# A PROXIMAL ORACLE SEQUENTIAL NLP ALGORITHM FOR CONSTRAINED NONSMOOTH OPTIMIZATION*

COSMIN G. PETRA†

**Abstract.** We develop a proximal sequential nonlinear programming (NLP) method for constrained finite-sum optimization with nonsmooth, nonconvex objective terms. Each nonsmooth summand is assumed only to be proper, lower semicontinuous, and prox-bounded, and is accessed through a proximal oracle requiring no derivatives or sensitivity information. The method applies Moreau envelope regularization separately to each summand, preserving separability while retaining the original smooth nonlinear constraints. We establish consistency of this distributed Moreau continuation: as the regularization parameter vanishes, optimal values and global minimizers of the regularized problems converge, along subsequences, to those of the original problem. We then propose a continuation algorithm whose inner loop solves sequential NLP majorization models. These models enforce the smooth constraints exactly and represent the regularized nonsmooth terms by proximal oracle subgradient linearizations plus quadratic stabilization, using the upper-$C^2$ property of Moreau envelopes; an acceptance-ratio test adaptively updates the stabilization parameter. For fixed regularization, the cluster points of the inner iteration satisfy the stationarity conditions for the regularized problem, and the inner loop terminates finitely for any positive stationarity tolerance. For prescribed positive regularization and stationarity tolerances, the full continuation scheme also terminates finitely and returns an approximately limiting sum-stationary point with an explicit error bound. As the regularization and tolerances vanish, cluster points satisfy first-order limiting stationarity for the original nonsmooth problem under standard constraint qualifications and an attentive convergence condition, which simplify for locally Lipschitz terms. Finally, we show how the algorithm applies when the nonsmooth summands are optimal value functions of parametric optimization problems, for which the proximal oracles can be implemented by quadratically stabilized subproblems.

## 1. Introduction.

We consider optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \quad F(x) := f(x) + \sum_{i\in\mathcal{K}} r_i(x), \tag{1.1}$$

$$\text{s.t.} \quad h(x) = 0, \;\; x_\ell \le x \le x_u, \tag{1.2}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^m$ are $C^2$, and each $r_i : \mathbb{R}^n \to \mathbb{R}$ is proper, lower semicontinuous (lsc), finite-valued, and possibly nonconvex and nonsmooth, for each $i$ in a finite index set $\mathcal{K}$. Problems with this structure arise in several applications; in particular, our main motivation comes from stochastic programming with recourse and security-constrained power system optimization, where the functions $r_i$ are optimal value functions of parametric optimization problems, as elaborated in Section 5.

We assume that the functions $r_i$ are accessed through an oracle. In many applications, evaluating $r_i(x)$ already requires solving an optimization subproblem, while derivatives of $r_i$ with respect to $x$ are unavailable, unreliable, or expensive to compute. Rather than requiring such sensitivities, we assume access to a proximal oracle, namely the value of the Moreau envelope [30] of $r_i$

$$e_\lambda r_i(x) \;:=\; \inf_{z\in\mathbb{R}^n} \Big\{ r_i(z) \;+\; \frac{1}{2\lambda}\,\|z-x\|^2 \Big\}, \tag{1.3}$$

together with a proximal point

$$w_i \in \operatorname{prox}_{\lambda} r_i(x) := \arg\min_{z\in\mathbb{R}^n} \Big\{ r_i(z) + \frac{1}{2\lambda}\|z-x\|^2 \Big\}, \; \forall i \in \mathcal{K},$$

*This work was supported by the U.S. Department of Energy through the Scientific Discovery through Advanced Computing (SciDAC) program under the ASCR–OE SLOPE-Grid partnership.

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA, USA. Email: petra1@llnl.gov. Release number: LLNL-JRNL-2019378

for $\lambda > 0$. When $r_i$ is defined through an optimization problem, existing numerical machinery can often be adapted to evaluate this oracle with little structural change.

Our starting point is the observation that the Moreau envelope regularizes each nonsmooth term. Under the mild assumption of prox-boundedness, $e_\lambda r_i$ converges pointwise to $r_i$ as $\lambda \downarrow 0$ and is locally Lipschitz and upper-$C^2$ for sufficiently small $\lambda$. This suggests replacing (1.1)–(1.2) by the family of regularized problems

$$\min_{x \in \mathbb{R}^n} \quad F_\lambda(x) := f(x) + \sum_{i \in \mathcal{K}} e_\lambda r_i(x) \tag{1.4}$$

$$\text{s.t.} \quad h(x) = 0 \ \text{ and } \ x_\ell \le x \le x_u, \tag{1.5}$$

and driving $\lambda \downarrow 0$ in a continuation scheme.

For fixed $\lambda$, the objective in (1.4) is a sum of a smooth term and upper-$C^2$ terms, and is therefore upper-$C^2$. At first sight, this places the fixed-$\lambda$ problem within the scope of the sequential quadratic programming (SQP) framework of Wang and Petra [33] for nonsmooth problems with upper-$C^2$ objectives. The present work, however, is not a direct application of that framework. In [33], the nonsmooth objective (1.4) is treated as an upper-$C^2$ function for which Clarke subgradients are assumed to be available; here, we do not assume access to subgradients of either $r_i$ or $e_\lambda r_i$. The only information used by the algorithm is a proximal oracle for each $r_i$.

The use of Moreau envelopes leads to a different stationarity mechanism, which ultimately provides limiting/Mordukhovich subgradients stationarity for the original problem and is stronger than the Clarke stationarity of [33] even for the regularized problems, despite apparently not using subgradients. The key idea is to use proximal points $w_i \in \operatorname{prox}_\lambda r_i(x)$, produced by the oracle, to form the vectors

$$v_i = \frac{1}{\lambda}(x - w_i).$$

For general lsc $r_i$, the vectors $v_i$ may not be limiting subgradients of $e_\lambda r_i$ at $x$, but:

(i) satisfy a majorization bound specific to upper-$C^2$ functions; and

(ii) are limiting subgradients of $r_i$ at the nearby proximal point $w_i$.

The property (ii) is noteworthy and plays an important role in the analysis of the algorithm proposed here. The property stems from the use of Moreau envelopes and is not present for general upper-$C^2$ functions. We refer to the vectors $v_i$ as *proximal oracle subgradients* to emphasize their Moreau envelope origin and to distinguish them from proximal subgradients previously used in the literature [30].

Consequently, property (i) allows us to develop a sequential proximal oracle NLP algorithm for the regularized, fixed-$\lambda$ constrained problems (1.4)–(1.5). The method builds convex quadratic stabilized models for the nonsmooth objective term(s) using only proximal points and uses an NLP model for step computations. The choice of an NLP model instead of an SQP model is made to enforce the smooth constraints and objective term exactly, as required in important classes of problems, such as power grid optimization, which motivates this work. An acceptance ratio test is also devised to adjust the stabilization parameter and obtain a simple descent-type of scheme. We prove that the inner, fixed-$\lambda$ iterations result in cluster points that satisfy stationarity conditions for the regularized problem and terminate finitely for any positive stationarity tolerance.

Furthermore, the property (ii) above yields an outer-loop convergence analysis for the continuation scheme that carries, as $\lambda \downarrow 0$, the stationarity certificates for the fixed-$\lambda$ problems to limiting subgradients-based stationarity of the original nonsmooth

problem. More specifically, under standard constraint qualifications (CQs) and a mild attentive convergence condition—which holds automatically in several important cases, including locally Lipschitz $r_i$—we show that every cluster point generated as $\lambda \downarrow 0$ satisfies first-order stationarity conditions for (1.1)–(1.2).

Applying Moreau regularization separately to individual summands $r_i$ preserves the finite-sum structure in the regularized problems, allows parallel proximal computations, and results in single-level decomposition with significant computational parallelism. For example, we have previously shown that this single-level decomposition is particularly well suited when each $r_i$ is the optimal value of a computationally expensive optimization subproblem and can scale up to exascale supercomputers [20]. However, since infimal convolution does not in general commute with summation, replacing each $r_i$ by $e_\lambda r_i$ is not merely a rewriting of the Moreau envelope of the full nonsmooth objective. We show that, under mild assumptions, the regularized problems (1.4)–(1.5) are nevertheless consistent with the original problem, in the sense that global minimizers of the regularized problems converge, along subsequences, to global minimizers of (1.1)–(1.2) as $\lambda \downarrow 0$.

**1.1. Related work.** Our approach is related to proximal point algorithms (PPAs) [4, 5, 26], but differs in an essential way. Classical PPAs regularize the entire objective and would require evaluating the proximal mapping of $F$, or equivalently solving subproblems involving $e_\lambda F$. In our setting, this would couple all objective terms and destroy the separable structure and the parallelization potential of the single-level decomposition associated with the nonsmooth part.

Compared with Douglas-Rachford splitting methods [21, 31, 32, 17] and related alternating direction method of multipliers schemes [34, 15, 1, 2, 16], the present framework is designed to work directly with the native structure of (1.1)–(1.2). Splitting methods are typically most natural after reformulating the problem into a two-block or linearly constrained consensus form, which may require aggregating the nonsmooth terms or introducing auxiliary variables and coupling constraints. In contrast, we keep the smooth nonlinear constraints explicit and preserve the finite-sum structure.

The closest broad framework is the constrained composite augmented Lagrangian approach of [12], which treats problems with a single nonsmooth composite term and general set-membership constraints. The framework of [12] is more general, but it does not take advantage of the particular oracle and separability structures that motivate this work. Our setting is deliberately more structured: each nonsmooth summand is accessed through its proximal oracle, and the original smooth nonlinear constraints are retained explicitly rather than absorbed into a reformulated composite constraint, which would obscure the term-wise subgradients needed in the continuation analysis.

The inner-loop sequential NLP approach here is related to nonsmooth SQP and gradient-sampling methods; see, for example, [35, 10, 6, 8, 9]. These methods build SQP models using sampled gradients or quasi-Newton approximations for locally Lipschitz nonsmooth constrained problems. However, our stationarity analysis relies on proximal points and deterministic nearby limiting subgradients rather than on Clarke subgradient sampling or penalty-SQP stationarity. Moreover, our setting assumes no derivative or sampled-gradient access to the nonsmooth summands.

**1.2. Contributions.** The paper's main contribution is a proximal-oracle sequential NLP algorithm for constrained finite-sum nonsmooth, nonconvex optimization under very general assumptions: each nonsmooth summand need only be proper, lsc, and prox-bounded, with no convexity, weak convexity, prox-regularity, or differentiability required. The algorithm uses proximal oracle evaluations rather than

sensitivity information and preserves the single-level decomposition of the original problem. It provides finite inner-loop termination and stationarity for the regularized problems by using the upper-$C^2$ property and proximal oracle subgradients features specific to Moreau envelopes. The stationarity certificates from the regularized problems are general enough to be carried via an outer continuation analysis and to obtain limiting first-order sum-stationarity for the original nonsmooth problem under standard CQs and a mild attentive convergence condition. The algorithm applies naturally to and is well suited for nonsmooth terms defined by optimal value functions, where proximal oracles can be implemented via stabilized subproblems.

**1.3. Organization.** The remainder of the paper is organized as follows. Section 2 reviews the variational analysis tools used throughout, with emphasis on upper-$C^2$ functions and subdifferential properties of Moreau envelopes. Section 3 establishes the well-posedness and consistency of the Moreau envelope continuation approach as the regularization parameter tends to zero. Section 4 presents the algorithm and analyzes both the fixed-$\lambda$ inner iterations and the outer continuation scheme. Section 5 discusses how the framework applies naturally to parametric recourse optimization models, with a case study for electric power grids. The focus here is theoretical rather than computational, as the present paper establishes the prerequisites needed to justify and support future computational developments. Section 6 summarizes some of these developments and provides concluding remarks.

**2. Preliminary technical results.** Let $r : \mathbb{R}^n \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$, and let $\bar{x}$ be such $r(\bar{x}) \in \mathbb{R}$. A vector $v \in \mathbb{R}^n$ is in the regular (Fréchet) subdifferential [30] (denoted by $\hat{\partial} r(\bar{x})$) if $r(x) \geq r(\bar{x}) + \langle v, x - \bar{x}\rangle + o(\|x - \bar{x}\|)$.

Following again [30, Definition 8.2], we define the limiting (also known as Mordukhovich) subdifferential, denoted by $\partial r(\bar{x})$. A vector $v \in \mathbb{R}^n$ is a limiting subgradient of $r$ at $\bar{x}$ (written $v \in \partial f(\bar{x})$) if there exist sequences

$$x^\nu \to_r \bar{x} \text{ and } v^\nu \in \hat{\partial} r(x^\nu) \text{ such that } v^\nu \to v.$$

Here, $x^\nu \to_r \bar{x}$ denotes the $r$-attentive convergence and holds whenever $x^\nu \to \bar{x}$ and $r(x^\nu) \to r(\bar{x})$ [30, Def. 8.2]. The *horizon subdifferential* is

$$\partial^\infty r(\bar{x}) = \Big\{ v^\infty : \exists x^\nu \to_r \bar{x},\ t_\nu \downarrow 0, \text{ and } v^\nu \in \hat{\partial} r(x^\nu) \text{ such that } t^\nu v_\nu \to v^\infty, \Big\}.$$

For a closed set $C \subset \mathbb{R}^n$, the *limiting normal cone* is defined as $N_C(x) := \partial \delta_C(x)$, where $\delta_C$ is the indicator of $C$, defined as $\delta_C(x) = 0$ if $x \in C$ and $+\infty$ otherwise.

