Beyond Alternating Projections: Accelerated Alternating Linear Minimizations

Masterarbeit bei
Prof. Dr. Sebastian Pokutta

vorgelegt von
Garam Kim
Technische Universität Berlin
Fachbereich Mathematik

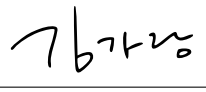
7. Dezember 2023

Sworn Affidavit

in accordance with section § 60 Abs. 8 AllgStuPO:

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, 07.12.2023



Garam Kim

Abstract

The alternating projections method involves finding a pair of points with minimal distance between two sets. While the resulting sequence ensures its convergence, the computational complexity of projection operators is potentially expensive, particularly in higher-dimensional problems. Addressing these challenges, Braun et al. (2023) introduced alternating linear minimizations (`ALM`), replacing projection oracles with linear minimization oracles (`LMO`). This algorithm inherits properties from `LMO`-based methods, such as sparsity of solutions and lower iteration costs. Despite the extensive research on `LMO`-based algorithms, little is known about `ALM` beyond its canonical convergence rate, which is generally non-improvable. This paper presents acceleration schemes for `ALM` by characterizing the geometry of various feasible regions. We propose that when a pair of points with minimal distance is unique, acceleration is possible regardless of how disjointed the two sets are. The results yield several accelerated convergence rates for `ALM`, enhancing its efficiency in solving feasibility problems.

Zusammenfassung

Die Methode der alternierenden Projektionen beinhaltet das Auffinden eines Punktpaares mit minimaler Distanz zwischen zwei Mengen. Obwohl die resultierende Sequenz ihre Konvergenz sicherstellt, ist die Rechenkomplexität der Projektionsoperatoren potenziell kostspielig, insbesondere bei Problemen höherer Dimensionen. Um diesen Herausforderungen zu begegnen, führten Braun et al. (2023) das Verfahren der alternierenden linearen Minimierungen (`ALM`) ein, bei dem Projektionsorakel durch lineare Minimierungsorakel (`LMO`) ersetzt werden. Dieser Algorithmus erbt Eigenschaften von `LMO`-basierten Methoden, wie beispielsweise die Sparsamkeit von Lösungen und niedrigere Iterationskosten. Trotz der umfangreichen Forschung zu `LMO`-basierten Algorithmen ist über `ALM` wenig bekannt, abgesehen von seiner kanonischen Konvergenzrate, die im Allgemeinen nicht verbessert werden kann. Diese Arbeit stellt Beschleunigungsschemata für `ALM` vor, indem die Geometrie verschiedener zulässiger Regionen charakterisiert wird. Wir schlagen vor, dass bei einem Paar Punkten mit minimaler Distanz die Beschleunigung möglich ist, unabhängig davon, wie weit voneinander die beiden Mengen entfernt sind. Die Ergebnisse liefern mehrere beschleunigte Konvergenzraten für `ALM`, was seine Effizienz bei der Lösung von Machbarkeitsproblemen verbessert.

Contents

List of Figures

1 Introduction

In this paper, we are interested in the feasibility problem

$$\text{Find} \quad \mathbf{x} \in \mathcal{P} \cap \mathcal{Q}$$

where $\mathcal{P}, \mathcal{Q} \subseteq \mathbb{R}^d$ are nonempty compact convex sets. This feasibility framework has found applications in various practical domains, including image recovery, signal processing, absolute value equations, and best approximation theory (Combettes, 1996; Combettes and Bondon, 1999; Alcantara et al., 2021; Bauschke and Borwein, 1996). A classical method to find such a point in the intersection of two sets is von Neumann's alternating projection algorithm (Neumann, 1949; Ginat, 2018), which projects a point alternatively onto one set and the other. When $\mathcal{P} \cap \mathcal{Q} \neq \emptyset$, such a point exists, and the algorithm guarantees its convergence in norm (Wiener, 1955; Sakai, 1995; Galantai, 2004) or in finitely many steps under favorable circumstances (Bauschke and Borwein, 1993; Lewis et al., 2007; Lewis and Malick, 2008; Gubin et al., 1967). In case where $\mathcal{P} \cap \mathcal{Q} = \emptyset$, the algorithm converges toward a pair of points with minimal distance between two sets, providing various convergence rates (Cheney and Goldstein, 1959; Censor and Zaknoon, 2018; Bui et al., 2021). Despite the wide literature on these topics, a notable drawback of the algorithm is its sensitivity to initial points, potentially resulting in slow convergence (Bertsekas, 1995; Lewis et al., 2007; Lewis and Malick, 2008). Additionally, the algorithm faces challenges when the sets are disjointed, failing to converge in finitely many steps (Drusvyatskiy et al., 2016; Bui et al., 2021; Behling et al., 2021).

Another significant drawback of alternating projections is their performance in high-dimensional spaces. As the problem dimensionality increases, the projection method encounters challenges regarding convergence speed and computational bottleneck across various applications (Boyd and Vandenberghe, 2004; Combettes and Pokutta, 2021). Furthermore, if the sets are only given implicitly, the traditional projection operators become inapplicable. Recognizing those challenges, Braun et al. (2023) introduced alternating linear minimizations (`ALM`) algorithm. In each iteration, `ALM` performs two linear minimizations instead of the conventional projection method. This approach not only reduces the computational cost per iteration but also produces sparse solutions, providing advantages in high-dimensional spaces (Jaggi, 2011; Clarkson, 2010). In addition, `ALM` ensures convergence and achieves a baseline convergence rate comparable to the alternating projections algorithm (Beck et al., 2015; Lacoste-Julien et al., 2013). In fact, `ALM` shares similarities with the Frank-Wolfe algorithm.

Algorithm 1: Frank-Wolfe algorithm (`FW`)

Input : $\mathbf{x}_0 \in \mathcal{C}$, step-size $\gamma_t \in [0,1]$ for $t \in \{0, \dots, T-1\}$.

for $t = 0, 1 \dots, T-1$ **do**
  $\mathbf{p}_t \in \arg\min_{\mathbf{p} \in \mathcal{C}} \langle \nabla f(\mathbf{x}_t), \mathbf{p} - \mathbf{x}_t \rangle$
  $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \gamma_t(\mathbf{p}_t - \mathbf{x}_t)$
end

The Frank-Wolfe (`FW`) algorithm (Frank and Wolfe, 1956; Levitin and Polyak, 1966) as illustrated in Algorithm 1, is a projection-free approach for solving the constrained convex optimization problem

$$\arg\min_{\mathbf{x} \in \mathcal{C}} f(\mathbf{x})$$

where $\mathcal{C} \subseteq \mathbb{R}^d$ is a compact convex set and $f : \mathcal{C} \to \mathbb{R}$ be a L-smooth convex function. The `FW` algorithm involves first-order access to the objective function f and a linear minimization oracle (`LMO`) for the feasible region $\mathcal{C}$. For given a vector $\mathbf{c} \in \mathbb{R}^d$, the `LMO` returns $\arg\min_{\mathbf{x} \in \mathcal{C}} \langle \mathbf{c}, \mathbf{x} \rangle$. Each iteration of `FW` calls one `LMO` to obtain a vertex $\mathbf{p}_t$ from the set of vertices of $\mathcal{C}$ and takes a step to obtain a new iterate $\mathbf{x}_{t+1} = \mathbf{x}_t + \gamma_t(\mathbf{p}_t - \mathbf{x}_t)$ where $\gamma_t \in [0,1]$ follows a specific step-size rule. Various step-size rules are available, including line-search $\gamma_t \in \arg\min_{\gamma \in [0,1]} f\left((1-\gamma)\mathbf{x}_t + \gamma \mathbf{p}_t\right)$, short-step $\gamma_t \in \min\left\{1, \frac{\langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{p}_t \rangle}{L \|\mathbf{x}_t - \mathbf{p}_t\|_2^2}\right\}$, and open-loop step-size rules $\gamma_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 1}$ (Wirth et al., 2023a,b; Dunn, 1978).

The FW algorithm exhibits a tight convergence rate of $\mathcal{O}(1/T)$ (Canon and Cullum, 1968; Jaggi, 2013; Lan, 2014). However, when the geometry of the feasible regions and the characteristics of the objective function are taken into account, FW can achieve faster convergence rates. For example, when the constrained set $\mathcal{C}$ is strongly convex and the norm of the gradient of the objective function is bounded away from zero, or when the optimum is in the interior of $\mathcal{C}$ and f is a strongly convex function, FW with line-search or short-step admits linear convergence rates (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970; Guélat and Marcotte, 1986). However, when the optimum gets close to the boundary of the feasible region, and both the objective function and the constrained set are strongly convex, the linear convergence regime deteriorates, interpolating a rate of order $\mathcal{O}(1/T)$ to $\mathcal{O}(1/T^2)$ (Garber and Hazan, 2015a). Kerdreux et al. (2021a) generalized their results to the family of uniformly convex sets, and Wirth et al. (2023a,b) further extended the results for FW with the open-loop step-size $\gamma_t = \frac{\ell}{t+\ell}$ for $\ell \in \mathbb{N}_{\geq 1}$.

Besides the extensive research on the accelerated convergence scheme for FW, the acceleration of ALM has not been explored to the best of our knowledge. This paper aims to address this gap by developing accelerated convergence rates for ALM with open-loop step-sizes by characterizing feasible regions.

1.1 Contributions

We develop our understanding of the algorithm of ALM and introduce accelerated convergence schemes of ALM with open-loop step-sizes. Specifically, our contributions are as follows:

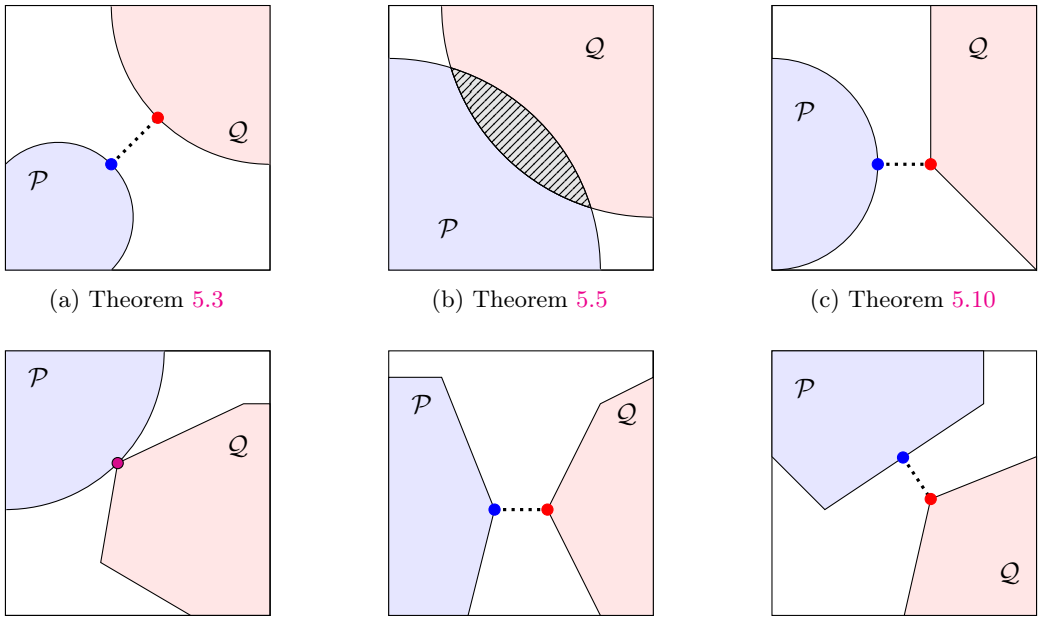
1. We integrate ALM introduced by Braun et al. (2023) with the acceleration developments of FW with open-loop step-sizes, the LMO-based algorithm over one set, proposed by Wirth et al. (2023a). This integration leads to accelerated versions of ALM. Our approach includes the proof of acceleration schemes achieved through categorizing feasible regions based on the geometry of the sets and the relation between them. Details are presented in Table 1.

2. We illustrate our main results through toy examples, highlighting the accelerated results achieved in efficiently finding a pair of points with minimal distance. These examples are provided in each subsection of Section 5, and the code is publicly available on GitHub.

3. In specific polytope settings under strict complementarity assumption where the lower bound captured in Guélat and Marcotte (1986), we demonstrate that ALM with open-loop step-sizes converges non-asymptotically faster than ALM with line-search or short-step.

Table 1: Comparison of convergence rates of ALM for various settings. Compact convexity of $\mathcal{P}$ and $\mathcal{Q}$ is always assumed. The objective function is $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. We denote the optimal solution by $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. "intersect*" indicates the case where two sets intersect at a single point. "str. convex" is an abbreviation for strongly convex and "open-loop" refers to open-loop step-size $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$ and $\nu \in]0, 1]$. Shading groups related regions and corresponding visual descriptions are presented in Figure 1.

References	Region $\mathcal{P}$	Region $\mathcal{Q}$	$\mathcal{P} \cap \mathcal{Q}$	Location of $\mathbf{x}^*$	Location of $\mathbf{y}^*$	Rate	Step-size rule
Proposition 4.8	-	-	-	-	-	$\mathcal{O}(1/t)$	open-loop
Proposition 4.9	-	-	-	-	-	$\mathcal{O}(1/t)$	line-search
Theorem 5.3	str. convex	str. convex	disjoint	boundary	boundary	$\mathcal{O}(1/t^\ell)$	open-loop
Theorem 5.5	str. convex	str. convex	intersect	unrestricted	unrestricted	$\mathcal{O}(1/t^2)$	open-loop
Theorem 5.10	str. convex	polytope	disjoint	boundary	vertex	$\mathcal{O}(1/t^\ell)$	open-loop
Theorem 5.11	str. convex	polytope	intersect*	boundary	vertex	$\mathcal{O}(1/t^2)$	open-loop
Theorem 5.14	polytope	polytope	disjoint	vertex	vertex	finite	open-loop
Theorem 5.23	polytope	polytope	disjoint	interior of face	vertex	$\mathcal{O}(1/t^2)$	open-loop



(a) Theorem 5.3 (b) Theorem 5.5 (c) Theorem 5.10

(d) Theorem 5.11 (e) Theorem 5.14 (f) Theorem 5.23

Figure 1: Case studies of 2-dimensional feasibility problem under different feasible regions. The blue and red regions represent $\mathcal{P}$ and $\mathcal{Q}$, respectively, and the gray cross-hatched area denotes the intersection of $\mathcal{P}$ and $\mathcal{Q}$. The blue and red dots refer to a pair of points of minimum distance in $\mathcal{P}$ and $\mathcal{Q}$, respectively, and the purple dot denotes a single point in the intersection of two sets. Corresponding theorems are captioned in each case.

1.2 Outline

In Section 2, we provide preliminaries for our presentation. Section 3 revisits von Neumann's alternating projection algorithm over convex sets (`POCS`) and presents its convergence rate. In Section 4, we introduce alternating linear minimizations (`ALM`) analogous to the algorithm in Section 3 and discuss the attributes and properties of the objective function. Subsequently, we derive baseline convergence rates for `ALM` with open-loop step-sizes and line-search, drawing a comparison with convergence rates observed in Section 3. In Section 5, we prove the accelerated convergence schemes of `ALM` with open-loop step-sizes by categorizing the feasible regions, and present toy examples to demonstrate results.

2 Preliminaries

Throughout this work, let $\mathcal{C} \subseteq \mathbb{R}^d$ be a compact convex set, let I_n be the identity matrix of size $n \in \mathbb{N}$, let $\mathbb{0} \in \mathbb{R}^d$ denote the all-zeros vector, let $\mathbb{1} \in \mathbb{R}^d$ denote the all-ones vector and let $e^{(i)} \in \mathbb{R}^d$ be the ith unit vector such that $e_i^{(i)} = 1$ and $e_j^{(i)} = 0$ for all $j \in \{1, \dots, d\} \setminus \{i\}$. Given a matrix $A \in \mathbb{R}^{n \times n}$, let $\det(A)$ denote the determinant of A. Let $\Pi_{\mathcal{C}}$ denote the orthogonal projector onto a set $\mathcal{C} \subseteq \mathbb{R}^d$, where for given $\mathbf{z} \in \mathbb{R}^d$ we define $\Pi_{\mathcal{C}}(\mathbf{z}) = \arg\min_{\mathbf{x} \in \mathcal{C}} \|\mathbf{x} - \mathbf{z}\|_2$. Given a set $\mathcal{C} \subseteq \mathbb{R}^d$, let $\operatorname{aff}(\mathcal{C}), \operatorname{span}(\mathcal{C}), \operatorname{conv}(\mathcal{C})$ and $\operatorname{vert}(\mathcal{C})$ denote the affine hull, the span, the convex hull and the set of vertices of $\mathcal{C}$, respectively.

We introduce necessary notions and definitions.

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

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

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

Definition 2.2 (Smooth function) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{P} \times \mathcal{Q}$, and let $L > 0$. We say that f is L-smooth function over $\mathcal{P} \times \mathcal{Q}$ with respect to $\|\cdot\|_2$ if for all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{P} \times \mathcal{Q}$, it holds that*

$$f(\mathbf{x}_1, \mathbf{y}_1) \leq f(\mathbf{x}_2, \mathbf{y}_2) + \left\langle \nabla f(\mathbf{x}_2, \mathbf{y}_2), \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{bmatrix} \right\rangle + \frac{L}{2} \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{bmatrix} \right\|_2^2.$$

Definition 2.3 (Strongly convex function) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{P} \times \mathcal{Q}$, and let $\alpha_f \geq 0$. We say that f is α_f-strongly convex over $\mathcal{P} \times \mathcal{Q}$ with respect to $\|\cdot\|_2$ if for $\alpha_f > 0$ and for all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{P} \times \mathcal{Q}$, it holds that*

$$f(\mathbf{x}_1, \mathbf{y}_1) \geq f(\mathbf{x}_2, \mathbf{y}_2) + \left\langle \nabla f(\mathbf{x}_2, \mathbf{y}_2), \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{bmatrix} \right\rangle + \frac{\alpha_f}{2} \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{bmatrix} \right\|_2^2. \tag{1}$$

We refer that f is convex over $\mathcal{P} \times \mathcal{Q}$ if (1) holds for $\alpha_f = 0$.

The following lemma characterizes a strongly convex and a convex function.

Lemma 2.4 *(Boyd and Vandenberghe, 2004, Section 9.1.2) Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be twice differentiable in an open set containing $\mathcal{P} \times \mathcal{Q}$, let $I_{2d} \in \mathbb{R}^{2d \times 2d}$ be the identity matrix and let $\nabla^2 f(\mathbf{x}, \mathbf{y})$ be the Hessian matrix of f. Then, f is strongly convex if and only if there exists $m > 0$ such that $\nabla^2 f(\mathbf{x}, \mathbf{y}) - m I_{2d}$ is positive semi-definite. When $m = 0$, f is convex but not strongly convex.*

Definition 2.5 (Polyak Łojasiewicz inequality) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{P} \times \mathcal{Q}$, and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x}, \mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set with optimal function value $f^* = f(\mathbf{x}^*, \mathbf{y}^*)$. We say that f satisfies the Polyak Łojasiewicz inequality if for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{P} \times \mathcal{Q}$, there exists $\mu > 0$ such that*

$$\frac{1}{2} \|\nabla f(\mathbf{x}, \mathbf{y})\|_2^2 \geq \mu \left(f(\mathbf{x}, \mathbf{y}) - f^* \right). \tag{PL}$$

Definition 2.6 (Quadratic growth) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be differentiable in an open set containing $\mathcal{P} \times \mathcal{Q}$, and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x}, \mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set with optimal function value $f^* = f(\mathbf{x}^*, \mathbf{y}^*)$. We say that f satisfies the ξ-quadratic growth if for all $(\mathbf{x}, \mathbf{y}) \in \mathcal{P} \times \mathcal{Q}$, there exists $\xi > 0$ such that*

$$\frac{2}{\xi}(f(\mathbf{x}, \mathbf{y}) - f^*) \geq \min_{(\mathbf{z}, \mathbf{w}) \in \Omega^*} \left\| \begin{bmatrix} \mathbf{x} - \mathbf{z} \\ \mathbf{y} - \mathbf{w} \end{bmatrix} \right\|_2^2. \tag{QG}$$

3 von Neumann's alternating projections

We begin by revisiting von Neumann's alternating projections algorithm (Neumann, 1949) and present a comprehensive convergence rate. Particularly in convex settings, this algorithm is commonly referred to as projections onto convex sets (POCS).

Consider two convex sets $\mathcal{P}, \mathcal{Q} \subseteq \mathbb{R}^d$ with associated projectors $\Pi_{\mathcal{P}}$ and $\Pi_{\mathcal{Q}}$ where $\Pi_{\mathcal{P}}(\mathbf{z})$ and $\Pi_{\mathcal{Q}}(\mathbf{z})$ are the projections of $\mathbf{z} \in \mathbb{R}^d$ onto $\mathcal{P}$ and $\mathcal{Q}$, respectively. At each iteration of POCS, the iterates are projected onto $\mathcal{P}$ and $\mathcal{Q}$, alternatively.

