Chapter 1
Bundle method for non-convex minimization with inexact subgradients and function values

Dominikus Noll

Abstract We discuss a bundle method to minimize locally Lipschitz functions which are both non-convex and non-smooth. We analyze situations where only inexact subgradients or function values are available. For suitable classes of such non-smooth functions we prove convergence of our algorithm to approximate critical points.

1.1 Introduction

We consider optimization programs of the form

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz but neither differentiable nor convex. We present a bundle algorithm which converges to a critical point of (1.1) if exact function and subgradient evaluation of f are provided, and to an approximate critical point if subgradients or function values are inexact. Here $\bar{x} \in \mathbb{R}^n$ is approximate critical if

$$\operatorname{dist}(0, \partial f(\bar{x})) \leq \varepsilon, \tag{1.2}$$

where $\partial f(x)$ is the Clarke subdifferential of f at x.

The method discussed here extends the classical bundle concept to the non-convex setting by using down-shifted tangents as a substitute for cutting planes. This idea was already used in the 1980s in Lemaréchal's M2FC1 code [34] or in Zowe's BT codes [50,56]. Its convergence properties can be assessed by the model-based bundle techniques [6,7] and [43,44]. Recent numerical experiments using the

Dominikus Noll
Université Paul Sabatier, Institut de Mathématiques, Toulouse, France e-mail: noll@mip.ups-tlse.fr

down-shift mechanism are reported in [8,21,52]. In the original paper of Schramm and Zowe [50] down-shift is discussed for a hybrid method combining bundling, trust region and line-search elements.

For convex programs (1.1) bundle methods which can deal with inexact function values or subgradients have been discussed at least since 1985, see Kiwiel [28,30]. More recently, the topic has been revived by Hintermüller [24], who presented a method with exact function values but inexact subgradients $g \in \partial_\varepsilon f(x)$, where ε remains unknown to the user. Kiwiel [32] expands on this idea and presents an algorithm which deals with inexact function values and subgradients, both with unknown errors bounds. Kiwiel and Lemaréchal [33] extend the idea further to address column generation. Incremental methods to address large problems in stochastic programming or Lagrangian relaxation can be interpreted in the framework of inexact values and subgradients, see e.g. Emiel and Sagastizábal [17,18], Kiwiel [31]. In [41] Nedic and Bertsekas consider approximate functions and subgradients which are in addition affected by deterministic noise.

Nonsmooth methods without convexity have been considered by Wolfe [54], Shor [51], Mifflin [40], Schramm and Zowe [50], and more recently by Lukšan and Vlček [37], Noll and Apkarian [42], Fuduli *et al.* [19,20], Apkarian *et al.* [6], Noll *et al.* [43], Hare and Sagastizábal [23], Sagastizábal [49], Lewis and Wright [35], Noll [44]. In the context of control applications, early contributions are Polak and Wardi [46], Mayne and Polak [38,39], Kiwiel [29], Polak [45], Apkarian et al. [1–7], Bompart *et al.* [9]. All these approaches use exact knowledge of function values and subgradients.

The structure of the paper is as follows. In section 1.2 we explain the concept of an approximate subgradient. Section 1.3 discusses the elements of the algorithm, acceptance, tangent program, aggregation, cutting planes, recycling, and the management of proximity control. Section 1.4 presents the algorithm. Section 1.5 analyses the inner loop in the case of exact function values and inexact subgradients. Section 1.6 gives convergence of the outer loop. Section 1.7 extends to the case where function values are also inexact. Section 1.8 uses the convergence theory of sections 1.5 – 1.7 to derive a practical stopping test. Section 1.9 concludes with a motivating example from control.

1.2 Preparation

Approximate subgradients in convex bundle methods refer to the ε-subdifferential [26]:

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : g^\top (y-x) \leq f(y) - f(x) + \varepsilon \text{ for all } y \in \mathbb{R}^n\}, \tag{1.3}$$

whose central property is that $0 \in \partial f_\varepsilon(\bar{x})$ implies ε-minimality of $\bar{x}$, i.e., $f(\bar{x}) \leq \min f + \varepsilon$. Without convexity we cannot expect a tool with similar global properties. We shall work with the following very natural approximate subdifferential

$$\partial_{[\varepsilon]} f(x) = \partial f(x) + \varepsilon B, \tag{1.4}$$

where B is the unit ball in some fixed Euclidian norm, and $\partial f(x)$ is the Clarke subdifferential of f. The present section motivates this choice.

The first observation concerns the optimality condition (1.2) arising from the choice (1.4). Namely $0 \in \partial_{[\varepsilon]} f(\bar{x})$ can also be written as $0 \in \partial(f + \varepsilon\| \cdot - x\|)(x)$, meaning that a small perturbation of f is critical at x.

We can also derive a weak form of ε-optimality from $0 \in \partial_{[\varepsilon]} f(x)$ for composite functions $f = g \circ F$ with g convex and F smooth, or more generally, for lower C^2 functions, see [47], which have such a representation locally.

Lemma 1. *Let $f = g \circ F$ where g is convex and F is of class C^2, and suppose $0 \in \partial_{[\varepsilon]} f(x)$. Fix $r > 0$, and define*

$$c_r := \max_{\|d\|=1} \max_{\|x'-x\| \le r} \max_{\phi \in \partial g(F(x))} \phi^\top D^2 F(x')[d,d].$$

Then x is $(r\varepsilon + r^2 c_r/2)$-optimal on the ball $B(x,r)$.

Proof. We have to prove $f(x) \le f(x^+) + r\varepsilon + r^2 c_r/2$ for every $x^+ \in B(x,r)$. Write $x^+ = x + td$ for some $\|d\| = 1$ and $t \le r$. Since $0 \in \partial_{[\varepsilon]} f(x)$, and since $\partial f(x) = DF(x)^* \partial g(F(x))$, there exists $\phi \in \partial g(F(x))$ such that $\|DF(x)^* \phi\| \le \varepsilon$. In other words, $\|\phi^\top DF(x) d\| \le \varepsilon$ because $\|d\| = 1$. By the subgradient inequality we have

$$\phi^\top (F(x+td) - F(x)) \le g(F(x+td)) - g(F(x)) = f(x^+) - f(x). \tag{1.5}$$

Second-order Taylor expansion of $t \mapsto \phi^\top F(x+td)$ at $t = 0$ gives

$$\phi^\top F(x+td) = \phi^\top F(x) + t\phi^\top DF(x)d + \tfrac{t^2}{2}\phi^\top D^2 F(x_t)[d,d]$$

for some x_t on the segment $[x, x+td]$. Substituting this into (1.5) and using the definition of c_r gives

$$f(x) \le f(x^+) + t\|\phi^\top DF(x)d\| + \tfrac{t^2}{2}\|\phi^\top D^2 F(x_t)[d,d]\| \le f(x^+) + r\varepsilon + \tfrac{r^2}{2} c_r,$$

hence the claim. □

Remark 1. For convex f we can try to relate the two approximate subdifferentials in the sense that

$$\partial_\varepsilon f(x) \subset \partial_{[\varepsilon']} f(x)$$

for a suitable $\varepsilon' = \varepsilon'(x,\varepsilon)$. For a convex quadratic function $f(x) = \frac{1}{2}x^\top Qx + q^\top x$ it is known that $\partial_\varepsilon f(x) = \{\nabla f(x) + Q^{1/2}z : \frac{1}{2}\|z\|^2 \le \varepsilon\}$, [26], so that $\partial_\varepsilon f(x) \subset \partial f(x) + \varepsilon' B = \partial_{[\varepsilon']} f(x)$ for $\varepsilon' = \sup\{\|Q^{1/2}z\| : \frac{1}{2}\|z\|^2 \le \varepsilon\}$, which means that $\varepsilon'(x,\varepsilon)$ is independent of x and behaves as $\varepsilon' = \mathscr{O}(\varepsilon^{1/2})$. We expect this type of relation to hold as soon as f has curvature information around x. On the other hand, if $f(x) = |x|$, then $\partial f_\varepsilon(x) = \partial f(x) + \frac{\varepsilon}{|x|}B$ for $x \neq 0$ (and $\partial_\varepsilon f(0) = \partial f(0)$), which means that the relationship $\varepsilon' = \varepsilon/|x|$ is now linear in ε for fixed $x \neq 0$. In general it is difficult to relate ε to ε'. See Hiriart-Urruty and Seeger [25] for more information on this question.

Remark 2. For composite functions $f = g \circ F$ with g convex and F of class C^1 we can introduce

$$\partial_\varepsilon f(x) = DF(x)^* \partial_\varepsilon g(F(x)),$$

where $\partial_\varepsilon g(y)$ is the usual convex ε-subdifferential (1.3) of g and $DF(x)^*$ is the adjoint of the differential of F at x. Since the corresponding chain rule is valid in the case of an affine F, $\partial_\varepsilon f(x)$ is consistent with (1.3). Without convexity $\partial f_\varepsilon(x)$ no longer preserves the global properties of (1.3). Yet, for composite functions $f = g \circ F$ a slightly more general version of Lemma 1 combining $\partial_{[\sigma]} f$ and $\partial_\varepsilon f$ can be proved along the lines of [42, Lemma 2]. In that reference the result is shown for the particular case $g = \lambda_1$, but an extension can be obtained by reasoning as in Lemma 1.

Remark 3. For convex f the set $\partial_{[\varepsilon]} f(x)$ coincides with the Fréchet ε-subdifferential $\partial_\varepsilon^F f(x)$. According to [36, Cor. 3.2] the same remains true for approximate convex functions. For the latter see section 1.5.

1.3 Elements of the algorithm

1.3.1 Local model

Let x be the current iterate of the outer loop. The inner loop with counter k generates a sequence y^k of trial steps, one of which is eventually accepted to become the new serious step x^+. At each instant k we dispose of a convex working model $\phi_k(\cdot,x)$, which approximates f in a neighborhood of x. We suppose that we know at least one approximate subgradient $g(x) \in \partial_{[\varepsilon]} f(x)$. The affine function

$$m_0(\cdot,x) = f(x) + g(x)^\top(\cdot - x)$$

will be referred to as the exactness plane at x. For the moment we assume that it gives an exact value of f at x, but not an exact subgradient. The algorithm assures $\phi_k(\cdot,x) \geq m_0(\cdot,x)$ at all times k, so that $g(x) \in \partial\phi_k(x,x)$ for all k. In fact we construct $\phi_k(\cdot,x)$ in such a way that $\partial\phi_k(x,x) \subset \partial_{[\varepsilon]} f(x)$ at all times k.

Along with the first-order working model $\phi_k(\cdot,x)$ we also consider an associated second-order model of the form

$$\Phi_k(y,x) = \phi_k(y,x) + \tfrac{1}{2}(y-x)^\top Q(x)(y-x),$$

where $Q(x)$ depends on the serious iterate x, but is fixed during the inner loop k. We allow $Q(x)$ to be indefinite.

1.3.2 Cutting planes

Suppose y^k is a null step. Then model $\Phi_k(\cdot,x)$ which gave rise to y^k was not rich enough and we have to improve it at the next inner loop step $k+1$ in order to perform better. We do this by modifying the first-order part. In convex bundling one includes a cutting plane at y^k into the new model $\phi_{k+1}(\cdot,x)$. This remains the same with approximate subgradients and values (cf. [24,32]) as soon as the concept of cutting plane is suitably modified. Notice that we have access to $g_k \in \partial_{[\varepsilon]} f(y^k)$, which gives us an approximate tangent

$$t_k(\cdot) = f(y^k) + g_k^\top(\cdot - y^k)$$

at y^k. Since f is not convex, we cannot use $t_k(\cdot)$ directly as cutting plane. Instead we use a technique originally developed in Schramm and Zowe [50] and Lemaréchal [34], which consists in shifting $t_k(\cdot)$ downwards until it becomes useful for $\phi_{k+1}(\cdot,x)$. Fixing $c > 0$ once and for all, we call

$$s_k := [t_k(x) - f(x)]_+ + c\|y^k - x\|^2 \tag{1.6}$$

the down-shift and introduce

$$m_k(\cdot,x) = t_k(\cdot) - s_k,$$

called the down-shifted tangent.

We sometimes use the following more stringent notation, where no reference to the counter k is made. The approximate tangent is $t_{y,g}(\cdot) = f(y) + g^\top(\cdot - y)$, bearing a reference to the point y where it is taken and to the specific approximate subgradient $g \in \partial_{[\varepsilon]} f(y)$. The down-shifted tangent is then $m_{y,g}(\cdot,x) = t_{y,g}(\cdot) - s$, where $s = s(y,g,x) = [t_{y,g}(x) - f(x)]_+ + c\|y-x\|^2$ is the down-shift. Since this notation is fairly heavy, we will try to avoid it whenever possible and switch to the former, bearing in mind that $t_k(\cdot)$ depends both on y^k and the subgradient $g_k \in \partial_{[\varepsilon]} f(y^k)$. Similarly, the down-shifted tangent plane $m_k(\cdot,x)$ depends on y^k, g^k, and on x, as does the down-shift s_k. We use $m_k(\cdot,x)$ as a substitute for the classical cutting plane. For convenience we continue to call $m_k(\cdot,x)$ a cutting plane.

The cutting plane satisfies $m_k(x,x) \le f(x) - c\|y^k - x\|^2$, which assures that it does not interfere with the subdifferential of $\phi_{k+1}(\cdot,x)$ at x. We build $\phi_{k+1}(\cdot,x)$ in such a way that it has $m_k(\cdot,x)$ as an affine minorant.

Proposition 1. *Let* $\phi_{k+1}(\cdot,x) = \max\{m_\nu(\cdot,x) : \nu = 0,\dots,k\}$. *Then* $\partial\phi_{k+1}(x,x) \subset \partial_{[\varepsilon]} f(x)$.

