THE PROXIMAL BUNDLE ALGORITHM UNDER A FRANK-WOLFE PERSPECTIVE: AN IMPROVED COMPLEXITY ANALYSIS

DAVID FERSZTAND * AND XU ANDY SUN†

Abstract. The proximal bundle algorithm (PBA) is a fundamental and computationally effective algorithm for solving optimization problems with nonsmooth components. We investigate its convergence rate, focusing on composite settings where one function is smooth and the other is piecewise linear. We interpret a sequence of *null steps* of the PBA as a Frank-Wolfe algorithm on the Moreau envelope of the dual problem. In light of this correspondence, we first extend the linear convergence of Kelley's method on convex piecewise linear functions from the positive homogeneous to the general case. Building on this result, we propose a novel complexity analysis of PBA and derive a $\mathcal{O}(\epsilon^{-4/5})$ iteration complexity, improving upon the best known $\mathcal{O}(\epsilon^{-2})$ guarantee. This approach also unveils new insights on bundle management.

Key words. iteration-complexity, proximal bundle method, optimal complexity bound

MSC codes. 90C25, 90C30, 90C46, 90C60

1. Introduction.

1.1. Problem statement.

Central to numerous machine learning algorithms [43, 39, 27] and optimization algorithms [9, 42, 8, 22] lies the problem of minimizing the sum of a smooth (Lipschitz gradient) convex function g and a piecewise linear convex function f:

$$(P) \quad \min_{x\in\mathbb{R}^n} g(x) + f(x). \tag{1.1}$$

Composite problems (1.1) are pervasive in adaptive robust and stochastic optimization [9, 42, 8, 22], which motivates us to study the convergence guarantees of applicable algorithms.

Kelley's cutting plane method [20] is a fundamental algorithm for solving such nonsmooth problems (see Algorithm 2.1). Although this method was shown to be convergent for general nonsmooth functions, the complexity lower bound grows exponentially with the dimension of the problem [32]. [45] demonstrated that imposing more stringent hypotheses on (1.1), specifically requiring f to be *positive homogeneous* and g to be strongly convex, results in a linear convergence rate for Kelley's method. While it is feasible to relax the additional hypothesis on f (as discussed in section 2), the strong convexity of g remains a key requirement in the proof.

This leads us to investigate the proximal bundle algorithm (PBA) (see Algorithm 3.1), which is a refinement of Kelley's method. PBA introduces two additional features to Kelley's method: a quadratic proximal term in the objective and a *null-step test* to update the proximal center only if the algorithm has made sufficient progress. Steps in which the proximal center is updated are called *serious steps* while those where it remains unchanged are referred to as *null steps*. These two modifications stabilize the convergence dynamics of the algorithm. In contrast to Kelley's method, this approach incorporates strong convexity into the function g through the proximal term. This enhancement makes it possible to draw upon results about the convergence rates of the Frank-Wolfe algorithm.

*Operations Research Center, MIT Cambridge, MA

†Sloan School of Management and Operations Research Center, MIT, Cambridge, MA (sunx@mit.edu).

In this paper, we study the convergence rate of the PBA, analyzing the prevalent case where f is piecewise linear. We give an interpretation of the PBA's *null steps* as dual to the Fully Corrective Frank Wolfe (FCFW, Algorithm 2.2) algorithm applied to the Moreau envelope of the conjugate of g. This new perspective yields a convergence rate of $\mathcal{O}(\epsilon^{-4/5}\log(\frac{1}{\epsilon})^{2/5})$ for the PBA, a significant improvement over the previously established rate of $\mathcal{O}(\epsilon^{-2})$ within this framework.

1.2. Related Work. Kelley's method and bundle methods Kelley's cutting plane method (1960) [20] is a foundational algorithm for minimizing non-smooth convex objectives. It iteratively refines a piecewise linear lower approximation of the true function. The PBA [21, 11], the trust region bundle method [38], and the level set bundle method [28, 26] are variants of Kelley's method. Recently, [12] showed that their parallel PBA is on par with the performance of Pegasos [40], a state-of-the-art sub-gradient solver for binary SVMs.

Convergence rates of the proximal bundle algorithm [13] provided the first iteration complexity of the PBA for a strongly convex objective f, and g=0: $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$. [12] improved this analysis and provided convergence rates for all combinations of smoothness and strong convexity. Namely $\mathcal{O}(\frac{1}{\epsilon^2})$ when the function is only Lipschtiz continuous, $\mathcal{O}(\frac{1}{\epsilon})$ when the function is smooth or strongly convex, removing the log term of [13]. When the function is smooth and strongly convex, the convergence rate becomes $\mathcal{O}(\log\left(\frac{1}{\epsilon}\right))$. In [29] and [30] the authors studied the composite problem (1.1) and proposed a new type of *null step test* that we have adopted in this paper. Their analysis provides a $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}))$ convergence rate when f is only Lipschitz continuous, $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ when f is supposed to be smooth. When g is strongly convex the corresponding rates become respectively $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon})^2)$ and $\mathcal{O}(\log(\frac{1}{\epsilon})^2)$. Note that g may take on positive infinity, in contrast to both our setup and the setups in [13, 12], and thus include constraints.

Frank-Wolfe The Frank–Wolfe algorithm (1956) [14] is a first-order optimization method for constrained convex problems. It doesn't require projection onto the feasible region; rather, it progresses toward an extreme point determined by a linear minimization oracle, ensuring it stays within the feasible region [10, 15]. [18] presented a Fully Corrective variant of the algorithm (FCFW). Forty years later, [23] gives a linear convergence rate of Away-step Frank-Wolfe for the minimization of smooth and strongly convex objectives over polytopes, involving the pyramidal width, a geometry-dependent constant. This result extends to FCFW [24] and [7] showed that the linear convergence generalizes to the sum of a strongly convex function and a linear term.

Dual Correspondance Frank-Wolfe and Kelley's method [5] shows that a generalized version of the Frank-Wolfe algorithm has a dual correspondence with the mirror descent algorithm. In [4], Chapter 7 introduces methods to minimize the sum $f+g$ where f is piecewise linear positive homogeneous. It shows that Kelley's method is no other than FCFW applied to the dual of g on the convex hull of the slopes of f. In the same setup, [45] leverages this correspondence and the results from [24] to prove the linear convergence of Kelley's method when g is smooth and strongly convex.

1.3. Contributions. We can summarize our contributions as follows:

1. We first extend the linear convergence of Kelley's method to a broader class of problems. In particular, we prove that Kelley's method converges linearly for problem (1.1) with strongly convex and Lipschitz smooth g and general piecewise linear convex f, thus extending the previous analysis that requires the additional assumption that f is *positive homogeneous* (see section 2).
2. We study the PBA from the dual perspective to derive stronger convergence

rates than those available in the literature. More precisely, we interpret the *null steps* of the PBA as dual to FCFW applied to the Moreau envelope of the Fenchel conjugate of g. This dual perspective allows us to derive an $\mathcal{O}(\frac{1}{\epsilon}\log(\frac{1}{\epsilon}))$ convergence rate of the PBA for L-smooth g, thereby improving upon the best established convergence rate of $\mathcal{O}(\frac{1}{\epsilon^2}\log(\frac{1}{\epsilon}))$ (see sections 3.1-3.3).

3. We show via a novel analysis that we can further enhance the convergence rate of the PBA to $\mathcal{O}(\frac{1}{\epsilon^{4/5}}\log(\frac{1}{\epsilon})^{2/5})$ in the same setting (see subsection 3.4). The key arguments of the proof significantly differ from recent work on the PBA [13, 12, 29, 30]. The analysis draws upon our dual interpretation of the algorithm and employs novel geometric arguments. Moreover, it sheds new light on cut management for the PBA. Specifically, it shows the theoretical ground for retaining cuts in the model after a *serious step* (see section 4).

1.4. Organization of the paper. Subsection 1.5 formally describes the assumptions on problem (1.1). Subsection 1.6 presents definitions used throughout this paper. Section 2 extends the linear convergence of Kelley's method from *positive homogeneous* to general piecewise linear convex functions. Section 3 contains the main result of this paper, namely an $\mathcal{O}(\frac{1}{\epsilon^{4/5}}\log(\frac{1}{\epsilon})^{2/5})$ iteration complexity for PBA. In section 4, we perform a numerical experiment to evaluate various bundle management strategies, providing insights into the improved convergence rate of the PBA.

1.5. Assumptions. We suppose that problem (1.1) has at least one optimal solution $x^* \in \mathbb{R}^n$. We study the oracle complexity for finding an ϵ-optimal solution of (1.1), *i.e.* finding x such that $h(x) - h(x^*) \leq \epsilon$. Since both Kelley's method and the PBA are translation-invariant, we assume without loss of generality that $x^* = 0$.

Assumptions on f: We suppose that f is a convex piecewise linear function and that we have access to f through a Linear Maximization Oracle (LMO). That is, there exists a finite subset $\mathcal{V}$ of $\mathbb{R}^{n+1}$ such that

$$\forall x \in \mathbb{R}^n \quad f(x) = \max_{(v,b)\in\mathcal{V}} v^T x + b, \quad \text{with an LMO :} \quad (\hat{v},\hat{b}) \in \arg\max_{(v,b)\in\mathcal{V}} v^T x + b.$$

For $\mathcal{V}^k \subset \mathcal{V}$, we denote $f_k : x \mapsto \max_{(v,b)\in\mathcal{V}^k} v^T x + b$, which is a lower approximation (also called a model) of f. Let D denote the diameter of $\mathcal{V}$, γ its pyramidal width (a geometric quantity defined in section 3 of [24]), and M_f the Lipschitz constant of f.

Assumption on g: We suppose that g is convex and L_g-smooth on $\mathbb{R}^n$:

$$\|\nabla g(y) - \nabla g(x)\| \leq L_g\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

In section 2 only, g is also supposed to be μ_g-strongly convex :

$$g(y) \geq g(x) + \langle \nabla g(x), y - x\rangle + \tfrac{\mu_g}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

We further define $h = f + g$ and $h_k = f_k + g$.

Assumption on the oracles: In section 2, we suppose that an oracle can solve the following problem $\min_{x\in\mathbb{R}^n} g(x) + f_k(x)$. In section 3, the oracle is slightly modified to solve $\min_{x\in\mathbb{R}^n} g(x) + f_k(x) + \frac{1}{2\rho}\|x - x_k\|^2$, where $x_k \in \mathbb{R}^n$. Please refer to the bundle minimization problem (3.1) in Algorithm 3.1.

As f_k contains fewer pieces than f, for SVMs or adaptive robust optimization, these oracles typically consist of solving smaller-scale linear or quadratic programs than the original problem (1.1).

1.6. Notations. We define $\|\cdot\|$ the Euclidean norm and $\langle\cdot,\cdot\rangle$ the corresponding inner product. Let Id denote the identity of $\mathbb{R}$. For a set $\mathcal{V}$, $\text{conv}(\mathcal{V})$ denotes the convex hull of $\mathcal{V}$. The Fenchel conjugate [34] of a proper, convex function g is defined as $g^*(u) = \sup_{x\in\mathbb{R}^n}\{\langle x, u\rangle - g(x)\}$. The Fenchel conjugate of a convex L-smooth

function is $1/L$-strongly convex, and the Fenchel conjugate of a μ-strongly convex function is $1/\mu$-smooth (see for instance Proposition 2.6 of [3]).

For u, v two proper functions from $\mathbb{R}^n$ to $\mathbb{R} \cup \{+\infty\}$, $\square$ denotes the infimal convolution over $\mathbb{R}^n$ defined by $(u \square v)(x) = \inf_{z \in \mathbb{R}^n} \{u(x-z) + v(z)\}$.

The Moreau envelope of a proper, lower semicontinuous, convex function φ is a $\frac{1}{\rho}$-smooth convex function that approximates φ from below (see Definition 1.22 and Theorem 2.26 in [36]). It is defined as $\mathcal{M}_{\rho,\varphi} = \varphi \square \frac{1}{2\rho}\|\cdot\|^2$, where $\rho > 0$ is the smoothness parameter. Let ϕ be defined by $\phi(u) = g^*(-u)$.

2. Linear convergence of Kelley's method. This section studies the oracle complexity of the convergence in function value of Algorithm 2.1. This analysis is an extension of [45] in which the function f is *positive homogeneous* (i.e., its epigraph is a polyhedral cone) to any piecewise linear convex function. We first show in subsection 2.1 that Algorithm 2.1 and Algorithm 2.2 are dual of one another. Theorem 2.7 is the main result of this section and will be used to provide convergence rates for the PBA in section 3.

Assumption : For this section only, we suppose that g is μ_g-strongly convex.

Algorithm 2.1 Kelley's method

Require:
$\mathcal{V}^0 \subseteq \mathcal{V}$, $f_0 : x \mapsto \sup_{(v,b)\in\mathcal{V}^0} \{v^T x + b\}$.
for $k \geq 0$ **do**
Solve subproblem:

$$x_{k+1} \in \arg\min_{x\in\mathbb{R}^n} g(x) + f_k(x). \tag{2.1}$$

Add a new cut:

$$(\hat{v}_{k+1}, \hat{b}_{k+1}) \in \arg\max_{(v,b)\in\mathcal{V}} v^T x_{k+1} + b. \tag{2.2}$$

$\mathcal{V}^{(k+1)} \leftarrow \mathcal{V}^k \cup \{(\hat{v}_{k+1}, \hat{b}_{k+1})\}$.
if $f(x_{k+1}) - f_k(x_{k+1}) \leq \epsilon$
Terminate and return x_{k+1}.
end if
end for

Algorithm 2.2 Fully Corrective FW

Require:
$\mathcal{V}^0 \subseteq \mathcal{V}$, $\phi = g^*(-\cdot)$ a smooth convex function.
for $k \geq 0$ **do**
Fully Corrective Step:

$$(w_{k+1}, \beta_{k+1}) \in \arg\min_{(w,\beta)\in\mathrm{conv}(\mathcal{V}^k)} \phi(w) - \beta. \tag{2.3}$$

Frank-Wolfe step:

$$(v^{FW}_{k+1}, b^{FW}_{k+1}) \in \arg\max_{(v,b)\in\mathcal{V}} -v^T \nabla\phi(w_{k+1}) + b. \tag{2.4}$$

$\mathcal{V}^{(k+1)} \leftarrow \mathcal{V}^k \cup \{(v^{FW}_{k+1}, b^{FW}_{k+1})\}$.
if $\langle \nabla\phi(w_{k+1}), w_{k+1} - v^{FW}_{k+1}\rangle - \beta_{k+1} + b^{FW}_{k+1} \leq \epsilon$
Terminate and return $-\nabla\phi(w_{k+1})$.
end if
end for

2.1. Equivalence between Algorithm 2.1 and Algorithm 2.2. We give in Lemma 2.1 the Lagrangian dual of (1.1), solved by FCFW and show that Algorithm 2.1 and Algorithm 2.2 are applied to problems that are dual of one another. Lemma 2.1 also shows the correspondence between steps (2.1) in Algorithm 2.1 and (2.3) in Algorithm 2.2.