Algorithm 2: Projections onto convex sets (POCS)

Input : $\mathbf{y}_0 \in \mathbb{R}^d$, $\Pi_{\mathcal{P}}$ projector onto $\mathcal{P} \subseteq \mathbb{R}^d$ and $\Pi_{\mathcal{Q}}$ projector onto $\mathcal{Q} \subseteq \mathbb{R}^d$.

for $t = 0, \dots, T-1$ **do**
  $\mathbf{x}_{t+1} \leftarrow \Pi_{\mathcal{P}}(\mathbf{y}_t)$
  $\mathbf{y}_{t+1} \leftarrow \Pi_{\mathcal{Q}}(\mathbf{x}_{t+1})$
end

Proposition 3.1 *(Braun et al., 2023, Proposition 2.1) Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets and let $\mathcal{Q}_{min}$ be the set of points of $\mathcal{Q}$ with minimal distance to $\mathcal{P}$, and $\mathbf{d}$ the distance vector between $\mathcal{P}$ and $\mathcal{Q}$, i.e., $\mathbf{d} = \mathbf{z} - \Pi_{\mathcal{P}}(\mathbf{z})$ for all $\mathbf{z} \in \mathcal{Q}_{min}$. Then, for the iterates of Algorithm 2, it holds that*

$$\min_{t=0,\dots,T-1} \|\mathbf{y}_t - \mathbf{x}_{t+1} - \mathbf{d}\|_2^2 + \|\mathbf{x}_{t+1} - \mathbf{y}_{t+1} + \mathbf{d}\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \left(\|\mathbf{y}_t - \mathbf{x}_{t+1} - \mathbf{d}\|_2^2 + \|\mathbf{x}_{t+1} - \mathbf{y}_{t+1} + \mathbf{d}\|_2^2 \right) \leq \frac{\operatorname{dist}(\mathbf{y}_0, Q_{min})^2}{T}.$$

In particular, if $\mathcal{P}$ and $\mathcal{Q}$ intersect, then

$$\|\mathbf{x}_T - \mathbf{y}_T\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \left(\|\mathbf{y}_t - \mathbf{x}_{t+1}\|_2^2 + \|\mathbf{x}_{t+1} - \mathbf{y}_{t+1}\|_2^2 \right) \leq \frac{\operatorname{dist}(\mathbf{y}_0, P \cap Q)^2}{T}.$$

The convergence proofs for POCS, as discussed in the above proposition, demonstrate that the sequences of POCS converge in norm, meaning that $\|\mathbf{x}_t - \mathbf{y}_t\|_2$ converges to $\|\mathbf{d}\|_2 = \operatorname{dist}(\mathcal{P}, \mathcal{Q})$. Additionally, the individual iterates $\mathbf{x}_t$ and $\mathbf{y}_t$ converge to their accumulation point $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$ with $\mathbf{x} - \mathbf{y} = \mathbf{d}$. Refer to (Braun et al., 2023, Proposition 2.1) for a detailed proof.

Note that POCS requires prior knowledge of the projector onto the corresponding set, and the geometric complexities of the sets make the projection step computationally expensive (Combettes and Pokutta, 2021). However, FW offers a projection-free alternative, relying solely on access to the LMO for the feasible regions. Considering a compact convex set $\mathcal{C} \subseteq \mathbb{R}^d$ and a convex function $g : \mathcal{C} \to \mathbb{R}$ via $g(\mathbf{x}) := \frac{1}{2}\|\mathbf{x} - \mathbf{a}\|_2^2$ for $\mathbf{a} \in \mathbb{R}^d$, for example, FW generates iterates that progressively approach $\mathbf{x}^* \in \mathcal{C}$ such that $\|\mathbf{x}^* - \mathbf{a}\|_2 = \operatorname{dist}(\mathbf{a}, \mathcal{C})$. Thus, Braun et al. (2023) introduced an algorithm analogous to POCS, which calls the LMOs by replacing projections in each iteration with a quadratic objective function.

4 Alternating linear minimizations

This section is structured as follows. First, we introduce alternating linear minimizations (ALM) and a quadratic function along with its mathematical properties, which will be employed to substantiate our acceleration results in Section 5. Then, in Section 4.1, we propose the dual gap and establish a connection between the current and the true dual gap. Finally, in Section 4.2, we derive canonical convergence rates of order $\mathcal{O}(1/t)$ for ALM with open-loop step sizes and line-search, further leading to comparing their convergence rates with POCSs.

Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets. As the goal is to find a pair of points in them with minimal distance, we define $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ via $f(\mathbf{x}, \mathbf{y}) := \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$. Considering the optimization problem

$$\arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y}), \tag{OPT}$$

LMO-based approach computes $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$ such that $\|\mathbf{x} - \mathbf{y}\|_2 = \operatorname{dist}(\mathcal{P}, \mathcal{Q})$ via ALM.

Algorithm 3: Alternating linear minimizations (ALM)

Input : $\mathbf{x}_0 \in \mathcal{P}, \mathbf{y}_0 \in \mathcal{Q}$, step-sizes $\eta_{t,\mathcal{P}}, \eta_{t,\mathcal{Q}} \in [0, 1]$ for $t \in \{0, \dots, T-1\}$.

1 **for** $t = 0, \dots, T-1$ **do**
2 $\quad \mathbf{u}_t \leftarrow \arg\min_{\mathbf{u} \in \mathcal{P}} \langle \mathbf{x}_t - \mathbf{y}_t, \mathbf{u} \rangle$
3 $\quad \mathbf{x}_{t+1} \leftarrow \mathbf{x}_t + \eta_{t,\mathcal{P}}(\mathbf{u}_t - \mathbf{x}_t)$
4 $\quad \mathbf{v}_t \leftarrow \arg\min_{\mathbf{v} \in \mathcal{Q}} \langle \mathbf{y}_t - \mathbf{x}_{t+1}, \mathbf{v} \rangle$
5 $\quad \mathbf{y}_{t+1} \leftarrow \mathbf{y}_t + \eta_{t,\mathcal{Q}}(\mathbf{v}_t - \mathbf{y}_t)$
6 **end**

For $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$, let

$$\nabla f(\mathbf{x}, \mathbf{y}) = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{x}} \\ \frac{\partial f}{\partial \mathbf{y}} \end{bmatrix} = \begin{bmatrix} \mathbf{x} - \mathbf{y} \\ \mathbf{y} - \mathbf{x} \end{bmatrix} =: \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}) \\ \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}) \end{bmatrix}.$$

Then, we can write Line 2 and 4 in Algorithm 3 using gradients: $\mathbf{u}_t \leftarrow \arg\min_{\mathbf{u} \in \mathcal{P}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{u} \rangle$ and $\mathbf{v}_t \leftarrow \arg\min_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{v} \rangle$, respectively. By definition of $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ and $\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$, we obtain $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})\|_2 = \|\mathbf{x} - \mathbf{y}\|_2 = \|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|_2$ and $\|\nabla f(\mathbf{x}, \mathbf{y})\|_2^2 = 2\|\mathbf{x} - \mathbf{y}\|_2^2$. Thus, the two lemmas below show that f satisfies the (PL) inequality and L-smoothness.

Lemma 4.1 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$, i.e., $\|\mathbf{x}^* - \mathbf{y}^*\|_2 = \operatorname{dist}(\mathcal{P}, \mathcal{Q})$. Then, f satisfies $\frac{1}{2}\|\nabla f(\mathbf{x}, \mathbf{y})\|_2^2 \geq f(\mathbf{x}, \mathbf{y}) - f(\mathbf{x}^*, \mathbf{y}^*)$ which implies the (PL) inequality.*

Proof Since $\|\nabla f(\mathbf{x}, \mathbf{y})\|_2^2 = 2\|\mathbf{x} - \mathbf{y}\|_2^2$, it holds that $\|\nabla f(\mathbf{x}, \mathbf{y})\|_2^2 \geq \|\mathbf{x} - \mathbf{y}\|_2^2 - \operatorname{dist}(\mathcal{P}, \mathcal{Q})^2$. ■

Lemma 4.2 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. Then, f is 2-smooth over $\mathcal{P} \times \mathcal{Q}$.*

Proof We will show that f satisfies $\|\nabla f(\mathbf{x}_1, \mathbf{y}_1) - \nabla f(\mathbf{x}_2, \mathbf{y}_2)\|_2 \leq 2 \left\| \begin{bmatrix} \mathbf{x}_1 - \mathbf{x}_2 \\ \mathbf{y}_1 - \mathbf{y}_2 \end{bmatrix} \right\|_2$ for all $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{P} \times \mathcal{Q}$, which is equivalent to 2-smoothness of f. For $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathcal{P} \times \mathcal{Q}$, we have

$$\nabla f(\mathbf{x}_1, \mathbf{y}_1) - \nabla f(\mathbf{x}_2, \mathbf{y}_2) = \begin{bmatrix} \mathbf{x}_1 - \mathbf{y}_1 - (\mathbf{x}_2 - \mathbf{y}_2) \\ \mathbf{y}_1 - \mathbf{x}_1 - (\mathbf{y}_2 - \mathbf{x}_2) \end{bmatrix},$$

which gives

$$\begin{aligned}\|\nabla f(\mathbf{x}_1,\mathbf{y}_1)-\nabla f(\mathbf{x}_2,\mathbf{y}_2)\|_2^2 &= 2\|\mathbf{x}_1-\mathbf{x}_2-(\mathbf{y}_1-\mathbf{y}_2)\|_2^2\\ &= 2\|\mathbf{x}_1-\mathbf{x}_2\|_2^2-4\langle\mathbf{x}_1-\mathbf{x}_2,\mathbf{y}_1-\mathbf{y}_2\rangle+2\|\mathbf{y}_1-\mathbf{y}_2\|_2^2\\ &\le 2\|\mathbf{x}_1-\mathbf{x}_2\|_2^2+4\|\mathbf{x}_1-\mathbf{x}_2\|_2\|\mathbf{y}_1-\mathbf{y}_2\|_2+2\|\mathbf{y}_1-\mathbf{y}_2\|_2^2\\ &= 2\left(\|\mathbf{x}_1-\mathbf{x}_2\|_2+\|\mathbf{y}_1-\mathbf{y}_2\|_2\right)^2\\ &\le 4\left(\|\mathbf{x}_1-\mathbf{x}_2\|_2^2+\|\mathbf{y}_1-\mathbf{y}_2\|_2^2\right)\\ &= 4\left\|\begin{bmatrix}\mathbf{x}_1-\mathbf{x}_2\\ \mathbf{y}_1-\mathbf{y}_2\end{bmatrix}\right\|_2^2\end{aligned}$$

where the first inequality holds due to Cauchy-Schwarz inequality and the last inequality follows from $(a+b)^2\le 2(a^2+b^2)$ for $a,b\in\mathbb{R}$. ■

Many accelerated convergence rates are established under conditions where the objective function is strongly convex (Guélat and Marcotte, 1986; Garber and Hazan, 2015a; Bach, 2021; Wirth et al., 2023a). However, the objective function $f(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ is not strongly convex, thereby preventing it from enjoying the nice properties of strongly convex functions.

Lemma 4.3 *Let $\mathcal{P}\subseteq\mathbb{R}^d$ and $\mathcal{Q}\subseteq\mathbb{R}^d$ be compact convex sets and let $f(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$. Then, f is not strongly convex, but convex.*

Proof Let I_d and I_{2d} be the identity matrices of size d and $2d$, respectively and let $m\in\mathbb{R}$. The Hessian matrix of f is

$$\nabla^2 f(\mathbf{x},\mathbf{y})=\begin{bmatrix}I_d & -I_d\\ -I_d & I_d\end{bmatrix}\in\mathbb{R}^{2d\times 2d}.$$

Let $\lambda>0$ be the eigenvalue of $\nabla^2 f(\mathbf{x},\mathbf{y})-mI_{2d}$. Since $\nabla^2 f(\mathbf{x},\mathbf{y})-mI_{2d}$ is symmetric, non-negative eigenvalues of the matrix imply positive semi-definite of $\nabla^2 f(\mathbf{x},\mathbf{y})-mI_{2d}$. Thus, we obtain its eigenvalue by solving the characteristic polynomial, $\det\left(\nabla^2 f(\mathbf{x},\mathbf{y})-mI_{2d}-\lambda I_{2d}\right)=0$, that is,

$$\det\begin{bmatrix}(1-m-\lambda)I_d & -I_d\\ -I_d & (1-m-\lambda)I_d\end{bmatrix}=\det\left((2-m-\lambda)I_d\right)\det\left((-m-\lambda)I_d\right)=(2-m-\lambda)(-m-\lambda)=0$$

where the first equality holds due to $\det\begin{bmatrix}A & B\\ B & A\end{bmatrix}=\det(A-B)\cdot\det(A+B)$ for all $A,B\in\mathbb{R}^{d\times d}$ (Silvester, 2000). Thus, $\nabla^2 f(\mathbf{x},\mathbf{y})-mI_{2d}$ is positive semi-definite only if $m=0$. Hence, by Lemma 2.4, it concludes the claim. ■

On the other hand, the objective function $f(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ satisfies the quadratic growth, a weaker condition of strong convexity. The below lemma shows that when f satisfies the (PL), it also satisfies the (QG).

Lemma 4.4 *(Garrigos, 2023, Corollary 12) Let $\mathcal{P}\subseteq\mathbb{R}^d$ and $\mathcal{Q}\subseteq\mathbb{R}^d$ be compact convex sets, let $f:\mathcal{P}\times\mathcal{Q}\to\mathbb{R}$ be a convex and L-smooth function, and let $\Omega^*=\{(\mathbf{x}^*,\mathbf{y}^*)\in\mathcal{P}\times\mathcal{Q}\mid(\mathbf{x}^*,\mathbf{y}^*)\in\arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}}f(\mathbf{x},\mathbf{y})\}$ be the solution set. Suppose that f satisfies the (PL) with constant $\mu>0$. Then, f satisfies the μ-(QG) for all $\mathbf{x}\in\mathcal{P}$ and $\mathbf{y}\in\mathcal{Q}$.*

As the objective function $f(\mathbf{x},\mathbf{y})=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ satisfies the (PL) with $\mu=1$, it satisfies the 1-(QG), i.e.,

$$f(\mathbf{x},\mathbf{y})-f(\mathbf{x}^*,\mathbf{y}^*)\ge\frac{1}{2}\min_{(\mathbf{z},\mathbf{w})\in\Omega^*}\left\|\begin{bmatrix}\mathbf{x}-\mathbf{z}\\ \mathbf{y}-\mathbf{w}\end{bmatrix}\right\|_2^2. \tag{2}$$

Indeed, (2) implies that f satisfies a $\left(\frac{1}{2},\sqrt{2}\right)$-Hölderian error bound (Kerdreux et al., 2021b, Definition 3.5). For special cases when the optimal solution $(\mathbf{x}^*,\mathbf{y}^*)\in\Omega^*$ is unique, (2) becomes

$$f(\mathbf{x},\mathbf{y})-f(\mathbf{x}^*,\mathbf{y}^*)\ge\frac{1}{2}\left\|\begin{bmatrix}\mathbf{x}-\mathbf{x}^*\\ \mathbf{y}-\mathbf{y}^*\end{bmatrix}\right\|_2^2. \tag{3}$$

4.1 The dual gap

For the iterates of Algorithm 3, we denote the primal gap at iteration $t \in \mathbb{N}$ by $h_t := f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}^*)$, where $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. The dual gap of a function f at iteration $t \in \mathbb{N}$ is defined as

$$\begin{aligned} g(\mathbf{x}_t, \mathbf{y}_t) &:= \max_{\mathbf{u} \in \mathcal{P}, \mathbf{v} \in \mathcal{Q}} \left\langle \nabla f(\mathbf{x}_t, \mathbf{y}_t), \begin{bmatrix} \mathbf{x}_t - \mathbf{u} \\ \mathbf{y}_t - \mathbf{v} \end{bmatrix} \right\rangle \\ &= \max_{\mathbf{u} \in \mathcal{P}, \mathbf{v} \in \mathcal{Q}} \left\langle \begin{bmatrix} \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) \\ \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) \end{bmatrix}, \begin{bmatrix} \mathbf{x}_t - \mathbf{u} \\ \mathbf{y}_t - \mathbf{v} \end{bmatrix} \right\rangle \\ &= \max_{\mathbf{u} \in \mathcal{P}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u} \rangle + \max_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v} \rangle \\ &= g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t). \end{aligned}$$

where we define $g_{\mathbf{x}}(\mathbf{x}, \mathbf{y}) := \max_{\mathbf{u} \in \mathcal{P}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y}), \mathbf{x} - \mathbf{u} \rangle$ and $g_{\mathbf{y}}(\mathbf{x}, \mathbf{y}) := \max_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y}), \mathbf{y} - \mathbf{v} \rangle$. The lemma below guarantees that the dual gap is bounded by the primal gap and non-negativity.

Lemma 4.5 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets with diameters $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. Then, for the iterates of Algorithm 3 with any step-sizes, the following inequality holds*

$$g(\mathbf{x}_t, \mathbf{y}_t) \geq h(\mathbf{x}_t, \mathbf{y}_t) \geq 0.$$

Proof It holds that

$$g(\mathbf{x}_t, \mathbf{y}_t) = \max_{\mathbf{u} \in \mathcal{P}, \mathbf{v} \in \mathcal{Q}} \left\langle \nabla f(\mathbf{x}_t, \mathbf{y}_t), \begin{bmatrix} \mathbf{x}_t - \mathbf{u} \\ \mathbf{y}_t - \mathbf{v} \end{bmatrix} \right\rangle \geq \left\langle \nabla f(\mathbf{x}_t, \mathbf{y}_t), \begin{bmatrix} \mathbf{x}_t - \mathbf{x}^* \\ \mathbf{y}_t - \mathbf{y}^* \end{bmatrix} \right\rangle \geq f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}^*) = h(\mathbf{x}_t, \mathbf{y}_t).$$

■

Moreover, the dual gap with respect to $\mathbf{x} \in \mathcal{P}$ holds

$$g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) = \max_{\mathbf{u} \in \mathcal{P}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u} \rangle \geq \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \Pi_{\mathcal{P}}(\mathbf{y}_t) \rangle \geq f(\mathbf{x}_t, \mathbf{y}_t) - f(\Pi_{\mathcal{P}}(\mathbf{y}_t), \mathbf{y}_t) \geq 0$$

where the second inequality holds due to convexity of f, and the last inequality follows from $\Pi_{\mathcal{P}}(\mathbf{y}_t) = \arg\min_{\mathbf{x} \in \mathcal{P}} \|\mathbf{x} - \mathbf{y}_t\|_2$. Analogously, we have $g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) \geq 0$.

Because `ALM` is looking at two feasible sets alternatively at one iteration, what we actually obtain from `ALM` at iteration t is $g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) = \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle$ and $g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) = \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle$. Therefore, it does not know the current true dual gap $g(\mathbf{x}_t, \mathbf{y}_t) = g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_t, \mathbf{y}_t)$, otherwise additional `LMO` call is required to monitor this quantity. Thus, we choose a fixed error parameter $\nu \in]0, 1]$ such that the current $\mathbf{v}_t \in \arg\min_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t, \mathbf{v} \rangle$ attains the true dual gap up to an approximation error of ν with respect to $\mathbf{y} \in \mathcal{Q}$, that is,

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle \geq \nu \max_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v} \rangle.$$

When an approximate parameter is used together with the open-loop step-sizes, then the step-sizes $\eta_{t,\mathcal{P}}, \eta_{t,\mathcal{Q}}$ of Algorithm 3 need to be increased to $\frac{\ell}{\nu t + \ell}$ instead of $\frac{\ell}{t + \ell}$ (Lacoste-Julien et al., 2013, Appendix C). Thus, the current dual gap up to an approximation error of ν at iteration t is obtained

$$\begin{aligned} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) &= \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle + \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle \\ &\geq \nu \left(\max_{\mathbf{u} \in \mathcal{P}} \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u} \rangle + \max_{\mathbf{v} \in \mathcal{Q}} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v} \rangle \right) \\ &= \nu g_t. \end{aligned} \tag{4}$$

4.2 Convergence rate of order $\mathcal{O}(1/t)$

We begin with the analysis of ALM with open-loop step-sizes and line-search by deriving a base convergence rate of the algorithm. Since f is quadratic, the iterates of Algorithm 3 with any step-sizes $\eta_{t,\mathcal{P}}, \eta_{t,\mathcal{Q}} \in [0,1]$ satisfy

$$\begin{aligned} f(\mathbf{x}_{t+1}, \mathbf{y}_t) &= \frac{1}{2}\left\|\mathbf{x}_t + \eta_{t,\mathcal{P}}(\mathbf{u}_t - \mathbf{x}_t) - \mathbf{y}_t\right\|_2^2 \\ &= \frac{1}{2}\left\|\mathbf{x}_t - \mathbf{y}_t\right\|_2^2 - \eta_{t,\mathcal{P}}\langle \mathbf{x}_t - \mathbf{y}_t, \mathbf{x}_t - \mathbf{u}_t\rangle + \frac{\eta_{t,\mathcal{P}}^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \\ &= f(\mathbf{x}_t, \mathbf{y}_t) - \eta_{t,\mathcal{P}}\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t\rangle + \frac{\eta_{t,\mathcal{P}}^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \\ &= f(\mathbf{x}_t, \mathbf{y}_t) - \eta_{t,\mathcal{P}} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + \frac{\eta_{t,\mathcal{P}}^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2. \end{aligned} \tag{5}$$

Analogously, we obtain

$$f(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}) = f(\mathbf{x}_{t+1}, \mathbf{y}_t) - \eta_{t,\mathcal{Q}} g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) + \frac{\eta_{t,\mathcal{Q}}^2}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2. \tag{6}$$

Plugging (5) into (6) yields

$$h_{t+1} = h_t - \eta_{t,\mathcal{P}} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) - \eta_{t,\mathcal{Q}} g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) + \frac{\eta_{t,\mathcal{P}}^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \frac{\eta_{t,\mathcal{Q}}^2}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2. \tag{7}$$

Now, we derive a progress bound for ALM with open-loop step-sizes.

