A VU-algorithm for convex minimization

Robert Mifflin and Claudia Sagastizabal

July 2004
A $\mathcal{VU}$-algorithm for convex minimization

Robert Mifflin* and Claudia Sagastizábal†

July 14, 2004

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 lower semicontinuous 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. 301800/96. 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], and [Ous00]. 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 $\mathcal{V}$-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], [MS03b]. A general function of this type can have a primal-dual track leading to a (minimizing point, zero subgradient) pair; see [MS03a]. 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 $\mathcal{V}$-(or cutting-plane) model of $f$ it also can approximate a dual track point corresponding to a bundle iterate where the $\mathcal{V}$-model is sufficiently accurate. To complete the algorithm's $\mathcal{VU}$-model combination the dual vector goes into updating a $\mathcal{U}$-(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 (cf. (13) 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].

The 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 [MS02] 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 $|z_k| \leq \varepsilon |z_k|$ for all $k \geq k_\varepsilon$.
- $\zeta_k = O(|z_k|) \iff \exists C > 0$ such that $|z_k| \leq C |z_k|$ for all $k \geq 1$.

For algebraic purposes we consider (sub)gradients to be column vectors. For subgradients, the symbol $\partial$ stands for differentiation with respect to $x \in \mathbb{R}^n$, while $\nabla$ 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, $\text{lin} Y$ 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 [MS03a].

### 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

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

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 $\text{ri} \partial f(\bar{x})$, 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:

$$\mathbb{R}^n \ni x = \bar{U} (\bar{U}^T x) + \bar{V} \left( [\bar{V}^T \bar{V}]^{-1} \bar{V}^T x \right) = \bar{U} x_\mathcal{U} + \bar{V} x_\mathcal{V}$$

where $x_\mathcal{U} \oplus x_\mathcal{V} \in \mathbb{R}^{\dim \mathcal{U}} \times \mathbb{R}^{\dim \mathcal{V}}$.

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 $\bar{g}_\mathcal{V} = (\mathcal{V}^\top \mathcal{V})^{-1} \mathcal{V}^\top g$, the $\mathcal{U}$-Lagrangian of $f$, depending on $\bar{g}_\mathcal{V}$, is defined by
\[
\mathbb{R}^{\text{dim } \mathcal{U}} \ni u \mapsto L_\mathcal{U}(u; \bar{g}_\mathcal{V}) := \min_{v \in \mathbb{R}^{\text{dim } \mathcal{V}}} \left\{ f(\bar{x} + \bar{U}u + \bar{V} v) - \bar{g}^\top \bar{V} v \right\}.
\]
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 $\bar{g} \in \text{ri} \partial f(\bar{x})$, 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
\[
\nabla L_\mathcal{U}(0; \bar{g}_\mathcal{V}) = \bar{g}_\mathcal{U} = \bar{U}^\top \bar{g} = \bar{V}^\top g \quad \text{for all } \bar{g} \in \partial f(\bar{x}).
\]
A case of particular interest 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 $L_\mathcal{U}(u; \bar{g}_\mathcal{V})$, 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 algorithmically exploiting 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 $u \in \mathbb{R}^{\text{dim } \mathcal{U}}$ small enough
\[
\chi(u) := \bar{x} + u \oplus v(u) \quad \text{and}
\gamma(u) := \arg\min \left\{ |g|^2 : g \in \partial f(\chi(u)) \right\},
\]
where
1. $v : \mathbb{R}^{\text{dim } \mathcal{U}} \mapsto \mathbb{R}^{\text{dim } \mathcal{V}}$ is a $C^2$-function satisfying $\nabla v(u) \in W_\mathcal{U}(u; \bar{g}_\mathcal{V})$ for all $\bar{g} \in \text{ri} \partial f(\bar{x})$,
2. the Jacobian $J\chi(u)$ is a basis matrix for $\mathcal{V}(\chi(u))^\perp$, and
3. the particular $\mathcal{U}$-Lagrangian $L_\mathcal{U}(u; 0)$ is a $C^2$-function.

When we write $v(u)$ we implicitly assume that $\text{dim } \mathcal{U} \geq 1$. If $\text{dim } \mathcal{U} = 0$ we define the primal-dual track to be the point $(\bar{x}, 0)$. If $\text{dim } \mathcal{U} = n$ 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 4.3 in [MS03a] shows that if $0 \in \text{ri} \partial f(\bar{x})$ 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 4.3 in [MS03a] corresponds to $\bar{g} = 0$ and $\bar{f}$ equal to the convex function subdifferential $\partial f$. 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 Lemma 2.3 in [MS02] is incorrect and should be deleted. This does not affect any of the results in [MS02]. The fact that it is incorrect means that $\bar{g} \in \partial^f(\bar{x})$ should not be replaced by $\bar{g} \in \partial f(\bar{x})$ in part (i) of Definition 1.

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

$$v(u) = O(|u|^2),$$

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 $0 \in \partial^f(\bar{x})$. 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 $\gamma(u) = J\chi(u)[J\chi(u)^\top J\chi(u)]^{-1} \nabla L_{\mathcal{U}}(u; 0) \in \partial^f(\chi(u))$, and

(vi) $\gamma(u) = \bar{U} \bar{H} u + o(|u|)$.

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 $B_{\mathcal{U}}(u) = J\chi(u)$ gives item (iv) as well as the relation

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

Theorem 3.3 in [MS02] implies that

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

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 $L_d(u;0)$ with a Newton or quasi-Newton method. The difficulty with this approach is that $L_d$ 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 $\mu$ as discussed next.

2.3 Relating proximal points to primal track points

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

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

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

In our development we use the following properties, which result from the above definition, [RW98, Thm. 2.26] and [Roc76, Prop.1(c)]:

(i) $g_\mu(x) := \mu(x - p_\mu(x)) \in \partial f(p_\mu(x))$,

(ii) if $(\mu,x) \rightarrow (\bar{\mu},\bar{x})$ with $\bar{\mu} > 0$ then $p(x) \rightarrow p_\bar{\mu}(\bar{x})$, and

(iii) if $\bar{x}$ minimizes $f$ then $p_\mu(\bar{x}) = \bar{x}$ and $|p_\mu(x) - \bar{x}|^2 \leq |x - \bar{x}|^2 - |x - p_\mu(x)|^2$.

The following result, showing that primal tracks attract proximal points, constitutes a fundamental link between proximal point theory and $\mathcal{V}U$-theory. Since here we let $\mu$ vary in a compact set of positive numbers, this result is an extension of the fixed $\mu$ results in [MS02, Theorems 5.1 and 5.2]. It also can be seen as a convex version of Lemma 5 in [MS03c].

