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 structure of nonsmoothness and for developing algorithms with superlinear convergence in those challenging (for fast convergence) settings. We extend the theory by defining a certain bivariate $\mathcal{U}$-Lagrangian function and the 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, and 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. As an example, our method is applied to solving ℓ_1-regularized problems. We exhibit the explicit process of constructing a basis of the $\mathcal{U}$-space and of calculating the $\mathcal{U}$-Hessian. We conclude with numerical results illustrating 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

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

The $\mathcal{VU}$-theory introduced in [20, 22] (closely related to partial smoothness [24]) has been used for the study of smooth structures in nonsmooth functions in [30, 21, 32, 34, 7, 2]. As explained in [42, 28], 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 [39, 35, 11, 13, 3]; see also [25, 40]. The approach is to decompose the space $\mathbb{R}^n$ into two orthogonal subspaces called $\mathcal{V}$ and $\mathcal{U}$, depending on a point $\bar{x}$. The $\mathcal{V}$-space is defined to be the subspace parallel to the affine hull of the subdifferential $\partial f(\bar{x})$, and $\mathcal{U}$ consists of the directions such that the directional derivative $f'(\bar{x};\cdot)$ is linear. Roughly speaking, the $\mathcal{V}$ and $\mathcal{U}$ spaces are defined so that near the point $\bar{x}$ the nonsmoothness of f is captured in the $\mathcal{V}$-space and the smoothness of f is captured in the $\mathcal{U}$-space. Through a parametrized Lagrangian defined on the $\mathcal{U}$-space, 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 [22], 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 in the $\mathcal{U}$-subspace of the $\mathcal{U}$-Lagrangian (where the $\mathcal{U}$-Lagrangian looks smooth). The superlinear convergence requires the existence of a positive definite $\mathcal{U}$-Hessian. This conceptual approach did not address how to identify the $\mathcal{VU}$-geometry along an algorithmic procedure. In order to pass from theory

*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 Software, South China Normal University, Shishan Town, Nanhai District, Foshan, Guangdong, 528225, China (shuai0liu@gmail.com).

‡ Universidade Estadual de Campinas, Rua Sérgio Buarque de Holanda 651, Campinas, SP, 13083-859, Brazil sagastiz@unicamp.br).

§ IMPA - Instituto de Matemática Pura e Aplicada, Estrada Dona Castorina 110, Rio de Janeiro, RJ, 22460-320, Brazil (solodov@impa.br).

to computational implementation, it is important to examine the structure of (nonsmooth) functions. The specific $\mathcal{VU}$-theory for finite-max functions and the numerical analysis of the relevant $\mathcal{VU}$-objects were considered in [29] and [11]. A more general class of functions given in [30, 32], said to have primal-dual-gradient (PDG) structure, identifies a *fast track*, which are points from where fast Newton-type steps are possible. Links with the $\mathcal{VU}$-theory of partly smooth functions are the subject of [15, 26, 10, 14]. In particular, [14, 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 [31], where it is shown 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 [18] that approximates this step by computing proximal points of succesive cutting-planes models of f. Fully implementable (fast) $\mathcal{VU}$-algorithms for solving (1.1) are scarce, because they need to approximate sufficiently well both the (exact proximal) $\mathcal{V}$-step and the (exact) $\mathcal{U}$-Newton direction. Generally, this involves solving at least two quadratic programming problems per iteration. Such is the case of [35], the first $\mathcal{VU}$-algorithm for problems like (1.1), where no specific structure for f is required. The work [36] proposes two sequential Newtonian methods based on local parameterizations obtained from relating $\mathcal{VU}$-theory with Riemannian geometry. Like for the $\mathcal{VU}$-theory, considering a family of functions with specific properties leads to more targeted implementations. For maximum eigenvalue and convex finite-max functions, we mention [39] and [13]. How the $\mathcal{VU}$-decomposition can be iteratively constructed by bundle methods for a certain class that includes max-functions is explored in [7].

When f in (1.1) has additive structure as in (3.1) further below, subtle geometrical properties of the proximal operator allow [2] to asymptotically detect the correct $\mathcal{V}$-step by means of a proximal gradient (PG) method [4]. Depending on the nonsmooth term, this calculation is explicit, or entails solving a simple quadratic program. The $\mathcal{U}$-step corrects the PG iterate by a certain Newton-like direction, computed by solving a (possibly another, second) quadratic programming problem. When applied to the same class of functions, our proposal eliminates the latter second quadratic program, thanks to a suitable shifting of the optimal subgradient resulting from the PG iterate calculation. The full corresponding algorithm, named Proximal Gradient $\mathcal{VU}$ method or PGVU for short, is given in Algorithm 3.1 below.

In order to analyze convergence of the PGVU method, an important extension of the $\mathcal{VU}$-theory is needed. In all the mentioned studies, given a point $\bar{x} \in \mathbb{R}^n$, the $\mathcal{U}$-Lagrangian is a single-variable function, defined considering a subgradient $\bar{g} \in \partial f(\bar{x})$ as a parameter. But in the algorithmic setting we have to deal with a sequence of subgradients ($g^k \in \partial f(x^k)$ at iteration k), that change the parameter defining the $\mathcal{U}$-Lagrangian along iterations. As illustrated by Example 2.2 below, with more than one fast track converging to a minimizer $\bar{x}$, different subgradients yield $\mathcal{U}$-Lagrangians associated with different fast tracks. To prevent possible oscillatory behaviour, in our $\mathcal{U}$-Lagrangian definition, the subgradient is no longer a parameter, but another variable. Accordingly, we extend the theories to such bivariate $\mathcal{U}$-Lagrangian, defining a partial $\mathcal{U}$-Hessian as the general partial Hessian of the new $\mathcal{U}$-Lagrangian. Properties that hold for the single-variable $\mathcal{U}$-Lagrangian are now shown to hold for the bivariate $\mathcal{U}$-Lagrangian. Thanks to our extended $\mathcal{VU}$-theory, the proposed Proximal Gradient $\mathcal{VU}$-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 $0 \in \operatorname{ri} \partial f(\bar{x})$. Moreover, we show that any convex partly smooth function that satisfies $0 \in \operatorname{ri} \partial f(\bar{x})$ automatically has a partial $\mathcal{U}$-Hessian. Finally, we demonstrate the constructive process through the application of PGVU to ℓ_1-regularized minimization.

The rest of the paper is organized as follows. In the remaining part of Section 1 we introduce 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 give the definition and smoothness properties of the bivariate $\mathcal{U}$-Lagrangian function and show that computing the proximal point can serve as the $\mathcal{V}$-step. In Section 3, we give the details of our PGVU-method and show its global convergence. The definition of a partial $\mathcal{U}$-Hessian is given in Section 4 and, under the assumption of a positive definite partial $\mathcal{U}$-Hessian and $0 \in \operatorname{ri}\partial f(\bar{x})$, we prove that PGVU is superlinearly convergent. In Section 4.2, we show that all convex partly smooth functions satisfying $0 \in \operatorname{ri}\partial f(\bar{x})$ have a partial $\mathcal{U}$-Hessian at $\bar{x}$. Section 5 applies the proposed Proximal Gradient $\mathcal{VU}$-method to ℓ_1-regularized minimization. We first verify the existence of a $\mathcal{U}$-Hessian. Then we provide an inexact prox-step as the $\mathcal{V}$-step and construct a basis for the $\mathcal{U}$-space. Numerical results reported at the end of this section show that PGVU performs well both in terms of computational time and accuracy. Concluding remarks are given in Section 6.

Notation. We mostly follow [41]. 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 $f'(x; d)$ is 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 $\bar{x}$. The indicator function of a convex set C 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 [38] 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 the definition of the two subspaces in question.