LEMMA 2.1 (Lagrangian dual). *The dual of Problem* (1.1) *is :*

$$D(\mathcal{V}) : \quad \max_{(w,\beta)\in\mathrm{conv}(\mathcal{V})} -g^*(-w) + \beta \quad \Leftrightarrow \quad \min_{(w,\beta)\in\mathrm{conv}(\mathcal{V})} \phi(w) - \beta. \tag{2.5}$$

Similarly, the dual of the minimization (2.1) *at iteration k in Kelley's algorithm is given by* (2.3).

Proof. We write the Lagrangian of the primal problem (1.1):

$$\begin{aligned}\mathcal{L}(x,z,\lambda) &= g(x) + z + \sum_{(v,b)\in\mathcal{V}} \lambda_{v,b}(v^T x + b - z) \\ &= g(x) + z(1 - \sum_{(v,b)\in\mathcal{V}} \lambda_{v,b}) + \sum_{(v,b)\in\mathcal{V}} \lambda_{v,b}(v^T x + b).\end{aligned} \tag{2.6}$$

Because of the dual feasibility condition $\forall (v,b) \in \mathcal{V}, \lambda_{v,b} \geq 0$ and the stationarity on z, $\sum_{(v,b)\in\mathcal{V}} \lambda_{v,b} = 1$, the Lagrangian can be rewritten :

$$\mathcal{L}(x,w,\beta) = g(x) + w^T x + \beta \qquad \text{with } (w,\beta) \in \text{conv}(\mathcal{V}).$$

The dual $D(\mathcal{V})$ is :

$$D(\mathcal{V}) = \max_{(w,\beta)\in\text{conv}(\mathcal{V})} \min_{x\in\mathbb{R}^n} g(x) + w^T x + \beta = \max_{(w,\beta)\in\text{conv}(\mathcal{V})} -g^*(-w) + \beta.$$

The same argument is applied to the minimization step (2.1) by replacing $\mathcal{V}$ by $\mathcal{V}^k$ and x^* by x_{k+1}. ∎

The primal problem (1.1) and the dual problem (2.5) have the same finite optimal value. Indeed, the primal problem (1.1) can be rewritten

$$\min_{x\in\mathbb{R}^n} \left(g(x) + \max_{(w,b)\in\text{conv}(\mathcal{V})} (w^T x + b) \right).$$

As $x \mapsto g(x) + w^T x$ is convex, $(w,b) \mapsto x^T w + b$ is concave and $\text{conv}(\mathcal{V})$ is convex compact, according to Sion's minimax theorem [41], strong duality holds. The following lemma draws a correspondence between the variables x_k in Algorithm 2.1 and (w_k, β_k) in Algorithm 2.2.

LEMMA 2.2 (Primal-Dual variables correspondence). *Let (w_{k+1}, β_{k+1}) be optimal for the dual of minimization step* (2.3) *and x_{k+1} be the primal optimal solution of* (2.1). *We have*

$$\begin{cases} x_{k+1} &= (\nabla g)^{-1}(-w_{k+1}) = \nabla g^*(-w_{k+1}), \\ w_{k+1} &= -\nabla g(x_{k+1}). \end{cases}$$

Proof. Let $\hat{x}_{k+1} \coloneqq \nabla g^*(-w_{k+1})$. The iterate w_{k+1} can be written as the following convex combination : $\sum_{(v,b)\in\mathcal{V}^k} \lambda_{v,b} v$. Looking at the Lagrangian (2.6), we define $z_{k+1} \coloneqq (w_{k+1})^T \hat{x}_{k+1} + \beta_{k+1}$. We now check that the primal-dual solution $(\hat{x}_{k+1}, z_{k+1}, (\lambda_{v,b})_{(v,b)\in\mathcal{V}^k})$ satisfies the KKT conditions of (2.1).

As $(\lambda_{v,b})_{(v,b)\in\mathcal{V}^k}$ defines a convex combination, $\forall (v,b) \in \mathcal{V}^k, \lambda_{v,b} \geq 0$ (dual feasibility) and $\sum_{(v,b)\in\mathcal{V}^k} \lambda_{v,b} = 1$ (stationarity on z). We now prove the stationarity on x: g is supposed to be closed convex and proper, using Fenchel's identity [37, p. 25], $\hat{x}_{k+1} \coloneqq \nabla g^*(-w_{k+1}) = \nabla g^{-1}(-w_{k+1})$. We deduce the stationarity condition on x :$\sum_{(v,b)\in\mathcal{V}^k} \lambda_{v,b} \cdot v \coloneqq w_{k+1} = -\nabla g((\nabla g)^{-1}(-w_{k+1})) = -\nabla g(\hat{x}_{k+1})$.

We still need to check that the primal feasibility and the complementary slackness hold, namely:

$$\forall (v,b) \in \mathcal{V}^k, (z_{k+1} - v^T \hat{x}_{k+1} - b) \geq 0 \tag{2.7}$$

$$\forall (v,b) \in \mathcal{V}^k, \lambda_{v,b} \cdot (z_{k+1} - v^T \hat{x}_{k+1} - b) = 0. \tag{2.8}$$

The optimality of w_{k+1} ensures that for all feasible directions (u, a),

$$u^T \nabla g^*(-w_{k+1}) + a^T \nabla Id(\beta_{k+1}) = u^T \nabla g^*(-w_{k+1}) + a \leq 0.$$

For $(v, b) \in \mathcal{V}^k$, $d = (v - w_{k+1}, b - \beta_{k+1})$ is a feasible direction as $(w_{k+1}, \beta_{k+1}) + d \in \mathcal{V}^k$

$$\forall (v, b) \in \mathcal{V}^k, \quad (v - w_{k+1})^T \nabla g^*(-w_{k+1}) + b - \beta_{k+1} \leq 0.$$

We deduce: $z_{k+1} \coloneqq w_{k+1}^T \hat{x}_{k+1} + \beta_{k+1} \geq \max\limits_{(v,b)\in\mathcal{V}^k} (v^T \hat{x}_{k+1} + b)$, which gives the primal feasibility (2.7). The reciprocal inequality is trivial as $(w_{k+1}, \beta_{k+1}) \in \text{conv}(\mathcal{V}^k)$. This shows that $z_{k+1} = \max\limits_{(v,b)\in\mathcal{V}^{(k)}} (v^T \hat{x}_{k+1} + b)$ and yields

$$\forall (v, b) \in \mathcal{V}^k, w_{k+1}^T \hat{x}_{k+1} + \beta_{k+1} > v^T \hat{x}_{k+1} + b \Rightarrow \lambda_{v,b} = 0.$$

This, and the primal feasibility (2.7) prove the complementary slackness condition (2.8). We have shown that $\hat{x}_{k+1} = \nabla g^*(-w_{k+1})$ is optimal for (2.1). Thus, $\hat{x}_{k+1} = x_{k+1}$. □

Note that the (2.7) and (2.8) also yield the following correspondence for the lower model : $f_k(x_{k+1}) = w_{k+1}^T x_{k+1} + \beta_{k+1}$.

Definition 2.3 (Frank–Wolfe gap, Definition 1.11 in [10]).
Let $s : \text{conv}(\mathcal{V}) \mapsto \mathbb{R}^n$ and $x^* = \arg\min\limits_{x\in\text{conv}(\mathcal{V})} s(x)$. We define the primal gap: $d^{primal}(x) = s(x) - s(x^*)$. The Frank–Wolfe gap (sometimes called dual gap) of s at x is $d^{FW}(x) \stackrel{\text{def}}{=} \max_{v\in\mathcal{V}} \langle \nabla s(x), x - v\rangle$. Clearly $d^{FW}(x) \geq 0$, as $x \in \text{conv}(\mathcal{V})$. Using the convexity of s, we get

$$0 \leq d^{primal}(x) \leq \langle \nabla s(x), x - x^*\rangle \leq \max_{v\in\mathcal{V}} \langle \nabla s(x), x - v\rangle = d^{FW}(x). \tag{2.9}$$

The following lemma shows the correspondence between the stopping criterion of Algorithm 2.1 and Algorithm 2.2.

Lemma 2.4 (Correspondence of the stopping criterion in Kelley's method and FCFW). *The termination criterion in Algorithm* 2.1 *corresponds to the stopping criterion in Algorithm* 2.2 *which a comparison between the accuracy ϵ and the Frank-Wolfe gap (Definition* 2.3*) associated with the dual problem* (2.5).

Proof. We consider the k^{th} iterate in the execution of Algorithm 2.1.

$$\begin{aligned}
f(x_{k+1}) - f_k(x_{k+1}) &\stackrel{\text{def}}{=} \max_{(v,b)\in\mathcal{V}} \left\{v^T x_{k+1} + b\right\} - \max_{(v,b)\in\text{conv}(\mathcal{V}^k)} \left\{v^T x_{k+1} + b\right\} \\
&= (v_{k+1}^{FW})^T x_{k+1} + b_{k+1}^{FW} - \left(w_{k+1}^T x_{k+1} + \beta_{k+1}\right) && (2.10) \\
&= \langle \nabla\phi(w_{k+1}), w_{k+1} - v_{k+1}^{FW}\rangle - \beta_{k+1} + b_{k+1}^{FW} && (2.11) \\
&= \langle \nabla \begin{pmatrix} \phi(w_{k+1}) \\ -Id \end{pmatrix} (w_{k+1}, \beta_{k+1}), \begin{pmatrix} w_{k+1} \\ \beta_{k+1} \end{pmatrix} - \begin{pmatrix} v_{k+1}^{FW} \\ b_{k+1}^{FW} \end{pmatrix} \rangle. && (2.12)
\end{aligned}$$

We get (2.10) from (2.3) and (2.4). In (2.11) we replace x_{k+1} by $-\nabla\phi(w_{k+1})$ using Lemma 2.2 and remark that it is exactly the stopping criterion of Algorithm 2.2. (2.12) is the Frank-Wolfe gap (Definition 2.3) associated with problem (2.5) at (w_{k+1}, β_{k+1}). □

We establish in the next proposition that Algorithm 2.1 and Algorithm 2.2 are dual to one another. This correspondence allows us to directly translate the convergence results of Algorithm 2.2 to Algorithm 2.1.

PROPOSITION 2.5. *The iterates of Kelley's method (Algorithm* 2.1*) applied to problem* (1.1) *match the iterates of FCFW (Algorithm* 2.2*) applied to* (2.5).

Proof. The two algorithms start with the same $\mathcal{V}^0$. Let x_1 be the unique minimizer of $g + f_1$ (g is strongly convex). Following Lemma 2.2, for (w_1, β_1) an optimal solution of (2.3), $w_1 = -\nabla g(x_1)$. The optimality in (2.3) shows the unicity of (w_1, β_1).

Lemma 2.2 shows that $x_1 = \nabla g^*(-w_1)$. The update step (2.2) of $\mathcal{V}^{(1)}$ corresponds to the FW step in FCFW (2.4). Also, the stopping criterion of the two algorithms matches as shown in Lemma 2.4. Thus, by recursion, the iterates and the set $\mathcal{V}^{(k)}$ are identical in both algorithms. ∎

2.2. Iteration complexity of Algorithm 2.2. In the setup of [45], where the dual problem is smooth and strongly convex, FCFW on a polytope converges linearly [24]. Note that (2.5) is not a strongly convex problem as it includes the linear map β. Even when the objective is not strongly convex but the sum of a strongly convex function and a linear map, such as in our case, [24] shows that the linear convergence holds with an adapted geometric strong convexity constant. We define:

- $E_1^n = Diag(1, ..., 1, 0) \in \mathbb{R}^{(n+1)\times(n+1)}$ (the projection matrix onto the directions of the supporting hyperplanes of f)
- $e_{n+1} = (0, ..., 0, 1)^T \in \mathbb{R}^{n+1}$ (the vector that extracts β)
- $\psi(w, \beta) = g^*(-E_1^n(w, \beta)) + \langle(-e_{n+1}), (w, \beta)\rangle$ (the function on which FCFW is applied)
- $D = diam(\mathcal{V})$, $D_w = diam(E_1^n\mathcal{V})$ and $D_b = diam(e_{n+1}^T\mathcal{V})$

Proposition 2.5 shows that it suffices to analyze the convergence rate of FCFW applied to the dual problem (2.5) that can be written

$$\min_{(w,b)\in\mathcal{V}^k} g^*(-w) - \beta \Leftrightarrow \min_{(w,b)\in\mathcal{V}^k} g^*(-E_1^n(w, \beta)) + \langle(-e_{n+1}), (w, \beta)\rangle. \tag{2.13}$$

The proof of the linear convergence of FCFW in Theorem 11 of [24] directly uses a similar convergence result established for the Away-Step Frank-Wolfe (AFW) from [7]. In the following Lemma 2.6, we provide an adapted geometric strong convexity constant $\bar{\mu}_\psi$ specializing the proof of [7] to (2.5). Compared to [7], leveraging the structure of problem (2.5), we eliminate the dependence on a Hoffman constant (see [7]) that may be arbitrarily large. As in [24], we will extend the obtained convergence rate for AFW to Algorithm 2.2 by noticing that a step of FCFW makes at least as much progress as the corresponding descent step and all the following away steps of AFW. This will directly lead to Theorem 2.7 which provides a linear convergence of Algorithm 2.1.