Proof. As all the down-shifts s_k are positive, $\phi_{k+1}(y,x) = m_0(y,x)$ in a neighborhood of x, hence $\partial\phi_{k+1}(x,x) = \partial m_0(x,x) = \{g(x)\} \subset \partial_{[\varepsilon]} f(x)$. □

1.3.3 Tangent program

Given the local model $\Phi_k(\cdot,x) = \phi_k(\cdot,x) + \frac{1}{2}(\cdot - x)^\top Q(x)(\cdot - x)$ at serious iterate x and inner-loop counter k, we solve the tangent program

$$\min_{y\in\mathbb{R}^n} \Phi_k(y,x) + \tfrac{\tau_k}{2}\|y-x\|^2. \tag{1.7}$$

We assume that $Q(x) + \tau_k I \succ 0$, which means (1.7) is strictly convex and has a unique solution y^k, called a trial step. The optimality condition for (1.7) implies

$$(Q(x) + \tau_k I)(x - y^k) \in \partial\phi_k(y^k,x). \tag{1.8}$$

If $\phi_k(\cdot,x) = \max\{m_\nu(\cdot,x) : \nu = 0,\dots,k\}$, with $m_\nu(\cdot,x) = a_\nu + g_\nu^\top(\cdot - x)$, then we can find $\lambda_0 \geq 0,\dots,\lambda_k \geq 0$, summing up to 1, such that

$$g_k^* := (Q(x) + \tau_k I)(x - y^k) = \sum_{\nu=0}^{k} \lambda_\nu g_\nu.$$

Traditionally, g_k^* is called the aggregate subgradient at y^k. We build the aggregate plane

$$m_k^*(\cdot,x) = a_k^* + g_k^{*\top}(\cdot - x),$$

where $a_k^* = \sum_{\nu=1}^k \lambda_\nu a_\nu$. Keeping $m_k^*(\cdot,x)$ as an affine minorant of $\phi_{k+1}(\cdot,x)$ allows to drop some of the older cutting planes to avoid overflow. As $\partial\phi_k(y^k,x)$ is the subdifferential of a max-function, we know that $\lambda_\nu > 0$ precisely for those $m_\nu(\cdot,x)$ which are active at y^k. That is, $\sum_{\nu=1}^k \lambda_\nu m_\nu(y^k,x) = \phi_k(y^k,x)$. Therefore the aggregate plane satisfies

$$m_k^*(y^k,x) = \phi_k(y^k,x). \tag{1.9}$$

As our algorithm chooses ϕ_{k+1} such that $m_k^*(\cdot,x) \leq \phi_{k+1}(\cdot,x)$, we have $\phi_k(y^k,x) \leq \phi_{k+1}(y^k,x)$. All this follows the classical line originally proposed in Kiwiel [27]. Maintaining a model $\phi_k(\cdot,x)$ which contains aggregate subgradients from previous sweeps instead of *all* the older g_ν, $\nu = 0,\dots,k$ does not alter the statement of Proposition 1, nor of formula (1.9).

1.3.4 Testing acceptance

Having computed the k^{th} trial step y^k via (1.7), we have to decide whether it should be accepted as the new serious iterate x^+. We compute the test quotient

$$\rho_k = \frac{f(x) - f(y^k)}{f(x) - \Phi_k(y^k,x)}.$$

Fixing constants $0 < \gamma < \Gamma < 1$, we call y^k *bad* if $\rho_k < \gamma$, and *good* if $\rho_k \geq \Gamma$. If y^k is not bad, meaning $\rho_k \geq \gamma$, then it is accepted to become x^+. We refer to this as a serious step. Here the inner loop ends. On the other hand, if y^k is bad, then it is rejected and referred to as a null step. In this case the inner loop continues.

1.3.5 Management of τ in the inner loop

The most delicate point is the management of the proximity control parameter during the inner loop. Namely, it may turn out that the trial steps y^k proposed by the tangent program (1.7) are too far from the current x, so that no decrease below $f(x)$ can be achieved. In the convex case one relies entirely on the mechanism of cutting planes. Indeed, if y^k is a null step, then the convex cutting plane, when added to model $\phi_{k+1}(\cdot,x)$, will cut away the unsuccessful y^k, paving the way for a better y^{k+1} at the next sweep.

The situation is more complicated without convexity, where cutting planes are no longer tangents to f. In the case of down-shifted tangents the information stored in the ideal set of all theoretically available cutting planes may *not* be sufficient to represent f correctly when y^k is far away from x. This is when we have to force smaller steps by increasing τ, i.e., by tightening proximity control. As a means to decide when this has to happen, we use the parameter

$$\widetilde{\rho}_k = \frac{f(x) - M_k(y^k,x)}{f(x) - \Phi_k(y^k,x)}, \tag{1.10}$$

where $m_k(\cdot,x)$ is the new cutting plane drawn for y^k as in Section 1.3.1, and $M_k(\cdot,x) = m_k(\cdot,x) + \frac{1}{2}(\cdot - x)^\top Q(x)(\cdot - x)$. We fix a parameter $\widetilde{\gamma}$ with $\gamma < \widetilde{\gamma} < 1$ and make the following decision.

$$\tau_{k+1} = \begin{cases} 2\tau_k & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k \geq \widetilde{\gamma} \\ \tau_k & \text{if } \rho_k < \gamma \text{ and } \widetilde{\rho}_k < \widetilde{\gamma} \end{cases}. \tag{1.11}$$

The idea in (1.11) can be explained as follows. The quotient $\widetilde{\rho}_k$ in (1.10) can also be written as $\widetilde{\rho}_k = \left(f(x) - \Phi_{k+1}(y^k,x)\right) / \left(f(x) - \Phi_k(y^k,x)\right)$, because the cutting plane at stage k will be integrated into model Φ_{k+1} at stage $k+1$. If $\widetilde{\rho}_k \approx 1$, we can therefore conclude that adding the new cutting plane at the null step y^k hardly changes the situation. Put differently, had we known the cutting plane before computing y^k, the result would not have been much better. In this situation we decide to force smaller trial steps by increasing the τ-parameter. If on the other hand $\widetilde{\rho}_k \ll 1$, then the gain of information provided by the new cutting plane at y^k is substantial with regard to the information already stored in Φ_k. Here we continue to add cutting planes and aggregate planes only, hoping that we will still make progress *without* having to increase τ. The decision $\widetilde{\rho}_k \approx 1$ versus $\widetilde{\rho}_k \ll 1$ is formalized by the rule (1.11).

Remark 4. By construction $\widetilde{\rho}_k \geq 0$, because aggregation assures that $\phi_{k+1}(y^k,x) \geq \phi_k(y^k,x)$. Notice that in contrast ρ_k may be negative. Indeed, $\rho_k < 0$ means that the trial step y^k proposed by the tangent program (1.7) gives no descent in the function values, meaning that it is clearly a bad step.

1.3.6 Management of τ in the outer loop.

The proximity parameter τ will also be managed dynamically between serious steps $x \to x^+$. In our algorithm we use a memory parameter $\tau_j^\sharp$, which is specified at the end of the $(j-1)^{\text{st}}$ inner loop, and serves to initialize the j^{th} inner loop with $\tau_1 = \tau_j^\sharp$.

A first rule which we already mentioned is that we need $Q(x^j) + \tau_k I \succ 0$ for all k during the j^{th} inner loop. Since τ is never decreased during the inner loop, we can assure this if we initialize $\tau_1 > -\lambda_{\min}(Q(x^j))$.

A more important aspect is the following. Suppose the $(j-1)^{\text{st}}$ inner loop ended at inner loop counter k_{j-1}, i.e. $x^j = y^{k_{j-1}}$ with $\rho_{k_{j-1}} \geq \gamma$. If acceptance was good, i.e., $\rho_{k_{j-1}} \geq \Gamma$, then we can trust our model, and we account for this by storing a smaller parameter $\tau_j^\sharp = \frac{1}{2}\tau_{k_{j-1}} < \tau_{k_{j-1}}$ for the j^{th} outer loop. On the other hand, if acceptance of the $(j-1)^{\text{st}}$ step was neither good nor bad, meaning $\gamma \leq \rho_{k_{j-1}} \leq \Gamma$, then there is no reason to decrease τ for the next outer loop, so we memorize $\tau_{k_{j-1}}$, the value we had at the end of the $(j-1)^{\text{st}}$ inner loop. Altogether

$$\tau_j^\sharp = \begin{cases} \max\{\frac{1}{2}\tau_{k_{j-1}}, -\lambda_{\min}(Q(x^j)) + \zeta\} & \text{if } \rho_{k_{j-1}} \geq \Gamma \\ \max\{\tau_{k_{j-1}}, -\lambda_{\min}(Q(x^j)) + \zeta\} & \text{if } \gamma \leq \rho_{k_{j-1}} < \Gamma \end{cases} \tag{1.12}$$

where $\zeta > 0$ is some small threshold fixed once and for all.

1.3.7 Recycling of planes

In a convex bundle algorithm one keeps in principle all cutting planes in the model, using aggregation to avoid overflow. In the non-convex case this is no longer possible. Cutting planes are down-shifted tangents, which links them to the value $f(x)$ of the current iterate x. As we pass from x to a new serious iterate x^+, the cutting plane $m_{z,g}(\cdot,x) = a + g^\top(\cdot - x)$ with $g \in \partial_{[\varepsilon]} f(z)$ for some z cannot be used as such, because we have no guarantee whether $a + g^\top(x^+ - x) \leq f(x^+)$. But we can down-shift it again if need be. We recycle the plane as

$$m_{z,g}(\cdot,x^+) = a - s^+ + g^\top(\cdot - x), \quad s^+ = [m_{z,g}(x^+,x) - f(x^+)]_+ + c\|x^+ - z\|^2.$$

In addition one may also apply a test whether z is too far from x^+ to be of interest, in which case the plane should simply be removed from the stock.

1.4 Algorithm

Algorithm 1. Proximity control algorithm for (1.1).

Parameters: $0 < \gamma < \Gamma < 1$, $\gamma < \widetilde{\gamma} < 1$, $0 < q < \infty$, $q < T < \infty$, $\widetilde{\varepsilon} > 0$.

1: **Initialize outer loop**. Choose initial guess x^1 and an initial matrix $Q_1 = Q_1^\top$ with $-qI \preceq Q_1 \preceq qI$. Fix memory control parameter $\tau_1^\sharp$ such that $Q_1 + \tau_1^\sharp I \succ 0$. Put $j = 1$.

2: **Stopping test**. At outer loop counter j, stop if $0 \in \partial_{[\widetilde{\varepsilon}]} f(x^j)$. 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 $\phi_1(\cdot, x^j)$, possibly recycling some planes from previous sweep $j-1$, and let $\Phi_1(\cdot, x^j) = \phi_1(\cdot, x^j) + \frac{1}{2}(\cdot - x^j)^\top Q_j (\cdot - x^j)$.

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

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

The solution is the new trial step y^k.

5: **Acceptance test.** Check whether

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

If this is the case put $x^{j+1} = y^k$ (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**. Compute a cutting plane $m_k(\cdot, x^j)$ at x^j for the null step y^k. Let $M_k(\cdot, x^j) = m_k(\cdot, x^j) + \frac{1}{2}(\cdot - x^j)^\top Q_j (\cdot - x^j)$ and compute secondary control parameter

$$\widetilde{\rho}_k = \frac{f(x^j) - M_k(y^k, x^j)}{f(x^j) - \Phi_k(y^k, x^j)}.$$

Put $\tau_{k+1} = \begin{cases} \tau_k, & \text{if } \widetilde{\rho}_k < \widetilde{\gamma} \quad \text{(bad)} \\ 2\tau_k, & \text{if } \widetilde{\rho}_k \geq \widetilde{\gamma} \quad \text{(too bad)} \end{cases}$

7: **Update working model**. Build new convex working model $\phi_{k+1}(\cdot, x^j)$ based on null step y^k by adding the new cutting plane $m_k(\cdot, x^j)$ (and using aggregation to avoid overflow). Keep exactness plane in the working model. 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

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

Increase $\tau_{j+1}^\sharp$ if necessary to ensure $Q_{j+1} + \tau_{j+1}^\sharp I \succ 0$. If $\tau_{j+1}^\sharp > T$ then re-set $\tau_{j+1}^\sharp = T$. Increase outer loop counter j by 1 and loop back to step 2.

1.5 Analysis of the inner loop

In this section we analyze the inner loop and show that there are two possibilities. Either the inner loop terminates finitely with a step $x^+ = y^k$ satisfying $\rho_k \geq \gamma$. Or we get an infinite sequence of null steps y^k which converges to x. In the latter case, we conclude that $0 \in \partial_{[\widetilde{\varepsilon}]} f(x)$, i.e., that x is approximate optimal.

Suppose the inner loop turns forever. Then there are two possibilities. Either τ_k is increased infinitely often, so that $\tau_k \to \infty$, or τ_k is frozen, $\tau_k = \tau_{k_0}$ for some k_0 and all $k \geq k_0$. These scenarios will be analyzed in Lemmas 3 and 4. Since the matrix $Q(x)$ is fixed during the inner loop, we write it simply as Q.

To begin with, we need an auxiliary construction. We define the following convex function:

$$\phi(y,x) = \sup\{m_{z,g}(y,x) : z \in B(0,M), g \in \partial_{[\varepsilon]} f(y)\}, \tag{1.13}$$

where $B(0,M)$ is a fixed ball large enough to contain x and all trial steps encountered during the inner loop. Recall that $m_{z,g}(\cdot,x)$ is the cutting plane at z with approximate subgradient $g \in \partial_{[\varepsilon]} f(z)$ with respect to the serious iterate x. Due to boundedness of $B(0,M)$, $\phi(\cdot,x)$ is defined everywhere.

Lemma 2. *We have $\phi(x,x) = f(x)$, $\partial\phi(x,x) = \partial_{[\varepsilon]} f(x)$, and ϕ is jointly upper-semicontinuous. Moreover, if $y^k \in B(0,M)$ for all k, then $\phi_k(\cdot,x) \leq \phi(\cdot,x)$ for every first-order working model ϕ_k.*

Proof. 1) The first statement follows because every cutting plane drawn at some $z \neq x$ and $g \in \partial_{[\varepsilon]} f(z)$ satisfies $m_{z,g}(x,x) \leq f(x) - c\|x-z\|^2 < f(x)$, while cutting planes at x obviously have $m_{x,g}(x,x) = f(x)$.

2) Concerning the second statement, let us first prove $\partial_{[\varepsilon]} f(x) \subset \partial\phi(x,x)$. We consider the set of limiting subgradients

$$\partial^l f(x) = \{\lim_{k\to\infty} \nabla f(y^k) : y^k \to x, f \text{ is differentiable at } y^k\}.$$

Then $\operatorname{co}\partial^l f(x) = \partial f(x)$ by [14]. It therefore suffices to show $\partial^l f(x) + \varepsilon B \subset \partial\phi(x,x)$, because $\partial\phi(x,x)$ is convex and we then have $\partial\phi(x,x) \supset \operatorname{co}(\partial^l f(x) + \varepsilon B) = \operatorname{co}\partial^l f(x) + \varepsilon B = \partial f(x) + \varepsilon B$.

Let $g_a \in \partial^l f(x) + \varepsilon B$. We have to show $g_a \in \partial\phi(x,x)$. Choose $g \in \partial^l f(x)$ such that $\|g - g_a\| \leq \varepsilon$. Pick a sequence $y^k \to x$ and $g_k = \nabla f(y^k) \in \partial f(y^k)$ such that $g_k \to g$. Let $g_{a,k} = g_k + g_a - g$, then $g_{a,k} \in \partial_{[\varepsilon]} f(y^k)$ and $g_{a,k} \to g_a$. Let $m_k(\cdot,x)$ be the cutting plane drawn at y^k with approximate subgradient $g_{a,k}$, then $m_k(y^k,x) \leq \phi(y^k,x)$. By the definition of the downshift process

$$m_k(y,x) = f(y^k) + g_{a,k}^\top (y - y^k) - s_k$$

where s_k is the down-shift (1.6). There are two cases, $s_k = c\|y^k - x\|^2$, and $s_k = t_k(x) - f(x) + c\|y^k - x\|^2$, according to whether the term $[\dots]_+$ in (1.6) equals zero or not.

