Frank-Wolfe with Moreau Envelope Smoothing for Nonsmooth Nonconvex Problems

Antonio Silveti-Falls[1*†], Cesare Molinari[2†] and Zev Woodstock[3†]

[1*]CVN, CentraleSupélec, 3 Rue Joliot Curie, Gif-sur-Yvette, 91190, Essonne, France.

[2]Department of Mathematics, Università di Genova, Via Dodecaneso 35, Genova, 16146, Liguria, Italy.

[3]Department of Mathematics & Statistics, James Madison University, 60 Bluestone Drive, Harrisonburg, 22807, VA, USA.

*Corresponding author(s). E-mail(s): tonys.falls@gmail.com;
Contributing authors: cecio.molinari@gmail.com; woodstzc@jmu.edu;
†All authors contributed equally.

Abstract

We present and analyze Frank-Wolfe with Moreau Envelope Smoothing (FRAMES) for solving nonsmooth nonconvex constrained optimization problems, taking advantage of iterative smoothing via the Moreau envelope followed by one Frank-Wolfe step per iteration. The problem template we consider encompasses splitting problems with multiple convex constraint sets as well as problems with nonsmooth weakly convex regularizers like the MCP or SCAD penalties. We prove convergence, with rates, for both of these cases under a variety of mild assumptions, including inconsistent problems. Additionally, we highlight a new relationship between the Frank-Wolfe gap for a problem with nonsmooth objective and the Frank-Wolfe gap for a smoothed surrogate problem, demonstrating suboptimality of prior analyses. Numerical experiments are performed for matrix factorization problems and nonconvex quadratic splitting over multiple convex constraint sets, where the improvements in analysis are empirically observed.

Keywords: Frank-Wolfe, Moreau envelope smoothing, nonsmooth nonconvex optimization, inconsistent constraints, weakly convex regularization, Frank-Wolfe gap.

1 Introduction

The Frank-Wolfe algorithm [1] and its related extensions [2–5] provide an attractive first-order scheme for solving minimization problems posed over a constraint set $\mathcal{C}$ without requiring projections onto it. Instead of projecting onto $\mathcal{C}$ at each iteration, the Frank-Wolfe algorithm makes use of a *linear minimization oracle* (LMO) to construct its update in a way that guarantees the next iterate remains feasible in $\mathcal{C}$. This is useful for problems where projections (i.e., quadratic minimization oracles) are computationally expensive but linear minimization oracles are tractable, with recent work precisely detailing the differences in complexity between projection and LMO [6, 7].

Typically, Frank-Wolfe is applied to problems where the objective function is continuously differentiable, since at each iteration it requires computing a gradient to pass into the linear minimization oracle. However, many modern problems fit the following, more general constrained composite format

$$\min_{x \in \mathcal{C} \subset \mathbb{R}^n} f(x) + g(Tx), \tag{P}$$

where $\mathcal{C} \neq \varnothing$ is the set over which the linear minimization oracle is accessible, f is smooth but not assumed to be convex, T is a linear map, and g is some nonsmooth function whose proximal operator is tractable, typically representing a regularizer or an additional constraint set $\mathcal{D}$ whose projection is cheap to compute. Examples fitting into (P) include inverse problems when f models the data-fidelity of the recovered solution and g is some regularizer, e.g., the MCP or SCAD penalties, or the indicator of an additional convex constraint set, like the nonnegative orthant. Directly applying Frank-Wolfe to such problems is not simple; for instance, replacing the gradient with a subgradient and applying Frank-Wolfe does not necessarily lead to convergence, even on simple functions like $g(x) = \max(x_1, x_2)$ [8].

There exist nonsmooth extensions of the Frank-Wolfe algorithm [3, 8–11] that use the Moreau envelope to smooth the function g. The existing analyses in these cited works all assume that g is convex. Furthermore, all of them except [11] assume that f is convex, thus excluding nonconvex instances of (P) like matrix factorization problems. For these smooth nonconvex problems, first-order stationarity is typically measured using the Frank-Wolfe gap (an analog of the norm of the gradient for unconstrained problems), detailed in Section 4. However, the compatibility of the Frank-Wolfe gap with the Moreau envelope is much less studied. While [11] established asymptotic subsequential convergence to a stationary point of (P) is possible when g is the indicator function of a specific linear subspace, in the general setting of (P), it is not obvious how the Frank-Wolfe gap associated to a surrogate smoothed problem will certify stationarity of the original problem.

Approach

As in [3, 8, 10, 11], we replace the nonsmooth term g by its Moreau envelope with a vanishing smoothing parameter. More precisely, at iteration k we consider the smooth surrogate

$$\Phi_k(x) := f(x) + g^{\beta_k}(Tx),$$

where $\beta_k > 0$ decreases to zero, and we apply one Frank-Wolfe step to Φ_k over the set $\mathcal{C}$. This construction preserves the projection-free nature of the method with respect to $\mathcal{C}$, since each update only requires a linear minimization oracle over $\mathcal{C}$, while the nonsmooth term is handled

through the proximity operator of g. The use of a decreasing smoothing parameter is essential: for fixed β, the algorithm would only solve a smoothed approximation of (P), whereas letting $\beta_k \to 0$ forces the surrogate objectives to return to the original problem. At the same time, this creates a tradeoff, since smaller values of β_k typically lead to worse smoothness constants for Φ_k. The main goal of this paper is therefore to analyze this tradeoff and to determine when convergence of the Frank-Wolfe gaps associated with the smoothed objectives yields meaningful stationarity guarantees for the original nonsmooth problem (P) without assuming convexity.

Contributions

We present a Moreau-envelope Frank-Wolfe framework for the nonsmooth composite problem (P), together with convergence guarantees that distinguish between progress on the smoothed surrogate problems and stationarity of the original nonsmooth problem. Our main contributions are as follows.

- **Smoothed Frank-Wolfe gap convergence with vanishing smoothing.** We prove that, despite the objective Φ_k changing at every iteration, the average and best Frank-Wolfe gaps associated with the smoothed objectives admit explicit convergence rates. In particular, for $p, q \in]0, 1[$, step sizes $\gamma_k = (k+1)^{-p}$ and smoothing schedules $\beta_k = \beta_0(k+1)^{-q}$, Theorem 5.3 establishes an $\mathcal{O}(N^{-\min\{p-q,\, 1-p-q\}})$ bound on the average smoothed Frank-Wolfe gap, and hence on the best smoothed gap among the first N iterates. This result extends the usual smooth nonconvex Frank-Wolfe gap analysis to a setting with a nonsmooth composite term, including both indicator and Lipschitz weakly convex cases, a linear composition, and a vanishing smoothing parameter. Remark 6.16 later explains that the choice $p = 1/2$ and $q = 1/4$ optimizes the final nonsmooth stationarity rate to $\mathcal{O}(N^{-1/4})$.
- **Nonsmooth stationarity guarantee.** We also show in Theorem 5.7 and Theorem 5.8 that convergence of the smoothed Frank-Wolfe gaps along a subsequence guarantees cluster points of that subsequence are stationary points, which provides a generalization of [11] beyond the setting when g is the indicator function of a specific linear subspace. In general, one cannot expect convergence of the full sequence of iterates, since even in the smooth setting Frank-Wolfe iterates need not converge [12]. We also show in Theorem 6.9 that, if the underlying problem is inconsistent, the algorithm still produces a stationary point of a relaxed problem that minimizes the distance between the constraint sets.
- **Transfer from smoothed gaps to nonsmooth stationarity certificates.** We then analyze when small smoothed Frank-Wolfe gaps imply meaningful progress for the original nonsmooth problem (P). For the indicator case $g = \iota_{\mathcal{D}}$, Lemma 6.4 characterizes relationships between the smoothed gap, the feasibility violation $\operatorname{dist}_{\mathcal{D}}(Tx_k)$, the smoothing parameter β_k, and the Frank-Wolfe gap over the true feasible set $\widetilde{\mathcal{C}} := \mathcal{C} \cap T^{-1}(\mathcal{D})$. For the Lipschitz weakly convex case, our gap-transfer result, Lemma 6.14, bounds an appropriate nonsmooth Frank-Wolfe gap in terms of the smoothed gap and the smoothing parameter. These estimates lead to explicit rates on the nonsmooth Frank-Wolfe gap in Theorems 6.5 and 6.15, clarifying how the decay of β_k affects stationarity guarantees for the original problem (P).

  These bounds are also independent of our algorithm, and can be applied to any problem of the form (P) that uses Moreau smoothing. For instance, Remark 6.6 demonstrates how

the smoothed convergence rates of [11, 13] translate to nonsmooth convergence rates, revealing their step size and smoothing schedules to be suboptimal.

- **Numerical evidence and smoothing-parameter behavior.** We illustrate the theory on examples involving nonnegative matrix factorization, trend-filtered matrix factorization with nonconvex nonsmooth penalties, and a nonconvex splitting problem over intersecting constraint sets in Section 7. These experiments demonstrate the practical effect of the smoothing schedule and the sensitivity to the initial smoothing parameter. The splitting experiment is designed in such a way that the nonsmooth Frank-Wolfe gap is tractable, and the difference between smoothed gap convergence and progress on the original nonsmooth stationarity measures is shown.

Outline

We finish this section with a discussion of related work before moving on to Section 2, where we formally state the standing assumptions on (P) and present the main algorithm, **FRAMES**. In Section 3, we collect the elementary estimates on Moreau envelopes and the smoothed objectives that are used throughout the convergence analysis. In Section 4, we introduce the different Frank-Wolfe gaps and stationarity certificates that arise for the smoothed problem, the indicator case, and the Lipschitz weakly convex case. In Section 5, we prove convergence rates for the smoothed Frank-Wolfe gaps generated by the algorithm and provide guarantees for subsequential stationary points of (P). In Section 6, we relate the convergence of these smoothed gaps to stationarity certificates for the original, nonsmooth problem (P), including both finite-time gap-transfer results (yielding explicit convergence rates) and subsequential stationarity guarantees for inconsistent problems. Finally, in Section 7, we illustrate the behavior of the method on several numerical examples and highlight the effect of the smoothing schedule on the certificate of stationarity for (P).

1.1 Related Work

We separate the discussion of related work into categories based on three features: the use of a linear minimization oracle over $\mathcal{C}$, the presence of the nonsmooth term $g \circ T$, and the notion of stationarity used in the results.

Smooth Nonconvex Frank-Wolfe

For smooth constrained problems, i.e., (P) with $g = 0$, Frank-Wolfe has been studied in the nonconvex setting already, with the notable work of [4] establishing an $\mathcal{O}(N^{-1/2})$ convergence rate for the Frank-Wolfe gap using either a short step or a line search step size.

Moreau Envelope Smoothing and Convex Composite Frank-Wolfe

Several works considering convex composite problems used the Moreau envelope (sometimes under the name homotopy-based smoothing) in conjunction with the linear minimization oracle over $\mathcal{C}$ as in Frank-Wolfe. In [3, 8], the nonsmooth term in a composite objective is replaced by a Moreau envelope with a decreasing smoothing schedule, with one step of Frank-Wolfe applied to the smoothed surrogate problem at each iteration; [3] considered only Lipschitz nonsmooth terms while [8] considered both Lipschitz nonsmooth terms and indicator functions. Augmented-Lagrangian variants for handling affine constraints like $Ax = b$ separately from

$\mathcal{C}$ were developed in [9, 10], with inexact and stochastic versions appearing in [14]. Stochastic composite Frank-Wolfe methods were also studied in [15].

Other approaches based on the Moreau envelope with linear minimization oracle over $\mathcal{C}$ for nonsmooth convex optimization include MOPES and MOLES [16, 17]. These methods also use Moreau smoothing, but their goal is convex suboptimality or regret, and the LMO-based variant approximates projection-type operations through an inner Frank-Wolfe routine. They do not address nonconvex stationarity for the composite problem (P).

All the algorithms studied in these papers are similar to **FRAMES** in the sense that they combine a linear minimization oracle over $\mathcal{C}$ with the proximal operator of g. However, these analyses all require convexity and provide guarantees on functional gaps, primal-dual gaps, or regret. In contrast, we consider (P) with nonconvex f and nonsmooth weakly convex g, for which the appropriate notion of first-order stationary point is better captured by the Frank-Wolfe gap.

Frank-Wolfe Splitting

The closest nonconvex predecessor to our work is [11], which studies a Frank-Wolfe method for minimizing a smooth nonconvex function over a nonempty intersection of compact convex sets,

$$\min_{x \in \bigcap_{i=1}^{m} \mathcal{C}_i} f(x).$$

Their approach lifts the problem to the product space $\mathcal{C}_1 \times \cdots \times \mathcal{C}_m$ and enforces consensus by taking g to be the indicator of the diagonal subspace (à la Pierra [18]), yielding a special case of (P). By smoothing this indicator and applying Frank-Wolfe to the resulting surrogate problems, [11] proves a convergence rate for averaged smoothed Frank-Wolfe gaps in the nonconvex setting. A follow-up work [13] improves the corresponding smoothed-gap rate through a modified step-size choice. These splitting methods are highly relevant to our work, but they treat a specific consensus-constraint structure. Our analysis keeps the same Moreau-smoothing philosophy while allowing more general nonlinear constraints and Lipschitz weakly convex nonsmooth terms, and we explicitly relate the smoothed Frank-Wolfe gaps to stationarity certificates for the original nonsmooth problem.

Variable Smoothing and Weak Convexity

Variable smoothing methods based on the Moreau envelope have also been studied outside the Frank-Wolfe setting. Most relevant for us is [19], which considers unconstrained composite problems involving a weakly convex nonsmooth term and a decreasing Moreau smoothing parameter. The estimates used there are similar in spirit to several of the Moreau-envelope estimates used in our analysis. The main difference is algorithmic and geometric: [19] does not work in a projection-free Frank-Wolfe setting, nor does it analyze stationarity through Frank-Wolfe gaps. In contrast, our setting (P) allows for constrained problems and our method preserves the linear-minimization-oracle geometry over $\mathcal{C}$, uses only the proximal operator of g, and studies how stationarity of the smoothed surrogate transfers back to stationarity certificates for the original nonsmooth composite problem (P).

There is also a broader literature on projection-free methods for convex or weakly smooth composite problems, including conditional-gradient methods under Hölder or weak smoothness assumptions [20–22]. The convergence guarantees in these works require convexity and are not designed to address nonsmooth nonconvex stationarity for the composite problem (P).

Nonsmooth Frank-Wolfe Methods Without Smoothing

A complementary line of work develops Frank-Wolfe-type methods for nonsmooth problems without using Moreau smoothing. [23] studies a Frank-Wolfe method for upper-$C^{1,\alpha}$ functions over compact convex sets and proves rates toward Clarke stationarity. Although this setting is close to that of (P), the distinction is that [23] relies on an upper-$C^{1,\alpha}$ structure and a suitable generalized gradient selection, whereas our approach exploits the composite form $f + g \circ T$, prox access to g, and Moreau smoothing of the nonsmooth outer term.

Another recent direction is Frank-Wolfe for abs-smooth functions [24], where the algorithm uses an abs-linearization of the objective and a generalized Frank-Wolfe gap. This provides an alternative to smoothing and can recover smooth-like rates for the corresponding generalized gap. However, the method relies on the specialized abs-smooth structure and requires solving a more involved subproblem rather than a standard linear minimization oracle over $\mathcal{C}$. Thus, it is complementary to the present approach, which keeps the usual Frank-Wolfe linear oracle and instead handles nonsmoothness through the Moreau envelope of g.

In the convex nonsmooth setting, several projection-free approaches also avoid Moreau smoothing. Deterministic nonsmooth Frank-Wolfe methods with coreset guarantees were studied in [25], while methods based on uniform affine approximations for separable nonsmooth convex functions were developed in [26]. More recently, [27] studied nonsmooth projection-free convex optimization with functional constraints using subgradients and a separation-type scheme. These works show that nonsmooth projection-free optimization can be approached without smoothing, but their guarantees are for convex problems or rely on specialized structure that is different from the general nonconvex composite model (P).

Finally, [28] proposes a nonsmooth Frank-Wolfe algorithm through a dual cutting-plane perspective. Their approach is quite different from the Moreau-envelope methods discussed above: it reinterprets the fully corrective Frank-Wolfe algorithm as dual to a cutting-plane method with partial linearization, and then uses this dual viewpoint to develop a nonsmooth Frank-Wolfe-type scheme. This provides an important recent no-smoothing alternative, although it requires convexity and is algorithmically distinct from the single-loop Moreau-smoothed Frank-Wolfe method studied here.

2 Problem Setup and Algorithm

2.1 Standing Assumptions

We work under the following assumptions on (P) throughout the paper.

Assumption 2.1 (Smooth term). *The function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable on an open set containing $\mathcal{C}$, and its gradient is Lipschitz-continuous on $\mathcal{C}$ with constant $L_{\nabla f} > 0$.*

Assumption 2.2 (Linear map). *The operator $T : \mathbb{R}^n \to \mathbb{R}^m$ is linear.*

Assumption 2.3 (Frank-Wolfe feasible set). *The nonempty set $\mathcal{C} \subset \mathbb{R}^n$ is compact and convex, with diameter $\operatorname{diam}_{\mathcal{C}} := \sup_{x,y\in\mathcal{C}} \|x - y\|_2 < +\infty$.*

Assumption 2.4 (Nonsmooth term). *The function $g : \mathbb{R}^m \to]-\infty, +\infty]$ satisfies one of the following:*

(I) $g = \iota_{\mathcal{D}}$ *for some nonempty closed convex set* $\mathcal{D} \subset \mathbb{R}^m$.
  (a) (optional) Additionally, it holds $\mathcal{D} \cap T(\mathcal{C}) \neq \varnothing$.
  (b) (optional) Additionally, it holds $\mathrm{ri}(\mathcal{D}) \cap \mathrm{ri}(T(\mathcal{C})) \neq \varnothing$.

(II) $g\colon \mathbb{R}^m \to \mathbb{R}$ *is* L_g*-Lipschitz-continuous on* $\mathbb{R}^m$ *and* ρ*-weakly convex for some* $\rho \geq 0$*, i.e.,* $g + \frac{\rho}{2}\|\cdot\|_2^2$ *is convex.*

Without loss of generality, Assumption 2.4 provides that g is ρ-weakly convex; under Assumption 2.4(I) we use the convention $\rho = 0$ with $\rho^{-1} = +\infty$. We assume access to three computational primitives: the gradient ∇f, the linear minimization oracle (LMO), given some vector v, returns a point in $\arg\min_{s\in\mathcal{C}}\langle v, s\rangle$, and the proximal operator of g, $\mathrm{prox}_{\beta g}(y) := \mathrm{argmin}_{u\in\mathbb{R}^m}\left\{g(u) + \frac{1}{2\beta}\|u-y\|_2^2\right\}$. Under Assumption 2.4(I), for every $\beta > 0$, $\mathrm{prox}_{\beta g} = \mathrm{P}_{\mathcal{D}}$. Open-source repositories for proximal operators and LMOs can be found, e.g., at [29, 30].