For a Lipschitz continuous function $r$ we also use Clarke subdifferentials [30, 29, 7] defined by

$$\partial^C r(x) = \big\{ v \in \mathbb{R}^n \,\big|\, r^\circ(x; d) \geq \langle v, d\rangle \text{ for all } d \in \mathbb{R}^n \big\}.$$

where $r^\circ(x; d)$ is the directional Clarke derivative along direction $d \in \mathbb{R}^n$, defined as

$$r^\circ(x; d) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{r(y + td) - r(y)}{t}.$$

The function $r$ is called *prox-bounded* if there exists some $\lambda > 0$ such that $e_\lambda r(x) > -\infty$ for some $x \in \mathbb{R}^n$. The prox-threshold of $r$ is the supremum over all such $\lambda > 0$. For lsc, proper functions, prox-boundedness is equivalent to the function $x \mapsto r + \frac{1}{2}\alpha\|\cdot\|^2$ being bounded from below for some $\alpha > 0$ [30, Exercise 1.24]. We remark

that any function bounded from below is prox-bounded (with prox-threshold $+\infty$). Also, for every $x \in \mathbb{R}^n$, $\operatorname{prox}_\lambda r(x)$ is nonempty and compact [30, Theorem 1.25] for any $\lambda > 0$ smaller than the prox-threshold.

We build on the concept of upper-$C^k$ ($k \in \{1, 2, \ldots\}$) functions in this work. Following Rockafellar and Wets [30, Definition 10.29], a real-valued function $f$ on an open set $O \in \mathbb{R}^n$ is upper-$C^k$ if locally $r$ is the minimum of a family of $C^k$ functions indexed over a compact set, namely, $r(x) = \min_{t \in T} f_t(x)$ for all $x$ in a neighborhood of $\bar{x} \in O$, with $T$ compact, $f_t \in C^k$, and the values and all partial derivatives up to order $k$ depending jointly continuously on $(t, x)$. The minimum of finitely many $C^k$ functions is therefore upper-$C^k$, but may fail to be differentiable. There is no distinction for orders $k \geq 2$ since upper-$C^2$ implies upper-$C^\infty$ [30, Corollary 10.34]. Some prominent members of the upper-$C^2$ class are squared distance functions to closed sets and Moreau envelopes of lsc, proper, prox-bounded functions for sufficiently small regularization parameters.

A useful equivalent characterization of upper-$C^2$ functions is via local difference of smooth and finite convex function, namely a function $r$ is upper-$C^2$ if and only if locally $r = r_{sm} - r_{co}$ with $r_{co}$ a finite convex function and $r_{sm} \in C^2$ [30, Theorem 10.33]. In fact, one may take $r_{sm} = \frac{\rho}{2}\|x\|_2^2$ locally [30, Theorem 10.33], which is equivalent to $r - \frac{\rho}{2}\|\cdot\|_2^2$ being concave locally. Recently, this property has been exploited to devise simplified nonsmooth algorithms for upper-$C^2$ objectives that need not maintain bundles, *e.g.*, [33]; we also leverage it here (see quadratic majorization Lemma 4.3) to obtain convergence to stationary points of the regularized problems (see Proposition 4.5) using proximal oracle subgradients (see Lemma 2.1 below).

Conversely, a function $r$ is lower-$C^k$ if $-r$ is upper-$C^k$. Lower-$C^2$ functions enjoy better regularity than upper-$C^2$, in particular they are subdifferentially regular (also known as lower regular) and, for example, the concepts of regular, limiting, and Clarke subdifferentials coincide [30]. On the other hand, the strict inclusion $\hat{\partial} \subset \partial \subset \partial^C$ can hold for upper-$C^2$ functions in general.

A salient feature of Moreau envelopes of lsc, proper, and prox-bounded functions, not available for general upper-$C^2$ functions, is that their limiting subdifferentials can be characterized by means of proximal points as shown by the following lemma.

LEMMA 2.1. *Let $r : \mathbb{R}^n \to (-\infty, +\infty]$ be proper lsc and prox-bounded, and let $\lambda > 0$ be below the prox-threshold. We have the following characterizations of limiting subdifferentials of the Moreau envelopes at a given $x \in \mathbb{R}^n$:*

*(i)* $\operatorname{prox}_\lambda r(x) \subset (I + \lambda\,\partial r)^{-1}(x)$;

*(ii)* $\partial e_\lambda r(x) \subset \frac{1}{\lambda}\big(x - \operatorname{prox}_\lambda r(x)\big)$;

*(iii) For every $v \in \partial e_\lambda r(x)$ there exists $y \in \operatorname{prox}_\lambda r(x)$ such that $v = \frac{x-y}{\lambda}$ and $v \in \partial r(y)$.*

*Proof.* (i) and (ii) are Examples 10.2 and 10.32 from [30], respectively.

(iii) Consider $v \in \partial e_\lambda r(x)$. (ii) implies there exists $y \in \operatorname{prox}_\lambda r(x)$ such that $v = \frac{x-y}{\lambda}$. Also, (i) indicates that $x \in y + \lambda\partial r(y)$, which implies $v \in \partial r(y)$ and completes the proof. ∎

Lemma 2.1 (ii) above shows that any (limiting) subgradient $v \in \partial e_\lambda r(x)$ corresponds to a minimizer $y \in \operatorname{prox}_\lambda r(x)$ of the Moreau envelope optimization problem (1.3). However, the converse is not necessarily true as the inclusion in (ii) can be strict. Intuitively, this can occur, for example, when $\operatorname{prox}_\lambda r(x)$ is not a singleton, as illustrated by the following example.

*Example* 2.2. Consider $e_{\frac{1}{2}} r(x) = \frac{1}{2}x^2 - |x| + \frac{1}{2}$, the Moreau envelope of $r(z) =$

$\frac{1}{2} - z^2$ if $|z| \le \frac{1}{2}$ and $r(z) = (|z| - 1)^2$ if $|z| \ge \frac{1}{2}$. For this, we observe that

$$\partial e_{1/2} r(0) = \{-1, 1\} \subsetneq [-1, 1] = 2\big(0 - \operatorname{prox}_{1/2} r(0)\big).$$

A complication caused by the possibly strict inclusion in (ii) of Lemma 2.1 is that minimizers/proximal points $y \in \operatorname{prox}_\lambda r(x)$ of (1.3) do not necessarily provide a subgradient in $\partial e_\lambda r(x)$. In the algorithm proposed in Section 4, we circumvent this limitation by using the proximal oracle subgradients $v = \frac{1}{\lambda}(x - y)$ instead of the true limiting subgradients of $\partial e_\lambda r$ at $x$ to solve the Moreau regularized problem for fixed $\lambda > 0$. Notably, Lemma 2.1 (i) shows that these proximal oracle subgradients are limiting subgradients of $r$ at proximal points $y \in \operatorname{prox}_\lambda r(x)$. This fact and the convergence consistency between proximal points and algorithm iterates (see Lemma 4.8) are key ingredients in obtaining (asymptotic) convergence of the algorithm to a limiting stationary point to the target problem (1.1)–(1.2) without using the limiting subgradients of the Moreau envelope approximants from (1.4)–(1.5).

*Remark* 2.3. If $r$ is convex, or more generally weakly convex [11], then for $\lambda$ small enough, $\operatorname{prox}_\lambda r$ is single valued and $e_\lambda r$ is $C^1$, so Lemma 2.1 (ii) holds with equality. For prox-regular [28] or lower-$C^2$ functions, same conclusion holds locally on neighborhoods where the proximal mapping is single-valued [30].

In the remainder of the paper, we work under the following two assumptions.

*Assumption* 2.4. Functions $f$ and $h$ are $C^2$ and functions $r_i$ are lsc, finite valued, and prox-bounded.

We remark that prox-boundedness is a general assumption (for example, holds for any function bounded from below) and that we do not assume convexity or weak-convexity (*e.g.*, lower-$C^1$, lower-$C^2$), prox-regularity, or differentiability of $r_i$.

Let us define the shorthand

$$r_\lambda(x) = \sum_{i \in \mathcal{K}} e_\lambda r_i(x)$$

and denote the feasible set of the master problem (1.1)–(1.2) (also of (1.4)–(1.5)) by

$$C := \{x \in \mathbb{R}^n : \ h(x) = 0, \ x_\ell \le x \le x_u\}.$$

For simplicity, we assume that the box bounds $x_\ell$ and $x_u$ are finite so that $C$ is compact. A more general assumption would be that the feasible sublevel sets of $F$ are compact, but this would introduce considerable additional technical complexity without adding much insight.

*Assumption* 2.5. The constraint set $C$ is nonempty and compact.

**3. Moreau continuation: well-posedness and consistency.** This section is concerned with the well-posedness of the continuation and proves that the minimizers of the Moreau envelope-regularized problem (1.4)–(1.5) converge, and every cluster point is a minimizer of the original problem (1.1)–(1.2) as $\lambda \downarrow 0$.

Let us define the extended-real objectives

$$G(x) = F(x) + \delta_C(x) \text{ and } G_\lambda(x) = F_\lambda(x) + \delta_C(x).$$

THEOREM 3.1. *Choose $\bar{\lambda} > 0$ below the prox-thresholds of $r_i$, $i \in \mathcal{K}$, and consider a sequence $\{\lambda_k\}_{k \in \mathbb{N}}$ such that $\lambda_k \downarrow 0$ and $\lambda_k < \bar{\lambda}$. Then the following properties hold under Assumptions 2.4 and 2.5.*

*(i)* $\{G_{\lambda_k}\}$ *converges pointwise and monotonically to* $G$, $G_{\lambda_k}(x) \uparrow G(x)$ *for all* $x$.
*(ii)* $\{G_{\lambda_k}\}$ *epi-converges to* $G$, *denoted as* $G_{\lambda_k} \underset{e}{\to} G$.
*(iii) Optimal values converge,*

$$\inf_{x\in C} F_{\lambda_k}(x) = \inf_x G_{\lambda_k}(x) \to \inf_x G(x) = \inf_{x\in C} F(x).$$

*(iv) Moreover, for each* $k$, *the regularized problem* (1.4)–(1.5) *admits a global minimizer, and the original problem* (1.1)–(1.2) *admits a global minimizer. For any choice of minimizers* $x(\lambda_k)$ *of* (1.4)–(1.5), *the sequence* $\{x(\lambda_k)\}$ *is bounded and every cluster point* $x^\star$ *is a minimizer of* (1.1)–(1.2). *If* $x^\star$ *is the unique solution of* (1.1)–(1.2), *then* $x(\lambda_k) \to x^\star$.

*Proof.* (i) For $x \notin C$, one has $G_{\lambda_k}(x) = G(x) = \infty$. For $x \in C$, since $r_i$ are prox-bounded proper lsc functions, the Moreau envelope satisfies [30, Theorem 1.25]

$$e_{\lambda_k} r_i(x) \uparrow r_i(x) \quad \text{as } \lambda_k \downarrow 0,$$

which immediately proves $G_{\lambda_k}(x) \uparrow G(x)$.

(ii) The sequence $\{G_{\lambda_k}\}$ is pointwise nondecreasing and each $G_{\lambda_k}$ is lsc, being the sum of a smooth function, upper-$C^2$ functions, and the indicator function of the closed set $C$. The monotone epi-limit equals the pointwise supremum of the lsc functions [30, Proposition 7.4(d)], namely, $G_{\lambda_k} \underset{e}{\to} G$.

(iii) We first remark that the domains of $G_{\lambda_k}$ are the compact set $C$ for all $k$ and hence are uniformly bounded. This ensures that all $\{G_{\lambda_k}\}$ are eventually level-bounded [30, Exercise 7.32(a)]. Using epi-convergence and eventual level-boundedness, [30, Theorem 7.33] implies that $\inf G_{\lambda_k} \to \inf G$, which proves (iii).

(iv) Since C is compact and $G_{\lambda_k}$ and $G$ are lsc with effective domain C, the existence of minimizers for all regularized problems and for the original problem follows from the Weierstrass theorem. Under Assumption 2.5, $\{x(\lambda_k)\}$ is bounded and we consider any cluster point $x^\star$ of it. Theorem 7.33 of [30] also implies that $x^\star$ is a minimizer of $G$. If $\operatorname{argmin} G$ is a singleton, then every convergent subsequence has $x^\star$ as the cluster point, hence $\{x(\lambda_k)\} \to x^\star$. ∎

Theorem 3.1 can be viewed as a continuation consistency result, namely, *global* minimizers of the regularized problems converge (along subsequences) to *global* minimizers of the original problem as $\lambda \downarrow 0$. A stronger result holds when the Moreau envelope is over the whole objective, *i.e.*, one uses $e_\lambda(f + \sum_{i\in\mathcal{K}} r_i)$ instead of regularizing each term $r_i$. Minimizers of the whole-objective regularized problem coincide with minimizers of $f + \sum_{i\in\mathcal{K}} r_i$ for $\lambda > 0$ [30, Example 1.46] under the assumption of prox-boundedness. Furthermore, the local minimizers are also preserved under similar assumptions as recently proved in [19].

*Remark* 3.2. In our setup, with the Moreau envelope distributed inside the minimization, such results involving local minimizers do not necessarily hold in the nonconvex case as illustrated by the following counterexample. Take $f(x) = x^3 + x^4 - \frac{1}{2}x^2$, $r_1(x) = \frac{1}{2}x^2$, and $\mathcal{K} = \{1\}$. Using that $e_\lambda r_1(x) = \frac{x^2}{2(1+\lambda)}$, we see that $F_\lambda$ has three stationary points: a global minimizer $x_1(\lambda) = \frac{1}{8}\left(-3 - \sqrt{\frac{9+25\lambda}{1+\lambda}}\right)$, a local minimizer $x_2(\lambda) = \frac{1}{8}\left(-3 + \sqrt{\frac{9+25\lambda}{1+\lambda}}\right)$, and a local maximizer 0, for any $\lambda > 0$. As $\lambda \downarrow 0$, $x_1(\lambda)$ converges to $x^\star = -\frac{3}{4}$, the global minimizer of $F$, consistent with Theorem 3.1; however, the local minimizer $x_2(\lambda)$ collapses to zero, which is a stationary point of $F$.

**4. Proposed algorithm.** In this section, we start by describing the continuation algorithm, which is summarized in listing Algorithm 4.1. Then, in Sections 4.1 and 4.2, we analyze its convergence properties for the fixed-$\lambda$ problems (1.4)–(1.5) and of the overall continuation scheme, respectively.

