SPLITTING THE CONDITIONAL GRADIENT ALGORITHM

ZEV WOODSTOCK* AND SEBASTIAN POKUTTA† †

Abstract. We propose a novel generalization of the conditional gradient (CG / Frank-Wolfe) algorithm for minimizing a smooth function f under an intersection of compact convex sets, using a first-order oracle for ∇f and linear minimization oracles (LMOs) for the individual sets. Although this computational framework presents many advantages, there are only a small number of algorithms which require one LMO evaluation per set per iteration; furthermore, these algorithms require f to be convex. Our algorithm appears to be the first in this class which is proven to also converge in the nonconvex setting. Our approach combines a penalty method and a product-space relaxation. We show that one conditional gradient step is a sufficient subroutine for our penalty method to converge, and we provide several analytical results on the product-space relaxation's properties and connections to other problems in optimization. We prove that our average Frank-Wolfe gap converges at a rate of $\mathcal{O}(\ln t/\sqrt{t})$, – only a log factor worse than the vanilla CG algorithm with one set.

Key words. Conditional gradient, splitting, nonconvex, Frank-Wolfe, projection free

MSC codes. 46N10, 65K10, 90C25, 90C26, 90C30

1. Introduction. Given a smooth function f which maps from a real Hilbert space $\mathcal{H}$ to $\mathbb{R}$ and a finite collection of m nonempty compact convex subsets $(C_i)_{i\in I}$ of $\mathcal{H}$, we seek to solve the following:

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \bigcap_{i\in I} C_i, \tag{1.1}$$

which has many applications in imaging, signal processing, and data science [8, 11, 33]. Classical projection-based algorithms can be used to solve (1.1) if given access to the operator $\operatorname{Proj}_{\bigcap_{i\in I} C_i}$. However, in practice, computing a projection onto $\bigcap_{i\in I} C_i$ is either impossible or numerically costly, and utilizing the individual projection operators $(\operatorname{Proj}_{C_i})_{i\in I}$ is more tractable. This issue has given rise to the advent of *splitting* algorithms, which seek to solve (1.1) by utilizing operators associated with the individual sets – not their intersection. Projection-based splitting algorithms – which use the collection of operators $(\operatorname{Proj}_{C_i})_{i\in I}$ instead of $\operatorname{Proj}_{\bigcap_{i\in I} C_i}$ – have made previously-intractable problems of the form (1.1) solvable with simpler tools on a larger scale [8, 11, 12].

While splitting methods have successfully been applied to projection-based algorithms, relatively little has been done for the splitting of *conditional gradient* (CG / Frank-Wolfe) algorithms. Standard CG algorithms minimize a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$ over one closed convex constraint set $C \subset \mathbb{R}^n$. While the iterates of this algorithm do not converge in general [4], at iteration $t \in \mathbb{N}$, the average Frank-Wolfe gap (which is closely related to showing Clarke stationarity [6]) converges at a rate of $\mathcal{O}(1/\sqrt{t})$, and the primal gap converges at a rate of $\mathcal{O}(1/t)$ when f is convex [28]. A key ingredient of these algorithms is the *linear minimization oracle*, LMO_C, which computes for a linear objective $c \in \mathbb{R}^n$ a point in $\operatorname{Argmin}_{x\in C}\langle c, x\rangle$. Similarly to traditional projection-based methods, computing $\mathrm{LMO}_{\bigcap_{i\in I} C_i}$ is often prohibitively costly, so an algorithm which relies on the individual operators $(\mathrm{LMO}_{C_i})_{i\in I}$ would be more tractable.

*Department of AI in Society, Science, and Technology, Zuse Institute Berlin, Berlin, DE (woodstock@zib.de, https://zevwoodstock.github.io/).

†Institute of Mathematics, Technische Universität Berlin, Berlin, DE (pokutta@zib.de).—

In principle, if two sets C_1 and C_2 are polytopes, one could compute $\mathrm{LMO}_{C_1\cap C_2}$ by solving a linear program which incorporates the LP formulations of both C_1 and C_2. However, since the number of inequalities in an LP formulation can scale exponentially with dimension [6, 32], LPs are usually only used to implement a polyhedral LMO if there are no alternatives. In reality, many polyhedra used in applications, e.g., the Birkhoff polytope and the ℓ_1 ball, have highly specialized algorithms for computing their LMO which are faster than using a linear program [10]. Hence, splitting algorithms which rely on evaluating the specialized algorithms for $(\mathrm{LMO}_{C_i})_{i\in I}$ gain the favorable scalability of existing LMO implementations.

Conditional gradient methods have seen a resurgence in popularity since, particularly for high-dimensional settings, LMOs can be more computationally efficient than projections. For instance, a common constraint in matrix completion problems is the spectrahedron

$$S = \{x \in \mathbb{S}^n_+ \,|\, \mathrm{Trace}(x) = 1\}, \tag{1.2}$$

where $\mathbb{S}^n_+$ is the set of positive semidefinite $n \times n$ matrices. Evaluating Proj_S requires a full eigendecomposition, while computing LMO_S only requires determining a dominant eigenpair [16]. Clearly, there are high-dimensional settings where evaluating LMO_S is possible while Proj_S is too costly [10]. Thus, we are particularly motivated by high-dimensional problems in data science (e.g., cluster analysis, graph refinement, and matrix decomposition) with these LMO-advantaged constraints, e.g., the nuclear norm ball, the Birkhoff polytope of doubly stochastic matrices, and the ℓ_1 ball [6, 13, 16, 18, 30, 35, 37].

Example 1.1 (Sparse-and-low-rank decomposition). Let $\tau_1, \tau_2 > 0$, and consider the setting when $\mathcal{H} = \mathbb{R}^{n\times p}$, $m = 2$, C_1 is the ℓ_1 ball of radius τ_1, and C_2 is the nuclear norm ball of radius τ_2. This intersection describes a convex approximation of a set of simultaneously sparse and low-rank matrices, and it is used for covariance estimation, graph denoising, and link prediction (e.g., for protein interactions) [30, 18].

Inexact proximal splitting methods are a natural choice for solving (1.1) in our computational setting, since LMO-based subroutines can approximate a projection. In the convex case, this approach appears in [19, 20, 23, 26]. However, there is often no bound on the number of LMO calls required to meet the relative error tolerance required of the subroutine, e.g., in [26]. Methods which require increasingly-accurate approximations can drive the number of LMO calls in each subroutine to infinity [23], and even if a bound on the number of LMO calls exists, it often depends on the conditioning of the projection subproblem.

We are interested in algorithms with low iteration complexity, since they are more tractable on large-scale problems. It appears that, for this computational setting, the lowest iteration complexity currently requires one LMO per set per iteration [7, 18, 21, 27, 33, 36]. To the best of our knowledge, all algorithms in this class are restricted to the convex setting. The case when $f = 0$ is addressed by [7], the case when $(C_i)_{i\in I}$ have additional structure is addressed in [21], and a matrix recovery problem is addressed in [27]. The approaches in [18, 33, 36] essentially show that one CG step is a sufficient subroutine for an inexact augmented Lagrangian (AL) approach. These methods prove convergence of different optimality criteria at various rates, e.g., arbitrarily close to $\mathcal{O}(t^{-1/3})$ [33], $\mathcal{O}(1/\sqrt{t})$ [36, 21], and (under restrictions on m or $(C_i)_{i\in I}$) $\mathcal{O}(1/t)$ [7, 18, 27]. All of these methods, similarly to many projection-based splitting algorithms, achieve approximate feasibility in the sense that a point in the intersection $\bigcap_{i\in I} C_i$ is only found asymptotically.

Our contributions are as follows. We propose a new algorithm in this class for solving (1.1) which requires one LMO per set per iteration. Our algorithm generalizes the vanilla CG algorithm in the sense that, when $m = 1$, both algorithms are identical. It appears that our algorithm is the first in this class possessing convergence guarantees for solving (1.1) in the setting when f is nonconvex. As is standard in the CG literature, we analyze convergence of the average of Frank-Wolfe gaps, and we prove a rate of $\mathcal{O}(\ln t/\sqrt{t})$ – only a log factor slower than the rate for nonconvex CG over a single constraint ($m = 1$) [28]. We also prove primal gap convergence for the convex case. Our theory deviates from the AL approach and shows convergence with direct CG analysis, without imposing additional structure on our problem. By recasting (1.1) in a product space, we derive a penalized relaxation which is tractable with the vanilla CG algorithm. The convergence rates in our analysis pertain to our penalized function, which includes both primal function value and feasibility terms. At each step of our algorithm, we perform one vanilla CG step on our product space relaxation; then, we update the objective function via a penalty. We provide an analytical and geometric exploration about the properties of this subproblem as its penalty changes, as well as its relationships to (1.1) and related optimization problems. In particular, we show that for any sequence of penalty parameters which approach infinity, our subproblems converge (in several ways) to the original problem.

Our method combines two classical tools from optimization: a product-space reformulation and a penalty method. Penalty methods with CG-based subroutines received some attention several decades ago [9, 25]. Our algorithm is related the Regularized Frank-Wolfe algorithm of [25], however their requirements do not apply in our setting.

In the remainder of this section, we introduce notation, background, and standing assumptions. In Section 2, we demonstrate our product space approach and we establish analytical results. In Section 3, we introduce our algorithm and prove it converges.

1.1. Notation, standing assumptions, and auxiliary results. Let $\mathcal{H}$ be a real Hilbert space with inner product $\langle \cdot \mid \cdot \rangle$ and identity operator Id. A closed ball centered at $x \in \mathcal{H}$ of radius $\varepsilon > 0$ is denoted $B(x;\varepsilon)$. Let $m \in \mathbb{N}$, set $I = \{1, \ldots, m\}$, and let $\{\omega_i\}_{i\in I} \subset \,]0,1]$ satisfy $\sum_{i\in I} \omega_i = 1$ (e.g., $\omega_i \equiv 1/m$).

$$\text{(1.3)}\quad \boldsymbol{\mathcal{H}} = \mathcal{H}^m \text{ is the real Hilbert space with inner product } \langle \cdot \mid \cdot \rangle_{\boldsymbol{\mathcal{H}}} = \sum_{i\in I} \omega_i \langle \cdot \mid \cdot \rangle_{\mathcal{H}}.$$

Points in $\boldsymbol{\mathcal{H}}$ and their subcomponents are denoted $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2, \ldots, \boldsymbol{x}^m)$. We call $\boldsymbol{D} = \big\{\boldsymbol{x} \in \boldsymbol{\mathcal{H}} \;\big|\; \boldsymbol{x}^1 = \boldsymbol{x}^2 = \ldots = \boldsymbol{x}^m\big\}$ the *diagonal subspace* of $\boldsymbol{\mathcal{H}}$. The block averaging operation and its adjoint are

$$A\colon \boldsymbol{\mathcal{H}} \to \mathcal{H}\colon \boldsymbol{x} \mapsto \sum_{i\in I} \omega_i \boldsymbol{x}^i \quad \text{and} \quad A^*\colon \mathcal{H} \to \boldsymbol{\mathcal{H}}\colon x \mapsto (x, \ldots, x). \tag{1.4}$$

The *projection* operator onto a closed convex set $C \subset \mathcal{H}$ is denoted $\operatorname{Proj}_C\colon \mathcal{H} \to \mathcal{H}\colon x \mapsto \operatorname{Argmin}_{c\in C}\|x - c\|$. The *distance* and *indicator* functions of the set $\boldsymbol{D}$ are denoted $\operatorname{dist}_{\boldsymbol{D}}\colon \boldsymbol{\mathcal{H}} \to \mathbb{R}\colon \boldsymbol{x} \mapsto \inf_{\boldsymbol{z}\in\boldsymbol{D}} \|\boldsymbol{x} - \boldsymbol{z}\|$ and

$$\iota_{\boldsymbol{D}}\colon \boldsymbol{\mathcal{H}} \to [0,+\infty] : \boldsymbol{x} \mapsto \begin{cases} 0 & \text{if } \boldsymbol{x} \in \boldsymbol{D} \\ +\infty & \text{if } \boldsymbol{x} \notin \boldsymbol{D}, \end{cases} \tag{1.5}$$

respectively. Note that [1, Cor. 12.31] $\nabla \mathrm{dist}_{\boldsymbol{D}}^2/2 = \mathrm{Id} - \mathrm{Proj}_{\boldsymbol{D}}$, and via (1.3),

$$\frac{1}{2}\mathrm{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}) = \frac{1}{2}\sum_{i\in I}\omega_i\|A\boldsymbol{x} - \boldsymbol{x}^i\|^2, \quad \text{so} \quad A^*A = \mathrm{Id} - \nabla\left(\frac{1}{2}\mathrm{dist}_{\boldsymbol{D}}^2\right) = \mathrm{Proj}_{\boldsymbol{D}} \tag{1.6}$$

and $\|A\| \leq 1$ [1, Sec. 4]. Unless otherwise stated, let $(C_i)_{i\in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$, let $L_f > 0$, and let $f\colon \mathcal{H} \to \mathbb{R}$ be a Fréchet differentiable function which is L_f-*smooth*,

$$(\forall (x,y) \in \mathcal{H}^2) \quad f(y) - f(x) \leq \langle \nabla f(x) \mid y - x\rangle + \frac{L_f}{2}\|y - x\|^2 \tag{1.7}$$

and, when restricted to Section 3.1, also *convex*,

$$(\forall (x,y) \in \mathcal{H}^2) \quad \langle \nabla f(x) \mid y - x\rangle \leq f(y) - f(x). \tag{1.8}$$

Technically (1.7) assumes smoothness of f on $\mathcal{H}$, although this work only requires (1.7) to hold on the Minkowski sum $\sum_{i\in I}\omega_i C_i$. This assumption excludes the use of functions only defined on the interior, e.g., some logarithmic barriers.