2.2 Notation

For general background see, e.g., [31]. We denote $\mathbb{N} := \{0, 1, 2, \ldots\}$ and $\mathbb{N}^* := \mathbb{N} \setminus \{0\}$. For a proper function $g : \mathbb{R}^m \to]-\infty, +\infty]$, its *domain* is $\mathrm{dom}(g) := \{x \in \mathbb{R}^m : g(x) < +\infty\}$. The *indicator function* of a set S is $\iota_S(x) := 0$ if $x \in S$; and $+\infty$ if $x \notin S$. For a closed set S, $\mathrm{dist}_S(y) := \inf_{u\in S}\|y-u\|_2$ and $\mathrm{dist}_S^2(y) := (\mathrm{dist}_S(y))^2$ for the *distance* (or its square) to S. The *projection* onto a nonempty closed convex set $\mathcal{D}$ is denoted by $\mathrm{P}_{\mathcal{D}}$. Since $\mathcal{C}$ is compact and T is linear, $T(\mathcal{C})$ is compact, hence the *diameter* $\mathrm{diam}_{T(\mathcal{C})} := \sup_{x,y\in\mathcal{C}}\|Tx - Ty\|_2$ is finite. Under Assumption 2.4(I), the *one-sided excess* of $T(\mathcal{C})$ from $\mathcal{D}$, $e_{\mathcal{D},T(\mathcal{C})} := \max_{z\in T(\mathcal{C})}\mathrm{dist}_{\mathcal{D}}(z)$, is finite since $T(\mathcal{C})$ is compact by Assumption 2.2 and Assumption 2.3. Moreover, under the additional hypotheses in either Assumption 2.4(I)(a) or Assumption 2.4(I)(b), the intersection $\mathcal{D} \cap T(\mathcal{C})$ is nonempty and therefore $e_{\mathcal{D},T(\mathcal{C})} \leq \mathrm{diam}_{T(\mathcal{C})}$. Finally, since f is continuously differentiable on an open set containing the compact set $\mathcal{C}$, the quantity

$$L_f := \sup_{x\in\mathcal{C}}\|\nabla f(x)\|_2$$

is finite. Hence, for all $x, y \in \mathcal{C}$, $f(x) - f(y) \leq L_f\|x-y\|_2 \leq L_f\mathrm{diam}_{\mathcal{C}}$. When g is Lipschitz-continuous, we will denote the *Clarke subdifferential* of g at x by $\partial g(x)$. When g is convex and Lipschitz-continuous, this coincides with the classical convex subdifferential. When $g = \iota_S$ is the indicator function for a closed convex set S and $x \in S$, ∂g will denote the *normal cone*, defined as $N_S(x) := \{v : \langle v, s - x\rangle \leq 0 \quad \forall s \in S\}$, with $N_S = \varnothing$ if $x \notin S$. Here and throughout, ri denotes *relative interior* and $T(\mathcal{C}) := \{Tx\colon x \in \mathcal{C}\}$. For $\beta > 0$, the Moreau envelope of g is

$$g^\beta(y) := \min_{u\in\mathbb{R}^m} g(u) + \frac{1}{2\beta}\|u-y\|_2^2. \tag{2.1}$$

This work only considers the choice $\beta < \rho^{-1}$ which guarantees $\mathrm{prox}_{\beta g}$ is unique. With this choice of β, under Assumption 2.4, g^β is also continuously differentiable [32, Corollary 3.4], and $\nabla g^\beta(y) = \frac{1}{\beta}\left(y - \mathrm{prox}_{\beta g}(y)\right)$.

2.3 Algorithm

At iteration k, we smooth g using the Moreau envelope to get the following smoothed objective and gradient

$$\Phi_k(x) := f(x) + g^{\beta_k}(Tx) \text{ with } \nabla\Phi_k(x) = \nabla f(x) + \frac{1}{\beta_k} T^* \left(Tx - \operatorname{prox}_{\beta_k g}(Tx)\right). \quad (2.2)$$

The **FRAMES** algorithm applies one Frank-Wolfe step to the smoothed objective Φ_k (only smoothing g rather than $g \circ T$) over $\mathcal{C}$ at each iteration. Since $\mathcal{C}$ is convex and the step size $\gamma_k \in]0, 1]$, the update in Step 3 guarantees $x_k \in \mathcal{C}$ for all k whenever $x_0 \in \mathcal{C}$. However, under Assumption 2.4(I), the same is not guaranteed for the constraint $T^{-1}(\mathcal{D})$; the iterates may not satisfy $Tx_k \in \mathcal{D}$ for any finite k.

Algorithm 1: **Fra**nk-Wolfe with **M**oreau **E**nvelope **S**moothing (**FRAMES**)

Input: $x_0 \in \mathcal{C}$, step sizes $\{\gamma_k\}_{k\in\mathbb{N}} \subset]0, 1]$, and smoothing schedule $\{\beta_k\}_{k\in\mathbb{N}} \subset]0, \rho^{-1}[$
For $k = 0, 1, 2, \ldots$ **do**

1. Compute the gradient of the smoothed objective
$$\nabla\Phi_k(x_k) = \nabla f(x_k) + \frac{1}{\beta_k} T^* \left(Tx_k - \operatorname{prox}_{\beta_k g}(Tx_k)\right).$$
2. Compute the LMO for this direction
$$s_k \in \arg\min_{s\in\mathcal{C}} \langle \nabla\Phi_k(x_k), s\rangle.$$
3. Update the iterate
$$x_{k+1} = x_k + \gamma_k(s_k - x_k).$$

End for

Remark 2.5. Even if $\operatorname{prox}_{\beta g}$ is available for $\beta > 0$, unless T has special structure (like orthogonality), evaluating $\operatorname{prox}_{g\circ T}$ is typically not possible in closed-form. Hence, by smoothing only the outer function g instead of the composition $g \circ T$, **FRAMES** only needs access to $\operatorname{prox}_{\beta g}$ for $\beta \in \left]0, \rho^{-1}\right[$ to compute $\nabla\Phi_k(x_k)$ at each iteration.

The main results in this paper will be stated for the open-loop schedules $\gamma_k = (k+1)^{-1/2}$ and $\beta_k = \beta_0(k+1)^{-1/4}$, where $\beta_0 \in \left]0, \rho^{-1}\right[$ in the Lipschitz weakly convex case, with the convention $\rho^{-1} = +\infty$ when $\rho = 0$. Both algorithms appearing in [11, 13] arise as special cases of **FRAMES** under the special case of Assumption 2.4(I) when $\mathcal{D}$ is a specific linear subspace. These suggested schedules of $\gamma_k = \mathcal{O}(k^{-1/2})$ and $\beta_k = \mathcal{O}(1/\log(k))$ will still yield convergence but our analysis shows that this will lead to inferior performance compared

to the main schedules proposed in this work. Some intermediate results are stated under the following, more general assumptions.

Assumption 2.6. *The step size sequence* $\{\gamma_k\}_{k\in\mathbb{N}} \subset]0,1]$ *is nonincreasing and* $\lim_{k\to\infty} \gamma_k = 0$.

Assumption 2.7. *The smoothing parameter sequence* $\{\beta_k\}_{k\in\mathbb{N}} \subset]0, \rho^{-1}[$ *is nonincreasing and* $\lim_{k\to\infty} \beta_k = 0$.

3 Preliminaries and Basic Estimates

The main purpose of this section is to record how the Moreau envelope behaves as the smoothing parameter varies, and to derive uniform bounds on $\Phi_k(x) = f(x) + g^{\beta_k}(Tx)$.

3.1 Moreau Envelope Estimates

Proposition 3.1 (Moreau envelope calculus). *Let* g *satisfy* Assumption 2.4*, and let* $\beta \in \left]0, \rho^{-1}\right[$*. Then* g^β *is continuously differentiable on* $\mathbb{R}^m$ *and* $(\forall y \in \mathbb{R}^m) \quad \nabla g^\beta(y) = \frac{1}{\beta}\left(y - \operatorname{prox}_{\beta g}(y)\right)$*. Moreover,* ∇g^β *is Lipschitz-continuous with constant* $L_{g,\beta} := \max\left\{\beta^{-1}, \frac{\rho}{1-\rho\beta}\right\} \leq \frac{1}{\beta(1-\rho\beta)}$*. In particular, for every* $y \in \mathbb{R}^m$*, under* Assumption 2.4*(I),* $g^\beta(y) = \frac{1}{2\beta}\operatorname{dist}^2_{\mathcal{D}}(y)$ *and* $\nabla g^\beta(y) = \frac{1}{\beta}\left(y - \mathrm{P}_{\mathcal{D}}(y)\right)$*; under* Assumption 2.4*(II),* $\left\|\nabla g^\beta(y)\right\|_2 \leq L_g$.

Proof The differentiability of g^β, the gradient formula, and the Lipschitz estimate for ∇g^β follow from [32, Corollary 3.4] and the choice $\beta \in \left]0, \rho^{-1}\right[$. The formula under (I) follows from the identity $\operatorname{prox}_{\beta\iota_{\mathcal{D}}} = \mathrm{P}_{\mathcal{D}}$. Finally, under (II), the optimality condition for $\operatorname{prox}_{\beta g}$ yields

$$\nabla g^\beta(y) = \frac{1}{\beta}\left(y - \operatorname{prox}_{\beta g}(y)\right) \in \partial g(\operatorname{prox}_{\beta g}(y)),$$

and the Clarke subgradients of an L_g-Lipschitz function have norm at most L_g [33]. □

Proposition 3.2 (Lipschitz displacement bound). *Suppose* g *satisfies* Assumption 2.4*(II). Then, for every* $\beta \in \left]0, \rho^{-1}\right[$ *and every* $y \in \mathbb{R}^m$*,*

$$\left\|y - \operatorname{prox}_{\beta g}(y)\right\|_2 \leq \beta L_g.$$

Moreover, if $\{\beta_k\}_{k\in\mathbb{N}}$ *satisfies* Assumption 2.7*, then for every sequence* $\{x_k\}_{k\in\mathbb{N}} \subset \mathbb{R}^n$*,* $\left\|Tx_k - \operatorname{prox}_{\beta_k g}(Tx_k)\right\|_2 \leq \beta_k L_g \longrightarrow 0$.

Proof Using the expression and bounds for ∇g^β given in Proposition 3.1,

$$\left\|y - \operatorname{prox}_{\beta g}(y)\right\|_2 = \beta\left\|\nabla g^\beta(y)\right\|_2 \leq \beta L_g. \tag{3.1}$$

The second claim follows by taking $y = Tx_k$ in (3.1) and using $\beta_k \to 0$ from Assumption 2.7. □

Proposition 3.3 (Bound on $\mathrm{Id} - \mathrm{prox}_{\beta g}$ over $T(\mathcal{C})$). *Let Assumptions 2.2, 2.3, and 2.4 hold. For $\beta \in \left]0, \rho^{-1}\right[$, $\max_{z \in T(\mathcal{C})} \left\|z - \mathrm{prox}_{\beta g}(z)\right\|_2 < +\infty$. Moreover, if $\{\beta_k\}_{k\in\mathbb{N}}$ satisfies* Assumption 2.7*, then for all $k \in \mathbb{N}^*$ it holds*

$$\max_{z \in T(\mathcal{C})} \left\|z - \mathrm{prox}_{\beta_k g}(z)\right\|_2 \leq B_{\mathrm{prox}} := \begin{cases} e_{\mathcal{D},T(\mathcal{C})} & \text{if } g \text{ satisfies Assumption 2.4(I),} \\ \beta_0 L_g & \text{if } g \text{ satisfies Assumption 2.4(II).} \end{cases} \tag{3.2}$$

Proof Since $T(\mathcal{C})$ is compact and $\mathrm{prox}_{\beta g}$ is continuous for $\beta \in \left]0, \rho^{-1}\right[$ since $\mathrm{prox}_{\beta g} = \mathrm{Id} - \beta \nabla g^{\beta}$, the map $z \mapsto \left\|z - \mathrm{prox}_{\beta g}(z)\right\|_2$ attains its maximum on $T(\mathcal{C})$. This proves finiteness. Under Assumption 2.4(I), for every $\beta > 0$, $\mathrm{prox}_{\beta g} = \mathrm{P}_{\mathcal{D}}$ and $\max_{z \in T(\mathcal{C})} \|z - \mathrm{P}_{\mathcal{D}}(z)\|_2 = e_{\mathcal{D},T(\mathcal{C})}$. Under Assumption 2.4(II), Proposition 3.2 and Assumption 2.7 provide $\max_{z \in T(\mathcal{C})} \left\|z - \mathrm{prox}_{\beta g}(z)\right\|_2 \leq \beta_k L_g \leq \beta_0 L_g$ because $\{\beta_k\}_{k\in\mathbb{N}}$ is nonincreasing. □

3.2 Estimates for the Smoothed Objectives

Proposition 3.4 (Descent lemma applied to Φ_k). *Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold. For every $\beta \in \left]0, \rho^{-1}\right[$, define $\Phi_\beta(x) := f(x) + g^\beta(Tx)$ and set*

$$L_\beta := L_{\nabla f} + \|T\|_{\mathrm{op}}^2 \max\left\{\beta^{-1}, \frac{\rho}{1-\rho\beta}\right\}. \tag{3.3}$$

Then, for all $x, y \in \mathcal{C}$,

$$\Phi_\beta(y) \leq \Phi_\beta(x) + \left\langle \nabla f(x) + T^* \nabla g^\beta(Tx), y - x \right\rangle + \frac{L_\beta}{2} \|y - x\|_2^2 .$$

Proof By Assumption 2.1, ∇f is $L_{\nabla f}$-Lipschitz on $\mathcal{C}$. By Proposition 3.1, ∇g^β is Lipschitz-continuous with constant $\max\{\beta^{-1}, \frac{\rho}{1-\rho\beta}\}$. Therefore, $\nabla \Phi_\beta$ is Lipschitz-continuous on $\mathcal{C}$ with constant L_β. The standard descent lemma for $C^{1,1}$ functions then gives the claim [34]. □

Proposition 3.4 provides an estimated smoothness constant of Φ_k in (3.3):

$$L_k := L_{\nabla f} + \|T\|_{\mathrm{op}}^2 \max\left\{\beta_k^{-1}, \frac{\rho}{1-\rho\beta_k}\right\}. \tag{3.4}$$

Lemma 3.5 (Variation of Φ_k with respect to k). *Let Assumptions 2.1, 2.2, 2.3, 2.4, and 2.7 hold, let Φ_k be given by (2.2), and let B_{prox} be given by (3.2). Then,*

$$(\forall k \in \mathbb{N})(\forall x \in \mathcal{C}) \qquad \Phi_{k+1}(x) - \Phi_k(x) \leq \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right) B_{\mathrm{prox}}^2. \tag{3.5}$$

Proof For every $x \in \mathcal{C}$, $\Phi_{k+1}(x) - \Phi_k(x) = g^{\beta_{k+1}}(Tx) - g^{\beta_k}(Tx)$.

Suppose that Assumption 2.4(I) holds. Then $g^{\beta_k}(Tx) = \frac{1}{2\beta_k}\mathrm{dist}^2_{\mathcal{D}}(Tx)$, so since $\mathrm{dist}_{\mathcal{D}}(\cdot) \leq e_{\mathcal{D},T(\mathcal{C})} = B_{\mathrm{prox}}$ (see Section 2.2 and (3.2)), Assumption 2.7 yields

$$\Phi_{k+1}(x) - \Phi_k(x) = \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)\mathrm{dist}^2_{\mathcal{D}}(Tx) \leq \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)B^2_{\mathrm{prox}}.$$

Suppose now that Assumption 2.4(II) holds. By Assumption 2.7, $\beta_k \geq \beta_{k+1}$, so for every $y \in \mathbb{R}^m$, $g^{\beta_{k+1}}(y) \geq g^{\beta_k}(y)$. By substituting $\mathrm{prox}_{\beta_k g}(y)$ in the definition of $g^{\beta_{k+1}}(y)$ in (2.1), we find

$$\begin{aligned} g^{\beta_{k+1}}(y) - g^{\beta_k}(y) &\leq \left(g(\mathrm{prox}_{\beta_k g}(y)) + \frac{1}{2\beta_{k+1}}\left\|y - \mathrm{prox}_{\beta_k g}(y)\right\|_2^2\right) \\ &\quad - \left(g(\mathrm{prox}_{\beta_k g}(y)) + \frac{1}{2\beta_k}\left\|y - \mathrm{prox}_{\beta_k g}(y)\right\|_2^2\right) \\ &= \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)\left\|y - \mathrm{prox}_{\beta_k g}(y)\right\|_2^2 \\ &\leq \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)\beta_k^2 L_g^2, \end{aligned}$$

where we have used Proposition 3.2 in the last inequality. Applying this to the variation of Φ gives

$$\Phi_{k+1}(x) - \Phi_k(x) \leq \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)\beta_k^2 L_g^2 \leq \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right)B^2_{\mathrm{prox}},$$

where the last inequality uses $\beta_{k+1} \leq \beta_k$ (Assumption 2.7) and Proposition 3.3. □

Lemma 3.6 (Range bound for Φ_k over $\mathcal{C}$). *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.6, and 2.7 hold, let Φ_k be given by (2.2) and let B_{prox} be given by (3.2). Then, for all $x, y \in \mathcal{C}$*

$$(\forall k \in \mathbb{N}) \qquad \Phi_k(x) - \Phi_k(y) \leq L_f \mathrm{diam}_{\mathcal{C}} + \beta_k^{-1} B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}.$$

Proof For $x, y \in \mathcal{C}$, Section 2.2 provides

$$f(x) - f(y) \leq L_f \mathrm{diam}_{\mathcal{C}}. \tag{3.6}$$

Suppose Assumption 2.4(I) holds. Since $\nabla \mathrm{dist}^2_{\mathcal{D}}(z) = 2\left(z - \mathrm{P}_{\mathcal{D}}(z)\right)$, and for every $z \in T(\mathcal{C})$, $\|z - \mathrm{P}_{\mathcal{D}}(z)\|_2 \leq e_{\mathcal{D},T(\mathcal{C})}$, the function $\mathrm{dist}^2_{\mathcal{D}}$ is $2e_{\mathcal{D},T(\mathcal{C})}$-Lipschitz on the convex set $T(\mathcal{C})$. Hence

$$\begin{aligned} g^{\beta_k}(Tx) - g^{\beta_k}(Ty) &= \frac{1}{2\beta_k}\left(\mathrm{dist}^2_{\mathcal{D}}(Tx) - \mathrm{dist}^2_{\mathcal{D}}(Ty)\right) \\ &\leq \frac{e_{\mathcal{D},T(\mathcal{C})}}{\beta_k}\|Tx - Ty\|_2 \\ &\leq \beta_k^{-1} e_{\mathcal{D},T(\mathcal{C})} \mathrm{diam}_{T(\mathcal{C})}. \end{aligned} \tag{3.7}$$

Suppose now that Assumption 2.4(II) holds. By Proposition 3.1, for every $z \in R^m$, $\left\|\nabla g^{\beta_k}(z)\right\|_2 \leq L_g$. Thus g^{β_k} is L_g-Lipschitz on $\mathbb{R}^m$, and

$$g^{\beta_k}(Tx) - g^{\beta_k}(Ty) \leq L_g\|Tx - Ty\|_2 \leq L_g \mathrm{diam}_{T(\mathcal{C})}. \tag{3.8}$$

Since Assumption 2.7 provides $\beta_k \leq \beta_0$, this implies $g^{\beta_k}(Tx) - g^{\beta_k}(Ty) \leq \beta_k^{-1}\beta_0 L_g \mathrm{diam}_{T(\mathcal{C})}$. Combining (3.6), (3.7), and (3.8) gives the claim. □

4 Frank-Wolfe Gaps and Stationarity Certificates

If h is differentiable and the constraint set is $\mathcal{C}$, the *Frank-Wolfe gap* is

$$\max_{s\in\mathcal{C}} \langle \nabla h(x), x - s\rangle \,. \tag{4.1}$$

For $x \in \mathcal{C}$, the quantity in (4.1) is nonnegative, and it vanishes at $x^* \in C$ if and only if

$$0 \in \nabla h(x^*) + N_{\mathcal{C}}(x^*) \qquad \Leftrightarrow \qquad x^* \text{ is a stationary point of } \min_{x\in\mathcal{C}} h(x), \tag{4.2}$$

where $N_{\mathcal{C}}(x^\star)$ is the normal cone of $\mathcal{C}$ at x^*. Thus, for *smooth* nonconvex objective functions, the Frank-Wolfe gap is a natural first-order stationarity certificate [34]. However, the objective in (P) is generally *nonsmooth*. This section delineates the smoothed gaps from the more meaningful stationarity certificates for the original problem (P).

