Cutting plane oracles to minimize non-smooth non-convex functions

Dominikus Noll*

Abstract

We discuss a bundle method for non-smooth non-convex optimization programs. In the absence of convexity, a substitute for the cutting plane mechanism has to be found. We propose such a mechanism and prove convergence of our method in the sense that every accumulation point of the sequence of serious iterates is critical.

Keywords. Non-smooth optimization, bundle method, cutting plane oracle.
AMS classification. 49J52, 90C26, 90C22, 93B36.

1 Introduction

We consider optimization programs of the form

$$\min_{x\in\mathbb{R}^n} f(x), \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is non-smooth, non-convex, and locally Lipschitz. We present a bundle algorithm for (1), which converges globally in the sense that every accumulation point $\bar{x}$ of the sequence of serious steps is critical, that is, satisfies

$$0\in\partial f(\bar{x}), \tag{2}$$

where $\partial f(x)$ is the Clarke subdifferential of f at x. Non-convexity of (1) leads to three major difficulties. Firstly, cutting planes can no longer be used as in the convex case, and a substitute has to be found. Secondly, recycling of affine support planes between serious steps has to be adapted to the new context. And thirdly, a more sophisticated management of the proximity control mechanism is required to obtain a satisfactory convergence theory. We will show in which way these elements can be addressed and combined into a successful algorithm.

The present work follows a line of investigation initiated in [2, 5] and continued in [3, 4, 6, 7], where non-smooth algorithms of bundle type are used to solve difficult problems in feedback control design. In [17] we have presented a non-convex bundle algorithm in a more abstract form. Here we expand on [17] in at least two important points. We refine the management of the proximity control parameter, and we introduce the concept of a cutting plane oracle in order to give convergence proofs for several heuristic techniques in non-smooth optimization.

*Université Paul Sabatier, Institut de Mathématiques, Toulouse, France

2 First and second order models

The concept of a local model $\phi(\cdot, x)$ in the neighbourhood of the current iterate x will be central for our approach.

Definition 1. *A function $\phi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a first-order model of f on $\Omega \subset \mathbb{R}^n$ if $\phi(\cdot, x)$ is convex for every fixed $x \in \Omega$, and if the following axioms are satisfied:*

(M_1) *$\phi(x, x) = f(x)$ and $\partial_1\phi(x, x) \subset \partial f(x)$.*

(M_2) *For every sequence $y_j \to x$ there exist $\epsilon_j \to 0^+$ such that $f(y_j) \leq \phi(y_j, x) + \epsilon_j\|y_j - x\|$ for all $j \in \mathbb{N}$.*

(M_3) *For sequences $y_j \to y$ in $\mathbb{R}^n$ and $x_j \to x$ in Ω one has $\limsup_{j\to\infty} \phi(y_j, x_j) \leq \phi(y, x)$.*

If $\Omega = \mathbb{R}^n$ then we simply call ϕ a first-order model of f.

Remark 1. Every locally Lipschitz function has the standard (or Clarke) model

$$\phi^\sharp(y, x) = f(x) + f^0(x, y - x),$$

where $f^0(x, d)$ is the Clarke directional derivative of f at x in direction d. Indeed, axiom (M_1) is immediate because $\partial_1\phi^\sharp(x, x) = \partial_2 f^0(x, 0) = \partial f(x)$. Axiom ($M_2$) follows directly from the definition of the Clarke directional derivative, and (M_3) uses upper semicontinuity of ∂f.

Lemma 1. *Every first-order model ϕ satisfies $\partial_1\phi(x, x) = \partial f(x)$.*

Proof: (a) Let us first prove that if f is differentiable at x, then $\nabla f(x) \in \partial_1\phi(x, x)$. Indeed, for fixed $d \in \mathbb{R}^n$ and $t > 0$ we have

$$t^{-1}\left(f(x + td) - f(x)\right) \leq t^{-1}\left(\phi(x + td, x) + \epsilon_t t\|d\| - f(x)\right)$$

by axiom (M_2), where $\epsilon_t \to 0$ as $t \to 0^+$. Passing to the limit we find $\nabla f(x)^\top d \leq \phi'(x, x, d)$, where $\phi'(x, x, d)$ is the directional derivative of $\phi(\cdot, x)$ at x in direction d. Since the left hand term is linear in d, and since $\phi'(x, x, \cdot)$ is the support function of $\partial_1\phi(x, x)$, we have $\nabla f(x) \in \partial_1\phi(x, x)$.

(b) Let us now show $\partial f(x) \subset \partial_1\phi(x, x)$. Since $\partial f(x)$ is the convex hull of the limiting subdifferential $\partial_a f(x)$, it suffices to show $\partial_a f(x) \subset \partial_1\phi(x, x)$. Let $g \in \partial_a f(x)$. Then there exists a sequence $x_j \to x$ such that f is differentiable at x_j and $g_j = \nabla f(x_j) \to g$. By part (a) we know that $g_j \in \partial_1\phi(x_j, x_j)$. By the subgradient inequality this means $g_j^\top d \leq \phi(x_j + d, x_j) - \phi(x_j, x_j)$ for every test vector d. Now $\phi(x_j, x_j) = f(x_j)$ by (M_1), so passing to the limit using $f(x) = \phi(x, x)$ gives $g^\top d \leq \limsup_{j\to\infty} \phi(x_j + d, x_j) - \phi(x, x) \leq \phi(x + d, x) - \phi(x, x)$, where the last estimate uses axiom (M_3). Since this holds for every d, we have shown $g \in \partial_1\phi(x, x)$. □

Lemma 2. *The standard model $\phi^\sharp$ is the smallest first-order model of f. That is, any other model ϕ satisfies $\phi^\sharp \leq \phi$.*

Proof: It follows from convexity of $\phi(\cdot, x)$ and (M_1) that $\phi(y,x) \geq f(x)+\phi'(x,x,y-x)$ for every y, where $\phi'(x,x,d)$ is the directional derivative of $\phi(\cdot,x)$ at x in direction d. But $\phi'(x,x,d) = f^0(x,d)$ by Lemma 1. Namely, $\phi'(x,x,\cdot)$ is the support function of $\partial_1\phi(x,x)$, $f^0(x,\cdot)$ the support function of $\partial f(x)$, and the two sets coincide by Lemma 1. Therefore $\phi(y,x) \geq f(x)+f^0(x,y-x) = \phi^\sharp(y,x)$ for every y. □

We will also need the following variants of definition 1.

Definition 2. *A first-order model $\phi(y,x)$ of f on Ω is called* strict *if axiom (M_2) is replaced by the stronger axiom*

$(\widehat{M_2})$ *Given sequences $y_j \to x$ in $\mathbb{R}^n$ and $x_j \to x$ in Ω there exists a sequence $\epsilon_j \to 0^+$ such that $f(y_j) \leq \phi(y_j,x_j) + \epsilon_j\|y_j - x_j\|$.*

The first-order model is called strong *if the following even stronger axiom is satisfied:*

$(\widetilde{M_2})$ *Given sequences $y_j \to x$, $x_j \to x$ in Ω, there exists $L > 0$ such that $f(y_j) \leq \phi(y_j,x_j)+L\|y_j - x_j\|^2$ for every $j \in \mathbb{N}$.*

Remark 2. Notice that (M_2) means $f(y) \leq \phi(y,x) + \mathrm{o}(\|y-x\|)$ for $y \to x$. Strictness $(\widehat{M_2})$ is $f(y) \leq \phi(y,x) + \mathrm{o}(\|y-x\|)$ as $y - x \to 0$ uniformly on bounded sets. Strongness $(\widetilde{M_2})$ means $f(y) \leq \phi(y,x) + \mathrm{O}(\|y-x\|^2)$ as $y - x \to 0$ uniformly on bounded sets. The difference between $(\widehat{M_2})$ and $(\widetilde{M_2})$ is therefore analogous to the difference between the first-order Taylor-Lagrange and the Taylor-Young formulas.

Remark 3. Suppose f is differentiable, then $\phi(y,x) = f(x) + \nabla f(x)^\top(y-x)$ seems a natural candidate for a model. But it is *not* a model in general, because axiom (M_3) fails. In order to have (M_3) one needs class C^1, and then ϕ is indeed the standard model. Since class C^1 means f is strict differentiability, we see that this model is then even strict. (This motivated the choice of the term strict for models with property $(\widehat{M_2})$). Finally, the Taylor expansion is a strong model as soon as f is class of $C^{1,1}$.

Remark 4. Every strong model is strict, and every strict model is a model. None of these are reversible. Consider $f(x) = x^2 \sin x^{-1}$ on the real line, then $\phi^\sharp$ is not strict on any neighbourhood of 0. On the other hand, if $f \in C^1 \setminus C^{1,1}$, then the standard model $\phi^\sharp(y,x) = f(x)+\nabla f(x)^\top(y-x)$ is strict but not strong.

Remark 5. If f is convex, then $\phi(\cdot,x) = f$ is a strong model. We say that a convex function is its own strong model.

Remark 6. How about concave functions? Consider $-f$, where f is convex. Let $g \in \partial f(x)$, then the subgradient inequality gives $g^\top(y-x) \leq f(y) - f(x)$, or what is the same,

$$-f(y) \leq -f(x) + g^\top(x-y) \leq -f(x) + f^0(x,x-y) = -f(x) + (-f)^0(x,y-x) = \phi^\sharp(y,x).$$

That means that the standard model is strong for concave functions. More generally, we have the following result.

Proposition 1. *Suppose f is upper C^2. Then its standard model $\phi^\sharp$ is strong. If f is upper C^1, then $\phi^\sharp$ is strict.*

Proof: 1) Recall that f is upper C^k if $-f$ is lower C^k. For the definition of lower C^k see [18].

2) Let us discuss the upper C^2 case first. According to [18, Prop. 13.33] $(-f)$ is prox-regular at every $\bar{x} \in \mathbb{R}^n$ with respect to every $\bar{g} \in \partial(-f)(\bar{x})$. That means there exists $\epsilon > 0$ and $r > 0$ such that for all $x, x' \in B(\bar{x}, \epsilon)$ and every $g(x) \in \partial(-f)(x)$ with $\|g(x) - \bar{g}\| \leq \epsilon$ one has

$$|f(x) - f(\bar{x})| < \epsilon \implies -f(x') \geq -f(x) + g(x)^\top (x' - x) - \tfrac{r}{2}\|x' - x\|^2.$$

That is the same as

$$\begin{aligned} f(x') &\leq f(x) - g(x)^\top (x' - x) + \tfrac{r}{2}\|x' - x\|^2 \\ &\leq f(x) + \sup_{-g \in \partial f(x)} (-g)^\top (x' - x) + \tfrac{r}{2}\|x' - x\|^2 = f(x) + f^0(x, x' - x) + \tfrac{r}{2}\|x' - x\|^2, \end{aligned}$$

which proves $(\widetilde{M}_2)$ with $L = \frac{r}{2}$ on the ball $B(\bar{x}, \epsilon)$.

3) Now consider the case where $(-f)$ is lower C^1. According to Daniilidis and Georgiev [12] $(-f)$ has the following property called approximate convexity: For every $\bar{x} \in \mathbb{R}^n$ and $\epsilon > 0$ there exists $\delta > 0$ such that $(-f)(ty + (1-t)x) \leq t(-f)(y) + (1-t)(-f)(x) + \epsilon t(1-t)\|x - y\|$ for all $x, y \in B(\bar{x}, \delta)$ and $0 \leq t \leq 1$. This can be re-arranged into

$$f(y) \leq f(x) + \frac{f(x + t(y - x)) - f(x)}{t} + \epsilon(1 - t)\|x - y\|.$$

Passing to the limit $t \to 0^+$ gives the estimate $f(y) \leq \phi^\sharp(y, x) + \epsilon\|x - y\|$ for every $y \in B(x, \delta)$, hence $(\widehat{M}_2)$. □

Remark 7. Suppose f is a composite function $f = h \circ F$, where h is convex and F of class C^1. Then a natural first order model for f is

$$\phi(y, x) = h\left(F(x) + F'(x)(y - x)\right).$$

Notice that ϕ is strict because $F(y) - [F(x) + F'(x)(y - x)] = \mathrm{o}(\|y - x\|)$ and because h is locally Lipschitz. We sometimes call ϕ the *natural* model of f. Observe that ϕ is strong if $F \in C^{1,1}$.

A typical application of this is eigenvalue optimization $f = \lambda_1 \circ F$, where F is usually smooth. The natural strong model is then $\phi(y, x) = \lambda_1\left(F(x) + F'(x)(y - x)\right)$. The standard model $\phi^\sharp$ in contrast is only strong at those x where the maximum eigenvalue of $F(x)$ has multiplicity 1.

Remark 8. Consider a lower C^2 function. For every bounded set B there exists a constant $\mu_0 > 0$ such that for every $x \in B$ and $\mu \geq \mu_0$ the function $\phi_\mu(y, x) = f(y) + \mu\|y - x\|^2$ is convex in $y \in B$ [18, Prop. 13.33]. Therefore each ϕ_μ with $\mu \geq \mu_0$ is a strong model of f on B.

Notice that f can be re-written as $f = h_\mu \circ F_\mu$ with h_μ convex and F_μ of class C^2, namely, $h_\mu(x, y) = y + f(x) + \frac{\mu}{2}\|x\|^2$ and $F_\mu(x) = (x, -\frac{\mu}{2}\|x\|^2)$, where $\mu \geq \mu_0$. Then ϕ_μ turns out to be the natural model of $h_\mu \circ F_\mu$ on B in the sense of remark 7.

Remark 9. Let $f = f_1 + f_2$, where f_1 is lower C^2 and f_2 is upper C^2. This applies in particular to convex differences. Then a natural candidate for a strong model of f is

$$\phi_\mu(y, x) = f_1(y) + \mu\|y - x\|^2 + f_2(x) + f_2^0(x, y - x), \tag{3}$$

because $f_2(x) + f_2^0(x, y - x)$ is the standard model of f_2, which is strong by Proposition 1, and because $f_1 + \mu\| \cdot -x\|^2$ is convex and therefore a strong model of f_1 if μ is as in remark 8. Is ϕ_μ a strong model of f?

What is missing to guarantee this is not strongness, but the property $\partial_1\phi_\mu(x, x) \subset \partial f(x)$. We have $\partial_1\phi_\mu(x, x) \subset \partial f_1(x) + \partial f_2(x)$. Now everything would be fine if we had $\partial f_1(x) + \partial f_2(x) \subset \partial(f_1 + f_2)(x)$. Unfortunately the sum rule for the Clarke generalized derivative goes the opposite way $\partial f(x) \subset \partial f_1(x) + \partial f_2(x)$, see [9, Prop. 2.3.3]. Equality holds if either f_1 or f_2 is strictly differentiable [9, Cor. 1, page 39] or if f_1, f_2 are regular in the sense of Clarke. The latter is hopeless here, because f_2 has no chance to be regular in the sense of Clarke.

Is this where (3) is finished off? Not at all, because we can consider that the sum rule holds with equality except some pathological cases. Moreover, the following argument may be put forward in favour of (3). If one defines $\bar{x}$ to be a cd-critical point of $f = f_1 + f_2$ if $\partial f_1(\bar{x}) \cap (-\partial f_2(\bar{x})) \neq \emptyset$, then all the remaining theory will be valid for model (3), only the algorithm will be stopped if an iterate x^j with $\partial f_1(x^j) \cap (-\partial f_2(x^j)) \neq \emptyset$ will be found, and the inner loop will only be initiated if x^j is not a cd-critical point of f. The sequence of serious iterates x^j will converge to such a cd-critical point.

We conclude this section with the definition of a second-order model.

Definition 3. *Let $\phi(y, x)$ be a first-order model of f. Then $\Phi(y, x) = \phi(y, x) + \frac{1}{2}(y-x)^\top Q(x)(y-x)$ is called a second-order model of f if $Q(x) \in \mathbb{S}^n$ is bounded on bounded sets of x.*

3 First- and second order working model

Having defined first- and second-order models $\phi(y, x)$, $\Phi(y, x)$, the difficulty is that computations based on $\phi(y, x)$ may be too costly. We therefore use an approximation $\phi_k(y, x)$ of $\phi(y, x)$, which we call the first-order working model. $\phi_k(y, x)$ is indexed by $k \in \mathbb{N}$, because it is updated iteratively during the inner loop with counter k. We sometimes call $\phi(y, x)$ the ideal first-order model, because it is what we would ideally like to use. For the working model ϕ_k we require the following conditions:

Definition 4. *The function ϕ_k is a first order working model of f at x associated with the ideal first-order model ϕ of f if $\phi_k(\cdot, x)$ is convex and satisfies $\phi_k(\cdot, x) \leq \phi(\cdot, x)$, $\phi_k(x, x) = f(x)$.*

Notice that $\partial_1\phi_k(x, x) \subset \partial f(x)$ for every k as a consequence of the fact that $\partial_1\phi(x, x) = \partial f(x)$ and $\phi_k \leq \phi$. An advantage of the working model is that $\partial_1\phi_k(x, x)$ can be a very small subset of $\partial f(x)$, while $\partial_1\phi(x, x) = \partial f(x)$ by Lemma 1. This is important in cases where $\partial f(x)$ is too large to be computed efficiently.

