PROXIMAL GRADIENT $\mathcal{VU}$-METHOD WITH SUPERLINEAR CONVERGENCE FOR NONSMOOTH CONVEX OPTIMIZATION*

SHUAI LIU†, CLAUDIA SAGASTIZÁBAL‡, AND MIKHAIL SOLODOV§

Abstract. The $\mathcal{VU}$-theory for nonsmooth functions and the associated space decomposition have been used for studying the smooth substructures and for developing algorithms with superlinear convergence in those settings, which are challenging for fast convergence. We extend the theory by defining a certain bivariate $\mathcal{U}$-Lagrangian function and partial $\mathcal{U}$-Hessian. Utilizing smoothness properties of the new $\mathcal{U}$-Lagrangian we develop the Proximal Gradient $\mathcal{VU}$-method for continuous nonsmooth convex optimization, and show its superlinear convergence under natural assumptions. The framework consists of a $\mathcal{V}$-step which is a prox-gradient step, followed by a $\mathcal{U}$-step which can be considered as a quasi-Newton step applied to the $\mathcal{U}$-Lagrangian. We show that partial $\mathcal{U}$-Hessians exist for most partly smooth functions. We exhibit the explicit process of constructing a basis of the $\mathcal{U}$-space and of calculating the $\mathcal{U}$-Hessian for ℓ_1-regularized problems. Numerical results illustrate the method's performance.

Key words. $\mathcal{VU}$-decomposition, proximal gradient $\mathcal{VU}$-method, quasi-Newton method, proximal point, nonsmooth optimization, superlinear convergence

MSC codes. 49J52, 90C53, 49J53, 90C99, 58C20

1. Introduction.

We consider the problem

$$\min_{x\in\mathbb{R}^n} f(x)\,, \text{ where } f\colon \mathbb{R}^n \to \mathbb{R} \text{ is a nonsmooth convex function}\,. \tag{1.1}$$

The $\mathcal{VU}$-theory introduced in [22, 24] (closely related to partial smoothness [26]) has been used for the study of smooth structures in nonsmooth functions in [33, 23, 35, 37, 10, 2]. As explained in [44, 31], nonsmoothness is particularly difficult for fast (i.e., superlinear) convergence. Despite this challenging context, the $\mathcal{VU}$-theory provides a favorable setting for the development of superlinearly convergent algorithms [41, 38, 13, 15, 3]; see also [28, 42]. The approach decomposes $\mathbb{R}^n$ into two orthogonal subspaces called $\mathcal{V}$ and $\mathcal{U}$ that capture, respectively, the nonsmoothness and smoothness of f near a given point $\bar{x}$. Through a parametrized function, depending on a subgradient $\bar{g} \in \partial f(\bar{x})$, called the $\mathcal{U}$-Lagrangian, second-order Taylor expansions of f in $\mathcal{U}$ can be obtained if a generalized Hessian (called $\mathcal{U}$-Hessian) exists for the $\mathcal{U}$-Lagrangian.

In the original $\mathcal{VU}$-algorithm [24], the $\mathcal{V}$-step minimizes a prox-regularization of f in the $\mathcal{V}$-subspace, and the $\mathcal{U}$-step makes a Newton-type step where the $\mathcal{U}$-Lagrangian looks smooth. This conceptual approach did not address how to identify the $\mathcal{VU}$-geometry along an algorithmic procedure. In order to pass from theory to computational implementation, it is important to examine the structure of nonsmoothness. Specific $\mathcal{VU}$-theory for finite-max functions and the numerical analysis of the relevant $\mathcal{VU}$-objects were considered in [32] and [13]. A more general class of functions given in [33, 35], said to have primal-dual-gradient structure, identifies a *fast track*, which are points from where fast Newton-type

*Submitted to the editors DATE.

Funding: Research of the first author was supported by the National Natural Science Foundation of China (Grant No. 12001208) and the São Paulo Research Foundation (FAPESP) Grant 2017/15936-2. The second author is supported by CNPq Grant 307509/2023-0 and by PRONEX–Optimization. The third author is supported in part by CNPq Grant 306775/2023-9, by FAPERJ Grant E-26/200.347/2023, and by PRONEX–Optimization.

†School of Accounting, Guangzhou College of Technology and Business, 166 Sanhua Road, Foshan, Guangdong, 528137, China. Phone: 86 757 8831 3138 (shuai0liu@gmail.com).

‡ Universidade Estadual de Campinas, Rua Sérgio Buarque de Holanda 651, Campinas, SP, 13083-859, Brazil. Phone: 55 21 24 33 68 45 (sagastiz@unicamp.br).

§ IMPA - Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, 22460-320, Brazil Phone: 55 21 25 29 52 28 (solodov@impa.br).

steps are possible. Links with the $\mathcal{VU}$-theory of partly smooth functions are the subject of [17, 29, 12, 16]. In particular, [16, Theorem 3.2] establishes a one-to-one correspondence with the fast track and the active manifold of a partly smooth function.

An important step towards implementable $\mathcal{VU}$-algorithms is [34], showing that proximal points are on the fast track. This result suggests that the $\mathcal{V}$-step can be implemented by a prox-step on f when it is easy to compute, or by a bundle method [20] that approximates this step with proximal points of succesive (cutting-plane) models of f. A bundle approach in [10] constructs the $\mathcal{VU}$-decomposition for a certain class that includes max-functions. When f in (1.1) has additive structure as in (3.1) further below, subtle geometrical properties of the proximal operator allow [3, 2] to detect asymptotically the manifold of smoothness (i.e., points on the fast track) by means of a proximal gradient (PG) method [4].

For a nonsmooth optimization method to be fast, it is not enough to define a $\mathcal{V}$-step that lands in the fast track. To accelerate convergence, it is necessary to locally approximate the curvature of the manifold and to perform a suitable correction in the $\mathcal{U}$-subspace. The $\mathcal{U}$-Hessian appears as an appropriate tool in this sense. Fully implementable (fast) $\mathcal{VU}$-algorithms for solving (1.1) are scarce for this reason. Generally, approximating sufficiently well both the (exact proximal) $\mathcal{V}$-step and the (exact) $\mathcal{U}$-Newton direction involves solving at least two quadratic programming problems per iteration. Such is the case of [38], the first $\mathcal{VU}$-algorithm for problems like (1.1), where no specific structure for f is required.

Identifying the fast track and its curvature over iterations, without entering into an oscillatory behaviour, can be tricky. As shown later by Figure 2.1 in Example 2.2, a function can have more than one smooth manifold near a minimizer $\bar{x}$. This happens when the $\mathcal{V}$-component of $\bar{g}$, the subgradient defining the $\mathcal{U}$-Lagrangian, generates multiple fast tracks. Since partial smoothness fixes the manifold and focuses on the $\mathcal{U}$-components, the concept is blind to this effect. To handle this situation properly, we extend the $\mathcal{VU}$-theory as follows. In all the studies mentioned, the $\mathcal{U}$-Lagrangian is a single-variable function, defined by considering $\bar{g} \in \partial f(\bar{x})$ as a parameter. But in an algorithmic framework, we have to deal with a sequence of subgradients ($g^k \in \partial f(x^k)$ at iteration k), which modify the parameter defining the $\mathcal{U}$-Lagrangian over the iterations. To ensure an appropriate management of the $\mathcal{VU}$-objects, we introduce a bivariate $\mathcal{U}$-Lagrangian, where the subgradient is no longer a parameter, but a variable. The general partial Hessian of the new $\mathcal{U}$-Lagrangian defines a partial $\mathcal{U}$-Hessian, that is shown to exist for any convex partly smooth function that satisfies $0 \in \operatorname{ri} \partial f(\bar{x})$.

Our extended $\mathcal{VU}$-theory opens the way to an implementable algorithm, called the Proximal Gradient $\mathcal{VU}$-method or PGVU for short, given in Algorithm 3.1. When applied to functions as in (3.1) with a proximable nonsmooth term, our proposal eliminates the need to solve quadratic programs. This is achieved by an appropriate shift of the optimal subgradient resulting from the PG iterate computation. The structured nonsmoothness of the proximable term also allows us to estimate the $\mathcal{U}$-subspace and its curvature. As shown in our numerical validation, this is essential for fast speed. The proposed PGVU method has superlinear convergence, requiring only the (natural, for potential fast convergence) assumptions of the existence of a positive-definite $\mathcal{U}$-Hessian at a solution $\bar{x}$ such that $0 \in \operatorname{ri} \partial f(\bar{x})$.

In the remainder, Section 1 concludes with the notation. In Section 2, we lay out the foundation of $\mathcal{VU}$-theory for the development of our Proximal Gradient $\mathcal{VU}$-method. We define and study smoothness properties of the bivariate $\mathcal{U}$-Lagrangian function. We show that computing the proximal point can serve as the $\mathcal{V}$-step. In Section 3, the PGVU-method is detailed and its global convergence is proven. Section 4. defines the partial $\mathcal{U}$-Hessian and shows that PGVU is superlinearly convergent under appropriate assumptions. Some of the technical results in Sections 2 and 4 are given in the Appendix. Section 5 applies the method to ℓ_1-regularized minimization and to a multidimensional extension of the function in Example 2.2. We end with concluding remarks in Section 6.

Notation. We mostly follow [43]. Let $\overline{\mathbb{R}} = [-\infty, \infty]$. By $\partial f(x)$ we denote the limiting subdifferential of f at x, and by $\partial^\infty f$ the horizon subdifferential of f. This is needed to refer to some cited results. Of course, for a convex finite-valued f, $\partial f(x)$ is is the usual subdifferential in Convex Analysis. The notation $f'(x;d)$ stands for the directional derivative at x in the direction d. For a smooth bivariate function $f(x,y)$, $\nabla_x f(x,y)$ and $\nabla^2_{xx} f(x,y)$ are the partial gradient and partial Hessian of f with respect to the variable x. For given points $\bar{x}$ and $\bar{y}$, the partial subdifferential $\partial_x f(\bar{x},\bar{y})$ is defined to be the subdifferential of $f(\cdot,\bar{y})$ at $\bar{x}$. For a convex set C, its indicator function is $\delta_C(\cdot)$, and its interior and relative interior are, respectively, $\operatorname{int} C$ and $\operatorname{ri} C$. The distance of a point x to a set C is $\operatorname{dist}(x;C) := \inf_{z\in C}\|z-x\|$. The Euclidean closed ball in $\mathbb{R}^n$ centered at $\bar{x}$ with radius $\epsilon \geq 0$ is denoted by $B(\bar{x},\epsilon)$ and the ball in $\mathbb{R}^m$ is $B^m(\bar{x},\epsilon)$. For a function f, its minimal value is denoted by f^* and its set of minimizers by S. The vector $e^j \in \mathbb{R}^n$ has all of its components null, except for $e^j_j = 1$. Regarding convergence rates, the notation "little o" in [40] for scalars is used for vectors, as follows. For vector sequences $\mathbb{R}^n \supset \{x^k\} \to \bar{x}$ and $\mathbb{R}^m \supset \{y^k\} \to \bar{y}$, and for $\|\cdot\|$ the Euclidean norm in the corresponding space, $x^k = o(y^k)$ is short hand for "$\forall \varepsilon > 0\,, \exists K : \|x^k\| \leq \varepsilon\|y^k\|$ for all $k \geq K$". The notation for "big O" term is used in a similar manner. The class of twice continuously differentiable functions is C^2.

2. Elements of the $\mathcal{VU}$-theory.

We start with defining the two subspaces in question.

Definition 2.1 ($\mathcal{VU}$-decomposition). *Given a convex function $f\colon \mathbb{R}^n \to \mathbb{R}$ and a point $\bar{x}$, the* $\mathcal{VU}$-decomposition *of $\mathbb{R}^n$ associated with f and $\bar{x}$ is defined by the subspaces*

$$\mathcal{V}(\bar{x}) = \operatorname{span}(\partial f(\bar{x}) - g),\ \mathcal{U}(\bar{x}) = \mathcal{V}(\bar{x})^\perp\,, \text{where } g \text{ is an arbitrary subgradient in } \partial f(\bar{x}).$$

The respective dimensions are $m = \dim \mathcal{U}(\bar{x})$ and $n - m = \dim \mathcal{V}(\bar{x})$. As vector spaces, $\mathcal{V}(\bar{x})$ and $\mathcal{U}(\bar{x})$ are endowed with a scalar product and norm induced from $\mathbb{R}^n$. When clear from the context, the short forms $\mathcal{V}$ and $\mathcal{U}$ are used below.

The algebraic form of the decomposition depends on two matrices:

$$\begin{array}{rcl} \mathbb{R}^{n\times m} \ni U: & = & \text{a matrix whose columns form an orthonormal basis for } \mathcal{U} \\ \mathbb{R}^{n\times(n-m)} \ni V: & = & \text{a matrix whose columns form a basis for } \mathcal{V}\,, \\ & & \text{with Moore-Penrose pseudo-inverse } V^\dagger := (V^\mathsf{T}V)^{-1}V^\mathsf{T}\,. \end{array}$$

More specifically, any $x \in \mathbb{R}^n$ has $\mathcal{U}$-and $\mathcal{V}$-components $x_u := U^\mathsf{T}x$ and $x_v := V^\dagger x$. By the definitions of $\mathcal{V}$ and $\mathcal{U}$, the set $U^\mathsf{T}\partial f(\bar{x})$ is a singleton, and hence,

$$\bar{g}_u := U^\mathsf{T} g \text{ for any } g \in \partial f(\bar{x})\,. \tag{2.1}$$

2.1. The $\mathcal{U}$-Lagrangian.

Given $\bar{g} \in \partial f(\bar{x})$, the single-variable $\mathcal{U}$-*Lagrangian* of f is

$$\mathbb{R}^m \ni u \mapsto L^{\bar{g}}_U(u) := \inf_{w\in\mathbb{R}^{n-m}} \left\{ f(\bar{x} + Uu + Vw) - \left\langle \bar{g}_v, V^\mathsf{T}Vw \right\rangle \right\}\,.$$

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

$$W^{\bar{g}}(u) := \left\{ w \in \mathbb{R}^{n-m} \colon L^{\bar{g}}_U(u) = f(\bar{x} + Uu + Vw) - \left\langle \bar{g}_v, V^\mathsf{T}Vw \right\rangle \right\}\,.$$

By its definition, the $\mathcal{U}$-Lagrangian is finite-valued and convex on $\mathbb{R}^n$. When, in addition, $\bar{g}_v \in V^\dagger \operatorname{ri} \partial f(\bar{x})$, it is shown in [24, Theorem 3.2, Theorem 3.3(ii)] that

$$W^{\bar{g}}(0) = \{0\}\,, L^{\bar{g}}_U(0) = f(\bar{x}), L^{\bar{g}}_U \text{ is differentiable at 0 with } \nabla L^{\bar{g}}_U(0) = \bar{g}_u\,. \tag{2.2}$$

A special first-order expansion for f is obtained through $\mathcal{V}$-minimizers:

$$\forall w^{\bar{g}}(u) \in W^{\bar{g}}(u),\ f(\bar{x}+Uu+Vw^{\bar{g}}(u)) = f(\bar{x}) + \langle \bar{g}_u, u\rangle + \left\langle \bar{g}_v, V^\top V w^{\bar{g}}(u)\right\rangle + o(Uu)\,.$$

When a second-order approximation for f exists along the $\mathcal{U}$-subspace, a Newton-like step is possible, opening the way to superlinearly convergent schemes; see Sections 2.3 and 4.

We introduce in this work an important advance with respect to the previous $\mathcal{VU}$-literature. It has to do with the following considerations. The original $\mathcal{U}$-Lagrangian was defined in [24] for some fixed $\bar{g} \in \operatorname{ri}\partial f(\bar{x})$. By (2.1), the $\mathcal{U}$-component $\bar{g}_u$ is the same for all $\bar{g} \in \partial f(\bar{x})$, but the relation is not valid for the $\mathcal{V}$-component $\bar{g}_v$. If, as in the example below, the value of $\bar{g}_v$ modifies the $\mathcal{V}$-minimizer, different $\mathcal{U}$-Lagrangians emerge from different $\bar{g}_v$.