LEMMA 2.6. *Let (w^*, β^*) be the (unique) optimal solution of problem* (2.5). *The generalized geometric strong convexity constant of ψ that satisfies for all $(w, \beta) \in \mathcal{V}$:*

$$\psi(w^*, \beta^*) - \psi(w, \beta) - 2\langle\nabla\psi(w, \beta), ((w^*)^T, \beta^*)^T - (w^T, \beta)^T\rangle \geq 2\bar{\mu}_\psi \left\| \begin{pmatrix} w^* \\ \beta^* \end{pmatrix} - \begin{pmatrix} w \\ \beta \end{pmatrix} \right\|^2$$

is such that

$$\bar{\mu}_\psi = \frac{1}{2}\left(D_b + 3\frac{8M_f^2}{\mu_g} + 6M_f\|x^*\| + 2L_g\left[\left(\frac{4M_f}{\mu_g} + \|x^*\|\right)^2 + 1\right]\right)^{-1}. \tag{2.14}$$

Proof. For completeness, we will adapt the proof of Lemma 2.5 of [7] to our setup. Let $G = \max_{w\in E_1^n\mathcal{V}}(\|\nabla g^*(-w)\|)$. It will be shown later that G is finite. We first use

the L_g^{-1}-strong convexity of g^* to derive the following inequality (see Lemma A.1) :

(2.15)
$$\psi(w^*,\beta^*) - \psi(w,\beta) - \langle \nabla\psi(w,\beta), ((w^*)^T,\beta^*)^T - (w^T,\beta)^T\rangle \geq \frac{1}{2L_g}\left\|w-w^*\right\|^2.$$

We bound the distance $|\beta^*-\beta|$ by the difference in function value $\psi(w,\beta)-\psi(w^*,\beta^*)$:

$$\begin{aligned}
\beta-\beta^* &= \langle e_{n+1}, (w^T,\beta)^T - ((w^*)^T,\beta^*)^T\rangle \\
&= \langle \nabla\psi(w^*,\beta^*), (w^T,\beta)^T - ((w^*)^T,\beta^*)^T\rangle - \langle \nabla g^*(-w^*), w-w^*\rangle \qquad (2.16)\\
&\leq \psi(w,\beta)-\psi(w^*,\beta^*) + \|\nabla g^*(-w^*)\|\|w-w^*\| \\
&\leq \psi(w,\beta)-\psi(w^*,\beta^*) + G\|w-w^*\|. \qquad (2.17)
\end{aligned}$$

Also, using (2.16), by Cauchy-Schwarz and the optimality of (w^*,β^*) on the convex set $\mathcal{V}$, we get $\beta^*-\beta \leq G\|w-w^*\|$, which we combine with (2.17) to deduce

$$\begin{aligned}
&(\beta-\beta^*)^2 \\
&\leq (\psi(w,\beta)-\psi(w^*,\beta^*) + G\|w-w^*\|)^2 \\
&= (\psi(w,\beta)-\psi(w^*,\beta^*))^2 + 2G\|w-w^*\|(\psi(w,\beta)-\psi(w^*,\beta^*)) + G^2\|w-w^*\|^2 \\
&\leq (\psi(w,\beta)-\psi(w^*,\beta^*))(\|\nabla g^*(-w)\|D_w + D_b + 2GD_w) \qquad (2.18)\\
&\quad + G^2 2L_g(\psi(w,\beta)-\psi(w^*,\beta^*))) \\
&\leq (\psi(w,\beta)-\psi(w^*,\beta^*))(D_b + 3GD_w + 2L_gG^2) \\
&\leq -\langle \nabla\psi(w,\beta), ((w^*)^T,\beta^*)^T - (w^T,\beta)^T\rangle(D_b+3GD_w+2L_gG^2). \qquad (2.19)
\end{aligned}$$

Inequality (2.18) uses Lemma A.1. For (2.19) we used the convexity of ψ. Adding (2.19) and (2.15), we get :

$$\begin{aligned}
&\psi(w^*,\beta^*) - \psi(w,\beta) - 2\langle \nabla\psi(w,\beta), ((w^*)^T,\beta^*)^T - (w^T,\beta)^T\rangle \\
&\quad \geq \max(D_b+3GD_w+2L_gG^2, 2L_g)^{-1}(\|w-w^*\|^2 + (\beta-\beta^*)^2) \\
&\quad \geq (D_b+3GD_w+2L_g(G^2+1))^{-1}\left\|\begin{pmatrix} w^* \\ \beta^*\end{pmatrix} - \begin{pmatrix} w \\ \beta\end{pmatrix}\right\|^2. \qquad (2.20)
\end{aligned}$$

Upper bound on G : Let $w \in E_1^n\mathcal{V}$. There exists $x \in \mathbb{R}$ such that :

$$g^*(-w) = -w^Tx - g(x) \quad \text{and } w = -\nabla g(x). \qquad \text{(Fenchel conjugate)}$$

Using the strong convexity of $f+g$:

$$\begin{aligned}
\|\nabla g(x) - \nabla g(x^*)\| + 2M_f &\geq \mu_g\|x-x^*\| && \text{(Strong convexity)}\\
\frac{\|w^*-w\|+2M_f}{\mu_g} &\geq \|\nabla g^*(\nabla g(x)) - x^*\| && (\ (\partial g)^{-1} = \partial g^*)\\
&\geq \|\nabla g^*(\nabla g(x))\| - \|x^*\|.
\end{aligned}$$

After rearranging, we have $\|\nabla g^*(-w)\| \leq \frac{D_w+2M_f}{\mu_g} + \|x^*\|$. As this is inequality valid for any $w \in E_1^n\mathcal{V}$:

$$G \leq \frac{D_w+2M_f}{\mu_g} + \|x^*\|. \qquad (2.21)$$

This leads to the claimed inequality using (2.20) and replacing G by (2.21) :

$$\begin{aligned}
&\psi(w^*,\beta^*) - \psi(w,\beta) - 2\langle \nabla\psi(w,\beta), ((w^*)^T,\beta^*)^T - (w^T,\beta)^T\rangle \\
&\geq \left(D_b + 3\frac{D_w^2 + 2D_w M_f}{\mu_g} + 3D_w\|x^*\| + 2L_g\left[\left(\frac{D_w + 2M_f}{\mu_g} + \|x^*\|\right)^2 + 1\right]\right)^{-1} \\
&\quad \times \left\| \begin{pmatrix} w^* \\ \beta^* \end{pmatrix} - \begin{pmatrix} w \\ \beta \end{pmatrix} \right\|^2 \\
&\geq \left(D_b + 3\frac{8M_f^2}{\mu_g} + 6M_f\|x^*\| + 2L_g\left[\left(\frac{4M_f}{\mu_g} + \|x^*\|\right)^2 + 1\right]\right)^{-1} \left\| \begin{pmatrix} w^* \\ \beta^* \end{pmatrix} - \begin{pmatrix} w \\ \beta \end{pmatrix} \right\|^2 . \square
\end{aligned}$$

THEOREM 2.7 (Linear convergence of Kelley's algorithm). *Let f be a piecewise linear convex function, g be L_g-smooth and μ_g-strongly convex. Algorithm* 2.1 *applied to* (1.1) *finds an ϵ-optimal solution in at most the following number of iterations :*

$$1 + \max\left\{2, \frac{D^2}{\bar{\mu}_\psi \mu_g \gamma^2}\right\} \log\left(\frac{D^2}{2\epsilon\mu_g}\right), \tag{2.22}$$

where $\bar{\mu}_\psi$ is the generalized geometric strong convexity constant defined in Lemma 2.6 *and γ is the pyramidal width of $\mathcal{V}$ defined in [24, section 3].*

Proof. We leverage the correspondence between Kelley's method and FCFW shown in Lemma 2.2, thereby bounding the iteration complexity of Kelley's method. Indeed, Theorem 8 in [24] has shown the upper bound of FCFW oracle complexity stated in Theorem 2.7 above, with the constant $\bar{\mu}_\psi$ as defined in Lemma 2.6, a lower bound of the generalized geometric strong convexity constant of ψ. ∎

Remark 2.8. Note that the number of iterations given in Theorem 2.7 is insufficient to guarantee that the stopping criterion will be satisfied. However, the stopping criterion can be satisfied with linear complexity as well (see Lemma 3.6).

3. Proximal bundle method. This section studies the convergence rate of the PBA by leveraging its dual formulation. When g is not strongly convex, g^* is not smooth. In this setup, we turn our attention to the PBA that adds strong convexity to g through a quadratic proximal term. This section contains the main result of this paper: Theorem 3.14, a $\mathcal{O}(\frac{1}{\epsilon^{4/5}})$ convergence rate, up to log terms, when g is smooth but not necessarily strongly convex.

3.1. Correspondance between Algorithm 3.1 and Algorithm 3.2. We write the Lagrangian dual of the minimization step (3.1). Suppose that we are at step k. Compared to (2.1), the function g is replaced by $\hat{g}_k := g + \frac{\rho}{2}\|\cdot - x_k\|^2$. The conjugate of $\hat{g}_k$ is the Moreau envelope of g^* (see [34] Theorem 16.4). This is formalized in Lemma 3.1 that shows that the dual of (3.1) is (3.2) up to constant terms.

LEMMA 3.1. *Suppose that we are at step k. The dual of problem* (3.1) *is*

$$D'(\mathcal{V}^k, x_k) : - \min_{(w,\beta)\in \mathrm{conv}(\mathcal{V}^k)} \{\mathcal{M}_{\rho,g^*}(\rho x_k - w) - \beta\} + \frac{\rho}{2}\|x_k\|^2. \tag{3.5}$$

Proof. We compute $\hat{g}_k^*(-w)$ for $w \in \mathbb{R}^n$. For fundamental properties of the conjugacy operation used below, see Proposition 1.3.1 and Corollary 2.1.3 in [17].

$$\hat{g}_k^*(-w) = \left(g + \frac{\rho}{2}\|\cdot - x_k\|^2\right)^*(-w) = \left(g^* \square \left(\frac{\rho}{2}\|\cdot - x_k\|^2\right)^*\right)(-w)$$

Algorithm 3.1 Prox. Bundle Alg.

Require:
$x_0 \in \mathbb{R}^n$, $\mathcal{V}^0 \subset \mathcal{V}$, $\delta = \epsilon/2$, $\rho > 0$
for $k \geq 0$ **do**
Compute y_{k+1} by solving :

$$\arg\min_{x \in \mathbb{R}^n} g(x) + f_k(x) + \frac{\rho}{2}\|x - x_k\|^2 \tag{3.1}$$

Null step test :
if $f(y_{k+1}) - f_k(y_{k+1}) \leq \delta$
$x_{k+1} \leftarrow y_{k+1}$ ▷ (serious step)
else
$x_{k+1} \leftarrow x_k$ ▷ (null step)
end if

$$(\hat{v}_{k+1}, \hat{b}_{k+1}) \in \arg\max_{(v,b) \in \mathcal{V}} (v^T y_{k+1} + b) \tag{3.4}$$

Update $\mathcal{V}^{(k+1)} \leftarrow \mathcal{V}^k \cup \{(\hat{v}_{k+1}, \hat{b}_{k+1})\}$
end for

Algorithm 3.2 Modified FCFW

Require:$x_0 \in \mathbb{R}^n$, $\mathcal{V}^0 \subset \mathcal{V}$, $\delta = \epsilon/2$, $\rho > 0$
for $k \geq 0$ **do**
Compute (w_{k+1}, β_{k+1}) by solving :

$$\min_{(w,\beta) \in \mathrm{conv}(\mathcal{V}^k)} \mathcal{M}_{\rho,\phi}(w - \rho x_k) - \beta \tag{3.2}$$

$$y_{k+1} \leftarrow -\nabla \mathcal{M}_{\rho,\phi}(w_{k+1} - \rho x_k) \tag{3.3}$$

if $\langle y_{k+1}, v_k^{FW} - w_k \rangle + b_k^{FW} - \beta_k \leq \delta$
$x_{k+1} \leftarrow y_{k+1}$
else
$x_{k+1} \leftarrow x_k$
end if

$$(v_{k+1}^{FW}, b_{k+1}^{FW}) \in \arg\max_{(v,b) \in \mathcal{V}} (v^T y_{k+1} + b) \tag{3.4}$$

Update $\mathcal{V}^{(k+1)} \leftarrow \mathcal{V}^k \cup \{(v_{k+1}^{FW}, b_{k+1}^{FW})\}$
end for

$$= \left(g^* \square (\frac{1}{2\rho}\|\cdot\|^2 + \langle \cdot, x_k \rangle)\right)(-w) = \left(g^* \square (\frac{1}{2\rho}\|\cdot + \rho x_k\|^2)\right)(-w) - \frac{\rho}{2}\|x_k\|^2$$

$$= \mathcal{M}_{\rho,g^*}(\rho x_k - w) - \frac{\rho}{2}\|x_k\|^2. \tag{3.6}$$

Replacing g by $\hat{g}_k$ in (2.5) leads to (3.5). ∎

Strong duality holds between (3.1) and (3.2) due to the compactness of $\mathrm{conv}(\mathcal{V}^k)$.

The following Lemma shows the correspondence between the iterates of Algorithm 3.1 and Algorithm 3.2.

LEMMA 3.2 (PBA iterates correspondence). *The dual minimization* (3.5) *has a unique optimal solution. Let* (w_{k+1}, β_{k+1}) *be this optimal solution. Let* y_{k+1} *be the candidate iterate at the* k^{th} *step and* x_k *the corresponding proximal center. Then if Algorithm* 3.1 *and Algorithm* 3.2 *share the same set* $\mathcal{V}^{(k)}$ *at iteration* k *:*

$$y_{k+1} = -\nabla \mathcal{M}_{\rho,\phi}(w_{k+1} - \rho x_k). \tag{3.7}$$

Proof. The argument is the same as in Proposition 2.5 by replacing g by $\hat{g}_k$ which is smooth and ρ-strongly convex. Applying Lemma 2.2 by replacing g by $\hat{g}_k$ and using the expression of the conjugate of $\hat{g}_k$ given in (3.6) leads to (3.7). ∎

The *null step test* $f(y_{k+1}) - f_k(y_{k+1}) \leq \delta$ differs from the standard test for the PBA [28, 13, 12]: $\beta(f(x_k) - f_k(y_{k+1})) \leq f(x_k) - f(y_{k+1})$, with $\beta \in (0, 1)$, which compares the progress made between the iterates x_k and y_{k+1} to the expected progress. The test in Algorithm 3.1 is inspired by the test introduced in [29, 30]: $f(y_{k+1}) - f_k(y_{k+1}) \leq \frac{\rho}{2}\|y_{k+1} - x_k\|^2 + \delta$. Lemma 3.3 shows that this new test is chosen to correspond to the stopping criterion of the Frank-Wolfe algorithm.

For k, the index of a *serious step*, we define $\Psi_k(w, \beta) \stackrel{\text{def}}{=} \mathcal{M}_{\rho,\phi}(w - \rho x_k) - \beta$ and $\underline{\Psi}_k$ its minimum on $\mathrm{conv}(\mathcal{V})$. We have $\nabla \Psi_k(w, \beta) = \binom{\nabla \mathcal{M}_{\rho,\phi}(w - \rho x_k))}{-1}$. Note that

problem (3.5) can be equivalently written

$$\min_{(w,\beta)\in \mathrm{conv}(\mathcal{V})} \Psi_k(w,\beta). \tag{3.8}$$

LEMMA 3.3. *Let $t \geq k$ be an iterate in the sequence of null steps following the serious step k. The null step test in Algorithm* 3.1 *corresponds to a comparison between δ, the desired accuracy for the proximal problem, and*

$$\langle \nabla \Psi_k(w_{t+1},\beta_{t+1}), \begin{pmatrix} w_{t+1} \\ \beta_{t+1} \end{pmatrix} - \begin{pmatrix} v_{t+1}^{FW} \\ b_{t+1}^{FW} \end{pmatrix} \rangle, \tag{3.9}$$

the Frank-Wolfe gap of problem (3.8) *at* (w_{t+1},β_{t+1}).

Proof. We remark that a sequence of *null steps* corresponds to applying Algorithm 2.1 to the proximal problem. The *null step test* corresponds to the stopping criterion of Algorithm 2.1 with accuracy δ. The stated result follows directly from the application of Lemma 2.4. ∎

Using Lemma 3.2, (3.2) yields the dual variable y_{k+1} through (3.3). Thus, we have the correspondence between step (3.1) of Algorithm 3.1 and step (3.3) of Algorithm 3.2, which shows that the iterates of Algorithm 3.1 match those of Algorithm 3.2.

