CUTTING PLANE ORACLES FOR NON-SMOOTH TRUST-REGIONS

DOMINIKUS NOLL

Abstract. We prove global convergence of a bundle trust region algorithm for non-smooth non-convex optimization, where cutting planes are generated by oracles respecting four basic rules. The benefit is that convergence theory applies to a large variety of methods encountered in practice. This includes in particular the method of downshifted tangents, for which previously no convergence result in the trust region framework was known. We also show that certain splitting techniques can be seen as special cases of bundle trust region techniques.

1. Introduction

We consider optimization problems of the form

$$\min_{x\in C} f(x) \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz, but neither differentiable, nor convex, and where C is a closed convex set in $\mathbb{R}^n$, which is typically of a simple structure, such as a polyhedron, or a semi-definite set. We discuss a non-smooth trust region algorithm, which generates local models of f by accumulating information from cutting planes at unsuccessful trial steps. We prove global convergence of our algorithm in the sense that all accumulation points x^* of the sequence of serious iterates x^j are critical points of (1.1), i.e., satisfy $0 \in \partial f(x^*) + N_C(x^*)$.

In [13, 14] we had discussed a trust region method, where model-based cutting planes were used. While this type of oracle is suited for a number of applications, we presently cover a larger variety of oracles encountered in practice. This includes in particular the well-know method of downshifted tangents, for which convergence in the trust regions framework had so far not been established. Our global convergence result applies to all cutting plane oracles satisfying four basic rules.

We answer the question posed in [60], whether aggregation developed in [36] for the convex bundle method to limit the number of cuts, could also be used in trust regions for the same purpose. In [52] we had justified aggregation for the non-convex bundle method, and in [14] for model based

2010 *Mathematics Subject Classification.* 65K05,90C56,90C90.

Key words and phrases. Non-smooth optimization · non-smooth trust region method · cutting plane oracle · global convergence · forward-backward splitting.

oracles in trust regions under the additional hypothesis that a positive definite second-order term was used in the working model. Here we complete our analysis by showing that this hypothesis cannot be removed. See [13, Remark 10] and also [62, 69, 30, 61] for this topic.

Convergence of bundle or trust region bundle methods requires that the objective f admits a *strict model*. Models of locally Lipschitz functions were introduced in [52] and may be understood as non-smooth substitutes for the first-order Taylor expansion. Every model gives rise to a cutting plane oracle, hence to a trust region or bundle method, and conversely, every cutting plane oracle generates a model as its upper envelope, so both notions are closely related. Strictness of a model in the non-smooth setting is the analogue to strict differentiability in the smooth setting. Strictness of an oracle decides whether or not the corresponding bundle or trust region method converges.

While the present work concentrates on theoretical convergence aspects, we have applied non-convex non-smooth bundle and trust region methods successfully to applications in mechanical contact problems, control and system theory, and operations research. For instance [29] applies a bundle technique with downshift to design a flight controller for a civil aircraft. In [25] a model-based bundle method is used for partial eigenstructure assignment, with applications to a launcher in atmospheric flight and to decoupling motions of a civil aircraft. In [8, 14, 58] the model-based trust-region method is applied to H_∞-control of infinite-dimensional systems, including boundary control of PDEs. In [26] we use our model-based technique in a delamination study in destructive testing in the material sciences. A non-smooth approach to reliable flight control is [9], where we also give a survey of recent relevant non-smooth methods in control. In [27] a bundle method with downshift is used to minimize the memory stored in a system, and an evacuation scenario of a fairground in case of an emergency is solved. In [11, 12] a non-convex spectral bundle method is used to solve bilinear matrix inequalities. This corresponds to cases where an infinity of cutting planes arise simultaneously. In [63] a frequency shaping control technique is developed using model-based bundling. Details on how non-linear constraints are handled are given in [29, 63, 9]. The first non-convex bundle method with convergence certificate under inexact subgradients and function values is [51], where downshifted tangents are used. The model-based trust-region bundle method has been used for parametric and mixed robust control, one of the most challenging problems in feedback control design [13, 10, 2, 1]. We have also used the trust region bundle method within a branch-and-bound approach to global optimization problems like computing the distance to instability of a controlled system, see [56, 55]. All these techniques can be seen as special cases of the present abstract framework.

The structure of the paper is as follows. In section 2 we discuss when a function f admits a strict model. Section 3 introduces rules for cutting plane oracles and relates them to models. We recall how cutting planes are used to

build working models. Section 4 presents the bundle trust region algorithm and its elements. A major difference with classical trust regions occurs in our management of the trust region radius. The central section 5 gives the global convergence proof for the algorithm. In section 6 we show that splitting techniques, the proximal point method, but also classical gradient oriented and quasi Newton methods are special cases of our oracle based bundle trust region concept. Extensions to constraint programs $\min\{f(x) : c(x) \leq 0, x \in C\}$ can be obtained by similar techniques via suitable progress functions in the sense of [46, 47, 48], or [29, 63, 9, 23], or by multi-objective methods [63, 6, 7].

2. The model concept

The basic tool in bundle and bundle trust-region methods is the cutting plane. Its rationale [60, 40, 41] is that if an unsuccessful trial step y (called a null step) in the neighborhood of the current serious iterate x is made, then a cutting plane to f at y should be included in the working model in order to orient the search for a new x^+ away from y in the next trial. In the convex case cuts are simply tangents to f at y, but for non-convex f, different ways to obtain cuts are needed. In order to understand how cutting planes are generated, we shall need the notion of a *model* of f near an iterate x, which can be considered as a non-smooth analogue of the first-order Taylor expansion of f at x. In this chapter, the theoretical background of the model concept is recalled.

2.1. **Model axioms.** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a locally Lipschitz function. Following [52, 49], a function $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a *model* of f if it satisfies the following properties:

- (M_1) $\phi(\cdot, x)$ is convex, $\phi(x, x) = f(x)$, and $\partial_1\phi(x, x) \subset \partial f(x)$.
- (M_2) If $y_k \to x$, then there exist $\epsilon_k \to 0^+$ such that $f(y_k) \leq \phi(y_k, x) + \epsilon_k\|y_k - x\|$.
- (M_3) If $x_k \to x, y_k \to y$, then $\limsup_{k\to\infty} \phi(y_k, x_k) \leq \phi(y, x)$.

Remark 2.1. We may interpret $\phi(\cdot, x)$ as a substitute for the first-order Taylor expansion of f at x. This is highlighted by the fact that every locally Lipschitz function f has the so-called *standard model*

$$\phi^\sharp(\cdot, x) = f(x) + f^\circ(x, \cdot - x),$$

where $f^\circ(x, d)$ is the Clarke directional derivative at x in direction d. However, a function f may have several models, so the notion of a model, unlike the Taylor expansion, claims no uniqueness.

Remark 2.2. Composite functions $f = \psi \circ F$ with convex ψ and F of class C^1 admit the so-called *natural model* $\phi(\cdot, x) = \psi\left(F(x) + F'(x)(\cdot - x)\right)$, see [11, 12, 52, 49]. Consequently, every lower-C^2 function admits a strict model, because by [59] it can be represented in the form $f = \psi \circ F$ on every bounded set.

Remark 2.3. When f is convex, its natural model is $\phi(\cdot, x) = f$. We say that a convex function is its own natural model. Note that convex f have still their standard model $\phi^\sharp$, which is different from f unless f is affine.

Proposition 2.4. (See [49, Lemmas 1,2]). *The standard model $\phi^\sharp$ is the smallest model of f, i.e., $\phi^\sharp \leq \phi$ for any other model ϕ of f.* □

Remark 2.5. The following operation on models is sometimes useful. Suppose f has two models, ϕ_1, ϕ_2, then $\phi = \max\{\phi_1, \phi_2\}$ is again a model. Equally useful is the following. Suppose $\phi_1, \ldots, \phi_m$ satisfy (M_1) and (M_3). Moreover, suppose ϕ_i satisfies (M_2) at points $x \in \Omega_i$ such that $\Omega_1 \cup \cdots \cup \Omega_m = \mathbb{R}^n$. Then $\phi = \max\{\phi_1, \ldots, \phi_m\}$ is a model of f.

Remark 2.6. There is another operation, which is more problematic. Suppose $f = f_1 + f_2$ where f_i has model ϕ_i. Then it would be convenient were $\phi = \phi_1 + \phi_2$ a model of f. This fails in general due to axiom (M_1). Indeed, we have $\partial_1\phi(x, x) = \partial_1\phi_1(x, x) + \partial_1\phi_2(x, x) \subset \partial f_1(x) + \partial f_2(x)$, but unfortunately, $\partial f_1(x) + \partial f_2(x) \not\subset \partial f(x)$ in general. In this situation we can modify the model concept as follows: We use $\partial^T f(x) := \partial f_1(x) + \partial f_2(x)$ in axiom (M_1) instead of the Clarke subdifferential $\partial(f_1 + f_2)(x)$. This is tolerable since every Clarke critical point satisfies also $0 \in \partial^T f(x)$, i.e., is critical in this extended sense. Toland [67, 68] was the first to use this type of subdifferential for dc-functions, i.e., $f = g - h$ with g, h convex, which is why we use the notation ∂^T.

2.2. **Strict models.** For convergence theory we will need a slightly stronger type of model, which is given by the following:

Definition 2.7. A model ϕ of f is called *strict* at $x \in \mathbb{R}^n$ if it satisfies the following strict version of axiom (M_2):

$(\widehat{M_2})$ If $x_k, y_k \to x$, then there exist $\epsilon_k \to 0^+$ such that $f(y_k) \leq \phi(y_k, x_k) + \epsilon_k \|y_k - x_k\|$.

The model ϕ is called strict if it is strict at every x. □

Remark 2.8. The difference between (M_2) and the strict version $(\widehat{M_2})$ is analogous to the difference between differentiability and strict differentiability, hence the nomenclature.

One may ask which locally Lipschitz functions f admit strict models. Since every f has its standard model $\phi^\sharp$, it is natural to ask first whether or when $\phi^\sharp$ is strict.

Recall that a locally Lipschitz function f is upper C^1 at $\bar{x}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in B(\bar{x}, \delta)$ and *every* $g \in \partial f(x)$ one has $f(y) \leq f(x) + g^\mathsf{T}(y - x) + \epsilon\|y - x\|$, where the latter could also be written as

$$f(y) \leq f(x) - f^\circ(x, x - y) + \epsilon\|y - x\|.$$

We weaken this by saying that f is weakly upper C^1 at $\bar{x}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, y \in B(\bar{x}, \delta)$ and *some* $g \in \partial f(x)$ one

has $f(y) \leq f(x) + g^{\mathsf{T}}(y-x) + \epsilon\|y-x\|$, or what is the same

$$f(y) \leq f(x) + f^{\circ}(x, y-x) + \epsilon\|x-y\|.$$

That is precisely strictness of $\phi^\sharp$, so we have (see [50]):

Proposition 2.9. *If f is weakly upper C^1 at x, then the standard model $\phi^\sharp$ of f is strict at x. In that case, every model ϕ of f is strict at x.* □

Remark 2.10. A consequence is that if f is differentiable at x, then the standard model is strict at x if and only if f is strictly differentiable at x.

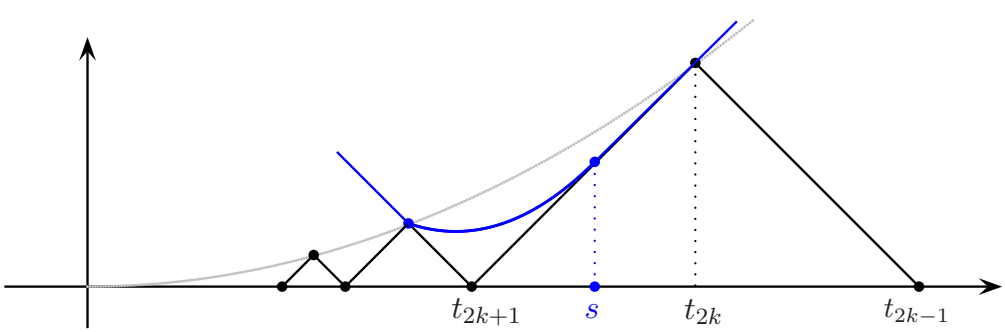
Example 2.11. Expanding on this, suppose f is differentiable but not strictly so, can it still have a strict model ϕ other than $\phi^\sharp$? For instance, can $f(x) = x^2 \sin x^{-1}$ have a model which is everywhere strict, including the origin? The answer is surprisingly 'yes'.

We consider for technical reasons a discrete version of $x^2(\sin x^{-1} + 1)$. Define $f : [-1, 1] \to \mathbb{R}^+$ as follows: start at $t_1 = 1$ with value $f(t_1) = 0$, then choose slope -1 on $[t_2, t_1]$, where $t_2 = \frac{\sqrt{5}-1}{2}$ with value $f(t_2) = t_2^2$. Next let f have slope $+1$ on $[t_3, t_2]$ with value $f(t_3) = 0$, etc. This leads to the recursions

$$t_{2k+1} = t_{2k} - t_{2k}^2, \quad t_{2k} = \frac{\sqrt{1 + 4t_{2k-1}} - 1}{2}, \quad f(t_{2k}) = t_{2k}^2, \quad f(t_{2k+1}) = 0,$$

with f piecewise linear with slope ± 1 in between the t_i. In fact, f zig-zags between the axis $y = 0$ and the parabola $y = x^2$. We extend the function symmetrically to $[-1, 0]$. Note that f is differentiable at the origin with $f'(0) = 0$, but is not strictly differentiable, because $\partial f(0) = [-1, 1]$.

We show that the standard model $\phi^\sharp$ of f is not strict. Suppose it were, then $f(t_{2k}) \leq \phi^\sharp(t_{2k}, t_{2k+1} - \delta_k) + \mathrm{o}(t_{2k} - t_{2k+1} + \delta_k)$, where $\delta_k > 0$ with $t_{2k+1} - \delta_k > t_{2k+2}$. Since on $[t_{2k+2}, t_{2k+1}]$ the slope is -1, this amounts to $t_{2k}^2 \leq f(t_{2k+1} - \delta_k) + (-1)(t_{2k} - t_{2k+1} + \delta_k) + \mathrm{o}(t_{2k} - t_{2k+1} + \delta_k) = \delta_k + t_{2k+1} - t_{2k} - \delta_k + \mathrm{o}(t_{2k} - t_{2k+1} + \delta_k) = -t_{2k}^2 + \mathrm{o}(t_{2k} - t_{2k+1} + \delta_k)$. So we would have $2t_{2k}^2 \leq \mathrm{o}(t_{2k} - t_{2k+1} + \delta_k)$, and since δ_k can be chosen arbitrarily small, we would need $t_{2k}^2 \leq \mathrm{o}(t_{2k+1} - t_{2k}) = \mathrm{o}(t_{2k}^2)$, which is wrong. Hence $\phi^\sharp$ is not strict at 0.



We now show that f admits a strict model ϕ. For peaks (t_{2k}, t_{2k}^2) let $\phi(\cdot, t_{2k}) = t_{2k}^2 + |\cdot - t_{2k}|$. Then $\partial_1\phi(t_{2k}, t_{2k}) = [-1, 1] = \partial f(t_{2k})$. The

same for every bottom points t_{2k+1}, including the origin. Hence again $\partial_1\phi(t_{2k+1}, t_{2k+1}) = [-1, 1] = \partial f(t_{2k+1})$. For intermediate points we define ϕ as follows. Let $t_{2k+1} < s < t_{2k}$ (as in the figure). Then $\phi(\cdot, s)$ consists of three arcs (shown in blue). Fit a parabola through the points (t_{2k+2}, t_{2k+2}^2) and $(s, f(s))$ such that its slope at s equals $f'(s) = 1$. To the left of t_{2k+2} extend by a line with slope -1, to the right of s by a line of slope 1. This is a continuous convex function, which is differentiable everywhere, except at t_{2k+2}. Moreover, $\phi(\cdot, s) \geq f$. For $s \in (t_{2k+2}, t_{2k+1})$ we proceed symmetrically, where now the non-differentiability occurs at t_{2k}. Note that $\phi(\cdot, s)$ depends continuously on s for $s \in (t_{2k+2}, t_{2k})$, where at $s = t_{2k+1}$ the three arc function degrades to the vee-shaped function $\phi(\cdot, t_{2k+1}) = |\cdot - t_{2k+1}|$. Only at the peaks $s = t_{2k}$ is the construction discontinuous, but it remains upper semi-continuous, so that (M_3) is satisfied at the t_{2k}. Finally, we also observe that we have upper semi-continuity at $s = 0$. This assures (M_3) also at 0. Strictness $(\widehat{M}_2)$ follows because $f \leq \phi(\cdot, s)$.

The function f is not weakly upper semi-smooth in the sense of [47, 48] at 0, nor is it upper C^1, so it is not amenable to linesearch methods. Yet the fact that it admits a strict model shows that it is amenable to non-smooth optimization techniques in our framework. □

Remark 2.12. Borwein and Moors [18, 19] construct locally Lipschitz functions where subdifferential integrability fails. Consider in particular a function $f : \mathbb{R} \to \mathbb{R}$ where $\partial f(x) = [-1, 1]$ for every x. Then the standard model of f is strict, but f is not upper C^1. See also [20].

Remark 2.13. The natural model $\phi(\cdot, x) = \psi(F(x) + F'(x)(\cdot - x))$ of a composite function $f = \psi \circ F$ with ψ convex and F of class C^1 is strict, because $F(y) = F(x) + F'(x)(y - x) + \mathrm{o}(\|y - x\|)$ as $y - x \to 0$, hence $f(y) - \phi(y, x) = \psi(F(y)) - \psi(F(x) + F'(x)(y - x)) = \mathrm{o}(\|y - x\|)$, as ψ is locally Lipschitz. In particular, every convex f is its own strict model $\phi(\cdot, x) = f$. Another consequence is that every lower C^2-function has a strict model.

