A $\mathcal{VU}$-algorithm for convex minimization

Robert Mifflin* and Claudia Sagastizábal†

revised April 19, 2005

Abstract

For convex minimization we introduce an algorithm based on $\mathcal{VU}$-space decomposition. The method uses a bundle subroutine to generate a sequence of approximate proximal points. When a primal-dual track leading to a solution and zero subgradient pair exists, these points approximate the primal track points and give the algorithm's $\mathcal{V}$, or corrector, steps. The subroutine also approximates dual track points that are $\mathcal{U}$-gradients needed for the method's $\mathcal{U}$-Newton predictor steps. With the inclusion of a simple line search the resulting algorithm is proved to be globally convergent. The convergence is superlinear if the primal-dual track points and the objective's $\mathcal{U}$-Hessian are approximated well enough.

Keywords Convex minimization, proximal points, bundle methods, $\mathcal{VU}$-decomposition, superlinear convergence.

1 Introduction and motivation

We consider the problem

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

where f is a finite-valued convex function. A conceptual algorithm to solve this problem is the proximal point method; see [Mor65] and [Roc76]. Implementable forms of the method can be obtained by means of a bundle technique, alternating serious steps with sequences of null steps [Aus87], [Fuk84], [HUL93]. The last decade has produced a "new generation" of proximal bundle methods, designed to seek faster convergence; see [LS94], [LS96], [QC97], [Mif96], [LS97b], [MSQ98], [CF99], [RF00]. Essentially, these methods introduce second-order information via f's Moreau-Yosida regularization F.

*Department of Mathematics, Washington State University, Pullman, WA 99164-3113. Research supported by the National Science Foundation under Grant No. DMS-0071459 and by CNPq (Brazil) under Grant No. 452966/2003-5. mifflin@math.wsu.edu

†IMPA, Estrada Dona Castorina 110, Jardim Botânico, Rio de Janeiro RJ 22460-320, Brazil. On leave from INRIA Rocquencourt, France. Research supported by FAPERJ (Brazil) under Grant No.E26/150.581/00 and by CNPq (Brazil) under Grant No. 383066/2004-2. sagastiz@impa.br

More recently, new conceptual schemes have been developed from an approach that is somewhat different from Moreau-Yosida regularization. These are based on the $\mathcal{VU}$-theory introduced in [LOS00] for convex functions; see also [MS99], [MS00b], [MS00a], [Ous00] and [MM04]. The idea is to decompose $\mathbb{R}^n$ into two orthogonal subspaces $\mathcal{V}$ and $\mathcal{U}$ depending on a point in such a way that, near the point, f's nonsmoothness is essentially due to its V-shaped graph on the $\mathcal{V}$-subspace. When f satisfies certain structural properties, it is possible to find a smooth trajectory, tangent to $\mathcal{U}$, yielding a second-order expansion for f. The very conceptual $\mathcal{VU}$-algorithm in [LOS00] finds points on such a trajectory, by generating minimizing steps in the $\mathcal{V}$-subspace. Alternating with these corrector steps are $\mathcal{U}$-Newton predictor steps that provide for superlinear convergence. However, since such an algorithm relies on knowing the subspaces $\mathcal{V}$ and $\mathcal{U}$ and converges only locally, it needs significant modification for implementation.

In [MS02] we establish a fundamental result for implementability by showing that, near a minimizer, a proximal point sequence follows a particular smooth trajectory that is called a *fast track*. This relation opens the way for defining a $\mathcal{VU}$-algorithm where $\mathcal{V}$-steps are replaced by proximal steps that can be estimated with a bundle technique that also approximates the unknown $\mathcal{V}$ and $\mathcal{U}$ subspaces as a computational by-product.

Also shown in [MS02] is the result that a convex function with a strongly transversal *primal-dual gradient* (`pdg`) structure has a fast track; see also [MS00b], [MS03]. A general function of this type can have a *primal-dual track* leading to a (minimizing point, zero subgradient) pair; see [MS04]. On the primal track such a function is C^2, while the dual track corresponds to a C^1 subgradient function defined pointwise as the minimum norm vector in the subdifferential at a primal track point. For our convex f a primal track is a fast track which, in turn, is a proximal point track, each of whose points can be approximated arbitrarily well by a bundle algorithm subroutine. Since such a subroutine collects subgradients for constructing a V-(or cutting-plane) model of f it also can approximate a dual track point corresponding to a bundle iterate where the V-model is sufficiently accurate. To complete the algorithm's VU-model combination the dual vector goes into updating a U-(or quadratic) model of a $\mathcal{U}$-Lagrangian [LOS00] that equals f on the primal track.

Our resulting bundle-based $\mathcal{VU}$-algorithm does not need to know `pdg` structure or related tracks to operate nor is existence of such structure needed for showing global convergence. Minimizing convergence from any starting point is accomplished by embedding a simple possible line search in the above framework. This algorithm represents a new type of bundle method as it has a new kind of bundle subroutine exit test that simultaneously involves the primal and dual (and associated $\mathcal{U}$-basis matrix) iterate estimates (see (14) below).

Another interesting feature of our algorithm is that it also can be considered as a way to speed up the proximal point method, because it adds a second-order step to each proximal iterate. This second-order step is done only relative to a $\mathcal{U}$-subspace estimate, unlike second-order Moreau-Yosida algorithms that employ estimates of Hessians of F relative to the whole space. In addition, for global convergence our method does not require line searches on F, but instead on the natural merit function, f itself, as in [CF99].

This paper is organized as follows. Section 2 gathers together all of the relevant results concerning $\mathcal{VU}$-theory. In Section 2.3 we review the main properties of proximal points and give the crucial relation between them and primal track points. Section 3 describes a bundle subprocedure needed by the main algorithm for computing good approximations of primal-dual track points. Then, in Section 4, we introduce our $\mathcal{VU}$-algorithm. Section 5 is devoted to proving both global and superlinear convergence, with the latter under reasonable conditions contained in assumptions **(S)(c)-(e)**. Some preliminary numerical results are reported in Section 6. The final section contains some concluding remarks.

Our notation follows that of [MS04] and [RW98]. In addition, given a sequence of vectors $\{z_k\}$ converging to 0,

– $\zeta_k = o(|z_k|) \iff \forall \varepsilon > 0\ \exists k_\varepsilon > 0$ such that $|\zeta_k| \le \varepsilon |z_k|$ for all $k \ge k_\varepsilon$;

– $\zeta_k = O(|z_k|) \iff \exists C > 0$ such that $|\zeta_k| \le C|z_k|$ for all $k \ge 1$.

For algebraic purposes we consider (sub)gradients to be column vectors. The symbol ∂ stands for subdifferentiation with respect to $x \in \mathbb{R}^n$, while ∇ indicates differentiation with respect to $u \in \mathbb{R}^{dim\, \mathcal{U}}$. For a vector function $v(\cdot)$, its Jacobian $Jv(\cdot)$ is a matrix, each row of which is the transposed gradient of the corresponding component of $v(\cdot)$. Finally, lin$Y$ denotes the linear hull of a set Y.

2 $\mathcal{VU}$-theory

We start by reviewing $\mathcal{VU}$-space decomposition and associated $\mathcal{U}$-Lagrangians from [LOS00] and defining primal-dual tracks relative to results from [MS02] and [MS04].

2.1 $\mathcal{VU}$-space decomposition and $\mathcal{U}$-Lagrangians

Throughout this paper we assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function. Let g be any subgradient in $\partial f(x)$, the subdifferential of f at $x \in \mathbb{R}^n$. Then the orthogonal subspaces

$$\mathcal{V}(x) := \mathrm{lin}(\partial f(x) - g) \qquad \text{and} \qquad \mathcal{U}(x) := \mathcal{V}(x)^\perp$$

written with $x = \bar{x}$ define the $\mathcal{VU}$-space decomposition at $\bar{x}$ from [LOS00, §2]. More precisely, $\mathbb{R}^n = \mathcal{U} \oplus \mathcal{V}$, where $\mathcal{V} := \mathcal{V}(\bar{x})$ and $\mathcal{U} := \mathcal{U}(\bar{x})$. From this definition, the relative interior of $\partial f(\bar{x})$, denoted by ri$\partial f(\bar{x})$, is the interior of $\partial f(\bar{x})$ relative to its affine hull, a manifold that is parallel to $\mathcal{V}$ (cf. [LOS00, relation 2.1 and Prop. 2.2]).

By letting $\bar{V}$ be a basis matrix for $\mathcal{V}$ and $\bar{U}$ be an orthonormal basis matrix for $\mathcal{U}$, every $x \in \mathbb{R}^n$ can be decomposed into components $x_\mathcal{U}$ and $x_\mathcal{V}$ as follows:

$$\begin{array}{rcccccl} \mathbb{R}^n \ni x & = & \bar{U} & (\bar{U}^\top x) & + & \bar{V} & \left([\bar{V}^\top \bar{V}]^{-1} \bar{V}^\top x\right) \\ & = & \bar{U} & x_\mathcal{U} & + & \bar{V} & x_\mathcal{V} \\ & = & & x_\mathcal{U} & \oplus & & x_\mathcal{V} \qquad \in \mathbb{R}^{dim\, \mathcal{U}} \times \mathbb{R}^{dim\, \mathcal{V}}. \end{array}$$

The reason why $\bar{V}$ is not assumed to be orthonormal too is because typical $\mathcal{V}$-basis matrix approximations made by minimization algorithms are not orthonormal.

Given a subgradient $\bar{g} \in \partial f(\bar{x})$ with $\mathcal{V}$-component $\bar{g}_{\mathcal{V}} = ([\bar{V}^\top \bar{V}]^{-1}\bar{V}^\top)\bar{g}$, the $\mathcal{U}$-Lagrangian of f, depending on $\bar{g}_{\mathcal{V}}$, is defined by

$$\mathbb{R}^{dim\,\mathcal{U}} \ni u \mapsto L_{\mathcal{U}}(u;\bar{g}_{\mathcal{V}}) := \min_{v\in\mathbb{R}^{dim\,\mathcal{V}}} \{f(\bar{x}+\bar{U}u+\bar{V}v) - \bar{g}^\top\bar{V}v\}\,.$$

The associated set of $\mathcal{V}$-space minimizers is defined by

$$W(u;\bar{g}_{\mathcal{V}}) := \{\bar{V}v : L_{\mathcal{U}}(u;\bar{g}_{\mathcal{V}}) = f(\bar{x}+\bar{U}u+\bar{V}v) - \bar{g}^\top\bar{V}v\}\,.$$

From [LOS00], $W(u;\bar{g}_{\mathcal{V}})$ is nonempty if $\bar{g} \in \mathrm{ri}\partial f(\bar{x})$, but this is not a necessary condition; see [MS99] and [MS00b]. Each $\mathcal{U}$-Lagrangian is a convex function that is differentiable at $u = 0$ with

$$\nabla L_{\mathcal{U}}(0;\bar{g}_{\mathcal{V}}) = \bar{g}_{\mathcal{U}} = \bar{U}^\top\bar{g} = \bar{U}^\top g \quad \text{ for all } g \in \partial f(\bar{x}).$$

The case of interest here is when $\bar{x}$ is a minimizer. In this case, $0 \in \partial f(\bar{x})$, so for all $\bar{g} \in \partial f(\bar{x})$, $\nabla L_{\mathcal{U}}(0;\bar{g}_{\mathcal{V}}) = 0$, $u = 0$ minimizes $L_U(u;\bar{g}_{\mathcal{V}})$, and $L_{\mathcal{U}}(0;0) = f(\bar{x})$.

2.2 Primal-dual tracks

When $L_{\mathcal{U}}(u;0)$ has a Hessian at $u = 0$, this $\mathcal{U}$-Lagrangian can be expanded up to second order. For the purpose of algorithmic exploitation of a related second-order expansion of f, we next define and analyze a particular pair of trajectories.

Definition 1 We say that $(\chi(u),\gamma(u))$ is a *primal-dual track* leading to $(\bar{x},0)$, a minimizer of f and zero subgradient pair, if for all $u \in \mathbb{R}^{dim\,\mathcal{U}}$ small enough

$$\text{the } \textit{primal track } \chi(u) = \bar{x} + u \oplus v(u) \quad \text{ and}$$
$$\text{the } \textit{dual track } \gamma(u) = \operatorname{argmin}\left\{|g|^2 : g \in \partial f(\chi(u))\right\}, \tag{1}$$

satisfy the following:

(i) $v : \mathbb{R}^{dim\,\mathcal{U}} \mapsto \mathbb{R}^{dim\,\mathcal{V}}$ is a C^2-function satisfying $\bar{V}v(u) \in W_{\mathcal{U}}(u;\bar{g}_{\mathcal{V}})$ for all $\bar{g} \in \mathrm{ri}\partial f(\bar{x})$,

(ii) the Jacobian $J\chi(u)$ is a basis matrix for $\mathcal{V}(\chi(u))^\perp$, and

(iii) the particular $\mathcal{U}$-Lagrangian $L_{\mathcal{U}}(u;0)$ is a C^2-function.