**Theorem 4** Let $\chi(u)$ be a primal track leading to a minimizer $\bar{x} \in \mathbb{R}^n$, as described in Definition 1, and suppose $0 \in \mathcal{R}f(\bar{x})$. Then for all $\mu \in [\underline{\mu},\bar{\mu}] \subset (0,\infty)$ and $x$ close enough to $\bar{x}$,

$$p_\mu(x) = \bar{x} + u_\mu(x) \oplus v(u_\mu(x)) \text{ where } u_\mu(x) := (p_\mu(x) - \bar{x})\mu$$

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

**Proof.** For $x$ close enough to $\bar{x}$, we write its proximal point using $\mathcal{V}U$ coordinates:

$$p_\mu(x) = \bar{x} + u_\mu(x) \oplus v_\mu(x) \text{ where } u_\mu(x) = (p_\mu(x) - \bar{x})\mu \text{ and } v_\mu(x) := (p_\mu(x) - \bar{x})\nu.$$ 

By Properties (3)(ii) and (iii), as $(\mu,x) \rightarrow (\bar{\mu},\bar{x})$ for any $\bar{\mu} > 0$, $x - p_\mu(x) \rightarrow \bar{x} - p_\bar{\mu}(\bar{x}) = 0$, and likewise, for the coordinates $u_\mu(x)$ and $v_\mu(x)$. Since $u_\mu(x) \rightarrow 0$, (2) implies $v(u_\mu(x)) \rightarrow 0$. As a result, the function $\omega_\mu : \mathbb{R}^n \rightarrow \mathbb{R}^{\dim \mathcal{V}}$ defined by

$$\omega_\mu(x) := \mu(\bar{x} - x)\nu - \frac{\mu}{2}(v(u_\mu(x)) + v_\mu(x)).$$
converges to 0 as $x \to \bar{x}$ for all $\mu \in [\bar{\mu}, \bar{\mu}]$. Since $0 \in \text{ri} \partial f(\bar{x})$, 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
\[
\omega := 0 \oplus \omega_\mu(x) \in \text{ri} \partial f(\bar{x}) \quad \text{for } x \text{ close enough to } \bar{x} \text{ and all } \mu \in [\bar{\mu}, \bar{\mu}].
\]
Thus, from the definition of $L_\mathcal{U}$ with $(u, \bar{\theta}) = (u_\mu(x), \omega \in \text{ri} \partial f(\bar{x}))$ and Definition 1,
\[
L_\mathcal{U}(u_\mu(x); \omega_\mu(x)) = f(\chi(u_\mu(x))) - \omega^\top \mathcal{V} v(u_\mu(x)).
\]
Since $v_{p_\mu}(x) \in \mathbb{R}^{\dim \mathcal{V}}$, $L_\mathcal{U}(u_\mu(x); \omega_\mu(x)) \leq f(\bar{x} + u_\mu(x) \oplus v_{p_\mu}(x)) - \omega^\top \mathcal{V} v_{p_\mu}(x)$. As a result,
\[
f(\chi(u_\mu(x))) - \omega^\top \mathcal{V} v(u_\mu(x)) \leq f(p_\mu(x)) - \omega^\top \mathcal{V} v_{p_\mu}(x). \tag{4}
\]
By the definition of the proximal point mapping,
\[
f(p_\mu(x)) + \frac{\mu}{2} |p_\mu(x) - x|^2 \leq f(\chi(u_\mu(x))) + \frac{\mu}{2} |\chi(u_\mu(x)) - x|^2. \tag{5}
\]
Combining the two inequalities above yields, after rearrangement of terms,
\[
0 \leq \frac{\mu}{2} (|\chi(u_\mu(x)) - x|^2 - |p_\mu(x) - x|^2) + \omega^\top \mathcal{V} (v(u_\mu(x)) - v_{p_\mu}(x)). \tag{6}
\]
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 $u_\mu$, $p_\mu$, $v(u_\mu)$, $v_{p_\mu}$, and $\omega_\mu$. First we expand the leading difference of squares term in (6) and use the fact that $\chi(u_\mu)$ and $p_\mu$ have the same $\mathcal{U}$-component:
\[
|\chi(u_\mu) - x|^2 - |p_\mu - x|^2 = (\chi(u_\mu) - p_\mu) ^\top (\chi(u_\mu) + p_\mu - 2x)
\]
\[
= (\mathcal{V} (v(u_\mu) - v_{p_\mu})) ^\top (\mathcal{V} (\chi(u_\mu) + p_\mu - 2x) \nu)
\]
\[
= (v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} (v(u_\mu) + v_{p_\mu} - 2(x - \bar{x}) \nu)
\]
\[
= v(u_\mu) ^\top \mathcal{V}^\top \mathcal{V} v(u_\mu) - v_{p_\mu} ^\top \mathcal{V}^\top \mathcal{V} v_{p_\mu} + 2(v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} (x - \bar{x}) \nu.
\]
Then
\[
\frac{\mu}{2} (|\chi(u_\mu) - x|^2 - |p_\mu - x|^2) = \frac{\mu}{2} (|\mathcal{V} (v(u_\mu))^2 - |\mathcal{V} v_{p_\mu}|^2) - \mu(v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} (x - \bar{x}) \nu.
\]
Now, since $\mathcal{V}^\top \omega = \mathcal{V}^\top \mathcal{V} \omega_\mu$, we use the definition of $\omega_\mu$ to write the second right hand side term in (6) as follows:
\[
\omega^\top \mathcal{V} (v(u_\mu) - v_{p_\mu}) = (v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \omega = (v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} \omega_\mu
\]
\[
= (v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} (\mu(\bar{x} - x) \nu - \frac{\mu}{2} (v(u_\mu) + v_{p_\mu}))
\]
\[
= \mu(v(u_\mu) - v_{p_\mu}) ^\top \mathcal{V}^\top \mathcal{V} (\bar{x} - x) \nu - \frac{\mu}{2} (v(u_\mu) ^\top \mathcal{V}^\top \mathcal{V} v(u_\mu) - v_{p_\mu} ^\top \mathcal{V}^\top \mathcal{V} v_{p_\mu}).
\]
Using these expressions in the right hand side in (6), we obtain that (6) holds with equality. Since the (in)equality 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 \( \mu \) is fixed, the above conclusion is extended to prox-regular and -bounded functions in [MS03c] and, with the addition of \( C^2 \)-substructure, in [Har03] and [MS03a].

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 \( \mathcal{V} \)-algorithm 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 these primal-dual approximations, via the solution of two quadratic programming problems, denoted below by \( \chi \)-QP and \( \gamma \)-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 \( \sigma \)-approximation of \( p_\mu(x) \), a bundle method accumulates information from past points \( y_i \) in the form

\[
\{(f(y_i), g_i \in \partial f(y_i))\}_{i \in \mathcal{B}},
\]

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

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

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

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

adding \( 0 = e_i - e_i \) to this inequality gives

\[
f(z) \geq f(x) + g_i^T(z - x) - e_i \quad \text{for all } z \in \mathbb{R}^n,
\]

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

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

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

\[
\varphi(p) := f(x) + \max_{i \in \mathcal{B}} \{-e_i + g_i^T(p - x)\} \quad \text{for } p \in \mathbb{R}^n.
\]
A pure cutting-plane algorithm [CG59], [Kel60] minimizes the polyhedral function $\varphi$ 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 $\mu$, which is added to the model function. In the bundle context, the prox-parameter $\mu$ 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 primal track point we solve a first quadratic programming subproblem $\chi$-QP, which has the following form and properties; see [BGLS03, Lemma 9.8]:

The problem

\[
\begin{aligned}
\min_{r \in \mathbb{R}, p \in \mathbb{R}^n} & \quad r + \frac{1}{2} \mu |p - x|^2 \\
\text{subject to} & \quad r \geq f(x) - e_i + g_i^T(p - x) \quad \text{for all } i \in \mathcal{B}. \\
\end{aligned}
\]

(\chi$\text{-QP}$)

has a dual

\[
\begin{aligned}
\min_{\alpha_i} & \quad \frac{1}{2} \sum_{i \in \mathcal{B}} \alpha_i g_i^2 + \mu \sum_{i \in \mathcal{B}} e_i \alpha_i \\
\text{subject to} & \quad \alpha \in \left\{ z \in [0, 1]^{\mathcal{B}} : \sum_{i \in \mathcal{B}} z_i = 1 \right\}.
\end{aligned}
\]

(8)

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

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

(9)

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

\[
\varphi(\hat{p}) = f(x) + \sum_{i \in \mathcal{B}} \hat{\alpha}_i (-e_i + g_i^T(\hat{p} - x)) = f(x) - \sum_{i \in \mathcal{B}} \hat{\alpha}_i e_i - \frac{1}{\mu} |\hat{g}|^2.
\]

(10)

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

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

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 $\hat{i}_+$, by letting $y_{\hat{i}_+} := \hat{p}$ and computing $f(\hat{p})$ and $g_{\hat{i}_+} \in \partial f(\hat{p})$. Note that since $f(\hat{p})$ is available, we can compute the V-model accuracy measure at $\hat{p}$ defined by

\[
\hat{e} := f(\hat{p}) - \varphi(\hat{p}) = f(\hat{p}) - \hat{r}.
\]

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

\[
\hat{B} := \left\{ i \in \mathcal{B} : \hat{r} = f(x) - e_i + g_i^T(\hat{p} - x) \right\} \cup \{ \hat{i}_+ \}.
\]
The second quadratic programming problem

$$\begin{align*}
\min_{r \in \mathbb{R}, p \in \mathbb{R}^n} & \quad \frac{1}{2} r + \frac{1}{2} |p - x|^2 \\
n & \quad r \geq g_i^T (p - x) \text{ for all } i \in \hat{B}
\end{align*}$$

(\gamma\text{-QP})

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

$$\begin{align*}
\min_{\alpha, \tilde{\alpha}, z} & \quad \frac{1}{2} \sum_{i \in \hat{B}} \alpha_i |g_i|^2 \\
\text{s.t.} & \quad \alpha \in \left\{ z \in [0, 1]^{\hat{B}} : \sum_{i \in \hat{B}} z_i = 1 \right\}
\end{align*}$$

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

$$\bar{p} - x = -\hat{s} \quad \text{where} \quad \hat{s} := \sum_{i \in \hat{B}} \bar{\alpha}_i g_i.$$  \hspace{1cm} (11)

Note that $\hat{s}$ is the vector with smallest Euclidean norm in the convex hull of $\{g_i : i \in \hat{B}\}$. Lemma 5 below will show that $\hat{s}$ is an $\epsilon$-subgradient of $f$ at $\bar{p}$. 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 $\hat{V}^T \hat{s} = 0$. Thus, if $p_\mu(x)$ is a primal track point $\chi(u)$ approximated by $\bar{p}$, then the convex hull of $\{g_i : i \in \hat{B}\}$ 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 $\hat{B}_{\text{act}} := \{ i \in \hat{B} : r = g_i^T (\bar{p} - x) \}$. Then, from (11),

$$\bar{r} = -g_i^T \hat{s} \quad \text{for all } i \in \hat{B}_{\text{act}},$$

so

$$\begin{align*}
(g_i - g_{\ell})^T \hat{s} = 0
\end{align*}$$

(12)

for all such $i$ and for a fixed $\ell \in \hat{B}_{\text{act}}$. Define a full column rank matrix $\hat{V}$ by choosing the largest number of indices $i$ satisfying (12) 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 $\hat{V}^\perp$ with $\hat{U} = I$ if $\hat{V}$ is vacuous.

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

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

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

$$\hat{\epsilon} \leq \frac{\sigma}{\mu} |\hat{s}|^2.$$
Otherwise, $\mathcal{B}$ above is replaced by $\hat{\mathcal{B}}$ and new iterate data are computed by solving updated subproblems ($\chi$-QP) and ($\gamma$-QP). This update, appending $(e_{i+}, g_{i+})$ to active data at the previous ($\chi$-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

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

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

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

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

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

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

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

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

In addition, for any parameter $m \in (0, 1)$, satisfaction of (13) implies

$$f(\hat{p}) - f(x) \leq -\frac{m}{2\mu} |\hat{g}|^2. \tag{14}$$

**Proof.** Since $g_{i+} \in \partial f(\hat{p})$ and $\varepsilon = f(\hat{p}) - \varphi(\hat{p}) \geq 0$, $g_{i+} \in \partial_{\varepsilon} f(\hat{p})$, 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 $i \neq i_+$ in $\hat{\mathcal{B}}$

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

so for all such $i$

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

Adding this result to (7) gives

$$f(z) \geq f(\hat{p}) + g_i^\top (z - \hat{p}) - \varepsilon \quad \text{for all } z \in \mathbb{R}^n, \tag{15}$$

and completes the proof of item (i).

Now item (ii) follows by multiplying each inequality in (15) by its corresponding multiplier $\bar{\alpha}_i \geq 0$, summing these results and then using the definition of $\hat{s}$ from (11) and the fact that these multipliers sum to one.

In a similar manner, this time using the multipliers $\bar{\alpha}_i$ that solve dual problem (8) and define
\( \hat{g} \) in (9), together with \( \hat{\alpha}_{i_+} := 0 \), we obtain the result that \( \hat{g} \in \partial_z f(\hat{p}) \). 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

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

so

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

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

\[
\hat{g} = -\mu(\hat{p} - x)
\]

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 (12), because these 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), (16) and the definition of \( \hat{B} \) imply that \( \mu(x - \hat{p}) = \hat{g} \) is in the convex hull of \( \{g_i : i \in \hat{B}\} \).

To show that for any \( m \in (0, 1) \) condition (14) holds when (13) holds, first note that, since \( \sigma \leq 1/2 \), we have \( \sigma \leq 1 - \frac{m}{2} \). Thus if (13) holds then \( \varepsilon \leq [(1 - \frac{m}{2})/\mu]\|\hat{s}\|^2 \). This inequality together with the definition of \( \hat{\varepsilon} \), (10) and the nonnegativity of \( \hat{\alpha}_i e_i \) gives

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

Finally, combining this inequality with item (v) gives (14). \( \square \)

In bundle terminology, (14) corresponds to declaring \( \hat{p} \) to be a “serious step”; see [HUL93, Ch. XIV-XV] for more details.

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

## 4 The \( \mathcal{U} \)-algorithm

Now we consider an algorithm depending on the \( \mathcal{U} \)-theory outlined in Section 2 and the potential primal-dual track point approximations from Section 3. When a primal track exists, the algorithm moves approximately along it by making a \( \mathcal{U} \)-Newton predictor step followed by a proximal point corrector step, or \( \mathcal{V} \)-step.