However, the following is more general:

Proposition 2.14. *Every lower C^1-function admits a strict model.*

A direct proof can be found in [49]. Here we will obtain a slightly stronger result from our more general theory of cutting plane oracles in section 3.3.

Remark 2.15. We call f a dc-function if it is the difference of two convex functions, i.e., $f = g - h$ for convex g, h. Let us call f a DC-function if it admits a dc-decomposition $f = g - h$ where $\partial f(x) = \partial g(x) - \partial h(x)$. Every DC-function has the strict model $\phi(\cdot, x) = g(\cdot) + \phi^\sharp(\cdot, x)$, where $\phi^\sharp$ is the standard model of the concave function $-h$. Note that if f is dc but not DC, we can still use $\partial^T f(x) = \partial g(x) - \partial h(x)$, then ϕ is automatically a strict ∂^T-model of f. For more information on dc-functions see [35].

Remark 2.16. Consider the Euclidian distance $f(x) = \frac{1}{2}d_S^2(x)$ to an arbitrary set $S \subset \mathbb{R}^n$. It follows with [16] that $h(x) = \frac{1}{2}|x|^2 - \frac{1}{2}d_S^2(x)$ is convex, hence $-f$ is lower C^2, and therefore f is upper C^2. This means both d_S^2 and $-d_S^2$ have strict models. For f we may use the standard oracle, which leads to the steepest descent method. For $-f$ we use downshifted tangents. Note that f is also a DC-function, because $\frac{1}{2}d_S^2(x) = \frac{1}{2}|x|^2 - \big(\frac{1}{2}|x|^2 - \frac{1}{2}d_S^2(x)\big)$, so the previous remark gives yet another strict model of f.

3. Cutting plane oracles

By a cutting plane oracle we understand, loosely, any procedure $\mathscr{O}$, which associates with every serious iterate x and unsuccessful trial step (null step) z near x one – or several – affine functions $m(\cdot, x) = a + g^{\mathsf{T}}(\cdot - x)$, which replace the tangent to f at y. These cuts are then accumulated to build a convex working model $\phi_k(\cdot, x)$ of f in the neighborhood of x. In the convex cutting plane or bundle method, $\mathscr{O}(z, x)$ consists of any of the tangents to f at z, i.e. $t_{z,g}(\cdot) = f(z) + g^{\mathsf{T}}(\cdot - z)$, where $g \in \partial f(z)$, so is in fact independent of the serious iterate x. In the non-convex case we may no longer proceed in this way, because tangents to f at z may pass above $f(x)$, and cannot be used directly as cutting planes.

Let $\mathscr{O} : \mathbb{R}^n \times \mathbb{R}^n \rightrightarrows \mathbb{R} \times \mathbb{R}^n$ be a set-valued operator mapping into the nonempty bounded subsets of $\mathbb{R} \times \mathbb{R}^n$, where $(a, g) \in \mathscr{O}(z, x)$ is understood as to represent the affine function $a + g^{\mathsf{T}}(\cdot - x)$. We call $\mathscr{O}$ a *cutting plane oracle* for f if it is bounded on bounded sets and satisfies the following axioms:

- ($\mathscr{O}_1$) If $(a, g) \in \mathscr{O}(z, x)$, then $a \leq f(x)$. Moreover, $\mathscr{O}(x, x)$ contains at least one element $(f(x), g)$ with $g \in \partial f(x)$.
- ($\mathscr{O}_2$) If $(a_j, g_j) \in \mathscr{O}(z_j, x)$ with $a_j \to f(x)$, $\|z_j - x\| \leq M$, and $g_j \to g$, then $g \in \partial f(x)$.
- ($\mathscr{O}_3$) If $z_j \to x$ there exist $(a_j, g_j) \in \mathscr{O}(z_j, x)$ and $\epsilon_j \to 0^+$ such that $f(z_j) \leq a_j + g_j^{\mathsf{T}}(z_j - x) + \epsilon_j \|z_j - x\|$.
- ($\mathscr{O}_4$) If $z_j \to z$, $y_j \to y$ and $x_j \to x$, and if $(a_j, g_j) \in \mathscr{O}(z_j, x_j)$, then there exists $z' \in B(x, M)$ and $(a', g') \in \mathscr{O}(z', x)$ such that $\limsup_{j\to\infty} a_j + g_j^{\mathsf{T}}(y_j - x_j) \leq a' + g'^{\mathsf{T}}(y - x)$.

For the $M > 0$ occurring in ($\mathscr{O}_2$), ($\mathscr{O}_4$) we define the envelope function of $\mathscr{O}$ by

$$\phi(\cdot, x) = \sup\{a + g^{\mathsf{T}}(\cdot - x) : (a, g) \in \mathscr{O}(z, x), \|z - x\| \leq M\}. \tag{3.1}$$

Proposition 3.1. *Suppose $\mathscr{O}$ satisfies axioms ($\mathscr{O}_1$) – ($\mathscr{O}_4$). Then the envelope function ϕ of $\mathscr{O}$ is a model of f.*

Proof. 1) Observe that $\mathscr{O}(x, x) \neq \emptyset$ by ($\mathscr{O}_1$). Pick $(a, g) \in \mathscr{O}(x, x)$, then $\phi(y, x) \geq a + g^{\mathsf{T}}(y - x) > -\infty$. This means $\phi(\cdot, x)$ is by construction convex and maps into $\mathbb{R} \cup \{\infty\}$. But $\phi(y, x) = \infty$ is impossible, because that would require a sequence $z_j \in B(x, M)$ and $(a_j, g_j) \in \mathscr{O}(z_j, x)$ such

that $a_j + g_j^{\mathsf{T}}(y-x) \to \infty$. Then (a_j, g_j) would have to be unbounded, contradicting the fact that $\mathscr{O}$ is by definition bounded on the bounded set $B(x,M)\times\{x\}$. This means $\phi(\cdot,x)$ is a convex function which is everywhere defined.

2) We have to check (M_1). Now by $(\mathscr{O}_1)$ we have $a \le f(x)$ for every $(a,g)\in\mathscr{O}(z,x)$, hence $\phi(x,x)\le f(x)$. But there exists $(f(x),g)\in\mathscr{O}(x,x)$, hence $\phi(x,x)\ge f(x)$, giving equality. It remains to check that every subgradient g of $\phi(\cdot,x)$ at x belongs to $\partial f(x)$.

Let g be a subgradient of $\phi(\cdot,x)$ at x, then $t(\cdot) := f(x) + g^{\mathsf{T}}(\cdot - x)$ is a tangent to $\phi(\cdot,x)$ at x, and by convexity $t(\cdot)\le\phi(\cdot,x)$. Now fix h and consider the convex function $t\mapsto\phi(x+th,x)$ on the real line. Its slope at $t=0$ is greater or equal than $g^{\mathsf{T}}h$. Now for every $t>0$ find $z_t\in B(x,M)$ and a cutting plane $(a_t,g_t)\in\mathscr{O}(z_t,x)$, represented as $m_t(\cdot,x)=a_t+g_t^{\mathsf{T}}(\cdot-x)$, such that $\phi(x+th,x)\ge m_t(x+th,x)$ with equality $\phi(x+th,x)=m_t(x+th,x)$. Observe that this implies $a_t\to f(x)$.

Now the slope of $\phi(\cdot,x)$ at $x+th$ is steeper than its slope at x (monotonicity), so $g^{\mathsf{T}}h\le g_t^{\mathsf{T}}h$. Choosing a subsequence $g_t\to\hat g$, $z_t\to z$, we get $\hat g\in\partial f(x)$ by axiom $(\mathscr{O}_2)$. That shows $g^{\top}h\le\hat g^{\mathsf{T}}h\le\max\{\tilde g^{\mathsf{T}}h:\tilde g\in\partial f(x)\}=f^{\circ}(x,h)$. But since h was arbitrary, we deduce $g\in\partial f(x)$ by Hahn-Banach. This completes the proof of (M_1).

3) Since (M_2) is clear from $(\mathscr{O}_3)$, it remains to check (M_3). Fix $y_k\to y$, $x_k\to x$, and let $z_k\in B(x_k,M)$ and $\epsilon_k\to 0^+$ such that $m_{z_k,g_k}(y_k,x_k) = a_k + g_k^{\mathsf{T}}(y_k-x_k)=\phi(y_k,x_k)$, where $(a_k,g_k)\in\mathscr{O}(z_k,x_k)$. Invoking $(\mathscr{O}_4)$, we get $z'\in B(x,M)$ and $(a',g')\in\mathscr{O}(z',x)$ such that $\limsup_{k\to\infty} a_k + g_k^{\mathsf{T}}(y_k-x_k)\le a' + g'^{\mathsf{T}}(y-x)\le\phi(y,x)$, and with the above this implies $\limsup_{k\to\infty}\phi(y_k,x_k)\le\phi(y,x)$. □

Definition 3.2. A cutting plane oracle $\mathscr{O}$ is called *strict* at x if the following strict version of $(\mathscr{O}_3)$ is satisfied:

- $(\widehat{\mathscr{O}}_3)$ Given $x_j, z_j\to x$ there exist $(a_j,g_j)\in\mathscr{O}(z_j,x_j)$ and $\epsilon_j\to 0^+$ such that $f(z_j)\le a_j + g_j^{\mathsf{T}}(z_j-x_j)+\epsilon_j\|z_j-x_j\|$.

□

3.1. **Model-based cutting plane oracles.** A natural way to generate cutting planes is when a model ϕ of f is available for computation. Let x be the serious iterate, z an unsuccessful trial step (a null step) at which we want to generate a cutting plane. We simply take an affine support function $m_{z,g}(\cdot,x)$ of $\phi(\cdot,x)$ at z, i.e. $m_{z,g}=\phi(z,x)+g^{\mathsf{T}}(\cdot-z)$ with $g\in\partial_1\phi(z,x)$. We use the notation $\mathscr{O}_\phi$ for the oracle generated by a model ϕ. If $\phi^\sharp$ is the standard model of f, then we write $\mathscr{O}^\sharp := \mathscr{O}_{\phi^\sharp}$, calling it the *standard oracle*.

Proposition 3.3. *Every locally Lipschitz function f admits a cutting plane oracle. It admits a strict cutting plane oracle if and only if it admits a strict model.*

Proof. The first part is clear, because every f has at least one model, $\phi^\sharp$, and every model ϕ generates an oracle $\mathscr{O}_\phi$. Note that if ϕ is strict, then so is

$\mathscr{O}_\phi$. Conversely, if $\mathscr{O}$ is strict, then its upper envelope (3.1) is a strict model, as $(\widehat{\mathscr{O}}_3)$ is easily seen to ensure $(\widehat{M}_2)$. □

3.2. **Cutting planes from downshifted tangents.** One prominent way to generate cutting planes is the downshifted tangent oracle $\mathscr{O}^{\downarrow}$, which we now discuss. Recall that for $g \in \partial f(z)$ the affine function $t_{z,g}(\cdot) = f(z) + g^{\mathsf{T}}(\cdot - z)$ is a tangent to f at z, or simply a tangent plane.

Definition 3.4. Let x be the current serious iterate, z a trial step. Let $t_{z,g}(\cdot)$ be a tangent of f at z. For a fixed constant $c > 0$ we define the *downshifted tangent* $m^{\downarrow}_{z,g}(\cdot, x)$ at serious iterate x and trial step z associated with the tangent $t_{z,g}(\cdot)$ as follows:

$$m^{\downarrow}_{z,g}(\cdot, x) = t_{z,g}(\cdot) - \left[t_{z,g}(x) - f(x) + c\|z - x\|^2\right]_+ . \tag{3.2}$$

The quantity appearing on the right

$$s(z, x, g) = \left[t_{z,g}(x) - f(x) + c\|z - x\|^2\right]_+ \tag{3.3}$$

is called the *downshift* of the tangent. The oracle is denoted $\mathscr{O}^{\downarrow}$. □

Remark 3.5. The explanation is that as long as the tangent $t_{z,g}(\cdot)$ passes below $f(x) - c\|z - x\|^2$ at x, we can use it directly as cutting plane. However, if $t_{z,g}(x) > f(x) - c\|z - x\|^2$, where $c > 0$ is the same fixed parameter, then we have to shift the tangent down to obtain the cutting plane $m^{\downarrow}_{z,g}(\cdot, x)$, so that now $m^{\downarrow}_{z,g}(x, x) = f(x) - c\|z - x\|^2$, as otherwise the oracle would not respect axiom $(\mathscr{O}_1)$. The term $c\|z - x\|^2$ is chosen for convenience. Any function $c(\cdot) : \mathbb{R} \to \mathbb{R}_+$ with $c(t) = \mathrm{o}(t)$ as $t \to 0$ would give a similar oracle with downshift $s = [t_{z,g}(x) - f(x) - c(\|z - x\|)]_+$.

Remark 3.6. If $z = x$, then every tangent $t_{x,g}(\cdot)$ at x is also a cutting plane $m^{\downarrow}_{x,g}(\cdot, x)$ at serious iterate x and trial step $z = x$, because $g \in \partial f(x)$ and because the downshift (3.3) at $z = x$ is automatically zero.

Definition 3.7. A cutting plane $m^{\downarrow}_{x,g}(\cdot, x)$ at serious iterate x and trial step $z = x$ is called an *exactness plane*. The same terminology is used for any other oracle $\mathscr{O}$. □

3.3. **Strictness of the downshift oracle $\mathscr{O}^{\downarrow}$.** Recall that a locally Lipschitz function f is lower C^1 at x if for $x_j, y_j \to x$ and *every* $g_j \in \partial f(y_j)$ there exist $\epsilon_j \to 0^+$ such that $f(y_j) \le f(x_j) + g_j^{\mathsf{T}}(y_j - x_j) + \epsilon_j\|y_j - x_j\|$, [64].

Proposition 3.8. *The downshift operator $\mathscr{O}^{\downarrow}$ satisfies axioms $(\mathscr{O}_1)$, $(\mathscr{O}_2)$ and $(\mathscr{O}_4)$. If f is Clarke regular at x, then $\mathscr{O}^{\downarrow}$ satisfies $(\mathscr{O}_3)$ at x. Suppose f is in addition lower C^1 at x. Then the downshift operator is strict at x, i.e., satisfied $(\widehat{\mathscr{O}}_3)$ at x.*

Proof. 1) Axiom $(\mathscr{O}_1)$ is clear. We check $(\mathscr{O}_2)$. Let $(a_j, g_j) \in \mathscr{O}^{\downarrow}(z_j, x)$ with $a_j \to f(x)$, $g_j \to g$, and z_j bounded. Passing to a subsequence, assume $z_j \to z$. Now $m^{\downarrow}_j(\cdot, x) = a_j + g_j^{\mathsf{T}}(\cdot - x) = f(z_j) + g_j^{\mathsf{T}}(\cdot - z_j) - s_j$, where

$s_j = s(x, z_j, g)$ is the corresponding downshift (3.3). From $a_j \to f(x)$ we get $f(z_j) + g_j^\mathsf{T}(x - z_j) - s_j \to f(x) = f(z) + g^\mathsf{T}(x - z) - s$, where $s = \lim s_j$ and $g \in \partial f(z)$. There are two cases, either $s_j \to 0$, or $s_j \to s > 0$. In the first case $s = 0$ we obtain

$$s_j = \Big[f(z_j) + g_j^\mathsf{T}(x - z_j) - f(x) + c\|z_j - x\|^2\Big]_+ \to 0. \tag{3.4}$$

But since $f(z_j) + g_j^\mathsf{T}(x - z_j) - f(x) \to 0$ due to $s = 0$, we must also have $c\|z_j - x\|^2 \to 0$, proving $z_j \to x$. Hence $z = x$, and so $g \in \partial f(x)$. Now in the case $s > 0$ we get $s_j > 0$ from some counter onward, hence $f(z_j) + g_j^\mathsf{T}(x - z_j) - f(z_j) - g_j^\mathsf{T}(x - z_j) + f(x) - c\|z_j - x\|^2 \to f(x) = f(z) + g^\mathsf{T}(x - z) - f(z) - g^\mathsf{T}(x - z) + f(x) - c\|z - x\|^2$, which implies $f(x) + c\|z_j - x\|^2 \to f(x)$, hence again $z_j \to x$, leading to the same conclusion.

2) Let us check ($\mathscr{O}_4$). This is obviously the analogue of the model axiom (M_3). Fix $z_j \to z$, $y_j \to y$ and $x_j \to x$, and $(a_j, g_j) \in \mathscr{O}^\downarrow(z_j, x_j)$. Then $m_j^\downarrow(\cdot, x_j) = a_j + g_j^\mathsf{T}(\cdot - x_j) = f(z_j) + g_j^\mathsf{T}(\cdot - z_j) - s_j$ is the downshifted tangent at z_j, i.e.

$$s_j = \Big[f(z_j) + g_j^\mathsf{T}(x_j - z_j) - f(x) + c\|z_j - x_j\|^2\Big]_+ . \tag{3.5}$$

Choose a subsequence $j \in J$ such that $\ell := \lim\limits_{j \in J} m_j^\downarrow(y_j, x_j) = \limsup\limits_{j \to \infty} m_j^\downarrow(y_j, x_j)$. Then this limit is $\ell = f(z) + g^\mathsf{T}(y - x) - s$, where $s = \lim_{j \in J} s_j$. We have to find z' such that $\ell \le a' + g'^\mathsf{T}(y - x)$ for $(a', g') \in \mathscr{O}(z', x)$. We simply choose $z' = z$ and $g' = g$, because then $f(z) + g^\mathsf{T}(\cdot - z) - s$ should be the downshifted tangent at z. For that to be confirmed, we have just to show that $s \le s(z, x, g)$, in other words, we have to show upper semi-continuity $\lim_{j \in J} s(z_j, x_j, g_j) \le s(z, x, g)$ of the downshift. But that follows immediately here due to convergence of all the elements in (3.5).

