(x^* \in \arg\min_{x \in \mathcal{C}} f(x)\), and suppose that there exists \(\kappa > 0\) such that Assumption 1 is satisfied. Let \(S = \lceil 8L\delta^2 (2\mu L\delta/\kappa)^{1/\theta} \rceil\), \(T \in \mathbb{N}\), and \(\eta_t = \frac{1}{1+t}\) for all \(t \in \mathbb{Z}\). Then, for the iterates of Algorithm 1 with step-size \(\eta_t\), it holds that \(p_t \in \text{vert}(\mathcal{C}^*)\) for all \(t \in \{S, \ldots, T-1\}\).

**Proof.** Let \(t \in \{S, \ldots, T-1\}\). Note that in Line 2 of Algorithm 1, \(p_t \in \text{argmin}_{p \in \mathcal{C}} \langle \nabla f(x_t), p - x_t \rangle\) can always be chosen such that \(p_t \in \text{argmin}_{p \in \text{vert}(\mathcal{C})} \langle \nabla f(x_t), p - x_t \rangle\). For \(p \in \text{vert}(\mathcal{C})\), it holds that

\[
\langle \nabla f(x_t), p - x_t \rangle = \langle \nabla f(x_t) - \nabla f(x^*), p - x^* + x^* - x_t \rangle
\]

\[
= \langle \nabla f(x_t) - \nabla f(x^*), p - x_t \rangle + \langle \nabla f(x^*), p - x_t \rangle + \langle \nabla f(x^*), x^* - x_t \rangle.
\]

We distinguish between vertices \(p \in \text{vert}(\mathcal{C}) \setminus \mathcal{C}^*\) and vertices \(p \in \text{vert}(\mathcal{C}^*)\). First, suppose that \(p \in \text{vert}(\mathcal{C}) \setminus \mathcal{C}^*\). Using strict complementarity, Cauchy-Schwarz, \(L\)-smoothness, and (HEB) to bound (4.1) yields

\[
\langle \nabla f(x_t), p - x_t \rangle \geq -\|\nabla f(x_t) - \nabla f(x^*)\|_2 \|p - x_t\|_2 + \kappa + \langle \nabla f(x^*), x^* - x_t \rangle
\]

\[
\geq \kappa - L\delta \|x_t - x^*\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle
\]

\[
\geq \kappa - \mu L\delta h_t^\theta + \langle \nabla f(x^*), x^* - x_t \rangle.
\]

Next, suppose that \(p \in \text{vert}(\mathcal{C}^*)\). Using strict complementarity, Cauchy-Schwarz, \(L\)-smoothness, and (HEB) to bound (4.1) yields

\[
\langle \nabla f(x_t), p - x_t \rangle \leq \|\nabla f(x_t) - \nabla f(x^*)\|_2 \|p - x_t\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle
\]

\[
\leq L\delta \|x_t - x^*\|_2 + \langle \nabla f(x^*), x^* - x_t \rangle
\]

\[
\leq \mu L\delta h_t^\theta + \langle \nabla f(x^*), x^* - x_t \rangle.
\]

By Proposition 3.1, \(\mu L\delta h_t^\theta \leq \mu L\delta h_t^\theta \leq \mu L\delta \left(\frac{8L^2S}{8L^2(2\mu L\delta/\kappa)^{1/\theta}}\right)^6 < \frac{\kappa}{2}\). Hence, for \(t \in \{S, \ldots, T-1\}\),

\[
\langle \nabla f(x_t), p - x_t \rangle = \begin{cases}
\geq \frac{\kappa}{2} + \langle \nabla f(x^*), x^* - x_t \rangle, & \text{if } p \in \text{vert}(\mathcal{C}) \setminus \mathcal{C}^* \\
< \frac{\kappa}{2} + \langle \nabla f(x^*), x^* - x_t \rangle, & \text{if } p \in \text{vert}(\mathcal{C}^*).
\end{cases}
\]
Then, by optimality of \( p_t \), for all iterations \( t \in \{S, \ldots, T - 1\} \) of Algorithm 1, it holds that \( p_t \in \text{vert}(C^*) \). □

In addition, we assume the optimal solution \( x^* \in \arg\min_{x \in C} f(x) \) to be in the relative interior of an at least one-dimensional face \( C^* \) of \( C \).

**Assumption 2** (Optimal solution in the relative interior of a face of \( C \)). Let \( C \subseteq \mathbb{R}^d \) be a polytope and let \( f: C \rightarrow \mathbb{R} \). Suppose that \( x^* \in \arg\min_{x \in C} f(x) \) is unique and contained in the relative interior of an at least one-dimensional face \( C^* \) of \( C \), that is, there exists \( \beta > 0 \) such that \( \emptyset \neq B_\beta(x^*) \cap \text{aff}(C^*) \subseteq C \).

Using Assumption 2, Bach (2021) derived the following scaling inequality, a variation of (Scaling-INT).

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

\[
\langle \nabla f(x), x - p \rangle \geq \beta \|\Pi \nabla f(x)\|_2,
\]

where \( \Pi x \) denotes the orthogonal projection of \( x \in \mathbb{R}^d \) onto the span of \( \{x^* - p \mid p \in C^*\} \).

**Proof.** Suppose that \( x \in C \) such that \( p \in \arg\min_{v \in C} \langle \nabla f(x), v \rangle \subseteq C^* \). Then,

\[
\langle \nabla f(x), x - p \rangle = \max_{v \in C} \langle \nabla f(x), x - v \rangle \\
\geq \langle \nabla f(x), x - x^* \rangle + \langle \nabla f(x), \beta \frac{\Pi \nabla f(x)}{\|\Pi \nabla f(x)\|_2} \rangle \\
= \langle \nabla f(x), x - x^* \rangle + \langle \Pi \nabla f(x) + (I - \Pi) \nabla f(x), \beta \frac{\Pi \nabla f(x)}{\|\Pi \nabla f(x)\|_2} \rangle \\
= \langle \nabla f(x), x - x^* \rangle + \beta \|\Pi \nabla f(x)\|_2 \\
\geq \beta \|\Pi \nabla f(x)\|_2,
\]

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

To derive the final scaling inequality, we next bound the distance between \( x_t \) and the optimal face \( C^* \).

**Lemma 4.3** (Distance to optimal face). Let \( C \subseteq \mathbb{R}^d \) be a polytope of diameter \( \delta > 0 \), let \( f: C \rightarrow \mathbb{R} \) be a convex and \( L \)-smooth function satisfying a \((\mu, \theta)\)-(HEB) for some \( \mu > 0 \) and \( \theta \in [0, 1/2] \) with unique minimizer \( x^* \in \arg\min_{x \in C} f(x) \), and suppose that there exist \( \beta, \kappa > 0 \) such that Assumptions 1 and 2 are satisfied. Let \( S = \max\{[8L\delta^2 (\mu/\beta)^{1/\theta}], [8L^2 (2\mu L\delta/\kappa)^{1/\theta}]\} \), \( T \in \mathbb{N} \), and \( \eta_t = \frac{4}{t + 3} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 1 with step-size \( \eta_t \), it holds that

\[
\|(I - \Pi)(x_t - x^*)\|_2 \leq \frac{\eta_t}{\eta_S^4} \beta \quad (4.2)
\]

for all \( t \in \{S, \ldots, T - 1\} \), where \( \Pi x \) denotes the orthogonal projection of \( x \in \mathbb{R}^d \) onto the span of \( \{x^* - p \mid p \in C^*\} \).
Proof. Let \( t \in \{S, \ldots, T-1\} \). By Lemma 4.1, \( p_t \in \text{vert}(C^*) \). Thus, \((I-\Pi)(p_t - x^*) = 0,
\)
\[
\begin{align*}
(I-\Pi)(x_{t+1} - x^*) &= (1 - \eta_t)(I-\Pi)(x_t - x^*) + \eta_t(I-\Pi)(p_t - x^*) \\
&= (1 - \eta_t)(I-\Pi)(x_t - x^*) \\
&= \prod_{i=S}^{t} (1 - \eta_i)(I-\Pi)(x_S - x^*) \\
&= \frac{S(S+1)(S+2)(S+3)}{(t+1)(t+2)(t+3)(t+4)} (I-\Pi)(x_S - x^*),
\end{align*}
\]
and \( \|(I-\Pi)(x_{t+1} - x^*)\|_2 \leq \frac{\eta_t^{t+1}}{\eta_s^2} \|(I-\Pi)(x_S - x^*)\|_2 \leq \frac{\eta_t^{t+1}}{\eta_s^2} \beta \), where the last inequality follows from Lemma 3.3.

We derive the second scaling inequality below.

**Lemma 4.4.** Let \( C \subseteq \mathbb{R}^d \) be a polytope of diameter \( \delta > 0 \), let \( f: C \to \mathbb{R} \) be an \( \alpha_f \)-strongly convex and \( L \)-smooth function with unique minimizer \( x^* \in \text{argmin}_{x \in C} f(x) \), and suppose that there exist \( \beta, \kappa > 0 \) such that Assumptions 1 and 2 are satisfied. Let \( M = \max_{x \in C} \|\nabla f(x)\|_2, \quad S = \max\{16L^2/\alpha_f \beta^2, [64L^3 \delta^4/\alpha_f \kappa^2]\} \), \( T \in \mathbb{N} \), and \( \eta_t = \frac{4}{t^2} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 1 with step-size \( \eta_t \) and \( t \in \{S, \ldots, T-1\} \), it holds that \( h_t \leq \frac{\eta_t^4}{\eta_S^2} \beta M \) or

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

(Scaling-CVX)

where \( \Pi x \) denotes the orthogonal projection of \( x \in \mathbb{R}^d \) onto the span of \( \{x^* - p \mid p \in C^*\} \).

**Proof.** Given a vector \( x \in \mathbb{R}^d \), let \( \Pi_{\text{aff}(C^*)} x \) denote the projection of \( x \) onto \( \text{aff}(C^*) \), that is, \( \Pi_{\text{aff}(C^*)} x \in \text{argmin}_{y \in \text{aff}(C^*)} \|y - x\|_2 \). We first demonstrate how to express \( \Pi_{\text{aff}(C^*)} x \) using \( \Pi \). Since \( \text{aff}(C^*) = x^* + \text{span}\{x^* - p \mid p \in C^*\} \), there has to exist some \( y \in \mathbb{R}^d \) such that \( \Pi_{\text{aff}(C^*)} x = (I-\Pi)x^* + \Pi x + \Pi y \). By orthogonality of \( \Pi \), we have \( \|\Pi_{\text{aff}(C^*)} x - x\|_2 = \|(I-\Pi)x^*-(I-\Pi)\Pi x + \Pi y\|_2 = \|(I-\Pi)x^*-(I-\Pi)x\|_2 + \|\Pi y\|_2 \). The right-hand side is minimized when \( \Pi y = 0 \). Thus, \( \Pi_{\text{aff}(C^*)} x = (I-\Pi)x^* + \Pi x \in \text{argmin}_{y \in \text{aff}(C^*)} \|y - x\|_2 \).