Let us start with the second case, where $t_k(x) > f(x)$. Then $s_k = f(y^k) + g_{a,k}^\top(x - y^k) - f(x) + c\|y^k - x\|^2$ and

$$\begin{aligned} m_k(y,x) &= f(y^k) + g_{a,k}^\top(y - y^k) - f(y^k) - g_{a,k}^\top(x - y^k) + f(x) - c\|y^k - x\|^2 \\ &= f(x) + g_{a,k}^\top(y - x) - c\|y^k - x\|^2. \end{aligned}$$

Therefore

$$\phi(y,x) - \phi(x,x) \geq m_k(y,x) - f(x) = g_{a,k}^\top(y - x) - c\|y^k - x\|^2.$$

Passing to the limit using $y^k \to x$ and $g_{a,k} \to g_a$ proves $g_a \in \partial\phi(x,x)$.

It remains to discuss the first case, where $t_k(x) \leq f(x)$, so that $s_k = c\|y^k - x\|^2$. Then

$$m_k(\cdot,x) = f(y^k) + g_{a,k}^\top(\cdot - y^k) - c\|y^k - x\|^2.$$

Therefore

$$\begin{aligned} \phi(y,x) - \phi(x,x) &\geq m_k(y,x) - f(x) \\ &= f(y^k) - f(x) + g_{a,k}^\top(y - y^k) - c\|y^k - x\|^2 \\ &= f(y^k) - f(x) + g_{a,k}^\top(x - y^k) + g_{a,k}^\top(y - x) - c\|y^k - x\|^2. \end{aligned}$$

As y is arbitrary, we have $g_{a,k} \in \partial_{|\zeta_k|}\phi(x,x)$, where $\zeta_k = f(y^k) - f(x) + g_{a,k}^\top(x - y^k) - c\|y^k - x\|^2$. Since $\zeta_k \to 0$, $y^k \to x$ and $g_{a,k} \to g_a$, we deduce again $g_a \in \partial\phi(x,x)$. Altogether for the two cases $[\ldots]_+ = 0$ and $[\ldots]_+ > 0$ we have shown $\partial^l f(x) + \varepsilon B \subset \partial\phi(x,x)$.

3) Let us now prove $\partial\phi(x,x) \subset \partial f(x) + \varepsilon B$. Let $g \in \partial\phi(x,x)$ and $m(\cdot,x) = f(x) + g^\top(\cdot - x)$ the tangent plane to the graph of $\phi(\cdot,x)$ at x associated with g. By convexity $m(\cdot,x) \leq \phi(\cdot,x)$. We fix $h \in \mathbb{R}^n$ and consider the values $\phi(x+th,x)$ for $t > 0$. According to the definition of $\phi(\cdot,x)$ we have $\phi(x+th,x) = m_{z_t,g_t}(x+th,x)$, where $m_{z_t,g_t}(\cdot,x)$ is a cutting plane drawn at some $z_t \in B(0,M)$ with $g_t \in \partial_{[\varepsilon]} f(z_t)$. The slope of the cutting plane along the ray $x + \mathbb{R}_+ h$ is $g_t^\top h$. Now the cutting plane passes through $\phi(x+th,x) \geq m(x+th,x)$, which means that its value at $x+th$ is above the value of the tangent. On the other hand, according to the downshift process, the cutting plane satisfies $m_{z_t,g_t}(x,x) \leq f(x) - c\|x - z_t\|^2$. Its value at x is therefore below the value of $m(x,x) = f(x)$. These two fact together tell us that $m_{z_t,g_t}(\cdot,x)$ is steeper than $m(\cdot,x)$ along the ray $x + \mathbb{R}_+ h$. In other words, $g^\top h \leq g_t^\top h$. Next observe that $\phi(x+th,x) \to \phi(x,x) = f(x)$ as $t \to 0^+$. That implies $m_{z_t,g_t}(x+th,x) \to f(x)$. Since by the definition of downshift $m_{z_t,g_t}(x+th,x) \leq f(x) - c\|x - z_t\|^2$, it follows that we must have $\|x - z_t\|^2 \to 0$, i.e., $z_t \to x$ as $t \to 0^+$. Passing to a subsequence, we may assume $g_t \to \widehat{g}$ for some $\widehat{g}$. With $z_t \to x$ it follows from upper semicontinuity of the Clarke subdifferential that $\widehat{g} \in \partial_{[\varepsilon]} f(x)$. On the other hand, $g^\top h \leq g_t^\top h$ for all t implies $g^\top h \leq \widehat{g}^\top h$. Therefore $g^\top h \leq \sigma_K(h) = \max\{\tilde{g}^\top h : \tilde{g} \in K\}$, where σ_K is the support function of $K = \partial_{[\varepsilon]} f(x)$. Given that h was arbitrary, and as K is closed convex, this implies $g \in K$ by Hahn-Banach.

4) Upper semi-continuity of ϕ follows from upper semi-continuity of the Clarke subdifferential. Indeed, let $x_j \to x$, $y_j \to y$. Using the definition (1.13) of ϕ, find cutting planes $m_{z_j,g_j}(\cdot,x_j) = t_{z_j}(\cdot) - s_j$ at serious iterate x_j, drawn at z_j with $g_j \in \partial_{[\varepsilon]} f(z_j)$, such that $\phi(y_j,x_j) \leq m_{z_j,g_j}(y_j,x_j) + \varepsilon_j$ and $\varepsilon_j \to 0$. We have $t_{z_j}(y) = f(z_j) + g_j^\top(y - z_j)$. Passing to a subsequence, we may assume $z_j \to z$ and $g_j \to g \in \partial_{[\varepsilon]} f(z)$. That means $t_{z_j}(\cdot) \to t_z(\cdot)$, and since $y_j \to y$ also $t_{z_j}(y_j) \to t_z(y)$. In order to conclude for the $m_{z_j,g_j}(\cdot,x_j)$ we have to see how the down-shift behaves. We have indeed $s_j \to s$, where s is the down-shift at z with respect to the approximate subgradient g and serious iterate x. Therefore $m_{z,g}(\cdot,x) = t_z(\cdot) - s$. This shows $m_{z_j,g_j}(\cdot,x_j) = t_{z_j}(\cdot) - s_j \to t_z(\cdot) - s = m_{z,g}(\cdot,x)$ as $j \to \infty$, and then also $m_{z_j,g_j}(y_j,x_j) = t_{z_j}(y_j) - s_j \to t_z(y) - s = m_{z,g}(y,x)$, where uniformity comes from boundedness of the g_j. This implies $\lim m_{z_j,g_j}(y_j,x_j) = m_{z,g}(y,x) \leq \phi(y,x)$ as required.

5) The inequality $\phi_k \leq \phi$ is clear, because $\phi_k(\cdot,x)$ is built from cutting planes $m_k(\cdot,x)$, and all these cutting planes are below the envelope $\phi(\cdot,x)$. □

Remark 5. In [43, 44] the case $\varepsilon = 0$ is discussed and a function $\phi(\cdot,x)$ with the properties in Lemma 2 is called a first-order model of f at x. It can be understood as a generalized first-order Taylor expansion of f at x. Every locally Lipschitz function f 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 at x. In the present situation it is reasonable to call $\phi(\cdot,x)$ an ε-model of f at x.

Following [36] a function f is called ε-convex on an open convex set U if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) + \varepsilon t(1-t)\|x-y\|$ for all $x,y \in U$ and $0 \leq t \leq 1$. Every ε-convex function satisfies $f'(y,x-y) \leq f(x) - f(y) + \varepsilon\|x-y\|$, hence for $g \in \partial f(y)$:

$$g^\top(x-y) \leq f(x) - f(y) + \varepsilon\|x-y\|. \tag{1.14}$$

A function f is called approximate convex if for every x and $\varepsilon > 0$ there exists $\delta > 0$ such that f is ε-convex on $B(x,\delta)$. Using results from [16] and [36] one may show that approximate convex functions coincide with lower C^1 function in the sense of Spingarn [53].

Lemma 3. *Suppose the inner loop turns forever and $\tau_k \to \infty$.*

1. *If f is ε'-convex on a set containing all y^k, $k \geq k_0$, then $0 \in \partial_{[\widetilde{\varepsilon}]} f(x)$, where $\widetilde{\varepsilon} = \varepsilon + (\varepsilon' + \varepsilon)/(\widetilde{\gamma} - \gamma)$.*
2. *If f is lower C^1, then $0 \in \partial_{[\alpha\varepsilon]} f(x)$, where $\alpha = 1 + (\widetilde{\gamma} - \gamma)^{-1}$.*

Proof. i) The second statement follows from the first, because every lower C^1 function is approximate convex, hence ε'-convex on a suitable neighborhood of x. We therefore concentrate on the first statement.

ii) By assumption none of the trial steps is accepted, so that $\rho_k < \gamma$ for all $k \in \mathbb{N}$. Since τ_k is increased infinitely often, there are infinitely many inner loop instances

k where $\widetilde{\rho}_k \geq \widetilde{\gamma}$. Let us prove that under these circumstances $y^k \to x$. Recall that $g_k^* = (Q + \tau_k I)(x - y^k) \in \partial \phi_k(y^k, x)$. By the subgradient inequality this gives

$$g_k^{*\top}(x - y^k) \leq \phi_k(x,x) - \phi_k(y^k, x). \tag{1.15}$$

Now use $\phi_k(x,x) = f(x)$ and observe that $m_0(y^k,x) \leq \phi_k(y^k,x)$, where $m_0(\cdot,x)$ is the exactness plane. Since $m_0(y,x) = f(x) + g(x)^\top(y - x)$ for some $g(x) \in \partial_{[\varepsilon]} f(x)$, expanding the term on the left of (1.15) gives

$$(x - y^k)^\top (Q + \tau_k I)(x - y^k) \leq g(x)^\top (x - y^k) \leq \|g(x)\| \|x - y^k\|. \tag{1.16}$$

Since $\tau_k \to \infty$, the term on the left hand side of (1.16) behaves asymptotically like $\tau_k \|x - y^k\|^2$. Dividing (1.16) by $\|x - y^k\|$ therefore shows that $\tau_k \|x - y^k\|$ is bounded by $\|g(x)\|$. As $\tau_k \to \infty$, this could only mean $y^k \to x$.

iii) Let us use $y^k \to x$ and go back to formula (1.15). Since the left hand side of (1.15) tends to 0 and $\phi_k(x,x) = f(x)$, we see that the limit superior of $\phi_k(y^k,x)$ is $f(x)$. On the other hand, $\phi_k(y^k,x) \geq m_0(y^k,x)$, where $m_0(\cdot,x)$ is the exactness plane. Since clearly $m_0(y^k,x) \to m_0(x,x) = f(x)$, the limit inferior is also $f(x)$, and we conclude that $\phi_k(y^k,x) \to f(x)$.

Keeping this in mind, let us use the subgradient inequality (1.15) again and subtract a term $\frac{1}{2}(x - y^k)^\top Q(x - y^k)$ from both sides. That gives the estimate

$$\tfrac{1}{2}(x - y^k)^\top Q(x - y^k) + \tau_k \|x - y^k\|^2 \leq f(x) - \Phi_k(y^k, x).$$

Fix $0 < \zeta < 1$. Using $\tau_k \to \infty$ we have

$$(1 - \zeta)\tau_k \|x - y^k\| \leq \|g_k^*\| \leq (1 + \zeta)\tau_k \|x - y^k\|$$

and also

$$\tfrac{1}{2}(x - y^k)^\top Q(x - y^k) + \tau_k \|x - y^k\|^2 \geq (1 - \zeta)\tau_k \|x - y^k\|^2$$

for sufficiently large k. Therefore,

$$f(x) - \Phi_k(y^k, x) \geq \tfrac{1-\zeta}{1+\zeta} \|g_k^*\| \|x - y^k\| \tag{1.17}$$

for k large enough.

iv) Now let $\eta_k := \operatorname{dist}\big(g_k^*, \partial \phi(x,x)\big)$. We argue that $\eta_k \to 0$. Indeed, using the subgradient inequality at y^k in tandem with $\phi(\cdot,x) \geq \phi_k(\cdot,x)$, we have for all $y \in \mathbb{R}^n$:

$$\phi(y,x) \geq \phi_k(y^k,x) + g_k^{*\top}(y - y^k).$$

Here our upper envelope function (1.13) is defined such that the ball $B(0,M)$ contains x and all trial points y^k at which cutting planes are drawn.

Since the subgradients g_k^* are bounded by part ii), there exists an infinite subsequence $\mathcal{N} \subset \mathbb{N}$ such that $g_k^* \to g^*$, $k \in \mathcal{N}$, for some g^*. Passing to the limit $k \in \mathcal{N}$ and using $y^k \to x$ and $\phi_k(y^k,x) \to f(x) = \phi(x,x)$, we have $\phi(y,x) \geq \phi(x,x) + g^{*\top}(y -$

$x)$ for all y. Hence $g^* \in \partial\phi(x,x)$, which means $\eta_k = \mathrm{dist}(g_k^*, \partial\phi(x,x)) \le \|g_k^* - g^*\| \to 0$, $k \in \mathscr{N}$, proving the argument.

v) Using the definition of η_k, choose $\tilde g_k \in \partial\phi(x,x)$ such that $\|g_k^* - \tilde g_k\| = \eta_k$. Now let $\mathrm{dist}(0,\partial\phi(x,x)) = \eta$. Then $\|\tilde g_k\| \ge \eta$ for all $k \in \mathscr{N}$. Hence $\|g_k^*\| \ge \eta - \eta_k > (1-\zeta)\eta$ for $k \in \mathscr{N}$ large enough, given that $\eta_k \to 0$ by iv). Going back with this to (1.17) we deduce

$$f(x) - \Phi_k(y^k, x) \ge \tfrac{(1-\zeta)^2}{1+\zeta}\eta\|x - y^k\| \tag{1.18}$$

for $k \in \mathscr{N}$ large enough.

vi) We claim that $f(y^k) \le M_k(y^k,x) + (1+\zeta)(\varepsilon' + \varepsilon)\|x - y^k\|$ for all k sufficiently large. Indeed, we have $m_k(\cdot,x) = t_k(\cdot) - s_k$, where s_k is the down-shift of the approximate tangent $t_k(\cdot)$ at y^k, $g_{\varepsilon k} \in \partial_{[\varepsilon]} f(y^k)$, and with regard to the serious iterate x. There are two cases. Assume first that $t_k(x) > f(x)$. Then

$$\begin{aligned} m_k(y,x) &= f(y^k) + g_{\varepsilon k}^\top (y - y^k) - s_k \\ &= f(y^k) + g_{\varepsilon k}^\top (y - y^k) - c\|x - y^k\|^2 - t_k(x) + f(x) \\ &= f(x) + g_{\varepsilon k}^\top (y - x) - c\|x - y^k\|^2. \end{aligned}$$

In consequence

$$\begin{aligned} f(y^k) - m_k(y^k,x) &= f(y^k) - f(x) - g_{\varepsilon k}^\top (y^k - x) + c\|x - y^k\|^2 \\ &= f(y^k) - f(x) - g_k^\top (y^k - x) + (g_k - g_{\varepsilon k})^\top (x - y^k) + c\|x - y^k\|^2. \end{aligned}$$

Now since f is ε'-convex, estimate (1.14) is valid under the form

$$g_k^\top (x - y^k) \le f(x) - f(y^k) + \varepsilon'\|x - y^k\|.$$

We therefore get

$$f(y^k) - m_k(y^k,x) \le (\varepsilon' + \varepsilon)\|x - y^k\| + c\|x - y^k\|^2.$$

Subtracting a term $\frac{1}{2}(x - y^k)^\top Q(x - y^k)$ on both sides gives

$$f(y^k) - M_k(y^k,x) \le (\varepsilon' + \varepsilon + \nu_k)\|x - y^k\|,$$

where $\nu_k := c\|x - y^k\|^2 - \frac{1}{2}(x - y^k)^\top Q(x - y^k) \to 0$, and $M_k(y,x) = m_k(y,x) + \frac{1}{2}(y - x)^\top Q(y - x)$. Therefore

$$f(y^k) - M_k(y^k,x) \le (1+\zeta)(\varepsilon' + \varepsilon)\|x - y^k\| \tag{1.19}$$

for k large enough.