When we write $v(u)$ we implicitly assume that $dim\,\mathcal{U} \geq 1$. If $dim\,\mathcal{U} = 0$ we define the primal-dual track to be the point $(\bar{x},0)$. If $dim\,\mathcal{U} = n$ then $(\chi(u),\gamma(u)) = (\bar{x}+u,\nabla f(\bar{x}+u))$ for all u in a ball about $0 \in \mathbb{R}^n$. □

Theorem 4.2 in [MS02] combined with Corollary 6 in [MS04] shows that if $0 \in \mathrm{ri}\partial f(\bar{x})$ and f has a `pdg` structure about $\bar{x}$ satisfying strong transversality, as defined in [MS00b], then f has a primal-dual track leading to $(\bar{x},0)$. The primal track defined here is called a fast track in [MS02] and the required specialization of Corollary 6 in [MS04] corresponds to $\bar{g} = 0$ and $\overline{\partial} f$ equal to the convex function subdifferential ∂f. The class of `pdg`-structured functions appears to be rather large, including general max-functions, such as maximum eigenvalue functions, and integral functions with max-function integrands.

Remark 2 *We take this opportunity to point out that Remark 2.3 in [MS02] is incorrect and should be deleted. This has no effect on any of the results in [MS02]. The fact that it is incorrect means that $\bar{g} \in \mathrm{ri}\partial f(\bar{x})$ should not be replaced by $\bar{g} \in \partial f(\bar{x})$ in part (i) of Definition 1 (Definition 2.1 in [MS02]).*

Whenever condition (i) in Definition 1 holds, from [LOS00, Corollary 3.5], $v(0) = 0$, $Jv(0) = 0$ and

$$v(u) = O(|u|^2)\,, \tag{2}$$

so $\chi(u)$ is a trajectory that is tangent to $\mathcal{U}$ at $\bar{x}$. Item (ii) in Definition 1 is such a tangency condition for the entire primal track. As for the dual track, we show in item (v) of the next Lemma that it is a C^1 "$\mathcal{U}$-gradient" that is tangent to the primal trajectory.

Lemma 3 *Let $(\chi(u), \gamma(u))$ be a primal-dual track leading to $(\bar{x}, 0)$ and let $\bar{H} := \nabla^2 L_{\mathcal{U}}(0;0)$. Suppose $0 \in \mathrm{ri}\partial f(\bar{x})$. Then for all u sufficiently small the following hold:*

(i) $\chi(u)$ is a C^2-function with $J\chi(u) = \bar{U} + O(|u|)$,

(ii) $L_{\mathcal{U}}(u;0) = f(\bar{x} + u \oplus v(u)) = f(\bar{x}) + \frac{1}{2}u^\top \bar{H} u + o(|u|^2)$,

(iii) $\nabla L_{\mathcal{U}}(u;0) = \bar{H}u + o(|u|)$,

(iv) $\chi(u)$ is the unique minimizer of f on the affine set $\chi(u) + \mathcal{V}(\chi(u))$,

(v) $\gamma(u)$ is a C^1-function with $\gamma(u) = J\chi(u)[J\chi(u)^\top J\chi(u)]^{-1}\nabla L_{\mathcal{U}}(u;0) \in \mathrm{ri}\partial f(\chi(u))$, and

(vi) $\gamma(u) = \bar{U}\bar{H}u + o(|u|) = \bar{U}\nabla L_{\mathcal{U}}(u;0) + o(|u|)$.

Proof. Since $v(u)$ is C^2, so is $\chi(u)$. Because $Jv(0) = 0$, $J\chi(0) = \bar{U}$, and, since $Jv(u)$ is C^1, $J\chi(u)$ is the same and item (i) follows. Items (ii) and (iii) follow from expansions of $L_{\mathcal{U}}$ and its gradient, since $L_{\mathcal{U}}(0;0) = f(\bar{x})$, $\nabla L_{\mathcal{U}}(0;0) = 0$, and $\bar{H} = \nabla^2 L_{\mathcal{U}}(0;0)$. Since $J\chi(u)$ is a basis for $\mathcal{V}(\chi(u))^\perp$, Theorem 3.4 in [MS02] with $\mathcal{B}_{\mathcal{U}}(u) = J\chi(u)$ gives item (iv) as well as the relation

$$s(u) \in \mathrm{ri}\partial f(\chi(u)) \quad \text{where } s(u) := J\chi(u)[J\chi(u)^\top J\chi(u)]^{-1}\nabla L_{\mathcal{U}}(u;0).$$

Theorem 3.3 in [MS02] implies that

$$J\chi(u)^\top g = \nabla L_{\mathcal{U}}(u;0) \quad \text{for all } g \in \partial f(\chi(u)).$$

Thus, the $\mathcal{V}(\chi(u))^\perp$-component of any such g does not depend on g and the minimization in (1), the definition of $\gamma(u)$, only minimizes the $\mathcal{V}(\chi(u))$-component of g. Now, since $s(u) \in \partial f(\chi(u))$ has a zero $\mathcal{V}(\chi(u))$-component, $\gamma(u) = s(u)$. Item (v) then follows from item (i), because $\bar{U}$ has full column rank. The expression for $J\chi(u)$ in item (i) implies that

$J\chi(u)[J\chi(u)^\top J\chi(u)]^{-1} = \bar{U} + O(|u|)$. The final result follows, using the expression for $\gamma(u)$ in item (v) and item (iii). □

The basic algorithm idea in this paper is to minimize f by minimizing the C^2 function $L_{\mathcal{U}}(u;0)$ with a Newton or quasi-Newton method. The difficulty with this approach is that $L_{\mathcal{U}}$ and associated quantities in Lemma 3 depending on u are unknown. Our approximation/decomposition algorithm defined later addresses this issue by estimating $(\chi(u), \gamma(u))$ in a manner such that the track parameter u depends on a prox-center x and a prox-parameter μ as discussed next.

2.3 Relating proximal points to primal track points

Our $\mathcal{VU}$-space decomposition algorithm defined in Section 4 below approximates primal track points by approximating equivalent proximal points.

Given a positive scalar parameter μ, the proximal point function depending on f, is defined by

$$p_\mu(x) := \text{argmin}_{p\in\mathbb{R}^n}\{f(p) + \frac{1}{2}\mu|p-x|^2\} \quad \text{for } x \in \mathbb{R}^n.$$

In our development we use the following properties, resulting from the above definition and [Roc76, Prop.1(c)]:

$$\begin{array}{ll} (i) & g_\mu(x) := \mu(x - p_\mu(x)) \in \partial f(p_\mu(x)), \textit{and} \\ (ii) & \text{if } \bar{x} \text{ minimizes } f \text{ then } p_\mu(\bar{x}) = \bar{x} \text{ and } |p_\mu(x) - \bar{x}|^2 \le |x-\bar{x}|^2 - |x - p_\mu(x)|^2. \end{array} \tag{3}$$

The following result, showing that primal tracks attract proximal points, constitutes a fundamental link between proximal point theory and $\mathcal{VU}$-theory. Since here we let μ vary, via possible dependence on x, this result is an extension of the fixed μ results in [MS02, Theorems 5.1 and 5.2]. It also can be seen as a convex version of Theorem 3.5 in [MS05].

Theorem 4 *Let $\chi(u)$ be a primal track leading to a minimizer $\bar{x} \in \mathbb{R}^n$, as described in Definition 1. Suppose $0 \in \text{ri}\partial f(\bar{x})$ and for all x close enough to $\bar{x}$, $\mu = \mu(x) > 0$ with $\mu(x)|x-\bar{x}| \to 0$ as $x \to \bar{x}$.*

Then, for all x close enough to $\bar{x}$,

$$p_\mu(x) = \chi(u_\mu(x)) = \bar{x} + u_\mu(x) \oplus v(u_\mu(x)) \textit{ where } u_\mu(x) := (p_\mu(x) - \bar{x})_{\mathcal{U}}$$

and $u_\mu(x) \to 0$ as $x \to \bar{x}$.

Proof. For x close enough to $\bar{x}$, we write its proximal point using $\mathcal{VU}$ coordinates: $p_\mu(x) = \bar{x} + u_\mu(x) \oplus v_{p_\mu}(x)$ where $u_\mu(x) = (p_\mu(x) - \bar{x})_{\mathcal{U}}$ and $v_{p_\mu}(x) := (p_\mu(x) - \bar{x})_{\mathcal{V}}$. By Property (3)(ii), $|p_\mu(x) - \bar{x}| \le |x - \bar{x}|$, so $u_\mu(x) \to 0$ as $x \to \bar{x}$. Furthermore, $\mu(x)|p_\mu(x) - \bar{x}| \le \mu(x)|x - \bar{x}|$ and, since, by assumption, $\mu(x)|x-\bar{x}| \to 0$ as $x \to \bar{x}$, we have that $\mu(x)(\bar{x} - x)_{\mathcal{V}}$, $\mu(x)v_{p_\mu}(x)$, and $\mu(x)u_\mu(x)$ all converge to zero.

Then (2) implies that $\mu(x)v(u_\mu(x)) \to 0$ as $x \to \bar{x}$. As a result, the function $\omega_\mu : \mathbb{R}^n \to \mathbb{R}^{dim\,\mathcal{V}}$ defined by

$$\omega_\mu(x) := \mu(x)(\bar{x} - x)_\mathcal{V} - \frac{\mu(x)}{2}\Big(v(u_\mu(x)) + v_{p_\mu}(x)\Big)$$

converges to 0 as $x \to \bar{x}$. Since $0 \in \mathrm{ri}\partial f(\bar{x})$, the interior of $\partial f(\bar{x})$ relative to its affine hull, a manifold that is parallel to $\mathcal{V}$ (cf. [LOS00, Prop. 2.2]), we obtain that

$$\omega := 0 \oplus \omega_\mu(x) \in \mathrm{ri}\partial f(\bar{x}) \quad \text{for } x \text{ close enough to } \bar{x}.$$

Thus, from the definition of $L_\mathcal{U}$ with $(u, \bar{g}) = (u_\mu(x), \omega) \in \mathbb{R}^{dim\,\mathcal{U}} \times \mathrm{ri}\partial f(\bar{x})$ and Definition 1,

$$L_\mathcal{U}\Big(u_\mu(x); \omega_\mu(x)\Big) = f(\chi(u_\mu(x))) - \omega^\top \bar{V} v(u_\mu(x)).$$

Since $v_{p_\mu}(x) \in \mathbb{R}^{dim\,\mathcal{V}}$, $L_\mathcal{U}\Big(u_\mu(x); \omega_\mu(x)\Big) \le f(\bar{x} + u_\mu(x) \oplus v_{p_\mu}(x)) - \omega^\top \bar{V} v_{p_\mu}(x)$. As a result,

$$f(\chi(u_\mu(x))) - \omega^\top \bar{V} v(u_\mu(x)) \le f(p_\mu(x)) - \omega^\top \bar{V} v_{p_\mu}(x). \tag{4}$$

By the definition of the proximal point mapping,

$$f(p_\mu(x)) + \frac{\mu}{2}|p_\mu(x) - x|^2 \le f(\chi(u_\mu(x))) + \frac{\mu}{2}|\chi(u_\mu(x)) - x|^2. \tag{5}$$

Combining the two inequalities above yields, after rearrangement of terms,

$$0 \le \frac{\mu}{2}\Big(|\chi(u_\mu(x)) - x|^2 - |p_\mu(x) - x|^2\Big) + \omega^\top \bar{V}(v(u_\mu(x)) - v_{p_\mu}(x)). \tag{6}$$

We now show that the inequality above is in fact an equality. To abbreviate notation, we drop the argument "(x)" in $u_\mu(x)$, $p_\mu(x)$, $v(u_\mu(x))$, $v_{p_\mu}(x)$, and $\omega_\mu(x)$, and write instead u_μ, p_μ, $v(u_\mu)$, v_{p_μ}, and ω_μ. First we expand the leading difference of squares term in (6) and use the fact that $\chi(u_\mu)$ and p_μ have the same $\mathcal{U}$-component:

$$\begin{aligned}
|\chi(u_\mu) - x|^2 - |p_\mu - x|^2 &= (\chi(u_\mu) - p_\mu)^\top(\chi(u_\mu) + p_\mu - 2x) \\
&= \Big(\bar{V}(v(u_\mu) - v_{p_\mu})\Big)^\top\Big(\bar{V}(\chi(u_\mu) + p_\mu - 2x)_\mathcal{V}\Big) \\
&= (v(u_\mu) - v_{p_\mu})^\top \bar{V}^\top \bar{V}\Big(v(u_\mu) + v_{p_\mu} - 2(x - \bar{x})_\mathcal{V}\Big) \\
&= v(u_\mu)^\top \bar{V}^\top \bar{V} v(u_\mu) - v_{p_\mu}^\top \bar{V}^\top \bar{V} v_{p_\mu} + 2(v(u_\mu) - v_{p_\mu})^\top \bar{V}^\top \bar{V}(x - \bar{x})_\mathcal{V}.
\end{aligned}$$