3) We have to verify axiom ($\mathscr{O}_3$) at x with f regular at x. Consider a sequence $y^k \to x$. Let $m_k(\cdot, x) := m^\downarrow_{y^k, g_k}(\cdot, x)$ be a cutting plane at serious iterate x and trial point y^k, where $g_k \in \partial f(y^k)$. We have to find $\epsilon_k \to 0^+$ such that $f(y^k) \le m_k(y^k, x) + \epsilon_k\|y^k - x\|$. Now let $t_{y^k, g_k}(\cdot)$ be the tangent to f from which m_k is downshifted, and let $s_k = s(y^k, x, g_k)$ be the downshift. Consider the case $s_k > 0$. Then $m_k(\cdot, x) = f(x) + g_k^\mathsf{T}(\cdot - x) - c\|y^k - x\|^2$, and we have to find $\epsilon_k \to 0^+$ such that

$$f(y^k) \le f(x) + g_k^\mathsf{T}(y^k - x) - c\|y^k - x\|^2 + \epsilon_k\|y^k - x\|.$$

If we put $h_k = (y^k - x)/\|y^k - x\|$, $t_k = \|y^k - x\|$, then this may be re-written as

$$\frac{f(x + t_k h_k) - f(x)}{t_k} \le g_k^\mathsf{T} h_k + \epsilon_k. \tag{3.6}$$

Passing to a subsequence, we may assume $h_k \to h$ for some $\|h\| = 1$, and then since f is locally Lipschitz, the left hand term converges to the Dini derivative $f'(x, h)$. On the right hand side, on the other hand, we have

$\limsup_{k\to\infty} g_k^\mathsf{T} h_k \le f^\circ(x,h)$, and by regularity, the two limits coincide. In other words, we find that $\epsilon_k = \max\{0, t_k^{-1}(f(x+t_kh_k)-f(x)) - g_k^\mathsf{T} h_k\} \to 0^+$ does the job. That proves $(\mathscr{O}_3)$.

4) It remains to verify $(\widehat{\mathscr{O}}_3)$ when f is lower C^1 at x. Fix $z_j, x_j \to x$ and let $(a_j, g_j) \in \mathscr{O}^\downarrow(z_j, x_j)$. Then by hypothesis there exist $\epsilon_j \to 0^+$ such that $f(z_j) \le f(x_j) + g_j^\mathsf{T}(z_j - x_j) + \epsilon_j\|z_j - x_j\|$. Adding $g_j^\mathsf{T}(\cdot - z_j)$ on both sides gives $t_{z_j,g_j}(\cdot) \le f(x_j) + g_j^\mathsf{T}(\cdot - x_j) + \epsilon_j\|z_j - x_j\|$. Hence the downshift satisfies $s_j = \left[t_{z_j,g_j}(x_j) - f(x_j) + c\|x_j - z_j\|^2\right]_+ \le \epsilon_j\|z_j - x_j\| + c\|x_j - z_j\|^2 =: \widetilde{\epsilon}_j(\|z_j - x_j\|)$ with $\widetilde{\epsilon}_j = \epsilon_j + c\|z_j - x_j\| \to 0^+$. Then $f(z_j) = t_{z_j,g_j}(z_j) = m^\downarrow_{z_j,g_j}(z_j, x_j) + s_j \le m^\downarrow_{z_j,g_j}(z_j, x_j) + \widetilde{\epsilon}_j\|z_j - x_j\|$ shows strictness. □

A different way to say that f is lower C^1 at $\bar{x}$ is that for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, z \in B(\bar{x}, \delta)$ and *every* $g \in \partial f(z)$ one has $f(x) - f(z) \ge g^\mathsf{T}(x - z) - \epsilon\|x - z\|$, which may also be written as

$$f(x) - f(z) \ge f^\circ(z, x - z) - \epsilon\|x - z\|.$$

We weaken this as follows: We say that f is weakly lower C^1 at $\bar{x}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for all $x, z \in B(\bar{x}, \delta)$ and *some* $g \in \partial f(z)$ we have $f(x) - f(z) \ge g^\mathsf{T}(x - z) - \epsilon\|x - z\|$, which may be written as

$$f(x) - f(z) \ge -f^\circ(z, z - x) - \epsilon\|x - z\|. \tag{3.7}$$

There is an oracle associated with this property, which is slightly stronger than $\mathscr{O}^\downarrow$, and which we denote $\mathscr{O}^{\downarrow\downarrow}$. For $\mathscr{O}^{\downarrow\downarrow}$ we do not take an arbitrary tangent to f at z, but the specific one $t_{z,g'}(\cdot)$ where $g' \in \partial f(z)$ satisfies $g'^\mathsf{T}(x - z) = -f^\circ(z, z - x)$. Then we downshift $t_{z,g'}$ as before to obtain the $\mathscr{O}^{\downarrow\downarrow}$-cutting plane.

Proposition 3.9. *Suppose f is weakly lower C^1 at x. Then $\mathscr{O}^{\downarrow\downarrow}$ is strict at x. In consequence, every weakly lower C^1 function has a strict model.*

Proof. We have to check $(\widehat{\mathscr{O}}_3)$. Let $z_j, x_j \to x$, then $(a_j, g'_j) = \mathscr{O}^{\downarrow\downarrow}(z_j, x_j)$ satisfies $g_j'^\mathsf{T}(x_j - z_j) = \inf\{g^\mathsf{T}(x_j - z_j) : g \in \partial f(z_j)\}$. By (3.7) we have $f(x_j) - f(z_j) \ge g_j'^\mathsf{T}(x_j - z_j) - \epsilon_j\|x_j - z_j\|$ for certain $\epsilon_j \to 0^+$. That can be written $f(x_j) \ge t_{z_j,g'_j}(x_j) - \epsilon_j\|z_j - x_j\|$, hence $\epsilon_j\|z_j - x_j\| + c\|z_j - x_j\|^2 \ge t_{z_jg'_j}(x_j) - f(x_j) + c\|z_j - x_j\|^2$. This shows that the downshift (3.3) is $s_j \le \widetilde{\epsilon}_j\|z_j - x_j\|$, where $\widetilde{\epsilon}_j = \epsilon_j + c\|z_j - x_j\| \to 0^+$. Hence with the argument of part 4) above, $\mathscr{O}^{\downarrow\downarrow}$ is strict at x. □

Proposition 3.10. *Suppose the function f is weakly upper C^1 or weakly lower C^1 at every x. Then f admits a strict cutting plane oracle, and consequently has a strict model.*

Proof. We can use the idea of remark 2.5. The standard oracle $\mathscr{O}^\sharp$ is strict at those x where f is weakly upper C^1, and $\mathscr{O}^{\downarrow\downarrow}$ is strict at those x where f is weakly lower C^1, so the oracle $\mathscr{O}^{\sharp\downarrow\downarrow}(z, x)$ which among the two possible planes takes the one which has the larger value at z will be strict. Naturally,

one could also take the oracle $\mathscr{O}^\sharp \vee \mathscr{O}^{\downarrow\downarrow}$ which takes both planes. The corresponding envelope models are then strict. $\square$

Example 3.11. For f in example 2.11 neither $\mathscr{O}^\downarrow$ nor $\mathscr{O}^{\downarrow\downarrow}$ is strict at 0. With the same notation, for $s_k \in (t_{2k}, t_{2k-1})$ the slope of the tangent at s_k is -1, so the cutting plane $\mathscr{O}^\downarrow(s_k, 0)$ is the line with slope -1 passing through the point $(0, -cs_k^2)$, i.e., $m^\downarrow(s_k, 0) = -cs_k^2 - s_k$. For $(\mathscr{O}_3)$ at 0 we would require $f(s_k) \le m^\downarrow(s_k, 0) + \mathrm{o}(s_k)$, hence $t_{2k}^2 - (s_k - t_{2k}) \le -cs_k^2 - s_k + \mathrm{o}(s_k) = -s_k + \mathrm{o}(s_k)$. Choose s_k such that $s_k - t_{2k} = t_{2k}^2/2$, then $s_k = t_{2k} + t_{2k}^2/2$, so $\mathrm{o}(s_k) = \mathrm{o}(t_{2k})$, and we have to assure $t_{2k}^2/2 \le -t_{2k} - t_{2k}^2/2 + \mathrm{o}(t_{2k})$, and that is impossible, as the right hand side is asymptotically < 0. Here the downshift operator $\mathscr{O}^\downarrow$ is not even an oracle at 0, let alone a strict one. A similar argument applies to $f(x) = x^2 \sin x^{-1}$.

Remark 3.12. Downshift is used in the early versions of the bundle technique to deal heuristically with non-convex cases. See e.g. the codes N1CV2 and N2BN1 by Lemaréchal and Sagastizábal [43], or in Zowe's BT package [62]. Mifflin [48] justifies the downshift oracle theoretically for a linesearch algorithm. For our justification of downshift in the bundle method see e.g. [49, 29].

3.4. **Oracles with infinitely many cuts.** We consider an eigenvalue optimization problem

$$\min\{\lambda_1\left(F(x)\right) : x \in C\},$$

where $F : \mathbb{R}^n \to \mathbb{S}^m$ is a C^1-mapping into the space $\mathbb{S}^m$ of $m \times m$ symmetric or Hermitian matrices, and $\lambda_1 : \mathbb{S}^m \to \mathbb{R}$ is the maximum eigenvalue. Suppose we use the natural model $\phi(\cdot, x) = \lambda_1\left(F(x) + F'(x)(\cdot - x)\right)$ of $f = \lambda_1 \circ F$. Let $\mathscr{G} = \{G \in \mathbb{S}^m : G \succeq 0, \mathrm{Tr}(G) = 1\}$, then $\lambda_1(X) = \max\{G \bullet X : G \in \mathcal{G}\}$, where $G \bullet X = \mathrm{Tr}(GX)$ is the scalar product in $\mathbb{S}^m$. Let z^k be a null step and suppose the multiplicity of $\phi(z^k, x) = \lambda_1\left(F(x) + F'(x)(z^k - x)\right)$ is $r > 1$. Let Q_k be a $m \times r$-matrix whose r columns form an orthonormal basis of the maximum eigenspace of $F(x) + F'(x)(z^k - x)$. Put $\mathcal{G}_k = \{Q_k^\mathsf{T} Y Q_k : Y \in \mathbb{S}^r, Y \succeq 0, \mathrm{Tr}(Y) = 1\}$, then $\mathcal{G}_k \subset \mathcal{G}$ and

$$\begin{aligned}
&\max\left\{G' \bullet \left[F(x) + F'(x)(\cdot - x)\right] : G' \in \mathcal{G}_k\right\} \\
&\quad = \max\left\{Y \bullet \left[Q_k(F(x) + F'(x)(\cdot - x)\right] Q_k^\mathsf{T} : Y \succeq 0, \mathrm{Tr}(Y) = 1\right\} \\
&\quad = \lambda_1\left(Q_k\left[F(x) + F'(x)(\cdot - x)\right] Q_k^\mathsf{T}\right).
\end{aligned}$$

This means we choose as oracle the infinite set $\mathscr{O}^{\mathrm{spec}}(z^k, x) = \{(a(Y), g(Y)) : Y \succeq 0, \mathrm{Tr}(Y) = 1\}$, where $a(Y) = Y \bullet Q_k F(x) Q_k^\mathsf{T}$ and $g(Y) = F'(x)^* Q_k^\mathsf{T} Y Q_k$, with $Q_k^\mathsf{T} Y Q_k \in \mathbb{S}^m$ and $F'(x)^* : \mathbb{S}^m \to \mathbb{R}^n$ the adjoint of $F'(x) : \mathbb{R}^n \to \mathbb{S}^m$. For practical aspects of this type of oracle, which leads to spectral bundle methods, see [11, 12, 31, 32, 33, 42, 3, 4, 15, 65, 66]. The key observation is that the tangent program (4.1) in this approach will be a convex SDP.

3.5. **Working models.** In our trust-region method the tangent program is based on a *working model* $\phi_k(\cdot, x)$ of f at serious iterate x. In the bundle literature originating from Lemaréchal's [40, 41], this model is traditionally denoted as $\check{f}_k$ and referred to as a model of f at x. In our approach we distinguish between model ϕ and working model ϕ_k from a reason which will become clear shortly.

Let $\mathscr{O}$ be a cutting plane oracle for f. Then the working model at serious iterate x and inner loop counter k has the form $\phi_k(\cdot, x) = \sup\{a + g^\mathsf{T}(\cdot - x) : (a, g) \in \mathscr{W}_k\}$, where the sets $\mathscr{W}_k$ are generated recursively through $\mathscr{O}$: Suppose z^k is an unsuccessful trial step (a null step) obtained via (4.2) from the solution y^k of the tangent program (4.1) based on the kth working model $\phi_k(\cdot, x)$. Then the $(k+1)^{\text{st}}$ working model $\phi_{k+1}(\cdot, x) = \sup\{a + g^\mathsf{T}(\cdot - x) : (a, g) \in \mathscr{W}_{k+1}\}$ is obtained from ϕ_k by the following rules:

- (W_1) At least one exactness plane $f(x) + g^\mathsf{T}(\cdot - x)$ with $(f(x), g) \in \mathscr{O}(x, x)$ is included in ϕ_{k+1}. That is, $(f(x), g) \in \mathscr{W}_{k+1}$ for some $(f(x), g) \in \mathscr{O}(x, x)$.
- (W_2) The aggregate plane $a_k^* + g_k^{*\mathsf{T}}(\cdot - x)$ associated with the solution y^k of the kth-tangent program (4.1) is included in ϕ_{k+1}, i.e., $(a_k^*, g_k^*) \in \mathscr{W}_{k+1}$. (See section 4.1).
- (W_3) All cutting planes $(a, g) \in \mathscr{O}(z^k, x)$ are simultaneously included in ϕ_{k+1}, that is, $\mathscr{O}(z^k, x) \subset \mathscr{W}_{k+1}$.
- (W_4) Planes contributing to $\phi_{k+1}(\cdot, x)$ via $\mathscr{W}_{k+1}$ other than those in $(W_1)-(W_3)$ must already have been elements of $\mathscr{W}_k$, but not all $(a, g) \in \mathscr{W}_k$ are kept in $\mathscr{W}_{k+1}$.

The initialization of $\phi_1(\cdot, x)$ is as follows. We request that $\phi_1(\cdot, x)$ contains at least one exactness plane generated by $\mathscr{O}$, i.e., $f(x) + g^\mathsf{T}(\cdot - x) \le \phi_1(\cdot, x)$ for some $(f(x), g) \in \mathscr{O}(x, x)$. In other words, $(f(x), g) \in \mathscr{W}_1 \subset \mathscr{O}(x, x)$.

If in addition a positive semi-definite symmetric matrix $Q(x) \succeq 0$ is available as a substitute for the Hessian of f at x, then we call $\Phi_k(\cdot, x) = \phi_k(\cdot, x) + \frac{1}{2}(\cdot - x)^\mathsf{T} Q(x)(\cdot - x)$ a *second-order working model* of f at serious iterate x. We shall occasionally use the semi-norm $|y|_Q^2 = y^\mathsf{T} Q y$, so that $\Phi_k(\cdot, x) = \phi_k(\cdot, x) + \frac{1}{2}|\cdot - x|_Q^2$.

Remark 3.13. By construction we have $\phi_k(\cdot, x) \le \phi(\cdot, x)$ at all inner loop instants k, where ϕ is the envelope model (3.1) of $\mathscr{O}$. By convexity we automatically have $\partial_1 \phi_k(x, x) \subset \partial_1 \phi(x, x) \subset \partial f(x)$, so that $0 \in \partial(\phi_k(\cdot, x) + i_C)(x)$ implies that x is a Clarke critical point of (1.1). This is crucial for practice, as we can stop the algorithm as soon as the tangent program based on the working model $\phi_k(\cdot, x)$ finds no model reduction step.

Remark 3.14. Note that we do not necessarily have $\mathscr{W}_k \subset \mathscr{W}_{k+1}$, as this would lead to tangent programs of increasing size. Rules $(W_1) - (W_4)$ only require that: a) An exactness plane assures $\phi_{k+1}(x, x) = f(x)$; b) the aggregate plane (a_k^*, g_k^*) assures $\phi_{k+1}(y^k, x) \ge \phi_k(y^k, x)$, and c) the cutting planes $(a_k, g_k) \in \mathscr{O}(z^k, x)$ are intended to bring the value $\phi_{k+1}(z^k, x)$ as close as

possible to the unduly large value $f(z^k)$, because that value was much larger than the predicted value $\Phi_k(z^k, x)$, having caused the failure of z^k.

Example 3.15. Consider $f(x) = \max_{i\in I} f_i(x)$, where I is infinite and the f_i are class C^1. The natural model is $\phi(\cdot, x) = \max_{i\in I} f_i(x) + \nabla f_i(x)^\mathsf{T}(\cdot - x)$, but computing it might be costly for large size I, and we might prefer a working model $\phi_k(\cdot, x) = \max_{i\in I_k} f_i(x) + \nabla f_i(x)^\mathsf{T}(\cdot - x)$ with some small subset I_k of I. In H_∞-optimization [5, 6, 11, 12] we have observed that diligent choices of an initial set $I_1 \subset I$ may greatly influence the performance of the method, see also [22, 21]. This is encouraged by rule (W_4).

4. Non-smooth trust-region method

In this section we present the main elements of the algorithm.