The \( \mathcal{U} \)-step depends on
- the \( \varepsilon \)-subgradient \( \hat{s} = \hat{U}\hat{U}^\top \hat{s} \) estimating a gradient of \( L_\mathcal{U} \) and
– an estimate of $\nabla^2 L_{d}(0; 0)$ (if relevant second order information is not available, a quasi-Newton method could be used).

The $V$-step uses the estimate $\hat{p}$ given by the bundle subprocedure output. The sum of these two steps gives what we call a “candidate” primal track step. It is declared “good enough” if it satisfies an $f$-value descent condition that is essentially testing for descent of $L_d$; see (17) 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 indexed by $k$.

Algorithm 6

**Initialization.** Choose positive parameters $\varepsilon$, $\mu$ 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 $U$-basis. Set $s_0 := g_0$ and $k := 0$.

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

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

**$U$-Step.** Compute an approximate $U$-Newton step by solving the linear system

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

Set $x_{k+1}^c := p_k + U_k \Delta u_k = p_k - U_k H_k^{-1} U_k^T s_k$.

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

Compute recursively

$$(\hat{p}, \hat{r}) = \chi \text{-QP} \left( \mu_{k+1}, x, \{ (e_i, g_i) \}_{i \in \mathcal{B}} \right), \quad \hat{\varepsilon} = f(\hat{p}) - \hat{r}, \quad \hat{B}, \quad \hat{\mathcal{G}}$$

until satisfaction of (13) with $\sigma = \sigma_{k+1}$.

Then set $(\varepsilon_{k+1}^c, p_{k+1}^c, s_{k+1}^c, U_{k+1}^c) := (\hat{\varepsilon}, \hat{p}, \hat{s}, \hat{U})$.

**Candidate evaluation and new iterate data determination.** If

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

then declare a successful candidate and set

$$(x_{k+1}, e_{k+1}, p_{k+1}, s_{k+1}, U_{k+1}) := (x_{k+1}^c, e_{k+1}^c, p_{k+1}^c, s_{k+1}^c, U_{k+1}^c).$$

Otherwise, execute a line search on the line determined by $p_k$ and $p_{k+1}^c$ 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 $(\varepsilon, \hat{p}, \hat{s}, \hat{U})$. Then set $(\varepsilon_{k+1}, p_{k+1}, s_{k+1}, U_{k+1}) := (\hat{\varepsilon}, \hat{p}, \hat{s}, \hat{U})$. Replace $k$ by $k + 1$ and go to **Stopping test.**

$\square$
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 \( \max \{ |s|^2, \frac{\mu}{\sigma} \} \leq \varepsilon \).

(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 \( p_\mu(x) \), here \( x_{k+1} \) 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
\[
H_{k+1} U_{k+1}^T (p_{k+1} - p_k) = U_{k+1}^T (s_{k+1} - s_k). \tag{18}
\]
A way to deal with the case when the \( U \)-matrix changes dimension would be to use a limited memory method [GL89], [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 \( x_{k+1}^e \) 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 \( x_{k+1}^e \) does not satisfy a Wolfe increase condition [BGLS03, Sec. 3.4]. If not satisfied, then \( f(x_{k+1}^e) < f(p_k) \) and safeguarded quadratic extrapolation could be executed to find a point satisfying such a derivative increase condition. Then \( x_{k+1}^e \) 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 \( f(x_{k+1}^e) \) sufficiently smaller than \( f(p_k) \). Information gained from this could go into the choice of \( \mu_{k+1} \) and initialization of \( \mathcal{B} \).

(v) For the purpose of early detection of a smooth function with \( \text{dim } \mathcal{U} = n \) a termination test for the bundle procedure, based on (17), could be included inside the bundle procedure as follows: Immediately after \( x_{k+1}^e \) and \( \mu_{k+1} \) are generated the bundle subprocedure would initialize the index \( i_+ \) so that \( y_{i+} = x_{k+1}^e \) 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
\[
(x_{k+1}^e, p_{k+1}^e, s_{k+1}^e, U_{k+1}^e) := (0, y_{i+}, g_{i+}, I),
\]
so as to generate a successful candidate satisfying (17) and the subsequent results in (19), (20) and (21). 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 \( p_{k+1}^e \) is not a successful candidate can be executed very simply by choosing \( x_{k+1} := \arg\min \{ f(p_k), f(p_{k+1}^e) \} \), 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 \( p_{k+1}^e \) is a successful candidate

\[
f(p_{k+1}) - f(p_k) \leq -\frac{m}{2\mu_{k+1}}|s_{k+1}|^2. \tag{19}\]

In the successful case, (19) is the same as (17). Otherwise, (14) and Lemma 5(v) with \((\mu, x, \hat{p}, \hat{s}, \hat{g}) = (\mu_{k+1}, x_{k+1}, p_{k+1}, s_{k+1}, g_{k+1})\) imply that

\[
f(p_{k+1}) - f(x_{k+1}) \leq -\frac{m}{2\mu_{k+1}}|g_{k+1}|^2 \leq -\frac{m}{2\mu_{k+1}}|s_{k+1}|^2,
\]

and the line search gives \( f(x_{k+1}) \leq f(p_k) \), so (19) also holds in the unsuccessful case. A necessary, but not sufficient, condition for the unsuccessful case is \( f(x_{k+1}^e) > f(p_k) \). □

5 Convergence properties of 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 (13) 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 \( p_\mu(x) - x \in V(p_\mu(x)) \).

**Proof.** The recursion in the bundle subprocedure replacing \( \mathcal{B} \) by \( \bar{\mathcal{B}} \) (i.e., appending \((e_i, g_i)\) to active data at the previous (\( \chi \)-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 (13) does not hold, the sequence of
$|s|\$-values also converges to 0. Thus, item (iii) in Lemma 5 implies that \( \{\hat{p}\} \rightarrow p(\mu(x)) \). 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 (13) implies \( \hat{\epsilon} = 0 \) 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
\[
0 - \mu(x - p(\mu(x))) \in \mathcal{V}(p(\mu(x))) \text{ and the final result follows, since } \mu \neq 0. \] 

If either case in Theorem 8 above holds then there is a minimizer \( \bar{x} = p(\mu(x)) \) such that \( \bar{x} \to x \in \mathcal{V}(\bar{x}) \), 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, we do not know of a proof of rapid convergence 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 \( \epsilon = 0 \) implies that infinite sequences
\[
\left\{(\mu_k, \sigma_k), \left(x_k^\epsilon, \epsilon_k, p_k, s_k, U_k \right)\right\}
\]
are generated such that, for all \( k \), \( s_k = U_k U_k^\epsilon s_k \) is not the zero vector, (19) holds and, from Lemma 5(ii) and (13) with (\( \mu, \sigma, \hat{\epsilon}, \hat{\mu}, \hat{s} \)) = (\( \mu_k, \sigma_k, \epsilon_k, p_k, s_k \)),
\[
f(p_k) + s_k^\epsilon (z - p_k) - \epsilon_k \leq f(z) \quad \text{for all } z \in \mathbb{R}^n \quad \text{and} \quad (20)
\]
\[
\epsilon_k \leq \frac{\sigma_k |s_k|^2}{\mu_k} \quad \text{ (21)}
\]

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 \( \{f(p_k)\} \rightarrow -\infty \) or \( \{|s_k|\} \) and \( \{\epsilon_k\} \) both converge to 0;

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

**Proof.** Since \( |s_k| \neq 0 \), (19) implies that \( \{f(p_k)\} \) is decreasing. Suppose \( \{f(p_k)\} \not\rightarrow -\infty \). Then summing (19) over \( k \) and using the fact that \( m \frac{m}{2\mu_k} \geq m \frac{m}{2\bar{\mu}} \) for all \( k \) implies that \( \{|s_k|\} \rightarrow 0 \).