Then

$$\frac{\mu}{2}\Big(|\chi(u_\mu) - x|^2 - |p_\mu - x|^2\Big) = \frac{\mu}{2}\Big(|\bar{V}v(u_\mu)|^2 - |\bar{V}v_{p_\mu}|^2\Big) - \mu(v(u_\mu) - v_{p_\mu})^\top \bar{V}^\top \bar{V}(\bar{x} - x)_\mathcal{V}.$$

Now, since $\bar{V}^\top\omega = \bar{V}^\top\bar{V}\omega_\mu$, we use the definition of ω_μ to write the second right hand side term in (6) as follows:

$$\begin{aligned}\omega^\top\bar{V}(v(u_\mu)-v_{p_\mu}) &= (v(u_\mu)-v_{p_\mu})^\top\bar{V}^\top\omega = (v(u_\mu)-v_{p_\mu})^\top\bar{V}^\top\bar{V}\omega_\mu\\ &= (v(u_\mu)-v_{p_\mu})^\top\bar{V}^\top\bar{V}\Big(\mu(\bar{x}-x)_{\mathcal{V}}-\frac{\mu}{2}(v(u_\mu)+v_{p_\mu})\Big)\\ &= \mu(v(u_\mu)-v_{p_\mu})^\top\bar{V}^\top\bar{V}(\bar{x}-x)_{\mathcal{V}}-\frac{\mu}{2}\Big(v(u_\mu)^\top\bar{V}^\top\bar{V}v(u_\mu)-v_{p_\mu}^\top\bar{V}^\top\bar{V}v_{p_\mu}\Big).\end{aligned}$$

Using these expressions in the right hand side in (6), we obtain that (6) holds with equality. Since the inequality in (6) cannot be strict, we deduce that neither the inequality in (4) nor the one in (5) can be strictly satisfied. In particular, since $p_\mu(x)$ is unique, from (5) we obtain that $p_\mu(x) = \chi(u_\mu(x))$, i.e., that $v_{p_\mu}(x) = v(u_\mu(x))$. □

When μ is fixed, the above conclusion is extended to prox-regular and -bounded functions in [MS05] and, with the addition of C^2-substructure, in [Har03] and [MS04].

3 Approximating primal-dual track points

It is known that a sequence of null steps from a bundle mechanism can approximate a proximal point with any desired accuracy. For our $\mathcal{VU}$-algorithm, when a primal-dual track exists, we also need to approximate points on the dual trajectory $\gamma(u)$ defined in (1) and basis matrices for the corresponding subspaces $\mathcal{V}(\chi(u))^\perp$. We now show that a bundle subroutine can provide such primal-dual approximations via the solution of two quadratic programming problems, denoted below by χ-QP and γ-QP.

3.1 The bundle subroutine

Given a tolerance $\sigma \in (0, 1/2]$, a prox-parameter $\mu > 0$ and a prox-center $x \in \mathbb{R}^n$, to find a σ-approximation of $p_\mu(x)$, our bundle subroutine accumulates information from past points y_i in the form

$$\Big\{\big(f(y_i)\,, g_i \in \partial f(y_i)\big)\Big\}_{i\in\mathcal{B}},$$

where $\mathcal{B}$ is some index set containing an index j such that $y_j = x$. A handier representation for this data set is given by introducing the linearization errors

$$e_i := e(x, y_i) := f(x) - f(y_i) - g_i^\top(x - y_i) \quad \text{for } i \in \mathcal{B},$$

which are nonnegative due to the convexity of f. Also, since

$$f(z) \geq f(y_i) + g_i^\top(z - y_i) \quad \text{for all } z \in \mathbb{R}^n,$$

adding $0 = e_i - e_i$ to this inequality gives

$$f(z) \geq f(x) + g_i^\top(z - x) - e_i \quad \text{for all } z \in \mathbb{R}^n, \tag{7}$$

which means that each g_i is an e_i-subgradient of f at x, i.e., $g_i \in \partial_{e_i} f(x)$ for $i \in \mathcal{B}$. The corresponding bundle of information

$$\Big\{(e_i\,, g_i \in \partial_{e_i} f(x))\Big\}_{i\in\mathcal{B}}$$

is used at each iteration to define a *V-model* underestimating f via the cutting-plane function

$$\varphi(z) := f(x) + \max_{i\in\mathcal{B}} \{-e_i + g_i^\top (z-x)\} \quad \text{for } z \in \mathbb{R}^n.$$

A pure cutting-plane algorithm [CG59], [Kel60] minimizes the polyhedral function φ to define its next iterate. When $\mathcal{V} \neq \mathbb{R}^n$, this method can converge extremely slowly and become unstable, essentially due to its over-estimation of the dimension of $\mathcal{V}$. Bundle methods, [HUL93, BGLS03], are stabilized cutting-plane algorithms. In their proximal form they employ a quadratic term, depending on μ, which is added to the model function. In the bundle context, the prox-parameter μ can be thought of as a "tightness" parameter where larger values give a more compact bundle (see the expression for $\hat{p}$ in (9) below).

To approximate a proximal point we solve a first quadratic programming subproblem χ-QP, which has the following form and properties; see [BGLS03, Lemma 9.8]:
The problem

$$\min\Big\{r + \frac{1}{2}\mu|p-x|^2 : (r,p) \in \mathbb{R}^{1+n}\,, r \geq f(x) - e_i + g_i^\top (p-x) \quad \text{for all } i \in \mathcal{B}\Big\} \qquad (\chi\text{-QP})$$

has a dual

$$\min\left\{\frac{1}{2\mu}\Big|\sum_{i\in\mathcal{B}} \alpha_i g_i\Big|^2 + \sum_{i\in\mathcal{B}} e_i\alpha_i : \alpha_i \geq 0 \quad \text{for } i \in \mathcal{B},\ \sum_{i\in\mathcal{B}} \alpha_i = 1\right\}. \qquad (8)$$

Their respective solutions, denoted by $(\hat{r},\hat{p})$ and $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_{|\mathcal{B}|})$, satisfy

$$\hat{r} = \varphi(\hat{p}) \quad \text{and} \quad \hat{p} = x - \frac{1}{\mu}\hat{g} \qquad \text{where} \quad \hat{g} := \sum_{i\in\mathcal{B}} \hat{\alpha}_i g_i. \qquad (9)$$

In addition, $\hat{\alpha}_i = 0$ for all $i \in \mathcal{B}$ such that $\hat{r} > f(x) - e_i + g_i^\top(\hat{p}-x)$ and

$$\varphi(\hat{p}) = f(x) + \sum_{i\in\mathcal{B}} \hat{\alpha}_i \left(-e_i + g_i^\top(\hat{p}-x)\right) = f(x) - \sum_{i\in\mathcal{B}} \hat{\alpha}_i e_i - \frac{1}{\mu}|\hat{g}|^2. \qquad (10)$$

For convenience, in the sequel we denote the output of these calculations by

$$(\hat{p},\hat{r}) = \chi\text{-QP}\Big(\mu, x, \{(e_i,g_i)\}_{i\in\mathcal{B}}\Big).$$

The vector $\hat{p}$ is an estimate of a proximal point and, hence, approximates a primal track point when the latter exists. To proceed further we define new data, corresponding to a new index i_+, by letting $y_{i_+} := \hat{p}$ and computing $f(\hat{p})$ and $g_{i_+} \in \partial f(\hat{p})$. Note that since $f(\hat{p})$ is available, we can compute the V-model accuracy measure at $\hat{p}$ defined by

$$\hat{\varepsilon} := f(\hat{p}) - \varphi(\hat{p}) = f(\hat{p}) - \hat{r}.$$

An approximate dual track point, denoted by $\hat{s}$, is constructed by solving a second quadratic problem, that depends on a new index set

$$\hat{\mathcal{B}} := \Big\{ i \in \mathcal{B} : \hat{r} = f(x) - e_i + g_i^\top (\hat{p} - x) \Big\} \cup \Big\{ i_+ \Big\}. \tag{11}$$

The second quadratic programming problem

$$\min \left\{ r + \frac{1}{2}|p - x|^2 : (r, p) \in \mathbb{R}^{1+n}, r \geq g_i^\top (p - x) \quad \text{for all } i \in \hat{\mathcal{B}} \right\} \tag{γ-QP}$$

has a dual problem similar to (8), but without linearization error terms:

$$\min \left\{ \frac{1}{2} \Big| \sum_{i \in \hat{\mathcal{B}}} \alpha_i g_i \Big|^2 : \alpha_i \geq 0 \quad \text{for } i \in \hat{\mathcal{B}}, \sum_{i \in \hat{\mathcal{B}}} \alpha_i = 1 \right\}.$$

Similar to (9), the respective solutions, denoted by $(\bar{r}, \bar{p})$ and $\bar{\alpha}$, satisfy

$$\bar{p} - x = -\hat{s} \quad \text{where} \quad \hat{s} := \sum_{i \in \hat{\mathcal{B}}} \bar{\alpha}_i g_i. \tag{12}$$

Note that $\hat{s}$ is the vector with smallest Euclidean norm in the convex hull of $\{g_i : i \in \hat{\mathcal{B}}\}$. Lemma 5 below will show that $\hat{s}$ is an $\hat{\varepsilon}$-subgradient of f at $\hat{p}$. Furthermore, a by-product of the above minimization is a basis matrix $[\hat{U}\hat{V}]$ for $\mathbb{R}^n$ such that $\hat{U}$ has orthonormal columns and $\hat{V}^\top \hat{s} = 0$. Thus, if $p_\mu(x)$ is a primal track point $\chi(u)$ approximated by $\hat{p}$, then the convex hull of $\{g_i : i \in \hat{\mathcal{B}}\}$ approximates $\partial f(\chi(u))$, so from (1) the corresponding $\gamma(u)$ is estimated by $\hat{s}$, and $\hat{U}$ approximates a basis matrix for $\mathcal{V}(\chi(u))^\perp$. See also Definition 1(ii) and Lemma 3(iv) and (v) for this motivation. The matrix construction is as follows: Let a (nonempty) active index set be defined by $\hat{\mathcal{B}}_{act} := \{i \in \hat{\mathcal{B}} : \bar{r} = g_i^\top (\bar{p} - x)\}$. Then, from (12), $\bar{r} = -g_i^\top \hat{s}$ for all $i \in \hat{\mathcal{B}}_{act}$, so

$$(g_i - g_\ell)^\top \hat{s} = 0 \tag{13}$$

for all such i and for a fixed $\ell \in \hat{\mathcal{B}}_{act}$. Define a full column rank matrix $\hat{V}$ by choosing the largest number of indices i satisfying (13) such that the corresponding vectors $g_i - g_\ell$ are linearly independent and by letting these vectors be the columns of $\hat{V}$. Then let $\hat{U}$ be a matrix whose columns form an orthonormal basis for the null-space of $\hat{V}^\top$ with $\hat{U} = I$ if $\hat{V}$ is vacuous.

For convenience, in the sequel we denote the output from these calculations by

$$\Big(\hat{s}, \hat{U}\Big) = \gamma\text{-QP}\Big(\{g_i\}_{i \in \hat{\mathcal{B}}}\Big).$$

The bundle subprocedure is terminated and $\hat{p}$ is declared to be a σ-approximation of $p_\mu(x)$ if

$$\hat{\varepsilon} \le \frac{\sigma}{\mu}|\hat{s}|^2. \tag{14}$$

Otherwise, $\mathcal{B}$ above is replaced by $\hat{\mathcal{B}}$ and new iterate data are computed by solving updated subproblems (χ-QP) and (γ-QP). This update, appending (e_{i_+}, g_{i_+}) to active data at the previous (χ-QP) solution, ensures convergence to a minimizing point in case of nontermination; see Theorem 8 in Section 5 below.