Let \( t \in \{S, \ldots, T-1\} \). By Lemma 3.3, \( \|x_t - x^*\|_2 \leq \beta \) and, thus, by Assumption 2, \( \Pi_{\text{aff}(C^*)} x_t \in C^* \). By \( L \)-smoothness of \( f \), it holds that \( \|\nabla f(x_t) - \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \leq L \|x_t - \Pi_{\text{aff}(C^*)} x_t\|_2 = L \|(I-\Pi)(x_t - x^*)\|_2 \).

By Lemma 4.3, it then holds that

\[
\|\nabla f(x_t) - \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \leq \frac{\eta_t^4}{\eta_S^2} L \beta.
\]

(4.3)

Since for any \( x \in \mathbb{R}^d \), we have that \( \|\Pi x\|_2 \leq \|x\|_2 \). Inequality (4.3) implies that \( \|\Pi \nabla f(x_t) - \Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \leq \frac{\eta_t^4}{\eta_S^2} L \beta \). Combined with the triangle inequality, \( \|\Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \leq \|\Pi \nabla f(x_t)\|_2 + \|\Pi \nabla f(x_t) - \Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \leq \|\Pi \nabla f(x_t)\|_2 + \frac{\eta_t^4}{\eta_S^2} L \beta \), which we rearrange to

\[
\|\Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 - \frac{\eta_t^4}{\eta_S^2} L \beta \leq \|\Pi \nabla f(x_t)\|_2.
\]

(4.4)

For the remainder of the proof, we bound \( \|\Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \) from below. To do so, define the function \( g: C \cap B_\beta(x^*) \to \mathbb{R} \) via \( g(x) := f(\Pi_{\text{aff}(C^*)} x) = f((I-\Pi)x^* + \Pi x) \). The gradient of \( g \) at \( x \in C \cap B_\beta(x^*) \) is \( \nabla g(x) = \Pi \nabla f(\Pi_{\text{aff}(C^*)} x) = \Pi \nabla f((I-\Pi)x^* + \Pi x) \). Since \( f \) is \( \alpha_f \)-strongly convex in \( C \) and \( g(x) = f(x) \) for all \( x \in \text{aff}(C^*) \cap B_\beta(x^*) \), \( g \) is \( \alpha_f \)-strongly convex in \( \text{aff}(C^*) \cap B_\beta(x^*) \). Since the projection onto \( \text{aff}(C^*) \) is idempotent, \( \Pi_{\text{aff}(C^*)} x_t \in \text{aff}(C^*) \cap B_\beta(x^*) \), and \( g \) is \( \alpha_f \)-strongly convex in \( \text{aff}(C^*) \cap B_\beta(x^*) \), it holds that \( \|\Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 = \|\Pi \nabla f(\Pi_{\text{aff}(C^*)} x_t)\|_2 \geq \sqrt{\frac{\alpha_f}{2}} \sqrt{g(\Pi_{\text{aff}(C^*)} x_t) - g(x^*)} = \sqrt{\frac{\alpha_f}{2}} \sqrt{g(x_t) - g(x^*)} \).
\[ \sqrt{\frac{\alpha_f}{2}} \sqrt{\text{aff}(\Pi) x_t - f(x^*)}. \]

Suppose that \( h_t \geq \frac{\alpha_f}{\eta_S} \beta M \). Then, by Lemma 4.3 and Cauchy-Schwarz, we obtain \( h_t - \langle \nabla f(x_t), (I - \Pi)(x_t - x^*) \rangle \geq h_t - \frac{\alpha_f}{\eta_S} \beta M \geq 0 \). Combined with convexity of \( f \), it holds that

\[ \begin{align*}
\| \Pi \nabla f(\Pi \text{aff}(C^*) x_t) \|_2 & \geq \sqrt{\frac{\alpha_f}{2}} \sqrt{\langle f(x_t) + (\Pi \text{aff}(C^*) x_t - x_t), \Pi \text{aff}(C^*) x_t - x_t \rangle - f(x^*)} \\
& = \sqrt{\frac{\alpha_f}{2}} \sqrt{h_t - \langle \nabla f(x_t), (I - \Pi)(x_t - x^*) \rangle} \\
& \geq \sqrt{\frac{\alpha_f}{2}} \sqrt{h_t - \frac{\eta_t}{\eta_S} \beta M}.
\end{align*} \]

Since for \( a, b \in \mathbb{R} \) with \( a \geq b \geq 0 \), we have \( \sqrt{a - b} \geq \sqrt{a} - \sqrt{b} \), we obtain \( \| \Pi \nabla f(\Pi \text{aff}(C^*) x_t) \|_2 \geq \sqrt{\frac{\alpha_f}{2}} (\sqrt{h_t} - \sqrt{\frac{\eta_t}{\eta_S} \beta M}) \). Combined with (4.4), we obtain (Scaling-CVX).

\[ \square \]

Finally, we prove that when the feasible region \( C \) is a polytope, the objective function \( f \) is strongly convex, and the unique minimizer \( x^* \in \text{argmin}_{x \in C} f(x) \) lies in the relative interior of an at least one-dimensional face \( C^* \) of \( C \), FW with the open-loop step-size \( \eta_t = \frac{1}{t+1} \) converges at a rate of order \( O(1/t) \) for iterations \( t \leq S \) and at a non-asymptotic rate of order \( O(1/t^2) \) for iterations \( t \geq S \), where \( S \) is defined as in Lemma 4.4.

**Theorem 4.5 (Optimal solution in the relative interior of a face of \( C \)).** Let \( C \subset \mathbb{R}^d \) be a polytope of diameter \( \delta > 0 \), let \( f : C \to \mathbb{R} \) be an \( \alpha_f \)-strongly convex and \( L \)-smooth function with unique minimizer \( x^* \in \text{argmin}_{x \in C} f(x) \), and suppose that there exist \( \beta, \kappa > 0 \) such that Assumptions 1 and 2 are satisfied. Let \( M = \max_{x \in C} \| \nabla f(x) \|_2, S = \max \{ ([16L^2 \delta^4]/(\alpha_f \beta^2)], ([64L^2 \delta^4]/(\alpha_f \kappa^2]) \}, T \in \mathbb{N}, \) and \( \eta_t = \frac{1}{t+1} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 1 with step-size \( \eta_t \), it holds that

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

for all \( t \in \{ S, \ldots, T \} \), where

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

**Proof.** Let \( t \in \{ S, \ldots, T - 1 \} \) and suppose that \( h_t \geq \frac{\alpha_f}{\eta_S} \beta M \). Combine (3.1) and (Progress-Bound) to obtain

\[ h_{t+1} \leq (1 - \eta_t) h_t - \eta_t \frac{\beta}{2} (\nabla f(x_t), x_t - p_t) + \eta_t^2 \frac{L \delta^2}{2} \sqrt{h_t - \frac{\eta_t}{\eta_S} \beta M} \].

Plugging (Scaling-BOR) and (Scaling-CVX) into this inequality results in \( h_{t+1} \leq (1 - \eta_t) h_t - \eta_t \frac{\beta}{2} (\sqrt{h_t} - \sqrt{\frac{h_t}{\eta_S} \beta M} - \frac{\eta_t}{\eta_S} \beta) + \frac{\eta_t^2 L \delta^2}{2} \). Since \( \eta_t/\eta_S \leq 1 \) for all \( t \in \{ S, \ldots, T - 1 \} \), it holds that

\[ h_{t+1} \leq (1 - \eta_t) h_t - \eta_t \frac{\beta}{2} \sqrt{h_t + \eta_t^2 \frac{L \delta^2}{2} + \frac{\beta}{\eta_S^2} \beta M} \]

Let \( A, B, C \) as in (4.6), \( C_t = 1 \) for all \( t \in \{ S, \ldots, T - 1 \} \), and \( \psi = 1/2 \). Ideally, we could now apply Lemma 3.5. However, Inequality (4.7) is only guaranteed to hold in case that \( h_t \geq \frac{\alpha_f}{\eta_S} \beta M \). Thus, we have to extend the proof of Lemma 3.5 for the case that \( h_t \leq \frac{\alpha_f}{\eta_S} \beta M \). In case \( h_t \leq \frac{\alpha_f}{\eta_S} \beta M \), (3.1) implies that

\[ h_{t+1} \leq (1 - \eta_t) h_t + \eta_t^2 \frac{L \delta^2}{2} \sqrt{h_t} \]

and \( E \leq \frac{L \delta^2}{2} \). Thus, in the proof of Lemma 3.5, the induction assumption (3.5) has to be replaced by \( h_t \leq \max \left\{ \eta_t h_t, \frac{h_t - \frac{\beta}{\eta_S} \beta M}{A^2} + h_t - \frac{\beta}{\eta_S} \beta M \right\} \). Then, using the same analysis as in Lemma 3.5, extended by the case that \( h_t \leq \frac{\beta}{\eta_S} \beta M \), proves that (4.5) holds for all \( t \in \{ S, \ldots, T \} \).
Figure 2: Comparison of FW with different step-sizes when the feasible region \( C \subseteq \mathbb{R}^{100} \) is the probability simplex, the objective \( f(x) = \frac{1}{2} \| x - \rho \bar{x} \|_2^2 \), where \( \rho \in \{ \frac{1}{4}, 2 \} \), is strongly convex, and the optimal solution \( x^* = \arg\min_{x \in C} f(x) \) lies in the relative interior of an at least one-dimensional face of \( C \). The y-axis represents the minimum primal gap. For both settings, FW with open-loop step-sizes converges at a rate of order \( O(1/t^2) \) when \( \ell > 1 \) and at a rate of order \( O(1/t) \) when \( \ell = 1 \). FW with line-search converges at a rate of order \( O(1/t) \) when \( \rho = \frac{1}{4} \) and linearly when \( \rho = 2 \). In the latter setting, FW with line-search solves the problem exactly after \( |\text{supp}(x^*)| \) iterations.

In the following remark to Theorem 4.5, we discuss how to relax strict complementarity.

**Remark 4.6** (Relaxation of strict complementarity). In the proof of Theorem 4.5, strict complementarity is only needed to guarantee that after a specific iteration \( S \in \{1, \ldots, T - 1\} \), for all \( t \in \{S, \ldots, T - 1\} \), it holds that \( p_t \in \text{vert}(\mathcal{C}^*) \), 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 \( t \in \{S, \ldots, T - 1\} \): Consider, for example, the minimization of \( f(x) = \frac{1}{2} \| x - b \|_2^2 \) for \( b = (0, 1/2, 1/2)^T \in \mathbb{R}^3 \) over the probability simplex \( \mathcal{C} = \text{conv} \{ (e^{(1)}, e^{(2)}, e^{(3)}) \} \). Note that \( \mathcal{C}^* = \text{conv} \{ (e^{(2)}, e^{(3)}) \} \). It holds that \( x^* = b \) and \( \nabla f(x^*) = (0, 0, 0)^T \in \mathbb{R}^3 \). Thus, strict complementarity is violated. However, for any \( x_t = (u, v, w)^T \in \mathbb{R}^3 \) 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.