4.1 The Smoothed Frank-Wolfe Gap

For $\beta \in \left]0, \rho^{-1}\right[$ and $x \in \mathcal{C}$, we define the *smoothed Frank-Wolfe gap* by

$$\mathrm{gap}^\beta(x) := \max_{s\in\mathcal{C}} \left\langle \nabla f(x) + T^*\nabla g^\beta(Tx), x - s\right\rangle . \tag{4.3}$$

Equivalently, $\mathrm{gap}^\beta(x)$ is the usual Frank-Wolfe gap (4.1) with $h = \Phi_k$ given by (2.2). In view of (4.2), gap^β characterizes stationarity for the smoothed problem $\min_{x\in\mathcal{C}} \Phi_\beta(x)$. Although Step 2 of **FRAMES** allows one to compute $\mathrm{gap}^\beta(x_k) = \langle \nabla\Phi_k(x_k),\, x_k - s_k\rangle$, this is not a stationarity measure for (P) a priori.

4.2 A Nonsmooth Frank-Wolfe Gap for the Indicator Case

Suppose first that Assumption 2.4(I)(a) holds, so that $g = \iota_{\mathcal{D}}$ and there is a minimizer for (P). Then $\widetilde{\mathcal{C}} := \mathcal{C}\cap T^{-1}(\mathcal{D})$ is nonempty, compact, and convex, and (P) is equivalent to minimizing f over $\widetilde{\mathcal{C}}$. The natural Frank-Wolfe stationarity certificate for this problem is below.

Definition 4.1 (Signed nonsmooth Frank-Wolfe gap). Under Assumption 2.1, 2.2, 2.3, and 2.4(I)(a), the *signed nonsmooth Frank-Wolfe gap* (or *signed gap* for short) at $x \in \mathbb{R}^n$ is

$$\widetilde{\mathrm{gap}}(x) := \max_{s\in\widetilde{\mathcal{C}}} \langle \nabla f(x), x - s\rangle \,. \tag{4.4}$$

Remark 4.2. Unlike the smoothed gap (4.3), for $x \in \mathcal{C}$, $\widetilde{\mathrm{gap}}(x)$ need not be nonnegative. Although the iterates of **FRAMES** always reside in $\mathcal{C}$, they are not guaranteed to satisfy $Tx_k \in \mathcal{D}$ for finite $k \in \mathbb{N}$. Thus, $\widetilde{\mathrm{gap}}$ should be interpreted as a signed stationarity measure.

Lemma 4.3 (Indicator stationarity certificate). *Suppose* Assumption 2.4*(I)(a) holds and define the signed gap as in* Definition 4.1*. If $\bar{x} \in \widetilde{\mathcal{C}}$, then $\bar{x}$ is a stationary point of $\min_{x\in\widetilde{\mathcal{C}}} f(x)$ if and only if*

$$\widetilde{\mathrm{gap}}(\bar{x}) = 0 \Leftrightarrow 0 \in \nabla f(\bar{x}) + N_{\widetilde{\mathcal{C}}}(\bar{x}).$$

Moreover, under Assumption 2.4*(I)(b), this condition can be written*

$$0 \in \nabla f(\bar{x}) + N_{\mathcal{C}}(\bar{x}) + T^* N_{\mathcal{D}}(T\bar{x}).$$

Proof The equivalence between $\widetilde{\mathrm{gap}}(x) = 0$ and $0 \in \nabla f(x) + N_{\widetilde{\mathcal{C}}}(x)$ when $x \in \widetilde{\mathcal{C}}$ is the standard Frank-Wolfe gap characterization (4.2). The second claim follows from [35, Proposition 6.19 & Theorem 16.47] which shows $N_{\widetilde{\mathcal{C}}} = N_{\mathcal{C}} + T^* \circ N_{\mathcal{D}} \circ T$. □

The indicator case therefore necessitates showing that two quantities vanish: feasibility, measured by $\mathrm{dist}_{\mathcal{D}}(Tx)$, and stationarity over the true feasible set $\widetilde{\mathcal{C}}$, measured by $\widetilde{\mathrm{gap}}(x)$. We finish this section with a nonlinear generalization of [11, Lemma 3.5] that provides a relationship between gap^β, $\mathrm{dist}_D \circ T$, and $\widetilde{\mathrm{gap}}$.

Lemma 4.4 (Indicator gap bound). *Let* gap^β *and* $\widetilde{\mathrm{gap}}$ *as in* (4.3) *and* (4.4) *respectively. If* Assumption 2.4*(I)(a) holds, then*

$$(\forall x \in \mathcal{C})(\forall \beta > 0) \qquad \widetilde{\mathrm{gap}}(x) + \frac{1}{\beta}\mathrm{dist}^2_{\mathcal{D}}(Tx) \leq \mathrm{gap}^\beta(x). \tag{4.5}$$

Proof Since $g = \iota_{\mathcal{D}}$, Proposition 3.1 provides $\nabla(g^\beta \circ T)(x) = \frac{1}{\beta} T^* \left(Tx - \mathrm{P}_{\mathcal{D}}(Tx)\right)$. Thus,

$$\begin{aligned} -\mathrm{gap}^\beta(x) &= \min_{s \in \mathcal{C}} \left\langle \nabla f(x) + \frac{1}{\beta} T^* \big(Tx - \mathrm{P}_{\mathcal{D}}(Tx)\big),\, s - x \right\rangle \\ &\leq \min_{s \in \widetilde{\mathcal{C}}} \left\langle \nabla f(x) + \frac{1}{\beta} T^* \big(Tx - \mathrm{P}_{\mathcal{D}}(Tx)\big),\, s - x \right\rangle \\ &= \min_{s \in \widetilde{\mathcal{C}}} \left\langle \nabla f(x),\, s - x \right\rangle + \frac{1}{\beta} \langle Tx - \mathrm{P}_{\mathcal{D}}(Tx),\, Ts - Tx \rangle. \end{aligned} \tag{4.6}$$

For every $s \in \widetilde{\mathcal{C}}$, one has $Ts \in \mathcal{D}$. By [35, Theorem 3.16], $\langle Tx - \mathrm{P}_{\mathcal{D}}(Tx), Ts - \mathrm{P}_{\mathcal{D}}(Tx) \rangle \leq 0$. Consequently, for all $s \in \widetilde{\mathcal{C}}$,

$$\begin{aligned} \langle Tx - \mathrm{P}_{\mathcal{D}}(Tx), Ts - Tx \rangle &= \langle Tx - \mathrm{P}_{\mathcal{D}}(Tx), Ts - \mathrm{P}_{\mathcal{D}}(Tx) \rangle - \|Tx - \mathrm{P}_{\mathcal{D}}(Tx)\|_2^2 \\ &\leq -\mathrm{dist}^2_{\mathcal{D}}(Tx). \end{aligned} \tag{4.7}$$

Combining (4.6) and (4.7) gives $-\mathrm{gap}^\beta(x) \leq -\widetilde{\mathrm{gap}}(x) - \frac{1}{\beta}\mathrm{dist}^2_{\mathcal{D}}(Tx)$, i.e., (4.5) holds. □

4.3 A Nonsmooth Frank-Wolfe Gap for Lipschitz Weakly Convex g

We now suppose that Assumption 2.4(II) holds, which implies that g has full domain and is locally Lipschitz. Since g is weakly convex and defined on a finite-dimensional space, it is also *subdifferentially regular* in the sense of [31, Definition 7.25], since g is the sum of two subdifferentially regular functions (the convex function [31, Example 7.27] $h(y) := g(y) + \frac{\rho}{2}\|y\|_2^2$ and the twice continuously differentiable function $y \mapsto -\frac{\rho}{2}\|y\|_2^2$) satisfying the usual qualification conditions [31, Corollary 10.9]. Moreover, for locally Lipschitz subdifferentially regular functions, the regular, limiting, and Clarke subdifferentials are compatible in the sense of [31, Definition 7.25, Corollary 8.11, Theorem 8.49]. Thus, throughout this case, we write simply ∂g for the Clarke subdifferential.

We first record the stationarity condition for (P) in this setting.

Proposition 4.5 (Stationarity condition in the Lipschitz weakly convex case). *Suppose* Assumption 2.4*(II) holds. Then, for every* $x \in \mathcal{C}$*,*

$$\partial \left(f + g \circ T + \iota_{\mathcal{C}}\right)(x) = \nabla f(x) + T^* \partial g(Tx) + N_{\mathcal{C}}(x), \tag{4.8}$$

where ∂g *denotes the Clarke subdifferential. As a result,* $x \in \mathcal{C}$ *is a stationary point of* (P) *if and only if*

$$0 \in \nabla f(x) + T^* \partial g(Tx) + N_{\mathcal{C}}(x). \tag{4.9}$$

Equivalently, $x \in \mathcal{C}$ *is stationary if and only if there exists* $\xi \in \partial g(Tx)$ *such that*

$$-\left(\nabla f(x) + T^* \xi\right) \in N_{\mathcal{C}}(x). \tag{4.10}$$

Proof Since $f \in C^1$ on a neighborhood of $\mathcal{C}$, the smooth sum rule (e.g., [31, Exercise 10.10]) gives

$$\partial \left(f + g \circ T + \iota_{\mathcal{C}}\right)(x) = \nabla f(x) + \partial \left(g \circ T + \iota_{\mathcal{C}}\right)(x). \tag{4.11}$$

Since $\iota_{\mathcal{C}}$ is subdifferentially regular at every $x \in \mathcal{C}$ [31, Example 7.27], and $g \circ T$ is subdifferentially regular because g is subdifferentially regular and T is linear, the sum rule for regular functions [31, Corollary 10.9] gives

$$\partial \left(g \circ T + \iota_{\mathcal{C}}\right)(x) = \partial (g \circ T)(x) + N_{\mathcal{C}}(x). \tag{4.12}$$

Finally, the Clarke chain rule for composition with a linear map gives

$$\partial (g \circ T)(x) = T^* \partial g(Tx), \tag{4.13}$$

see [33, Theorem 2.3.10]. Combining (4.11), (4.12), and (4.13) yields (4.8). The stationarity condition (4.9) follows from the definition of a stationary point, and (4.10) is the same inclusion written with a selected subgradient $\xi \in \partial g(Tx)$. □

Proposition 4.5 motivates the following nonsmooth notion of a Frank-Wolfe gap.

Definition 4.6 (Nonsmooth Frank-Wolfe subgradient gap). Under Assumption 2.1, 2.2, 2.3, and 2.4(II), we define the *nonsmooth Frank-Wolfe subgradient gap* (or *subgradient gap* for short) at $x \in \mathbb{R}^n$ with subgradient $\xi \in \partial g(Tx)$ to be

$$\operatorname{gap}(x; \xi) := \max_{s \in \mathcal{C}} \left\langle \nabla f(x) + T^* \xi, x - s \right\rangle.$$

Lemma 4.7 (Lipschitz weakly convex stationarity certificate). *Suppose* Assumption 2.4*(II) holds and define the subgradient gap as in* Definition 4.6*. A point* $x \in \mathcal{C}$ *satisfies the stationarity condition*

$$0 \in \nabla f(x) + T^* \partial g(Tx) + N_{\mathcal{C}}(x) \tag{4.14}$$

if and only if there exists $\xi \in \partial g(Tx)$ *such that*

$$\operatorname{gap}(x; \xi) = 0.$$

Proof Fix $x \in \mathcal{C}$ and $\xi \in \partial g(Tx)$. Since $x \in \mathcal{C}$, the choice $s = x$ gives $\operatorname{gap}(x; \xi) \geq 0$. Moreover, $\operatorname{gap}(x; \xi) = 0 \Leftrightarrow \forall s \in \mathcal{C} \;\; \langle \nabla f(x) + T^* \xi, x - s \rangle \leq 0 \Leftrightarrow \forall s \in \mathcal{C} \;\; \langle -(\nabla f(x) + T^* \xi), s - x \rangle \leq 0$, which is precisely the condition $-(\nabla f(x) + T^* \xi) \in N_{\mathcal{C}}(x)$. Thus, $\operatorname{gap}(x; \xi) = 0$ for some $\xi \in \partial g(Tx)$ if and only if (4.14) holds. □

The smoothed gap and the nonsmooth gap are related but not identical. Indeed, the Moreau envelope produces the vector

$$\xi_\beta(x) := \nabla g^\beta(Tx) = \frac{Tx - \mathrm{prox}_{\beta g}(Tx)}{\beta}.$$

By the proximal optimality condition, $\xi_\beta(x) \in \partial g(\mathrm{prox}_{\beta g}(Tx))$. Thus, $\xi_\beta(x)$ is a Clarke subgradient of g at $\mathrm{prox}_{\beta g}(Tx)$, not necessarily at Tx. Due to this, $\mathrm{gap}^\beta(x)$ is not simply the subgradient gap, $\mathrm{gap}(x;\xi)$, evaluated at $\xi = \xi_\beta(x)$. This mismatch is one of the main motivators for the gap-transfer analysis in Section 6.

Remark 4.8 (The two nonsmooth gaps are distinct). The signed gap given in Definition 4.1 and the subgradient gap given in Definition 4.6 are different stationarity certificates. Using Definition 4.6 under Assumption 2.4(I) would require an element of $\partial \iota_{\mathcal{D}}(Tx) = N_{\mathcal{D}}(Tx)$, which is empty whenever $Tx \notin \mathcal{D}$. In view of Remark 4.2, since **FRAMES** may generate iterates satisfying $x_k \in \mathcal{C}$ but $Tx_k \notin \mathcal{D}$, the subgradient gap is not informative, as it would be identically $+\infty$. On the other hand, Definition 4.1 provides a finite signed quantity that can be analyzed in tandem with the feasibility measure $\mathrm{dist}_{\mathcal{D}}(Tx_k)$.

5 Smoothed Convergence Rates and Asymptotic Stationarity

In this section, we prove that the average and best smoothed gaps generated by **FRAMES** converge to zero at an explicit rate. The main difference from the standard smooth Frank-Wolfe analysis is that the objective changes with k, since the smoothing parameter β_k is decreasing to zero. The estimates from Section 3 allow us to control both the usual Frank-Wolfe descent term and the variation of the smoothed objectives Φ_k.

5.1 An Energy Estimate for the Smoothed Gaps

We begin with the basic descent estimate for one iteration of **FRAMES**, using the definition of smoothed Frank-Wolfe gap from (4.3) throughout.

Lemma 5.1 (One-step smoothed descent). *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.6, and 2.7 hold and let $\{x_k\}_{k\in\mathbb{N}}$ be generated by **FRAMES**. Then, for every $k \in \mathbb{N}$,*

$$\Phi_k(x_{k+1}) \le \Phi_k(x_k) - \gamma_k \mathrm{gap}^{\beta_k}(x_k) + \frac{L_k \gamma_k^2}{2} \mathrm{diam}_{\mathcal{C}}^2, \tag{5.1}$$

where L_k is the estimated smoothness constant defined in (3.4). Moreover,

$$\Phi_{k+1}(x_{k+1}) \le \Phi_k(x_k) - \gamma_k \mathrm{gap}^{\beta_k}(x_k) + \frac{L_k \gamma_k^2}{2} \mathrm{diam}_{\mathcal{C}}^2 + \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right) B_{\mathrm{prox}}^2. \tag{5.2}$$

where B_{prox} is defined in (3.2).

Proof By Proposition 3.4, applied to Φ_k at x_k and x_{k+1},

$$\Phi_k(x_{k+1}) \leq \Phi_k(x_k) + \langle \nabla \Phi_k(x_k), x_{k+1} - x_k \rangle + \frac{L_k}{2} \|x_{k+1} - x_k\|_2^2 .$$

Since $x_{k+1} = x_k + \gamma_k(s_k - x_k)$, we have

$$\langle \nabla \Phi_k(x_k), x_{k+1} - x_k \rangle = -\gamma_k \langle \nabla \Phi_k(x_k), x_k - s_k \rangle = -\gamma_k \mathrm{gap}^{\beta_k}(x_k),$$

where the last equality follows from the definition of s_k in **FRAMES**. Moreover, since $x_k, s_k \in \mathcal{C}$,

$$\|x_{k+1} - x_k\|_2 = \gamma_k \|s_k - x_k\|_2 \leq \gamma_k \mathrm{diam}_{\mathcal{C}}.$$

This proves (5.1). Adding (3.5) of Lemma 3.5 to (5.1) yields (5.2). □

Summing Lemma 5.1 over $k \in \mathbb{N}$ yields the main estimate used to find convergence rates.

Lemma 5.2 (Energy estimate for smoothed gaps). *Let Assumptions 2.1, 2.2, 2.3, 2.4, 2.6, and 2.7 hold and let $\{x_k\}_{k\in\mathbb{N}}$ be generated by **FRAMES**. Then, for every $N \in \mathbb{N}^*$,*

$$0 \leq \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \leq \frac{L_f \mathrm{diam}_{\mathcal{C}} + B\beta_N^{-1}}{\gamma_{N-1}} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \sum_{k=0}^{N-1} \gamma_k \left(L_{\nabla f} + \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0} \beta_k^{-1} \right),$$

where $B := \max\{\frac{1}{2} B_{\mathrm{prox}}^2,\ B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}\}$ and $M_{\rho,\beta_0} := \max\{1, \frac{\rho\beta_0}{1-\rho\beta_0}\}$.