Definition 5. *Let ϕ_k be a first-order working model for f associated with the ideal first-order model ϕ. Let $\Phi(y, x) = \phi(y, x) + \frac{1}{2}(y - x)^\top Q(x)(y - x)$ be a second-order model associated with ϕ. Then $\Phi_k(y, x) = \phi_k(y, x) + \frac{1}{2}(y - x)^\top Q(x)(y - x)$ is the corresponding second-order working model.*

Remark 10. Our notation suggests an important detail. The first-order working model ϕ_k is updated iteratively in the inner loop k, while the second order term $Q(x)$ depends only on the serious iterate x and remains unchanged during the inner loop k. Updating $Q(x) \to Q(x^+)$ happens only in the outer loop when a serious step $x \to x^+$ is made. The reason for this will become clear in section 9.

4 Elements of the algorithm

In this section we will discuss the different elements of the algorithm, which itself will be presented in section 5.

4.1 Tangent program

Let $x \in \mathbb{R}^n$ be the current serious iterate. The inner loop now turns until a new serious iterate x^+ is found. At inner loop counter k we dispose of a first-order working model $\phi_k(\cdot, x)$ and the corresponding second order working model $\Phi_k(y, x) = \phi_k(y, x) + \frac{1}{2}(y-x)^\top Q(x)(y-x)$, where $Q(x)$ depends on x, but is fixed during the inner loop k. Now we compute a solution y^{k+1} to the tangent program with proximity control

$$\min_{y \in \mathbb{R}^n} \Phi_k(y, x) + \tfrac{\tau_k}{2}\|y - x\|^2, \tag{4}$$

where $\tau_k > 0$ is the so-called *proximity control parameter*. We assume throughout that $Q(x)+\tau_k I \succ 0$, so that (4) is strictly convex and has a unique solution y^{k+1}. We refer to y^{k+1} as the *trial step*. The optimality condition $0 \in \partial_1 \Phi_k(y^{k+1}, x)$ can be re-written as

$$g^*_{k+1} := (Q(x) + \tau_k I)\left(x - y^{k+1}\right) \in \partial_1 \phi_k(y^{k+1}, x). \tag{5}$$

It is standard to call g^*_{k+1} the *aggregate subgradient*. Its use is explained in section 4.4 below.

4.2 Acceptance

In order to decide whether the trial step y^{k+1} is acceptable to become the next serious iterate x^+, step 5 of the algorithm computes the quotient

$$\rho_k = \frac{f(x) - f(y^{k+1})}{f(x) - \Phi_k(y^{k+1}, x)}, \tag{6}$$

which tests agreement between f and $\Phi_k(\cdot, x)$ at y^{k+1}. If agreement is good, we expect $\rho_k \approx 1$. In order to obtain a decision we fix constants $0 < \gamma < \Gamma < 1$. We say that the agreement between f and Φ_k at y^{k+1} is *good* if $\rho_k > \Gamma$, where the reader might for instance imagine $\Gamma = \frac{3}{4}$. On the other hand, we say that the agreement is *bad* if $\rho_k < \gamma$, where $\gamma = \frac{1}{4}$ makes sense. We accept the trial step y^{k+1} already as the new serious iterate x^+ if $\rho_k \geq \gamma$, that is, if it is *not bad*. Standard terminology in bundling refers to this as a *serious step*. In that case the inner loop ends.

The delicate case is when agreement is bad, that is, $\rho_k < \gamma$. Here y^{k+1} is rejected and referred to as a null step. Now the inner loop k has to continue. In order to do better at the next sweep, model Φ_{k+1} has to improve over Φ_k. This is achieved by two mechanisms, known as *aggregation* and *cutting planes*. A third mechanism, which is not needed in the convex case, but is mandatory in the absence of convexity is updating the proximity control parameter τ_k in an intelligent way. Having arranged these three elements, we will increase counter k and solve (4) again, hoping for a better trial step y^{k+2} at the next iteration.

4.3 Exactness

Our working models $\phi_k(\cdot, x)$ need to satisfy $\partial_1 \phi_k(x, x) \subset \partial f(x)$ and $\phi_k(x, x) = f(x)$ at all times k. To guarantee this, we pick a subgradient $g(x) \in \partial f(x)$ and assure that $m_e(y, x) = f(x) + g(x)^\top (y - x)$ is an affine minorant of $\phi_k(\cdot, x)$ at all times k. We refer to $m_e(\cdot, x)$ as the exactness plane at x.

4.4 Aggregation

Let us now explain aggregation. As we have seen, optimality (5) gives the aggregate subgradient g^*_{k+1}. We also refer to

$$m^*_{k+1}(y,x) = \phi_k(y^{k+1},x) + g^{*\top}_{k+1}(y-y^{k+1})$$

as the *aggregate plane*. It is an affine support function of $\phi_k(\cdot,x)$ at y^{k+1} and could also be written as $m^*_{k+1}(y,x) = a^*_{k+1} + g^{*\top}_{k+1}(y-x)$, where $a^*_{k+1} = \phi_k(y^{k+1},x) + g^{*\top}_{k+1}(x-y^{k+1})$. We will assure that the new working model $\phi_{k+1}(\cdot,x)$ has $m^*_{k+1}(\cdot,x)$ as an affine minorant. The consequence is:

Lemma 3. *Suppose the new first-order working model $\phi_{k+1}(\cdot,x)$ satisfies $m^*_{k+1}(\cdot,x) \le \phi_{k+1}(\cdot,x)$. Then $\phi_{k+1}(y^{k+1},x) \ge \phi_k(y^{k+1},x)$. Moreover, in that case condition (5) is satisfied.*

Aggregation follows the usual lines as introduced in [10].

4.5 Cutting planes

The next element which is fundamental in bundle methods is inclusion of a cutting plane $m_{k+1}(\cdot,x)$ among the affine support planes of the new working model $\phi_{k+1}(\cdot,x)$. The role of the cutting plane $m_{k+1}(\cdot,x)$ is to cut away the unsuccessful trial step y^{k+1}. The idea is that if we let $\phi_{k+1}(\cdot,x) \ge m_{k+1}(\cdot,x)$, then y^{k+1} is no longer solution of the tangent program. If f is convex, then this is indeed what happens. The cutting plane is simply an affine support function of f at y^{k+1}. We put $m_{k+1}(y,x) = f(y^{k+1}) + g^\top_{k+1}(y-y^{k+1})$, where $g_{k+1} \in \partial f(y^{k+1})$, and the information $(f(y^{k+1}), g_{k+1})$ is referred to as the oracle of f at y^{k+1}.

Without convexity tangents to f at y^{k+1} may be useless, so cutting planes cannot be obtained that way. We need to elaborate a substitute, and this is the first major complication due to non-convexity. In [17] the following approach was chosen: Use the ideal first-order model $\phi(\cdot,y)$ as a substitute and take tangents of $\phi(\cdot,x)$ at y^{k+1} rather than tangents of f. Here we go even further and consider cutting planes as of furnished by an abstract process satisfying certain axioms. The advantage of this will become more evident when we discuss applications.

Definition 6. *A cutting plane oracle for f on Ω is a bounded mapping which with every triplet $(k,y^+,x) \in \mathbb{N}\times\mathbb{R}^n\times\Omega$ associates an affine function $m_{k,y^+,x}(y) = a + g^\top(y-x)$, called a cutting plane at serious step x, trial step y^+, and counter k, such that the following conditions are satisfied:*

(C_1) *For $y^+ = x$ we have $a = f(x)$ and $g \in \partial f(x)$ for every $k \in \mathbb{N}$.*

(C_2) *Suppose $y^+_j \to x$, and $k_j \in \mathbb{N}$. Suppose the cutting plane is drawn at trial point y^+_j for serious point x at counter k_j. Then there exist $\epsilon_j \to 0^+$ such that $f(y^+_j) \le m_{k_j,y^+_j,x}(y^+_j) + \epsilon_j\|y^+_j - x\|$ for every j.*

(C_3) *Suppose $y^+_j \to y^+$, $y_j \to y$, and $x_j \to x$. Then there exist z^+, bounded in x,y,y^+, and $\ell \in \mathbb{N}$ such that one has $\limsup_{j\to\infty} m_{k_j,y^+_j,x_j}(y_j) \le m_{\ell,z^+,x}(y)$.*

We say that the cutting plane oracle is strict *if the following stronger version of (C_2) is satisfied:*

($\widehat{C}_2$) *Suppose $y^+_j \to x$, $x_j \to x$, and $k_j \in \mathbb{N}$. Suppose the cutting plane is drawn at trial point y^+_j for center point x_j at instant k_j. Then there exist $\epsilon_j \to 0^+$ such that $f(y^+_j) \le m_{k_j,y^+_j,x_j}(y^+_j) + \epsilon_j\|y^+_j - x_j\|$ for every j.*

The cutting plane oracle is called strong, *if the following even stronger variant of* (C_2) *holds:*

$(\widetilde{C}_2)$ *Suppose* $y_j^+ \to x$, $x_j \to x$, *and* $k_j \in \mathbb{N}$. *Suppose the cutting plane is drawn at trial point* y_j^+ *for center point* x_j *at instant* k_j. *Then there exists* $L > 0$ *such that* $f(y_j^+) \le m_{k_j,y_j^+,x_j}(y_j^+) + L\|y_j^+ - x_j\|^2$ *for every* j.

Consider all possible cutting planes $m_{k,y^+,x}(\cdot)$ arising at the current serious iterate x and at *all* possible trial points y^+ in a large but bounded neighbourhood $B(x, M)$ of x and *all* counters $k \in \mathbb{N}$. Then for fixed x the expression

$$\phi^\uparrow(y, x) := \sup\{m_{k,y^+,x}(y) : y^+ \in \mathbb{R}^n, \|y^+ - x\| \le M, k \in \mathbb{N}\}$$

defines a convex function of the argument y, the convex envelope of all oracle planes at x. Since the plane $m_{k,y^+,x}$ and also z^+ in (C_3) are bounded in the data (y^+, x) independently of k, $\phi^\uparrow$ is finite everywhere. Axioms (C_1) - (C_3) assure that $\phi^\uparrow$ is a first-order model of f. If axiom $(\widehat{C}_2)$ is satisfied, then $\phi^\uparrow$ is strict, and if $(\widetilde{C}_3)$ is satisfied, then $\phi^\uparrow$ is strong.

Definition 7. *We call* $\phi^\uparrow$ *the upper envelope model associated with the cutting plane oracle, or simply the upper envelope model.*

In order to understand our axiomatic approach, we need to discuss examples.

Example 1. Tangent cutting planes. Given an ideal model $\phi(y, x)$ of f, the most natural way to draw cutting planes is to take the tangents of ϕ. That is, $m_{k,y^+,x}(y^+) = \phi(y^+, x)$ and $\nabla m_{k,y^+,x} \in \partial_1\phi(y^+, x)$. This approach is discussed in [17]. Here the upper envelope model $\phi^\uparrow$ coincides with ϕ. For convex $f = \phi(\cdot, x)$ we recover the standard form of the oracle, where all cutting planes are tangent planes of f.

Example 2. Cutting planes for f lower C^2. Our next example is motivated by [17]. Suppose f is lower C^2. Then for every x there exists a neighbourhood U of x and $\mu_0 > 0$ such that for every $x' \in U$ and $\mu \ge \mu_0$ $\phi_\mu(y, x') = f(y) + \mu\|y - x'\|^2$ is convex in y. That means, each of these $\phi_\mu(\cdot, x')$ could be chosen as ideal model in the sense of the previous example. One would obviously like to adapt μ at each step, taking the smallest μ which convexifies f. This approach is proposed in [17], and underlies the algorithms in Sagastizábal and Hare [19] and Sagastizábal [20] for lower C^2 functions.

So the situation is slightly more elaborate than in example 1 in so far as we have a whole family of models ϕ_μ at our disposal, and we will draw tangent planes from any one of them, changing μ between the null steps of the inner loop if need be. All that is required is that the set of μ visited during this procedure remains bounded. If this is the case, then $\phi^\uparrow(y, x) = \sup\{\phi_\mu(y, x) : \mu$ used in the procedure$\}$ is the upper envelope model in the sense of Definition 7. The reader will notice the difference with example 1. The cutting planes are all below $\phi^\uparrow$, *but need not be tangents to* $\phi^\uparrow$. Yet $\phi^\uparrow$ has the same structure as the individual ϕ_μ, so we could at any moment use $\phi^\uparrow$ itself to draw tangents, which would bring us straight back to example 1.

Example 3. Downshift planes. Our third example is motivated by Schramm and Zowe [21] and also by heuristics used in various convex bundle codes like Lemaréchal's M2FC1 [15] and in the BT codes [24]. Let f be non-convex and let x the current iterate, y^+ a trial step. Taking $g_+ \in \partial f(y^+)$ gives a tangent plane $m_t(y) = f(y^+) + g_+^\top(y - y^+)$ to f at y^+. However, $m_t(\cdot)$ has no reason to be below f, and we do not know whether it is useful to build our working model ϕ_k.

It is not even clear whether $m_t(x) \leq f(x)$, which is the least we would expect. Fixing a small parameter $c > 0$, we therefore put

$$s := [m_t(x) - f(x)]_+ + c\|y^+ - x\|^2,$$

and call this the *down shift*. We then build the cutting plane by $m_{k,y^+,x}(y) = m_t(y) - s$, so that $m_{k,y^+,x}(x) \leq f(x) - c\|y^+ - x\|^2$. Put more explicitly, if the tangent plane $m_t(\cdot)$ does not pass below $f(x)$ at x, then we shift it first down by the gap $f(x) - m_t(x) > 0$. Then we add the small additional down shift $c\|y^+ - x\|^2$, which is needed to assure that the oracle model is contingent with f at x. Defining the upper envelope model as

$$\phi^\uparrow(y, x) = \sup\{m_{k,x,y^+}(y) : y^+ \text{ possible trial step in } B(x, M) \text{ at counter } k\},$$

contingency $\partial_1\phi^\uparrow(x, x) \subset \partial f(x)$ is assured. (For convenience the ball $B(x, M)$ is chosen large enough so that it contains also the elements z^+ arising in axiom (C_3)). The details will be discussed in section 10.

What is the difference of this construction with examples 1 and 2? We know that $m_{k,x,y^+}(y^+) \leq \phi^\uparrow(y^+, x)$, but there might be another oracle plane $m_{k',x,z^+}(\cdot)$, originating from another trial point z^+ at another counter $k' \in \mathbb{N}$, which satisfies $m_{k',x,z^+}(y^+) > m_{k,x,y^+}(y^+)$. Then $\phi^\uparrow(y^+, x) > m_{k,x,y^+}(y^+)$. So the difference with example 1 is that oracle planes are not necessarily tangents of $\phi^\uparrow$.

Neither was this the case in example 2. So what is the difference with example 2? The definition of the model $\phi^\uparrow$ as an upper envelope of convex cuts makes $\phi^\uparrow$ much more difficult to use than in example 2. We are not able to draw tangents (or obtain cutting planes) from it. $\phi^\uparrow(y, x)$ cannot be used in the algorithm, but remains a convenient theoretical tool in the convergence analysis.

Example 4. Using memory. The examples above do not use the counter k of the inner loop. The action taken to generate the cutting plane depends only on x and y^+. But it is beneficial to allow dependence on k. The simplest case where this may happen is when we use memory and keep some of the cutting planes from previous steps $k-1$, $k-2, \ldots, k-t$. The construction in Example 3 could be modified as follows. Construct m_{x,y^+} as above, which does not depend on k. Then take the last t planes stored from the previous steps and take the maximum of all these planes at y^+. This is now dependent on k.

Example 5. Including heuristics. Our axiomatic provides a fairly general mechanism, which allows users a maximum degree of freedom to build their own cutting plane oracle. Users might have additional knowledge how to add heuristic planes $m_h(\cdot)$ to the working model. Obviously they would like to build their working model ϕ_k such that $\phi_k(y^+, x) \geq m_h(y^+)$. But what to do if the heuristic plane $m_h(\cdot)$ drawn at (k, y^+, x) satisfies $m_h(y^+) > m_{k,y^+,x}(y^+)$? Should one in that case reject m_h? The answer is no, as long as $m_h(x) \leq f(x) - c\|y^+ - x\|^2$. One simply replaces the oracle plane by $m_h(\cdot)$, but keeps the old oracle plane in the working model. Increasing the oracle value at y^+ does no harm to axioms (C_2), $(\widehat{C}_2)$, $(\widetilde{C}_2)$. One has to assure that axiom (C_3) is still satisfied.

Example 6. Automatic control and the H_∞-norm. This application has been extremely useful for the development of our theory. Here the function (1) is of the form

$$f(x) = \max_{\omega \in [0,\infty]} f(x, \omega), \qquad f(x, \omega) = \sigma_{\max}\left(F(x, \omega)\right),$$

where $\sigma_{\max}$ is the maximum singular value of a matrix. The operator $F : \mathbb{R}^n \times [0,\infty] \to \mathbb{C}^{p\times m}$ is jointly of class C^∞ and maps into a space of $p \times m$ matrices. A natural candidate for a first-order model of f is

$$\phi(y,x) = \sigma_{\max}\left(F(x,\omega) + F'(x,\omega)(y-x)\right),$$

where derivatives refer to x. Strongness of ϕ follows from remark 7.