Fact 1.2. Since $\nabla \mathrm{dist}_{\boldsymbol{D}}^2/2 = \mathrm{Id} - \mathrm{Proj}_{\boldsymbol{D}} = \mathrm{Proj}_{\boldsymbol{D}^\perp}$ is a projection operator onto a nonempty closed convex set, it is 1-Lipschitz continuous and therefore $\mathrm{dist}_{\boldsymbol{D}}^2/2$ is 1-smooth [1, Corollary 12.31, Section 4].

For every $i \in I$ and every $x \in \mathcal{H}$, the operation LMO_i returns a point in $\mathrm{Argmin}_{z\in C_i}\langle x \mid z\rangle$. The *Frank-Wolfe gap* (F-W gap) of f over a compact convex set $C \subset \mathcal{H}$ at $x \in \mathcal{H}$ is

$$G_{f,C}(x) := \sup_{v\in C}\langle \nabla f(x) \mid x - v\rangle = \langle \nabla f(x) \mid x - \mathrm{LMO}_C(\nabla f(x))\rangle. \tag{1.9}$$

Note that, for every $x \in \mathcal{H}$ [6],

$$x \text{ is a stationary point of } \underset{x\in C}{\text{minimize}}\ f(x) \quad \Leftrightarrow \quad \begin{cases} x \in C \\ G_{f,C}(x) \leq 0. \end{cases} \tag{1.10}$$

Note that if $x \in C$, we always have $G_{f,C}(x) \geq 0$.

LEMMA 1.3. *Let f and h be real-valued functions on a nonempty set $C \subset \mathcal{H}$, let $\lambda, \Delta > 0$, and suppose that*

$$x \in \underset{x\in C}{\mathrm{Argmin}}\, f(x) + \lambda h(x) \quad \textit{and} \quad z \in \underset{z\in C}{\mathrm{Argmin}}\, f(z) + (\lambda + \Delta)h(z).$$

Then $f(x) \leq f(z)$ and $h(z) \leq h(x)$.

Proof. Since x and z are optimal solutions, we have $f(x) + \lambda h(x) \leq f(z) + \lambda h(z)$ and $f(z) + (\lambda + \Delta)h(z) \leq f(x) + (\lambda + \Delta)h(x)$, so in particular,

$$(\lambda + \Delta)(h(z) - h(x)) \leq f(x) - f(z) \leq \lambda(h(z) - h(x)). \tag{1.11}$$

Subtracting $\lambda(h(z) - h(x))$ from (1.11) implies that $h(z) - h(x) \leq 0$ which, in view of (1.11), yields $f(x) - f(z) \leq 0$. ∎

We assume the ability to compute ∇f, $(\mathrm{LMO}_i)_{i\in I}$, and basic linear algebra operations, e.g., those in (1.4). Let $f\colon \mathcal{H} \to \left]-\infty, +\infty\right]$. The *subdifferential* of f at $x \in \mathcal{H}$ is given by $\partial f(x) = \{u \in \mathcal{H} \mid (\forall y \in \mathcal{H}) \quad f(x) + \langle u \mid y - x\rangle \leq f(y)\}$. The *epigraph* of f is $\mathrm{epi}\, f = \{(x,\eta) \in \mathcal{H} \times \mathbb{R} \mid f(x) \leq \eta\}$. The *graph* of an operator $M\colon \mathcal{H} \to 2^{\mathcal{H}}$ is $\mathrm{gra}\, M = \{(x,u) \in \mathcal{H}^2 \mid u \in M(x)\}$. Some of our analytical results rely on the theory of convergence of sets and set-valued operators; for a broad review, see [31].

DEFINITION 1.4. *Let $(C_n)_{n\in\mathbb{N}}$ be a sequence of subsets of $\mathbb{R}^n$, and let $(f_n)_{n\in\mathbb{N}}$ be functions on $\mathbb{R}^n$. The* outer limit *and* inner limit *of $(C_n)_{n\in\mathbb{N}}$ are*

$$\begin{aligned} &\limsup_{n\in\mathbb{N}}(C_n)_{n\in\mathbb{N}} = \big\{x\in\mathbb{R}^n \;\big|\; \lim\sup_{n\to+\infty} \operatorname{dist}_{C_n}(x) = 0\big\} \\ \text{and}\quad &\liminf_{n\in\mathbb{N}}(C_n)_{n\in\mathbb{N}} = \big\{x\in\mathbb{R}^n \;\big|\; \lim\inf_{n\to+\infty} \operatorname{dist}_{C_n}(x) = 0\big\}, \end{aligned} \tag{1.12}$$

respectively [31, Ex. 4.2]. If both limits exist and coincide, this set is the limit *of $(C_n)_{n\in\mathbb{N}}$. The sequence $(f_n)_{n\in\mathbb{N}}$* converges epigraphically *to a function f on $\mathbb{R}^n$ if the sequence of epigraphs $(\operatorname{epi} f_n)_{n\in\mathbb{N}}$ converge to $\operatorname{epi} f$. The sequence $(\partial f_n)_{n\in\mathbb{N}}$* converges graphically *to ∂f if $(\operatorname{gra}\partial f_n)_{n\in\mathbb{N}}$ converges to $\operatorname{gra}\partial f$.*

2. Splitting constraints with a product space.

This section outlines our algorithm and provides additional analysis relating our approach to similar problems in optimization.

2.1. Algorithm design.

The vanilla conditional gradient algorithm solves

$$\underset{x\in C}{\text{minimize}}\;\; f(x) \tag{2.1}$$

using LMO_C and gradients of f. However, one of the central hurdles in designing a tractable CG-based splitting algorithm is finding a way to enforce membership in the constraint $\bigcap_{i\in I} C_i$ without access to its projection or LMO. Our approach to solving this issue comes from the following construction on the product space $\boldsymbol{\mathcal{H}}$ (see Section 1.1 for notation and Fig. 1 for visualization).

PROPOSITION 2.1. *Let $(C_i)_{i\in I}$ be a collection of nonempty subsets of $\mathcal{H}$, and let $\boldsymbol{D}\subset\boldsymbol{\mathcal{H}}$ denote the diagonal subspace. Then*

$$(\forall x\in\mathcal{H})\quad (x,\ldots,x)\in \boldsymbol{D}\cap \bigtimes_{i\in I} C_i \quad\Leftrightarrow\quad x\in\bigcap_{i\in I} C_i \tag{2.2}$$

$$(\forall \boldsymbol{x}\in\boldsymbol{\mathcal{H}})\qquad \boldsymbol{x}\in \boldsymbol{D}\cap \bigtimes_{i\in I} C_i \quad\Leftrightarrow\quad (\exists\, x\in\mathcal{H})\quad \begin{cases} \boldsymbol{x} = (x,\ldots,x) \\ x\in\bigcap_{i\in I} C_i, \end{cases} \tag{2.3}$$

Proof. Clear from construction.[1] ∎

Proposition 2.1 provides a decomposition of the split feasibility constraint $\bigcap_{i\in I} C_i$ in terms of two simpler sets $\boldsymbol{D}$ and $\bigtimes_{i\in I} C_i$. This yields a product space reformulation of (1.1)

$$\underset{\boldsymbol{x}\in\bigtimes_{i\in I} C_i}{\text{minimize}}\;\; f(A\boldsymbol{x}) + \iota_{\boldsymbol{D}}(\boldsymbol{x}). \tag{2.4}$$

The constraints $\boldsymbol{D}$ and $\bigtimes_{i\in I} C_i$ are simpler in the sense that, even in our restricted computational setting, we can compute operators to enforce them. In particular, the projection onto $\boldsymbol{D}$ is computed by simply repeating the average of all components in every component $\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x} = A^*(\sum_{i\in I}\omega_i\boldsymbol{x}^i)$. Critically, this operation is cheap, so one can actually evaluate the gradient $\nabla\operatorname{dist}^2_{\boldsymbol{D}}/2 = \mathrm{Id} - \operatorname{Proj}_D$ even though it involves a projection. The constraint $\bigtimes_{i\in I} C_i$ is readily processed using the following property.

Fact 2.2. Let $(C_i)_{i\in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$. Then

$$(\forall \boldsymbol{x}\in\boldsymbol{\mathcal{H}})\quad \mathrm{LMO}_{(\bigtimes_{i\in I} C_i)}(\boldsymbol{x}) = (\mathrm{LMO}_{C_1}\boldsymbol{x}^1,\ldots,\mathrm{LMO}_{C_m}\boldsymbol{x}^m). \tag{2.5}$$

[1] The type of construction in Proposition 2.1 goes back to the work of Pierra [29].

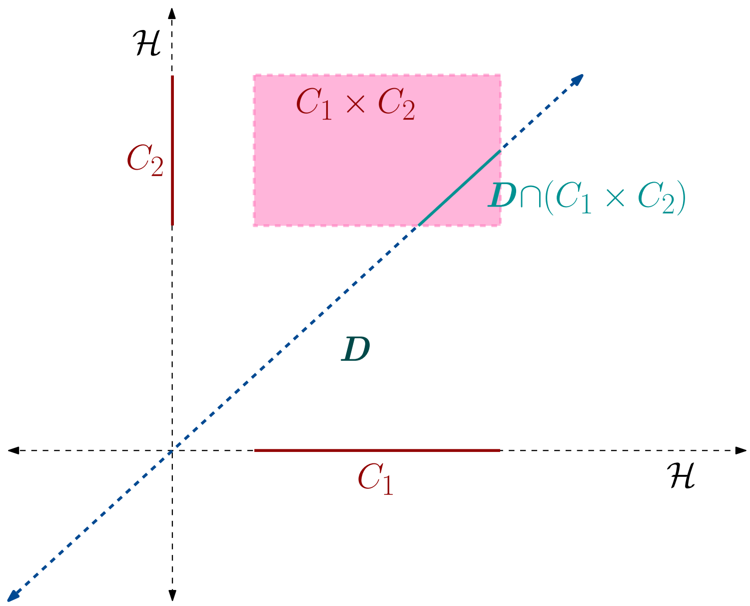

Fig. 1. *Visualization of the product space for $\mathcal{H} = \mathbb{R}$ and $m = 2$. Our algorithm produces iterates $\boldsymbol{x}_t$ which are always inside the shaded constraint set, and their averages $A^*A\boldsymbol{x}_t$ are always on the diagonal subspace $\boldsymbol{D}$. The solid segment where $C_1 \times C_2$ and $\boldsymbol{D}$ intersect corresponds precisely to our split feasibility constraint via Proposition 2.1.*

In particular, to evaluate an LMO for the product $\times_{i\in I} C_i$, it suffices to evaluate the individual operators $(\mathrm{LMO}_i)_{i\in I}$ once.

With these ideas in mind, let us introduce the penalized function

$$F_\lambda \colon \boldsymbol{\mathcal{H}} \to \mathbb{R} \colon \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \frac{1}{2} \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}), \tag{2.6}$$

which, for every $\lambda \geq 0$, is $(L_f + \lambda)$-smooth (cf. Fact 1.2). We observe that for every penalty parameter $\lambda_t \geq 0$, even under our restricted computational setting, the following relaxation of (2.4) is still tractable with the vanilla CG algorithm

$$\underset{\boldsymbol{x} \in \times_{i\in I} C_i}{\text{minimize}} \;\; F_{\lambda_t}(\boldsymbol{x}). \tag{2.7}$$

Indeed, vanilla CG requires the ability to compute the gradient of the objective function and the LMO of the constraint. Computing $\nabla F_\lambda = \nabla f + \lambda(\mathrm{Id} - \mathrm{Proj}_{\boldsymbol{D}})$ amounts to one evaluation of ∇f, computing one average, and some algebraic manipulations. By promoting membership of $\boldsymbol{D}$ via the objective function, we are left with the LMO-amenable constraint $\times_{i\in I} C_i$.