The results in Figure 2, see Section 7.1.2 for details, show that when the feasible region \( C \) is a polytope, \( f = \frac{1}{2} \| x - \rho \bar{x} \|_2^2 \), where \( \rho \in \{ \frac{1}{4}, 2 \} \), is strongly convex, the constrained optimal solution \( x^* = \arg\min_{x \in C} f(x) \) lies in the relative interior of an at least one-dimensional face of \( C \), FW with open-loop step-sizes converges at a rate of order \( O(1/t^2) \), where \( \ell \in \mathbb{N}_{\geq 2} \), converges at a rate of order \( O(1/t^2) \) and FW with open-loop step-size \( \eta_t = \frac{1}{t+1} \) converges at a rate of order \( O(1/t) \). For the same setting, FW with line-search either converges at a rate of order \( O(1/t) \) when \( \rho = \frac{1}{4} \) or linearly when \( \rho = 2 \). We have thus demonstrated both theoretically and in practice that there exist settings for which FW with open-loop step-sizes converges non-asymptotically faster than FW with line-search or short-step.

**5. Algorithmic variants**

In Section 4, we established that when the feasible region \( C \) is a polytope, the objective \( f \) is strongly convex, and the unique minimizer \( x^* = \arg\min_{x \in C} f(x) \) lies in the relative interior of an at least one-dimensional face \( \mathcal{C}^* \) of \( C \), FW with open-loop step-size \( \eta_t = \frac{1}{t+1} \) converges at a rate of order \( O(1/t^2) \). Combined with the convergence-rate lower bound of \( \Omega(1/t^{1+\epsilon}) \) for any \( \epsilon > 0 \) for FW with line-search or short-step by Wolfe (1970), this characterizes a problem setting for which FW with open-loop step-sizes converges non-asymptotically faster than FW with line-search or short-step. However, our accelerated convergence rate only holds when
strict complementarity or similar assumptions, see Remark 4.6, hold. Similarly, the accelerated convergence rate of MFW (Li et al., 2021) in the described setting also relies on the assumption of strict complementarity.

Here, we address this gap in the literature and present two FW variants employing open-loop step-sizes that admit convergence rates of order $O(1/t^2)$ in the setting of the lower bound due to Wolfe (1970) without relying on the assumption of strict complementarity.

### 5.1 Decomposition-invariant pairwise Frank-Wolfe algorithm

Using the proof blueprint from Section 3.2, we derive accelerated convergence rates for the decomposition-invariant pairwise Frank-Wolfe algorithm (DIFW) (Garber and Meshi, 2016) in the setting of the lower bound due to Wolfe (1970). DIFW with line-search or step-size as in Option 1 in Garber and Meshi (2016, Algorithm 3) converges linearly when the feasible region is a specific type of polytope and the objective function is strongly convex. Benefits of DIFW are that the convergence rate does not depend on the dimension of the problem but the sparsity of the optimal solution $x^* \in \text{argmin}_{x \in C} f(x)$, that is, $|\text{supp}(x^*)| = |\{x^*_i \neq 0 \mid i \in \{1, \ldots, d\}\}| \ll d$, and it is not necessary to maintain a convex combination of the iterate $x_t$ throughout the algorithm’s execution. The latter property leads to reduced memory overhead compared to other variants of FW that admit linear convergence rates in the setting of Wolfe (1970). The main drawback of DIFW is that the method is not applicable to general polytopes, but only feasible regions that are similar to the simplex, that is, of the form described below.

**Definition 5.1 (Simplex-like polytope (SLP)).** Let $C \subseteq \mathbb{R}^d$ be a polytope such that $C$ can be described as $C = \{x \in \mathbb{R}^d \mid Ax = b\}$ for $A \in \mathbb{R}^{m \times d}$ and $b \in \mathbb{R}^m$ for some $m \in \mathbb{N}$ and all vertices of $C$ lie on the Boolean hypercube $\{0, 1\}^d$. Then, we refer to $C$ as a *simplex-like polytope* (SLP).

Examples of SLPs are the probability simplex and the flow, perfect matchings, and marginal polytopes, see Garber and Meshi (2016) and references therein for more details. In this section, we show that DIFW with open-loop step-size $\eta_t = \frac{8}{t+8}$ admits a convergence rate of order up to $O(1/t^2)$ when optimizing a function satisfying (HEB) over a SLP.

**Algorithm 4:** Decomposition-invariant pairwise Frank-Wolfe algorithm (DIFW) (Garber and Meshi, 2016)

| Input: $x_0 \in C$, step-sizes $\eta_t \in [0, 1]$ for $t \in \{0, \ldots, T - 1\}$. |
|---|
| 1 $x_1 \in \text{argmin}_{p \in C} \langle \nabla f(x_0), p - x_0 \rangle$ |
| 2 for $t = 0, \ldots, T - 1$ do |
| 3 $p_t^+ \in \text{argmin}_{p \in C} \langle \nabla f(x_t), p - x_t \rangle$ |
| 4 Define the vector $\nabla f(x_t) \in \mathbb{R}^d$ entry-wise for all $i \in \{1, \ldots, d\}$: |
| $$(\nabla f(x_t))_i = \begin{cases} (\nabla f(x_t))_i, & \text{if } (x_t)_i > 0 \\ -\infty, & \text{if } (x_t)_i = 0. \end{cases}$$ |
| 5 $p_t^- \in \text{argmin}_{p \in C} \langle -\nabla f(x_t), p - x_t \rangle$ |
| 6 Let $\delta_t$ be the smallest natural number such that $2^{-\delta_t} \leq \eta_t$, and define the new step-size $\gamma_t \leftarrow 2^{-\delta_t}$. |
| 7 $x_{t+1} \leftarrow x_t + \gamma_t(p_t^+ - p_t^-)$ |
| end |

#### 5.1.1 Algorithm overview

We refer to $p_t^+$ and $p_t^-$ as the FW vertex and away vertex, respectively. At iteration $t \in \{0, \ldots, T\}$, consider the representation of $x_t$ as a convex combination of vertices of $C$, that is, $x_t = \sum_{i=0}^{t-1} \lambda_{p_i,t} p_i$, where $p_i \in \text{vert}(C)$ and $\lambda_{p_i,t} \geq 0$ for all $i \in \{0, \ldots, t - 1\}$ and $\sum_{i=0}^{t-1} \lambda_{p_i,t} = 1$. DIFW takes a step in the direction $\frac{p_t^+ - p_t^-}{\|p_t^+ - p_t^-\|_2}$.
which moves weight from the away vertex $p_i^-$ to the FW vertex $p_i^+$. Note that DIFW does not need to actively maintain a convex combination of $x_t$ because of the assumption that the feasible region is a SLP.

5.1.2 Convergence rate of order $O(1/t)$

We first derive a baseline convergence rate of order $O(1/t)$ for DIFW with open-loop step-size $\eta_t = \frac{8}{t+8}$.

**Proposition 5.2** (Convergence rate of order $O(1/t)$). Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a SLP of diameter $\delta > 0$ and let $f: \mathcal{C} \to \mathbb{R}$ be a convex and $L$-smooth function with unique minimizer $x^* \in \text{argmin}_{x \in \mathcal{C}} f(x)$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{8}{t+8}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 4 with open-loop step-size $\eta_t$, it holds that $h_t \leq \frac{32L\delta^2}{t+8} = \eta_{t-4}L\delta^2$ for all $t \in \{1, \ldots, T\}$.

**Proof.** Let $t \in \{0, \ldots, T-1\}$. Feasibility of $x_t$ follows from Lemma 1 in Garber and Meshi (2016). Further, in the proof of Lemma 3 in Garber and Meshi (2016), it is shown that

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

Consider an irreducible representation of $x_t$ as a convex sum of vertices of $\mathcal{C}$, that is, $x_t = \sum_{i=0}^{k} \lambda_{p_i,t} p_i$ such that $p_i \in \text{vert}(\mathcal{C})$ and $\lambda_{p_i,t} > 0$ for all $i \in \{0, \ldots, k\}$, where $k \in \mathbb{N}$. By Observation 1 in Garber and Meshi (2016), it holds that $\langle \nabla f(x_t), p_i \rangle \leq \langle \nabla f(x_t), p_t^- \rangle$ for all $i \in \{0, \ldots, k\}$. Thus, $\langle \nabla f(x_t), x_t - p_t^- \rangle \leq \langle \nabla f(x_t), x_t - p_t^+ \rangle \leq \langle \nabla f(x_t), x_t - x_t \rangle = 0$. Plugging this inequality into (5.1), using $\langle \nabla f(x_t), p_t^+ - x_t \rangle \leq -h_t$, and using $h_1 \leq L\delta^2$, which is derived in the proof of Theorem 1 in Garber and Meshi (2016), we obtain

$$h_{t+1} \leq h_t + \eta_t \langle \nabla f(x_t), p_t^+ - x_t \rangle + \frac{\eta_t \langle \nabla f(x_t), x_t - p_t^- \rangle + \eta_t^2 L \delta^2}{2} \leq \frac{(1 - \frac{\eta_t}{2})h_t + \frac{\eta_t^2 L \delta^2}{2}}{1 - \frac{\eta_t}{2}} \leq \prod_{i=1}^{t} \left(1 - \frac{\eta_i}{2}\right) h_1 + \frac{L \delta^2}{2} \left(\prod_{i=1}^{t} \left(1 - \frac{\eta_i}{2}\right)\right) \leq \frac{5 \cdot 6 \cdot 7 \cdot 8}{(t+5)(t+6)(t+7)(t+8)} h_1 + \frac{L \delta^2}{2} \left(\prod_{i=1}^{t} \left(1 - \frac{\eta_i}{2}\right)\right) \leq 64L \delta^2 \frac{1}{(t+7)(t+8)} + \frac{t}{(t+7)(t+8)} \leq 32L \delta^2 \frac{1}{t+8}.$$ 

\[\square\]

5.1.3 Convergence rate of order up to $O(1/t^2)$

Then, acceleration follows almost immediately from the analysis performed in Garber and Meshi (2016).