Lemma 5 *Each iteration of the above bundle subprocedure, with output data*

$$\begin{array}{rcl} (\hat{p}, \hat{r}) & = & \chi\text{-QP}\Big(\mu, x, \{(e_i, g_i)\}_{i\in\mathcal{B}}\Big), \\ \Big(\hat{s}, \hat{U}\Big) & = & \gamma\text{-QP}\Big(\{g_i\}_{i\in\hat{\mathcal{B}}}\Big), \end{array}$$

and $\hat{\varepsilon} = f(\hat{p}) - \hat{r}$ satisfies the following:

(i) each g_i for $i \in \hat{\mathcal{B}}$ is an $\hat{\varepsilon}$-subgradient of f at $\hat{p}$;

(ii) $\hat{s}$ is an $\hat{\varepsilon}$-subgradient of f at $\hat{p}$;

(iii) $\mu|\hat{p} - p_\mu(x)|^2 \le \hat{\varepsilon}$;

(iv) $\hat{s} = \hat{U}\hat{U}^\top\hat{s}$, with $\hat{s} = 0$ if $\hat{U}$ is vacuous;

(v) $|\hat{s}| \le |\hat{g}|$ where $\hat{g} = \mu(x - \hat{p})$;

In addition, for any parameter $m \in (0, 1)$, satisfaction of (14) implies

$$f(\hat{p}) - f(x) \le -\frac{m}{2\mu}|\hat{g}|^2. \tag{15}$$

Proof. Since $g_{i_+} \in \partial f(\hat{p})$ and $\hat{\varepsilon} = f(\hat{p}) - \varphi(\hat{p}) \ge 0$, $g_{i_+} \in \partial_{\hat{\varepsilon}} f(\hat{p})$, so the result of item (i) holds for $i = i_+$. From the definitions of $\hat{p}$, $\hat{r}$ and $\hat{\mathcal{B}}$ we have that for all $i \ne i_+$ in $\hat{\mathcal{B}}$

$$\varphi(\hat{p}) = \hat{r} = f(x) - e_i + g_i^\top(\hat{p} - x),$$

so for all such i

$$\hat{\varepsilon} = f(\hat{p}) - \varphi(\hat{p}) = f(\hat{p}) - f(x) + e_i - g_i^\top(\hat{p} - x).$$

Adding this result to (7) gives

$$f(z) \ge f(\hat{p}) + g_i^\top(z - \hat{p}) - \hat{\varepsilon} \quad \text{for all } z \in \mathbb{R}^n, \tag{16}$$

and completes the proof of item (i).
Now item (ii) follows by multiplying each inequality in (16) by its corresponding multiplier $\bar{\alpha}_i \geq 0$, summing these results and then using the definition of $\hat{s}$ from (12) and the fact that these multipliers sum to one.
In a similar manner, this time using the multipliers $\hat{\alpha}_i$ that solve dual problem (8) and define $\hat{g}$ in (9), together with $\hat{\alpha}_{i_+} := 0$, we obtain the result that $\hat{g} \in \partial_{\hat{\varepsilon}} f(\hat{p})$. This fact combined with the Property (3)(i) result $g_\mu(x) = \mu(x - p_\mu(x)) \in \partial f(p_\mu(x))$ and the convexity of f gives

$$f(\hat{p}) + \hat{g}^\top (p_\mu(x) - \hat{p}) - \hat{\varepsilon} \leq f(p_\mu(x)) \leq f(\hat{p}) - g_\mu(x)^\top (\hat{p} - p_\mu(x)),$$

so

$$(\hat{g} - g_\mu(x))^\top (p_\mu(x) - \hat{p}) - \hat{\varepsilon} \leq 0.$$

Then, since the expression for $\hat{g}$ from (9) written in the form

$$\hat{g} = -\mu(\hat{p} - x) \tag{17}$$

combined with the definition of $g_\mu(x)$ from Property (3)(i) implies that $\hat{g} - g_\mu(x) = \mu(p_\mu(x) - \hat{p})$, we obtain satisfaction of item (iii).
Item (iv) follows from the definitions of $\hat{V}$ and $\hat{U}$ and (13), because these items imply the identity $I = \hat{U}\hat{U}^\top + \hat{V}[\hat{V}^\top \hat{V}]^{-1} \hat{V}^\top$ and the result that $\hat{V}^\top \hat{s} = 0$.
Item (v) follows from the minimum norm property of $\hat{s}$, because (9), (8), (17) and the definition of $\hat{\mathcal{B}}$ imply that $\mu(x - \hat{p}) = \hat{g}$ is in the convex hull of $\{g_i : i \in \hat{\mathcal{B}}\}$.
To show that for any $m \in (0, 1)$ condition (15) holds when (14) holds, first note that, since $\sigma \leq 1/2$, we have $\sigma \leq 1 - \frac{m}{2}$. Thus if (14) holds then $\hat{\varepsilon} \leq [(1 - \frac{m}{2})/\mu]|\hat{s}|^2$. This inequality together with the definition of $\hat{\varepsilon}$, (10) and the nonnegativity of $\bar{\alpha}_i e_i$ gives

$$\begin{aligned} f(\hat{p}) - f(x) &= \hat{\varepsilon} + \varphi(\hat{p}) - f(x) \\ &= \hat{\varepsilon} - \sum_{i \in \hat{\mathcal{B}}} \bar{\alpha}_i e_i - \frac{1}{\mu}|\hat{g}|^2 \\ &\leq [(1 - \tfrac{m}{2})/\mu]|\hat{s}|^2 - \frac{1}{\mu}|\hat{g}|^2. \end{aligned}$$

Finally, combining this inequality with item (v) gives (15). □

In bundle terminology, (15) corresponds to declaring $\hat{p}$ to be a "serious step" rather than a "null step"; see [HUL93, Ch. XIV-XV] for more details.

The main algorithm depending on the above bundle subprocedure is defined next.

4 The $\mathcal{VU}$-algorithm

Now we consider an algorithm depending on the $\mathcal{VU}$-theory outlined in Section 2 and the potential primal-dual track point approximations from Section 3. When the tracks exist, the

algorithm moves approximately along the primal track by making a $\mathcal{U}$-Newton predictor step followed by a corrector step, or $\mathcal{V}$-step, coming from a bundle subroutine proximal point estimate.

Each $\mathcal{U}$-step is an approximate Newton-step for minimizing $L_{\mathcal{U}}(u;0)$. The k^{th} one corresponds to a certain u and depends on:

– an ε_k-subgradient s_k approximating $\gamma(u) = \bar{U}\nabla L_{\mathcal{U}}(u;0) + o(|u|)$,

– a matrix U_k approximating a basis for $\mathcal{V}(\chi(u))^\top$, and

– a matrix H_k approximating $\nabla^2 L_{\mathcal{U}}(u;0)$ (if relevant second order information is not available, a quasi-Newton method could be used).

This $\mathcal{U}$-step is followed by the $\mathcal{V}$-step numbered $k+1$. The sum of these two steps gives what we call a "candidate" primal track point p^c_{k+1}. This candidate is declared "good enough" if it satisfies an f-value descent condition that is essentially testing for descent of $L_{\mathcal{U}}$; see (18) below. To deal with "bad" candidates, and ensure convergence to some minimizer from any starting point, a possible line search is included. Nonsatisfaction of the descent condition results in a second bundle subroutine run in a major iteration.

Algorithm 6

Initialization. Choose positive parameters $\varepsilon\,, \underline{\mu}$ and m with $m < 1$. Let $p_0 \in \mathbb{R}^n$ and $g_0 \in \partial f(p_0)$, respectively, be an initial point and subgradient. Also, let U_0 be a matrix with orthonormal n-dimensional columns estimating an optimal $\mathcal{U}$-basis. Set $s_0 := g_0$ and $k := 0$.

Stopping test. Stop if $|s_k|^2 \leq \varepsilon$.

$\mathcal{U}$-Hessian. Choose an $n_k \times n_k$ positive definite matrix H_k, where n_k is the number of columns of U_k.

$\mathcal{U}$-Step. Compute an approximate $\mathcal{U}$-Newton step by solving the linear system

$$H_k \Delta u = -U_k^\top s_k \quad \text{for } \Delta u = \Delta u_k \in \mathbb{R}^{n_k}.$$

Set $x^c_{k+1} := p_k + U_k \Delta u_k = p_k - U_k H_k^{-1} U_k^\top s_k$.

Candidate primal-dual track data. Choose $\mu_{k+1} \geq \underline{\mu}$, $\sigma_{k+1} \in (0, 1/2]$, initialize $\mathcal{B}$ and run the following bundle subprocedure with $x = x^c_{k+1}$:

Compute recursively

$$\begin{aligned} (\hat{p}, \hat{r}) &= \chi\text{-QP}\Big(\mu_{k+1}, x, \{(e_i, g_i)\}_{i\in\mathcal{B}}\Big), \\ \hat{\varepsilon} &= f(\hat{p}) - \hat{r}, \quad \hat{\mathcal{B}} \text{ given by (11), and} \\ \Big(\hat{s}, \hat{U}\Big) &= \gamma\text{-QP}\Big(\{g_i\}_{i\in\hat{\mathcal{B}}}\Big) \end{aligned}$$

until satisfaction of (14) with $(\sigma/\mu) = (\sigma_{k+1}/\mu_{k+1})$.

Then set $\Big(\varepsilon^c_{k+1}, p^c_{k+1}, s^c_{k+1}, U^c_{k+1}\Big) := (\hat{\varepsilon}, \hat{p}, \hat{s}, \hat{U})$.

Candidate evaluation and new iterate data determination. If

$$f(p^c_{k+1}) - f(p_k) \leq -\frac{m}{2\mu_{k+1}} |s^c_{k+1}|^2 \tag{18}$$

then declare a successful candidate and set

$$\Big(x_{k+1}, \varepsilon_{k+1}, p_{k+1}, s_{k+1}, U_{k+1}\Big) := \Big(x^c_{k+1}, \varepsilon^c_{k+1}, p^c_{k+1}, s^c_{k+1}, U^c_{k+1}\Big).$$

Otherwise, execute a line search on the line determined by p_k and p^c_{k+1} to find x_{k+1} thereon satisfying $f(x_{k+1}) \leq f(p_k)$; reinitialize $\mathcal{B}$ and rerun the above bundle subroutine, but with $x = x_{k+1}$, to find new values for $(\hat{\varepsilon}, \hat{p}, \hat{s}, \hat{U})$; then set $\Big(\varepsilon_{k+1}, p_{k+1}, s_{k+1}, U_{k+1}\Big) := (\hat{\varepsilon}, \hat{p}, \hat{s}, \hat{U})$.

Loop. Replace k by $k+1$ and go to *Stopping test.*

□

Remark 7 The following items concerning Algorithm 6 should be noted.

(*i*) An overall stopping test also should be placed inside the bundle procedure. For example, to be consistent, it could be of the following form:
Stop if $max\{|\hat{s}|^2, \frac{\mu}{\sigma}\hat{\varepsilon}\} \leq \varepsilon$.

(*ii*) As for the methods in [FQ96], [LS97b], [MSQ98], [CF99], this algorithm also can be considered as a way to speed up the proximal point method. Instead of setting the next iterate to be an approximation of a proximal point, here x^c_{k+1} is set equal to such an approximation plus a non-null $\mathcal{U}$-step. Note that we do not make a Newton-step in the full space $\mathbb{R}^n$ as is done in the methods in the above four references. This means that we take advantage of nonsmoothness to reduce the dimension of the space for which second derivatives need to be estimated.

(*iii*) To have a $\mathcal{U}$-quasi-Newton method, H_{k+1} should be chosen to be positive definite, close to H_k, and close to satisfying the secant equation

$$H_{k+1}U^\top_{k+1}(p_{k+1} - p_k) = U^\top_{k+1}(s_{k+1} - s_k)\,. \tag{19}$$

A way to deal with the case when the U-matrix changes dimension would be to use a limited memory method [GL89], [LN89], [BNS94], [KON98] which stores a certain number of past difference vectors $p_j - p_{j-1}$ and $s_j - s_{j-1}$ for $j \leq k+1$ and determines H_{k+1} by projecting all of them using U_{k+1}.

(*iv*) In addition to the stated possible line search it first may be beneficial to redefine x^c_{k+1} to be some other point on the half-line from p_k in the direction $U_k\Delta u_k$. This change could be very helpful if a directional derivative underestimate at x^c_{k+1} does not satisfy a Wolfe increase condition [BGLS03, Sec. 3.4]. If not satisfied, then $f(x^c_{k+1}) < f(p_k)$ and safeguarded quadratic extrapolation could be executed to find a point satisfying such a derivative increase condition. Then x^c_{k+1} would be redefined, if necessary, to be a search point found with least f-value. If $n_k = n$ then a more sophisticated line search should be executed to attempt to get $f(x^c_{k+1})$ sufficiently smaller than $f(p_k)$. Information gained from this could go into the choice of μ_{k+1} and initialization of $\mathcal{B}$.

(*v*) For the purpose of early detection of a smooth function with $dim\,\mathcal{U} = n$ a termination test for the bundle procedure, based on (18), could be included inside the bundle procedure as follows: Immediately after x^c_{k+1} and μ_{k+1} are generated the bundle subprocedure would initialize the index i_+ so that $y_{i_+} = x^c_{k+1}$ and g_{i_+} is a subgradient of f at this point, as well as including i_+ in the initialization of $\mathcal{B}$. Then, if

$$f(y_{i_+}) - f(p_k) \leq -\frac{m}{2\mu_{k+1}}|g_{i_+}|^2,$$

the bundle procedure would be terminated and followed by the setting

$$\left(\varepsilon^c_{k+1}, p^c_{k+1}, s^c_{k+1}, U^c_{k+1}\right) := (0, y_{i_+}, g_{i_+}, I),$$

so as to generate a successful candidate satisfying (18) and the subsequent results given below in (20), (21) and (22). This setting does not have a proximal point interpretation, but it does have a primal track one, since when $dim\,\mathcal{U} = n$ the primal track is a full-dimensional ball about $\bar{x}$.