Example 2.2 (A nonsmooth function with multiple fast tracks). Given a scalar $a > 0$,

$$\text{for } (u,v) \in \mathbb{R}^2 \text{ the function } F(u,v) = \frac{a}{2}u^2 + \max\left\{\frac{a}{2}u^2, |v|\right\}$$

is differentiable everywhere except for points satisfying the equation $|v| = \frac{a}{2}u^2$. Its unique minimizer is $\bar{x} = (0,0)$, where the subdifferential is $\partial F(\bar{x}) = \{0\} \times [-1,1]$. Figure 2.1

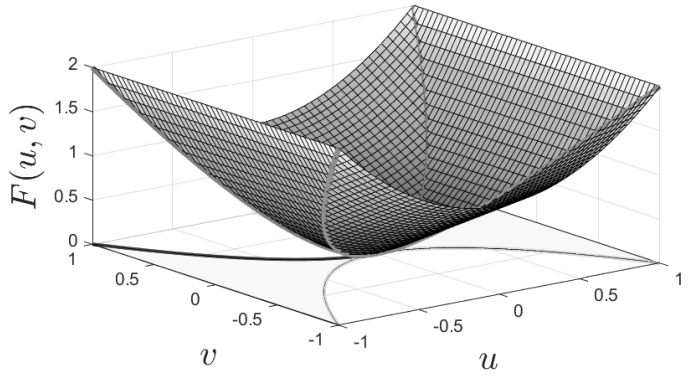

Fig. 2.1. *Function F in Example 2.2, has two $\mathcal{V}$-minimizer sets, and two fast tracks*

shows that the graph of F is $\mathcal{U}$-shaped along the u-axis and $\mathcal{V}$-shaped along the v-axis. The $\mathcal{VU}$-decomposition at $\bar{x}$ gives $\mathcal{U} = \mathbb{R}\times\{0\}$ and $\mathcal{V} = \{0\}\times\mathbb{R}$. Then for any $\bar{g} \in \operatorname{ri}\partial F(\bar{x}) = \{0\}\times(-1,1)$ we have $\bar{g}_v \in (-1,1)$. (When clear that a point is in $\mathbb{R}\times 0$ or $0\times\mathbb{R}$, we omit the 0 component.) Working out the calculations of the three cases for the $\mathcal{V}$-minimizers,

$$W^{\bar{g}}(u) = \begin{cases} \left\{\frac{a}{2}u^2\right\}, & \text{if } \bar{g}_v \in (0,1),\\ \left\{v\colon |v| \le \frac{a}{2}u^2\right\}, & \text{if } \bar{g}_v = 0,\\ \left\{-\frac{a}{2}u^2\right\}, & \text{if } \bar{g}_v \in (-1,0). \end{cases} \quad\Longrightarrow\quad L_U^{\bar{g}}(u) = (2-|\bar{g}_v|)\frac{a}{2}u^2\,.$$

Notice the dependence of the $\mathcal{U}$-Lagrangian on the chosen subgradient.

The hidden smoothness in F relates to the trajectory below:

$$\chi^{\bar{g}}(u) = \left\{\bar{x} + (u, v^{\bar{g}}(u))\,:\ \text{for } u\in\mathbb{R} \text{ and } v^{\bar{g}}(u) = \frac{a}{2}\operatorname{sign}(\bar{g}_v)u^2 \in W^{\bar{g}}(u)\right\}\,.$$

This defines a *fast track* [34], along which F can be expanded up to second order:

$$F(\chi^{\bar{g}}(u)) = au^2,\ \nabla_u F(\chi^{\bar{g}}(u)) = 2au, \text{ and } \nabla^2_{uu}F(\chi^{\bar{g}}(u)) = 2a\,.$$

Taking $\bar{g}_v = \pm\frac{1}{2}$ gives two different activity manifolds (as defined by the concept of partial smoothness), with the same curvature $(2a)$.

The situation illustrated by Example 2.2 (trajectories of smoothness depending on the choice of the subgradient $\bar{g}$), motivates the consideration of a *bivariate* $\mathcal{U}$-Lagrangian, where subgradients are treated as variables (rather than parameters).

2.2. Two-variable $\mathcal{U}$-Lagrangian. Pursuing further the analysis of the impact of g_v on the $\mathcal{U}$-objects, we now consider extending the $\mathcal{VU}$-decomposition theory to a setting in which g_v is an *argument* of the $\mathcal{U}$-Lagrangian. Rather than depending only on u, the function has primal and dual variables, that is, $(u, \bar{g}_v) \in \mathbb{R}^m \times \mathbb{R}^{n-m}$.

Definition 2.3 ($\mathcal{U}$-Lagrangian with two variables). *The bivariate* $\mathcal{U}$-Lagrangian *of* f *is defined from* $\mathbb{R}^m \times V^\dagger \partial f(\bar{x})$ *to* $\overline{\mathbb{R}}$ *as follows:*

$$\mathbb{R}^m \times V^\dagger \partial f(\bar{x}) \ni (u, g_v) \mapsto L_U(u, g_v) := \inf_{w \in \mathbb{R}^{n-m}} \left\{ f(\bar{x} + Uu + Vw) - \left\langle g_v, V^\mathsf{T} V w \right\rangle \right\},$$

and the associated set of $\mathcal{V}$*-space minimizers is*

$$W(u, g_v) := \left\{ w \in \mathbb{R}^{n-m} \colon L_U(u, g_v) = f(\bar{x} + Uu + Vw) - \left\langle g_v, V^\mathsf{T} V w \right\rangle \right\}.$$

The $L_U(u, \bar{g}_v)$ notation in this work should not be confused with [34], where only u is a variable, and the semicolon in $L_U(u; \bar{g}_v)$ was used to expose that $\bar{g}_v$ is a parameter.

Consider $g \in \operatorname{ri} \partial f(\bar{x})$, the subdifferential relative interior defined by $\operatorname{ri} \partial f(\bar{x}) := \{g \in \mathbb{R}^n \colon g + \operatorname{int} B(0, \eta) \cap \mathcal{V} \subset \partial f(\bar{x}) \text{ for some } \eta > 0\}$. Then we have $U\bar{g}_u + Vg_v + \frac{\eta Vw}{\|Vw\|} \in \partial f(\bar{x})$ for all $w \in \mathbb{R}^{n-m}$ and the convexity of f implies that, for any $(u, w) \in \mathbb{R}^m \times \mathbb{R}^{n-m}$,

$$f(\bar{x} + Uu + Vw) \geq f(\bar{x}) + \langle \bar{g}_u, u \rangle + \left\langle g_v, V^\mathsf{T} V w \right\rangle + \eta \|Vw\|. \tag{2.3}$$

In order to deal with variations on the $\mathcal{U}$-component properly, in a manner that is uniform relative to interior subgradients, for a small positive number η we define the closed subset

$$G_v(\bar{x}) := V^\dagger (\eta\text{-}\operatorname{ri}) \partial f(\bar{x}) \text{ for } (\eta\text{-}\operatorname{ri}) \partial f(\bar{x}) := \{g \in \mathbb{R}^n \colon g + B(0, \eta) \cap \mathcal{V} \subset \partial f(\bar{x})\}. \tag{2.4}$$

In Theorem 2.4 we replicate in our new bivariate context several important relations known for the single-variable $\mathcal{U}$-Lagrangian. We refer to the Appendix for its proof.

Theorem 2.4 (Smoothness of bivariate $\mathcal{U}$-objects). *Given* $G_v(\bar{x})$ *from* (2.4)*, the bivariate* $\mathcal{U}$*-Lagrangian and the* $\mathcal{V}$*-space minimizer set from Definition 2.3 satisfy the following properties.*

(i) L_U *is finite-valued on* $\mathbb{R}^m \times G_v(\bar{x})$*;*
(ii) W *is outer semi-continuous and locally bounded on* $\mathbb{R}^m \times G_v(\bar{x})$*;*
(iii) $W(0, g_v) = \{0\}$ *and* W *is continuous at* $(0, g_v)$ *for any* $g_v \in G_v(\bar{x})$*;*
(iv) L_U *is locally Lipschitz continuous on the interior of* $\mathbb{R}^m \times G_v(\bar{x})$*;*
(v) L_U *is differentiable at* $(0, g_v)$ *for any* g_v *in the interior of* $G_v(\bar{x})$*, with* $\nabla L_U(0, g_v) = (\bar{g}_u, 0)$*;*
(vi) *For all* $(u, g_v) \in \mathbb{R}^m \times V^\dagger \operatorname{ri} \partial f(\bar{x})$*, and an arbitrary point* w *in* $W(u, g_v)$*,*

$$\partial_u L_U(u, g_v) = \{s_u \colon s \in \partial f(\bar{x} + Uu + Vw),\ s_v = g_v\}.$$

(vii) $L_U(u, g_v) = f(\bar{x}) + \langle \bar{g}_u, u \rangle + o\Big(Uu + V(g_v - \bar{g}_v)\Big)$ *for any* $\bar{g} \in \operatorname{ri} \partial f(\bar{x})$*,* $g_v \in G_v(\bar{x})$*.*

All the properties listed in (2.2), with only u considered a variable, can now be compared with the statements in items (iii) and (v) of Theorem 2.4, shown in the bivariate setting.

2.3. $\mathcal{V}$-minimizers and proximal points. When the special trajectory associated with $\mathcal{V}$-minimizers is identified, the function appears smooth along the resulting $\mathcal{U}$-subspace, and a $\mathcal{U}$-Newton step is possible. For a $\mathcal{VU}$-method to be superlinearly convergent, the fast $\mathcal{U}$-step should dominate over the $\mathcal{V}$-step. In this respect, the behavior of $\mathcal{V}$-minimizers in the set $W(u, g_v)$ from Definition 2.3 is crucial.

For the original $\mathcal{U}$-Lagrangian, [24, Corollary 3.5] shows that $\mathcal{V}$-minimizers are *tangent* to the $\mathcal{U}$-subspace. The same important result holds for our new bivariate $\mathcal{U}$-Lagrangian, as combining item (vii) in Theorem 2.4 together with the $\mathcal{U}$-Lagrangian Definition 2.3, we obtain the following first-order expansion for f:

$$f(\bar{x} + Uu + Vw(u, g_v)) = f(\bar{x}) + \langle \bar{g}_u, u\rangle + \left\langle g_v, V^\mathsf{T} V w(u, g_v)\right\rangle + o\Big(Uu + Vg_v\Big)\,.$$

When compared with the relation given in Section 2.1 for the single-variable setting, we see that for fast convergence purposes, not only the $\mathcal{U}$-component (u) should dominate eventually, but also the $\mathcal{V}$-component of the subgradient (g_v) should vanish. In Algorithm 3.1 below, this is achieved by adding the $\mathcal{U}$-Newton step to the proximal gradient update.

Having highlighted the importance of $\mathcal{V}$-minimizers, we now show that they can be identified by means of the proximal point mapping. For any function $h\colon \mathbb{R}^n \to \mathbb{R}$ and a real number $\mu > 0$, recall that the proximal mapping is given by

$$\operatorname{prox}_{h,\mu}(x) := \operatorname*{arg\,min}_{y\in\mathbb{R}^n} \left\{h(y) + \frac{\mu}{2}\|y - x\|^2\right\}\,.$$

We now provide a mechanism that makes the $\mathcal{V}$-step be tangential to the $\mathcal{U}$-subspace.

Lemma 2.5 (Characterization of $\mathcal{V}$-minimizers). *Let $g \in \operatorname{ri}\partial f(\bar{x})$, $\bar{x} \in \mathbb{R}^n$. For any $p \in \mathbb{R}^n$ and the corresponding $\mathcal{VU}$-components $u(p) := (p-\bar{x})_u$ and $v(p) := (p-\bar{x})_v$, it holds that $v(p) \in W(u(p), g_v)$ if and only if $g_v \in V^\dagger \partial f(p)$, in which case $g_u \in \partial_u L_U(u(p), g_v)$. If, in addition, there is $g' \in \partial f(p)$ such that $g'_v = g_v$, then $g'_u \in \partial_u L_U(u(p), g_v)$.*

Proof. The convex function $\mathbb{R}^{n-m} \ni v \mapsto h(v) := f(\bar{x} + Uu(p) + Vv)$ has the subdifferential $\partial h(v) = V^\mathsf{T}\partial f(\bar{x} + Uu(p) + Vv)$. The necessary and sufficient optimality condition for $v(p) \in W(u(p); g_v)$ is $0 \in \partial h(v(p)) - V^\mathsf{T} V g_v = V^\mathsf{T}\partial f(\bar{x} + Uu(p) + Vv(p)) - V^\mathsf{T} V g_v = V^\mathsf{T}\partial f(p) - V^\mathsf{T} V g_v$ and the equivalence follows from the definition of the pseudo-inverse $V^\dagger$. To show that $g_u \in \partial_u L_U(u(p), g_v)$, note that if $g' \in \partial f(p)$ and $g'_v = g_v$, then $g_v \in V^\dagger \partial f(p)$. Hence, the expression of $\partial_u L_U(\cdot, g_v)$ in Theorem 2.4 is verified by g'_u. ∎

Proximal points are related to $\mathcal{V}$-minimizers through Lemma 2.5, by taking the subgradient in the optimality condition of the proximal point problem. Specifically, for given $z \in \mathbb{R}^n$ and $\mu > 0$, the result is applied with $p = \operatorname{prox}_{f,\mu}(z)$ and $g' = \mu(z - p)$; see Theorem 3.1(i). In the algorithm given in next section, however, the $\mathcal{V}$-step does not compute exact proximal points. Rather, having a model function for f, the proximal point of the model is computed. By exploiting structural properties of the function to be minimized, given in (3.1) below, we can rewrite the model proximal point as an *exact* proximal point of f, by shifting the prox-center; see Theorem 3.1. Thanks to this shifting, the result in Lemma 2.5 applies.

3. The algorithm and its global convergence. We next focus our attention on the function in (1.1) having the following additive structure:

(3.1)

$f(x) \equiv q(x) + h(x)$	
for $q\colon \mathbb{R}^n \to \mathbb{R}$	convex, C^2-smooth
	with locally Lipschitz continuous gradient, and
$h\colon \mathbb{R}^n \to \mathbb{R}$	a continuous convex function, possibly nonsmooth,
	with an easy-to-compute proximal point.

In the considered context, for the family of model functions

$$m(x;y) := q(y) + \langle \nabla q(y), x - y\rangle + h(x)\,, \tag{3.2}$$

it is easy to compute the proximal point of $m(\cdot\,; y)$, for any parameter $y \in \mathbb{R}^n$.

3.1. $\mathcal{V}$-step: proximal gradient iterations. To minimize a function f as in (3.1), the well-known proximal gradient algorithm [39, 5] computes the proximal point of the model (3.2). At iteration k, given the current iterate x^k and a prox-parameter μ_k, the next point is $x^{k+1} = \operatorname{prox}_{m(\cdot;x^k),\mu_k}(x^k)$. If μ_k is greater than or equal to ∇q's global Lipschitz constant and f satisfies an error bound, the PG iterates converge with linear rate [11, Theorems 3.1 and 5.5]; see also [1]. To achieve superlinear speed, in Algorithm 3.1 those iterations are corrected by a suitable $\mathcal{U}$-step. The PG iterative Procedure 1 corresponds to our $\mathcal{V}$-step.