4.1. **Tangent program, aggregate plane and trial steps.** Once a second-order working model $\Phi_k(\cdot, x) = \phi_k(\cdot, x) + \frac{1}{2}(\cdot - x)^\mathsf{T} Q(x)(\cdot - x)$ at serious iterate x and inner loop counter k is fixed, we solve the trust-region tangent program

$$\begin{array}{ll} \text{minimize} & \Phi_k(y, x) \\ \text{subject to} & y \in C \\ & \|y - x\| \le R_k, \end{array} \tag{4.1}$$

where $R_k > 0$ is the current trust-region radius. We obtain a solution $y^k \in C$ of (4.1), which is unique in case $Q(x) \succ 0$. For a polyhedral norm like $\|\cdot\| = |\cdot|_\infty$ or $\|\cdot\| = |\cdot|_1$ and a polyhedron C, program (4.1) reduces to a convex quadratic program, as long as the first-order working models are polyhedral, or even an LP if $Q(x) = 0$. In the spectral bundle method [11, 12], where ϕ_k contains an infinity of cuts, (see section 3.4), the tangent program is a convex SDP, and C may then be allowed to be an SDP-constrained set. In this case the choice $\|\cdot\| = |\cdot|_2$ is acceptable, as the trust-region can be turned into a conical constraint, so that the tangent program is a convex SDP.

Note that by the necessary optimality conditions for the tangent program in step 4 of the algorithm, there exists a subgradient $g_k^* \in \partial\left(\phi_k(\cdot, x) + i_C\right)(y^k)$ such that $g_k^* + Q(y^k - x) + v_k = 0$, where v_k is in the normal cone to the trust-region norm ball $B(x, R_k)$ at y^k, and where i_C is the indicator function of the convex set C.

Definition 4.1. We call g_k^* the *aggregate subgradient*. The affine function $m_k^*(\cdot, x) = \phi_k(y^k, x) + g_k^{*\mathsf{T}}(\cdot - y^k)$ is called the *aggregate plane*. □

Now along with the solution y^k of the tangent program we consider a larger set of admissible trial steps z^k. Fixing constants $\Theta \ge 1$ and $0 < \theta \ll 1$, we admit as trial point any $z^k \in C$ satisfying

$$f(x) - \Phi_k(z^k, x) \ge \theta\left(f(x) - \Phi_k(y^k, x)\right) \text{ and } \|z^k - x\| \le \Theta\|y^k - x\|. \tag{4.2}$$

4.2. **Acceptance and building the new working model.** Once a trial point z^k associated with a solution y^k of the tangent program (4.1) has been determined as in (4.2), acceptance is tested by the standard test

$$\rho_k = \frac{f(x) - f(z^k)}{f(x) - \Phi_k(z^k, x)} \overset{?}{\geq} \gamma \tag{4.3}$$

for fixed $0 < \gamma \ll 1$. If $\rho_k < \gamma$ the trial point z^k is rejected (a null step). Then we have to keep the inner loop going. Applying rules $(W_1) - (W_4)$ we generate cutting planes $(a, g) \in \mathscr{O}(z^k, x)$ at z^k, the aggregate plane (a^*, g^*) at y^k, and add those into the new $\phi_{k+1}(\cdot, x)$, tapering out the old $\phi_k(\cdot, x)$ by dropping some of the older cuts contributing to the aggregate plane. We also keep at least one exactness plane $(a_0, g_0) \in \mathscr{O}(x, x)$ in the model $\phi_{k+1}(\cdot, x)$, which may, or may not be, the one used at counter k.

4.3. **Management of the trust-region radius.** Understanding step 7 of the algorithm is crucial, because here a main difference with the classical trust-region method occurs. Namely, in the case of an unsuccessful trial step z^k in the inner loop at x we do *not* automatically reduce the trust-region radius. We first call the oracle to provide cutting planes at z^k. Then the secondary test in step 7

$$\widetilde{\rho}_k = \frac{f(x) - \phi_{k+1}(z^k, x)}{f(x) - \Phi_k(z^k, x)} \overset{?}{\geq} \widetilde{\gamma} \tag{4.4}$$

(with $0 < \gamma < \widetilde{\gamma} < 1$) serves to decide whether or not to reduce R_k at the next sweep $k + 1$ of the inner loop. The rationale of this test is as follows: If $\widetilde{\rho}_k < \widetilde{\gamma}$, then the effect of the cutting plane(s) at z^k is sensible, i.e., it is reasonable to believe that we could have performed better, had we already included this cutting plane in the working model ϕ_k. In that case we keep enriching the working model by cuts and maintain R_k unchanged, being reluctant to reduce R_k prematurely, as this leads to unnecessarily small steps. In the opposite case $\widetilde{\rho}_k \geq \widetilde{\gamma}$ the new cutting plane(s) do(es) not seem to contribute anything substantial at z^k, and here we reduce the trust-region radius in order to get closer to the current x, where progress is ultimately possible (due to $0 \notin \partial f(x) + N_C(x)$).

Remark 4.2. The secondary test appears for the first time in the first nonconvex version of the bundle method [52], see also [11, 12]. It can also be used with $\Phi_{k+1}(z^k, x)$ instead of $\phi_{k+1}(z^k, x)$ in the numerator.

4.4. **Algorithm.** We are now ready to present the bundle trust-region algorithm. (See algorithm 1).

Algorithm 1. Non-smooth trust-region algorithm

Parameters: $0 < \gamma < \widetilde{\gamma} < 1$, $0 < \gamma < \Gamma \leq 1$, $0 < \theta \ll 1$, $\Theta \geq 1$, $q > 0$.

▷ **Step 1** (`Initialize outer loop`)**.** Fix initial iterate $x^1 \in C$ and memory trust-region radius $R_1^\sharp > 0$. Initialize $Q_1 \succeq 0$ with $\|Q_1\| \leq q$. Put outer loop counter $j = 1$.

▷ **Step 2** (`Stopping test`)**.** At outer loop counter j, stop if x^j is a critical point of (1.1). Otherwise go to inner loop.

▷ **Step 3** (`Initialize inner loop`)**.** Put inner loop counter $k = 1$ and initialize trust-region radius as $R_1 = R_j^\sharp$. Build first-order working model $\phi_1(\cdot, x^j)$, including at least one exactness plane at x^j. Possibly add recycled planes from previous steps. Build second-order working model $\Phi_1(\cdot, x^j) = \phi_1(\cdot, x^j) + \frac{1}{2}(\cdot - x^j)^\mathsf{T} Q_j(\cdot - x^j)$.

▷ **Step 4** (`Trial step generation`)**.** At inner loop counter k compute solution y^k of trust-region tangent program

$$\begin{array}{ll} \text{minimize} & \Phi_k(y, x^j) \\ \text{subject to} & \|y - x^j\| \leq R_k, y \in C \end{array}$$

Admit any $z^k \in C$ satisfying $\|z^k - x^j\| \leq \Theta \|y^k - x^j\|$ and $f(x^j) - \Phi_k(z^k, x^j) \geq \theta \left(f(x^j) - \Phi_k(y^k, x^j)\right)$ as trial step.

▷ **Step 5** (`Acceptance test`)**.** If

$$\rho_k = \frac{f(x^j) - f(z^k)}{f(x^j) - \Phi_k(z^k, x^j)} \geq \gamma$$

put $x^{j+1} = z^k$ (serious step), quit inner loop and goto step 8. Otherwise (null step), continue inner loop with step 6.

▷ **Step 6** (`Update working model`)**.** Use aggregation to taper out model $\phi_k(\cdot, x^j)$. Then call oracle $\mathscr{O}$ to generate cutting planes $m_k(\cdot, x^j)$ at z^k. Include aggregate plane and cutting planes in new first-order working model $\phi_{k+1}(\cdot, x^j)$. Generate second-order working model $\Phi_{k+1}(\cdot, x^j) = \phi_{k+1}(\cdot, x^j) + \frac{1}{2}(\cdot - x^j)^\mathsf{T} Q_j(\cdot - x^j)$.

▷ **Step 7** (`Update trust-region radius`)**.** Compute secondary control parameter

$$\widetilde{\rho}_k = \frac{f(x^j) - \phi_{k+1}(z^k, x^j)}{f(x^j) - \Phi_k(z^k, x^j)}$$

and put

$$R_{k+1} = \begin{cases} R_k & \text{if } \widetilde{\rho}_k < \widetilde{\gamma} \\ \frac{1}{2} R_k & \text{if } \widetilde{\rho}_k \geq \widetilde{\gamma} \end{cases}$$

Increase inner loop counter k and go back to step 4.

▷ **Step 8** (`Update memory radius`)**.** Store new memory trust-region radius

$$R_{j+1}^\sharp = \begin{cases} R_k & \text{if } \rho_k < \Gamma \\ 2R_k & \text{if } \rho_k \geq \Gamma \end{cases}$$

Update $Q_j \to Q_{j+1}$ respecting $Q_{j+1} \succeq 0$ and $\|Q_{j+1}\| \leq q$. Increase outer loop counter j and go back to step 2.

4.5. **Recycling of planes.** When a new serious iterate x^+ is found, then in the next sweep in step 3 a new initial first-order working model $\phi_1(\cdot, x^+)$ has to be generated. Contrary to the convex case, we cannot keep cuts generated around x to stay in the model for x^+. At least in the case of the downshift oracle $\mathscr{O}^{\downarrow}$ we can recycle them as follows: If $(a, g) \in \mathscr{O}^{\downarrow}(z, x)$ is in the model $\phi_k(\cdot, x)$ at acceptance, then include $(a^+, g) \in \mathscr{O}^{\downarrow}(z, x^+)$ in $\phi_1(\cdot, x^+)$ where $a^+ = f(z) + g^{\mathsf{T}}(x^+ - z) - s(z, x^+, g)$.

4.6. **Practical aspects.** Consider the case where the first-order working model has the form $\phi_k(\cdot, x^j) = \max_{i \in I_k} a_i + g_i^{\mathsf{T}}(\cdot - x^j)$ for some finite set I_k. Suppose the trust-region norm is the maximum norm, and $C = \{x : Ax \le b\}$ is a polyhedron. Then the tangent program is the following CQP

$$\begin{array}{ll} \text{minimize} & t + \frac{1}{2}(y - x^j)^{\mathsf{T}} Q_j (y - x^j) \\ \text{subject to} & a_i + g_i^{\mathsf{T}}(y - x^j) \le t, \quad i \in I_k \\ & Ay \le b \\ & -R_k \le y_i - x_i^j \le R_k \quad , i = 1, \dots, n \end{array} \tag{4.5}$$

with decision variable $(t, y) \in \mathbb{R} \times \mathbb{R}^n$, giving rise to the solution y^k in step 4. If $Q_j = 0$, then the tangent program is even a LP.

Remark 4.3. Axioms (W_1), (W_4) can be further relaxed by allowing so-called *anticipated cutting planes*. That means we can call the oracle $\mathscr{O}(z, x)$ at points z other than y^k, z^k and include those cutting planes into the working models. This still gives convergence and allows to exploit the specific structure of a problem to speed up acceptance in the inner loop.

5. Convergence

In this section we prove convergence of the trust-region algorithm toward a critical point $0 \in \partial f(x^*) + N_C(x^*)$. We prepare with three technical lemmas in section 5.1 and prove termination of the inner loop in section 5.2. During this part of the proof we write $x = x^j$ and $Q = Q_j$, as those elements are fixed during the inner loop at counter j. We write $|\cdot|_Q$ for the seminorm $|x|_Q^2 = x^{\mathsf{T}} Q x$.

5.1. Three technical lemmas.

Lemma 5.1. *There exists $\sigma > 0$, depending only on the trust-region norm $\|\cdot\|$, such that the solution y^k of the trust-region tangent program in step 4, with the corresponding aggregate subgradient $g_k^* \in \partial\left(\phi_k(\cdot, x) + i_C\right)(y^k)$, satisfies the estimate*

$$f(x) - \phi_k(y^k, x) \ge \sigma \|g_k^* + Q(y^k - x)\| \|y^k - x\|. \tag{5.1}$$

Proof. 1) Let $\|\cdot\|$ be the norm used in the trust-region tangent program, $|\cdot|$ the standard Euclidian norm. There exists $\epsilon > 0$ such that $|u| \le \epsilon$ implies $\|u\| \le 1$. Now if $\|u\| = 1$ and if v is in the normal cone to the $\|\cdot\|$-unit ball at u, we have $v^{\mathsf{T}}(u - u') \ge 0$ for every $\|u'\| \le 1$ by the normal cone criterion.

Hence $v^\mathsf{T}(u-u') \geq 0$ for every $|u'| \leq \epsilon$ by the above, and using $u' = \epsilon v/|v|$ that implies $v^\mathsf{T} u \geq \epsilon |v|$.

2) Since y^k is an optimal solution of (4.1), we have $0 = g_k^* + Q(y^k - x) + v_k$ by the optimality condition, where $g_k^* \in \partial\left(\phi_k(\cdot, x) + i_C\right)(y^k)$ is the aggregate subgradient and v_k a normal vector to the $\|\cdot\|$-norm ball $B(x, R_k)$ at y^k. By the subgradient inequality, using $x, y^k \in C$, we have

$$g_k^{*\mathsf{T}}(x - y^k) \leq \phi_k(x, x) - \phi_k(y^k, x) = f(x) - \phi_k(y^k, x).$$

Now by part 1), on putting $u_k = (y^k - x)/\|y^k - x\|$, we have $v_k^\mathsf{T} u_k \geq \epsilon |v_k|$ independently of k, because v_k, being normal to the $\|\cdot\|$-ball of radius $\|y^k - x\|$ and center 0 at $y^k - x$, is also normal to the $\|\cdot\|$-unit ball at u_k. But then $g_k^{*\mathsf{T}}(x - y^k) = v_k^\mathsf{T}(y^k - x) + (x - y^k)^\mathsf{T} Q(x - y^k) \geq v_k^\mathsf{T}(y^k - x) \geq \epsilon |v_k| \|y^k - x\| \geq \epsilon^2 \|v_k\| \|y^k - x\| = \epsilon^2 \|g_k^* + Q(y^k - x)\| \|y^k - x\|$, where we have used part 1). This proves the result with $\sigma = \epsilon^2$. □

Lemma 5.2. *There exists constants $\sigma' > 0, \sigma'' \geq 0$, depending only on the trust region norm $\|\cdot\|$ and the parameters θ, Θ, q used in the algorithm, such that the trial points z^k in step 4 of the algorithm, associated with the solutions y^k of the tangent program (4.1) and aggregate subgradients $g_k^* \in \partial\left(\phi_k(\cdot, x) + i_C\right)(y^k)$, satisfy the estimate*

$$f(x) - \Phi_k(z^k, x) \geq \left(\sigma' \|g_k^*\| - \sigma'' \|y^k - x\|\right) \|y^k - x\|. \tag{5.2}$$

Proof. Subtracting $\frac{1}{2}|y^k - x|_Q^2$ from both sides of (5.1) and using $|y^k - x|_Q^2 \leq \|Q\| \|y^k - x\|^2$ gives

$$f(x) - \Phi_k(y^k, x) \geq \left(\epsilon^2 \|g_k^* + Q(y^k - x)\| - \tfrac{1}{2}\|Q\| \|y^k - x\|\right) \|y^k - x\|.$$

Hence by (4.2), the triangle inequality, and $\|Q\| \leq q$ in step 8 of the algorithm, we have

$$\begin{aligned} f(x) - \Phi_k(z^k, x) &\geq \theta \left(\epsilon^2 \|g_k^* + Q(y^k - x)\| - \tfrac{1}{2}\|Q\| \|y^k - x\|\right) \|y^k - x\| \\ &\geq \theta \epsilon^2 \|g_k^*\| \|y^k - x\| - (\tfrac{1}{2} - \theta\epsilon^2) \|Q\| \|y^k - x\|^2 \\ &\geq \left(\theta \epsilon^2 \|g_k^*\| - (\tfrac{1}{2} - \theta\epsilon^2) q \|y^k - x\|\right) \|y^k - x\|. \end{aligned} \tag{5.3}$$

This is (5.2) with $\sigma' = \theta\epsilon^2$ and $\sigma'' = \max\{0, (\frac{1}{2} - \theta\epsilon^2) q\}$. □

Lemma 5.3. *Suppose $\Delta_k = \|y^k - x\|/\|g_k^*\| \to 0$ as $k \to \infty$. Then there exists a constant $\sigma > 0$, depending only on θ, Θ, q and the trust region norm $\|\cdot\|$, such that from some counter k_0 onward,*

$$f(x) - \Phi_k(z^k, x) \geq \sigma \|g_k^*\| \|z^k - x\|. \tag{5.4}$$

The counter k_0 can be chosen smallest with the property that $\Delta_k < \frac{1}{2}\theta\epsilon^2/(\frac{1}{2} - \theta\epsilon^2) q$ for all $k \geq k_0$.

Proof. The estimate is obvious from the previous Lemma if $\frac{1}{2} \leq \theta\epsilon^2$, as then $\sigma'' = 0$ in (5.2). Otherwise, since $\|y^k - x\|/\|g_k^*\| \to 0$, there exists an index k_0 such that

$$\Delta_k = \frac{\|y^k - x\|}{\|g_k^*\|} < \frac{\frac{1}{2}\theta\epsilon^2}{\left(\frac{1}{2} - \theta\epsilon^2\right) q} \tag{5.5}$$

for all $k \geq k_0$. Then from (5.3)

$$\begin{aligned} f(x) - \Phi_k(z^k, x) &\geq \left(\theta\epsilon^2\|g_k^*\| - (\tfrac{1}{2} - \theta\epsilon^2)q\|y^k - x\|\right)\|y^k - x\| \\ &\geq \tfrac{1}{2}\theta\epsilon^2\|g_k^*\|\|y^k - x\| && \text{(using (5.5))} \\ &\geq \tfrac{1}{2}\theta\epsilon^2\Theta^{-1}\|g_k^*\|\|z^k - x\| && \text{(using (4.2))} \end{aligned}$$

for all $k \geq k_0$. This is the requested estimate with $\sigma = \frac{1}{2}\theta\epsilon^2\Theta^{-1}$. □

5.2. **Finite termination of the inner loop.** We now investigate whether the inner loop at the current serious iterate x can find a new serious iterate x^+ satisfying the acceptance condition (4.3), or whether it fails and turns infinitely. The hypothesis for the inner loop is that the cutting plane oracle satisfies $(\mathscr{O}_1) - (\mathscr{O}_4)$. Strictness $(\widehat{\mathscr{O}}_3)$ of $\mathscr{O}$ is not needed for termination of the inner loop.