The continuation strategy consists of using a sequence of regularization parameters $\lambda_s \downarrow 0$ and iteratively solving fixed-$\lambda$ problems (1.4)–(1.5) to approximate stationarity. In what follows, for each $\lambda_s$ the iterations of the algorithm are denoted by $\tilde{x}^k$, $k = \{0, 1, 2, \ldots\}$ and referred to as inner iterations. For better readability, the iterates corresponding to each $\lambda_s$ are denoted as $x^s$ and referred to as outer iterates.

Consider at some fixed $\lambda_s > 0$ the inner iterate $\tilde{x}^k$. A new inner iteration $\tilde{x}^{k+1}$ is obtained by means of an inexact stationary point $\bar{x}^{k+1}$ of a NLP model approximation of (1.4)–(1.5) in the form

$$\min_{x \in \mathbb{R}^n} M(x; \tilde{x}^k, g^k; \lambda_s, \alpha_k) \;:=\; f(x) + r_{\lambda_s}(\tilde{x}^k) + (g^k)^\top (x - \tilde{x}^k) + \alpha_k \|x - \tilde{x}^k\|^2 \tag{4.1}$$

$$\text{s.t.} \quad h(x) = 0,\; x_\ell \le x \le x_u. \tag{4.2}$$

Similar to the SQP approach of Wang and Petra [33] we use a convex quadratic model for the nonsmooth objective that uses first-order derivative information $g^k$ defined below in (4.4) and a quadratic penalty term for step/update stabilization controlled by the scalar $\alpha_k > 0$. At variance with the SQP approach, we enforce the smooth constraints (4.2) exactly. This is done for practical considerations since it allows using off-the-shelf NLP solvers to compute the new iterate, but it also simplifies the analysis. We believe that the SQP approach [33] can be revisited and used to replace the NLP approach present here at the cost of increased complexity in the analysis and implementation since it requires using feasibility restoration to deal with potentially inconsistent SQP linearizations of the constraints. In contrast, the sequential NLP approach here leaves this complication to the underlying NLP solver.

We do not require that the candidate $\bar{x}^{k+1}$ be a local or global minimizer of the NLPs (4.1) – (4.2) since this may be unrealistic. Instead, we only require a stationary point $\tilde{x}^{k+1}$ of the NLP model that satisfies a weak descent condition

$$M(\bar{x}^{k+1}; \tilde{x}^k, g^k; \lambda_s, \alpha_k) \le M(\tilde{x}^k; \tilde{x}^k, g^k; \lambda_s, \alpha_k). \tag{4.3}$$

We remark that there is always a path in $C$ along which the descent condition holds, unless $\tilde{x}^k$ is a stationary point. Indeed, if $\tilde{x}^k$ is not a critical point of the NLP model, then there must exist a direction $d$ in the tangent cone of $C$ to $\tilde{x}^k$ such that $d^\top \nabla M(\tilde{x}^k) < 0$; otherwise, if every tangent direction $d$ satisfies $d^\top \nabla M(\tilde{x}^k) \ge 0$, we would have $-\nabla M(\tilde{x}^k) \in N_C(\tilde{x}^k)$, meaning $\tilde{x}^k$ is stationary. Moreover, using definition of tangents, there is a path $x(\beta) = \tilde{x}^k + \beta d + o(\beta) \in C$, $\beta \in [0, \bar{\beta}]$ such that $M(x(\beta); \tilde{x}^k, g_k; \lambda, \alpha) = M(\tilde{x}^k; \tilde{x}^k, g_k; \lambda, \alpha) + \beta d^\top \nabla M(\tilde{x}^k) + o(\beta) < M(\tilde{x}^k)$ for all $\beta \in [0, \bar{\beta}]$ and small enough $\bar{\beta} > 0$. Therefore, the NLP model has a feasible path of descent and the descent condition (4.3) is a mild requirement for the underlying NLP solver and less intrusive than requiring a minimizer.

Above, $g^k$ encapsulates "derivative" information, namely,

$$g^k = \sum_{i \in \mathcal{K}} g_i^k \text{ with } g_i^k = \frac{1}{\lambda_s}(\tilde{x}^k - w_i^k) \text{ and } w_i^k \in \operatorname{prox}_{\lambda_s r_i}(\tilde{x}^k). \tag{4.4}$$

We use slightly less regular and more general subgradients than Wang and Petra [33], namely, we rely on proximal oracle subgradients $g_i^k$, which are limiting subgradients

at "nearby" points (see Lemma 2.1) and are defined for discontinuous functions, while the mentioned previous work employs Clarke subgradients. The proximal oracle subgradients used in $g^k$ above are a subset of the Clarke subgradients of the Moreau envelopes in general, see Theorem 4.6 and its proof of Clarke stationarity. The choice of proximal oracle subgradients $g_i^k$s (as opposed to choosing $g_i^k$ more restrictively, $g_i^k \in \partial e_{\lambda_s} r_i(\tilde{x}^k)$) stems from the representations of subdifferentials of Moreau envelopes of general functions satisfying Assumption 2.4, as discussed after Lemma 2.1.

We use an accept-reject rule based on a majorization-ratio test that updates the penalty parameter $\alpha^k > 0$ in the NLP model solved at each iteration. Upon solving the NLP model (4.1)–(4.2), we define the predicted and actual reductions at the candidate $\bar{x}^{k+1}$ as

$$\begin{aligned} &\mathrm{Pred}_k(\bar{x}^{k+1}) = M(\tilde{x}^k; \tilde{x}^k, g^k; \lambda, \alpha_k) - M(\bar{x}^{k+1}; \tilde{x}^k, g^k; \lambda, \alpha_k) \text{ and} \\ &\mathrm{Ared}_k(\bar{x}^{k+1}) = F_\lambda(\tilde{x}^k) - F_\lambda(\bar{x}^{k+1}). \end{aligned} \tag{4.5}$$

We define the ratio

$$\rho_k = \frac{\mathrm{Ared}_k(\bar{x}^{k+1})}{\mathrm{Pred}_k(\bar{x}^{k+1}) + \sigma\|\bar{x}^{k+1} - \tilde{x}^k\|^2}.$$

This is analogous to trust-region acceptance, but specialized to our exact constraints model. The iterate $\tilde{x}^{k+1} = \bar{x}^{k+1}$ is accepted whenever $\rho_k \geq \bar{\rho}$ with $\bar{\rho} \in (0, 1)$; otherwise, the step is rejected and $\alpha_k$ increases by a factor of two.

At each outer iteration $s$, the inner iterates $\tilde{x}^k$ of Algorithm 4.1 (steps 5 – 20) aim to reach an approximate stationary point for (1.4)–(1.5). Inexactness is controlled by the tolerance $\varepsilon_s$ and has the goal to balance inner solve accuracy with the continuation error. The stopping criterion of the inner iterates use the lagged stationarity measure

$$\hat{R}(\tilde{x}^k) := \nabla f(\tilde{x}^k) + g^{k-1} + \nabla h(\tilde{x}^k)^\top \nu^k - \gamma^k + \theta^k,$$

where $\nu^k \in \mathbb{R}^m$, $\gamma^k \geq 0$, and $\theta^k \geq 0$ are the KKT multipliers corresponding to the optimality conditions (see (4.11)–(4.13) for the exact form) of the model NLPs (4.1) – (4.2) under the general MFCQ for $C$. From these KKT conditions, we remark that $\hat{R}(\tilde{x}^k) = -2\alpha_{k-1}(\tilde{x}^k - \tilde{x}^{k-1})$, which makes this lagged stationarity both computationally cheap and practical (since it does not require dual information from the numerical solver used for the master problem). Therefore, the inner loop stopping criterion is

$$\|\hat{R}(\tilde{x}^k)\|_\infty = 2\alpha_{k-1}\|\tilde{x}^k - \tilde{x}^{k-1}\|_\infty \leq \varepsilon_s. \tag{4.6}$$

Later, in Proposition 4.5 we show that the inner loop reaches (4.6) for $\varepsilon_s > 0$ in a finite number of iterations. Theorem 4.6 analyzes the infinite inner loop (without the stopping criterion) and shows that the inner iterates $\tilde{x}^k$ are bounded and any of their accumulation or cluster points $x^{\star\lambda}$ are stationary points of (1.4)–(1.5).

Algorithm 4.1 repeats the sequential NLP steps discussed above for a given sequence $\lambda_s \downarrow 0$. It is shown in Theorem 4.11 that under mild conditions the outer iterates $x^s$ of Algorithm 4.1 are bounded and any of their accumulation points $x^\star$ satisfy stationarity conditions

$$0 \in \nabla f(x^\star) + \sum_{i \in \mathcal{K}} \partial r_i(x^\star) + \nabla h(x^\star)^\top \nu^\star - \gamma^\star + \theta^\star, \tag{4.7}$$

$$h(x^\star) = 0, \quad 0 \leq \gamma^\star \perp (x^\star - \ell) \geq 0, \quad 0 \leq \theta^\star \perp (u - x^\star) \geq 0, \tag{4.8}$$

for some multipliers $\nu^\star \in \mathbb{R}^m$, $\gamma^\star \geq 0$, and $\theta^\star \geq 0$ under MFCQ for the feasible set $C$. The limiting stationarity condition (4.7) targeted here is analogous in form to standard

first-order necessary optimality conditions, see, for example, [30]. Consequently, we use the following basic CQ (BCQ) specific to non-Lipschitz functions, counterpart of the horizon BCQ from [30, Theorem 8.15].

**BCQ for non-Lipschitz objectives.** If there exist $v_i \in \partial^\infty r_i(x^\star)$, $i \in \mathcal{K}$, such that $-\sum_{i\in\mathcal{K}} v_i \in N_C(x^\star)$, then $v_i = 0$ for all $i \in \mathcal{K}$. Under MFCQ, the BCQ is equivalent to: if there exist $v_i \in \partial^\infty r_i(x^\star)$, $i \in \mathcal{K}$, and $\nu \in \mathbb{R}^m$, $\gamma \geq 0$, and $\theta \geq 0$ satisfying complementarity conditions from (4.8) such that $\sum_{i\in\mathcal{K}} v_i + \nabla h(x^\star)^\top \nu - \gamma + \theta = 0$, then $v_i = 0$ for all $i \in \mathcal{K}$.

**Algorithm 4.1** Proximal oracle sequential NLP continuation algorithm

Initial point $x^0 \in C$ and $\lambda_0$ below the prox-thresholds of all $r_i$; parameters $\bar\rho \in (0, 1]$, $\sigma > 0$, and $\underline{\alpha} > 0$; sequences of tolerances $\varepsilon_s \downarrow 0$ and regularizations $\lambda_s \downarrow 0$.
$s \leftarrow 0$
**while** true **do**
  Inner loop initializations: $k \leftarrow 0$, $\tilde x^0 \leftarrow x^s$, $\alpha_k \leftarrow \max\{\underline{\alpha}, \frac{|K|}{2\lambda_s}\}$
  **while** true **do**
    **for** $i \in \mathcal{K}$ **do**
      Compute $w_i^k \in \operatorname{prox}_{\lambda_s} r_i(\tilde x^k)$ and $g_i^k = (\tilde x^k - w_i^k)/\lambda_s$
    **end for**
    $g^k \leftarrow \sum_{i\in\mathcal{K}} g_i^k$
    Compute a stationary point $\bar x^{k+1}$ of $M(\cdot; \tilde x^k, g^k; \lambda_s, \alpha_k)$ from (4.1)–(4.2) satisfying (4.3)
    Compute $\rho_k(\bar x^{k+1}) = \mathrm{Ared}_k(\bar x^{k+1})/(\mathrm{Pred}_k(\bar x^{k+1}) + \sigma\|\bar x^{k+1} - \tilde x^k\|^2)$ using (4.5) if $\bar x^{k+1} \neq \tilde x^k$, otherwise set $\rho_k(\bar x^{k+1}) = 1$
    **if** $\rho_k(\bar x^{k+1}) \geq \bar\rho$ **then**
      $\tilde x^{k+1} \leftarrow \bar x^{k+1}$, $\alpha_{k+1} \leftarrow \alpha_k$, $k \leftarrow k + 1$ {*accept iterate*}
    **else**
      $\tilde x^{k+1} \leftarrow \tilde x^k$, $\alpha_{k+1} \leftarrow 2\alpha_k$ {*reject iterate*}
      $g^{k+1} \leftarrow g^k$, $g_i^{k+1} \leftarrow g_i^k$, $w_i^{k+1} \leftarrow w_i^k$, $i \in \mathcal{K}$, $k \leftarrow k + 1$
      Return to step 10
    **end if**
    **if** $\|\hat R(\tilde x^k)\|_\infty \leq \varepsilon_s$ **then**
      **break** {*stopping criterion for inner loop*}
    **end if**
  **end while**
  $s \leftarrow s + 1$, $x^s \leftarrow \tilde x^k$, $g_i^s \leftarrow g_i^{k-1}$, $w_i^s \leftarrow w_i^{k-1}$ for all $i \in \mathcal{K}$
**end while**

BCQ rules out abnormal stationarity conditions and is a prerequisite for necessary first-order optimality conditions (4.7)–(4.8) to hold for constrained problems with nonsmooth and nonconvex objectives. In this respect, it can be viewed as the counterpart of KKT conditions, which rule out Fritz-John points with zero objective multipliers, in smooth constrained optimization. Geometrically, near $x^\star$, limiting subgradients can become arbitrarily large, and their normalized limiting directions belong to the horizon subgradients $\partial^\infty r_i(x^\star)$ (for some $i \in \mathcal{K}$). If one such direction exists and is opposite to the normal cone of $C$ at some boundary point, or, equivalently, cancels out the multiplier terms in the BCQ above, then the constraint boundary can absorb this unbounded tilt. In such cases general first-order stationarity conditions such as [30, Theorem 8.15] may fail at minimizers as shown in Examples 4.1

and 4.2. The BCQ excludes exactly this pathology. We remark that the BCQ holds automatically for locally Lipschitz functions $r_i$, since $\partial^\infty r_i(x^\star) = \{0\}$.