Procedure 1: Proximal Gradient (the $\mathcal{V}$-step)

Input: f as in (3.1), $x^k \in \mathbb{R}^n$, $\mu_k > 0$, and $p = \operatorname{prox}_{m(\cdot;x^k),\mu_k}(x^k)$ for model (3.2).
while $q(p) > q(x^k) + \left\langle \nabla q(x^k), p - x^k \right\rangle + \frac{\mu_k}{2}\|p - x^k\|^2$ **do**
  declare a null step: $\mu_k \leftarrow 2\mu_k$
  compute $p = \operatorname{prox}_{m(\cdot;x^k),\mu_k}(x^k)$
end
Output: μ_k and $p^k = p$.

The assumptions in (3.1) ensure that this procedure is finite. To see this,

$$\begin{array}{l} \text{considering the output } p^k \text{ of Procedure 1 and its input } x^k, \\ \text{and given a neighborhood } \mathcal{N}(x^k) \text{ containing } p^k, \\ \text{the corresponding local Lipschitz constant of } \nabla q \text{ is denoted by } \beta_k\,. \end{array} \tag{3.3}$$

Then, since p^k satisfies $q(p^k) \le q(x^k) + \left\langle \nabla q(x^k), p^k - x^k \right\rangle + \frac{\mu_k}{2}\|p^k - x^k\|^2$, by the model definition (3.2), the procedure will terminate once $\mu_k \ge \beta_k$ (in the bundle method terminology [20], the sequence of null steps is always finite).

Additionally, note that if in (3.1) there is no smooth term, then $q \equiv 0$ and $h \equiv f$. In this case the model is the same function (assuming the prox-calculation of f is easy), the $\mathcal{V}$-step performs an exact proximal step for f and there are no null steps.

The procedure output is the proximal point of the model (3.2) at x^k. In order to apply Lemma 2.5, and in this way identify the output with a $\mathcal{V}$-minimizer, p^k needs to be the proximal point of the function, and not of its model. This is shown in our next result, where we exhibit p^k to be the *exact* proximal point of f at a certain shifted point.

THEOREM 3.1 (Shifting a model proximal point to exact proximal point of the function). *Given the output p^k of Procedure 1, define*

$$g^k := \mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k) \quad \textit{and} \quad z^k := p^k + \frac{1}{\mu_k} g^k\,. \tag{3.4}$$

Then it holds that

$$p^k = \operatorname{prox}_{f,\mu_k}(z^k)\,, \quad z^k = x^k + \frac{1}{\mu_k}(\nabla q(p^k) - \nabla q(x^k))\,, \quad g^k \in \partial f(p^k)\,. \tag{3.5}$$

Therefore, for $\bar{x} \in S$ a minimizer of f in (1.1), the following holds.

(i) *Setting $u^k := (p^k - \bar{x})_u$, the corresponding $\mathcal{V}$-component $v^k := (p^k - \bar{x})_v \in W(u^k, g_v^k)$ if and only if $g_v^k \in V^\dagger \operatorname{ri} \partial f(\bar{x})$, in which case $g_u^k \in \partial_u L_U(u^k, g_v^k)$.*

(ii) *Furthermore, whenever $\mu_k > \beta_k$ from (3.3), it holds that*

$$\|p^k - \bar{x}\| \leq \frac{\mu_k + \beta_k}{\mu_k - \beta_k}\|x^k - \bar{x}\| \quad \textit{and} \quad \|g^k\| \leq \frac{2\mu_k(\mu_k + \beta_k)}{\mu_k - \beta_k}\|x^k - \bar{x}\|\,.$$

Proof. We have that $\mu_k(x^k - p^k) \in \partial m(p^k; x^k) = \nabla q(x^k) + \partial h(p^k)$, which yields $\mu_k(x^k - p^k) - \nabla q(x^k) \in \partial h(p^k)$. Therefore, $\mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k) \in \nabla q(p^k) + \partial h(p^k) = \partial f(p^k)$. The remaining assertions in (3.5) follow from the optimality condition $0 \in \partial f(p^k) + \mu_k(p^k - z^k)$, i.e., $\mu_k(z^k - p^k) \in \partial f(p^k)$ which by (3.4) is just $g^k \in \partial f(p^k)$. Item (i) follows from (3.5) and Lemma 2.5, written with p, g' therein replaced by p^k, g^k. Finally, for (ii) first note that, by (3.4), it holds that

$$\|g^k\| \leq (\mu_k + \beta_k)\|x^k - p^k\| \leq (\mu_k + \beta_k)\Big(\|x^k - \bar{x}\| + \|\bar{x} - p^k\|\Big)\,, \tag{3.6}$$

Next, by (3.5) and the nonexpansiveness of the proximal operator,

$$\|\bar{x} - p^k\| = \|\operatorname{prox}_{f,\mu_k}(\bar{x}) - \operatorname{prox}_{f,\mu_k}(z^k)\| \leq \|\bar{x} - z^k\|\,.$$

Using the expression for z^k in (3.5) and the assumption of local Lipschitz continuity of ∇q, we obtain that

$$\|\bar{x} - p^k\| \leq \|\bar{x} - x^k\| + \frac{1}{\mu_k}\|\nabla q(p^k) - \nabla q(x^k)\| \leq \|\bar{x} - x^k\| + \frac{\beta_k}{\mu_k}\|p^k - x^k\|\,.$$

Adding $0 = \bar{x} - \bar{x}$ in the right-most term, gives

$$\|\bar{x} - p^k\| \leq \|\bar{x} - x^k\| + \frac{\beta_k}{\mu_k}\|p^k - \bar{x}\| + \frac{\beta_k}{\mu_k}\|\bar{x} - x^k\|\,.$$

After some rearrangements of terms, we obtain that

$$\|\bar{x} - p^k\| \leq \frac{1 + \frac{\beta_k}{\mu_k}}{1 - \frac{\beta_k}{\mu_k}}\|\bar{x} - x^k\| \;= \frac{\mu_k + \beta_k}{\mu_k - \beta_k}\|\bar{x} - x^k\|\,.$$

The last inequality in (3.6) yields the final result. ∎

The explicit shifting in Theorem 3.1 is possible thanks to the structure of f in (3.1). Note that the tangential property in (i) depends on g_v^k eventually becoming (the $V^\dagger$ component of) an interior subgradient at a minimizer. To achieve this, the $\mathcal{VU}$-algorithm drives to zero the subgradient g^k from (3.4). Accordingly, the Proximal Gradient $\mathcal{VU}$-method given in Algorithm 3.1, stops when $\|g^k\| \leq \text{TOL}$ for a given tolerance parameter TOL.

3.2. $\mathcal{U}$-step and the algorithm statement. After the $\mathcal{V}$-step is done, the output of Procedure 1 is corrected as follows:

$$x^{k+1} = p^k - U_k Q_k U_k^\mathsf{T} g^k\,, \tag{3.7}$$

where U_k is a certain orthonormal matrix and Q_k is positive semidefinite. The purpose of this correction is to track eventually a trajectory where the function behaves smoothly, through the relation of $\mathcal{V}$-minimizers with proximal points. Thus, g^k in (3.7) is the shifted gradient from (3.4), the inverse of Q_k approximates asymptotically the so-called *partial $\mathcal{U}$-Hessian* (a second-order object related to the $\mathcal{U}$-Lagrangian defined in Section 4), and the orthonormal matrix U_k is a basis for a subspace $\mathcal{U}_k$ that approximates $\mathcal{U}(p^k)$. For the latter, see [10], and also Section 5 concerning the ℓ_1-regularized setting which is one illustration in this paper.

Algorithm 3.1 Proximal Gradient $\mathcal{VU}$-method (PGVU)

Data: f as in (3.1), $x^0 \in \mathbb{R}^n$, $\mu_0 > 0$, and a stopping tolerance $\text{TOL} \geq 0$. Set $k = 0$.
repeat
  Obtain μ_k and p^k from Procedure 1.
    Define (shift the subgradient) $g^k := \mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k)$.
    Compute an orthonormal basis $U_k \in \mathbb{R}^{n\times n_k}$.
    Choose a symmetric positive semidefinite matrix $Q_k \in \mathbb{R}^{n_k\times n_k}$.
  Update $x^{k+1} = p^k - U_k Q_k U_k^\top g^k$, set $k = k + 1$.
until $\|g^k\| \leq \text{TOL}$.

The full proximal gradient $\mathcal{VU}$-method is given in Algorithm 3.1, where the stopping criterion is justified by the fact that $g^k \to 0$, shown in Theorem 3.3.

When f is differentiable, $\mathcal{U}(p^k)$ is the whole space, its basis is the identity matrix, and Q_k can be defined as usual for quasi-Newton methods; see, e.g., [21]. Otherwise, the structural properties for f are essential to define suitable matrices in (3.7). Namely, as $\partial f = \nabla q + \partial h$, in Definition 2.1 the $\mathcal{V}$-subspaces of f, h, and the model m in (3.2) are all identical:

$$\mathcal{V}(p) := \mathcal{V}f(p) = \mathcal{V}h(p) = \mathcal{V}m(p; y) \quad \text{for all } p, y \in \mathbb{R}^n .$$

Hence, also $\mathcal{U}(p) := \mathcal{U}f(p) = \mathcal{U}h(p) = \mathcal{U}m(p; y)$.

To give an insight/illustration, we consider again our function from Example 2.2, and compare the performance of three methods, according to the possible choices.

Example 3.2 (Proximal, PG and PGVU algorithms). For the function F from Example 2.2, the gradient of the smooth term has the global Lipschitz constant $\beta = a = 2$.

We consider minimizing F with the proximal point method (P), the proximal gradient algorithm, and PGVU as given in Algorithm 3.1, with the following specifications:

- The implementation of both P and PG follows Procedure 1, with respective model functions $F(\cdot)$ and $m(\cdot, x^k)$ from (3.2), and stopping test $\max(\|x^k\|, \|p^k\|) \leq \text{TOL}$.
- For PGVU, there is the additional $\mathcal{U}$-step, requiring the matrices in (3.7).
  - Outside of the fast track, that is when $p^k \notin \{(p_1, p_2) : |p_2| = \frac{a}{2}(p_1)^2\}$, the function is differentiable, $\mathcal{U}(p^k) = \mathbb{R}^2$ and U_k is the identity matrix. The matrix Q_k is set to $\begin{bmatrix} 2a & 0 \\ 0 & 0 \end{bmatrix}^\dagger$ if $|p_2^k| < \frac{a}{2}(p_1^k)^2$, and $\begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}^\dagger$ if $|p_2^k| > \frac{a}{2}(p_1^k)^2$.
  - When p^k is on the fast track, its $\mathcal{V}(p^k)$ subspace is generated by $\partial h(p^k) = \{(\xi a p_1^k, (1-\xi)\operatorname{sign}(p_2^k)) : \xi \in [0, 1]\}$ except at $\bar{x}$. In this case, U_k is the orthonormal vector generated from $\left(1, \operatorname{sign}(p_2^k) a p_1^k\right)$ and $Q_k = (2a)^{-1}$.

The methods were run with the same μ_0 and x^0, until the termination tolerance $\text{TOL} = 10^{-8}$ or a maximum number of iterations set to 50 is attained. Table 3.1 reports the number of iterations and accuracy of the three algorithms, for two different initial values of μ_0 and x^0.

When μ_0 is small, the performances of P, PG, and PGVU are similar. This is explained by the "null-step" inner loop in Procedure 1, which makes all the three methods increase μ_k until a value larger than $\beta = 2.0$ is attained. By contrast, the runs with the large value of μ_0 are troublesome for P and PG. Procedure 1 always accepts the large value $\mu_k = \mu_0 = 10$, there is no backtracking mechanism to reduce μ_k. This is prejudicial for P and PG, as confirmed when examining the three trajectories generated for one of the starting points in Figure 3.1.

These plots highlight the following important point: *for a nonsmooth optimization method to achieve high accuracy, identifying the fast track is necessary, but it is not sufficient.* Curvature information, the $\mathcal{U}$-step in PGVU, is also needed. Both P and PG iterates eventually

TABLE 3.1
Methods' performance

$\mu^0 = 0.18$	$x^0 = (-1.2, 2.2)^\intercal$			$x^0 = (-1.1, 0.76)^\intercal$		
	P	PG	PGVU	P	PG	PGVU
Iterations	50	50	10	50	50	50
Digits	7	8	8	7	8	8
$\mu^0 = 10$	$x^0 = (-1.2, 2.2)^\intercal$			$x^0 = (-1.1, 0.76)^\intercal$		
	P	PG	PGVU	P	PG	PGVU
Iterations	50	50	24	50	50	9
Digits	5	5	10	5	6	8

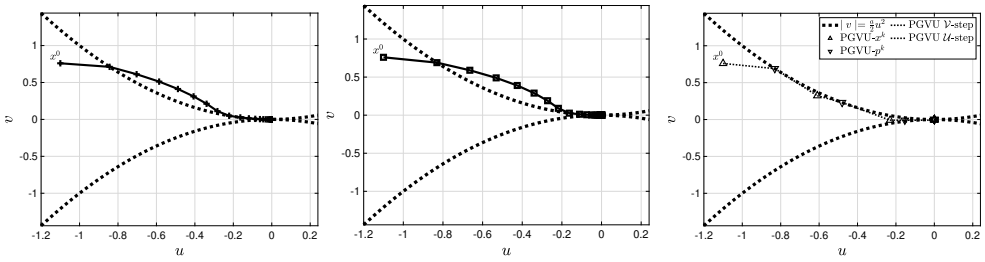

FIG. 3.1. *P, PG, and PGVU trajectories (left, middle, right), when minimizing F from $x^0 = (-1.1, 0.76)^\intercal$, starting with $\mu_0 = 10$. Because $\mu_k = \mu_0$ is too large, P and PG take very small steps along the fast track, which hinders their performance. By contrast, thanks to its $\mathcal{U}$-step, PGVU approaches the fast track tangentially and reaches rapidly a much better degree of accuracy.*

land on the smooth trajectory defined by the fast track. This is consistent with the theory in [34]; see also [3, Theorem 3.1]. As illustrated by the left and middle plots in Figure 3.1, because $\mu_k = \mu_0$ is too large, the progress between consecutive iterates becomes too small. On the right plot we see that the too large $\mu_k = \mu_0$ also makes PGVU take a relatively small $\mathcal{V}$-step, but afterwards its $\mathcal{U}$-step is long, in the "right direction" to solution, but also deviating the iterate from the fast track. Subsequent $\mathcal{V}$-steps bring iterates back to the fast track, where a new fast $\mathcal{U}$-step is possible. The overall process can be interpreted to work in a "predictor-corrector" fashion. A first conclusion of this basic comparison is that when accuracy is important, PGVU is the approach to be preferred. If accuracy is not a concern, the simpler PG technique is sufficient. Of course, one PGVU iteration involves more computational work. Nevertheless, for the larger instances in Subsection 5.2, tested giving to PGVU the same time budget as to PG, PGVU remains more accurate (PG reaches in general 2 or 4 digits of accuracy at most).

3.3. Global convergence and some other asymptotic results. In Theorem 3.3 below about the global convergence result, we do not show that the sequence $\{x^k\}$ is bounded (has accumulation points). This is in part related to the comment above that in some naturally included special cases which do not involve nonsmoothness, our method can actually reduce to the usual smooth quasi-Newton technique. It is known that proving boundedness of the usual quasi-Newton updates (say, BFGS) without very technical complex modifications to the update, is not possible; see, e.g., the discussion in [21, Chapter 2.2.2]. Naturally, it is the same in our setting. Also, proving properties about accumulation points of algorithmic sequences is quite standard in the literature, separately from their existence.