Then (21) with \( \sigma_k \leq 1 \) and \( \mu_k \geq \mu > 0 \) implies that \( \{\epsilon_k\} \rightarrow 0 \), 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 \( \{\epsilon_k\} \) converge to 0 by item (i), (20) 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. \( \square \)

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 $0 \in \partial f(\bar{x})$ and the corresponding $U$-Lagrangian $L_{\bar{u}}(u; 0)$ has a Hessian at $u = 0$ that is positive definite. \[ \square \]

A first consequence of $\bar{x}$ being a strong minimizer, following from positive definiteness of $\nabla^2 L_{\bar{u}}(0; 0)$, is that $u = 0$ is the unique minimizer of $L_{\bar{u}}(u; 0)$. This result, together with [Thm.][L001], implies that $\bar{x}$ is the unique minimizer of $f$. In addition, from [Roc76, Thms. 27.1(d)-(f) and 8.4]

for any $\alpha \geq f(\bar{x})$ the level set $\{ x \in \mathbb{R}^n : f(x) \leq \alpha \}$ is compact. \[ (22) \]

**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

\[ \{ H^{-1}_k \} \text{ is bounded,} \]

then $\{ x_{k+1}^e \} \text{ and } \{ x_k \}$ both converge to $\bar{x}$ and $\{ s_{k+1}^e \}$ converges to $0 \in \mathbb{R}^n$. \[ (23) \]

**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 $\{ y : f(y) \leq f(p_0) \}$ is compact by (22). Thus, $\{ p_k \}$ is bounded and, from item (ii) in Theorem 9, any accumulation point of $\{ p_k \}$ is $\bar{x}$.

Now suppose (23) holds. Then, since each $U_k$ has orthonormal columns, the sequence $\{ x_{k+1}^e - p_k \} = \{ -U_k H^{-1}_k U_k^T s_k \}$, if infinite, converges to $0 \in \mathbb{R}^n$ because, by Theorem 9(i), $\{ s_k \}$ does the same. Thus, $\{ x_{k+1}^e \}$ has the same limit as $\{ p_k \}$, namely $\bar{x}$. Next consider any infinite subsequence of $\{ x_{k+1} \}$ such that $x_{k+1} \neq x_{k+1}^e$. 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 $\{ x_{k+1}^e \} \rightarrow \bar{x}$. This also implies the final result that $\{ s_{k+1}^e \} \rightarrow 0$ because, from (14) and Lemma 5(a) with $(\chi, \bar{\nu}, \bar{e}, s) = (x_{k+1}^e, p_k^e, g_{k+1}^e, s_{k+1}^e)$ and $\nu = \mu_{k+1} \leq \bar{\nu}$, $f(\bar{x}) - f(x_{k+1}^e) \leq f(p_k^e) - f(x_{k+1}^e) \leq \frac{m}{2 \mu_{k+1}} |g_{k+1}^e|^2 \leq -\frac{m}{2 \bar{\nu}} |s_{k+1}^e|^2$. \[ \square \]

### 5.2 Superlinear convergence

In order to prepare for our local convergence analysis we first establish some technical results. The initial suppositions that we need for showing superlinear convergence are the following:

**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) $\{ H^{-1}_k \}$ is bounded, i.e., (23) holds;
(iii) \(\{\sigma_k\} \to 0\) as \(k \to \infty\), for example, by choosing \(\sigma_{k+1} \leq 1/(k+2)\).

Next we combine Theorems 4 and 9(iii) to obtain 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 \(u_{k+1}^\epsilon := u_{\mu_{k+1}}(x_{k+1}^\epsilon)\). Then the following hold for all \(k\) sufficiently large:

(i) \(p_\mu(x) = \bar{x} + u_\mu(x) \oplus v(u_\mu(x))\) for \((\mu, x)\) equal to \((\mu_k, x_k)\) and to \((\mu_{k+1}, x_{k+1}^\epsilon)\);

(ii) \(\{u_k\}\) and \(\{u_{k+1}^\epsilon\}\) both converge to \(0 \in \mathbb{R}^\text{dim} \mathcal{U}\);

(iii) \(f(p_{\mu_k}(x_k)) - f(p_k) = |p_k - p_{\mu_k}(x_k)||O(|u_k|); \text{ and}

(iv) \(f(p_{\mu_{k+1}}(x_{k+1}^\epsilon)) - f(p_{\mu_k}(x_k)) \leq -\frac{1}{2}\lambda_{\min}(\bar{H})|u_k|^2 + O\left(|u_{k+1}^\epsilon|^2 \right) + o(|u_k|^2),

where \(\lambda_{\min}(\bar{H})\) is the smallest eigenvalue of \(\bar{H} = \nabla^2 L_{\mu}(0; 0)\).

**Proof.** Via Corollary 11, items (i) and (ii) follow from Theorem 4 with \((\mu, x)\) therein equal to \((\mu_k, x_k)\) and to \((\mu_{k+1}, x_{k+1}^\epsilon)\).

From Lemma 3(v), \(\gamma(u_k) \in \partial f(\chi(u_k))\) where \(\chi(u_k) = \bar{x} + u_k \oplus v(u_k)\). From item (i) with \((\mu, x) = (\mu_k, x_k)\) we have \(p_{\mu_k}(x_k) = \chi(u_k)\), so \(\gamma_k \in \partial f(p_{\mu_k}(x_k))\). By convexity of \(f\),

\[ f(p_{\mu_k}(x_k)) + \gamma(u_k) \cdot (p_k - p_{\mu_k}(x_k)) \leq f(p_k). \]

Thus,

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

and item (iii) then follows because, from Lemma 3 (vi), \(\gamma(u_k) = O(|u_k|)\).