The core idea of our algorithm is, at each iteration $t \in \mathbb{N}$, to perform one Frank-Wolfe step to the relaxed subproblem (2.7). Then, between iterations, we update the objective function in (2.7) via λ_t to promote feasibility. Although (2.7) is a relaxation of the intractable problem (2.4), taking $\lambda_t \to \infty$ suffices to show convergence in F-W gap (and primal gap, in the convex case) to solutions of (2.4) and hence (1.1); this is substantiated in Sections 2.2.2 and 3. For every $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$, the ith component of the gradient is given by $\nabla F_\lambda(\boldsymbol{x})^i = \nabla f(A\boldsymbol{x}) + \lambda(\boldsymbol{x}^i - A\boldsymbol{x})$. So, a CG step applied to (2.7) yields Algorithm 2.1. While Section 3 contains the precise schedules for Lines 3 and 4, the parameters behave like $(\lambda_t, \gamma_t) = (\mathcal{O}(\ln t), \mathcal{O}(1/\sqrt{t}))$.

Algorithm 2.1 Split conditional gradient (SCG) algorithm

Require: Smooth function f, weights $\{\omega_i\}_{i\in I} \subset]0,1]$ such that $\sum_{i\in I}\omega_i = 1$, point $\boldsymbol{x}_0 \in \bigtimes_{i\in I} C_i$

1: $x_0 \leftarrow \sum_{i\in I}\omega_i \boldsymbol{x}_0^i$
2: **for** $t = 0, 1$ **to** $\ldots$ **do**
3: Choose penalty parameter $\lambda_t \in]0, +\infty[$
4: Choose step size $\gamma_t \in]0, 1]$
5: $g_t \leftarrow \nabla f(x_t)$ # Store $\nabla f(A\boldsymbol{x}_t)$ for CG step on (2.7)
6: **for** $i = 1$ **to** m **do**
7: $\boldsymbol{v}_t^i \leftarrow \mathrm{LMO}_i(g_t + \lambda_t(\boldsymbol{x}_t^i - x_t))$ # LMO applied to $\nabla F_{\lambda_t}(\boldsymbol{x}_t)^i$
8: $\boldsymbol{x}_{t+1}^i \leftarrow \boldsymbol{x}_t^i + \gamma_t(\boldsymbol{v}_t^i - \boldsymbol{x}_t^i)$ # CG step in ith component
9: **end for**
10: $x_{t+1} \leftarrow \sum_{i\in I}\omega_i \boldsymbol{x}_{t+1}^i$ # Approximate solution by averaging
11: **end for**

CG-based algorithms possess the advantage that, at every iteration, the iterates are feasible (i.e., for (2.1), $x_t \in C$). Our approach inherits this familiar property; however, since we solve a product space relaxation, $\boldsymbol{x}_t \in \bigtimes_{i\in I} C_i$ and hence, for every $i \in I$, the ith component of our sequence is feasible for the ith constraint, i.e., $(\boldsymbol{x}_t^i)_{t\in\mathbb{N}} \in C_i$. Importantly, this does not guarantee that any subcomponent $\boldsymbol{x}_t^i$ resides in $\bigcap_{i\in I} C_i$, so they are not feasible for the splitting problem (1.1); feasibility in $\bigcap_{i\in I} C_i$ is acquired "in the limit", by showing that $\boldsymbol{x}_t \in \bigtimes_{i\in I} C_i$ and $\operatorname{dist}_{\boldsymbol{D}}(\boldsymbol{x}_t) \to 0$ (proven in Section 3).

In practice, one needs a route to construct an approximate solution to (1.1) in $\mathcal{H}$ from an iterate of Algorithm 2.1 in the product space $\boldsymbol{\mathcal{H}}$. Instead of taking a component, our approximation is the average computed in Line 10, since

$$(\forall \boldsymbol{x} \in \boldsymbol{\mathcal{H}}) \qquad \boldsymbol{x} \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i \quad \Rightarrow \quad A\boldsymbol{x} \in \bigcap_{i\in I} C_i \tag{2.8}$$

is a strict implication. Hence the condition $A\boldsymbol{x} \in \bigcap_{i\in I} C_i$ is easier to satisfy than $\boldsymbol{x} \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i$ (see also Sec. 2.2.1).

Remark 2.3. If we have only $m = 1$ set constraint, then $A = \mathrm{Id}$, and $\mathcal{H} = \boldsymbol{\mathcal{H}} = \boldsymbol{D}$, so at every iteration $t \in \mathbb{N}$, $F_{\lambda_t} = f(x)$. Therefore, the classical CG algorithm is a special case of Algorithm 2.1.

Remark 2.4. The convex weights $(\omega_i)_{i\in I}$ in (1.3)–(1.4) can be used to preferentially promote membership of the approximate solution $A\boldsymbol{x}_t$ into some constraint(s) over others. If all constraints are equally important, we suggest $\omega_i \equiv 1/m$; if one constraint C_j is more important, then by selecting $\omega_j > \omega_i$ for all $i \in I \setminus \{j\}$, the weighted average $A\boldsymbol{x}_t$ is closer (in a Euclidean sense) to $\boldsymbol{x}_t^j \in C_j$ than the components of $\boldsymbol{x}_t$ in other sets $(C_i)_{i\in I\setminus\{j\}}$.

2.2. Analysis. Here we gather analytical results pertaining to our algorithm, the geometry of our product-space construction, and how our relaxed problem relates to other classical problems in optimization. While these results are interesting in their own right, many are also used to show convergence in Section 3.

2.2.1. Geometry (and tractability) of penalty functions on the Cartesian product. As seen in Section 2.1, Algorithm 2.1 promotes split feasibility by, at every iteration $t \in \mathbb{N}$, requiring that $\boldsymbol{x}_t \in \bigtimes_{i\in I} C_i$ and penalizing the distance from $\boldsymbol{x}_t$ to $\boldsymbol{D}$. However, as seen in (2.8), $\operatorname{dist}_{\boldsymbol{D}}(\boldsymbol{x}_t) = 0$ is a sufficient (but not necessary) condition to acquire a feasible average $A\boldsymbol{x}_t \in \bigcap_{i\in I} C_i$; see Fig. 2. In this section, we present a penalty function which precisely characterizes this condition. Via a simple geometric argument based on the projection theorem, we guarantee that although utilizing this penalty is not computationally tractable, it is nonetheless minimized when $\operatorname{dist}_{\boldsymbol{D}}$ vanishes. These results also further substantiate the claim that $\boldsymbol{x} \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i$ is a stricter condition than $A\boldsymbol{x} \in \bigcap_{i\in I} C_i$, which is our motivation to use the average in Line 10 of Algorithm 2.1 as our approximate solution to (1.1).

Proposition 2.5. *Let $(C_i)_{i\in I}$ be a collection of nonempty closed convex subsets of $\mathcal{H}$, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ denote the diagonal subspace, and set*

$$d\colon \boldsymbol{\mathcal{H}} \to \left]-\infty, +\infty\right] : \boldsymbol{x} \mapsto \sum_{i\in I} \omega_i \operatorname{dist}^2_{C_i}(A\boldsymbol{x}). \tag{2.9}$$

Then, for every $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$, the following are equivalent.

(i) $d(\boldsymbol{x}) = 0$.
(ii) $A\boldsymbol{x} \in \bigcap_{i\in I} C_i$.
(iii) $\operatorname{Proj}_{\boldsymbol{D}}(\boldsymbol{x}) \in \bigtimes_{i\in I} C_i$.

Proof. (i)⇒(ii): For every $i \in I$, $0 \le \omega_i \operatorname{dist}^2_{C_i}(A\boldsymbol{x}) \le d(\boldsymbol{x}) = 0$. Since $\omega_i > 0$, it follows that $\operatorname{dist}^2_{C_i}(A\boldsymbol{x}) = 0$ and hence $A\boldsymbol{x} \in C_i$.

(ii)⇒(iii): By applying A^* to the inclusion (ii), (1.6) implies that

$$\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x} = A^*A\boldsymbol{x} \in A^* \bigcap_{i\in I} C_i = \left\{ (x, \ldots, x) \in \boldsymbol{\mathcal{H}} \;\middle|\; x \in \bigcap_{i\in I} C_i \right\}. \tag{2.10}$$

So, by Proposition 2.1, $\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x} \in \left\{ \boldsymbol{x} \in \boldsymbol{\mathcal{H}} \;\middle|\; \boldsymbol{x} \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i \right\} \subset \bigtimes_{i\in I} C_i$.

(iii)⇒(i): We begin by observing that

$$\operatorname{Proj}_{\bigtimes_{i\in I} C_i}(\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x}) = \operatorname*{Argmin}_{\boldsymbol{c}\in\bigtimes_{i\in I} C_i} \|\boldsymbol{c} - A^*A\boldsymbol{x}\|^2_{\boldsymbol{\mathcal{H}}} = \operatorname*{Argmin}_{\boldsymbol{c}\in\bigtimes_{i\in I} C_i} \sum_{i\in I} \omega_i \|\boldsymbol{c}^i - A\boldsymbol{x}\|^2_{\mathcal{H}} \tag{2.11}$$

is a separable problem whose solution is $(\operatorname{Proj}_{C_1}(A\boldsymbol{x}), \ldots, \operatorname{Proj}_{C_m}(A\boldsymbol{x}))$. Therefore,

$$d(\boldsymbol{x}) = \sum_{i\in I} \omega_i \|A\boldsymbol{x} - \operatorname{Proj}_{C_i}(A\boldsymbol{x})\|^2_{\mathcal{H}} = \|\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x} - \operatorname{Proj}_{\bigtimes_{i\in I} C_i}(\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x})\|^2_{\boldsymbol{\mathcal{H}}}. \tag{2.12}$$

Since $\operatorname{Proj}_{\boldsymbol{D}}(\boldsymbol{x}) = \operatorname{Proj}_{\bigtimes_{i\in I} C_i}(\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x})$, we conclude $d(\boldsymbol{x}) = 0$. ∎

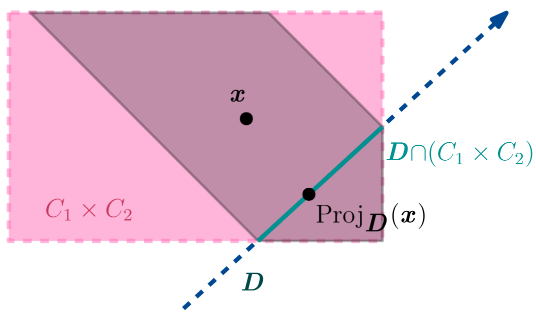

FIG. 2. *Zoomed view of Fig. 1. The darker shaded area is the collection of points* $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$ *for which* $\operatorname{Proj}_{\boldsymbol{D}}(\boldsymbol{x})$ *remains in* $\bigtimes_{i\in I} C_i$*. By Proposition 2.5, this is the set of points satisfying* $A\boldsymbol{x} \in \bigcap_{i\in I} C_i$*. This exemplifies that the implication* (2.8) *is strict.*

Since we do not assume the ability to project onto the sets $(C_i)_{i\in I}$, evaluating $\nabla d = 2\sum_{i\in I} \omega_i(\mathrm{Id} - \operatorname{Proj}_{C_i})$ is not possible. Therefore, replacing F_λ in (2.7) with the composite function $f(A\boldsymbol{x}) + \lambda d(\boldsymbol{x})$ is not tractable with a vanilla CG-based approach. However, d is closely related to our penalty function $\operatorname{dist}^2_{\boldsymbol{D}}$ via the following result.

COROLLARY 2.6. *In the setting of Proposition 2.5, let* $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$*, set* $\boldsymbol{y} = \operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x}$ *and set* $\boldsymbol{p} = \operatorname{Proj}_{(\bigtimes_{i\in I} C_i)}(\boldsymbol{y})$*. Then*

$$d(\boldsymbol{x}) = \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}) - \|\boldsymbol{x} - \boldsymbol{p}\|^2 + 2\underbrace{\langle \boldsymbol{x} - \boldsymbol{p} \mid \boldsymbol{y} - \boldsymbol{p}\rangle}_{\leq 0}. \tag{2.13}$$

In consequence, $0 \leq d(\boldsymbol{x}) \leq \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})$.