**Theorem 5.3** (Convergence rate of order up to $O(1/t^2)$). Let $\mathcal{C} \subseteq \mathbb{R}^d$ be a SLP of diameter $\delta > 0$ and let $f: \mathcal{C} \to \mathbb{R}$ be a convex and $L$-smooth function satisfying a $(\mu, \theta)$-HEB for some $\mu > 0$ and $\theta \in [0, 1/2]$. Let $T \in \mathbb{N}$ and $\eta_t = \frac{8}{t+8}$ for all $t \in \mathbb{Z}$. Then, for the iterates of Algorithm 4 with open-loop step-size $\eta_t$, it holds that

$$h_t \leq \max \left\{ \frac{1}{(1-t)} \frac{L \delta^2}{2}, \left(\eta_{t-2}2\mu L \delta^2 \sqrt{\text{supp}(x^*)}\right)^{1/\theta} + \frac{\eta_{t-2}^2 L \delta^2}{2} \right\}.$$ 

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

18
Proof. Let $t \in \{1, \ldots, T - 1\}$. We can extend Lemma 3 in Garber and Meshi (2016) from $\alpha_f$-strongly convex functions to convex functions satisfying (HEB). Strong convexity is only used to show that $\Delta_t := \sqrt{\frac{2\text{supp}(x^*)}{\alpha_f}} h_t$ satisfies $\Delta_t \geq \frac{1}{2} \text{supp}(x^*) \|x_t - x^*\|_2$. Here, we instead define $\Delta_t := \sqrt{\text{supp}(x^*)} \mu h_t^\theta$ for a function $f$ satisfying a $(\mu, \theta)$-(HEB). Then, $\Delta_t \geq \frac{1}{2} \text{supp}(x^*) \|x_t - x^*\|_2$. By Lemma 3 in Garber and Meshi (2016), we have $h_{t+1} \leq h_t - \frac{\eta t^1}{4\mu \sqrt{\text{supp}(x^*)}} + \eta t^2 \delta^2$. Combined with (5.2),

$$h_{t+1} \leq \left(1 - \frac{\eta}{4}\right) h_t - \frac{\eta t^1}{4\mu \sqrt{\text{supp}(x^*)}} + \eta t^2 \delta^2.$$ (5.4)

Using the same proof technique as in Lemma 3.5, we prove that

$$h_t \leq \max \left\{ \left(\frac{\eta - 2\eta t^1}{2}\right)^{1/(2(1-\theta))} \frac{L\delta^2}{2}, \left(\eta - 2\eta t^1 \left(2\mu L\delta^2 \sqrt{\text{supp}(x^*)}\right)^2 \right)^{1/(2(1-\theta))} + \eta t^2 \delta^2 \right\}.$$ (5.5)

for all $t \in \{1, \ldots, T\}$, which then implies (5.3). For $t = 1$, $h_1 \leq \frac{L\delta^2}{2}$ and (5.5) holds. Suppose that (5.5) is satisfied for a specific iteration $t \in \{1, \ldots, T - 1\}$. We distinguish between two cases. First, suppose that $h_t \leq (\eta t^2 \mu L\delta^2 \sqrt{\text{supp}(x^*)})^{1/(1-\theta)}$. Plugging this bound on $h_t$ into (5.4) yields $h_{t+1} \leq (\eta t^2 \mu L\delta^2 \sqrt{\text{supp}(x^*)})^{1/(1-\theta)} + \eta t^2 \delta^2$. Next, suppose that $h_t \geq (\eta t^2 \mu L\delta^2 \sqrt{\text{supp}(x^*)})^{1/(1-\theta)}$. Plugging this bound on $h_t$ into (5.4) and using the induction assumption yields

$$h_{t+1} \leq \left(1 - \frac{\eta}{4}\right) h_t + 0$$ (5.6)

$$= \frac{t + 6}{t + 8} h_t$$

$$\leq \frac{\eta}{\eta - 2} \left(\frac{\eta - 2\eta t^1}{2}\right)^{1/(2(1-\theta))} \frac{L\delta^2}{2}, \left(\eta - 2\eta t^1 \left(2\mu L\delta^2 \sqrt{\text{supp}(x^*)}\right)^2 \right)^{1/(2(1-\theta))} + \eta t^2 \delta^2 \right\}.$$ (5.5)

where the last inequality holds due to $\frac{\eta}{\eta - 2} (\eta - 2\eta t^1)^{1/(2(1-\theta))} \leq (\eta - 1\eta)^{1/(2(1-\theta))}$ for $\frac{\eta}{\eta - 2} \in [0, 1]$ and $1/(2(1-\theta)) \in [1/2, 1]$. In either case, (5.5) is satisfied for $t + 1$. By induction, the theorem follows. \qed

Below, we discuss the technical necessity for $\eta_t = \frac{8}{t+8}$ instead of $\eta_t = \frac{4}{t+4}$ in Theorem 5.3.

Remark 5.4 (Necessity of $\eta_t = \frac{8}{t+8}$). Note that Inequality (5.4) is responsible for making our usual proof with $\eta_t = \frac{4}{t+4}$, $t \in \mathbb{Z}$, impossible. Indeed, for $\eta_t = \frac{4}{t+4}$, $\left(1 - \frac{\eta}{4}\right) = \frac{t+3}{t+4}$, which is not enough progress in, for example, (5.6) assuming that $\theta = \frac{1}{2}$, to obtain a convergence rate of order $O(1/t^2)$.

5.2 Away-step Frank-Wolfe algorithm

In this section, we derive a version of the away-step Frank-Wolfe algorithm (AFW) (Guélat and Marcotte, 1986; Lacoste-Julien and Jaggi, 2015) with step-size $\eta_t = \frac{4}{t+4}$ that admits a convergence rate of order up to $O(1/t^2)$ when optimizing a function satisfying (HEB) over a polytope.

5.2.1 Algorithm overview

For better understanding, we first discuss AFW with line-search, which is presented in Algorithm 6. At iteration $t \in \{0, \ldots, T\}$, we can write $x_t = \sum_{i=t-1}^{t-1} \lambda_{p,t} p_i$, where $p_i \in \text{vert}(C)$ and $\lambda_{p,t} \geq 0$ for all $i \in \{0, \ldots, t - 1\}$ and $\sum_{i=0}^{t-1} \lambda_{p,t} = 1$. We refer to $S_t := \{p_i \mid \lambda_{p,t} > 0\}$ as the active set at iteration $t$. Note that maintaining the active set can incur a significant memory overhead. However, with AFW, instead of being limited to
Algorithm 5: Away-step Frank-Wolfe algorithm (AFW) with open-loop step-sizes

**Input:** $x_0 \in \text{vert}(C)$, step-sizes $\eta_t \in [0,1]$ for $t \in \{0,\ldots,T-1\}$.

1. $S_0 \leftarrow \{x_0\}$
2. $\lambda_{p,0} \leftarrow \begin{cases} 1, & \text{if } p = x_0 \\ 0, & \text{if } p \in \text{vert}(C) \setminus \{x_0\} \end{cases}$
3. $\ell_0 \leftarrow 0$ \quad $\triangleright \ell_t$: number of progress steps performed before iteration $t$
4. for $t = 0,\ldots,T-1$ do
   5. $p_t^{FW} \in \arg\min_{p \in C} \langle \nabla f(x_t), p - x_t \rangle$
   6. $p_t^A \in \arg\max_{p \in S_t} \langle \nabla f(x_t), p - x_t \rangle$
   7. if $\langle \nabla f(x_t), p_t^{FW} - x_t \rangle \leq \langle \nabla f(x_t), x_t - p_t^A \rangle$ then
      8. $d_t \leftarrow p_t^{FW} - x_t; \eta_{t,\text{max}} \leftarrow 1$
   else
      9. $d_t \leftarrow x_t - p_t^A; \eta_{t,\text{max}} \leftarrow \frac{\lambda_t^{A,t}}{1 - \lambda_t^{A,t}}$
   end
   10. $\gamma_t \leftarrow \min \{\eta_t, \eta_{t,\text{max}}\}$
   11. $x_{t+1} \leftarrow x_t + \gamma_td_t$
   12. if $\langle \nabla f(x_t), p_t^{FW} - x_t \rangle \leq \langle \nabla f(x_t), x_t - p_t^A \rangle$ then
      13. $\lambda_{p,t+1} \leftarrow \begin{cases} (1 - \gamma_t)\lambda_{p,t} + \gamma_t, & \text{if } p = p_t^{FW} \\ (1 - \gamma_t)\lambda_{p,t}, & \text{if } p \in \text{vert}(C) \setminus \{p_t^{FW}\} \end{cases}$
   else
      14. $\lambda_{p,t+1} \leftarrow \begin{cases} (1 + \gamma_t)\lambda_{p,t} - \gamma_t, & \text{if } p = p_t^A \\ (1 + \gamma_t)\lambda_{p,t}, & \text{if } p \in \text{vert}(C) \setminus \{p_t^A\} \end{cases}$
   end
   15. $S_{t+1} \leftarrow \{p \in \text{vert}(C) \mid \lambda_{p,t+1} > 0\}$
   16. if $(\eta_t - \gamma_t)\langle \nabla f(x_t), p_t^A - p_t^{FW} \rangle \leq (\eta_t^2 - \gamma_t^2)L\delta^2$ then
      17. $\ell_{t+1} \leftarrow \ell_t + 1$ \quad $\triangleright$ progress step
   else
      18. $\ell_{t+1} \leftarrow \ell_t$ \quad $\triangleright$ non-progress step
   end
5. end

Algorithm 6: Away-step Frank-Wolfe algorithm (AFW) with line-search (Guélat and Marcotte, 1986)

1. Identical to Algorithm 5, except that Lines 3, 20, 21, 22, 23, and 24 have to be deleted and Line 12 has to be replaced by $\gamma_t \leftarrow \arg\min_{\gamma \in [0,\eta_{t,\text{max}}]} f(x_t + \gamma d_t)$.

Taking a step in the direction of a vertex $p_t^{FW} \in \text{vert}(C)$ as in Line 2 of vanilla FW, we are also able to take an away step: Compute $p_t^A \in \arg\max_{p \in S_t} \langle \nabla f(x_t), p - x_t \rangle$ and take a step away from vertex $p_t^A$, removing weight from vertex $p_t^A$ and adding it to all other vertices in the active set. Away steps facilitate the option of taking drop steps. A drop step occurs when a vertex gets removed from the active set. In case $x^*$ lies in the relative interior of an at least one-dimensional face $C^*$ of $C$, drop steps allow AFW to get rid of bad vertices in the convex combination representing $x_t$, that is, vertices not in $C^*$. As soon as the optimal face is reached, that is, $x_t \in C^*$, the problem becomes that of having the optimal solution in the relative interior of $C^*$, for which FW with line-search admits linear convergence rates.

We next explain AFW with step-size $\eta_t = \frac{4}{t+4}$ presented in Algorithm 5, which requires a slight modification of the version presented in Lacoste-Julien and Jaggi (2015). The main idea is to replace
line-search with the open-loop step-size \( \eta_t = \frac{1}{t + t} \). However, as we motivate in detail below, at iteration \( t \in \{0, \ldots, T - 1\} \), AFW’s step-length is \( \eta_t \), where \( 0 = \ell_0 \leq \ell_1 \leq \ldots \leq \ell_T - 1 \leq T - 1 \), that is, AFW may perform multiple steps of the same length. Let \( t \in \{0, \ldots, T - 1\} \). Note that for \( d_t \) obtained from either Line 8 or Line 10 in Algorithm 5, it holds that \( \langle \nabla f(x_t), d_t \rangle \leq \langle \nabla f(x_t), p_t^{FW} - p_t^A \rangle / 2 \). By L-smoothness,

\[
h_{t+1} \leq h_t - \gamma_t (\nabla f(x_t), p_t^A - p_t^{FW}) + \frac{\gamma_t^2 L \delta^2}{2}.
\]

(5.7)

Working towards a convergence rate of order up to \( O(1/t^2) \), we need to characterize a subsequence of steps for which an inequality of the form (3.3) holds. To do so, let

\[
g_t(\gamma) := -\frac{\gamma (\nabla f(x_t), p_t^A - p_t^{FW})}{2} + \frac{\gamma^2 L \delta^2}{2} \quad \text{for } \gamma \in [0, 1].
\]

We refer to all iterations \( t \in \{0, \ldots, T - 1\} \) such that \( g_t(\gamma_t) \leq g_t(\eta_t) \) as progress steps and denote the number of progress steps performed before iteration \( t \in \{0, \ldots, T\} \) by \( \ell_t \), see Lines 3, 12, and 20-24 of Algorithm 5. Thus, a progress step occurs during iteration \( t \) if and only if the inequality in Line 20 is satisfied, which necessitates the computation of the smoothness constant \( L \) of \( f \) prior to the execution of the algorithm. A non-drop step is always a progress step as \( \gamma_t = \eta_t \) and the following lemma shows that drop steps which are non-progress steps do not increase the primal gap.