Lemma 4.6 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$, let $\nu \in]0,1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 1}$ and let $\epsilon \in [0, \ell]$. Then, for iterates of Algorithm 3 with step-size η_t, we have*

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t \frac{\epsilon}{\ell}\left(g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)\right) + \frac{\eta_t^2}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right). \tag{8}$$

Proof Since $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t$, by (7), we obtain

$$\begin{aligned} h_{t+1} &= h_t - \eta_t\left(g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)\right) + \frac{\eta_t^2}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) &(9)\\ &\leq h_t - \nu\eta_t g_t + \frac{\eta_t^2}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) \\ &\leq (1 - \nu\eta_t) h_t + \frac{\eta_t^2}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) &(10) \end{aligned}$$

where the first inequality follows from (4) and the last inequality holds due to Lemma 4.5. Combining (9) with (10) for some $\epsilon \in [0, \ell]$ concludes the claim. ■

The following technical lemma, a variant of Lemma 3.3 derived by Wirth et al. (2023b), is used in the proof of Proposition 4.8.

Lemma 4.7 *Let $\nu \in]0,1]$, let $S \in \mathbb{N}$ and let $\eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 1}$. Then, for $t \geq S$, it holds that*

$$\prod_{i=S}^{t}(1 - \nu\eta_i) \leq \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell}.$$

Proof The proof is presented in Appendix A.1. ■

Having Lemma 4.6 and 4.7 at hand, we prove a baseline convergence rate for ALM with open-loop step-sizes.

Proposition 4.8 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets with diameters $D_\mathcal{P}, D_\mathcal{Q} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$, let $\nu \in]0, 1]$ be an approximation parameter, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$. Then, for iterates of Algorithm 3 with step-size η_t, we have*

$$h_t \leq \frac{(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell^2}{2\nu\,(\nu(t-1)+\ell)} = \eta_{t-1}\frac{(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell}{2\nu}.$$

Proof The proof is usually done in the literature with $\ell = 2$ and $\nu = 1$, e.g., Beck et al. (2015) and Braun et al. (2023). Here, we extend the proofs to general open-loop step-sizes, $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, $\epsilon \in [0, \ell]$ and $\nu \in]0, 1]$, adapting the straight-forward proof of Jaggi (2013). By (9), we have $h_1 \leq h_0 + \frac{1}{2}(D_\mathcal{P}^2 + D_\mathcal{Q}^2)$ and from (10) with $\|\mathbf{x}_t - \mathbf{u}_t\|_2 \leq D_\mathcal{P}$, $\|\mathbf{y}_t - \mathbf{v}_t\|_2 \leq D_\mathcal{Q}$, it holds that

$$\begin{aligned}
h_{t+1} &\leq (1-\nu\eta_t)\,h_t + \frac{\eta_t^2}{2}(D_\mathcal{P}^2 + D_\mathcal{Q}^2)\\
&\leq \prod_{i=1}^{t}(1-\nu\eta_i)\,h_1 + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}\sum_{i=1}^{t}\eta_i^2\prod_{j=i+1}^{t}(1-\nu\eta_j)\\
&\leq \left(h_0 + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}\right)\prod_{i=1}^{t}(1-\nu\eta_i) + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}\sum_{i=1}^{t}\eta_i^2\prod_{j=i+1}^{t}(1-\nu\eta_j)\\
&\leq \left(h_0 + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}\right)\left(\prod_{i=1}^{t}(1-\nu\eta_i) + \sum_{i=1}^{t}\eta_i^2\prod_{j=i+1}^{t}(1-\nu\eta_j)\right). \qquad (11)
\end{aligned}$$

By Lemma 4.7, it holds that

$$\prod_{i=1}^{t}(1-\nu\eta_i) + \sum_{i=1}^{t}\eta_i^2\prod_{j=i+1}^{t}(1-\nu\eta_j) \leq \eta_t^\ell + \sum_{i=1}^{t}\eta_i^2\left(\frac{\eta_t}{\eta_i}\right)^\ell \leq \eta_t^\ell + \sum_{i=1}^{t}\eta_t^2 \leq (t+1)\eta_t^2 \leq \frac{\ell^2}{\nu(\nu t+\ell)}.$$

where the last equality holds due to $\ell \in \mathbb{N}_{\geq 2}$ and $\nu \in]0, 1]$. Plugging this bound into (11) concludes the claim. ■

Minimizing the right-hand sides of (5) and (6) over $\eta_{t,\mathcal{P}}, \eta_{t,\mathcal{Q}} \in [0, 1]$, respectively, leads to the optimal choices of step-sizes,

$$\eta_{t,\mathcal{P}} = \max\left\{\frac{g_\mathbf{x}(\mathbf{x}_t, \mathbf{y}_t)}{\|\mathbf{x}_t - \mathbf{u}_t\|_2^2}, 1\right\}, \qquad \eta_{t,\mathcal{Q}} = \max\left\{\frac{g_\mathbf{y}(\mathbf{x}_{t+1}, \mathbf{y}_t)}{\|\mathbf{y}_t - \mathbf{v}_t\|_2^2}, 1\right\},$$

which states ALM with short-step, guaranteeing monotone decreasing function values. Since f is quadratic for fixed $\mathbf{x} \in \mathcal{P}$ or $\mathbf{y} \in \mathcal{Q}$, performing line-search and short-step are identical. Now, we prove a baseline convergence of ALM with line-search.

Proposition 4.9 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact convex sets with diameters $D_\mathcal{P}, D_\mathcal{Q} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$, let $\nu \in]0, 1]$ be an approximation parameter and let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error. Then, for iterates of Algorithm 3 with line-search, we have*

$$h_t \leq \frac{4(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\max\{D_\mathcal{P}, D_\mathcal{Q}\}^2}{\nu^2(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)(t-1) + 8\max\{D_\mathcal{P}, D_\mathcal{Q}\}^2}.$$

Proof From (7), we obtain

$$\begin{aligned} h_{t+1} &= h_t - \eta_{t,\mathcal{P}} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) - \eta_{t,\mathcal{Q}} g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) + \frac{\eta_{t,\mathcal{P}}^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \frac{\eta_{t,\mathcal{Q}}^2}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2 \\ &= h_t - \frac{g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t)^2}{2\|\mathbf{x}_t - \mathbf{u}_t\|_2^2} - \frac{g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)^2}{2\|\mathbf{y}_t - \mathbf{v}_t\|_2^2} \\ &\leq h_t - \frac{g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t)^2}{2D_{\mathcal{P}}^2} - \frac{g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)^2}{2D_{\mathcal{Q}}^2}. \end{aligned} \tag{12}$$

where the last inequality follows from $\|\mathbf{x}_t - \mathbf{u}_t\|_2 \leq D_{\mathcal{P}}$ and $\|\mathbf{y}_t - \mathbf{v}_t\|_2 \leq D_{\mathcal{Q}}$. Rearranging (12) yields

$$h_t - h_{t+1} \geq \frac{g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t)^2 + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)^2}{2\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2} \geq \frac{\left(g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t)\right)^2}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2} \geq \frac{\nu^2 g_t^2}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2} \geq \frac{\nu^2 h_t^2}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2} \tag{13}$$

where the second inequality holds due to $(a+b)^2 \leq 2(a^2+b^2)$ for $a, b \in \mathbb{R}$, the third inequality follows from (4) and the last inequality holds due to Lemma 4.5. We then divide both sides of (13) by $h_t h_{t+1}$, yielding

$$\frac{1}{h_{t+1}} - \frac{1}{h_t} \geq \frac{\nu^2}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2}\frac{h_t}{h_{t+1}} \geq \frac{\nu^2}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2}$$

where the last inequality holds due to the fact that h_t is a non-increasing sequence by (12). Summing up both sides yields

$$\frac{1}{h_{t+1}} \geq \frac{\nu^2 t}{4\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2} + \frac{1}{h_1}.$$

Combined with $h_1 \leq h_0 + \frac{1}{2}(D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)$, it concludes the claim. ■

To end Section 4.2, it is necessary to compare the convergence rates derived in Proposition 3.1 for `POCS` with Proposition 4.8 and 4.9 for `ALM` with open-loop step-sizes and line-search, respectively.

Remark 4.10 (Comparison to `POCS`) Let us consider the case $\mathcal{P} \cap Q \neq \emptyset$. Let $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$ be diameter of $\mathcal{P}$ and $\mathcal{Q}$, respectively. Let $\nu \in]0, 1]$ be an approximation error and let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error at the starting point. Proposition 3.1 guarantees that `POCS` converge after T many iterations at a rate of

$$\|\mathbf{x}_T - \mathbf{y}_T\|_2^2 \leq \frac{\operatorname{dist}(\mathbf{y}_0, \mathcal{P} \cap \mathcal{Q})^2}{T}.$$

In contrast to this, the `LMO`-based approach guarantees via Proposition 4.8 for `ALM` with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$ after T many iterations

$$\|\mathbf{x}_T - \mathbf{y}_T\|_2^2 \leq \frac{(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2}{2\nu^2\,(T - 1 + \ell/\nu)}$$

and via Proposition 4.9 for `ALM` with line-search after T many iterations

$$\|\mathbf{x}_T - \mathbf{y}_T\|_2^2 \leq \frac{8(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2}{\nu^2(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)(T-1) + 8\max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2}.$$

All approaches ensure a convergence rate of $\mathcal{O}(1/T)$, while the `LMO`-based process requires no projection, resulting in low iteration complexity.

5 Accelerated convergence rates for ALM with open-loop step-sizes

In this section, we characterize feasible regions for acceleration schemes of ALM with open-loop step-sizes. We begin by demonstrating accelerated convergence rates under the setting where both feasible regions are strongly convex in Section 5.1. Then, we consider the scenario where the feasible regions are a strongly convex set and a polytope leading to further improved convergence rates in Section 5.2. Finally, we investigate acceleration in Section 5.3 for the case where both feasible sets are polytopes. At the end of each section, we present several toy examples to demonstrate results and raise open questions.

5.1 $\mathcal{P}$ and $\mathcal{Q}$ are strongly convex

A faster convergence of the LMO based algorithms can be achieved by the geometry of the feasible region and assumptions on the objective function. Specifically, when the feasible set is strongly convex, and the norm of the gradient is bounded away from zero, FW with line-search enjoys linear convergence (Levitin and Polyak, 1966; Demyanov and Rubinov, 1970). When both the objective function and the feasible region are strongly convex, the convergence rate of FW with line-search or short-step can be improved from the rate of $\mathcal{O}(1/t)$ to $\mathcal{O}(1/t^2)$ (Garber and Hazan, 2015a). Kerdreux et al. (2021a) generalized the acceleration to the family of uniformly convex sets, further assuming that the objective function satisfies Hölderian error bound with a unique optimum. Wirth et al. (2023b) extended the accelerated results to FW with open-loop step-size of the form of $\eta_t = \frac{\ell}{t+\ell}$ where $\ell \in \mathbb{N}_{\geq 4}$, establishing the rate of $\mathcal{O}(1/t^\ell)$.

Similar approaches lead to the accelerated convergence rate for ALM with open-loop step-sizes over strongly convex feasible regions without assuming a strong convexity of the objective function. Instead, we have that the objective function $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ satisfies the (PL) inequality, which is a weaker condition than the strong convexity. The scaling inequality below exploits the geometry of the feasible region.

Lemma 5.1 *(Kerdreux et al., 2021a, Lemma 2.1) Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact $\alpha_\mathcal{P}, \alpha_\mathcal{Q}$-strongly convex sets, respectively, and let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be a convex and L-smooth function. Then, for all $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$, it holds that*

$$\langle \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}), \mathbf{x} - \mathbf{u} \rangle \geq \frac{\alpha_\mathcal{P}}{2} \|\mathbf{x} - \mathbf{u}\|_2^2 \|\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y})\|_2$$
$$\langle \nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}), \mathbf{y} - \mathbf{v} \rangle \geq \frac{\alpha_\mathcal{Q}}{2} \|\mathbf{y} - \mathbf{v}\|_2^2 \|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y})\|_2,$$

where $\mathbf{u} \in \arg\min_{\mathbf{p} \in \mathcal{P}} \langle \nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y}), \mathbf{x} - \mathbf{p} \rangle$ and $\mathbf{v} \in \arg\min_{\mathbf{q} \in \mathcal{Q}} \langle \nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y}), \mathbf{y} - \mathbf{q} \rangle$, respectively.

The following technical lemma is required to derive an improved convergence rate for ALM with open-loop step-sizes, a variant of Lemma 3.5 derived by Wirth et al. (2023b).

Lemma 5.2 *Let $\epsilon \in [0, \ell]$, let $\nu \in]0, 1]$ and for $t \geq S$, let $\eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$. Then, it holds that*

$$\prod_{i=S}^{t} \left(1 - \left(1 - \frac{\epsilon}{\ell}\right) \nu \eta_i\right) \leq \exp\left(\frac{\epsilon(\ell+1)}{S + \left\lfloor \frac{\ell}{\nu} \right\rfloor - \ell}\right) \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell - \epsilon}.$$

Proof The proof is presented in Appendix A.2. ■

We begin with our presentation when the feasible regions are disjoint strongly convex sets, see Figure 1a for a graphical representation. In this setting, the distance between $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$ is lower bounded by the minimal distance between two sets. Thus, it holds that

$$\|\nabla_\mathbf{x} f(\mathbf{x}, \mathbf{y})\|_2 = \|\nabla_\mathbf{y} f(\mathbf{x}, \mathbf{y})\|_2 = \|\mathbf{x} - \mathbf{y}\|_2 \geq \operatorname{dist}(\mathcal{P}, \mathcal{Q}) \tag{14}$$

for all $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$, the norm of the gradient of the objective is bounded away from the positive constant. Under this setting, we can apply Lemma 5.1 to achieve a faster convergence rate.

Theorem 5.3 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact $\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}$-strongly convex sets of diameters $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, such that $\mathcal{P} \cap Q = \emptyset$. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x}-\mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$, let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $\epsilon \in]0, \ell]$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \left\lceil \ell^2 / \left(\epsilon \nu \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\} \mathrm{dist}(\mathcal{P}, \mathcal{Q})\right)\right\rceil$. Then, for the iterates of Algorithm 3 with step-size η_t, it holds that*

$$h_t \leq \begin{cases} \eta_{t-1} \frac{\left(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2\right)\ell}{2\nu} = \mathcal{O}(1/t), & t \leq S \\ \eta_{t-1}^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \left\lfloor \frac{\ell}{\nu} \right\rfloor - \ell}\right) \frac{h_S}{\eta_{S-1}^{\ell-\epsilon}} = \mathcal{O}(1/t^{\ell-\epsilon}), & t \geq S. \end{cases}$$

Proof Since $\mathcal{P}$ and $\mathcal{Q}$ are $\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}$-strongly convex sets, respectively, it holds that by Lemma 5.1,

$$\begin{aligned} g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) &= \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle + \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle \\ &\geq \frac{\alpha_{\mathcal{P}}}{2} \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 + \frac{\alpha_{\mathcal{Q}}}{2} \|\mathbf{y}_t - \mathbf{v}_t\|_2^2 \|\nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t)\|_2 \\ &\geq \frac{\alpha_{\mathcal{P}}}{2} \mathrm{dist}(\mathcal{P}, \mathcal{Q}) \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \frac{\alpha_{\mathcal{Q}}}{2} \mathrm{dist}(\mathcal{P}, \mathcal{Q}) \|\mathbf{y}_t - \mathbf{v}_t\|_2^2 \\ &\geq \frac{\min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\}}{2} \mathrm{dist}(\mathcal{P}, \mathcal{Q}) \left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right). \end{aligned}$$

where the second inequality holds due to (14). Plugging this bound into Lemma 4.6 gives

$$\begin{aligned} h_{t+1} &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t \frac{\epsilon}{2\ell} \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\} \mathrm{dist}(\mathcal{P}, \mathcal{Q}) \left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) + \frac{\eta_t^2}{2} \left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) \\ &= \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t + \frac{\eta_t}{2} \left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) \left(\eta_t - \frac{\epsilon}{\ell} \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\} \mathrm{dist}(\mathcal{P}, \mathcal{Q})\right). \end{aligned}$$

Let $S = \left\lceil \frac{\ell^2}{\epsilon\nu \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\} \mathrm{dist}(\mathcal{P}, \mathcal{Q})} \right\rceil$. Then, for $t \geq S$, it holds that $\eta_t - \frac{\epsilon}{\ell} \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\} \mathrm{dist}(\mathcal{P}, \mathcal{Q}) \leq 0$. Hence, for all $t \geq S$, by Lemma 5.2, it holds that

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t \leq \prod_{i=S}^{t} \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_i\right) h_S \leq \exp\left(\frac{\epsilon(\ell+1)}{S + \left\lfloor \frac{\ell}{\nu} \right\rfloor - \ell}\right) \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell-\epsilon} h_S.$$

■

Next, we consider the setting when $\mathcal{P}$ and $\mathcal{Q}$ are strongly convex and no assumptions are made regarding the relation between the feasible regions. In order to analyze this scenario, Lemma 5.4 is essential to show that `ALM` with open-loop step-sizes achieves a convergence rate of $\mathcal{O}(t^{-\ell+\epsilon} + t^{-2})$ for $\epsilon \in [0, \ell]$, after the iteration S. Note that the rate is bounded to $\mathcal{O}(t^{-2})$ when $\ell - \epsilon \geq 2$, while the rate $\mathcal{O}(t^{-\ell+\epsilon})$ is attained when $\ell - \epsilon < 2$.

Lemma 5.4 *Let $\eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $\epsilon \in [0, \ell]$, let $\nu \in]0, 1]$, and let $S \in \mathbb{N}$. Suppose that there exists $A, B, C > 0$, a nonnegative sequence $\{C_t\}_{t=S}^{\infty}$ such that $0 \leq C_t \leq C$, and the sequence $\{h_t\}_{t=S}^{\infty}$ satisfies*

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t A C_t h_t^{\frac{1}{2}} + \eta_t^2 B C_t \tag{15}$$

for all $t \geq S$. Then, for all $t \geq S$, it holds that

$$\begin{aligned} h_t &\leq \max\left\{ \left(\frac{\eta_{t-1}}{\eta_{S-1}}\right)^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \left\lfloor \frac{\ell}{\nu} \right\rfloor - \ell}\right) h_S, \eta_{t-1}^2 \left(\frac{B^2}{A^2} + BC\right) \right\} \\ &= \mathcal{O}(t^{-\ell+\epsilon} + t^{-2}). \end{aligned} \tag{16}$$

Proof The proof is a straightforward modification of the proof of Proposition 2.2 in Bach (2021), and we extend the lemma for general $\ell \in \mathbb{N}_{\geq 2}$. We prove by induction on t. Let $t \geq S$. The base case $t = S$ is clear. Suppose (16) holds for some $t \geq S$. We distinguish between two cases.

1. Suppose that

$$h_t \leq \left(\frac{\eta_t B}{A}\right)^2. \tag{17}$$

Combined with an upper bound on (15), we obtain

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

2. Suppose that (17) does not hold. Then, from (15), we have

$$\begin{aligned} h_{t+1} &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t A C_t h_t^{\frac{1}{2}} + \eta_t^2 BC_t \\ &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t \\ &= \prod_{i=S}^{t} \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_i\right) h_S \\ &\leq \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor \frac{\ell}{\nu} \rfloor - \ell}\right) h_S \end{aligned}$$

where the last inequality holds due to Lemma 5.2.

In both cases, the bound (16) holds for $t+1$. Thus, (16) holds for all $t \geq S$. ■

Finally, we prove that `ALM` with the open-loop step-sizes admits improved convergence rates over strongly convex sets, without further assumptions on the lower bound of gradients.