3.2. Interpretation of the dual of the Proximal Bundle Method. In this subsection, we give a new interpretation of Algorithm 3.2. As pointed out by [30], the PBA can be interpreted as an inexact proximal point algorithm [31]. Rockafellar [35] showed that the proximal point algorithm is the Augmented Lagrangian Method (ALM) [16, 33] applied to the dual problem. In particular, Algorithm 3.2 can be viewed as an inexact ALM for which the minimization subproblem is solved approximately using FCFW. The PBA's *serious step* corresponds to a dual ascent step in ALM. To see more clearly, we rewrite the dual problem (2.5) using the following notations: $\lambda = \rho^{-1}$ and $A = -I_n$, the identity matrix of size n.

LEMMA 3.4. *The dual problem* (2.5) *can be written*

$$\max_{(w,\beta)\in \mathrm{conv}(\mathcal{V})} -g^*(-w)+\beta \qquad \Leftrightarrow \qquad \min_{v\in\mathbb{R}^n, u\in\mathbb{R}^n} f^*(u)+\phi(v) \qquad s.t \quad Au+v=0.$$

Proof. We start from the primal problem (1.1) $\min_{x\in\mathbb{R}^n} f(x)-(-g(x))$, and we check that $(-g)$ is a proper concave function on $\mathbb{R}^n$, f is a proper convex function on $\mathbb{R}^n$. Moreover, $\mathrm{relint}(\mathrm{dom} f) \cap \mathrm{relint}(\mathrm{dom}(-g)) = \mathbb{R}^n \neq \varnothing$. Following Fenchel's Duality Theorem ([34] Theorem 31.1):

$$\begin{aligned} \inf_{x\in\mathbb{R}^n} g(x)+f(x) &= \sup_{u\in\mathbb{R}^n} (-g(u))^* - f^*(u) = -\inf_{u\in\mathbb{R}^n} g^*(-u)+f^*(u) \\ &= -\min_{u\in\mathbb{R}^n} \phi(u)+f^*(u) = -\min_{v\in\mathbb{R}^n, u\in\mathbb{R}^n} \phi(v)+f^*(u) \quad \text{s.t } Au+v=0. \end{aligned}$$

For the second equality, we use $(-g(u))^* = -g^*(-u)$. We can replace the inf by a min by using the strong convexity of $\phi + f^*$. ∎

We consider $(k_i)_{i\geq 0}$ the sequence of *serious steps*. The sequence of *null steps* $\{k_{i-1}+1, ..., k_i - 1\}$ corresponds to applying FCFW to $D'(\mathcal{V}, x_{k_{i-1}})$ until finding an approximate solution. This sequence is followed by the update of the proximal center given by (3.3). Thus, Algorithm 3.2 can be reformulated as the following inexact ALM, where (3.10) is approximately solved by FCFW:

$$(v_{k_i}, u_{k_i}) \approx \operatorname*{arg\,min}_{v\in\mathbb{R}^n, u\in\mathbb{R}^n} \left\{ \phi(v) + f^*(u) + \langle x_{k_{i-1}}, Au+v\rangle + \frac{1}{2\rho}\|Au+v\|^2 \right\}, \tag{3.10}$$

$$x_{k_i} = x_{k_{i-1}} + \lambda(Au_{k_i} + v_{k_i}) \qquad \text{(dual ascent step)}. \tag{3.11}$$

3.3. First convergence rate analysis.

3.3.1. Number of serious steps. We consider the iteration k of Algorithm 3.1, which we suppose is a *serious step*. x_k is the proximal center yielding y_{k+1}. As iteration k is a *serious step* : $x_{k+1} = y_{k+1} = -\nabla \mathcal{M}_{\rho,\phi}(w_{k+1} - \rho x_k)$.

As x_{k+1} is optimal for problem (3.1), which is the minimization of a ρ-strongly convex function :

$$h_k(x_{k+1}) + \frac{\rho}{2}\|x_{k+1} - x_k\|^2 \leq h_k(x^*) + \frac{\rho}{2}\|x^* - x_k\|^2 - \frac{\rho}{2}\|x_{k+1} - x^*\|^2. \tag{3.12}$$

Rearranging, using $h_k \leq h$, the *serious step* condition $h(x_k) - h_k(x_{k+1}) \leq \frac{\epsilon}{2}$ and summing over *serious step* index k_i:

$$\min_{i=1,\dots,T}\{h(x_{k_i}) - h(x^*)\} \leq \frac{\rho}{2T}\|x_0 - x^*\|^2 - \frac{\rho}{2T}\sum_{i=0}^{T-1}\|x_{k_{i+1}} - x_{k_i}\|^2 + \frac{\epsilon}{2}. \tag{3.13}$$

Inequality (3.13) leads to the following lemma that bounds the number of *serious steps*.

LEMMA 3.5 (Proposition 2 in [30]). *Algorithm* 3.1 *finds an ϵ-optimal solution to* (1.1) *in at most $\|x^* - x_0\|^2\rho/\epsilon + 1$ serious steps.*

3.3.2. Number of null steps. The following lemma bounds the number of consecutive *null steps* between two *serious steps*. During a sequence of *null steps*, the proximal center is never updated. The minimization step (3.1) in Algorithm 3.1 can be viewed as the update step (2.1) (Algorithm 2.1) applied to $\hat{g}_k$ where k is the last *serious step*. As g is smooth, $\hat{g}_k$ is smooth and strongly convex. Theorem 2.7 shows the linear convergence of Kelley's method applied to $\hat{g}_k + f$. The sequence of *null steps* ends when the *null step test* $f(y_{k+1}) - f_k(y_{k+1}) \leq \delta = \frac{\epsilon}{2}$ is satisfied. We showed in Lemma 3.3 that this *null step test* is exactly the Frank-Wolfe gap of $D'(\mathcal{V}, x_k)$ (see Definition 2.3). We finally upper bound this Frank-Wolfe gap using the primal gap of $D'(\mathcal{V}, x_k)$. This will show that a *serious step* occurs after at most $\mathcal{O}(\log(1/\epsilon))$ *null steps*.

The iteration complexity for Algorithm 2.1 given in Theorem 2.7 depends on $\|x^*\|$ through the generalized strong convexity constant given in (2.14). The bound on the number of consecutive *null steps* would thus depend on the norm of the optimal solution of the proximal problem $\arg\min_{x\in\mathbb{R}^n} h(x) + \frac{1}{2\rho}\|x - x_k\|^2$. We provide a uniform bound on this quantity in Lemma B.1.

LEMMA 3.6. *Algorithm* 3.1 *performs at most the following number of null steps between two serious steps*

$$1 + \max\left\{2, \frac{D^2}{\rho\gamma^2\bar{\mu}_{\psi,\rho}}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right), \tag{3.14}$$

with $\bar{\mu}_{\psi,\rho}$ derived from Lemma 2.6 *with adapted parameters :*

$$\begin{aligned}\frac{1}{2}\bar{\mu}_{\psi,\rho}^{-1} &= D_b + 3\frac{8M_f^2}{\rho} + 6M_f(\|x^*\| + 2\sqrt{(1+\rho^2)}\|x^* - x_0\| + 2\sqrt{\rho\epsilon}) \\ &\quad + 2L_g\left[\left(\frac{4M_f}{\rho} + \|x^*\| + 2\sqrt{(1+\rho^2)}\|x^* - x_0\| + 2\sqrt{\rho\epsilon}\right)^2 + 1\right].\end{aligned} \tag{3.15}$$

Proof. We consider k the index of a *serious step* and $t \geq k$ an index in the sequence of *null steps* following k. Because ϕ is convex, we can apply Theorem 6.60 in [6] (smoothness of the Moreau envelope) to deduce that $\mathcal{M}_{\rho,\phi}(\cdot - \rho x_k)$ is $\frac{1}{\rho}$-smooth. For any $(w_1, \beta_1), (w_2, \beta_2) \in \text{conv}(\mathcal{V})$:

$$\begin{aligned}\|\nabla\Psi_k(w_1,\beta_1) - \nabla\Psi_k(w_2,\beta_2)\| &= \left\| \begin{pmatrix} \nabla\mathcal{M}_{\rho,\phi}(w_1 - \rho x_k) - \nabla\mathcal{M}_{\rho,\phi}(w_2 - \rho x_k)) \\ 0 \end{pmatrix} \right\| \\ &= \|\nabla\mathcal{M}_{\rho,\phi}(w_1 - \rho x_k) - \nabla\mathcal{M}_{\rho,\phi}(w_2 - \rho x_k))\| \leq \frac{1}{\rho}\|w_1 - w_2\| \\ &\leq \frac{1}{\rho}\sqrt{\|w_1 - w_2\|^2 + (\beta_1 - \beta_2)^2} = \frac{1}{\rho}\|(w_1,\beta_1)^T - (w_2,\beta_2)^T\|.\end{aligned}$$

We conclude that Ψ_k is $\frac{1}{\rho}$-smooth. Also, $\mathcal{M}_{\rho,\phi}(w - \rho x_k) = (\underbrace{g + \frac{\rho}{2}\|\cdot\|^2}_{(\rho+L_g)-smooth})^*(\rho x_k - w)$.

The conjugate of a σ-smooth function is $\frac{1}{\sigma}$-strongly convex. So $\mathcal{M}_{\rho,\phi}(\cdot - \rho x_k)$ is $(\rho + L_g)^{-1}$-strongly convex. According to Definition 2.3, we define the primal gap

$$d_k^{primal}(w_{t+1},\beta_{t+1}) \stackrel{\text{def}}{=} \Psi_k(w_{t+1},\beta_{t+1}) - \underline{\Psi}_k, \tag{3.16}$$

for problem (3.8) at the point (w_{t+1}, β_{t+1}) of (3.9) and the corresponding Frank-Wolfe gap $d_k^{FW}(w_{t+1}, \beta_{t+1})$. We use Theorem 2 in [24] to upper-bound the Frank-Wolfe gap for a primal gap sufficiently small: $d_k^{primal}(w_{t+1}, \beta_{t+1}) \leq D^2/(2\rho)$

$$d_k^{FW}(w_{t+1},\beta_{t+1}) \leq D\sqrt{2\rho^{-1}d_k^{primal}(w_{t+1},\beta_{t+1})} \quad \text{(as } \Psi_k \text{ is } \rho^{-1}\text{-smooth)}. \tag{3.17}$$

Note that the *null step test* quantity $f(x_{t+1}) - f_t(x_{t+1})$ is exactly this Frank-Wolfe gap. If $d_k^{FW}(w_{t+1}, \beta_{t+1}) \leq \delta$, then a *serious step* is reached. Using (3.17), a $\frac{\delta^2\rho}{2D^2}$-optimal solution to (3.8) would guarantee a Frank-Wolfe gap of at most δ.

During a sequence of *null steps*, the PBA and Kelley's method match. Proposition 2.5 shows that these iterates also match with FCFW on problem (3.8), with their optimality gap (primal gap) matching as well.

We define $x_k^* = \underset{x\in\mathbb{R}^n}{\arg\min}\, h(x) + \frac{1}{2\rho}\|x - x_k\|^2$. The geometric strong convexity constant given in Lemma 2.6 involves $\|x^*\|$. To leverage the result of Theorem 2.7, we use the uniform bound on $\|x_k^*\|$ given in Lemma B.1.

According to Theorem 2.7, Kelley's algorithm finds an $\frac{\delta^2\rho}{2D^2}$-optimal solution to (3.8) in at most the following number of iterations:

$$1 + \max\left\{2, \frac{D^2}{\rho\gamma^2\bar{\mu}_{\psi,\rho}}\right\}\log\left(\frac{D^2}{2\frac{\delta^2\bar{\mu}_{\psi,\rho}}{2D^2}\bar{\mu}_{\psi,\rho}}\right) = 1 + \max\left\{2, \frac{D^2}{\rho\gamma^2\bar{\mu}_{\psi,\rho}}\right\}\log\left(\frac{D^4}{\delta^2\bar{\mu}_{\psi,\rho}^2}\right).$$

With $\bar{\mu}_{\psi,\rho}$ the generalized geometric strong convexity constant given by Lemma 2.6 by replacing L_g by $\rho + L_g$, μ_g by ρ and $\|x^*\|$ by $\|x^*\| + 2\sqrt{(1+\rho^2)}\|x^* - x_0\| + 2\sqrt{\rho\epsilon}$. ∎

We combine Lemma 3.5 and Lemma 3.6 to derive Theorem 3.7.

THEOREM 3.7. *Algorithm* 3.1 *with* $\delta = \epsilon/2$ *finds an* ϵ*-optimal solution to* (1.1) *in at most the following number of iterations*

$$\left(1 + \frac{\|x^* - x_0\|^2\rho}{\epsilon}\right)\left[1 + \max\left\{2, \frac{D^2}{\bar{\mu}_{\psi,\rho}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right)\right].$$

3.4. Improved convergence rate analysis. The previous analysis does not take advantage of the progress made in the discovery of the feasible set $\mathcal{V}$ as FCFW unfolds. Intuitively, if the consecutive proximal centers x_{k-1} and x_k are close, the cuts added to the bundle $\mathcal{V}^k$ while solving $D'(\mathcal{V}, x_{k-1})$ are relevant for solving $D'(\mathcal{V}, x_k)$ and this latter problem may be solved in fewer iterations. Determining a threshold on $\|x_{k-1} - x_k\|$ to achieve a constant number of iterations for solving $D'(\mathcal{V}, x_k)$ is not straightforward. A pivotal aspect of this proof lies in considering proximal centers that are close enough, ensuring the optimality of $D'(\mathcal{V}, x_k)$ after just one step. In that case, x_{k+1} will also be a serious iterate.

We categorize *serious steps* into three distinct (non-exclusive) types. Firstly, "consecutive" *serious steps* occur when a *serious step* is immediately followed by another *serious step*. Secondly, "distant center" steps arise when the new proximal center is sufficiently far from the previous proximal center. Lastly "small norm" refers to steps where the norm $\|x_k\|$ is small enough to meet a specific threshold. In Lemmas 3.8-3.13, we will establish bounds on the number of each of these step types. In Theorem 3.14 we will add these bounds to derive the overall convergence rate.

We first show in Lemma 3.8 that, when FCFW is applied on a polytope, a small angle between the gradients of two iterates guarantees that the algorithm has found an optimal solution.

We then derive in Lemma 3.11 the conditions under which the angle between two consecutive iterates y_k and y_{k+1} is small enough to guarantee that w_{k+1} is optimal and conclude that x_{k+1} is a serious iterate. The number of consecutive *serious steps* will be upper bounded by the total number of *serious steps* given in Lemma 3.5.

The third part of the proof consists in bounding the number of *serious steps* followed by *null steps*. Looking at (3.13), a large distance between two consecutive proximal centers will translate into primal progress: Lemma 3.12 and Lemma 3.13 show that these steps are an order of magnitude fewer than the total number of *serious steps*. According to Lemma 3.6, these *serious steps* will be followed by at most $\mathcal{O}(\log(\epsilon^{-1}))$ *null steps*.