**Lemma 5.5** (Drop-step characterization). Let \( g : [0,1] \to \mathbb{R} \) be defined via \( g(\eta) := -\eta A + \eta^2 B \), where \( A, B > 0 \). For \( t \in \mathbb{N} \), let \( \eta_t = \frac{1}{t + t} \) and \( \gamma_t \in [0, \eta_t] \). Then, \( g(\gamma_t) \leq g(0) \) or \( g(\gamma_t) \leq g(\eta_t) \).

**Proof.** By case distinction. Let \( t \in \mathbb{N} \). Case 1: \( g(\eta_t) \leq g(0) \). By convexity, \( g(\gamma_t) = g(\lambda \eta_t + (1 - \lambda) 0) \leq \lambda g(\eta_t) + (1 - \lambda) g(0) \leq g(0) = 0 \) where \( \lambda \in [0, 1] \). Case 2: \( g(\eta_t) > g(0) \). Then, \( \eta_t > \eta^* \in \text{argmin}_{\eta_t \in [0, \eta_t]} g(\eta) \), as \( g \) is monotonously decreasing in the interval \([0, \eta^*] \). If \( \eta^* \leq \gamma_t \), then \( g(\gamma_t) \leq g(\eta_t) \) due to \( g \) being monotonously increasing in \([\eta^*, \eta_t] \). If \( \eta^* \geq \gamma_t \), then \( g(\gamma_t) \leq g(0) \), as \( g \) is monotonously decreasing in \([0, \eta^*] \).

Thus, a drop step is either a progress step and \( h_{t+1} \leq h_t + g_t(\eta_t) \), or \( h_{t+1} \leq h_t \).

**Lemma 5.6** (Number of progress steps). Let \( C \subseteq \mathbb{R}^d \) be a compact convex set of diameter \( \delta > 0 \), let \( f : C \to \mathbb{R} \) be a convex and L-smooth function. Let \( T \in \mathbb{N} \) and \( \eta_t = \frac{1}{t+1} \) for all \( t \in \mathbb{Z} \). Then, for all iterations \( t \in \{0, \ldots, T\} \) of Algorithm 5 with step-size \( \eta_t \), it holds that \( \ell_t \geq \lceil t/2 \rceil \geq t/2 \).

**Proof.** Since all non-drop steps are progress steps and \( S_t \), where \( t \in \{0, \ldots, T\} \), has to contain at least one vertex of \( C \), there cannot occur more drop steps than non-drop steps. Thus, \( \ell_t \geq \lceil t/2 \rceil \geq t/2 \).

5.2.2 Convergence rate of order \( O(1/t) \)

We first derive a baseline convergence rate of order \( O(1/t) \) for AFW with step-size \( \eta_t = \frac{t}{t+1} \).

**Proposition 5.7** (Convergence rate of order \( O(1/t) \)). Let \( C \subseteq \mathbb{R}^d \) be a compact convex set of diameter \( \delta > 0 \), let \( f : C \to \mathbb{R} \) be a convex and L-smooth function. Let \( T \in \mathbb{N} \) and \( \eta_t = \frac{1}{t+1} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 5 with step-size \( \eta_t \), it holds that \( h_t \leq \frac{16L\delta^2}{t+6} = \eta_t^2 4L\delta^2 \) for all \( t \in \{1, \ldots, T\} \).

**Proof.** Let \( t \in \{0, \ldots, T - 1\} \) and suppose that during iteration \( t \), we perform a progress step. Either \( d_t = p_t^{FW} - x_t \), or \( d_t = x_t - p_t^A \) and by Line 7 of Algorithm 5, \( \langle \nabla f(x_t), x_t - p_t^A \rangle \leq \langle \nabla f(x_t), p_t^{FW} - x_t \rangle \). In either case, by L-smoothness,

\[
h_{t+1} \leq h_t - \gamma_t (\nabla f(x_t), x_t - p_t^{FW}) + \frac{\gamma_t^2 L \delta^2}{2} \leq (1 - \gamma_t) h_t + \frac{\gamma_t^2 L \delta^2}{2}.
\]

(5.8)
By Lemma 5.5, since non-progress steps do not increase the primal gap, we can limit our analysis to the subsequence of iterations corresponding to progress steps, \( \{t(k)\}_{k \in \{0, \ldots, T\}} \), for which, by (5.8), it holds that

\[
h_{t(k)} \leq (1 - \eta_{t(k)})h_{t(k)} + \frac{\eta_{t(k)}^2 L \delta^2}{2} = (1 - \eta_{k})h_{t(k)} + \frac{\eta_{k}^2 L \delta^2}{2} \tag{5.9}
\]

for all \( k \in \{0, \ldots, T - 1\} \). Since the first step is a non-drop step and thus a progress step, \( h_{t(1)} \leq h_1 \leq \frac{L \delta^2}{2} \).

By similar arguments as in the proof of Proposition 3.1 starting with (3.1), we obtain the bound \( h_{t(k)} \leq \frac{8L \delta^2}{\mu + \vartheta} \) for all \( k \in \{1, \ldots, T\} \). Since non-progress steps do not increase the primal gap and by Lemma 5.6, \( h_t \leq h_{t(k)} \leq \frac{8L \delta^2}{\mu + \vartheta} \leq \frac{16L \delta^2}{\mu + \vartheta} = \eta_{k+2}L \delta^2 \) for all \( t \in \{1, \ldots, T\} \).

\[\Box\]

### 5.2.3 Convergence rate of order up to \( O(1/t^2) \)

The introduction of away steps introduces another type of scaling inequality based on the pyramidal width, a constant depending on the feasible region, see Lacoste-Julien and Jaggi (2015) for more details.

**Lemma 5.8** (Lacoste-Julien and Jaggi, 2015). Let \( C \subseteq \mathbb{R}^d \) be a polytope with pyramidal width \( \omega > 0 \) and let \( f: C \to \mathbb{R} \) be a convex function with unique minimizer \( x^* \in \text{argmin}_{x \in C} f(x) \). Let \( p^{FW} \in \text{argmin}_{p \in C}(\nabla f(x), p) \) and \( p^L \in \text{argmax}_{p \in S}(\nabla f(x), p) \) for some \( S \subseteq \text{vert}(C) \) such that \( x \in \text{conv}(S) \). Then, it holds that

\[
\frac{\langle \nabla f(x), p^L - p^{FW} \rangle}{\omega} \geq \frac{\langle \nabla f(x), x - x^* \rangle}{\|x - x^*\|_2}.
\]

(Scaling-A)

For example, the pyramidal width of the unit cube in \( \mathbb{R}^d \) satisfies \( \omega \geq 2/\sqrt{d} \) (Lacoste-Julien and Jaggi, 2015) and the pyramidal width of the \( \ell_1 \)-ball in \( \mathbb{R}^d \) satisfies \( \omega \geq 1/\sqrt{d - 1} \) (Wirth et al., 2023). Combining (Scaling-A) and (Scaling-HEB) leads to a subsequence of primal gaps of the form (3.3) and a convergence rate of order up to \( O(1/t^2) \) for Algorithm 5.

**Theorem 5.9** (Convergence rate of order up to \( O(1/t^2) \)). Let \( C \subseteq \mathbb{R}^d \) be a polytope of diameter \( \delta > 0 \) and pyramidal width \( \omega > 0 \) and let \( f: C \to \mathbb{R} \) be a convex and \( L \)-smooth function satisfying a \((\mu, \theta)\)-HEB for some \( \mu > 0 \) and \( \theta \in [0, 1/2] \) with unique minimizer \( x^* \in \text{argmin}_{x \in C} f(x) \). Let \( T \in \mathbb{N} \) and \( \eta_t = \frac{4}{t + 4} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 5 with step-size \( \eta_t \), it holds that

\[
h_t \leq \max \left\{ \eta_{t/2 - 2}^{1/(1 - \theta)} \frac{L \delta^2}{\omega}, \left( \eta_{t/2 - 2}^{1/(1 - \theta)} \frac{2 \mu L \delta^2}{\omega} \right)^{1/(1 - \theta)} + \eta_{t/2 - 2}^{1/(1 - \theta)} \frac{L \delta^2}{2} \right\} \tag{5.10}
\]

for all \( t \in \{1, \ldots, T\} \).

**Proof.** Let \( t \in \{0, \ldots, T - 1\} \). By (5.7), (Scaling-A), convexity of \( f \), and (Scaling-HEB), it holds that

\[
h_{t+1} \leq h_t - \frac{\gamma \omega}{2\|\nabla f(x_t) - x_t - x^*\|_2} + \frac{\gamma^2 \delta^2}{2} \leq h_t - \frac{\gamma \omega}{2\mu} h_t^{1 - \theta} + \frac{\gamma^2 \delta^2}{2}.
\]

Thus, by Lemma 5.5, non-progress steps satisfy \( h_{t+1} \leq h_t \) and progress steps satisfy

\[
h_{t+1} \leq h_t - \frac{\eta_t \omega}{2\mu} h_t^{1 - \theta} + \frac{\eta_t L \delta^2}{2}. \tag{5.11}
\]

Since non-progress steps do not increase the primal gap, we can limit our analysis to the subsequence of iterations corresponding to progress steps, \( \{t(k)\}_{k \in \{0, \ldots, T\}} \), for which, by (5.11), it holds that

\[
h_{t(k+1)} \leq h_{t(k)} - \frac{\eta_{t(k)} \omega}{2\mu} h_{t(k)}^{1 - \theta} + \frac{\eta_{t(k)}^2 L \delta^2}{2} = h_{t(k)} - \frac{\eta_k \omega}{2\mu} h_{t(k)}^{1 - \theta} + \frac{\eta_k^2 L \delta^2}{2}.
\]

Combined with (5.9), it thus holds that

\[
h_{t(k+1)} \leq (1 - \frac{\eta_k}{2})h_{t(k)} - \frac{\eta_k \omega}{4\mu} h_{t(k)}^{1 - \theta} + \frac{\eta_k^2 L \delta^2}{2}. \tag{5.12}
\]
for all $k \in \{1, \ldots, \ell_T - 1\}$. Since the first step is a non-drop step and thus a progress step, $h_{\ell_T} = h_1 \leq \frac{L\delta^2}{2}$. Inequality 5.12 allows us to apply Lemma 3.5 with $A = \frac{\omega}{d_n}$, $B = \frac{L\delta^2}{2}$, $C = 1$, $C_{\ell_T} = 1$ for all $k \in \{1, \ldots, \ell_T - 1\}$, $\psi = \theta$, and $S = 1$, resulting in $h_{\ell_T} \leq \max \left\{ \frac{\nu_1}{\eta_1} \left( 1 - \theta \right) \frac{L\delta^2}{2} \right\} \left( \frac{m - 2\mu L\delta^2}{\omega} \right)^{1/(1-\theta)} + \eta_1 \frac{L\delta^2}{2} \right\}$ for all $k \in \{1, \ldots, \ell_T\}$, where we used that $\eta_1 \geq \eta_0 = 1$. Since non-progress steps do not increase the primal gap and by Lemma 5.6, (5.10) holds for all $t \in \{1, \ldots, T\}$. 