Theorem 5.5 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be compact $\alpha_\mathcal{P}, \alpha_\mathcal{Q}$-strongly convex sets of diameters $D_\mathcal{P}, D_\mathcal{Q} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$, let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, and let $\epsilon \in]0, \ell]$. Then, for the iterates of Algorithm 3 with step-size η_t, it holds that*

$$\begin{aligned} h_t &\leq \max\left\{\eta_{t-1}^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{1 + \lfloor \frac{\ell}{\nu} \rfloor - \ell}\right) h_1, \eta_{t-1}^2 \left(\frac{\epsilon \alpha_\mathcal{Q} D_\mathcal{P}}{2\ell} + 2\right)\left(\frac{2\epsilon\alpha_\mathcal{Q} D_\mathcal{P}\ell + 8\ell^2}{\epsilon^2 \min\{\alpha_\mathcal{P}, \alpha_\mathcal{Q}\}^2} + \max\{D_\mathcal{P}, D_\mathcal{Q}\}^2\right)\right\} \\ &= \mathcal{O}(t^{-\ell+\epsilon} + t^{-2}). \end{aligned}$$

Proof Since f satisfies the (PL) by Lemma 4.1, we obtain

$$\|\nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 = \|\nabla_\mathbf{y} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 = \frac{1}{\sqrt{2}}\|\nabla f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \geq h_t^{\frac{1}{2}}. \tag{18}$$

From strong convexity of $\mathcal{P}$, we apply Lemma 5.1 and get

$$g_\mathbf{x}(\mathbf{x}_t, \mathbf{y}_t) = \langle \nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t\rangle \geq \frac{\alpha_\mathcal{P}}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \|\nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \geq \frac{\alpha_\mathcal{P}}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 h_t^{\frac{1}{2}} \tag{19}$$

where the last inequality holds due to (18). By triangle inequality, it holds that

$$\|\mathbf{y}_t - \mathbf{x}_t\|_2 = \|(\mathbf{y}_t - \mathbf{x}_{t+1}) + (\mathbf{x}_{t+1} - \mathbf{x}_t)\|_2 \leq \|\mathbf{y}_t - \mathbf{x}_{t+1}\|_2 + \eta_t\|\mathbf{u}_t - \mathbf{x}_t\|_2 \leq \|\mathbf{y}_t - \mathbf{x}_{t+1}\|_2 + \eta_t D_\mathcal{P},$$

which implies

$$\|\nabla_\mathbf{y} f(\mathbf{x}_{t+1}, \mathbf{y}_t)\|_2 = \|\mathbf{y}_t - \mathbf{x}_{t+1}\|_2 \geq \|\mathbf{y}_t - \mathbf{x}_t\|_2 - \eta_t D_\mathcal{P} = \|\nabla_\mathbf{y} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 - \eta_t D_\mathcal{P} \geq h_t^{\frac{1}{2}} - \eta_t D_\mathcal{P} \tag{20}$$

where the last inequality follows from (18). From $\alpha_\mathcal{Q}$-strong convexity of $\mathcal{Q}$, we obtain

$$\begin{aligned} g_\mathbf{y}(\mathbf{x}_{t+1}, \mathbf{y}_t) &= \langle \nabla_\mathbf{y} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t\rangle \\ &\geq \frac{\alpha_\mathcal{Q}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2 \|\nabla_\mathbf{y} f(\mathbf{x}_{t+1}, \mathbf{y}_t)\|_2 \\ &\geq \frac{\alpha_\mathcal{Q}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2 h_t^{\frac{1}{2}} - \eta_t \frac{\alpha_\mathcal{Q} D_\mathcal{P}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2 \end{aligned} \tag{21}$$

where the last inequality holds due to (20). Thus, combining (19) with (21) gives

$$g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) \geq \left(\frac{\alpha_{\mathcal{P}}}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \frac{\alpha_{\mathcal{Q}}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) h_t^{\frac{1}{2}} - \eta_t \frac{\alpha_{\mathcal{Q}} D_{\mathcal{P}}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2$$
$$\geq \frac{\min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\}}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) h_t^{\frac{1}{2}} - \eta_t \frac{\alpha_{\mathcal{Q}} D_{\mathcal{P}}}{2}\|\mathbf{y}_t - \mathbf{v}_t\|_2^2$$

Plugging this bound into Lemma 4.6 yields

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t \frac{\epsilon}{2\ell} \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right) h_t^{\frac{1}{2}}$$
$$+ \frac{\eta_t^2}{2}\left(\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \left(\frac{\epsilon \alpha_{\mathcal{Q}} D_{\mathcal{P}}}{\ell} + 1\right)\|\mathbf{y}_t - \mathbf{v}_t\|_2^2\right)$$

Without loss of generality, suppose that $\|\mathbf{x}_t - \mathbf{u}_t\|_2 \geq \|\mathbf{y}_t - \mathbf{v}_t\|_2$. Then, we have

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t \frac{\epsilon}{2\ell} \min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 h_t^{\frac{1}{2}} + \frac{\eta_t^2}{2}\left(\frac{\epsilon \alpha_{\mathcal{Q}} D_{\mathcal{P}}}{\ell} + 2\right)\|\mathbf{x}_t - \mathbf{u}_t\|_2^2.$$

We then apply Lemma 5.4 for $S = 1$ with $A = \frac{\epsilon}{2\ell}\min\{\alpha_{\mathcal{P}}, \alpha_{\mathcal{Q}}\}, B = \frac{\epsilon \alpha_{\mathcal{Q}} D_{\mathcal{P}}}{2\ell} + 2, C = \max\{D_{\mathcal{P}}, D_{\mathcal{Q}}\}^2$ and $C_t = \|\mathbf{x}_t - \mathbf{u}_t\|_2^2$, which concludes the claim. ■

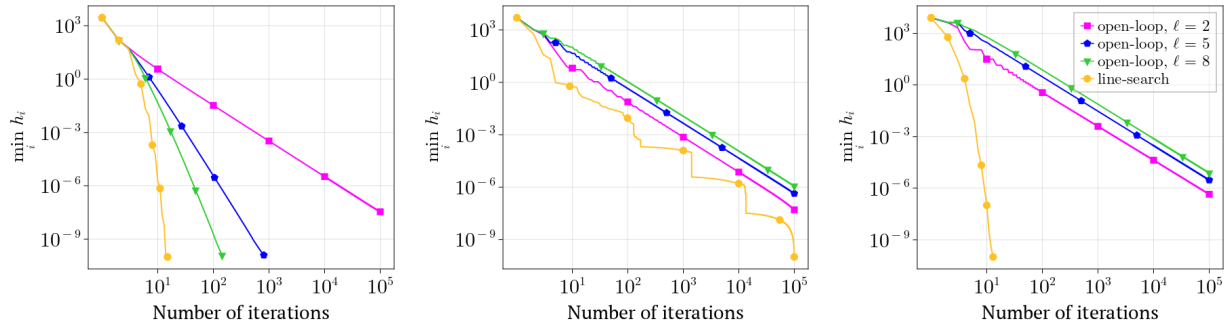
We revisit definition of the ℓ_p-norm of $\mathbf{z} \in \mathbb{R}^d$, when $p \geq 1$:

$$\|\mathbf{z}\|_p = \left(\sum_{i=1}^{d} |\mathbf{z}_i|^p\right)^{\frac{1}{p}}.$$

The ℓ_p-ball is then characterized as the set of points satisfying $\|\mathbf{z}\|_p \leq 1$ for $\mathbf{z} \in \mathbb{R}^d$. Specifically, when $p \in]1, 2]$, the ℓ_p-ball represents $\frac{p-1}{4}$-strongly convex sets (Hanner, 1956). In the forthcoming example, we explore scenarios involving two ℓ_2-balls.

Example 5.6 (Strongly convex setting) Let $\mathcal{P} = \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x} - \mathbf{c}_{\mathcal{P}}\|_2 \leq r_{\mathcal{P}}\right\}$ be an ℓ_2-balls with radius $r_{\mathcal{P}} > 0$ shifted by $\mathbf{c}_{\mathcal{P}} \in \mathbb{R}^d$, and $\mathcal{Q} = \left\{\mathbf{y} \in \mathbb{R}^d \mid \|\mathbf{y} - \mathbf{c}_{\mathcal{Q}}\|_2 \leq r_{\mathcal{Q}}\right\}$ be an ℓ_2-balls with radius $r_{\mathcal{Q}} > 0$ shifted by $\mathbf{c}_{\mathcal{Q}} \in \mathbb{R}^d$. Then, both feasible sets $\mathcal{P}$ and $\mathcal{Q}$ are $\frac{1}{4}$-strongly convex sets. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$. When $r_{\mathcal{Q}} \in]0, \|\mathbf{c}_{\mathcal{P}} - \mathbf{c}_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}[$, then two sets are disjoint, and the norm of the gradient of f is bounded away from $\mathrm{dist}(\mathcal{P}, \mathcal{Q}) = \|\mathbf{c}_{\mathcal{P}} - \mathbf{c}_{\mathcal{Q}}\|_2 - r_{\mathcal{P}} - r_{\mathcal{Q}} > 0$. On the other hand, when $r_{\mathcal{Q}} \geq \|\mathbf{c}_{\mathcal{P}} - \mathbf{c}_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}$, then two sets intersect.

Experimental settings of Figure 2. In the setting of Example 5.6, for $d = 100$, random vectors $\mathbf{c}_{\mathcal{P}}, \mathbf{c}_{\mathcal{Q}}$ and random number $r_{\mathcal{P}}$, we compare `ALM` with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t + \ell}$ for $\ell \in \{2, 5, 8\}$ and line-search, starting with $\mathbf{x}_0 = \mathbf{c}_{\mathcal{P}} + r_{\mathcal{P}} e^{(1)}$ and $\mathbf{y}_0 = \mathbf{c}_{\mathcal{Q}} + r_{\mathcal{Q}} e^{(d)}$. We choose $\nu = 1$ and $r_{\mathcal{Q}} \in \left\{\frac{1}{2}\|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}, \|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}, \frac{3}{2}\|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}\right\}$ for the cases where two sets are disjoint (Figure 2a), two sets intersect at a single point (Figure 2b) and two sets intersect at multiple points (Figure 2c), respectively. The results of the experiments are plotted in log-log plots in Figure 2.



(a) Disjoint. (b) Intersect at a single point. (c) Intersect at multiple points.

Figure 2: Convergence rate comparison of ALM with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \{2, 5, 8\}$ with approximation error $\nu = 1$. For $d = 100$, the feasible regions $\mathcal{P} \subseteq \mathbb{R}^d$ is an ℓ_2-balls with radius $r_{\mathcal{P}} > 0$ shifted by $\mathbf{c}_{\mathcal{P}} \in \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ is an ℓ_2-balls with radius $r_{\mathcal{Q}} > 0$ shifted by $\mathbf{c}_{\mathcal{Q}} \in \mathbb{R}^d$. For random vectors $\mathbf{c}_{\mathcal{P}}, \mathbf{c}_{\mathcal{Q}}$ and the random number $r_{\mathcal{P}}$, $r_{\mathcal{Q}} \in \left\{\frac{1}{2}\|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}, \|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}, \frac{3}{2}\|C_{\mathcal{P}} - C_{\mathcal{Q}}\|_2 - r_{\mathcal{P}}\right\}$ correspond to the case where two regions are disjoint, intersect at a single point and intersect at multiple points, respectively.

In figure 2, we observe that ALM with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \mathbb{N}_{\geq 2}$ converges at an expected rate of order $\mathcal{O}(1/t^{\ell})$ for the disjoint case, and at a rate of $\mathcal{O}(1/t^2)$ for the intersect cases, irrespective of the number of points in the intersection. However, ALM with line-search exhibits different convergence behavior when two sets intersect at a single point, whereas it converges linearly in other cases, posing an open question. Additionally, the effect of different values of approximation error $\nu \in]0, 1]$ on the performance of ALM with open-loop step-sizes is yet to be discussed. For further studies, refer to Section 6 for the ablation study of approximation error.

5.2 $\mathcal{P}$ is a strongly convex set and $\mathcal{Q}$ is a polytope

In Section 5.1, we established the setting when both feasible regions are strongly convex, ALM with open-loop step-sizes achieves improved convergence rates, as illustrated in Figure 2. In this section, we change one of the feasible regions from a strongly convex set to another and analyze its convergence rate. Specifically, when $\mathcal{P}$ is a strongly convex set and $\mathcal{Q}$ is a polytope, we consider scenarios where these sets are either disjoint or intersect at a single point.

When working with structured feasible regions like polytopes, the convergence rates of the LMO based algorithm and its variants often explicitly depend on the dimension of the problem (Lacoste-Julien and Jaggi, 2015; Garber and Hazan, 2015b). To circumvent this dependency, strict complementarity condition is introduced in the early works of Guélat and Marcotte (1986), which is commonly assumed in the FW literature (Garber, 2020; Carderera et al., 2021; Wirth et al., 2023a).

Assumption 5.7 (Strict complementarity) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ be a compact convex set, let $\mathcal{Q} \subseteq \mathbb{R}^d$ be a polytope, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be a convex and L-smooth function, let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set, and let $\mathcal{Q}^*$ be the optimal face of $\mathcal{Q}$. Then, for all $(\mathbf{x}^*, \mathbf{y}^*) \in \Omega^*$ and $\mathbf{v} \in \mathcal{Q}$, we have that $\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{v} - \mathbf{y}^* \rangle = 0$ if and only if $\mathbf{v} \in \mathcal{Q}^*$. Or stated equivalently, there exists $\kappa_{\mathcal{Q}} > 0$ such that $\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{v} - \mathbf{y}^* \rangle \geq \kappa_{\mathcal{Q}}$ for $\mathbf{v} \in \operatorname{vert}(\mathcal{Q}) \setminus \mathcal{Q}^*$.*

As the iterates of ALM approach the optimal face of the polytope, the LMO gives a vertex within the set of vertices on the optimal face. The lemma below demonstrates the existence of such iterate $S \in \mathbb{N}$, ensuring that ALM with open-loop step-sizes identifies the optimal face after such iteration S is surpassed. We generalize prior results that assume f satisfies strongly convex or Hölderian error bound with unique optimal solution (Braun et al., 2022; Wirth et al., 2023a) to the objective function satisfies the (QG).

Lemma 5.8 (Active set identification) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ be a compact convex set and $\mathcal{Q} \subseteq \mathbb{R}^d$ be a polytope of diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$, and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set. Suppose that there exists $\kappa_{\mathcal{Q}} > 0$ such that Assumption 5.7 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error and let $S = \left\lceil 4\left(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2\right)\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\right\rceil$. Then, for the iterates of Algorithm 3 with step-size η_t, it holds that $\mathbf{v}_t \in \mathrm{vert}(\mathcal{Q}^*)$ for all $t \geq S$.*

Proof For every iteration $t \in \mathbb{N}$, we denote $(\mathbf{x}_t^*, \mathbf{y}_t^*) \in \arg\min_{(\mathbf{z},\mathbf{w})\in\Omega^*} \left\| \begin{bmatrix} \mathbf{x}_t - \mathbf{z} \\ \mathbf{y}_t - \mathbf{w} \end{bmatrix} \right\|_2$ the pair of points in Ω^* closest to the iterate $(\mathbf{x}_t, \mathbf{y}_t)$. For $\mathbf{v} \in \mathrm{vert}(\mathcal{Q})$, we write

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{v} - \mathbf{y}_t^* \rangle = \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{v} - \mathbf{y}_t^* \rangle + \langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{v} - \mathbf{y}_t^* \rangle . \tag{22}$$

By Cauchy-Schwarz inequality, we have

$$|\langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{y}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{v} - \mathbf{y}_t^* \rangle| \leq \|\mathbf{y}_t - \mathbf{y}_t^*\|_2 \|\mathbf{v} - \mathbf{y}_t^*\|_2 \leq \|\mathbf{y}_t - \mathbf{y}_t^*\|_2 D_{\mathcal{Q}} < \sqrt{2} D_{\mathcal{Q}} h_t^{\frac{1}{2}} < \frac{\kappa_Q}{2} \tag{23}$$

where the second inequality holds due to $\|\mathbf{v} - \mathbf{y}_t^*\|_2 \leq D_{\mathcal{Q}}$, and third inequality follows from $\|\mathbf{y}_t - \mathbf{y}_t^*\|_2 \leq \left\| \begin{bmatrix} \mathbf{x}_t - \mathbf{x}_t^* \\ \mathbf{y}_t - \mathbf{y}_t^* \end{bmatrix} \right\|_2 \leq \sqrt{2} h_t^{\frac{1}{2}}$ by (2), and the last inequality holds due to Proposition 4.8 for $t \geq S$, i.e.,

$$\sqrt{2} D_{\mathcal{Q}} h_t^{\frac{1}{2}} \leq \sqrt{2} D_{\mathcal{Q}} \left(\frac{(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2}{2\nu^2 (t - 1 + \ell/\nu)} \right)^{\frac{1}{2}} \leq \sqrt{2} D_{\mathcal{Q}} \left(\frac{(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2}{2\nu^2 \left(\frac{(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2}{2\nu^2} \right) \left(\frac{2\sqrt{2} D_{\mathcal{Q}}}{\kappa_{\mathcal{Q}}} \right)^2} \right)^{\frac{1}{2}} < \frac{\kappa_{\mathcal{Q}}}{2}.$$

Now, we distinguish between two cases.
First, suppose that $\mathbf{v} \notin \mathrm{vert}(\mathcal{Q}^*)$. Then, by (22), (23) and by strict complementarity, we have

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{v} - \mathbf{y}_t^* \rangle > \kappa_{\mathcal{Q}} - \frac{\kappa_{\mathcal{Q}}}{2} = \frac{\kappa_{\mathcal{Q}}}{2}.$$

Next, suppose that $\mathbf{v} \in \mathrm{vert}(\mathcal{Q}^*)$. By similar analysis, we obtain

$$\langle \nabla_{\mathbf{y}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{v} - \mathbf{y}_t^* \rangle < \frac{\kappa_{\mathcal{Q}}}{2}. \tag{24}$$

for all $t \geq S$.

We use proof by contradiction. Assume that there exists an iterate $n \geq S$ such that $\mathbf{v}_n \in \arg\min_{\mathbf{v}\in\mathrm{vert}(\mathcal{Q})} \langle \nabla_{\mathbf{y}} f(\mathbf{x}_n, \mathbf{y}_n), \mathbf{v} \rangle$ but $\mathbf{v}_n \notin \mathrm{vert}(\mathcal{Q}^*)$. Then, we obtain $\langle \nabla_{\mathbf{y}} f(\mathbf{x}_n, \mathbf{y}_n), \mathbf{v}_n - \mathbf{y}_n^* \rangle > \frac{\kappa_{\mathcal{Q}}}{2}$. However, there exists $\bar{\mathbf{v}} \in \mathrm{vert}(\mathcal{Q}^*)$ such that (24) holds, which contradicts optimality of $\mathbf{v}_n$. Hence, $\mathbf{v}_t \in \mathrm{vert}(\mathcal{Q}^*)$ for all $t \geq S$. ■

In this section, we assume that the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is a pair of points where $\mathbf{x}^*$ lies on the boundary of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$, thus the optimal solution is unique. From the lemma below, we derive that there exists a certain iterate $S \in \mathbb{N}$ such that for $t \geq S$, all $\mathbf{y}_t \in \mathcal{Q}$ reach the optimum $\mathbf{y}^*$ when $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$.

Lemma 5.9 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ be a compact convex set and $\mathcal{Q} \subseteq \mathbb{R}^d$ be a polytope with diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with unique $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. Suppose that $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$ and that there exists $\kappa_{\mathcal{Q}} > 0$ such that Assumption 5.7 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \left\lceil \left(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2\right)\ell^2/(2\nu^2\kappa_{\mathcal{Q}}) \right\rceil$. Then, for the iterates of Algorithm 3 with step-size η_t, all $\mathbf{y}_t$ reach $\mathbf{y}^*$ for all $t \geq S$.*

Proof Assume that there exists an iterate $n \geq S$ such that $\mathbf{y}_n \neq \mathbf{y}^*$, which follows that $\mathbf{y}_n \in \operatorname{conv}(\operatorname{vert}(\mathcal{Q}) \setminus \{\mathbf{y}^*\})$. From convexity of f, it holds that

$$h_n = f(\mathbf{x}_n, \mathbf{y}_n) - f(\mathbf{x}^*, \mathbf{y}^*) \geq \left\langle \nabla f(\mathbf{x}^*, \mathbf{y}^*), \begin{bmatrix} \mathbf{x}_n - \mathbf{x}^* \\ \mathbf{y}_n - \mathbf{y}^* \end{bmatrix} \right\rangle \geq \langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{y}_n - \mathbf{y}^* \rangle \geq \kappa_{\mathcal{Q}}$$

where the second inequality follows from $\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{x}_n - \mathbf{x}^* \rangle = \langle \mathbf{x}^* - \mathbf{y}^*, \mathbf{x}_n - \mathbf{x}^* \rangle \geq 0$ by Dattorro (2004, Theorem E.9.1.0.2) and the last inequality holds due to Assumption 5.7. However, we have $h_t < \kappa_{\mathcal{Q}}$ for all $t \geq S$, which contradicts the assumption. Thus, $\mathbf{y}_t = \mathbf{y}^*$ for all $t \geq S$. ■

With Lemma 5.9, we present our first result when $\mathbf{x}^*$ is on the boundary of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$ such that $\mathcal{P} \cap \mathcal{Q} = \emptyset$, `ALM` with open-loop step-sizes exhibits an accelerated convergence rate, see Figure 1c for a graphical representation. Recall that for $\mathcal{P} \cap \mathcal{Q} = \emptyset$, $\|\nabla_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})\|_2 = \|\mathbf{x} - \mathbf{y}\|_2 \geq \operatorname{dist}(\mathcal{P}, \mathcal{Q})$ for all $\mathbf{x} \in \mathcal{P}$ and $\mathbf{y} \in \mathcal{Q}$, the norm of the gradient of the objective function is bounded away from the positive constant.