Writing Lemma 3(ii) for the primal track points \(p_{\mu_{k+1}}(x_{k+1}^\epsilon)\) and \(p_{\mu_k}(x_k)\) yields, respectively,

\[ f(p_{\mu_{k+1}}(x_{k+1}^\epsilon)) = f(\bar{x}) + \frac{1}{2}(u_{k+1}^\epsilon + \epsilon H) + o(|u_{k+1}^\epsilon|^2) \]

and \(f(p_{\mu_k}(x_k)) = f(\bar{x}) + \frac{1}{2}u_k H u_k + o(|u_k|^2)\). Therefore,

\[ f(p_{\mu_{k+1}}(x_{k+1}^\epsilon)) - f(p_{\mu_k}(x_k)) \leq -\frac{1}{2}\lambda_{\min}(\bar{H})|u_k|^2 - o(|u_k|^2) + \frac{1}{2}\lambda_{\max}(\bar{H})|u_{k+1}^\epsilon|^2 + o(|u_{k+1}^\epsilon|^2) \]

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

To give rate of convergence results, in addition to assuming there is a primal track leading to a strong minimizer, we make the following algorithm assumptions stating how well the dual track points and corresponding \(\mathcal{U}\)-basis and -Hessian matrices need to be approximated. Since estimates may come either from a successful candidate, or from a line search, we introduce the following notation, to be used with \((\mu, x)\) equal to \((\mu_k, x_k)\) and to \((\mu_{k+1}, x_{k+1}^\epsilon)\):

\[
\left(\tilde{p}_\mu(x), \tilde{s}_\mu(x)\right) := (\tilde{p}, \tilde{s}), \text{ output from } \chi\text{-QP/} \gamma\text{-QP satisfying (13)}. \quad (24)
\]
(S)(c) For all $k$ sufficiently large and for $(\mu, x)$ equal to $(\mu_k, x_k)$ and to $(\mu_{k+1}, x_{k+1}^c)$ the $\gamma$-QP output $\tilde{s}_{\mu}(x)$, correspondingly equal to $s_k$ and to $s_{k+1}^c$, satisfies

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

(25)

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 $\{U_kQ_k^T\}$ converging to $\bar{U}$.

(S)(e) $[Q_kH_kQ_k^T - \bar{H}]u_k = o(|s_k|) + o(|u_k|)$ 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 the approximation of $\gamma(u_{\mu}(x))$ by $\tilde{s}_{\mu}(x)$ is as good as that of $g_{\mu}(x)$ by $\hat{g}_{\mu}(x) : = \mu(x - \hat{p}_{\mu}(x))$ in the sense that

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

then (25) in (S)(c) holds. This follows, because Lemma 5(iii) and (13) combined with $\hat{g}_{\mu}(x) - g_{\mu}(x) = \mu(p_{\mu}(x) - \hat{p}_{\mu}(x))$ imply $|\hat{g}_{\mu}(x) - g_{\mu}(x)| \leq \sqrt{\sigma}|\tilde{s}_{\mu}(x)|$, 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 $Q^TQ = I$). This supposition means that all accumulation matrices of the bounded set $\{U_k\}$ are orthonormal basis matrices for $\mathcal{U}$. The matrix $U_kQ_k^T$ 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 (18), because inserting $I = Q_{k+1}^TQ_{k+1}$ into this equation and multiplying by $Q_{k+1}$ on the left gives the equivalent quasi-Newton equation

$$(Q_{k+1}H_{k+1}Q_{k+1}^T)(U_{k+1}Q_{k+1}^T)^T(p_{k+1} - p_k) = (U_{k+1}Q_{k+1}^T)^T(s_{k+1} - s_k).$$

The next lemma uses the additional assumptions (S)(b)(iii) and (c), concerning adequate approximation of dual track points $\gamma(u_k)$, to show correspondingly accurate approximation of the primal track points $p_{\mu_k}(x_k)$ and to 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 $u_{k+1}^c$, that are useful for rate of convergence analysis, rather than their linearly-related computed quantities $s_k$ and $s_{k+1}^c$, that are good for rate of computational progress observation.
Lemma 14 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 \( (\mu, x) \) equal to \( (\mu, x_k) \) and to \( (\mu_{k+1}, x^c_{k+1}) \)

(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 \( \hat{p}^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1}) = O(|u_{k+1}|)O(|u^c_{k+1}|) \).

Proof. For large \( k \) and \( (\mu, x) \) equal to \( (\mu_k, x_k) \) or to \( (\mu_{k+1}, x^c_{k+1}) \), (25) in (S)(c) and Lemma 3 (vi) with \( u = u_\mu(x) \) gives

\[
\hat{s}_\mu(x) - o(|\hat{s}_\mu(x)|) = \gamma(u_\mu(x)) + o(|u_\mu(x)|) = O(|u_\mu(x)|) + o(|u_\mu(x)|).
\]

This implies that item (i) holds, which together with (25) gives item (ii).

In addition, for such \( (\mu, x) \) values and corresponding \( \sigma \) values, Lemma 5(iii) and (13), and \( \mu \geq \mu \), give

\[
|\hat{p}_\mu(x) - p_\mu(x)|^2 \leq \frac{\sigma}{\mu^2} |\hat{s}_\mu(x)|^2 \leq \frac{\sigma}{\mu^2} |\hat{s}_\mu(x)|^2,
\]

which together with (S)(b)(iii) and item (i) implies the validity of item (iii).

Moreover, for \( (\mu, x, \sigma) = (\mu_{k+1}, x^c_{k+1}, \sigma_{k+1}) \) (26) becomes

\[
\left| \hat{p}_{\mu_{k+1}}(x^c_{k+1}) - p_{\mu_{k+1}}(x^c_{k+1}) \right|^2 \leq \frac{\sigma_{k+1}}{\mu^2} |\hat{s}_{\mu_{k+1}}(x^c_{k+1})|^2.
\]

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

\[
\left| \hat{p}^c_{k+1} - p_{\mu_{k+1}}(x^c_{k+1}) \right| \leq \frac{\sqrt{\sigma_{k+1}}}{\mu} O(|u^c_{k+1}|).
\]

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 (27). \( \square \)

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 \( Q_k U_k^T \gamma(u_k) \) and its approximant \( Q_k U_k^T s_k \) behave like the \( U \)-gradient \( \nabla I_{\xi_k}(u_k; 0) \) that, by Lemma 3(iii), equals \( \tilde{H} u_k + o(|u_k|) \). The second part is concerned with the \( u \)-rate of convergence of the candidate data.
Lemma 15 Suppose that (S)(a)-(e) hold and let \( \{u_k\} \) and \( \{u_{k+1}^c\} \) be the zero-convergent sequences from Lemma 12. Then the sequences \( \{U_kH_k^{-1}U_k^\top\} \) and \( \{U_kH_k^{-1}Q_k^\top\} \) are bounded and for all \( k \) sufficiently large the following hold:

(i) \( Q_k U_k^\top \gamma(u_k) = \tilde{H} u_k + o(|u_k|) \);

(ii) \( x_{k+1}^c - \bar{x} = o(|u_k|), u_{k+1}^c = o(|u_k|), p_{k+1}^c - \bar{x} = o(|u_k|), \) and \( s_{k+1}^c = o(|u_k|) \).

Proof. Since the matrices \( U_k \) and \( Q_k \) have orthonormal columns, assumption (S)(b)(ii) implies that \( \{U_kH_k^{-1}U_k^\top\} \) and \( \{U_kH_k^{-1}Q_k^\top\} \) are bounded.

By assumption (S)(d), \( Q_k U_k^\top \rightarrow \bar{U}^\top \), so by Lemma 3(vi) with \( u = u_k \), item (i) holds for all \( k \) sufficiently large.