Proof. Follows from Lemma 2.12 and Theorem 3.16 of [1]. ∎

Since the iterates of Algorithm 2.1 always reside in $\bigtimes_{i\in I} C_i$, Corollary 2.6 reinforces our choice of $A\boldsymbol{x}_t$ as our approximate solution of (1.1). Firstly, its implication that $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}) = 0 \Rightarrow d(\boldsymbol{x}) = 0$ underlines the observation from (2.8) that $A\boldsymbol{x}_t \in \bigcap_{i\in I} C_i$ is easier to satisfy than $\boldsymbol{x} \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i$. Furthermore, by characterizing the gap between d and $\operatorname{dist}^2_{\boldsymbol{D}}$, we see that there are plenty of points for which the inequality between d and $\operatorname{dist}^2_{\boldsymbol{D}}$ is strict, e.g., those $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$ for which $\boldsymbol{x} \neq \boldsymbol{p}$ (see also Fig. 2). Due to this strictness, $d(\boldsymbol{x}_t)$ may vanish far *before* $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_t)$ vanishes over the iterations of Algorithm 2.1. This is consistent with our preliminary numerical observations that $A\boldsymbol{x}_t \in \bigcap_{i\in I} C_i$ often occurs before $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_t)$ vanishes.

Remark 2.7. Another natural penalty to consider is

$$g\colon \boldsymbol{x} \mapsto \operatorname{dist}^2_{\bigcap_{i\in I} C_i}(A\boldsymbol{x}) = \|\operatorname{Proj}_{\boldsymbol{D}}\boldsymbol{x} - \operatorname{Proj}_{\boldsymbol{D}\cap\bigtimes_{i\in I} C_i}\boldsymbol{x}\|^2, \tag{2.14}$$

although evaluating $\nabla g = 2A^*(\mathrm{Id} - \operatorname{Proj}_{\bigcap_{i\in I} C_i})(A\boldsymbol{x})$ involves computing an intractable projection. While, for every $\boldsymbol{x} \in \mathcal{H}$, g and d (see (2.9)) have the order

$$d(\boldsymbol{x}) = \sum_{i\in I} \omega_i \inf_{c\in C_i} \|A\boldsymbol{x} - c\|^2 \leq \sum_{i\in I} \omega_i \inf_{c\in\bigcap_{i\in I} C_i} \|A\boldsymbol{x} - c\|^2 = g(\boldsymbol{x}), \tag{2.15}$$

there is no general ordering between g and our penalty $\operatorname{dist}^2_{\boldsymbol{D}}$ for $\dim(\mathcal{H}) \geq 2$. How-

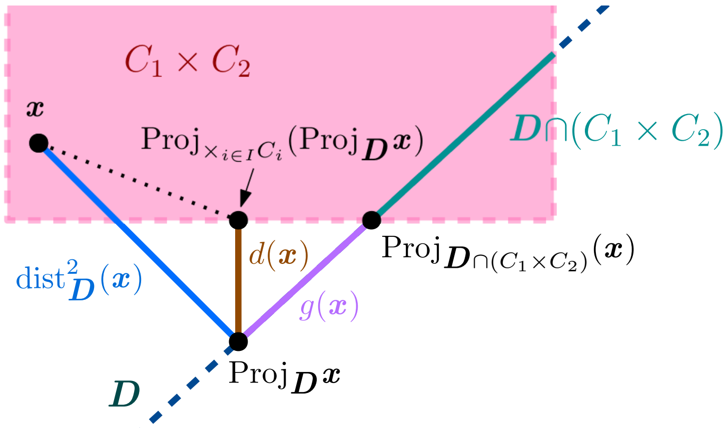

Fig. 3. *Zoomed view of Fig. 1 displaying geometric relationships between* $\mathrm{dist}^2_{\boldsymbol{D}}$*,* d*, and* g *(see* (2.12) *and Remark 2.7). This shows a feasible point* $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$ *and several vectors whose squared magnitude are equal to the given labels. Corollary 2.6 describes the relative magnitude of* $\mathrm{dist}^2_{\boldsymbol{D}}$ *and* d*, as well as the obtuse angle between* $\boldsymbol{x}$, $\mathrm{Proj}_{\times_{i\in I} C_i}(\mathrm{Proj}_{\boldsymbol{D}}\boldsymbol{x})$*, and* $\mathrm{Proj}_{\boldsymbol{D}}\boldsymbol{x}$*. Remark 2.7 describes the orthogonality seen in the angle between* $\boldsymbol{x}$, $\mathrm{Proj}_{\boldsymbol{D}}\boldsymbol{x}$*, and* $\mathrm{Proj}_{\boldsymbol{D}\cap\times_{i\in I} C_i}(\boldsymbol{x})$*. We also see* $\mathrm{Proj}_{\boldsymbol{D}\cap\times_{i\in I} C_i}\boldsymbol{x} = \mathrm{Proj}_{\boldsymbol{D}\cap\times_{i\in I} C_i}(\mathrm{Proj}_{\boldsymbol{D}}\boldsymbol{x})$*, which holds in general [1, Prop. 24.18].*

ever, using (1.3)–(1.4) reveals that they are related in the following geometric sense
(2.16)

$$\begin{aligned}\sum_{i\in I}\omega_i\|\boldsymbol{x}^i - \mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x})\|^2 &= \|\boldsymbol{x} - A^*A\boldsymbol{x} + A^*A\boldsymbol{x} - A^*\mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x})\|^2 \\ &= g(\boldsymbol{x}) + \mathrm{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}) - \underbrace{2\langle A^*\mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x}) - A^*A\boldsymbol{x} \mid \boldsymbol{x} - A^*A\boldsymbol{x}\rangle}_{=0}.\end{aligned}$$

Since the lefthand and righthand vectors in the scalar product are in $\boldsymbol{D}$ and $\boldsymbol{D}^\perp$ respectively, g and $\mathrm{dist}^2_{\boldsymbol{D}}$ describe the squared magnitude of two orthogonal vectors.

2.2.2. Interpolating constraints: From the Minkowski sum to the intersection. This section presents an analysis of how our subproblem (2.7) changes with the parameter λ. In addition to their utility in Section 3 to prove that our sequence of relaxations (2.7) actually solves the correct problem (2.4), the results in this section show that (2.7) connects two classical problems in optimization.

From a certain perspective, (2.7) "interpolates" from the following problem (when $\lambda = 0$) over the Minkowski sum

$$\underset{x\in\sum_{i\in I}\omega_i C_i}{\mathrm{minimize}}\; f(x), \quad \text{where} \quad \sum_{i\in I}\omega_i C_i = \left\{\sum_{i\in I}\omega_i c^i \;\middle|\; (\forall i\in I)\;\; c^i \in C_i\right\}, \tag{2.17}$$

to the splitting problem (1.1) (when $\lambda \nearrow +\infty$). We shall make this latter observation precise via several notions of convergence in Proposition 2.13.

Remark 2.8. While this article is predominantly focused on (1.1), it is worth noting that, when $\lambda = 0$, the problems (2.7) and (2.17) coincide in the sense that, for every solution $\boldsymbol{x}^*$ of (2.7), $A\boldsymbol{x}^*$ solves (2.17) (and for every solution $\sum_{i\in I}\omega_i x^i$ of (2.17), $(x^i)_{i\in I}$ solves (2.7)). Therefore, Fact 2.2 leads to a Frank-Wolfe approach to solving (2.17). The Minkowski sum constraint arises in Bayesian learning, placement problems, and robot motion planning [3, 14, 22, 24].

We begin with the following observations about how F_λ relates as λ varies.

LEMMA 2.9. *Let $f\colon \mathcal{H} \to \mathbb{R}$, let $\lambda, \Delta \in \mathbb{R}$, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ be nonempty, and set $F_\lambda\colon \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$. Then,*

$$(\forall \boldsymbol{x} \in \boldsymbol{\mathcal{H}}) \quad F_\lambda(\boldsymbol{x}) = F_{\lambda+\Delta}(\boldsymbol{x}) - \Delta \frac{1}{2} \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}). \tag{2.18}$$

In consequence, if $\Delta \geq 0$, then $F_\lambda(\boldsymbol{x}) \leq F_{\lambda+\Delta}(\boldsymbol{x})$ and

$$\inf F_\lambda(\bigtimes_{i\in I} C_i) \leq \inf F_{\lambda+\Delta}(\bigtimes_{i\in I} C_i).$$

Proof. $F_\lambda(\boldsymbol{x}) = f(A\boldsymbol{x}) + (\lambda+\Delta)\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2 - \Delta \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2 =$
$F_{\lambda+\Delta}(\boldsymbol{x}) - \Delta \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$. ◻

Next, we show that the optimal value of (2.7) is sandwiched between that of the splitting problem (1.1) and the Minkowski sum problem (2.17).

PROPOSITION 2.10. *Let $f\colon \mathcal{H} \to \mathbb{R}$, let $\lambda \geq 0$, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ be nonempty, set $F_\lambda\colon \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$, and let $(C_i)_{i\in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$ such that $\bigcap_{i\in I} C_i \neq \varnothing$. Then*

$$\inf_{x\in\bigcap_{i\in I} C_i} f(x) \geq \inf_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} F_\lambda \geq \inf_{x\in\sum_{i\in I}\omega_i C_i} f(x). \tag{2.19}$$

Proof. To show the first inequality, we note that for every $\boldsymbol{x} \in \boldsymbol{D}$, $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}) = 0$, so using the product space formulation (2.4) of (1.1),

$$\inf_{x\in\bigcap_{i\in I} C_i} f(x) = \inf_{\boldsymbol{x}\in\boldsymbol{D}\cap\bigtimes_{i\in I} C_i} f(A\boldsymbol{x}) + \frac{\lambda}{2}\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}) \geq \inf_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} F_\lambda(\boldsymbol{x}). \tag{2.20}$$

The second inequality follows from the observation that (2.17) coincides with $\inf_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} F_0(\boldsymbol{x})$, so by Lemma 2.9 we have $\inf_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} F_\lambda(\boldsymbol{x}) \geq \inf_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} F_0(\boldsymbol{x})$. ◻

It turns out that, for an increasing sequence of penalty parameters $(\lambda_n)_{n\in\mathbb{N}}$, the ordering of Proposition 2.10 is preserved if we only consider the optimal values of f (instead of F_λ). Intuitively, the order is reversed when we compare optimal values of the penalty $\operatorname{dist}^2_{\boldsymbol{D}}$.

COROLLARY 2.11. *Let $f\colon \mathcal{H} \to \mathbb{R}$, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ be nonempty, set $F_\lambda\colon \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$, and let $(C_i)_{i\in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$ such that $\bigcap_{i\in I} C_i \neq \varnothing$. Suppose that $(\lambda_t)_{t\in\mathbb{N}}$ is an increasing sequence of nonnegative real numbers and, for every $t \in \mathbb{N}$, let $\boldsymbol{x}_t^*$ be a minimizer of F_{λ_t} over $\bigtimes_{i\in I} C_i$. Then*

$$\inf_{x\in\bigcap_{i\in I} C_i} f(x) \geq f\left(A\boldsymbol{x}_{t+1}^*\right) \geq f\left(A\boldsymbol{x}_t^*\right) \geq \inf_{x\in\sum_{i\in I}\omega_i C_i} f(x). \tag{2.21}$$

If $\boldsymbol{z} \in \operatorname{Argmin}_{\boldsymbol{x}\in\bigtimes_{i\in I} C_i} f(A\boldsymbol{x})$ (i.e., $A\boldsymbol{z}$ solves (2.17)), then

$$0 \leq \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t+1}^*) \leq \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_t^*) \leq \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{z}). \tag{2.22}$$

Proof. Follows from Lemma 1.3 and Proposition 2.10. ◻

The following example demonstrates that the penalty sequence $(\lambda_t)_{t\in\mathbb{N}}$ may need to tend to $+\infty$ in order for the solutions of (2.7) and (1.1) to coincide.

Example 2.12. Set $\mathcal{H} = \mathbb{R}$, set $f = \|x\|^2/2$, let $z \geq 0$, set $C_1 = \{z\}$, and set $C_2 = [-z-1, z+1]$. Clearly, $z = \operatorname{Argmin}_{x \in C_1 \cap C_2} f(x)$. However, it is straightforward to verify that, for every $\lambda \geq 0$, $\boldsymbol{x}_\lambda^* = ((\lambda-1)z/(1+\lambda), z)$ is the unique minimizer of F_λ over $C_1 \times C_2$. Since $A\boldsymbol{x}_\lambda^* = \lambda z/(1+\lambda) \neq z$, the solutions of (2.7) and (1.1) (via (2.4)) do not coincide for finite λ; taking $\lambda \to +\infty$ implies $A\boldsymbol{x}_\lambda^* \to z$.

The following result establishes three notions of convergence (see Definition 1.4) relating the problems (2.7) and (1.1) (via its equivalent product space formulation (2.4)). For this result, we rely on the fact that every constrained optimization problem can be described using a single objective function via the use of indicator functions.