Now consider the second case $t_k(x) \le f(x)$. Here we get an even better estimate than (1.19), because $s_k = c\|x - y^k\|^2$, so that $f(y^k) - m_k(y^k,x) = c\|x - y^k\|^2 \le \varepsilon\|x - y^k\|$ for k large enough.

vii) To conclude, using (1.18) and (1.19) we expand the coefficient $\widetilde{\rho}_k$ as

$$\begin{aligned}\widetilde{\rho}_k &= \rho_k + \frac{f(y^k) - M_k(y^k,x)}{f(x) - \Phi_k(y^k,x)} \\ &\leq \rho_k + \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)\|x-y^k\|}{(1-\zeta)^2\eta\|x-y^k\|} = \rho_k + \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2\eta}.\end{aligned}$$

This shows

$$\eta < \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2(\widetilde{\gamma}-\gamma)}.$$

For suppose we had $\eta \geq \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2(\widetilde{\gamma}-\gamma)}$, then $\widetilde{\rho}_k \leq \rho_k + (\widetilde{\gamma}-\gamma) \leq \widetilde{\gamma}$ for all k, contradicting $\widetilde{\rho}_k > \widetilde{\gamma}$ for infinitely many k. As $0 < \zeta < 1$ was arbitrary, we have the estimate $\eta \leq \frac{\varepsilon'+\varepsilon}{\widetilde{\gamma}-\gamma}$. Since $\partial\phi(x,x) = \partial f(x) + \varepsilon B$ by Lemma 2, we deduce $0 \in \partial\phi(x,x) + \eta B \subset \partial f(x) + (\varepsilon+\eta)B$, and this is the result claimed in statement 1. □

Remark 6. Suppose we choose γ very small and $\widetilde{\gamma}$ close to 1, then $\alpha = 2 + \xi$ for some small ξ, so roughly $\alpha \approx 2$.

Lemma 4. *Suppose the inner loop turns forever and τ_k is frozen. Then $y^k \to x$ and $0 \in \partial_{[\varepsilon]} f(x)$.*

Proof. i) The control parameter is frozen from counter k_0 onwards, and we put $\tau := \tau_k$, $k \geq k_0$. This means that $\rho_k < \gamma$ and $\widetilde{\rho}_k < \widetilde{\gamma}$ for all $k \geq k_0$.

ii) We prove that the sequence of trial steps y^k is bounded. Notice that

$$g_k^{*\top}(x - y^k) \leq \phi_k(x,x) - \phi_k(y^k,x)$$

by the subgradient inequality at y^k and the definition of the aggregate subgradient. Now observe that $\phi_k(x,x) = f(x)$ and $\phi_k(y^k,x) \geq m_0(y^k,x)$. Therefore, using the definition of g_k^*, we have

$$(x-y^k)^\top(Q+\tau I)(x-y^k) \leq f(x) - m_0(y^k,x) = g(x)^\top(x-y^k) \leq \|g(x)\|\|x-y^k\|.$$

Since the τ-parameter is frozen and $Q + \tau I \succ 0$, the expression on the left is the square $\|x-y^k\|^2_{Q+\tau I}$ of the Euclidean norm derived from $Q+\tau I$. Since both norms are equivalent, we deduce after dividing by $\|x-y^k\|$ that $\|x-y^k\|_{Q+\tau I} \leq C\|g(x)\|$ for some constant $C > 0$ and all k. This proves the claim.

iii) Let us introduce the objective function of tangent program (1.7) for $k \geq k_0$:

$$\psi_k(\cdot,x) = \phi_k(\cdot,x) + \tfrac{1}{2}(\cdot - x)^\top(Q+\tau I)(\cdot - x).$$

Let $m_k^*(\cdot,x)$ be the aggregate plane, then $\phi_k(y^k,x) = m_k^*(y^k,x)$ by (1.9) and therefore also

$$\psi_k(y^k,x) = m_k^*(y^k,x) + \tfrac{1}{2}(y^k-x)^\top(Q+\tau I)(y^k-x).$$

We introduce the quadratic function $\psi_k^*(\cdot,x) = m_k^*(\cdot,x) + \frac{1}{2}(\cdot - x)^\top(Q+\tau I)(\cdot - x)$. Then

$$\psi_k(y^k,x) = \psi_k^*(y^k,x) \tag{1.20}$$

by what we have just seen. By construction of model $\phi_{k+1}(\cdot,x)$ we have $m_k^*(y,x) \leq \phi_{k+1}(y,x)$, so that

$$\psi_k^*(y,x) \leq \psi_{k+1}(y,x). \tag{1.21}$$

Notice that $\nabla \psi_k^*(y,x) = \nabla m_k^*(y,x) + (Q+\tau I)(y-x) = g_k^* + (Q+\tau I)(y-x)$, so that $\nabla \psi_k^*(y^k,x) = 0$ by (1.8). We therefore have the relation

$$\psi_k^*(y,x) = \psi_k^*(y^k,x) + \tfrac{1}{2}(y-y^k)^\top (Q+\tau I)(y-y^k), \tag{1.22}$$

which is obtained by Taylor expansion of $\psi_k^*(\cdot,x)$ at y^k. Recall that step 8 of the algorithm assures $Q+\tau I \succ 0$, so that the quadratic expression defines the Euclidean norm $\|\cdot\|_{Q+\tau I}$.

iv) From the previous point iii) we now have

$$\begin{aligned}
\psi_k(y^k,x) &\leq \psi_k^*(y^k,x) + \tfrac{1}{2}\|y^k - y^{k+1}\|^2_{Q+\tau I} && \text{(using (1.20))}\\
&= \psi_k^*(y^{k+1},x) && \text{(using (1.22))}\\
&\leq \psi_{k+1}(y^{k+1},x) && \text{(using (1.21))}\\
&\leq \psi_{k+1}(x,x) && (y^{k+1} \text{ minimizer of } \psi_{k+1})\\
&= \phi_{k+1}(x,x) = f(x).
\end{aligned} \tag{1.23}$$

We deduce that the sequence $\psi_k(y^k,x)$ is monotonically increasing and bounded above by $f(x)$. It therefore converges to some value $\psi^* \leq f(x)$.

Going back to (1.23) with this information shows that the term $\frac{1}{2}\|y^k - y^{k+1}\|^2_{Q+\tau I}$ is squeezed in between two convergent terms with the same limit, ψ^*, which implies $\frac{1}{2}\|y^k - y^{k+1}\|^2_{Q+\tau I} \to 0$. Consequently, $\|y^k - x\|^2_{Q+\tau I} - \|y^{k+1} - x\|^2_{Q+\tau I}$ also tends to 0, because the sequence of trial steps y^k is bounded by part ii).

Recalling $\phi_k(y,x) = \psi_k(y,x) - \frac{1}{2}\|y-x\|^2_{Q+\tau I}$, we deduce, using both convergence results, that

$$\begin{aligned}
&\phi_{k+1}(y^{k+1},x) - \phi_k(y^k,x) = \\
&\quad \psi_{k+1}(y^{k+1},x) - \psi_k(y^k,x) - \tfrac{1}{2}\|y^{k+1} - x\|^2_{Q+\tau I} + \tfrac{1}{2}\|y^k - x\|^2_{Q+\tau I} \to 0.
\end{aligned} \tag{1.24}$$

v) We want to show that $\phi_k(y^k,x) - \phi_{k+1}(y^k,x) \to 0$, and then of course also $\Phi_k(y^k,x) - \Phi_{k+1}(y^k,x) \to 0$.

Recall that by construction the cutting plane $m_k(\cdot,x)$ is an affine support function of $\phi_{k+1}(\cdot,x)$ at y^k. By the subgradient inequality this implies

$$g_k^\top (y-y^k) \leq \phi_{k+1}(y,x) - \phi_{k+1}(y^k,x) \tag{1.25}$$

for all y. Therefore

$$\begin{aligned}
0 &\le \phi_{k+1}(y^k,x) - \phi_k(y^k,x) && \text{(using aggregation)}\\
&= \phi_{k+1}(y^k,x) + g_k^\top(y^{k+1}-y^k) - \phi_k(y^k,x) - g_k^\top(y^{k+1}-y^k)\\
&\le \phi_{k+1}(y^{k+1},x) - \phi_k(y^k,x) + \|g_k\|\|y^{k+1}-y^k\| && \text{(using (1.25))}
\end{aligned}$$

and this term converges to 0, because of (1.24), because the g_k are bounded, and because $y^k - y^{k+1} \to 0$ according to part iv) above. Boundedness of the g_k follows from boundedness of the trial steps y^k shown in part ii). Indeed, $g_k \in \partial f(y^k) + \varepsilon B$, and the subdifferential of f is uniformly bounded on the bounded set $\{y^k : k \in \mathbb{N}\}$. We deduce that $\phi_{k+1}(y^k,x) - \phi_k(y^k,x) \to 0$. Obviously, that also gives $\Phi_{k+1}(y^k,x) - \Phi_k(y^k,x) \to 0$.

vi) We now proceed to prove $\Phi_k(y^k,x) \to f(x)$, and then also $\Phi_{k+1}(y^k,x) \to f(x)$. Assume this is not the case, then $\limsup_{k\to\infty} f(x) - \Phi_k(y^k,x) =: \eta > 0$. Choose $\delta > 0$ such that $\delta < (1-\widetilde{\gamma})\eta$. It follows from v) above that there exists $k_1 \ge k_0$ such that

$$\Phi_{k+1}(y^k,x) - \delta \le \Phi_k(y^k,x)$$

for all $k \ge k_1$. Using $\widetilde{\rho}_k \le \widetilde{\gamma}$ for all $k \ge k_0$ then gives

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

Passing to the limit implies $-\widetilde{\gamma}\eta \le -\eta + \delta$, contradicting the choice of δ. This proves $\eta = 0$.

vii) Having shown $\Phi_k(y^k,x) \to f(x)$, and therefore also $\Phi_{k+1}(y^k,x) \to f(x)$, we now argue that $y^k \to x$. This follows from the definition of ψ_k, because

$$\Phi_k(y^k,x) \le \psi_k(y^k,x) = \Phi_k(y^k,x) + \tfrac{\tau}{2}\|y^k - x\|^2 \le \psi^* \le f(x).$$

Since $\Phi_k(y^k,x) \to f(x)$ by part vi), we deduce $\frac{\tau}{2}\|y^k - x\|^2 \to 0$ using a sandwich argument, which also proves *en passant* that $\psi^* = f(x)$ and $\phi_k(y^k,x) \to f(x)$.

To finish the proof, let us now show $0 \in \partial_{[\varepsilon]} f(x)$. Remember that by the necessary optimality condition for (1.7) we have $(Q+\tau I)(x-y^k) \in \partial\phi_k(y^k,x)$. By the subgradient inequality,

$$\begin{aligned}
(x-y^k)^\top(Q+\tau I)(y-y^k) &\le \phi_k(y,x) - \phi_k(y^k,x)\\
&\le \phi(y,x) - \phi_k(y^k,x),
\end{aligned}$$

where ϕ is the upper envelope (1.13) of all cutting planes drawn at $z \in B(0,M)$, $g \in \partial_{[\varepsilon]} f(z)$, which we choose large enough to contain the bounded set $\{x\} \cup \{y^k : k \in \mathbb{N}\}$, a fact which assures $\phi_k(\cdot,x) \le \phi(\cdot,x)$ for all k (see Lemma 2). Passing to the limit, observing $\|x-y^k\|^2_{Q+\tau I} \to 0$ and $\phi_k(y^k,x) \to f(x) = \phi(x,x)$, we obtain:

$$0 \le \phi(y,x) - \phi(x,x)$$

for all y. This proves $0 \in \partial\phi(x,x)$. Since $\partial\phi(x,x) \subset \partial_{[\varepsilon]} f(x)$ by Lemma 2, we have shown $0 \in \partial_{[\varepsilon]} f(x)$. □

1.6 Convergence of the outer loop

In this section we prove subsequence convergence of our algorithm for the case where function values are exact and subgradients are in $\partial_{[\varepsilon]} f(y^k)$. We write $Q_j = Q(x^j)$ for the matrix of the second order model, which depends on the serious iterates x^j.

Theorem 1. *Let x^1 be such that $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded. Suppose f is ε'-convex on Ω and that subgradients are drawn from $\partial_{[\varepsilon]} f(y)$, whereas function values are exact. Then every accumulation point $\bar{x}$ of the sequence of serious iterates x^j satisfies $0 \in \partial_{[\widetilde{\varepsilon}]} f(\bar{x})$, where $\widetilde{\varepsilon} = \varepsilon + (\varepsilon' + \varepsilon)/(\gamma - \widetilde{\gamma})$.*

Proof. i) From the analysis in section 1.5 we know that if we apply the stopping test in step 2 with $\widetilde{\varepsilon} = \varepsilon + (\varepsilon' + \varepsilon)/(\gamma - \widetilde{\gamma})$, then the inner loop 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 $0 \in \partial_{[\widetilde{\varepsilon}]} f(x)$. Let us exclude this case, and let x^j denote the infinite sequence of serious iterates. 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, $y^{k_j} = x^{j+1}$ passes the acceptance test in step 5 of the algorithm and becomes a serious iterate, while the y^k with $k < k_j$ are null steps. That means

$$f(x^j) - f(x^{j+1}) \geq \gamma \left(f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) \right). \tag{1.26}$$

Now recall that $(Q_j + \tau_{k_j} I)(x^j - x^{j+1}) \in \partial\phi_{k_j}(x^{j+1}, x^j)$ by optimality of the tangent program (1.7). The subgradient inequality for $\phi_{k_j}(\cdot, x^j)$ at x^{j+1} therefore gives

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

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 have

$$\tfrac{1}{2}\|x^{j+1} - x^j\|^2_{Q_j + \tau_{k_j} I} \leq f(x^j) - \Phi_{k_j}(x^{j+1}, x^j) \leq \gamma^{-1} \left(f(x^j) - f(x^{j+1}) \right), \tag{1.27}$$

using (1.26). Summing (1.27) from $j = 1$ to $j = J$ gives

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

Here the right hand side is bounded above because our method is of descent type in the serious steps and Ω is bounded. Consequently the series on the left is summable, and therefore $\|x^{j+1} - x^j\|^2_{Q_j + \tau_{k_j} I} \to 0$ as $j \to \infty$. Let $\bar{x}$ be an accumulation point of the sequence x^j. We have to prove $0 \in \partial_{[\tilde{\varepsilon}]} f(\bar{x})$. We select a subsequence $j \in J$ such that $x^j \to \bar{x}$, $j \in J$. There are now two cases. The first is discussed in part ii), the second is more complicated and will be discussed in iii) – ix).

ii) Suppose there exists an infinite subsequence J' of J such that $g_j := (Q_j + \tau_{k_j} I)\left(x^j - x^{j+1}\right)$ converges to 0, $j \in J'$. We will show that in this case $0 \in \partial_{[\tilde{\varepsilon}]} f(\bar{x})$.