THEOREM 3.3 (Global convergence). *In Algorithm 3.1, for all k sufficiently large, suppose $\{U_k\}$ and $\{Q_k\}$ are bounded and $\mu_{\max} \geq \mu_k \geq \beta_k$ from (3.3). Then every accumulation point $\bar{x}$ of the sequence $\{x^k\}$ is a minimizer of f, and the condition $\mu_k \geq \beta_k$ can be replaced by $\mu_k \geq \bar{\beta}$, the Lipschitz constant of ∇q associated with a neighborhood $\mathcal{N}(\bar{x})$ of $\bar{x}$.*

Proof. Let $\bar{x}$ be an accumulation point of $\{x^k\}$, i.e., there exists a convergent subsequence of $\{x^k\}$, indexed by $k' \in \mathcal{K}$, such that $\{x^k\} \xrightarrow{\mathcal{K}} \bar{x}$. We claim that $\{p^k - x^k\} \xrightarrow{\mathcal{K}} 0$ and $g^k \xrightarrow{\mathcal{K}} \bar{g} = 0$. As U_k, Q_k and μ_k are assumed to be bounded, without loss of generality, passing onto a further subsequence in $\mathcal{K}$ if necessary, we can assume that $\left(U_k, Q_k, \mu_k\right) \xrightarrow{\mathcal{K}} \left(U, Q, \bar{\mu}\right)$. Define $H = UQU^\intercal$. With the assumptions in (3.1), the models in (3.2) converge continously in the second argument: $m(\cdot; x^k) \to m(\cdot; \bar{x})$. Then, by [43, Thms. 7.11 and 12.35], the models epi-converge to $m(\cdot; \bar{x})$, and the proximal mappings $\operatorname{prox}_{m(\cdot;x^k),\mu_k}(\cdot)$ converge to $\operatorname{prox}_{m(\cdot;\bar{x}),\bar{\mu}}(\cdot)$ uniformly on bounded sets. Together with [43, Theorem 7.14], $p^k \to \operatorname{prox}_{m(\cdot;\bar{x}),\bar{\mu}}(\bar{x}) =: \bar{p}$. Passing onto the limit in (3.7) with g^k from (3.4), we obtain that

$$\bar{x} = \bar{p} + H(\bar{\mu}(\bar{p} - \bar{x}) + \nabla q(\bar{x}) - \nabla q(\bar{p}))\,.$$

By the mean-value theorem it holds that $\nabla q(\bar{x}) - \nabla q(\bar{p}) = \nabla^2 q(\bar{y})(\bar{x} - \bar{p})$ for some intermediate point $\bar{y}$. Then, after some direct algebraic manipulations,

$$(I + \bar{\mu}H - H\nabla^2 q(\bar{y}))(\bar{p} - \bar{x}) = (I + H\left[\bar{\mu}I - \nabla^2 q(\bar{y})\right])(\bar{p} - \bar{x}) = 0\,.$$

Because $\mu_k \geq \beta_k$, we have that $\bar{\mu} \geq \bar{\beta}$, the local Lipschitz constant of ∇q associated with $\mathcal{N}(\bar{x})$ containing $\bar{p}$ and $\bar{y}$. Accordingly, the matrix $\bar{\mu}I - \nabla^2 q(\bar{y})$ is positive semidefinite and, by positive semidefiniteness of Q, the matrix H is positive semidefinite. As a result, $I + H\left[\bar{\mu}I - \nabla^2 q(\bar{y})\right]$ is positive definite and, hence, $\bar{p} = \bar{x}$. The definition of g^k in (3.4) and the continuity of ∇q, readily give $g^k \to 0$. By (3.5), this means that $\bar{p} = \bar{x}$ is a minimizer of f. To conclude, note that, because $\{x^k\} \xrightarrow{\mathcal{K}} \bar{x}$ and $\{p^k - x^k\} \xrightarrow{\mathcal{K}} 0$, the neighbourhoods associated with β_k in (3.3) can all be taken as $\mathcal{N}(\bar{x})$ for k sufficiently large. Therefore, $\beta_k \equiv \bar{\beta}$, and the condition $\mu_k \geq \beta_k$ can be replaced by $\mu_k \geq \bar{\beta}$, as stated. □

Thanks to the fact that $g^k \to \bar{g} = 0$, we now show that convergent subsequences eventually generate $\mathcal{V}$-minimizers and identify smooth trajectories associated with the $\mathcal{U}$-Lagrangian.

COROLLARY 3.4 (Asymptotic $\mathcal{U}$-Lagrangian identification and rates). *Under the assumptions in Theorem 3.3, suppose $\bar{g} = 0 \in \operatorname{ri} \partial f(\bar{x})$ and let $\mathcal{K}$ denote the indices of the subsequence converging to $\bar{x}$. Then, for $k \in \mathcal{K}$ sufficiently large and (u^k, v^k) defined in Theorem 3.1, the following hold.*

(i) $g_u^k \in \partial_u L_U(u^k, g_v^k)$;

(ii) *If, in addition, $\mu_k > \beta_k$ from (3.3), then $Vv^k = o(x^k - \bar{x})$.*

Proof. Throughout the proof, $k \in \mathcal{K}$. Notice that, by Theorem 3.3, the subsequence $p^k \to \bar{p} = \bar{x}$, and, thus, both u^k and $g^k \to 0$. By (2.4), $g_v^k \in G_v(\bar{x}) \subset V^\dagger \operatorname{ri} \partial f(\bar{x})$, so eventually g_v^k lies in the relative interior of $V^\dagger \partial f(\bar{x})$, and the statement in (i) corresponds to Theorem 3.1(i), where it is also shown that $v^k \in W(u^k, g_v^k)$. To show the second item, recall that, since $\|Uu^k\| \leq \|p^k - \bar{x}\|$ and $\|Vg_v^k\| \leq \|g^k\|$, by Theorem 3.1(ii) we obtain that

$$\|Uu^k + Vg_v^k\| \leq \frac{3\mu_k(\mu_k + \beta_k)}{\mu_k - \beta_k}\|x^k - \bar{x}\|\,. \tag{3.8}$$

Definition 2.3 combined with the expression in Theorem 2.4, item (vii) with $\bar{g} = 0$, yields

$$f(\bar{x} + Uu^k + Vv^k) - \left\langle g_v^k, V^\intercal V v^k \right\rangle = L_U(u^k, g_v^k) = f(\bar{x}) + o\Big(Uu^k + Vg_v^k\Big)\,.$$

The left-hand side is bounded below by $f(\bar{x})+\eta\|Vv^k\|$ (writing (2.3) with $w = v^k$ and $\bar{g}_u = 0$). Hence, $\eta\|Vv^k\| \leq o\left(Uu^k + Vg_v^k\right)$, and (3.8) concludes the proof. ∎

4. $\mathcal{U}$-Hessians and superlinear convergence. The $\mathcal{V}$-minimizers exhibit first-order expansions for f. To proceed further, a generalized notion of a Hessian [19] is needed.

4.1. The partial $\mathcal{U}$-Hessian of partly smooth functions. The $\mathcal{U}$-Hessian introduced in [24, § 3.3] for the single-variable $\mathcal{U}$-Lagrangian is the basis to define a partial $\mathcal{U}$-Hessian, obtained when differentiating the bivariate $\mathcal{U}$-Lagrangian in the first variable.

Definition 4.1 (partial $\mathcal{U}$-Hessian). *Given $\bar{x}$ and $\bar{g} \in \operatorname{ri}\partial f(\bar{x})$, we say that f has at $\bar{x}$ a* partial $\mathcal{U}$-Hessian $H(\bar{x};\bar{g}_v)$ *associated with $\bar{g}_v$ if*

$$\partial_u L_U(u, \bar{g}_v + z) \subset \bar{g}_u + H(\bar{x};\bar{g}_v)u + B^m(0, o(u,z)), \tag{4.1}$$

where $\bar{g}_u$ is defined in (2.1) and $z \in \mathbb{R}^{n-m}$ is such that $\bar{g}_v + z \in V^\dagger \partial f(\bar{x})$.

Lemma 4.2 (relation with single-variable $\mathcal{U}$-Hessian). *The partial $\mathcal{U}$-Hessian in (4.1) is also a $\mathcal{U}$-Hessian associated with single-variable $\mathcal{U}$-Lagrangian, with $\bar{g}$ being a parameter:*

$$\partial L_U^{\bar{g}}(u) \subset \nabla L_U^{\bar{g}}(0) + H^{\bar{g}}(\bar{x})u + B^m(0, o(\|u\|)), \text{ where } H^{\bar{g}}(\bar{x}) = H(\bar{x};\bar{g}_v).$$

Furthermore, if $0 \in \operatorname{ri}\partial f(\bar{x})$ and f has a partial $\mathcal{U}$-Hessian at $\bar{x}$ associated with $\bar{g}_v = 0$, then the following hold for $\overline{H} = H(\bar{x}, 0)$.

(i) *Under the assumptions in Theorem 3.3, the shifted gradients in Algorithm 3.1 satisfy the inclusion $g_u^k \in \overline{H}u^k + o(u^k)$;*
(ii) *If, in addition, $\overline{H}$ is positive definite, then $\bar{x}$ is the unique minimizer in (1.1).*

Proof. For the identification with the single-variable $\mathcal{U}$-Hessian it suffices to recall (2.2) and write (4.1) with $z \equiv 0$. Item (i) also follows from (4.1), because with our assumptions $g_u^k \in \partial_u L_U(u^k, g_v^k)$ by Corollary 3.4(i). To show (ii), note that if $0 \in \operatorname{ri}\partial f(\bar{x})$ then $\bar{g}_u = 0$ and we can take $\bar{g}_v = 0$. By the identification between the partial and single-variable $\mathcal{U}$-Hessians, $\overline{H} = H^{\bar{g}=0}(\bar{x})$. Therefore, $\bar{u} = 0$ is the unique minimizer of the single-variable $\mathcal{U}$-Lagrangian $L_U^{\bar{g}}(u)$ from Definition 2.1 and the result follows from [23, Theorem 1]. ∎

Partial smoothness, a concept recalled in the Appendix, is related to the existence of partial $\mathcal{U}$-Hessians in Theorem 4.3. Roughly speaking, the $\mathcal{U}$-Hessian corresponds to the curvature of the manifold of smoothness. However, as discussed in the Introduction, the correspondence can be established with only one manifold near $\bar{x}$. A sufficient condition in this sense is that the set of $\mathcal{V}$-minimizers in Definition 2.3 does not change with g_v. The function F in Example 2.2 illustrates well the situation. Its bivariate $\mathcal{U}$-Lagrangian is $L(u,\bar{g}_v) = (2 - |\bar{g}_v|)\frac{a^2}{2}$, so F has the partial $\mathcal{U}$-Hessian $\overline{H} = 2a > 0$, but there are two distinct activity manifolds. These are the two fast tracks in Figure 2.1, generated by the different $\mathcal{V}$-minimizers that emanate from taking a positive or a negative $\bar{g}_v$, i.e., from $W(u,\bar{g}_v) = \left\{\frac{a}{2}\operatorname{sign}(\bar{g}_v)u^2\right\}$.

This is the reason why in our next result we associate the manifold of partial smoothness with a special $\mathcal{V}$-minimizer, that *is the same* for all interior subgradients (the same function was considered in [37, Theorem 6] to characterize fast tracks for prox-regular functions). When specializing the result to the setting (3.1), it is then possible to derive an explicit expression for the $\mathcal{U}$-Hessian when the nonsmooth function in (3.1) is polyhedral, as in regularized regression problems. We refer to the Appendix for the proof.

Theorem 4.3 (Special $\mathcal{V}$-minimizers from partial smoothness and explicit $\mathcal{U}$-Hessian). *Let f be a convex function that is partly smooth at the point $\bar{x}$ relative to a non-singleton set $\mathcal{M} \subset \mathbb{R}^n$. Then, for all small u, there exists a C^2-function $v_{\partial f}$ such that*

$$\forall g \in \operatorname{ri}\partial f(\bar{x}), v_{\partial f}(u) \in W(u, g_v), \text{and } v_{\partial f}(u) = O(\|u\|^2).$$

As a result, there exist neighborhoods $\mathcal{N} := \mathcal{N}(\bar{x})$ of $\bar{x} \in \mathbb{R}^n$ and $\mathcal{N}(0)$ of $0 \in \mathbb{R}^m$ such that

$$L_U(u, g_v) = f\big|_{\mathcal{M}\cap\mathcal{N}}(u) - \left\langle g_v, V^\mathsf{T} V v_{\partial f}(u)\right\rangle \text{ and } \mathcal{M}\cap\mathcal{N} = \left\{\bar{x} + Uu + Vv_{\partial f}(u) \colon u \in \mathcal{N}(0)\right\},$$

where $f\big|_{\mathcal{M}\cap\mathcal{N}}$ is considered a composite function of u.

If, in addition, $0 \in \operatorname{ri}\partial f(\bar{x})$, a partial $\mathcal{U}$-Hessian of f at $\bar{x}$ associated with $\bar{g} = 0$ is given by the Hessian restriction: $\overline{H} = \nabla^2_{uu} f\big|_{\mathcal{M}\cap\mathcal{N}}(0)$. In particular, for a function as in (3.1), if the nonsmooth function h is finite-valued and polyhedral, then the partial $\mathcal{U}$-Hessian at $\bar{x}$ associated with $\bar{g} = 0 \in \operatorname{ri}\partial f(\bar{x})$ has the explicit expression $\overline{H} = U^\top \nabla^2 q(\bar{x}) U$.

Partly smooth functions with the structure in (3.1) are considered in [3] to show that the proximal gradient method can identify the smooth manifold at a minimizer. This manifold is actually the fast track, which have been shown to be equivalent objects, for convex functions in [14] and for prox-regular functions in [29]. In the method proposed in [3], after identifying the manifold via the proximal gradient mapping, certain Riemannian gradient and Hessian are employed to compute a $\mathcal{U}$-Newton direction and boost convergence speed.

4.2. Superlinear convergence of the PGVU method. For superlinear convergence, naturally, properties of the partial $\mathcal{U}$-Hessian at $\bar{x}$ associated with $\bar{g}_v = 0$ are important. This matrix is assumed to be positive definite. Also, the Dennis-Moré-type condition below, typical in quasi-Newton methods (see, e.g., [21]), is required:

$$(U_k Q_k U_k^\mathsf{T} - U\overline{W}U^\mathsf{T}) g^k = o(U g_u^k)\,, \text{ where } \overline{W}^{-1} = \overline{H} := H(\bar{x}; 0)\,, \tag{4.2}$$

and the matrix U spans the $\mathcal{U}(\bar{x})$-subspace.

THEOREM 4.4 (Superlinear rate). *Suppose f has a positive definite partial $\mathcal{U}$-Hessian $\overline{H}$ at $\bar{x}$ associated with $\bar{g}_v = 0$ and that (4.2) holds. Under the assumptions in Theorem 3.3, let $\bar{x}$ be an accumulation point of $\{x^k\}$ such that $\bar{g} = 0 \in \operatorname{ri}\partial f(\bar{x})$. Then $\bar{x}$ is the unique minimizer of f, and both $\{x^k\}$ and $\{p^k\}$ converge to $\bar{x}$. Furthermore, if $\mu_k > \bar{\beta}$, the local Lipschitz constant of ∇q associated with a neighborhood $\mathcal{N}(\bar{x})$ of $\bar{x}$, then $\|x^{k+1} - \bar{x}\| = o(x^k - \bar{x})$, i.e., the iterates generated by Algorithm 3.1 converge superlinearly.*

Proof. Because $0 \in \operatorname{ri}\partial f(\bar{x})$ and f has a positive definite partial $\mathcal{U}$-Hessian at $\bar{x}$, we have from Lemma 4.2 that $\bar{x}$ is the unique minimizer of f. By Theorem 3.3, every accumulation point of the sequence $\{x^k\}$ generated by Algorithm 3.1 is a minimizer of f. Consequently, $\bar{x}$ is the unique accumulation point of $\{x^k\}$, with both $\{x^k\}$ and $\{p^k\}$ converging to $\bar{x}$.