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

$$\mathcal{V}(\bar{x}) = \operatorname{span}(\partial f(\bar{x}) - g), \quad \mathcal{U}(\bar{x}) = \mathcal{V}(\bar{x})^\perp,$$

where g 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:

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

More specifically, the $\mathcal{U}$-and $\mathcal{V}$-components of any $x \in \mathbb{R}^n$ are defined by

$$x_u := U^\intercal x, \qquad x_v := V^\dagger x.$$

By the definitions of $\mathcal{V}$ and $\mathcal{U}$, the set $U^\intercal \partial f(\bar{x})$ is a singleton, and hence,

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

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

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

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

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

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 [22, Theorem 3.2, Theorem 3.3(ii)] that

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

Evaluating the $\mathcal{U}$-Lagrangian at some $\mathcal{V}$-minimizer yields the following special first-order expansion for f:

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

When, in addition, 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 below.

In the sequel, we shall introduce an important advance with respect to the previous $\mathcal{VU}$-literature. It has to do with the following considerations. Note that the original $\mathcal{U}$-Lagrangian from [22] was defined for some fixed $\bar{g} \in \operatorname{ri} \partial f(\bar{x})$. While it is true that (2.1) guarantees that the $\mathcal{U}$-component $\bar{g}_u$ is the same for all $\bar{g} \in \partial f(\bar{x})$, the argument 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 function with structured nonsmoothness). Given a scalar $a > 0$, for $(u, v) \in \mathbb{R}^2$ the function

$$F(u, v) = \max \left\{ \frac{a}{2} u^2, |v| \right\} = \frac{a}{2} u^2 + \max \left\{ 0, |v| - \frac{a}{2} u^2 \right\}$$

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 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, we obtain that

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

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

For functions with structured nonsmoothness, the $\mathcal{VU}$-decomposition is useful to reveal hidden smoothness. For F, this relates to the trajectory below:

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

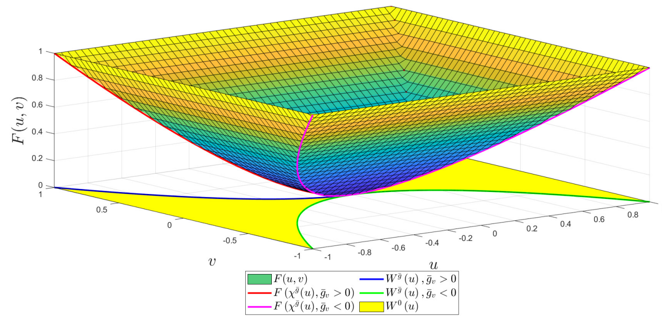

Fig. 2.1: Function F in Example 2.2, the associated $\mathcal{V}$-minimizers, and fast tracks

In the parlance of [31], this is a *fast track* along which F can be expanded up to second order:

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

The fast track $\{(u, \frac{a}{2}u^2)\}$ in Figure 2.1 is obtained with $\bar g = 0$.

The situation illustrated by Example 2.2, with trajectories of smoothness depending on the choice of the subgradient $\bar g$, motivates the consideration of a *bivariate* $\mathcal{U}$-Lagrangian, introduced in this work.

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:*

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

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

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

The notation in this work, $L_U(u, \bar g_v)$, should not be confused with [31], where only u is a variable, and the semicolon in $L_U(u; \bar g_v)$ is used to expose that $\bar g_v$ is a parameter.

The set $\operatorname{ri}\partial f(\bar x) := \{g \in \mathbb{R}^n \colon g + \operatorname{int} B(0,\eta) \cap \mathcal{V} \subset \partial f(\bar x) \text{ for some } \eta > 0\}$ defines the subdifferential relative interior. For each $g \in \operatorname{ri}\partial f(\bar x)$, we have $U\bar g_u + V g_v + \frac{\eta V w}{\|Vw\|} \in \partial f(\bar x)$ for all $w \in \mathbb{R}^{n-m}$ and the convexity of f implies that, for any $(u, w) \in \mathbb{R}^m \times \mathbb{R}^{n-m}$, it holds that

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

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

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

To show some continuity and smoothness properties of the $\mathcal{U}$-objects, we consider (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 2.4. *Given $G_v(\bar{x})$ from* (2.4)*, the mapping*

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

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 [41, 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 (and can be unbounded) only when $g_v \in G_v(\bar{x})$, in which case

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

In view of (2.3), for any such g_v and $(u, w) \in \mathbb{R}^{m+p}$ with w an element of $S_\alpha(u, g_v)$, we have $\alpha \ge f(\bar{x} + Uu + Vw) - \langle g_v, V^{\mathsf{T}} V w\rangle \ge f(\bar{x}) + \langle \bar{g}_u, u\rangle + \eta\|Vw\|$, 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 [41, 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 $w \in S_\alpha(u, g_v) \subset T(u, g_v)$. □

As stated, minimizing the mappings in Lemma 2.4 in the first variable yields, for all $(u, g_v) \in \mathbb{R}^m \times G_v(\bar{x})$, the bivariate $\mathcal{U}$-objects in Definition 2.3:

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

Thanks to Lemma 2.4, several important relations known for the single-variable $\mathcal{U}$-Lagrangian hold in our new bivariate context.

THEOREM 2.5 (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.*

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

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

Proof. For notational convenience, we define $G := G_v(\bar{x})$. All the subsequent references in this proof are from the book [41], noting that the assumptions in the invoked statements hold thanks to Lemma 2.4. To see item (i), apply first Theorem 1.17(a), to show that L_U is proper and lsc on $\mathbb{R}^n$. 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

$$\bigcup_{w \in W(u, g_v)} \left\{(s^1, s^2) \colon (0, s^1, s^2) \in \partial^\infty \Phi(w, u, g_v)\right\} = \{(0, 0)\} \ . \tag{2.5}$$

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 $\overline{\mathbb{R}}$:

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

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

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

where the last equality holds because the horizon subdifferential of the finite-valued convex function f is null. As claimed, (2.5) 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

$$\left\{(y^1, y^2) \colon (0, y^1, y^2) \in \partial \Phi(0, 0, g_v)\right\} = \{(\bar{g}_u, 0)\} \ . \tag{2.6}$$

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

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

The expression in (2.6) then follows from (2.1), concluding the proof of item (v). Finally, 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 $\Phi g_v(w, u) := f(\bar{x} + Uu + Vw) - \langle g_v, V^\intercal V w\rangle$. In view of Lemma 2.4, $\Phi g_v(w, u)$ 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 $\partial \Phi g_v(w, u) = \{(V^\intercal s - V^\intercal V g_v, U^\intercal s) \colon s \in \partial f(\bar{x} + Uu + Vw)\}$. We also have that $0 \in V^\intercal s - V^\intercal V g_v$ equivalent to $(V^\intercal V)^{-1} V^\intercal s = g_v$. Consequently, item (vi) holds. ∎

All the properties listed in (2.2), with only u considered a variable, can now be compared with the statements in items (iii) and (iv) of Theorem 2.5, 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, [22, 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 established next.

Lemma 2.6 (Tangential trajectories). *Let $\bar g \in \operatorname{ri}\partial f(\bar x)$, $\bar x \in \mathbb{R}^n$. With the notation and assumptions of Lemma 2.4, we have that*

$$L_U(u, g_v) = f(\bar x) + \langle \bar g_u, u\rangle + o\Big(Uu + V(g_v - \bar g_v)\Big) . \tag{2.7}$$

Proof. By item (v) in Theorem 2.5, the function 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 (2.7). ∎

Together with the $\mathcal{U}$-Lagrangian given in Definition 2.3, from (2.7) we obtain the following first-order expansion for f:

$$f(\bar x + Uu + Vw(u, g_v)) = f(\bar x) + \langle \bar g_u, u\rangle + \Big\langle g_v, V^\top V w(u, g_v)\Big\rangle + o\Big(Uu + Vg_v\Big) .$$

When compared with the relation given in Section 2.1 for the single-variable setting, we see that for fast convergenge 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

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

Combined with Lemma 2.6, the following result provides a mechanism that makes the $\mathcal{V}$-step be tangential to the $\mathcal{U}$-subspace.