PROPOSITION 2.13. *Let $f \colon \mathcal{H} \to \mathbb{R}$, let $(C_i)_{i \in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$ such that $\bigcap_{i \in I} C_i \neq \varnothing$, and let $\boldsymbol{D}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$. Suppose that $(\lambda_t)_{t \in \mathbb{N}} \to +\infty$ and, for every $t \in \mathbb{N}$, set $\boldsymbol{f}_t = f \circ A + \lambda_t \operatorname{dist}_{\boldsymbol{D}}^2/2 + \iota_{\times_{i \in I} C_i}$. Then the following hold.*

(i) *$\boldsymbol{f}_t$ converges pointwise to $f \circ A + \iota_{\boldsymbol{D} \cap \times_{i \in I} C_i}$.*
(ii) *Suppose $\mathcal{H} = \mathbb{R}^n$. Then $\boldsymbol{f}_t$ converges epigraphically to $f \circ A + \iota_{\boldsymbol{D} \cap \times_{i \in I} C_i}$.*
(iii) *Suppose $\mathcal{H} = \mathbb{R}^n$ and f is convex. Then $\partial \boldsymbol{f}_n$ converges graphically to $\partial(f \circ A + \iota_{\boldsymbol{D} \cap \times_{i \in I} C_i})$.*

Proof. Since

$$\iota_{\times_{i \in I} C_i} + \iota_D = \iota_{D \cap \times_{i \in I} C_i}, \tag{2.23}$$

it suffices to show that $\lambda_t \operatorname{dist}_{\boldsymbol{D}}^2/2$ converges to ι_D under each notion of convergence.
(i): Let $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$. If $\boldsymbol{x} \in \boldsymbol{D}$, then for every $n \in \mathbb{N}$, $\lambda_t \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x})/2 = 0 = \iota_{\boldsymbol{D}}(\boldsymbol{x})$. On the other hand, if $\boldsymbol{x} \notin \boldsymbol{D}$, then $0 < \lambda_t \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x})/2 \to +\infty = \iota_D(\boldsymbol{x})$.
(ii): Let $\boldsymbol{x} \in \boldsymbol{\mathcal{H}}$. By [31, Proposition 7.2], it suffices to show both of the following.

$$\text{For some sequence } (\boldsymbol{x}_t)_{t \in \mathbb{N}} \text{ converging to } \boldsymbol{x}, \quad \limsup_{t \in \mathbb{N}} \frac{\lambda_t}{2} \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}_t) \leq \iota_{\boldsymbol{D}}(\boldsymbol{x}). \tag{2.24}$$

$$\text{For every sequence } (\boldsymbol{x}_t)_{t \in \mathbb{N}} \text{ converging to } \boldsymbol{x}, \quad \liminf_{t \in \mathbb{N}} \frac{\lambda_t}{2} \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}_t) \geq \iota_{\boldsymbol{D}}(\boldsymbol{x}). \tag{2.25}$$

To realize (2.24), we consider the constant sequence $(\boldsymbol{x}_t)_{t \in \mathbb{N}} \equiv \boldsymbol{x}$. By (i),

$$\limsup_{t \in \mathbb{N}} \lambda_t \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}_t)/2 = \lim_{t \in \mathbb{N}} \lambda_t \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x})/2 = \iota_{\boldsymbol{D}}(\boldsymbol{x}), \tag{2.26}$$

so this is always satisfied with equality. To show (2.25), let $(\boldsymbol{x}_t)_{t \in \mathbb{N}}$ be a sequence converging to $\boldsymbol{x}$. If $\boldsymbol{x} \in \boldsymbol{D}$, then since $\operatorname{dist}_{\boldsymbol{D}}^2 \geq 0$ and $\iota_{\boldsymbol{D}}(\boldsymbol{x}) = 0$, (2.25) holds. Otherwise, if $\boldsymbol{x} \notin \boldsymbol{D}$, then there exists a radius $\varepsilon > 0$ such that $B(\boldsymbol{x}; \varepsilon) \cap \boldsymbol{D} = \varnothing$. Since $\operatorname{dist}_{\boldsymbol{D}}^2$ is continuous and only vanishes on $\boldsymbol{D}$, we know $\eta := \inf_{\boldsymbol{y} \in B(\boldsymbol{x}; \varepsilon/2)} \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{y})/2 > 0$. Therefore, since $\boldsymbol{x}_t \to \boldsymbol{x}$, we have that, for some $N \in \mathbb{N}$, $n > N$ implies that $\boldsymbol{x}_t \in B(\boldsymbol{x}; \varepsilon/2)$, hence

$$\frac{\lambda_t}{2} \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}_t) \geq \lambda_t \eta \to +\infty. \tag{2.27}$$

In particular, $\lim_{t \in \mathbb{N}} \lambda_t \operatorname{dist}_{\boldsymbol{D}}^2(\boldsymbol{x}_t)/2 = +\infty = \iota_{\boldsymbol{D}}(\boldsymbol{x})$ so we are done.
(iii): Follows from (ii) and Attouch's Theorem [31, Theorem 12.35]. ∎

In general, the functions in Proposition 2.13 do not converge uniformly[2]. In spite of this, it turns out that one can nonetheless commute the limit with an infimum, hence

[2] Uniform convergence for extended-real valued functions is defined in [31].

showing that the optimal values of our subproblems (2.7) converge to the optimal value of (1.1).

PROPOSITION 2.14. *Let $f\colon \mathcal{H} \to \mathbb{R}$, let $(C_i)_{i\in I}$ be a collection of nonempty compact convex subsets of $\mathcal{H}$, let $\boldsymbol{D}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$, and for every $\lambda \geq 0$, set $F_\lambda\colon \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x})/2$. Suppose that $(\lambda_n)_{n\in\mathbb{N}} \nearrow +\infty$. Then*

$$\lim_{t\to+\infty}\left(\inf_{\boldsymbol{x}\in\times_{i\in I} C_i} F_{\lambda_t}(\boldsymbol{x})\right) \to \inf_{\boldsymbol{x}\in\times_{i\in I} C_i}\left(\lim_{t\to\infty} F_{\lambda_t}(\boldsymbol{x})\right) = \inf_{x\in\bigcap_{i\in I} C_i} f(x). \tag{2.28}$$

Proof. First, we point out that the equality in (2.28) follows from Proposition 2.13 and the fact that the minimal values of (1.1) and (2.4) coincide. Let

$$\mu < \inf_{x\in\bigcap_{i\in I} C_i} f(x) = \inf_{\boldsymbol{x}\in\times_{i\in I} C_i} f(A\boldsymbol{x}) + \iota_{\boldsymbol{D}}(\boldsymbol{x}).$$

By Proposition 2.13, for every $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$, $\lim_{t\to\infty} F_{\lambda_t}(\boldsymbol{x}) = f(A\boldsymbol{x}) + \iota_{\boldsymbol{D}}(\boldsymbol{x}) > \mu$. Since $\bigtimes_{i\in I} C_i$ is compact, for $t \in \mathbb{N}$ sufficiently large, $\inf_{\boldsymbol{x}\in\times_{i\in I} C_i} F_{\lambda_t}(\boldsymbol{x}) \geq \mu$, which implies (via Proposition 2.10 for the second inequality)

$$\mu \leq \lim_{t\to\infty}\left(\inf_{\boldsymbol{x}\in\times_{i\in I} C_i} F_{\lambda_t}(\boldsymbol{x})\right) \leq \inf_{x\in\bigcap_{i\in I} C_i} f(x). \tag{2.29}$$

Taking $\mu \uparrow \inf_{x\in\bigcap_{i\in I} C_i} f(x)$ completes the result. □

3. Convergence of Algorithm 2.1.

We first prove that Algorithm 2.1 converges in function value when f is convex (Section 3.1). Then, we establish guarantees for stationarity in general (Section 3.2). We begin with an estimate which is used for both settings.

LEMMA 3.1. *Let $(C_i)_{i\in I}$ be a finite collection of nonempty compact convex subsets of $\mathcal{H}$ with diameters $\{R_i\}_{i\in I} \subset [0, +\infty[$, and let $\boldsymbol{D}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$. Suppose that $\bigcap_{i\in I} C_i \neq \varnothing$. Then*

$$\left(\forall \boldsymbol{x}, \boldsymbol{y} \in \bigtimes_{i\in I} C_i\right) \quad \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \leq \sum_{i\in I} \omega_i R_i^2 \quad \textit{and} \quad \|\boldsymbol{x} - \boldsymbol{y}\|^2 \leq \sum_{i\in I} \omega_i R_i^2. \tag{3.1}$$

Proof. Since, for every $i \in I$, $\mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x}) \in C_i$, (2.16) yields the upper bound

$$\mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \leq \sum_{i\in I} \omega_i \|\boldsymbol{x}^i - \mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x})\|^2 \leq \sum_{i\in I} \omega_i R_i^2. \tag{3.2}$$

For the second bound, $\|\boldsymbol{x} - \boldsymbol{y}\|^2 = \sum_{i\in I} \omega_i \|\boldsymbol{x}^i - \boldsymbol{y}^i\|^2 \leq \sum_{i\in I} \omega_i R_i^2$. □

3.1. Convex setting.

Here we show that, if f is convex, Algorithm 2.1 achieves an $\mathcal{O}(\ln t/\sqrt{t})$ convergence rate in terms of the primal value gap of our subproblems (2.7). In tandem with Proposition 2.13, this establishes function value convergence. Unlike the Augmented Lagrangian approaches [18, 33, 36], our analysis does not require further assumptions concerning the relative interiors of $(C_i)_{i\in I}$, making it consistent with traditional Frank-Wolfe theory [6, Section 2.1].

LEMMA 3.2. *Let f be convex and L_f-smooth, let $\boldsymbol{D}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$, let $(C_i)_{i\in I}$ be a finite collection of nonempty compact convex subsets of $\mathcal{H}$ with diameters $\{R_i\}_{i\in I} \subset [0, +\infty[$ such that $\bigcap_{i\in I} C_i \neq \varnothing$, and for every $\lambda \geq 0$, set*

$F_\lambda \colon \boldsymbol{\mathcal{H}} \to \left]-\infty,+\infty\right] : \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$, *set* $\boldsymbol{x}_t^* \in \operatorname{Argmin}_{\boldsymbol{x}\in\times_{i\in I} C_i} F_{\lambda_t}(\boldsymbol{x})$, *and set* $H_t = F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_t}(\boldsymbol{x}_t^*)$. *Suppose that* $(\lambda_t)_{t\in\mathbb{N}}$ *is an increasing sequence. Then the iterates of Algorithm 2.1 satisfy*

$$H_{t+1} \leq (1-\gamma_t)H_t + \frac{(\lambda_{t+1}-\lambda_t)}{2}\sum_{i\in I}\omega_i R_i^2 + \gamma_t^2 \frac{(\lambda_t + L_f)}{2}\sum_{i\in I}\omega_i R_i^2. \tag{3.3}$$

Proof. Let us begin by observing that F_{λ_t} is convex and $L_f + \lambda_t$-smooth (cf. Fact 1.2). Since Algorithm 2.1 performs one step of the vanilla CG algorithm to (2.7), a standard CG argument [6] (relying on smoothness (1.7), Line 7 and Fact 2.2, then convexity (1.8)) shows

$$F_{\lambda_t}(\boldsymbol{x}_{t+1}) - F_{\lambda_t}(\boldsymbol{x}_t) \leq \gamma_t\Big(F_{\lambda_t}(\boldsymbol{x}_t^*) - F_{\lambda_t}(\boldsymbol{x}_t)\Big) + \gamma_t^2\frac{L_f+\lambda_t}{2}\sum_{i\in I}\omega_i R_i^2. \tag{3.4}$$

Using Lemma 2.9, then adding $F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_t}(\boldsymbol{x}_t^*)$ to both sides of (3.4) reveals

$$H_{t+1} \leq F_{\lambda_{t+1}}(\boldsymbol{x}_{t+1}) - F_{\lambda_t}(\boldsymbol{x}_t^*) \tag{3.5}$$

$$= F_{\lambda_t}(\boldsymbol{x}_{t+1}) - F_{\lambda_t}(\boldsymbol{x}_t^*) + \frac{\lambda_{t+1}-\lambda_t}{2}\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t+1}) \tag{3.6}$$

$$\leq (1-\gamma_t)H_t + \frac{\lambda_{t+1}-\lambda_t}{2}\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t+1}) + \gamma_t^2\frac{L_f+\lambda_t}{2}\sum_{i\in I}\omega_i R_i^2. \tag{3.7}$$

Finally, Lemma 3.1 finishes the result. ∎

THEOREM 3.3. *In the setting of Lemma 3.2, for every* $t \geq 0$ *set* $\gamma_t = 2/(\sqrt{t}+2)$. *Let* $\lambda_0 > 0$ *and for every* $t \geq 1$ *set* $\lambda_{t+1} = \lambda_t + \lambda_0(\sqrt{t}+2)^{-2}$. *Then, for every* $t \in \mathbb{N}$, *the iterates of Algorithm 2.1 satisfy*

$$0 \leq H_t \leq 2\sum_{i\in I}\omega_i R_i^2\left(\frac{\lambda_0(2\ln(\sqrt{t}+2)+\frac{1}{4}) + L_f}{\sqrt{t}+2} + \frac{4\lambda_0}{(\sqrt{t}+2)^2}\right). \tag{3.8}$$

In particular, $F_{\lambda_t}(\boldsymbol{x}_t) \to \inf_{x\in\bigcap_{i\in I}C_i} f(x)$ *and* $\operatorname{dist}_{\boldsymbol{D}}(\boldsymbol{x}_t) \to 0$. *Furthermore, every accumulation point* $\boldsymbol{x}_\infty$ *of* $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ *produces a solution* $A\boldsymbol{x}_\infty \in \bigcap_{i\in I} C_i$ *such that* $f(A\boldsymbol{x}_\infty) = \inf_{x\in\bigcap_{i\in I}C_i} f(x)$.