6. Kernel herding

In this section, we explain why FW with open-loop step-sizes converges at a rate of order $O(1/t^2)$ in the kernel-herding setting of Bach et al. (2012, Section 5.1 and Figure 3, right).

6.1 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 $\Phi: \mathcal{Y} \to \mathcal{H}$ 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 $\Phi$ is denoted by $k: (y, z) \mapsto k(y, z) = \langle \Phi(y), \Phi(z) \rangle_{\mathcal{H}}$ for $y, z \in \mathcal{Y}$. In kernel herding, the feasible region is usually the marginal polytope $C$, the convex hull of all functions $\Phi(y)$ for $y \in \mathcal{Y}$, that is, $C = \text{conv} \{ \{ \Phi(y) \mid y \in \mathcal{Y} \} \} \subseteq \mathcal{H}$. We consider a fixed probability distribution $p$ over $\mathcal{Y}$ and denote the associated mean element by $\mu = \mathbb{E}_{p(y)} \Phi(y) \in C$, where $\mu \in C$ 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 $\eta_t = \frac{1}{t+1}$:

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

(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-sizes to solve (OPT-KH) (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 $C$ 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 $\Phi(y) \in C$ for $y \in \mathcal{Y}$. Furthermore, FW constructs iterates of the form $x_t = \sum_{i=1}^{t} v_i \Phi(y_i)$, where $v = (v_1, \ldots, v_t)^T$ is a weight vector, that is, $\sum_{i=1}^{t} v_i = 1$ and $v_i \geq 0$ for all $i \in \{1, \ldots, t\}$, and $x_t$ corresponds to an empirical distribution $\tilde{p}_t$ over $\mathcal{Y}$ with empirical mean $\tilde{\mu}_t = \mathbb{E}_{\tilde{p}_t} \Phi(y) = \sum_{i=1}^{t} v_i \Phi(y_i) = x_t \in C$. Then, according to Bach et al. (2012), $\sup_{x \in H, \|x\|_H = 1} | \mathbb{E}_{p(y)} x(y) - \mathbb{E}_{\tilde{p}_t} x(y) | = \| \mu - \tilde{\mu}_t \|_{\mathcal{H}}$. Thus, a bound on $\| \mu - \tilde{\mu}_t \|_{\mathcal{H}}$ implies control on the error in computing the expectation for all $x \in H$ such that $\|x\|_H = 1$. In kernel herding, since the objective function is a quadratic, line-search and short-step are identical.

6.2 Explaining the phenomenon in Bach et al. (2012)

We briefly recall the infinite-dimensional kernel-herding setting of Bach et al. (2012, Section 5.1 and Figure 3, right), see also Wahba (1990, Section 2.1). Let $\mathcal{Y} = [0, 1]$ and

$$\mathcal{H} = \{ x: [0, 1] \to \mathbb{R} \mid x'(y) \in L^2([0, 1]), x(y) = \sum_{j=1}^{\infty} (a_j \cos(2\pi j y) + b_j \sin(2\pi j y)), a_j, b_j \in \mathbb{R} \}. \quad (6.1)$$

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

$$k(y, z) = \sum_{j=1}^{\infty} \frac{2}{(2\pi j)^2} \cos(2\pi j(y - z)) = \frac{1}{2} B_2(y - z - [y - z]) = \frac{1}{2} B_2([y - z]), \quad \text{(Bernoulli-kernel)}$$

23
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 \( \mu(z) = \int_{[0,1]} k(z,y)p(y)dy = \int_{[0,1]} \sum_{j=1}^{\infty} \frac{2}{(2\pi j)^2} \cos(2\pi j(z-y)) \cdot 1dy = \sum_{j=1}^{\infty} 0 = 0 \), where the integral and the sum can be interchanged due to the theorem of Fubini, see, for example, Reynen 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, \( f(x) = \frac{1}{2} \|x\|_H^2 \). For this setting, Bach et al. (2012) observed empirically that FW with open-loop step-size \( \eta_t = \frac{1}{\sqrt{t}} \) converges at a rate of order \( O(1/t^2) \), whereas FW with line-search converges at a rate of order \( O(1/t) \), see the reproduced plot in Figure 3a. The theorem below explains the accelerated convergence rate for FW with step-size \( \eta_t = \frac{1}{t^{1+a}} \).

**Theorem 6.1 (Kernel herding).** Let \( \mathcal{H} \) be the Hilbert space defined in (6.1), let \( k: \mathbb{R} \times \mathbb{R} \to \mathcal{H} \) be the kernel defined in (Bernoulli-kernel), let \( \Phi: [0,1] \to \mathcal{H} \) be the feature map associated with \( k \) restricted to \([0,1] \times \{0,1\} \), let \( C = \text{conv}\{\{\Phi(y) \mid y \in [0,1]\}\} \) be the marginal polytope, and let \( \mu = 0 \) such that \( f(x) = \frac{1}{2} \|x\|_H^2 \). Let \( T \in \mathbb{N} \) and \( \eta_t = \frac{1}{t^{1+a}} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 1 with step-size \( \eta_t \) and the LMO satisfying Assumption 3 (a tie-breaking rule), it holds that \( f(x_t) = 1/(24t^2) \) for all \( t \in \{1, \ldots, T\} \) such that \( t = 2^m \) for some \( m \in \mathbb{N} \).

We first provide a proof sketch for Theorem 6.1 and subsequently prove the theorem in detail.

**Sketch of proof for Theorem 6.1.** The main idea behind the proof is that FW with \( \eta_t = \frac{1}{t^{1+a}} \) leads to iterates \( x_t = \frac{1}{4} \sum_{i=1}^{t} \Phi(y_i) \) with \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i = 1, \ldots, t\} \) 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 2 of Algorithm 1 becomes \( p_t \in \text{argmin}_{p \in C} Df(x_t)(p - x_t) = \text{argmin}_{p \in C} Df(x_t)(p) \), 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 \( x_t = \frac{1}{4} \sum_{i=1}^{t} \Phi(y_i) \) for \( y_1, \ldots, y_t \in [0,1] \), it holds that \( Df(x_t)(x) = \langle \frac{1}{4} \sum_{i=1}^{t} \Phi(y_i), x \rangle_{\mathcal{H}} \). Then, for \( y \in [0,1] \), let

\[
g_t(y) := \langle \frac{1}{4} \sum_{i=1}^{t} \Phi(y_i), \Phi(y) \rangle_{\mathcal{H}} = \frac{1}{t} \sum_{i=1}^{t} k(y_i, y). \tag{6.2}
\]

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 \( \min_{p \in C} Df(x_t)(p) = \min_{p \in [0,1]} g_t(y) \) and the vertex returned by the LMO during iteration \( t \) is contained in the set \( \{\Phi(z) \mid z \in \text{argmin}_{y \in [0,1]} g_t(y)\} \). 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 \( \text{argmin}_{p \in C} Df(x_t)(p) \geq 2 \).

**Assumption 3.** The LMO of FW always returns \( p_t \in \text{argmin}_{p \in C} Df(x_t)(p) \) such that \( p_t = \Phi(z) \) for \( z = \min(\text{argmin}_{y \in [0,1]} g_t(y)) \).

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 6.4, without loss of generality, we can assume that FW starts at iterate \( x_1 = \Phi(y_1) \), where \( y_1 = 0 \). \( \square \)

To rigorously prove Theorem 6.1, we require the following four technical lemmas. In the lemma below, we prove several technical properties of kernel \( k \) as in (Bernoulli-kernel).

**Lemma 6.2.** Let \( \mathcal{H} \) be the Hilbert space defined in (6.1) and let \( k: \mathbb{R} \times \mathbb{R} \to \mathcal{H} \) be the kernel defined in (Bernoulli-kernel). For \( y, z \in [0,1] \) and \( n \in \mathbb{Z} \), it holds that \( k(y, z) = k(z, y) = k(|y - z|, 0) = \frac{1}{2} B_2(|y - z|) \) and \( k(y, z) = k(y, z + n) \).

**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,

\[
[a] = a, \quad [-a] = 1 - a, \quad B_2([a]) = a^2 - a + \frac{1}{6} = (1 - a)^2 - (1 - a) + \frac{1}{6} = B_2([-a]), \quad (6.3)
\]

\[
[1] = 0, \quad [-1] = 0, \quad B_2([1]) = B_2([-1]). \quad (6.4)
\]
By (6.3) and (6.4), for any \( y, z \in [0, 1] \), it holds that 
\[
k(y, z) = \frac{1}{2} B_2([y - z]) = \frac{1}{2} B_2([z - y]) = k(z, y).
\]

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 (6.3), 
\[
k(y, z) = \frac{1}{2} B_2([y - z]) = \frac{1}{2} B_2([y - z]) = k([y - z], 0).
\]

By (6.3) and (6.4), for any \( y, z \in [0, 1] \) such that \( [y - z] = 1 \). Then, by (6.4), 
\[
k(y, z) = \frac{1}{2} B_2([y - z]) = \frac{1}{2} B_2([y - z]) = k([y - z], 0).
\]

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, 
\[
k(y, z) = \frac{1}{2} B_2(y - z - [y - z]) = \frac{1}{2} B_2(y - z - n - [y - z - n]) = k(y, z + n).
\]

In the two lemmas below, we characterize \( \arg\min_{y \in [0, 1]} g_t(y) \), where \( g_t \) is defined as in (6.2).

**Lemma 6.3.** Let \( H \) be the Hilbert space defined in (6.1), let \( k: \mathbb{R} \times \mathbb{R} \rightarrow H \) be the kernel defined in (Bernoulli-kernel), let \( \Phi: [0, 1] \rightarrow H \) be the feature map associated with \( k \) restricted to \( [0, 1] \times [0, 1] \), let \( t \in \mathbb{N} \), let \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \ldots, t\}\} \), and let \( g_t \) be defined as in (6.2), that is, \( g_t(y) = \frac{1}{t} \sum_{i=1}^{t} k(y_i, y) \). Then, it holds that \( \arg\min_{y \in [0, 1]} g_t(y) = \{y_i + \frac{1}{2t} \mid i \in \{1, \ldots, t\}\} \).