(*vi*) The line search required when p^c_{k+1} is not a successful candidate can be executed very simply by choosing $x_{k+1} := argmin\{f(p_k), f(p^c_{k+1})\}$, a choice that is similar to an ordinary serious step. However, a more expensive line search based on [LM82] has the possibility of finding a better setting for x_{k+1} with better bundle initialization information.

(*vii*) The following reasoning shows that whether or not p^c_{k+1} is a successful candidate

$$f(p_{k+1}) - f(p_k) \leq -\frac{m}{2\mu_{k+1}}|s_{k+1}|^2\,. \tag{20}$$

In the successful case, (20) is the same as (18). Otherwise, (15) and Lemma 5(v) with $(\mu, x, \hat{p}, \hat{s}) = (\mu_{k+1}, x_{k+1}, p_{k+1}, s_{k+1})$ imply that

$$f(p_{k+1}) - f(x_{k+1}) \leq -\frac{m}{2\mu_{k+1}}|\hat{g}|^2 \leq -\frac{m}{2\mu_{k+1}}|s_{k+1}|^2\,,$$

and the line search gives $f(x_{k+1}) \leq f(p_k)$, so (20) also holds in the unsuccessful case. A necessary, but not sufficient, condition for the unsuccessful case is $f(x^c_{k+1}) > f(p_k)$. □

5 Convergence properties of the algorithm

Throughout this section we assume that $\varepsilon = 0$ and that Algorithm 6 does not terminate.

5.1 Global convergence

We first show that if some execution of the bundle procedure defined in Section 3 continues indefinitely, then there is convergence to a minimizer of f.

Theorem 8 *If the bundle procedure does not terminate, i.e., if* (14) *never holds, then the sequence of* $\hat{p}$*-values converges to* $p_\mu(x)$ *and* $p_\mu(x)$ *minimizes* f*. If the procedure terminates with* $\hat{s} = 0$*, then the corresponding* $\hat{p}$ *equals* $p_\mu(x)$ *and minimizes* f*. In both of these cases* $p_\mu(x) - x \in \mathcal{V}(p_\mu(x))$*.*

Proof. The recursion in the bundle subprocedure replacing $\mathcal{B}$ by $\hat{\mathcal{B}}$ (i.e., appending (e_{i_+}, g_{i_+}) to active data at the previous (χ-QP) solution) satisfies conditions (4.7) to (4.9) in [CL93]. By Proposition 4.3 therein, if this procedure does not terminate then it generates an infinite sequence of $\hat{\varepsilon}$-values converging to zero. Since (14) does not hold, the sequence of $|\hat{s}|$-values also converges to 0. Thus, item (iii) in Lemma 5 implies that $\{\hat{p}\} \to p_\mu(x)$. Then the continuity of f and Lemma 5(ii) gives

$$f(z) \geq f(p_\mu(x)) \quad \text{for all } z \in \mathbb{R}^n.$$

The termination case with $\hat{s} = 0$ follows in a similar manner, since (14) implies $\hat{\varepsilon} = 0$ in this case. In either case, by the minimality of $p_\mu(x)$, $0 \in \partial f(p_\mu(x))$. From Property (3)(i), $g_\mu(x) = \mu(x - p_\mu(x)) \in \partial f(p_\mu(x))$, so differencing these two subgradients at $p_\mu(x)$ gives $0 - \mu(x - p_\mu(x)) \in \mathcal{V}(p_\mu(x))$ and the final result follows, since $\mu \neq 0$. □

If either case in Theorem 8 above holds then there is a minimizer $\bar{x} = p_\mu(x)$ such that $\bar{x} - x \in \mathcal{V}(\bar{x})$, so the net move from x to $\bar{x}$ is in a subspace on which the V-approximation (i.e., cutting-plane) aspect of bundling should be very efficient. However, this issue is an open question except in the $n = 1$ case; see [LM82] and [Mif91].

From here on we assume that all executions of the bundle procedure terminate. Then nontermination of Algorithm 6 with $\varepsilon = 0$ implies that infinite sequences

$$\left\{(\mu_k, \sigma_k), \left(x_k^c, \varepsilon_k^c, p_k^c, s_k^c, U_k^c\right), \left(x_k, \varepsilon_k, p_k, s_k, U_k\right)\right\}$$

are generated such that, for all k, $s_k = U_k U_k^\top s_k$ is not the zero vector, (20) holds and, from Lemma 5(ii) and (14) with $(\mu, \sigma, \hat{\varepsilon}, \hat{p}, \hat{s}) = (\mu_k, \sigma_k, \varepsilon_k, p_k, s_k)$,

$$f(p_k) + s_k^\top (z - p_k) - \varepsilon_k \leq f(z) \quad \text{for all } z \in \mathbb{R}^n \quad \text{and} \tag{21}$$

$$\varepsilon_k \leq \frac{\sigma_k}{\mu_k} |s_k|^2. \tag{22}$$

Our next theorem shows minimizing convergence from any initial point without assuming the existence of a primal track.

Theorem 9 *Suppose that the Algorithm 6 sequence $\{\mu_k\}$ is bounded above by $\bar{\mu}$. Then the following hold:*

(i) the sequence $\{f(p_k)\}$ is decreasing and either $\{f(p_k)\} \to -\infty$ or $\{|s_k|\}$ and $\{\varepsilon_k\}$ both converge to 0;

(ii) if f is bounded from below, then any accumulation point of $\{p_k\}$ minimizes f.

Proof. Since $|s_k| \neq 0$, (20) implies that $\{f(p_k)\}$ is decreasing. Suppose $\{f(p_k)\} \not\to -\infty$. Then summing (20) over k and using the fact that $\dfrac{m}{2\mu_k} \geq \dfrac{m}{2\bar{\mu}}$ for all k implies that $\{|s_k|\} \to 0$. Then (22) with $\sigma_k \leq 1/2$ and $\mu_k \geq \underline{\mu} > 0$ implies that $\{\varepsilon_k\} \to 0$, which establishes (i).
Now suppose f is bounded from below and $\bar{p}$ is any accumulation point of $\{p_k\}$. Then, because $\{|s_k|\}$ and $\{\varepsilon_k\}$ converge to 0 by item (i), (21) together with the continuity of f implies that $f(\bar{p}) \leq f(z)$ for all $z \in \mathbb{R}^n$ and (ii) is proved. □

In order to obtain convergence of the whole sequence $\{p_k\}$, we need the concept of a strong minimizer:

Definition 10 *We say that $\bar{x}$ is a* strong minimizer *of f if $0 \in \mathrm{ri}\partial f(\bar{x})$ and the corresponding $\mathcal{U}$-Lagrangian $L_{\mathcal{U}}(u;0)$ has a Hessian at $u = 0$ that is positive definite.* □

A first consequence of $\bar{x}$ being a strong minimizer, following from positive definiteness of $\nabla^2 L_{\mathcal{U}}(0;0)$, is that $u = 0$ is the unique minimizer of $L_{\mathcal{U}}(u;0)$. This result, together with [LO01, Thm. 1], implies that $\bar{x}$ is the unique minimizer of f. In addition, from [Roc76, Thms. 27.1(d)-(f) and 8.4]

$$\text{for any } \alpha \geq f(\bar{x}) \text{ the level set } \{z \in \mathbb{R}^n : f(z) \leq \alpha\} \text{ is compact.} \tag{23}$$

Corollary 11 *Suppose that $\bar{x}$ is a strong minimizer of f, as in Definition 10 and that the Algorithm 6 sequence $\{\mu_k\}$ is bounded above by $\bar{\mu}$. Then $\{p_k\}$ converges to $\bar{x}$. If, in addition, the sequence*

$$\{H_k^{-1}\} \text{ is bounded,} \tag{24}$$

then $\{x_{k+1}^c\}$ and $\{x_k\}$ both converge to $\bar{x}$ and $\{s_{k+1}^c\}$ converges to $0 \in \mathbb{R}^n$.

Proof. Since $\bar{x}$ is a strong minimizer, it is the unique minimizer of f, f is bounded from below by $f(\bar{x})$, and $\{z : f(z) \leq f(p_0)\}$ is compact by (23). Thus, $\{p_k\}$ is bounded and, from item (ii) in Theorem 9, any accumulation point of $\{p_k\}$ is $\bar{x}$.
Now suppose (24) holds. Then, since each U_k has orthonormal columns, the sequence $\{x_{k+1}^c - p_k\} = \{-U_k H_k^{-1} U_k^\top s_k\}$, if infinite, converges to $0 \in \mathbb{R}^n$ because, by Theorem 9(i), $\{s_k\}$ does the same. Thus, $\{x_{k+1}^c\}$ has the same limit as $\{p_k\}$, namely $\bar{x}$. Next consider any infinite subsequence of $\{x_{k+1}\}$ such that $x_{k+1} \neq x_{k+1}^c$. From the line search definition of x_{k+1}

$$f(\bar{x}) \leq f(x_{k+1}) \leq f(p_k) \leq f(p_0),$$

so, as in the above proof for $\{p_k\}$, this subsequence of $\{x_{k+1}\}$ converges $\bar{x}$, which then holds for the entire sequence, because $\{x^c_{k+1}\} \to \bar{x}$. This also implies the final result that $\{s^c_{k+1}\} \to 0$ because, from (15) and Lemma 5(v) with $(x, \hat{p}, \hat{g}, \hat{s}) = (x^c_{k+1}, p^c_{k+1}, g^c_{k+1}, s^c_{k+1})$ and $\mu = \mu_{k+1} \leq \bar{\mu}$,
$f(\bar{x}) - f(x^c_{k+1}) \leq f(p^c_{k+1}) - f(x^c_{k+1}) \leq -\dfrac{m}{2\mu_{k+1}}|g^c_{k+1}|^2 \leq -\dfrac{m}{2\bar{\mu}}|s^c_{k+1}|^2$. □

5.2 Superlinear convergence

For our local convergence analysis we need some technical results that are rather involved. We give their proofs in Appendix A. In this subsection we summarize these intermediate results before giving our main result on rate of convergence.

We first discuss the assumptions needed to show superlinear convergence. The initial set of suppositions is related to existence of primal-dual tracks, and other basic algorithmic requirements:

(S)(a) There exists a primal-dual track $(\chi(u), \gamma(u))$ leading to $(\bar{x}, 0)$, as described in Definition 1, with $\bar{x}$ a strong minimizer, as in Definition 10.

(S)(b) The Algorithm 6 sequences satisfy the following:

- **(i)** $\{\mu_k\}$ is bounded above by $\bar{\mu}$;
- **(ii)** $\{H_k^{-1}\}$ is bounded, i.e., (24) holds;
- **(iii)** $\{\sigma_k\} \to 0$ as $k \to \infty$, for example, by choosing $\sigma_{k+1} \leq 1/(k+2)$.

Combining these assumptions with Theorem 4 and Corollary 11 in a straightforward manner gives the following primal track related results:

Lemma 12 *Suppose that* **(S)(a), (S)(b)(i) and (ii)** *hold. Let $u_\mu(x)$ be the function defined in Theorem 4 and, for all k sufficiently large, let $u_k := u_{\mu_k}(x_k)$ and $u^c_{k+1} := u_{\mu_{k+1}}(x^c_{k+1})$. Then for all k sufficiently large*

(i) $p_\mu(x) = \bar{x} + u_\mu(x) \oplus v(u_\mu(x))$ for (μ, x) equal to (μ_k, x_k) and to (μ_{k+1}, x^c_{k+1}) and

(ii) $\{u_k\}$ and $\{u^c_{k+1}\}$ both converge to $0 \in \mathbb{R}^{\dim \mathcal{U}}$. □

In addition to **(S)(a)-(b)**, we make the following algorithm assumptions stating how well the dual track points $\gamma(u)$ and corresponding $\mathcal{U}$-basis and -Hessian matrices need to be approximated. Since we are interested in a step from p_k, depending on x_k, to p^c_{k+1} via x^c_{k+1}, we introduce the following notation to be used with (μ, x) equal to (μ_k, x_k) and to (μ_{k+1}, x^c_{k+1}) as in Lemma 12(i):

$$\left(\hat{p}_\mu(x), \hat{s}_\mu(x)\right) \;:=\; (\hat{p}, \hat{s}), \quad \text{output from } \chi\text{-QP}/\gamma\text{-QP satisfying (14).} \tag{25}$$

(S)(c) For all k sufficiently large and for (μ, x) equal to (μ_k, x_k) and to (μ_{k+1}, x_{k+1}^c) the γ-QP output $\hat{s}_\mu(x)$, correspondingly equal to s_k and to s_{k+1}^c, satisfies

$$\hat{s}_\mu(x) - \gamma(u_\mu(x)) = o(|\hat{s}_\mu(x)|) + o(|u_\mu(x)|) \tag{26}$$

where $u_\mu(x)$ is defined in Theorem 4.