Proof. For notational convenience, set $R = \sum_{i\in I}\omega_i R_i^2$ and $\xi\colon \mathbb{R}\to\mathbb{R}\colon s \mapsto 2\ln(\sqrt{s}+2) + 4/(\sqrt{s}+2)$. By calculus, for every $t\in\mathbb{N}$ such that $t \geq 1$, $\lambda_t - \lambda_0 \leq \lambda_0\xi(t) - \lambda_0\xi(0)$, so $\lambda_t \leq \lambda_0\xi(t)$. We shall proceed by induction. The base case for $t = 0$ follows from (1.7), (1.10), and Lemma 3.1. Next, we suppose that (3.8) holds for $t\in\mathbb{N}$. Our inductive hypothesis, bound on λ_t, and (3.3) yield

$$H_{t+1} \leq (1-\gamma_t)\left(2R\frac{\lambda_0\xi(t) + L_f + \frac{\lambda_0}{4}}{\sqrt{t}+2}\right) + \frac{\lambda_{t+1}-\lambda_t}{2}R + \gamma_t^2\frac{(L_f+\lambda_0\xi(t))R}{2} \tag{3.9}$$

$$= \frac{\sqrt{t}}{\sqrt{t}+2}\left(2R\frac{\lambda_0\xi(t)+L_f+\frac{\lambda_0}{4}}{\sqrt{t}+2}\right) + 2R\left(\frac{L_f+\frac{\lambda_0}{4}}{(\sqrt{t}+2)^2} + \frac{\lambda_0\xi(t)}{(\sqrt{t}+2)^2}\right) \tag{3.10}$$

$$\leq \frac{\sqrt{t}+1}{(\sqrt{t}+2)^2}\big(2R(L_f + \frac{\lambda_0}{4} + \lambda_0\xi(t+1))\big) \tag{3.11}$$

$$\leq \frac{1}{\sqrt{t+1}+2}\left(2R(L_f+\frac{\lambda_0}{4}+\lambda_0\xi(t+1))\right), \tag{3.12}$$

where (3.11) is because ξ is increasing and (3.12) is because and $(\sqrt{t}+1)(\sqrt{t+1}+2) \le (\sqrt{t}+2)^2$. Having shown (3.8), we point out that Proposition 2.14 implies $\lim_{t\to\infty} F_{\lambda_t}(\boldsymbol{x}_t^*) = \inf_{x\in\bigcap_{i\in I} C_i} f(x)$. Hence $\lim_{t\to\infty} F_{\lambda_t}(\boldsymbol{x}_t)$ exists and, via (3.8), is equal to $\inf_{x\in\bigcap_{i\in I} C_i} f(x)$. Since $\lambda_t \to \infty$, it must be that $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_t) \to 0$. Therefore, every accumulation point $x_\infty \in \bigtimes_{i\in I} C_i$ must also reside in $\boldsymbol{D}$, so $A\boldsymbol{x}_\infty \in \bigcap_{i\in I} C_i$. Passing to a subsequence, since f is continuous we have

$$\inf_{x\in\bigcap_{i\in I} C_i} f(x) \le f(A\boldsymbol{x}_\infty) = \lim_{k\to\infty} f(A\boldsymbol{x}_{t_k}) \le \lim_{k\to\infty} F_{\lambda_{t_k}}(\boldsymbol{x}_{t_k}) = \inf_{x\in\bigcap_{i\in I} C_i} f(x). \quad (3.13)$$

□

Note that, although Theorem 3.3 shows convergence of the primal gaps of the subproblem (2.7), these gaps are never actually computed in practice, since $\boldsymbol{x}_t^*$ is inaccessible. We also point out that, for the choice of $\lambda_0 = L_f$, our convergence rate becomes scale-invariant.

The convergence rate in Theorem 3.3 is atypical of CG algorithms with convex objective functions, because they usually have an $\mathcal{O}(1/t)$ convergence rate. This was achieved in the split-LMO setting under the condition $m = 2$ in [7, 27] and with a Slater-type condition in [18] by choosing stepsizes of magnitude $\gamma_t = \mathcal{O}(1/t)$. However, in order to achieve convergence in the proof of Theorem 3.3 with this larger stepsize, this would necessitate that $\lambda_{t+1} - \lambda_t \le \mathcal{O}(1/t^2)$, i.e., $\lambda_t \not\to \infty$. Since Example 2.12 establishes that $\lambda_t \to \infty$ can be necessary (supported also by Proposition 2.13), we would no longer be able to show that the sequence of relaxed subproblems (2.7) converges to the original splitting problem (1.1). So, using a faster stepsize schedule would still yield a convergent algorithm, but it would not necessarily solve (1.1). We shall consider the topic of achieving a faster rate with extra assumptions in future work.

Remark 3.4. Without additional assumptions, Algorithm 2.1 does not guarantee iterate convergence of $(x_t)_{t\in\mathbb{N}}$, which is consistent with other CG methods [4]. If, for instance, f is also μ-strongly convex, then Theorem 3.3 can be strengthened to provide convergence of the averages, because $A\boldsymbol{x}_t^*$ converges to the unique solution x^* of (1.1) and $0 \le \mu\|A\boldsymbol{x}_t - A\boldsymbol{x}_t^*\|/2 \le F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_t}(\boldsymbol{x}_t^*) \to 0$, so $A\boldsymbol{x}_t \to x^*$ as well.

3.2. Nonconvex setting. For CG methods which address (2.1) in the case when f is nonconvex, it is standard to show that the Frank-Wolfe gap at $x \in \mathcal{H}$, $G_{f,C}(x) := \sup_{v\in C} \langle \nabla f(x) \mid x - v\rangle$, converges to zero, because f is stationary at $x \in C$ whenever the F-W gap vanishes (1.10) [6]. Since F-W gaps are highly variable between iterations, convergence rates are typically derived for the average of F-W gaps. In this section, we consider the F-W gaps for our subproblems (2.7) which converge to (1.1) (in the sense of Proposition 2.13).

We begin by connecting the F-W gaps of our subproblems (2.7) to that of the original problem (1.1). In particular, for every $\lambda \ge 0$ the Frank-Wolfe gaps of our subproblems at $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$ provide an upper bound to *both* the penalty $\lambda\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})$ and the F-W gap of the original problem (1.1) at $A\boldsymbol{x}$. Although $G_{F_\lambda,\times_{i\in I} C_i}(\boldsymbol{x}_t) \ge 0$ is guaranteed, it is interesting to note that the F-W gap for the splitting problem (1.1), namely $G_{f,\bigcap_{i\in I} C_i}(A\boldsymbol{x}_t)$, may actually be negative since $A\boldsymbol{x}_t$ is not guaranteed to reside in $\bigcap_{i\in I} C_i$ after a finite number of iterations.

LEMMA 3.5. *Let f be smooth, set $\beta_f = \sup_{\boldsymbol{x}\in\times_{i\in I} C_i} \|\nabla f(A\boldsymbol{x})\|$, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$, let $(C_i)_{i\in I}$ be a finite collection of nonempty compact convex subsets of $\mathcal{H}$ with diameters $\{R_i\}_{i\in I} \subset [0,+\infty[$ such that $\bigcap_{i\in I} C_i \ne \varnothing$, and for every $\lambda \ge 0$, set $F_\lambda \colon \boldsymbol{\mathcal{H}} \to]-\infty,+\infty] : \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x})/2$. Then, for* *every* $\boldsymbol{x} \in \bigtimes_{i\in I} C_i$,

$$\sup_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \langle \nabla F_\lambda(\boldsymbol{x}) \mid \boldsymbol{x} - \boldsymbol{v}\rangle \geq \sup_{v\in\bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}) \mid A\boldsymbol{x} - v\rangle + \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \geq -\beta_f \sum_{i\in I} \omega_i R_i. \tag{3.14}$$

Proof. First, by infimizing over a subset of $\bigtimes_{i\in I} C_i$, we find

$$\inf_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \langle \nabla F_\lambda(\boldsymbol{x}) \mid \boldsymbol{v} - \boldsymbol{x}\rangle = \inf_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \langle A^*\nabla f(A\boldsymbol{x}) + \lambda(\boldsymbol{x} - A^*A\boldsymbol{x}) \mid \boldsymbol{v} - \boldsymbol{x}\rangle \tag{3.15}$$

$$\leq \inf_{\boldsymbol{v}\in\boldsymbol{D}\cap\bigtimes_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}) \mid A\boldsymbol{v} - A\boldsymbol{x}\rangle + \lambda\langle \boldsymbol{x} - A^*A\boldsymbol{x} \mid \boldsymbol{v} - \boldsymbol{x}\rangle. \tag{3.16}$$

Since $\boldsymbol{x} - A^*A\boldsymbol{x} \in \boldsymbol{D}^\perp$ and $A^*A\boldsymbol{x} \in \boldsymbol{D}$, we have the following identity for every $\boldsymbol{v} \in \boldsymbol{D}$

$$\langle \boldsymbol{x} - A^*A\boldsymbol{x} \mid \boldsymbol{v} - \boldsymbol{x}\rangle = \langle \boldsymbol{x} - A^*A\boldsymbol{x} \mid -A^*A\boldsymbol{x} - (\boldsymbol{x} - A^*A\boldsymbol{x})\rangle = -\|\boldsymbol{x} - A^*A\boldsymbol{x}\|^2. \tag{3.17}$$

So, using Proposition 2.1 for a change of variables, we set $p = \mathrm{Proj}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x})$ to find that

$$\inf_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \langle \nabla F_\lambda(\boldsymbol{x}) \mid \boldsymbol{v} - \boldsymbol{x}\rangle \leq \inf_{v\in\bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}) \mid v - A\boldsymbol{x}\rangle - \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \tag{3.18}$$

$$\leq \langle \nabla f(A\boldsymbol{x}) \mid p - A\boldsymbol{x}\rangle - \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \tag{3.19}$$

$$\leq \beta_f \mathrm{dist}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x}) - \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x}) \tag{3.20}$$

$$\leq \beta_f \sum_{i\in I} \omega_i R_i, \tag{3.21}$$

since $\mathrm{dist}_{\bigcap_{i\in I} C_i}(A\boldsymbol{x}) = \|\sum_{i\in I} \omega_i(\boldsymbol{x}^i - p)\| \leq \sum_{i\in I} \omega_i R_i$. Finally, negation yields (3.14). ∎

With these results in-hand, we can now prove our main result.