Next, using (3.7), and recalling the definition of (u^k, v^k) in Theorem 3.1, we obtain that

$$\begin{aligned}
x^{k+1} - \bar{x} &= p^k - U_k Q_k U_k^\mathsf{T} g^k - \bar{x} \\
&= -U_k Q_k U_k^\mathsf{T} g^k + p^k - \bar{x} + U\overline{W}U^\mathsf{T} g^k - U\overline{W}U^\mathsf{T} g^k \\
&= \left(U\overline{W}U^\mathsf{T} - U_k Q_k U_k^\mathsf{T}\right) g^k - U\overline{W}U^\mathsf{T} g^k + (p^k - \bar{x})_u + (p^k - \bar{x})_v \\
&= \left(U\overline{W}U^\mathsf{T} - U_k Q_k U_k^\mathsf{T}\right) g^k + \left(Uu^k - U\overline{W}U^\mathsf{T} g^k\right) + v^k\,.
\end{aligned}$$

The third term above is of the order $o(x^k - \bar{x})$, by item (ii) in Corollary 3.4. The same holds for the first term, as $\left(U\overline{W}U^\mathsf{T} - U_k Q_k U_k^\mathsf{T}\right) g^k = o(g^k)$ because of (4.2), and $g^k = O(x^k - \bar{x})$, by Theorem 3.1(ii). To conclude the proof, it remains to show that $T_2 := \left(Uu^k - U\overline{W}U^\mathsf{T} g^k\right) = o(x^k - \bar{x})$. Since $T_2 = U\left(u^k - \overline{W}U^\mathsf{T} g^k\right) = U\left(u^k - \overline{W} g_u^k\right)$, after multiplying on the left by $\overline{H}U^\mathsf{T}$, we see that $\overline{H}U^\mathsf{T} T_2 = \overline{H}u^k - g_u^k$. Lemma 4.2(i) then ensures that $\overline{H}U^\mathsf{T} T_2 = o(u^k)$. The result follows, because $u^k = O(x^k - \bar{x})$, by Theorem 3.1(ii). ∎

5. Numerical assessment. In our numerical validation, we first consider a nonsmooth term that is a multiple of the polyhedral function in (5.1), in which case computing proximal points of h is straightforward. The second benchmark is done with an n-dimensional extension of Example 2.2, having a non-polyhedral manifold of partial smoothness (requiring more involved computations for obtaining the proximal point in the $\mathcal{V}$-step). We analyze the beneficial impact of the $\mathcal{U}$-step in both cases.

5.1. ℓ_1-regularized minimization. The ℓ_1-norm is shown in [45, Sec. 3.1] to be partly smooth relative to an affine or linear set, with an explicit $\mathcal{U}$-subspace:

$$h(x) = \|x\|_1 \text{ then } \mathcal{U}(p) = \operatorname{span}\left\{e^j : j \in J(p)\right\} \text{ for } J(p) = \left\{i \le n : |p_j| > 0\right\} . \tag{5.1}$$

We consider problems such that

$$\text{in (3.1) the nonsmooth function is } h(x) = \lambda\|x\|_1 , \text{ for a positive parameter } \lambda, \tag{5.2}$$

so the $\mathcal{V}$-step in Procedure 1 computes

$$p^k = \max\left\{0, w^k - \frac{\lambda}{\mu_k}\right\} - \max\left\{0, -w^k - \frac{\lambda}{\mu_k}\right\} , \text{ for } w^k = x^k - \frac{1}{\mu_k}\nabla q(x^k) . \tag{5.3}$$

The remaining two calculations that need to be specified in Algorithm 3.1 refer to the orthonormal basis U_k, and the positive semidefinite matrix Q_k.

5.1.1. Defining Q_k and U_k and the full algorithm. For PGVU global convergence in Theorem 3.3, the matrices only need to be bounded. The superlinear rate in Theorem 4.4 requires a quite standard Dennis-Moré-type condition, natural in quasi-Newton frameworks. There are various choices that are compatible with the theory. For better numerical performance, it is preferable that matrices do not change too abruptly along consecutive iterations.

Regarding the second-order information along the $\mathcal{U}$-subspace, [45, Example 10] shows that the ℓ_1-norm is partly smooth at any $p \in \mathbb{R}^n$ relative to $\mathcal{U}(p)$ defined in (5.1). As the same holds for f when h is as in (5.2), by Theorem 4.3, if $0 \in V^\dagger \operatorname{ri} \partial f(\bar{x})$ then

$$\overline{H} := U^\top \nabla^2 q(\bar{x}) U$$

is a partial $\mathcal{U}$-Hessian of f at $\bar{x}$ associated with $\bar{g}_v = 0$. Thus, a natural choice for the quasi-Newton matrices in the $\mathcal{U}$-step is to take $Q_k^{-1} = U_k^\top \nabla^2 q(x^k) U_k$.

The choice of the matrices U_k is more delicate. Since $\partial f(x) = \nabla q(x) + \lambda \partial h(x)$, the $\mathcal{VU}$-decomposition is determined by the $\mathcal{V}$-subspace associated with h. Then, at first glance, defining the basis matrix U_k for the subspaces $\mathcal{U}(p^k)$ from (5.1) might appear straightfoward. Nevertheless, given $\varepsilon \ge 0$, we consider instead the smaller subspaces

$$\mathcal{U}_\varepsilon(x) = \operatorname{span}\left\{e^j : j \in J_\varepsilon(x)\right\} \text{ for } J_\varepsilon(x) := \left\{i \le n : |x_i| > \frac{\varepsilon}{2}\right\} . \tag{5.4}$$

This $\mathcal{V}_\varepsilon\mathcal{U}_\varepsilon$-decomposition was introduced in [30] to deal with the lack of continuity of the subdifferential as a multifunction. Unlike the $\mathcal{VU}$-decomposition, the following important continuity property, [30, equation (5.13)], holds for the ε-counterpart:

$$\lim_{\substack{(x,\varepsilon)\to(\bar{x},0)\\ \varepsilon\ge 0}} \mathcal{V}_\varepsilon(x) = \mathcal{V}(\bar{x}) \quad \text{and} \quad \lim_{\substack{(x,\varepsilon)\to(\bar{x},0)\\ \varepsilon\ge 0}} \mathcal{U}_\varepsilon(x) = \mathcal{U}(\bar{x}) .$$

Taking U_k as an orthonormal basis for $\left[e^j : j \in J_{\varepsilon^k}(x^k)\right]$ with $\varepsilon^k \to 0$ and $x^k \to \bar{x}$, ensures that $U_k \to U$, and the Dennis-Moré condition (4.2) in Theorem 4.4 is satisfied.

Three different (natural) choices for ε in (5.4), all driven to 0 by PGVU, are:

$$\varepsilon_0^k = 10^{-6}\,, \quad \varepsilon_1^k = f(x^k) - f(p^k) - \mu_k\|p^k - x^k\|^2\,, \text{ and } \varepsilon_2^k = \mu_k\|p^k - x^k\|\,.$$

The first option corresponds to the PGVU runs in Example 3.2. The second one transports $\mu_k(x^k - p^k)$, a subgradient of the model at p^k, to $\mu_k(x^k - p^k) \in \partial_{\varepsilon_k} f(x^k)$, [20, Prop. XI.4.2.2]. Since computing ε_1^k can be time consuming (it requires to evaluate f at two points), the third option appears a good alternative (Theorem 3.3 shows that $x^k - p^k \to 0$).

We repeated the runs in Example 3.2, for each PGVU variant, respectively labeled PGVU-0,1,2, in a reference to the value of ε^k employed in (5.4). The comparison of the number of iterations and digits of accuracy indicate ε_2^k as the best option. In particular, for $\mu^0 = 0.18$ and $x^0 = (-1.1, 0.76)^\intercal$, as reported in Table 3.1, PGVU-0 took 50 iterations to reach 8 digits, while PGVU-2 attained the same accuracy in only 7 iterations.

We are now ready to introduce the algorithm.

Algorithm 5.1 Proximal Gradient $\mathcal{VU}$-method for ℓ_1-regularized minimization (PGVU-2)

Data: f as in (5.2), $x^0 \in \mathbb{R}^n$, $\mu_0 > 0$, and a stopping tolerance TOL ≥ 0. Set $k = 0$.
repeat
  Apply Procedure 1 with p^k defined in (5.3).
  Shift the gradient $g^k = \mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k)$.
  Compute $U_k = \left[e^j : |x_j^k| > \frac{\varepsilon_k}{2}, 1 \leq j \leq n\right]$ for $\varepsilon_k := \max\left(\text{TOL}, \mu_k\|p^k - x^k\|\right)$.
  Obtain d^k by solving the linear system $U_k^\intercal W_k U_k d = -U_k^\intercal g^k$ for $W_k \approx \nabla^2 q(x^k)$.
  Update $x^{k+1} = p^k + U_k d^k$, set $k = k + 1$.
until $\|g^k\| \leq$ TOL;

Being a special instance of PGVU, Algorithm 5.1 has global and superlinear convergence if the conditions in Theorems 3.3 and 4.4 are satisfied. With our definitions, such is the case if $\mu_k > \bar{\beta}$ and, for $\bar{x}$ an accumulation point of $\{x^k\}$ generated by Algorithm 5.1, it holds that $0 \in \operatorname{ri} \partial f(\bar{x})$ and $\overline{H} = U^\intercal \nabla^2 q(\bar{x}) U$ is positive definite.

5.1.2. Test functions and parameters. We consider regularized least-square problems:

$$f(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \text{ for } A \in \mathbb{R}^{m\times n}\,, b \in \mathbb{R}^m\,, \text{ and } \lambda = 0.1\|A^\top b\|_\infty\,.$$

The smooth term in (3.1) is quadratic, so the gradient global Lipschitz is the largest singular value of the matrix A.

There are two sets of problems: 5000 mid-size randomly generated instances, and 95 large-size statistical classification and regression instances in the webpage [1] of LIBSVM, a library for support vector machines. The mid-size problems, referred below as QUAD, have random dimensions $m \in [10, 1000]$ and n an integer in $[0.1m, 2m]$. For half of the QUAD problems, that is 2500 runs, we set $A = -\frac{1}{\sqrt{2n}}A'$ for a random matrix A'. The outcome vector $b = Ab' + 0.0001\xi$ for random $\xi \in [0, 1]$ uniformly distributed. The vector b' has non-null components set to ± 1, depending on a sparsity parameter randomly chosen. The second half of the QUAD problems, sets A as a random matrix with normalized columns and $b = Ab' + \sqrt{0.002}\xi$ for b' and ξ as above. The problems from LIBSVM are all scaled to [-1,1] or [0,1] and referred to as LIB.

[1] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/

In Procedure 1, we set $x^0 = 0$, and $\mu_0 = {g^0}^\top g^0/(2\max\{1, |f(x^0)|\})$. The maximum number of iterations was set to 100, and the stopping tolerance is $\text{TOL} = 10^{-6}$. After the $\mathcal{U}$-step, the prox-parameter is updated according to [25], i.e., $\mu_{k+1} = \frac{{y^k}^\top y^k \mu_k}{{y^k}^\top y^k + \mu_k {y^k}^\top s^k}$ where $y^k := g^{k+1} - g^k$ and $s^k := x^{k+1} - x^k$; see also [6, Ch.10.3.3]. (This update is used with safeguards, to ensure boundedness.) In the $\mathcal{U}$-step of PGVU-2, computing the $\mathcal{U}$-Newton direction requires solving a linear system at each iteration. For mid-size instances like QUAD. taking $Q^k = (U_k^\top (A^\top A + \text{TOL}\, I) U_k))^{-1}$ is sound. For LIB's large-size problems, we provide two options for solving the linear system. The first one sets $Q^k = (U_k^\top \operatorname{diag}(A^\top A + \text{TOL}\, I) U_k)^{-1}$, inverting only the positive diagonal elements. The second option is just to solve for d^k the system $U_k^\top A^\top A U_k d = -U_k^\top g^k$ with the backslash (\) in Matlab and if the warning of singular matrix or nearly singular matrix is detected then discard the $\mathcal{U}$-step.

5.1.3. Solvers and figures with evaluation measures. The benchmark compares MATLAB implementations of the following solvers:

(i) PGVU-2, as in Algorithm 5.1.
(ii) SpaRSA 2.0, the sparse reconstruction by separable approximation [46] [2].
(iii) FISTA, the fast iterative shrinkage-thresholding algorithm with backtracking [4] [3].
(iv) ADMM, alternating direction method of multipliers [7], with $(\rho, \alpha) = (10^{-4}, 1.5)$ [4].
(v) qNVU, the $\mathcal{VU}$-algorithm from [38].

The experiments were performed on an Intel Core i7 computer with 12 cores and 32 GB RAM, running under Ubuntu 22.02.2 LTS.

The performance is measured by comparing the accuracy and the computing time of each solver, separately for the QUAD and LIB problems. We proceed as follows. First, for the accuracy criterion, if `best` is the smallest functional value found on a given instance for all the solvers, the accuracy of solver s is

$$\text{ACC}(s) = \min\left(-\log\left(\frac{|f^*(s) - \texttt{best}|}{|\texttt{best}|}\right), 16\right).$$

Solvers having achieved at least $\texttt{cutoff} \in \{2, 4, 6\}$ digits of accuracy are considered successful. The corresponding values are reported in Table 5.1, where the high achieved accuracy, the differential of both $\mathcal{VU}$-based methods, becomes evident. The accuracy achieved by FISTA is also impressive, being slightly inferior to $\mathcal{VU}$ for the large LIB instances.

TABLE 5.1
Successful runs for each solver

		PGVU-2	SpaRSA	FISTA	ADMM	qNVU
QUAD	ACC ≥ 2	5000	4998	5000	516	5000
mid-size	ACC ≥ 4	4999	4933	5000	5	5000
(5000 runs)	ACC ≥ 6	4998	4493	5000	1	5000
LIB	ACC ≥ 2	86	91	85	8	86
large-size	ACC ≥ 4	81	51	79	2	86
(95 runs)	ACC ≥ 6	74	26	70	0	86

SpaRSA has good accuracy for the mid-size instances, but performs less well for the LIB problems. On these runs, and with the considered parameters, ADMM did not perform well.

[2] http://www.lx.it.pt/~mtf/SpaRSA/
[3] https://github.com/tiepvupsu/FISTA
[4] https://web.stanford.edu/~boyd/papers/admm/lasso/lasso.html

Regarding computing times, for low accuracy (`cutoff`=2), SpaRSA is always the fastest solver. The profile in Figure 5.1 compares computing times among the successful runs, for the value of `cutoff`=4. In the right plot in Figure 5.1, qNVU exhibits a slower performance.

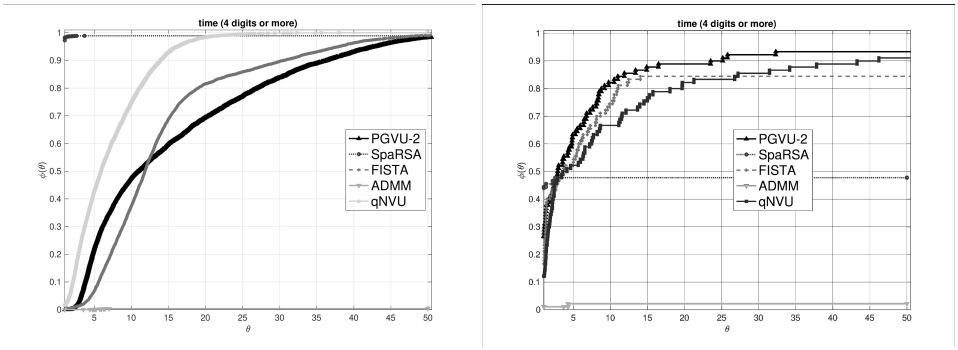

FIG. 5.1. *Profiles for computing time of successful runs, for the QUAD and LIB instances (left and right).*

This is because at each iteration the qNVU method [38] solves two quadratic programming problems, a computationally expensive calculation for the large LIB instances. For the mid-size instances, SpaRSA remains the fastest solver, but not for the LIB problems. A solver-to-solver comparison clarifies this situation with four plots reported in Figure 5.2, comparing PGVU-2 to SpaRSA and FISTA, when LIB problems were solved with at least 2 or 4 digits.