Theorem 5.10 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ be a $\alpha_{\mathcal{P}}$-strongly convex set and $\mathcal{Q} \subseteq \mathbb{R}^d$ be a polytope with diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, such that $\mathcal{P} \cap \mathcal{Q} = \emptyset$. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. Suppose that $\mathbf{x}^*$ lies on the boundary of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$, and that there exists $\kappa_{\mathcal{Q}} > 0$ such that Assumption 5.7 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $\epsilon \in]0, \ell]$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let*

$$S = \max\left\{ \left\lceil \ell^2 / \left(\epsilon \nu \alpha_{\mathcal{P}} \operatorname{dist}(\mathcal{P}, \mathcal{Q})\right) \right\rceil, \left\lceil 4(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2 D_{\mathcal{Q}}^2 / (\nu^2 \kappa_{\mathcal{Q}}^2) \right\rceil, \left\lceil (2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2 / (2\nu^2 \kappa_{\mathcal{Q}}) \right\rceil \right\}.$$

Then, for the iterates of Algorithm 3 with step-size η_t, it holds that

$$h_t \leq \begin{cases} \eta_{t-1} \frac{(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell}{2\nu} = \mathcal{O}(1/t), & t \leq S \\ \eta_{t-1}^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor \frac{\ell}{\nu} \rfloor - \ell} \right) \frac{h_S}{\eta_{S-1}^{\ell-\epsilon}} = \mathcal{O}(1/t^{\ell-\epsilon}), & t \geq S. \end{cases}$$

Proof From strong convexity of $\mathcal{P}$, we obtain by Lemma 5.1

$$g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) = \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle \geq \frac{\alpha_{\mathcal{P}}}{2} \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \geq \frac{\alpha_{\mathcal{P}}}{2} \operatorname{dist}(\mathcal{P}, \mathcal{Q}) \|\mathbf{x}_t - \mathbf{u}_t\|_2^2$$

where the last inequality holds due to (14). Let $t \geq S$. Since Lemma 5.8 implies $\mathbf{v}_t \in \operatorname{vert}(\mathcal{Q}^*) = \{\mathbf{y}^*\}$ and Lemma 5.9 gives $\mathbf{y}_t = \mathbf{y}^*$, we obtain $\|\mathbf{y}_t - \mathbf{v}_t\|_2 = 0$. Thus, it holds that $g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) = \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle = 0$, yielding $g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) + g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) \geq \frac{\alpha_{\mathcal{P}}}{2} \operatorname{dist}(\mathcal{P}, \mathcal{Q}) \|\mathbf{x}_t - \mathbf{u}_t\|_2^2$. Plugging this bound into Lemma 4.6 with $\|\mathbf{y}_t - \mathbf{v}_t\|_2 = 0$ gives

$$\begin{aligned} h_{t+1} &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right) \nu \eta_t\right) h_t - \eta_t \frac{\epsilon \alpha_{\mathcal{P}}}{2\ell} \operatorname{dist}(\mathcal{P}, \mathcal{Q}) \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 + \frac{\eta_t^2}{2} \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \\ &= \left(1 - \left(1 - \frac{\epsilon}{\ell}\right) \nu \eta_t\right) h_t + \frac{\eta_t}{2\ell} \|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \left(\ell \eta_t - \epsilon \alpha_{\mathcal{P}} \operatorname{dist}(\mathcal{P}, \mathcal{Q})\right) \end{aligned}$$

Since $\ell \eta_t - \epsilon \alpha_{\mathcal{P}} \operatorname{dist}(\mathcal{P}, \mathcal{Q}) \leq 0$ for $t \geq S$, we conclude the claim by Lemma 5.2

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right) \nu \eta_t\right) h_t = \prod_{i=S}^{t} \left(1 - \left(1 - \frac{\epsilon}{\ell}\right) \nu \eta_i\right) h_S \leq \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor \frac{\ell}{\nu} \rfloor - \ell}\right) h_S.$$

■

We can now apply a similar approach to that in Theorem 5.3 to derive a convergence rate when two sets intersect at a single point, see Figure 1d for a graphical representation. To this end, we simply exploit the geometry of a set, the (PL) property of f and Lemma 5.9.

Theorem 5.11 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ be a $\alpha_{\mathcal{P}}$-strongly convex set and $\mathcal{Q} \subseteq \mathbb{R}^d$ be a polytope with diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ with $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. Suppose that $\mathcal{P}$ and $\mathcal{Q}$ intersect at a single point where $\mathbf{x}^*$ lies on the boundary of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$, and that there exists $\kappa_{\mathcal{Q}} > 0$ such that Assumption 5.7 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $\epsilon \in]0, \ell]$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \max\left\{\left\lceil 4(2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\right\rceil, \left\lceil (2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2)\ell^2/(2\nu^2\kappa_{\mathcal{Q}})\right\rceil\right\}$. Then, for the iterates of Algorithm 3 with step-size η_t, for all $t \geq S$, it holds that*

$$h_t \leq \max\left\{\eta_{t-1}^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor \frac{\ell}{\nu} \rfloor - \ell}\right) \frac{h_S}{\eta_{S-1}^{\ell-\epsilon}}, \eta_{t-1}^2 \left(\frac{\ell^2}{\epsilon^2 \alpha_{\mathcal{P}}^2} + \frac{D_{\mathcal{P}}^2}{2}\right)\right\} = \mathcal{O}(t^{-\ell+\epsilon} + t^{-2}).$$

Proof Since $\mathcal{P}$ is strongly convex, from Lemma 5.1, we obtain

$$g_{\mathbf{x}}(\mathbf{x}_t, \mathbf{y}_t) = \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle \geq \frac{\alpha_{\mathcal{P}}}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 \|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \geq \frac{\alpha_{\mathcal{P}}}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 h_t^{\frac{1}{2}}$$

where the last inequality follows from the fact that $\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\|_2 = \frac{1}{\sqrt{2}}\|\nabla f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \geq h_t^{\frac{1}{2}}$ since f satisfies the (PL) by Lemma 4.1. Let $t \geq S$. As we demonstrate in the proof of Theorem 5.10 when $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$, it holds that $g_{\mathbf{y}}(\mathbf{x}_{t+1}, \mathbf{y}_t) = \langle \nabla_{\mathbf{y}} f(\mathbf{x}_{t+1}, \mathbf{y}_t), \mathbf{y}_t - \mathbf{v}_t \rangle = 0$ and $\|\mathbf{y}_t - \mathbf{v}_t\|_2 = 0$ by Lemma 5.7 and 5.9. Plugging this bound into Lemma 4.6 with $\|\mathbf{y}_t - \mathbf{v}_t\|_2 = 0$ gives

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t \frac{\epsilon\alpha_{\mathcal{P}}}{2\ell}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2 h_t^{\frac{1}{2}} + \frac{\eta_t^2}{2}\|\mathbf{x}_t - \mathbf{u}_t\|_2^2$$

We then apply Lemma 5.4 with $A = \frac{\epsilon\alpha_{\mathcal{P}}}{2\ell}, B = \frac{1}{2}, C = D_{\mathcal{P}}^2$ and $C_t = \|\mathbf{x}_t - \mathbf{u}_t\|_2^2$, which concludes the claim. ■

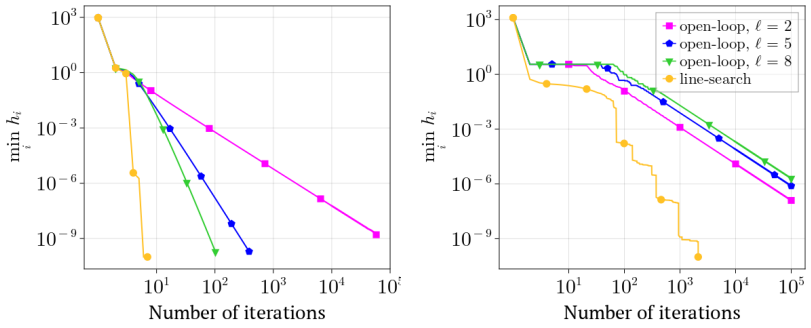
Let $\{e^{(1)}, \ldots, e^{(d)}\}$ be the standard basis vectors and $\mathbb{0}$ be the zero-vector in $\mathbb{R}^d$. The unit simplex is defined as

$$\left\{\mathbf{z} \in \mathbb{R}^d \mid \langle \mathbf{z}, \mathbb{1} \rangle \leq 1, \mathbf{z} \geq 0\right\} = \operatorname{conv}(\mathbb{0}, e^{(1)}, \ldots, e^{(d)})$$

the convex combination of the standard basis vectors and the zero vector. The following example investigates scenarios involving the ℓ_2-ball and the unit simplex, representing a $\frac{1}{4}$-strongly convex set and a polytope.

Example 5.12 (Strongly convex and polytope setting) Let $\sigma > 0$, let $\mathcal{P} = \left\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 \leq r\right\}$ be an ℓ_2-balls with radius $r > 0$, and let $\mathcal{Q} = \left\{\mathbf{y} \in \mathbb{R}^d \mid \langle \mathbf{y} - \sigma\mathbb{1}, \mathbb{1} \rangle \leq \sigma, \mathbf{y} \geq 0\right\}$ be an unit simplex shifted by $\sigma\mathbb{1}$ and scaled by σ. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, and let $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{P} \times \mathcal{Q}$ be a pair of points with minimal distance between $\mathcal{P}$ and $\mathcal{Q}$. Then, $(\mathbf{x}^*, \mathbf{y}^*) = (r\mathbb{1}, \sigma\mathbb{1})$ where $\mathbf{x}^*$ lies on the boundary of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$. When $r < \sigma\|\mathbb{1}\|_2$, then two sets are disjoint, and the norm of the gradient of f is bounded away from $\operatorname{dist}(\mathcal{P}, \mathcal{Q}) = \sigma\|\mathbb{1}\|_2 - r > 0$. When $r = \sigma\|\mathbb{1}\|_2$, two sets intersect a single point.

Experimental settings of Figure 3. In the setting of Example 5.12, for $d = 100$, we compare `ALM` with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t + \ell}$ for $\ell \in \{2, 5, 8\}$, starting with $\mathbf{x}_0 = e^{(1)}$ and $\mathbf{y}_0 = \sigma\mathbb{1} + \frac{\sigma}{2}e^{(2)}$. We choose $\sigma = 5$, $\nu = 1$ and $r \in \left\{\frac{\sigma}{2}\|\mathbb{1}\|_2, \sigma\|\mathbb{1}\|_2\right\}$ for the cases where two sets are disjoint (Figure 3a), and two set intersect at a single point (Figure 3b), respectively. The results of the experiments are plotted in log-log plots in Figure 3.



(a) $r = \frac{\rho}{2}\|\mathbb{1}\|_2$. (b) $r = \rho\|\mathbb{1}\|_2$.

Figure 3: Convergence rate comparison of ALM with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \{2, 5, 8\}$ with approximation error $\nu = 1$. For $d = 100$, the feasible region $\mathcal{P} \subseteq \mathbb{R}^d$ is the ℓ_2-ball with a radius $r > 0$, and $\mathcal{Q} \subseteq \mathbb{R}^d$ is the unit simplex shifted by $\sigma\mathbb{1}$ and scaled by $\sigma > 0$, where $r \in \left\{\frac{\sigma}{2}\|\mathbb{1}\|_2, \sigma\|\mathbb{1}\|_2\right\}$ corresponding to the case where two regions are disjoint and intersect at a single point, respectively.

In Figure 3, we observe that ALM with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \mathbb{N}_{\geq 2}$ converges at the expected rates of order $\mathcal{O}(1/t^{\ell})$ for the disjoint case, and of order $\mathcal{O}(1/t^2)$ for the intersect at a single point case after an initial burn-in phase. Furthermore, a linear convergence of ALM with line-search is shown in Figure 3a, while the experimental results of ALM with line-search in Figure 3b poses an open question similar to that depicted in Figure 2b. Additionally, the ablation study investigating the effect of approximation error $\nu \in]0, 1]$ is presented in Section 6.

5.3 $\mathcal{P}$ and $\mathcal{Q}$ are polytopes

The convergence rates of some LMO based algorithms over a polytope can vary with the location of the optimal solution. Specifically, the vanilla FW with line-search or short-step converges linearly when the optimum of the strongly convex function is contained in the interior of the polytope (Guélat and Marcotte, 1986). However, when the optimum lies inside a face under mild assumptions, the iterates move in progressively inefficient directions. As a result, a lower bound of the convergence rate is slightly worse than a rate of order $\mathcal{O}(1/t)$, not only in practice but in theoretical analysis (Wolfe, 1976; Canon and Cullum, 1968). On the other hand, FW with open-loop step-sizes converges at a rate of $\mathcal{O}(1/t^2)$ (Bach, 2021), demonstrating that FW with open-loop step-sizes is faster than FW with line-search or short-step in several settings (Wirth et al., 2023a,b).

In this section, we further investigate accelerated convergence rates over two disjoint polytopes with two possible scenarios of the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$. First, $\mathbf{x}^*$ and $\mathbf{y}^*$ are vertices of each polytope. Second, $\mathbf{x}^*$ lies in the relative interior of at least one-dimensional face $\mathcal{P}^*$ of $\mathcal{P}$, and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$. For the latter case, in particular, we remark whether ALM with open-loop step-sizes demonstrates faster convergence rates in comparison to ALM with line-search, similar to the phenomenon observed in the vanilla FW algorithm.

We extend Assumption 5.7 for the case of two polytopes.

Assumption 5.13 (Strict complementarity for two polytopes) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be a convex and L-smooth function, and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set. Let $\mathcal{P}^*$ and $\mathcal{Q}^*$ be the optimal faces of $\mathcal{P}$ and $\mathcal{Q}$, respectively. Then, for all $(\mathbf{x}^*, \mathbf{y}^*) \in \Omega^*$ and $\mathbf{u} \in \mathcal{P}$, we have that $\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{u} - \mathbf{x}^* \rangle = 0$ if and only if $\mathbf{u} \in \mathcal{P}^*$. Otherwise, there exists $\kappa_{\mathcal{P}} > 0$ such that $\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{u} - \mathbf{x}^* \rangle \geq \kappa_{\mathcal{P}}$ for $\mathbf{u} \in \operatorname{vert}(\mathcal{P}) \setminus \mathcal{P}^*$. Moreover, for all $(\mathbf{x}^*, \mathbf{y}^*) \in \Omega^*$ and $\mathbf{v} \in \mathcal{Q}$, we have that $\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{v} - \mathbf{y}^* \rangle = 0$ if and only if $\mathbf{v} \in \mathcal{Q}^*$. Otherwise, there exists $\kappa_{\mathcal{Q}} > 0$ such that $\langle \nabla_{\mathbf{y}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{v} - \mathbf{y}^* \rangle \geq \kappa_{\mathcal{Q}}$ for $\mathbf{v} \in \operatorname{vert}(\mathcal{Q}) \setminus \mathcal{Q}^*$.*

Optimal solution - a pair of vertices

We first discuss the case where the optimal solution of f is a pair of vertices of two polytopes. In this setting, when we denote $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ is the optimal solution set, two possible scenarios arise: either $|\Omega^*| = 1$ or $|\Omega^*| \geq 2$. The latter case implies that at least one-dimensional optimal faces of two polytopes, including pairs of vertices, are parallel. Therefore, multiple optimal solutions exist. In this section, however, we specifically consider where a pair of vertices is the unique optimal solution facing one another, see Figure 1e for a graphical representation. In this special case, `ALM` with open-loop step-sizes finds the optimal solution in a finite number of iterations, without requiring the objective function to satisfy additional properties.

Theorem 5.14 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes of diameter $D_\mathcal{P}, D_\mathcal{Q} > 0$ respectively, such that $\mathcal{P} \cap \mathcal{Q} = \emptyset$, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. Suppose that $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is unique and a pair of vertices of $\mathcal{P} \times \mathcal{Q}$ and that there exists $\kappa_\mathcal{P}, \kappa_\mathcal{Q} > 0$ such that Assumption 5.13 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$ and let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error. Then, Algorithm 3 with step-size η_t finds the optimal solution in $\mathcal{O}\left(\left\lceil(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell^2/(2\nu^2 \min\{\kappa_\mathcal{P}, \kappa_\mathcal{Q}\})\right\rceil\right)$ iterations.*

Proof We use proof by contradiction. Assume that Algorithm 3 with step-size η_t runs for $T - 1$ iterations and the final iterate satisfies $\mathbf{x}_T \neq \mathbf{x}^*$ or $\mathbf{y}_T \neq \mathbf{y}^*$, which implies $\mathbf{x}_T \in \text{conv}(\text{vert}(\mathcal{P}) \setminus \{\mathbf{x}^*\})$ or $\mathbf{y}_T \in \text{conv}(\text{vert}(\mathcal{Q}) \setminus \{\mathbf{y}^*\})$. From convexity of f, it holds that

$$h_T = f(\mathbf{x}_T, \mathbf{y}_T) - f(\mathbf{x}^*, \mathbf{y}^*) \geq \left\langle \nabla f(\mathbf{x}^*, \mathbf{y}^*), \begin{bmatrix} \mathbf{x}_T - \mathbf{x}^* \\ \mathbf{y}_T - \mathbf{y}^* \end{bmatrix} \right\rangle = \langle \nabla_\mathbf{x} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{x}_T - \mathbf{x}^* \rangle + \langle \nabla_\mathbf{y} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{y}_T - \mathbf{y}^* \rangle.$$

By Assumption 5.13, it holds that either $h_T \geq \min\{\kappa_\mathcal{P}, \kappa_\mathcal{Q}\}$ if $\mathbf{x}_T \neq \mathbf{x}^*$ and $\mathbf{y}_T = \mathbf{y}^*$ without loss of generality, or $h_T \geq \kappa_\mathcal{P} + \kappa_\mathcal{Q}$ if $\mathbf{x}_T \neq \mathbf{x}^*$ and $\mathbf{y}_T \neq \mathbf{y}^*$. Thus, $h_T \geq \min\{\kappa_\mathcal{P}, \kappa_\mathcal{Q}\}$ in both cases. However, by Proposition 4.8, it follows that after $T = \left\lceil(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell^2/(2\nu^2 \min\{\kappa_\mathcal{P}, \kappa_\mathcal{Q}\})\right\rceil$ iterations, $h_T < \min\{\kappa_\mathcal{P}, \kappa_\mathcal{Q}\}$ which contradicts the assumption. Thus, the final iterate satisfies $\mathbf{x}_T = \mathbf{x}^*$ and $\mathbf{y}_T = \mathbf{y}^*$. ■

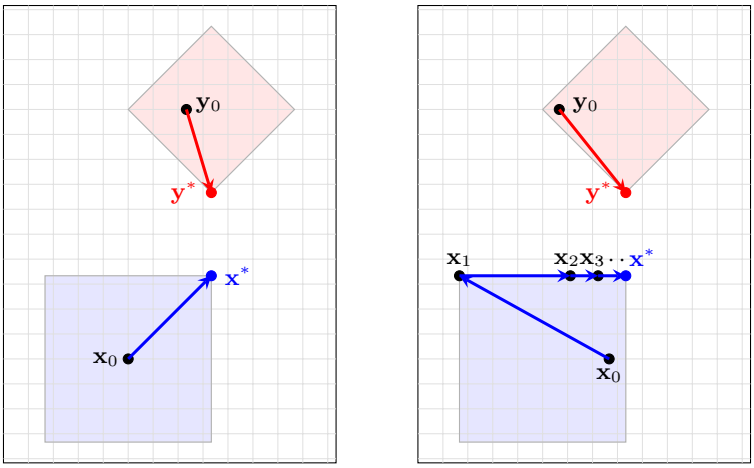
We revisit definition of the ℓ_∞-norm of $\mathbf{z} \in \mathbb{R}^d$:

$$\|\mathbf{z}\|_\infty = \max_{i\in\{1,\ldots,d\}} |\mathbf{z}_i|.$$

The ℓ_∞-ball is the set of points satisfying $\|\mathbf{z}\|_\infty \leq 1$, the d-dimensional hypercube $[-1, 1]^d$. The following example considers the ℓ_∞-ball and the ℓ_1-ball, representing two polytopes.

Example 5.15 (Case study of the finite convergence) Let $\mathcal{P} = \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_\infty \leq 1\}$ be the ℓ_∞-ball, and let $\mathcal{Q} = \{\mathbf{y} \in \mathbb{R}^d \mid \|\mathbf{y} - \mathbf{c}\|_1 \leq 1\}$ be an ℓ_1-ball shifted by $\mathbf{c} \in \mathbb{R}^d$. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, and let $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ be a pair of points with minimal distance between $\mathcal{P}$ and $\mathcal{Q}$. Then, both feasible sets are convex polytopes.