THEOREM 3.6. *Let f be L_f-smooth, let $\boldsymbol{D} \subset \boldsymbol{\mathcal{H}}$ denote the diagonal subspace of $\boldsymbol{\mathcal{H}}$, let $(C_i)_{i\in I}$ be a finite collection of nonempty compact convex subsets of $\mathcal{H}$ with diameters $\{R_i\}_{i\in I} \subset [0, +\infty[$ such that $\bigcap_{i\in I} C_i \neq \varnothing$, and for every $\lambda \geq 0$, set $F_\lambda \colon \boldsymbol{\mathcal{H}} \to]-\infty, +\infty] : \boldsymbol{x} \mapsto f(A\boldsymbol{x}) + \lambda \mathrm{dist}\,^2_{\boldsymbol{D}}(\boldsymbol{x})/2$. Set $\gamma_t = 1/\sqrt{t+1}$, let $\lambda_0 > 0$, and for every $t \geq 1$, set $\lambda_t = \lambda_0 \sum_{k=0}^{t-1} 1/(k+1)$. Then, for every $t \geq 1$, the iterates of Algorithm 2.1 satisfy*[3]

$$0 \leq \frac{1}{t}\sum_{k=0}^{t-1} \sup_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \left\langle \nabla F_{\lambda_k}(\boldsymbol{x}_k) \;\middle|\; \boldsymbol{x}_k - \boldsymbol{v}\right\rangle \leq \mathcal{O}\left(\frac{\ln t}{\sqrt{t}} + \frac{1}{\sqrt{t}}\right). \tag{3.22}$$

In particular, there exists a subsequence $(t_k)_{k\in\mathbb{N}}$ such that

$$\left(\sup_{\boldsymbol{v}\in\bigtimes_{i\in I} C_i} \langle \nabla F_{\lambda_{t_k}}(\boldsymbol{x}_{t_k}) \mid \boldsymbol{x}_{t_k} - \boldsymbol{v}\rangle\right)_{k\in\mathbb{N}} \to 0. \tag{3.23}$$

Furthermore, every accumulation point $\boldsymbol{x}_\infty$ of $(\boldsymbol{x}_{t_k})_{k\in\mathbb{N}}$ yields a point in the intersection via $A\boldsymbol{x}_\infty \in \bigcap_{i\in I} C_i$ which satisfies $G_{f,\bigcap_{i\in I} C_i}(A\boldsymbol{x}_t) = 0$, i.e., $A\boldsymbol{x}_\infty$ is a stationary point of the problem (1.1).

[3] Precise constants for (3.22) are in (3.43).

Proof. For notational convenience, set $R = \sum_{i\in I}\omega_i R_i^2$, $R_A = \sum_{i\in I}\omega_i R_i$, and $B = \max\{\beta_p\sqrt{R}, R\}$; for every $t\in\mathbb{N}$, let $\boldsymbol{x}_t^*$ be a minimizer of F_{λ_t} over $\bigtimes_{i\in I} C_i$, set $H_t = F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_t}(\boldsymbol{x}_t^*)$, and set $\boldsymbol{v}_t = (\boldsymbol{v}_t^i)_{i\in I}\in\bigtimes_{i\in I} C_i$ (cf. Line 7). Let us recall that F_{λ_t} is $(L_f+\lambda_t)$-smooth. By the optimality of $\boldsymbol{v}_t$ (Fact 2.2 and Line 7), construction in Line 8, and the smoothness inequality (1.7), we have

$$0 \le \gamma_t\langle\nabla F_{\lambda_t}(\boldsymbol{x}_t)\mid \boldsymbol{x}_t-\boldsymbol{v}_t\rangle \le F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_t}(\boldsymbol{x}_{t+1}) + \gamma_t^2\frac{L_f+\lambda_t}{2}\|\boldsymbol{v}_t-\boldsymbol{x}_t\|^2. \tag{3.24}$$

So, using Lemma 2.9 and Lemma 3.1 twice,

$$\begin{aligned}
0 &\le \langle\nabla F_{\lambda_t}(\boldsymbol{x}_t)\mid \boldsymbol{x}_t-\boldsymbol{v}_t\rangle && (3.25)\\
&\le \frac{F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_{t+1}}(\boldsymbol{x}_{t+1})}{\gamma_t} + \frac{\lambda_{t+1}-\lambda_t}{\gamma_t}\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t+1}) + \gamma_t\frac{L_f+\lambda_t}{2}R && (3.26)\\
&\le \frac{F_{\lambda_t}(\boldsymbol{x}_t) - F_{\lambda_{t+1}}(\boldsymbol{x}_{t+1})}{\gamma_t} + \frac{\lambda_{t+1}-\lambda_t}{\gamma_t}R + \gamma_t\frac{L_f+\lambda_t}{2}R. && (3.27)
\end{aligned}$$

Furthermore, since f and $\operatorname{dist}^2_{\boldsymbol{D}}/2$ are smooth and $\sum_{i\in I}\omega_i C_i$ and $\bigtimes_{i\in I} C_i$ are compact, it follows that their gradients are bounded. Hence, f and $\operatorname{dist}^2_{\boldsymbol{D}}/2$ are Lipschitz continuous on these sets, with constants $\beta_f := \sup_{c\in\sum_{i\in I}\omega_i C_i}\|\nabla f(c)\|$ and $\beta_p := \sup_{\boldsymbol{c}\in\bigtimes_{i\in I} C_i}\|\nabla\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{c})/2\|$ respectively. Therefore, we find that by Jensen's inequality and Lemma 3.1,

$$\begin{aligned}
H_t &\le \beta_f\|A\boldsymbol{x}_t - A\boldsymbol{x}_t^*\| + \lambda_t\beta_p\|\boldsymbol{x}_t-\boldsymbol{x}_t^*\| && (3.28)\\
&\le \beta_f\sum_{i\in I}\omega_i R_i + \lambda_t\beta_p\sqrt{\sum_{i\in I}\omega_i R_i^2} && (3.29)\\
&= \beta_f R_A + \lambda_t\beta_p\sqrt{R}. && (3.30)
\end{aligned}$$

By Lemma 2.9, we have $\gamma_t^{-1}(F_{\lambda_{t+1}}(\boldsymbol{x}_{t+1}^*) - F_{\lambda_t}(\boldsymbol{x}_t^*)) \ge 0$. Combining all of these facts, we find

$$\begin{aligned}
0 &\le \sum_{k=0}^{t-1}\left\langle\nabla F_{\lambda_k}(\boldsymbol{x}_k)\,\middle|\,\boldsymbol{x}_k-\boldsymbol{v}_k\right\rangle && (3.31)\\
&\le \sum_{k=0}^{t-1}\left(\frac{F_{\lambda_k}(\boldsymbol{x}_k) - F_{\lambda_{k+1}}(\boldsymbol{x}_{k+1})}{\gamma_k} + \frac{\lambda_{k+1}-\lambda_k}{\gamma_k}R + \gamma_k\frac{L_f+\lambda_k}{2}R\right) && (3.32)\\
&\le \sum_{k=0}^{t-1}\left(\frac{H_k - H_{k+1}}{\gamma_k} + \frac{\lambda_{k+1}-\lambda_k}{\gamma_k}R + \gamma_k\frac{L_f+\lambda_k}{2}R\right) && (3.33)
\end{aligned}$$

where we use Lemma 2.9 in (3.33). However, the upper bound in (3.33) is equal to

$$\frac{H_0}{\gamma_0} - \frac{H_t}{\gamma_{t-1}} + \sum_{k=1}^{t-1}\left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right)H_k + \sum_{k=0}^{t-1}\left(\frac{\lambda_{k+1}-\lambda_k}{\gamma_k}R + \gamma_k\frac{L_f+\lambda_k}{2}R\right). \tag{3.34}$$

Continuing from (3.33), we drop a negative term and use (3.30) in (3.34); we then use the construction B, and simplify to reveal

$$0 \leq \sum_{k=0}^{t-1} \left\langle \nabla F_{\lambda_k}(\boldsymbol{x}_k) \,\middle|\, \boldsymbol{x}_k - \boldsymbol{v}_k \right\rangle \tag{3.35}$$

$$\leq \frac{\beta_f R_A + \lambda_0 \beta_p \sqrt{R}}{\gamma_0} + \sum_{k=1}^{t-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) \left(\beta_f R_A + \lambda_k \beta_p \sqrt{R} \right) + \sum_{k=0}^{t-1} \left(\frac{\lambda_{k+1} - \lambda_k}{\gamma_k} R + \gamma_k \frac{L_f + \lambda_k}{2} R \right) \tag{3.36}$$

$$\leq \frac{\beta_f R_A + \lambda_0 B}{\gamma_0} + \sum_{k=1}^{t-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) (\beta_f R_A + \lambda_k B) + \sum_{k=0}^{t-1} \frac{\lambda_{k+1} - \lambda_k}{\gamma_k} B + \sum_{k=0}^{t-1} \gamma_k \frac{L_f + \lambda_k}{2} R \tag{3.37}$$

$$= \frac{\beta_f R_A + \lambda_t B}{\gamma_{t-1}} + \sum_{k=0}^{t-1} \gamma_k \frac{L_f + \lambda_k}{2} R. \tag{3.38}$$

Next, we note that $\sum_{k=0}^{t-1} \gamma_k \leq 2\sqrt{t}$ and $\lambda_t \leq \lambda_0(\ln(t+1)+1)$, so

$$0 \leq \frac{1}{t} \sum_{k=0}^{t-1} \left\langle \nabla F_{\lambda_k}(\boldsymbol{x}_k) \,\middle|\, \boldsymbol{x}_k - \boldsymbol{v}_k \right\rangle \tag{3.39}$$

$$\leq \frac{\beta_f R_A + \lambda_t B}{\sqrt{t}} + \frac{1}{t} \sum_{k=0}^{t-1} \gamma_k \frac{L_f + \lambda_k}{2} R \tag{3.40}$$

$$\leq \frac{\beta_f R_A + \lambda_t B}{\sqrt{t}} + \frac{1}{t}(L_f + \lambda_{t-1}) \sum_{k=0}^{t-1} \gamma_k \frac{1}{2} R \tag{3.41}$$

$$\leq \frac{\beta_f R_A + \lambda_0(\ln(t+1)+1)B}{\sqrt{t}} + \frac{1}{\sqrt{t}}(L_f + \lambda_0(\ln(t)+1))R \tag{3.42}$$

$$\leq \frac{1}{\sqrt{t}} \Big(\beta_f \sum_{i\in I} \omega_i R_i + (L_f + \lambda_0) \sum_{i \in I} \omega_i R_i^2 + \lambda_0 B \Big) + \frac{\ln(t+1)}{\sqrt{t}} \lambda_0 \Big(\sum_{i\in I} \omega_i R_i^2 + B \Big), \tag{3.43}$$

which establishes (3.22). Since the Frank-Wolfe gaps $(\langle \nabla F_{\lambda_t}(\boldsymbol{x}_t) \mid \boldsymbol{x}_t - \boldsymbol{v}_t \rangle)_{t\in\mathbb{N}}$ are positive and the sequence of averages goes to zero, the existence of a subsequence $(t_k)_{k\in\mathbb{N}}$ such that $\langle \nabla F_{\lambda_t}(\boldsymbol{x}_{t_k}) \mid \boldsymbol{x}_{t_k} - \boldsymbol{v}_{t_k} \rangle \to 0$ follows. Lemma 3.5 implies that

$$\left(\sup_{v \in \bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}_{t_k}) \mid A\boldsymbol{x}_{t_k} - v \rangle + \lambda_{t_k} \operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t_k}) \right)_{k\in\mathbb{N}} \tag{3.44}$$

is bounded. So, since $\lambda_{t_k} \to \infty$, we must have $\operatorname{dist}^2_{\boldsymbol{D}}(\boldsymbol{x}_{t_k}) \to 0$. Therefore, for every accumulation point $\boldsymbol{x}_\infty$ of $(\boldsymbol{x}_{t_k})_{k\in\mathbb{N}}$, $\boldsymbol{x}_\infty \in \boldsymbol{D} \cap \bigtimes_{i\in I} C_i$, so $A\boldsymbol{x}_\infty \in \bigcap_{i\in I} C_i$ and

$$0 \leq \sup_{v \in \bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}_\infty) \mid A\boldsymbol{x}_\infty - \boldsymbol{v} \rangle. \tag{3.45}$$

Finally, we can bound the gap above using continuity and Lemma 3.5:

$$\sup_{v\in\bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}_\infty) \mid A\boldsymbol{x}_\infty - \boldsymbol{v}\rangle \leq \limsup_{k\to\infty} \left(\sup_{v\in\bigcap_{i\in I} C_i} \langle \nabla f(A\boldsymbol{x}_{t_k}) \mid A\boldsymbol{x}_{t_k} - \boldsymbol{v}\rangle \right) \tag{3.46}$$

$$\leq \limsup_{k\to\infty} \left(\langle \nabla F_{\lambda_{t_k}}(\boldsymbol{x}_{t_k}) \mid \boldsymbol{x}_{t_k} - \boldsymbol{v}_{t_k}\rangle \right) \tag{3.47}$$

$$= 0. \tag{3.48}$$

Since $G_{f,\bigcap_{i\in I} C_i}(A\boldsymbol{x}_\infty) = 0$, we conclude from (1.10) that $A\boldsymbol{x}_\infty$ is a stationary point. □

Remark 3.7. We emphasize that, for the cost of one extra inner product, the Frank-Wolfe gap $\langle \nabla F_{\lambda_t}(\boldsymbol{x}_t) \mid \boldsymbol{x}_t - \boldsymbol{v}_t\rangle$ can be computed while Algorithm 2.1 is running. So, checking for stationarity in the subproblems (2.7) is tractable in practice. Also, similarly to the convex-case, the choice of $\lambda_0 = L_f$ makes our convergence rate in (3.43) scale-invariant.