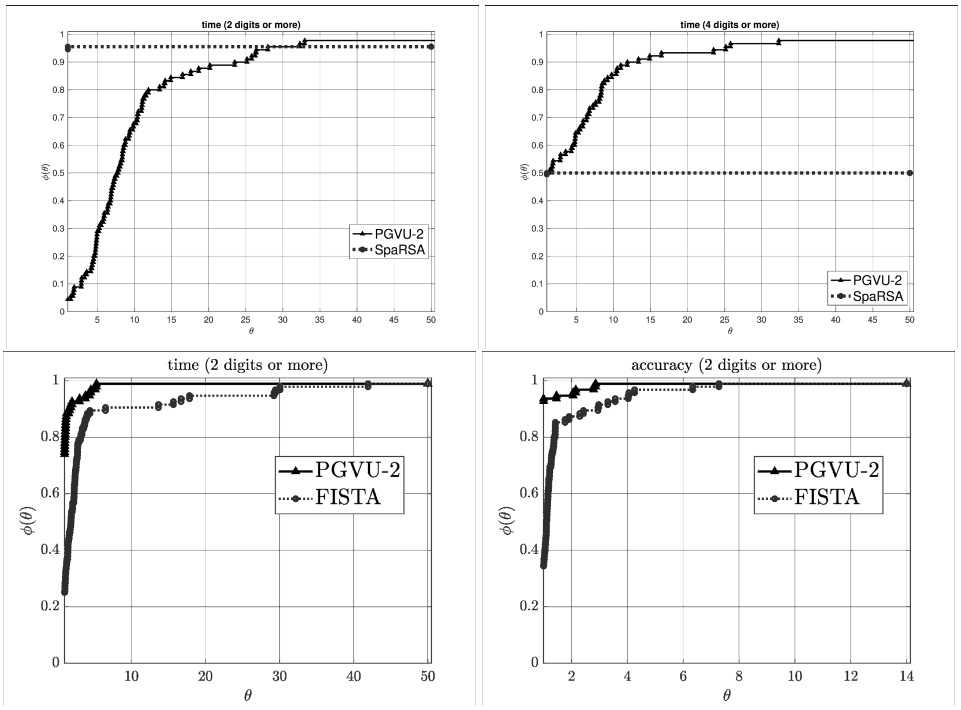

FIG. 5.2. *Solver-to-solver comparisons over LIB instances for different accuracy.*

On the top left plot in Figure 5.2, we notice that to reach 2 digits of accuracy, SpaRSA is faster than PGVU-2. To get 4 digits, the right top plot shows that PGVU-2 is faster. The

bottom plots show the profiles of PGVU-2 and FISTA. In this specific test, we used the backslash (\) in Matlab to solve the linear system for d^k. As we can see, PGVU outperforms FISTA in both time and accuracy.

5.2. Importance of the $\mathcal{U}$-step. The previous benchmark illustrates the global improvement of performance when PGVU-2 is compared to several solvers. To put in evidence the beneficial effect of the $\mathcal{U}$-step regarding local convergence rates, we generate random multidimensional versions of the function in Example 2.2, all having optimal value $\bar{f} = 0$.

Let $m > 0$, a vector $a \in \mathbb{R}^m$ with positive components, and $\kappa \geq 0$ be given. For all $x \in \mathbb{R}^n$ with $n = 2m$, the functions defining f in (3.1) are

$$\text{(5.5)} \qquad \begin{cases} q(x) &= \dfrac{1}{2}\displaystyle\sum_{j=1}^{m} a_j x_j^2 + \dfrac{\kappa}{n}\|x\|^4 \quad \text{and, for } \lambda > 0, \\ h(x) &= \lambda \displaystyle\sum_{j=1}^{m} h_j(x) \qquad\qquad \text{where } h_j(x) = \max(\tfrac{1}{2}a_j x_j^2, |x_{j+m}|)\,. \end{cases}$$

The instance with $m = 1$, $\kappa = 0$, and $\lambda = 1$ yields the function in Example 2.2. Now the u and v variables are two vectors of dimension $m = n/2$, with respective components $u = (x_1, x_2, \ldots, x_m)$ and $v = (x_{m+1}, x_{m+2}, \ldots, x_{2m=n})$.

Because nonsmoothness is driven by the function h, at the minimizer $\bar{x} = 0$ the subspace $\mathcal{U}(\bar{x})$ is spanned by canonical vectors $e^j \in \mathbb{R}^n$ for $j = \{1, 2, \ldots, m\}$. Also, as the nonsmooth term is weighted by λ, we expect problems with larger values of that parameter to be more difficult to solve. This is clear in Figure 5.3, plotting for four different parameters λ the graphs a function as in (5.5), with $n = 4$ and $\kappa = 0$, varying components 1 and 4 close to $\bar{x}$.

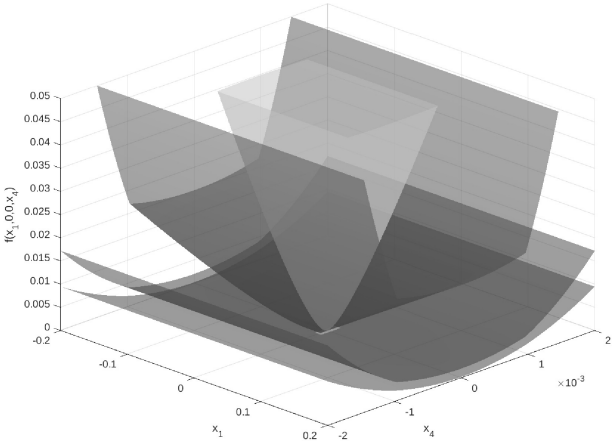

FIG. 5.3. *Some 2-D plots near $\bar{x} = 0 \in \mathbb{R}^4$ for f as in (5.5) defined for $n = 4$ and $\kappa = 0$, taking $\lambda \in \{0.1, 1, 10, 50\}$. The flatter surface corresponds to $\lambda = 0.1$. Larger values of this parameter create large functional variations near the optimum $\bar{x}$, progressively curving the surface. The pointed one corresponds to $\lambda = 50$.*

In the figure, it is noticeable that, as λ increases, the slope of the manifold of smoothness becomes severely steep near the optimum. Since in this case it is more difficult to achieve high accuracy when following the fast track, we expect the beneficial effect of PGVU to be more perceptible for instances with larger values of λ (the expectation is confirmed by the trajectories in the bottom plots in Figure 5.4).

We examine how fast PGVU's functional values tend to zero, when compared to PG and to another nonsmooth method incorporating second-order information. For the latter, we chose the BFGS phase of HANSO 3.0[5] [8], which combines a quasi-Newton direction defined using subgradients as gradients with a weak Wolfe line-search [27]. We did not include qNVU because the experiment focuses on the accuracy reached when solvers are given the same time budget. The two quadratic programming programs solved by qNVU at each iteration are a handicap in this sense (the BFGS updates in HANSO, by contrast, are very fast).

The benchmark is designed according to the following principles.

- Given $n \in \{100, 1000\}$ and $\lambda \in \{0.1, 1, 10, 50\}$, fifty vectors $a \in (0,1)^{n/2}$ and starting points $x^0 \in [-1,1]^n$ were generated using a uniform random distribution.
- Half of the 50 instances is run with $\kappa = 0$ and the other half with $\kappa = 10$. Positive values of κ make the smooth term more different from a quadratic function and, hence, more difficult to approximate with a quasi-Newton update.
- HANSO was run with its default parameters. Since this is a general nonsmooth method, which does not exploit f's particular structure, its running time was taken as reference. For each instance, the time that it took for HANSO to make 200 iterations defines the time budget that PG and PGVU have on the same instance.
- To put on equal ground PG and PGVU (regarding their proximal point calculation and μ-update), we run our implementation of PGVU with and without its $\mathcal{U}$-step. The latter variant is indeed PG, as x^k is updated just by Procedure 1.
- When compared to the least-square problems in the previous section, the calculation of proximal points is more involved than applying the shrinkage operator (5.3). Specifically, p^k results from m separate prox-computations, one for each term h_j in (5.5). Similarly to Example 2.2, computing $\text{prox}_{h_j,\mu_k}(x^k)$ amounts to solving cubic equations associated with the optimality conditions of

$$\begin{cases} \min_{(r,y,z)\in\mathbb{R}^3} & r + \frac{1}{2}\mu_k(y - x_j^k)^2 + \frac{1}{2}\mu_k(z - x_{j+m}^k)^2 \\ \text{s.t.} & r \geq \frac{1}{2}a_j y^2 \,, r \geq z \,, \text{ and } r \geq -z \,. \end{cases}$$

  The (y,z)-components of the solution give p_j^k and p_{j+m}^k in Procedure 1.
- Regarding the $\mathcal{U}$-step in PGVU:
  - the basis U_k has columns $e^l \in \mathbb{R}^n$ for $l \in L_{\text{sq}}(p^k) \cup L_{\text{abs}}(p^k)$, where

$$L_{\text{sq}}(p) = \left\{ j \leq m : \tfrac{1}{2}a_j p_j^2 \geq h_j(p) - \varepsilon_2 \right\} \text{ and}$$
$$L_{\text{abs}}(p) = \left\{ j + m : j \leq m \text{ and } |p_{j+m}| \geq h_j(p) - \varepsilon_2 \right\} ,$$

    for $\varepsilon_2 = \max\left(\text{TOL}, \mu_k \|p^k - x^k\|\right)$.
  - The initial quasi-Newton matrix $Q_0^{-1} = [\nabla^2 q(x^0)]^{-1}$ is updated by the BFGS formula, using `bfgsHessianInverse.m` in GRANSO [9].
  - Because PGVU is a nonmonotone method, the $\mathcal{U}$-step is dropped whenever

$$f(p^k + U_k d^k) \geq f(p^k). \tag{5.6}$$

    In this case, PGVU sets $x^{k+1} = p^k$ as in PG.

The mechanism in (5.6) either takes or discards a complete step in the $\mathcal{U}$-direction, without any damping. We implemented this procedure to not give PGVU the advantage of an elaborate line-search. The results reported below show that, even if PGVU discards $\mathcal{U}$-steps, the retained ones improve the performance of PGVU over that of PG.

[5] http://www.cs.nyu.edu/faculty/overton/software/hanso/index.html

Comparing PG and PGVU, we provide in Figure 5.4 a numerical illustration that the successful identification of the manifold obtained by PG is not sufficient: to achieve fast convergence, it is essential to correct the output of Procedure 1 with a $\mathcal{U}$-quasi-Newton move. Furthermore, as hinted in Example 3.2, and known in smooth optimization, accuracy depends on properly identifying curvature. The comparison of HANSO and PGVU in the same figure shows that taking into account curvature information in nonsmooth structured optimization is profitable only if the correction is performed in the $\mathcal{U}$-space. These remarks arise from the examination of the slopes of the trajectories corresponding respectively to HANSO, PG and PGVU, displayed in Figure 5.4, whose meaning is explained next.

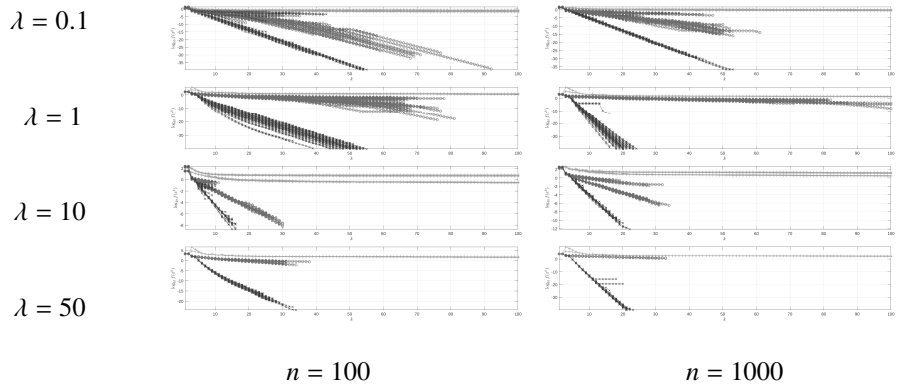

FIG. 5.4. *HANSO and PG against PGVU, comparing 400 instances of f as in (5.5) with $\kappa \in \{0, 10\}$, setting $n = 100$ and $n = 1000$ on the left and right columns. Rows correspond to values of $\lambda \in \{0.1, 1, 10, 50\}$. crosses, cercles and x's correspond to HANSO, PG, and PGVU, respectively. Each graph plots the values $\{\log_{10} f(x^k)\}_{k \le 100}$ generated for a given function, until reaching the time it took HANSO to perform 200 iterations.*

Figure 5.4 has eight graphs, the output obtained when varying the parameters λ is displayed in four rows, while the two columns show the results for $n = 100$ and $n = 1000$. As mentioned, from the nonsmooth point of view the easiest (most difficult) configuration is the one with smallest (largest) λ and n, shown in the top left (bottom right) plot.

Since $\bar{f} = 0$ for all the instances, the ordinates $\log_{10} f(x^k)$ give a good indication of the convergence rate. Trajectories generated by HANSO, PG, and PGVU are displayed with crosses, cercles and x's, running the three solvers on the same 400 instances, given the time needed by HANSO to perform 200 iterations on the instance (0.1 s in average). Because of the prox-calculation, PG and PGVU need more time per iteration than HANSO, that manages to make more iterations using the same bugdet of time. This is reflected by the long tails near the top of each graph (this is why we only plot the first 100 iterates of HANSO).

The ordinates in Figure 5.4, descending from 0 to -35, represent the negative of the number of digits reached by the solvers in the different runs. The slope trajetories indicate how fast were the solvers. The most accurate one is PGVU, whose trajectories have the steepest slopes in all the tested instances. HANSO is fast to iterate, but it is not precise. This is not a surprise because (like qNVU) HANSO is a general solver, which does not exploit the structural properties of the functions to be minimized, as PG or PGVU do.

PG performs better than HANSO but is truly accurate only for the easiest configuration ($\lambda = 0.1$, $n = 100$). Because of the $\mathcal{U}$-step correction, PGVU spends more time per iteration

and performs less iterations than PG. In compensation, the extra computational work allows PGVU to achieve a higher accuracy than PG, using the same budget of time. This is especially true when nonsmoothness is more prevalent, for $\lambda = 50$. In both dimensions, the bottom row in the figure confirms that PG struggles to achieve high accuracy. For this test-set at least, the interest of performing $\mathcal{U}$-Newton steps appears clear.

HANSO trajectories look more uniform than those generated by PG or PGVU on the different instances. However, when $\lambda = 10$ there are two distinct paths generated by HANSO, corresponding to the quadratic and non-quadratic instances, i.e., with $\kappa = 0$ or 10 (the option that activates the Oren-Levenberg scaling in the first iteration had no effect in this respect).

Iterates in HANSO and PGVU involve a quasi-Newton move that, unlike PG, incorporates second-order-like information. For smooth functions HANSO behaves locally in a Newtonian manner. The infinity norm of all the starting points is less than one, so the runs do not start too far from $\bar{f}$. Nevertheless, over the 200 runs with $n = 100$, the smallest functional value attained by HANSO was 0.005. Actually, HANSO achieved 2 digits of accuracy only in the easiest configuration, where PGVU reached more than 30 digits. For the tested instances, it appears that making a curvature correction only in the $\mathcal{U}$-space is extremely beneficial.

As with any Newtonian approach, PGVU could be improved by equipping the $\mathcal{U}$-step with a suitable line-search. In average, about 25% of the iterations satisfied (5.6) and the $\mathcal{U}$-correction was dropped. Nevertheless, the crude approach used for our proof-of-concept benchmark already shows a significant gain in terms of accuracy on the 400 tested instances.