In order to prove this claim, notice first that since $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ 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. We can therefore use the convex upper envelope function ϕ of (1.13), where $B(0,M)$ contains Ω and also *all* the trial points y^k visited during all inner loops j.

Indeed, the set of x^j being bounded, so are the $\|g(x^j)\|$, where $g(x^j) \in \partial_{[\varepsilon]} f(x^j)$ is the exactness subgradient of the jth inner loop. From (1.16) we know that $\|x^j - y^k\|_{Q_j + \tau_k I} \leq \|g(x^j)\|$ for every j and every trial step y^k arising in the j^{th} inner loop at some instant k. From the management of the τ-parameter in the outer loop (1.12) we know that $Q_j + \tau_k I \succ \zeta I$ for some $\zeta > 0$, so $\|x^j - y^k\| \leq \zeta^{-1}\|g(x^j)\| \leq C < \infty$, meaning the y^k are bounded. During the following the properties of ϕ obtained in Lemma 2 will be applied at every $x = x^j$.

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:

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

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

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

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

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

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

$$\tfrac{1}{2}\|x^j - x^{j+1}\|^2_{Q_j+\tau_{k_j}I} + g_j^\top h' \le \phi(x^j+h',x^j) - f(x^j) + \gamma^{-1}\left(f(x^j)-f(x^{j+1})\right). \tag{1.28}$$

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

$$0 \le \phi(\bar{x}+h',\bar{x}) - \phi(\bar{x},\bar{x}). \tag{1.29}$$

In (1.28) the rightmost term $f(x^j) - f(x^{j+1}) \to 0$ converges by monotonicity, convergence of the leftmost term was shown in part i), and $g_j \to 0$ is the working hypothesis. Now the test vector h' in (1.29) is arbitrary, which shows $0 \in \partial\phi(\bar{x},\bar{x})$. By Lemma 2 we have $0 \in \partial_{[\varepsilon]} f(\bar{x}) \subset \partial_{[\tilde{\varepsilon}]} f(\bar{x})$.

iii) The second more complicated case is when $\|g_j\| = \|(Q_j + \tau_{k_j}I)(x^j - x^{j+1})\| \ge \mu > 0$ for some $\mu > 0$ and every $j \in J$. The remainder of this proof will be entirely dedicated to this.

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 and boundedness of the serious steps, there exists then an infinite subsequence $j \in J'$ of J such that Q_j, τ_{k_j} and $x^j - x^{j+1}$ converge respectively to $\bar{Q}$, $\bar{\tau}$ and $\delta\bar{x}$ as $j \in J'$. This implies that the corresponding subsequence of g_j converges to $(\bar{Q}+\bar{\tau}I)\delta\bar{x}$, where $\|(\bar{Q}+\bar{\tau}I)\delta\bar{x}\| \ge \mu > 0$. Similarly, $(x^j - x^{j+1})^\top(Q_j + \tau_{k_j}I)(x^j - x^{j+1}) \to \delta\bar{x}^\top(\bar{Q}+\bar{\tau}I)\delta\bar{x}$. By part i) of the proof we have $g_j^\top(x^{j+1}-x^j) = \|x^{j+1}-x^j\|^2_{Q_j+\tau_{k_j}I} \to 0$, which means $\delta\bar{x}^\top(\bar{Q}+\bar{\tau}I)\delta\bar{x} = 0$. Since $\bar{Q}+\bar{\tau}I$ is symmetric and $\bar{Q}+\bar{\tau}I \succeq 0$, we deduce $(\bar{Q}+\bar{\tau}I)\delta\bar{x} = 0$, contradicting $\|(\bar{Q}+\bar{\tau}I)\delta\bar{x}\| \ge \mu > 0$. 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

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

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

v) 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 $\tau_{k_j} \to \infty$, $\|g_j\| \ge \mu > 0$, 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 $\nu_j \ge 1$. Then

$$\tau_{k_j} = \tau_{k_j-1} = \cdots = \tau_{k_j-\nu_j+1} = 2\tau_{k_j-\nu_j}. \tag{1.30}$$

According to step 6 of the algorithm, the increase at counter $k_j - \nu_j$ is due to the fact that

$$\rho_{k_j-\nu_j} < \gamma \quad \text{and} \quad \widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma}. \tag{1.31}$$

This case is labelled *too bad* in step 6 of the algorithm.

vi) Condition (1.31) means that there are infinitely many $j \in J^+$ satisfying

$$\rho_{k_j-\nu_j} = \frac{f(x^j) - f(y^{k_j-\nu_j})}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j)} < \gamma$$

and

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

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

$$\begin{aligned}(x^j - y^{k_j-\nu_j})^\top (Q_j + \tau_{k_j-\nu_j} I)(x^j - y^{k_j-\nu_j}) &\leq \phi_{k_j-\nu_j}(x^j, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \\ &\leq f(x^j) - m_0(y^{k_j-\nu_j}, x^j) \\ &= g(x^j)^\top (x^j - y^{k_j-\nu_j}) \\ &\leq \|g(x^j)\| \|x^j - y^{k_j-\nu_j}\|,\end{aligned} \tag{1.32}$$

where $m_0(\cdot, x^j) = f(x^j) + g(x^j)^\top(\cdot - x^j)$ is the exactness plane at x^j. As $\tau_{k_j} \to \infty$, we have $\tau_{k_j-\nu_j} = \frac{1}{2}\tau_{k_j} \to \infty$, too, so the left hand side of (1.32) behaves asymptotically like constant times $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j}\|^2$. 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 $\|x^j - y^{k_j-\nu_j}\|$. This shows boundedness of $\tau_{k_j-\nu_j}\|x^j - y^{k_j-\nu_j}\|$, and therefore $x^j - y^{k_j-\nu_j} \to 0$, because $\tau_{k_j-\nu_j} \to \infty$.

vii) Recall that $x^j \to \bar{x}$, $j \in J$. By vi) we know that $y^{k_j-\nu_j} \to \bar{x}$, $j \in J$. Passing to a subsequence J' of J, we may assume $\widetilde{g}_j \to \widetilde{g}$ for some $\widetilde{g}$. We show $\widetilde{g} \in \partial\phi(\bar{x}, \bar{x})$.

For a test vector h and $j \in J'$:

$$\begin{aligned}\widetilde{g}_j^\top h &\leq \phi_{k_j-\nu_j}(y^{k_j-\nu_j} + h, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \\ &\leq \phi(y^{k_j-\nu_j} + h, x^j) - \phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j).\end{aligned} \tag{1.33}$$

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

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

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

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

Combining this and estimate (1.33) gives

$$\widetilde{g}_j^\top h \leq \phi(y^{k_j-\nu_j}+h,x^j)-f(x^j)+\widetilde{\gamma}^{-1}\left(f(x^j)-M_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)\right) +\tfrac{1}{2}(y^{k_j-\nu_j}-x^j)^\top Q_j(y^{k_j-\nu_j}-x^j). \quad (1.34)$$

As we have seen $y^{k_j-\nu_j}-x^j\to 0$, hence the rightmost term in (1.34) converges to 0 by boundedness of Q_j. Moreover, we claim that $\lim f(x^j)-M_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)=0$, so the term $\widetilde{\gamma}^{-1}(\dots)$ on the right hand side of (1.34) converges to 0. Indeed, to see this claim, notice first that it suffices to show $f(x^j)-m_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)\to 0$, because the second order term converges to 0. Since $m_{k_j-\nu_j}(\cdot,x^j)$ is a cutting plane at x^j, we have $m_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)\leq f(y^{k_j-\nu_j})$ by definition of the down-shift. So it suffices to show $\liminf m_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)\geq f(\bar{x})$. Now this follows from the definition of the down-shift s_j at $y^{k_j-\nu_j}$ with regard to x^j. Recall that for the tangent $t_{k_j-\nu_j}(\cdot)$ at $y^{k_j-\nu_j}$, approximate subgradient $\widetilde{g}_j$, and serious iterate x^j, we have

$$s_j=[t_{k_j-\nu_j}(x^j)-f(x^j)]_+ + c\|y^{k_j-\nu_j}-x^j\|^2.$$

We can clearly concentrate on proving $t_{k_j-\nu_j}(x^j)-f(x^j)\to 0$. Now $t_{k_j-\nu_j}(x^j)-f(x^j)=f(y^{k_j-\nu_j})-f(x^j)+\widetilde{g}_j^\top(x^j-y^{k_j-\nu_j})$, and since $y^{k_j-\nu_j}\to\bar{x}$, $x^j\to\bar{x}$ and the $\widetilde{g}_j$ are bounded, our claim follows.

Going back to (1.34) with the information $\widetilde{g}_j^\top h\to\widetilde{g}^\top h$, it remains to prove $\limsup\phi(y^{k_j-\nu_j}+h,x^j)\leq\phi(\bar{x}+h,\bar{x})$. Indeed, once this is proved, passing to the limit in (1.34) shows $\widetilde{g}^\top h\leq\phi(\bar{x}+h,\bar{x})-f(\bar{x})=\phi(\bar{x}+h,\bar{x})-\phi(\bar{x},\bar{x})$. This proves $\widetilde{g}\in\partial\phi(\bar{x},\bar{x})$, and then $\widetilde{g}\in\partial_{[\varepsilon]}f(\bar{x})$ by Lemma 2.

What remains to be shown is obviously joint upper semi-continuity of ϕ at $(\bar{x}+h,\bar{x})$, and this follows from Lemma 2, hence our claim $\widetilde{g}\in\partial_{[\varepsilon]}f(\bar{x})$ is proved.

viii) Let $\eta:=\operatorname{dist}(0,\partial\phi(\bar{x},\bar{x}))$. Then $\|\widetilde{g}\|\geq\eta$ by vii) above. Let us fix $0<\zeta<1$, then, as $\widetilde{g}_j\to\widetilde{g}$, we have $\|\widetilde{g}_j\|\geq(1-\zeta)\eta$ for $j\in J'$ large enough.

Now, assuming first $[\dots]_+>0$ in the downshift, we have

$$\begin{aligned} m_{k_j-\nu_j}(\cdot,x^j) &= f(y^{k_j-\nu_j})+\widetilde{g}_j^\top(\cdot-y^{k_j-\nu_j})-s_j \\ &= f(y^{k_j-\nu_j})+\widetilde{g}_j^\top(\cdot-y^{k_j-\nu_j})-c\|y^{k_j-\nu_j}-x^j\|^2-t_{k_j-\nu_j}(x^j)+f(x^j) \\ &= f(x^j)+\widetilde{g}_j^\top(\cdot-x^j)-c\|y^{k_j-\nu_j}-x^j\|^2, \end{aligned}$$

for $\widetilde{g}_j\in\partial_{[\varepsilon]}f(y^{k_j-\nu_j})$ as above. Pick $g_j\in\partial f(y^{k_j-\nu_j})$ such that $\|g_j-\widetilde{g}_j\|\leq\varepsilon$. Then

$$\begin{aligned} f(y^{k_j-\nu_j}) - m_{k_j-\nu_j}(y^{k_j-\nu_j},x^j) &= f(y^{k_j-\nu_j}) - f(x^j) - \widetilde{g}_j^\top(y^{k_j-\nu_j}-x^j) \\ &\qquad + c\|y^{k_j-\nu_j}-x^j\|^2 \\ &= f(y^{k_j-\nu_j}) - f(x^j) - g_j^\top(y^{k_j-\nu_j}-x^j) \\ &\qquad + (\widetilde{g}_j - g_j)(y^{k_j-\nu_j}-x^j) \\ &\qquad + c\|y^{k_j-\nu_j}-x^j\|^2. \end{aligned}$$

Since f is ε'-convex, we have $g_j^\top(y^{k_j-\nu_j}-x^j) \leq f(x^j) - f(y^{k_j-\nu_j}) + \varepsilon'\|y^{k_j-\nu_j}-x^j\|$. Substituting this we get

$$f(y^{k_j-\nu_j}) - m_{k_j-\nu_j}(y^{k_j-\nu_j},x^j) \leq (\varepsilon'+\varepsilon)\|y^{k_j-\nu_j}-x^j\| + c\|y^{k_j-\nu_j}-x^j\|^2. \tag{1.35}$$

In the case $[\dots]_+ = 0$ an even better estimate is obtained, so that (1.35) covers both cases. Subtracting a term $\frac{1}{2}(y^{k_j-\nu_j}-x^j)^\top Q_j(y^{k_j-\nu_j}-x^j)$ on both sides of (1.35) and using $y^{k_j-\nu_j}-x^j \to 0$, we get

$$f(y^{k_j-\nu_j}) - M_{k_j-\nu_j}(y^{k_j-\nu_j},x^j) \leq (\varepsilon'+\varepsilon+\nu_j)\|y^{k_j-\nu_j}-x^j\|,$$

where $\nu_j \to 0$. In consequence

$$f(y^{k_j-\nu_j}) - M_{k_j-\nu_j}(y^{k_j-\nu_j},x^j) \leq (1+\zeta)(\varepsilon'+\varepsilon)\|y^{k_j-\nu_j}-x^j\| \tag{1.36}$$

for j large enough. Recall that $\widetilde{g}_j = (Q_j + \frac{1}{2}\tau_{k_j}I)(x^j - y^{k_j-\nu_j}) \in \partial\phi_{k_j-\nu_j}(y^{k_j-\nu_j},x^j)$ by (1.8) and (1.30). Hence by the subgradient inequality

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

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

$$\tfrac{1}{2}(x^j-y^{k_j-\nu_j})^\top Q_j(x^j-y^{k_j-\nu_j}) + \tfrac{1}{2}\tau_{k_j}\|x^j-y^{k_j-\nu_j}\|^2 \leq f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j},x^j). \tag{1.37}$$