(S)(d) For all k sufficiently large $n_k = dim\, \mathcal{U}$ and there exists a $dim\, \mathcal{U} \times dim\, \mathcal{U}$ orthogonal matrix Q_k with the corresponding product sequence $\{U_k Q_k^\top\}$ converging to $\bar{U}$.

(S)(e) For all k sufficiently large $[Q_k H_k Q_k^\top - \nabla^2 L_{\mathcal{U}}(0;0)]u_k = o(|s_k|) + o(|u_k|)$ where $u_k = u_{\mu_k}(x_k)$ as in Lemma 12.

Remark 13 Some comments on the assumptions above are in order:

(*i*) If **(S)(b)(iii)** holds and if the approximation of $\gamma(u_\mu(x))$ by $\hat{s}_\mu(x)$ is as good as that of $g_\mu(x)$ by $\hat{g}_\mu(x) := \mu(x - \hat{p}_\mu(x))$ in the sense that

$$\hat{s}_\mu(x) - \gamma(u_\mu(x)) = O(|\hat{g}_\mu(x) - g_\mu(x)|)$$

then (26) in **(S)(c)** holds. This follows, because Lemma 5(iii) and (14) combined with $\hat{g}_\mu(x) - g_\mu(x) = \mu(p_\mu(x) - \hat{p}_\mu(x))$ imply $|\hat{g}_\mu(x) - g_\mu(x)| \leq \sqrt{\sigma}|\hat{s}_\mu(x)|$, and **(S)(b)(iii)** implies $\sigma \to 0$ as $x \to \bar{x}$.

(*ii*) **(S)(d)** allows an accumulation matrix of $\{U_k\}$ to be a basis matrix for $\mathcal{U}$ other than $\bar{U}$, which then equals $\bar{U}Q$ for some orthogonal matrix Q (i.e., satisfying $Q^\top Q = I$). This supposition means that all accumulation matrices of the bounded set $\{U_k\}$ are orthonormal basis matrices for $\mathcal{U}$. The matrix $U_k Q_k^\top$ is intended to approximate a basis matrix for $\mathcal{V}(p_{\mu_k}(x_k))^\perp$ similar to $J\chi(u_k)$ which, by item (i) in Lemma 3, converges to $\bar{U}$ linearly in u_k. **(S)(d)** only asks for convergence with no particular speed thereof.

(*iii*) Assumptions **(S)(d)** and **(e)** are consistent with the quasi-Newton equation (19), because inserting $I = Q_{k+1}^\top Q_{k+1}$ into this equation and multiplying by Q_{k+1} on the left gives the equivalent quasi-Newton equation

$$\left(Q_{k+1} H_{k+1} Q_{k+1}^\top\right)\left(U_{k+1} Q_{k+1}^\top\right)^\top (p_{k+1} - p_k) = \left(U_{k+1} Q_{k+1}^\top\right)^\top (s_{k+1} - s_k)\,. \square$$

The following proposition summarizes some technical results following from our assumptions. These involve accuracy of primal track point approximation and a strengthening of **(S)(b)(iii)** to ensure candidate success. They depend on the definition of $\hat{p}_\mu(x)$ in (25) and that of $u_\mu(x)$ in Theorem 4.

Proposition 14 *Consider the sequences generated by Algorithm 6 with k sufficiently large for the results of Lemma 12 to hold.*

(i) If **(S)(a)-(c)** *hold, then* $\hat{p}_\mu(x) - p_\mu(x) = o(|u_\mu(x)|)$ *for* (μ, x) *equal to* (μ_k, x_k) *and to* (μ_{k+1}, x^c_{k+1}).

(ii) If **(S)(a)-(e)** *hold, then* $p^c_{k+1} - \bar{x} = o(|u_k|)$.

(iii) If **(S)(a)-(e)** *hold and* $\sigma_{k+1} \leq min\{1/(k+2), |s_k|^2/|s_0|^2\}$ *for all* $k \geq 0$, *then* (18) *holds and* $p_{k+1} = p^c_{k+1}$.

Proof. The three stated items correspond, respectively, to Lemmas 17(iii), 18(ii), and 19 in Appendix A. □

We conclude this subsection by giving our main rate of convergence result.

Theorem 15 *Suppose that* **(S)(a)-(e)** *hold. Then for all* k *sufficiently large*

(i) $p^c_{k+1} - \bar{x} = o(|p_k - \bar{x}|)$ *and*

(ii) if, in addition, $\sigma_{k+1} \leq min\{1/(k+2), |s_k|^2/|s_0|^2\}$, *then* $p_{k+1} - \bar{x} = o(|p_k - \bar{x}|)$, *i.e., the sequence* $\{p_k\}$ *converges to* $\bar{x}$ *superlinearly.*

Proof. Since $p_k - \bar{x} = p_k - p_{\mu_k}(x_k) + p_{\mu_k}(x_k) - \bar{x} = p_k - p_{\mu_k}(x_k) + u_k \oplus v(u_k)$, by Lemma 12(i) with $(\mu, x) = (\mu_k, x_k)$, its $\mathcal{U}$-component can be written as $(p_k - \bar{x})_\mathcal{U} = (p_k - p_{\mu_k}(x_k))_\mathcal{U} + u_k$. By Proposition 14(i), $(p_k - p_{\mu_k}(x_k))_\mathcal{U} = o(|u_k|)$, so, $|(p_k - \bar{x})_\mathcal{U}| = |o(|u_k|) + u_k|$. Combining this with Proposition 14(ii) gives the ratio

$$\frac{p^c_{k+1} - \bar{x}}{|(p_k - \bar{x})_\mathcal{U}|} = \frac{o(|u_k|)}{|o(|u_k|) + u_k|}$$

which converges to 0 as $u_k \to 0$. So, $p^c_{k+1} - \bar{x} = o(|(p_k - \bar{x})_\mathcal{U}|)$ and the first result follows, because $|(p_k - \bar{x})_\mathcal{U}| \leq |p_k - \bar{x}|$.
The second result then follows from Proposition 14(iii). □

6 Preliminary numerical results

In order to begin validation of our approach we wrote a MATLAB implementation of a simplified version of Algorithm 6, suitable for objective functions of the form

$$f = \max_{j \in J} f_j \text{ where } J \text{ is finite and each } f_j \text{ is } C^2 \text{ on } \mathbb{R}^n.$$

For each bundled subgradient g_i we use a gradient of f_{j_i} at y_i where j_i is an active index such that $f_{j_i}(y_i) = f(y_i)$. In order to test these max-functions with good second order information we do the following: When $s_k = \hat{s}$ in the algorithm then the matrix H_k is set equal to $U_k^\top(\sum_{i \in \hat{\mathcal{B}}} \bar{\alpha}_i H^i) U_k$ where H^i is the Hessian of f_{j_i} at y_i and the multipliers $\bar{\alpha}_i$ correspond to $\hat{s}$ via (12). Here we are mainly testing to see if primal-dual points and associated basis matrices can be approximated well enough. Future work will deal with less precise $\mathcal{U}$-Hessian

estimation. When this setting of H_k is not positive definite the $n_k \times n_k$ diagonal matrix $\min\{1/k, |s_k|\}I$ is added to it.

For our runs we used the following functions:

– `F2d`, the function in [LS97a], defined for $x \in \mathbb{R}^2$ by $\texttt{F2d}(x) := \max\left\{\frac{1}{2}(x_1^2 + x_2^2) - x_2, x_2\right\}$;

– `F3d-U`ν, four functions of three variables, where $\nu = 3, 2, 1, 0$ denotes the corresponding dimension of the $\mathcal{U}$-subspace. Given $e := (0, 1, 1)^\top$ and four parameter vectors $\beta^\nu \in \mathbb{R}^4$, for $x \in \mathbb{R}^3$

$$\texttt{F3d-U}\nu(x) := \max\left\{\frac{1}{2}(x_1^2 + x_2^2 + 0.1x_3^2) - e^\top x - \beta_1^\nu, x_1^2 - 3x_1 - \beta_2^\nu, x_2 - \beta_3^\nu, x_2 - \beta_4^\nu\right\},$$

where $\beta^3 := (-5.5, 10, 11, 20)$, $\beta^2 := (-5, 10, 0, 10)$, $\beta^1 := (0, 10, 0, 0)$ and $\beta^0 := (0.5, -2, 0, 0)$;

– `MAXQUAD`, the piecewise quadratic function described in [BGLS03, p. 131].

Table 1 shows some additional relevant data for the problems, including the dimensions of $\mathcal{V}$ and $\mathcal{U}$, the optimal values and solutions, and the starting points.

Table 1: Problem data

Name	n	$dim\,\mathcal{V}$	$dim\,\mathcal{U}$	$f(\bar{x})$	$\bar{x}$	Starting point
`F2d`	2	1	1	0.	$(0, 0)$	$\bar{x} + (0.9, 1.9)$
`F3d-U3`	3	0	3	0.	$(0, 1, 10)$	$\bar{x} + (100, 33, -100)$
`F3d-U2`	3	1	2	0.	$(0, 0, 10)$	$\bar{x} + (100, 33, -100)$
`F3d-U1`	3	2	1	0.	$(0, 0, 0)$	$\bar{x} + (100, 33, -100)$
`F3d-U0`	3	3	0	0.	$(1, 0, 0)$	$\bar{x} + (100, 33, -100)$
`MAXQUAD`	10	3	7	-0.8414083	see [BGLS03, p. 131]	all components equal to 1

For comparison purposes, we also solved these problems using N1CV2, the proximal bundle method from [LS97b] (with quadratic programming subproblems solved by the method described in [Kiw86]), available upon request at `http://www-rocq.inria.fr/estime/modulopt/optimization-routines/n1cv2.html`.

The settings of the Algorithm 6 parameters are $\varepsilon := 10^{-8}$, $m := 10^{-1}$ and U_0 equal to the $n \times n$ identity matrix. We take $\sigma_k := 1/(k^2 + 1)$ and μ_k to be a safeguarded version of the *reversal quasi-Newton* scalar update in N1CV2; see [BGLS03, § 9.3.3]. More precisely,

$$\mu_{k+1} := \min\left(10\mu_k, \max\left(\mu_{k+1}^{\text{N1CV2}}, \max\left(\underline{\mu}, 0.1\mu_k\right)\right)\right) \text{ for } k \geq 1,\ \mu_1 := 4,\ \underline{\mu} := 0.01\mu_1, \text{ and}$$

$$\mu_{k+1}^{\text{N1CV2}} := \frac{|g(p_k) - g(p_{k-1}|^2}{(g(p_k) - g(p_{k-1})^\top (p_k - p_{k-1})}.$$

For N1CV2 $\mu_{k+1} := \mu_{k+1}^{\text{N1CV2}}$ with p_k equal to the k^{th} serious step point and $\mu_1 := \dfrac{5|g(p_0)|^2}{|f(p_0)|}$.

The stopping test in N1CV2 is chosen to correspond to the one in Algorithm 6 with the above setting of ε. Both codes use a bundle management strategy that keeps only active elements.

For Algorithm 6, after x_{k+1}^c is generated, $\mathcal{B}$ is initialized by appending an index corresponding to x_{k+1}^c-data to the $\hat{\mathcal{B}}$-set associated with $(\hat{p}, \hat{s}) = (p_k, s_k)$ as defined in Section 3. If p_{k+1}^c is not a successful candidate then the simple setting $x_{k+1} := \text{argmin}\{f(p_k), f(p_{k+1}^c)\}$ is made. Then, if $x_{k+1} = p_k$(p_{k+1}^c, resp.) $\mathcal{B}$ is reinitialized by appending an index corresponding to p_{k+1}^c-data to the $\hat{\mathcal{B}}$-set associated with p_k(p_{k+1}^c, resp.).

Our numerical results are reported in Table 2 below. For each run of both algorithms, we give the total number of evaluations of f and one gradient (and one Hessian in the case of Algorithm 6) and an accuracy measure equal to the number of correct optimal objective value digits after the decimal point.

Table 2: Summary of the results

	F2d		F3d-U3		F3d-U2		F3d-U1		F3d-U0		MAXQUAD	
	#f/g	Acc	#f/g	Acc	#f/g	Acc	#f/g	Acc	#f/g	Acc	#f/g	Acc
ALG 6	20	9	5	16	24	10	15	13	20	12	79	14
N1CV2	30	4	45	6	43	4	31	4	45	6	156	8

For all six functions Algorithm 6 generated final basis matrix approximations having the correct $\mathcal{VU}$ dimensions as given in Table 1.

In order to obtain a good picture of superlinear convergence, for the MAXQUAD function we computed and plotted relative errors $\left\{\left|\frac{f(p_k)-f_{\text{best}}}{f_{\text{best}}}\right|\right\}$, where $f_{\text{best}} = -0.8414083345964012$ is our best known function value. For each algorithm Figure 1 on the next page shows relative errors generated by 79 function evaluations. For Algorithm 6 this number of evaluations generates 6 primal track estimates p_k, while for N1CV2 it gives 25 serious step points p_k. The plots with a logarithmic vertical scale clearly show linear convergence of the p_k sequence for N1CV2 and superlinear behavior for Algorithm 6.