The crucial point about model ϕ is that it is more expensive to compute than f, sometimes up to a factor 27. An alternative was therefore developed in [6]. This strongly motivated the current approach, because the solution of [6] matches the concept of a cutting plane oracle.

Definition 8. *Suppose $m_{x,y^+,k}$ is a cutting plane oracle in the sense of Definition 6. Then $M_{x,y^+,k}(y) = m_{x,y^+,k}(y)+\frac{1}{2}(y-x)^\top Q(x)(y-x)$ is the associated second order oracle. If $\phi^\uparrow$ is the upper envelope model associated with the oracle $m_{x,y^+,k}$, then $\Phi^\uparrow(y,x) = \phi^\uparrow(y,x)+\frac{1}{2}(y-x)^\top Q(x)(y-x)$ denotes the corresponding second order upper envelope model.*

4.6 Management of proximity control

The management of the proximity control parameter τ_k marks the second major difference between the non-convex and the convex case. In the convex case proximity control may without harm be fixed once and for all. Early variants of the bundle method fixed τ_k indeed, while later versions allowed updates of τ_k, which remained optional. In the non-convex case a sophisticated management of τ_k is mandatory to establish convergence.

In order to explain the idea, let us temporarily consider the case of the tangent oracle (Example 1), where cutting planes are tangents of the ideal model $\phi(y,x)$. If the trial step y^{k+1} fails to make progress over x, this is because the current working model $\Phi_k(\cdot,x)$ does not agree well with the real f at y^{k+1}. We know that aggregation and cutting planes keep improving the model during inner steps k, but unlike the convex case, these elements only drive Φ_k closer to the ideal model Φ, not directly to f. If $\Phi_k(y^{k+1},x)$ is already close to $\Phi(y^{k+1},x)$ and we still do not make progress, then this must be because $\Phi(y^{k+1},x)$ is by itself too far from $f(y^{k+1})$. Since $\Phi(\cdot,x)$ is in some sense the best model we have (the ideal model), we can only make progress by making smaller trial steps $\|x-y^{k+2}\| < \|x-y^{k+1}\|$. This is arranged by increasing τ_k and referred to as *tightening proximity control*.

In order to decide when to increase τ_k and when not, we draw the cutting plane $m_{k+1}(\cdot,x) := m_{k,x,y^{k+1}}$ at x, counter k and trial point y^{k+1}. Then we build $M_{k+1}(y,x) = m_{k+1}(y,x) + \frac{1}{2}(y-x)^\top Q(x)(y-x)$ and use $M_{k+1}(y^{k+1},x)$ as a substitute for $\Phi(y^{k+1},x)$. In step 6 of the algorithm we compute the secondary control parameter

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

which tests agreement between $\Phi_k(\cdot,x)$ and $M_{k+1}(\cdot,x)$ at y^{k+1}. If agreement between f and Φ_k at y^{k+1} is bad ($\rho_k < \gamma$), but at the same time agreement between Φ_k and M_{k+1} is not bad ($\widetilde{\rho}_k \geq \widetilde{\gamma}$), then we decide that aggregation and cutting planes *alone* will not do the job, because they will only bring Φ_k closer and closer to M_{k+1}, without making progress toward f. This is when we tighten proximity control by increasing τ_k.

On the other hand, when Φ_k is far from f at y^{k+1}, ($\rho_k < \gamma$), but also Φ_k far from M_{k+1}, ($\widetilde{\rho}_k < \widetilde{\gamma}$), then nothing seems decided as yet. We then continue to rely on cutting planes and aggregation

alone, being reluctant to increase τ_k prematurely. This was the strategy analysed in [17]. Here we propose an extension which uses a third control parameter:

$$\widehat{\rho}_k = \frac{f(x) - f(y^{k+1})}{f(x) - M_{k+1}(y^{k+1}, x)} = \frac{\rho_k}{\widetilde{\rho}_k}, \tag{7}$$

whose use was in fact established in numerical tests in [1]. Fixing a third threshold $\widehat{\gamma} < 1$, we now have the following decision:

$$\begin{cases} \text{if } \rho_k < \gamma, \widetilde{\rho}_k < \widetilde{\gamma}, \widehat{\rho}_k < \widehat{\gamma}, & \begin{cases} \text{and } f(x) < M_{k+1}(y^{k+1}, x) \text{ then increase } \tau_k \\ \text{and } f(x) \geq M_{k+1}(y^{k+1}, x) \text{ then leave } \tau_k \text{ as is} \end{cases} \\ \text{if } \rho_k < \gamma, \widetilde{\rho}_k < \widetilde{\gamma}, \widehat{\rho}_k \geq \widehat{\gamma} & \text{then leave } \tau_k \text{ as is} \\ \text{if } \rho_k < \gamma, \widetilde{\rho}_k \geq \widetilde{\gamma} & \text{then increase } \tau_k \end{cases}$$

The entire decision hierarchy can also be seen in Table 1.

Remark 11. The decision in step 6 of the algorithm looks technical and needs some explanation. Notice that in [17] the simpler alternative

$$\tau_{k+1} = \begin{cases} \tau_k, & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k < \widetilde{\gamma} \\ 2\tau_k & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k \geq \widetilde{\gamma} \end{cases}$$

was used. What we do here is a fine analysis of the case $\widetilde{\rho}_k < \widetilde{\gamma}$. We observed in experiments [1,17,22] that it sometimes happens that after a long series of aggregation and cutting plane steps with frozen τ_k, Φ_k comes close enough to M_{k+1} to have $\widetilde{\rho}_k \geq \widetilde{\gamma}$. Then the τ-parameter ultimately *has* to be increased. However, delaying the increase of the τ-parameter is done with the sole intention that serious steps $x \to x^+$ should not become too small. Therefore, if the above happens, this goal is missed. The new test $\widehat{\rho}_k \overset{?}{<} \widehat{\gamma}$ is supposed to accelerate the increase of τ when it cannot be avoided.

Primary	Secondary	Ternary	Decision	Action	Quality
$\rho_k > \Gamma$			accept y^{k+1} serious step	τ_k decreased recycle planes	good
$\gamma \leq \rho_k \leq \Gamma$			accept y^{k+1} serious step	τ_k unchanged recycle planes	not bad
$\rho_k < \gamma$	$\widetilde{\rho}_k \geq \widetilde{\gamma}$		reject y^{k+1} null step	agg. + cutting pl. τ_k increased	too bad
$\rho_k < \gamma$	$\widetilde{\rho}_k < \widetilde{\gamma}$	$\widehat{\rho}_k < \widehat{\gamma}$ and $f(x) > M_{k+1}$	reject y^{k+1} null step	agg.+ cutting pl. τ_k increased	too bad
$\rho_k < \gamma$	$\widetilde{\rho}_k < \widetilde{\gamma}$	else	reject y^{k+1} null step	agg. + cutting pl. τ_k unchanged	bad

TABLE 1: Decision scheme. Secondary and ternary tests are only used in case $\rho_k < \gamma$ (null step) and help to decide whether τ_k is increased or kept fixed. In column three $M_{k+1} := M_{k+1}(y^{k+1}, x)$ and the case *else* includes *either* $\widehat{\rho}_k \geq \widehat{\gamma}$ *or* $\widehat{\rho}_k < \widehat{\gamma}$ in tandem with $f(x) > M_{k+1}$.

Remark 12. We also need to comment on the decision in step 8. If $\rho_k \geq \Gamma$, then agreement between f and Φ_k at $y^{k+1} = x^+$ is *good* and we can trust our model. In trust region methods this is accounted for by increasing the trust region radius. Here we do the same by reducing τ_k in that case, so we *relax proximity control*. Since the inner loop ends in that situation, we pass the information to the next sweep via the memory element $\tau_j^\sharp$ (see steps 3 and 8 of the algorithm). Notice, however, that we have to assure $Q_{j+1} + \tau_1 P_{j+1} \succ 0$ according to what was said in section 4.2. The decision is shown in table 1.

4.7 Recycling of planes

When a new inner loop starts in step 3, a new working model $\phi_1(\cdot, x^+)$ is formed. In the convex case this model does not start from scratch, because one typically recycles some of the cutting and aggregate planes from the previous step x. This happens naturally because these planes are affine minorants of f and remain such as we go from x to x^+.

This is no longer the case when f is non-convex. In principle it may be impossible to use information from the previous serious step x at the new x^+. In this case our strategy to memorize the τ-parameter becomes doubtful, because it presumes some sort of history in time in the working model. Fortunately, in many cases recycling of planes between serious steps $x \to x^+$ *is* possible.

Consider the case where $f = h \circ F$, with h convex and F of class C^1. Suppose $m(y, x) = a + g^\top(y - x)$ is one of the planes used at x. That means $a \leq \phi(x, x) = f(x)$. Can we recycle m at x^+? According to the chain rule we know that $g = F'(x)^* \widetilde{g}$ for some subgradient $\widetilde{g} \in \partial h\,(F(x))$. Since h is convex, $\widetilde{g}$ is still useful even though we pass from x to x^+. We therefore put $g^+ := F'(x^+)^* \widetilde{g}$, and we build the plane $m^+(y, x^+) = a^+ + g^{+\top}(y - x^+)$, where $a^+ = f(x^+)$. The procedure does not interfere with our convergence analysis, so the user can do this in a sophisticate way if the particular structure of the application can be exploited.

A general way to recycle planes which applies without any specific structure of f is to downshift them in exactly the same way as done in Example 3 of section 4.5. That is, if $m(y) = a + g^\top(y - x)$ is a plane used at x, and if the new serious iterate x^+ arrives, then compute the shift $s = [f(x^+) - m(x^+)]_+ + c\|x - x^+\|^2$ and recycle the plane under the new guise $m^+(y) = m(y) - s$. This is less sophisticated than the previous procedure, because it only involves a shift, while tilting plus shifting was used before.

4.8 Memorizing τ

As we have seen, non-convexity makes a dynamic management of the proximity control parameter mandatory. Step 8 of the algorithm regulates how τ is memorized between serious steps.

Definition 9. *We shall say that the τ-parameter is fully memorized between serious steps if $T = \infty$. If $T < \infty$, then we shall say that the large multiplier safeguard rule is applied.*

Except for the constraint $Q_{j+1} + \tau_{j+1}^\sharp P_{j+1} \succ 0$, $T = \infty$ means that we are free to memorize the last value τ_{k+1} if the step was not bad, and $\frac{1}{2}\tau_{k+1}$ if the step was good. The case $T < \infty$ means that $\tau_{j+1}^\sharp$ is corrected as soon as it gets larger than T.

5 Algorithm

Algorithm 1. Proximity control algorithm for (1).

Parameters: $0 < \gamma < \widetilde{\gamma} < \Gamma < 1$, $\widehat{\gamma} < 1$, $0 < q < \infty$, $0 < c < C < \infty$, $q/c < T \leq \infty$.

1: **Initialize outer loop**. Choose initial guess x^1 and an initial matrix $Q_1 = Q_1^\top$ with $-qI \preceq Q_1 \preceq qI$. Choose Euclidian norm $\|y\|_1^2 = y^\top P_1 y$ with $c\|\cdot\| \leq \|\cdot\|_1 \leq C\|\cdot\|$. Fix memory control parameter $\tau_1^\sharp$ such that $Q_1 + \tau_1^\sharp P_1 \succ 0$. Put $j = 1$.

2: **Stopping test**. At outer loop counter j, stop if $0 \in \partial f(x^j)$. Otherwise goto inner loop.

3: **Initialize inner loop**. Put inner loop counter $k = 1$ and initialize τ-parameter using the memory element, i.e., $\tau_1 = \tau_j^\sharp$. Choose initial convex working model $\phi_1(\cdot, x^j)$, and let $\Phi_1(y, x^j) = \phi_1(y, x^j) + \frac{1}{2}(y - x^j)^\top Q_j (y - x^j)$. The Euclidian norm is $\|y\|_j^2 = y^\top P_j y$.

4: **Trial step generation.** At inner loop counter k solve tangent program

$$\min_{y \in \mathbb{R}^n} \Phi_k(y, x^j) + \tfrac{\tau_k}{2}\|y - x^j\|_j^2.$$

The solution is the new trial step y^{k+1}.

5: **Acceptance test.** Check whether

$$\rho_k = \frac{f(x^j) - f(y^{k+1})}{f(x^j) - \Phi_k(y^{k+1}, x^j)} \geq \gamma.$$

If this is the case put $x^{j+1} = y^{k+1}$ (serious step), quit inner loop and goto step 8. If this is not the case (null step) continue inner loop with step 6.

6: **Update proximity parameter**. Call cutting plane oracle $m_{k+1}(\cdot, x^j)$ at x^j, trial point at y^{k+1}, and counter k. Then let $M_{k+1}(y, x^j) = m_{k+1}(y, x^j) + \frac{1}{2}(y - x^j)^\top Q_j (y - x^j)$ and compute secondary and ternary control parameters

$$\widetilde{\rho}_k = \frac{f(x^j) - M_{k+1}(y^{k+1}, x^j)}{f(x^j) - \Phi_k(y^{k+1}, x^j)}, \qquad \widehat{\rho}_k = \frac{f(x^j) - f(y^{k+1})}{f(x) - M_{k+1}(y^{k+1}, x^j)} = \frac{\rho_k}{\widetilde{\rho}_k}$$

$$\text{Put } \tau_{k+1} = \begin{cases} \tau_k, & \text{if } \widetilde{\rho}_k < \widetilde{\gamma} \text{ .and. } \big(\widehat{\rho}_k \geq \widehat{\gamma} \text{ .or. } f(x) < M_{k+1}(y^{k+1}, x)\big) \quad \text{(bad)} \\ 2\tau_k, & \text{if } \widetilde{\rho}_k < \widetilde{\gamma} \text{ .and. } \widehat{\rho}_k < \widehat{\gamma} \text{ .and. } M_{k+1}(y^{k+1}, x) < f(x) \quad \text{(too bad)} \\ 2\tau_k, & \text{if } \widetilde{\rho}_k \geq \widetilde{\gamma} \quad \text{(too bad)} \end{cases}$$

7: **Update working model**. Build new convex working model $\phi_{k+1}(\cdot, x^j)$ based on null step y^{k+1} by respecting the three rules (exactness, cutting plane, aggregation). Then increase inner loop counter k and continue inner loop with step 4.

8: **Update Q_j and memory element**. Update matrix $Q_j \to Q_{j+1}$ respecting $Q_{j+1} = Q_{j+1}^\top$ and $-qI \preceq Q_{j+1} \preceq qI$. Then store new memory element

$$\tau_{j+1}^\sharp = \begin{cases} \tau_{k+1}, & \text{if } \gamma \leq \rho_k < \Gamma \quad \text{(not bad)} \\ \\ \frac{1}{2}\tau_{k+1}, & \text{if } \rho_k \geq \Gamma \quad \text{(good)} \end{cases}$$

If $\tau_{j+1}^\sharp > T$ then re-set $\tau_{j+1}^\sharp = T$. Choose new Euclidian norm $\|\cdot\|_{j+1}$ with $c\|\cdot\| \leq \|\cdot\|_{j+1} \leq C\|\cdot\|$. Increase $\tau_{j+1}^\sharp$ if necessary to ensure $Q_{j+1} + \tau_{j+1}^\sharp P_{j+1} \succ 0$. Increase outer loop counter j by 1 and loop back to step 2.

6 Analysis of the inner loop

In this section we prove finite termination of the inner loop. This requires three Lemmas. Since the Euclidean norm $\|y\|_P^2 = y^\top P y$ is fixed in the inner loop, we suppress the index and simply write it as $\|\cdot\|$. The matrix $Q(x)$ is also fixed, and we write Q.

Our first result shows that the third alternative $\rho_k < \gamma$, $\widetilde{\rho}_k \geq \widetilde{\gamma}$ in step 6 (third line in table 1) cannot occur infinitely often.