Lemma 2.7 (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 $\partial h(v) = V^\top \partial f(\bar x + Uu(p) + Vv)$. The necessary and sufficient optimality condition for $v(p) \in W(u(p); g_v)$ is $0 \in \partial h(v(p)) - V^\top V g_v = V^\top \partial f(\bar x + Uu(p) + Vv(p)) - V^\top V g_v = V^\top \partial f(p) - V^\top V g_v$ 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 2.5 is verified by g'_u. ∎

Proximal points are related to $\mathcal{V}$-minimizers through Lemma 2.7, 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) below. 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 2.7 applies.

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

$$
\boxed{\begin{array}{rl} f(x) \equiv q(x) + h(x) & \\ \text{for } q\colon \mathbb{R}^n \to \mathbb{R} & \text{convex, } C^2\text{-smooth} \\ & \text{with gradient Lipschitz constant denoted by } \beta \\ h\colon \mathbb{R}^n \to \mathbb{R} & \text{a continuous convex function, possibly nonsmooth,} \\ & \text{with an easy-to-compute proximal point.} \end{array}} \tag{3.1}
$$

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 [37, 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 $x^{k+1} = \text{prox}_{m(\cdot;x^k),\mu_k}(x^k)$. If $\mu_k \ge \beta$ and f satisfies an error bound, the proximal gradient iterates converge with linear rate [9, Theorems 3.1 and 5.5]; see also [1]. To achieve superlinear speed, in our method those iterations are corrected by a suitable $\mathcal{U}$-step; see Algorithm 3.1 below. The proximal gradient iteration in 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 $p = \text{prox}_{m(\cdot;x^k),\mu_k}(x^k)$ for the 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: set $\mu_k := 2\mu_k$
  compute $p = \text{prox}_{m(\cdot;x^k),\mu_k}(x^k)$
end
Output: μ_k and $p^k = p$.

Some comments regarding Procedure 1 are in order. The output p^k satisfies $q(p^k) \le q(x^k) + \left\langle \nabla q(x^k), p^k - x^k \right\rangle + \frac{\mu_k}{2}\|p^k - x^k\|^2$. With our assumptions in (3.1), this ensures that, once $\mu_k \ge \beta$, the procedure will terminate (in the bundle methods terminology [18], 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 $h \equiv f$. Since 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 2.7, 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 proximal point of the model to exact proximal point of the function).
Given the output p^k of Procedure 1, define

$$
g^k := \mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k) \quad \textit{and} \quad z^k := p^k + \frac{1}{\mu_k} g^k\,. \tag{3.3}
$$

Then it holds that

$$p^k = \operatorname{prox}_{f,\mu_k}(z^k)\,, \quad z^k = x^k + \frac{1}{\mu_k}(\nabla q(p^k) - \nabla q(x^k))\,, \quad g^k \in \partial f(p^k)\,. \tag{3.4}$$

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 $v^k := (p^k - \bar{x})_v \in W(u^k, g^k_v)$ if and only if $g^k_v \in V^\dagger \operatorname{ri} \partial f(\bar{x})$, in which case $g^k_u \in \partial_u L_U(u^k, g^k_v)$.*

(ii) *Furthermore, whenever $\mu_k > \beta$, it holds that*

$$\|p^k - \bar{x}\| \le \frac{\mu + \beta}{\mu - \beta}\|x^k - \bar{x}\| \quad \textit{and} \quad \|g^k\| \le \frac{2\mu(\mu + \beta)}{\mu - \beta}\|x^k - \bar{x}\|\,.$$

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 $\mu_k(x^k - p^k) - \nabla q(x^k) \in \partial h(p^k)$. Ttherefore, $\mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k) \in \nabla q(p^k) + \partial h(p^k) = \partial f(p^k)$. The remaining assertions in (3.4) 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 (3.3) is just $g^k \in \partial f(p^k)$. Item (i) follows from (3.4) and Lemma 2.7, written with p, g' therein replaced by p^k, g^k.

For the final item, first note that, by (3.3), it holds that

$$\|g^k\| \le (\mu_k + \beta)\|x^k - p^k\| \le (\mu_k + \beta)\Big(\|x^k - \bar{x}\| + \|\bar{x} - p^k\|\Big)\,, \tag{3.5}$$

Next, by (3.4) and the nonexpansiveness of the proximal operator,

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

Using the expression for z^k in (3.4) and the bound for ∇q from (3.1), we obtain that

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

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

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

After some rearrangements of terms we obtain that

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

The last inequality in (3.5) 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 depends on g^k_v 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 (3.3). Accordingly, the Proximal Gradient $\mathcal{VU}$-method given in Algorithm 3.1, stops when $\|g^k\| \le$ TOL for a given tolerance 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:

$$x^{k+1} = p^k - U_k Q_k U_k^\top g^k\,, \tag{3.6}$$

where U_k is a certain orthonormal matrix and Q_k is positive semidefinite. The purpose of this correction is to eventually track a trajectory where the function behaves smoothly, through the relation of $\mathcal{V}$-minimizers with proximal points. Thus, in (3.6) g^k is the shifted gradient from (3.3), Q^k asymptotically approximates the so-called $\mathcal{U}$-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 [7], and also Section 5 concerning the ℓ_1-regularized setting which is our illustration in this paper.

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.

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

Data: f as in (3.1), starting point $x^0 \in \mathbb{R}^n$, prox-parameter $\mu_0 > 0$, and a stopping tolerance TOL ≥ 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 $\|g^k\| \leq$ *TOL*;

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., [19]. Otherwise, the structural properties for f are essential to define suitable matrices in (3.6). 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, proximal gradient and proximal gradient VU algorithms). For the function F from Example 2.2, the smooth function in (3.1) is $q(u,v) = \frac{a}{2}u^2$ and the Lipschitz constant is $\beta = a = 2$.

We consider minimizing F with the proximal point method (P), the proximal gradient algorithm (PG), 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 $\max(\|x^k\|, \|p^k\|) \leq$ TOL.
- For PGVU, there is the additional $\mathcal{U}$-step, requiring the matrices in (3.6).
  - Outside of the fast track, that is when $p^k \notin \left\{(p_1, p_2) : |p_2| = \frac{a}{2}(p_1)^2\right\}$, the function is differentiable, $\mathcal{U}(p^k)$ is the whole space, and U_k is just the identity matrix. The matrix Q_k is set to $\begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix}^\dagger$ if $|p_2^k| < \frac{a}{2}(p_1^k)^2$, and $Q_k = 0$ 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 $\partial F(p^k) = \left\{(\xi a p_1^k, 1 - \xi)^\top : \xi \in [0,1]\right\}$. In this case, U_k is the orthonormal vector generated from $\left(sign(p_1^k), sign(p_2^k) a p_1^k\right)$ and $Q_k = a^{-1}$.

The methods were run with the same initial μ_0 and x^0, until the termination tolerance TOL $= 10^{-10}$ or a maximum number of iterations set to 50 achieved. 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.

Table 3.1: Methods' performance

$\mu^0 = 0.18$	$x^0 = (-1.2, 2.2)^\top$			$x^0 = (-1.1, 0.9)^\top$		
	P	PG	PGVU	P	PG	PGVU
Iterations	10	16	11	8	14	11
Digits	9	11	11	7	10	10
$\mu^0 = 10$	$x^0 = (-1.2, 2.2)^\top$			$x^0 = (-1.1, 0.9)^\top$		
	P	PG	PGVU	P	PG	PGVU
Iterations	50	50	45	50	50	16
Digits	2	3	23	3	4	23

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, and can be explained by observing the plots in Fig. 3.1, with the three trajectories generated for one of the starting points.