Lemma 5.4. *Suppose the inner loop at serious iterate x turns infinitely with $\liminf_{k\to\infty} R_k = 0$. Then x is a critical point of* (1.1)*, i.e., satisfies $0 \in \partial f(x) + N_C(x)$.*

Proof. According to step 7 of the algorithm we have $\widetilde{\rho}_k \geq \widetilde{\gamma}$ for infinitely many $k \in \mathcal{K}$. Since R_k is never increased during the inner loop, that implies $R_k \to 0$. Hence $y^k, z^k \to x$ as $k \to \infty$, where we use the trial step generation rule (4.2) in step 4 of the algorithm. We argue that this implies $\phi_k(z^k, x) \to f(x)$.

Indeed, $\limsup_{k\to\infty} \phi_k(z^k, x) \leq \limsup_{k\to\infty} \phi(z^k, x) = \lim_{k\to\infty} \phi(z^k, x) = f(x)$ is always true due to $\phi_k \leq \phi$ and axiom (M_1), and where ϕ is the envelope model of the cutting plane oracle $\mathscr{O}$. On the other hand, ϕ_k includes (i.e. dominates) an exactness plane $m_0(\cdot, x) = f(x) + g_0^\top(\cdot - x)$ by rule (W_1), hence $f(x) = \lim_{k\to\infty} m_0(z^k, x) \leq \liminf_{k\to\infty} \phi_k(z^k, x)$. These two together show $\phi_k(z^k, x) \to f(x)$, and then immediately also $\Phi_k(z^k, x) \to f(x)$. We also readily obtain $\phi_k(y^k, x) \to f(x)$ from the link (4.2) between z^k, y^k in step 4 of the algorithm.

We now prove that $\liminf_{k\to\infty} \|g_k^*\| = 0$, where g_k^* are the aggregate subgradients (definition 4.1) at the y^k. Assume on the contrary that $\|g_k^*\| \geq \eta > 0$ for all k. Since $y^k \to x$, by (5.4) there exists a constant $\sigma > 0$ such that for k large enough,

$$f(x) - \Phi_k(z^k, x) \geq \sigma\eta\|z^k - x\|. \tag{5.6}$$

Next observe that since by rule (W_3) all cutting planes at z^k are integrated in the new model $\phi_{k+1}(\cdot, x)$, we have $\phi_{k+1}(z^k, x) \geq m_k(z^k, x) = a_k + g_k^\top(z_k -$

x), where $(a_k, g_k) \in \mathscr{O}(z^k, x)$. But by $(\mathscr{O}_3)$ there exist $\epsilon_k \to 0^+$ and some such plane $m_k(\cdot, x)$ such that $f(z^k) \le m_k(z^k, x) + \epsilon_k \|z^k - x\|$. Hence

$$f(z^k) \le m_k(z^k, x) + \epsilon_k \|z^k - x\| \le \phi_{k+1}(z^k, x) + \epsilon_k \|z^k - x\|. \tag{5.7}$$

Now using (5.6) and (5.7) we estimate

$$\widetilde{\rho}_k = \rho_k + \frac{f(z^k) - \phi_{k+1}(z^k, x)}{f(x) - \Phi_k(z^k, x)} \le \rho_k + \frac{\epsilon_k \|z^k - x\|}{\sigma\eta \|z^k - x\|} = \rho_k + \epsilon_k/(\sigma\eta).$$

Since $\epsilon_k \to 0$ and $\rho_k < \gamma$, we have $\limsup \widetilde{\rho}_k \le \gamma < \widetilde{\gamma}$, a contradiction with $\widetilde{\rho}_k > \widetilde{\gamma}$ for the infinitely many $k \in \mathcal{K}$. That proves $g_k^* \to 0$ for a subsequence $k \in \mathcal{N} \subset \mathcal{K}$.

Write $g_k^* = p_k + q_k$ with $p_k \in \partial_1 \phi_k(y^k, x)$ and $q_k \in N_C(y^k)$. Using boundedness of the y^k, and hence boundedness of the p_k, we extract another subsequence $k \in \mathcal{K}' \subset \mathcal{N}$ such that $p_k \to p$, $q_k \to q$, $p + q = 0$. Since $y^k \to x$, we have $q \in N_C(x)$. We argue that $p \in \partial f(x)$. Indeed, for any test vector h the subgradient inequality gives

$$p_k^\mathsf{T} h \le \phi_k(y^k + h, x) - \phi_k(y^k, x) \le \phi(y^k + h, x) - \phi_k(y^k, x).$$

Since $\phi_k(y^k, x) \to f(x) = \phi(x, x)$, passing to the limit using $y^k \to x$ gives

$$p^\mathsf{T} h \le \phi(x + h, x) - \phi(x, x),$$

proving $p \in \partial_1 \phi(x, x) \subset \partial f(x)$ by axiom (M_1). This proves $0 = p + q \in \partial f(x) + N_C(x)$, hence x is a critical point of (1.1). □

Remark 5.5. Kiwiel's aggregation rule [36] for the convex bundle method allows to limit the bundle to any fixed number ≥ 3 planes. In [52] we have shown that the rule carries over to our non-convex bundle method, see also [49]. Ruszczyński [60] had asked whether aggregation could also be used in the convex trust region method. In [14] we had justified the rule for the non-convex non-smooth trust region method under the hypothesis that $Q \succ 0$. Here we shall show that this hypothesis cannot be removed.

The following result justifies the use of aggregation in the first place for the special case $z^k = y^k$ and $Q \succ 0$. Note that the trivial choice $z^k = y^k$ in step 4 is always authorized (due to $\Theta \ge 1$ and $\theta \le 1$ in rule (4.2)), but of course we want to use the additional freedom offered by z^k to improve performance of our method, so $z^k = y^k$ is rather restrictive, and we will remove it later.

Lemma 5.6. *Suppose $Q \succ 0$, the inner loop at x turns infinitely, and the trust-region radius R_k stays bounded away from* 0*. Suppose the y^k are chosen as trial steps. Then x is a critical point of* (1.1)*.*

Proof. Since the trust-region radius is frozen $R_k = R_{k_0}$ from some counter k_0 onwards, we write $R := R_{k_0}$. According to step 7 of the algorithm that means $\widetilde{\rho}_k < \widetilde{\gamma}$ for $k \ge k_0$. The only progress in the working model as we update $\phi_k \to \phi_{k+1}$ is now the addition of the cutting plane(s) and the aggregate plane. The working models ϕ_k now contain at least three planes,

an exactness plane, at least one cut from the last unsuccessful trial step, and the aggregate plane. They may contain more planes, but those will not be used in our argument below.

Since R_k stays bounded away from 0, it is not a priori clear whether $y^k \to x$, and we have to work to prove this. Since $Q \succ 0$, $|x|_Q^2 = x^\mathsf{T} Q x$ is a Euclidian norm. We write the objective of the tangent program as

$$\Phi_k(\cdot, x) = \phi_k(\cdot, x) + \tfrac{1}{2}|\cdot - x|_Q^2.$$

We know that the aggregate plane $m_k^*(\cdot, x)$ satisfies $m_k^*(y^k, x) = \phi_k(y^k, x)$, so it memorizes the value $\phi_k(y^k, x)$. The latter gives

$$\Phi_k(y^k, x) = m_k^*(y^k, x) + \tfrac{1}{2}|y^k - x|_Q^2. \tag{5.8}$$

Now we introduce the quadratic function

$$\Phi_k^*(\cdot, x) = m_k^*(\cdot, x) + \tfrac{1}{2}|\cdot - x|_Q^2,$$

then from what we have just seen in (5.8)

$$\Phi_k^*(y^k, x) = \Phi_k(y^k, x). \tag{5.9}$$

Moreover, we have

$$\Phi_k^*(\cdot, x) \le \Phi_{k+1}(\cdot, x), \tag{5.10}$$

because according to the aggregation rule (W_2) we include the aggregate plane $m_k^*(\cdot, x)$ in the new model ϕ_{k+1}, that is, we have $m_k^*(\cdot, x) \le \phi_{k+1}(\cdot, x)$, and hence (5.10). Expanding the quadratic function $\Phi_k^*(\cdot, x)$ at y^k gives

$$\Phi_k^*(\cdot, x) = \Phi_k^*(y^k, x) + \nabla\Phi_k^*(y^k, x)^\mathsf{T}(\cdot - y^k) + \tfrac{1}{2}|\cdot - y^k|_Q^2,$$

where $\nabla\Phi_k^* = g_k^* + Q(y^k - x)$. From the optimality condition of the tangent program at y^k we get $g_k^* + Q(y^k - x) = -v_k$ with v_k in the normal cone to the ball $B(x, R)$ at y^k, hence

$$\Phi_k^*(\cdot, x) = \Phi_k^*(y^k, x) - v_k^\mathsf{T}(\cdot - y^k) + \tfrac{1}{2}|\cdot - y^k|_Q^2. \tag{5.11}$$

Now we argue as follows:

(5.12)

$$\begin{aligned}
\Phi_k(y^k, x) &= \Phi_k^*(y^k, x) && \text{(by (5.9))}\\
&\le \Phi_k^*(y^k, x) + \tfrac{1}{2}|y^{k+1} - y^k|_Q^2 \\
&= \Phi_k^*(y^{k+1}, x) + v_k^\mathsf{T}(y^{k+1} - y^k) && \text{(by (5.11))}\\
&\le \Phi_k^*(y^{k+1}, x) && \text{(since } v_k^\mathsf{T}(y^{k+1} - y^k) \le 0)\\
&\le \Phi_{k+1}(y^{k+1}, x) && \text{(by (5.10))}\\
&\le \Phi_{k+1}(x, x) && (y^{k+1} \text{ minimizer of } \Phi_{k+1}(\cdot, x))\\
&= \phi(x, x) = f(x).
\end{aligned}$$

Therefore the sequence $\Phi_k(y^k, x)$ is increasing and bounded above, and converges to a limit $\Phi^* \leq f(x)$. Going back with this information to the estimation chain (5.12) shows $\frac{1}{2}|y^{k+1} - y^k|_Q^2 \to 0$ and also $v_k^\mathsf{T}(y^{k+1} - y^k) \to 0$. Since $Q \succ 0$, we deduce $y^{k+1} - y^k \to 0$. Then also

$$\tfrac{1}{2}|y^{k+1} - x|_Q^2 - \tfrac{1}{2}|y^k - x|_Q^2 \to 0,$$

by the triangle inequality. In consequence

$$\begin{aligned}\phi_{k+1}(y^{k+1}, x) - \phi_k(y^k, x) = \Phi_{k+1}(y^{k+1}, x) - \Phi_k(y^k, x) \\ - \tfrac{1}{2}|y^{k+1} - x|_Q^2 + \tfrac{1}{2}|y^k - x|_Q^2 \to 0.\end{aligned} \tag{5.13}$$

So far we have not made use of the cutting plane oracle, and we need it now. Observe that one of the cutting planes $m_k(\cdot, x)$ furnished by the oracle $\mathscr{O}$ at y^k must be an affine support function of $\phi_{k+1}(\cdot, x)$ at y^k. Indeed, $\phi_{k+1}(\cdot, x)$ is the envelope of the exactness plane, the aggregate plane, and the cutting planes, but here one of the cutting planes must be active at y^k, i.e., $\phi_{k+1}(y^k, x) = m_{y^k, g_k}(y^k, x) = a_k + g_k^\mathsf{T}(y^k - x)$ for some $(a_k, g_k) \in \mathscr{O}(y^k, x)$, because if the values of the planes furnished by the oracle are below the values of the other planes maintained in $\phi_{k+1}(\cdot, x)$, then $\widetilde{\rho}_k \geq 1$. But that cannot happen here since we are in the case where R_k is frozen, and we know that when $\widetilde{\rho}_k \geq 1$, the trust-region radius R_k is reduced due to the secondary test in step 7. Hence by the subgradient inequality

$$\phi_{k+1}(y^k, x) + g_k^\mathsf{T}(\cdot - y^k) \leq \phi_{k+1}(\cdot, x). \tag{5.14}$$

Now using (5.14) and the fact that $\phi_k(y^k, x) \leq \phi_{k+1}(y^k, x)$, we estimate as follows:

$$\begin{aligned}0 &\leq \phi_{k+1}(y^k, x) - \phi_k(y^k, x) \\ &= \phi_{k+1}(y^k, x) + g_k^\mathsf{T}(y^{k+1} - y^k) - \phi_k(y^k, x) - g_k^\mathsf{T}(y^{k+1} - y^k) \\ &\leq \phi_{k+1}(y^{k+1}, x) - \phi_k(y^k, x) - g_k^\mathsf{T}(y^{k+1} - y^k).\end{aligned}$$

Since the g_k are bounded and $y^{k+1} - y^k \to 0$, we have $g_k^\mathsf{T}(y^{k+1} - y^k) \to 0$, hence using (5.13) we deduce $\phi_{k+1}(y^k, x) - \phi_k(y^k, x) \to 0$, and also $\Phi_{k+1}(y^k, x) - \Phi_k(y^k, x) \to 0$.

Now we claim that $\phi_k(y^k, x) \to f(x)$. Since $\phi_k(y^k, x) \leq \Phi_k(y^k, x) \to \Phi^* \leq f(x)$, it remains to prove $\liminf \phi_k(y^k, x) \geq f(x)$. Suppose that this is not the case, and let $\phi_k(y^k, x) \to f(x) - \eta$ for a subsequence and some $\eta > 0$. Then also $\phi_{k+1}(y^k, x) \to f(x) - \eta$ for that subsequence (using $\phi_k(y^k, x) - \phi_{k+1}(y^k, x) \to 0$ proved above). Passing to yet another subsequence, and using boundedness of the y^k, we may assume $\frac{1}{2}|y^k - x|_Q^2 \to \ell \geq 0$. Choose $\delta > 0$ such that $\delta < (1 - \widetilde{\gamma})\eta$. From what we have just seen there exists k_1 such that

$$\phi_{k+1}(y^k, x) - \phi_k(y^k, x) < \delta$$

for all $k \geq k_1$. Now recall that $\widetilde{\rho}_k \leq \widetilde{\gamma}$ for every $k \geq k_0$, hence

$$\widetilde{\gamma}\left(\Phi_k(y^k, x) - f(x)\right) \leq \phi_{k+1}(y^k, x) - f(x) \leq \phi_k(y^k, x) - f(x) + \delta.$$

Passing to the limit gives $-\widetilde{\gamma}\eta + \widetilde{\gamma}\ell \le -\eta + \delta$, hence $(1-\widetilde{\gamma})\eta + \ell\widetilde{\gamma} \le \delta$, which contradicts the choice of δ. Hence $\phi_k(y^k, x) \to f(x)$. We immediately deduce that $\Phi_k(y^k, x) \to f(x)$ and $\Phi_{k+1}(y^k, x) \to f(x)$. This also shows $|y^k - x|_Q \to 0$, as follows from the estimates

$$\phi_k(y^k, x) \le \Phi_k(y^k, x) = \phi_k(y^k, x) + \tfrac{1}{2}|y^k - x|_Q^2 \le \Phi^* \le f(x),$$

where due to $\phi_k(y^k, x) \to f(x)$ all terms go to $f(x)$, implying $|y^k - x|_Q \to 0$.

Let us now show that x must be a critical point of (1.1). Let $\Phi(\cdot, x) = \phi(\cdot, x) + \frac{1}{2}|\cdot - x|_Q^2$ be the second order model associated with the envelope model ϕ of $\mathscr{O}$. Let $y \in C \cap B(x, R)$, then $\phi(y, x) \ge \phi_k(y, x)$ by construction of the working models, hence $\Phi(y, x) \ge \Phi_k(y, x) \ge \Phi_k(y^k, x)$, the latter since y^k is the minimizer of $\Phi_k(\cdot, x)$ over $C \cap B(x, R)$. Since $\Phi_k(y^k, x) \to f(x)$ by what we have seen above, we deduce $\Phi(y, x) \ge f(x)$ for every $y \in C \cap B(x, R)$. But $\Phi(x, x) = f(x)$, which means x is a minimizer of $\Phi(\cdot, x)$ over $C \cap B(x, R)$, hence over C. In consequence there exists $g \in \partial_1 \Phi(x, x)$ and $v \in N_C(x)$ such that $g + v = 0$. Since $\partial_1 \Phi(x, x) = \partial_1 \phi(x, x) \subset \partial f(x)$, this proves that x is a critical point of (1.1). □

Remark 5.7. Let us now see our first method to justify aggregation in the case $Q \succ 0$, $z^k \ne y^k$. In the algorithm we first allow z^k as a trial point in step 4. If acceptance in step 5 fails for z^k, then we include the cutting plane at z^k and continue with step 7 for z^k. However, if step 7 for z^k gives no reduction of R_k, then we are in the difficult case not covered by Lemma 5.4. We then do the following. We fall back on y^k as the trial point, i.e., we forget about z^k. When y^k is not accepted, we proceed with step 6, now for y^k, and apply aggregation. This is justified, because we are in the situation covered by Lemma 5.6. Note that the additional work required in steps 6 and 7 is marginal, so we do not waste time by this evasive maneuver. We could even perform this maneuver as default (i.e. checking y^k whenever z^k fails). We have proved the following

Lemma 5.8. *Suppose the inner loop at x turns infinitely. Let $Q \succ 0$ and accept to fall back on y^k if z^k fails in step 5 with $\widetilde{\rho}_k < \widetilde{\gamma}$ in step 7. Then $0 \in \partial f(x) + N_C(x)$.* □

Let us now discuss a second way to justify aggregation in the case $Q \succ 0$, $z^k \ne y^k$, which does *not* require the falling back procedure above. The idea proposed in [34] is to sharpen condition (4.2) as follows. Fix a sequence $\nu_k \to 0^+$ and require that the trial points z^k satisfy

$$z^k \in B(y^k, \nu_k) \text{ and } (4.2). \tag{5.15}$$

Then the technique of Lemma 5.6 allows a direct justification of aggregation with the aggregate plane taken at y^k, and the cutting plane taken at z^k.