Lemma 3.8 presents a key geometric argument. It studies the execution of the Fully Corrective Frank-Wolfe (FCFW) algorithm on a polytope and establishes a lower bound on the angle between the gradients obtained during the iterations, which are the directions in which the Linear Maximization Oracle (LMO) is called. The main idea of this lemma is that the same direction cannot be explored twice in the FCFW algorithm, as this would indicate that no progress is being made, implying that an optimal solution has already been found. Lemma 3.8 generalizes this observation by demonstrating that this principle holds even when the two directions are not exactly the same but are sufficiently close to each other. This result leverages the boundedness of the set and the polyhedral nature of the feasible region.

LEMMA 3.8. *Let ψ be a convex function and u_k be obtained in step k of FCFW by* (2.3)*, i.e., $u_k \in \arg\min_{u \in \mathrm{conv}(\mathcal{V}^{k-1})} \psi(u)$, and $\tilde{v}_k^{FW}$ be obtained by the corresponding FW step* (2.4)*, i.e. $\tilde{v}_k^{FW} \in \arg\max_{v \in \mathcal{V}} \{d_k^\top v\}$, where $d_k = -\nabla\psi(u_k)/\|\nabla\psi(u_k)\|$. The following claims hold:*

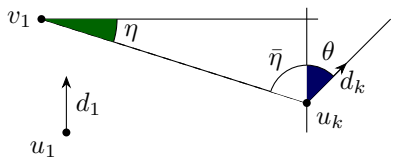
1. *If $\tilde{v}_k^{FW} \in \mathcal{V}^{k-1}$, then u_k is optimal for the original problem $\min_{u \in \mathrm{conv}(\mathcal{V})} \psi(u)$.*
2. *Suppose $d_1 \in \mathbb{R}^n$ is a unit vector and the FW step $\tilde{v}_1^{FW} \in \arg\max_{v \in \mathcal{V}} \{d_1^T v\}$ satisfies $\tilde{v}_1^{FW} \in \mathcal{V}^{k-1}$. Let θ be the positive angle between d_1 and d_k. Then, if θ is small enough, i.e. $\sin(\theta) < \frac{\gamma}{3D}$, u_k is again optimal for the original problem.*

Proof. We first prove claim 1. Suppose that $\tilde{v}_k^{FW} \in \mathcal{V}^{k-1}$. Inequality (2.9) and

the first-order optimality of u_k in the feasible direction $\tilde{v}_k^{FW} - u_k$ of $\mathcal{V}^{k-1}$ implies

$$d^{primal}(u_k) \leq d^{FW}(u_k) = \langle d_k, \tilde{v}_k^{FW} - u_k \rangle \leq 0,$$

which shows that u_k is optimal for the original problem $\min_{u \in \text{conv}(\mathcal{V})} \psi(u)$.
To prove claim 2 by the contrapositive, suppose u_k is not optimal, then by claim 1, we know $\tilde{v}_k^{FW} \notin \mathcal{V}^{k-1}$. Suppose $\theta \leq \frac{\pi}{2}$. The definition of the Pyramidal Width and the assumption that $\tilde{v}_k^{FW} \notin \mathcal{V}^{k-1}$ gives $\langle d_k, \tilde{v}_k^{FW} - u_k \rangle \geq \gamma$.



FCFW : Explored directions

$$\begin{aligned} \langle d_1, \tilde{v}_1^{FW} - u_k \rangle \geq \langle d_1, \tilde{v}_k^{FW} - u_k \rangle &= \langle d_k, \tilde{v}_k^{FW} - u_k \rangle + \langle d_1 - d_k, \tilde{v}_k^{FW} - u_k \rangle \\ &\geq \gamma - \|d_1 - d_k\| \|\tilde{v}_k^{FW} - u_k\| \geq \gamma - \|d_1 - d_k\| D. \end{aligned} \tag{3.18}$$

The first inequality uses the optimality of $\tilde{v}_1^{FW}$, the next inequality uses Cauchy-Schwarz and the pyramidal width. The last inequality uses $u_k, \tilde{v}_k^{FW} \in \mathcal{V}$.

If $u_k = \tilde{v}_1^{FW}$, then from inequality (3.18), $\|d_1 - d_k\| \geq \gamma/D$. We get $\gamma/D \leq \|d_1 - d_k\| = 2\|\frac{d_1+d_k}{2} - d_k\| = 2\sin\left(\frac{\theta}{2}\right) \leq 2\sin(\theta)$, which proves the result in this case.

We now suppose that $\|u_k - \tilde{v}_1^{FW}\| > 0$ and lower bound this quantity. Let Π be the plane containing u_k and the directions d_1 and d_k. Let v_1 be the orthogonal projection of $\tilde{v}_1^{FW}$ onto Π. Then, since projections on convex sets are contractive,

$$\|u_k - \tilde{v}_1^{FW}\| \geq \|u_k - v_1\|. \tag{3.19}$$

We define the angle η such that $\bar{\eta} \stackrel{\text{def}}{=} \pi/2 - \eta$ is the positive angle between d_1 and $v_1 - u_k$. Since v_1 is the orthogonal projection of $\tilde{v}_1^{FW}$ on Π, we have $\langle d_k, \tilde{v}_1^{FW} - u_k \rangle = \langle d_k, v_1 - u_k \rangle$. Thus, $\langle d_k, v_1 - u_k \rangle = \langle d_k, \tilde{v}_1^{FW} - u_k \rangle \leq 0$. The inequality follows from the optimality of $u_k \in \arg\min_{u \in \text{conv}(\mathcal{V}^{k-1})} \psi(u)$ and $\tilde{v}_1^{FW} \in \mathcal{V}^{k-1}$. Remark that $\cos(\bar{\eta} + \theta)\|v_1 - u_k\| = \langle d_k, v_1 - u_k \rangle \leq 0$. Hence, $\bar{\eta} + \theta \geq \pi/2$. We conclude that $\eta \leq \theta$.

$$\langle d_1, v_1 - u_k \rangle = \cos(\frac{\pi}{2} - \eta)\|u_k - v_1\| = \sin(\eta)\|u_k - v_1\| \leq \sin(\theta)\|u_k - v_1\|. \tag{3.20}$$

Combining (3.18), (3.19), (3.20), we have

$$\sin(\theta)\|u_k - \tilde{v}_1^{FW}\| \geq \langle d_1, \tilde{v}_1^{FW} - u_k \rangle \geq (\gamma - \|d_1 - d_k\| D). \tag{3.21}$$

Now rearrange and apply (3.21) and (3.18),

$$\begin{aligned} \sin(\theta) &\geq \frac{\langle d_1, \tilde{v}_1^{FW} - u_k \rangle}{\|u_k - \tilde{v}_1^{FW}\|} \geq \frac{\gamma}{D} - \frac{D\|d_1 - d_k\|}{\|u_k - \tilde{v}_1^{FW}\|} \\ &\geq \frac{\gamma}{D} - \frac{D\|d_1 - d_k\|}{(\gamma - \|d_1 - d_k\| D)\sin(\theta)^{-1}} = \frac{\gamma}{D} - \frac{\sin(\theta) D\|d_1 - d_k\|}{\gamma - \|d_1 - d_k\| D}, \end{aligned} \tag{3.22}$$

where (3.22) uses (3.21) a second time.

Rearranging and using $\|d_1 - d_k\| = 2\|\frac{d_1+d_k}{2} - d_k\| = 2\sin\left(\frac{\theta}{2}\right) \leq 2\sin(\theta)$,

$$\sin(\theta)\left(1 + \frac{D\|d_1 - d_k\|}{\gamma - \|d_1 - d_k\| D}\right) \geq \frac{\gamma}{D},$$

$$\implies \sin(\theta)\left(1+\frac{2\sin(\theta)D}{\gamma-2\sin(\theta)D}\right)\geq\frac{\gamma}{D}.$$

If $\gamma-2\sin(\theta)D\leq 0$, then $\sin(\theta)\geq\frac{\gamma}{D}$.

Otherwise,

$$\gamma\sin(\theta)\geq\frac{\gamma}{D}(\gamma-2\sin(\theta)D),$$
$$\text{hence,}\quad \sin(\theta)\geq\frac{\gamma}{3D}.$$

In both cases, $\sin(\theta)\geq\frac{\gamma}{3D}$. ∎

For $k\in\mathbb{N}$, we define $\hat{s}_k:(v,w)\mapsto\phi(v)+\frac{1}{2\rho}\|v+\rho x_k-w\|^2$.

Let $s_k:(v,w,\beta)\mapsto\hat{s}_k(v,w)-\beta$. We define $v_{k+1}\overset{\text{def}}{=}\arg\min_{v\in\mathbb{R}^n}s_k(v,w_{k+1},\beta_{k+1})$. Note that $s_k(v_{k+1},w_{k+1},\beta_{k+1})=\mathcal{M}_{\rho,\phi}(w_{k+1}-\rho x_k)-\beta_{k+1}$.

LEMMA 3.9. *$\hat{s}_k$ is α-strongly convex, with*

$$\alpha=-\frac{1}{2}\sqrt{\frac{1}{L_g^2}+\frac{4}{\rho^2}}+\frac{1}{2L_g}+\frac{1}{\rho}. \tag{3.23}$$

Proof. As g is L_g-smooth, ϕ is L_g^{-1}-strongly convex. We only need to show that $(v,w)\mapsto\frac{1}{2L_g}\|v\|^2+\frac{1}{2\rho}\|v+\rho x_j-w\|^2$ is strongly convex. We define H, the Hessian of this function:

$$H=\frac{1}{L_g}\begin{bmatrix}I_n & 0\\ 0 & 0\end{bmatrix}+\frac{1}{\rho}\begin{bmatrix}I_n & -I_n\\ -I_n & I_n\end{bmatrix}.$$

A direct algebraic derivation of the smallest eigenvalue of H would prove (3.23). Instead, we relate this computation to a more general result on the smallest eigenvalue of the sum of two projections. The Hessian H can be written $H=aP_1+bP_2$, with $a=\frac{1}{L_g}$, $b=\frac{2}{\rho}$, P_1, the orthogonal projection matrix onto the $\mathbb{R}_n^{2n}$, the first n coordinates of $\mathbb{R}^{2n}$ and P_2, the orthogonal projection matrix onto D_{2n} the subspace $\{(u,-u)\mid u\in\mathbb{R}^n\}\subset\mathbb{R}^{2n}$. Following [19, p. 77], we know that the smallest eigenvalue of aP_1+bP_2 is $\frac{1}{2}(a+b)-\sqrt{(a+b)^2-4ab\sin^2\theta_1}$, where $\theta_1=\pi/4$ is the Friedrichs angle between $\mathbb{R}_n^{2n}$ and D_{2n}. This yields (3.23). ∎

We consider step $(k-1)$, which we suppose is a *serious step* (so that $x_k=y_k$). The following lemma establishes a crucial link: when two consecutive serious primal iterates, x_k and x_{k-1}, are close, the corresponding gradients associated with the fully corrective step (3.2) are also close. Our proof draws upon Lemma 3.9, the serious nature of step x_k, and the inherent smoothness of the Moreau envelope.

LEMMA 3.10. *Assume step $(k-1)$ is a serious step. Let $(v_{i+1},w_{i+1},\beta_{i+1})$ be the optimal solution of (3.2) in step i with proximal center x_i, for $i=k-1,k$. Then,*

$$\|\nabla\mathcal{M}_{\rho,\phi}(w_k-\rho x_{k-1})-\nabla\mathcal{M}_{\rho,\phi}(w_{k+1}-\rho x_k)\|\leq(1+\frac{5}{\alpha\rho})\max(\sqrt{\frac{\epsilon\alpha}{2}},\|x_{k-1}-x_k\|).$$

Proof. By Lemma 3.9, $\hat{s}_k$ is α-strongly convex. We apply Lemma A.1 to $p(v,w)+q(\beta)$, where $p(v,w)=\hat{s}_k(v,w)$ and $q(\beta)=-\beta$,

$$\hat{s}_k(v_{k+1},w_{k+1})-\beta_{k+1}$$

$$
\begin{aligned}
&\leq \hat{s}_k(v_k, w_k) - \beta_k - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) && (3.24)\\
&= \phi(v_k) + \frac{1}{2\rho}\|v_k + \rho x_{k-1} - w_k\|^2 + \langle x_k - x_{k-1}, v_k + \rho x_{k-1} - w_k\rangle \\
&\quad + \frac{\rho}{2}\|x_k - x_{k-1}\|^2 - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) - \beta_k \\
&= s_{k-1}(v_k, w_k, \beta_k) + \langle x_k - x_{k-1}, v_k + \rho x_{k-1} - w_k\rangle + \frac{\rho}{2}\|x_k - x_{k-1}\|^2 \\
&\quad - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2). && (3.25)
\end{aligned}
$$

The first inequality (3.24) uses Lemma A.1 and the optimality of $(v_{k+1}, w_{k+1}, \beta_{k+1})$ for (3.2) with proximal center x_k over the convex set $\mathbb{R}^n \times \mathcal{V}^k \supset \mathbb{R}^n \times \mathcal{V}^{k-1}$. Following the definition of the FW primal gap (3.16), define

$$
d_{k-1}^{primal}(w_k, \beta_k) \stackrel{\text{def}}{=} s_{k-1}(v_k, w_k, \beta_k) - \min_{(v,w,\beta)\in\mathbb{R}^n\times\text{conv}(\mathcal{V})} s_{k-1}(v, w, \beta).
$$

As pointed out in Definition 2.3, $d_{k-1}^{primal}(w_k, \beta_k) \leq d_{k-1}^{FW}(w_k, \beta_k)$. Step $(k-1)$ is a *serious step* and Lemma 3.3 shows that the Frank-Wolfe gap corresponds to the *null step test* quantity. We get $d_{k-1}^{primal}(w_k, \beta_k) \leq d_{k-1}^{FW}(w_k, \beta_k) \leq \delta = \frac{\epsilon}{2}$.

$$
\begin{aligned}
s_{k-1}(v_k, w_k, \beta_k) &= \min_{(v,w,\beta)\in\mathbb{R}^n\times\text{conv}(\mathcal{V})} s_{k-1}(v, w, \beta) + d_{k-1}^{primal}(w_k, \beta_k) \\
&\leq \min_{(v,w,\beta)\in\mathbb{R}^n\times\text{conv}(\mathcal{V})} s_{k-1}(v, w, \beta) + \frac{\epsilon}{2} \\
&\leq s_{k-1}(v_{k+1}, w_{k+1}, \beta_{k+1}) + \frac{\epsilon}{2}. && (3.26)
\end{aligned}
$$

We combine (3.25) and (3.26), we develop and reorganize as follows:

$$
\begin{aligned}
0 &\leq s_{k-1}(v_{k+1}, w_{k+1}, \beta_{k+1}) - s_k(v_{k+1}, w_{k+1}, \beta_{k+1}) + \frac{\epsilon}{2} + \langle x_k - x_{k-1}, v_k + \rho x_{k-1} - w_k\rangle \\
&\quad + \frac{\rho}{2}\|x_k - x_{k-1}\|^2 - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) \\
&= \frac{1}{2\rho}\|v_{k+1} + \rho x_{k-1} - w_{k+1}\|^2 - \frac{1}{2\rho}\|v_{k+1} + \rho x_k - w_{k+1}\|^2 + \frac{\epsilon}{2} + \frac{\rho}{2}\|x_k - x_{k-1}\|^2 \\
&\quad + \langle x_k - x_{k-1}, v_k + \rho x_{k-1} - w_k\rangle - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) \\
&= \langle x_{k-1} - x_k, (v_{k+1} - v_k) - (w_{k+1} - w_k)\rangle - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) + \frac{\epsilon}{2} \\
&\leq \|x_{k-1} - x_k\|(\|v_{k+1} - v_k\| + \|w_{k+1} - w_k\|) - \frac{\alpha}{2}(\|v_k - v_{k+1}\|^2 + \|w_k - w_{k+1}\|^2) + \frac{\epsilon}{2} \\
&\leq -\frac{\alpha}{4}(\|v_k - v_{k+1}\| + \|w_k - w_{k+1}\|)^2 + \|x_{k-1} - x_k\|(\|v_{k+1} - v_k\| + \|w_{k+1} - w_k\|) + \frac{\epsilon}{2}. && (3.27)
\end{aligned}
$$

The penultimate inequality follows from the Cauchy-Schwarz inequality, while the final inequality leverages the identity $2(a^2 + b^2) \geq (a + b)^2$.