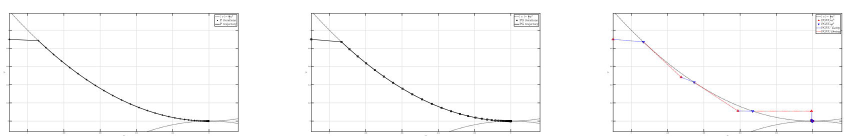

Fig. 3.1: P, PG, and PGVU iterations (left, middle, right), when minimizing F from $x^0 = (-1.1, 0.9)^\top$, starting with $\mu_0 = 10$. The dotted parabola is the fast track, rapidly identified by all the algorithms. Because $\mu_k = \mu_0$ is too large, P and PG take very small steps along the fast 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.

Figure 3.1 highlights the following important point: *for a nonsmooth optimization method to achieve high accuracy, identifying the fast track is necessary, but it is not sufficient.* Both P and PG iterates land soon on the smooth trajectory defined by the fast track. This is consistent with the theory in [31]; see also [3, Theorem 3.1]. Once on the fast track, P and PG both remain there. But because $\mu_k = \mu_0$ is too large, the progress between consecutive iterates becomes too small and, as illustrated by the left and middle plots in Fig. 3.1, those two methods see their performance severely impaired. On the other hand, on the right plot we see that the too large $\mu_k = \mu_0$ also makes PGVU take a too small $\mathcal{V}$-step, but afterwards its $\mathcal{U}$-step is long, in the "right direction" to solution, but also driving the iterate far from the fast track. The remedy is that the subsequent $\mathcal{V}$-step takes the iterate to the fast track again, and a new fast $\mathcal{U}$-step is possible again. The overall iterative process can be interpreted to work in a "predictor-corrector" fashion. Observe that PGVU achieves a much higher accuracy, in less iterations. Noting, of course, that one PGVU iteration involves more computational work.

The conclusion is that when accuracy is important, PGVU is the right approach. If accuracy is not a concern, simpler techniques would be preferred.

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., [19]. 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, let the matrices $\{U_k\}$ and $\{Q_k\}$ be bounded, and the prox-parameters satisfy $\mu_{\max} \geq \mu_k \geq \beta$ for all k sufficiently large.*

Then every accumulation point $\bar{x}$ of the sequence $\{x^k\}$ generated by Algorithm 3.1 is a minimizer of f.

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 $\{x^k\}_{k'\in\mathcal{K}} \to \bar{x}$. We claim that $\{p^k - x^k\}_{k'\in\mathcal{K}} \to 0$ and $\{g^k_{k'\in\mathcal{K}}\} \to \bar{g} = 0$.

Let U, Q and $\bar{\mu}$ denote limit points of the corresponding subsequences of matrices and prox-parameters. 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, combining [41, Thms. 7.11 and 12.35], the models epi-converge to $m(\cdot; \bar{x})$, and the proximal mappings $\text{prox}_{m(\cdot;x^k),\mu_k}(\cdot)$ converge to $\text{prox}_{m(\cdot;\bar{x}),\bar{\mu}}(\cdot)$ uniformly on bounded sets. By epi-convergence of the models and [41, Theorem 7.14], the proximal point sequence converges continuously: $p^k \to \text{prox}_{m(\cdot;\bar{x}),\bar{\mu}}(\bar{x}) =: \bar{p}$. Passing onto the limit in (3.6) with g^k from (3.3),

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

Since the Hessian $\nabla^2 q(\cdot)$ is positive semidefinite by assumption, by the mean-value theorem $\nabla q(\bar{x}) - \nabla q(\bar{p}) = \nabla^2 q(\bar{y})(\bar{x} - \bar{p})$ for some intermediate point $\bar{y}$. Then, after some direct algebraic manipulations, we obtain that

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

Because $\mu_k \geq \beta$ for large k, 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 (3.3) and the continuity of ∇q, readily give $g^k \to 0$. By (3.4), this means that $\bar{p} = \bar{x}$ is a minimizer of f, as stated. ∎

Thanks to the fact that $g^k \to \bar{g} = 0$, we are now in position to 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 $\bar{g} = 0 \in \text{ri}\,\partial f(\bar{x})$. Then, for $k \in \mathcal{K}$ sufficiently large and (u^k, v^k) defined in Theorem 3.1, the following holds.*

(i) *$g^k_u \in \partial_u L_U(u^k, g^k_v)$;*

(ii) *If, in addition, $\mu_k > \beta$, 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), $g^k_v \in G_v(\bar{x}) \subset V^\dagger\,\text{ri}\,\partial f(\bar{x})$, 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 final result recall that, since $\|Uu^k\| \leq \|p^k - \bar{x}\|$ and $\|Vg_v^k\| \leq \|g^k\|$, by Theorem 3.1(iii) we obtain that

$$\|Uu^k + Vg_v^k\| \leq \frac{3\mu(\mu+\beta)}{\mu-\beta}\|x^k - \bar{x}\|\,. \tag{3.7}$$

The $\mathcal{U}$-Lagrangian Definition 2.3 combined with the expression (2.7) 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 inequality (2.3) written with $w = v^k$ and $\bar{g}_u = 0$ gives the lower bound $f(\bar{x}) + \eta\|Vv^k\|$ for the left-hand side in the relations above. Hence, $\eta\|Vv^k\| \leq o\Big(Uu^k + Vg_v^k\Big)$, and (3.7) 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 [17] is needed.

4.1. A partial second-order object for the bivariate $\mathcal{U}$-Lagrangian. The $\mathcal{U}$-Hessian introduced in [22, § 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 $\bar{g} \in \mathrm{ri}\,\partial f(\bar{x})$, 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 the $\mathcal{U}$-Hessian associated with single-variable $\mathcal{U}$-Lagrangian, with $\bar{g}$ being a parameter:*

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

Furthermore, if $0 \in \mathrm{ri}\,\partial f(\bar{x})$ and f has a partial $\mathcal{U}$-Hessian at $\bar{x}$ associated with $\bar{g}_v = 0$, then the following holds 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) *For small $d \in \mathbb{R}^n$, $f(\bar{x} + d) = f(\bar{x}) + \frac{1}{2}\left\langle \overline{H}d_u, d_u\right\rangle + o(\|d_u^2\|)$. As a result, if $\overline{H}$ is positive definite,*

$$\exists c > 0 : d \in \mathbb{R}^n \textit{ small } \implies f(\bar{x} + d) \geq f(\bar{x}) + \frac{c}{2}\|d\|^2\,.$$

In particular, $\bar{x}$ is the unique solution 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 $g_u^k \in \partial_u L_{\mathcal{U}}(u^k, g_v^k)$ by Corollary 3.4(i). To show (ii), note that if $0 \in \mathrm{ri}\,\partial f(\bar{x})$ 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, $\overline{H} = H^{\bar{g}=0}(\bar{x})$. Thus, f has a $\mathcal{U}$-Hessian at $\bar{x}$, and [22, Theorem 3.9] gives the second-order expansion. If $\overline{H}$ is positive definite, then [21, Corollary 1] gives the lower bound for all small d. Uniqueness of $\bar{x}$, called a strong minimizer in [35], follows from f's convexity. □

The lower bound for f that holds when the partial $\mathcal{U}$-Hessian is positive definite is called *local subdifferential convexity* in [9, Theorem 6.2]. The property is shown to be equivalent to both tilt stability and to strong metric regularity of the subdifferential at $\bar{x}$, a stable strong minimizer for (1.1); see also [8].

4.2. The partial $\mathcal{U}$-Hessian of partly smooth functions. When the function in (1.1) is partly smooth [24, Definition 2.7], it is shown in [3] that Riemannian Newton-like methods can be combined with proximal gradient steps to boost convergence speed. Under the assumption of partial smoothness, we now study conditions for the existence of a partial $\mathcal{U}$-Hessian.