As $\tau_{k_j} \to \infty$, we have

$$(1-\zeta)\tfrac{1}{2}\tau_{k_j}\|x^j-y^{k_j-\nu_j}\| \leq \|\widetilde{g}_j\| \leq (1+\zeta)\tfrac{1}{2}\tau_{k_j}\|x^j-y^{k_j-\nu_j}\| \tag{1.38}$$

and

$$\tfrac{1}{2}(x^j-y^{k_j-\nu_j})^\top Q_j(x^j-y^{k_j-\nu_j}) + \tfrac{1}{2}\tau_{k_j}\|x^j-y^{k_j-\nu_j}\|^2 \geq (1-\zeta)\tfrac{1}{2}\tau_{k_j}\|x^j-y^{k_j-\nu_j}\|^2 \tag{1.39}$$

both for j large enough. Therefore, plugging (1.38) and (1.39) into (1.37) gives

$$f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j},x^j) \geq \tfrac{1-\zeta}{1+\zeta}\|\widetilde{g}_j\|\|x^j-y^{k_j-\nu_j}\|$$

for j large enough. Since $\|\widetilde{g}_j\| \geq (1-\zeta)\eta$ for j large enough, we deduce

$$f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \geq \tfrac{(1-\zeta)^2}{1+\zeta}\eta\,\|x^j - y^{k_j-\nu_j}\|. \tag{1.40}$$

ix) Combining (1.36) and (1.40) gives the estimate

$$\begin{aligned}\widetilde{\rho}_{k_j-\nu_j} &= \rho_{k_j-\nu_j} + \frac{f(y^{k_j-\nu_j}) - M_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j)}{f(x^j) - \Phi_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j)} \\ &\leq \rho_{k_j-\nu_j} + \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)\|y^{k_j-\nu_j} - x^j\|}{(1-\zeta)^2\eta\|y^{k_j-\nu_j} - x^j\|}.\end{aligned} \tag{1.41}$$

This proves

$$\eta \leq \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2(\widetilde{\gamma}-\gamma)}.$$

For suppose we had $\eta > \frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2(\widetilde{\gamma}-\gamma)}$, then $\frac{(1+\zeta)^2(\varepsilon'+\varepsilon)}{(1-\zeta)^2\eta} < \widetilde{\gamma} - \gamma$, which gave $\widetilde{\rho}_{k_j-\nu_j} \leq \rho_{k_j-\nu_j} + \widetilde{\gamma} - \gamma < \widetilde{\gamma}$ for all j, contradicting $\widetilde{\rho}_{k_j-\nu_j} \geq \widetilde{\gamma}$ for infinitely many $j \in J$.

Since ζ in the above discussion was arbitrary, we have shown $\eta \leq \frac{\varepsilon'+\varepsilon}{\widetilde{\gamma}-\gamma}$. Recall that $\eta = \operatorname{dist}\big(0, \partial_{[\varepsilon]} f(\bar{x})\big)$. We therefore have shown $0 \in \partial_{[\widetilde{\varepsilon}]} f(\bar{x})$, where $\widetilde{\varepsilon} = \varepsilon + \eta$. This is what is claimed. □

Corollary 1. *Suppose $\Omega = \{x \in \mathbb{R}^n : f(x) \leq f(x^1)\}$ is bounded and f is lower C^1. Let approximate subgradients be drawn from $\partial_{[\varepsilon]} f(y)$, whereas function values are exact. Then every accumulation point $\bar{x}$ of the sequence of serious iterates x^j satisfies $0 \in \partial_{[\alpha\varepsilon]} f(\bar{x})$, where $\alpha = 1 + (\widetilde{\gamma} - \gamma)^{-1}$.* □

Remark 7. At first glance one might consider the class of lower C^1 functions used in Corollary 1 as too restrictive to offer sufficient scope. This misapprehension might be aggravated, or even induced, by the fact that lower C^1 functions are *approximately convex* [16, 36], an unfortunate nomenclature which erroneously suggests something close to a convex function. We therefore stress that lower C^1 is a large class which includes all examples we have so far encountered in practice. Indeed, applications are as a rule even lower C^2, or *amenable* in the sense of Rockafellar [47], a much smaller class, yet widely accepted as of covering all applications of interest.

Recent approaches to non-convex non-smooth optimization like [23, 35, 49] all work with composite (and therefore lower C^2) functions. This is in contrast with our own approach [21,22,43,44], which works for lower C^1 and is currently the only one I am aware of that *has* the technical machinery to go beyond lower C^2. On second glance one will therefore argue that it is rather the class of lower C^2-functions which does not offer sufficient scope to justify the development of a new theory, because the chapter on nonsmooth composite convex functions $f = g \circ F$ in [48] covers this class nicely and leaves little space for new contributions, and because one *can* do things for lower C^1.

1.7 Extension to inexact values

In this section we discuss what happens when we not only have inexact subgradients, but also inexact function values. In the previous sections we assumed that for every approximate subgradient g_a of f at x, there exists an exact subgradient $g \in \partial f(x)$ such that $\|g_a - g\| \leq \varepsilon$. Similarly, we will assume that approximate function values $f_a(x)$ satisfy $|f_a(x) - f(x)| \leq \bar{\varepsilon}$ for a fixed error tolerance $\bar{\varepsilon}$. We do not assume any link between ε and $\bar{\varepsilon}$.

Let us notice the following fundamental difference between the convex and the non-convex case, where it is often reasonable to assume $f_a \leq f$, see e.g. [32, 33]. Suppose f is convex, x is the current iterate, and an approximate value $f(x) - \bar{\varepsilon} \leq f_a(x) \leq f(x)$ is known. Suppose y^k is a null step, so that we draw an approximate tangent plane $t_k(\cdot) = f_a(y^k) + g_k^\top(\cdot - y^k)$ at y^k with respect to $g_k \in \partial_{[\varepsilon]} f(y^k)$. If we follow [32, 33], then $t_k(\cdot)$, while not a support plane, is still an affine minorant of f. It may then happen that $t_k(x) = f_a(y^k) + g_k^\top(x - y^k) > f_a(x)$, because $f_a(x), f_a(y^k)$ are approximations only. Now the approximate cutting plane gives us viable information as to the fact that the true value $f(x)$ satisfies $f(x) \geq t_k(x) > f_a(x)$. We shall say that *we can trust the value* $t_k(x) > f_a(x)$.

What should we do if we find a value $t_k(x)$ in which we can trust, and which reveals our estimate $f_a(x)$ as too low? Should we correct $f_a(x)$ and replace it by the better estimate now available? If we do this we create trouble. Namely, we have previously rejected trial steps y^k during the inner loop at x based on the incorrect information $f_a(x)$. Some of these steps might have been acceptable, had we used $t_k(x)$ instead. But on the other hand, x was accepted as serious step in the inner loop at x^- because $f_a(x)$ was sufficiently below $f_a(x^-)$. If we correct the approximate value at x, then acceptance of x may become unsound as well. For short, correcting values as soon as better estimates arrive is not a good idea, because we might be forced to go repeatedly back all the way through the history of our algorithm.

In order to avoid this backtracking, Kiwiel [32] proposes the following original idea. If $f_a(x)$, being too low, still allows progress in the sense that x^+ with $f_a(x^+) < f_a(x)$ can be found, then why waste time and correct the value $f_a(x)$? After all, there is still progress! On the other hand, if the under-estimation $f_a(x)$ is so severe that the algorithm will stop, then we should be sure that no further decrease within the error tolerances $\bar{\varepsilon}, \varepsilon$ is possible. Namely, if this is the case, then we can stop in all conscience. To check this, Kiwiel progressively relaxes proximity control in the inner loop, until it becomes clear that the model of all possible approximate cutting planes itself does not allow to descend below $f_a(x)$, and therefore, does not allow to descend more than $\bar{\varepsilon}$ below $f(x)$.

The situation outlined is heavily based on convexity and does not appear to carry over to non-convex problems. The principal difficulty is that without convexity we cannot trust values $t_{y,g}(x) > f_a(x)$ even in the case of *exact* tangent planes, $g \in \partial f(y)$. We know that tangents have to be downshifted, and without the exact knowledge of $f(x)$ the only available reference value to organize the downshift is $f_a(x)$. Naturally, as soon as we downshift with reference to $f_a(x)$, cutting planes $m_{y,g}(\cdot, x)$ satisfying

$m_{y,g}(x,x) > f_a(x)$ can no longer occur. This removes one of the difficulties. However, it creates, as we shall see, a new one.

In order to proceed with inexact function values, we will need the following property of the cutting plane $m_k(\cdot,x) := t_k(\cdot) - s_k$ at null step y^k and approximate subgradient $g_k \in \partial_{[\varepsilon]} f(y^k)$. We need to find $\widetilde{\varepsilon} > 0$ such that $f_a(y^k) \leq m_k(y^k,x) + \widetilde{\varepsilon}\|x-y^k\|$. More explicitly, this requires

$$f_a(y^k) \leq f_a(x) + g_k^\top (y^k - x) + \widetilde{\varepsilon}\|x-y^k\|.$$

If f is ε'-convex, then

$$\begin{aligned} f(y^k) &\leq f(x) + g^\top (y^k - x) + \varepsilon'\|x-y^k\| \\ &\leq f(x) + g_k^\top (y^k - x) + (\varepsilon' + \varepsilon)\|x-y^k\| \end{aligned}$$

for $g \in \partial f(y^k)$ and $\|g - g_k\| \leq \varepsilon$. That means

$$f(y^k) - (f(x) - f_a(x)) \leq f_a(x) + g_k^\top (y^k - x) + (\varepsilon + \varepsilon')\|x-y^k\|.$$

So what we need in addition is something like

$$f_a(y^k) \leq f(y^k) - (f(x) - f_a(x)) + \varepsilon''\|x-y^k\|,$$

because then we get the desired relation with $\widetilde{\varepsilon} = \varepsilon + \varepsilon' + \varepsilon''$. The condition can still be slightly relaxed to make it more useful in practice. The axiom we need is that there exist $\delta_k \to 0^+$ such that

$$f(x) - f_a(x) \leq f(y^k) - f_a(y^k) + (\varepsilon'' + \delta_k)\|x-y^k\| \tag{1.42}$$

for every $k \in \mathbb{N}$. Put differently, as $y^k \to x$, the error me make at y^k by underestimating $f(y^k)$ by $f_a(y^k)$ is larger than the corresponding underestimation error at x, up to a term proportional to $\|x-y^k\|$. The case of exact values $f = f_a$ corresponds to $\varepsilon'' = 0, \delta_k = 0$.

Remark 8. As f is continuous at x, condition (1.42) implies upper semi-continuity of f_a at serious iterates, i.e., $\limsup f_a(y^k) \leq f_a(x)$.

We are now ready to modify our algorithm and then run through the proofs of Lemmas 3, 4 and Theorem 1 and see what changes need to be made to account for the new situation. As far as the algorithm is concerned, the changes are easy. We replace $f(y^k)$ and $f(x)$ by $f_a(y^k)$ and $f_a(x)$. The rest of the procedure is the same.

We consider the same convex envelope function $\phi(\cdot,x)$ defined in (1.13). We have the following

Lemma 5. *The upper envelope model satisfies $\phi(x,x) = f_a(x)$, $\phi_k \leq \phi$. ϕ is jointly upper $2\bar{\varepsilon}$-semicontinuous, and $\partial\phi(x,x) \subset \partial_{[\varepsilon]} f(x) \subset \partial_{2\bar{\varepsilon}}\phi(x,x)$, where $\partial_{2\bar{\varepsilon}}\phi(x,x)$ is the $2\bar{\varepsilon}$-subdifferential of $\phi(\cdot,x)$ at x in the usual convex sense.*

Proof. 1) Any cutting plane $m_{z,g}(\cdot,x)$ satisfies $m_{z,g}(x,x) \leq f_a(x) - c\|x-z\|^2$. This shows $\phi(x,x) \leq f_a(x)$, and if we take $z = x$, we get equality $\phi(x,x) = f_a(x)$.

2) We prove $\partial_{[\varepsilon]} f(x) \subset \partial_{2\bar{\varepsilon}} \phi(x,x)$. Let $g \in \partial f(x)$ be a limiting subgradient, and choose $y^k \to x$, where f is differentiable at y^k with $g_k = \nabla f(y^k) \in \partial f(y^k)$ such that $g_k \to g$. Let g_a be an approximate subgradient such that $\|g - g_a\| \leq \varepsilon$. We have to prove $g_a \in \partial_{2\bar{\varepsilon}} \phi(x,x)$. Putting $g_{a,k} := g_k + g_a - g \in \partial_{[\varepsilon]} f(y^k)$ we have $g_{a,k} \to g_a$. Let $m_k(\cdot,x)$ be the cutting plane drawn at y^k with approximate subgradient $g_{a,k}$. That is, $m_k(\cdot,x) = m_{y^k,g_{a,k}}(\cdot,x)$. Then

$$m_k(y,x) = f_a(y^k) + g_{a,k}^\top (y - y^k) - s_k,$$

where $s_k = [f_a(x) - t_k(x)]_+ + c\|x - y^k\|^2$ is the down-shift, and where $t_k(\cdot)$ is the approximate tangent at y^k with respect to $g_{a,k}$. There are two cases, $s_k = c\|x - y^k\|^2$, and $s_k = f_a(x) + t_k(x) + c\|x - y^k\|^2$, according to whether $[\dots]_+ = 0$ or $[\dots]_+ > 0$. Let us start with the case $t_k(x) > f_a(x)$. Then

$$s_k = f_a(y^k) + g_{a,k}^\top (x - y^k) + c\|x - y^k\|^2$$

and

$$m_k(y,x) = f_a(y^k) + g_{a,k}^\top (y - y^k) - f_a(y^k) - g_{a,k}^\top (x - y^k) + f_a(x) - c\|x - y^k\|^2.$$

Therefore

$$\phi(y,x) - \phi(x,x) \geq m_k(y^k,x) - f_a(x) = g_{a,k}^\top (y - x) - c\|x - y^k\|^2.$$

Passing to the limit $k \to \infty$ proves $g_a \in \partial \phi(x,x)$, so in this case a stronger statement holds.

Let us next discuss the case where $t_k(x) \leq f_a(x)$, so that $s_k = c\|x - y^k\|^2$. Then

$$m_k(y,x) = f_a(y^k) + g_{a,k}^\top (y - y^k) - c\|x - y^k\|^2.$$

Therefore

$$\begin{aligned}
\phi(y,x) - \phi(x,x) &\geq m_k(y^k,x) - f_a(x) \\
&= f_a(y^k) - f_a(x) + g_{a,k}^\top (y - y^k) - c\|x - y^k\|^2 \\
&= f_a(y^k) - f_a(x) + g_{a,k}^\top (x - y^k) - c\|x - y^k\|^2 + g_{a,k}^\top (y - x).
\end{aligned}$$

Put $\zeta_k := g_{a,k}^\top (x - y^k) - c\|x - y^k\|^2 + (g_{a,k} - g_a)^\top (y - x)$, then

$$\phi(y,x) - \phi(x,x) \geq f_a(y^k) - f_a(x) + \zeta_k + g_a^\top (y - x).$$

Notice that $\lim \zeta_k = 0$, because $g_{a,k} \to g_a$ and $y^k \to x$. Let $F_a(x) := \liminf_{k \to \infty} f_a(y^k)$, then we obtain

$$\phi(y,x) - \phi(x,x) \geq F_a(x) - f_a(x) + g_a^\top (y - x).$$

Putting $\varepsilon(x) := [f_a(x) - F_a(x)]_+$, we therefore have shown

$$\phi(y,x) - \phi(x,x) \geq -\varepsilon(x) + g_a^\top(y-x),$$

which means $g_a \in \partial_{\varepsilon(x)}\phi(x,x)$. Since approximate values f_a are within $\bar{\varepsilon}$ of exact values f, we have $|f_a(x) - F_a(x)| \leq 2\bar{\varepsilon}$, hence $\varepsilon(x) \leq 2\bar{\varepsilon}$. That shows $g_a \in \partial_{\varepsilon(x)}\phi(x,x) \subset \partial_{2\bar{\varepsilon}}\phi(x,x)$.