*Example* 4.1. Consider $\min_x\{r_1(x) := x^{1/3} \mid x \geq 0\}$ with the (global) minimizer $x^\star = 0$. This problem fails BCQ above since there exists $v = \frac{1}{3} \in \partial^\infty r_1(0)$ such that $v - \gamma = 0$, where $0 \leq x \perp \gamma \geq 0$. We remark that $\partial r_1(0) = \emptyset$ and therefore the necessary first-order stationary conditions (4.7)–(4.8) for $f = 0$ or ones in [30][Theorem 8.15]) do not hold at this minimizer.

*Example* 4.2. The following example shows that (global) minimizers fail to satisfy first-order stationarity when the BCQ does not hold even when the subdifferential is nonempty and the feasible set satisfies MFCQ. Consider

$$\min_{x\in\mathbb{R}^2}\{r_1(x_1, x_2) := x_1^2 + x_1 + \sqrt{\max(x_2, 0)} \mid x \in C\},$$

with $C := \{x \in \mathbb{R}^2 : \ h(x) := x_1^2 - x_2 \leq 0\}$. Since $r_1(x_1, x_2) = x_1^2 + x_1 + \sqrt{x_2} \geq x_1^2 + x_1 + |x_1|$ over $x = (x_1, x_2) \in C$, we see that $x^\star = (0, 0)$ is a strict local/global minimizer over $C$. We remark that MFCQ holds at $x^\star$ and that $\partial r_1(0, 0) = \hat\partial r(0, 0) = \{1\} \times \mathbb{R}_+$.

BCQ fails at $x^*$ since $0 \neq v = (0, 1) \in \partial^\infty r(0, 0)$ satisfies $v + \nu(0, -1) = 0$ with $0 \leq \nu \perp -h(0, 0) \geq 0$ (for $\nu = 1$). Here we remark that $\nu$ is signed since $h(x)$ is involved in an inequality constraint and, also, the bound multipliers $\gamma$ and $\theta$ are not present. To see that $(0, 1) \in \partial^\infty r_1(0, 0)$, take $x_\tau = (0, \tau)$ with $\tau \downarrow 0$. As $x_\tau \to (0, 0)$, we have $r_1(x_\tau) = \sqrt{\tau} \to 0$ and $v^\tau := \nabla r(0, \tau) = (1, \frac{1}{2\sqrt{\tau}}) \in \hat\partial r_1(x_\tau)$. By choosing $t_\tau = 2\sqrt{\tau} \to 0$ we obtain $t_\tau v^\tau = (2\sqrt{\tau}, 1) \to (0, 1)$. Then $(0, 1) \in \partial^\infty r_1(0, 0)$ by definition of horizon subdifferential from Section 2. In fact, one can prove that $\partial^\infty r_1(0, 0) = \{0\} \times \mathbb{R}_+$.

Even though $(0, 0)$ is a minimizer the stationarity conditions do not hold since

$$0 \notin \partial r_1(0, 0) + \nu(0, -1) = \{1\} \times \mathbb{R}_+ + \nu(0, -1)$$

for any $\nu$ such that $0 \leq \nu \perp h(0, 0) \geq 0$.

We use the concept of sum-stationarity using $\sum_{i\in\mathcal{K}} \partial r_i$, which comes naturally in our composite objective, instead of $\partial(\sum_{i\in\mathcal{K}} r_i)$ commonly used with weakly convex summands. To the best of our knowledge, no inclusion is guaranteed between the two sets in our general setup without additional qualifications. We show in Lemma 4.12 that $\sum_{i\in\mathcal{K}} \partial r_i \subseteq \partial(\sum_{i\in\mathcal{K}} r_i)$ for subdifferentially regular functions, which cover weakly convex functions, hence the stationarity conditions used in this work characterize the typical limiting stationarity for this class of problems.

**4.1. Inner-loop convergence.** We start with a majorization lemma specific to upper-$C^2$ functions [33]. Here we provide a stronger version for Moreau envelopes with an explicit bound in terms of the regularization $\lambda$ and proximal oracle subgradients.

LEMMA 4.3. *For any $\lambda > 0$ below the prox-thresholds of $r_i(\cdot)$ and a given $\bar x$, the following majorization bound holds: for any $\bar g_i \in \frac{1}{\lambda}(\bar x - \operatorname{prox}_{\lambda r_i}(\bar x))$, $i \in \mathcal{K}$,*

$$r_\lambda(x) \leq r_\lambda(\bar x) + \langle \bar g, x - \bar x\rangle + Q_\lambda\|x - \bar x\|^2, \ \forall x, \ \text{where } \bar g = \sum_{i\in\mathcal{K}} \bar g_i \text{ and } Q_\lambda = \frac{|\mathcal{K}|}{2\lambda}.$$

*Proof.* Fix $i \in \mathcal{K}$ and consider $\bar g_i \in \frac{1}{\lambda}(\bar x - \operatorname{prox}_{\lambda r_i}(\bar x))$, namely, there exists $w_i \in \operatorname{prox}_{\lambda r_i}(\bar x)$ such that $\bar g_i = \frac{1}{\lambda}(\bar x - w_i)$. The optimality of $w_i$ with respect to the minimization problem defining the Moreau envelope implies $e_{\lambda_i} r_i(x) \leq r_i(w_i) + \frac{1}{2\lambda}\|x -$

$w_i||^2$ for all $x$. Also, for $x = \bar{x}$ we have $e_{\lambda_i} r_i(\bar{x}) = r_i(w_i) + \frac{1}{2\lambda}||\bar{x} - w_i||^2$. By subtracting these two conditions, we obtain after some manipulations of the quadratic terms that

$$\begin{aligned} e_{\lambda_i} r_i(x) &\leq e_{\lambda_i} r_i(\bar{x}) + \frac{1}{2\lambda}\langle x - \bar{x} + \bar{x} - w_i, x - \bar{x} + \bar{x} - w_i\rangle - \frac{1}{2\lambda}||\bar{x} - w_i||^2 \\ &= e_{\lambda_i} r_i(\bar{x}) + \langle \frac{1}{\lambda}(\bar{x} - w_i), x - \bar{x}\rangle + \frac{1}{2\lambda}\|x - \bar{x}\|^2. \end{aligned}$$

Finally, summing over $i \in \mathcal{K}$ gives the desired majorization bound. ∎

The above lemma is the key to proving the convergence of the proposed algorithm. First, we use it to prove that the candidate iterates $\bar{x}^{k+1}$ are eventually accepted.

LEMMA 4.4. *Fix an outer iteration $s$. Then only finitely many inner steps can be rejected in Algorithm 4.1. Consequently, the penalty parameters satisfy $\alpha_k \in [\underline{\alpha}, \bar{\alpha}_s]$, with $\bar{\alpha}_s = \max\{\underline{\alpha}, 2(Q_\lambda + \bar{\rho}\sigma)\}$, for all $k$ greater than some iteration threshold index.*

*Proof.* We first show that whenever $\alpha_k$ becomes larger than $Q_\lambda + \bar{\rho}\sigma$, all subsequent iterations are accepted. Assume that at iterate $k$, the trial point $\bar{x}^{k+1}$ is computed with $\alpha_k > Q_\lambda + \bar{\rho}\sigma$. We use the majorization of Lemma (4.3) to write

$$\begin{aligned} F_\lambda(\bar{x}^{k+1}) &= f(\bar{x}^{k+1}) + r_\lambda(\bar{x}^{k+1}) \leq f(\bar{x}^{k+1}) + r_\lambda(\tilde{x}^k) + \langle g^k, \bar{x}^{k+1} - \tilde{x}^k\rangle + \\ &\quad Q_\lambda\|\bar{x}^{k+1} - \tilde{x}^k)\|^2 = M_k(\bar{x}^{k+1}; \tilde{x}^k, g^k; \lambda, \alpha_k) + (Q_\lambda - \alpha_k)\|\bar{x}^{k+1} - \tilde{x}^k)\|^2. \end{aligned}$$

By subtracting the above from $F_\lambda(\tilde{x}^k) = M_k(\tilde{x}^k; \tilde{x}^k, g^k; \lambda, \alpha_k)$ and using our working assumption that $\alpha_k > Q_\lambda + \bar{\rho}\sigma$, we obtain

$$\begin{aligned} \text{Ared}_k(\bar{x}^{k+1}) &\geq \text{Pred}_k(\bar{x}^{k+1}) + (\alpha_k - Q_\lambda)\|\bar{x}^{k+1} - \tilde{x}^k)\|^2 \\ &> \text{Pred}_k(\bar{x}^{k+1}) + \bar{\rho}\sigma\|\bar{x}^{k+1} - \tilde{x}^k)\|^2 \geq \bar{\rho}(\text{Pred}_k(\bar{x}^{k+1}) + \sigma\|\bar{x}^{k+1} - \tilde{x}^k)\|^2), \end{aligned}$$

with the last inequality due to the algorithmic parameter $\bar{\rho} \leq 1$ and $\text{Pred}_k(\bar{x}^{k+1}) \geq 0$. Therefore, $\rho_k(\bar{x}^{k+1}) \geq \bar{\rho}$ and $\bar{x}^{k+1}$ will be accepted in the acceptance test from step 13 of Algorithm 4.1, namely $\tilde{x}^{k+1} = \bar{x}^{k+1}$, and, also, $\alpha_{k+1} = \alpha_k$. Proceeding by induction through the next iterations, it should be apparent that all iterations $k + j$ with $j \geq 1$ of Algorithm 4.1 accept $\bar{x}^{k+j}$ and $\alpha_{k+j} = \alpha_k$ for all $j \geq 1$.

We remark that the inner loop of Algorithm 4.1 may satisfy the termination criteria in step 20 and exit after finite number of iterations without increasing the penalty parameters $\alpha_k$ above the $Q_\lambda + \bar{\rho}\sigma$ threshold. In this case, it should be apparent that $\alpha_k \in [\underline{\alpha}, Q_\lambda + \bar{\rho}\sigma]$ for all $k$. If this is not the case, let $k_0$ denote the smallest iteration index such that $\alpha_{k_0} > Q_\lambda + \bar{\rho}\sigma$. This index must be finite because of the geometric increase of penalties $\alpha_k$ for the rejected iterates.

If $k_0 = 0$, then $\alpha_{k_0} = \max\{\underline{\alpha}, Q_\lambda\} \in [\underline{\alpha}, \max\{\underline{\alpha}, Q_\lambda\}]$ due to initialization. If $k_0 \geq 1$, then we immediately see that the previous step must have been rejected, therefore, $\alpha_{k_0 - 1} = \frac{1}{2}\alpha_{k_0} < Q_\lambda + \bar{\rho}\sigma$, which implies that $\alpha_{k_0} < 2(Q_\lambda + \bar{\rho}\sigma)$. Finally, we remark that $\alpha_{k_0} = \alpha_{k_0 + j} = \ldots < 2(Q_\lambda + \bar{\rho}\sigma)$ for all $j \geq 1$, as per the first part of the proof.

The expression for $\bar{\alpha}_s$ follows after collecting the lower and upper bounds for $\alpha_k$ over the above three cases. ∎

PROPOSITION 4.5. *At any outer iteration $s$, the inner iterates $\{\tilde{x}^k\}$ of Algorithm 4.1 are bounded and*

*(i) $\{F_\lambda(\tilde{x}^k\}$ is decreasing monotonically:*

$$F_\lambda(\tilde{x}^{k+1}) - F_\lambda(\tilde{x}^k) \;\leq\; -\bar{\rho}\sigma\|\tilde{x}^{k+1} - \tilde{x}^k\|^2, \forall k \geq 0;$$

*(ii) the updates $\Delta^k = \tilde{x}^{k+1} - \tilde{x}^k$ are square-summable, that is, $\sum_k \|\Delta_k\|^2 < \infty$, and, thus, $\Delta_k \to 0$;*

*(iii) for any tolerance $\varepsilon_s > 0$, the inner-loop stopping criterion (4.6) is reached in a finite number of inner iterations $k_s$.*

*Proof.* We first remark that the sequence $\{\tilde{x}^k\}$ is bounded since it is generated by the algorithm in the feasible set $C$, which is compact by the Assumption 2.5.

(i) By Lemma 4.4 all iterates $k$ greater than some index $k_0$ are accepted in step 13 of the algorithm, hence

$$F_\lambda(\tilde{x}^k) - F_\lambda(\tilde{x}^{k+1}) \geq \bar{\rho}\left(\mathrm{Pred}_k(\tilde{x}^{k+1}) + \sigma\|\tilde{x}^{k+1} - \tilde{x}^k\|^2\right) \geq \bar{\rho}\,\sigma\|\tilde{x}^{k+1} - \tilde{x}^k\|^2.$$

The iterates 0 to $k_0$ also satisfy the monotonicity: the accepted iterates in the form above, while the rejected iterates trivially, since $\tilde{x}^{k+1} = \tilde{x}^k$.

(ii) Let us sum over $k \in [k_0, \bar{k}]$ the inequality from (i):

$$\bar{\rho}\sigma \sum_{k=k_0}^{\bar{k}} \|\Delta_k\|^2 \leq F_\lambda(\tilde{x}^{k_0}) - F_\lambda(\tilde{x}^{\bar{k}}).$$

Since $f(\cdot)$ is $C^2$ and the Moreau envelopes defining $r_\lambda(\cdot)$ are upper-$C^2$, hence Lipschitz continuous, the function $F_\lambda(\cdot)$ is continuous. Therefore, boundedness of $\{\tilde{x}^k\}$ implies that $\{F_\lambda(\tilde{x}^{\bar{k}})\}$ is bounded. Taking the limit $\bar{k} \to \infty$ we obtain that $\bar{\rho}\sum_k \|\Delta_k\|^2 < \infty$. This completes the proof of (ii) since $\bar{\rho}\sigma > 0$.

(iii) First observe that $\|\hat{R}(\tilde{x}^k)\| = 2\alpha_k\|\tilde{x}^k - \tilde{x}^{k-1}\| \leq 2\bar{\alpha}_s\|\Delta_{k-1}\|$ by Lemma 4.4. Using $\Delta_k \to 0$ from (ii), stopping criterion (4.6) is going to be met for sufficiently large $k$, so the inner loop terminates after finitely many iterations. ∎

The convergence behavior of the inner loop (for fixed $\lambda$) of Algorithm 4.1 fits under a standard subsequence convergence and only requires a standard CQ on the feasible set of the target problem. We assume MFCQ.