These favorable results demonstrate that it is worthwhile to continue development of the $\mathcal{VU}$-algorithm. This will entail adding line searches and then determining good choices for μ_k (not depending on N1CV2), σ_k and H_k.

Concluding remarks

This paper has provided a minimization algorithm based on combining a V-model of the objective with a U-model of a corresponding subspace Lagrangian. These two basic models are

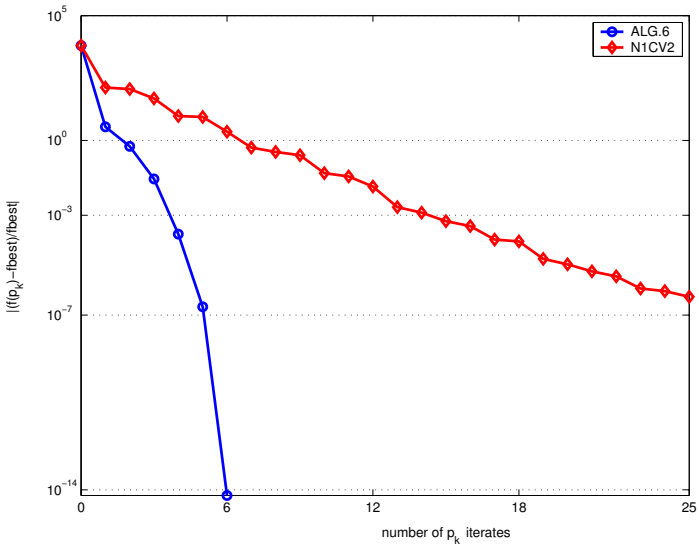

Figure 1: Relative errors for 79 `MAXQUAD` function evaluations

connected both theoretically (by proximal point and primal-dual track theory) and practically (via a bundle subroutine that approximates related primal-dual track points). The method can operate well on the large class of `pdg`-structured objective functions and at least converge for more general convex functions. These are two among the several advantages of a bundle-based approach. Moreover, the algorithm does not need explicit knowledge of existing `pdg`-structure in order to exploit such a framework. It is broadly applicable, as many constraints can be handled with penalty functions and it can be extended for solving large scale problems by aggregating ([Kiw83], [CL93], [HUL93]) its χ-QP-subproblem constraints and/or using a limited memory quasi-Newton method for estimating the $\mathcal{U}$-Hessian; see Remark 7(iii). However, including either one of the last two techniques most certainly will result in the loss of superlinear convergence. Thus, large scale problems with known separable structure should be handled, instead, by decomposition techniques [BGLS03, Ch. 10]. These produce a smaller dimensional, typically nonsmooth, outer objective function whose evaluation separates into independent optimization subproblems that, if necessary, can be solved in parallel with a grid of computational devices. Our algorithm is especially suited for this kind of objective function because it requires only one subgradient value with each function evaluation and the class of pdg-structured functions contains many "infinitely-defined" max-functions. Another way to view this algorithm is as an extension of a quasi-Newton method to the nonsmooth convex case where, via exact penalty functions, nonsmoothness also encompasses constrained problems, for example those with conic functions [LVBL98]. This will be a subject of future work dealing with finding general `pdg`-structured functions satisfying conditions **(S)(c)-(e)**.

Acknowledgements

We thank the referees for beneficial comments.

References

[Aus87] A. Auslender. Numerical methods for nondifferentiable convex optimization. *Math. Programming Stud.*, 30:102–126, 1987.

[BGLS03] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization. Theoretical and Practical Aspects.* Universitext. Springer-Verlag, Berlin, 2003. xiv+423 pp.

[BNS94] R.H. Byrd, J. Nocedal, and R.B. Schnabel. Representations of quasi-Newton matrices and their use in limited-memory methods. *Math. Program.*, 63:129–156, 1994.

[CF99] X. Chen and M. Fukushima. Proximal quasi-Newton methods for nondifferentiable convex optimization. *Math. Program.*, 85(2, Ser. A):313–334, 1999.

[CG59] E. Cheney and A. Goldstein. Newton's method for convex programming and Tchebycheff approximations. *Numerische Mathematik*, 1:253–268, 1959.

[CL93] R. Correa and C. Lemaréchal. Convergence of some algorithms for convex minimization. *Math. Program.*, 62(2):261–275, 1993.

[FQ96] M. Fukushima and L. Qi. A globally and superlinearly convergent algorithm for nonsmooth convex minimization. *SIAM Journal on Optimization*, 6:1106–1120, 1996.

[Fuk84] M. Fukushima. A descent algorithm for nonsmooth convex optimization. *Math. Program.*, 30:163–175, 1984.

[GL89] J.Ch. Gilbert and C. Lemaréchal. Some numerical experiments with variable-storage quasi-Newton algorithms. *Math. Program.*, 45:407–435, 1989.

[Har03] W.L. Hare. *Nonsmooth Optimization with Smooth Substructure.* PhD thesis, Department of Mathematics, Simon Fraser University, 2003. Preprint available at `http://www.cecm.sfu.ca/~whare`.

[HUL93] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms.* Number 305-306 in Grund. der math. Wiss. Springer-Verlag, 1993.

[Kel60] J. E. Kelley. The cutting plane method for solving convex programs. *J. Soc. Indust. Appl. Math.*, 8:703–712, 1960.

[Kiw83] K.C. Kiwiel. An aggregate subgradient method for nonsmooth convex minimization. *Math. Program.*, 27:320–341, 1983.

[Kiw86] K.C. Kiwiel. A method for solving certain quadratic programming problems arising in nonsmooth optimization. *IMA Journal of Numerical Analysis*, 6:137–152, 1986.

[KON98] T.G. Kolda, D.P. O'Leary, and J.L. Nazareth. BFGS with update skipping and varying memory. *SIAM J. on Optimization*, 8:1060–1083, 1998.

[LM82] C. Lemaréchal and R. Mifflin. Global and superlinear convergence of an algorithm for one-dimensional minimization of convex functions. *Math. Program.*, 24:241–256, 1982.

[LN89] D.C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Program.*, 45:503–520, 1989.

[LO01] C. Lemaréchal and F. Oustry. Growth conditions and $\mathcal{U}$-Lagrangians. *Set-Valued Analysis*, 9(1/2):123–129, 2001.

[LOS00] C. Lemaréchal, F. Oustry, and C. Sagastizábal. The $\mathcal{U}$-Lagrangian of a convex function. *Trans. Amer. Math. Soc.*, 352(2):711–729, 2000.

[LS94] C. Lemaréchal and C. Sagastizábal. An approach to variable metric bundle methods. In J. Henry and J-P. Yvon, editors, *Systems Modelling and Optimization*, number 197 in Lecture Notes in Control and Information Sciences, pages 144–162. Springer-Verlag, 1994.

[LS96] C. Lemaréchal and C. Sagastizábal. More than first-order developments of convex functions: primal-dual relations. *Journal of Convex Analysis*, 3(2):1–14, 1996.

[LS97a] C. Lemaréchal and C. Sagastizábal. Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):367–385, 1997.

[LS97b] C. Lemaréchal and C. Sagastizábal. Variable metric bundle methods: from conceptual to implementable forms. *Math. Program., Ser. A*, 76:393–410, 1997.

[LVBL98] M. Lobo, L. Vandenberghe, S. Boyd and H. Lebret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

[Mif91] R. Mifflin. On superlinear convergence in univariate nonsmooth minimization. *Math. Program.*, 49:273–279, 1991.

[Mif96] R. Mifflin. A quasi-second-order proximal bundle algorithm. *Math. Program.*, 73(1):51–72, 1996.

[MM04] S. A. Miller and J. Malick. Connections between $\mathcal{U}$-Lagrangian, Riemannian Newton, and SQP Methods. *Rapport de Recherche INRIA* N^o 5185, May 2004.

[Mor65] J.J. Moreau. Proximité et dualité dans un espace Hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[MS99] R. Mifflin and C. Sagastizábal. $\mathcal{VU}$-decomposition derivatives for convex max-functions. In R. Tichatschke and M. Théra, editors, *Ill-posed Variational Problems and Regularization Techniques*, number 477 in Lecture Notes in Economics and Mathematical Systems, pages 167–186. Springer-Verlag Berlin Heidelberg, 1999.

[MS00a] R. Mifflin and C. Sagastizábal. Functions with primal-dual gradient structure and $\mathcal{U}$-Hessians. In G. Di Pillo and F. Giannessi, editors, *Nonlinear Optimization and Related Topics*, number 36 in Applied Optimization, pages 219–233. Kluwer Academic Publishers B.V., 2000.

[MS00b] R. Mifflin and C. Sagastizábal. On $\mathcal{VU}$-theory for functions with primal-dual gradient structure. *SIAM Journal on Optimization*, 11(2):547–571, 2000.

[MS02] R. Mifflin and C. Sagastizábal. Proximal points are on the fast track. *Journal of Convex Analysis*, 9(2):563–579, 2002.

[MS03] R. Mifflin and C. Sagastizábal. Primal-Dual Gradient Structured Functions: second-order results; links to epi-derivatives and partly smooth functions. *SIAM Journal on Optimization*, 13(4):1174–1194, 2003.

[MS04] R. Mifflin and C. Sagastizábal. $\mathcal{VU}$-Smoothness and Proximal Point Results for Some Nonconvex Functions. *Optimization Methods and Software*, 19(5):463–478, 2004.

[MS05] R. Mifflin and C. Sagastizábal. Relating $\mathcal{U}$-Lagrangians to Second Order Epi-derivatives and Proximal Tracks. *Journal of Convex Analysis*, 12(1):81-93, 2005.

[MSQ98] R. Mifflin, D.F. Sun, and L.Q. Qi. Quasi-Newton bundle-type methods for non-differentiable convex optimization. *SIAM Journal on Optimization*, 8(2):583–603, 1998.

[Ous00] F. Oustry. A second-order bundle method to minimize the maximum eigenvalue function. *Math. Program.*, 89(1, Ser. A):1–33, 2000.

[QC97] L. Qi and X. Chen. A preconditioning proximal Newton method for nondifferentiable convex optimization. *Math. Program.*, 76(3, Ser. B):411–429, 1997.

[RF00] A.I. Rauf and M. Fukushima. Globally convergent BFGS method for nonsmooth convex optimization. *J. Optim. Theory Appl.*, 104(3):539–558, 2000.

[Roc76] R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14:877–898, 1976.

[RW98] R.T. Rockafellar and R.J.-B. Wets. *Variational Analysis*. Number 317 in Grund. der Math. Wiss. Springer-Verlag, 1998.

Appendix A: Technical results for superlinear convergence

The initial set of suppositions, **(S)(a)-(b)**, combined with Lemma 12 give the following primal track related results:

Lemma 16 *Suppose that* **(S)(a), (S)(b)(i) and (ii)** *hold and let* $\{u_k\}$ *and* $\{u^c_{k+1}\}$ *be the zero-convergent sequences from Lemma 12 . Then for all* k *sufficiently large*

(i) $f(p_{\mu_k}(x_k)) - f(p_k) = |p_k - p_{\mu_k}(x_k)|O(|u_k|)$, *and*

(ii) $f(p_{\mu_{k+1}}(x^c_{k+1})) - f(p_{\mu_k}(x_k)) \leq -\frac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + O\left(|u^c_{k+1}|^2\right) + o(|u_k|^2)\,,$
where $\lambda_{min}(\bar{H})$ *is the smallest eigenvalue of* $\bar{H} = \nabla^2 L_{\mathcal{U}}(0;0)$.

Proof. Lemma 12 with $(\mu, x) = (\mu_k, x_k)$ and the definition of $\chi(u)$ imply that $p_{\mu_k}(x_k) = \bar{x} + u_k \oplus v(u_k) = \chi(u_k)$ so, by Lemma 3(v), $\gamma(u_k) \in \partial f(p_{\mu_k}(x_k))$. By convexity of f,

$$f(p_{\mu_k}(x_k)) + \gamma(u_k)^\top (p_k - p_{\mu_k}(x_k)) \leq f(p_k)\,.$$

Thus,

$$f(p_{\mu_k}(x_k)) - f(p_k) \leq |\gamma(u_k)|\ |p_k - p_{\mu_k}(x_k)|,$$

and item (i) then follows because, from Lemma 3 (vi), $\gamma(u_k) = O(|u_k|)$.
Writing Lemma 3(ii) for the primal track points $p_{\mu_{k+1}}(x^c_{k+1})$ and $p_{\mu_k}(x_k)$ yields, respectively, $f(p_{\mu_{k+1}}(x^c_{k+1})) = f(\bar{x}) + \frac{1}{2}(u^c_{k+1})^\top \bar{H} u^c_{k+1} + o(|u^c_{k+1}|^2)$ and $f(p_{\mu_k}(x_k)) = f(\bar{x}) + \frac{1}{2}u_k^\top \bar{H} u_k + o(|u_k|^2)$. Therefore,

$$f(p_{\mu_{k+1}}(x^c_{k+1})) - f(p_{\mu_k}(x_k)) \leq -\frac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 - o(|u_k|^2) + \frac{1}{2}\lambda_{max}(\bar{H})|u^c_{k+1}|^2 + o(|u^c_{k+1}|^2)$$

where $\lambda_{max}(\bar{H})$ is the largest eigenvalue of $\bar{H}$. This implies the inequality in (ii) and completes the proof. □

By means of additional assumptions, **(S)(b)(iii)-(c)**, concerning adequate approximation of dual track points $\gamma(u_k)$, we now show correspondingly accurate approximation of the primal track points $p_{\mu_k}(x_k)$ and give a related result to be used later on for showing candidate success. These results are all in terms of the unknown primal track quantities u_k and u^c_{k+1}, that are useful for rate of convergence analysis, rather than their linearly-related computed quantities s_k and s^c_{k+1}, that are good for rate of computational progress observation.