Definition 4.3. *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) in a neighbourhood X of x, the restriction of f to $\mathcal{M}$, $f\big|_{\mathcal{M}\cap X}$, 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 $\mathcal{M}$.*

The function F in Example 2.2 has the partial $\mathcal{U}$-Hessian $\overline{H} = a > 0$, corresponding to the bivariate $\mathcal{U}$-Lagrangian $L(u, \bar g_v) = (1-|\bar g_v|)\frac{a^2}{2}$. However, F is not partly smooth, because near $\bar x = 0$ there are two distinct activity manifolds $\mathcal{M}$. These are the two fast tracks displayed in Figure 2.1, generated by the different $\mathcal{V}$-minimizers that emanate from taking a positive or a negative $\bar g_v$ in Definition 2.3, i.e., $W(u, \bar g_v) = \left\{\frac{a}{2}\, sign(\bar g_v)u^2\right\}$. By contrast, the simple modification of Example 2.2 given by $\tilde F(u, v) = \frac{a}{2}u^2 + \max\left\{0, v - \frac{a}{2}u^2\right\}$, is partly smooth at $\bar x$. The fundamental difference is that the $\mathcal{V}$-minimizers of $\tilde F$ are $W(u, \bar g_v) = \left\{\frac{a}{2}u^2\right\}$, *the same* for all $\bar g_v$. Now the $\mathcal{U}$-Lagrangian $L_U(u, \bar g_v) = (1 - \bar g_v)\frac{a}{2}u^2$ provides the single activity manifold $\mathcal{M} := \left\{(v, u) \colon v = \frac{a}{2}u^2\right\}$, where the partial $\mathcal{U}$-Hessian is again $\overline{H} = a > 0$.

Our next result states a similar relation in the general setting, by connecting partial smoothness and $\mathcal{VU}$-analysis, thanks to [24, Theorem 6.1]. We associate the manifold of partial smoothness with a special $\mathcal{V}$-minimizer that is a C^2-function and *is the same* for all interior subgradients (the same function was considered in [34, Theorem 6] to characterize fast tracks for prox-regular functions).

Theorem 4.4 (Special $\mathcal{V}$-minimizers from partial smoothness). *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*

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

As a result, there exist a neighborhood $X \subset \mathbb{R}^n$ of $\bar x$, a neighborhood $Y \subset \mathbb{R}^m$ of 0 such that

$$L_U(u, g_v) = f\big|_{\mathcal{M}\cap X}(u) - \left\langle g_v, V^\top V v_{\partial f}(u)\right\rangle\,. \tag{4.2}$$

$$\mathcal{M} \cap X = \left\{\bar x + Uu + V v_{\partial f}(u) \colon u \in Y\right\}\,, \tag{4.3}$$

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

Proof. From the property (ii) in Definition 4.3, 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 [24, Theorem 6.1], there exist a neighborhood $X \subset \mathbb{R}^n$ of $\bar x$, a neighborhood $Y \subset \mathbb{R}^m$ of 0, and a function $v_{\partial f} \colon \mathbb{R}^m \to \mathbb{R}^{n-m}$ such that for all $u \in Y$,

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

From the last relation, the restriction in Definition 4.3(i) has the expression $f\big|_{\mathcal{M}\cap X} = f(\bar x + Uu + V v_{\partial f}(u))$. Again by [24, 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. The identity (4.2) follows 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. ∎

When specializing Theorem 4.4 to the setting (3.1), it is 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.

COROLLARY 4.5 (explicit partial $\mathcal{U}$-Hessian). *When, under the assumptions in Theorem 4.4,* $0 \in \operatorname{ri} \partial f(\bar{x})$*, a partial* $\mathcal{U}$*-Hessian of* f *at* $\bar{x}$ *associated with* $\bar{g} = 0$ *is given by the Hessian restriction:* $\overline{H} = \nabla^2_{uu} f\big|_{\mathcal{M}\cap X}(0)$.

If, in addition, in (3.1) *the nonsmooth function* h *is finite-valued and polyhedral, then* $\overline{H} = U^\top \nabla^2 q(\bar{x}) U$.

Proof. In (4.2), the linear term defines the bivariate function on $Y \times V^\dagger \operatorname{ri} \partial f(\bar{x})$ given by

$$P(u, g_v) := \left\langle g_v, V^\intercal V v(u) \right\rangle = g^\top V v_{\partial f}(u)\,,$$

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

$$\mathcal{J}_u P(u, g_v) = \mathcal{J} v_{\partial f}(u)^\intercal V^\intercal g = \mathcal{J} v_{\partial f}(u)^\intercal V^\intercal V g_v\,. \tag{4.4}$$

Since $W(u, g_v) \ni v_{\partial f}(u) = O(\|u\|^2)$ by Theorem 4.4, combining Theorem 2.5(iii) and the fact that

$$\mathcal{J} v(u) = \mathcal{J} v(0) + O(u) = O(u)\,,$$

gives the following expansion for the gradient of the $\mathcal{U}$-Lagrangian from (4.2):

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

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 (4.4) and (4.2), the cross-derivative has the form

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

Recalling that $\nabla_u L_U(0, g_v) = \bar{g}_u$, gives the desired expression for the partial $\mathcal{U}$-Hessian. In view of (4.4),

$$\mathcal{J}_u P(u, g_v) = o(u, g_v)\,. \tag{4.5}$$

In particular, $\mathcal{J}_u P(0, g_v) = 0$. Hence, by (4.2) and the fact that $f\big|_{\mathcal{M}\cap X}$ is a composite function of u, we obtain that $\nabla_u f\big|_{\mathcal{M}\cap X}(0) = \nabla_u L_U(0, g_v) + \nabla_u P(0, g_v) = \bar{g}_u$. Then the smoothness of $f\big|_{\mathcal{M}\cap X}$ yields $\nabla_u f\big|_{\mathcal{M}\cap X}(u) = \bar{g}_u + \nabla^2_{uu} f\big|_{\mathcal{M}\cap X}(0) u + o(u)$. Consequently, by (4.2) and (4.5),

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

As $o(u) - o(u, g_v) = o(u, g_v)$, we see from Prop. 4.2 that $\nabla^2_{uu} f\big|_{\mathcal{M}\cap X}(0)$ 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

$$h(x) = \max_{i \in I} \left\{ \left\langle a^i, x \right\rangle + b^i \right\} \quad \text{for some finite index set } I \neq \emptyset.$$

Then, for the "active" index set $I(x) = \left\{ i \in I \colon \left\langle a^i, x \right\rangle + b^i = f(x) \right\}$,

$$\partial h(x) = \left\{ \sum_{i \in I(x)} \alpha_i a^i : \sum_{i \in I(x)} \alpha_i = 1,\ \alpha_i \geq 0\ (i \in I(x)). \right\} \tag{4.6}$$

The function h is partly smooth at any $\bar{x}$ relative to $\mathcal{M}_{\bar{x}} := \{x \in \mathbb{R}^n : I(x) = I(\bar{x})\}$, [24, Example 3.4]. From [24, 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 [12, 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 $\bar{x} + Uu + V\bar{v} \in \mathcal{M} \cap X$, it must hold that $\bar{v} = 0$; so $\mathcal{M} \cap X = \{\bar{x} + Uu : u \in Y\}$, for Y a neigbourhood of $0 \in \mathbb{R}^m$. To show the claim, consider $i \in I(\bar{x})$ and note that, because $\bar{x} + Uu + V\bar{v} \in \mathcal{M}_{\bar{x}}$, it must be that $I(\bar{x} + Uu + V\bar{v}) = I(\bar{x})$, that is $h(\bar{x} + Uu + V\bar{v}) = \left\langle a^i, \bar{x} + Uu + V\bar{v}\right\rangle = \left\langle a^i, \bar{x}\right\rangle + b^i + \left\langle a^i, Uu + V\bar{v}\right\rangle = h(\bar{x}) + \left\langle a^i, Uu + V\bar{v}\right\rangle$. Therefore,

$$h(\bar{x} + Uu + V\bar{v}) - h(\bar{x}) = \left\langle a^i, Uu + V\bar{v}\right\rangle . \tag{4.7}$$