Proof For each $k \in \mathbb{N}$, define $\Phi_k^\star := \min_{x\in\mathcal{C}} \Phi_k(x)$ and $r_k := \Phi_k(x_k) - \Phi_k^\star$. By Lemma 5.1, for every $k \in \mathbb{N}$,

$$\gamma_k \mathrm{gap}^{\beta_k}(x_k) \leq \Phi_k(x_k) - \Phi_{k+1}(x_{k+1}) + \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right) B_{\mathrm{prox}}^2 + \frac{L_k \gamma_k^2}{2} \mathrm{diam}_{\mathcal{C}}^2.$$

Since $\beta_{k+1} \leq \beta_k$, the Moreau envelopes are pointwise nondecreasing in k, and hence $\forall x \in \mathbb{R}^n$ $\Phi_k(x) \leq \Phi_{k+1}(x)$. Therefore $\Phi_k^\star \leq \Phi_{k+1}^\star$, and so $\Phi_k(x_k) - \Phi_{k+1}(x_{k+1}) \leq r_k - r_{k+1}$. Thus,

$$\gamma_k \mathrm{gap}^{\beta_k}(x_k) \leq r_k - r_{k+1} + \frac{1}{2}\left(\beta_{k+1}^{-1} - \beta_k^{-1}\right) B_{\mathrm{prox}}^2 + \frac{L_k \gamma_k^2}{2} \mathrm{diam}_{\mathcal{C}}^2.$$

Dividing by γ_k and summing from $k = 0$ to $N-1$ gives

$$\sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \leq \sum_{k=0}^{N-1} \frac{r_k - r_{k+1}}{\gamma_k} + \frac{1}{2} B_{\mathrm{prox}}^2 \sum_{k=0}^{N-1} \frac{\beta_{k+1}^{-1} - \beta_k^{-1}}{\gamma_k} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \sum_{k=0}^{N-1} \gamma_k L_k.$$

We now bound the first two sums. By summation by parts,

$$\sum_{k=0}^{N-1} \frac{r_k - r_{k+1}}{\gamma_k} = \frac{r_0}{\gamma_0} + \sum_{k=1}^{N-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) r_k - \frac{r_N}{\gamma_{N-1}}.$$

Since $r_N \geq 0$ and $\{\gamma_k\}_{k\in\mathbb{N}}$ is nonincreasing, we get

$$\sum_{k=0}^{N-1} \frac{r_k - r_{k+1}}{\gamma_k} \leq \frac{r_0}{\gamma_0} + \sum_{k=1}^{N-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}} \right) r_k.$$

By Lemma 3.6, for every $k \in \mathbb{N}$, $r_k = \Phi_k(x_k) - \Phi_k^\star \le L_f \mathrm{diam}_{\mathcal{C}} + \beta_k^{-1} B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}$. Using $B := \max\{\frac{1}{2} B_{\mathrm{prox}}^2,\ B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}\}$, we obtain

$$\sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \le \frac{L_f \mathrm{diam}_{\mathcal{C}} + B\beta_0^{-1}}{\gamma_0} + \sum_{k=1}^{N-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) \left(L_f \mathrm{diam}_{\mathcal{C}} + B\beta_k^{-1}\right) + B \sum_{k=0}^{N-1} \frac{\beta_{k+1}^{-1} - \beta_k^{-1}}{\gamma_k} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \sum_{k=0}^{N-1} \gamma_k L_k.$$

The terms involving $L_f \mathrm{diam}_{\mathcal{C}}$ telescope:

$$\frac{L_f \mathrm{diam}_{\mathcal{C}}}{\gamma_0} + \sum_{k=1}^{N-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) L_f \mathrm{diam}_{\mathcal{C}} = \frac{L_f \mathrm{diam}_{\mathcal{C}}}{\gamma_{N-1}}.$$

The terms involving B also telescope, giving

$$\frac{B\beta_0^{-1}}{\gamma_0} + B \sum_{k=1}^{N-1} \left(\frac{1}{\gamma_k} - \frac{1}{\gamma_{k-1}}\right) \beta_k^{-1} + B \sum_{k=0}^{N-1} \frac{\beta_{k+1}^{-1} - \beta_k^{-1}}{\gamma_k} = \frac{B\beta_N^{-1}}{\gamma_{N-1}}.$$

Combining these estimates yields

$$\sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \le \frac{L_f \mathrm{diam}_{\mathcal{C}} + B\beta_N^{-1}}{\gamma_{N-1}} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \sum_{k=0}^{N-1} \gamma_k L_k. \tag{5.3}$$

Since $\beta_k \le \beta_0$, we have $\frac{\rho}{1-\rho\beta_k} \le \frac{\rho\beta_0}{1-\rho\beta_0} \beta_k^{-1}$, and therefore $\max\left\{\beta_k^{-1}, \frac{\rho}{1-\rho\beta_k}\right\} \le M_{\rho,\beta_0} \beta_k^{-1}$. Consequently, $L_k \le L_{\nabla f} + \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0} \beta_k^{-1}$; with (5.3), this shows the upper bound. The lower bound follows from nonnegativity of the smoothed gap and the fact that $x_k \in \mathcal{C}$. □

5.2 Convergence Rate for the Smoothed Gaps

We now specialize the energy estimate to specific open-loop schedules in order to derive rates.

Theorem 5.3 (smoothed gap rates for power schedules). *Fix $p, q \in]0,1[$ satisfying $q < p$ and $p + q < 1$ and set $\gamma_k = (k+1)^{-p}$ and $\beta_k = \beta_0 (k+1)^{-q}$ with $\beta_0 \in]0, \rho^{-1}[$. Let Assumptions 2.1, 2.2, 2.3, and 2.4 hold and let $\{x_k\}_{k\in\mathbb{N}}$ be generated by* ***FRAMES****. Then, there exists some $C_{p,q} \ge 0$, defined in (5.11), such that, for every $N \in \mathbb{N}^*$,*

$$0 \le \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \le C_{p,q} N^{-\min\{p-q,\ 1-p-q\}}. \tag{5.4}$$

Consequently,

$$0 \le \min_{0\le k\le N-1} \mathrm{gap}^{\beta_k}(x_k) \le C_{p,q} N^{-\min\{p-q,\ 1-p-q\}}. \tag{5.5}$$

In particular, for $p = 1/2$ and $q = 1/4$ one has $\min\{p-q,\ 1-p-q\} = 1/4$ and a constant $C \ge 0$, defined in (5.12), independent of N such that

$$0 \le \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \le C N^{-1/4} \qquad \textit{and} \qquad 0 \le \min_{0\le k\le N-1} \mathrm{gap}^{\beta_k}(x_k) \le C N^{-1/4}.$$

Proof The chosen schedules satisfy Assumptions 2.6 and 2.7. By Lemma 5.2, we have

$$0 \leq \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \leq \frac{L_f \mathrm{diam}_{\mathcal{C}} + B\beta_N^{-1}}{\gamma_{N-1}} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \sum_{k=0}^{N-1} \gamma_k \left(L_{\nabla f} + \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0} \beta_k^{-1} \right), \quad (5.6)$$

where $B := \max\{\frac{1}{2} B_{\mathrm{prox}}^2,\ B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}\}$. For the power schedules, $\gamma_{N-1}^{-1} = N^p$ and $\beta_N^{-1} = \beta_0^{-1}(N+1)^q$. Since $N \geq 1$ and $q \in]0,1[$, we have $(N+1)^q \leq 2^q N^q$ and $N^{p-1} \leq N^{p+q-1}$. Therefore

$$\begin{aligned} \frac{L_f \mathrm{diam}_{\mathcal{C}} + B\beta_N^{-1}}{N\gamma_{N-1}} &= L_f \mathrm{diam}_{\mathcal{C}} N^{p-1} + B\beta_0^{-1} N^{p-1} (N+1)^q \\ &\leq L_f \mathrm{diam}_{\mathcal{C}} N^{p+q-1} + 2^q B \beta_0^{-1} N^{p+q-1} \\ &= \left(L_f \mathrm{diam}_{\mathcal{C}} + 2^q B \beta_0^{-1} \right) N^{-(1-p-q)}. \end{aligned} \quad (5.7)$$

It remains to consider the second term in the right hand side of (5.6). Since $\gamma_k = (k+1)^{-p}$ and $\beta_k^{-1} = \beta_0^{-1}(k+1)^q$ we obtain that $\frac{1}{N}\sum_{k=0}^{N-1} \gamma_k (L_{\nabla f} + \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_k^{-1})$ is equal to

$$\frac{L_{\nabla f}}{N} \sum_{k=0}^{N-1} (k+1)^{-p} + \frac{\|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}}{N} \sum_{k=0}^{N-1} (k+1)^{-(p-q)}. \quad (5.8)$$

Since $p \in]0,1[$ and $p-q \in]0,1[$, the integral test gives

$$\sum_{k=0}^{N-1} (k+1)^{-p} = \sum_{j=1}^{N} j^{-p} \leq \frac{N^{1-p}}{1-p} \text{ and } \sum_{k=0}^{N-1} (k+1)^{-(p-q)} = \sum_{j=1}^{N} j^{-(p-q)} \leq \frac{N^{1-p+q}}{1-p+q}. \quad (5.9)$$

Combining the expression (5.8) with (5.9) and the fact $N^{-p} \leq N^{-(p-q)}$ yields

$$\begin{aligned} \frac{1}{N} \sum_{k=0}^{N-1} \gamma_k \left(L_{\nabla f} + \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_k^{-1} \right) &\leq \frac{L_{\nabla f}}{1-p} N^{-p} + \frac{\|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}}{1-p+q} N^{-(p-q)} \\ &\leq \left(\frac{L_{\nabla f}}{1-p} + \frac{\|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}}{1-p+q} \right) N^{-(p-q)}. \end{aligned} \quad (5.10)$$

Dividing (5.6) by N and using (5.7) and (5.10), we get

$$\begin{aligned} 0 \leq \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k) \leq \left(L_f \mathrm{diam}_{\mathcal{C}} + 2^q B \beta_0^{-1} \right) N^{-(1-p-q)} + \\ \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \left(\frac{L_{\nabla f}}{1-p} + \frac{\|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}}{1-p+q} \right) N^{-(p-q)}. \end{aligned}$$

Hence (5.4) holds, where $B := \max\{\frac{1}{2} B_{\mathrm{prox}}^2,\ B_{\mathrm{prox}} \mathrm{diam}_{T(\mathcal{C})}\}$, $M_{\rho,\beta_0} = \max\{1,\ \frac{\rho\beta_0}{1-\rho\beta_0}\}$, and

$$C_{p,q} := L_f \mathrm{diam}_{\mathcal{C}} + 2^q B \beta_0^{-1} + \frac{\mathrm{diam}_{\mathcal{C}}^2}{2} \left(\frac{L_{\nabla f}}{1-p} + \frac{\|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}}{1-p+q} \right). \quad (5.11)$$

The best-gap estimate (5.5) follows from nonnegativity of the smoothed gaps

$$0 \leq \min_{0 \leq k \leq N-1} \mathrm{gap}^{\beta_k}(x_k) \leq \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{gap}^{\beta_k}(x_k).$$

Finally, substituting $p = 1/2$ and $q = 1/4$ gives the claimed $N^{-1/4}$ bounds with

$$C := L_f \mathrm{diam}_{\mathcal{C}} + 2^{1/4} B \beta_0^{-1} + L_{\nabla f} \mathrm{diam}_{\mathcal{C}}^2 + \frac{2\mathrm{diam}_{\mathcal{C}}^2}{3} \|T\|_{\mathrm{op}}^2 M_{\rho,\beta_0}\beta_0^{-1}. \quad (5.12)$$

□

Remark 5.4 (Smoothed gaps versus nonsmooth certificates). For fixed q, the smoothed gap exponent in Theorem 5.3 is maximized by balancing $p - q = 1 - p - q$, which gives $p = 1/2$ and the smoothed gap exponent $1/2 - q$. Thus, if one only tries to optimize the smoothed gaps, it is tempting to choose q very small, obtaining a smoothed gap rate close to $N^{-1/2}$. This perspective is natural in smoothing-based analyses, as was done in [11, 13].

However, the smoothed gap does not directly certify stationarity for (P). In order to transfer the convergence rate of the smoothed gaps to determine a convergence rate for the nonsmooth gaps (see Section 4), it turns out the nonsmooth convergence rate also depends on the smoothing error $\beta_k = \mathcal{O}(k^{-q})$, as shown later in Section 6. Hence a schedule that is favorable for the smoothed gaps alone can smooth too slowly for the final nonsmooth stationarity certificate. Later, Remark 6.16 demonstrates $p = 1/2$ and $q = 1/4$ to be optimal.

Corollary 5.5 (Last-half best iterate). *Under the same assumptions as* Theorem 5.3, *let* $N \geq 2$, $p = 1/2$, $q = 1/4$, *and choose*

$$k_N^\star \in \operatorname*{argmin}_{\lfloor N/2 \rfloor \leq k \leq N-1} \operatorname{gap}^{\beta_k}(x_k).$$

Then,

$$\operatorname{gap}^{\beta_{k_N^\star}}(x_{k_N^\star}) \leq 2CN^{-1/4} \quad \textit{and} \quad \beta_{k_N^\star} \leq 2^{1/4}\beta_0 N^{-1/4}.$$

Proof Since the set $\{\lfloor N/2 \rfloor, \ldots, N-1\}$ has cardinality at least $N/2$, we have

$$\operatorname{gap}^{\beta_{k_N^\star}}(x_{k_N^\star}) \leq \frac{2}{N} \sum_{k=\lfloor N/2 \rfloor}^{N-1} \operatorname{gap}^{\beta_k}(x_k) \leq \frac{2}{N} \sum_{k=0}^{N-1} \operatorname{gap}^{\beta_k}(x_k) \leq 2CN^{-1/4}.$$

Moreover, since $k_N^\star \geq \lfloor N/2 \rfloor$,

$$\beta_{k_N^\star} = \beta_0 (k_N^\star + 1)^{-1/4} \leq \beta_0 \left(\lfloor N/2 \rfloor + 1\right)^{-1/4} \leq 2^{1/4}\beta_0 N^{-1/4}.$$

□

5.3 Subsequential Convergence to Stationary Points of (P)

Corollary 5.6 (Smoothed gap-vanishing subsequences). *Under the assumptions of* Theorem 5.3, *there exists an increasing sequence* $\{k_j\}_{j \in \mathbb{N}} \subset \mathbb{N}$ *such that*

$$x_{k_j} \to \bar{x} \in \mathcal{C} \quad \textit{and} \quad \operatorname{gap}^{\beta_{k_j}}(x_{k_j}) \to 0.$$

Proof Since the smoothed gaps are nonnegative and their average converges to zero by Theorem 5.3, we have

$$\liminf_{k \to +\infty} \operatorname{gap}^{\beta_k}(x_k) = 0.$$

Thus, there exists an increasing sequence $\{k_j\}_{j \in \mathbb{N}}$ such that

$$\operatorname{gap}^{\beta_{k_j}}(x_{k_j}) \to 0.$$

The compactness of $\mathcal{C}$ and the fact that $x_k \in \mathcal{C}$ for every k imply that $\{x_{k_j}\}_{j \in \mathbb{N}}$ has a convergent subsequence; picking this subsequence and relabeling indices gives the final result. □

Theorem 5.3 shows that **FRAMES** controls the computable smoothed certificates $\mathrm{gap}^{\beta_k}(x_k)$ with few assumptions. The convergence results obtained are in fact enough to ensure stationarity for (P) asymptotically, which we show next, starting with the indicator case.

Theorem 5.7 (Indicator subsequential stationarity). *Suppose* Assumption 2.1, 2.2, 2.3, *and* 2.4(I)(a) *hold and suppose that* $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{C}$ *with* $\{\alpha_j\}_{j\in\mathbb{N}} \subset]0, +\infty[$ *define a convergent smoothed gap-vanishing sequence, i.e.,*

$$\alpha_j \to 0, \qquad u_j \to \bar{u}, \qquad \textit{and} \qquad \mathrm{gap}^{\alpha_j}(u_j) \to 0.$$

Then $\bar{u} \in \widetilde{\mathcal{C}}$ *and* $\widetilde{\mathrm{gap}}(\bar{u}) = 0$, *using the definition of signed gap given in* Definition 4.1. *Equivalently,* $0 \in \nabla f(\bar{u}) + N_{\widetilde{\mathcal{C}}}(\bar{u})$. *Under* Assumption 2.4(I)(b), *this can be rewritten as*

$$0 \in \nabla f(\bar{u}) + N_{\mathcal{C}}(\bar{u}) + T^* N_{\mathcal{D}}(T\bar{u}).$$

Proof Since $u_j \in \mathcal{C}$ for every $j \in \mathbb{N}^*$ and $\mathcal{C}$ is compact, the limit satisfies $\bar{u} \in \mathcal{C}$.

We first prove feasibility. By Lemma 4.4,

$$\widetilde{\mathrm{gap}}(u_j) + \alpha_j^{-1}\mathrm{dist}_{\mathcal{D}}^2(Tu_j) \le \mathrm{gap}^{\alpha_j}(u_j).$$

For every $u \in \mathcal{C}$ and every $s \in \widetilde{\mathcal{C}}$, $\widetilde{\mathrm{gap}}(u) \ge \langle \nabla f(u), u - s\rangle \ge -L_f \mathrm{diam}_{\mathcal{C}}$. and therefore

$$\alpha_j^{-1}\mathrm{dist}_{\mathcal{D}}^2(Tu_j) \le \mathrm{gap}^{\alpha_j}(u_j) + L_f \mathrm{diam}_{\mathcal{C}}.$$

Since $\alpha_j \to 0$ and $\mathrm{gap}^{\alpha_j}(u_j) \to 0$, it follows that $\mathrm{dist}_{\mathcal{D}}(Tu_j) \to 0$. By continuity of $u \mapsto \mathrm{dist}_{\mathcal{D}}(Tu)$, we get $\mathrm{dist}_{\mathcal{D}}(T\bar{u}) = 0$. Thus $T\bar{u} \in \mathcal{D}$, and so $\bar{u} \in \widetilde{\mathcal{C}}$.