Lemma 4. *Let $0 \notin \partial f(x)$. Let axioms (C_1) and (C_2) be satisfied. Suppose none of the trial steps y^{k+1} is accepted, i.e., $\rho_k < \gamma$ for all k. Then there exists k_0 such that $\widetilde{\rho}_k < \widetilde{\gamma}$ for all $k \geq k_0$. In other words, the third alternative in step 6 of the algorithm can only occur for finitely many k.*

Proof: This is essentially the same as Lemma 4 in [17]. A slight difference is that $\Phi(y^{k+1}, x)$ in the definition of $\widetilde{\rho}_k$ in [17] has to be replaced by $M_{k+1}(y^{k+1}, x)$. We therefore need to replace estimate (14) in [17] by an estimate of the form $f(y^{k+1}) - M_{k+1}(y^{k+1}, x) \leq \widetilde{\omega}_k \|x - y^{k+1}\|$ with $\widetilde{\omega}_k \to 0$. This is readily obtained from axiom (C_2). The remainder of the proof is unchanged. □

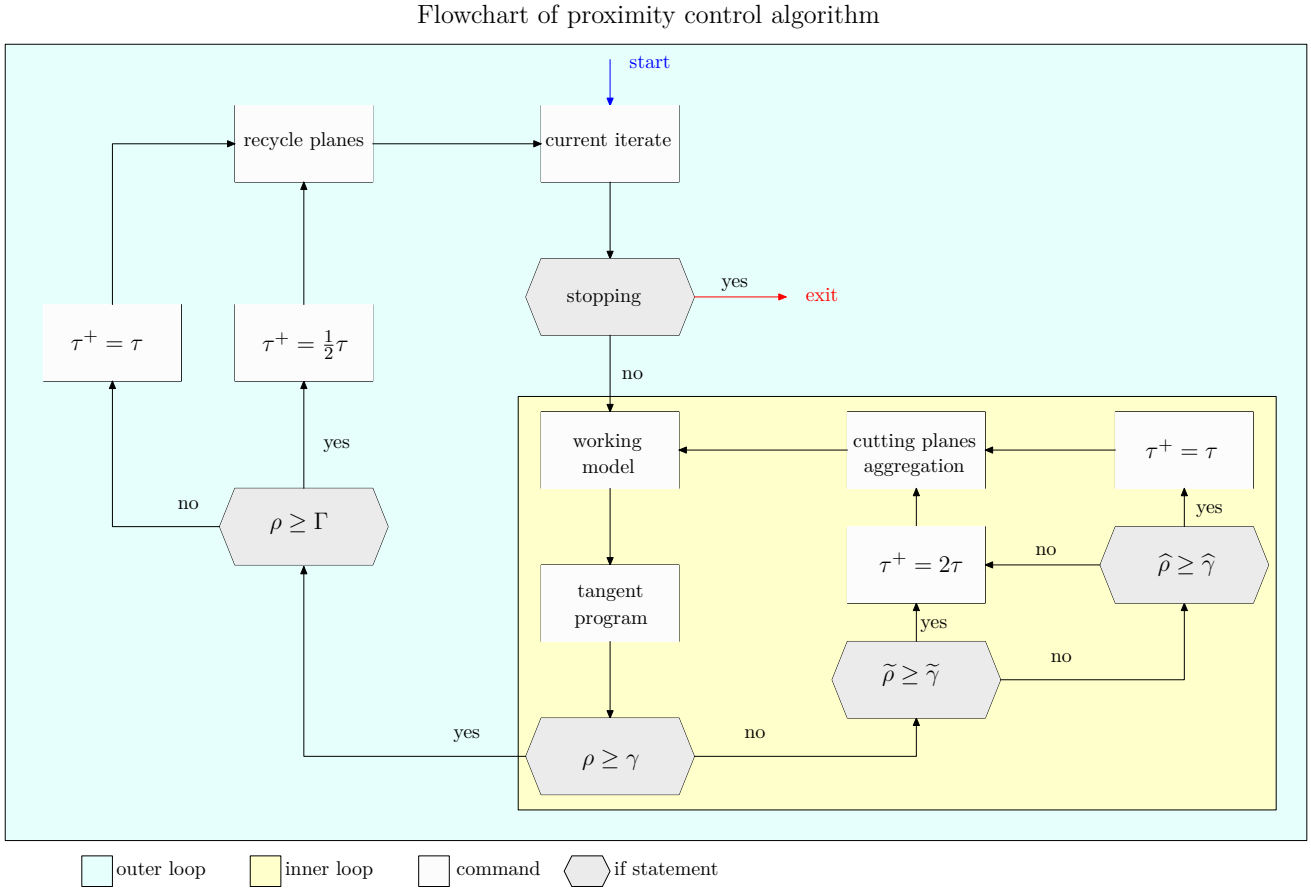

FIGURE 1: Algorithm 1 represented as a flow chart.

Lemma 4 shows that from some counter k_0 onwards we must have $\rho_k < \gamma$ and $\widetilde{\rho}_k < \widetilde{\gamma}$ for all $k \geq k_0$. In other words, the third alternative in the update formula of step 6 (third line in table 1)

can no longer occur. That means, during the remaining infinitely many inner loop steps k either line 4 or line 5 in table 1 must occur. This is where the third control parameter $\widehat{\rho}_k$ enters the scene. There are two logical possibilities, according to step 6 of the algorithm. Case 1 is $\rho_k < \gamma$, $\widetilde{\rho}_k < \widetilde{\gamma}$ from k_0 onwards together with $\widehat{\rho}_k < \widehat{\gamma}$ and $M_{k+1}(y^{k+1}, x) < f(x)$ for infinitely many k. Case 2 is $\rho_k < \gamma$, $\widetilde{\rho}_k < \widetilde{\gamma}$ from k_0 onwards, but now combined with either $\widehat{\rho}_k \geq \widehat{\gamma}$ or $M_{k+1}(y^{k+1}, x) \geq f(x)$ for all k from some counter k_1 onwards. Case 1 corresponds to the second alternative in step 6, (line 5 in table 1), where τ_k is increased infinitely often. Case 2 is the first alternative in step 6, (line 6 in table 1), and this is where τ_k is frozen from some counter k_1 onwards. The following Lemma handles case 1. Case 2 will be discussed in Lemma 6.

Lemma 5. *Let axioms (C_1) and (C_2) be satisfied. Let $0 \notin \partial f(x)$ and suppose the inner loop turns forever, i.e., $\rho_k < \gamma$ for all k. Then there exists k_0 such that for all $k \geq k_0$ $\rho_k < \gamma$, $\widetilde{\rho}_k < \widetilde{\gamma}$ and either $\widehat{\rho}_k \geq \widehat{\gamma}$ or $M_{k+1}(y^{k+1}, x) \geq f(x)$. In other words, the second alternative in step 6 can only occur for finitely many k.*

Proof: (i) By the previous Lemma 4 we know that $\rho_k < \gamma$ and $\widetilde{\rho}_k < \widetilde{\gamma}$ from some counter k_0 onwards. In other words, the third line in step 6 no longer occurs for $k \geq k_0$. Assume contrary to what is claimed that there exist infinitely many $k \in \mathcal{K}$ such that $\widehat{\rho}_k < \widehat{\gamma}$ and at the same time $M_{k+1}(y^{k+1}, x) < f(x)$. Then the second line of the rule in step 6 of the algorithm (line 5 in table 1) is applied infinitely often. Since τ_k is increased in this case, we have $\tau_k \to \infty$. We argue that this implies $y^{k+1} \to x$.

By definition of the aggregate subgradient we have $g^*_{k+1} = (Q + \tau_k P)(x - y^{k+1}) \in \partial_1 \phi_k(y^{k+1}, x)$. By the subgradient inequality this gives

$$g^{*\top}_{k+1}(x - y^{k+1}) \leq \phi_k(x, x) - \phi_k(y^{k+1}, x). \tag{8}$$

Now use the fact that $\phi_k(x, x) = f(x)$ and that the exactness plane $m_e(\cdot, x)$ satisfies $m_e(y^{k+1}, x) \leq \phi_k(y^{k+1}, x)$. Recall that $m_e(y, x) = f(x) + g(x)^\top(y - x)$ for some $g(x) \in \partial f(x)$. Then (8) becomes

$$(x - y^{k+1})^\top (Q + \tau_k P)(x - y^{k+1}) \leq g(x)^\top (x - y^{k+1}) \leq \|g(x)\| \|x - y^{k+1}\|. \tag{9}$$

Since $\tau_k \to \infty$, the term on the left hand side behaves asymptotically like $\tau_k \|x - y^{k+1}\|^2$. Dividing (9) by a factor $\|x - y^{k+1}\|$ then shows $\tau_k \|x - y^{k+1}\| \leq C \|g(x)\|$ for some constant C, which can only happen when $\|x - y^{k+1}\| \to 0$. This proves indeed $y^{k+1} \to x$.

An important consequence of the above estimate is that the sequence g^*_{k+1} is bounded. This follows because $\|g^*_{k+1}\|$ is proportional to $\tau_k \|x - y^{k+1}\|$ for large k.

(ii) We claim that $\liminf_{k \in \mathcal{K}} \dfrac{f(y^{k+1}) - f(x)}{\|y^{k+1} - x\|} \geq 0$. By assumption we have $\widehat{\rho}_k < \widehat{\gamma}$, and at the same time $f(x) - M_{k+1}(y^{k+1}, x) > 0$ for $k \in \mathcal{K}$. Therefore we have

$$\begin{aligned} f(x) - f(y^{k+1}) &\leq \widehat{\gamma} \left(f(x) - M_{k+1}(y^{k+1}, x) \right) \\ &= \widehat{\gamma} \left(f(x) - m_{k+1}(y^{k+1}, x) - \tfrac{1}{2}(y^{k+1} - x)^\top Q (y^{k+1} - x) \right). \end{aligned}$$

By axiom (C_2) there exist $\epsilon_k \to 0^+$ such that $f(y^{k+1}) \leq m_{k+1}(y^{k+1}, x) + \epsilon_k \|y^{k+1} - x\|$. Substituting this gives

$$f(x) - f(y^{k+1}) \leq \widehat{\gamma} \left(f(x) - f(y^{k+1}) + \widetilde{\epsilon}_k \|y^{k+1} - x\| \right),$$

where $\widetilde{\epsilon}_k = \epsilon_k - \frac{1}{2}(y^{k+1} - x)^\top Q (y^{k+1} - x)/\|y^{k+1} - x\| \to 0$. Dividing by $\|y^{k+1} - x\|$ shows

$$(1 - \widehat{\gamma}) \frac{f(y^{k+1}) - f(x)}{\|y^{k+1} - x\|} \geq -\widehat{\gamma} \widetilde{\epsilon}_k,$$

hence $\liminf_{k\in\mathcal{K}} \frac{f(y^{k+1})-f(x)}{\|y^{k+1}-x\|} \geq 0$ as claimed.

(iii) Let $m_e(y,x) = f(x) + g(x)^\top(y-x)$ be the exactness plane at x. Then $g(x) \in \partial f(x)$. Put $d_k = \frac{y^{k+1}-x}{\|y^{k+1}-x\|}$ and assume without loss that $d_k \to d$, passing to a subsequence of $\mathcal{K}$ if necessary. Then $g(x)^\top d \geq \liminf_{k\in\mathcal{K}} \frac{f(y^{k+1})-f(x)}{\|y^{k+1}-x\|} \geq 0$, which follows from the definition of the Clarke subdifferential applied to $-f$ and ii) above.

(iv) Since y^{k+1} solves the tangent program, we know $\psi_k(y^{k+1},x) < \psi_k(x,x) = f(x)$, where $\psi_k(y,x) = \Phi_k(y,x) + \frac{\tau_k}{2}\|y-x\|^2$ is the objective function of the tangent program. The exactness plane satisfies $m_e(\cdot,x) \leq \phi_k(\cdot,x)$. This implies $m_e(y,x)+\frac{1}{2}(y-x)^\top Q(y-x)+\frac{\tau_k}{2}\|y-x\|^2 \leq \Phi_k(y,x)+\frac{\tau_k}{2}\|y-x\|^2 = \psi_k(y,x)$, hence $m_e(y^{k+1},x)+\frac{1}{2}(y^{k+1}-x)^\top(Q+\tau_k P)(y^{k+1}-x) \leq \psi_k(y^{k+1},x) < f(x)$. By the definition of the exactness plane, we deduce $g(x)^\top(y^{k+1}-x)+\frac{1}{2}\|y^{k+1}-x\|^2_{Q+\tau_k P} < 0$. Dividing by $\|y^{k+1}-x\|$ gives then $g(x)^\top d_k + \frac{1}{2}\|y^{k+1}-x\|^2_{Q+\tau_k P}/\|y^{k+1}-x\| < 0$. But $\tau_k \to \infty$, so asymptotically the right hand term is $\sim \tau_k\|y^{k+1}-x\|$, which in turn is $\sim \|g^*_{k+1}\|$, the norm of the aggregate subgradient. Using $\liminf g(x)^\top d_k \geq 0$ proved in part (iii) now implies $\|g^*_{k+1}\| \to 0$, $k \in \mathcal{K}$.

(v) Our next step is to prove $\phi_k(y^{k+1},x) \to f(x)$. From (8) we see that $\liminf f(x) - \phi_k(y^{k+1},x) \geq 0$, because the left hand side of (8) converges to 0. Here we use boundedness of the sequence g^*_{k+1} in tandem with with $y^{k+1} \to x$; see part i). On the other hand, the exactness plane satisfies $\phi_k(y^{k+1},x) \geq m_e(y^{k+1},x) \to m_e(x,x) = f(x)$ as $k \to \infty$, which gives $\liminf \phi_k(y^{k+1},x) - f(x) \geq 0$. Together these two show $\phi_k(y^{k+1},x) \to f(x)$.

(vi) Recall that g^*_{k+1} is a subgradient of $\phi_k(\cdot,x)$ at y^{k+1}. Therefore, for every test vector y, we have by the subgradient inequality

$$\begin{aligned} g^{*\top}_{k+1}(y-y^{k+1}) &\leq \phi_k(y,x) - \phi_k(y^{k+1},x) \\ &\leq \phi^\uparrow(y,x) - \phi_k(y^{k+1},x). \end{aligned}$$

The left hand side converges to 0, $k \in \mathcal{K}$, the right hand side converges to $\phi^\uparrow(y,x) - f(x)$, because $\phi_k(y^{k+1},x) \to f(x)$ from part (v). Together this proves $0 \in \partial_1\phi^\uparrow(x,x)$, hence $0 \in \partial f(x)$. This contradiction proves our result. □

In Lemma 4 we have shown that the second line in the update formula for τ_k in step 6 of the algorithm (or line 5 in table 1) can only occur for finitely many k. From Lemma 5 we know that the same is true for the last line in step 6 (line 4 in table 1). Since by our standing assumption the inner loop turns forever, this means that the first line in the formula of step 6 (last line in table 1) must be active from some counter k_1 onwards. This means that the τ-parameter is frozen from this counter k_1 onwards. The consequences of this case are given by the following

Lemma 6. *Suppose the cutting plane oracle satisfies axioms (C_1) and (C_2). Let $0 \notin \partial f(x)$. Then the inner loop finds a serious step after a finite number of trials k.*

Proof: Assume contrary to what is claimed that the inner loop turns forever. By Lemmas 4 and 5 there exists k_1 such that for all $k \geq k_1$ we have $\rho_k < \gamma$, $\widetilde{\rho}_k < \widetilde{\gamma}$ and *either* $\widehat{\rho}_k \geq \widehat{\gamma}$ *or* $M_{k+1}(y^{k+1},x) \geq f(x)$. By step 6 of the algorithm this means the parameter τ_k is frozen from k_1 onwards, and only cutting planes and aggregation are at work. This is the case which was analysed in [17, Lemma 5]. Earlier work with similar results for convex bundle methods (in the case $Q = 0$) is for instance [11, Proposition 4.3], or [14, Chapter XV], or part II of [8]. □

Remark 13. The stopping test $0 \in \partial f(x)$ in step 2 of the algorithm may seem unrealistic, in particular in large scale applications like Lagrangian relaxation, where $\partial f(x)$ is too expensive to compute at each step x. If we dispense with it and enter the inner loop directly after arrival of a new serious iterate, then the inner loop may turn forever. The proofs of Lemmas 4–6 cover this case as well. They tell us that in this event the sequence y^{k+1} converges to x^j and $0 \in \partial f(x^j)$. What is needed then is a good stopping test for the inner loop, based on slow progress or proximity of y^{k+1} to x^j, allowing us to halt with the correct diagnostic $0 \in \partial f(x^j)$. Since in practice such a test is needed anyway, keeping step 2 in its present form is in no way restrictive. Besides, we can always interpret this as of entrusting the stopping test to the inner loop.