THEOREM 4.6. *Assume MFCQ holds for the feasible set $C$, in addition to Assumptions 2.4 and 2.5. Fix an outer iteration $s$ and consider the sequence of inner iterates $\{\tilde{x}^k\}$ generated by the inner loop of of Algorithm 4.1 without the stopping test in step 20, i.e., the infinite loop consisting of steps 5–17. Then $\{\tilde{x}^k\}$ is bounded and all its cluster points $x^{\star\lambda}$ satisfy the proximal limiting stationarity conditions*

$$0 = \nabla f(x^{\star\lambda}) + \sum_{i\in\mathcal{K}} g_i^{\star\lambda} + \nabla h(x^{\star\lambda})^\top \nu^{\star\lambda} - \gamma^{\star\lambda} + \theta^{\star\lambda}, \tag{4.9}$$

$$h(x^{\star\lambda}) = 0, \quad 0 \leq \gamma^{\star\lambda} \perp (x^{\star\lambda} - \ell) \geq 0, \quad 0 \leq \theta^{\star\lambda} \perp (u - x^{\star\lambda}) \geq 0. \tag{4.10}$$

*for some multipliers $\nu^{\star\lambda} \in \mathbb{R}^m$, $\gamma^{\star\lambda} \geq 0$, and $\theta^{\star\lambda} \geq 0$ and with $g_i^{\star\lambda} = \frac{1}{\lambda}(x^{\star\lambda} - w_i^{\star\lambda}) \in \partial r_i(w_i^{\star\lambda})$, where $w_i^{\star\lambda} \in \mathrm{prox}_\lambda r_i(x^{\star\lambda})$.*

*Moreover, $x^{\star\lambda}$ also satisfies Clarke stationarity conditions*

$$0 \in \nabla f(x^{\star\lambda}) + \partial^C r_\lambda(x^{\star\lambda}) + \nabla h(x^{\star\lambda})^\top \nu^{\star\lambda} - \gamma^{\star\lambda} + \theta^{\star\lambda},$$
$$h(x^{\star\lambda}) = 0, \quad 0 \leq \gamma^{\star\lambda} \perp (x^{\star\lambda} - \ell) \geq 0, \quad 0 \leq \theta^{\star\lambda} \perp (u - x^{\star\lambda}) \geq 0.$$

*for some multipliers $\nu^{\star\lambda}$, $\gamma^{\star\lambda} \geq 0$, and $\theta^{\star\lambda} \geq 0$.*

*Proof.* The boundedness of $\{\tilde{x}^k\}$ from Proposition 4.5 shows that cluster points of $\{\tilde{x}^k\}$ exist. Let $x^{*\lambda}$ be an arbitrary such point and, without re-indexing, let $\tilde{x}^k \to x^{*\lambda}$.

By Lemma 4.4, $\tilde{x}^{k+1}$ becomes an accepted iterate for large enough $k$ and therefore satisfies the KKT conditions for the sequential NLP model (4.1)-(4.2), namely, there exist multipliers $\nu^{k+1}$, $\gamma^{k+1}$, and $\theta^{k+1}$ such that

$$\nabla f(\tilde{x}^{k+1}) + g^k + 2\alpha_k \Delta^k + \nabla h(\tilde{x}^{k+1})^\top \nu^{k+1} - \gamma^{k+1} + \theta^{k+1} = 0, \tag{4.11}$$

$$h(\tilde{x}^{k+1}) = 0, \tag{4.12}$$

$$0 \le \gamma^{k+1} \perp (\tilde{x}^{k+1} - \ell) \ge 0,\ 0 \le \theta^{k+1} \perp (u - \tilde{x}^{k+1}) \ge 0. \tag{4.13}$$

The main steps of the proof of limiting stationarity equations (4.9)–(4.10) are:

(a) show the sequence $\{g^k\}_k$ is bounded and its cluster points $g^{\lambda\star}$ satisfy $g^{\lambda\star} = \sum_{i\in\mathcal{K}} g_i^{\lambda\star}$ with $g_i^{\lambda\star} = \frac{1}{\lambda}(x^{\lambda\star} - w_i^{\lambda\star}) \in \partial r_i(w_i^{\lambda\star})$ and $w_i^{\lambda\star} \in \mathrm{prox}_{\lambda r_i}(x^{\lambda\star})$, for all $i \in \mathcal{K}$.

(b) show the sequences $\{\nu^{k+1}\}_k$, $\{\gamma^{k+1}\}_k$, and $\{\theta^{k+1}\}_k$ of multipliers are bounded as well;

(c) use (a) and (b) to take the limit $k \to \infty$ in the above KKT system to obtain the conclusion.

To prove (a), we first look at each $g_i^k$ term from the expression of $g^k$ in steps 7 and 9 of Algorithm 4.1 (see also (4.4)). We use a well-known proximal behavior, namely, if $\tilde{x}_k \to x^{\lambda\star}$ and $w_i^k \in \mathrm{prox}_{\lambda r_i}(\tilde{x}^k)$, then the sequence $\{w_i^k\}$ is bounded and its cluster points $w_i^{\lambda\star}$ lie in $\mathrm{prox}_{\lambda r_i}(x^{\lambda\star})$ [30, Proposition 1.25]. Hence the sequence of $g_i^k = \frac{1}{\lambda}(\tilde{x}^k - w_i^k)$ is bounded, and so is the sequence of their sum, which is $g^k$. Similarly, cluster points satisfy the relationship from (i), with the observation that the relationship $g_i^{\lambda\star} \in \partial r_i(w_i^{\lambda\star})$ follows from (i) of Lemma 2.1.

We prove (b) by contradiction, namely assume there exists an unbounded subsequence (not relabeled) such that $t_k := \|(\nu^k, \gamma^k, \theta^k)\| \to \infty$. Since $\|(\frac{\nu^k}{t_k}, \frac{\gamma^k}{t_k}, \frac{\theta^k}{t_k})\| = 1$ for all $k$, there must exist a cluster point and a subsequence (not relabeled) such that

$$\left(\frac{\nu^k}{t_k}, \frac{\gamma^k}{t_k}, \frac{\theta^k}{t_k}\right) \to (\hat{\nu}, \hat{\gamma}, \hat{\theta}) \text{ as } k \to \infty \text{ with } \|(\hat{\nu}, \hat{\gamma}, \hat{\theta})\| = 1. \tag{4.14}$$

Next consider (4.11) rescaled by $t_{k+1}$, namely

$$\frac{1}{t_{k+1}}\left(\nabla f(\tilde{x}^{k+1}) + g^k + 2\alpha_k \Delta^k\right) + \nabla h(\tilde{x}^{k+1})^\top \frac{\nu^{k+1}}{t_{k+1}} - \frac{\gamma^{k+1}}{t_{k+1}} + \frac{\theta^{k+1}}{t_{k+1}} = 0.$$

Recall that we work with a convergent (sub)sequence $\{\tilde{x}^{k+1}\} \to x^{\lambda *}$, which makes the term $\nabla f(\tilde{x}^{k+1}) + g^k + 2\alpha_k \Delta^k$ bounded/convergent due to $\nabla f$ being continuous (by Assumption 2.4), $\{g_k\}$ being bounded (by (i) above), $\Delta^k \to 0$ (by Proposition 4.5), and $\alpha_k \le \bar{\alpha}_s < \infty$ (by Lemma 4.4). Also $\nabla h(\tilde{x}^{k+1}) \to \nabla h(x^{\lambda\star})$ since $\nabla h$ is continuous per Assumption 2.4. We see that by also using (4.14), we can take the limit to obtain

$$\nabla h(x^{\lambda\star})^\top \hat{\nu} - \hat{\gamma} + \hat{\theta} = 0.$$

Similar continuity arguments can be applied to (4.13) to obtain that

$$0 \le \hat{\gamma} \perp (x^{\lambda\star} - \ell) \ge 0,\ 0 \le \hat{\theta} \perp (u - x^{\lambda\star}) \ge 0.$$

By MFCQ, the only multipliers $\hat{\nu}$, $\hat{\gamma}$, $\hat{\theta}$ satisfying the above relationships are the trivial ones, $(\hat{\nu}, \hat{\gamma}, \hat{\theta}) = (0, 0, 0)$, which contradicts (4.14). Therefore, $\{\nu^{k+1}\}_k$, $\{\gamma^{k+1}\}_k$, and $\{\theta^{k+1}\}_k$ are bounded.

We now prove (c). First consider any cluster point $(\nu^{\lambda\star}, \gamma^{\lambda\star}, \theta^{\lambda\star})$ of $(\nu^k, \gamma^k, \theta^k)$ and cluster point $g^{\lambda\star}$ of $g^k$. Since $\Delta_k \to 0$ by Proposition 4.5, $\alpha_k \leq \bar{\alpha}_s \leq \infty$ by Lemma 4.4, and $\nabla f$, $\nabla h$, and $h$ are continuous by Assumption 2.4, we take the limit over any converging subsequence in (4.11)-(4.13) to obtain (4.9)–(4.10).

The Clarke stationarity follows from (4.9)–(4.10) since, as we next show, $\sum_{i\in\mathcal{K}} g_i^{\star\lambda} \in \partial^C r_\lambda(x^{\star\lambda})$. A general characterization [18][Proposition 3.1 and Theorem 3.2] for the Clarke subdifferentials of Moreau envelopes $e_\lambda r_i(\cdot)$ is available via convex hulls involving proximal points, namely,

$$\partial^C e_\lambda r_i(x) = co\left\{\frac{1}{\lambda}(x-w) \,:\, w \in \operatorname{prox}_\lambda r_i(x)\right\}.$$

Therefore, $g_i^{\star\lambda}$ defined in (4.9)–(4.10) are Clarke subgradients of the Moreau envelopes,

$$g_i^{\star\lambda} \in \partial^C e_\lambda f_i(x^{\star\lambda}),\ \forall i \in \mathcal{K}.$$

By Lemma A.1, we have that $\sum_{i\in\mathcal{K}} g_i^{\star\lambda} \in \partial^C \left(\sum_{i\in\mathcal{K}} e_\lambda r_i(x^{\star\lambda})\right) = \partial^C r_\lambda(x^{\star\lambda})$, which concludes the proof. ∎

*Remark* 4.7. Clarke stationarity conditions (ii) of Theorem 4.6 are weaker than the (limiting) sum-stationarity from (i) since the Clarke subdifferential is the convex hull of the limiting subdifferential in our general setup. This is a major improvement over our previous algorithm for upper-$C^2$ [33], mainly possible because we exploit the additional structure that Moreau envelopes have over general upper-$C^2$ functions, namely, the relationships between proximal points (minimizers of the Moreau envelope) and limiting subdifferentials from Lemma 2.1. In this respect, (4.9)–(4.10) of Theorem 4.6 can be viewed as stationarity conditions involving limiting subgradients at "nearby" points. More important, they are instrumental in proving the (subsequence) convergence of the outer iterates in the limit of Moreau parameter $\lambda \downarrow 0$. Clarke stationarity conditions (ii) are too weak to obtain the same convergence result.

**4.2. Outer-loop convergence.** In this section we elucidate the convergence behavior of the outer loop of Algorithm 4.1. We start with a technical lemma showing that proximal points of convergent sequences are also convergent.

LEMMA 4.8. *Consider a sequence $x^s$ that converges to $x^\star$ and a sequence $\lambda_s \downarrow 0$. For any choice of proximal points $w_i^s \in \operatorname{prox}_{\lambda_s} r_i(x^s)$, the sequences $\{w_i^s\}$ converge to $x^\star$ for all $i \in \mathcal{K}$.*

*Proof.* The proof is common to all $i \in \mathcal{K}$, therefore let $r := r_i$ and $w_s := w_i^s$. Prox-boundedness of $r$ implies that there exists a fixed $\bar{\lambda} > 0$ such that $\bar{m} := e_{\bar{\lambda}}(x^\star) > -\infty$. We then use the definition (1.3) of the Moreau envelope to write $r(w^s) \geq \bar{m} - \frac{1}{2\bar{\lambda}}\|w^s - x^\star\|^2$. By the choice of $w^s \in \operatorname{prox}_{\lambda_s} r(x^s)$, for $\lambda_s < \bar{\lambda}/2$, we can write

$$r(w^s) + \frac{1}{2\lambda_s}\|w^s - x^s\|^2 \leq r(x^\star) + \frac{1}{2\lambda_s}\|x^s - x^\star\|^2. \tag{4.15}$$

The previous two inequalities imply that

$$\bar{m} - \frac{1}{2\bar{\lambda}}\|w^s - x^\star\|^2 + \frac{1}{2\lambda_s}\|w^s - x^s\|^2 \leq r(x^\star) + \frac{1}{2\lambda_s}\|x^s - x^\star\|^2.$$

Now use $\|w^s - x^s\|^2 = \|(w^s - x^\star) - (x^s - x^\star)\|^2 \geq \frac{1}{2}\|w^s - x^\star\|^2 - \|x^s - x^\star\|^2$ and bring the terms involving $\|x^s - x^\star\|^2$ to the left to obtain

$$\bar{m} + \left(\frac{1}{4\lambda_s} - \frac{1}{2\bar{\lambda}}\right)\|w^s - x^\star\|^2 \leq r(x^\star) + \frac{1}{\lambda_s}\|x^s - x^\star\|^2.$$

Consequently, $\|w^s - x^\star\|^2 \le \frac{4\bar\lambda}{\bar\lambda - 2\lambda_s}\left[\lambda_s(r(x^\star) - \bar m) + \|x^s - x^\star\|^2\right]$, which proves $w^s \to x^\star$. ∎

One would expect the function values $r_i(w_i^s) \to r_i(x^\star)$ as well; however, this is not necessarily the case when $r_i(\cdot)$ has a jump discontinuity at $x^\star$. For example, $r(x) = 1$ for $x \neq 0$ and $r(0) = 0$, the Moreau envelope is $e_\lambda(x) = \min\{1, \frac{|x|^2}{2\lambda}\}$; also, $\operatorname{prox}_{\lambda r}(x) = \{0\}$ for $|x| < \sqrt{2\lambda}$, $\operatorname{prox}_{\lambda r}(x) = \{0, x\}$ for $|x| = \sqrt{2\lambda}$, and $\operatorname{prox}_{\lambda r}(x) = \{x\}$ for $|x| > \sqrt{2\lambda}$. For simplicity, let us work with only one term, $\mathcal{K} = \{1\}$, and set $r_i(x) = r(x)$. Assume that Algorithm 4.1 generates $x^s = \sqrt{2\lambda_s}$ and that the proximal oracle always returns $w_i^s = \sqrt{2\lambda_s} \in \operatorname{prox}_{\lambda_s r}(x) = \{0, \sqrt{2\lambda_s}\}$. We observe that $r_i(w_i^s) \to 1 \neq r_i(0)$, even though $w_i^s \to x^\star = 0$. This pathology has previously been recognized in [5], where an attentive convergence condition is used (see their assumption (H3)) to prove convergence of their PPA framework, namely the condition below.