We now prove stationarity over $\widetilde{\mathcal{C}}$. From Lemma 4.4, $\widetilde{\mathrm{gap}}(u_j) \le \mathrm{gap}^{\alpha_j}(u_j)$. Taking the limit superior gives $\limsup_{j\to+\infty} \widetilde{\mathrm{gap}}(u_j) \le 0$. Since $\widetilde{\mathcal{C}}$ is compact and ∇f is continuous, the function $\widetilde{\mathrm{gap}}$ is continuous. Hence $\widetilde{\mathrm{gap}}(\bar{u}) \le 0$. But $\bar{u} \in \widetilde{\mathcal{C}}$, so choosing $s = \bar{u}$ in the definition of $\widetilde{\mathrm{gap}}$ gives $\widetilde{\mathrm{gap}}(\bar{u}) \ge 0$. Therefore $\widetilde{\mathrm{gap}}(\bar{u}) = 0$. The equivalence with the normal cone stationarity condition follows from Lemma 4.3. □

Theorem 5.8 (Lipschitz weakly convex subsequential stationarity). *Suppose* Assumption 2.1, 2.2, 2.3, *and* 2.4(II) *hold and suppose that* $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{C}$ *with* $\{\alpha_j\}_{j\in\mathbb{N}} \subset]0, \rho^{-1}[$ *define a convergent smoothed gap-vanishing sequence, i.e.,*

$$\alpha_j \to 0, \qquad u_j \to \bar{u}, \qquad \textit{and} \qquad \mathrm{gap}^{\alpha_j}(u_j) \to 0.$$

Then, using the definition of subgradient gap in Definition 4.6, *there exists* $\bar{\xi} \in \partial g(T\bar{u})$ *such that* $\mathrm{gap}(\bar{u}, \bar{\xi}) = 0$ *and thus*

$$0 \in \nabla f(\bar{u}) + T^*\bar{\xi} + N_{\mathcal{C}}(\bar{u}), \quad \textit{i.e.,} \quad 0 \in \nabla f(\bar{u}) + T^*\partial g(T\bar{u}) + N_{\mathcal{C}}(\bar{u}).$$

Proof For each $j \in \mathbb{N}$, define $\xi_j := \nabla g^{\alpha_j}(Tu_j)$. By Proposition 3.2,

$$\left\| Tu_j - \mathrm{prox}_{\alpha_j g}(Tu_j) \right\|_2 \le \alpha_j L_g \to 0. \tag{5.13}$$

Since $u_j \to \bar{u}$ and T is linear, this implies $\operatorname{prox}_{\alpha_j g}(Tu_j) \to T\bar{u}$. Moreover, Proposition 3.1 gives

$$\|\xi_j\|_2 \le L_g \qquad \forall j \in \mathbb{N}.$$

Hence, after passing to a subsequence if necessary, there exists $\bar{\xi} \in \mathbb{R}^m$ such that $\xi_j \to \bar{\xi}$. For every j, the proximal optimality condition gives

$$\xi_j \in \partial g(\operatorname{prox}_{\alpha_j g}(Tu_j)).$$

Since (5.13) yields $\operatorname{prox}_{\alpha_j g}(Tu_j) \to T\bar{u}$, $\xi_j \to \bar{\xi}$, and since the Clarke subdifferential is outer semicontinuous for locally Lipschitz functions [33, Proposition 2.1.5], we obtain $\bar{\xi} \in \partial g(T\bar{u})$.

It remains to show the normal cone condition. For every $s \in \mathcal{C}$, (4.3) gives

$$\langle \nabla f(u_j) + T^* \xi_j, u_j - s \rangle \le \operatorname{gap}^{\alpha_j}(u_j).$$

Passing to the limit along the chosen subsequence yields

$$\langle \nabla f(\bar{u}) + T^* \bar{\xi}, \bar{u} - s \rangle \le 0 \qquad \forall s \in \mathcal{C}.$$

Equivalently, $-\left(\nabla f(\bar{u}) + T^* \bar{\xi}\right) \in N_{\mathcal{C}}(\bar{u})$. Thus $0 \in \nabla f(\bar{u}) + T^* \bar{\xi} + N_{\mathcal{C}}(\bar{u})$ as claimed. □

Combining Corollary 5.6 with Theorem 5.7 and Theorem 5.8 gives the following.

Corollary 5.9 (Existence of stationary cluster points). *Under the assumptions of* Theorem 5.3 *with* Assumption 2.4*(I)(a) if* $g = \iota_{\mathcal{D}}$*, the sequence* $\{x_k\}_{k\in\mathbb{N}}$ *generated by* ***FRAMES*** *admits a convergent smoothed gap-vanishing subsequence with limit point* $\bar{x} \in \mathcal{C}$ *that is stationary for* (P)*, i.e.,*

$$\begin{cases} 0 \in \nabla f(\bar{x}) + N_{\widetilde{\mathcal{C}}}(\bar{x}) & \text{under Assumption 2.4(I)(a),} \\ 0 \in \nabla f(\bar{x}) + T^* \partial g(T\bar{x}) + N_{\mathcal{C}}(\bar{x}) & \text{under Assumption 2.4(II) or Assumption 2.4(I)(b).} \end{cases}$$

The next section explains how this smoothed information can be transferred to stronger, finite-time guarantees on the nonsmooth stationarity certificates introduced in Section 4 under slightly stronger assumptions.

6 Transferring Convergence Rates from Smoothed Gaps to Nonsmooth Gaps

Theorem 5.3 shows that the average and best smoothed gaps converge to zero at the rate $\mathcal{O}(N^{-1/4})$. However, as discussed in Section 4, $\operatorname{gap}^{\beta_k}(x_k)$ certifies stationarity only for the smoothed problem with objective $\Phi_k = f + g^{\beta_k} \circ T$. In this section, we relate this smoothed information to stationarity certificates for the original nonsmooth problem (P). The analysis naturally splits into the two cases of Assumption 2.4.

6.1 The Indicator Case

Throughout this subsection, we suppose that Assumption 2.4(I) holds so that $g = \iota_{\mathcal{D}}$ with either (a) or (b) as well. In Section 5 we showed that Lemma 4.4 was enough to prove qualitative stationarity of smoothed gap-vanishing subsequences for (P). To obtain explicit finite-time rates for feasibility and for the signed gap, we use the following error-bound.

Assumption 6.1 (Indicator error bound). *There exists a constant* $\kappa_{\mathrm{ind}} > 0$ *such that, for every* $x \in \mathcal{C}$*,*

$$\mathrm{dist}_{\widetilde{\mathcal{C}}}(x) \le \kappa_{\mathrm{ind}} \mathrm{dist}_{\mathcal{D}}(Tx).$$

Proposition 6.2. *Suppose* Assumption 2.2*, 2.3, and 2.4(I)(b) hold and assume* T *is surjective. Then,* Assumption 6.1 *holds.*

Proof Fix $\bar{y} \in \mathrm{ri}(\mathcal{D}) \cap \mathrm{ri}(T(\mathcal{C}))$. Since $T(\mathrm{ri}(\mathcal{C})) = \mathrm{ri}(T(\mathcal{C}))$ under our assumptions, there exists $\bar{x} \in \mathrm{ri}(\mathcal{C})$ such that $T\bar{x} = \bar{y}$. By Assumption 2.4(I)(b), it holds $\mathrm{ri}(T(\mathcal{C})) \cap \mathcal{D} \neq \varnothing$ and thus $\mathrm{range}(T) \cap \mathrm{ri}(\mathcal{D}) \neq \varnothing$ since $\mathrm{ri}(T(\mathcal{C})) \subset \mathrm{range}(T)$. From this we deduce that $T^{-1}(\mathrm{ri}(\mathcal{D})) = \mathrm{ri}(T^{-1}(\mathcal{D}))$ and therefore $\bar{x} \in \mathrm{ri}(T^{-1}(\mathcal{D}))$. Hence, $\mathrm{ri}(\mathcal{C}) \cap \mathrm{ri}(T^{-1}(\mathcal{D})) \neq \varnothing$. By [36, Corollary 6], $\exists\, \kappa_0 > 0$ such that

$$\mathrm{dist}_{\mathcal{C} \cap T^{-1}(\mathcal{D})}(x) \le \kappa_0 \mathrm{dist}_{T^{-1}(\mathcal{D})}(x), \quad \forall x \in \mathcal{C}. \tag{6.1}$$

Since T is surjective, $TT^{\dagger} = \mathrm{Id}_{\mathbb{R}^m}$ with $T^{\dagger}$ the *Moore-Penrose right inverse* $T^{\dagger} = T^*(TT^*)^{-1}$. Fix $x \in \mathcal{C}$ and let $y = \mathrm{P}_{\mathcal{D}}(Tx)$ so $\|Tx - y\| = \mathrm{dist}_{\mathcal{D}}(Tx)$. Define $\hat{x} = x + T^{\dagger}(y - Tx)$; then

$$T\hat{x} = Tx + TT^{\dagger}(y - Tx) = Tx + (y - Tx) = y \in \mathcal{D}$$

and so $\hat{x} \in T^{-1}(\mathcal{D})$. Therefore,

$$\mathrm{dist}_{T^{-1}(\mathcal{D})}(x) \le \|x - \hat{x}\| = \|T^{\dagger}(y - Tx)\| \le \|T^{\dagger}\|_{\mathrm{op}} \|y - Tx\| = \|T^{\dagger}\|_{\mathrm{op}} \mathrm{dist}_{\mathcal{D}}(Tx).$$

Combining with the previous estimate gives

$$\mathrm{dist}_{\widetilde{\mathcal{C}}}(x) = \mathrm{dist}_{\mathcal{C} \cap T^{-1}(\mathcal{D})}(x) \le \kappa_0 \mathrm{dist}_{T^{-1}(\mathcal{D})}(x) \le \kappa_0 \|T^{\dagger}\|_{\mathrm{op}} \mathrm{dist}_{\mathcal{D}}(Tx).$$

□

Remark 6.3. Assumption 6.1 is a standard metric regularity condition for the system $x \in \mathcal{C}$, $Tx \in \mathcal{D}$. When T is surjective, a classical sufficient condition is the relative interior qualification already present in Assumption 2.4(I)(b), as demonstrated in Proposition 6.2. While surjectivity of T together with Assumption 2.4(I)(b) is sufficient to guarantee 6.1 holds, surjectivity is not necessary. Also, note that Assumption 6.1 is not required for the qualitative subsequential stationarity result in Theorem 5.7.

Lemma 6.4 (Indicator finite-time transfer). *Suppose* Assumption 2.4*(I)(a) and* Assumption 6.1 *hold and let* $x \in \mathcal{C}$ *and* $\beta > 0$*. Then*

$$\mathrm{dist}_{\mathcal{D}}(Tx) \le \frac{L_f \kappa_{\mathrm{ind}} \beta}{2} + \frac{1}{2}\sqrt{L_f^2 \kappa_{\mathrm{ind}}^2 \beta^2 + 4\beta \mathrm{gap}^{\beta}(x)} \tag{6.2}$$

and, for the signed gap given in Definition 4.1*,*

$$|\widetilde{\mathrm{gap}}(x)| \le \max\{\mathrm{gap}^{\beta}(x), L_f \kappa_{\mathrm{ind}} \mathrm{dist}_{\mathcal{D}}(Tx)\}. \tag{6.3}$$

Moreover,

$$\mathrm{dist}_{\mathcal{D}}(Tx) = \mathcal{O}\Big(\max\big\{\beta, \sqrt{\beta \mathrm{gap}^{\beta}(x)}\big\}\Big) \quad \textit{and} \quad |\widetilde{\mathrm{gap}}(x)| = \mathcal{O}\left(\max\left\{\beta, \mathrm{gap}^{\beta}(x)\right\}\right).$$

Proof For any $x \in \mathcal{C}$, the projection satisfies $\mathrm{P}_{\widetilde{\mathcal{C}}}(x) \in \widetilde{\mathcal{C}}$ and the definition of $\widetilde{\mathrm{gap}}$ gives

$$\begin{aligned}\widetilde{\mathrm{gap}}(x) = \max_{s \in \widetilde{\mathcal{C}}} \langle \nabla f(x), x - s \rangle &\geq \left\langle \nabla f(x), x - \mathrm{P}_{\widetilde{\mathcal{C}}}(x) \right\rangle \\ &\geq -\|\nabla f(x)\|_2 \left\| x - \mathrm{P}_{\widetilde{\mathcal{C}}}(x) \right\|_2 \\ &\geq -L_f \mathrm{dist}_{\widetilde{\mathcal{C}}}(x) \\ &\geq -L_f \kappa_{\mathrm{ind}} \mathrm{dist}_{\mathcal{D}}(Tx),\end{aligned} \tag{6.4}$$

where the last inequality uses Assumption 6.1. Combining this lower bound with Lemma 4.4 yields

$$\frac{1}{\beta}\mathrm{dist}_{\mathcal{D}}^2(Tx) - L_f \kappa_{\mathrm{ind}} \mathrm{dist}_{\mathcal{D}}(Tx) \leq \mathrm{gap}^{\beta}(x).$$

Letting $\delta := \mathrm{dist}_{\mathcal{D}}(Tx)$, this is equivalent to $\delta^2 - L_f \kappa_{\mathrm{ind}} \beta \delta - \beta \mathrm{gap}^{\beta}(x) \leq 0$. Since $\delta \geq 0$, it is bounded above by the larger root of the quadratic, which gives (6.2). Next, Lemma 4.4 gives $\widetilde{\mathrm{gap}}(x) \leq \mathrm{gap}^{\beta}(x)$, which with (6.4) implies (6.3). The asymptotic estimates follow from (6.2), (6.3), and the inequality $\sqrt{ab} \leq \max\{a, b\}$ for $a, b \geq 0$. □

Theorem 6.5 (Indicator nonsmooth certificate rate). *Suppose* Assumption 2.4*(I)(a)* *and* Assumption 6.1 *hold and let* $\{x_k\}_{k \in \mathbb{N}}$ *be generated by* ***FRAMES*** *with*

$$\gamma_k = (k+1)^{-1/2}, \qquad \beta_k = \beta_0 (k+1)^{-1/4}.$$

For $N \geq 2$, *let* $k_N^\star \in \underset{\lfloor N/2 \rfloor \leq k \leq N-1}{\mathrm{argmin}} \mathrm{gap}^{\beta_k}(x_k)$. *Then the last-half best iterate satisfies*

$$\mathrm{dist}_{\mathcal{D}}(Tx_{k_N^\star}) = \mathcal{O}(N^{-1/4})$$

and, for the signed gap given in Definition 4.1,

$$\left|\widetilde{\mathrm{gap}}(x_{k_N^\star})\right| = \mathcal{O}(N^{-1/4}).$$

Proof By Corollary 5.5, $\mathrm{gap}^{\beta_{k_N^\star}}(x_{k_N^\star}) = \mathcal{O}(N^{-1/4})$ and $\beta_{k_N^\star} = \mathcal{O}(N^{-1/4})$. Applying Lemma 6.4 with $x = x_{k_N^\star}$ and $\beta = \beta_{k_N^\star}$ gives

$$\mathrm{dist}_{\mathcal{D}}(Tx_{k_N^\star}) = \mathcal{O}\left(\max\left\{\beta_{k_N^\star}, \sqrt{\beta_{k_N^\star} \mathrm{gap}^{\beta_{k_N^\star}}(x_{k_N^\star})}\right\}\right) = \mathcal{O}(N^{-1/4}).$$

The estimate on $\left|\widetilde{\mathrm{gap}}(x_{k_N^\star})\right|$ follows from (6.3). □

Remark 6.6 (Logarithmic smoothing is slower). Just like how Lemma 6.4 yields the $\mathcal{O}(N^{-1/4})$ results in Theorem 6.5, a straightforward application of Lemma 6.4 with the schedules $\beta_k = \mathcal{O}(1/\log(k+2))$ and $\gamma_k = \mathcal{O}(1/\sqrt{k})$ suggested in [11, 13] provides a nonsmooth convergence rate of $\mathcal{O}(\log(N+2)^{-1})$ for the average absolute value of the signed gaps or the last-half best iterates $|\widetilde{\mathrm{gap}}(x_{k_N^\star})|$, which is much slower. This is because the analysis in [11, 13] only characterizes the effect of the smoothing schedule on the rate of convergence of the smoothed gaps to 0 and does not take into account the difference between the smooth gap (associated to the smoothed problem) and the signed gap (associated to (P)).

6.2 The Indicator Case with Inconsistent Sets

In the most general case of Assumption 2.4(I), $T^{-1}(\mathcal{D}) \cap \mathcal{C}$ may be empty. Then one cannot hope to show a stationarity result for (P), since the original problem has no feasible points. Instead, the cluster points of smoothed gap-vanishing subsequences solve a best-approximate feasibility problem in the image space, namely they minimize $x \mapsto \operatorname{dist}_{\mathcal{D}}(Tx)$ over $\mathcal{C}$.

Assumption 6.7 (Inconsistent system). *Assumption 2.4(I) holds and* $\mathcal{D} \cap T(\mathcal{C}) = \varnothing$.

Under Assumptions 2.3, 2.4(I), and 6.7, compactness of $T(\mathcal{C})$ and closedness of $\mathcal{D}$ give $\min_{x\in\mathcal{C}} \operatorname{dist}_{\mathcal{D}}(Tx) := \delta > 0$. Since $\mathcal{C}$ is bounded, [35, Prop. 11.15, Prop. 27.1, & Cor. 4.24] yield that the following *closest-point set* is nonempty, compact, and convex

$$\mathcal{C}^\dagger := \operatorname*{argmin}_{x\in\mathcal{C}} \tfrac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(Tx) = \operatorname*{argmin}_{x\in\mathcal{C}} \operatorname{dist}_{\mathcal{D}}(Tx). \tag{6.5}$$

Lemma 6.8 (Inconsistent gap-transfer). *Suppose* Assumption 2.1, 2.2, 2.3, 2.4(I) *and* 6.7 *hold. Let* $\delta := \min_{\bar{x}\in\mathcal{C}} \operatorname{dist}_{\mathcal{D}}(T\bar{x})$, $x \in \mathcal{C}$, $\beta > 0$, *and let* $C^\dagger$ *be given by* (6.5). *Then*

$$0 \le \delta(\operatorname{dist}_{\mathcal{D}}(Tx) - \delta) \le \operatorname{dist}^2_{\mathcal{D}}(Tx) - \delta^2 \le 2\beta\operatorname{gap}^\beta(x) + 2\beta L_f \operatorname{diam}_{\mathcal{C}}. \tag{6.6}$$

Moreover, for every $s \in \mathcal{C}^\dagger$,

$$\langle \nabla f(x), x - s\rangle \le \operatorname{gap}^\beta(x). \tag{6.7}$$

Proof For every $z \in \mathcal{C}$, using $\Phi_\beta = f + \beta^{-1}(\frac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(T\cdot))$ gives

$$\begin{aligned}\left\langle T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - z\right\rangle &= \beta\left\langle \nabla f(x) + \beta^{-1}T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - z\right\rangle - \beta\left\langle \nabla f(x), x - z\right\rangle \\ &\le \beta\operatorname{gap}^\beta(x) + \beta L_f \operatorname{diam}_{\mathcal{C}}.\end{aligned}$$

Let $x^\dagger \in \mathcal{C}^\dagger$. By convexity of $x \mapsto \frac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(Tx)$,

$$\operatorname{dist}^2_{\mathcal{D}}(Tx) - \delta^2 \le 2\left\langle T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - x^\dagger\right\rangle \le 2\beta\operatorname{gap}^\beta(x) + 2\beta L_f \operatorname{diam}_{\mathcal{C}}.$$

Since $\operatorname{dist}_{\mathcal{D}}(Tx) \ge \delta$, we also have

$$\delta(\operatorname{dist}_{\mathcal{D}}(Tx) - \delta) \le (\operatorname{dist}_{\mathcal{D}}(Tx) - \delta)(\operatorname{dist}_{\mathcal{D}}(Tx) + \delta) = \operatorname{dist}^2_{\mathcal{D}}(Tx) - \delta^2,$$

which proves (6.6). Now fix $s \in \mathcal{C}^\dagger$. By convexity of $x \mapsto \frac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(Tx)$,

$$\tfrac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(Ts) \ge \tfrac{1}{2}\operatorname{dist}^2_{\mathcal{D}}(Tx) + \left\langle T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), s - x\right\rangle.$$

Since $\operatorname{dist}_{\mathcal{D}}(Ts) = \delta \le \operatorname{dist}_{\mathcal{D}}(Tx)$, it follows that $\langle T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - s\rangle \ge 0$. Therefore,

$$\begin{aligned}\langle \nabla f(x), x - s\rangle &= \left\langle \nabla f(x) + \beta^{-1}T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - s\right\rangle - \beta^{-1}\left\langle T^*(Tx - \mathrm{P}_{\mathcal{D}}(Tx)), x - s\right\rangle \\ &\le \operatorname{gap}^\beta(x),\end{aligned}$$

which proves (6.7). □

Theorem 6.9 (Inconsistent gap-vanishing sequences). *Suppose* Assumption 2.1, 2.2, 2.3, *and* 6.7 *hold, let* $\delta := \min_{\bar{x}\in\mathcal{C}} \mathrm{dist}_{\mathcal{D}}(T\bar{x})$, $C^\dagger$ *be given by* (6.5), *and suppose* $\{u_j\}_{j\in\mathbb{N}} \subset \mathcal{C}$ *with* $\{\alpha_j\}_{j\in\mathbb{N}} \subset]0, +\infty[$ *define a convergent smoothed gap-vanishing sequence, i.e.,*

$$\alpha_j \to 0, \qquad u_j \to \bar{u}, \qquad \textit{and} \qquad \mathrm{gap}^{\alpha_j}(u_j) \to 0.$$

Then $\mathrm{dist}_{\mathcal{D}}(Tu_j) \to \delta$ *and* $\mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \to 0$. *Moreover,*

$$\max_{s\in\mathcal{C}^\dagger} \langle \nabla f(u_j), u_j - s\rangle \to 0. \tag{6.8}$$

Consequently, every cluster point $\bar{u}$ *of* $\{u_j\}_{j\in\mathbb{N}}$ *belongs to* $\mathcal{C}^\dagger$ *and satisfies*

$$0 \in T^*(T\bar{u} - \mathrm{P}_{\mathcal{D}}(T\bar{u})) + N_{\mathcal{C}}(\bar{u}).$$

Furthermore, $\bar{u}$ *is stationary for the secondary problem*

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

i.e., $0 \in \nabla f(\bar{u}) + N_{\mathcal{C}^\dagger}(\bar{u})$ *and, equivalently in terms of the Frank-Wolfe gap,*

$$\max_{s\in\mathcal{C}^\dagger} \langle \nabla f(\bar{u}), \bar{u} - s\rangle = 0. \tag{6.9}$$

In particular, if f *is convex on* $\mathcal{C}^\dagger$ *then* $\bar{u} \in \mathrm{argmin}_{x\in\mathcal{C}^\dagger} f(x)$.