**Proof.** Let \( t \in \mathbb{N} \) and \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} \mid i \in \{1, \ldots, t\}\} \). We stress that this does not imply that for all \( i \in \{1, \ldots, t\} \), \( y_i = \frac{i-1}{t} \). By Lemma 6.2, for all \( y \in [0, 1] \), it holds that 
\[
g_t(y) = \frac{1}{t} \sum_{i=1}^{t} k(y_i, y) = \frac{1}{2t} \sum_{i=1}^{t} (y_i - y)^2 - \frac{1}{2t} y^2 + \frac{1}{2t} \sum_{i=1}^{t} y_i^2 \]  
For \( y \in \left[ \frac{i-1}{t}, \frac{i}{t} \right] \), where \( i \in \{1, \ldots, t\} \), it holds that 
\[
g_t'(y) = \frac{1}{2} (2y - \frac{t-1}{t} - \frac{1}{t} \left| \{y_i < y : i \in \{1, \ldots, t\}\} \right| + \frac{1}{t} \left| \{y_i > y : i \in \{1, \ldots, t\}\} \right|).
\]
For \( y \in \left[ \frac{i-1}{t}, \frac{i}{t} \right] \), it holds that 
\[
g_t'(y) = \frac{1}{2} (2y - \frac{t-1}{t} - \frac{1}{t} \left| \{y_i < y : i \in \{1, \ldots, t\}\} \right| + \frac{1}{t} \left| \{y_i > y : i \in \{1, \ldots, t\}\} \right|).
\]
For \( y = \frac{i-1}{t} \), it holds that 
\[
g_t'(y) = \frac{1}{2} (2y - \frac{t-1}{t} - \frac{1}{t} \left| \{y_i < y : i \in \{1, \ldots, t\}\} \right| + \frac{1}{t} \left| \{y_i > y : i \in \{1, \ldots, t\}\} \right|).
\]
for all \( j \in \{1, \ldots, t-1\} \). Thus, \( g_t(\frac{i-1}{t} + \frac{1}{2t}) = \frac{1}{2t} \sum_{i=1}^{t} k(y_i, y) \) for all \( j \in \{1, \ldots, t-1\} \). Thus, \( g_t(\frac{i-1}{t} + \frac{1}{2t}) = \frac{1}{2t} \sum_{i=1}^{t} k(y_i, y) \) for all \( i, j \in \{1, \ldots, t\} \), proving the lemma.

**Lemma 6.4.** Let \( H \) be the Hilbert space defined in (6.1), let \( k: \mathbb{R} \times \mathbb{R} \rightarrow H \) be the kernel defined in (Bernoulli-kernel), let \( \Phi: [0, 1] \rightarrow H \) 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 (6.2), that is, 
\[
g_t(y) = \frac{1}{t} \sum_{i=1}^{t} k(y, y_i).
\]
Suppose that \( \arg\min_{y \in [0, 1]} g_t(y) = \{z_1, \ldots, z_k\} \subseteq [0, 1] \) for some \( k \in \mathbb{N} \). Let \( c \in \mathbb{R} \), let \( y_i = [y_i + c] \) for all \( i \in \{1, \ldots, t\} \), and let \( \tilde{g}_t(y) = \frac{1}{t} \sum_{i=1}^{t} k(y, y_i) \). Then, \( \arg\min_{z \in [0, 1]} \tilde{g}_t(z) = \{[z_1 + c], \ldots, [z_k + c]\} \).
Proof. It holds that
\[
\argmin_{z \in [0,1]} \tilde{g}_t(z) = \argmin_{z = [y+c], y \in \mathbb{R}} \tilde{g}_t(z)
\]
\[
= \argmin_{z = [y+c], y \in \mathbb{R}} \frac{1}{2t} \sum_{i=1}^{t} B_2([y_i + c] - [y + c])
\]
\[
= \argmin_{z = [y+c], y \in \mathbb{R}} \frac{1}{2t} \sum_{i=1}^{t} B_2([y_i + c - (y + c) - (-[y + c])])
\]
\[
= \argmin_{z = [y+c], y \in \mathbb{R}} \frac{1}{2t} \sum_{i=1}^{t} B_2([y_i - y - [y_i + c] + [y + c]])
\]
\[
= \argmin_{z = [y+c], y \in \mathbb{R}} \frac{1}{2t} \sum_{i=1}^{t} B_2([y_i - y])
\]
\[
= \{[z_1 + c], \ldots, [z_k + c]\},
\]
where the second-to-last equality is due to Lemma 6.2. \qed

In the lemma below, we leverage the previous lemmas to prove that FW with step-size \( \eta_t = \frac{1}{\frac{1}{2}t} \) leads to iterates \( x_t = \frac{1}{2} \sum_{i=1}^{t} \Phi(y_i) \) with \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} | i \in \{1, \ldots, t\}\} \) for all \( t = 2^m \), where \( m \in \mathbb{N} \).

Lemma 6.5. Let \( \mathcal{H} \) be the Hilbert space defined in (6.1), let \( k: \mathbb{R} \times \mathbb{R} \rightarrow \mathcal{H} \) be the kernel defined in (Bernoulli-kernel), let \( \Phi: [0,1] \rightarrow \mathcal{H} \) be the feature map associated with \( k \) restricted to \([0,1] \times [0,1] \), let \( C = \text{conv} \{\Phi(y) | y \in [0,1]\} \) be the marginal polytope, and let \( \mu = 0 \) such that \( f(x) = \frac{1}{2} \|x\|_2^2 \). Let \( T \in \mathbb{N} \) and \( \eta_t = \frac{1}{\frac{1}{2}t} \) for all \( t \in \mathbb{Z} \). Then, for the iterates of Algorithm 1 with step-size \( \eta_t \) and the LMO satisfying Assumption 3 it holds that \( x_t = \frac{1}{2} \sum_{i=1}^{t} \Phi(y_i) \) with \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} | i \in \{1, \ldots, t\}\} \) for all \( t \in \{1, \ldots, T\} \) such that \( t = 2^m \) for some \( m \in \mathbb{N} \).

Proof. Since \( \eta_0 = 1 \), it holds that \( x_1 = \Phi(y_1) \). By Lemma 6.4, 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 \( \eta_t = \frac{1}{\frac{1}{2}t} \), we obtain uniform weights, that is, \( x_i = \frac{1}{2} \sum_{j=1}^{i} \Phi(y_j) \), where \( y_j \in [0,1] \) for all \( i \in \{1, \ldots, t\} \). Suppose that \( t = 2^m \) for some \( m \in \mathbb{N} \). The proof that it holds that \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} | i \in \{1, \ldots, t\}\} \) 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 \( \{y_1, \ldots, y_t\} = \{\frac{i-1}{t} | i \in \{1, \ldots, t\}\} \). If we show that
\[
\{y_1, \ldots, y_{2t}\} = \{\frac{i-1}{2t} | i \in \{1, \ldots, 2t\}\},
\]
the statement of the lemma follows from induction. (6.5) 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 6.3 and Assumption 3, 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 \( \tilde{g}_{t}(y) = \left(\frac{1}{2t} \sum_{i=1}^{t} \Phi(y_i + \frac{1}{2t}) \right) \Phi(y) \), that is, we consider the decomposition \( g_{t+\ell}(y) = \frac{1}{2t} g_t(y) + \frac{1}{2t} \tilde{g}_t(y) \). By Lemma 6.3, \( \text{argmin}_{y \in [0,1]} g_t(y) = \{y_i + \frac{1}{2t} | i \in \{1, \ldots, t\}\} \subseteq [0,1] \) and by Assumption 3, \( y_{t+1} = \text{min} \left(\text{argmin}_{y \in [0,1]} g_t(y) \right) \). Thus, by Lemma 6.4, it holds that \( \text{min} \left(\text{argmin}_{y \in [0,1]} \tilde{g}_t(y) \right) = \text{min} \left(\text{argmin}_{y \in [0,1]} g_t(y) + \frac{1}{2t} \right) = y_{t+1} + \frac{1}{2} \in \{y_i + \frac{1}{2t} | i \in \{1, \ldots, t\}\} \). Thus, \( \text{min} \left(\text{argmin}_{y \in [0,1]} \tilde{g}_t(y) \right) \in \text{argmin}_{y \in [0,1]} g_t(y) \) and
\[
y_{t+\ell+1} = \text{min} \left(\text{argmin}_{y \in [0,1]} g_{t+\ell}(y) \right) = \text{min} \left(\text{argmin}_{y \in [0,1]} \tilde{g}_t(y) \right) = y_{t+1} + \frac{1}{2t}.
\]
By induction, \( y_{t+j} = y_j + \frac{1}{2t} \) for all \( j \in \{1, \ldots, t\} \), as required to conclude the proof. \qed

Finally, we prove Theorem 6.1.
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. The equivalence of FW with kernel herding leads to the study and discovery of new convergence rates for FW.

Proof of Theorem 6.1. By Lemma 6.5, \( x_t = \frac{1}{t} \sum_{i=1}^{t} \Phi(\frac{i-1}{t}) \) and, since \( \mu = 0 \), we have \( f(x_t) = \frac{1}{2} \| x_t \|^2_f = \frac{1}{2t} \sum_{j=1}^{t} \sum_{i=1}^{t} k(\frac{i-1}{t}, \frac{j-1}{t}) = \frac{1}{2t} \sum_{i=1}^{t} k(\frac{i-1}{t}, 1) \), where the last equality follows from repeatedly applying

\[
\sum_{i=1}^{t} k(\frac{i-1}{t}, \frac{j-1}{t}) = \sum_{i=1}^{t} k(\frac{i-1}{t}, \frac{j}{t}),
\]

where \( j \in \{1, \ldots, t\} \). To see that (6.6) holds, recall that by Lemma 6.2, it holds that

\[
\sum_{i=1}^{t} k(\frac{i-1}{t}, \frac{j-1}{t}) - \sum_{i=1}^{t} k(\frac{i}{t}, \frac{j}{t}) = \sum_{i=1}^{t} k(\frac{i}{t}, \frac{j}{t}) - \sum_{i=1}^{t} k(\frac{i-1}{t}, \frac{j}{t}) = k(\frac{j}{t}) - k(0, \frac{j}{t}) = 0
\]

for all \( j \in \{1, \ldots, t\} \). Thus, \( f(x_t) = \frac{1}{t} \sum_{j=1}^{t} k(\frac{j}{t}, 1) = \frac{1}{t} \sum_{j=1}^{t} k(\frac{j}{t}, 0) = \frac{1}{t} \sum_{j=1}^{t} ((\frac{j}{t})^2 - \frac{j}{t} + \frac{1}{6}) \), where the second, third, and fourth equalities are due to Lemma 6.2. Since \( \sum_{i=1}^{t} i = \frac{t(t+1)}{2} \) and \( \sum_{i=1}^{t} i^2 = \frac{t(t+1)(2t+1)}{6} \), it holds that \( f(x_t) = \frac{1}{t} \left( \frac{2t^3 + t^2}{6} - \frac{t+1}{2} + \frac{1}{6} \right) = \frac{1}{t} \left( \frac{1}{2} + \frac{1}{6} \right) \).

The proof of Theorem 6.1 implies that the iterates of FW with open-loop step-size \( \eta_t = \frac{1}{t} \) 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 order \( O(1/t^2) \) (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.

The results in Figure 3, see Section 7.1.3 for details, show that in the kernel-herding setting of Section 6.2, for RKHS \( \mathcal{H} \) as in (6.1), kernel \( k \) as in (Bernoulli-kernel), and both uniform and non-uniform densities over \( \mathcal{Y} = [0, 1] \), FW with open-loop step-sizes \( \eta_t = \frac{1}{t} \), where \( t \in \mathbb{N}_{\geq 1} \), converges at a rate of order \( O(1/t^2) \) and FW with line-search converges at a rate of order \( O(1/t) \). It remains an open problem to extend Theorem 6.1 to non-uniform densities.