*Condition* 4.9. For convergent sequences $x^s \to x^\star$ and corresponding $w_i^s \in \operatorname{prox}_{\lambda_s r_i}(x^s)$, attentive convergence holds along $\{w_i^s\}$, namely $r_i(w_i^s) \to r_i(x^\star)$, for all $i \in \mathcal{K}$.

This condition is quite general and is satisfied, for example, when the functions $r_i(\cdot)$ are continuous. It should also be noted that it holds in the lsc case whenever the outer iterates $x^s$, as returned by the inner loop of Algorithm 4.1, are exact minimizers of the regularized problem (1.4)–(1.5) as $\lambda_s \downarrow 0$. The latter fact is proved in Lemma 4.15.

Given two tolerances $\varepsilon > 0$ and $\delta > 0$, a point $x \in C$ is $(\varepsilon, \delta)$-approximately sum-stationary if there exist points $w_i$, subgradients $g_i \in \partial r_i(w_i)$, and a normal vector $\eta \in N_C(x)$ such that

$$\|w_i - x\|_\infty \le \delta,\ \forall i \in \mathcal{K}, \quad \text{and} \quad \|\nabla f(x) + \sum_{i\in\mathcal{K}} g_i + \eta\|_\infty \le \varepsilon.$$

In words, $x \in C$ is $(\varepsilon, \delta)$-approximately sum-stationary if it is stationary up to a residual norm $\varepsilon$ with subgradients sampled at points no further than $\delta$ from $x$.

Our main and general convergence results are provided by the following proposition. In the following, let $k_s$ be the number of inner iterations at each outer iteration $s$ in Algorithm 4.1. We recall that $x^s = \tilde x^{k_s}$ according to our notation convention; also, let $x_-^s = \tilde x^{k_s - 1}$ and $\Delta^s = x^s - x_-^s$.

PROPOSITION 4.10. *Consider the iterates sequence $\{x^s\}$ generated by Algorithm 4.1 under a sequence of regularization parameters $\lambda_s \downarrow 0$ and inner-loop tolerances $\varepsilon_s \downarrow 0$. Assume that Condition 4.9 holds along subsequences $x^s \to x^\star$, that the conditions of Theorem 4.6 hold, that BCQ holds at cluster points $x^\star$, and that Assumptions 2.4 and 2.5 are satisfied. Then:*

*(i) Each outer-loop iterate $x^s$ satisfies*

$$\|\nabla f(x^s) + \sum_{i\in\mathcal{K}} g_i^s + \nabla h(x^s)^\top \nu^s - \gamma^s + \theta^s\|_\infty \le \varepsilon_s,$$

*together with the KKT feasibility $x^s \in C$ and KKT complementarity conditions. As a result $x^s$ is $(\varepsilon_s, \delta_s)$-approximately sum-stationary, with*

$$\delta_s := \max_{i\in\mathcal{K}} \|x^s - w_i^s\|_\infty \le \frac{\varepsilon_s}{2\underline\alpha} + \lambda_s \max_{i\in\mathcal{K}} \|g_i^s\|_\infty.$$

*(ii) The sequences $\{g_i^s\}$ and $\{\eta^s\}$, where $\eta^s := \nabla h(x^s)^\top \nu^s - \gamma^s + \theta^s$, are bounded. Consequently, $\delta_s \le \frac{\varepsilon_s}{2\underline\alpha} + M_g\lambda_s$ with $M_g := \sup_s \max_{i\in\mathcal{K}} \|g_i^s\|_\infty < \infty$ and, therefore $\delta_s \downarrow 0$.*

*(iii) For any prescribed tolerances $\lambda_{tol} > 0$ and $\varepsilon_{tol} > 0$, Algorithm 4.1 with stopping criteria $\lambda_s \le \lambda_{tol}$ and $\varepsilon_s \le \varepsilon_{tol}$ terminates after finitely many iterations and returns a point $x^{\bar{s}}$ that is $(\varepsilon_{\bar{s}}, \delta_{\bar{s}})$-approximately sum-stationary, with $\delta_{\bar{s}} \le \frac{\varepsilon_{tol}}{2\underline{\alpha}} + M_g \lambda_{tol}$.*

*Proof.* (i) Fix an outer iteration $s$. At the terminating inner iteration index $k_s$, which is finite per Proposition 4.5, the KKT conditions of the model NLP at this iteration (see (4.11)–(4.13)) give

$$\nabla f(x^s) + \sum_{i \in \mathcal{K}} g_i^s + \eta^s = -2\alpha_{k_s - 1}(x^s - x_-^s),$$

where $\eta^s$ is the shorthand defined in (ii) above. We also remark that $\hat{R}(x^s) = \nabla f(x^s) + \sum_{i \in \mathcal{K}} g_i^s + \eta^s$ simply by the form of $\eta^s$ and definition of $\hat{R}(x^s)$ (see above (4.6)). By the algorithm's inner loop stopping criterion (4.6) at iteration $k_s$, we have that $\|\hat{R}(x^s)\|_\infty \le \varepsilon_s$, yielding the stationarity inequality from (i). The feasibility and complementarity conditions follow directly from the aforementioned KKT conditions of the model NLP.

The inner loop stopping criterion $\|\hat{R}(x^s)\|_\infty \le \varepsilon_s$ also implies

$$\|x^s - x_-^s\|_\infty \le \frac{\varepsilon_s}{2\alpha_{k_s - 1}} \le \frac{\varepsilon_s}{2\underline{\alpha}},$$

where the last inequality is due to the lower bound for penalty parameters $\alpha_k$ from Lemma 4.4.

On the other hand, since $x_-^s - w_i^s = \lambda_s g_i^s$ for all $i \in \mathcal{K}$, we can write

$$\|x^s - w_i^s\|_\infty \le \|x^s - x_-^s\|_\infty + \|x_-^s - w_i^s\|_\infty = \|x^s - x_-^s\|_\infty + \lambda_s \|g_i^s\|_\infty,$$

which proves the $(\varepsilon_s, \delta_s)$-approximate sum-stationarity of $x^s$.

(ii) Since $\{x^s\} \in C$, Assumption 2.5 ensures $\{x^s\}$ is bounded. Let

$$e^s = \nabla f(x^s) + \sum_{i \in \mathcal{K}} g_i^s + \eta^s.$$

Suppose by contradiction that the family $\{g_i^s\}_s$, $i \in \mathcal{K}$ and $\{\eta^s\}_s$ is unbounded as $\lambda_s \to 0$. Passing to an unbounded subsequence and relabeling if necessary, we may assume $\sum_{i \in \mathcal{K}} \|g_i^s\| + \|\eta^s\| \to +\infty$. Define the scaling

$$t_s = \frac{1}{1 + \sum_{i \in \mathcal{K}} \|g_i^s\| + \|\eta^s\|} \downarrow 0,$$

and multiply stationarity identity above by $t_s$ to obtain

$$t_s \nabla f(x^s) + \sum_{i \in \mathcal{K}} t_s g_i^s + t_s \eta^s = t_s e^s.$$

Since $x^s$ is bounded and $\nabla f(\cdot)$ is continuous, $\{\nabla f(x^s)\}$ must be bounded, hence $t_s \nabla f(x^s) \to 0$. Similarly, since $\|e^s\|_\infty = \|\hat{R}(x^s)\|_\infty \le \varepsilon_s$, we must have $t_s e^s \to 0$. Hence,

$$\sum_{i \in \mathcal{K}} t_s g_i^s + t_s \eta^s \to 0. \tag{4.16}$$

On the other hand, simply by the definition of $t_s$ and the unboundedness hypothesis, we observe that

$$\sum_{i\in\mathcal{K}} \|t_s g_i^s\| + \|t_s \eta^s\| = \frac{\sum_{i\in\mathcal{K}} \|g_i^s\| + \|\eta^s\|}{1 + \sum_{i\in\mathcal{K}} \|g_i^s\| + \|\eta^s\|} \to 1,$$

and, therefore, $\{(t_s g_i^s)_{i\in\mathcal{K}},\, t_s\eta^s\}$ has a nonzero cluster point. Passing to a subsequence if needed, we obtain

$$t_s g_i^s \to v_i^\infty \;\forall i \in \mathcal{K} \text{ and } t_s\eta^s \to \eta^\infty,$$

with at least one of the $v_i^\infty$ and $\eta^\infty$ being nonzero.

Consider a convergent subsequence, still not relabeled, such that $x^s \to x^\star$. From part (i), $\|x^s - x_-^s\| \le \frac{\varepsilon_s}{2\underline{\alpha}}$, hence $x_-^s \to x^\star$ as well. Lemma 4.8 applied to $\{x_-^s\}$ implies $w_i^s \to x^\star$; also $r_i(w_i^s) \to r_i(x^*)$ under the $r_i$-attentive convergence of $\{w_i^s\}$ from Condition 4.9.

We next prove $v_i^\infty \in \partial^\infty r_i(x^\star)$. We first pass to regular subgradients required by the definition of horizon subdifferentials. Since $g_i^s \in \partial r_i(w_i^s)$, definition of the limiting subdifferential yields points $\hat{w}_i^s$ and $\hat{g}_i^s \in \hat{\partial} r_i(\hat{w}_i^s)$ for all $s$ such that

$$\|\hat{w}_i^s - w_i^s\| \le t_s,\; |r_i(\hat{w}_i^s) - r_i(w_i^s| \le t_s, \text{ and } \|\hat{g}_i^s - g_i^s\| \le t_s.$$

Then $\hat{w}_i^s \to x^\star$, $r_i(\hat{w}_i^s) \to r_i(x^\star)$, and

$$\|t_s\hat{g}_i^s - v_i^\infty\| \le t_s\|\hat{g}_i^s - g_i^s\| + \|t_s g_i^s - v_i^\infty\| \to 0.$$

With these regular subgradients, the definition of the horizon subdifferential provides

$$v_i^\infty \in \partial^\infty r_i(x^\star),\; \forall i \in \mathcal{K}.$$

Similarly, since each $N_C(x^s)$ is a cone and $\eta^s \in N_C(x^s)$, we have $t_s\eta^s \in N_C(x^s)$. By outer semicontinuity of the limiting normal cone [30, Proposition 6.6], $x^s \to x^\star$ and $t_s\eta^s \to \eta^\infty$ imply $\eta^\infty \in N_C(x^\star)$. Taking the limit $t_s \downarrow 0$ in (4.16), we obtain $\sum_{i\in\mathcal{K}} v_i^\infty + \eta^\infty = 0$. BCQ at $x^\star$ then forces $v_i^\infty = 0$ for all $i$ and $\eta^\infty = 0$, contradicting that at least one cluster is nonzero. Therefore $\{g_i^s\}$ and $\{\eta^s\}_s$ are bounded.

(iii) Since $\lambda_s \downarrow 0$ and $\varepsilon_s \downarrow 0$, there exists a finite outer index $\bar{s}$ with $\lambda_{\bar{s}} \le \lambda_{tol}$ and $\varepsilon_{\bar{s}} \le \varepsilon_{tol}$. By Proposition 4.5 for all outer iterations $s \le \bar{s}$ the inner iterations terminate finitely, therefore $x^{\bar{s}}$ is reached in a finite number of iterations. The $(\varepsilon_{\bar{s}}, \delta_{\bar{s}})$-approximate stationarity follows from (i) and (ii) above. ∎

An explicit coupling between inner-loop tolerance $\varepsilon_s$ and regularization $\lambda_s$ from Proposition 4.10 should be used by a numerical method based on Algorithm 4.1 to properly balance accuracy of the inner and outer solves. The key bound is given by Proposition 4.10:

$$\delta_s \le \frac{\varepsilon_s}{2\underline{\alpha}} + \lambda_s \max_{i\in\mathcal{K}} \|g_i^s\|_\infty = \frac{\varepsilon_s}{2\underline{\alpha}} + \max_{i\in\mathcal{K}} \|x_-^s - w_i^s\|_\infty.$$

A natural choice is to set $\varepsilon_s$ the same order as the proximal displacement $B_s := \max_{i\in\mathcal{K}} \|x_-^s - w_i^s\|_\infty$. Otherwise, if $\varepsilon_s \ll B_s$, the regularized subproblem (1.4)–(1.5) is oversolved during the inner iterations; and, if $B_s \ll \varepsilon_s$, the overall error is dominated by the inner-loop accuracy. A practical, a posteriori rule for the inner-loop termination (step 20 in the algorithm) would therefore be

$$\|\hat{R}(\tilde{x}^k)\|_\infty \le B \max_{i\in\mathcal{K}} \|\tilde{x}^k - w_i^k\|_\infty \text{ or, equivalently, } \|\hat{R}(\tilde{x}^k)\|_\infty \le B\lambda_s \max_{i\in\mathcal{K}} \|g_i^k\|_\infty,$$

for some prescribed $B > 0$.

THEOREM 4.11. *Consider Algorithm 4.1 under a sequence of regularizations* $\lambda_s \downarrow 0$ *and inner-loop tolerances* $\varepsilon_s \downarrow 0$. *Under the same assumptions as in Proposition 4.10, any cluster point* $x^\star$ *of the bounded sequence* $\{x^s\}$ *satisfies the sum-stationarity conditions* (4.7)–(4.8) *for the target problem* (1.1)–(1.2).