Experimental settings of Figure 4. In the setting of Example 5.15, for $d = 2$, $\alpha, \beta \in \mathbb{R}$ and $\mathbf{c} = e^{(1)} + 3e^{(2)} \in \mathbb{R}^d$, we compare `ALM` with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \mathbb{N}_{\geq 2}$ with different initial points $(\mathbf{x}_0, \mathbf{y}_0) = \left(\alpha e^{(1)}, \beta e^{(1)} + 3e^{(2)}\right) \in \mathcal{P} \times \mathcal{Q}$. We choose $\ell = 2$ and $\nu = 1$ in the experiment. Then, the optimal solution is $(\mathbf{x}^*, \mathbf{y}^*) = \left(e^{(1)} + e^{(2)}, e^{(1)} + 2e^{(2)}\right)$, where $\mathbf{x}^*$ and $\mathbf{y}^*$ is a vertex of $\mathcal{P}$ and $\mathcal{Q}$, respectively. Feasible regions $\mathcal{P}$ and $\mathcal{Q}$, as well as the iterates and the directions of `ALM` with open-loop step-sizes are depicted in Figure 4.



(a) 1-step convergence. (b) Finite convergence.

Figure 4: Initialization dependence of ALM with open-loop step-sizes on finite convergence. The small example allows only a limited number of directions, which clearly illustrate the trajectory. The feasible regions are convex polytopes $\mathcal{P}$ and $\mathcal{Q}$, depicted in blue and red regions, respectively. The trajectories of $\mathbf{x}_t$ and $\mathbf{y}_t$ are illustrated by blue and red arrows within the regions, while the blue and red dots correspond to the optimal solutions $\mathbf{x}^*$ and $\mathbf{y}^*$, respectively.

In Figure 4, the finite convergence of ALM with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ with $\ell = 2$ and $\nu = 1$, is illustrated. Specifically, when initialized at $(\mathbf{x}_0, \mathbf{y}_0) = \left(\alpha e^{(1)}, \beta e^{(1)} + 3e^{(2)}\right) \in \mathcal{P} \times \mathcal{Q}$ such that $\alpha < \beta$, ALM with open-loop step-sizes achieves convergence within a single step, as $\mathbf{u}_0 \in \arg\min_{\mathbf{u}\in\mathcal{P}} \langle \mathbf{x}_0 - \mathbf{y}_0, \mathbf{u}\rangle = \mathbf{x}^*$ and consequently, $\mathbf{x}_1 = \mathbf{x}^*$. On the other hand, when $\alpha \geq \beta$, ALM with open-loop step-sizes converges in a finite number of iterations, as $\mathbf{x}_t$ need to approach $\mathbf{x}^*$ through the update rule $\mathbf{x}_{t+1} = \mathbf{x}_t + \eta_{t,\mathcal{P}}(\mathbf{u}_t - \mathbf{x}_t)$. Thus, when $\alpha \geq \beta$, for $\mathbf{x}_1 = -e^{(1)} + e^{(2)}$, $\mathbf{u}_1 = e^{(1)} + e^{(2)}$ and $\eta_{i,\mathcal{P}} = \frac{2}{i+2}$, we obtain

$$
\begin{aligned}
\mathbf{x}_{t+1} &= (1 - \eta_{t,\mathcal{P}})\mathbf{x}_t + \eta_{t,\mathcal{P}}\mathbf{u}_t \\
&= \prod_{i=1}^{t} \frac{i}{i+2}\mathbf{x}_1 + \sum_{i=1}^{t} \frac{2}{i+2} \prod_{j=i+1}^{t} \frac{j}{j+2}\mathbf{u}_1 \\
&= \left(-\frac{2}{(t+1)(t+2)} + \sum_{i=1}^{t} \frac{2}{i+2} \frac{(i+1)(i+2)}{(t+1)(t+2)} \right) e^{(1)} + e^{(2)} \\
&= \left(\frac{2}{(t+1)(t+2)} \left(-1 + \sum_{i=1}^{t} (i+1) \right) \right) e^{(1)} + e^{(2)} \\
&= \frac{t^2 + 3t - 2}{(t+1)(t+2)} e^{(1)} + e^{(2)} \\
&\geq \mathbf{x}^* - \left(\frac{2}{t+1} \right)^2 e^{(1)}.
\end{aligned}
$$

Let $\epsilon > 0$ be a sufficiently small error. Hence, it requires $2/\sqrt{\epsilon}$ iterations to converge to the optimal solution up to an additive error ϵ. Here, strict complementarity condition is not required to prove a finite convergence, and it is also violated in Example 5.15 of the setting of Figure 4.

Remark 5.16 (Sufficient but not necessary of strict complementarity.) In the proof of Theorem 5.14, strict complementarity is assumed to ensure that all vertices from the LMO lie in the corresponding optimal face after a specific iteration $S \in \mathbb{N}$. However, we have that $\mathbf{u}_t \in \mathcal{P}^*$ for $t \geq 1$ without strict complementarity being satisfied. Furthermore, since $(\mathbf{x}^*, \mathbf{y}^*) = \left(e^{(1)} + e^{(2)}, e^{(1)} + 2e^{(2)}\right)$ is a vertex of $\mathcal{P} \times \mathcal{Q}$, the optimal faces containing $\mathbf{x}^*$ and $\mathbf{y}^*$ are singleton sets $\mathcal{P}^* = \{\mathbf{x}^*\}$ and $\mathcal{Q}^* = \{\mathbf{y}^*\}$, respectively. Consequently, for

$\mathbf{u} = -e^{(1)} + e^{(2)}$, it follows that $\langle \nabla_{\mathbf{x}} f(\mathbf{x}^*, \mathbf{y}^*), \mathbf{u} - \mathbf{x}^* \rangle = \langle e^{(2)}, 2e^{(1)} \rangle = 0$, but $\mathbf{u} \notin \mathcal{P}^*$, thereby violating strict complementarity. Hence, strict complementarity is sufficient but not necessary to ensure that all vertices lie in the optimal face after a certain iteration, remaining a discussion.

Optimal solution - a pair consisting of a vertex and a point in the relative interior

In this section, we address the latter scenario involving two disjoint polytopes, where the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is a pair of points, consisting of a point contained in at least one-dimensional optimal face of $\mathcal{P}$ and a vertex contained in a zero-dimensional optimal face, see Figure 1f for a graphical representation. The lemma below assumes strict complementarity and demonstrates that after a certain iterate $S \in \mathbb{N}$, all vertices $\mathbf{u}_t \in \mathcal{P}$ and $\mathbf{v}_t \in \mathcal{Q}$ from `ALM` with open-loop step-sizes lie in the corresponding optimal face $\mathcal{P}^*$ of $\mathcal{P}$ and $\mathcal{Q}^*$ of $\mathcal{Q}$, respectively, a variation of Lemma 5.8.

Lemma 5.17 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes of diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$ and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set. Suppose that there exists $\kappa_{\mathcal{P}}, \kappa_{\mathcal{Q}} > 0$ such that Assumption 5.13 is satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \max\left\{\left\lceil 4\tau\ell^2 D_{\mathcal{P}}^2/(\nu^2\kappa_{\mathcal{P}}^2)\right\rceil, \left\lceil 4\tau\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\right\rceil\right\}$ where $\tau = 2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2$. Then, for the iterates of Algorithm 3 with step-size η_t, it holds that $\mathbf{u}_t \in \operatorname{vert}(\mathcal{P}^*)$ and $\mathbf{v}_t \in \operatorname{vert}(\mathcal{Q}^*)$ for all $t \geq S$.*

Proof The proof closely follows the proof of Lemma the proof of Lemma 5.8 but is adapted for integration with respect to $\mathbf{x} \in \mathcal{P}$. Since a convex polytope is a compact convex set with specific geometries, it allows us to exploit Lemma 5.8. Clearly, $\mathbf{v}_t \in \operatorname{vert}(\mathcal{Q}^*)$ for all $t \geq S_1$ where $S_1 = \left\lceil 4\tau\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\right\rceil$ by Lemma 5.8. We will now extend this result by establishing the existence of $S_2 \in \mathbb{N}$ such that for all $t \geq S_2$, $\mathbf{u}_t \in \operatorname{vert}(\mathcal{P}^*)$.

For every iteration $t \in \mathbb{N}$, we denote $(\mathbf{x}_t^*, \mathbf{y}_t^*) \in \arg\min_{(\mathbf{z},\mathbf{w}) \in \Omega^*} \left\| \begin{bmatrix} \mathbf{x}_t - \mathbf{z} \\ \mathbf{y}_t - \mathbf{w} \end{bmatrix} \right\|_2$ the pair of points in Ω^* closest to the iterate $(\mathbf{x}_t, \mathbf{y}_t)$. For $\mathbf{u} \in \operatorname{vert}(\mathcal{P})$, we write

$$\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{u} - \mathbf{x}_t^* \rangle = \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{u} - \mathbf{x}_t^* \rangle + \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{u} - \mathbf{x}_t^* \rangle .$$

By Cauchy-Schwarz inequality, we have

$$|\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{u} - \mathbf{y}_t^* \rangle| \leq \|\mathbf{x}_t - \mathbf{x}_t^*\|_2 \|\mathbf{u} - \mathbf{x}_t^*\|_2 \leq \left\| \begin{bmatrix} \mathbf{x}_t - \mathbf{x}_t^* \\ \mathbf{y}_t - \mathbf{y}_t^* \end{bmatrix} \right\|_2 D_{\mathcal{P}} \leq \sqrt{2} D_{\mathcal{P}} h_t^{\frac{1}{2}} < \frac{\kappa_{\mathcal{P}}}{2}$$

where the third inequality holds due to the fact that f satisfies the 1-(QG) by Lemma 4.4 and the last inequality follows from Proposition 4.8 for all $t \geq S_2$, where $S_2 = \left\lceil 4\tau\ell^2 D_{\mathcal{P}}^2/(\nu^2\kappa_{\mathcal{P}}^2)\right\rceil$. Thus, for all $t \geq S_2$, we obtain $|\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t) - \nabla_{\mathbf{x}} f(\mathbf{x}_t^*, \mathbf{y}_t^*), \mathbf{u} - \mathbf{x}_t^* \rangle| \leq \frac{\kappa_{\mathcal{P}}}{2}$ and the subsequent steps of the proof follow the same analysis as in Lemma 5.8. Thus, for all $t \geq S_2$, it holds that $\mathbf{v}_t \in \operatorname{vert}(\mathcal{Q}^*)$. Hence, we conclude that for all $t \geq S$ where $S = \max\left\{\left\lceil 4\tau\ell^2 D_{\mathcal{P}}^2/(\nu^2\kappa_{\mathcal{P}}^2)\right\rceil, \left\lceil 4\tau\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\right\rceil\right\}$, it holds that $\mathbf{u}_t \in \operatorname{vert}(P^*)$ and $\mathbf{v}_t \in \operatorname{vert}(Q^*)$. ■

In addition, we assume that $\mathbf{x}^*$ lies in the relative interior of at least the one-dimensional face $\mathcal{P}^*$ of $\mathcal{P}$, where $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ represents the solution set. Note that $\Pi_{\mathcal{P}}(\mathbf{y}^*) = \mathbf{x}^*$ for $(\mathbf{x}^*, \mathbf{y}^*) \in \Omega^*$.

Assumption 5.18 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes of diameter $D_{\mathcal{P}}, D_{\mathcal{Q}} > 0$, respectively, let $f : \mathcal{P} \times \mathcal{Q} \to \mathbb{R}$ be a convex and L-smooth function and let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set. For all $(\mathbf{x}^*, \mathbf{y}^*) \in \Omega^*$, suppose that $\mathbf{x}^* = \Pi_{\mathcal{P}}(\mathbf{y}^*)$ is contained in the relative interior of at least one-dimensional face $\mathcal{P}^*$ of $\mathcal{P}$, that is, for all $\mathbf{y}^* \in \mathcal{Q}$, there exists $\beta > 0$ such that $\emptyset \neq B_\beta(\mathbf{x}^*) \cap \operatorname{aff}(\mathcal{P}^*) \subseteq \mathcal{P}$ where $\mathbf{x}^* = \Pi_{\mathcal{P}}(\mathbf{y}^*)$.*

We require technical ingredients and generalize upon its counterpart in Wirth et al. (2023a) by replacing Hölderian error bound with quadratic growth.

Lemma 5.19 (Distance to the optimum) *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes of diameter $D_\mathcal{P}, D_\mathcal{Q} > 0$, respectively, let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$, let $\Omega^* = \{(\mathbf{x}^*, \mathbf{y}^*) \mid (\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})\}$ be the solution set. Suppose that there exists $\kappa_\mathcal{P}, \kappa_\mathcal{Q}, \beta > 0$ such that Assumption 5.13 and 5.18 are satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \max\left\{\left\lceil \tau\ell^2/(\beta^2\nu^2)\right\rceil, \left\lceil 4\tau\ell^2 D_\mathcal{P}^2/(\nu^2\kappa_\mathcal{P}^2)\right\rceil, \left\lceil 4\tau\ell^2 D_\mathcal{Q}^2/(\nu^2\kappa_\mathcal{Q}^2)\right\rceil\right\}$ where $\tau = 2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2$. Then, for the iterates of Algorithm 3 with step-size η_t, for all $t \geq S$, it holds that*

$$\left\|\mathbf{x}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t)\right\|_2 \leq \eta_t^\ell \frac{\beta}{\eta_S^\ell}$$

where $\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{z})$ is the orthogonal projection of $\mathbf{z} \in \mathbb{R}^d$ onto $\mathrm{aff}(\mathcal{P}^*)$.

Proof Let $t \geq S$. By Lemma 5.17, $\mathbf{u}_t \in \mathrm{vert}(\mathcal{P}^*)$. Hence, it holds that $\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{u}_t) = \mathbf{u}_t$, and we obtain

$$\begin{aligned}
\mathbf{x}_{t+1} - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_{t+1}) &= (1-\eta_t)\left(\mathbf{x}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t)\right) + \eta_t\left(\mathbf{u}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{u}_t)\right) \\
&= (1-\eta_t)\left(\mathbf{x}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t)\right) \\
&= \prod_{i=S}^{t}(1-\eta_i)\left(\mathbf{x}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_S)\right) \\
&\leq \frac{\eta_{t+1}^\ell}{\eta_S^\ell}\left(\mathbf{x}_t - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_S)\right) \qquad (25)
\end{aligned}$$

where the last inequality holds due to $\prod_{i=S}^{t}(1-\eta_i) \leq \prod_{i=S}^{t}(1-\nu\eta_i) \leq \left(\frac{\eta_t}{\eta_{S-1}}\right)^\ell \leq \left(\frac{\eta_{t+1}}{\eta_S}\right)^\ell$ by Lemma 4.7 for $\nu \in]0, 1]$ and Lemma A.1 with $H = \frac{\ell}{\nu}$. For every iterate $t \in \mathbb{N}$, we denote $(\mathbf{x}_t^*, \mathbf{y}_t^*) \in \arg\min_{(\mathbf{z},\mathbf{w})\in\Omega^*} \left\|\begin{bmatrix}\mathbf{x}_t - \mathbf{z} \\ \mathbf{y}_t - \mathbf{w}\end{bmatrix}\right\|_2$ the pair of points in Ω^* closest to the iterate $(\mathbf{x}_t, \mathbf{y}_t)$. Since f satisfies the 1-(QG) by Lemma 4.4, we obtain for all $t \geq S$,

$$\|\mathbf{x}_t - \mathbf{x}_t^*\|_2 \leq \left\|\begin{bmatrix}\mathbf{x}_t - \mathbf{x}_t^* \\ \mathbf{y}_t - \mathbf{y}_t^*\end{bmatrix}\right\|_2 \leq \sqrt{2}h_t^{\frac{1}{2}} \leq \sqrt{2}\left(\frac{(2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell^2}{2\nu^2((2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2)\ell^2/(2\nu^2)(\sqrt{2}/\beta)^2}\right)^{\frac{1}{2}} \leq \beta.$$

Thus, it holds that for all $t \geq S$,

$$\left\|\mathbf{x}_{t+1} - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_{t+1})\right\|_2 \leq \frac{\eta_{t+1}^\ell}{\eta_S^\ell}\left\|\mathbf{x}_S - \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_S)\right\|_2 \leq \frac{\eta_{t+1}^\ell}{\eta_S^\ell}\|\mathbf{x}_S - \mathbf{x}_S^*\|_2 \leq \eta_{t+1}^\ell\frac{\beta}{\eta_S^\ell}.$$

■

Using assumption 5.18, we derive the following scaling inequality. Note that the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ consists of a pair of points - $\mathbf{x}^*$ lying in at least one-dimensional optimal face $\mathcal{P}^*$ of $\mathcal{P}$ and $\mathbf{y}^*$ being a vertex in a zero-dimensional optimal face $\mathcal{Q}^*$ of $\mathcal{Q}$. This implies that the optimal solution is unique and, as $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$, we can apply Lemma 5.9 which states that the existence of a certain iterate $S \in \mathbb{N}$ such that all iterates $\mathbf{y}_t$ reach $\mathbf{y}^*$ for all $t \geq S$.

Lemma 5.20 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. Suppose that $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is unique such that $\mathbf{x}^*$ lies in the relative interior of an at least one-dimensional optimal face of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex contained a zero-dimensional optimal face of $\mathcal{Q}$. Suppose that there exists $\kappa_\mathcal{P}, \kappa_\mathcal{Q}, \beta > 0$ such that Assumption 5.13 and 5.18 are satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error and let $S = \max\left\{\left\lceil \tau\ell^2/(\beta^2\nu^2)\right\rceil, \left\lceil \tau\ell^2/(2\nu^2\kappa_\mathcal{Q})\right\rceil\right\}$ where $\tau = 2h_0 + D_\mathcal{P}^2 + D_\mathcal{Q}^2$. Then, for the iterates of Algorithm 3 with step-size η_t, for all $t \geq S$, it holds that*

$$\langle \nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t\rangle \geq \beta\left\|\Pi_S\left(\nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 \qquad (26)$$

where $\Pi_S(\mathbf{z})$ denotes the orthogonal projection of $\mathbf{z} \in \mathbb{R}^d$ onto $S = \mathrm{span}\{\mathbf{x}^* - \mathbf{x} \mid \mathbf{x} \in \mathcal{P}^*\}$.

Proof Let $t \geq S$. Since there exists $\beta > 0$ such that $\emptyset \neq B_\beta(\mathbf{x}^*) \cap \text{aff}(\mathcal{P}^*) \subseteq \mathcal{P}$ by Assumption 5.18, we let $\mathbf{z} \in B_\beta(\mathbf{x}^*) \cap \text{aff}(\mathcal{P}^*)$ via $\mathbf{z} := \mathbf{x}^* - \beta \frac{\Pi_S(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t))}{\|\Pi_S(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t))\|_2}$. Thus, by optimality of $\mathbf{u}_t$, it holds that

$$\begin{aligned}
\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t \rangle &\geq \langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + \left\langle \nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t), \beta \frac{\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)}{\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\|_2} \right\rangle \\
&\geq f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}_t) + \left\langle \Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right) + (\mathrm{I} - \Pi_S)(\nabla_{\mathbf{x}}\left(f(\mathbf{x}_t, \mathbf{y}_t)\right)), \beta \frac{\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)}{\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\|_2} \right\rangle \\
&\geq f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}_t) + \beta \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 \\
&\geq \beta \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2
\end{aligned}$$

where the second inequality holds by the convexity of f, the third inequality follows from the fact that $(\mathrm{I} - \Pi_S)(\mathbf{z})$ and $\Pi_S(\mathbf{z})$ are orthogonal for any $\mathbf{z} \in \mathbb{R}^d$, and the last inequality follows from

$$f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}_t) \geq \langle \nabla_{\mathbf{x}}(\mathbf{x}^*, \mathbf{y}_t), \mathbf{x}_t - \mathbf{x}^* \rangle = \langle \mathbf{x}^* - \mathbf{y}^*, \mathbf{x}_t - \mathbf{x}^* \rangle + \langle \mathbf{y}^* - \mathbf{y}_t, \mathbf{x}_t - \mathbf{x}^* \rangle \geq 0$$

where the last inequality holds due to $\langle \mathbf{x}^* - \mathbf{y}^*, \mathbf{x}_t - \mathbf{x}^* \rangle \geq 0$ by Dattorro (2004, Theorem E.9.1.0.2) and $\mathbf{y}_t = \mathbf{y}^*$ by Lemma 5.9. ■

We derive the final technical lemma below.

Lemma 5.21 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. Suppose that $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is unique such that $\mathbf{x}^*$ lies in the relative interior of an at least one-dimensional optimal face of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex contained a zero-dimensional optimal face of $\mathcal{Q}$. Suppose that there exists $\kappa_{\mathcal{P}}, \kappa_{\mathcal{Q}}, \beta > 0$ such that Assumption 5.13 and 5.18 are satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, let $S = \max\left\{\lceil 4\tau\ell^2/(\beta^2\nu^2)\rceil, \lceil 4\tau\ell^2 D_{\mathcal{P}}^2/(\nu^2\kappa_{\mathcal{P}}^2)\rceil, \lceil 4\tau\ell^2 D_{\mathcal{Q}}^2/(\nu^2\kappa_{\mathcal{Q}}^2)\rceil, \lceil \tau\ell^2/(2\nu^2\kappa_{\mathcal{Q}})\rceil\right\}$ where $\tau = 2h_0 + D_{\mathcal{P}}^2 + D_{\mathcal{Q}}^2$, and let $M = \max_{(\mathbf{x},\mathbf{y}) \in \mathcal{P} \times \mathcal{Q}} \|\mathbf{x} - \mathbf{y}\|_2$. Then, for the iterates of Algorithm 3 with step-size η_t, for all $t \geq S$, it holds that $h_t \leq \eta_t^\ell \frac{\beta M}{\eta_S^\ell}$ or*

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 \geq \sqrt{2}\left(h_t^{\frac{1}{2}} - \eta_t^{\frac{\ell}{2}} \frac{\sqrt{\beta M}}{\eta_S^{\ell/2}} - \eta_t^\ell \frac{\beta}{\sqrt{2}\eta_S^\ell}\right)$$

where $\Pi_S(\mathbf{z})$ denotes the orthogonal projection of $\mathbf{z} \in \mathbb{R}^d$ onto $S = \text{span}\{\mathbf{x}^* - \mathbf{x} \mid \mathbf{x} \in \mathcal{P}^*\}$.