7 Convergence of the outer loop

This central part of the paper shows subsequence convergence of our algorithm. We assume that $Q_j = Q(x^j)$ is the matrix of the second order model, which depends on the serious iterates x^j. We assume that at every instance $j \in \mathbb{N}$ of the outer loop a Euclidean metric $\|y\|_j^2 = y^\top P_j y$ is chosen. This means the tangent program (4) at x^j and inner-loop counter k takes the form

$$\min_{y \in \mathbb{R}^n} \Phi_k(y, x^j) + \tfrac{\tau_k}{2}(y - x^j)^\top P_j (y - x^j). \tag{10}$$

The necessary optimality condition is $\tau_k P_j(x^j - y^{k+1}) \in \partial_1 \Phi_k(y^{k+1}, x^j)$, or what is the same

$$(Q_j + \tau_k P_j)\,(x^j - y^{k+1}) \in \partial_1 \phi_k(y^{k+1}, x^j). \tag{11}$$

Recall that the norms P_j are assumed uniformly equivalent, that is, there exist $c, C > 0$ such that $c\|y\| \le \|y\|_j \le C\|y\|$ for all y and all j. Now we are ready to state

Theorem 1. *Let x^1 be such that $\Omega = \{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$ is bounded. Suppose f has a strict cutting plane oracle on Ω. Suppose the algorithm is operated with the large multiplier safeguard rule, (i.e., $T < \infty$). Then every accumulation point of the sequence of serious iterates is critical.*

Proof: i) From the analysis in section 6 we know that the inner loop always ends after a finite number of steps k with a new x^+ satisfying the acceptance test in step 5, unless we have finite termination due to $0 \in \partial f(x)$. Let us exclude this case, and let x^j denote the infinite sequence of serious steps. We assume that at outer loop counter j the inner loop finds a serious step at inner loop counter $k = k_j$. In other words, $y^{k_j+1} = x^{j+1}$ passes the acceptance test in step 5 of the algorithm and becomes a serious iterate, while the y^{k+1} with $k < k_j$ are null steps. That means

$$f(x^j) - f(x^{j+1}) \ge \gamma \left(f(x^j) - \Phi_{k_j}(x^{j+1}, x^j)\right). \tag{12}$$

Now recall that $(Q_j + \tau_{k_j} P_j)(x^j - x^{j+1}) \in \partial_1 \phi_{k_j}(x^{j+1}, x^j)$ by (5) respectively (11). The subgradient inequality for $\phi_{k_j}(\cdot, x^j)$ at x^{j+1} therefore gives

$$\left(x^j - x^{j+1}\right)^\top (Q_j + \tau_{k_j} P_j)(x^j - x^{j+1}) \;\le\; \phi_{k_j}(x^j, x^j) - \phi_{k_j}(x^{j+1}, x^j) = f(x^j) - \phi_{k_j}(x^{j+1}, x^j),$$

using $\phi_{k_j}(x^j, x^j) = f(x^j)$. With $\Phi_k(y, x^j) = \phi_k(y, x^j) + \frac{1}{2}(y - x^j)^\top Q_j (y - x^j)$ we therefore have

$$\tfrac{1}{2}\|x^{j+1} - x^j\|^2_{Q_j + \tau_{k_j} P_j} \le f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) \le \gamma^{-1}\left(f(x^j) - f(x^{j+1})\right), \tag{13}$$

using (12). Summing (13) from $j = 1$ to $j = J$ gives

$$\sum_{j=1}^{J} \|x^{j+1} - x^j\|^2_{Q_j+\tau_{k_j}P_j} \le \gamma^{-1} \sum_{j=1}^{J} \left(f(x^j) - f(x^{j+1})\right) = \gamma^{-1} \left(f(x^1) - f(x^{J+1})\right).$$

Here the right hand side is bounded above because our method is of descent type in the serious steps. Consequently the series on the left is summable, and therefore $\|x^{j+1} - x^j\|^2_{Q_j+\tau_{k_j}P_j} \to 0$ as $j \to \infty$. Let $\bar{x}$ be an accumulation point of the sequence x^j and select a subsequence $j \in J$ such that $x^j \to \bar{x}$, $j \in J$. We have to prove $0 \in \partial f(\bar{x})$.

ii) Suppose there exists an infinite subsequence $j \in J'$ of $j \in J$ such that $g_j := (Q_j + \tau_{k_j}P_j)(x^j - x^{j+1}) \to 0$, $j \in J'$. Here we claim that $0 \in \partial f(\bar{x})$, so that we are done.

In order to prove this claim, notice first that since $\{x \in \mathbb{R}^n : f(x) \le f(x^1)\}$ is bounded by hypothesis, and since our algorithm is of descent type in the serious steps, the sequence x^j, $j \in \mathbb{N}$ is bounded.

Since g_j is a subgradient of $\phi_{k_j}(\cdot, x^j)$ at $x^{j+1} = y^{k_j+1}$, we have for every test vector h:

$$\begin{aligned} g_j^\top h &\le \phi_{k_j}(x^{j+1} + h, x^j) - \phi_{k_j}(x^{j+1}, x^j) \\ &\le \phi^\uparrow(x^{j+1} + h, x^j) - \phi_{k_j}(x^{j+1}, x^j) \qquad (\text{using } \phi_{k_j} \le \phi^\uparrow). \end{aligned}$$

Now we use the fact that $y^{k_j+1} = x^{j+1}$ was accepted in step 5 of the algorithm, which means

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

Combining these two estimates for a fixed test vector h gives:

$$\begin{aligned} g_j^\top h &\le \phi^\uparrow(x^{j+1} + h, x^j) - f(x^j) + f(x^j) - \phi_{k_j}(x^{j+1}, x^j) \\ &= \phi^\uparrow(x^{j+1} + h, x^j) - f(x^j) + f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) + \tfrac{1}{2}(x^j - x^{j+1})^\top Q_j (x^j - x^{j+1}) \\ &\le \phi^\uparrow(x^{j+1} + h, x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1})\right) + \tfrac{1}{2}(x^j - x^{j+1})^\top Q_j (x^j - x^{j+1}) \\ &= \phi^\uparrow(x^{j+1} + h, x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1})\right) + \\ &\qquad + \tfrac{1}{2}(x^j - x^{j+1})^\top (Q_j + \tau_{k_j}P_j)(x^j - x^{j+1}) - \tfrac{\tau_{k_j}}{2}\|x^j - x^{j+1}\|_j^2 \\ &\le \phi^\uparrow(x^{j+1} + h, x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1})\right) + \tfrac{1}{2}(x^j - x^{j+1})^\top (Q_j + \tau_{k_j}P_j)(x^j - x^{j+1}). \end{aligned}$$

Now fix $h' \in \mathbb{R}^n$. Plugging $h = x^j - x^{j+1} + h'$ in the above estimate gives

$$\tfrac{1}{2}\|x^j - x^{j+1}\|^2_{Q_j+\tau_{k_j}P_j} + g_j^\top h' \le \phi^\uparrow(x^j + h', x^j) - f(x^j) + \gamma^{-1} \left(f(x^j) - f(x^{j+1})\right). \tag{14}$$

Passing to the limit $j \in J'$ and using, in the order named, $\|x^j - x^{j+1}\|^2_{Q_j+\tau_{k_j}P_j} \to 0$, $g_j \to 0$, $x^j \to \bar{x}$, $f(x^j) \to f(\bar{x}) = \phi^\uparrow(\bar{x}, \bar{x})$ and $f(x^j) - f(x^{j+1}) \to 0$, we obtain:

$$0 \le \phi^\uparrow(\bar{x} + h', \bar{x}) - \phi^\uparrow(\bar{x}, \bar{x}). \tag{15}$$

Here the rightmost term $f(x^j) - f(x^{j+1}) \to 0$ converges by monotonicity, while convergence of the leftmost term was shown in part i). Now the test vector h' in (15) is arbitrary, which shows $0 \in \partial_1 \phi^\uparrow(\bar{x}, \bar{x})$. By axiom (M_1) we have $0 \in \partial f(\bar{x})$.

iii) As a consequence of part ii), we are now left to deal with the case where $\|g_j\| = \|(Q_j + \tau_{k_j}P_j)(x^j - x^{j+1})\| \ge \mu > 0$ for some $\mu > 0$ and every $j \in J$. The remainder of this proof will be entirely dedicated to this case.

We notice first that under this assumption the τ_{k_j}, $j \in J$, must be unbounded. Indeed, assume on the contrary that the τ_{k_j}, $j \in J$, are bounded. By boundedness of Q_j, P_j and boundedness of the serious steps, there exists then an infinite subsequence $j \in J'$ of J such that Q_j, P_j, τ_{k_j} and $x^j - x^{j+1}$ converge respectively to $\bar{Q}$, $\bar{P}$, $\bar{\tau}$ and $\delta\bar{x}$ as $j \in J'$. This implies that the corresponding subsequence of g_j converges to $(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x}$, where $\|(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x}\| \geq \mu > 0$. Similarly, $(x^j-x^{j+1})^\top(Q_j+\tau_{k_j}P_j)(x^j-x^{j+1}) \to \delta\bar{x}^\top(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x}$. By part i) of the proof we have $g_j^\top(x^{j+1}-x^j) = \|x^{j+1}-x^j\|^2_{Q_j+\tau_{k_j}P_j} \to 0$, which means $\delta\bar{x}^\top(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x} = 0$. Since $\bar{Q}+\bar{\tau}\bar{P}$ is symmetric and $\bar{Q}+\bar{\tau}\bar{P} \succeq 0$, we deduce $(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x} = 0$, contradicting $\|(\bar{Q}+\bar{\tau}\bar{P})\delta\bar{x}\| \geq \mu > 0$. This argument proves that the τ_{k_j}, $j \in J$, are unbounded.

iv) Having shown that the sequence τ_{k_j}, $j \in J$ is unbounded, we can without loss assume that $\tau_{k_j} \to \infty$, $j \in J$, passing to a subsequence if required. Let us now distinguish two types of indices $j \in J$. We let J^+ be the set of those $j \in J$ for which the τ-parameter was increased at least once during the j^{th} inner loop. The remaining indices $j \in J^-$ are those where the τ-parameter remained unchanged during the j^{th} inner loop. Since the j^{th} inner loop starts at $\tau_j^\sharp$ and ends at τ_{k_j}, we have

$$J^+ = \{j \in J : \tau_{k_j} < \tau_j^\sharp\} \quad \text{and} \quad J^- = \{j \in J : \tau_{k_j} = \tau_j^\sharp\}.$$

We claim that the set J^- must be finite. For suppose J^- is infinite, then $\tau_{k_j} \to \infty$, $j \in J^-$. Then also $\tau_j^\sharp \to \infty$, $j \in J^-$. But this contradicts the large multiplier safeguard rule in step 8 of the algorithm, which forces $\tau_j^\sharp \leq T$. This contradiction shows that J^+ is cofinal in J.

iv) Remember that we are still in the case whose discussion started in point iii). We are now dealing with an infinite subsequence $j \in J^+$ of $j \in J$ such that $\tau_{k_j} \to \infty$, $\|g_j\| \geq \mu > 0$, and such that the τ-parameter was increased at least once during the j^{th} inner loop. Suppose this happened for the last time at stage $k_j - \nu_j$ for some $\nu_j \geq 1$. Then

$$\tau_{k_j} = \tau_{k_j-1} = \cdots = \tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}. \tag{16}$$

According to step 6 of the algorithm, the increase at counter $k_j-\nu_j$ may have been for two different reasons. Either

$$\rho_{k_j-\nu_j} < \gamma \quad \text{and} \quad \widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma} \tag{17}$$

or

$$\rho_{k_j-\nu_j} < \gamma, \quad \widetilde{\rho}_{k_j-\nu_j} < \widetilde{\gamma} \quad \text{and} \quad \widehat{\rho}_{k_j-\nu_j} < \widehat{\gamma}, \tag{18}$$

the latter condition in tandem with $M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j) < f(x^j)$. These are the two cases labelled *too bad* in step 6 of the algorithm (see also Table 1). In part v) below we will discuss the consequences of (17). Case (18) will be considered in part vi).

v) We continue by discussing case (17), where infinitely many $j \in J^+$ satisfy

$$\rho_{k_j-\nu_j} = \frac{f(x^j) - f(y^{k_j-\nu_j+1})}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} < \gamma \quad \text{and} \quad \widetilde{\rho}_{k_j-\nu_j} = \frac{f(x^j) - M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j)}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} \geq \widetilde{\gamma}.$$

Notice first that as $\tau_{k_j} \to \infty$ and $\tau_{k_j} = 2\tau_{k_j-\nu_j}$, boundedness of the subgradients $\widetilde{g}_j := (Q_j + \frac{1}{2}\tau_{k_j}P_j)(x^j - y^{k_j-\nu_j+1}) \in \partial_1\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)$ shows $y^{k_j-\nu_j+1} \to \bar{x}$. Indeed, boundedness follows

from the subgradient inequality

$$\begin{aligned}(x^j - y^{k_j-\nu_j+1})^\top (Q_j + \tau_{k_j-\nu_j} P_j)(x^j - y^{k_j-\nu_j+1}) &\leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\\ &\leq f(x^j) - m_{e,j}(y^{k_j-\nu_j+1})\\ &= g(x^j)^\top (x^j - y^{k_j-\nu_j+1})\\ &\leq \|g(x^j)\| \|x^j - y^{k_j-\nu_j+1}\|,\end{aligned}$$

where $m_{e,j}(y) = f(x^j) + g(x^j)^\top (y - x^j)$ is the exactness plane at x^j. Now as $\tau_{k_j} \to \infty$, the left hand side behaves asymptotically like constant times $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j+1}\|^2$, because $\tau_{k_j-\nu_j} = \frac{1}{2}\tau_{k_j} \to \infty$. On the other hand the $x^j \in \Omega$ are bounded, hence so are the $g(x^j)$. The right hand side therefore behaves asymptotically like constant times $\|x^j - y^{k_j-\nu_j+1}\|$. This shows boundedness of $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j+1}\|$, and therefore $x^j - y^{k_j-\nu_j+1} \to 0$.

We now have to discuss two logical possibilities. Either there exists an infinite subset J' of J^+ such that $\widetilde{g}_j \to 0$, $j \in J'$, or $\|\widetilde{g}_j\| \geq \eta > 0$ for some $\eta > 0$ and all $j \in J^+$.

Suppose first that there exists an infinite subsequence J' of J^+ such that $\|\widetilde{g}_{j'}\| \to 0$, $j' \in J'$. Then all is well, as we now argue. Namely, for a test vector h and $j \in J'$:

$$\begin{aligned}\widetilde{g}_j^\top h &\leq \phi_{k_j-\nu_j}(y^{k_j+\nu_j+1} + h, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\\ &\leq \phi^\uparrow(y^{k_j+\nu_j+1} + h, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j).\end{aligned} \tag{19}$$

Using the fact that $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma}$, we have

$$f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \leq \widetilde{\gamma}^{-1}\left(f(x^j) - M_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\right).$$

Adding $\frac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j (y^{k_j-\nu_j+1} - x^j)$ on both sides gives

$$\begin{aligned}&f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\\ &\qquad \leq \widetilde{\gamma}^{-1}\left(f(x^j) - M_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\right) + \tfrac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j (y^{k_j-\nu_j+1} - x^j).\end{aligned}$$

Combining this and estimate (19) gives

$$\begin{aligned}\widetilde{g}_j^\top h \leq \phi^\uparrow(y^{k_j-\nu_j+1} + h, x^j) - f(x^j) + \widetilde{\gamma}^{-1}&\left(f(x^j) - M_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)\right)\\ &+\tfrac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j (y^{k_j-\nu_j+1} - x^j).\end{aligned} \tag{20}$$

As we have seen $y^{k_j-\nu_j+1} - x^j \to 0$, hence the rightmost term converges to 0 by boundedness of Q_j. Moreover, we claim that $\lim f(x^j) - M_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) = 0$, so the term $\widetilde{\gamma}^{-1}(\dots)$ on the right hand side of (20) converges to 0. Indeed, to see this claim, notice that since $y^{k_j-\nu_j+1} - x^j \to 0$ and $x^j \to \bar{x}$, axiom (C_3) gives $\limsup m_j(y^{k_j-\nu_j+1}, x^j) \leq \limsup \phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \leq \phi^\uparrow(\bar{x}, \bar{x}) = f(\bar{x})$. Since the oracle is strict, so is $\phi^\uparrow$, and axiom $(\widehat{M}_2)$ gives $\epsilon_j \to 0$ such that

$$f(y^{k_j-\nu_j+1}) - \phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \leq \epsilon_j \|y^{k_j-\nu_j+1} - x^j\|. \tag{21}$$

Passing to the limit in (21) implies $\liminf \phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \geq f(\bar{x})$, so the two estimates together show $f(x^j) - \phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \to 0$. Since the quadratic term converges to 0, we deduce $f(x^j) - \Phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \to 0$, proving the claim. Going back with this information to the above subgradient inequality and passing to the limit shows

$$0 \leq \phi^\uparrow(\bar{x} + h, \bar{x}) - f(\bar{x}) = \phi^\uparrow(\bar{x} + h, \bar{x}) - \phi^\uparrow(\bar{x}, \bar{x}),$$

where we apply axiom (M_3) for $\phi^\uparrow$ to the first term on the right hand side. This proves $0 \in \partial_1\phi^\uparrow(\bar{x},\bar{x})$, because h was arbitrary. Therefore $0 \in \partial f(\bar{x})$ by model axiom (M_1). This shows that indeed all is well in the case of a subsequence $j \in J'$ with $\widetilde{g}_j \to 0$.

To continue, let us now consider the second logical alternative $\|\widetilde{g}_j\| \geq \eta$ for some $\eta > 0$ and all $j \in J^+$. As we shall see, this case can be ruled out. Indeed, we argue that there exists $\theta > 0$ such that

$$f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \geq \theta\|y^{k_j-\nu_j+1} - x^j\| \tag{22}$$

for all $j \in J^+$ sufficiently large. Namely, by the subgradient inequality we have

$$\widetilde{g}_j^\top(x^j - y^{k_j-\nu_j+1}) \;\leq\; \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) = f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j).$$

Subtracting $\frac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j(y^{k_j-\nu_j+1} - x^j)$ from both sides gives

$$\tfrac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top(Q_j + \tau_{k_j}P_j)(y^{k_j-\nu_j+1} - x^j) \leq f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j).$$

Now as $\tau_{k_j} \to \infty$, we have $\frac{1}{4}\|\widetilde{g}_j\|\|y^{k_j-\nu_j+1} - x^j\| \leq \frac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top(Q_j + \tau_{k_j}P_j)(y^{k_j-\nu_j+1} - x^j)$ for $j \in J^+$ large enough, which proves formula (22) with $\theta = \frac{1}{4}\eta$.