4. Conclusion and Future Work. Theorem 3.6 appears to be the first convergence guarantee for solving (1.1) in the nonconvex split-LMO setting. Furthermore, our rate of convergence is only one log factor less than the rate of CG for one set constraint ($m = 1$) [28]. While it is unclear if this log factor can be removed for the nonconvex setting, we believe that the analysis for the convex rate can be improved since typically the nonconvex average-F-W-gap rate is quadratically slower than the convex primal gap rate [28]. This speed-up has been achieved in some settings with algorithms which require one LMO call per iteration [18, 27], but it appears that the question of whether or not $\mathcal{O}(1/t)$ convergence is possible in the split-LMO setting without additional assumptions remains open.

While our analysis shows that Algorithm 2.1 asymptotically solves (1.1) (in the sense of Theorems 3.3 and 3.6), one drawback is that the *rates* of convergence concern the penalized functions F_{λ_t}, which contain both the primal function value and the penalized feasibility term $\lambda_t \mathrm{dist}^2_{\boldsymbol{D}}$. This means that, outside of the case $m = 1$, both primal suboptimality and feasibility are analyzed in one, composite quantity. An improvement to our analysis would be to translate convergence rates of the composite penalty into separate convergence rates for both primal suboptimality and feasibility, e.g., as was done for a method of sequential averaging in the recent preprint [17].

Based on preliminary numerical experiments, there may be more to discover for Algorithm 2.1. We have observed that the algorithm performance (in terms of feasibility and F-W gap) can be highly dependent on the initial value λ_0 when using the parameter schedules $(\lambda_t, \gamma_t) = (\mathcal{O}(\ln t), \mathcal{O}(1/\sqrt{t}))$ which are proven to work. However, based on rough experimentation, we are hopeful that it may be possible to use Algorithm 2.1 with an adaptive strategy for $(\lambda_t)_{t\in\mathbb{N}}$ in conjunction with a faster step-size $(\gamma_t)_{t\in\mathbb{N}}$, for instance, $\mathcal{O}(1/t)$ or short-step selections similar to [28]. In fact, the proofs of Theorems 3.3 and 3.6 can easily be extended to a short-step selection for γ_t by minimizing the upper bound arising from (1.7); however, thus-far, we have not been able to properly analyze a new adaptive scheme for $(\lambda_t)_{t\in\mathbb{N}}$. This is a topic of future work. At least for now, it appears our contribution is predominantly of theoretical interest.

In addition to the questions above, there are several interesting theoretical and numerical investigations to be performed. For instance, one could investigate Algorithm 2.1 under additional assumptions on the objective or constraints. For instance, CG algorithms possess accelerated convergence rates when the objective function or constraints are strongly convex [15, 34]; one could extend this analysis to Algorithm 2.1. Arguments similar to that of [6, Theorem 2.6] appear fruitful for allowing a variant of Algorithm 2.1 which admits approximate LMO evaluations in the convex setting of Theorem 3.3, provided the accuracy of computing $\boldsymbol{v}_t$ is bounded by $\mathcal{O}(1/\sqrt{t})$; more generally investigating approximate LMO implementations for this setting appears to be an open problem. Many projection-based splitting methods have an advantage of being block-iterative, i.e., instead of requiring a computation for all constraints indexed by I (as is required in the for loop in Algorithm 2.1, Line 6) at every iteration t, only a subset $I_t \subset I$ of updates are performed. This can significantly reduce the computational load per iteration, and block-iterative projection methods enjoy convergence under very mild assumptions on the blocks $(I_t)_{t\in\mathbb{N}}$ [11, 12]. It is worth noting that the inner loop of Algorithm 2.1 can be parallelized, and a block-iterative capability would further improve the per-iteration cost. Several LMO-based block-iterative algorithms have been proposed for solving problems like the relaxation (2.7) [2, 5], but extending them to solve (1.1) remains to be done.

Acknowledgments. The work for this article has been supported by MODAL-Synlab, and took place on the Research Campus MODAL funded by the German Federal Ministry of Education and Research (BMBF) (fund numbers 05M14ZAM, 05M20ZBM). This research was also supported by the DFG Cluster of Excellence MATH+ (EXC-2046/1, project ID 390685689) funded by the Deutsche Forschungsgemeinschaft (DFG).

We thank the reviewers who have significantly improved the quality of this article through their thoughtful comments. We also thank Kamiar Asgari, Gábor Braun, Mathieu Besançon, Ibrahim Ozaslan, Christophe Roux, Antonio Silveti-Falls, David Martínez-Rubio, and Elias Wirth for their valuable feedback and discussions.

REFERENCES

[1] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd ed.*, Springer, 2017.

[2] A. Beck, E. Pauwels, and S. Sabach, *The cyclic block conditional gradient method for convex optimization problems*, SIAM J. Optim., 25 (2015), pp. 2024–2049.

[3] T. Bernholt, F. Eisenbrand, and T. Hofmeister, *Constrained Minkowski sums: A geometric framework for solving interval problems in computational biology efficiently*, Discrete Comput. Geom., 42 (2009), pp. 22–36.

[4] J. Bolte, C. W. Combettes, and E. Pauwels, *The iterates of the Frank-Wolfe algorithm may not converge*, Math. Oper. Res., (to appear).

[5] I. Bomze, F. Rinaldi, and D. Zeffiro, *Projection free methods on product domains*, 2023, https://arxiv.org/abs/2302.04839.

[6] G. Braun, A. Carderera, C. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta, *Conditional gradient methods*, 2022, https://arxiv.org/abs/2211.14103.

[7] G. Braun, S. Pokutta, and R. Weismantel, *Alternating linear minimization: Revisiting von Neumann's alternating projections*, 2022, https://arxiv.org/abs/2212.02933.

[8] Y. Censor and A. Cegielski, *Projection methods: an annotated bibliography of books and reviews*, Optimization, 64 (2015), pp. 2343–2358, https://doi.org/10.1080/02331934.2014.957701.

[9] I. Chryssoverghi, A. Bacopoulos, B. Kokkinis, and J. Coletsos, *Mixed Frank–Wolfe penalty method with applications to nonconvex optimal control problems*, J. Optim. Theory Appl., 94 (1997), pp. 311–334.

[10] C. W. Combettes and S. Pokutta, *Complexity of linear minimization and projection on some sets*, Oper. Res. Lett., 49 (2021), pp. 565–571.

[11] P. L. Combettes and J.-C. Pesquet, *Proximal Splitting Methods in Signal Processing*, Springer New York, New York, NY, 2011, https://doi.org/10.1007/978-1-4419-9569-8_10.

[12] P. L. Combettes and Z. C. Woodstock, *Reconstruction of functions from prescribed proximal points*, J. Approx. Theory, 268 (2021), p. 105606, https://doi.org/https://doi.org/10.1016/

j.jat.2021.105606.

[13] T. Ding, D. Lim, R. Vidal, and B. D. Haeffele, *Understanding doubly stochastic clustering*, in Proceedings of the 39th International Conference on Machine Learning, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds., vol. 162 of Proceedings of Machine Learning Research, PMLR, 17–23 Jul 2022, pp. 5153–5165.

[14] L. L. Duan, A. L. Young, A. Nishimura, and D. B. Dunson, *Bayesian constraint relaxation*, Biometrika, 107 (2019), pp. 191–204, https://doi.org/10.1093/biomet/asz069.

[15] D. Garber and E. Hazan, *Faster rates for the Frank-Wolfe method over strongly-convex sets*, in Proceedings of the 32nd International Conference on Machine Learning, F. Bach and D. Blei, eds., vol. 37 of Proceedings of Machine Learning Research, Lille, France, 07–09 Jul 2015, PMLR, pp. 541–549.

[16] D. Garber, A. Kaplan, and S. Sabach, *Improved complexities of conditional gradient-type methods with applications to robust matrix recovery problems*, Math. Program., 186 (2021), pp. 185–208.

[17] K.-H. Giang-Tran, N. Ho-Nguyen, and D. Lee, *Projection-free methods for solving convex bilevel optimization problems*, 2023, https://arxiv.org/abs/2311.09738.

[18] G. Gidel, F. Pedregosa, and S. Lacoste-Julien, *Frank-Wolfe splitting via augmented Lagrangian method*, in Proceedings of the 21st International Conference on Artificial Intelligence and Statistics, A. Storkey and F. Perez-Cruz, eds., vol. 84 of Proceedings of Machine Learning Research, PMLR, 09–11 Apr 2018, pp. 1456–1465.

[19] N. He and Z. Harchaoui, *Semi-proximal mirror-prox for nonsmooth composite minimization*, in Advances in Neural Information Processing Systems, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds., vol. 28, Curran Associates, Inc., 2015.

[20] V. Kolmogorov and T. Pock, *One-sided Frank-Wolfe algorithms for saddle problems*, in Proceedings of the 38th International Conference on Machine Learning, M. Meila and T. Zhang, eds., vol. 139 of Proceedings of Machine Learning Research, PMLR, 18–24 Jul 2021, pp. 5665–5675.

[21] G. Lan, E. Romeijn, and Z. Zhou, *Conditional gradient methods for convex optimization with general affine and nonlinear constraints*, SIAM J. Optim., 31 (2021), pp. 2307–2339.

[22] K. Lange, J.-H. Won, and J. Xu, *Projection onto Minkowski sums with application to constrained learning*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, PMLR, 09–15 Jun 2019, pp. 3642–3651.

[23] Y.-F. Liu, X. Liu, and S. Ma, *On the nonergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming*, Math. Oper. Res., 44 (2019), pp. 632–650.

[24] T. Lozano-Pérez and M. A. Wesley, *An algorithm for planning collision-free paths among polyhedral obstacles*, Commun. ACM, 22 (1979), pp. 560–570.

[25] A. Migdalas, *A regularization of the Frank—Wolfe method and unification of certain nonlinear programming methods*, Math. Program., 65 (1994), pp. 331–345.

[26] R. D. Millán, O. P. Ferreira, and L. F. Prudente, *Alternating conditional gradient method for convex feasibility problems*, 80 (2021), pp. 245—269, https://doi.org/10.1080/10556788.2013.796683.

[27] C. Mu, Y. Zhang, J. Wright, and D. Goldfarb, *Scalable robust matrix recovery: Frank-Wolfe meets proximal methods*, SIAM J. Sci. Comput., 38 (2016), pp. A3291–A3317.

[28] F. Pedregosa, G. Negiar, A. Askari, and M. Jaggi, *Linearly convergent Frank-Wolfe with backtracking line-search*, in International conference on artificial intelligence and statistics, PMLR, 2020, pp. 1–10.

[29] G. Pierra, *Decomposition through formalization in a product space*, Math. Program., 28 (1984), pp. 96–115.

[30] E. Richard, P. Savalle, and N. Vayatis, *Estimation of simultaneously sparse and low rank matrices*, in Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012, icml.cc / Omnipress, 2012.

[31] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317, Springer Science & Business Media, 2009.

[32] T. Rothvoss, *The matching polytope has exponential extension complexity*, J. ACM, 64 (2017), https://doi.org/10.1145/3127497.

[33] A. Silveti-Falls, C. Molinari, and J. Fadili, *Generalized conditional gradient with augmented Lagrangian for composite minimization*, SIAM J. Optim., 30 (2020), pp. 2687–2725, https://doi.org/10.1137/19M1240460.

[34] E. Wirth, T. Kerdreux, and S. Pokutta, *Acceleration of Frank-Wolfe algorithms with open-loop step-sizes*, in Proceedings of The 26th International Conference on Artificial

Intelligence and Statistics, F. Ruiz, J. Dy, and J.-W. van de Meent, eds., vol. 206 of Proceedings of Machine Learning Research, PMLR, 25–27 Apr 2023, pp. 77–100.

[35] Z. Yang, J. Corander, and E. Oja, *Low-rank doubly stochastic matrix decomposition for cluster analysis*, J. Mach. Learn. Res., 17 (2016), pp. 6454–6478.

[36] A. Yurtsever, O. Fercoq, and V. Cevher, *A conditional-gradient-based augmented Lagrangian framework*, in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, eds., vol. 97 of Proceedings of Machine Learning Research, PMLR, 09–15 Jun 2019, pp. 7272–7281.

[37] Z. Zhang, Z. Zhai, and L. Li, *Graph refinement via simultaneously low-rank and sparse approximation*, SIAM J. Sci. Comput., 44 (2022), pp. A1525–A1553.