6. Concluding remarks. We have extended the $\mathcal{VU}$-theory for convex functions by defining the two variable $\mathcal{U}$-Lagrangian and the partial $\mathcal{U}$-Hessian. We showed that $\mathcal{V}$-minimizers are tangent to the $\mathcal{U}$-subspace if $\bar{g}_v \to 0$, an important property leading to superlinear convergence of the Proximal Gradient $\mathcal{VU}$-method. For functions with primal-dual gradient structure (including ℓ_1-regularization), the Hessian of the single-variable $\mathcal{U}$-Lagrangian exists along a certain fast-track [36, Theorem 4.1]. We extend this result to our bivariate $\mathcal{U}$-Lagrangian, so that a Newtonian step can be performed as the $\mathcal{U}$-step.

The PGVU method is applicable to structured convex optimization problems, and converges superlinearly despite nonsmoothness. Numerical experiments verify that the method is particularly useful when high accuracy is desired. The comparison with PG showed that the manifold identification needs to be complemented with a quasi-Newton step. But, as shown in the comparison with HANSO, such second-order-like step must be done in the $\mathcal{U}$-subspace.

Considering a specific family of functions, as in (3.1), leads to more targeted implementations. This was the case for the $\mathcal{VU}$-methods [41, 15, 3], respectively taylored for maximum eigenvalue, convex finite-max, and additive optimization. The latter, in particular, takes a $\mathcal{V}$-step as in PGVU, but requires solving a quadratic programming problem to define the $\mathcal{U}$-Newton direction. We avoid that computation by our result in Theorem 3.1, that shifts the subgradient of the model proximal point.

Originally defined for convex functions, the $\mathcal{VU}$-theory has been generalized to the nonconvex setting [36, 18, 29]. In [29] a localized version of $\mathcal{U}$-Lagrangian and the notion of fast track are defined for prox-regular functions, and the correspondence between an active manifold of a partly smooth function and a fast track is given. In [12], under the condition called tilt stability, the smoothness properties of the function f restricted to the fast track are shown. Combining those theoretical results with a suitable line-search, developing nonconvex versions of PGVU might be a subject for future research.

Acknowledgments. The authors are grateful for the constructive suggestions provided by the reviewers and the associate editor.

REFERENCES

[1] F. Atenas, C. Sagastizábal, P. J. S. Silva, and M. Solodov, *A unified analysis of descent sequences in weakly convex optimization, including convergence rates for bundle methods*, SIAM J. Optim., 33 (2023), pp. 89–115, https://doi.org/10.1137/21M1465445.

[2] G. Bareilles, F. Iutzeler, and J. Malick, *Harnessing structure in composite nonsmooth minimization*, SIAM J. Optim., 33 (2023), pp. 2222–2247, https://doi.org/10.1137/22M1505827.

[3] G. Bareilles, F. Iutzeler, and J. Malick, *Newton acceleration on manifolds identified by proximal gradient methods*, Math. Program., 200 (2023), pp. 37–70, https://doi.org/10.1007/s10107-022-01873-w.

[4] A. Beck and M. Teboulle, *Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems*, IEEE Trans. Image Process., 18 (2009), pp. 2419–2434.

[5] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202, https://doi.org/10.1137/080716542.

[6] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal, *Numerical Optimization. Theoretical and Practical Aspects*, Springer Berlin Heidelberg, 2006, https://doi.org/10.1007/978-3-540-35447-5.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122, https://doi.org/10.1561/2200000016.

[8] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. S. oes, *Gradient sampling methods for nonsmooth optimization*, in Numerical Nonsmooth Optimization: State of the Art Algorithms, A. M. Bagirov, M. Gaudioso, N. Karmitsa, M. M. Mäkelä, and S. Taheri, eds., Springer International Publishing, 2020, pp. 201–225.

[9] F. E. Curtis, T. Mitchell, and M. L. Overton, *A BFGS-SQP method for nonsmooth, nonconvex, constrained optimization and its evaluation using relative minimization profiles*, Optim. Methods Softw., 32 (2017), pp. 148–181, https://doi.org/10.1080/10556788.2016.1208749.

[10] A. Daniilidis, C. Sagastizábal, and M. Solodov, *Identifying structure of nonsmooth convex functions by the bundle technique*, SIAM J. Optim., 20 (2009), pp. 820–840, https://doi.org/10.1137/080729864.

[11] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Math. Oper. Res., 43 (2018), pp. 919–948, https://doi.org/10.1287/moor.2017.0889.

[12] A. C. Eberhard, Y. Luo, and S. Liu, *On partial smoothness, tilt stability and the $\mathcal{VU}$-decomposition*, Math. Program., 175 (2019), pp. 155–196, https://doi.org/10.1007/s10107-018-1238-8.

[13] W. Hare, *Numerical analysis of VU-decomposition, U-gradient, and U-hessian approximations*, SIAM J. Optim., 24 (2014), pp. 1890–1913.

[14] W. Hare and A. Lewis, *Identifying active constraints via partial smoothness and prox-regularity*, J. Convex Anal., 11 (2004), pp. 251–266.

[15] W. Hare, C. Planiden, and C. Sagastizábal, *A derivative-free VU-algorithm for convex finite-max problems*, Optim. Methods Softw., 0 (2019), pp. 1–39, https://doi.org/10.1080/10556788.2019.1668944.

[16] W. Hare, C. Planiden, and C. Sagastizábal, *The chain rule for VU-decompositions of nonsmooth functions*, J. Convex Anal., 27 (2020), pp. 335–360.

[17] W. L. Hare, *Nonsmooth optimization with smooth substructure*, PhD thesis, Simon Fraser University, 2004.

[18] W. L. Hare and R. Poliquin, *The quadratic sub-Lagrangian of a prox-regular function*, Nonlinear Anal., 47 (2001), pp. 1117–1128, https://doi.org/10.1016/s0362-546x(01)00251-6.

[19] J. B. Hiriart-Urruty, *The approximate first-order and second-order directional derivatives for a convex function*, in Mathematical Theories of Optimization, J. P. Cecconi and T. Zolezzi, eds., Berlin, Heidelberg, 1983, Springer Berlin Heidelberg, pp. 144–177.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms II*, vol. 306 of Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg, New York, 1993, https://doi.org/10.1007/978-3-662-06409-2.

[21] A. F. Izmailov and M. V. Solodov, *Newton-type methods for optimization and variational problems*, Springer Series in Operations Research and Financial Engineering, Springer, Cham, 2014, https://doi.org/10.1007/978-3-319-04247-3.

[22] C. Lemaréchal and C. Sagastizábal, *Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries*, SIAM J. Optim., 7 (1997), pp. 367–385, http://link.aip.org/link/?SJE/7/367/1.

[23] C. Lemaréchal and F. Oustry, *Growth conditions and U-Lagrangians*, Set-Valued Anal., 9 (2001), pp. 123–129, https://doi.org/10.1023/A:1011267019516.

[24] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The U-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.

[25] C. Lemaréchal and C. Sagastizábal, *Variable metric bundle methods: From conceptual to implementable forms*, Math. Program., 76 (1997), pp. 393–410, https://doi.org/10.1007/BF02614390.

[26] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM J. Optim., 13 (2002), pp. 702–725, https://doi.org/10.1137/s1052623401387623.

[27] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., (2013), pp. 135–163.

[28] J. Liang, J. Fadili, and G. Peyré, *Activity identification and local linear convergence of forward–backward-*

type methods, SIAM J. Optim., 27 (2017), pp. 408–437, https://doi.org/10.1137/16M106340X.
[29] S. Liu, A. Eberhard, and Y. Luo, *The U-Lagrangian, fast track, and partial smoothness of a prox-regular function*, Set-Valued Var. Anal., 28 (2020), pp. 369–394, https://doi.org/10.1007/s11228-019-00518-z.
[30] S. Liu, C. Sagastizábal, and M. Solodov, *Subdifferential enlargements and continuity properties of the VU-decomposition in convex optimization*, in Nonsmooth Optimization and Its Applications, S. Hosseini, B. S. Mordukhovich, and A. Uschmajew, eds., Springer International Publishing, Cham, 2019, pp. 55–87, https://doi.org/10.1007/978-3-030-11370-4_4.
[31] R. Mifflin and C. Sagastizábal, *Optimization Stories*, vol. Extra Volume ISMP 2012, ed. by M. Grötschel, DOCUMENTA MATHEMATICA, 2012, ch. A Science Fiction Story in Nonsmooth Optimization Originating at IIASA, p. 460.
[32] R. Mifflin and C. Sagastizábal, *VU-decomposition derivatives for convex max-functions*, Ill-Posed Variational Problems and Regularization Techniques, 477 (1999), pp. 167–186.
[33] R. Mifflin and C. Sagastizábal, *On VU-theory for functions with primal-dual gradient structure*, SIAM J. Optim., 11 (2000), pp. 547–571.
[34] R. Mifflin and C. Sagastizábal, *Proximal points are on the fast track*, J. Convex Anal., 9 (2002), pp. 563–580.
[35] R. Mifflin and C. Sagastizábal, *Primal-dual gradient structured functions: second-order results; links to epi-derivatives and partly smooth functions*, SIAM J. Optim., 13 (2003), pp. 1174–1194.
[36] R. Mifflin and C. Sagastizábal, *UV-smoothness and proximal point results for some nonconvex functions*, Optim. Methods Softw., 19 (2004), pp. 463–478, https://doi.org/10.1080/10556780410001704902.
[37] R. Mifflin and C. Sagastizábal, *Relating U-Lagrangians to second order epiderivatives and proximal tracks*, J. Convex Anal., 12 (2005), pp. 81–93.
[38] R. Mifflin and C. Sagastizábal, *A VU-algorithm for convex minimization*, Math. Program., 104 (2005), pp. 583–608.
[39] Y. Nesterov, *Smooth Convex Optimization*, Springer US, 2004, pp. 51–110, https://doi.org/10.1007/978-1-4419-8853-9_2.
[40] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[41] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Program., 89 (2000), pp. 1–33, https://doi.org/10.1007/s101070000166.
[42] C. Planiden and T. Rajapaksha, *Linear convergence of the derivative-free proximal bundle method on convex nonsmooth functions, with application to the derivative-free $\mathcal{VU}$-algorithm*, Set-Valued Var. Anal., 32 (2024), https://doi.org/10.1007/s11228-024-00718-2.
[43] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 1998.
[44] C. Sagastizábal, *A $\mathcal{VU}$-point of view of nonsmooth optimization*, in Proceedings of the International Congress of Mathematicians 2018- Invited Lectures, vol. 3, 2018, pp. 3785–3806.
[45] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal, *The degrees of freedom of partly smooth regularizers*, Ann. Inst. Statist. Math., 69 (2017), pp. 791–832, https://doi.org/10.1007/s10463-016-0563-z.
[46] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Trans. Signal Process., 57 (2009), pp. 2479–2493, https://doi.org/10.1109/TSP.2009.2016892.

Appendix A. Technical results and concepts required to show Theorems 2.4 and 4.3.

Some continuity and smoothness properties of the bivariate $\mathcal{U}$-Lagrangian depend on Lemma A.1 below, that considers the pair (u, g_v) as a perturbation parameter in a family of parametric minimization problems with value function equal to the $\mathcal{U}$-Lagrangian, and solution mapping equal to the set of $\mathcal{V}$-minimizers.

Lemma A.1. *Given $G_v(\bar{x})$ from (2.4), the mapping*

$$\Phi(w, u, g_v) := f(\bar{x} + Uu + Vw) - \langle g_v, V^{\mathsf{T}} V w \rangle + \delta_{G_v(\bar{x})}(g_v)\,,$$

defined from $\mathbb{R}^{n-m} \times \mathbb{R}^m \times \mathbb{R}^{n-m}$ to $\overline{\mathbb{R}}$, is proper, lsc, and level-bounded in w, locally uniformly *in (u, g_v); see [43, Definition 1.16].*

Proof. Clearly, Φ is proper because f is finite-valued and Φ is lsc by the continuity of f and closedness of $G_v(\bar{x})$, which follows from the definitions in (2.4). To show the property of uniform level-boundedness, for all $\alpha \in \mathbb{R}$ consider the mapping $S_\alpha(u, g_v) := \{w :$

$\Phi(w,u,g_v) \le \alpha\}$. This mapping is nonempty (possibly unbounded) only when $g_v \in G_v(\bar{x})$,

$$S_\alpha(u,g_v) = \{w : f(\bar{x}+Uu+Vw) - \left\langle g_v, V^\top V w\right\rangle \le \alpha\}\,.$$

In view of (2.3), for any such g_v and $(u,w) \in \mathbb{R}^m \times \mathbb{R}^{n-m}$ with w an element of $S_\alpha(u,g_v)$, we have $\alpha \ge f(\bar{x}+Uu+Vw) - \langle g_v, V^\top V w\rangle \ge f(\bar{x}) + \langle \bar{g}_u, u\rangle + \eta\|Vw\|$, and therefore, $S_\alpha(u,g_v) \subset T(u,g_v) := \{w : f(\bar{x}) + \langle \bar{g}_u, u\rangle + \eta\|Vw\| \le \alpha\}$. By [43, Example 5.17(b)], it suffices to show that S_α is uniformly bounded below. We first derive a uniform local bound for the larger mapping T, by considering $(u,g_v) \in B(\bar{u},\epsilon)\times G_v(\bar{x}) \cap B(\bar{g}_v,\epsilon)$, for some $(\bar{u},\bar{g}_v) \in \mathbb{R}^m \times G_v(\bar{x})$. As $\bar{g}_u$ is fixed, the uniform bound follows, because $\langle \bar{g}_u, u\rangle \ge -\|\bar{g}_u\|\Big(\epsilon + \|\bar{u}\|\Big)$, which in particular yields that $\eta\|Vw\| \le \alpha - f(\bar{x}) + \|\bar{g}_u\|\Big(\epsilon + \|\bar{u}\|\Big)$ for all $w \in S_\alpha(u,g_v) \subset T(u,g_v)$. □

Lemma A.1 is useful to show Theorem 2.4 because for all $(u,g_v) \in \mathbb{R}^m \times G_v(\bar{x})$,

$$L_U(u,g_v) = \inf_w \Phi(w,u,g_v) \quad \text{and} \quad W(u,g_v) = \arg\min_w \Phi(w,u,g_v)\,.$$

Proof of Theorem 2.4.