Proof By Lemma 6.8,

$$0 \le \mathrm{dist}^2_{\mathcal{D}}(Tu_j) - \delta^2 \le 2\alpha_j \mathrm{gap}^{\alpha_j}(u_j) + 2\alpha_j L_f \mathrm{diam}_{\mathcal{C}} \to 0.$$

Thus $\mathrm{dist}_{\mathcal{D}}(Tu_j) \to \delta$.

We next show that $\mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \to 0$. Assume for the sake of contradiction that there exist $\varepsilon > 0$ and a subsequence such that $\mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \ge \varepsilon$. Since $u_j \to \bar{u} \in \mathcal{C}$, the subsequence also converges to $\bar{u}$. Continuity of $\mathrm{dist}_{\mathcal{D}} \circ T$ gives $\mathrm{dist}_{\mathcal{D}}(T\bar{u}) = \delta$, hence $\bar{u} \in \mathcal{C}^\dagger$, contradicting $\mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \ge \varepsilon$.

For all $j \in \mathbb{N}$, let

$$G_j := \max_{s\in\mathcal{C}^\dagger} \left\langle \nabla f(u_j), u_j - s\right\rangle.$$

Taking the maximum over $s \in C^\dagger$ in (6.7) gives $G_j \le \mathrm{gap}^{\alpha_j}(u_j) \to 0$. For the lower bound, let $p_j := \mathrm{P}_{\mathcal{C}^\dagger}(u_j)$. Since $p_j \in \mathcal{C}^\dagger$,

$$G_j \ge \left\langle \nabla f(u_j), u_j - p_j\right\rangle \ge -L_f \left\|u_j - p_j\right\|_2 = -L_f \mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \to 0.$$

Therefore $G_j \to 0$, proving (6.8).

Let $\bar{u}$ be a cluster point of $\{u_j\}_{j\in\mathbb{N}}$. Since $\mathrm{dist}_{\mathcal{C}^\dagger}(u_j) \to 0$ and $\mathcal{C}^\dagger$ is closed, $\bar{u} \in \mathcal{C}^\dagger$. Since $\bar{u}$ minimizes $x \mapsto \frac{1}{2}\mathrm{dist}^2_{\mathcal{D}}(Tx)$ over $\mathcal{C}$, the first-order optimality condition gives

$$0 \in T^*(T\bar{u} - \mathrm{P}_{\mathcal{D}}(T\bar{u})) + N_{\mathcal{C}}(\bar{u}).$$

Passing to the limit in (6.8), using continuity of ∇f and compactness of $\mathcal{C}^\dagger$, gives (6.9). The normal cone inclusion $0 \in \nabla f(\bar{u}) + N_{\mathcal{C}^\dagger}(\bar{u})$ is the standard Frank-Wolfe gap characterization on the nonempty compact convex set $\mathcal{C}^\dagger$ [34].

Finally, if f is convex on $\mathcal{C}^\dagger$, then for every $s \in \mathcal{C}^\dagger$,

$$f(s) \geq f(\bar{u}) + \langle \nabla f(\bar{u}), s - \bar{u} \rangle \geq f(\bar{u}),$$

so $\bar{u}$ minimizes f over $\mathcal{C}^\dagger$. □

Corollary 6.10 (Last-half best iterate in the inconsistent case). *Fix $p, q \in]0,1[$ satisfying $q < p$ and $p + q < 1$ and set $\gamma_k = (k+1)^{-p}$ and $\beta_k = \beta_0(k+1)^{-q}$ with $\beta_0 \in \left]0, \rho^{-1}\right[$. Suppose* Assumption 2.1, 2.2, 2.3, *and* 6.7 *hold and let $\{x_k\}_{k\in\mathbb{N}}$ be generated by* **FRAMES**. *For $N \geq 2$, let*

$$k_N^\star \in \operatorname*{argmin}_{\lfloor N/2 \rfloor \leq k \leq N-1} \operatorname{gap}^{\beta_k}(x_k).$$

Then, as $N \to \infty$,

$$\operatorname{dist}_{\mathcal{D}}(Tx_{k_N^\star}) \to \delta \qquad \text{and} \qquad \operatorname{dist}_{\mathcal{C}^\dagger}(x_{k_N^\star}) \to 0.$$

Moreover,

$$\max_{s \in \mathcal{C}^\dagger} \left\langle \nabla f(x_{k_N^\star}), x_{k_N^\star} - s \right\rangle \to 0. \tag{6.10}$$

Every cluster point of $\{x_{k_N^\star}\}_{k\in\mathbb{N}}$ is stationary for $\min_{x\in\mathcal{C}^\dagger} f(x)$. If f is convex on $\mathcal{C}^\dagger$, then every such cluster point minimizes f over $\mathcal{C}^\dagger$.

Proof By Corollary 5.5, $\operatorname{gap}^{\beta_{k_N^\star}}(x_{k_N^\star}) \to 0$. Since $k_N^\star \geq \lfloor N/2 \rfloor$ and $\beta_k \to 0$, we also have $\beta_{k_N^\star} \to 0$. The result follows from Theorem 6.9 applied with $u_N = x_{k_N^\star}$ and $\alpha_N = \beta_{k_N^\star}$. □

Remark 6.11. Under Assumption 6.7, **FRAMES** cannot generate a feasible limit point for the original problem (P), because no such point exists. The conclusion above is therefore a best-approximation statement: smoothed gap-vanishing sequences approach the closest-point set $\mathcal{C}^\dagger$, and their cluster points are stationary for f restricted to that set.

6.3 The Lipschitz Weakly Convex Case

We now suppose that Assumption 2.4(II) holds, making (P) the constrained analog of the setting considered in [19] (because of $\mathcal{C}$). In this case, the smoothed gradient

$$\xi_\beta(x) := \nabla g^\beta(Tx)$$

belongs to $\partial g(\operatorname{prox}_{\beta g}(Tx))$, not necessarily to $\partial g(Tx)$. This prevents the smoothed gap from being directly identified with the nonsmooth gap, $\operatorname{gap}(x;\xi)$. The following assumption gives a finite-time transfer under an additional lifting assumption, which is similar to the surjectivity assumption made in [19].

Assumption 6.12 (Proximal lift). *There exists a constant $M_{\mathrm{lift}} > 0$ such that, for every $x \in \mathcal{C}$ and every $\beta \in \left]0, \rho^{-1}\right[$, there exists $z = z(x,\beta) \in \mathcal{C}$ satisfying*

$$Tz = \operatorname{prox}_{\beta g}(Tx)$$

and

$$\|z - x\|_2 \le M_{\text{lift}} \|Tz - Tx\|_2 .$$

Before continuing, we emphasize that Assumption 6.12 is used only for the finite-time gap-transfer bounds in Lemma 6.14 and Theorem 6.15. The qualitative subsequential stationarity results in Theorem 5.8 and Corollary 5.9 do not require it.

Remark 6.13. Assumption 6.12 asks that the proximal point $\text{prox}_{\beta g}(Tx)$ can be lifted back into $\mathcal{C}$ through some $z \in \mathcal{C}$ satisfying

$$Tz = \text{prox}_{\beta g}(Tx) \qquad \text{and} \qquad \|z - x\|_2 \le M_{\text{lift}} \|Tz - Tx\|_2 .$$

This is an additional restriction on the triple $(g, T, \mathcal{C})$ and does not hold in general. We mention two simple settings where it is satisfied.

1. *Identity operator with componentwise-shrinking prox.* Let $T = \text{Id}$ and let $\mathcal{C} = \{x : \|x\|_p \le \tau\}$ be an ℓ^p-norm ball for some $p \in [1, \infty]$. If $\text{prox}_{\beta g}$ acts componentwise and satisfies
$$|\text{prox}_{\beta g}(y)_i| \le |y_i| \qquad \forall i,$$
then $z := \text{prox}_{\beta g}(x)$ satisfies $\|z\|_p \le \|x\|_p \le \tau$ so $z \in \mathcal{C}$. This holds for $g = \lambda \|\cdot\|_1$, whose proximal operator is soft-thresholding, and for the usual separable SCAD and MCP penalties, whose proximal maps shrink each coordinate toward zero. In this case one can take $M_{\text{lift}} = 1$.
2. *Invertible T with prox-invariant image constraint.* Suppose that T is invertible and that $\mathcal{C} = T^{-1}(\mathcal{K})$ for some set $\mathcal{K} \subset \mathbb{R}^m$ satisfying
$$\text{prox}_{\beta g}(\mathcal{K}) \subset \mathcal{K} \qquad \forall \beta \in \left]0, \rho^{-1}\right[.$$
Then, for $x \in \mathcal{C}$, we have $Tx \in \mathcal{K}$ and hence $\text{prox}_{\beta g}(Tx) \in \mathcal{K}$. Defining $z := T^{-1}\text{prox}_{\beta g}(Tx)$ gives $z \in \mathcal{C}$ and $Tz = \text{prox}_{\beta g}(Tx)$. Moreover,
$$\|z - x\|_2 = \left\|T^{-1}\left(\text{prox}_{\beta g}(Tx) - Tx\right)\right\|_2 \le \left\|T^{-1}\right\|_{\text{op}} \left\|\text{prox}_{\beta g}(Tx) - Tx\right\|_2 .$$
Thus Assumption 6.12 holds with $M_{\text{lift}} = \left\|T^{-1}\right\|_{\text{op}}$. A simple way to guarantee prox-invariance is to take $\mathcal{K}$ to be a sublevel set of g, i.e., $\mathcal{K} = \{y : g(y) \le \alpha\}$, since the definition of the proximal point gives
$$g(\text{prox}_{\beta g}(y)) + \frac{1}{2\beta}\left\|\text{prox}_{\beta g}(y) - y\right\|_2^2 \le g(y)$$
and therefore $g(\text{prox}_{\beta g}(y)) \le g(y)$.

Lemma 6.14 (Lipschitz gap-transfer). *Suppose* Assumption 2.1, 2.2, 2.3, 2.4(II) *and* Assumption 6.12 *all hold. Let $x \in \mathcal{C}$, $\beta \in \left]0, \rho^{-1}\right[$, and choose $z = z(x, \beta) \in \mathcal{C}$ as in* Assumption 6.12. *Define $\xi_\beta(x) := \nabla g^\beta(Tx)$ and recall the subgradient gap given in*

Definition 4.6. *Then* $\xi_\beta(x) \in \partial g(Tz)$ *and, denoting* $C_{\text{lift}} := M_{\text{lift}} L_g (L_f + L_g \|T^*\|_{\text{op}} + L_{\nabla f} \text{diam}_{\mathcal{C}})$,

$$\text{gap}(z; \xi_\beta(x)) \leq \text{gap}^\beta(x) + C_{\text{lift}} \beta. \tag{6.11}$$

Proof Since $Tz = \text{prox}_{\beta g}(Tx)$, the proximal optimality condition gives

$$\xi_\beta(x) = \frac{Tx - \text{prox}_{\beta g}(Tx)}{\beta} \in \partial g(\text{prox}_{\beta g}(Tx)) = \partial g(Tz).$$

Therefore $\text{gap}(z; \xi_\beta(x))$ is well-defined.

For any $s \in \mathcal{C}$, add and subtract x and $\nabla f(x)$ to obtain

$$\begin{aligned} \left\langle \nabla f(z) + T^* \xi_\beta(x), z - s \right\rangle = &\left\langle \nabla f(x) + T^* \xi_\beta(x), x - s \right\rangle \\ &+ \left\langle \nabla f(x) + T^* \xi_\beta(x), z - x \right\rangle \\ &+ \left\langle \nabla f(z) - \nabla f(x), z - s \right\rangle . \end{aligned}$$

Taking the maximum over $s \in \mathcal{C}$ gives

$$\text{gap}(z; \xi_\beta(x)) \leq \text{gap}^\beta(x) + \left\| \nabla f(x) + T^* \xi_\beta(x) \right\|_2 \|z - x\|_2 + L_{\nabla f} \|z - x\|_2 \, \text{diam}_{\mathcal{C}}. \tag{6.12}$$

By the boundedness of ∇f on $\mathcal{C}$ and Proposition 3.1,

$$\left\| \nabla f(x) + T^* \xi_\beta(x) \right\|_2 \leq L_f + L_g \left\| T^* \right\|_{\text{op}}. \tag{6.13}$$

Moreover, by Assumption 6.12 and Proposition 3.2,

$$\|z - x\|_2 \leq M_{\text{lift}} \left\| \text{prox}_{\beta g}(Tx) - Tx \right\|_2 \leq M_{\text{lift}} \beta L_g. \tag{6.14}$$

Substituting (6.13) and (6.14) into (6.12) yields (6.11). □

Theorem 6.15 (Lipschitz nonsmooth gap rate). *Suppose* Assumption 2.1, 2.2, 2.3, 2.4(II), 2.6, 2.7, *and* Assumption 6.12 *hold and let* $\{x_k\}_{k \in \mathbb{N}}$ *be generated by* ***FRAMES*** *with*

$$\gamma_k = (k+1)^{-1/2}, \qquad \beta_k = \beta_0 (k+1)^{-1/4},$$

where $\beta_0 \in \left]0, \rho^{-1}\right[$. *For each* $k \in \mathbb{N}$, *let* $z_k := z(x_k, \beta_k)$ *be chosen as in* Assumption 6.12, *set* $\xi_k := \nabla g^{\beta_k}(Tx_k)$, *and recall the subgradient gap given in* Definition 4.6. *Then there exists a constant* $C_{\text{ns}} > 0$ *such that, for every* $N \in \mathbb{N}^*$,

$$0 \leq \frac{1}{N} \sum_{k=0}^{N-1} \text{gap}(z_k; \xi_k) \leq C_{\text{ns}} N^{-1/4}.$$

Consequently,

$$0 \leq \min_{0 \leq k \leq N-1} \text{gap}(z_k; \xi_k) \leq C_{\text{ns}} N^{-1/4}.$$

Moreover, for last-half best iterate with index $k_N^\star \in \underset{\lfloor N/2 \rfloor \leq k \leq N-1}{\text{argmin}} \text{gap}^{\beta_k}(x_k)$ *as in* Corollary 5.5, *it holds*

$$\text{gap}(z_{k_N^\star}; \xi_{k_N^\star}) = \mathcal{O}(N^{-1/4}).$$

Proof By Lemma 6.14, for every $k \in \mathbb{N}$,

$$\operatorname{gap}(z_k; \xi_k) \le \operatorname{gap}^{\beta_k}(x_k) + C_{\mathrm{lift}}\beta_k.$$

Averaging this inequality and using Theorem 5.3 gives

$$\frac{1}{N}\sum_{k=0}^{N-1} \operatorname{gap}(z_k; \xi_k) \le CN^{-1/4} + \frac{C_{\mathrm{lift}}}{N}\sum_{k=0}^{N-1} \beta_k.$$

Since $\beta_k = \beta_0(k+1)^{-1/4}$,

$$\sum_{k=0}^{N-1} \beta_k \le \beta_0\left(1 + \frac{4}{3}N^{3/4}\right).$$

Thus the average nonsmooth gap is $\mathcal{O}(N^{-1/4})$. The minimum bound follows because the gaps are nonnegative. The final statement for the last-half best iterate corresponding to $k_N^\star$ follows by combining Lemma 6.14 with Corollary 5.5. □

Remark 6.16 (Optimal balance among power schedules). We now return to the power schedules from Theorem 5.3. For $\gamma_k = (k+1)^{-p}$ and $\beta_k = \beta_0(k+1)^{-q}$ with $0 < q < p$ and $p + q < 1$, the smoothed gap exponent is $\min\{p - q,\ 1 - p - q\}$. The finite-time transfer estimates above show there is an additional smoothing error $\beta_k = \mathcal{O}(k^{-q})$ to account for. Thus the exponent governing the final nonsmooth stationarity certificates is

$$\min\{q, \min\{p - q,\ 1 - p - q\}\} = \min\{q,\ p - q,\ 1 - p - q\}. \tag{6.15}$$

In the Lipschitz weakly convex case, this follows directly from Lemma 6.14; in the indicator case, the signed gap is controlled by the same bottleneck through Lemma 6.4. Maximizing (6.15) over $0 < q < p < 1$ leads directly to the choices $p = 1/2$ and $q = 1/4$.

7 Applications and Numerical Experiments

In this section, we apply **FRAMES** to several problems that illustrate its flexibility across assumptions on g. Section 7.1 considers a nonconvex splitting problem where the nonsmooth gap is computable in closed form, providing a direct verification of the convergence theory. Section 7.2.1 and Section 7.2.2 share the same smooth objective and constraint set but differ in the nonsmooth term: the first uses a nonnegativity indicator (Assumption 2.4(I)), the second uses trend filtering with weakly convex penalties (Assumption 2.4(II)). Finally, in Section 7.3 we explore what happens for inconsistent problems where $\mathcal{D} \cap T(\mathcal{C}) = \varnothing$ as in Assumption 6.7.