7. Numerical experiments

In this section, we present the numerical experiments. Numerical experiments corroborating our results in Sections 3.2, 3.4, and 5 are omitted since the studies do not provide new insights or highlight unexplained
convergence rates. All of our numerical experiments are implemented in Python and performed on an Nvidia GeForce RTX 3080 GPU with 10GB RAM and an Intel Core i7 11700K 8x CPU at 3.60GHz with 64 GB RAM. Our code is publicly available on GitHub. For all numerical experiments, to avoid the oscillating behavior of the primal gap, the y-axis represents min_{i \in \{1, \ldots, t\}} h_i, where t denotes the number of iterations and h_i the primal gap.

7.1 Detailed setups for the numerical experiments in Figures 1, 2, and 3

Throughout the paper, we present several toy examples in Figures 1, 2, and 3 to illustrate results and raise open questions. For completeness, we present the detailed setups for these experiments below.

7.1.1 Detailed setup for numerical experiments in Figure 1

For d = 100, we address (OPT) with FW for C ⊆ R^d the ℓ_p-ball, f(x) = \frac{1}{2} ||Ax - b||_2^2, where A ⊆ R^{100 \times 100} and b ∈ R^{100} are a random matrix and vector, respectively, such that f is not strongly convex, the unconstrained optimal solution argmin_{x \in \mathbb{R}^d} f(x) lies in the exterior of the feasible region and, thus, ||∇f(x)||_2 ≥ \lambda > 0 for all x \in C and some λ > 0. For p ∈ {2, 3, 5}, we compare FW with open-loop step-sizes η_t = \ell_t + ℓ_p, where \ell ∈ {1, 2, 4, 6}, and the constant step-size introduced in Remark 3.11, starting with x_0 = e^{(1)}. We plot the results of the experiments in log-log plots in Figure 1.

7.1.2 Detailed setup for numerical experiments in Figure 2

For d = 100, we address (OPT) with FW for C ⊆ R^d the probability simplex and f(x) = \frac{1}{2} ||x - \bar{1}||_2^2, where ρ ≥ \frac{2}{d} and \bar{1} is the vector with zeros for the first \lfloor d/2 \rfloor entries and ones for the remaining entries. Then, \frac{2}{\rho} \bar{1} = x^* = \operatorname{argmin}_{x \in C} f(x) is the unique minimizer of f. For ρ ∈ \{1/4, 2\}, we compare FW with line-search and open-loop step-sizes η_t = \ell_t, where \ell ∈ \{1, 2, 4, 6\}, starting with x_0 = e^{(1)}. Here, short-step is identical to line-search and, thus, omitted. We plot the results of the experiments in log-log plots in Figure 2.

7.1.3 Detailed setup for numerical experiments in Figure 3

We consider the kernel-herding setting of Section 6.2 over \mathcal{Y} = [0, 1], that is, \mathcal{H} is the RKHS as in (6.1) and k is the kernel as in (Bernoulli-kernel). Given either the uniform density or a random non-uniform density of the form p(y) ~ (\sum_{i=1}^{n} (a_i \cos(2\pi i y) + b_i \sin(2\pi i y)))^2 with n ≤ 5 and a_i, b_i ∈ R for all i ∈ {1, \ldots, n} such that \int_{[0,1]} p(y)dy = 1, we address (OPT-KH) with FW with line-search and open-loop step-sizes η_t = \ell_t, where \ell ∈ \{1, 2\}. The LMO is implemented as an exhaustive search over [0, 1] and run for 1,000 iterations. We plot the results of the experiments in log-log plots in Figure 3.
7.2 Logistic regression

We consider the problem of logistic regression, which for feature vectors $a_1, \ldots, a_m \in \mathbb{R}^d$, label vector $b \in \{-1, +1\}^m$, $p \in \mathbb{R}_{\geq 1}$, and radius $r > 0$, leads to the problem formulation

$$
\min_{x \in \mathbb{R}^d} \frac{1}{m} \sum_{i=1}^{m} \log(1 + \exp(-b_i a_i^\top x))
$$

subject to $\|x\|_p \leq r$.

Note that the feasible region is an $\ell_p$-ball and when $p = 1$, the problem formulation is that of sparsity-constrained logistic regression, which induces sparsity in the iterates of FW variants. For $p \in \{1, 2, 5\}$, we compare FW, PAFW, and MFW, with open-loop step-sizes $\eta_t = \ell_t + \ell$, where $\ell \in \{2, 6\}$, on the Z-score normalized Gisette dataset\(^2\) (Guyon and Elisseeff, 2003). The number of features is $d = 5,000$, we use $m = 2,000$ samples of the dataset, and we set $r = 1$. We plot the results of the experiments in log-log plots in Figure 4.

PAFW and MFW seem to enjoy the same accelerated convergence rates as FW with step-sizes $\eta_t = \ell_t + \ell$, where $\ell \in \mathbb{N}_{\geq 1}$. This includes the rates of order $O(1/t^\ell)$ when $p \in \{2, 5\}$, see also Remark 3.11. This raises the question whether PAFW (Lan, 2013; Kerdreux et al., 2021a) and MFW (Li et al., 2021) admit accelerated convergence rates due to the exploitation of momentum, as indicated in the respective works, or due to the specific choice of open-loop step-size. Furthermore, MFW seems to converge at an accelerated rate earlier than FW, which converges at an accelerated rate earlier than PAFW. However, for $p = 5$, MFW converges quickly during early iterations but then converges at a slower rate than FW and PAFW, especially for step-size $\eta_t = \frac{2}{t+2}$. For $p = 1$, all methods converge at the same rate of order $O(1/t^2)$.

7.3 Collaborative filtering

We consider the problem of collaborative filtering. In particular, let $A \in \mathbb{R}^{m \times d}$ be a matrix with only partially observed entries, that is, there exists a subset of indices $\mathcal{I} \subseteq \{1, \ldots, m\} \times \{1, \ldots, d\}$ such that only the entries $A_{i,j}$ with $(i, j) \in \mathcal{I}$ are observed. The task is to predict the unobserved entries of $A$. Let $H_\rho$ be the Huber loss with parameter $\rho > 0$ (Huber, 1992):

$$
H_\rho: x \in \mathbb{R} \mapsto \begin{cases} 
\frac{x^2}{2}, & \text{if } |x| \leq \rho \\
\rho |x| - \frac{\rho^2}{2}, & \text{if } |x| > \rho,
\end{cases}
$$

---

\(^2\) Available online at https://archive.ics.uci.edu/ml/datasets/Gisette.
\[ \|\cdot\|_{\text{nuc}} : X \in \mathbb{R}^{m \times d} \mapsto \text{tr}(\sqrt{X^T X}) \] the nuclear norm, and \( r > 0 \) the radius of the nuclear norm ball. Since we assume the solution to be low rank, the approach of Mehta et al. (2007) leads to the problem formulation

\[
\min_{X \in \mathbb{R}^{m \times d}} \frac{1}{|I|} \sum_{(i,j) \in I} H_{\rho}(A_{i,j} - X_{i,j})
\]

subject to \( \|X\|_{\text{nuc}} \leq r. \)

We compare FW, PAFW, and MFW, with open-loop step-sizes \( \eta_t = \ell_t + \ell \), where \( \ell \in \{2, 6\} \), on the MovieLens 100k dataset (Harper and Konstan, 2015) with \( m = 943 \), \( d = 1682 \), and \( |I| = 10,000 \), and we set \( \rho = 1 \) and \( r = 2,000 \). We plot the results of the experiments in a log-log plot in Figure 5.

All algorithms with any step-size ultimately converge at a rate of order \( O(1/t^2) \), except for MFW with step-size \( \eta_t = 6t + 6 \), which appears to converge at a rate of order \( O(1/t^6) \). The latter phenomenon is not currently motivated by results in this paper or Li et al. (2021). Among the different methods, MFW admits the fastest rate of convergence, followed by FW.

### 8. Discussion and open questions

We investigated settings in which FW with open-loop step-sizes achieves accelerated convergence rates. Specifically, we observed in Figures 1 and 4 that FW with step-size \( \eta_t = \ell_t + \ell \), where \( \ell \in \mathbb{N}_{\geq 1} \), converges at a rate of order \( O(1/t^\ell) \) when the feasible region \( \mathcal{C} \) is strongly convex and the norm of the gradient of \( f \) is bounded from below by a nonnegative constant. These rates are better than the rates of order \( O(1/t^{\ell/2}) \) derived in Remark 3.11, which raises the question whether this gap between theory and practice can be closed. Furthermore, it remains to investigate the accelerated rates of order up to \( O(1/t^\ell) \) when \( \mathcal{C} \) is only uniformly convex instead of strongly convex, see Figures 1b and 1c. Furthermore, these convergence guarantees of order \( O(1/t^{\ell/2}) \) are significantly better than the convergence guarantees of order up to \( O(1/t^2) \) of FW variants PAFW (Lan, 2013; Kerdreux et al., 2021a) and MFW (Li et al., 2021), which are designed to perform well in this setting. We thus conducted numerical experiments to investigate whether PAFW and MFW also achieve accelerated rates depending on the choice of open-loop step-size. According to the logistic-regression experiments in Figure 4, it appears that they do, which raises the question whether the accelerated convergence rates of PAFW and MFW stem from exploitation of momentum, as suggested in the respective works, or are in fact due to the choice of the open-loop step-size. The latter explanation is further supported by the unexplained convergence rate of order \( O(1/t^\ell) \) of MFW with step-size \( \eta_t = \frac{6}{t+6} \) in the collaborative filtering experiment in Figure 5. Further, we proved that FW with open-loop step-sizes achieves faster convergence rates than FW with line-search or short-step in the setting of the lower bound due to Wolfe (1970), assuming strict complementarity is satisfied. In case strict complementarity or similar assumptions are not satisfied, we proved that DIFW and AFW with open-loop step-sizes always converge at accelerated rates. We also answered the open question in Bach et al. (2012) by demonstrating that FW with open-loop step-size \( \eta_t = \frac{1}{t+1} \) achieves accelerated convergence rates in the setting of Section 6.2 for the uniform density in Theorem 6.1. Numerical experiments in Figure 3b indicate that acceleration also holds for non-uniform densities, an observation which is currently not backed by theoretical results. Finally, an important limitation of our study is that the proofs rely on norms, which are affine variant, whereas FW is known to be affine invariant. We plan to address this limitation in future work.

**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).

---

3. Available online at [https://grouplens.org/datasets/movielens/100k/](https://grouplens.org/datasets/movielens/100k/).
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.

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., Carderera, A., Combettes, C. W., Hassani, H., Karbasi, A., Mokhtari, A., and Pokutta, S. (2022). Conditional gradient methods. *arXiv preprint arXiv:2211.14103*.

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 Meshi, O. (2016). Linear-memory and decomposition-invariant linearly convergent conditional gradient algorithm for structured polytopes. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 1009–1017. PMLR.

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.
Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3(Mar):1157–1182.

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

Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19.

Huber, P. J. (1992). Robust estimation of a location parameter. *Breakthroughs in statistics: Methodology and distribution*, pages 492–518.

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:2102.05194*.

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.

Mehta, B., Hofmann, T., and Nejdl, W. (2007). Robust collaborative filtering. In *Proceedings of the ACM Conference on Recommender Systems*, pages 49–56.

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.

Wirth, E., Kera, H., and Pokutta, S. (2023). Approximate vanishing ideal computations at scale. In *Proceedings of the International Conference on Learning Representations*.

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