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, $\left\langle a^i + \nabla q(\bar{x}), Uu\right\rangle = 0$ and $\left\langle a^i, Uu\right\rangle = -\langle \nabla q(\bar{x}), Uu\rangle$. And (4.7) yields $\left\langle V\bar{v}, a^i\right\rangle = h(\bar{x}+Uu+V\bar{v}) - h(\bar{x}) + \langle \nabla q(\bar{x}), Uu\rangle$. 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, $\left\langle V\bar{v}, a^i + \nabla q(\bar{x})\right\rangle = 0$. Because $\mathcal{V} = \text{span}(\partial h(\bar{x}) + \nabla q(\bar{x})) = \text{span}\left\{a^i + \nabla q(\bar{x}) : i \in I(\bar{x})\right\}$, 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 $u \in Y$. Since $\mathcal{M} \cap X = \{\bar{x} + Uu : u \in Y\} = \mathcal{M}_{\bar{x}} \cap X$ and $I(\bar{x} + Uu) = I(\bar{x})$, from the characterization of $\partial h(x)$ in (4.6), 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 (4.3) in Thm. 4.4 we can take X and Y sufficently small such that $v_{\partial f}(u) \equiv 0$. Consequently, the restriction of f on $\mathcal{M} \cap X$ 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 $\nabla^2_{uu} f(\bar{x} + Uu) = U^\top \nabla^2 q(\bar{x} + Uu)U$. ∎

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 has been shown to be equivalent objects, for convex functions in [12] and for prox-regular functions in [26]. 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.

4.3. 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., [19]), is required:

$$(U_k Q_k U_k^\top - U\overline{W}U^\top)g^k = o(Ug_u^k)\,, \text{ where } \overline{W}^{-1} = \overline{H} := H(\bar{x}; 0)\,. \tag{4.8}$$

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

Theorem 4.6 (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* (4.8) *holds. Under the assumptions in Theorem 3.3, let $\bar{x}$ be an accumulation point of $\left\{x^k\right\}$ such that $\bar{g} = 0 \in \text{ri}\,\partial f(\bar{x})$. Then $\bar{x}$ is the unique minimizer of f, and both $\left\{x^k\right\}$ and $\left\{p^k\right\}$ converge to $\bar{x}$. Furthermore, if $\mu_k > \beta$, then $\|x^{k+1} - \bar{x}\| = o(x^k - \bar{x})$, i.e., the iterates generated by Algorithm 3.1 converge superlinearly.*

Proof. Because $0 \in \text{ri}\,\partial f(\bar{x})$ 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. Recall from Theorem 3.3 that every accumulation point of the sequence $\left\{x^k\right\}$ generated by Algorithm 3.1 is a minimizer of f. Consequently, $\bar{x}$ is the unique accumulation point of $\left\{x^k\right\}$, with both $\left\{x^k\right\}$ and $\left\{p^k\right\}$ converging to $\bar{x}$, as stated.

Next, using (3.6), adding $\pm U\overline{W}U^\intercal g^k$, and recalling the definition of (u^k, v^k) in Theorem 3.1:

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

The third term above is of the order $o(x^k - \bar{x})$, by item (ii) in Corollary3.4. The same holds for the first term, as $\Big(U\overline{W}U^\intercal - U_k Q_k U_k^\intercal\Big) g^k = o(g^k)$ because of (4.8), and $g^k = O(x^k - \bar{x})$, by Theorem 3.1(ii). To conclude the proof, it remains to show that $T_2 := \Big(Uu^k - U\overline{W}U^\intercal g^k\Big) = o(x^k - \bar{x})$. Since $T_2 = U\Big(u^k - \overline{W}U^\intercal g^k\Big) = U\Big(u^k - \overline{W}g_u^k\Big)$, after multiplying on the left by $\overline{H}U^\intercal$, we see that $\overline{H}U^\intercal T_2 = \overline{H}u^k - g_u^k$. Lemma 4.2(i) then ensures that $\overline{H}U^\intercal T = o(u^k)$. The result follows, because $u^k = O(x^k - \bar{x})$, by Theorem 3.1(ii). ∎

The following two examples of polyhedral functions that are partly smooth relative to an affine or linear set from [43, Sec. 3.1], are common illustrations in regularized regression problems. The corresponding $\mathcal{U}$-subspaces have then an explicit expression. If

$$h(x) = \|x\|_1 \text{ then } \mathcal{U}(p) = \operatorname{lin}\left\{e^j : j \in J(p)\right\} \text{ for } J(p) = \left\{i \le n : |p_j| > 0\right\} \,. \tag{4.9}$$

Likewise, if $h(x) = \|x\|_\infty$, then $\mathcal{U}(p) = \{x \colon x_J = k \operatorname{sign}(p_J),\ k \in \mathbb{R}\}$ for the activity index set $J = J(p) = \{i \le n : |p_i| = \|p\|_\infty\}$.

For our numerical validation, we now consider problems having h as in (4.9).

5. Application to ℓ_1-regularized minimization.

We now apply the PGVU method (Algorithm 3.1), as its illustration, to solve ℓ_1-regularized problems. So,

$$\text{in (3.1) the nonsmooth function is } h(x) = \lambda\|x\|_1 \,, \text{ for a positive parameter } \lambda, \tag{5.1}$$

and the proximal points for h are easy to compute. Accordingly, the $\mathcal{V}$-step in Procedure 1 computes

$$p^k = \max\{0, w^k - \frac{\lambda}{\mu_k}\} - \max\{0, -w^k - \frac{\lambda}{\mu_k}\}\,, \text{ for } w^k = x^k - \frac{1}{\mu_k}\nabla q(x^k)\,. \tag{5.2}$$

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. Defining Q_k and U_k.

For PGVU global convergence in Theorem 3.3, the matrices only need to be bounded. The superlinear rate in Theorem 4.6 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, [43, 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 (4.9). As the same holds for f when h is as in (5.1), by Theorem 4.4, 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 (4.9) might appear straightfoward. Nevertheless, given $\varepsilon \geq 0$, we consider instead the smaller subspaces

$$\mathcal{U}_\varepsilon(x) = \operatorname{span}\left\{e^j : j \in J_\varepsilon(x)\right\} \text{ for } J_\varepsilon(x) := \left\{i \leq n : |x_i| > \frac{\varepsilon}{2}\right\}. \tag{5.3}$$

This $\mathcal{V}_\varepsilon\mathcal{U}_\varepsilon$-decomposition was introduced in [27] to deal with the lack of continuity of the subdifferential as a multifunction. Unlike the $\mathcal{VU}$-decomposition, the following important continuity property, [27, equation (5.13)], holds for the ε-counterpart:

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

Thanks to this property, 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 $U_k \to U$, as needed to satisfy the Dennis-Moré condition (4.8) in Theorem 4.6.

The choice of the parameter ε should ensure that it is driven to zero by the algorithmic process. We discuss the impact of such choices on our previous example function.

Example 5.1 (Choosing ε). For F from Example 2.2, we run PGVU considering for the $\mathcal{U}$-step three different (natural) choices for ε in (5.3):

$$\varepsilon_0^k = 0, \quad \varepsilon_1^k = f(x^k) - f(p^k) - \mu_k\|p^k - x^k\|^2, \text{ and } \varepsilon_2^k = \mu_k\|p^k - x^k\|^2.$$

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 $\mu_k(x^k - p^k) \in \partial_{\varepsilon_k} f(x^k)$, [18, 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 $x^k - p^k \to 0$).

For the two initial values of μ_0 and x^0 in Table 3.1, we run each PGVU variant, respectively labeled PGVU-0,1,2, in a reference to the value of ε^k employed in (5.3). The comparison of the number of iterations and digits of accuracy indicate ε_2^k as the best option, as illustrated by the trajectories in Figure 5.1, generated with ε_0^k on the left plot and with ε_2^k on the right.