Proof. For notational convenience, we define $G := G_v(\bar{x})$. All the subsequent references in this proof are from the book [43], noting that the assumptions in the invoked statements hold thanks to Lemma A.1. To see item (i), apply first Theorem 1.17(a), to show that L_U is proper and lsc on $\mathbb{R}^m \times \mathbb{R}^{n-m}$. For each $(u,g_v) \in \mathbb{R}^m \times G$, $L_U(u,g_v) \le \Phi(0,u,g_v) = f(\bar{x}+Uu) < +\infty$. Consequently, L_U is finite-valued on $\mathbb{R}^m \times G$. From Theorem 1.17(c), we have that L_U is continuous on $\mathbb{R}^m \times G$, as for any $\bar{w} \in W(\bar{u},\bar{g}_v)$ the function $\Phi(\bar{w},\cdot)$ is continuous in (u,g_v) at $(\bar{u},\bar{g}_v)$ relative to $\mathbb{R}^m \times G$. Consequently, item (ii) follows from Theorem 7.41(b). Item (iii) is derived from Example 5.22, by exhibiting a point $(\bar{u},\bar{g}_v) \in \mathbb{R}^m \times G$ such that $W(\bar{u},\bar{g}_v)$ is single valued with $L_U(\bar{u},\bar{g}_v)$ continuous relative to $\mathbb{R}^m \times G$. The latter condition is ensured by item (i) while the former is achieved by taking $\bar{u} = 0$ and $\bar{g}_v \in G$ and applying (2.2), noting that $W(0,\bar{g}_v) = W^{\bar{g}}(0) = \{0\}$. To show the next item, according to Corollary 10.14(a), we need to verify that for any $(u,g_v) \in \operatorname{int}\mathbb{R}^m \times G$, it holds that

$$\bigcup_{w\in W(u,g_v)} \left\{(s^1,s^2) \colon (0,s^1,s^2) \in \partial^\infty \Phi(w,u,g_v)\right\} = \{(0,0)\}\,. \tag{A.1}$$

To this end, define $\Phi = h_1 + h_2$ for the following two functions from $\mathbb{R}^{n-m}\times\mathbb{R}^m\times\mathbb{R}^{n-m}$ to $\overline{\mathbb{R}}$:

$$h_1(w,u,g_v) := f(\bar{x}+Uu+Vw) + \delta_G(g_v) \quad \text{and} \quad h_2(w,u,g_v) := -\left\langle g_v, V^\top V w\right\rangle + \delta_G(g_v)\,.$$

Since h_1 is proper, lsc and convex and h_2 is strictly differentiable at (w,u,g_v) for any $g_v \in \operatorname{int} G$, applying Exercises 10.10 and 10.7, we obtain that

$$\partial^\infty\Phi(w,u,g_v) = \partial^\infty h_1(w,u,g_v) = \left\{(V^\top s, U^\top s, 0) \colon s \in \partial^\infty f(\bar{x}+Uu+Vw)\right\} = \{(0,0,0)\}\,,$$

where the last equality holds because the horizon subdifferential of the finite-valued convex function f is null. As claimed, (A.1) holds and (iv) follows. Item (v) is derived from the result $W(0,g_v) = \{0\}$, item (iv), and Corollary 10.14(b), by showing that

$$\left\{(y^1,y^2) \colon (0,y^1,y^2) \in \partial\Phi(0,0,g_v)\right\} = \{(\bar{g}_u, 0)\}\,. \tag{A.2}$$

Once more, from Exercises 10.10 and 10.7, we get that

$$\partial\Phi(0,0,g_v) = \left\{(V^\top y - V^\top V g_v, U^\top y, 0) \colon y \in \partial f(\bar{x})\right\}\,.$$

The expression in (A.2) then follows from (2.1), concluding the proof of item (v). Next, to see item (vi), fix any $g_v \in V^\dagger \operatorname{ri} \partial f(\bar{x})$. For each $u \in \mathbb{R}^m$, the partial subdifferential $\partial_u L_U(u, g_v)$ is defined to be $\partial L_U^g(u)$. For each $g_v \in V^\dagger \operatorname{ri} \partial f(\bar{x})$, $L_U^g(u) = \inf_w \Phi g_v(w, u)$, where $\Phi g_v(w, u) := f(\bar{x} + Uu + Vw) - \langle g_v, V^\mathsf{T} V w\rangle$. In view of Lemma A.1, $\Phi g_v(w, u)$ is level-bounded in w locally uniformly in u because in this case g_v is a parameter and (2.3) holds for some η_{g_v}. Noting that $L_U^{g_v}(u)$ is convex, we can apply Corollary 10.13 to obtain that $\partial L_U^{g_v}(u) = \{y \in \mathbb{R}^m \colon (0, y) \in \partial \Phi g_v(w, u)\}$ for any $w \in W^{g_v}(u)$. It is next seen that $\partial \Phi g_v(w, u) = \{(V^\mathsf{T} s - V^\mathsf{T} V g_v, U^\mathsf{T} s) \colon s \in \partial f(\bar{x} + Uu + Vw)\}$. We also have that $0 \in V^\mathsf{T} s - V^\mathsf{T} V g_v$ equivalent to $(V^\mathsf{T} V)^{-1} V^\mathsf{T} s = g_v$. Consequently, item (vi) holds. Finally, to see item (vii), recall that by Theorem 2.4(v), L_U is differentiable at $(0, \bar{g}_v)$, and hence,

$$L_U(u, g_v) = L_U(0, \bar{g}_v) + \langle \nabla L_U(0, \bar{g}_v), (u, g_v) - (0, \bar{g}_v)\rangle + o\Big(Uu + V(g_v - \bar{g}_v)\Big)\,.$$

Substituting $L_U(0, \bar{g}_v)$ and its gradient by their explicit expressions gives the final item. □

Definition of partial smoothness. A convex function f is said to be partly smooth at x relative to a set $\mathcal{M}$ if $\mathcal{M}$ is a manifold around x and the following three properties hold:

(i) (restricted smoothness) the restriction $f\big|_{\mathcal{M}\cap\mathcal{N}(x)}$ of f to $\mathcal{M}$ in a neighborhood $\mathcal{N}(x)$ of x is of class C^2;
(ii) (normals parallel to subdifferential) $N_{\mathcal{M}}(x) = \mathcal{V}(x)$;
(iii) (subgradient continuity) the subdifferential ∂f is continuous at x relative to $\mathcal{M}$.

Proof of Theorem 4.3.

Proof. From property (ii) in the partial smoothness definition, when the set $\mathcal{M}$ is not a singleton, the subspaces tangent and normal to the manifold at $\bar{x}$ coincide respectively with $\mathcal{U}$ and $\mathcal{V}$. Then, by [26, Theorem 6.1], there exist a neighborhood $\mathcal{N} \subset \mathbb{R}^n$ of $\bar{x}$, a neighborhood $\mathcal{N}(0) \subset \mathbb{R}^m$ of 0, and a function $v_{\partial f} \colon \mathbb{R}^m \to \mathbb{R}^{n-m}$ such that for all $u \in \mathcal{N}(0)$,

$$v_{\partial f}(u) \text{ is of class } C^2\,,\ v_{\partial f}(u) = O(\|u\|^2)\,,\ \text{and } \mathcal{M} \cap \mathcal{N} = \big\{\bar{x} + Uu + V v_{\partial f}(u) \colon u \in \mathcal{N}(0)\big\}\,.$$

From the last relation, the restriction in the partial smoothness definition (i) has the expression $f\big|_{\mathcal{M}\cap\mathcal{N}} = f(\bar{x}+Uu+Vv_{\partial f}(u))$. Again by [26, Theorem 6.1], for all $g \in \operatorname{ri} \partial f(\bar{x})$, the function $h^g(w) := f(\bar{x} + Uu + Vw) - \langle g, \bar{x} + Uu + Vw\rangle$ has $v_{\partial f}$ as a sharp minimizer. Then

$$L_U(u, g_v) = f\big|_{\mathcal{M}\cap\mathcal{N}}(u) - \big\langle g_v, V^\mathsf{T} V v_{\partial f}(u)\big\rangle \tag{A.3}$$

because $\langle g_v, Vw\rangle = \langle g, \bar{x} + Uu + Vw\rangle$, which shows that $h^g(w)$ is the minimand defining the bivariate $\mathcal{U}$-Lagrangian. To show

$$\mathcal{M} \cap \mathcal{N} = \big\{\bar{x} + Uu + V v_{\partial f}(u) \colon u \in \mathcal{N}(0)\big\}\,, \tag{A.4}$$

suppose now $\bar{g} \in \operatorname{ri} \partial f(\bar{x})$. In (A.3), the linear term defines the bivariate function on $\mathcal{N}(0) \times V^\dagger \operatorname{ri} \partial f(\bar{x})$ given by

$$P(u, g_v) := \big\langle g_v, V^\mathsf{T} V v_{\partial f}(u)\big\rangle = g^\mathsf{T} V v_{\partial f}(u)\,,$$

by definition of the $\mathcal{V}$-components. The gradient of this function on the first component is

$$\nabla_u P(u, g_v) = \mathcal{J} v_{\partial f}(u)^\mathsf{T} V^\mathsf{T} g = \mathcal{J} v_{\partial f}(u)^\mathsf{T} V^\mathsf{T} V g_v\,. \tag{A.5}$$

where $\mathcal{J} v_{\partial f}(u)$ is the Jacobian of $v_{\partial f}(u)$. Since $W(u, g_v) \ni v_{\partial f}(u) = O(\|u\|^2)$ by Theorem 4.3, combining Theorem 2.4(iii) and the fact that

$$\mathcal{J} v_{\partial f}(u) = \mathcal{J} v_{\partial f}(0) + O(u) = O(u)\,,$$

gives the following expansion for the gradient of the $\mathcal{U}$-Lagrangian from (A.3):

$$\nabla_u L_U(u, g_v + z) = \nabla_u L_U(0, g_v) + \nabla^2_{uu} L_U(0, g_v)u + \frac{\partial^2}{\partial u \partial g_v} L_U(0, g_v)z + o(u, z)\,,$$

for any $z \in \mathbb{R}^{n-m}$ small enough such that $g_v + z \in V^\dagger \operatorname{ri} \partial f(\bar{x})$. In this expansion, by (A.5) and (A.3), the cross-derivative has the form

$$\frac{\partial^2}{\partial u \partial g_v} L_U(0, g_v) = -\frac{\partial^2}{\partial u \partial g_v} P(0, g_v) = -\mathcal{J} v_{\partial f}(0)^\top V^\top V = 0\,.$$

The expression for the partial $\mathcal{U}$-Hessian follows, because $\nabla_u L_U(0, g_v) = \bar{g}_u$. By (A.5),

$$\nabla_u P(u, g_v) = o(u, g_v) \text{and, in particular, } \nabla_u P(0, g_v) = 0\,. \tag{A.6}$$

Hence, by (A.3) and the fact that $f\big|_{\mathcal{M}\cap\mathcal{N}}$ is a composite function of u, $\nabla_u f\big|_{\mathcal{M}\cap\mathcal{N}}(0) = \nabla_u L_U(0, g_v) + \nabla_u P(0, g_v) = \bar{g}_u$. Then the smoothness of $f\big|_{\mathcal{M}\cap\mathcal{N}}$ yields $\nabla_u f\big|_{\mathcal{M}\cap\mathcal{N}}(u) = \bar{g}_u + \nabla^2_{uu} f\big|_{\mathcal{M}\cap\mathcal{N}}(0)u + o(u)$. Consequently, by (A.3) and (A.6),

$$\nabla_u L_U(u, g_v) = \nabla_u f\big|_{\mathcal{M}\cap\mathcal{N}}(u) - \nabla_u P(u, g_v) = \bar{g}_u + \nabla^2_{uu} f\big|_{\mathcal{M}\cap\mathcal{N}}(0)u + o(u) - o(u, g_v)\,.$$

As $o(u) - o(u, g_v) = o(u, g_v)$, we see from Prop. 4.2 that $\nabla^2_{uu} f\big|_{\mathcal{M}\cap\mathcal{N}}(0)$ is a partial $\mathcal{U}$-Hessian of f at $\bar{x}$ associated with 0.

Now consider the special setting of f in (3.1), with h finite-valued polyhedral, so that $h(x) = \max_{i\in I}\left\{\langle a^i, x\rangle + b^i\right\}$ for some finite index set $I \neq \emptyset$. and, for the "activity" index set $I(x) = \left\{i \in I\colon \langle a^i, x\rangle + b^i = f(x)\right\}$,

$$\partial h(x) = \left\{ \sum_{i\in I(x)} \alpha_i a^i \colon \sum_{i\in I(x)} \alpha_i = 1,\ \alpha_i \geq 0\ (i \in I(x)) \right\}. \tag{A.7}$$

The function h is partly smooth at any $\bar{x}$ relative to $\mathcal{M}_{\bar{x}} := \{x \in \mathbb{R}^n \colon I(x) = I(\bar{x})\}$, [26, Example 3.4]. From [26, Corollary 4.7] we see that f is partly smooth at $\bar{x}$ relative to $\mathcal{M}_{\bar{x}}$ and that h is partly smooth at $\bar{x}$ relative to $\mathcal{M}$. By [14, Corollary 4.2], the active manifold in the definition of partial smoothness is unique. Hence, near $\bar{x}$ we have $\mathcal{M} \equiv \mathcal{M}_{\bar{x}}$.

Next, we show that whenever a vector $\bar{v} \in \mathbb{R}^{n-m}$ satisfies $\bar{x} + Uu + V\bar{v} \in \mathcal{M} \cap \mathcal{N}$, it must hold that $\bar{v} = 0$; so $\mathcal{M} \cap \mathcal{N} = \{\bar{x} + Uu \colon u \in \mathcal{N}(0)\}$. To show the claim, consider $i \in I(\bar{x})$ and note that $I(\bar{x} + Uu + V\bar{v}) = I(\bar{x})$, because $\bar{x} + Uu + V\bar{v} \in \mathcal{M}_{\bar{x}}$. That is, $h(\bar{x}+Uu+V\bar{v}) = \langle a^i, \bar{x} + Uu + V\bar{v}\rangle + b^i = \langle a^i, \bar{x}\rangle + b^i + \langle a^i, Uu + V\bar{v}\rangle = h(\bar{x}) + \langle a^i, Uu + V\bar{v}\rangle$. Therefore,

$$h(\bar{x} + Uu + V\bar{v}) - h(\bar{x}) = \langle a^i, Uu + V\bar{v}\rangle\,. \tag{A.8}$$

Because $0 \in \partial f(\bar{x})$, we have that $-\nabla q(\bar{x}) \in \partial h(\bar{x})$ and $a^i + \nabla q(\bar{x}) \in \mathcal{V}$. As a result, $\langle a^i + \nabla q(\bar{x}), Uu\rangle = 0$ and $\langle a^i, Uu\rangle = -\langle \nabla q(\bar{x}), Uu\rangle$. And (A.8) yields $\langle V\bar{v}, a^i\rangle = h(\bar{x}+Uu+V\bar{v}) - h(\bar{x}) + \langle \nabla q(\bar{x}), Uu\rangle$. The expression $-\nabla q(\bar{x}) = \sum_{i\in I(\bar{x})} \bar{\alpha}_i a^i$ with $\sum_{i\in I(\bar{x})} \alpha_i = 1$, $\bar{\alpha}_i \geq 0$ $(i \in I(\bar{x}))$ implies that $\langle V\bar{v}, -\nabla q(\bar{x})\rangle = h(\bar{x} + Uu + V\bar{v}) - h(\bar{x}) + \langle \nabla q(\bar{x}), Uu\rangle$, and hence, $\langle V\bar{v}, a^i + \nabla q(\bar{x})\rangle = 0$. Because $\mathcal{V} = \operatorname{span}(\partial h(\bar{x}) + \nabla q(\bar{x})) = \operatorname{span}\left\{a^i + \nabla q(\bar{x}) \colon i \in I(\bar{x})\right\}$, we actually have that $\langle V\bar{v}, z\rangle = 0$ for all $z \in \mathcal{V}$ and our claim that $\bar{v} = 0$ follows.

Consider $u \in \mathcal{N}(0)$. Since $\mathcal{M} \cap \mathcal{N} = \{\bar{x} + Uu \colon u \in \mathcal{N}(0)\} = \mathcal{M}_{\bar{x}} \cap \mathcal{N}$ and $I(\bar{x} + Uu) = I(\bar{x})$, By (A.7), it holds that $\partial h(\bar{x}+Uu) = \partial h(\bar{x})$. Consequently, $U^\top \partial h(\bar{x}+Uu) = U^\top \partial h(\bar{x}) = \bar{g}_u$. On the other hand, in view of (A.4), we can take $\mathcal{N}$ and $\mathcal{N}(0)$ sufficiently small such that $v_{\partial f}(u) \equiv 0$. Consequently, the restriction of f on $\mathcal{M}\cap\mathcal{N}$ is $f(\bar{x}+Uu) = q(\bar{x}+Uu) + h(\bar{x}+Uu)$ and, therefore, $\nabla_u f(\bar{x}+Uu) = U^\top \nabla q(\bar{x}+Uu) + U^\top \partial h(\bar{x}+Uu)$. Because $U^\top \partial h(\bar{x}+Uu) = \bar{g}_u$, this completes the proof, as then $\nabla^2_{uu} f(\bar{x} + Uu) = U^\top \nabla^2 q(\bar{x} + Uu)U$. □