3) The proof of $\partial\phi(x,x) \subset \partial_{[\varepsilon]}f(x)$ remains the same, after replacing $f(x)$ by $f_a(x)$.

4) If a sequence of planes $m_r(\cdot)$, $r \in \mathbb{N}$, contributes to the envelope function $\phi(\cdot,x)$, and if $m_r(\cdot) \to m(\cdot)$ in the pointwise sense, then $m(\cdot)$ also contributes to $\phi(\cdot,x)$, because the graph of $\phi(\cdot,x)$ is closed. On the other hand, we may expect discontinuities as $x_j \to x$. We obtain $\limsup_{j\to\infty}\phi(y_j,x_j) \leq \phi(y,x) + \bar{\varepsilon}$ for $y_j \to y$, $x_j \to x$. □

Remark 9. If approximate function values are under-estimations, $f_a \leq f$, as is often the case, then $|F_a - f_a| \leq \bar{\varepsilon}$ and the result holds with $\partial\phi(x,x) \subset \partial_{[\varepsilon]}f(x) \subset \partial_{\bar{\varepsilon}}\phi(x,x)$.

Corollary 2. *Under the hypotheses of Lemma 5, if x is a point of continuity of f_a, then $\partial\phi(x,x) = \partial_{[\varepsilon]}f(x)$ and ϕ is jointly upper semicontinuous at (x,x).*

Proof. Indeed, as follows from part 2) of the proof above, for a point of continuity x of f_a we have $\varepsilon(x) = 0$. □

Lemma 6. *Suppose the inner loop at serious iterate x turns forever and $\tau_k \to \infty$. Suppose f is ε'-convex on a set containing all y^k, $k \geq k_0$, and let (1.42) be satisfied. Then $0 \in \partial_{[\widetilde{\varepsilon}]}f(x)$, where $\widetilde{\varepsilon} = \varepsilon + (\varepsilon'' + \varepsilon' + \varepsilon)/(\widetilde{\gamma} - \gamma)$.*

Proof. We go through the proof of Lemma 3 and indicate the changes caused by using approximate values $f_a(y^k)$, $f_a(x)$. Part ii) remains the same, except that $\phi(x,x) = f_a(x)$. The exactness subgradient has still $g(x) \in \partial_{[\varepsilon]}f(x)$. Part iii) leading to formula (1.17) remains the same with $f_a(x)$ instead of $f(x)$. Part iv) remains the same, and we obtain the analogue of (1.18) with $f(x)$ replaced by $f_a(x)$.

Substantial changes occur in part v) of the proof leading to formula (1.19). Indeed, consider without loss the case where $t_k(x) > f_a(x)$. Then

$$\begin{aligned} m_k(y,x) &= f_a(y^k) + g_{\varepsilon k}^\top(y-y^k) - s_k \\ &= f_a(x) + g_{\varepsilon k}^\top(y-x) - c\|x-y^k\|^2, \end{aligned}$$

as in the proof of Lemma 3, and therefore

$$f_a(y^k) - m_k(y^k,x) = f_a(y^k) - f_a(x) - g_k^\top(y^k-x) + (g_k - g_{\varepsilon k})^\top(x-y^k) + c\|x-y^k\|^2.$$

Since f is ε'-convex, we have $g_k^\top(x-y^k) \leq f(x) - f(y^k) + \varepsilon'\|x-y^k\|$. Hence

$$f_a(y^k) - m_k(y^k,x) \leq f(x) - f_a(x) - \Big(f(y^k) - f_a(y^k)\Big) + (\varepsilon' + \varepsilon + \nu_k)\|x-y^k\|,$$

where $\nu_k \to 0$. Now we use axiom (1.42), which gives

$$f_a(y^k) - m_k(y^k,x) \leq (\varepsilon'' + \varepsilon' + \varepsilon + \delta_k + \nu_k)\|x-y^k\|,$$

for $\delta_k, \nu_k \to 0$. Subtracting the usual quadratic expression on both sides gives $f_a(y^k) - M_k(y^k,x) \leq (\varepsilon'' + \varepsilon' + \varepsilon + \delta_k + \widetilde{\nu}_k)\|x - y^k\|$ with $\delta_k, \widetilde{\nu}_k \to 0$. Going back with this estimation to the expansion $\widetilde{\rho}_k \leq \rho_k + \frac{\varepsilon''+\varepsilon'+\varepsilon}{\eta}$ shows $\eta < \frac{\varepsilon''+\varepsilon'+\varepsilon}{\widetilde{\gamma}-\gamma}$ as in the proof of Lemma 3, where $\eta = \text{dist}(0, \partial\phi(x,x))$. Since $\partial\phi(x,x) \subset \partial_{[\varepsilon]}f(x)$ by Lemma 5, we have $0 \in \partial_{[\varepsilon+\eta]}f(x)$. This proves the result. □

Lemma 7. *Suppose the inner loop turns forever and τ_k is frozen from some counter k onwards. Then $0 \in \partial_{[\varepsilon]}f(x)$.*

Proof. Replacing $f(x)$ by $f_a(x)$, the proof proceeds in exactly the same fashion as the proof of Lemma 4. We obtain $0 \in \partial\phi(x,x)$ and use Lemma 5 to conclude $0 \in \partial_{[\varepsilon]}f(x)$. □

As we have seen, axiom (1.42) was necessary to deal with the case $\tau_k \to \infty$ in Lemma 6, while Lemma 7 gets by without this condition. Altogether, that means we have to adjust the stopping test in step 2 of the algorithm to $0 \in \partial_{[\widetilde{\varepsilon}]}f(x^j)$, where $\widetilde{\varepsilon} = \varepsilon + (\varepsilon'' + \varepsilon' + \varepsilon)/(\widetilde{\gamma} - \gamma)$. As in the case of exact function values, we may delegate the stopping test to the inner loop, so if the latter halts due to insufficient progress, we interpret this as $0 \in \partial_{[\widetilde{\varepsilon}]}f(x^j)$, which is the precision we can hope for. Section 1.8 below gives more details.

Let us now scan through the proof of Theorem 1 and see what changes occur through the use of inexact function values $f_a(y^k)$, $f_a(x^j)$.

Theorem 2. *Let x^1 be such that $\Omega' = \{x \in \mathbb{R}^n : f(x) \leq f(x^1) + 2\bar{\varepsilon}\}$ is bounded. Suppose f is ε'-convex on Ω, that subgradients are drawn from $\partial_{[\varepsilon]}f(y)$, and that inexact function values $f_a(y)$ satisfy $|f(y) - f_a(y)| \leq \bar{\varepsilon}$. Suppose axiom (1.42) is satisfied. Then every accumulation point $\bar{x}$ of the sequence x^j satisfies $0 \in \partial_{[\widetilde{\varepsilon}]}f(\bar{x})$, where $\widetilde{\varepsilon} = \varepsilon + (\varepsilon'' + \varepsilon' + \varepsilon)/(\widetilde{\gamma} - \gamma)$.*

Proof. Notice that $\widetilde{\varepsilon}$ used in the stopping test has a different meaning than in Theorem 2. Replacing $f(x^j)$ by $f_a(x^j)$ and $f(y^{k_j})$ by $f_a(y^{k_j})$, we follow the proof of Theorem 1. Part i) is still valid with these changes. Notice that $\Omega = \{x : f_a(x) \leq f_a(x^1)\} \subset \Omega'$, and Ω' is bounded by hypothesis, so Ω is bounded.

As in the proof of Theorem 1 the set of all trial points $y^1, \ldots, y^{k_j}$ visited during all the inner loops j is bounded. However, a major change occurs in part ii). Observe that the accumulation point $\bar{x}$ used in the proof of Theorem 1 is neither among the trial points nor the serious iterates. Therefore, $f_a(\bar{x})$ is never called for in the algorithm. Now observe that the sequence $f_a(x^j)$ is decreasing and by boundedness of Ω converges to a limit $F_a(\bar{x})$. We re-define $f_a(\bar{x}) = F_a(\bar{x})$, which is consistent with the condition $|f_a(\bar{x}) - f(\bar{x})| \leq \bar{\varepsilon}$, because $f_a(x^j) \geq f(x^j) - \bar{\varepsilon}$, so that $F_a(\bar{x}) \geq f(\bar{x}) - \bar{\varepsilon}$.

The consequences of the re-definition of $f_a(\bar{x})$ are that the upper envelope model ϕ is now jointly upper semicontinuous at $(\bar{x},\bar{x})$, and that the argument leading to formula (1.29) remains unchanged, because $f_a(x^j) \to \phi(\bar{x},\bar{x})$.

Let us now look at the longer argument carried out in parts iii) - ix) of the proof of Theorem 1, which deals with the case where $\|g_j\| \geq \mu > 0$ for all j. Parts iii)

- vii) are adapted without difficulty. Joint upper semi-continuity of ϕ at $(\bar{x}+h,\bar{x})$ is used at the end of vii), and this is assured as a consequence of the re-definition $f_a(\bar{x}) = F_a(\bar{x})$ of f_a at $\bar{x}$.

Let us next look at part viii). In Theorem 1 we use ε'-convexity. Since the latter is in terms of exact values, we need axiom (1.42) for the sequence $y^{k_j-\nu_j} \to \bar{x}$, similarly to the way it was used in Lemma 5. We have to check that despite the re-definition of f_a at $\bar{x}$ axiom (1.42) is still satisfied. To see this, observe that $y^{k_j-\nu_j}$ is a trial step which is rejected in the j^{th} inner loop, so that its approximate function value is too large. In particular, $f_a(y^{k_j-\nu_j}) \geq f_a(x^{j+1})$, because x^{j+1} is the first trial step accepted. This estimate shows that (1.42) is satisfied at $\bar{x}$.

Using (1.42) we get the analogue of (1.36), which is

$$f_a(y^{k_j-\nu_j}) - M_{k_j-\nu_j}(y^{k_j-\nu_j}, x^j) \leq (\varepsilon'' + \varepsilon' + \nu_j + \delta_j)\|y^{k_j-\nu_j} - x^j\|$$

for certain $\nu_j, \delta_j \to 0$. Estimate (1.40) remains unchanged, so we can combine the two estimates to obtain the analogue of (1.41) in part ix), which is

$$\widetilde{\rho}_{k_j-\nu_j} \leq \rho_{k_j-\nu_j} + \frac{(1+\zeta^2)(\varepsilon''+\varepsilon'+\varepsilon)}{(1-\zeta)^2\eta}.$$

Using the same argument as in the proof of Theorem 1, we deduce

$$\eta \leq \frac{(1+\zeta)^2(\varepsilon''+\varepsilon'+\varepsilon)}{(1-\zeta)^2(\widetilde{\gamma}-\gamma)}$$

for $\eta = \operatorname{dist}(0, \partial\phi(x,x))$. Since $0 < \zeta < 1$ was arbitrary, we obtain $\eta \leq \frac{\varepsilon''+\varepsilon'+\varepsilon}{\widetilde{\gamma}-\gamma}$. Now as $\bar{x}$ is a point of continuity of f_a, Corollary 2 tells us that $\eta = \operatorname{dist}(0, \partial_{[\varepsilon]} f(\bar{x}))$. Therefore $0 \in \partial_{[\varepsilon+\eta]} f(\bar{x})$. Since $\varepsilon + \eta = \widetilde{\varepsilon}$, we are done. □

1.8 Stopping

In this section we address the practical problem of stopping the algorithm. The idea is to use tests which are based on the convergence theory developed in the previous sections.

In order to safe time, the stopping test in step 2 of the algorithm is usually delegated to the inner loop. This is based on Lemmas 3, 4 and the following

Lemma 8. *Suppose tangent program (1.7) has the solution $y^k = x$. Then $0 \in \partial_{[\varepsilon]} f(x)$.*

Proof. From (1.8) we have $0 \in \partial\phi_k(x,x) \subset \partial\phi(x,x) \subset \partial_{[\varepsilon]} f(x)$ by Lemma 5. □

In [22] we use the following two-stage stopping test. Fixing a tolerance level tol > 0, if x^+ is the serious step accepted by the inner loop at x, and if x^+ satisfies

$$\frac{\|x - x^+\|}{1+\|x\|} < \text{tol},$$

then we stop the outer loop and accept x^+ as the solution, the justification being Lemma 8. On the other hand, if the inner loop at x fails to find x^+ and either exceeds a maximum number of allowed inner iterations or provides three consecutive trial steps y^k satisfying

$$\frac{\|x-y^k\|}{1+\|x\|} < \text{tol},$$

then we stop the inner loop and the algorithm and accept x as the final solution. Here the justification comes from Lemmas 3,4.

Remark 10. An interesting aspect of inexactness theory with unknown precisions $\varepsilon, \varepsilon', \varepsilon''$ are the following two scenarios, which may require different handling. The first is when functions and subgradients are inexact or noisy, but we do not take this into account and proceed as if information were exact. The second scenario is when we deliberately use inexact information in order to gain speed or deal with problems of very large size. In the first case we typically arrange all elements of the algorithm like in the exact case, including situations where we are not even aware that information is inexact. In the second case we might introduce new elements which make the most of the fact that data are inexact.

As an example of the latter, in [32] where f is convex, the author does not use downshift with respect to $f_a(x)$, and as a consequence one may have $\phi_k(x,x) > f_a(x)$, so that the tangent program (1.7) may fail to find a predicted descent step y^k at x. The author then uses a sub-loop of the inner loop, where the τ-parameter is decreased until *either* a predicted descent step is found, *or* optimality within the allowed tolerance of function values is established.

1.9 Example from control

Optimizing the H_∞-norm [1,7,21,22] is a typical application of (1.1) where inexact function and subgradient evaluations may arise. The objective function is of the form

$$f(x) = \max_{\omega\in\mathbb{R}} \overline{\sigma}\left(G(x, j\omega)\right), \tag{1.43}$$

where $G(x,s) = C(x)\left(sI - A(x)\right)^{-1} B(x) + D(x)$ is defined on the open set $S = \{x \in \mathbb{R}^n : A(x) \text{ stable}\}$, and where $A(x)$, $B(x)$, $C(x)$, $D(x)$ are matrix valued mappings depending smoothly on $x \in \mathbb{R}^n$. In other words, for $x \in S$ each $G(x,s)$ is a stable real-rational transfer matrix.