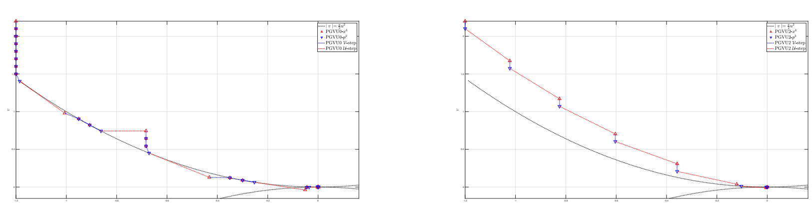

Fig. 5.1: Trajectories of PGVU-0 and PGVU-2 iterations, when minimizing F from $x^0 = (-1.2, 2.2)^\top$, starting with $\mu_0 = 10$. Both variants stopped having reached more than 20 digits of accuracy, but PGVU-2 needed much less iterations. When the $\mathcal{U}_{\varepsilon^k}(p^k)$ subspace is determined with $\varepsilon^k = 0$, as on the left plot, only 4 $\mathcal{U}$-steps are done, and PGVU-0 needed 45 iterations to trigger the stopping test. For the trajectory on the right, by contrast, that was generated with ε_2^k, it sufficed to perform 9 iterations that involved 8 $\mathcal{U}$-steps.

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

Data: f from (5.1), starting point x^0, prox-parameter μ_0, and a stopping tolerance TOL. Set $k = 0$.

repeat

Apply Procedure 1 with p^k defined in (5.2).

Shift the gradient $g^k = \mu_k(x^k - p^k) + \nabla q(p^k) - \nabla q(x^k)$.

Compute $U_k = \left[e^j : |x^k_j| > \frac{\varepsilon_k}{2}, 1 \le j \le n\right]$ for $\varepsilon_k := \max\Big(\text{TOL}, \mu_k\|p^k - x^k\|\Big)$.

Obtain the direction $d^k = -W^k U_k^\intercal g^k$ for $W^k \approx (U_k^\intercal \nabla^2 q(x^k) U_k)^\dagger$.

Update $x^{k+1} = p^k + U_k d^k$, set $k = k + 1$

until $\|g^k\| \le$ *TOL*;

5.2. Algorithm statement and numerical experiments.

We are now ready to introduce the algorithm.

Being a special instance of PGVU, Algorithm 5.1 has global and superlinear convergence if the conditions in Theorems 3.3 and 4.6 are satisfied. With our definitions, such is the case if $\mu_k > \beta$ and, for $\bar{x}$ an accumulation point of $\{x^k\}$ generated by Algorithm 5.1, it holds that $0 \in \mathrm{ri}\,\partial f(\bar{x})$ and $\overline{H} = U^\intercal \nabla^2 q(\bar{x}) U$ is positive definite.

5.2.1. Test functions and parameters.

The performance of Algorithm 5.1 is assessed on regularized least-square problems:

$$\text{in (1.1)}, \quad f(x) = \frac{1}{2}\|Ax - b\|^2 + \lambda\|x\|_1 \text{ for } A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, \text{ and } \lambda = 0.1\|A^\top b\|_\infty .$$

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, refered below as QUAD, have random dimensions $m \in [10, 1000]$ and $n \in [0.1m, 2m]$. 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 support vector machine (SVM) problems are all scaled to [-1,1] or [0,1].

In Procedure 1, we set $x^0 = 0$, $\sigma = 10^{-4}$ and $\mu_0 = \frac{\nabla f(x^o)^\top \nabla f(x^0)}{2\max\{1, |f(x^0)|\}}$. The maximal number of iterations was set to 100, and the stopping tolerance is TOL $= 10^{-6}$. After the $\mathcal{U}$-step, the prox-parameter is updated according to [23], 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 $y^k := \nabla f(x^{k+1}) - \nabla f(x^k)$ and $s^k := x^{k+1} - x^k$. In the $\mathcal{U}$-step, to compute d^k, since $\nabla^2 q(x) = A^\intercal A$ for all x, we let Id denote the identity matrix of order n and define

$$W^k = \begin{cases} (A^\intercal A + \text{TOL}\,\mathrm{Id})^{-1} & \text{for QUAD,} \\ diag(A^\intercal A + \text{TOL}\,\mathrm{Id})^{-1} & \text{for SVM.} \end{cases}$$

5.2.2. Solvers and figures with evaluation measures.

The benchmark compares MATLAB implementations of the following solvers:

1. PGVU-2, as in Algorithm 5.1.
2. SpaRSA 2.0, the sparse reconstruction by separable approximation[44] [2].

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

[2] http://www.lx.it.pt/~mtf/SpaRSA/

3. FISTA, the fast iterative shrinkage-thresholding algorithm [4] [3].
4. ADMM, the alternating direction method of multipliers[6], with parameters $(\rho, \alpha) = (0.0001, 1.5)$ [4].
5. qNVU, the $\mathcal{VU}$-algorithm from [35].

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 SVM 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

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

Solvers having achieved at least `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 SVM instances.

Table 5.1: Successful runs for each solver

		PGVU-2	SpaRSA	FISTA	ADMM	qNVU
QUAD	$\textsc{acc} \geq 2$	5000	4998	5000	516	5000
mid-size	$\textsc{acc} \geq 4$	4999	4933	5000	5	5000
(5000 runs)	$\textsc{acc} \geq 6$	4998	4493	5000	1	5000
SVM	$\textsc{acc} \geq 2$	86	91	85	8	86
large-size	$\textsc{acc} \geq 4$	81	51	79	2	86
(95 runs)	$\textsc{acc} \geq 6$	74	26	70	0	86

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

Regarding computing times, for low accuracy (`cutoff`=2), SpaRSA is always the fastest solver. The profile in Figure 5.2 compares computing times among the successful runs, for the value of `cutoff`=4. In the right plot in Figure 5.2, qNVU exhibits a slower performance. This is because at each iteration the qNVU method [35] solves two quadratic programming problems, a computationally expensive calculation for the large SVM instances. For the mid-size instances, SpaRSA remains the fastest solver, but not for the SVM problems. A solver-to-solver comparison clarifies this situation with four plots reported in Figure 5.3, comparing PGVU-2 to SpaRSA and FISTA, when SVM problems were solved with at least 2 or 4 digits.

On the top left plot in Figure 5.3, we notice that to reach 2 digits of accuracy, SpaRSA is faster than PGVU-2, which is in turn faster than FISTA (right top plot). To get 4 digits, the bottom plots show that PGVU-2 always wins, reaching the accuracy level faster than both SpaRSA and FISTA.

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 $\mathcal{V}$-

[3] https://github.com/tiepvupsu/FISTA
[4] https://web.stanford.edu/~boyd/papers/admm/lasso/lasso.html

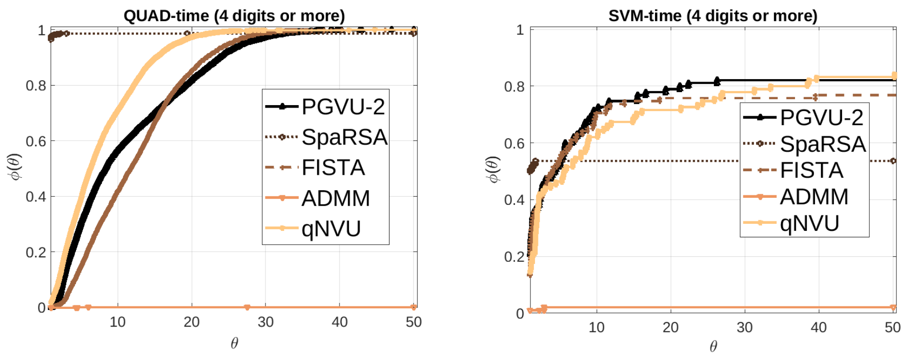

Fig. 5.2: Profiles for computing time for the successful runs for all solvers.