Lemma 5.9. *Suppose trial points z^k in step 4 of the algorithm satisfy the stronger condition* (5.15)*. Suppose in case of a null step z^k cutting planes*

are generated at z^k and the aggregate plane is generated at y^k. Let $Q \succ 0$. Suppose the inner loop turns infinitely. Then x is a critical point of (1.1).

Proof. We follow the line of proof in Lemma 5.6. Up to formula (5.13) only properties of the aggregate plane are used, so we arrive at that same formula. Now we use that $\|z^k - y^k\| \le \nu_k \to 0$, and then since the $\phi_k(\cdot, x)$ have a joint Lipschitz constant L on the bounded set $B(x, R)$, we deduce $|\phi_k(z^k, x) - \phi_k(y^k, x)| \le L\|z^k - y^k\| \to 0$. In view of (5.13) this implies

$$\phi_{k+1}(z^{k+1}, x) - \phi_k(z^k, x) \to 0.$$

Now observe that we must have $\phi_{k+1}(z^k, x) \ge \phi_k(z^k, x)$, because otherwise $\widetilde{\rho}_k < \widetilde{\gamma}$ could not be satisfied. Hence one of the cutting planes $m_k(\cdot, x)$ is an affine support plane of $\phi_{k+1}(\cdot, x)$ at z^k, i.e., $\phi_{k+1}(z^k, x) = m_k(z^k, x) = a_k + g_k^\mathsf{T}(z^k - x)$ for at least one $(a_k, g_k) \in \mathscr{O}(z^k, x)$. Then from the subgradient inequality at z^k,

$$\phi_{k+1}(z^k, x) + g_k^\mathsf{T}(\cdot - x) \le \phi_{k+1}(\cdot, x).$$

Consequently, if we apply this at z^{k+1},

$$\begin{aligned} 0 &\le \phi_{k+1}(z^k, x) - \phi_k(z^k, x) \\ &= \phi_{k+1}(z^k, x) + g_k^\mathsf{T}(z^{k+1} - z^k) - \phi_k(z^k, x) - g_k^\mathsf{T}(z^{k+1} - z^k) \\ &\le \phi_{k+1}(z^{k+1}, x) - \phi_k(z^k, x) - g_k^\mathsf{T}(z^{k+1} - z^k). \end{aligned}$$

Since $y^{k+1} - y^k \to 0$ by (5.12) and $z^k - y^k \to 0$ by (5.15), we deduce $z^{k+1} - z^k \to 0$. Since the g_k are bounded, we have $g_k^\mathsf{T}(z^{k+1} - z^k) \to 0$. On the other hand, we know that $\phi_{k+1}(z^{k+1}, x) - \phi_k(z^k, x) \to 0$, so altogether we deduce

$$\phi_{k+1}(z^k, x) - \phi_k(z^k, x) \to 0, \tag{5.16}$$

and then also $\Phi_{k+1}(z^k, x) - \Phi_k(z^k, x) \to 0$.

Now we prove that $\phi_k(z^k, x) \to f(x)$. Since $\phi_k(z^k, x) - \phi_k(y^k, x) \to 0$ and by (5.12) $\phi_k(y^k, x) \to \Phi^* \le f(x)$, we have $\lim_k \phi_k(z^k, x) = \Phi^* \le f(x)$, and it remains to show the opposite estimate.

Suppose that contrary to what is claimed $\phi_k(z^k, x) \to f(x) - \eta$ for some $\eta > 0$ and a subsequence. Then also $\phi_{k+1}(z^k, x) \to f(x) - \eta$ as a consequence of (5.16). Passing to yet another subsequence, and using boundedness of the z^k, we may assume $\frac{1}{2}|z^k - x|_Q^2 \to \ell \ge 0$. Choose $\delta > 0$ such that $\delta < (1 - \widetilde{\gamma})\eta$. From what we have just seen there exists k_1 such that

$$\phi_{k+1}(z^k, x) - \phi_k(z^k, x) < \delta$$

for all $k \ge k_1$. Now recall that $\widetilde{\rho}_k \le \widetilde{\gamma}$ for every $k \ge k_0$, hence

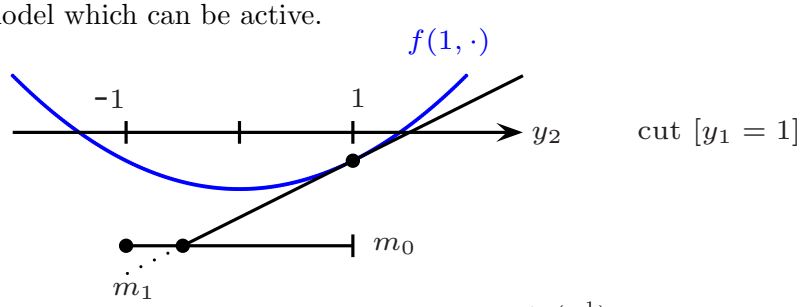
$$\widetilde{\gamma}\Big(\Phi_k(z^k, x) - f(x)\Big) \le \phi_{k+1}(z^k, x) - f(x) \le \phi_k(z^k, x) - f(x) + \delta.$$

Passing to the limit gives $-\widetilde{\gamma}\eta + \widetilde{\gamma}\ell \le -\eta + \delta$, hence $(1 - \widetilde{\gamma})\eta + \ell\widetilde{\gamma} \le \delta$, which contradicts the choice of δ. Hence $\phi_k(z^k, x) \to f(x)$.

Using $z^k - y^k \to 0$, we deduce $\phi_k(y^k, x) \to f(x)$ and then from (5.12) we get $\Phi_k(y^k, x) \to f(x)$. That implies $|y^k - x|_Q^2 \to 0$, and from $Q \succ 0$ follows $y^k \to x$.

Let us now show that x must be a critical point of (1.1). Let $\Phi(\cdot, x) = \phi(\cdot, x) + \frac{1}{2}|\cdot - x|_Q^2$ be the second order model associated with the envelope model ϕ of $\mathscr{O}$. Let $y \in C \cap B(x, R)$, then $\phi(y, x) \geq \phi_k(y, x)$ by construction of the working models, hence $\Phi(y, x) \geq \Phi_k(y, x) \geq \Phi_k(y^k, x)$, the latter since y^k is the minimizer of $\Phi_k(\cdot, x)$ over $C \cap B(x, R)$. Since $\Phi_k(y^k, x) \to f(x)$ by what we have seen above, we deduce $\Phi(y, x) \geq f(x)$ for every $y \in C \cap B(x, R)$. But $\Phi(x, x) = f(x)$, which means x is a minimizer of $\Phi(\cdot, x)$ over $C \cap B(x, R)$, hence over C. In consequence there exists $g \in \partial_1 \Phi(x, x)$ and $v \in N_C(x)$ such that $g + v = 0$. Since $\partial_1 \Phi(x, x) = \partial_1 \phi(x, x) \subset \partial f(x)$, this proves that x is a critical point of (1.1). $\square$

Example 5.10. We show by way of an example that the hypothesis $Q \succ 0$ cannot be removed neither from Lemmas 5.6,5.8, nor from Lemma 5.9. Consider (1.1) with $f(x) = \frac{1}{2}x_1^2 - x_1 + \frac{1}{4}x_2^2$, $C = \mathbb{R}^2$. In the algorithm choose $\gamma = \frac{1}{2}$, $Q = 0$, the trust region norm $\|\cdot\| = |\cdot|_\infty$. Suppose the current iterate is $x = (x_1, x_2) = (0, 0)$ with $f(0, 0) = 0$, $g_0 = (-1, 0) \in \partial f(0, 0)$. The exactness plane is $m_0(z, x) = -z_1$, and we let $\phi_1 = m_0$. Suppose the current trust region radius is $R_1 = 1$, then a minimizer of the tangent program is $y^1 = (1, 1)$. In fact, the entire segment $\{1\} \times [-1, 1]$ is minimizing, but we choose y^1 at a corner. Take $z^1 = y^1$. The predicted progress at y^1 is 1, because $m_0(y^1, x) = \phi_1(y^1, x) = -1$. The aggregate plane $m_1^*(\cdot, x)$ at y^1 coincides with $m_0(\cdot, x)$, because there is only one plane in the working model which can be active.



We have $f(y^1) = -\frac{1}{4}$. Therefore $\rho_1 = \frac{0-(-\frac{1}{4})}{0-(-1)} = \frac{1}{4} < \gamma$ so y^1 is a null step. The function being convex, the cutting plane at y^1 is the tangent plane $m_1(z, x) = \frac{1}{4} + \frac{1}{2}(z_2 - 1)$. Since f is convex, the cutting plane always gives an improvement due to $\rho = \widetilde{\rho}$, so we keep $R_2 = R_1$, and we have $\phi_2(\cdot, x) = \max\{m_0(\cdot, x), m_1(\cdot, x)\} = \max\left\{-z_1, \frac{1}{4} + \frac{1}{2}(z_2 - 1)\right\}$. Now a minimizer of $\phi_2(\cdot, x)$ over the trust region is $y^2 = (1, -1)$. (The figure shows the trace of the graph of ϕ_1, ϕ_2, f in the plane $y_1 = 1$). Since $m_0((1, -1), x) = -1$ and $m_1((1, -1), x) = -\frac{5}{4}$, only m_0 is active at y^1, hence the aggregate plane is again $m_2^* = m_0$. Since $\rho_2 = \widetilde{\rho}_2 = \frac{1}{4}$, the situation is the same as in the first trial, (the figure being the one reflected by $y_2 = 0$), and we need a cutting

plane at y^2. This is now $m_2(z,x) = -\frac{1}{4} - \frac{1}{2}(z_2+1)$. Now if we do not keep the plane m_1 in the working model, the tangent program is minimizing $\phi_3(\cdot,x) = \max\{m_0(\cdot,x), m_2(\cdot,x)\} = \max\{-z_1, -\frac{1}{4} - \frac{1}{2}(z_2+1)\}$ over $[-1,1]^2$, and a solution is $y^3 = (1,1)$. This means, the inner loop oscillates between $y^{2k+1} = y^1$ and $y^{2k} = y^2$, and no progress occurs.

Remark 5.11. The phenomenon in this example cannot occur if $\|\cdot\| = |\cdot|_2$. However, since the choice of this norm is not as useful as in the classical trust region method due to the affine constraints in ϕ_k, we do not present the details of the result here.

Remark 5.12. In the general case $Q \succeq 0$ we can still limit the size of the working model to $n+2$ using Carathéodory's theorem. See [13] for the details.

Remark 5.13. In summary, justification of the aggregate rule in the case $Q \succ 0$ can be based either on the technique of remark 5.7, or on adopting the stronger rule (5.15) for trial steps.

Remark 5.14. We investigate whether backtracking steps $z_\alpha = x + \alpha(y^k - x)$ for $0 < \alpha < 1$ could be trial steps in the sense of (4.2) or (5.15), as this would allow to use linesearch in case of rejection of y^k.

Let $\Delta := f(x) - \Phi_k(y^k, x) > 0$. By convexity the line joining $(x, f(x))$ and $(y^k, f(x) - \Delta)$ is above the curve $\alpha \mapsto \Phi_k(z_\alpha, x)$. Therefore z_θ satisfies $f(x) - \Phi_k(z_\theta, x) \geq \theta\Delta$. Since $z_\theta \in B(x, \|x - y^k\|)$ is clear, every z_α with $\theta \leq \alpha \leq 1$ is a trial point in the sense of (4.2). But not necessarily in the sense of (5.15), because $\|z_\theta - y^k\| = (1-\theta)\|x - y^k\|$. If we impose $\|z_\theta - y^k\| \leq \nu_k$, then shorter steps might be forced. This is an argument for the fall back method.

5.3. **Convergence of the outer loop.** We are now ready to prove the main convergence result for our optimization method. This is where strictness of the oracle ($\widehat{\mathscr{O}}_3$) is needed.

Theorem 5.15. *Suppose $x^1 \in C$ is such that the level set $\{x \in C : f(x) \leq f(x^1)\}$ is bounded. Let $x^j \in C$ be the sequence of serious iterates generated by the bundle trust-region algorithm based on a strict cutting plane oracle. Then every accumulation point x^* of the x^j is a critical point of* (1.1).

Proof. 1) Without loss we consider the case where the algorithm generates an infinite sequence $x^j \in C$ of serious iterates. Suppose that at outer loop counter j the inner loop finds a successful trial step at inner loop counter k_j, that is, $z^{k_j} = x^{j+1}$, where the corresponding solution of the tangent program is $\tilde{x}^{j+1} = y^{k_j}$. Then $\rho_{k_j} \geq \gamma$, which means

$$f(x^j) - f(x^{j+1}) \geq \gamma \left(f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) \right). \tag{5.17}$$

Moreover, by condition (4.2) we have $\|\tilde{x}^{j+1} - x^j\| \leq \Theta \|x^{j+1} - x^j\|$ and

$$f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) \geq \theta \left(f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j) \right), \tag{5.18}$$

and combining (5.17) and (5.18) gives

$$f(x^j) - f(x^{j+1}) \geq \gamma\theta \left(f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j)\right). \tag{5.19}$$

Since $y^{k_j} = \tilde{x}^{j+1}$ is a solution of the k_j^{th} tangent program (4.1) of the jth inner loop, by the optimality condition there exist $g_j^* \in \partial\left(\phi_{k_j}(\cdot, x^j) + i_C\right)(\tilde{x}^{j+1})$ and a unit normal vector v_j to the ball $B(x^j, R_{k_j})$ at $\tilde{x}^{j+1}$ such that

$$g_j^* + Q_j(\tilde{x}^{j+1} - x^j) + \lambda_j v_j = 0, \tag{5.20}$$

where $\lambda_j = \|g_j^* + Q_j(\tilde{x}^{j+1} - x^j)\|$. Decomposing further, there exist $p_j \in \partial_1\phi_{k_j}(\tilde{x}^{j+1}, x^j)$ and $q_j \in N_C(\tilde{x}^{j+1})$ such that

$$0 = g_j^* + Q_j(\tilde{x}^{j+1} - x^j) + \lambda_j v_j = p_j + q_j + Q_j(\tilde{x}^{j+1} - x^j) + \lambda_j v_j. \tag{5.21}$$

By the subgradient inequality, applied to $p_j \in \partial\phi_{k_j}(\cdot, x^j)(\tilde{x}^{j+1})$, we have

$$\begin{aligned}
&-(q_j + \lambda_j v_j)^\mathsf{T}(x^j - \tilde{x}^{j+1}) + |\tilde{x}^{j+1} - x^j|_{Q_j}^2 \\
&\qquad = p_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) && \text{(using (5.21))} \\
&\qquad \leq \phi_{k_j}(x^j, x^j) - \phi_{k_j}(\tilde{x}^{j+1}, x^j) && \text{(subgradient inequality)} \\
&\qquad = f(x^j) - \phi_{k_j}(\tilde{x}^{j+1}, x^j) \\
&\qquad = f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j) + \tfrac{1}{2}|\tilde{x}^{j+1} - x^j|_{Q_j}^2 \\
&\qquad \leq \gamma^{-1}\theta^{-1}\left(f(x^j) - f(x^{j+1})\right) + \tfrac{1}{2}|\tilde{x}^{j+1} - x^j|_{Q_j}^2 && \text{(by (5.19))}.
\end{aligned}$$

Re-arranging, we obtain

$$(q_j + \lambda_j v_j)^\mathsf{T}(\tilde{x}^{j+1} - x^j) + \tfrac{1}{2}|\tilde{x}^{j+1} - x^j|_{Q_j}^2 \leq \gamma^{-1}\theta^{-1}\left(f(x^j) - f(x^{j+1})\right). \tag{5.22}$$

By the normal cone condition for $C \cap B(x^j, R_{k_j})$ at $\tilde{x}^{j+1}$ we have $(q_j + \lambda_j v_j)^\mathsf{T}(\tilde{x}^{j+1} - x^j) \geq 0$, hence both terms on the left of (5.22) are nonnegative. But the term on the right of (5.22) is telescoping, hence summable due to convergence of $f(x^j)$, so we deduce summability of $\sum_{j\in\mathbb{N}}(q_j + \lambda_j v_j)^T(\tilde{x}^{j+1} - x^j) < \infty$ and of $\sum_{j\in\mathbb{N}} |\tilde{x}^{j+1} - x^j|_{Q_j}^2 < \infty$. From (5.21) we then get summability of $\sum_{j\in\mathbb{N}} p_j^\mathsf{T}(x^j - \tilde{x}^{j+1})$. Hence $p_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) \to 0$, $(q_j + \lambda_j v_j)^\mathsf{T}(x^j - \tilde{x}^{j+1}) \to 0$, $|x^j - \tilde{x}^{j+1}|_{Q_j} \to 0$. Moreover, we know by local boundedness of the subdifferential that the sequence p_j is bounded, hence the sequence $q_j + \lambda_j v_j$ is also bounded.