To show item (ii) we write \( x_{k+1}^c \) from Algorithm 6 using notation (24) and Lemma 14(iii):

\[
x_{k+1}^c = \bar{p}_{\mu_k}(x_k) - U_kH_k^{-1}U_k^\top s_{\mu_k}(x_k) = p_{\mu_k}(x_k) - U_kH_k^{-1}U_k^\top s_{\mu_k}(x_k) + o(|u_k|),
\]

(28)

where \( u_k = u_{\mu_k}(x_k) \). We now rewrite the second right hand side term above, using successively, Lemma 14(ii), \( Q_k^\top Q_k = I \) together with the boundedness of \( \{U_kH_k^{-1}U_k^\top\} \), and item (i) together with the boundedness of \( \{U_kH_k^{-1}Q_k^\top\} \):

\[
U_kH_k^{-1}U_k^\top s_{\mu_k}(x_k) = U_kH_k^{-1}U_k^\top \gamma(u_k) + U_kH_k^{-1}U_k^\top o(|u_k|) = U_kH_k^{-1}(Q_k^\top Q_k)U_k^\top \gamma(u_k) + o(|u_k|) = U_kH_k^{-1}Q_k^\top \bar{H} u_k + o(|u_k|).
\]

In turn, this last expression can be rewritten using (S)(e) together with Lemma 14(i), the expression \( H_k^{-1}Q_k^\top Q_k H_k = I \), and (S)(d):

\[
U_kH_k^{-1}Q_k^\top \bar{H} u_k + o(|u_k|) = (U_kH_k^{-1}Q_k^\top)Q_k H_k Q_k^\top u_k + o(|u_k|) = U_k Q_k^\top u_k + \bar{U} u_k - \bar{U} u_k + o(|u_k|) = \bar{U} u_k + o(|u_k|).
\]

As a result, from (28) we obtain

\[
x_{k+1}^c = p_{\mu_k}(x_k) - \bar{U} u_k + o(|u_k|).
\]

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

\[
x_{k+1}^c - \bar{x} = (u_k \oplus v(u_k)) - \bar{U} u_k + o(|u_k|)
\]

\[
= o(|u_k|),
\]

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 \( p_{\mu_{k+1}}(x_{k+1}^c) - \bar{x} = o(|u_k|) \), by Property (3)(iii). From Lemma 12(i) with \( u_{\mu_{k+1}}(x_{k+1}^c) = u_{k+1}^c, p_{\mu_{k+1}}(x_{k+1}^c) - \bar{x} = u_{k+1}^c \oplus v(u_{k+1}^c) \). Thus, \( |u_{k+1}^c| \leq |p_{\mu_{k+1}}(x_{k+1}^c) - \bar{x}| \) and, hence, \( u_{k+1}^c = o(|u_k|) \). Next, since \( p_{k+1}^c - \bar{x} =
\[ p_{k+1}^c - p_{\mu_{k+1}}(x_{k+1}^c) + p_{\mu_{k+1}}(x_{k+1}^c) - \bar{x} \quad \text{and} \quad p_{k+1}^c = \hat{p}_{\mu_{k+1}}(x_{k+1}^c) \] it follows from Lemma 14(iii) that \( p_{k+1}^c - \bar{x} = o(|u_{k+1}^c|) + o(|u_k|) = o(|u_k|) \). Finally, from Lemma 14(i), \( s_{k+1}^c = \hat{s}_{\mu_{k+1}}(x_{k+1}^c) = O(|u_{k+1}^c|) = o(|u_k|) \). \[ \square \]

In order to have \( p_{k+1}^c \) be a successful candidate, we can strengthen assumption \( (S)(b)(i) \) by choosing \( \sigma_{k+1} \) to be bounded above by a constant multiple of \( |s_k|^2 \) as follows:

\[ \text{Lemma 16} \quad \text{Suppose that } (S)(a)-(e) \text{ hold and } \sigma_{k+1} \leq \min\{1/(k+2), |s_k|^2/|s_0|^2 \} \text{ for all } k \geq 0. \text{ Then for all } k \text{ sufficiently large, (17) holds and } p_{k+1} = p_{k+1}^c. \]

**Proof.** We start by writing \( f(p_{k+1}^c) - f(p_k) \) as the sum of three difference terms:

\[
(f(p_{k+1}^c) - f(p_{\mu_{k+1}}(x_{k+1}^c))) + (f(p_{\mu_{k+1}}(x_{k+1}^c)) - f(p_{\mu_k}(x_k))) + (f(p_{\mu_k}(x_k)) - f(p_k)).
\]

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

\[
|f(p_{k+1}^c) - f(p_{\mu_{k+1}}(x_{k+1}^c))| \leq L_f |p_{k+1}^c - p_{\mu_{k+1}}(x_{k+1}^c)|
= L_f O(|u_{k+1}^c|) O(|u_k|)
= O(|u_k|) o(|u_k|)
= o(|u_k|^2),
\]

where we used \( \sigma_{k+1} \leq |s_k|^2/|s_0|^2 \) together with Lemma 14(iv), and the fact that \( u_{k+1}^c = o(|u_k|) \) by Lemma 15(ii). The bounds for the other terms are given by Lemma 12. By item (iv) therein and the Lemma 15 result \( u_{k+1}^c = o(|u_k|) \),

\[
f(p_{\mu_{k+1}}(x_{k+1}^c)) - f(p_{\mu_k}(x_k)) \leq -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 + O(|u_{k+1}^c|^2) + o(|u_k|^2)
= -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 + O(o(|u_k|^2)) + o(|u_k|^2)
= -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 + o(|u_k|^2).
\]

While, by item (iii) of Lemma 12 and Lemma 14(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

\[
f(p_{k+1}^c) - f(p_k) \leq -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 + o(|u_k|^2).
\]

By Lemma 15, \( s_{k+1}^c = o(|u_k|)^2 = o(|u_k|^2) \). Therefore, since \( \mu_{k+1} \geq \mu \),

\[
f(p_{k+1}^c) - f(p_k) + \frac{m}{2\mu_{k+1}} |s_{k+1}^c|^2 \leq -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 + o(|u_k|^2).
\]

Since the right hand side of this inequality is dominated by the negative term \( -\frac{1}{2} \lambda_{\min}(\tilde{H}) |u_k|^2 \) as \( u_k \to 0 \), inequality (17) is satisfied and the result is established. \[ \square \]

We conclude by giving our main rate of convergence result.
Theorem 17 Suppose that (S)(a)-(e) hold. Then for all $k$ sufficiently large

(i) $p_{k+1}^e - \bar{x} = o(|p_k - \bar{x}|)$ 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), its $U$-component can be written as $(p_k - \bar{x})_U = (p_k - p_{\mu_k}(x_k))_U + u_k$. By Lemma 14(iii), $(p_k - p_{\mu_k}(x_k))_U = o(|u_k|)$, so, $|(p_k - \bar{x})_U| = o(|u_k|) + u_k$. Combining this with the Lemma 15(ii) result $p_{k+1}^e - \bar{x} = o(|u_k|)$ gives the ratio

$$\frac{p_{k+1}^e - \bar{x}}{|(p_k - \bar{x})_U|} = \frac{o(|u_k|)}{|o(|u_k|) + u_k|}$$

which converges to 0 as $u_k \to 0$. So, $p_{k+1}^e - \bar{x} = o((p_k - \bar{x})_U)$ and the first result follows, because $|(p_k - \bar{x})_U| \leq |p_k - \bar{x}|$.