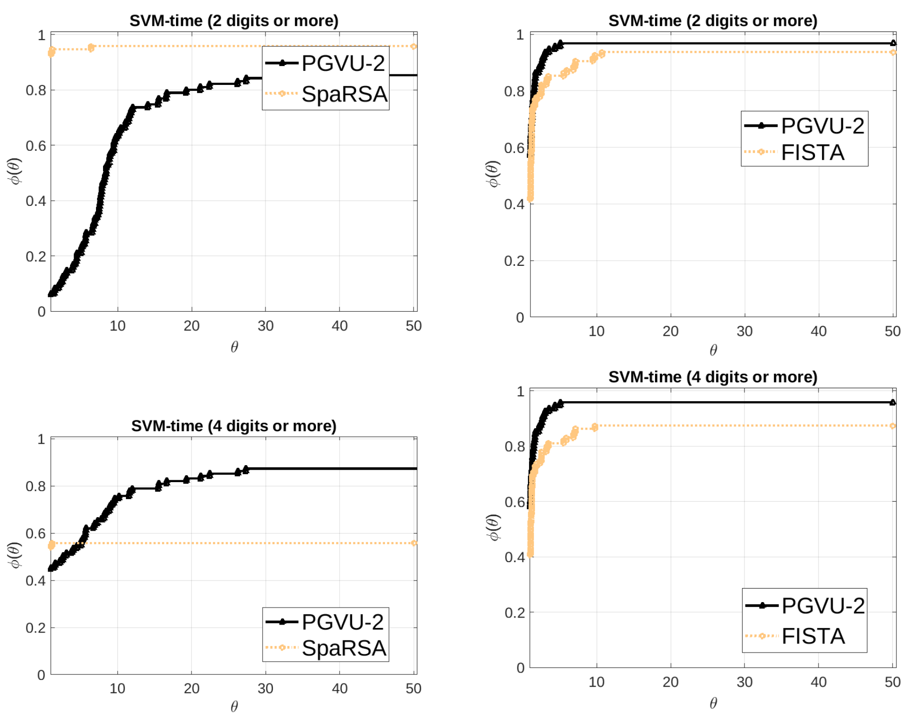

Fig. 5.3: Solver-to-solver time comparisons over SVM instances for different accuracy.

minimizers are *tangent* to the $\mathcal{U}$-subspace, an important property leading to superlinear convergence of the Proximal Gradient $\mathcal{VU}$-method, under natural assumptions.

For PDG-structured functions (including ℓ_1-regularization), the Hessian of the single-variable $\mathcal{U}$-Lagrangian exists along a certain fast-track [33, 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. In particular, we proved that partly smooth functions satisfying $0 \in \operatorname{ri} \partial f(\bar{x})$ always have a partial $\mathcal{U}$-Hessian at $\bar{x}$.

We introduced the Proximal Gradient $\mathcal{VU}$ method, applicable to various structured convex optimization problems, with superlinear convergence despite the nonsmoothness. Numerical experiments verify that the method is particularly useful when high accuracy is desired.

Originally defined for convex functions, the $\mathcal{VU}$-theory has been generalized to the nonconvex setting [33, 16, 26]. In [26] a localized version of $\mathcal{U}$-Lagrangian and the notion of fast track are defined for a type of nonconvex functions called prox-regular functions [41], and the correspondence between an active manifold of a partly smooth function and a fast track is given. In [10], 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.

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 Journal on Optimization, 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 Journal on Optimization, 33 (2023), pp. 2222–2247, https://doi.org/10.1137/22M1505827, https://doi.org/10.1137/22M1505827, https://arxiv.org/abs/https://doi.org/10.1137/22M1505827.

[3] G. Bareilles, F. Iutzeler, and J. Malick, *Newton acceleration on manifolds identified by proximal gradient methods*, Mathematical Programming, 200 (2023), pp. 37–70, https://doi.org/10.1007/s10107-022-01873-w, 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 Transactions on Image Processing, 18 (2009), pp. 2419–2434.

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

[6] 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), p. 1–122, https://doi.org/10.1561/2200000016, https://doi.org/10.1561/2200000016.

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

[8] D. Drusvyatskiy and A. S. Lewis, *Tilt stability, uniform quadratic growth, and strong metric regularity of the subdifferential*, SIAM Journal on Optimization, 23 (2013), pp. 256–267, https://doi.org/10.1137/120876551.

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

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

[11] W. Hare, *Numerical analysis of VU-decomposition, U-gradient, and U-hessian approximations*, SIAM Journal on Optimization, 24 (2014), pp. 1890–1913.

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

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

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

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

[16] W. L. Hare and R. Poliquin, *The quadratic sub-Lagrangian of a prox-regular function*, Nonlinear Analysis: Theory, Methods & Applications, 47 (2001-08), pp. 1117–1128, https://doi.org/10.1016/s0362-546x(01)00251-6.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The U-Lagrangian of a convex function*, Transactions of the American Mathematical Society, 352 (2000), pp. 711–729.

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

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

[25] J. Liang, J. Fadili, and G. Peyré, *Activity identification and local linear convergence of forward–backward-type methods*, SIAM Journal on Optimization, 27 (2017), pp. 408–437, https://doi.org/10.1137/16M106340X, https://doi.org/10.1137/16M106340X, https://arxiv.org/abs/https://doi.org/10.1137/16M106340X.

[26] S. Liu, A. Eberhard, and Y. Luo, *The U-Lagrangian, fast track, and partial smoothness of a prox-regular function*, Set-Valued and Variational Analysis, 28 (2020), pp. 369–394, https://doi.org/10.1007/s11228-019-00518-z.

[27] 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.

[28] 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.

[29] R. Mifflin and C. Sagastizábal, *VU-decomposition derivatives for convex max-functions*, Ill-Posed Variational Problems and Regularization Techniques, 477 (1999), pp. 167–186.

[30] R. Mifflin and C. Sagastizábal, *On VU-theory for functions with primal-dual gradient structure*, SIAM Journal on Optimization, 11 (2000), pp. 547–571.

[31] R. Mifflin and C. Sagastizábal, *Proximal points are on the fast track*, Journal of Convex Analysis, 9 (2002), pp. 563–580.

[32] R. Mifflin and C. Sagastizábal, *Primal-dual gradient structured functions: second-order results; links to epi-derivatives and partly smooth functions*, SIAM Journal on Optimization, 13 (2003), pp. 1174–1194.

[33] R. Mifflin and C. Sagastizábal, *UV-smoothness and proximal point results for some nonconvex functions*, Optimization Methods and Software, 19 (2004), pp. 463–478, https://doi.org/10.1080/10556780410001704902.

[34] R. Mifflin and C. Sagastizábal, *Relating U-Lagrangians to second order epiderivatives and proximal tracks*, Journal of Convex Analysis, 12 (2005), pp. 81–93.

[35] R. Mifflin and C. Sagastizábal, *A VU-algorithm for convex minimization*, Mathematical Programming, 104 (2005), pp. 583–608.

[36] S. A. Miller and J. Malick, *Newton methods for nonsmooth convex minimization: connections among U-Lagrangian, Riemannian Newton and SQP methods*, Mathematical Programming, 104 (2005), pp. 609–633.

[37] Y. Nesterov, *Smooth Convex Optimization*, Springer US, 2004, pp. 51–110, https://doi.org/10.1007/978-1-4419-8853-9_2.

[38] J. Ortega and W. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[39] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Mathematical Programming, Series B, 89 (2000), pp. 1–33, https://doi.org/10.1007/s101070000166, https://www.scopus.com/inward/record.uri?eid=2-s2.0-0002928322&partnerID=40&md5=4ff600667e238f59008a3320f8e82dae.

[40] 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 and Variational Analysis, 32 (2024), https://doi.org/10.1007/s11228-024-00718-2, https://doi.org/10.1007/s11228-024-00718-2.

[41] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, vol. 317 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 1998.

[42] 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.
[43] S. Vaiter, C. Deledalle, J. Fadili, G. Peyré, and C. Dossal, *The degrees of freedom of partly smooth regularizers*, Annals of the Institute of Statistical Mathematics, 69 (2017), pp. 791–832, https://doi.org/10.1007/s10463-016-0563-z.
[44] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, *Sparse reconstruction by separable approximation*, IEEE Transactions on Signal Processing, 57 (2009), pp. 2479–2493, https://doi.org/10.1109/TSP.2009.2016892.