2) Now let x^* be an accumulation point of the sequence x^j. We have to show that x^* is a critical point of (1.1). Fix a subsequence $j \in J$ such that $x^j \to x^*$, and $p_j \to p$, $q_j + \lambda_j v_j \to q$, $\tilde{x}^{j+1} \to \tilde{x}$ for suitable limits $p, q, \tilde{x}$. We shall now analyze two types of infinite subsequences $j \in J$, those where the trust-region constraint is active at $\tilde{x}^{j+1}$ and the Lagrange multiplier of the trust-region constraint is nonzero, i.e. $\lambda_j > 0$ in (5.20), and those where the Lagrange multiplier of the trust-region constraint vanishes, i.e., $\lambda_j = 0$ in (5.20).

3) Let us start with the simpler case of an infinite subsequence x^j, $j \in J$, where the Lagrange multiplier of the trust-region constraint vanishes, i.e., $\lambda_j = 0$ in (5.20). That occurs either when $\|x^j - \tilde{x}^{j+1}\| < R_{k_j}$, i.e., where the trust-region constraint is inactive at acceptance, or when it is active but with vanishing multiplier. In this case equation (5.21) simplifies to

$$0 = g_j^* + Q_j(\tilde{x}^{j+1} - x^j) = p_j + q_j + Q_j(\tilde{x}^{j+1} - x^j).$$

This means $p_j \to p$, $q_j \to q$ and $p + q = 0$, bearing in mind that we have $|\tilde{x}^{j+1} - x^j|_{Q_j} \to 0$.

Now let h be any test vector, then from the subgradient inequality, the acceptance condition, and using $\phi_k \le \phi$, we have

$$\begin{aligned} p_j^\mathsf{T} h &\le \phi_{k_j}(\tilde{x}^{j+1} + h, x^j) - \phi_{k_j}(\tilde{x}^{j+1}, x^j) \\ &\le \phi(\tilde{x}^{j+1} + h, x^j) - f(x^j) + f(x^j) - \phi_{k_j}(\tilde{x}^{j+1}, x^j) \\ &= \phi(\tilde{x}^{j+1} + h, x^j) - f(x^j) + f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j) + \tfrac{1}{2}|x^j - \tilde{x}^{j+1}|_{Q_j}^2 \\ &\le \phi(\tilde{x}^{j+1} + h, x^j) - f(x^j) + \gamma^{-1}\theta^{-1}\left(f(x^j) - f(x^{j+1})\right) + \tfrac{1}{2}|x^j - \tilde{x}^{j+1}|_{Q_j}^2. \end{aligned}$$

Let h' be another test vector and put $h = x^j - \tilde{x}^{j+1} + h'$. On substituting this expression above we obtain

$$\begin{aligned} p_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) + p_j^\mathsf{T} h' \le \phi(x^j + h', x^j) - f(x^j) \\ + \gamma^{-1}\theta^{-1}\left(f(x^j) - f(x^{j+1})\right) + \tfrac{1}{2}|x^j - \tilde{x}^{j+1}|_{Q_j}^2. \end{aligned}$$

Passing to the limit $j \in J$, we have $p_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) \to 0$ and $|x^j - \tilde{x}^{j+1}|_{Q_j} \to 0$ by 1) above, and $f(x^j) - f(x^{j+1}) \to 0$ by the construction of the descent method. Moreover, $\limsup_{j \in J} \phi(x^j + h', x^j) \le \phi(x^* + h', x^*)$ by $x^j \to x^*$ and axiom (M_3). Since $p_j \to p$, we get

$$p^\mathsf{T} h' \le \phi(x^* + h', x^*) - f(x^*) = \phi(x^* + h', x^*) - \phi(x^*, x^*).$$

Since h' was arbitrary and $\phi(\cdot, x^*)$ is convex, we deduce $p \in \partial_1 \phi(x^*, x^*)$, hence $p \in \partial f(x^*)$ by axiom (M_1).

Now we have to show that $q \in N_C(x^*)$. Since $q_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) \to 0$ due to $\lambda_j = 0$ and by part 1) above, and since $q_j \to q$, we have $q^\mathsf{T}(x^* - \tilde{x}) = 0$. Since $q_j \in N_C(\tilde{x}^{j+1})$ and $\tilde{x}^{j+1} \to \tilde{x}$, we know that $q \in N_C(\tilde{x})$. Hence for any element $x \in C$ we have $q^\mathsf{T}(\tilde{x} - x) \ge 0$ by the normal cone criterion. Hence $q^\mathsf{T}(x^* - x) = q^\mathsf{T}(\tilde{x} - x) + q^\mathsf{T}(x^* - \tilde{x}) = q^\mathsf{T}(\tilde{x} - x) \ge 0$, so the normal cone criterion holds also at x^*, proving $q \in N_C(x^*)$. We have shown that $0 = p + q \in \partial\left(\phi(\cdot, x^*) + i_C\right)(x^*) \subset \partial f(x^*) + N_C(x^*)$, hence x^* is a critical point of (1.1).

4) Let us now consider the more complicated case of an infinite subsequence, where $\|x^j - \tilde{x}^{j+1}\| = R_{k_j}$ with $\lambda_j > 0$, corresponding to the case of a non-vanishing multiplier in (5.20). Recall that $x^j \to x^*$, $j \in J$, and that we have to show that x^* is critical.

We shall now have to distinguish two subcases. Either $R_{k_j} \geq R_0 > 0$ for some $R_0 > 0$ and all $j \in J$, or there exists a subsequence $J' \subset J$ such that $R_{k_j} \to 0$ as $j \in J'$. The first case is discussed in 5), the second case will be handled in 6) - 7).

5) Let us consider the sub-case of an infinite subsequence $j \in J$ where $\|x^j - \tilde{x}^{j+1}\| = R_{k_j} \geq R_0 > 0$ for every $j \in J$. Recall from the general considerations in part 1) that $|\tilde{x}^{j+1} - x^j|_{Q_j} \to 0$. Since on the other hand we are in the case where $\|\tilde{x}^{j+1} - x^j\| = R_{k_j} \geq R_0 > 0$ for $j \in J$, Lemma 5.1 provides a constant $\sigma > 0$ independent of j such that

$$f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j) \geq \sigma \|g_j^*\| \|\tilde{x}^{j+1} - x^j\| \geq \sigma R_0 \|g_j^*\|,$$

for all $j \in J$. By acceptance $\rho_{k_j} \geq \gamma$ and (4.2) that gives

$$f(x^j) - f(x^{j+1}) \geq \gamma\theta \left(f(x^j) - \Phi_{k_j}(\tilde{x}^{j+1}, x^j)\right) \geq \gamma\theta\sigma R_0 \|g_j^*\|$$

and implies $g_j^* \to 0$. Splitting $g_j^* = p_j + q_j \to 0$ with $p_j \in \partial_1 \phi_{k_j}(\tilde{x}^{j+1}, x^j)$ and $q_j \in N_C(\tilde{x}^{j+1})$, recall that $p_j \to p$ and $q_j + \lambda_j v_j \to q$, and $0 = g_j^* + Q_j(\tilde{x}^{j+1} - x^j) + \lambda_j v_j \to p + q$, hence $p + q = 0$ and $\lambda_j v_j \to 0$, i.e., $\lambda_j \to 0$, and we find $q_j \to q$. Since $\tilde{x}^{j+1} \to \tilde{x}$, we infer $q \in N_C(\tilde{x})$. From here on we can argue as in part 2), which means x^* is critical, where in the last argument we infer $q_j^\mathsf{T}(x^j - \tilde{x}^{j+1}) \to 0$ from $\lambda_j \to 0$.

6) It remains to discuss the most complicated sub-case of an infinite subsequence $j \in J$, where the trust-region constraint is active with non-vanishing multiplier $\lambda_j > 0$ and $R_{k_j} \to 0$. This needs two sub-sub-cases. The first of these is a sequence $j \in J$ where in each jth outer loop the trust-region radius was reduced at least once. The second sub-sub-case are infinite subsequences where the trust-region radius stayed frozen ($R_j^\sharp = R_{k_j}$) throughout the jth inner loop for every $j \in J$. This is discussed in 6) below.

Let us first consider the case of an infinite sequence $j \in J$ where R_{k_j} is active at $\tilde{x}^{j+1}$ with $\lambda_j > 0$, and $R_{k_j} \to 0$, $j \in J$, and where during the jth inner loop the trust-region radius was reduced at least once. Suppose this happened the last time before acceptance at inner loop counter $k_j - \nu_j$ for some $\nu_j \geq 1$. Then for $j \in J$,

$$R_{k_j} = R_{k_j - 1} = \cdots = R_{k_j - \nu_j + 1} = \tfrac{1}{2} R_{k_j - \nu_j}.$$

By step 7 of the algorithm, that implies

$$\widetilde{\rho}_{k_j - \nu_j} \geq \widetilde{\gamma}, \quad \rho_{k_j - \nu_j} < \gamma.$$

Now $\|\tilde{x}^{j+1} - x^j\| = R_{k_j}$, $\|x^{j+1} - x^j\| \leq \Theta R_{k_j}$ and $\|z^{k_j - \nu_j} - x^j\| \leq \Theta R_{k_j - \nu_j - 1} = 2\Theta R_{k_j}$, hence $x^{j+1} - z^{k_j - \nu_j} \to 0$, $x^j - z^{k_j - \nu_j} \to 0$, $j \in J$.

By strictness ($\widehat{\mathscr{O}}_3$), and due to convergence $z^{k_j - \nu_j} \to x^*$ as well as $x^j \to x^*$, there exists at least one cutting plane $m_j(\cdot, x^j) = a_j + g_j^\mathsf{T}(\cdot - x^j)$ with $(a_j, g_j) \in \mathscr{O}(z^{k_j - \nu_j}, x^j)$ in tandem with $\epsilon_j \to 0^+$ such that $f(z^{k_j - \nu_j}) \leq m_j(z^{k_j - \nu_j}, x^j) + \epsilon_j \|z^{k_j - \nu_j} - x^j\|$. Since by rule ($W_3$) this cutting plane is

included in the next working model $\phi_{k_j-\nu_j+1}(\cdot, x^j)$, we have $m_j(\cdot, x^j) \leq \phi_{k_j-\nu_j+1}(\cdot, x^j)$, hence we obtain

$$f(z^{k_j-\nu_j}) \leq \phi_{k_j-\nu_j+1}(z^{k_j-\nu_j}, x^j) + \epsilon_j \|z^{k_j-\nu_j} - x^j\|.$$

Let $\widetilde{g}_j \in \partial\left(\phi_{k_j-\nu_j}(\cdot, x^j) + i_C\right)(y^{k_j-\nu_j})$ denote the aggregate subgradient at $y^{k_j-\nu_j}$. By Lemma 5.3 we have $f(x^j) - \Phi_{k_j-\nu_j}(z^{k_j-\nu_j}, x^j) \geq \sigma\|\widetilde{g}_j\|\|x^j - z^{k_j-\nu_j}\|$ for a constant σ independent of j. Now recall that $x^j \to x^*$ and that we have to show that x^* is critical. It suffices to show that there is a subsequence $j \in J'$ with $\widetilde{g}_j \to 0$.

Assume on the contrary that $\|\widetilde{g}_j\| \geq \eta > 0$ for every $j \in J$. Then

$$f(x^j) - \Phi_{k_j-\nu_j}(z^{k_j-\nu_j}, x^j) \geq \eta\sigma\|z^{k_j-\nu_j} - x^j\|.$$

Now expanding the test quotient at $z^{k_j-\nu_j}$ gives

$$\begin{aligned}\widetilde{\rho}_{k_j-\nu_j} &= \rho_{k_j-\nu_j} + \frac{f(z^{k_j-\nu_j}) - \phi_{k_j-\nu_j+1}(z^{k_j-\nu_j}, x^j)}{f(x^j) - \Phi_{k_j-\nu_j}(z^{k_j-\nu_j}, x^j)} \\ &\leq \rho_{k_j-\nu_j} + \frac{\epsilon_j\|z^{k_j-\nu_j} - x^j\|}{\eta\sigma\|z^{k_j-\nu_j} - x^j\|} < \widetilde{\gamma}\end{aligned}$$

for $j \in J$ sufficiently large, contradicting $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma}$. This shows that there must exist a subsequence J' of J such that $\widetilde{g}_j \to 0$, $j \in J'$. Passing to the limit $j \in J'$, we use the argument of part 2) to show that $0 \in \partial\left(\phi(\cdot, x^*) + i_C\right)(x^*)$, hence x^* is critical for (1.1).

7) Now consider an infinite subsequence $j \in J$ where $x^j \to x^*$, the trust-region radius R_{k_j} was active at $\widetilde{x}^{j+1}$ with non-zero multiplier $\lambda_j > 0$ when x^{j+1} was accepted, where $R_{k_j} \to 0$, but where during the jth inner loop the trust-region radius was never reduced. Since $R_{k_j} \to 0$, the work to bring the radius to 0 must be put about somewhere else outside J. For every $j \in J$ define $j' \in \mathbb{N}$ to be the largest index $j' < j$ such that in the j'th inner loop, the trust-region radius was reduced at least once. This means that in none of the loops $j'+1, \ldots, j$ was the trust-region radius reduced. As a consequence,

$$R_{k_{j'}} \leq R_{k_{j'+1}} \leq \cdots \leq R_{k_j} \to 0,$$

so all the trust region radii at acceptance in any of the loops between j' and j tend to 0. Indeed, all that can happen is that due to good acceptance the trust-region radii are increased at acceptance (see step 8), so the largest among them is R_{k_j}.

Let $J' = \{j' : j \in J\}$, where we understand $j \mapsto j'$ as a function. Passing to a subsequence of J, J', we may assume that $x^{j'} \to x'$ and $g^*_{j'} \to 0$, the latter because the sequence J' corresponds to one of the cases discussed in parts 2) - 5). Passing to yet another subsequence, we may arrange that the sequences J, J' are interlaced. That is, $j' < j < j'^+ < j^+ < j'^{++} < j^{++} < \cdots \to \infty$. This since j' tends to ∞ as a function of j.

Now assume that there exists $\eta > 0$ such that $\|g_j^*\| \geq \eta$ for all $j \in J$. Then since $x^j \to x^*$, we also have $x^{j+1} \to x^*$ due to $R_{k_j} \to 0$ and the second part of rule (4.2). Fix $\epsilon > 0$ with $\epsilon < \eta$. For $j \in J$ large enough we have $\|g_{j'}^*\| < \epsilon$, because $g_{j'}^* \to 0$, $j' \in J'$, and as j gets larger, so does j'. That means in the interval $[j', j)$ there exists an index $j'' \in \mathbb{N}$ such that

$$\|g_{j''}^*\| < \epsilon, \quad \|g_i^*\| \geq \epsilon \text{ for all } i = j''+1, \ldots, j.$$

The index j'' may coincide with j', it might also be larger, but it precedes j. In any case, $j \mapsto j''$ is again a function on J and defines another infinite index set J'' still interlaced with J.

Now since $|x^i - x^{i+1}|_{Q_i} \to 0$, and at the same time $\|g_i^*\| \geq \epsilon$ for $i = j''+1, \ldots, j$, we can use Lemma 5.3 to get estimates of the form $\sigma\|g_i^*\|\|x^i - x^{i+1}\| \leq f(x^i) - f(x^{i+1})$ with σ independent of $i \in [j''+1, \ldots, j]$ and j large enough. Summation gives

$$\sigma \sum_{i=j''+1}^{j} \|g_i^*\|\|x^i - x^{i+1}\| \leq f(x^{j''+1}) - f(x^{j+1}) \to 0 \qquad (j \in J, j \to \infty, j \mapsto j'').$$

Since $\|g_i^*\| \geq \epsilon$ for all $i \in [j''+1, \ldots, j]$, the sequence $\sum_{i=j''+1}^{j} \|x^i - x^{i+1}\| \to 0$ converges as $j \in J, j \to \infty$. By the triangle inequality, $x^{j''+1} - x^{j+1} \to 0$. Therefore $x^{j''+1} \to x^*$, and also $\tilde{x}^{j''+1} \to x^*$ because $\|x^{j''+1} - \tilde{x}^{j''+1}\| \leq \|x^{j''+1} - x^{j''}\| + \|x^{j''} - \tilde{x}^{j''+1}\| \leq \Theta R_{k_{j''}} + R_{k_{j''}} \to 0$.

Since as an aggregate subgradient $g_{j''}^* \in \partial\left(\phi(\cdot, x^{j''}) + i_C\right)(\tilde{x}^{j''+1})$, passing to yet another subsequence and using local boundedness of ∂f, we get $g_{j''}^* \to g^*$. But with $x^{j''} \to x^*$ and model property (M_3) we get $g^* \in \partial(\phi(\cdot, x^*) + i_C)(x^*)$. It follows that $\partial(\phi(\cdot, x^*) + i_C)(x^*)$ contains an element g^* of norm less than or equal to ϵ. As $\epsilon < \eta$ was arbitrary, we conclude that $0 \in \partial(\phi(\cdot, x^*) + i_C)(x^*) \subset \partial f(x^*) + N_C(x^*)$. That settles the remaining case. □

Remark 5.16. As the proof reveals, strictness of the cutting plane oracle is not needed if the trust region radius stays bounded away from 0, or if the trust region constraint is not strongly active.

6. Applications

A natural question is why the secondary test (4.4) in step 7 of the algorithm is required. A partial answer is that convergence fails if the test is removed without substitute, as shown in our examples in [50, 13]. In exchange, sometimes the secondary test (4.4) becomes redundant. There are at least three situations where this happens.

The first case is of course the convex tangent plane oracle $\mathscr{O}^{\tan}$, where cutting planes are tangents to f and $Q = 0$. Here $\rho_k = \widetilde{\rho}_k$, so for a null step the trust region radius is fixed. This is corroborated by [60], where

convergence is proved with a trust region radius fixed all along. The resulting method then resembles the cutting plane method.