Proof Let $t \geq S$. By definition of $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$ and by Lemma 5.19, it holds that

$$\left\|\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}^*) - \nabla_{\mathbf{x}} f(\Pi_{\text{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right\|_2 = \left\|\mathbf{x}_t - \Pi_{\text{aff}(\mathcal{P}^*)}(\mathbf{x}_t)\right\|_2 \leq \eta_t^\ell \frac{\beta}{\eta_S^\ell}. \tag{27}$$

Since for any $\mathbf{z} \in \mathbb{R}^d$, we have $\|\Pi_S(\mathbf{z})\|_2 \leq \|\mathbf{z}\|_2$. Then, inequality (27) implies

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}^*)\right) - \Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\text{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 \leq \eta_t^\ell \frac{\beta}{\eta_S^\ell}.$$

By the triangle inequality, it holds that

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\text{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 - \eta_t^\ell \frac{\beta}{\eta_S^\ell} \leq \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}^*)\right)\right\|_2. \tag{28}$$

Again, by triangle inequality and by definition of $\nabla_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$, we obtain

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}^*)\right)\right\|_2 = \left\|\Pi_S\left(\mathbf{x}_t - \mathbf{y}^*\right)\right\| \leq \left\|\Pi_S\left(\mathbf{x}_t - \mathbf{y}_t\right)\right\|_2 + \left\|\Pi_S\left(\mathbf{y}_t - \mathbf{y}^*\right)\right\|_2 = \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 + \left\|\Pi_S\left(\mathbf{y}_t - \mathbf{y}^*\right)\right\|_2.$$

Combining this bound with (28) yields

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\text{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 - \eta_t^\ell \frac{\beta}{\eta_S^\ell} \leq \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 + \left\|\Pi_S\left(\mathbf{y}_t - \mathbf{y}^*\right)\right\|_2 = \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\mathbf{x}_t, \mathbf{y}_t)\right)\right\|_2 \tag{29}$$

where the last inequality holds due to $\|\Pi_S(\mathbf{y}_t - \mathbf{y}^*)\|_2 = 0$ by Lemma 5.9.

For the remainder of the proof, we bound $\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\|_2$ below. Define $g : \mathcal{P} \cap B_\beta(\mathbf{x}^*) \to \mathbb{R}$ via $g(\mathbf{x}) := f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}), \mathbf{y}^*)$ and the gradient of g at $\mathbf{x} \in \mathcal{P} \cap B_\beta(\mathbf{x}^*)$ is $\nabla g(\mathbf{x}) = \Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}), \mathbf{y}^*)\right)$. Note that $g(\mathbf{x}) = f(\mathbf{x}, \mathbf{y}^*)$ for all $\mathbf{x} \in \mathrm{aff}(\mathcal{P}^*) \cap B_\beta(\mathbf{x}^*)$, and as f satisfies the (PL) by Lemma 4.1, g satisfies the (PL) for all $\mathbf{x} \in \mathrm{aff}(\mathcal{P}^*) \cap B_\beta(\mathbf{x}^*)$. Therefore, since a projection is idempotent and $\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t) \in \mathrm{aff}(\mathcal{P}^*) \cap B_\beta(\mathbf{x}^*)$ for all $t \geq S$, we obtain

$$\begin{aligned}
\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 &= \left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi^2_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 \\
&= \left\|\nabla g(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t))\right\|_2 \\
&\geq \sqrt{2}\left(g(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t)) - g(\mathbf{x}^*)\right)^{\frac{1}{2}} \\
&= \sqrt{2}\left(f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*) - f(\mathbf{x}^*, \mathbf{y}^*)\right)^{\frac{1}{2}} \\
&= \sqrt{2}\left(h_t + f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*) - f(\mathbf{x}_t, \mathbf{y}_t)\right)^{\frac{1}{2}} \\
&\geq \sqrt{2}\left(h_t + \left\langle \nabla f(\mathbf{x}_t, \mathbf{y}_t), \begin{bmatrix} \Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t) - \mathbf{x}_t \\ \mathbf{y}^* - \mathbf{y}_t \end{bmatrix} \right\rangle\right)^{\frac{1}{2}} \\
&\geq \sqrt{2}\left(h_t - \|\nabla f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \sqrt{\left\|\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t) - \mathbf{x}_t\right\|_2^2 + \|\mathbf{y}^* - \mathbf{y}_t\|_2^2}\right)^{\frac{1}{2}}
\end{aligned}$$

where the first inequality holds due to the (PL) and the second inequality follows from convexity of f, and the last holds due to Cauchy-Schwarz inequality. By Lemma 5.9 and 5.21, we obtain $\|\mathbf{y}_t - \mathbf{y}^*\|_2 = 0$ and $\left\|\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t) - \mathbf{x}_t\right\|_2 \leq \eta_t^\ell \frac{\beta}{\eta_S^\ell}$, respectively, resulting in $h_t - \|\nabla f(\mathbf{x}_t, \mathbf{y}_t)\|_2 \sqrt{\left\|\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t) - \mathbf{x}_t\right\|_2^2 + \|\mathbf{y}^* - \mathbf{y}_t\|_2^2} \geq h_t - \eta_t^\ell \frac{\beta M}{\eta_S^\ell}$. Suppose that $h_t \geq \eta_t^\ell \frac{\beta M}{\eta_S^\ell}$. Then, it holds that

$$\left\|\Pi_S\left(\nabla_{\mathbf{x}} f(\Pi_{\mathrm{aff}(\mathcal{P}^*)}(\mathbf{x}_t), \mathbf{y}^*)\right)\right\|_2 \geq \sqrt{2}\left(h_t - \eta_t^\ell \frac{\beta M}{\eta_S^\ell}\right)^{\frac{1}{2}} \geq \sqrt{2}\left(h_t^{\frac{1}{2}} - \eta_t^{\frac{\ell}{2}} \frac{\sqrt{\beta M}}{\eta_S^{\ell/2}}\right)$$

where the last inequality follows from the fact that $\sqrt{a-b} \geq \sqrt{a} - \sqrt{b}$ holds for nonnegative $a, b \in \mathbb{R}$. Plugging this bound into (29) concludes the claim. ■

We require the following corollary, a modification of Lemma 5.4 with an additional assumption.

Corollary 5.22 *Let $\nu \in]0, 1]$, let $\eta_t = \frac{\ell}{t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $\epsilon \in [0, \ell]$, and let $S \in \mathbb{N}$. Suppose that there exists $A, B, C, E > 0$, a nonnegative sequence $\{C_t\}_{t=S}^\infty$ such that $0 \leq C_t \leq C$, and the sequence $\{h_t\}_{t=S}^\infty$ satisfies*

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t - \eta_t A C_t h_t^{\frac{1}{2}} + \eta_t^2 B C_t \tag{30}$$

when $h_t \geq \eta_t^\ell \frac{E}{\eta_S^\ell}$, otherwise $h_t \leq \eta_t^\ell \frac{E}{\eta_S^\ell}$ for all $t \geq S$. Then, for all $t \geq S$, it holds that

$$h_t \leq \max\left\{\left(\frac{\eta_{t-1}}{\eta_{S-1}}\right)^{\ell-\epsilon} \exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor \frac{\ell}{\nu} \rfloor - \ell}\right) h_S, \eta_{t-1}^2\left(\frac{B^2}{A^2} + BC\right), \eta_{t-1}^2\left(\frac{E}{\eta_S^2} + BC\right)\right\} = \mathcal{O}(t^{-\ell+\epsilon} + t^{-2}).$$

Proof The base case $t = S$ is clear. We extend Lemma 5.4 by proving that the case $h_t \leq \eta_t^\ell \frac{E}{\eta_S^\ell}$ satisfies the induction assumption.

Suppose that $h_t \leq \eta_t^\ell \frac{E}{\eta_S^\ell}$. Plugging this bound into (30), we obtain

$$h_{t+1} \leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\nu\eta_t\right) h_t + \eta_t^2 BC \leq h_t + \eta_t^2 BC \leq \eta_t^\ell \frac{E}{\eta_S^\ell} + \eta_t^2 BC \leq \eta_t^2\left(\frac{E}{\eta_S^2} + BC\right).$$

The remainder of the proof of the case $h_t \geq \eta_t^\ell \frac{E}{\eta_S^\ell}$ is the same analysis as in Lemma 5.4. ■

Finally, we prove our last presentation of a two disjointed polytope setting.

Theorem 5.23 *Let $\mathcal{P} \subseteq \mathbb{R}^d$ and $\mathcal{Q} \subseteq \mathbb{R}^d$ be polytopes, and let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ over $\mathcal{P} \times \mathcal{Q}$. Suppose that $(\mathbf{x}^*, \mathbf{y}^*) \in \arg\min_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}} f(\mathbf{x}, \mathbf{y})$ is unique such that $\mathbf{x}^*$ lies in the relative interior of an at least one-dimensional optimal face of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex contained a zero-dimensional optimal face of $\mathcal{Q}$. Suppose that there exists $\kappa_\mathcal{P}, \kappa_\mathcal{Q}, \beta > 0$ such that Assumption 5.13 and 5.18 are satisfied. Let $\nu \in]0, 1]$ be an approximation parameter, let $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \eta_t = \frac{\ell}{\nu t + \ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$, let $h_0 = f(\mathbf{x}_0, \mathbf{y}_0) - f(\mathbf{x}^*, \mathbf{y}^*)$ be the initial error, and let $S = \max\left\{\left\lceil 4\tau\ell^2/(\beta^2\nu^2)\right\rceil, \left\lceil 4\tau\ell^2 D_\mathcal{P}^2/(\nu^2\kappa_\mathcal{P}^2)\right\rceil, \left\lceil 4\tau\ell^2 D_\mathcal{Q}^2/(\nu^2\kappa_\mathcal{Q}^2)\right\rceil, \left\lceil \tau\ell^2/(2\nu^2\kappa_\mathcal{Q})\right\rceil\right\}$, and let $M = \max_{(\mathbf{x},\mathbf{y})\in\mathcal{P}\times\mathcal{Q}}\|\mathbf{x} - \mathbf{y}\|_2$. Then, for the iterates of Algorithm 3 with step-size η_t, for all $t \geq S$, it holds that*

$$h_t \leq \max\left\{\eta_{t-1}^{\ell-\epsilon}\exp\left(\frac{\epsilon(\ell+1)}{S + \lfloor\frac{\ell}{\nu}\rfloor - \ell}\right)\frac{h_S}{\eta_{S-1}^{\ell-\epsilon}}, \eta_{t-1}^2\left(\frac{\rho^2\ell^2}{2\beta^2\epsilon^2} + \rho\right), \eta_{t-1}^2\left(\frac{\beta M}{\eta_S^2} + \rho\right)\right\} = \mathcal{O}(t^{-\ell+\epsilon} + t^{-2}).$$

where $\rho = \frac{\beta\epsilon(\sqrt{2\beta M}+\beta)}{\ell\eta_S} + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}$.

Proof Let $t \geq S$. Suppose that $h_t \geq \eta_t^\ell \frac{\beta M}{\eta_S^\ell}$. Then, combined Lemma 5.20 with 5.21, it holds that

$$g_\mathbf{x}(\mathbf{x}_t, \mathbf{y}_t) + g_\mathbf{y}(\mathbf{x}_{t+1}, \mathbf{y}_t) \geq \langle\nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}_t - \mathbf{u}_t\rangle \geq \beta\left(\|\Pi_S\left(\nabla_\mathbf{x} f(\mathbf{x}_t, \mathbf{y}_t)\right)\|_2\right) \geq \sqrt{2}\beta\left(h_t^{\frac{1}{2}} - \eta_t^{\frac{\ell}{2}}\frac{\sqrt{\beta M}}{\eta_S^{\ell/2}} - \eta_t^\ell\frac{\beta}{\sqrt{2}\eta_S^\ell}\right),$$

otherwise we have $h_t \leq \eta_t^\ell \frac{\beta M}{\eta_S^\ell}$. Plugging this bound with $\|\mathbf{x}_t - \mathbf{u}_t\|_2 \leq D_\mathcal{P}$ and $\|\mathbf{y}_t - \mathbf{v}_t\|_2 \leq D_\mathcal{Q}$ into Lemma 4.6 yields

$$\begin{aligned} h_{t+1} &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\eta_t\right)h_t - \eta_t\frac{\sqrt{2}\beta\epsilon}{\ell}\left(h_t^{\frac{1}{2}} - \eta_t^{\frac{\ell}{2}}\frac{\sqrt{\beta M}}{\eta_S^{\ell/2}} - \eta_t^\ell\frac{\beta}{\sqrt{2}\eta_S^\ell}\right) + \frac{\eta_t^2}{2}(D_\mathcal{P}^2 + D_\mathcal{Q}^2) \\ &\leq \left(1 - \left(1 - \frac{\epsilon}{\ell}\right)\eta_t\right)h_t - \eta_t\frac{\sqrt{2}\beta\epsilon}{\ell}h_t^{\frac{1}{2}} + \eta_t^2\left(\frac{\beta\epsilon(\sqrt{2\beta M}+\beta)}{\ell\eta_S} + \frac{D_\mathcal{P}^2 + D_\mathcal{Q}^2}{2}\right). \end{aligned}$$

Then, applying Corollary 5.22 with $A = \frac{\sqrt{2}\beta\epsilon}{\ell}, B = \frac{\beta\epsilon(\sqrt{2\beta M}+\beta)}{\ell\eta_S} + \frac{D_\mathcal{P}^2+D_\mathcal{Q}^2}{2}, C = 1, C_t = 1$ and $E = \beta M$ concludes the claim. ■

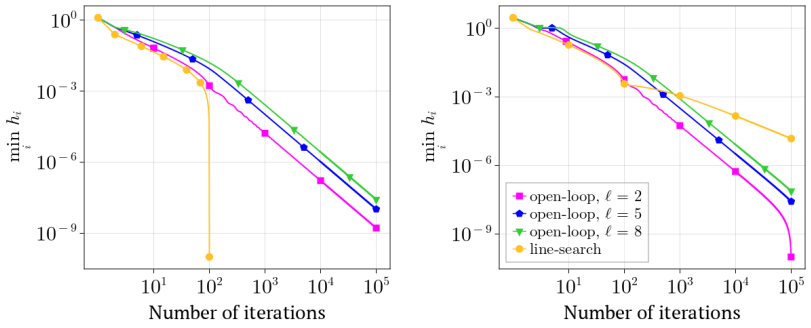
We revisit the standard basis vectors $\{e^{(1)}, \ldots, e^{(d)}\}$ in $\mathbb{R}^d$. The probability simplex is defined as

$$\triangle_d = \left\{\mathbf{z} \in \mathbb{R}^d \mid \langle\mathbf{z}, \mathbb{1}\rangle = 1, \mathbf{z} \geq 0\right\} = \mathrm{conv}(e^{(1)}, \ldots, e^{(d)}),$$

the convex combination of all standard basis vectors. Then, the probability simplex is a $(d-1)$-dimensional face of the d-dimensional unit simplex, where each vertex corresponds to a probability distribution. In the following example, we investigate scenarios involving two unit simplexes, representing a two polytopes case.

Example 5.24 (Polytopes setting) Let $\rho > 0$, let $\mathcal{P} = \left\{\mathbf{x} \in \mathbb{R}^d \mid \langle\mathbf{x}, \mathbb{1}\rangle \leq \rho, \mathbf{x} \geq 0\right\}$ be an unit simplex scaled by ρ, and let $\mathcal{Q} = \left\{\mathbf{y} \in \mathbb{R}^d \mid \langle\mathbf{y} - \mathbb{1}, \mathbb{1}\rangle \leq \frac{\rho}{2}, \mathbf{y} \geq 0\right\}$ be an unit simplex shifted by $\mathbb{1}$ and scaled by $\frac{\rho}{2}$. Let $f(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$, and let $(\mathbf{x}^*, \mathbf{y}^*) \in \mathcal{P} \times \mathcal{Q}$ be a pair of points with minimal distance between $\mathcal{P}$ and $\mathcal{Q}$. Then, $(\mathbf{x}^*, \mathbf{y}^*) = (\frac{\rho}{d}\mathbb{1}, \mathbb{1})$, where $\mathbf{x}^*$ lies in the relative interior of an optimal face of $\mathcal{P}$ and $\mathbf{y}^*$ is a vertex of $\mathcal{Q}$. Specifically, $\mathbf{x}^*$ lies in the relative interior of the $(d-1)$-dimensional probability simplex scaled by ρ, due to $\langle\mathbf{x}^*, \mathbb{1}\rangle = \langle\frac{\rho}{d}\mathbb{1}, \mathbb{1}\rangle = \rho$.

Experimental settings of Figure 5. In the setting of Example 5.24, for $d = 100$, we compare `ALM` with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t + \ell}$ for $\ell \in \{2, 5, 8\}$, starting with $\mathbf{x}_0 = e^{(1)}$ and $\mathbf{y}_0 = \mathbb{1} + \frac{\rho}{2}(e^{(1)} + e^{(2)})$. We choose $\nu = 1$ and $\rho \in \left\{1, \frac{7}{4}\right\}$ for the cases of Figure 5a and Figure 5b, respectively. The results of the experiments are plotted in log-log plots in Figure 5.



(a) $\rho = 1$. (b) $\rho = \frac{7}{4}$.

Figure 5: Convergence rate comparison of ALM with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell \in \{2, 5, 8\}$ with approximation error $\nu = 1$. For $d = 100$, the feasible regions $\mathcal{P} \subseteq \mathbb{R}^d$ is the unit simplex scaled by ρ, and $\mathcal{Q} \subseteq \mathbb{R}^d$ is the unit simplex shifted by $\mathbb{1}$ and scaled by $\frac{\rho}{2}$, where $\rho \in \left\{1, \frac{7}{4}\right\}$.

The results in Figure 5 demonstrate that ALM with open-loop step-sizes converges at a rate of order $\mathcal{O}(1/t^2)$ in both settings, whereas ALM with line-search converges either linearly or at a rate of order $\mathcal{O}(1/t)$. Note that the optimal solution $(\mathbf{x}^*, \mathbf{y}^*) = (\frac{\rho}{d}\mathbb{1}, \mathbb{1})$ is unique with the optimal function value $f(\mathbf{x}^*, \mathbf{y}^*) = \frac{1}{2}\left\|\frac{\rho}{d}\mathbb{1} - \mathbb{1}\right\|_2^2 = \frac{d}{2}\left(\frac{\rho}{d} - 1\right)^2$, and it is clear that $\mathbf{y}_t = \mathbf{y}^*$ for all $t \geq 1$. In the setting of Figure 5a, where $\rho = 1$ and $\mathbf{x}_0 = e^{(1)}$, after $t < d$ iterations, the LMO with respect to $\mathbf{x} \in \mathcal{P}$ returned at most $t+1$ out of d standard basis vectors. Thus, without loss of generality, it holds that

$$f(\mathbf{x}_t, \mathbf{y}_t) \geq \min_{\substack{\mathbf{x} \in \operatorname{conv}(S) \\ S \subseteq \{e^{(1)}, \ldots, e^{(d)}\} \\ |S| \leq t+1}} f(\mathbf{x}, \mathbb{1}) = \frac{1}{2}\left\|\sum_{i=1}^{t+1} \frac{1}{t+1} e^{(i)} - \mathbb{1}\right\|_2^2 = \frac{1}{2}\left(\left(\frac{t}{t+1}\right)^2 \left\|\sum_{i=1}^{t+1} e^{(i)}\right\|_2^2 + \left\|\sum_{i=t+2}^{d} e^{(i)}\right\|_2^2\right) = \frac{1}{2}\left(\frac{t^2}{t+1} + d - t - 1\right).$$

Thus, the primal gap after t iterations satisfies $f(\mathbf{x}_t, \mathbf{y}_t) - f(\mathbf{x}^*, \mathbf{y}^*) \geq \frac{1}{2}\left(\frac{t^2}{t+1} + d - t - 1\right) - \frac{d}{2}\left(\frac{1}{d} - 1\right)^2$, achieving optimality exactly at $t = d - 1$ iterations. The algorithm exhibits linear convergence, as depicted in Figure 5a, as late iterates all lie in the optimal face $\mathcal{P}^*$, i.e., the probability simplex. Conversely, in the setting of Figure 5b, ALM with line-search encounters the so-called zigzagging phenomenon.