7.1 Nonconvex Splitting over ℓ_1 Balls

We construct a problem fitting into the splitting framework of [8, 10, 11] and for which we can compute the signed gap in closed form. We first consider

$$\min_{x \in \mathcal{C}_1 \cap \mathcal{C}_2} f(x) = \tfrac{1}{2}x^\top Q x - b^\top x, \tag{7.1}$$

where $Q \in \mathbb{R}^{n\times n}$ is symmetric and indefinite (so that f is nonconvex but satisfies Assumption 2.1), $b \in \mathbb{R}^n$, $\mathcal{C}_1 = \{x : \|x - e_1\|_1 \le 2\}$, and $\mathcal{C}_2 = \{x : \|x + e_1\|_1 \le 2\}$ where

e_1 denotes the vector with 1 in the first entry and 0 everywhere else. One can verify that $\mathcal{C}_1 \cap \mathcal{C}_2 = B_1 := \{x : \|x\|_1 \leq 1\}$.

To obtain an instance of (P), we lift to the product space $\mathbf{x} = (x_1, x_2) \in \mathcal{C}_1 \times \mathcal{C}_2$ and enforce consensus among components

$$\min_{(x_1,x_2)\in\mathcal{C}_1\times\mathcal{C}_2} \underbrace{f(\tfrac{x_1+x_2}{2})}_{f(\mathbf{x})} + \underbrace{\iota_{\{0\}}(x_1 - x_2)}_{g(T\mathbf{x})},$$

with $T\mathbf{x} = x_1 - x_2$, $\mathcal{D} = \{0\}$, $\bar{x} = \frac{x_1+x_2}{2}$, and $\operatorname{prox}_{\beta g}(z) = P_{\{0\}}(z) = 0$. This satisfies Assumption 2.2, 2.3, and 2.4(I)(b). Moreover, the LMO is separable over the blocks,

$$\operatorname*{lmo}_{\mathcal{C}_1\times\mathcal{C}_2}((G_1, G_2)) = (\operatorname*{lmo}_{\mathcal{C}_1}(G_1),\ \operatorname*{lmo}_{\mathcal{C}_2}(G_2)),$$

and easily computed using the vector sign.

Computable Nonsmooth Frank-Wolfe Gap

Since $\mathcal{C}_1 \cap \mathcal{C}_2 = B_1$ is known, the feasible set of the original problem is $\tilde{\mathcal{C}} = \{(s, s) : s \in B_1\}$. The nonsmooth gap in this setting corresponds to a signed gap as given in Definition 4.1; at $\mathbf{x} = (x_1, x_2)$ is $\widetilde{\operatorname{gap}}(\mathbf{x}) = \max_{s\in\tilde{\mathcal{C}}} \langle \nabla f(\mathbf{x}), \mathbf{x} - s\rangle$. Since $\nabla f(\mathbf{x}) = \frac{1}{2}(Q\bar{x} - b,\ Q\bar{x} - b)$, substituting $s = (s', s')$ with $s' \in B_1$ yields

$$\widetilde{\operatorname{gap}}(\mathbf{x}) = \max_{s'\in B_1} \langle Q\bar{x} - b, \bar{x} - s'\rangle\,.$$

Denoting $q := Q\bar{x} - b$, this is the Frank-Wolfe gap of the original problem (7.1) evaluated at $\bar{x}$:

$$\widetilde{\operatorname{gap}}(\mathbf{x}) = \langle q,\ \bar{x} - \operatorname*{lmo}_{B_1}(q)\rangle.$$

Since $\operatorname{lmo}_{B_1}(q)$ returns $-\operatorname{sign}(q_{j^\star})e_{j^\star}$ where $j^\star \in \operatorname{argmax}_j |q_j|$, we have $\langle q, -\operatorname{lmo}_{B_1}(q)\rangle = \|q\|_\infty$, giving

$$\widetilde{\operatorname{gap}}(\mathbf{x}) = \langle Q\bar{x} - b,\ \bar{x}\rangle + \|Q\bar{x} - b\|_\infty, \tag{7.2}$$

which is easily computed in closed-form using the available variables while **FRAMES** runs.

Data Generation

We generate $Q = V\Lambda V^\top$ with V a random orthogonal matrix and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with eigenvalues drawn uniformly from $[0.5, 5]$ in magnitude, roughly 30% of them negated. The vector b is sampled randomly with components following the standard normal distribution. We use $n = 50$, and $N = 50,000$. The number of iterations N is intentionally large to demonstrate the difference in the schedules.

Results

Figure 2 displays the convergence using **FRAMES** with $\gamma_k = 1/(k+1)^{1/2}$ and various $\beta_0 \in \{0.25, 1/L_{\nabla f}, 0.5, 1, 2, 4\}$, where $L_{\nabla f} = \|Q\|_{\text{op}}/2 = 2.49$. We also compare the power schedule $\beta_k = \beta_0/(k+1)^{1/4}$ against the log schedule $\beta_k = \beta_0/\log(k+2)$ for each value

of β_0. This is shown in Figure 2, where dashed lines represent the log schedule and solid lines represent the power schedule.

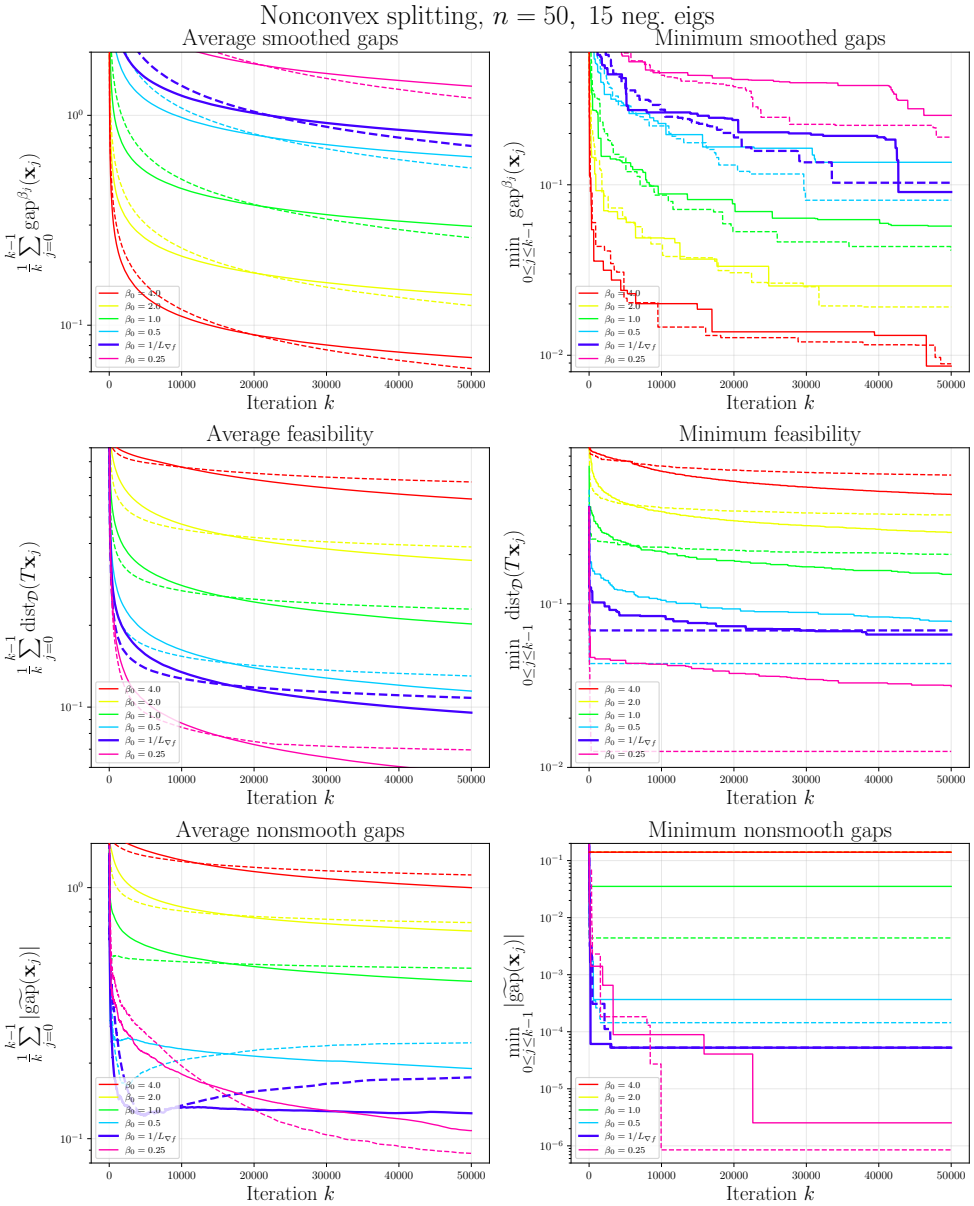

Fig. 2: Solid curves use the power schedule $\beta_k = \beta_0/(k+1)^{1/4}$ and dashed curves use the (natural) log schedule $\beta_k = \beta_0/\log(k+2)$, as suggested by prior work [11, 13]; both use the step size $\gamma_k = (k+1)^{-1/2}$. In the bottom right plot, the curves for $\beta_0 = 4$ and $\beta_0 = 2$ overlap.

- *Smoothed gaps.* The running minimum and average of the smoothed gaps decrease at the predicted $\mathcal{O}(k^{-1/4})$ rate for all values of β_0, confirming Theorem 5.3 in the nonconvex setting.
- *Feasibility.* The consensus violation $\|x_{1,k} - x_{2,k}\|^2$ decreases for all β_0, consistent with Lemma 6.4. Setting β_0 very small produces a trajectory which reaches the lowest infeasibility. On the other hand, the smaller β_0 is, the slower the smoothed gaps converge for both schedules.
- *Nonsmooth gap.* We plot $|\widetilde{\mathrm{gap}}(\mathbf{x}_k)|$ in the right panel, testing the predictions of Theorem 6.5. When $\bar{x}_k \notin B_1$, the nonsmooth gap (7.2) can be negative. Smaller β_0 seems to yield a smaller $|\widetilde{\mathrm{gap}}|$, with $\beta_0 = 0.25$ achieving the best values for both schedules. The log schedule produces a larger average and minimal $|\widetilde{\mathrm{gap}}|$ than the power schedule for most β_0 values, confirming that the slower smoothing decay is detrimental as predicted by Theorem 6.5 and Remark 6.6.

7.2 Low-rank Matrix Factorization

In this section, we consider two matrix factorization problems, focusing on nonnegative matrix factorization [37] in Section 7.2.1 and matrix factorization with trend filtering [38] in Section 7.2.2. Both experiments consider the smooth nonconvex objective

$$f(U, V) = \tfrac{1}{2}\|UV^\top - X^\star\|_F^2,$$

where $X^\star \in \mathbb{R}^{m\times n}$ is a given matrix and $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$ are matrix variables. The constraint set is the product of two spectral-norm balls,

$$\mathcal{C} = \left\{U \in \mathbb{R}^{m\times r} : \|U\|_{\mathrm{op}} \le \tau_U\right\} \times \left\{V \in \mathbb{R}^{n\times r} : \|V\|_{\mathrm{op}} \le \tau_V\right\},$$

so, given a gradient (G_U, G_V), the LMO decomposes across the two factor blocks, i.e.,

$$\operatorname*{lmo}_{\mathcal{C}}((G_U, G_V)) = \left(\operatorname*{lmo}_{\{W:\, \|W\|_{\mathrm{op}}\}\le\tau_U}(G_U), \quad \operatorname*{lmo}_{\{W:\, \|W\|_{\mathrm{op}}\}\le\tau_V}(G_V)\right). \tag{7.3}$$

Since f is a quartic polynomial in the entries of (U, V), it is smooth but nonconvex; restricted to the convex compact constraint set $\mathcal{C}$, Assumptions 2.1, 2.2, and 2.3 are satisfied.

For a spectral-norm ball $\{M : \|M\|_{\mathrm{op}} \le \tau\}$, the LMO applied to a gradient G returns $-\tau\cdot\mathrm{msign}(G)$, where $\mathrm{msign}(G) = L_G R_G^\top$ is the matrix sign obtained from the reduced SVD, $G = L_G \Sigma_G R_G^\top$, by replacing all nonzero singular values with ones. The optimization variable is represented as a single vector $x = [\mathrm{vec}(U),\ \mathrm{vec}(V)] \in \mathbb{R}^{(m+n)r}$. The cost of computing (7.3) using Newton-Schulz [39] is cheap compared to the projection operator of $\mathcal{C}$.

7.2.1 Nonnegativity Constraints

We consider the problem of factoring the matrix $X^\star$ into two rank $r = 20$ factors U and V with nonnegative entries,

$$\min_{\substack{\|U\|_{\mathrm{op}} \le \tau_U \\ \|V\|_{\mathrm{op}} \le \tau_V}} \frac{1}{2}\|UV^\top - X^\star\|_F^2 \; + \; \iota_{\mathbb{R}_+^{m\times r}\times\mathbb{R}_+^{n\times r}}(U, V) \tag{7.4}$$

which has many applications [37]. This is an instance of (P) with $T = \mathrm{Id}$ and $g = \iota_{\mathcal{D}}$, where

$$\mathcal{D} = \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{n\times r}$$

satisfies Assumption 2.4(I)(b). Thus, for all $\beta > 0$, $\mathrm{prox}_{\beta g}(x) = \mathrm{P}_{\mathcal{D}}(x) = \max(x, 0)$ where the maximum is applied componentwise.

Data Generation

We generate nonnegative ground-truth factors $U^\star, V^\star$ with i.i.d. entries drawn as $|\mathcal{N}(0, 1)|$, set $X^\star = U^\star(V^\star)^\top$, and fix the constraint radii as $\tau_U = 1.05\,\|U^\star\|_{\mathrm{op}}$, $\tau_V = 1.05\,\|V^\star\|_{\mathrm{op}}$, ensuring that the ground truth is strictly feasible in $\mathcal{C}$. Throughout, $m = n = 100$ and $N = 50,000$ iterations.

Sensitivity to β_0

Figure 3 displays the behavior of **FRAMES** across $\beta_0 \in \{1/L_{\nabla f}, 0.2, 0.5, 1, 2, 5\}$, where $L_{\nabla f}$ is estimated numerically to be 373.07. We use $\gamma_k = (k+1)^{-1/2}$ and $\beta_k = \beta_0(k+1)^{-1/4}$. The value $\beta_0 = 1/L_{\nabla f}$ makes the initial Lipschitz constant of the Moreau-envelope term comparable to that of the smooth matrix-factorization loss.

Results

The four panels in Figure 3 show average smoothed gaps, minimum smoothed gaps, feasibility, and relative reconstruction error.

- *Smoothed gaps.* The average and running minimum of the smoothed gaps decrease as predicted in Theorem 5.3 for all plotted values of β_0. Smaller values of β_0 place more weight on the Moreau envelope term and lead to larger smoothed gaps but smaller feasibility gaps, consistent with the dependence on β_k^{-1} in Theorem 5.3 and Theorem 6.5.
- *Feasibility.* The feasibility panel plots $\mathrm{dist}_{\mathcal{D}}(x_k)$ for the nonnegative orthant. The smallest value, $\beta_0 = 1/L_{\nabla f}$, enforces nonnegativity most aggressively and gives the lowest infeasibility among the runs, while slowing down convergence of the smoothed gaps.
- *Reconstruction error.* The relative error $\|U_k V_k^\top - X^\star\|_F / \|X^\star\|_F$ decreases for all values of β_0. We remark that smaller β_0 values give better feasibility and $\beta_0 = 0.2$ (one of the smallest β_0 values we tried) seems to perform the best. This illustrates the practical tradeoff between optimizing the smoothed objective and enforcing the nonsmooth constraint.

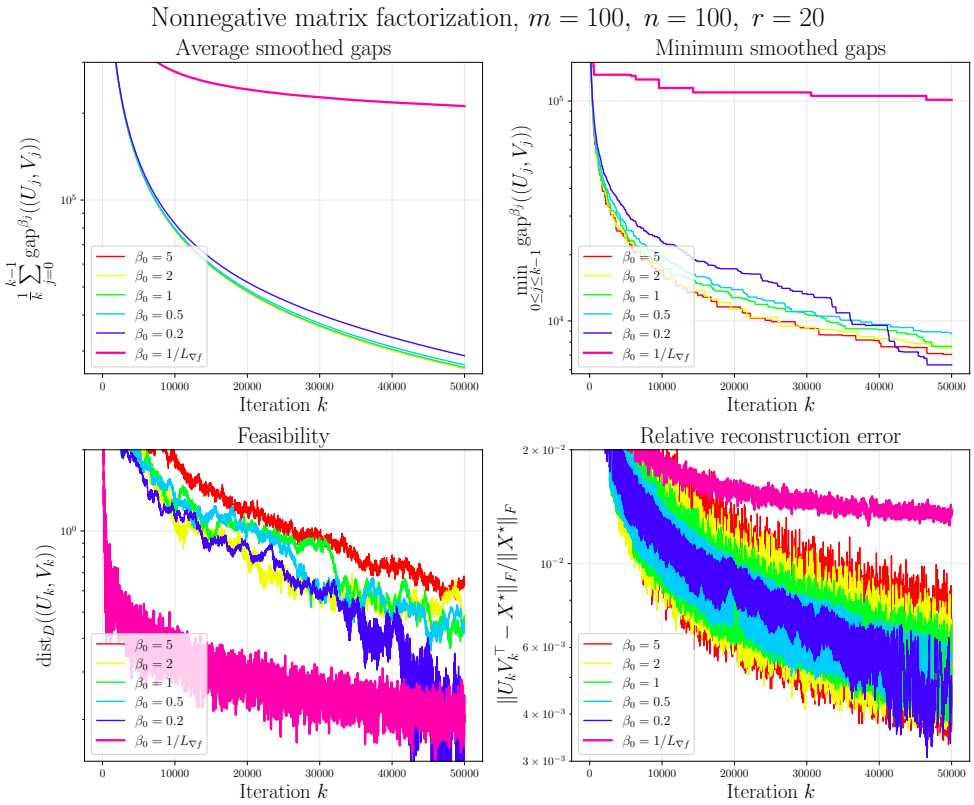

Fig. 3: Convergence profiles for **FRAMES** applied to nonnegative matrix factorization problem of Section 7.2.1 with $T = \mathrm{Id}$ and $g = \iota_{\mathcal{D}}$ for $\mathcal{D} = \mathbb{R}_+^{m\times r} \times \mathbb{R}_+^{n\times r}$.