6.1. **The all-tangents-oracle.** A second case is when a model ϕ is available to construct an oracle and we choose the working model as $\phi_k = \phi$ at all k. This is allowed and corresponds to choosing as oracle $\mathscr{O}^{\mathrm{all}}(z, x)$ the set of *all tangents to* $\phi(\cdot, x)$ at all points $z' \in B(x, M)$. Then always $\widetilde{\rho}_k \geq 1$, hence the trust region radius is *always* reduced at a null step. The price to pay for this simplification is that the tangent program is no longer of a simple structure.

Example 6.1. (Proximal point method). Let us show that splitting techniques fit nicely into the framework of the all-tangents-oracle $\mathscr{O}^{\mathrm{all}}$. Consider $f = g + h$ with g convex and h of class C^1. We know that $\phi(y, x) = g(y) + h(x) + \nabla h(x)(y - x)$ is a strict model of f, because here the Clarke subdifferential is additive due to strict differentiability of h. Now we use $\mathscr{O}^{\mathrm{all}}$, that is, we choose as working model $\phi_k(\cdot, x) = \phi(\cdot, x)$. As second order working model we choose

$$\Phi(y, x) = \phi(y, x) + \tfrac{1}{2r_x}\|y - x\|^2,$$

where $r_x > 0$ may depend on the serious iterate x. That corresponds to $Q(x) = (1/r_x)I$. Note that if we add the constant term $(r_x/2)\|\nabla h(x)\|^2 - h(x)$, then the tangent program (4.1) becomes

(6.1)
$$x^+ = \operatorname*{argmin}_y \Phi(y, x) - h(x) + \tfrac{r_x}{2}\|\nabla h(x)\|^2 = \operatorname*{argmin}_y g(y) + \tfrac{1}{2r_x}\|y - x + r_x \nabla h(x)\|^2$$

which is the convex proximal point method with splitting. This means convergence theory of the convex proximal point method with splitting is a very special case of our convergence theory. Convexity of h is not required in our approach. If h is in addition convex, convergence to a single minimum follows with Kiwiel's anchor technique [37], see also [38, 39].

Example 6.2. (Forward-backward splitting). In contrast with the typical splitting literature we do not require convexity of g. If g is lower C^2, so that $g + (1/2\ell_x)\|\cdot - x\|^2$ is convex on $B(x, R)$ for a suitable $\ell_x > 0$, then we choose as first-order model $\phi(y, x) = g(y) + (1/2\ell_x)\|y - x\|^2 + h(x) + \nabla h(x)(y - x)$. Now for $r_x < \ell_x$ we use the second order working model

$$\Phi(y, x) = \phi(y, x) + (1/2r_x - 1/2\ell_x)\|y - x\|^2,$$

then on adding the same constant term $-h(x) + (r_x/2)\|\nabla h(x)\|^2$, and using $\mathscr{O}^{\mathrm{all}}$, we end up with the same tangent program (6.1), where now $r_x < \ell_x$ is required to convexify g in the neighborhood $B(x, R)$ of the current x in which trial steps are taken, assuring a convex tangent program.

But we can do still better, because if g is only lower C^1, we can still use the oracle $\mathscr{O}^{\downarrow}$ for g and the standard model for h, so that a cutting plane is of the form $m^{\downarrow}(\cdot, x) + h(x) + \nabla h(x)(\cdot - x)$. This oracle is strict.

Remark 6.3. In [17, 28] the proximal point method is combined with uncertainty in the computation of $f, \nabla h$, respectively, of ∂g, where Monte-Carlo, respectively, stochastic subgradient techniques cause the uncertainty. This now turns out a special case of our approach [51] to inexact subgradients and values in the non-convex bundle method. In that work we also allow the more challenging situation when the level of uncertainty in the data remains unknown to the user.

6.2. **The standard model.** Let us consider the standard oracle $\mathscr{O}^\sharp = \mathscr{O}_{\phi^\sharp}$, where the cutting plane $(a, g) \in \mathscr{O}^\sharp(z, x)$ requires finding the $g \in \partial f(x)$ with $f^\circ(x, z - x) = g^\mathsf{T}(z - x)$. While this may still be hard to compute in some cases, we get a simplification if x is a point of strict differentiability of f, as then $\partial f(x) = \{\nabla f(x)\}$. In that case $\phi_k(z^k, x) = \phi^\sharp(z^k, x)$, and the secondary test is again redundant. See Borwein and Moors [18] for criteria when a function f is densely or almost everywhere strictly differentiable.

In that case, we automatically have $\phi_k = \phi^\sharp$, so the secondary test is again redundant. We have then arranged that the trust-region tangent program is

$$\min\{f(x^j) + \nabla f(x^j)(y - x) + \tfrac{1}{2}(y - x^j)^\mathsf{T} Q_j(y - x^j) : \|y - x^j\| \le R, y \in C\},$$

which is the classical trust-region method. Convergence of this method, which the user cannot distinguish from classical trust regions, hinges therefore on whether $\mathscr{O}^\sharp$ is strict at the accumulation points of the sequence x^j. As we know, this is almost never the case in non-smooth optimization, which gives us the explanation why classical methods as a rule do not work on non-smooth problems.

References

[1] R.S.S. de Aguiar, P. Apkarian, D. Noll. *Structured robust control against mixed uncertainty.* IEEE Transactions on Control System Technology, **PP(99)**:2017, 1-11.

[2] P. Apkarian, M.N. Dao, D. Noll. *Parametric robust structured control design.* IEEE Transactions on Automatic Control, **60**,(7):2015,1857-1869.

[3] D. Noll, P. Apkarian. *Spectral bundle method for nonconvex maximum eigenvalue functions: first-order methods.* Math. Programming, Series B, **104**:2005,701-727.

[4] D. Noll, P. Apkarian. *Spectral bundle method for nonconvex maximum eigenvalue functions: second-order methods.* Math. Programming, Series B, **104**:2005,729-747.

[5] P. Apkarian, D. Noll. *Non-smooth H_∞ synthesis.* IEEE Transactions on Automatic Control, **51**(1):2006,71-86.

[6] P. Apkarian, D. Noll. *Non-smooth optimization for multidisk H_∞-synthesis.* Eur. J. Control, **12**(3):2006,229-244.

[7] P. Apkarian, D. Noll, *Nonsmooth optimization for multiband frequency domain control design.* Automatica **43**:2007,724-731.

[8] P. Apkarian, D. Noll. *Structured H_∞-control of infinite-dimensional systems.* Int. J. Robust. Nonlin. Control, **28**(9):2018,3212-3238.

[9] P. Apkarian, D. Noll. *Optimization-based control design techniques and tools.* Encyclopedia of Systems and Control DOI 10.1007/978-1-4471-5102-9 _144-1. Springer-Verlag London 2014. Baillieul, John; Samad, Tariq (Editors-in-Chief).

[10] P. Apkarian, D. Noll. *Worst case stability and performance with mixed parametric and dynamic uncertainties.* International Journal of Robust and Nonlinear Control, **27**(8):2017,1284-1301.

[11] P. Apkarian, D. Noll, O. Prot. *A trust region spectral bundle method for non-convex eigenvalue optimization.* SIAM Journal on Optimization, **10**(1):2008,281-306.

[12] P. Apkarian, D. Noll, O. Prot. *A proximity control algorithm to minimize non-smooth and non-convex semi-infinite maximum eigenvalue functions.* Journal of Convex Analysis, **16**:2009,641-666.

[13] P. Apkarian, D. Noll, L. Ravanbod. *Non-smooth bundle trust-region algorithm with applications to robust stability.* Set-Valued and Variational Analysis, **24**(1):2016,115-148.

[14] P. Apkarian, D. Noll, L. Ravanbod. *Non-smooth optimization for robust control of infinite-dimensional systems.* Set-Valued and Variational Analysis, **26**(2):2018,405-429.

[15] P. Apkarian, D. Noll, J.-B. Thevenet and H.D. Tuan. *A spectral quadratic-SDP method with applications to fixed-order H_2 and H_∞ synthesis.* European Journal of Control, **10**(6):2004,527-538.

[16] E. Asplund. *Chebyshev sets in Hilbert space*, Trans. Amer. Math. Soc. **144**:1969,235-240.

[17] Y.F. Atchadé, G. Fort, E. Moulines. *On perturbed proximal gradient algorithms.* `arXiv:1402.2365v4 [math.ST] 19 Nov 2016`.

[18] J.M. Borwein, W.B. Moors. *A chain rule for essentially strictly differentiable Lipschitz functions.* SIAM J. Optim. **8**:1998,300-308.

[19] J. M. Borwein and W. B. Moors. *Null sets and essentially smooth Lipschitz functions.* SIAM Journal on Optimization **8**(2):1998,309-323

[20] J.M. Borwein, Q.J. Zhu. *A survey of subdifferential calculus with applications*, Nonlinear Anal., **38**:1999,687-773.

[21] S. Boyd, V. Balakrishnan. *A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm for computing its L_∞-norm.* Syst. Control Letters, **15**:1990,1-7.

[22] S. Boyd, V. Balakrishnan, P. Kabamba. *A bisection method for computing the H_∞-norm of a transfer matrix and related problems.* Mathematics of Control, Signals, and Systems, **2**:1989,207-219.

[23] M. N. Dao. *Bundle method for nonconvex nonsmooth constrained optimization.* Journal of Convex Analysis **22**(4):2015,1061-1090.

[24] A. Daniilidis, P. Georgiev. *Approximate convexity and submonotonicity.* J. Math. Anal. Appl., **291**:2004,117-144.

[25] M.N. Dao, P. Apkarian, D. Noll. *Robust eigenstructure clustering by non-smooth optimization.* International Journal of Control, **88**(8):2015,1441-1455.

[26] M.N. Dao, J. Gwinner, D. Noll, N. Ovcharova. *Non-convex bundle method with application to a delamination problem.* Computational Optimization and Applications, **65**(1):2016,173-203.

[27] M.N. Dao, D. Noll. *Minimizing memory effects of a system.* Mathematics of Control, Signals and Systems, **27**(1):2015,77-110.

[28] D. Davis, B. Grimmer. *Proximally guided stochastic subgradient method for non-smooth, nonconvex problems.* `arXiv:1707.03505v4[math.OC]17 Oct 2017`.

[29] M. Gabarrou, D. Alazard, D. Noll. *Design of a flight control architecture using a non-convex bundle method.* Mathematics of Control, Signals, and Systems, **25**(2):2013,257-290.

[30] W. Hare, C. Sagastizábal. *A redistributed proximal bundle method for non-convex optimization.* SIAM J. Optim., **20**(5):2010,2442 - 2473.

[31] C. Helmberg, K.C. Kiwiel. *A spectral bundle method with bounds.* Math. Programming, **93**:2002,173-194.

[32] C. Helmberg, F. Oustry. *Bundle methods to minimize the maximum eigenvalue function.* Handbook of Semidefinite Programming. Theory, Algorithms and Applications. L. Vandenberghe, R. Saigal, H. Wolkowitz (eds.), vol. 27, 2000.

[33] C. Helmberg, F. Rendl. *Spectral bundle method for semidefinite programming.* SIAM J. Optimization, **10**:2000,673-696.

[34] L. Hertlein, M. Ulbrich, *An inexact bundle algorithm for nonconvex nondifferentiable functions in Hilbert space*, Deutsche Forschungsgemeinschaft (DFG). Priority Program 1962. Preprint. SPP1962-084 (2018) 1–27.

[35] J.-B. Hiriart-Urruty. *Generalized differentiability, duality and optimization for problems dealing with differences of convex functions.* Lect. Notes in Economics and Math. Systems **256**:1985,37-70.

[36] C. Kiwiel. *An aggregate subgradient method for non-smooth convex minimization.* Math. Programming, **27**:1983,320-341.

[37] K.C. Kiwiel. *Proximity control in bundle methods for convex nondifferentiable optimization.* Math. Programming, **46**:1990,105-122.

[38] K.C. Kiwiel. *Methods of descent for nondifferentiable optimization*, vol. 1133 of Springer Lect. Notes in Math., Springer Verlag, 1985.

[39] K.C. Kiwiel. *A linearization algorithm for computing control systems subject to singular value inequalities.* IEEE Trans. Autom. Control, **AC-31**:1986,595-602.

[40] C. Lemaréchal. *An extension of Davidson's method to nondifferentiable problems.* Math. Programming Studies, Nondifferentiable Optimization, M.L. Balinski and P. Wolfe (eds.), North Holland, 1975, pp. 95 - 109.

[41] C. Lemaréchal. *Bundle methods in nonsmooth optimization.* Nonsmooth Optimization, Proc. IIASA Workshop 1977, C. Lemaréchal, R. Mifflin (eds.), 1978.

[42] C. Lemaréchal, F. Oustry. *Nonsmooth algorithms to solve semidefinite programs.* SIAM Series: Advances in Linear Matrix Inequality Methods in Control, pp 57 - 77, Society for Industrial and Applied Mathematics Philadelphia, PA, USA 2000.

[43] C. Lemaréchal, C. Sagastizábal. *Variable metric bundle methods: from conceptual to implementable forms.* Mathematical Programming **76**:1997, 393-410.

[44] D.T. Luc, H. Van Ngai, M. Théra. *On ϵ-monotonicity and ϵ-convexity. Calculus of Variations and Differential Equations*, (Haifa 1998), Chapman & Hall/CRC Research Notes in Mathematics, **410**:2000,82-100.

[45] D. Mayne, E. Polak. *Algorithms for the design of control systems subject to singular value inequalities.* Math. Programming Stud., **18**:1982,112-134.

[46] D. Mayne, E. Polak, A. Sangiovanni. *Computer aided design via optimization.* Automatica, **18**:1982,147-154.

[47] R. Mifflin. *An algorithm for constrained optimization with semi-smooth functions.* Mathematics of Operations Research, **2**(2):1977,191-207.

[48] R. Mifflin. *A modification and an extension of Lemaréchal's algorithm for nonsmooth optimization,* Math. Prog. Stud. **17**:1982,77-90.

[49] D. Noll. *Cutting plane oracles to minimize non-smooth non-convex functions.* Set-Valued Var. Anal., **18**(3-4):2010,531-568.

[50] D. Noll. *Convergence of non-smooth descent methods using the Kurdyka-Łojasiewicz inequality.* Journal of Optimization Theory and Applications, **160**(2):2014,553-572.

[51] D. Noll. *Bundle method for non-convex minimization with inexact subgradients and function values,* Computational and Analytical Mathematics. Springer Proceedings in Mathematics & Statistics, **50**:2013,555-592.

[52] D. Noll, O. Prot, A. Rondepierre. *A proximity control algorithm to minimize nonsmooth and non-convex functions.* Pac. J. Optim. **4**(3):2008,571-604.

[53] E. Polak, Y. Wardi. *A nondifferential optimization algorithm for the design of control systems subject to singular value inequalities over the frequency range.* Automatica, **18**:1982,267-283.

[54] E. Polak, S. Salcudean. *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in H_∞ space.* IEEE Trans. Autom. Control, **AC-34**:1989,268-276.

[55] L. Ravanbod, D. Noll, P. Apkarian. *Computing the structured distance to instability.* Proceedings of the SIAM Conference on Control and Applications, Paris, 2015, pp. 423 - 430.

[56] L. Ravanbod, D. Noll, P. Apkarian. *Branch and bound algorithm for the robustness analysis of uncertain systems.* Proc. 16th IFAC Workshop on Control Applications of Optimization. IFAC-PapersOnLine, **48**(25):2015,85-90.

[57] L. Ravanbod, D. Noll, P. Apkarian. *An extension of the linear quadratic Gaussian loop transfer recovery procedure.* IET Control Theory and Applications, **6**(14):2012,2269-2278.

[58] L. Ravanbod, D. Noll, J.-P. Raymond, J.-M. Buchot. *Robustified H_2-control of a system with large state dimension.* European Journal of Control, **31**:2016,59-71.

[59] R.T. Rockafellar, R.J.-B. Wets. *Variational Analysis.* Springer Verlag (2004).

[60] A. Ruszczyński. *Nonlinear Optimization.* Princeton University Press, 2007.

[61] C. Sagastizábal. *Composite proximal bundle method.* Math. Progr., **140**(1):2013,189-233.

[62] H. Schramm, J. Zowe. *A version of the bundle idea for minimizing a non-smooth function: conceptual ideas, convergence analysis, numerical results.* SIAM J. Opt., **2**:1992,121-152.

[63] A. Simoes, P. Apkarian, D. Noll. *Non-smooth multi-objective synthesis with applications.* Control Engineering Practice, **17**(11):2009,1338-1348.

[64] J.E. Spingarn. *Submonotone subdifferentials of Lipschitz functions.* Trans. Amer. Math. Soc., **264**:1981,77-89.

[65] J.-B. Thevenet, D. Noll and P. Apkarian. *Nonlinear spectral SDP method for BMI-constrained problems: applications to control design.* Informatics in Control, Automation and Robotics I, J. Braz, H. Arajo A. Viera and B. Encarnaco (eds.), Springer 2006, 61 - 72.

[66] J.-B. Thevenet, D. Noll and P. Apkarian. *Nonsmooth methods for large bilinear matrix inequalities: applications to feedback control.* Submitted.

[67] J. Toland. *Duality in nonconvex optimization.* J. Math. Anal. Appl., **66**:1978,399-415.

[68] J. Toland. *A duality principle for nonconvex optimization and the calculus of variations.* Arch. Rational Mech. Anal., **71**:1979,41-61.

[69] J. Zowe. *The BT-algorithm for minimizing a non-smooth functional subject to linear constraints.* Non-smooth Optimization and Related Topics, F.H. Clarke, V.F. Demyanov, F. Gianessi (eds.), Plenum Press, 1989.

(D. Noll) Université de Toulouse, Institut de Mathématiques, France
Email address: dominikus.noll@math.univ-toulouse.fr