*Proof.* The sequence $\{x^s\}$ is bounded since $C$ is compact by Assumption 2.5. We take an arbitrary accumulation point $x^\star$ and a convergent subsequence (not relabeled) $x^s \to x^\star$. By (ii) of Proposition 4.10, we can take subsequences, not relabeled, such that

$$g_i^s \to g_i^\star, \forall i \in \mathcal{K}, \text{ and } \eta^s \to \eta^\star.$$

Possibly after a relabeling, for $x^s \to x^\star$ we have $\nabla f(x^s) \to \nabla f(x^\star)$ by continuity of $\nabla f$ given by Assumption 2.4. Also by closedness of normal cones $\eta^\star \in N_C(x^\star)$.

Using a similar argument as in the proof of (ii) of Proposition 4.10, one can prove that $x_-^s \to x^\star$ as well. Lemma 4.8 applied to the sequence $\{x_-^s\}$ implies $w_i^s \to x^\star$. Since $g_i^s \in \partial r_i(w_i^s)$ and $g_i^s \to g_i^\star$, under the $r_i$-attentive convergence of $\{w_i^s\}$ from Condition 4.9, we must have $g_i^\star \in \partial r_i(x^\star)$ for each $i \in \mathcal{K}$, by the outer semicontinuity of limiting subdifferential [30, Proposition 8.7].

With the above convergence properties, we can take the limit of the stationarity condition from (i) of Proposition 4.10 to obtain

$$\nabla f(x^\star) + \sum_{i \in \mathcal{K}} g_i^\star + \eta^\star = 0, \quad \text{with } g_i^\star \in \partial r_i(x^\star) \text{ and } \eta^\star \in N_C(x^\star).$$

Under the MFCQ for the feasible set $C$ of the master problem (see Theorem 4.6), $\eta^\star \in N_C(x^\star)$ has the representation $\eta^\star = \nabla h(x^\star)^\top \nu^\star - \gamma^\star + \theta^\star$ with the multipliers $\nu^\star$, $\gamma^\star$, and $\theta^\star$ satisfying sign/complementarity conditions from (4.8). Finally, the stationarity equation above implies the stationarity condition (4.7), which completes the proof. ∎

A sharper iteration-complexity analysis via Kurdyka–Łojasiewicz (KL) assumption and arguments is not immediate in our framework, since our stationarity measure is based on proximal oracle subgradients, whereas the standard KL machinery is typically formulated in terms of the limiting subdifferential and its associated slope. Relating these two notions in a way that yields clean rate complexity estimates would very likely require additional regularity assumptions, so we defer such a KL-based analysis to future work.

**4.3. Stronger convergence guarantees under additional regularity.** For several classes of objective functions $r_i$ covered by Assumption 2.4, including lower-$C^1$, lower-$C^2$, and (lsc) convex functions, the sum-stationarity obtained above can be strengthened to the more familiar limiting subgradient stationarity. Moreover, the convergence prerequisites from the previous section hold automatically for broader classes of nonsmooth objectives, including locally Lipschitz functions (which encompass all of the aforementioned classes) and semismooth functions.

LEMMA 4.12. *If the functions* $r_i$ *satisfying Assumption 2.4 are also subdifferentially regular in the sense of Rockafellar and Wets [30, Definition 7.25] (also known as lower regular [24]), the sum-stationarity condition* (4.7) *from Theorem 4.11 takes the more commonly used form*

$$0 \in \nabla f(x^\star) + \partial \left( \sum_{i \in \mathcal{K}} r_i(x^\star) \right) + \nabla h(x^\star)^\top \nu^\star - \gamma^\star + \theta^\star.$$

*Proof.* The stationarity condition follows from

$$\sum_{i\in\mathcal{K}} \partial r_i(x^\star) = \sum_{i\in\mathcal{K}} \hat{\partial} r_i(x^\star) \subseteq \hat{\partial}\left(\sum_{i\in\mathcal{K}} r_i(x^\star)\right) \subseteq \partial\left(\sum_{i\in\mathcal{K}} r_i(x^\star)\right),$$

where the first inclusion is given by the sum rule for regular subgradients [30, Corollary 10.9], the second inclusion is due to the regular subdifferential being a subset of the limiting subdiferential, and the equality is due to $r_i$ being subdifferentially regular. ∎

We remark that if $r_i$ are also locally Lipschitz, then the horizon qualification from Corollary 10.9 of [30] implies the stronger relationship $\sum_{i\in\mathcal{K}} \partial r_i(x^\star) = \partial\left(\sum_{i\in\mathcal{K}} r_i(x^\star)\right)$.

Under the Lipschitzian property for objective terms $r_i$, the conditions needed for the finite termination and subsequence convergence of Proposition 4.10 and Theorem 4.11 can be considerably relaxed, as shown in the following proposition.

PROPOSITION 4.13. *If the functions $r_i$ are locally Lipschitz, the finite convergence properties of Proposition 4.10 and the asymptotic convergence of Theorem 4.11 hold only under Assumptions 2.4 and 2.5 and MFCQ of Theorem 4.6 since Condition 4.9 and BCQ hold automatically.*

*Proof.* The proof is immediate once we observe that (i) $\partial^\infty r_i(\bar{x}) = \{0\}$ by local Lipschitz continuity [25], hence BCQ is satisfied, and (ii) Condition 4.9 also holds due to continuity of $r_i$. ∎

*Remark* 4.14. Proposition 4.13 applies to semismooth functions [22], lower-$C^1$, lower-$C^2$, and (lsc) convex functions, which are locally Lipschitz.

The following lemma shows that Condition 4.9 holds for the general case ($r_i$ only satisfy Assumption 2.4 and are possibly discontinuous) if the inner iterations of Algorithm 4.1 finish with (approximate) minimizers $x^s$ of the inner, fixed-$\lambda$ problem (1.4)–(1.5) instead of (approximate) stationary points (in the sense of (i) of Proposition 4.10). This occurs, for example, when the feasible set $C$ and the objective $f$ in the master problem are convex and it may also occur as well when a global numerical solver is used for the inner problem.

LEMMA 4.15. *Consider Algorithm 4.1 under a sequence of regularization parameters $\lambda_s \downarrow 0$. Suppose that, for each outer iteration $s$, the inner loop terminates after a finite number $k_s$ of inner iterations and returns the outer iterate $x^s := \tilde{x}^{k_s} \in C$, satisfying*

$$F_{\lambda_s}(x^s) \leq \inf\{F_{\lambda_s}(x) : x \in C\} + \tau_s,$$

*for some sequence $\tau_s \downarrow 0$. Let $x^s \to x^\star$ be a convergent subsequence of such outer iterates. Then Condition 4.9 holds along this subsequence.*

*Proof.* Let $w_i^s \in \operatorname{prox}_{\lambda_s} r_i(x^s)$ and define

$$t_i^s := \frac{1}{2\lambda_s}\|x^s - w_i^s\|^2 \geq 0.$$

From the choice of $w_i^s$, we have $e_{\lambda_s} r_i(x^s) = r_i(w_i^s) + t_i^s$, and therefore $F_{\lambda_s}(x^s) = f(x^s) + \sum_{i\in K}\left(r_i(w_i^s) + t_i^s\right)$.

Since $x^\star \in C$ and $x^s$ is $\tau_s$-suboptimal for $F_{\lambda_s}$ over $C$,

$$F_{\lambda_s}(x^s) \leq \inf\{F_{\lambda_s}(x) : x \in C\} + \tau_s \leq F_{\lambda_s}(x^\star) + \tau_s.$$

Also, $e_{\lambda_s} r_i(x^\star) \le r_i(x^\star)$ by the monotonicity of the Moreau envelope (see Theorem 3.1), hence

$$f(x^s) + \sum_{i\in K} \big(r_i(w_i^s) + t_i^s\big) \le f(x^\star) + \sum_{i\in K} r_i(x^\star) + \tau_s.$$

Since $f$ is continuous, $x^s \to x^\star$, and $\tau_s \to 0$, we obtain

$$\limsup_{s\to\infty} \sum_{i\in K} \big(r_i(w_i^s) + t_i^s\big) \le \sum_{i\in K} r_i(x^\star). \tag{4.17}$$

On the other hand, Lemma 4.8 gives $w_i^s \to x^\star$, and lower semicontinuity of $r_i$ implies

$$r_i(x^\star) \le \liminf_{s\to\infty} r_i(w_i^s), \qquad \forall i \in K.$$

Summing over $i \in K$ and using $t_i^s \ge 0$, we get

$$\sum_{i\in K} r_i(x^\star) \le \sum_{i\in K} \liminf_{s\to\infty} r_i(w_i^s) \le \liminf_{s\to\infty} \sum_{i\in K} r_i(w_i^s) \le \liminf_{s\to\infty} \sum_{i\in K} \big(r_i(w_i^s) + t_i^s\big),$$

where the second inequality uses properties of lim inf (i.e., Fatou's Lemma). Combining this with (4.17) yields

$$\lim_{s\to\infty} \sum_{i\in K} \big(r_i(w_i^s) + t_i^s\big) = \sum_{i\in K} r_i(x^\star).$$

Finally, for each $i \in K$, the lower semicontinuity of $r_i$, nonnegativity of $t_i^s$, and $w_i^s \to x^\star$ imply

$$\liminf_{s\to\infty} \big(r_i(w_i^s) + t_i^s\big) \ge \liminf_{s\to\infty} r_i(w_i^s) \ge r_i(x^\star).$$

The last two relations imply $r_i(w_i^s) + t_i^s \to r_i(x^\star), \forall i \in K$. This and $w_i^s \to x^\star$ imply

$$\begin{aligned} 0 &\le \limsup_{s\to\infty} t_i^s = \limsup_{s\to\infty} [r_i(w_i^s) + t_i^s - r_i(w_i^s)] \le \limsup_{s\to\infty} [r_i(w^s) + t_i^s] - \liminf_{s\to\infty} r_i(w_i^s) \\ &= r_i(x^\star) - \liminf_{s\to\infty} r_i(w_i^s) \le 0, \end{aligned}$$

therefore $t_i^s \to 0$ and, consequently, $r_i(w_i^s) \to r_i(x^\star)$, proving Condition 4.9. ∎

**5. A case study in parametric optimization with applications to electrical power grids.** As mentioned in the introduction, a main motivation for the present framework comes from optimization models in which the nonsmooth terms $r_i$ are scenario-wise optimal value functions. A representative form is

$$\begin{aligned} r_i(x) := \min_{y_1, y_2}\ & c_i(y_1, y_2) \\ \text{s.t. } & x - \Delta_i \le y_1 \le x + \Delta_i,\ y_\ell \le y_2 \le y_u,\ h_i(y_1, y_2) = 0, \end{aligned} \tag{5.1}$$

where $x$ denotes the master decision, $y_1$ collects scenario variables constrained to remain close to $x$, $y_2$ collects scenario-local state variables, $\Delta_i > 0$ prescribes admissible post-scenario deviations, and $i \in K$ indexes scenarios, *e.g.*, system contingencies or uncertainty realizations. We use this bound-coupling form only for notational simplicity. The same discussion applies when $x$ also enters the objective or the general constraints of the scenario problem.

This is precisely the setting in which Algorithm 4.1 is useful. In many applications, evaluating $r_i(x)$ already requires solving a large NLP. The main difficulty is therefore not the function evaluation itself, but obtaining reliable first-order information for the value function $r_i$. Even when the scenario data are $C^2$, the marginal function $r_i$ is, in general, only lsc [30]. Under additional constraint qualifications such as MFCQ, it becomes locally Lipschitz [13, 14]. However, even under MFCQ, both Clarke and limiting subdifferentials may fail to admit multiplier-based representations in degenerate or unstable nonconvex cases, see for example, Example 1 in [13] and Theorem 4 in [23], respectively. Algorithm 4.1 avoids differentiating through the scenario NLP (5.1). Instead of relying on potentially numerically unavailable subgradients, it requires only a proximal point of $r_i$, which can be obtained by solving essentially the same scenario problem with an additional convex quadratic term. Concretely, one needs to solve

$$
\begin{aligned}
e_\lambda r_i(x) = \min_{y_1,y_2,w}\ & c_i(y_1,y_2) + \frac{1}{2\lambda}\|w-x\|^2 \\
\text{s.t.}\ \ & w-\Delta_i \le y_1 \le w+\Delta_i, \quad y_\ell \le y_2 \le y_u, \quad h_i(y_1,y_2)=0,
\end{aligned}
\tag{5.2}
$$

and any minimizer $w_i^\lambda$ yields a proximal oracle subgradient $g_i^\lambda = \frac{1}{\lambda}(x - w_i^\lambda)$. Thus, once a solver for the scenario NLP is available, the oracle required by Algorithm 4.1 can often be implemented with minimal structural changes, without differentiating through KKT multipliers or relying on sensitivity routines.

This viewpoint is particularly natural for preventive security-constrained optimal power flow models (SC-OPF) and related stochastic programming models relevant in power dispatch operations under stochastic renewable generation, loads, and operating conditions [3]. The master variable $x$ represents base-case power dispatch and control settings. Each scenario $i \in K$ corresponds to an $N-1$ contingency, such as the loss of a generator, line, or transformer. For fixed $x$, the scenario subproblem (5.1) computes a post-contingency operating point subject to power flow equations, nodal active/reactive balance, generator capability limits, emergency voltage and thermal limits, and contingency transition rules that describe the post-contingency response and grid state via the constraints in (5.1). The scenario objective $c_i$ is often convex, as it captures quadratic penalties associated with the post-contingency response. The NLP (5.1) is generally convex for direct-current (DC) OPF but nonconvex for alternating-current (AC) OPF. For DC models, computing numerical proximal subgradients with local optimizers applied to the proximal problem (5.2) is therefore straightforward; for AC models, nonconvexity can be problematic. In our extensive computational experience with scenario AC NLPs (5.1), local interior-point solvers consistently return high-quality solutions that are stable across a range of initializations. Since the proximal NLP (5.2) preserves the same scenario structure while adding convex quadratic regularization, this evidence suggests that such solvers are capable of recovering minimizers relevant to the Moreau envelope, and thereby provide valid proximal oracle points and subgradients. A systematic numerical investigation is deferred to future work.