The expression (3.27) is a quadratic polynomial P in $(\|v_k - v_{k+1}\| + \|w_k - w_{k+1}\|)$ with a negative leading coefficient. $P(0) > 0$ and $\|v_k - v_{k+1}\| + \|w_k - w_{k+1}\| \geq 0$. (3.27) implies that $\|v_k - v_{k+1}\| + \|w_k - w_{k+1}\| \leq x^+$ where x^+ is the largest root of P.

$$
\|v_k - v_{k+1}\| + \|w_k - w_{k+1}\| \leq 2\frac{\|x_{k-1} - x_k\| + \sqrt{\|x_{k-1} - x_k\|^2 + \epsilon\alpha/2}}{\alpha}
$$

$$\leq \frac{2(1+\sqrt{2})}{\alpha}\max(\sqrt{\frac{\epsilon\alpha}{2}}, \|x_{k-1}-x_k\|).$$

As $\|v_k - v_{k+1}\| \geq 0$, $\|w_k - w_{k+1}\| \leq \frac{5}{\alpha}\max(\sqrt{\frac{\epsilon\alpha}{2}}, \|x_{k-1}-x_k\|)$.

Using the $\frac{1}{\rho}$-smoothness of $\mathcal{M}_{\rho,\phi}$:

$$\begin{aligned}\|\nabla\mathcal{M}_{\rho,\phi}(w_k-\rho x_{k-1}) - \nabla\mathcal{M}_{\rho,\phi}(w_{k+1}-\rho x_k)\| &\leq \frac{1}{\rho}(\|w_k - w_{k+1}\| + \|\rho(x_{k-1}-x_k)\|)\\ &\leq (1+\frac{5}{\alpha\rho})\max(\sqrt{\frac{\epsilon\alpha}{2}}, \|x_{k-1}-x_k\|)\square\end{aligned}$$

The following lemma gives conditions that ensure that the angle between the consecutive iterates y_k and y_{k+1} is small enough to apply Lemma 3.8 and conclude that step k is a *serious step.*

LEMMA 3.11. *Let $A \geq \sqrt{\alpha/2}$. In the execution of Algorithm* 3.1*, suppose, step $(k-1)$ is a serious step, $\|y_k\| \geq \sqrt{\epsilon}\frac{6A(1+5(\alpha\rho)^{-1})D}{\gamma}$, and $\|x_{k-1}-x_k\| \leq A\sqrt{\epsilon}$. Then, the step k is also a serious step.*

Proof. Following (3.3) in Algorithm 3.2 :

$$y_k = -\nabla\mathcal{M}_{\rho,\phi}(w_k-\rho x_{k-1}) \quad \text{and} \quad y_{k+1} = -\nabla\mathcal{M}_{\rho,\phi}(w_{k+1}-\rho x_k).$$

Let θ be the angle between y_k and y_{k+1} :

$$\|y_k - y_{k+1}\|^2 = \|y_k\|^2\sin(\theta)^2 + (\|y_{k+1}\| - \|y_k\|\cos(\theta))^2 \geq \|y_k\|^2\sin(\theta)^2. \tag{3.28}$$

Starting from (3.28) and applying Lemma 3.10:

$$\begin{aligned}\sin(\theta) \leq \frac{\|y_k - y_{k+1}\|}{\|y_k\|} &\leq \frac{\gamma(1+5(\alpha\rho)^{-1})\max(\|x_j - x_{j+1}\|, \sqrt{\epsilon\alpha/2})}{6(1+5(\alpha\rho)^{-1})\sqrt{\epsilon}AD}\\ &= \frac{\gamma\max(\|x_{k-1}-x_k\|, \sqrt{\epsilon\alpha/2})}{6\sqrt{\epsilon}AD} \leq \frac{\gamma\max(1, \frac{\sqrt{\alpha/2}}{A})}{6D} \leq \frac{\gamma}{6D}.\end{aligned} \tag{3.29}$$

As $\gamma/(6D) < 1$ by the definition of pyramidal width, (3.29) leads to $\|y_k - y_{k+1}\| < \|y_k\|$, which shows that $\cos(\theta) > 0$. Thus $0 \leq \theta < \pi/2$. The inequality (3.29) lower-bounds the angle between y_k and y_{k+1}. At step $(k-1)$, the negative gradient direction $-\nabla\Psi_{k-1}(w_k, \beta_k) = \binom{-\nabla\mathcal{M}_{\rho,\phi}(w_k-\rho x_{k-1})}{\nabla(-Id)} = \binom{y_k}{-1}$ is explored. To fulfill the hypothesis of Lemma 3.8, we will upper bound the angle θ' between $\binom{y_k}{\nabla(-Id)} = \binom{y_k}{-1}$ and $\binom{y_{k+1}}{-1}$ by taking into account the $(n+1)^{th}$ dimension of this gradient. We can restrict ourselves to a 3-dimensional subspace of $\mathbb{R}^{n+1}$ that includes $\binom{y_k}{0}$, $\binom{y_{k+1}}{0}$, and $\binom{0_{\mathbb{R}^n}}{1}$. We denote $\tilde{y}_k$ and $\tilde{y}_{k+1}$ the restriction of $\binom{y_k}{-1}$ and $\binom{y_{k+1}}{-1}$ in this subspace. We can find an orthonormal basis of this subspace such that the coordinates of $\tilde{y}_k$ and $\tilde{y}_{k+1}$ are:

$$\tilde{y}_k = \begin{bmatrix} r \\ 0 \\ -1 \end{bmatrix} \quad \text{and} \quad \tilde{y}_{k+1} = \begin{bmatrix} r'\cos(\theta) \\ r'\sin(\theta) \\ -1 \end{bmatrix}. \quad \text{Note that} \quad \tilde{y}_k \times \tilde{y}_{k+1} = \begin{bmatrix} r'\sin(\theta) \\ r - r'\cos(\theta) \\ rr'\sin(\theta) \end{bmatrix},$$

where $\tilde{y}_k \times \tilde{y}_{k+1}$ denotes the cross product of $\tilde{y}_k$ and $\tilde{y}_{k+1}$.

$$\|\tilde{y}_k \times \tilde{y}_{k+1}\|^2 = r'^2 + r^2 - 2rr'\cos(\theta) + r^2r'^2\sin^2(\theta)$$

$$\implies \sin^2(\theta') = \frac{r'^2 + r^2 - 2rr'\cos(\theta) + r^2r'^2\sin^2(\theta)}{\|\tilde{y}_k\|^2\|\tilde{y}_{k+1}\|^2}$$

$$= \frac{(r-r')^2 + 4rr'\sin^2\left(\frac{\theta}{2}\right) + r^2r'^2\sin^2(\theta)}{(r^2+1)(r'^2+1)} \tag{3.29}$$

$$< \frac{(r-r')^2}{(r^2+1)\cdot 1} + \frac{4rr'}{r^2+r'^2}\sin^2(\theta) + \sin^2(\theta)$$

$$\leq \frac{\|y_k - y_{k+1}\|^2}{\|y_k\|^2} + 2\sin^2(\theta) + \sin^2(\theta) \tag{3.30}$$

$$\leq \left(\frac{\gamma}{6D}\right)^2 + 3\sin^2(\theta) \leq 4\left(\frac{\gamma}{6D}\right)^2.$$

Equation (3.29) uses $\cos(\theta) = 1 - 2\sin(\theta/2)^2$ and (3.30) uses $|r - r'| \leq \|y_k - y_{k+1}\|$. Taking the square root, we conclude that $\sin(\theta') < \frac{\gamma}{3D}$ with $0 \leq \theta < \frac{\pi}{2}$.

We consider the execution of the FCFW algorithm applied to the problem (3.8). At step k, when the fully corrective step (3.2) is computed, the direction $\tilde{y}_k$ has already been explored at the previous iteration k. We have shown that θ' the angle between $\tilde{y}_{k+1}$, the gradient of the iterate (w_{k+1}, β_{k+1}), and $\tilde{y}_k$ is such that $\sin(\theta') < \frac{\gamma}{3D}$ and $\theta' \leq \frac{\pi}{2}$. We apply Lemma 3.8 with $d_1 = \tilde{y}_k$ and the iterate $u_k = (w_{k+1}, \beta_{k+1})$ to deduce that u_k is optimal for the *null steps* problem (3.8). In particular, this shows the Frank-Wolfe gap is zero and that k is a *serious step*. ∎

The number of *serious steps* immediately followed by another *serious step* is upper-bounded by the total number of *serious steps* given in Lemma 3.5. We now bound the number of *serious steps* followed by at least one *null step*. This is done by bounding the number of *serious steps* for which one of the hypotheses of Lemma 3.11 is violated. Intuitively, looking at (3.13), a large distance between consecutive *serious steps* iterates ensures a sufficient decrease in function value. Also, when a small gradient y_k is encountered, the number of remaining *serious steps* is small. It is the subject of the next two lemmas. We use step k_i to denote the i^{th} *serious step*.

LEMMA 3.12. *In the execution of Algorithm* 3.1*, there are at most* $\frac{2\|x_0 - x^*\|^2}{\epsilon A^2} + \frac{2}{A^2\rho}$ *serious iterates such that* $\|x_{k-1} - x_k\| > A\sqrt{\epsilon}$ *before reaching an* ϵ*-optimal solution.*

Proof. Using (3.13), for all $T \in \mathbb{N} \setminus \{0\}$:

$$0 \leq \min_{i=1,\dots,T}\{h(x_{k_i}) - h(x^*)\} \leq \frac{\rho}{2T}\|x_0 - x^*\|^2 - \frac{\rho}{2T}\sum_{i=1}^{T}\|x_{k_i} - x_{k_i-1}\|^2 + \frac{\epsilon}{2}.$$

$$\sum_{i=1}^{T}\|x_{k_i} - x_{k_i-1}\|^2 \leq \frac{T\epsilon}{\rho} + \|x^* - x_0\|^2. \tag{3.31}$$

According to Lemma 3.5, we reach an ϵ-optimal solution in at most $T = \frac{\rho\|x_0 - x^*\|^2}{\epsilon} + 1$ *serious steps*. Let $K \in \mathbb{N}$ be the number of iterates such that $\|x_{k-1} - x_k\| > A\sqrt{\epsilon}$. Following (3.31), $KA^2\epsilon \leq \frac{T\epsilon}{\rho} + \|x^* - x_0\|^2$. We deduce that

$$K \leq \frac{T}{A^2\rho} + \frac{\|x^* - x_0\|^2}{A^2\epsilon} \leq \frac{2}{A^2\rho}\left(\frac{\rho\|x_0 - x^*\|^2}{\epsilon} + 1\right) = \frac{2\|x_0 - x^*\|^2}{\epsilon A^2} + \frac{2}{A^2\rho}. \quad ∎$$

LEMMA 3.13. *Let* $M > 0$*. In the execution of Algorithm* 3.1*, suppose that the serious iterate* y_k *is such that* $\|y_k\| \leq \sqrt{\epsilon}M$*, we find an* ϵ*-optimal solution after at most* $\lceil \rho M^2 \rceil$ *serious steps.*

Proof. We recall the convergence guarantee in terms of the number of *serious steps* given in (3.13). Suppose the iterate $y_k = x_{k_i}$ corresponds to the i^{th} *serious step*. For *serious steps* x_{k_l} and $x_{k_{l-1}}$, it holds

$$h(x_{k_l}) - h(x^*) \leq \frac{\rho}{2}(\|x^* - x_{k_{l-1}}\|^2 - \|x_{k_l} - x^*\|^2) + \delta - \frac{\rho}{2}\|x_{k_{l-1}} - x_{k_l}\|^2.$$

Let $T_i \geq 1$. We have

$$\frac{1}{T_i}\sum_{l=i}^{T_i+i-1} h(x_{k_l}) - h(x^*) \leq \frac{\rho}{2T_i}\|x^* - x_{k_i}\|^2 + \delta - \frac{1}{T_i}\sum_{l=i}^{T_i+i-1}\frac{\rho}{2}\|x_{k_{l-1}} - x_{k_l}\|^2,$$

$$\min_{l\in[1,T_i+i-1]}\{h(x_{k_l})\} - h(x^*) \leq \frac{\|x^* - x_{k_i}\|^2\rho}{2T_i} + \delta.$$

As $\delta = \frac{\epsilon}{2}$, an ϵ-optimal solution is found if $\frac{\|x^*-x_{k_i}\|^2\rho}{2T_i} \leq \frac{\epsilon}{2}$. The algorithm stops after at most T_i iterations with

$$\lfloor T_i \rfloor \leq \frac{\|x^* - x_{k_i}\|^2\rho}{\epsilon} = \frac{\|y_k\|^2\rho}{\epsilon} \leq \rho\left(\sqrt{\epsilon}M\right)^2\epsilon^{-1} = \rho M^2.$$

This inequality holds as we supposed that $x^* = 0$. We can upper bound the number of remaining iterations by $\lceil \rho M^2 \rceil$, concluding the proof of the lemma. ∎

We now give the main result of this section which improves upon the convergence rate given in Theorem 3.7. We distinguish between "consecutive" *serious steps* that satisfy the assumptions of Lemma 3.11 and will thus be immediately followed by another *serious step*, and "distant center" or "small norm" *serious steps* that will be followed by $\mathcal{O}(\log(1/\epsilon))$ *null steps* (Lemma 3.6). We find an optimal value for the constant A (which does not appear in the algorithm). We also give the optimal parameter ρ and the corresponding convergence rate.

THEOREM 3.14. *The Algorithm* 3.1*, with null step test parameter* $\delta = \frac{\epsilon}{2}$ *and first iterate* x_0*, finds an* ϵ*-optimal solution in at most* $\frac{\rho\|x_0-x^*\|^2}{\epsilon} + \mathcal{O}(\frac{1}{\rho^{3/2}\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right))$ *iterations, which is minimized to be* $\mathcal{O}(\frac{1}{\epsilon^{4/5}}\log(\frac{1}{\epsilon})^{2/5})$ *when* $\rho = \epsilon^{1/5}\log(\frac{1}{\epsilon})^{2/5}$.