Next using (21), and subtracting the usual $\frac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j(y^{k_j-\nu_j+1} - x^j)$ from both sides gives

$$f(x^j) - \Phi^\uparrow(y^{k_j-\nu_j+1}, x^j) \leq \widetilde{\epsilon}_j\|y^{k_j-\nu_j+1} - x^j\|, \tag{23}$$

where $\widetilde{\epsilon}_j := \epsilon_j + \frac{1}{2}\|Q_j\|\|y^{k_j-\nu_j+1} - x^j\| \to 0$. Combining (22) and (23) gives the estimate

$$\widetilde{\rho}_{k_j-\nu_j} \leq \rho_{k_j-\nu_j} + \frac{\widetilde{\epsilon}_j\|y^{k_j-\nu_j+1} - x^j\|}{\theta\|y^{k_j-\nu_j+1} - x^j\|}$$

which shows $\limsup \widetilde{\rho}_{k_j-\nu_j} \leq \limsup \rho_{k_j-\nu_j} \leq \gamma$, contradicting $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma} > \gamma$ for the infinitely many $j \in J'$. This contradiction shows that $\|\widetilde{g}_j\| \geq \eta > 0$ for all $j \in J^+$ was impossible, so some subsequence of $\widetilde{g}_j$ $(j \in J^+)$ *does* converge to 0, proving $0 \in \partial f(\bar{x})$. This ends the discussion of condition (17).

vi) It remains to discuss the consequences of (18), the case where the τ-parameter was increased for infinitely many $j \in J^+$ because of the second rule in step 6 of the algorithm:

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

For the indices $k = k_j - \nu_j + 1, \ldots, k_j - 1$ we have $\rho_k < \gamma$, $\widetilde{\rho}_k < \widetilde{\gamma}$ and $\widehat{\rho}_k \geq \widehat{\gamma}$, while of course $\rho_{k_j} \geq \gamma$, because the last iterate in the j^{th} inner loop $y^{k_j+1} = x^{j+1}$ was accepted as the serious step of the j^{th} outer loop.

Recall that in the case of the third rule we also know that $M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j) < f(x^j)$, so condition $\widehat{\rho}_{k_j-\nu_j+1} < \widehat{\gamma}$ becomes

$$\begin{aligned}
f(x^j) - f(y^{k_j-\nu_j+1}) &< \widehat{\gamma}\left(f(x^j) - M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j)\right) \\
&= \widehat{\gamma}\left(f(x^j) - m_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j) - \tfrac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top Q_j(y^{k_j-\nu_j+1} - x^j)\right) \\
&\leq \widehat{\gamma}\left(f(x^j) - f(y^{k_j-\nu_j+1}) + \epsilon_j\|y^{k_j-\nu_j+1} - x^j\| + \tfrac{1}{2}\|Q_j\|\|y^{k_j-\nu_j+1} - x^j\|^2\right) \\
&= \widehat{\gamma}\left(f(x^j) - f(y^{k_j-\nu_j+1}) + \widetilde{\epsilon}_j\|y^{k_j-\nu_j+1} - x^j\|\right), \qquad (24)
\end{aligned}$$

where $\widetilde{\epsilon}_j \to 0$. Here the third line is based on axiom $(\widehat{C}_2)$. In order to be allowed to use this axiom, we have to prove $y^{k_j-\nu_j+1} - x^j \to 0$ as $j \in \mathcal{N}'$. To see this observe first that the $y^{k_j-\nu_j+1}$ are bounded. This was already established in part iii) of the proof, and here the argument is quite similar. Namely, by the subgradient inequality we have

$$\begin{aligned}(x^j - y^{k_j-\nu_j+1})^\top (Q_j + \tau_{k_j-\nu_j} P_j)(x^j - y^{k_j-\nu_j+1}) &\leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \\ &\leq f(x^j) - m_{e,j}(y^{k_j-\nu_j+1}) \\ &= g(x^j)^\top (x^j - y^{k_j-\nu_j+1}) \\ &\leq \|g(x^j)\| \|x^j - y^{k_j-\nu_j+1}\|,\end{aligned}$$

where $m_{e,j}(y) = f(x^j) + g(x^j)^\top (y - x^j)$ is the exactness plane at x^j. Now the left hand side behaves asymptotically like $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j+1}\|^2$, because from the definition of ν_j we have $\tau_{k_j-\nu_j} = \frac{1}{2}\tau_{k_j} \to \infty$. On the other hand, since $x^j \in \Omega$, the sequence $g(x^j)$ is bounded, so the right hand side behaves like constant times $\|x^j - y^{k_j-\nu_j+1}\|$. This shows boundedness of $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j+1}\|$, and therefore also $x^j - y^{k_j-\nu_j+1} \to 0$.

Re-arranging estimate (24) and dividing by $\|y^{k_j-\nu_j+1} - x^j\|$ gives

$$(1 - \widehat{\gamma}) \frac{f(x^j) - f(y^{k_j-\nu_j+1})}{\|x^j - y^{k_j-\nu_j+1}\|} \leq \widetilde{\epsilon}_j,$$

hence passing to the limit $j \in J'$ using $\widetilde{\epsilon}_j \to 0$ leads to the estimate

$$\liminf_{j \in J'} \frac{f(y^{k_j-\nu_j+1}) - f(x^j)}{\|y^{k_j-\nu_j+1} - x^j\|} \geq 0. \tag{25}$$

Let J'' be a subsequence of J' such that $x^j \to \bar{x}$, $j \in J''$. We have to prove $0 \in \partial f(\bar{x})$. Let us put $d_j = \frac{y^{k_j-\nu_j+1} - x^j}{\|y^{k_j-\nu_j+1} - x^j\|}$. Passing to yet another subsequence of J'' if necessary, we may assume $d_j \to d$ and $g(x^j) \to g \in \partial f(\bar{x})$, the latter by upper semicontinuity of the Clarke subdifferential. If we apply the definition of the Clarke directional derivative to $-f$ we obtain, using (25), that

$$\liminf_{j \in J''} g(x^j)^\top d_j \geq \liminf_{j \in J''} \frac{f(y^{k_j-\nu_j+1}) - f(x^j)}{\|y^{k_j-\nu_j+1} - x^j\|} \geq 0. \tag{26}$$

This estimate will come in handily in a moment.

Recall that $y^{k_j-\nu_j+1}$ is the solution of program

$$\min_{y \in \mathbb{R}^n} \psi_{k_j-\nu_j}(y, x^j) = \Phi_{k_j-\nu_j}(y, x^j) + \tfrac{\tau_{k_j-\nu_j}}{2}\|y - x^j\|_j^2.$$

The exactness plane $m_{e,j}$ satisfies $m_{e,j}(y) + \frac{1}{2}(y - x^j)^\top Q_j (y - x^j) + \frac{\tau_{k_j-\nu_j}}{2}\|y - x^j\|_j^2 \leq \Phi_{k_j-\nu_j}(y, x^j) + \frac{\tau_{k_j-\nu_j}}{2}\|y - x^j\|_j^2 = \psi_{k_j-\nu_j}(y, x^j)$. Hence

$$\begin{aligned}m_{e,j}(y^{k_j-\nu_j+1}) + \tfrac{1}{2}(y^{k_j-\nu_j+1} - x^j)^\top (Q_j + \tau_{k_j-\nu_j} P_j)(y^{k_j-\nu_j+1} - x^j) \\ \leq \psi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \leq \psi_{k_j-\nu_j}(x^j, x^j) = f(x^j).\end{aligned}$$

By definition of the exactness plane, subtracting $f(x^j)$ on both sides gives

$$g(x^j)^\top (y^{k_j-\nu_j+1} - x^j) + \tfrac{1}{2}\|y^{k_j-\nu_j+1} - x^j\|^2_{Q_j + \tau_{k_j-\nu_j} P_j} \leq 0.$$

Dividing by $\|y^{k_j-\nu_j+1} - x^j\|$ and using the definition of d_j gives

$$g(x^j)^\top d_j + \tfrac{1}{2}\frac{\|y^{k_j-\nu_j+1} - x^j\|^2_{Q_j+\tau_{k_j-\nu_j}P_j}}{\|y^{k_j-\nu_j+1} - x^j\|} \leq 0. \tag{27}$$

Here the right hand expression behaves asymptotically like constant times $\tau_{k_j-\nu_j}\|y^{k_j-\nu_j+1} - x^j\|$, which in turn behaves like constant times $\|g_j^*\|$, where $g_j^* = (Q_j + \tau_{k_j-\nu_j}P_j)(x^j - y^{k_j-\nu_j+1})$ is the aggregate subgradient at inner loop stage $k_j - \nu_j$. Using $\liminf g(x^j)^\top d_j \geq 0$, proved in (26) above, we get $g(x^j)^\top d_j \to 0$. From estimate (27) it now also follows that $\|g_j^*\| \to 0$, because the middle term in (27) is proportional to $\|g_j^*\|$ and is squeezed in between two terms converging to 0. This is the key property, but before we can exploit it, we need to verify one more fact.

Notice that $\tau_{k_j} \to \infty$ in tandem with (16) shows $y^{k_j-\nu_j+1} \to \bar{x}$, $j \in J''$. We now claim that the following holds:

$$\lim_{j\in J''} \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) = f(\bar{x}).$$

In order to prove this, observe that the aggregate subgradient at inner loop instant $k_j - \nu_j$ is $g_j^* = (Q_j + \tau_{k_j-\nu_j}P_j)(x^j - y^{k_j-\nu_j+1})$. By the subgradient inequality we have

$$g_j^{*\top}(x^j - y^{k_j-\nu_j+1}) \leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j).$$

Since the left hand side converges to 0, we have $\limsup_{j\in J''} \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \leq f(\bar{x})$. So it remains to establish the reverse estimate. This uses properties of the exactness plane. Namely, $\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \geq m_{e,j}(y^{k_j-\nu_j+1})$, where $m_{e,j}$ is the exactness plane at x^j. But $m_{e,j}(y^{k_j-\nu_j+1}) = f(x^j) + g(x^j)^\top(y^{k_j-\nu_j+1} - x^j) \to f(\bar{x})$ as $j \in J''$, because $f(x^j) \to f(\bar{x})$ and $g(x^j) \to g \in \partial f(\bar{x})$, $y^{k_j-\nu_j+1} - x^j \to 0$. This proves $\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \to f(\bar{x})$.

Now we are ready to finish the proof of case (18). For an arbitrary test vector y we have by the subgradient inequality:

$$\begin{aligned} g_j^{*\top}(y - y^{k_j-\nu_j+1}) &\leq \phi_{k_j-\nu_j}(y, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \\ &\leq \phi^\uparrow(y, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j). \end{aligned}$$

Passing to the limit $j \in J''$ using $g_j^* \to 0$ on the left and $\phi_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j) \to f(\bar{x})$ on the right shows

$$0 \leq \limsup_{j\in J''} \phi^\uparrow(y, x^j) - \phi^\uparrow(\bar{x}, \bar{x}).$$

Now axiom (M_3) shows $\limsup \phi^\uparrow(y, x^j) \leq \phi^\uparrow(y, \bar{x})$. We have therefore proved $0 \in \partial_1\phi^\uparrow(\bar{x}, \bar{x})$, which gives $0 \in \partial f(\bar{x})$. This completes the proof. □

It is convenient to give the following

Definition 10. *If the memory parameter $\tau_{j+1}^\sharp$ in step 8 of the algorithm exceeds T, and is therefore re-set to T, then we shall say that the large multiplier safeguard rule is applied. Setting $T = \infty$ in the algorithm is therefore another way of saying that the large multiplier safeguard rule is not used.*

Corollary 1. *Suppose $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded and f has a strict cutting plane oracle on Ω. Suppose the algorithm is operated without the large multiplier safeguard rule (i.e. $T = \infty$). Then there exists at least one accumulation point of the sequence of serious iterates which is critical.*

Proof: In order to analyse this case let us introduce the following terminology. We call an outer loop index j a *drone* if the τ-parameter is never increased during the j^{th} inner loop. The large multiplier safeguard rule therefore excludes the existence of infinite subsequences $j \in J$ such that $\tau_{k_j} \to \infty$ and such that every $j \in J$ is a drone. Let us call such a subsequence *parasitic*.

The proof of Theorem 1 shows that whenever $\bar{x}$ is the accumulation point of a subsequence x^j which is *not* parasitic, then $\bar{x}$ is critical. However, having dispensed with the large multiplier safeguard rule, we cannot exclude the existence of parasitic subsequences. Fortunately, if a parasitic subsequence exists, then we can also find an infinite set J such that $\tau_{k_j} \to \infty$, and such that the τ-parameter was increased at least once during the j^{th} inner loop. For short, we can also find a non parasitic subsequence. Indeed, if the τ-parameter is unbounded, and since the drones do not do any work to increase it, some infinite subsequence where all the work is done must exist. Every accumulation point of this non parasitic sequence x^j, $j \in J$ is critical. □

Remark 14. The large multiplier safeguard rule does not altogether exclude the existence of infinite subsequences $j \in J$ with $\tau_{k_j} \to \infty$. It only forbids infinite subsequences consisting entirely of drones, i.e., parasitic subsequences. Neither does the large multiplier safeguard rule exclude the existence of drones.

8 Convergence with complete memory

In this section we discuss the question of convergence without the large multiplier safeguard rule in closer detail. It turns out that this rule is not needed when a strong oracle is available.

Theorem 2. *Let $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ be bounded and suppose f has a strong cutting plane oracle on Ω. Suppose the algorithm is operated without the large multiplier safeguard rule and also without the ternary test (i.e., $T = \infty$ and $\widehat{\gamma} = -\infty$). Then every accumulation point of the sequence of serious iterates is critical.*

Proof: We follow the line of Theorem 1. Parts i) – iii) of the proof can be adopted without modification. We have to deal with a sequence $j \in J$ such that $\tau_{k_j} \to \infty$ and $\|g_j\| \geq \mu > 0$. If the τ-parameter was increased at least once during the j^{th} inner loop, then we are in the case discussed in the proof of Theorem 1, so we can follow step (v) of that proof and conclude that $0 \in \partial f(\bar{x})$. The new situation we have to deal with is when the τ-parameter is never increased in the j^{th} inner loop, i.e., if the sequence is parasitic in the sense discussed in the proof of Corollary 1.

iv) We argue that in this case there exists another infinite subsequence $j \in J'$ with $\tau_{k_j} \to \infty$, such that in addition for each $j \in J'$, the doubling rule in step 6 of the algorithm is applied at least once before the step $x^{j+1} = y^{k_j+1}$ was accepted. Indeed, to construct J' we let, for every $j \in J$, $j' \leq j$ be that outer-loop instant where the τ-parameter was increased for the last time before j, and we put $J' := \{j' : j \in J\}$. It is possible that $j' = j$, but in general we may have $j' < j$, and we only know that

$$2\tau_{k_{j'-1}} \leq \tau_{k'_j} \quad \text{and} \quad \tau_{k_{j'}} \geq \tau_{k_{j'+1}} \geq \cdots \geq \tau_{k_j}.$$

Since $\tau_{k_j} \to \infty$, $j \in J$, we also get $\tau_{k_{j'}} \to \infty$, $j' \in J'$. Since the doubling rule was applied at least once at the outer-loop counter j', the set J' is as claimed.