Notice that f is a composite function of the form $f = \|\cdot\|_\infty \circ \mathscr{G}$, where $\|\cdot\|_\infty$ is the H_∞-norm, which turns the Hardy space $\mathscr{H}_\infty$ of functions G which are analytic and bounded in the open right-half plane [55, p. 100] into a Banach space,

$$\|G\|_\infty = \sup_{\omega\in\mathbb{R}} \overline{\sigma}\left(G(j\omega)\right),$$

and $\mathscr{G}: S \to \mathscr{H}_\infty$, $x \mapsto G(x,\cdot) = C(x)(\cdot I - A(x))^{-1}B(x) + D(x) \in \mathscr{H}_\infty$ is a smooth mapping, defined on the open subset $S = \{x \in \mathbb{R}^n : A(x) \text{ stable}\}$. Since composite functions of this form are lower C^2, and therefore also lower C^1, we are in business. For the convenience of the reader we also include a more direct argument proving the same result:

Lemma 9. *Let f be defined by (1.43), then f is lower C^2, and therefore also lower C^1, on the open set $S = \{x \in \mathbb{R}^n : A(x) \text{ stable}\}$.*

Proof. Recall that $\overline{\sigma}(G) = \max_{\|u\|=1} \max_{\|v\|=1} \operatorname{Re} u G v^H$, so that

$$f(x) = \max_{\omega \in \mathbb{S}^1} \max_{\|u\|=1} \max_{\|v\|=1} \operatorname{Re} u G(x, j\omega) v^H.$$

Here, for $x \in S$, the stability of $G(x,\cdot)$ assures that $G(x,s)$ is analytic in s on a band $\mathscr{B}$ on the Riemann sphere $\mathbb{C} \cup \{\infty\}$ containing the zero meridian $j\mathbb{S}^1$ with $\mathbb{S}^1 = \{\omega : \omega \in \mathbb{R} \cup \{\infty\}\}$, a compact set homeomorphic to the real 1-sphere. This shows that f is lower C^2 on the open set S. Indeed, $(x,\omega,u,v) \mapsto F(x,\omega,u,v) := \operatorname{Re} u G(x, j\omega) v^H$ is jointly continuous on $S \times \mathbb{S}^1 \times \mathbb{C}^m \times \mathbb{C}^p$ and smooth in x, and $f(x) = \max_{(\omega,u,v) \in K} F(x,\omega,u,v)$ for the compact set $K = \mathbb{S}^1 \times \{u \in \mathbb{C}^m : \|u\| = 1\} \times \{v \in \mathbb{C}^p : \|v\| = 1\|\}$. □

The evaluation of $f(x)$ is based on the iterative bisection method of Boyd *et al.* [11]. Efficient implementations use Boyd and Balakrishnan [12] or Bruisma and Steinbuch [13] and guarantee quadratic convergence. All these approaches are based on the Hamiltonian test from [11], which states that $f(x) > \gamma$ if and only if the Hamiltonian

$$H(x,\gamma) = \begin{bmatrix} A(x) & 0 \\ 0 & -A(x)^\top \end{bmatrix} - \begin{bmatrix} 0 & B(x) \\ C(x)^\top & 0(x) \end{bmatrix} \begin{bmatrix} \gamma I & D(x) \\ D(x)^\top & \gamma I \end{bmatrix}^{-1} \begin{bmatrix} C(x) & 0 \\ 0 & -B(x)^\top \end{bmatrix} \tag{1.44}$$

has purely imaginary eigenvalues $j\omega$. The bundle method of [7], which uses (1.44) to compute function values, can now be modified to use approximate values $f_a(y^k)$ for unsuccessful trial points y^k. Namely, if the trial step y^k is to become the new serious iterate x^+, its value $f(y^k)$ has to be below $f(x)$. Therefore, as soon as the Hamiltonian test (1.44) certifies $f(y^k) > f(x)$ even before the exact value $f(y^k)$ is known, we may dispense with the exact computation of $f(y^k)$. We may stop the Hamiltonian algorithm at the stage where the first γ with $f(y^k) > \gamma \geq f(x)$ occurs, compute the intervals where $\omega \mapsto \overline{\sigma}(G(x, j\omega))$ is above γ, take the midpoints of these intervals, say $\omega_1, \dots, \omega_r$, and pick the one where the frequency curve is maximum. If this is ω_ν, then $f_a(y^k) = \overline{\sigma}(G(x, j\omega_\nu))$. The approximate subgradient g_a is computed via the formulas of [1] with ω_ν replacing an active frequency. This procedure is trivially consistent with (1.42), because $f(x) = f_a(x)$ and $f_a(y) \leq f(y)$.

If we wish to allow inexact values not only at trial points y but also at serious iterates x, we can use the termination tolerance of the Hamiltonian algorithm [12]. The algorithm works with estimates $f_l(x) \leq f(x) \leq f_u(x)$ and terminates when

$f_u(x) - f_l(x) \leq 2\eta_x F(x)$, returning $f_a(x) := (f_l(x) + f_u(x))/2$, where we have the choice $F(x) \in \{f_l(x), f_u(x), f_a(x)\}$. Then $|f(x) - f_a(x)| \leq 2\eta_x |F(x)|$. As η_x is under control, we can arrange that $\eta_x |F(x)| \leq \eta_y |F(y)| + o(\|x-y\|)$ in order to assure condition (1.42).

Remark 11. The outlined method applies in various other cases in feedback control where function evaluations use iterative procedures, which one may stop short to save time. We mention IQC-theory [4], which uses complex Hamiltonians, [7] for related semi-infinite problems, or the multidisk problem [3], where several H_∞-criteria are combined in a progress function. The idea could be used quite naturally in the ε-subgradient approaches [38, 39], or in search methods like [2].

Acknowledgements The author acknowledges funding by Fondation d'Entreprise EADS under grant *Technicom*, and by Fondation de Recherche pour l'Aéronautique et l'Espace under grant *Survol*.

References

1. Apkarian, P., Noll, D.: Nonsmooth H_∞-synthesis. Transactions on Automatic Control, **51**, no. 1, 229 – 244 (2006).
2. Apkarian, P., Noll, D.: Controller design via nonsmooth multidirectional search. SIAM Journal on Control and Applications, **44**, no. 6, 1923 – 1949 (2006).
3. Apkarian, P., Noll, D.: Nonsmooth optimization for multidisk H_∞-synthesis. European Journal of Control, **12**, no. 3, 229 – 244 (2006).
4. Apkarian, P., Noll, D.: IQC analysis and synthesis via nonsmooth optimization. Systems and Control Letters, **55**, no. 12, 971 – 981 (2006).
5. Apkarian, P., Noll, D.: Nonsmooth optimization for multiband frequency domain control design. Automatica, **43**, no. 4, 724 – 731 (2007).
6. Apkarian, P., Noll, D., Prot, O.: A trust region spectral bundle method for non-convex eigenvalue optimization. SIAM Journal on Optimization, **19**, no. 1, 281 – 306 (2008).
7. Apkarian, P., Noll, D., Prot, O.: A proximity control algorithm to minimize non-smooth and non-convex semi-infinite maximum eigenvalue functions. Journal of Convex Analysis, **16**, 641 – 666 (2009).
8. Apkarian, P., Ravanbod-Hosseini, L., Noll, D.: Time-domain constrained structured H_∞-synthesis. International Journal of Robust and Nonlinear Control, **21**, no. 2, 197 – 217 (2010).
9. Bompart, V., Noll, D., Apkarian, P.: Second order nonsmooth optimization for H_∞-synthesis. Numerische Mathematik, **107**, no. 3, 433 – 454 (2007).
10. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**, no. 2, 556–572 (2007).
11. Boyd, S., Balakrishnan, V.: A regularity result for the singular values of a transfer matrix and a quadratically convergent algorithm to compute its L_∞-norm. Systems and Control Letters, **15**, 1 – 7 (1990).
12. Boyd, S., Balakrishnan, V., Kabamba, P. : A bisection method for computing the H_∞-norm of a transfer matrix and related problems. Mathematics of Control, Signals, and Systems, **2**, 207 – 219 (1989).
13. Bruisma, N., Steinbuch, M.: A fast algorithm to compute the H_∞ norm of a transfer function. Systems and Control Letters, **14**, 287 – 293 (1990).
14. Clarke, F.: Optimization and Nonsmooth Analysis. SIAM Publications (1983).

15. Cullum, J., Donath, W., Wolfe, P.: The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices. Math. Programming Study, **3**, 35 – 55 (1975).
16. Daniilidis, A., Georgiev, P.: Approximate convexity and submonotonicity. J. Math. Anal. Appl., **291**, 117 – 144 (2004).
17. Emiel, G., Sagastizábal, C.: Incremental-like bundle methods with application to energy planning. Computational Optimization and Applications, **46**, no. 2, 305 – 332 (2010).
18. Emiel, G., Sagastizábal, C.: Dynamic subgradient methods. Preprint 2010.
19. Fuduli, A., Gaudioso, M., Giallombardo, G.: A DC piecewise affine model and a bundling technique in nonconvex nonsmooth optimization. Optimization Methods and Software, **19**, 89 - 102 (2004).
20. Fuduli, A., Gaudioso, M., Giallombardo, G.: Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. SIAM J. Optim. **14**, 743 - 756 (2004).
21. Gabarrou, M., Alazard, D., Noll, D.: Structured flight control law design using non-smooth optimization. Proceedings of the 18^{th} IFAC Symposium on Automatic Control in AeroSpace (ACA), Nara, Japan, 2010.
22. Gabarrou, M., Alazard, D., Noll, D.: Design of a flight control architecture using a non-convex bundle method. Submitted 2012.
23. Hare, W.L., Sagastizábal, C.: Redistributed proximal bundle method for nonconvex optimization. SIAM J. Opt., **20**, no. 5, 2442 – 2473 (2010).
24. Hintermüller, M.: A proximal bundle method based on approximate subgradients. Computational Optimization and Applications, **20**, 245 – 266 (2001).
25. Hiriart-Urruty, J.-B., Seeger, A.. The second-order subdifferential and the Dupin indicatrix of a non-differentiable convex function. Proc. London Math. Soc. **58**, 351 – 365 (1989).
26. Hiriart-Urruty, J.-B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms II. Springer Verlag (1993).
27. Kiwiel, K.C.: An aggregate subgradient method for non-smooth convex minimization. Math. Programming, **27**, 320 - 341 (1983).
28. Kiwiel, K.C..: An algorithm for nonsmooth convex minimization with errors. Mathematics of Computation, **45**, 171 – 180 (1985).
29. Kiwiel, K.C.: A linearization algorithm for computing control systems subject to singular value inequalities. IEEE Trans. Autom. Control, **AC-31**, 595 - 602 (1986).
30. Kiwiel, K.C.: Approximations in proximal bundle methods and decomposition of convex programs. J. Optim. Theory Appl., **84**, no. 3, 529 – 548 (1995).
31. Kiwiel, K.C.: Convergence of approximate and incremental subgradient methods for convex optimization. SIAM J. Optim. **14**, no. 3, 807 – 840 (2003).
32. Kiwiel, K.C.: A proximal bundle method with approximate subgradient linearizations. SIAM J. on Optimization, **16**, 1007 – 1023 (2006).
33. Kiwiel, K.C., Lemaréchal, C. : An inexact bundle variant suited to column generation. Math. Programming, Series A, **118**, no. 1, 177 – 206 (2007).
34. Lemaréchal, C., Strodiot, J.-J., Bihain, A.: On a bundle algorithm for non-smooth optimization. In: Mangasarian, Meyer, Robinson (eds.), Nonlinear Programming 4, Academic Press, 245 – 282 (1981).
35. Lewis, A., Wright, S.: A proximal method for composite minimization. Preprint 2008.
36. Luc, D.T., Van Ngai, H., Théra, M.: On ε-monotonicity and ε-convexity. Calculus of Variations and Differential Equations (Haifa 1998), Chapman & Hall/CRC Research Notes in Mathematics, **410**, 82 – 100 (2000).
37. Lukšan, L., Vlček, J.: A bundle-Newton method for nonsmooth unconstrained minimization. Math. Programming, **83**, 373 - 391 (1998).
38. Mayne, D., Polak, E.: Algorithms for the design of control systems subject to singular value inequalities. Math. Programming Studies, **18**, 112 - 134 (1982).
39. Mayne, D., Polak, E. , Sangiovanni, A.: Computer aided design via optimization. Automatica, **18**, no. 2, 147 - 154 (1982).
40. Mifflin, R.: A quasi-second-order proximal bundle algorithm. Math. Programming Series A, **73**, no. 1, 51 – 72 (1996).

41. Nedić, A., Bertsekas, D.P.: The effect of deterministic noise in subgradient methods. Math. programming series A, **125**, no. 1, 75 – 99 (2010).
42. Noll, D., Apkarian, P.: Spectral bundle method for non-convex maximum eigenvalue functions: first-order methods. Math. Programming, Ser. B **104**, 701 – 727 (2005).
43. Noll, D., Prot, O., Rondepierre, A.: A proximity control algorithm to minimize non-smooth and non-convex functions. Pacific Journal of Optimization, **4**, no. 3, 569 – 602 (2008).
44. Noll, D.: Cutting plane oracles to minimize non-smooth non-convex functions. Set-Valued and Variational Analysis, **18**, no. 3-4, 531 – 568 (2010).
45. Polak, E.: On the mathematical foundations of nondifferentiable optimization in engineering design. SIAM Review, **29**, 21 - 89 (1987).
46. Polak, E., Wardi, Y.: A nondifferentiable optimization algorithm for the design of control systems subject to singular value inequalities over the frequency range. Automatica, **18**, 267 - 283 (1982).
47. Rockafellar, R.T., Wets, R.: Variational Analysis. Springer Verlag (2004).
48. Ruszczyǹski, A.: Nonlinear Optimization. Princeton University Press (2006).
49. Sagastizábal, C.: Composite proximal bundle method. Preprint 2009.
50. Schramm, H., Zowe, J.: A version of the bundle idea for minimizing nondifferentiable functions: conceptual idea, convergence analysis, numerical results. SIAM Journal on Optimization, **2**, 121 – 151 (1992).
51. Shor, N.: Minimization Methods for Nondifferentiable Functions. Springer Verlag (1985).
52. Simões, A., Apkarian, P., Noll, D.: Non-smooth multi-objective synthesis with applications. Control Engineering Practice, **17**, no. 11, 1338 – 1348 (2009).
53. Spingarn, J.E.: Submonotone subdifferentials of Lipschitz functions. Trans. Amer. Math. Soc., **264**, 77 – 89 (1981).
54. Wolfe, P.: A method of conjugate subgradients for minimizing nondifferentiable functions. Math. Programming Studies, **3**. Nondifferentiable Optimization, M.L. Balinski, P. Wolfe (eds.), North-Holland, 145 - 173 (1975).
55. Zhou, K., Doyle, J.C., Glover, K.: Robust and Optimal Control. Prentice Hall (1996).
56. Zowe, J.: 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).