Proof. The proof is made by case analysis. We first suppose that $A \geq \sqrt{\alpha/2}$. As long as all *serious step* iterates y_k remain sufficiently distant from $x^* = 0$, ensuring that no "small norm" step occurs (i.e., $\|y_k\| \geq \sqrt{\epsilon}\frac{6A(1+5(\alpha\rho)^{-1})D}{\gamma}$), the following holds:

1. According to Lemma 3.12 among these *serious steps*, at most $\frac{2\|x_0-x^*\|^2}{\epsilon A^2} + \frac{2}{A^2\rho}$ will be "distant center" *serious steps* (such that $\|x_{k-1} - x_k\| > A\sqrt{\epsilon}$).
2. We can apply Lemma 3.11 to the other *serious steps* and show that they are "consecutive". We upper bound the number "consecutive" *serious steps* by the total number of *serious steps* $\frac{\rho\|x_0-x^*\|}{\epsilon} + 1$ given in Lemma 3.5.

We now suppose that a "small norm" *serious step* y_k occurs. Lemma 3.13 shows that there will be at most $\rho\left(\frac{6(1+5(\alpha\rho)^{-1})AD}{\gamma}\right)^2$ *serious steps* after that.

We use the upper bound on the number of *null steps* between two *serious steps* given by Lemma 3.6. The total number of steps is upper bounded by

$$\underbrace{1 + \frac{\rho\|x_0 - x^*\|^2}{\epsilon}}_{\text{consecutive serious steps}} +$$

$$\left(1+\underbrace{\rho\left(\frac{6(1+5(\alpha\rho)^{-1})AD}{\gamma}\right)}_{\text{after small norm occurs}}^{2}+\underbrace{\frac{2\|x_0-x^*\|^2}{\epsilon A^2}+\frac{2}{A^2\rho}}_{\text{distant center steps}}\right)\underbrace{\left[1+\max\left\{2,\frac{D^2}{\bar{\mu}_{\psi,\rho}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right)\right]}_{\text{number of consecutive null steps}}$$

We choose A to minimize $\rho\left(\frac{6(1+5(\alpha\rho)^{-1})AD}{\gamma}\right)^2+\frac{2\|x_0-x^*\|^2}{\epsilon A^2}$:

$$A^2=\frac{\sqrt{2}\|x_0-x^*\|\gamma}{6(1+5(\alpha\rho)^{-1})D\sqrt{\rho}}\epsilon^{-1/2}, \tag{3.32}$$

The number of *serious steps* with *null steps* considered here is bounded by

$$K\frac{\sqrt{\rho}}{\sqrt{\epsilon}} \quad \text{with} \quad K\stackrel{\text{def}}{=}2\cdot\frac{6\sqrt{2}(1+5(\alpha\rho)^{-1})D\|x_0-x^*\|}{\gamma}.$$

Note that when $\rho\to 0$, $(\alpha\rho)^{-1}=\mathcal{O}(\frac{\rho}{2L_g})$. We now suppose that $A<\sqrt{\alpha/2}$ so that, following (3.32) :

$$\frac{\sqrt{2}\|x_0-x^*\|\gamma}{6(1+5(\alpha\rho)^{-1})D\sqrt{\rho}}\epsilon^{-1/2}<\frac{\alpha}{2}.$$

This gives an inequality involving the upper bound of the number of *null steps*.

$$\begin{aligned}\frac{\rho\|x_0-x^*\|^2}{\epsilon}&<\frac{6^2(1+5(\alpha\rho)^{-1})^2\alpha^2D^2\rho^2}{2\cdot 4\cdot\gamma^2}=\frac{9(\alpha\rho+5)^2D^2}{2\gamma^2}\\&\leq\frac{9(1+\frac{\rho}{2L_g}+5)^2D^2}{2\gamma^2}=\frac{D^2}{\gamma^2}\left(162+\frac{9\rho^2}{4L_g^2}\right).\end{aligned}$$

In this case, we upper bound the total number of iterations by

$$\begin{aligned}&\left(\frac{\rho\|x_0-x^*\|^2}{\epsilon}+1\right)\left[1+\max\left\{2,\frac{D^2}{\bar{\mu}_{\psi,\rho}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right)\right]\\&\quad\leq\frac{D^2}{\gamma^2}\left(163+\frac{9\rho^2}{4L_g^2}\right)\left[1+\max\left\{2,\frac{D^2}{\bar{\mu}_{\psi,\rho}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right)\right].\end{aligned}$$

Putting these two results together, an ϵ-optimal solution is obtained after at most the following number of iterations

$$\underbrace{1+\frac{\rho\|x_0-x^*\|^2}{\epsilon}}_{\text{consecutive serious steps}}+ \tag{3.33}$$

$$\underbrace{\left[1+\frac{K\|x_0-x^*\|\sqrt{\epsilon}}{2\sqrt{\rho}}+\frac{K\sqrt{\rho}}{\sqrt{\epsilon}}+\frac{D^2}{\gamma^2}\left(163+\frac{9\rho^2}{4L_g^2}\right)\right]\left[1+\max\left\{2,\frac{D^2}{\bar{\mu}_{\psi,\rho}\rho\gamma^2}\right\}\log\left(\frac{4D^4}{\epsilon^2\rho^2}\right)\right]}_{\text{Serious steps with } \textit{null steps} \text{ inbetween}}$$

As $\epsilon^{-1/2}\log\left(\frac{1}{\epsilon}\right)=o\left(\frac{1}{\epsilon}\right)$, the overall number of iterations is of the order of $\frac{\rho\|x_0-x^*\|^2}{\epsilon}$ in ϵ. While $\epsilon=o(\rho)$, the term $\frac{K\|x_0-x^*\|\sqrt{\epsilon}}{2\sqrt{\rho}}$ can be neglected in comparison to $\frac{K\sqrt{\rho}}{\sqrt{\epsilon}}$. Also, $\bar{\mu}_{\psi,\rho}^{-1}=\mathcal{O}(1/\rho)$ in the limit of small ρ. The convergence rate (3.33), as $\epsilon\to 0$, simplifies to $\frac{\rho\|x_0-x^*\|^2}{\epsilon}+\mathcal{O}\left(\frac{1}{\rho^{3/2}\sqrt{\epsilon}}\log\left(\frac{1}{\epsilon}\right)\right)$.

By selecting $\rho=\epsilon^{1/5}\log\left(\frac{1}{\epsilon}\right)^{2/5}$, this rate can be expressed as $\mathcal{O}\left(\frac{1}{\epsilon^{4/5}}\log\left(\frac{1}{\epsilon}\right)^{2/5}\right)$.□

4. Numerical Experiments for PBA. In this section, we compare three bundle management policies for PBA. The three policies mainly differ in how to handle the bundle of cuts after *serious steps*. We provide preliminary numerical results to illustrate that the *active-cut policy* that maintains a proper level of model accuracy by keeping *active cuts* after *both serious* and *null steps* has advantages over the *all-cut policy* that may keep too many cuts or the *single-cut policy* that may keep too few cuts. This result aligns with the improved convergence rate from Theorem 3.14 which requires keeping active cuts in the bundle after a *serious step* as well.

In Algorithm 2.1 and Algorithm 3.1, a cut $i \leq k$ is said to be *active* at iteration k if $f_k(y_{k+1}) = (v_i^{FW})^T y_{k+1} + b_i^{FW}$. These active cuts correspond to the supporting hyperplanes of f that have the maximal value at y_{k+1} among all cuts defining f_k.

Three bundle management policies

All-cut policy: retaining all the past cuts, i.e. both active and non-active cuts of the model f_k after every *serious* and *null step* as in the current form of Algorithm 3.1. This policy may lead to growing memory demands and escalate the difficulty in solving the associated minimization problems (3.1).

Single-cut policy: as noted in [12, 30], retaining some approximation accuracy of the previous model is only necessary after *null steps*. To keep the bundle as lean as possible, it is suggested in [29] to drop all but the *newest added cut* after every *serious step* (thus the name) and keep active cuts only after *null steps*.

Active-cut policy: retaining active cuts after *both serious* and *null steps*. Contrary to existing analyses, in the improved complexity bound in Theorem 3.14, we assume that after a *serious step* k, w_k lies within the convex hull of $\mathcal{V}^k$ (see (3.24) in Lemma 3.8), which requires retaining the active cuts after *serious steps*. Since at most $n+1$ cuts are retained during *null step* sequences [13, 45, 44], retaining active cuts after a *serious step* does not increase the maximum memory requirement. To the best of our knowledge, our analysis is the first to give a theoretical ground for retaining active cuts after a *serious step*. Indeed, the all-cut and single-cut policies prove to be less balanced and less effective than the active-cut policy as illustrated by the following numerical experiment.

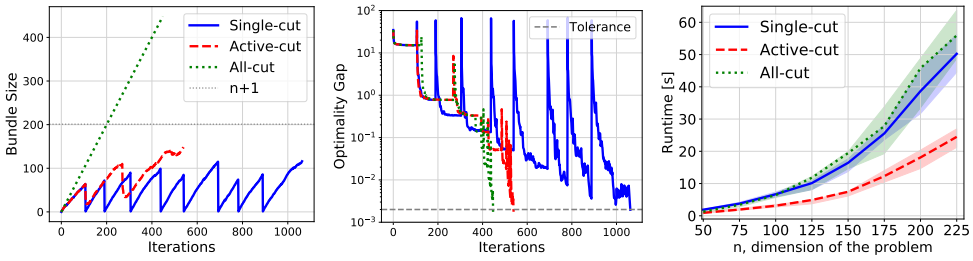


Fig. 1: The two leftmost plots compare the memory usage and the optimality gap across bundle management policies ($n = 200$). The last plot compares their runtime for varying problem dimension n with interquartiles.

Sythetic data experiment. We illustrate the performance gain of the active-cut policy through a numerical study on the problem: $\min \frac{1}{2}\|x\|^2 + \max\{x^T y + d : y \in [-1,1]^n, d \in [-1,1], Ay + cd \leq b\}$. The dimension of the problem, n, is five times the number of constraints m, and A, b, and c are chosen with uniformly random components in $[-1,1]$. The problem is solved with accuracy $\epsilon = 2 \cdot 10^{-3}$ and $\delta = 10^{-3}$,

while ρ is set to 1. The implementation [1] is done in Julia. All experiments were conducted on an Intel Xeon Platinum processor machine featuring 48 cores and 4 GB of RAM per core. This problem captures the essential structure of many composite nonsmooth problems, such as SVM and two-stage robust optimization.

Figure 1 illustrates this tradeoff by comparing the memory usage (bundle size), optimality gap, and total runtime of the three policies. For the two leftmost plots of Figure 1, which illustrate the memory requirement and iteration complexity across the three bundle management policies, we have chosen $n = 200$ and $m = 40$. The rightmost plot shows the runtime of the PBA for the three bundle management policies with instances of increasing problem size. The runtimes have been averaged over 20 random instances, with the shaded area representing the interquartile range.

Retaining more cuts reduces the total number of iterations but increases the time per iteration. The leftmost figure shows the fast growth of bundle size of the all-cut policy (green), the frequent cut cleansing in the single-cut policy (blue), and balanced bundle size of active-cut policy (red). In the middle figure, the all-cut policy has the fewest number of iterations, while cleansing the bundle after *serious steps* as in the single-cut policy leads to wild increase of optimality gap after *serious steps* and slow convergence. The right figure shows that fewer iterations in the all-cut policy does not result in a shorter runtime because the iterations become more computationally expensive. Retaining only the active cuts at *serious step* achieves a good balance and results in the best runtime.

5. Discussions. We compare our results to the complexity lower bounds for PBA and discuss their relation to the best known convergence rates for our setup.

5.1. Complexity lower bounds. Algorithm 2.2, Algorithm 3.2 (and their dual algorithms Algorithm 2.1 and Algorithm 3.1) are linear optimization-based convex programming (LCP) methods as defined in [25]. Algorithm 2.2 is an LCP that minimizes a $1/\mu_g$-smooth function. It is shown in [25] that the worst-case iteration complexity of LCP to solve smooth (even strongly-convex) problems cannot be smaller than $\min\left\{n/2, D^2/(4\mu_g\epsilon)\right\}$.

This aligns with the linear convergence rate given in Theorem 2.7. Indeed, the constants in this rate involve $\frac{D^2}{\bar{\mu}_\psi \mu_g \gamma^2} \geq \frac{D^2}{\gamma^2} \geq \lfloor \frac{n}{2} \rfloor$, with the final inequality regarding the condition number D^2/γ^2 being established in [1]. Likewise, the convergence rate (3.33) presented in the proof of Theorem 3.14 includes a term $163D^2/\gamma^2 \geq n/2$.

5.2. Best known convergence rates. In the setting of section 3, the analysis of Liang and Monteiro [30] provides an $\mathcal{O}(\epsilon^{-2}\log(1/\epsilon))$ iteration complexity. It is worth noting that in the regime where $n \gg \frac{1}{\epsilon}$ (which corresponds to high-dimensional settings or low-accuracy scenarios), the rate given in Theorem 3.7 and the analysis of Liang and Monteiro may be tighter than that of Theorem 3.14.

Theorem 3.14 builds upon [24], which demonstrates a geometry-dependent rate of $\mathcal{O}\left(D^2\gamma^{-2}\log\left(1/\epsilon\right)\right)$ for some Frank-Wolfe variants on polyhedral sets. When the problem must be solved with high accuracy ($\epsilon \ll 1/n$), their result is more favorable than the previously known, dimension-independent, rate of $\mathcal{O}(1/\epsilon)$ for general closed convex sets. Likewise, our geometry-dependent rate is a substantial progress and reveals favorable convergence behavior of PBA for seeking high-accuracy solutions for a broad class of convex composite nonsmooth problems.

[1] https://github.com/dfersztand/Improved-Complexity-Analysis-for-the-Proximal-Bundle-Algorithm-Under-a-Novel-Perspective

The improved the dependence on ϵ given in Theorem 3.14 is possible at the cost of including geometry-dependent constants that can be arbitrarily large and supposing that f is piecewise linear and g is smooth. From a dual perspective, this trade-off in dimension dependence for improved complexity in terms of ϵ is made possible by the structure of problem (2.5). It remains an open question whether a Frank-Wolfe-type algorithm with a similar trade-off can be designed when the feasible region is a general compact convex set. More generally, problem (2.5) is a strongly convex but nonsmooth problem for which none of the classic Frank-Wolfe variants is well suited. Designing a Frank-Wolfe type method with a $\mathcal{O}(1/\epsilon)$ convergence rate for problems like (2.5) (nonsmooth, strongly-convex) was recently identified as an open problem in [2].