Let us say that for $j \in J'$ the doubling rule was applied for the last time at stage $\tau_{k_j-\nu_j}$ for some $\nu_j \geq 1$. That means, $\tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}$, while the τ-parameter remained unchanged during

the following inner steps before acceptance:

$$\tau_{k_j} = \tau_{k_j-1} = \cdots = \tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}. \tag{28}$$

Now recall that in step 6 of the algorithm the doubling rule is applied for two different reasons. Either because $\rho < \gamma$ and $\widetilde{\rho} \geq \widetilde{\gamma}$, or because $\rho < \gamma$, $\widetilde{\rho} < \widetilde{\gamma}$, $\widehat{\rho} < \widehat{\gamma}$. But remember that we dispensed with the second case by putting $\widehat{\gamma} = -\infty$. Therefore, if the τ-parameter is increased, this is because $\rho_k < \gamma$ and $\widetilde{\rho}_k \geq \widetilde{\gamma}$. Since by assumption this is the case at stage $k_j - \nu_j$ we have

$$\rho_{k_j-\nu_j} = \frac{f(x^j) - f(y^{k_j-\nu_j+1})}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} < \gamma \quad \text{and} \quad \widetilde{\rho}_{k_j-\nu_j} = \frac{f(x^j) - M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j)}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} \geq \widetilde{\gamma}.$$

By (16) we now have $\widetilde{g}_j = (Q_j + \frac{1}{2}\tau_{k_j}P_j)\left(x^j - y^{k_j-\nu_j+1}\right) \in \partial_1\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)$. Using the subgradient inequality for $\phi_{k_j-\nu_j}(\cdot, x^j)$ at $y^{k_j-\nu_j+1}$ and $\phi_{k_j-\nu_j}(x^j, x^j) = f(x^j)$, we obtain

$$\begin{aligned}\left(x^j - y^{k_j-\nu_j+1}\right)^\top (Q_j + \tfrac{1}{2}\tau_{k_j}P_j)\left(x^j - y^{k_j-\nu_j+1}\right) &\leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \\ &= f(x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j),\end{aligned}$$

which on subtracting $\frac{1}{2}(x^j - y^{k_j-\nu_j+1})^\top Q_j(x^j - y^{k_j-\nu_j+1})$ from both sides becomes

$$\tfrac{1}{2}(x^j - y^{k_j-\nu_j+1})^\top(Q_j + \tau_{k_j}P_j)(x^j - y^{k_j-\nu_j+1}) \leq f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j).$$

Using $\|x^j - y^{k_j-\nu_j+1}\|^2_{Q_j+\tau_{k_j}P_j} \geq (\tau_{k_j}\|P_j\| - \|Q_j\|)\|x^j - y^{k_j-\nu_j+1}\|^2$, this could also be written as

$$\frac{(\tau_{k_j}\|P_j\| - \|Q_j\|)\|x^j - y^{k_j-\nu_j+1}\|^2}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} \leq 2. \tag{29}$$

Now, substituting (29) into the expression for $\widetilde{\rho}_{k_j-\nu_j}$ and expanding gives

$$\begin{aligned}\widetilde{\rho}_{k_j-\nu_j} &= \rho_{k_j-\nu_j} + \frac{f(y^{k_j-\nu_j+1}) - M_{k_j-\nu_j+1}(y^{k_j-\nu_j+1}, x^j)}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} \\ &\leq \rho_{k_j-\nu_j} + \frac{(L + \|Q_j\|)\|x^j - y^{k_j-\nu_j+1}\|^2}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)} \qquad \text{(using } (\widetilde{C}_2)\text{)} \\ &\leq \rho_{k_j-\nu_j} + 2\frac{L + \|Q_j\|}{\tau_{k_j}\|P_j\| - \|Q_j\|} \qquad \text{(using (29))}.\end{aligned}$$

Here the estimate in $(\widetilde{C}_2)$ is applied to the set $B = \{x^j, y^{k_j-\nu_j+1} : j \in J'\}$, which as we now argue is bounded. Indeed, to see this observe that $\tau_{k_j-\nu_j} = \frac{1}{2}\tau_{k_j} \to \infty$ as $j \in J'$. Applying the subgradient inequality to $\widetilde{g}_j = (Q_j + \tau_{k_j-\nu_j}P_j)(x^j - y^{k_j-\nu_j+1}) \in \partial_1\phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j)$ gives

$$\begin{aligned}(x^j - y^{k_j-\nu_j+1})^\top(Q_j + \tau_{k_j-\nu_j}P_j)(x^j - y^{k_j-\nu_j+1}) &\leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j+1}, x^j) \\ &\leq f(x^j) - m_{e,j}(y^{k_j-\nu_j+1}) = g(x^j)^\top(x^j - y^{k_j-\nu_j+1}) \\ &\leq \|g(x^j)\|\|x^j - y^{k_j-\nu_j+1}\|.\end{aligned}$$

Here $m_{e,j}(\cdot)$ is again the exactness plane at x^j. By (16) we have $\tau_{k_j-\nu_j} \to \infty$, so using boundedness of the x^j and boundedness of the Q_j, P_j, we deduce (via the argument already employed in the proof of Lemma 1) that the $y^{k_j-\nu_j+1}$ are bounded.

Going back to the above estimate involving $\widetilde{\rho}_{k_j-\nu_j}$ and $\rho_{k_j-\nu_j}$, notice that $\rho_{k_j-\nu_j} < \gamma$ and $(L + \|Q_j\|)/(\tau_{k_j}\|P_j\| - \|Q_j\|) \to 0$ imply $\limsup_{j\to\infty} \widetilde{\rho}_{k_j-\nu_j} \leq \gamma$, contradicting $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma} > \gamma$ for the infinitely many $j \in J'$. This proves that an infinite sequence $j \in J$ with $\|g_j\| \geq \mu > 0$ and $\tau_{k_j} \to \infty$ could not exist. Due to part ii) of the proof of Theorem 1, we infer that $0 \in \partial f(\bar{x})$ for every accumulation point of the sequence x^j, $j \in J$. This completes the proof. □

Remark 15. The proof of Theorem 2 shows that a strong oracle excludes parasitic subsequences by itself, i.e., without the large multiplier safeguard rule. However, our proof only works if the ternary test is dispensed with ($\widehat{\gamma} = -\infty$), and so far we have not been able to establish the same result if the ternary test is used.

9 Some examples

This section presents a variety of examples which show that our algorithm could also be understood as a fairly general method to establish convergence results. We shall also put forward situations where the choice of $Q(x)$ may lead to fast local convergence.

Example. Objective strictly differentiable. Suppose f is strictly differentiable. Then the standard model is $\phi^\sharp(y, x) = f(x) + \nabla f(x)^\top(y - x)$, which is then also the natural model. Suppose we choose $Q(x) = 0$ and let the working model coincide with the ideal model, i.e., $\phi_k = \phi^\sharp$ for every k. Then the tangent program is

$$\min_{y\in\mathbb{R}^n} f(x) + \nabla f(x)^\top(y - x) + \tfrac{\tau}{2}\|y - x\|^2,$$

which means $y = x - \tau^{-1}\nabla f(x)$. In other words, the trial step is a steepest descent step with steplength τ^{-1}. The acceptance test in step 5 of the algorithm is

$$\rho = \frac{f(x) - f(y)}{-\nabla f(x)^\top(y - x)} \geq \gamma,$$

which is nothing but the usual Armijo test with Armijo constant $0 < \gamma < 1$. Since $\widetilde{\rho} = 1$, we always increase τ when the Armijo condition is not satisfied. (The decision parameter $\widehat{\rho}$ is not needed here.) This corresponds to decreasing the step τ^{-1}, a backtracking linesearch. The large multiplier safeguard rule, if used, becomes a small safeguard rule against small stepsizes. Namely, if the j^{th} linesearch ended successfully with step t_j, then we start the $(j+1)^{\text{st}}$ linesearch at t_j if the accepted step was not bad, respectively at $2t_j$ if the accepted step was good. Naturally, the large multiplier safeguard rule, if applied, becomes a safeguard rule against $t_j^\sharp$ becoming too small, where we re-set $t_j^\sharp = T^{-1}$ if $t_j^\sharp < T^{-1}$. If $T = \infty$, then $T^{-1} = 0$ and no such rule is applied. Since $Q(x) = 0$, we are then in the case where $t_j^\sharp \in \{t_j, 2t_j\}$, and we refer to this as the step being fully memorized between iterates $x^j \to x^{j+1}$. The relation $t = \tau^{-1}$ has even more surprising consequences.

Proposition 2. *Let f be differentiable and suppose the standard first-order model $\phi^\sharp$ is used as working model $\phi^\sharp = \phi_k$ with $Q(x) = 0$. Then our non-smooth algorithm specializes to the steepest descent method with backtracking linesearch and Armijo acceptance condition. The outer loop produces the iterates x^j, the inner loop is the linesearch. Suppose one of the following scenarios occurs:*

1. *f is of class C^1 and the safeguard rule against small steps $t_j^\sharp \geq T^{-1}$ is applied.*

2. *f is of class $C^{1,1}$ and the steplength is fully memorized.*

Then every accumulation point of the sequence of serious iterates x^j is a critical point. □

Remark 16. In descent methods where second order steps are tempted, the line search starts with stepsize $t^\sharp = 1$ in order to allow full Newton or quasi Newton steps. (This corresponds to applying the safeguard rule against small steps). In contrast, in a pure first-order method we might memorize the successful steplength at stage j and start the $(j+1)^{\text{st}}$ linesearch there in order to safe time. This is exactly what our algorithm does, except for the fact that the memorized steplength $t^\sharp$ would be $t^\sharp = 2t_j$ if the accepted step x^j is good. If f is of class $C^{1,1}$, then the standard model is strong, and this is why we still converge by fully memorizing the stepsize $t = \tau^{-1}$ as in step 8 of algorithm 1.

If we allow a variable metric $\|y\|_j = \left(y^\top P_j y\right)^{1/2}$ at each outer step j, then the step becomes

$$x^{j+1} = x^j - \tau^{-1} P_j^{-\frac{1}{2}} \nabla f(x^j)^\top P_j^{-\frac{1}{2}}. \tag{30}$$

Here we are performing a backtracking linesearch with descent direction $d_j = -P_j^{-\frac{1}{2}} \nabla f(x^j)^\top P_j^{-\frac{1}{2}}$. Recall that we assume $c\|y\| \leq \|y\|_j \leq C\|y\|$ for certain $0 < c < C < \infty$ and all j. These descent directions d_j are therefore gradient oriented, i.e., the angle between d_j and $-\nabla f(x^j)$ stays bounded away from $\pm 90^0$, or what is the same, $-90^0 < -\alpha \leq \angle(d_j, -\nabla f(x^j)) \leq \alpha < 90^{\text{o}}$ for some fixed angle $0^0 < \alpha < 90^{\text{o}}$. Also, $c'\|d_j\| \leq \|\nabla f(x_j)\| \leq C'\|d_j\|$ for some $0 < c' < C' < \infty$ and all j.

Proposition 3. *Let f be of class C^1. Then every descent method with gradient oriented descent direction (30), Armijo condition, and backtracking linesearch can be interpreted as a special case of our non-smooth algorithm. Consequently, every accumulation point of the sequence of iterates so generated is critical if the linesearch is initialized at $t^\sharp = (\tau^\sharp)^{-1}$ with $t^\sharp \geq T^{-1}$. If f is of class $C^{1,1}$ then the successful stepsize of the line search can be fully memorized when passing from x^j to x^{j+1}.* □

Remark 17. To our knowledge Propositions 2 and 3 are new. Even when the line search is initialized at a stepsize $t^\sharp$ larger than some threshold $T^{-1} > 0$, convergence in the literature is usually proved for $C^{1,1}$ functions, while our proof shows C^1 is enough. Naturally, one would ask whether more practical backtracking procedures are covered. The rule $\tau_{k+1} = 2\tau_k$ in steps 6 the algorithm could at any moment be replaced by more flexible choices like $\tau_{k+1} = \Theta_k \tau_k$ for some $\Theta_k > 1$. All that is needed is that applying this rule infinitely often causes τ_k to converge to ∞.

Example. Non-smooth steepest descent. Let us consider a non-smooth f with the standard model $\phi^\sharp$ and $Q(x) = 0$. As before let the working model coincide with the ideal model. Then $\widetilde{\rho} = 1$, so the only action taken in the inner loop is reducing τ.

The tangent program is $\min_{y \in \mathbb{R}^n} f(x) + f^0(x, y-x) + \frac{\tau}{2}\|y-x\|^2$. Writing $h = y - x$ and omitting the constant term $f(x)$, this could be written as a minimax program $v := \min_h \max_{g \in \partial f(x)} g^\top h + \frac{\tau}{2}\|h\|^2$. Using Fenchel duality we may swap the min and max operators. Then

$$v = \max_{g \in \partial f(x)} \min_h g^\top h + \tfrac{\tau}{2}\|h\|^2 = \max_{g \in \partial f(x)} g^\top h(g) + \tfrac{\tau}{2}\|h(g)\|^2,$$

where the inner minimum over h is unconstrained and can therefore be solved explicitly. The solution is $h(g) = -\tau^{-1}g$, which we substitute back. This gives $v = \max_{g\in\partial f(x)} -\frac{1}{2\tau}\|g\|^2 = -\frac{1}{2\tau}\|g(x)\|^2$, where $g(x)$ is the steepest descent direction $g(x) = \arg\min\{\|g\| : g \in \partial f(x)\}$. This leads back to $y = x + h = x - \frac{1}{\tau}g(x)$.

Proposition 4. *Suppose the standard model $\phi^\sharp(y, x) = f(x) + f^0(x, y - x)$ is used as working model ϕ_k and $Q(x) = 0$. Then our method specializes to the non-smooth steepest descent method with backtracking linesearch and Armijo acceptance test. If $\phi^\sharp$ is strict, then every accumulation point of the sequence of iterates x^j is a critical point of f as long as the safeguard rule against small stepsizes is applied. If the standard model $\phi^\sharp$ is strong, then the stepsize τ^{-1} may be fully memorized between serious steps $x^j \to x^{j+1}$.* □

Remark 18. This result is complementary to classical statements where convergence of the non-smooth steepest descent method is established when $\tau_j^{-1} \to 0$ in tandem with $\sum_j \tau_j^{-1} = \infty$. Conditions of that type can neither be checked nor forced algorithmically. In contrast, our condition can be tested beforehand. It is for instance satisfied if f is upper C^1.

Example. Objective function is C^2. For f of class C^2 the standard model is $\phi^\sharp(y, x) = f(x) + \nabla f(x)^\top(y - x)$ and coincides with the natural model. Here it makes sense to let $Q(x) = \nabla^2 f(x)$.

We assume that the working model coincides with the ideal model. Then the tangent program (4) computes $y^{k+1} = \operatorname{argmin}_{y\in\mathbb{R}^n} f(x) + \nabla f(x)^\top(y - x) + \frac{1}{2}(y - x)^\top\left(\nabla^2 f(x) + \tau I\right)(y - x)$. This leads to $y = x - \left(\nabla^2 f(x) + \tau I\right)^{-1}\nabla f(x)$, which is a damped Newton step. The trial step y is accepted as the next serious step x^+ as soon as $\rho \geq \gamma$. This is equivalent to

$$\rho = \frac{f(x) - f(y)}{f(x) - \Phi(y, x)} = \frac{f(x) - f(y)}{-\nabla f(x)^\top(y - x) - \frac{1}{2}(y - x)^\top\nabla^2 f(x)(y - x)} \geq \gamma.$$

This test differs from the usual Armijo test, where one would require

$$\rho_{\text{armijo}} = \frac{f(x) - f(y)}{f(x) - \phi(y, x)} = \frac{f(x) - f(y)}{-\nabla f(x)^\top(y - x)} \geq \gamma$$

for the Armijo constant $0 < \gamma < 1$. Notice that $\rho \geq \rho_{\text{armijo}}$ if $\nabla^2 f(x) \succeq 0$, which means that in the convexity zone of f our test is easier to satisfy than the Armijo test. For $\gamma < \frac{1}{2}$ both tests are asymptotically equivalent, as shown by the argument of the Dennis-Moré theorem [13]. Their argument shows that $\rho \approx 1$ in the neighbourhood of a local minimum satisfying the second order sufficient optimality condition. In consequence, if the sequence of serious iterates starts in this neighbourhood, then eventually $\rho > \Gamma$, which has the consequence that the τ parameter converges to 0. We obtain the following

Proposition 5. *Suppose f is of class C^2. Suppose the standard first order model is used as first-order working model. Suppose $Q(x) = \nabla^2 f(x)$, so that $\Phi_k = \Phi$ is the second order Taylor polynomial of f at x. Let $\bar{x}$ a local minimum of (1) satisfying the second order sufficient optimality condition. Then there exist $\epsilon > 0$ such that whenever $x^1 \in B(\bar{x}, \epsilon)$, the sequence of serious iterates x^j generated by our algorithm satisfies $x^{j+1} \in B(\bar{x}, \epsilon)$ for all j. Moreover, $x^{j+1} = x^j - \left(\nabla^2 f(x^j) + 2^{-j}\tau_1 I\right)^{-1}\nabla f(x^j)$ is a damped Newton step, which converges superlinearly to $\bar{x}$. Each inner loop accepts the first trial step to become x^{j+1} with the good case $\rho \geq \Gamma$.* □

Example. Objective function is C^2 (continued). If the sufficient second order optimality condition is satisfied at the local minimum $\bar{x}$, then $\nabla^2 f(x) \succeq \epsilon I \succ 0$ in a neighbourhood of the minimum. For x in this neighbourhood one may therefore choose $Q(x) = \nabla^2 f(x) - \epsilon(x) I \succeq 0$, where $\epsilon(x) \approx \epsilon$ as x approaches $\bar{x}$. This allows $\tau_j^\sharp$ to become arbitrarily small, because the condition $Q(x^j)+\tau_j^\sharp I \succ 0$ in step 8 is no longer restrictive. In consequence, the tangent program computes the Newton step, and not just a damped version. This is clearly satisfactory, as our method includes an important classical situation. Notice that we do not have to know ϵ to get this. As soon as we are in the neighbourhood of attraction, the undamped Newton step becomes interpretable as generated by our algorithm. Naturally, unless ϵ is known, the difficulty is that we do not know when we are in the neighbourhood of attraction, so we do not know from what moment onward we are authorized to try a Newton step.

Example. Objective is convex. Suppose f is convex and the first order model $\phi(\cdot, x) = f$ is used. Then it makes sense to let $Q(x) = 0$, because $Q(x) \neq 0$ would produce a model Φ farther away from f than ϕ. With this choice our algorithm reproduces the classical form of the bundle algorithm for convex objectives.