Algorithm 4.1 matches the operational structure of grid models in several ways. First, the master constraints (1.2) are enforced exactly, so every accepted iterate is a feasible base-case operating point. This is especially attractive in real-time grid operations, which are subject to stringent time-to-solution requirements, where preserving feasibility and security is often more important than aggressively pursuing marginal cost decreases. Second, the proximal oracles are scenario-separable since for a given $x$, each $e_\lambda r_i$ can be evaluated independently, resulting in a computational pattern with

exceptional parallel efficiency at large scale [20]. Third, the method returns first-order information in a form that is immediately usable by the master algorithm, namely the proximal oracle subgradients $g_i^\lambda$, without requiring the delicate post-processing of dual information from the inner NLP solves. In [27] we successfully employed derivative-free convex proximal stabilizations for $r_i$, constructed using power systems engineering insight, to address the pervasive difficulties associated with derivative evaluations of $r_i$. This strategy, however, entails a loss of robustness and limits the method's applicability to other classes of optimization problems. In this sense, Algorithm 4.1 turns the proximal stabilization intuition behind decomposition-based SC-OPF heuristics from [27] into a principled continuation method with explicit stationarity guarantees.

The general regularity assumptions and results of Section 4 apply in this setting. In addition to being lsc, the value functions $r_i$ are proper and finite-valued whenever the scenario problems are feasible for all relevant $x$; furthermore, since the scenario objective $c_i$ from (5.1) is generally bounded from below, $r_i$ is also prox-bounded and therefore falls under Assumption 2.4. When MFCQ holds for the scenario NLP (5.1), $r_i$ is locally Lipschitz [14], so the stronger Proposition 4.13 applies (since Condition 4.9 and BCQ hold automatically). Thus, the parametric NLP setting provides a concrete class of problems compatible with the convergence guarantees presented above.

**6. Conclusions and future work.** We introduced a proximal oracle sequential NLP framework for nonsmooth constrained optimization problems in which the nonsmooth objective is a finite sum of proper lower semicontinuous nonconvex terms accessed through proximal oracles. The method applies Moreau regularization termwise to preserve rich decomposition properties of the original problem and builds exact-constraint NLP models from the proximal points $w_i \in \operatorname{prox}_\lambda r_i(x)$ and the associated proximal oracle subgradients $(x - w_i)/\lambda$. This construction avoids differentiating through scenario subproblems and does not rely on multiplier sensitivity calculations. We proved consistency of the distributed Moreau-envelope continuation at the level of global minimizers. For fixed $\lambda$, the inner sequential NLP iteration has cluster points satisfying proximal limiting stationarity and Clarke stationarity for the regularized problem. Along the outer continuation, cluster points satisfy first-order sum-stationarity for the original problem under the stated attentive-convergence and CQ assumptions; when the terms $r_i$ are locally Lipschitz, these additional convergence prerequisites hold automatically. The iterative algorithm terminates finitely for prescribed positive regularization and stationarity tolerances with an explicit error bound. We also show how the required proximal oracles can be computed with minimal modifications of the scenario NLP for the cases when $r_i$ are optimal value functions of parametric problems.

Several directions remain for future work. A primary direction is the development of a robust implementation of the proposed algorithm, including treatment of inexact proximal oracles, adaptive choices of regularization and inner-loop tolerances, and safeguards for nonconvex NLP oracles when solved by local numerical solvers. Such an implementation would enable systematic numerical evaluation on large-scale problems and would clarify the tradeoffs among the various algorithmic and computational choices that affect scalability, robustness, and applicability. Further work will also investigate sharper complexity guarantees, for example under a KL assumption. This would require relating the proximal oracle subgradient stationarity used here to a limiting subdifferential slope [5], possibly under additional structure such as prox-regularity [28] or a KL property [4, 5].

**Acknowledgments.** This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under contract DE–AC52–07NA27344.

**Appendix A. Sum rule for Clarke subdifferentials.** The following lemma establishes that the Clarke subdifferential satisfies the exact sum rule for upper-$C^2$ functions. We include a proof because we were not able to find it established previously in the literature.

LEMMA A.1. *Let $g_1, g_2 : \mathbb{R}^n \to \mathbb{R}$ be upper-$C^2$ in a neighborhood of $x$. Then*

$$\partial^C(g_1 + g_2)(x) = \partial^C g_1(x) + \partial^C g_2(x).$$

*Proof.* Since $g_1$ and $g_2$ are upper-$C^2$ near $x$, by the local representation theorem for upper-$C^2$ functions [30, Theorem 10.33], there exist a neighborhood $U$ of $x$, $C^2$ functions $\phi_i : U \to \mathbb{R}$, and convex functions $q_i : U \to \mathbb{R}$ such that $g_i = \phi_i - q_i$ on $U$, $i = \{1, 2\}$. Therefore $g_1 + g_2 = (\phi_1 + \phi_2) - (q_1 + q_2)$ on $U$, with $\phi_1 + \phi_2$ being $C^2$ and $q_1 + q_2$ being convex.

For any proper convex function $q$, the Clarke subdifferential coincides with the convex subdifferential [7], that is, $\partial^C q(x) = \partial q(x)$. Moreover, for a $C^1$ function $\phi$ and a locally Lipschitz function $\psi$, the Clarke sum rule [7] gives $\partial^C(\phi + \psi)(x) = \nabla\phi(x) + \partial^C\psi(x)$. Applying this with $\psi = -q$ and using $\partial^C(-q)(x) = -\partial^C q(x)$, we obtain $\partial^C(\phi - q)(x) = \nabla\phi(x) - \partial q(x)$. Therefore,

$$\partial^C(g_1 + g_2)(x) = \nabla(\phi_1 + \phi_2)(x) - \partial(q_1 + q_2)(x).$$

Since $q_1$ and $q_2$ are convex and finite on $U$, the convex subdifferential sum rule [30] yields $\partial(q_1 + q_2)(x) = \partial q_1(x) + \partial q_2(x)$. Consequently,

$$\begin{aligned}\partial^C(g_1 + g_2)(x) &= \nabla\phi_1(x) + \nabla\phi_2(x) - \big(\partial q_1(x) + \partial q_2(x)\big) \\ &= \big(\nabla\phi_1(x) - \partial q_1(x)\big) + \big(\nabla\phi_2(x) - \partial q_2(x)\big) = \partial^C g_1(x) + \partial^C g_2(x).\end{aligned}$$

This proves the result. ∎

## REFERENCES

[1] F. J. ARAGÓN ARTACHO AND J. M. BORWEIN, *Global convergence of a non-convex Douglas–Rachford iteration*, Journal of Global Optimization, 57 (2013), pp. 753–769.

[2] F. J. ARAGÓN ARTACHO, J. M. BORWEIN, AND M. K. TAM, *Global behavior of the Douglas–Rachford method for a nonconvex feasibility problem*, Journal of Global Optimization, 65 (2016), pp. 309–327.

[3] I. ARAVENA, D. K. MOLZAHN, S. ZHANG, C. G. PETRA, F. E. CURTIS, S. TU, A. WÄCHTER, E. WEI, E. WONG, A. GHOLAMI, K. SUN, X. A. SUN, S. T. ELBERT, J. T. HOLZER, AND A. VEERAMANY, *Recent developments in security-constrained AC optimal power flow: Overview of Challenge 1 in the ARPA-E Grid Optimization Competition*, Operations Research, 71 (2023), pp. 1997–2014.

[4] H. ATTOUCH AND J. BOLTE, *On the convergence of the proximal algorithm for nonsmooth functions involving analytic features*, Mathematical Programming, 116 (2009), pp. 5–16.

[5] H. ATTOUCH, J. BOLTE, AND B. F. SVAITER, *Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized Gauss–Seidel methods*, Mathematical Programming, 137 (2013), pp. 91–129.

[6] J. V. BURKE, F. E. CURTIS, A. S. LEWIS, M. L. OVERTON, AND L. E. A. SIMÕES, *Gradient sampling methods for nonsmooth optimization*, in Numerical Nonsmooth Optimization: State of the Art Algorithms, A. M. Bagirov, M. Gaudioso, N. Karmitsa, M. M. Mäkelä, and S. Taheri, eds., Springer, Cham, 2020, pp. 201–225.

[7] Y. Cui and J.-S. Pang, *Modern Nonconvex Nondifferentiable Optimization*, vol. 29 of MOS-SIAM Series on Optimization, Mathematical Optimization Society (MOS) and Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, Dec 2021.

[8] F. E. Curtis, T. Mitchell, and M. L. Overton, *A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles*, Optimiz. Meth. and Software, 32 (2017), pp. 148–181.

[9] F. E. Curtis and M. L. Overton, *A sequential quadratic programming algorithm for nonconvex, nonsmooth constrained optimization*, SIAM J. on Optimization, 22 (2012), pp. 474–500.

[10] F. E. Curtis, X. Qu, and D. P. Robinson, *A proximal-gradient method for solving regularized optimization problems with general constraints*, 2025, https://arxiv.org/abs/2512.23166. Lehigh ISE Technical Report 25T-023; arXiv version revised 15 January 2026.

[11] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM J. on Optimization, 29 (2019), pp. 207–239.

[12] A. De Marchi, X. Jia, C. Kanzow, and P. Mehlitz, *Constrained composite optimization and augmented Lagrangian methods*, Mathematical Programming, (2023).

[13] J. Gauvin, *The generalized gradient of a marginal function in mathematical programming*, Mathematics of Operations Research, 4 (1979), pp. 458–463.

[14] J. Gauvin and F. Dubeau, *Differential properties of the marginal function in mathematical programming*, in Optimality and Stability in Mathematical Programming, vol. 19 of Mathematical Programming Studies, Springer, 1982, pp. 101–119.

[15] K. Guo, D. Han, D. Z. W. Wang, and T. Wu, *Convergence of ADMM for multi-block nonconvex separable optimization models*, Frontiers of Mathematics in China, 12 (2017), pp. 1139–1162.

[16] R. Hesse and D. R. Luke, *Nonconvex notions of regularity and convergence of fundamental algorithms for feasibility problems*, SIAM J. on Optimization, 23 (2013), pp. 2397–2419.

[17] M. Hong, Z.-Q. Luo, and M. Razaviyayn, *Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems*, SIAM J. on Optimization, 26 (2016), pp. 337–364.

[18] A. Jourani, L. Thibault, and D. Zagrodny, *Differential properties of the Moreau envelope*, Journal of Functional Analysis, 266 (2014), pp. 1185–1237.

[19] P. D. Khanh, V. V. H. Khoa, B. S. Mordukhovich, and V. T. Phat, *Local minimizers of nonconvex functions in Banach spaces via Moreau envelopes*, Vietnam Journal of Mathematics, 53 (2025), pp. 803–813.

[20] Lawrence Livermore Natl. Lab., *LLNL-Developed Software Leverages Exascale Power to Optimize National Grid's Emergency Response*. HPCwire, Off the Wire Press Releases, Aug. 2023, https://www.hpcwire.com/off-the-wire/llnl-developed-software-leverages-exascale-power-to-optimize-national-grids-emergency-response/. Accessed: 2026-05-08.

[21] G. Li and T. K. Pong, *Douglas–Rachford splitting for nonconvex optimization with application to nonconvex feasibility problems*, Mathematical Programming, 159 (2016), pp. 371–401. Series A.

[22] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. on Control and Optimization, 15 (1977), pp. 959–972.

[23] B. S. Mordukhovich, N. M. Nam, and N. D. Yen, *Subgradients of marginal functions in parametric mathematical programming*, Mathematical Programming, 116 (2009), pp. 369–396.

[24] B. S. Mordukhovich and T. T. A. Nghia, *Subdifferentials of nonconvex supremum functions and their applications to semi-infinite and infinite programs with lipschitzian data*, SIAM Journal on Optimization, 23 (2013), pp. 406–431.

[25] B. S. Mordukhovich and R. T. Rockafellar, *Second-order subdifferential calculus with applications to tilt stability in optimization*, SIAM Journal on Optimization, 22 (2012), pp. 953–986.

[26] N. Parikh and S. Boyd, *Proximal algorithms*, Foundations and Trends® in Optimization, 1 (2014), pp. 127–239.

[27] C. G. Petra and I. Aravena, *A surrogate-based asynchronous decomposition technique for realistic security-constrained optimal power flow problems*, Operations Research, 71 (2023), pp. 2015–2030.

[28] R. A. Poliquin and R. T. Rockafellar, *Prox-regular functions in variational analysis*, Transactions of the American Mathematical Society, 348 (1996), pp. 1805–1838.

[29] Z. Qi, Y. Cui, Y. Liu, and J. Pang, *Asymptotic properties of stationary solutions of nonconvex nonsmooth empirical risk minimization*, Math. Oper. Res., 47 (2022), pp. 2034–2064.

[30] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, Berlin Hei-

delberg, 1998.
[31] A. Themelis and P. Patrinos, *Douglas–Rachford splitting and ADMM for nonconvex optimization: Tight convergence results*, SIAM J. on Optimization, 30 (2020), pp. 149–181.
[32] A. Themelis, L. Stella, and P. Patrinos, *Douglas–Rachford splitting and ADMM for nonconvex optimization: accelerated and Newton-type linesearch algorithms*, Computational Optimization and Applications, 82 (2022), pp. 395–440.
[33] J. Wang and C. G. Petra, *A sequential quadratic programming algorithm for nonsmooth problems with upper-$\mathcal{C}^2$ objective*, SIAM J. on Optimization, 33 (2023), pp. 2379–2405.
[34] Y. Wang, W. Yin, and J. Zeng, *Global convergence of ADMM in nonconvex nonsmooth optimization*, Journal of Scientific Computing, 78 (2019), pp. 29–63.
[35] G. D. Yalcin and F. E. Curtis, *Incremental quasi-Newton algorithms for solving a nonconvex, nonsmooth, finite-sum optimization*, Optimiz. Meth. and Software, 39 (2024), pp. 345–367.