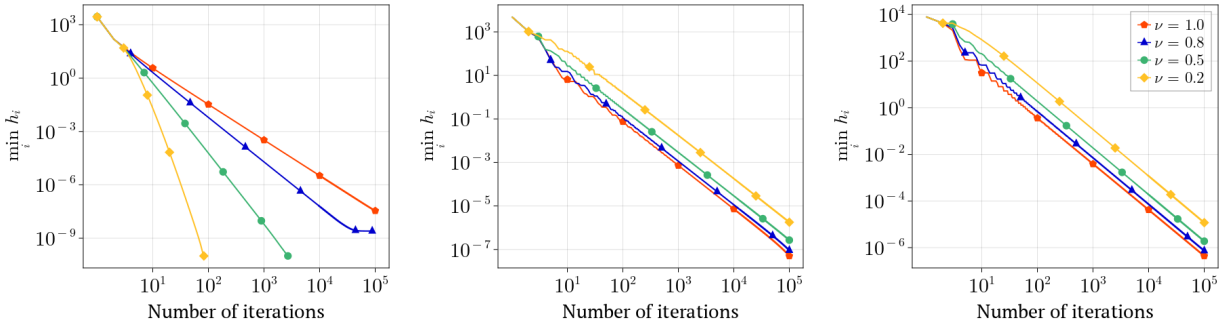
Remark 5.25 (A lower bound of LMO based algorithm.) Assuming a strongly convex objective function, Canon and Cullum (1968) proposed a lower bound $\Omega(1/t)$ of FW with line-search or short-step step-size, under the conditions where a late iterate lies outside of the optimal face and the function value at the iterate is less than the minimum value at a vertex. Consider the setting of Figure 5b, $\rho = \frac{7}{4}$, $\mathbf{x}_0 = e^{(1)}$ and $\mathbf{y}_0 = \mathbb{1} + \frac{\rho}{2}(e^{(1)} + e^{(2)})$, after mild computations, we obtain $\mathbf{u}_0 = \rho e^{(2)}$ and $\eta_{0,\mathcal{P}} = \langle \mathbf{x}_0 - \mathbf{y}_0, \mathbf{x}_0 - \mathbf{u}_0 \rangle / \|\mathbf{x}_0 - \mathbf{u}_0\|_2^2 = \frac{1}{2}$, resulting in $\mathbf{x}_1 = \frac{1}{2}\mathbf{x}_0 + \frac{1}{2}\mathbf{u}_0 = \frac{1}{2}e^{(1)} + \frac{\rho}{2}e^{(2)}$ but $\mathbf{x}_1 \notin \mathcal{P}^*$. Then, for some $\mathbf{u} \in \operatorname{vert}(\mathcal{P}) = \{\rho e^{(i)}\}_{i=1}^d$, it holds that

$$f(\mathbf{x}_1, \mathbf{y}_1) = \frac{1}{2}\left\|\frac{1}{2}e^{(1)} + \frac{\rho}{2}e^{(2)} - \mathbb{1}\right\|_2^2 = \frac{1}{2}\left(\frac{1}{4} + \left(\frac{\rho}{2} - 1\right)^2 + d - 2\right) < \frac{1}{2}\left((\rho - 1)^2 + d - 1\right) = \frac{1}{2}\|\mathbf{u} - \mathbb{1}\|_2^2 = f(\mathbf{u}, \mathbf{y}_1).$$

As $\mathbf{y}_t = \mathbb{1}$ for all $t \geq 1$, the objective function depends solely on $\mathbf{x} \in \mathcal{P}$. Thus, for all $t \geq 1$, the objective function is $f(\mathbf{x}_t, \mathbf{y}_t) = \frac{1}{2}\|\mathbf{x}_t - \mathbb{1}\|_2^2$, a strongly convex function. Consequently, by Canon and Cullum (1968), $h_t \geq \Omega(1/t)$ for infinitely many t.

We have thus demonstrated that the algorithm based on the LMO with open-loop step-sizes converges non-asymptotically faster than with line-search or short-step in specific polytope settings.

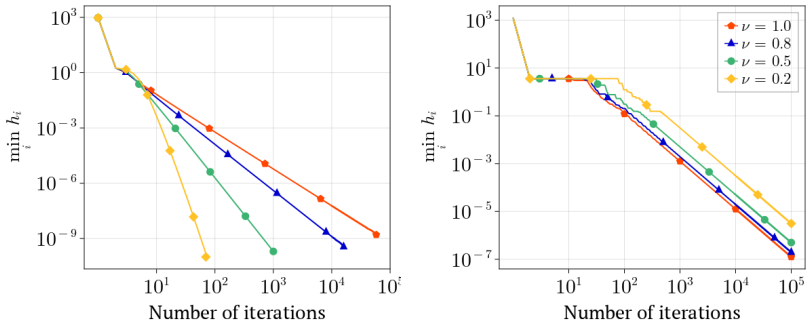
6 Ablation study for approximation error



(a) ℓ_2-balls disjoint.

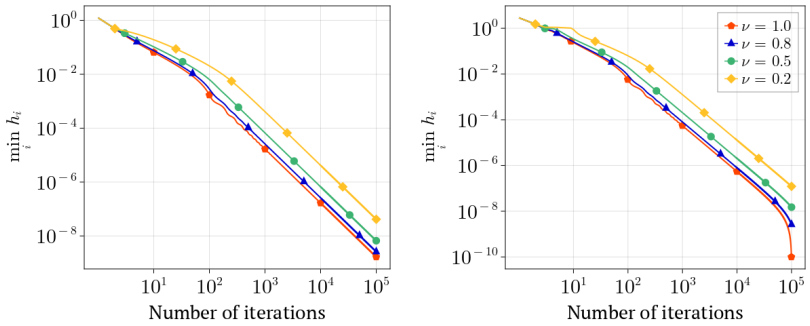
(b) ℓ_2-balls intersect at a single point.

(c) ℓ_2-balls, intersect at multiple points.



(d) ℓ_2-ball, unit simplex disjoint.

(e) ℓ_2-ball, unit simplex intersect at a single point.



(f) Unit simplices $\rho = 1$.

(g) Unit simplices $\rho = \frac{7}{4}$.

Figure 6: Convergence rate comparison of `ALM` with line-search and open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for $\ell = 2$ with approximation error $\nu \in \{1, \frac{4}{5}, \frac{1}{2}, \frac{1}{5}\}$. The comparison is performed for the settings of Figure 2, Figure 3 and Figure 5, corresponding to the first, second, and last row, respectively.

In Figure 6, we examine the impact of different values of approximation error $\nu \in]0, 1]$ in open-loop step-sizes $\frac{\ell}{\nu t+\ell}$ for $\ell \in \mathbb{N}_{\geq 2}$. The comparison is carried out for all toy examples, where the configurations of the first, second, and last row correspond to the scenarios in Figure 2, Figure 3, and Figure 5, respectively. For $\ell = 2$ and $\nu \in \{1, \frac{4}{5}, \frac{1}{2}, \frac{1}{5}\}$, the results in Figure 6a and 6d depict faster convergence rates than $\mathcal{O}(1/t^\ell)$ as demonstrated by Theorem 5.3 and 5.10. This suggests a potential gap between theory and practice. On the

other hand, the results for the remaining cases show consistent convergence rates of order $\mathcal{O}(1/t^2)$ since the convergence rate of the form of $\mathcal{O}(t^{-\ell+\epsilon} + t^{-2})$ is capped to $\mathcal{O}(1/t^2)$.

7 Discussion

Our study analyzed alternating linear minimizations (`ALM`) and established a sublinear convergence rate of $\mathcal{O}(1/t)$, which is standard for the linear minimization oracle (`LMO`) based algorithms. We explored that `ALM` with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$ with approximation error $\nu \in]0, 1]$ enjoys accelerated convergence rates in several settings. The results can be summarized in four-fold.

First, when both feasible regions are strongly convex sets, `ALM` with open-loop step-sizes $\eta_{t,\mathcal{P}} = \eta_{t,\mathcal{Q}} = \frac{\ell}{\nu t+\ell}$ for some $\ell \in \mathbb{N}_{\geq 2}$ with approximation error $\nu = 1$, demonstrates accelerations in both disjoint and intersecting cases. Specifically, when the sets are disjoint, the algorithm converges at a rate of $\mathcal{O}(1/t^\ell)$, while in intersecting sets, it achieves a rate of $\mathcal{O}(1/t^2)$. Second, quadratic improvements in the convergence rate of the form of $\mathcal{O}(1/t^2)$ are attained for `ALM` with open-loop step-sizes over a strongly convex set and a polytope with a unique optimal solution. Additionally, we established that `ALM` with open-loop step-sizes achieves faster convergence rates compared to `ALM` with line-search where two sets are disjointed polytopes in the setting of Figure 5. In this scenario, a lower bound of `LMO` based algorithm was proposed by Guélat and Marcotte (1986) under strict complementarity. Finally, our ablation study explores the impact of approximation error $\nu \in]0, 1]$ on the convergence rate of `ALM` with open-loop step-sizes. The results observed in Figure 6a and 6d indicated that `ALM` with open-loop step-sizes converges at faster rates as an approximation parameter ν decreases. These rates outperform the rate of $\mathcal{O}(1/t^\ell)$ derived in Theorem 5.3 and 5.10, posing an open question for further investigation.

In conclusion, our study has provided investigations into the acceleration of `ALM` with open-loop step-sizes when there exists a unique pair of points with minimal distance between two sets. However, open questions remain regarding accelerated results on non-unique pairs of points with minimal distance, as well as intersections in polytope settings. Additionally, while this study employed simple linear minimizations and straightforward updates, there are suggestions from the literature unexplored. This includes extensions to various variants of `LMO`-based algorithms that do not necessitate strict complementarity or depend on the location of optimal solutions. We plan to address open questions in future work.

References

Jan Harold Alcantara, Jein-Shan Chen, and Matthew K. Tam. Method of alternating projection for the absolute value equation, 2021.

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

H. Bauschke and Jonathan (Jon) Borwein. On the convergence of von neumann's alternating projection algorithm for two sets. *Set-Valued Analysis*, 1:185–212, 06 1993.

Heinz H. Bauschke and Jonathan M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996. ISSN 00361445.

Amir Beck, Edouard Pauwels, and Shoham Sabach. The cyclic block conditional gradient method for convex optimization problems. *SIAM Journal on Optimization*, 25(4):2024–2049, 2015.

Roger Behling, Yunier Bello-Cruz, and Luiz-Rafael Santos. Infeasibility and error bound imply finite convergence of alternating projections. *SIAM Journal on Optimization*, 31(4):2863–2892, 2021.

Dimitri Bertsekas. Nonlinear programming. *Athena Scientific*, 48, 01 1995.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Gábor Braun, Alejandro Carderera, Cyrille W. Combettes, Hamed Hassani, Amin Karbasi, Aryan Mokhtari, and Sebastian Pokutta. Conditional gradient methods, 2022.

Gábor Braun, Sebastian Pokutta, and Robert Weismantel. Alternating linear minimization: Revisiting von neumann's alternating projections, 2023.

Hoa T. Bui, Ryan Loxton, and Asghar Moeini. A note on the finite convergence of alternating projections. *Operations Research Letters*, 49(3):431–438, 2021. ISSN 0167-6377.

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

Alejandro Carderera, Jelena Diakonikolas, Cheuk Yin Lin, and Sebastian Pokutta. Parameter-free locally accelerated conditional gradients, 2021.

Yair Censor and Maroun Zaknoon. Algorithms and convergence results of projection methods for inconsistent feasibility problems: A review, 2018.

Ward Cheney and Allen A. Goldstein. Proximity maps for convex sets. *Proceedings of the American Mathematical Society*, 10(3):448–450, 1959. ISSN 00029939, 10886826.

Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Trans. Algorithms*, 6(4), sep 2010. ISSN 1549-6325.

Cyrille Combettes and Sebastian Pokutta. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 49, 06 2021.

Patrick L. Combettes. The convex feasibility problem in image recovery. *Advances in Imaging and Electron Physics*, 95:155–270, 1996.

P.L. Combettes and P. Bondon. Hard-constrained inconsistent signal feasibility problems. *IEEE Transactions on Signal Processing*, 47(9):2460–2468, 1999.

Jon Dattorro. *Convex Optimization & Euclidean Distance Geometry*. Meboo Publishing USA, 2004.

V. F. Demyanov and A. M. Rubinov. Approximate methods in optimization problems. *Modern Analytic and Computational Methods in Science and Mathematics*, 32, 1970.

Dmitriy Drusvyatskiy, Guoyin Li, and Henry Wolkowicz. A note on alternating projections for ill-posed semidefinite feasibility problems. *Mathematical Programming*, 162, 07 2016.

J. C. Dunn. Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals. In *1978 IEEE Conference on Decision and Control including the 17th Symposium on Adaptive Processes*, pages 915–920, 1978.

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

Aurel Galantai. *Projection Methods in Optimization*, pages 181–213. Springer New York, NY, 01 2004. ISBN 978-1-4613-4825-2.

Dan Garber. Revisiting frank-wolfe for polytopes: Strict complementarity and sparsity. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18883–18893. Curran Associates, Inc., 2020.

Dan Garber and Elad Hazan. Faster rates for the frank-wolfe method over strongly-convex sets. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 541–549. JMLR.org, 2015a.

Dan Garber and Elad Hazan. A linearly convergent conditional gradient algorithm with applications to online and stochastic optimization, 2015b.

Guillaume Garrigos. Square distance functions are polyak-łojasiewicz and vice-versa, 2023.

Omer Ginat. The method of alternating projections, 2018. PhD thesis.

L.G. Gubin, B.T. Polyak, and E.V. Raik. The method of projections for finding the common point of convex sets. *USSR Computational Mathematics and Mathematical Physics*, 7(6):1–24, 1967. ISSN 0041-5553.

J Guélat and P Marcotte. Some comments of wolfe's 'away step'. *Math. Program.*, 35(1):110–119, may 1986. ISSN 0025-5610.

Olof Hanner. On the uniform convexity of Lp and lp. *Arkiv för Matematik*, 3(3):239 – 244, 1956.

Martin Jaggi. Convex optimization without projection steps, 2011. PhD thesis.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 427–435, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Projection-free optimization on uniformly convex sets. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 19–27. PMLR, 13–15 Apr 2021a.

Thomas Kerdreux, Alexandre d'Aspremont, and Sebastian Pokutta. Restarting frank-wolfe: Faster rates under hölderian error bounds, 2021b.

Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of frank-wolfe optimization variants, 2015.

Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 53–61, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle, 2014.

E.S. Levitin and Boris Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6:1–50, 12 1966.

Adrian Lewis, Russell Luke, and Jerome Malick. Local convergence for alternating and averaged nonconvex projections, 2007.

Adrian S. Lewis and Jérôme Malick. Alternating projections on manifolds. *Mathematics of Operations Research*, 33(1):216–234, 2008. ISSN 0364765X, 15265471.

John Von Neumann. On rings of operators. reduction theory. *Annals of Mathematics*, 50(2):401–485, 1949. ISSN 0003486X.

Makoto Sakai. Strong convergence of infinite products of orthogonal projections in hilbert space. *Applicable Analysis*, 59(1-4):109–120, 1995.

John R. Silvester. Determinants of block matrices. *The Mathematical Gazette*, 84:460 – 467, 2000.

Norbert Wiener. On the factorization of matrices. *Commentarii mathematici Helvetici*, 29:97–111, 1955.

Elias Wirth, Thomas Kerdreux, and Sebastian Pokutta. Acceleration of frank-wolfe algorithms with open-loop step-sizes. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 77–100. PMLR, 25–27 Apr 2023a.

Elias Wirth, Javier Pena, and Sebastian Pokutta. Accelerated affine-invariant convergence rates of the frank-wolfe algorithm with open-loop step-sizes, 2023b.

Philip Wolfe. Finding the nearest point in a polytope. *Mathematical Programming*, 11:128–149, 1976.

Appendix A. Missing proofs

The following lemma provides an upper bound that will be used in the proof of Lemma 4.7.

Lemma A.1 *Let $H \geq 1$, and $S \in \mathbb{N}_{\geq 1}$. For $t \geq S$, it holds that $\frac{S+n}{t+1+n} \leq \frac{S+H}{t+1+H}$ for all $n \in [0, H-1]$.*

Proof Let $t \geq S$. We define a function $g(t) := \frac{t+1+n}{t+1+H}$, where $n \in [0, H-1]$. Since $g(t)$ is monotone increasing for all $t \geq 0$, it holds that $g(S-1) \leq g(t)$ for all $t \geq S$, that is, $\frac{S+n}{S+H} \leq \frac{t+1+n}{t+1+H}$. Rearranging the inequality concludes the claim. ■

A.1 Proof of Lemma 4.7

Since $\eta_i = \frac{\ell}{\nu i + \ell}$, it follows that

$$\prod_{i=S}^{t}(1-\nu\eta_i) = \prod_{i=S}^{t}\frac{i+\frac{\ell}{\nu}-\ell}{i+\frac{\ell}{\nu}} = \frac{\left(S+\frac{\ell}{\nu}-\ell\right)\cdots\left(S+\frac{\ell}{\nu}-1\right)}{\left(t+1+\frac{\ell}{\nu}-\ell\right)\cdots\left(t+\frac{\ell}{\nu}\right)} \leq \frac{\left(S+\frac{\ell}{\nu}-1\right)^{\ell}}{\left(t+\frac{\ell}{\nu}\right)^{\ell}} = \left(\frac{\nu(S-1)+\ell}{\nu t+\ell}\right)^{\ell} = \left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell}$$

where the first inequality follows from Lemma A.1 with $H = \frac{\ell}{\nu} - 1$.

A.2 Proof of Lemma 5.2

Let $\ell \in \mathbb{N}_{\geq 2}$, $\epsilon \in [0, \ell]$, $\nu \in]0, 1]$ and $\eta_i = \frac{\ell}{\nu i + \ell}$. It holds that

$$\begin{aligned}
\prod_{i=S}^{t}\left(1-\left(1-\frac{\epsilon}{\ell}\right)\nu\eta_i\right) &= \prod_{i=S}^{t}\frac{i+\frac{\ell}{\nu}-\ell+\epsilon}{i+\frac{\ell}{\nu}} \\
&= \left(\prod_{i=S}^{t}\frac{i+\frac{\ell}{\nu}-\ell}{i+\frac{\ell}{\nu}}\right)\left(\prod_{i=S}^{t}\frac{i+\frac{\ell}{\nu}-\ell+\epsilon}{i+\frac{\ell}{\nu}-\ell}\right) \\
&\leq \left(\frac{S+\frac{\ell}{\nu}-1}{t+\frac{\ell}{\nu}}\right)^{\ell}\underbrace{\prod_{i=S}^{t}\left(1+\frac{\epsilon}{i+\frac{\ell}{\nu}-\ell}\right)}_{\mathrm{A}} && \triangleright \text{ by Lemma 4.7} \\
&\leq \left(\frac{S+\frac{\ell}{\nu}-1}{t+\frac{\ell}{\nu}}\right)^{\ell}\exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\left(\frac{t+\frac{\ell}{\nu}}{S+\frac{\ell}{\nu}-1}\right)^{\epsilon} \\
&\leq \exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\left(\frac{S+\frac{\ell}{\nu}-1}{t+\frac{\ell}{\nu}}\right)^{\ell-\epsilon} \\
&= \exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\left(\frac{\eta_t}{\eta_{S-1}}\right)^{\ell-\epsilon}
\end{aligned}$$

where it holds that

$$
\begin{aligned}
A &= \prod_{i=S}^{t}\left(1+\frac{\epsilon}{i+\frac{\ell}{\nu}-\ell}\right)\\
&\le \prod_{i=S}^{t}\left(1+\frac{\epsilon}{i+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\\
&\le \exp\left(\epsilon\sum_{i=S}^{t}\frac{1}{i+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right) && \triangleright \text{ for } x\ge 0,\ 1+x\le\exp(x)\\
&\le \exp\left(\epsilon\left(\frac{1}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}+\cdots+\frac{1}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor}+\frac{1}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor+1}+\cdots+\frac{1}{t+\left\lfloor\frac{\ell}{\nu}\right\rfloor}\right)\right)\\
&\le \exp\left(\epsilon\left(\frac{\ell+1}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}+\frac{1}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor+1}+\cdots+\frac{1}{t+\left\lfloor\frac{\ell}{\nu}\right\rfloor}\right)\right)\\
&\le \exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\exp\left(\epsilon\int_{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor}^{t+\left\lfloor\frac{\ell}{\nu}\right\rfloor}\frac{1}{x}dx\right) && \triangleright \ \frac{1}{i} \text{ is monotone decreasing}\\
&\le \exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\left(\frac{t+\left\lfloor\frac{\ell}{\nu}\right\rfloor}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor}\right)^{\epsilon}\\
&\le \exp\left(\frac{\epsilon(\ell+1)}{S+\left\lfloor\frac{\ell}{\nu}\right\rfloor-\ell}\right)\left(\frac{t+\frac{\ell}{\nu}}{S+\frac{\ell}{\nu}-1}\right)^{\epsilon}. && \triangleright \text{ for any } x\in\mathbb{R},\ x-1<\lfloor x\rfloor\le x
\end{aligned}
$$