Example. Objective lower C^2. Consider the case where f is lower C^2, and choose $\mu > 0$ such that $\phi_\mu(y, x) = f(y) + \mu\|y - x\|^2$ is convex, hence a strong model for f. Our natural choice of the second order term is $\frac{1}{2}(y - x)^\top Q(x)(y - x) = -\mu\|y - x\|^2$, because this gives $\Phi(y, x) = f(y)$. We assume $\Phi_k = \Phi$ for all k, so the tangent program is

$$\min_{y \in \mathbb{R}^n} f(y) + \tfrac{\tau_k}{2}\|y - x\|^2. \tag{31}$$

Since we must have $Q(x) + \tau_k I \succeq 0$, we are only allowed proximity parameters satisfying $\tau_k/2 \geq \mu$. This is just saying that $f + \frac{\tau_k}{2}\|\cdot - x\|^2$ is convex. In other words, the tangent program computes a proximal step. The acceptance test is $\rho_k = \frac{f(x)-f(y)}{f(x)-f(y)-\frac{\tau_k}{2}\|y-x\|^2} \geq \gamma$, and if $\rho_k < \gamma$ then the proximity constant is increased, and it is decreased if a serious step with $\rho \geq \Gamma$ occurs. All the other actions in the algorithm (aggregate and cutting planes, decisions depending on $\widetilde{\rho}$ and $\widehat{\rho}$) are redundant.

Proposition 6. *For a lower C^2 function the proximal point algorithm based on (31) can be interpreted as a special case of our algorithm.* □

Notice in addition that we have the option to memorize the τ parameter between serious steps, because the model ϕ_μ is strong. Moreover, if we use the oracle of Example 2 in section 4.5, then μ can be adapted anew after each serious step. For instance, we may decrease μ as going from x to x^+ in cases where a smaller μ^+ suffices to convexify f at x^+. We can also be forced to increase μ if the opposite happens. In the latter case we might have to correct τ at the beginning of the inner loop according to step 8 to assure $\tau/2 \geq \mu$.

Can we decrease μ during the inner loop? According to the oracle of Example 2 in section 4.5 we can indeed. However, in that case we have to keep the aggregate planes, because they are no longer redundant. So here we no longer use a pure proximal point method.

Remark 19. We have to read the result above with some care, because $f + \mu\|\cdot - x\|^2$ is typically only convex in a neighbourhood of x. Since $\phi_\mu(\cdot, x)$ has to be convex *everywhere*, we may have to define it differently outside this neighbourhood, which is not in the spirit of a practical method. In assuming that the solution of the proximal point program $\min_y f(y) + \mu\|y - x\|^2$ gives the new trial step y, we therefore make the implicit assumption that y lies within the region of convexity of $f + \mu\|\cdot - x\|^2$.

Another restriction is that in order to converge superlinearly, the proximal point method needs $\tau_j = \tau_j^\sharp \to 0$ [18]. But this may be in conflict with the requirement $\tau_j/2 \geq \mu$. It certainly is if $\mu > 0$ is just fixed. We better adapt μ at every step, choosing $\mu(x)$ as small as possible to guarantee convexity of $\phi_{\mu(x)}(y, x) = f(y) + \mu(x)\|y - x\|^2$.

But this raises yet another interesting question. As x approaches a local minimum $\bar{x}$, can we render $\phi_{\mu(x)}(\cdot, x)$ convex with smaller and smaller $\mu(x)$, even with $\mu(x) \to 0$ as $x \to \bar{x}$? It appears that the answer should be positive, because functions get more and more convex as minima are approached. Unfortunately, this is so only for smooth functions. Looking at $f(x) = |x| - x^2$, which has a local minimum at 0, we see what may happen. Indeed, to convexify f we need $Q(x) = -I$, hence $\mu(x) \geq 1$ even for $x \to 0$.

Remark 20. The models $\phi_\mu(y, x) = f(y) + \mu\|y - x\|^2$ for prox-regular f appear for the first time in [17]. Later Sagastizábal and Hare [19], Sagastizábal [20], and also Lewis and Wright [16] use essentially the same construction. As we have seen, the delicate problem is the choice of μ. In order to assure that the solution of the tangent program lies in the zone of convexity of ϕ_μ, we want μ large, but on the other hand we want to use the smallest possible μ to convexify in order to keep our model close to the true f. The solution to this dilemma is to not use ϕ_μ at all. What we recommend instead is the downshift oracle. It is a much more flexible tool, which in addition applies to the much larger class of lower C^1 functions.

10 Application: Downshift oracle

In this section we apply our theory to the downshift oracle, which is Example 3 in Section 4.5. The best known convergence result for this oracle is Schramm and Zowe's [21, Theorem 3.1], where it is proved that *some* accumulation point of the sequence of serious iterates is critical. The algorithm of [21] differs from ours in the management of τ. In relating the downshift technique to our oracle concept, we can prove stronger convergence results.

Lemma 7. *The downshift oracle satisfies axiom* (C_1).

Proof: A tangent plane at $y^+ = x$ is of the form $m(y) = f(x) + g^\top(y - x)$ for some $g \in \partial f(x)$. Since the quadratic term $c\|x - x\|^2$ in the downshift vanishes, we have $s = 0$, so the cutting plane $m_{x,x}$ coincides with the tangent. This immediately gives axiom (C_1). □

Lemma 8. *Let f be a lower C^1 function. Then the downshift oracle satisfies axiom* $(\widehat{C}_2)$, *and therefore also axiom* (C_2).

Proof: i) Let $y_j \to x$, $x_j \to x$. We have to find $\epsilon_j \to 0^+$ such that $f(y_j) \leq m_{y_j,x_j}(y_j) + \epsilon_k\|y_j - x_j\|$. Now observe that $m_{y_j,x_j}(y_j) \geq f(y_j) - s_j$, where s_j is the downshift relating the tangent and the oracle plane at trial point y_j. The case $s_j = c\|y_j - x\|^2$ is when the tangent plane to f at y_j passes below $f(x_j)$. Here we simply let $\epsilon_j = c\|y_j - x_j\|$. In the case where the tangent plane m_t passes above $f(x_j)$ the down shift is $s_j = m_t(x_j) - f(x_j) + c\|y_j - x_j\|^2$. Now $f(y_j) \leq m_{y_j,x_j}(y_j) + \epsilon_j\|y_j - x_j\|$ is equivalent to $f(y_j) \leq f(x_j) + g_j^\top(y_j - x_j) - c\|y_j - x_j\|^2 + \epsilon_j\|y_j - x_j\|$, where $g_j \in \partial f(y_j)$. So equivalently, we have to find $\epsilon_j \to 0$ such that $g_j^\top(x_j - y_j) \leq f(x_j) - f(y_j) + \epsilon_j\|y_j - x_j\|$.

ii) Following Daniilidis and Georgiev [12] a lower C^1 function is approximately convex: For every $\bar{x}$ and $\epsilon > 0$ there exists $\delta > 0$ such that $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) + \epsilon t(1-t)\|x - y\|$

for all $x, y \in B(\bar{x}, \delta)$ and all $0 \leq t \leq 1$. Re-arranging this estimate gives

$$\frac{f(y + t(x - y)) - f(y)}{t} \leq f(x) - f(y) + \epsilon(1 - t)\|x - y\|.$$

By [18, Theorem 10.31] f is Clarke regular, so that passing to the limit $t \to 0$ gives the estimate

$$f^0(y, x - y) \leq f(x) - f(y) + \epsilon\|x - y\|.$$

Then for every $g \in \partial f(y)$, $g^\top(x - y) \leq f^0(y, x - y) \leq f(x) - f(y) + \epsilon\|x - y\|$. This is precisely the estimate we need to satisfy the condition of part i). □

Lower C^1 function have been introduced by Spingarn [23]. Lower C^k functions, $k \geq 1$, are discussed in Rockafellar-Wets [18].

Lemma 9. *The downshift oracle satisfies axiom* (C_3).

Proof: Let $y_j^+ \to y^+$, $y_j \to y$, $x_j \to x$. Then $m_{y_j^+, x_j} = m_{t,j} - s_j$, where $m_{t,j}$ is the tangent plane to f at y_j^+, that is $m_{t,j}(y) = f(y_j^+) + g_j^\top(y - y_j^+)$ for some $g_j \in \partial f(y_j^+)$. Passing to a subsequence if necessary, we may assume $g_j \to g^+ \in \partial f(y^+)$. Then $m_{t,j} \to m_t$, where m_t is a tangent to f at y^+ (upper semicontinuity of the Clarke subdifferential). Since the down shift s_j depends continuously on the data x_j, y_j^+ and g_j, and because of $g_j \to g^+$, we have for this subsequence $s_j \to s$, where s is the down shift associated with x, y^+ and g^+. In other words, $m_{t,j} - s_j = m_{y_j^+, x_j}$ and $m_t - s = m_{y^+, x}$. Then since $y_j \to y$ we clearly have $\lim_j m_{y_j^+, x_j}(y_j) = m_{y^+, x}(y)$, because the sequence of gradients $\nabla m_{y_j^+, x_j}$ is uniformly bounded. □

By what is proved so far we know that we are in business. The down shift oracle is strict as soon as f is lower C^1. It remains to settle the question when the oracle is strong. This is answered by the following

Lemma 10. *The following condition is equivalent to strongness of the downshift oracle: Whenever* $x_j \to x$, $y_j \to x$, $g_j \in \partial f(y_j)$, *then there exists* $L > 0$ *such that* $f(y_j) - f(x_j) - g_j^\top(x_j - y_j) \leq L\|y_j - x_j\|^2$ *for all* j.

Proof: We let $y_j \to x$, $x_j \to x$. Then $f(y_j) \leq m_{y_j, x_j}(y_j) + s_j$, where s_j is the downshift belonging to the trial point y_j at serious point x_j. That is $s_j = [m_{t,j}(x_j) - f(x_j)]_+ + c\|y_j - x_j\|^2$, where $m_{t,j}(y) = f(y_j) + g_j^\top(y - y_j)$ is a tangent to f at y_j. In particular, $g_j \in \partial f(y_j)$. Strongness of the oracle requires $s_j = \mathrm{O}(\|y_j - x_j\|^2)$. Since this is obvious for the quadratic term $c\|y_j - x_j\|^2$, everything hinges on whether $m_{t,j}(x_j) - f(x_j) \leq \mathrm{O}(\|y_j - x_j\|^2)$. Now this term equals $f(y_j) + g_j^\top(x_j - y_j) - f(x_j)$, so strongness is assured as soon as $f(x_j) - f(y_j) + g_j^\top(x_j - y_j) \geq -L\|y_j - x_j\|^2$ for some $L > 0$. □

To understand the condition of Lemma 10, consider

$$\Delta_{f,y,g}(h) := \frac{f(y + h) - f(y) - g^\top h}{\|h\|^2}$$

the second difference quotient of f at y with respect to $g \in \partial f(y)$. The condition of Lemma 10 above reads

$$\Delta_{f,y_j,g_j}(x_j - y_j) \geq -L. \tag{32}$$

Notice that if a function f is convex, then $\Delta_{f,y,g}(h) \geq 0$ for every h by the subgradient inequality. Therefore, (32) ought to be related to convexity. And indeed, observe that the second difference quotient of $x \mapsto L\|x-y\|^2$ is constant $\Delta_{L\|\cdot-y\|^2,x,2L(x-y)}(h) = L$ for all h. Therefore, $\Delta_{f,y_j,g_j}(x_j - y_j) \geq -L$ is equivalent to $\Delta_{f+L\|\cdot-y_j\|^2,y_j,g_j+2L(x_j-y_j)}(x_j - y_j) \geq 0$, which comes down to convexity of $f + L\|\cdot - y_j\|^2$. We summarize:

Proposition 7. *If f is lower C^2, then the downshift oracle is strong.*

Proof: The lower C^2 property implies that for every $\bar{x}$ we find $L > 0$ and a neighbourhood U of $\bar{x}$ such that for every $y \in U$ the function f may be convexified on U by adding $L\|\cdot - y\|^2$. Then $\Delta_{f+L\|\cdot-y\|^2,y,g_j+2L(x-y)}(x-y) \geq 0$ for all $x, y \in U$, hence we have (32). □

Theorem 3. *Let f be locally Lipschitz and suppose $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Suppose the downshift oracle with one of the following operational modes is applied:*

1. *The large multiplier safeguard rule is used ($T < \infty$), and f is lower C^1.*
2. *The proximity control parameter is fully memorized ($T = \infty$), the ternary test involving $\widehat{\rho}$ is dispensed with ($\widehat{\gamma} = -\infty$), and f is lower C^2.*

Then every accumulation point of the sequence of serious iterates x^j is a critical point of f. □

11 Conclusion

Cutting plane oracles have been introduced and used to develop bundling techniques for non-convex non-smooth optimization. The new concept expands naturally on existing techniques like convex bundling, model-based cutting planes [17], and techniques for composite functions like [16, 19, 20]. Downshift of tangent planes used in [15, 21, 24] can also be seen as a special instance of our cutting plane oracle. As a consequence, we obtain satisfactory and natural convergence proofs for these cases in the class of lower C^1 functions.

Acknowledgement

Funding by Fondation de Recherche pour l'Aéronautique et l'Espace under contract *Survol*, and by Fondation EADS under contract *Technicum* is gratefully acknowledged.

References

[1] P. Apkarian, L. Ravanbod-Hosseini, D. Noll. *Time-domain constrained structured H_∞-synthesis.* Int. J. Robust Nonlin. Control, to appear.

[2] P. Apkarian, D. Noll. *Non-smooth H_∞ synthesis.* IEEE Transactions on Automatic Control, vol. 51, no. 1, 2006, pp. 71 – 86.

[3] P. Apkarian, D. Noll. *Non-smooth optimization for multidisk H_∞ synthesis.* European Journal of Control, vol. 12, no. 3, 2006, pp. 229 – 244.

[4] P. Apkarian, D. Noll. *Non-smooth optimization for multiband frequency domain control design.* Automatica, vol. 43, no. 4, 2007, pp. 724 – 731.

[5] P. Apkarian, D. Noll. *Controller design with non-smooth multidirectional search.* SIAM Journal on Control and Optimization, vol. 44, no. 6, 2006, pp. 1923 – 1949.

[6] P. Apkarian, D. Noll, O. Prot. *A trust region spectral bundle method for non-convex eigenvalue optimization.* SIAM Journal on Optimization, vol. 19, no. 1, 2008, pp. 281 – 306.

[7] 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, vol. 16, 2009, pp. 641 – 666.

[8] F. Bonnans, J.-C. Gilbert, C. Lemaréchal, C. Sagastizábal. Numerical Optimization. Theoretic and Practical Aspects. Springer Verlag 2006.

[9] F. H. Clarke. Optimization and Non-smooth Analysis. SIAM Classics in Applied Mathematics. Philadelphia 1990.

[10] K.C. Kiwiel. *An aggregate subgradient method for non-smooth convex minimization.* Math. Programming, vol. 27, 1983, p. 320 - 341.

[11] R. Correa, and C. Lemaréchal. *Convergence of some algorithms for convex minimization.* Math. Programming, 62(2):261–275, 1993.

[12] A. Daniilidis, P. Georgiev. *Approximate convexity and submonotonicity.* J. Math. Anal. Appl., vol. 291, 2004, pp. 117 – 144.

[13] J. E. Dennis jr., R. Schnabel. Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice Hall Series in Computational Mathematics. 1983.

[14] J.-B. Hiriart-Urruty, C. Lemaréchal. *Convex analysis and minimization algorithms, vol I: and vol II: Advanced theory and bundle methods*, vol. 306 of Grundlehren der mathematischen Wissenschaften, Springer Verlag, New York, Heidelberg, Berlin, 1993.

[15] C. Lemaréchal, J.-J. Strodiot, A. Bihain. *On a bundle algorithm for non-smooth optimization.* In: Mangasarian, Meyer, Robinson (eds.), Nonlinear Programming 4, Academic Press, 1981, pp. 245 – 282.

[16] A. Lewis, S. Wright. *A proximal method for composite minimization.* Preprint 2008.

[17] D. Noll, O. Prot, A. Rondepierre. *A proximity control algorithm to minimize non-smooth and non-convex functions.* Pacific Journal of Optimization, vol. 4, no. 3, 2008, pp. 569 – 602.

[18] R.T. Rockafellar, R. J.-B. Wets. Variational Analysis. Grundlehren der mathematischen Wissenschaften, vol. 317, Springer Verlag, 1998.

[19] C. Sagastizábal, W. Hare. *A redistributed proximal bundle method for nonconvex optimization.* Preprint 2009.

[20] C. Sagastizábal. *Composite proximal bundle method.* Preprint 2009.

[21] H. Schramm, J. Zowe. *A version of the bundle idea for minimizing nondifferentiable functions: conceptual idea, convergence analysis, numerical results.* SIAM Journal on Optimization, vol. 2, 1992, pp. 121 – 151.

[22] Simões, P. Apkarian, D. Noll. *Non-smooth multi-objective synthesis with applications.* Control Engineering Practice, vol. 17, no. 11, 2009, pp. 1338 – 1348.

[23] J.E. Spingarn. *Submonotone subdifferentials of Lipschitz functions.* Trans. Amer. Math. Soc., vol. 264, 1981, pp. 77 – 89.

[24] 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.