The second result then follows from Lemma 16. □

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

$$f(x) = \max_{j \in J} f_j(x)$$

where $J$ is finite and each $f_j$ is $C^2$.

For the bundled subgradients we use $g_i = \nabla_x f_{j_i}(y_i)$ where $j_i$ is an active index such that $f_{j_i}(y_i) = f(y_i)$ and $\nabla_x$ indicates differentiation with respect to $x$. In order to test these max-functions with good second order information we do the following: When $s_k = \delta$ in the algorithm then the matrix $H_k$ is set equal to $U_k(\sum_{i \in B} \alpha_i \nabla^2 f_{j_i}(y_i))U_k$ where the multipliers $\alpha_i$ correspond to $\delta$ via (11). 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 $U$-Hessian estimation. When this setting of $H_k$ is not positive definite the $n_k \times n_k$ matrix $\min\{1/k, |s_k|\}I$ is added to it.

For our runs we used the following simple functions:

- F2d, the function in [LS97a], defined for $x \in \mathbb{R}^2$ by $F2d(x) := \max \left\{ \frac{1}{2}(x_1^2 + x_2^2) - x_2, x_2 \right\}$;
- F3d-$U\nu$, four functions of three variables, where $\nu = 3,2,1,0$ denotes the corresponding dimension of the $U$-subspace. Given $\nu := (0,1,1)^T$ and four parameter vectors $\beta^\nu \in \mathbb{R}^4$, for $x \in \mathbb{R}^3$

$$F3d-U\nu(x) := \max \left\{ \frac{1}{2}(x_1^2 + x_2^2 + 0.1x_3^2) - e^T x - \beta^\nu_1 x_1^2 + \beta^\nu_2 x_1^2 - 3x_1^2 - \beta^\nu_3 x_2 x_2 - \beta^\nu_4 x_2^2 - \beta^\nu_5 x_2^2 - \beta^\nu_6 x_2^2 \right\},$$

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>
<thead>
<tr>
<th>Name</th>
<th>$n$</th>
<th>$\dim \mathcal{V}$</th>
<th>$\dim \mathcal{U}$</th>
<th>$f(\bar{x})$</th>
<th>$\bar{x}$</th>
<th>Starting point</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2d</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0.</td>
<td>(0, 0)</td>
<td>$\bar{x} + (0.9, 1.9)$</td>
</tr>
<tr>
<td>F3d-U3</td>
<td>3</td>
<td>0</td>
<td>3</td>
<td>0.</td>
<td>(0, 1, 10)</td>
<td>$\bar{x} + (100, 33, -100)$</td>
</tr>
<tr>
<td>F3d-U2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0.</td>
<td>(0, 0, 10)</td>
<td>$\bar{x} + (100, 33, -100)$</td>
</tr>
<tr>
<td>F3d-U1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0.</td>
<td>(0, 0, 0)</td>
<td>$\bar{x} + (100, 33, -100)$</td>
</tr>
<tr>
<td>F3d-U0</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0.</td>
<td>(1, 0, 0)</td>
<td>$\bar{x} + (100, 33, -100)$</td>
</tr>
</tbody>
</table>

| MAXQUAD | 10 | 4 | 6 | -0.8414080 | see [BGLS03, p. 131] | $x_i = 0$ |

Table 1: Problem data

For comparison purposes, we also solved these problems using N1CV2, the proximal bundle method from [LS07b], available at http://www.inria.fr/lepor/lemanelearchal.

The settings of the Algorithm 6 parameters are $\varepsilon := 10^{-10}$, $m := 10^{-1}$ and $U_0$ equal to the $n \times n$ identity matrix. The sequence $\{\sigma_k\}$ is defined by $\sigma_k := 1/(k^2 + 1)$, and $\underline{\mu}$ and $\{\mu_k\}$ are chosen according to the procedure in N1CV2. The stopping test in N1CV2 is chosen to correspond to the one in Algorithm 6 with the above setting of $\varepsilon$. Both codes use the same quadratic programming problem subroutine as well as the same bundle management strategy, which keeps only active elements. After $x_{k+1}^\ell$ is generated, $\mathcal{B}$ is initialized by appending an index corresponding to $x_{k+1}^\ell$-data, to the $\mathcal{B}$-set associated with $(\hat{p}, \hat{s}) = (p_k, s_k)$ as defined in Section 3. If $p_{k+1}^\ell$ is not a successful candidate then the simple setting $x_{k+1} := \arg\min \{f(p_k), f(p_{k+1}^\ell)\}$ is made. Then, if $x_{k+1} = p_k(p_{k+1}^\ell$, resp.) $\mathcal{B}$ is reinitialized by appending an index corresponding to $p_{k+1}^\ell$-data to the $\mathcal{B}$-set associated with $p_k(p_{k+1}^\ell$, resp.). Since this implementation has no line searches there is the possibility of using an excess overall number of function evaluations.

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. The symbol $!!$ indicates that a maximum number of 500 evaluations was reached before activation of the $\varepsilon$-stopping test.

<table>
<thead>
<tr>
<th></th>
<th>F2d</th>
<th>F3d-U3</th>
<th>F3d-U2</th>
<th>F3d-U1</th>
<th>F3d-U0</th>
<th>MAXQUAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALG6</td>
<td>#f/g</td>
<td>Acc</td>
<td>#f/g</td>
<td>Acc</td>
<td>#f/g</td>
<td>Acc</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>8</td>
<td>5</td>
<td>16</td>
<td>23</td>
<td>9</td>
</tr>
<tr>
<td>44</td>
<td>44</td>
<td>5</td>
<td>48</td>
<td>8</td>
<td>56</td>
<td>4</td>
</tr>
<tr>
<td>N1CV2</td>
<td>#f/g</td>
<td>Acc</td>
<td>#f/g</td>
<td>Acc</td>
<td>#f/g</td>
<td>Acc</td>
</tr>
<tr>
<td>95</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 2: Summary of results

These favorable results demonstrate that it is worthwhile to continue development of this
algorithm, concentrating on the choices of $\mu_k$, $\sigma_k$ and $H_k$ together with the inclusion of line searches. Finally, we note that Algorithm 6 generated final approximations for the subspaces $V$ and $U$ with the proper dimensions for all problems except MAXQUAD, where the final obtained dimensions for $V$ and $U$ were, respectively, 3 and 7 (instead of 4 and 6).

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 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 $\chi$-QP-subproblem constraints and/or using a limited memory quasi-Newton method for estimating the $U$-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).

References


[CG59] E. Cheney and A. Goldstein. Newton's method for convex programming and Tchebycheff-


[GL89] J.C. Gilbert and C. Lemaréchal. Some numerical experiments with variable-storage quasi-

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


[KON98] T.G. Kolda, D.P. O'Leary and J.L. Nazareth. BFGS with update skipping and varying

Boston, MA, 1989.


J. Henry and J-P. Yvon, editors, Systems Modelling and Optimization, number 197 in Lecture