7.2.2 Trend Filtering Regularization

We now replace the nonnegativity constraint by a nonsmooth weakly convex regularizer acting on finite differences of the U factor, a variant of what is commonly used in *trend filtering* problems [38, 40, 41]. Let $D_{\mathrm{row}} \in \mathbb{R}^{(m-1)\times m}$ be the first-order difference matrix, i.e.,

$$(D_{\mathrm{row}}U)_{i,j} = U_{i+1,j} - U_{i,j}.$$

We consider the problem

$$\min_{\substack{\|U\|_{\mathrm{op}}\le\tau_U \\ \|V\|_{\mathrm{op}}\le\tau_V}} \frac{1}{2}\|UV^\top - X^\star\|_F^2 \ + \ \sum_{i,j} g_0\big((D_{\mathrm{row}}U)_{i,j}\big), \tag{7.5}$$

which factors an observed matrix $X^\star$. This fits (P) with the linear map

$$T(U,V) = D_{\mathrm{row}}U,$$

and with g being the separable sum of g_0 across the entries of $D_{\text{row}}U$. The effect of the regularizer is to promote piecewise-constant structure along the rows of the columns of U.

We compare two nonconvex penalties for g_0, each relying on two hyperparameters (λ, a):

- $g_0(t) = \text{SCAD}_{\lambda,a}(t) = \begin{cases} \lambda|t| & \text{if } |t| \le \lambda \\ \frac{-2a\lambda|t|+t^2+\lambda^2}{2(1-a)} & \text{if } \lambda < |t| \le a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{if } |t| > a\lambda \end{cases}$ [42];
- $g_0(t) = \text{MCP}_{\lambda,\gamma}(t) = \begin{cases} \lambda|t| - \frac{t^2}{2\gamma} & \text{if } |t| \le \gamma\lambda \\ \frac{\gamma\lambda^2}{2} & \text{if } |t| > \gamma\lambda \end{cases}$ [43].

Both penalties are Lipschitz and weakly convex, and both have closed-form proximity operators and formulas for ρ [42, 43]. Hence this experiment falls under Assumption 2.4(II).

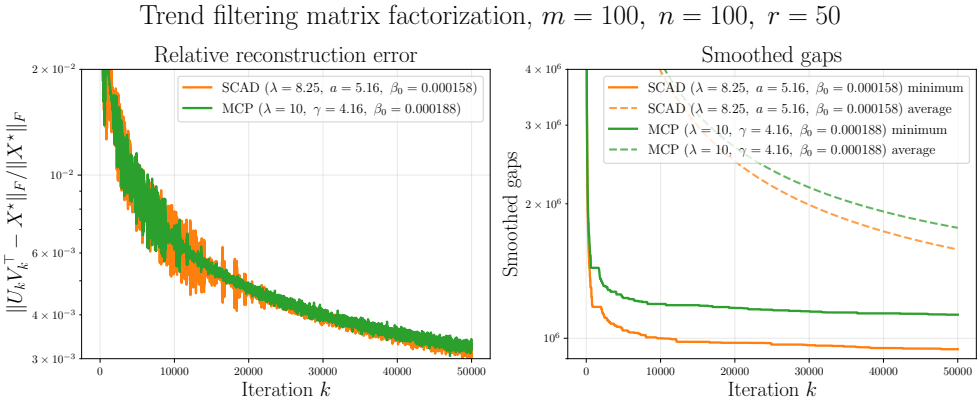

Fig. 4: FRAMES applied to trend-filtered matrix factorization problem of Section 7.2.2 with tuned MCP and SCAD penalties.

Data Generation

The ground-truth factor $U^\star$ is generated with five constant blocks per column, with block heights drawn independently and block boundaries shared across columns. Thus $D_{\text{row}}U^\star$ is sparse. The factor $V^\star$ has i.i.d. entries distributed as $|\mathcal{N}(0,1)|$, and $X^\star = U^\star(V^\star)^\top$. In this experiment we use $m = n = 100$, rank $r = 50$, and $N = 50,000$ iterations for the displayed trajectories. The displayed SCAD run uses the hyperparameters

$$\text{SCAD}: \quad \lambda = 8.25, \quad a = 5.16, \quad \beta_0 = 1.58 \cdot 10^{-4},$$

and the displayed MCP run uses the hyperparameters

$$\text{MCP}: \quad \lambda = 10, \quad \gamma = 4.16, \quad \beta_0 = 1.88 \cdot 10^{-4},$$

which were found with a simple grid search for the plot; we have included them here for reproducibility but they are otherwise unimportant for this experiment.

Results

Figure 4 compares the tuned SCAD and MCP runs.

- *Reconstruction error.* The left panel shows that both nonconvex trend-filtering penalties reduce the relative reconstruction error to the same range, with MCP slightly lower at the end of the run.
- *Smoothed gaps.* The right panel plots both the running minimal smoothed gap up to iteration k (solid curves) and the average of the smoothed gaps (dashed curves). The smoothed gaps decrease for both penalties at the rates predicted by Theorem 5.3, with SCAD producing the smaller smoothed gaps on this instance.

7.3 Inconsistent Problem over the ℓ_∞ Ball

We consider an example with inconsistent constraints in $\mathbb{R}^2$ (see Section 6.2). Let

$$\mathcal{C} = [-1, 1]^2, \; T = \begin{bmatrix} 0 & 1 \end{bmatrix}, \; \mathcal{D} = \{2\}.$$

Thus $T\colon \mathbb{R}^2 \to \mathbb{R}$ is the coordinate projection $T(x_1, x_2) = x_2$. Since $T(\mathcal{C}) = [-1, 1]$, we have $T(\mathcal{C}) \cap \mathcal{D} = \varnothing$, and the indicator problem

$$\min_{x \in \mathcal{C}} f(x) + \iota_{\mathcal{D}}(Tx)$$

has no feasible point. The closest-point set from (6.5) is

$$\begin{aligned} \mathcal{C}^\dagger &= \operatorname*{argmin}_{x \in \mathcal{C}} \operatorname{dist}_{\mathcal{D}}(Tx) \\ &= \{(x_1, 1) : x_1 \in [-1, 1]\}, \end{aligned} \tag{7.6}$$

and $\delta = \min_{x \in \mathcal{C}} \operatorname{dist}_{\mathcal{D}}(Tx) = 1$.

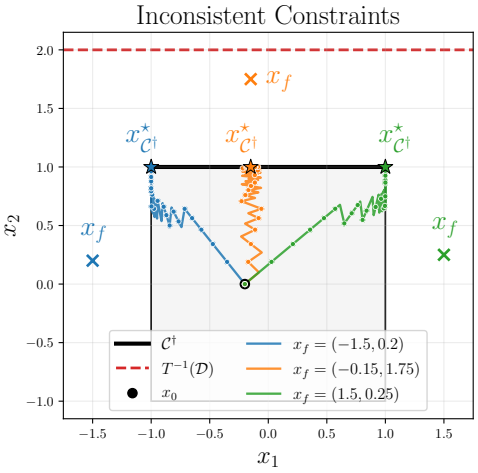

Fig. 5: Trajectories for FRAMES applied to the inconsistent problem with $\mathcal{C} = [-1, 1]^2$, $T(x_1, x_2) = x_2$, and $\mathcal{D} = \{2\}$. The dashed red line is $T^{-1}(\mathcal{D}) = \{(x_1, x_2) : x_2 = 2\}$, while the black segment is the closest-point set $\mathcal{C}^\dagger$. The three colored trajectories correspond to the three choices of x_f in $f(x) = \|x - x_f\|^2$.

Secondary Selection

To illustrate the selection of a point in $\mathcal{C}^\dagger$, we use the smooth objective

$$f(x) = \|x - x_f\|^2$$

with various choices of anchor point x_f: either $(-1.5, 0.2)$, $(-0.15, 1.75)$, or $(1.5, 0.25)$. For this choice of f, the secondary solution is

$$x^\star_{\mathcal{C}^\dagger} \in \operatorname*{argmin}_{s \in \mathcal{C}^\dagger} f(s) = \mathrm{P}_{\mathcal{C}^\dagger}(x_f),$$

which gives $x^\star_{\mathcal{C}^\dagger} = (-1, 1)$, $x^\star_{\mathcal{C}^\dagger} = (-0.15, 1)$, and $x^\star_{\mathcal{C}^\dagger} = (1, 1)$ for the three different choices of anchor point, respectively.

Algorithmic Details

We run **FRAMES** from $x_0 = (-0.2, 0)$ for $N = 1500$ iterations with $\beta_k = \beta_0(k+1)^{-1/4}$, $\beta_0 = 3$, and $\gamma_k = (k+100)^{-1/2}$. The shifted step size is used only to help visualize the initial part of the trajectory; the smoothing schedule is the same power schedule used throughout the experiments. Using Theorem 6.9, convergence for this schedule is also guaranteed using a similar argument to Corollary 6.10.

Results

Figure 5 shows that the iterates approach $\mathcal{C}^\dagger$ and then select the point in $\mathcal{C}^\dagger$ minimizing the smooth objective. Figure 6 plots $|\mathrm{dist}_{\mathcal{D}}(Tx_k) - \delta|$, which converges to zero because $\mathrm{dist}_{\mathcal{D}}(Tx_k) \to \delta = 1$, and $|f(x_k) - f(x^\star_{\mathcal{C}^\dagger})|$, showing convergence to the secondary minimizer predicted by Theorem 6.9. In this example, f is convex on $\mathcal{C}^\dagger$, so stationarity implies optimality.

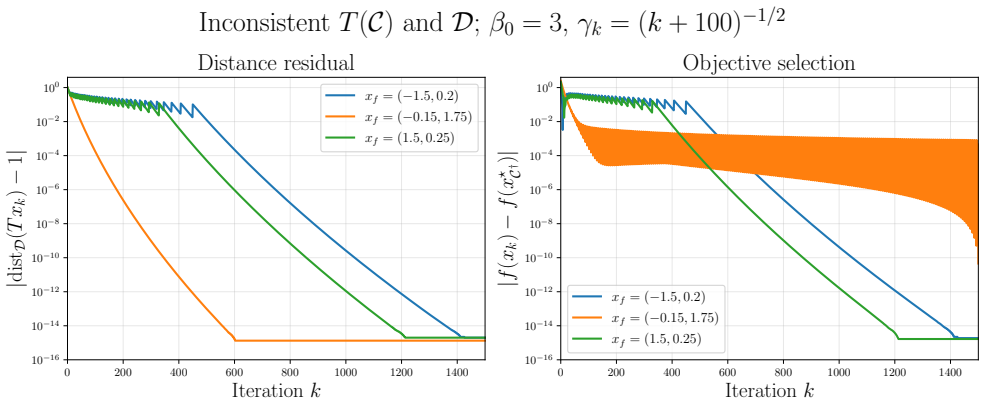

Fig. 6: Convergence profiles for **FRAMES** applied to the inconsistent problem where $\delta = 1$.

8 Conclusion

We have shown how to extend Frank-Wolfe to the nonsmooth nonconvex composite setting using the Moreau envelope. Our analysis reveals a delicate relationship between the step size and the smoothing parameter, with both having a strong effect on the rate of convergence for the Frank-Wolfe gap of the original problem (P). Larger smoothing parameters make the smoothed objectives easier to optimize but give poorer approximations of the original nonsmooth problem, while smaller smoothing parameters improve the approximation at the cost of worse smoothness constants and slower convergence of the stationarity criterion for (P). In future work, we would like to extend these ideas and their analysis to some unconstrained analogs of Frank-Wolfe, such as normalized steepest descent [44]. We also think it would be fruitful to develop a short step analysis of nonsmooth Frank-Wolfe. A main limitation of the finite-time Lipschitz-case gap-transfer result is the proximal-lift assumption, which requires a feasible point $z \in \mathcal{C}$ satisfying $Tz = \mathrm{prox}_{\beta g}(Tx)$. Removing or weakening this assumption is an important direction for future work.

Acknowledgements. This work was supported by the French National Research Agency (ANR) under grant ANR-25-CE23-3749 (project SIMPLES). The work of ZW was supported by the National Science Foundation under Grant DMS-2532423.

Competing Interests. Not applicable.

Availability of Data and Materials. All code is available on GitHub.

References

[1] Frank, M., Wolfe, P.: An algorithm for quadratic programming. Naval research logitistics quarterly **3**(1-2), 95–110 (1956)

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

[3] Argyriou, A., Signoretto, M., Suykens, J.: Hybrid conditional gradient-smoothing algorithms with applications to sparse and low rank regularization. Regularization, Optimization, Kernels, and Support Vector Machines, 53–82 (2014)

[4] Lacoste-Julien, S., Jaggi, M.: On the global linear convergence of Frank-Wolfe optimization variants. In: NIPS, pp. 496–504 (2015)

[5] Combettes, C., Pokutta, S.: Boosting Frank–Wolfe by chasing gradients. In: International Conference on Machine Learning, pp. 2111–2121 (2020). PMLR

[6] Combettes, C.W., Pokutta, S.: Complexity of linear minimization and projection on some sets. Operations Research Letters **49**(4), 565–571 (2021)

[7] Woodstock, Z.: High-precision linear minimization is no slower than projection. Optimization Letters, 1–7 (2026)

[8] Yurtsever, A., Fercoq, O., Locatello, F., Cevher, V.: A conditional gradient framework for composite convex minimization with applications to semidefinite programming. In: International Conference on Machine Learning, pp. 5727–5736 (2018). PMLR

[9] Yurtsever, A., Fercoq, O., Cevher, V.: A conditional-gradient-based augmented Lagrangian framework. In: International Conference on Machine Learning, pp. 7272–7281 (2019). PMLR

[10] Silveti-Falls, A., Molinari, C., Fadili, J.: Generalized conditional gradient with augmented Lagrangian for composite minimization. SIAM Journal on Optimization **30**(4), 2687–2725 (2020)

[11] Woodstock, Z., Pokutta, S.: Splitting the conditional gradient algorithm. SIAM Journal on Optimization **35**(1), 347–368 (2025)

[12] Bolte, J., Combettes, C.W., Pauwels, E.: The iterates of the Frank–Wolfe algorithm may not converge. Mathematics of Operations Research **49**(4), 2565–2578 (2024)

[13] Halbey, J., Rakotomandimby, S., Besançon, M., Designolle, S., Pokutta, S.: Efficient quadratic corrections for frank-wolfe algorithms. In: Proceedings of the Conference on Neural Information Processing Systems, vol. 38 (2025)

[14] Silveti-Falls, A., Molinari, C., Fadili, J.: Inexact and stochastic generalized conditional gradient with augmented Lagrangian and proximal step. Journal of Nonsmooth Analysis and Optimization **2**(Original research articles) (2021)

[15] Locatello, F., Yurtsever, A., Fercoq, O., Cevher, V.: Stochastic Conditional Gradient Method for Composite Convex Minimization. arXiv e-prints, 1901–10348 (2019) arXiv:1901.10348 [math.OC]

[16] Thekumparampil, K.K., Jain, P., Netrapalli, P., Oh, S.: Projection efficient subgradient method and optimal nonsmooth Frank–Wolfe method. Advances in neural information processing systems **33**, 12211–12224 (2020)

[17] Thekumparampil, K.K., Jain, P., Netrapalli, P., Oh, S.: Optimal nonsmooth Frank–Wolfe method for stochastic regret minimization. In: 12th Annual Workshop on Optimization for Machine Learning (2020)

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

[19] Böhm, A., Wright, S.J.: Variable smoothing for weakly convex composite functions. Journal of optimization theory and applications **188**(3), 628–649 (2021)

[20] Lan, G., Zhou, Y.: Conditional gradient sliding for convex optimization. SIAM Journal on Optimization **26**(2), 1379–1409 (2016)

[21] Ouyang, Y., Squires, T.: Universal conditional gradient sliding for convex optimization. SIAM Journal on Optimization **33**(4), 2962–2987 (2023)

[22] Ito, M., Lu, Z., He, C.: A parameter-free conditional gradient method for composite minimization under hölder condition. Journal of Machine Learning Research **24**(166), 1–34 (2023)

[23] De Oliveira, W.: Short paper-a note on the Frank–Wolfe algorithm for a class of nonconvex and nonsmooth optimization problems. Open Journal of Mathematical Optimization **4**, 1–10 (2023)

[24] Kreimeier, T., Pokutta, S., Walther, A., Woodstock, Z.: On a Frank–Wolfe approach for abs-smooth functions. Optimization Methods and Software, 1–27 (2024)

[25] Ravi, S.N., Collins, M.D., Singh, V.: A deterministic nonsmooth Frank–Wolfe algorithm with coreset guarantees. Informs Journal on Optimization **1**(2), 120–142 (2019)

[26] Cheung, E., Li, Y.: Solving separable nonsmooth problems using Frank–Wolfe with uniform affine approximations. In: IJCAI, pp. 2035–2041 (2018)

[27] Asgari, K., Neely, M.J.: Nonsmooth projection-free optimization with functional constraints. Computational Optimization and Applications **89**(3), 927–975 (2024)

[28] Mazanti, G., Moquet, T., Pfeiffer, L.: A nonsmooth Frank–Wolfe algorithm through a dual cutting-plane approach. Journal of Optimization Theory and Applications **207**(2), 29 (2025)

[29] Besançon, M., Carderera, A., Pokutta, S.: FrankWolfe.jl: A high-performance and flexible toolbox for Frank-Wolfe algorithms and conditional gradients. INFORMS Journal on Computing (2022)

[30] Chierchia, G., Chouzenoux, E., Combettes, P.L., Pesquet, J.-C.: The Proximity Operator Repository. https://proximity-operator.net/

[31] Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis vol. 317. Springer, Berlin (2009)

[32] Hoheisel, T., Laborde, M., Oberman, A.: On proximal point-type algorithms for weakly convex functions and their connection to the backward euler method. Optimization Online (2010)

[33] Clarke, F.H.: Optimization and Nonsmooth Analysis. SIAM, Philadelphia (1990)

[34] Braun, G., Carderera, A., Combettes, C.W., Hassani, H., Karbasi, A., Mokhtari, A., Pokutta, S.: Conditional Gradient Methods: From Core Principles to AI Applications. SIAM, Philadelphia (2025)

[35] Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces, 2nd edn. CMS Books in Mathematics. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-48311-5

[36] Bauschke, H.H., Borwein, J.M., Li, W.: Strong conical hull intersection property, bounded linear regularity, Jameson's property (G), and error bounds in convex optimization. Math. Program., Ser. A **86**(4), 135–160 (1999)

[37] Gillis, N.: Nonnegative Matrix Factorization. SIAM, Philadelphia (2020)

[38] Wang, Y.-X., Sharpnack, J., Smola, A.J., Tibshirani, R.J.: Trend filtering on graphs. Journal of Machine Learning Research **17**(105), 1–41 (2016)

[39] Amsel, N., Persson, D., Musco, C., Gower, R.M.: The polar express: Optimal matrix sign methods and their application to the muon algorithm. arXiv preprint arXiv:2505.16932 (2025)

[40] Kim, S.-J., Koh, K., Boyd, S., Gorinevsky, D.: ℓ_1 trend filtering. SIAM review **51**(2), 339–360 (2009)

[41] Fan, W., Liu, X., Jin, W., Zhao, X., Tang, J., Li, Q.: Graph trend filtering networks for recommendation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 112–121 (2022)

[42] Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association **96**(456), 1348–1360 (2001)

[43] Zhang, C.-H.: Nearly unbiased variable selection under minimax concave penalty (2010)

[44] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge university press, Cambridge (2004)