Lemma 17 *Suppose that* **(S)(a)-(c)** *hold and let* $\{u_k\}$ *and* $\{u^c_{k+1}\}$ *be the zero-convergent sequences from Lemma 12. Then for all* k *sufficiently large and for* (μ, x) *equal to* (μ_k, x_k) *and to* (μ_{k+1}, x^c_{k+1})

(i) $\hat{s}_\mu(x) = O(|u_\mu(x)|)$,

(ii) $\hat{s}_\mu(x) - \gamma(u_\mu(x)) = o(|u_\mu(x)|)$,

(iii) $\hat{p}_\mu(x) - p_\mu(x) = o(|u_\mu(x)|)$, *and*

(iv) if $\sigma_{k+1} = O(|s_k|^2)$ *then* $|p^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1})| = O(|u_k|)O(|u^c_{k+1}|)$.

Proof. For large k and (μ, x) equal to (μ_k, x_k) or to (μ_{k+1}, x^c_{k+1}), (26) in **(S)(c)** and Lemma 3 (vi) with $u = u_\mu(x)$ gives

$$\begin{aligned} \hat{s}_\mu(x) - o(|\hat{s}_\mu(x)|) &= \gamma(u_\mu(x)) + o(|u_\mu(x)|) \\ &= O(|u_\mu(x)|) + o(|u_\mu(x)|). \end{aligned}$$

This implies that item (i) holds, which together with (26) gives item (ii).
In addition, for such (μ, x) values and corresponding σ values, Lemma 5(iii) and(14), and $\mu \geq \underline{\mu}$ give

$$|\hat{p}_\mu(x) - p_\mu(x)|^2 \leq \frac{\sigma}{\mu^2}|\hat{s}_\mu(x)|^2 \leq \frac{\sigma}{\underline{\mu}^2}|\hat{s}_\mu(x)|^2, \tag{27}$$

which together with **(S)(b)(iii)** and item (i) implies the validity of item (iii).
Moreover, for $(\mu, x, \sigma) = (\mu_{k+1}, x^c_{k+1}, \sigma_{k+1})$ (27) becomes

$$|\hat{p}_{\mu_{k+1}}(x^c_{k+1}) - p_{\mu_{k+1}}(x^c_{k+1})|^2 \leq \frac{\sigma_{k+1}}{\underline{\mu}^2}|\hat{s}_{\mu_{k+1}}(x^c_{k+1})|^2,$$

which implies, by item (i) with $u_{\mu_{k+1}}(x^c_{k+1}) = u^c_{k+1}$, that

$$|p^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1})| \leq \frac{\sqrt{\sigma_{k+1}}}{\underline{\mu}} O(|u^c_{k+1}|). \tag{28}$$

Finally, if $\sigma_{k+1} = O(|s_k|^2)$ then, from item (i), this time with $(\mu, x) = (\mu_k, x_k)$ and $u_{\mu_k}(x_k) = u_k$, we have $\sqrt{\sigma_{k+1}} = O(|u_k|)$ and, so, item (iv) follows from (28). □

Next we append suppositions **(S)(d)** and **(e)**, concerning adequate approximation of basis and Hessian matrices, to obtain the main lemma for showing superlinear convergence. Its first part shows that $Q_k U_k^\top \gamma(u_k)$ and its approximant $Q_k U_k^\top s_k$ behave like the $\mathcal{U}$-gradient $\nabla L_{\mathcal{U}}(u_k; 0)$ that, by Lemma 3(iii), equals $\bar{H} u_k + o(|u_k|)$. The second part is concerned with the u-rate of convergence of the candidate data.

Lemma 18 *Suppose that* **(S)(a)-(e)** *hold and let* $\{u_k\}$ *and* $\{u^c_{k+1}\}$ *be the zero-convergent sequences from Lemma 12. Then the sequences* $\{U_k H_k^{-1} U_k^\top\}$ *and* $\{U_k H_k^{-1} Q_k^\top\}$ *are bounded and for all* k *sufficiently large*

(i) $Q_k U_k^\top \gamma(u_k) = \bar{H} u_k + o(|u_k|)$, *and*

(ii) $x^c_{k+1} - \bar{x} = o(|u_k|)$, $u^c_{k+1} = o(|u_k|)$, $p^c_{k+1} - \bar{x} = o(|u_k|)$, *and* $s^c_{k+1} = o(|u_k|)$.

Proof. Since the matrices U_k and Q_k have orthonormal columns, assumption **(S)(b)(ii)** implies that $\{U_k H_k^{-1} U_k^\top\}$ and$\{U_k H_k^{-1} Q_k^\top\}$ are bounded.
By assumption **(S)(d)**, $Q_k U_k^\top \to \bar{U}^\top$, so by Lemma 3(vi) with $u = u_k$, item (i) holds for all k sufficiently large.
To show item (ii) we write x^c_{k+1} from Algorithm 6 using notation (25) and Lemma 17(iii):

$$x^c_{k+1} = \hat{p}_{\mu_k}(x_k) - U_k H_k^{-1} U_k^\top \hat{s}_{\mu_k}(x_k) = p_{\mu_k}(x_k) - U_k H_k^{-1} U_k^\top \hat{s}_{\mu_k}(x_k) + o(|u_k|), \tag{29}$$

where $u_k = u_{\mu_k}(x_k)$. We now rewrite the second right hand side term above, using successively, Lemma 17(ii), $Q_k^\top Q_k = I$ together with the boundedness of $\{U_k H_k^{-1} U_k^\top\}$, and item (i) together with the boundedness of $\{U_k H_k^{-1} Q_k^\top\}$:

$$\begin{aligned} U_k H_k^{-1} U_k^\top \hat{s}_{\mu_k}(x_k) &= U_k H_k^{-1} U_k^\top \gamma(u_k) + U_k H_k^{-1} U_k^\top o(|u_k|) \\ &= U_k H_k^{-1} (Q_k^\top Q_k) U_k^\top \gamma(u_k) + o(|u_k|) \\ &= U_k H_k^{-1} Q_k^\top \bar{H} u_k + o(|u_k|). \end{aligned}$$

In turn, this last expression can be rewritten using **(S)(e)** together with Lemma 17(i), the expression $H_k^{-1} Q_k^\top Q_k H_k = I$, and **(S)(d)**:

$$\begin{aligned} U_k H_k^{-1} Q_k^\top \bar{H} u_k + o(|u_k|) &= (U_k H_k^{-1} Q_k^\top) Q_k H_k Q_k^\top u_k + o(|u_k|) \\ &= U_k Q_k^\top u_k + \bar{U} u_k - \bar{U} u_k + o(|u_k|) \\ &= \bar{U} u_k + o(|u_k|). \end{aligned}$$

As a result, from (29) we obtain

$$x^c_{k+1} = p_{\mu_k}(x_k) - \bar{U} u_k + o(|u_k|).$$

Now subtract $\bar{x}$ and use Lemma 12 item (i) to obtain

$$\begin{aligned} x^c_{k+1} - \bar{x} &= (u_k \oplus v(u_k)) - \bar{U} u_k + o(|u_k|) \\ &= o(|u_k|), \end{aligned}$$

where the last equality follows from (2) and the fact that $\bar{U}$ is the left basis matrix for the $\oplus$ decomposition.
We now show the last three equalities in item (ii). Note that $p_{\mu_{k+1}}(x^c_{k+1}) - \bar{x} = o(|u_k|)$, by Property (3)(ii). From Lemma 12(i) with $u_{\mu_{k+1}}(x^c_{k+1}) = u^c_{k+1}$, $p_{\mu_{k+1}}(x^c_{k+1}) - \bar{x} = u^c_{k+1} \oplus v(u^c_{k+1})$. Thus, $|u^c_{k+1}| \leq |p_{\mu_{k+1}}(x^c_{k+1}) - \bar{x}|$ and, hence, $u^c_{k+1} = o(|u_k|)$. Next, since $p^c_{k+1} - \bar{x} =$

$p^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1}) + p_{\mu_{k+1}}(x^c_{k+1}) - \bar{x}$ and $p^c_{k+1} = \hat{p}_{\mu_{k+1}}(x^c_{k+1})$, it follows from Lemma 17(iii) that $p^c_{k+1} - \bar{x} = o(|u^c_{k+1}|) + o(|u_k|) = o(|u_k|)$. Finally, from Lemma 17(i), $s^c_{k+1} = \hat{s}_{\mu_{k+1}}(x^c_{k+1}) = O(|u^c_{k+1}|) = o(|u_k|)$. □

In order to have p^c_{k+1} be a successful candidate, we can strengthen assumption **(S)(b)(iii)** by choosing σ_{k+1} to be bounded above by a constant multiple of $|s_k|^2$ as follows:

Lemma 19 *Suppose that* **(S)(a)-(e)** *hold and* $\sigma_{k+1} \leq min\{1/(k+2), |s_k|^2/|s_0|^2\}$ *for all* $k \geq 0$. *Then for all* k *sufficiently large,* (18) *holds and* $p_{k+1} = p^c_{k+1}$.

Proof. We start by writing $f(p^c_{k+1}) - f(p_k)$ as the sum of three difference terms:

$$\Big(f(p^c_{k+1}) - f(p_{\mu_{k+1}}(x^c_{k+1}))\Big) + \Big(f(p_{\mu_{k+1}}(x^c_{k+1})) - f(p_{\mu_k}(x_k))\Big) + \Big(f(p_{\mu_k}(x_k)) - f(p_k)\Big).$$

Next we proceed to bound each one of the three terms. Let L_f be a Lipschitz constant for f on a large enough ball about $\bar{x}$. Then

$$\begin{aligned} |f(p^c_{k+1}) - f(p_{\mu_{k+1}}(x^c_{k+1}))| &\leq L_f |p^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1})| \\ &= L_f O(|u_k|) O(|u^c_{k+1}|) \\ &= O(|u_k|) o(|u_k|) \\ &= o(|u_k|^2), \end{aligned}$$

where we used $\sigma_{k+1} \leq |s_k|^2/|s_0|^2$ together with Lemma 17(iv), and the fact that $u^c_{k+1} = o(|u_k|)$ by Lemma 18(ii). The bounds for the other terms are given by Lemma 16. By item (ii) therein and the Lemma 18 result $u^c_{k+1} = o(|u_k|)$,

$$\begin{aligned} f(p_{\mu_{k+1}}(x^c_{k+1})) - f(p_{\mu_k}(x_k)) &\leq -\tfrac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + O(|u^c_{k+1}|^2) + o(|u_k|^2) \\ &= -\tfrac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + O(o(|u_k|)^2) + o(|u_k|^2) \\ &= -\tfrac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + o(|u_k|^2). \end{aligned}$$

While, by item (i) of Lemma 16 and Lemma 17(iii) with $(\mu, x) = (\mu_k, x_k)$ and $\hat{p}_{\mu_k}(x_k) = p_k$,

$$f(p_{\mu_k}(x_k)) - f(p_k) \leq |p_k - p_{\mu_k}(x_k)| O(|u_k|) = o(|u_k|) O(|u_k|) = o(|u_k|^2)\,.$$

Altogether, we obtain the inequality

$$f(p^c_{k+1}) - f(p_k) \leq -\frac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + o(|u_k|^2)\,.$$

By Lemma 18, $|s^c_{k+1}|^2 = |o(|u_k|)|^2 = o(|u_k|^2)$. Therefore, since $\mu_{k+1} \geq \underline{\mu}$,

$$f(p^c_{k+1}) - f(p_k) + \frac{m}{2\mu_{k+1}}|s^c_{k+1}|^2 \leq -\frac{1}{2}\lambda_{min}(\bar{H})|u_k|^2 + o(|u_k|^2)\,.$$

Since the right hand side of this inequality is dominated by the negative term $-\frac{1}{2}\lambda_{min}(\bar{H})|u_k|^2$ as $u_k \to 0$, inequality (18) is satisfied for k sufficiently large and the result is established. □