Conclusions. Expanding on the duality between Kelley's method and Fully-Corrective Frank-Wolfe to non-homogeneous convex piecewise linear functions, our work addresses a broader class of problems. Our interpretation of *null steps* in the proximal bundle method as dual to the Frank-Wolfe algorithm leads to a convergence rate of $\mathcal{O}(\epsilon^{-1}\log(1/\epsilon))$, surpassing the previous $\mathcal{O}(\epsilon^{-2})$ rate. Further refinement achieves an iteration complexity of $\mathcal{O}(\epsilon^{-4/5}\log(1/\epsilon)^{2/5})$, a notable improvement. We also provide theoretical ground and preliminary numerical results to support the active-cut policy for bundle management.

Appendix A. Definitions and basic properties.

A.1. Technical lemma. It is well known that the sum of a convex function and a strongly convex function defined on the same domain is strongly convex. The following lemma presents a simple result that, to the best of our knowledge, has not been explicitly stated in the existing literature. Specifically, we demonstrate the strong convexity of the sum of a convex function and a strongly convex function defined on the Cartesian product of their domains.

LEMMA A.1. *Let $p : X \subset \mathbb{R}^n \to \mathbb{R}$ be an α-strongly convex differentiable function and $q : Y \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. If (x^*, y^*) minimizes $p(x) + q(y)$ over $(x, y) \in Q \subseteq X \times Y$ a convex set, then*

$$\forall (x,y) \in Q, \quad p(x) + q(y) \geq p(x^*) + q(y^*) + \frac{\alpha}{2}\|x - x^*\|^2. \tag{A.1}$$

Remark: Note that the domains of p and q do not lie in the same space. So $p(x) + q(y)$ is not strongly convex jointly in (x, y).

Proof. By the first order optimality condition, $\langle \nabla p(x^*), x - x^* \rangle + \langle \nabla q(y^*), y - y^* \rangle \geq 0$ for all $(x, y) \in Q$.

$$\begin{aligned}
p(x) + q(y) &= p(x^*) + \int_0^1 \nabla p(tx + (1-t)x^*)^T (x - x^*) dt + q(y) \\
&\geq p(x^*) + \int_0^1 \langle \nabla p(tx + (1-t)x^*) - \nabla p(x^*), (x - x^*) \rangle dt \\
&\quad + q(y^*) + \nabla q(y^*)^T (y - y^*) + \nabla p(x^*)^T (x - x^*) \\
&\geq p(x^*) + q(y^*) + \int_0^1 \frac{1}{t} \langle \nabla p(tx + (1-t)x^*) - \nabla p(x^*), (tx - tx^*) \rangle dt \\
&\geq p(x^*) + q(y^*) + \int_0^1 \frac{1}{t} \alpha \|tx - tx^*\|^2 dt = p(x^*) + q(y^*) + \frac{\alpha}{2}\|x - x^*\|^2.
\end{aligned}$$

The convexity of q gives the first inequality, the second inequality follows from the optimality of (x^*, y^*), and the third inequality uses the strong convexity of p. ∎

Appendix B. Proofs of subsection 3.3.

LEMMA B.1. *For $i \geq 0$, Let $x_i^* = \arg\min_{x \in \mathbb{R}^n} h(x) + \frac{\rho}{2}\|x - x_{j_i}\|^2$, where x_{j_i} are the serious steps iterates in the execution of Algorithm* 3.1. *Then,*

$$\|x_i^*\| \leq \|x^*\| + 2\sqrt{2}\|x^* - x_0\| + 2\sqrt{\frac{\epsilon}{\rho}}.$$

Proof. We use (3.12) for $1 \leq k \leq i$:

$$h_{j_k-1}(x_{j_k}) + \frac{\rho}{2}\|x_{j_k} - x_{j_k-1}\|^2 \leq h_{j_k-1}(x^*) + \frac{\rho}{2}(\|x^* - x_{j_k-1}\|^2 - \|x^* - x_{j_k}\|^2)$$

This leads to :

$$\begin{aligned}
\|x^* - x_{j_k}\|^2 - \|x^* - x_{j_k-1}\|^2 &\leq \frac{2}{\rho}(h_{j_k-1}(x^*) - h_{j_k-1}(x_{j_k})) - \|x_{j_k} - x_{j_k-1}\|^2 \\
&\leq \frac{2}{\rho}(h_{j_k-1}(x^*) - h_{j_k-1}(x_{j_k})) \leq \frac{2}{\rho}(h_{j_k-1}(x^*) + \frac{\epsilon}{2} - h(x_{j_k})) \qquad \text{(serious step)} \\
&\leq \frac{2}{\rho}(h(x^*) + \frac{\epsilon}{2} - h(x_{j_k})) \leq \frac{\epsilon}{\rho}.
\end{aligned}$$

We sum these inequalities from $k = 1$ to i. We use Lemma 3.5 to bound the number of *serious steps* by $\frac{\rho\|x_0 - x^*\|^2}{\epsilon} + 1$:

$$\begin{aligned}
\|x^* - x_{j_i}\|^2 &\leq \|x^* - x_0\|^2 + \frac{\epsilon}{\rho}\left(\frac{\rho\|x_0 - x^*\|^2}{\epsilon} + 1\right) \\
&= \|x^* - x_0\|^2 + \|x_0 - x^*\|^2 + \frac{\epsilon}{\rho} = 2\|x^* - x_0\|^2 + \frac{\epsilon}{\rho}.
\end{aligned}$$

By optimality of x^* and x_i^* for their respective problems:

$$h(x^*) + \frac{\rho}{2}\|x_i^* - x_{j_i}\|^2 \leq h(x_i^*) + \frac{\rho}{2}\|x_i^* - x_{j_i}\|^2 \leq h(x^*) + \frac{\rho}{2}\|x^* - x_{j_i}\|^2.$$

Thus, $\|x_i^* - x_{j_i}\| \leq \|x^* - x_{j_i}\|$. We combine these two inequalities :

$$\begin{aligned}
\|x_i^*\| &\leq \|x^*\| + \|x_i^* - x^*\| \leq \|x^*\| + \|x_i^* - x_{j_i}\| + \|x^* - x_{j_i}\| \leq \|x^*\| + 2\|x^* - x_{j_i}\| \\
&\leq \|x^*\| + 2\sqrt{2\|x^* - x_0\|^2 + \frac{\epsilon}{\rho}} \leq \|x^*\| + 2\sqrt{2}\|x^* - x_0\| + 2\sqrt{\frac{\epsilon}{\rho}}.
\end{aligned}$$

□

REFERENCES

[1] R. ALEXANDER, *The width and diameter of a simplex*, Geometriae Dedicata, 6 (1977), pp. 87–94, https://api.semanticscholar.org/CorpusID:123627093.

[2] K. ASGARI AND M. J. NEELY, *Projection-free non-smooth convex programming*, arXiv, (2023), https://arxiv.org/abs/2208.05127.

[3] D. AZÉ AND J.-P. PENOT, *Uniformly convex and uniformly smooth convex functions*, Annales de la Faculté des Sciences de Toulouse. Mathématiques. Série 6, 4 (1995), pp. 705–730, http://www.numdam.org/item?id=AFST_1995_6_4_4_705_0.

[4] F. BACH, *Learning with Submodular Functions: A Convex Optimization Perspective*, Now Publishers Inc., Hanover, MA, USA, 2013, https://doi.org/10.1561/2200000039.

[5] F. BACH, *Duality between subgradient and conditional gradient methods*, SIAM Journal on Optimization, 25 (2015), pp. 115–129, https://doi.org/10.1137/130941961.

[6] A. Beck, *First-Order Methods in Optimization*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017, https://doi.org/10.1137/1.9781611974997.
[7] A. Beck and S. Shtern, *Linearly convergent away-step conditional gradient for non-strongly convex functions*, Mathematical Programming, 164 (2017), pp. 1–27, https://doi.org/10.1007/s10107-016-1069-4.
[8] A. Ben-Tal, L. Ghaoui, and A. Nemirovski, *Robust Optimization*, Princeton University Press, 2009, https://doi.org/10.1515/9781400831050.
[9] D. Bertsimas and D. den Hertog, *Robust and Adaptive Optimization*, Dynamic Ideas LLC, 2022, https://books.google.com/books?id=V_RPzwEACAAJ.
[10] G. Braun, A. Carderera, C. W. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta, *Conditional gradient methods*, 2023, https://arxiv.org/abs/2211.14103.
[11] W. de Oliveira, C. Sagastizábal, and C. Lemaréchal, *Convex proximal bundle methods in depth: a unified analysis for inexact oracles*, Mathematical Programming, 148 (2014), pp. 241–277, https://doi.org/10.1007/s10107-014-0809-6.
[12] M. Diaz and B. Grimmer, *Optimal convergence rates for the proximal bundle method*, SIAM Journal on Optimization, (2023), pp. 424–454, https://doi.org/10.1137/21M1428601.
[13] Y. Du and A. Ruszczyński, *Rate of convergence of the bundle method*, Journal of Optimization Theory and Applications, 173 (2017), pp. 908–922, https://doi.org/10.1007/s10957-017-1108-1.
[14] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval research logistics quarterly., 3 (1956), https://doi.org/10.1002/nav.3800030109.
[15] R. M. Freund and P. Grigas, *New analysis and results for the Frank–Wolfe method*, Mathematical Programming, 155 (2016), pp. 199–230, https://doi.org/10.1007/s10107-014-0841-6.
[16] M. R. Hestenes, *Multiplier and gradient methods*, Journal of Optimization Theory and Applications, 4 (1969), pp. 303–320, https://doi.org/10.1007/BF00927479.
[17] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, Berlin, Heidelberg, 2001, https://doi.org/10.1007/978-3-642-56468-0.
[18] C. A. Holloway, *An extension of the Frank and Wolfe method of feasible directions*, Mathematical Programming, 6 (1974), pp. 14–27, https://doi.org/10.1007/BF01580219.
[19] A. Kaur and S. H. Lui, *New lower bounds on the minimum singular value of a matrix*, Linear Algebra and its Applications, 666 (2023), pp. 62–95, https://doi.org/10.1016/j.laa.2023.02.013.
[20] J. E. Kelley, Jr., *The cutting-plane method for solving convex programs*, Journal of the Society for Industrial and Applied Mathematics, 8 (1960), pp. 703–712, https://doi.org/10.1137/0108053.
[21] K. C. Kiwiel, *Proximity control in bundle methods for convex nondifferentiable minimization*, Mathematical Programming, 46 (1990), pp. 105–122, https://doi.org/10.1007/BF01585731.
[22] M. Kuchlbauer, F. Liers, and M. Stingl, *Adaptive bundle methods for nonlinear robust optimization*, INFORMS Journal on Computing, 34 (2022), pp. 2106–2124, https://doi.org/10.1287/ijoc.2021.1122.
[23] S. Lacoste-Julien and M. Jaggi, *An affine invariant linear convergence analysis for Frank-Wolfe algorithms*, NIPS 2013 Workshop on Greedy Algorithms, Frank-Wolfe and Friends, (2013), https://doi.org/10.48550/arXiv.1312.7864.
[24] S. Lacoste-Julien and M. Jaggi, *On the global linear convergence of frank-wolfe optimization variants*, in NIPS'15, 2015, p. 496–504.
[25] G. Lan, *The complexity of large-scale convex programming under a linear optimization oracle*, 2014, https://arxiv.org/abs/1309.5550.
[26] G. Lan, *Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization*, Mathematical Programming, 149 (2015), pp. 1–45, https://doi.org/10.1007/s10107-013-0737-x.
[27] Q. Le, A. Smola, and S. Vishwanathan, *Bundle methods for machine learning*, Advances in Neural Information Processing Systems, 20 (2007), https://proceedings.neurips.cc/paper_files/paper/2007/file/26337353b7962f533d78c762373b3318-Paper.pdf.
[28] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Mathematical Programming, 69 (1995), pp. 111–147, https://doi.org/10.1007/BF01585555.
[29] J. Liang and R. D. C. Monteiro, *A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes*, SIAM Journal on Optimization, 31 (2021), pp. 2955–2986, https://doi.org/10.1137/20M1327513.
[30] J. Liang and R. D. C. Monteiro, *A unified analysis of a class of proximal bundle methods for solving hybrid convex composite optimization problems*, Mathematics of Operations Research, (2023), https://doi.org/10.1287/moor.2023.1372.
[31] B. Martinet, *Algorithmes Pour La Résolution de Problèmes d'optimisation et de Minimax*,

PhD thesis, Université Scientifique et Médicale de Grenoble, 1972.

[32] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Applied Optimization, Springer New York, NY, 1 ed., 2004, https://doi.org/10.1007/978-1-4419-8853-9.

[33] M. J. D. Powell, *A method for nonlinear constraints in minimization problems*, Optimization: Symposium of the Institute of Mathematics and Its Applications, University of Keele, England, 1968, (1969), pp. 283–298.

[34] R. Rockafellar, *Convex Analysis*, Princeton Mathematics, Princeton University Press, Princeton, New Jersey, 28 ed., 1970, https://doi.org/doi:10.1515/9781400873173.

[35] R. T. Rockafellar, *The multiplier method of hestenes and powell applied to convex programming*, Journal of Optimization Theory and Applications, 12 (1973), pp. 555–562, https://doi.org/10.1007/BF00934777.

[36] R. T. Rockafellar and R. J. B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer, 1998, https://doi.org/10.1007/978-3-642-02431-3.

[37] E. K. Ryu and W. Yin, *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*, Cambridge University Press, Cambridge, UK, new ed., 2022.

[38] H. Schramm and J. Zowe, *Bundle method for nonsmooth optimization: Concept, convergence, and computation*, SIAM Journal on Optimization, 2 (1992), pp. 121–152, https://doi.org/10.1137/0802008.

[39] S. Shalev-Shwartz and Y. Singer, *Efficient learning of label ranking by soft projections onto polyhedra*, Journal of Machine Learning Research, 7 (2006), pp. 1567–1599, http://jmlr.org/papers/v7/shalev-shwartz06a.html.

[40] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: primal estimated sub-gradient solver for SVM*, Mathematical Programming, 127 (2011), pp. 3–30, https://doi.org/10.1007/s10107-010-0420-4.

[41] M. Sion, *On general minimax theorems*, Pacific Journal of Mathematics, 8 (1958), pp. 171–176, https://doi.org/10.2140/pjm.1958.8.171.

[42] X. A. Sun and A. J. Conejo, *Robust Optimization in Electric Energy Systems*, International Series in Operations Research & Management Science, Springer Cham, 2021, https://doi.org/10.1007/978-3-030-85128-6.

[43] C. H. Teo, S. Vishwanthan, A. J. Smola, and Q. V. Le, *Bundle methods for regularized risk minimization*, Journal of Machine Learning Research, 11 (2010), pp. 311–365, http://jmlr.org/papers/v11/teo10a.html.

[44] W. van Ackooij, V. Berge, W. de Oliveira, and C. Sagastizábal, *Probabilistic optimization via approximate p-efficient points and bundle methods*, Computers & Operations Research, (2017), https://doi.org/10.1016/j.cor.2016.08.002.

[45] S. Zhou, S. Gupta, and M. Udell, *Limited memory Kelley's method converges for composite convex and submodular objectives*, in Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18, Red Hook, NY, USA, 2018, pp. 4419–4429, https://doi.org/10.48550/arXiv.1807.07531.