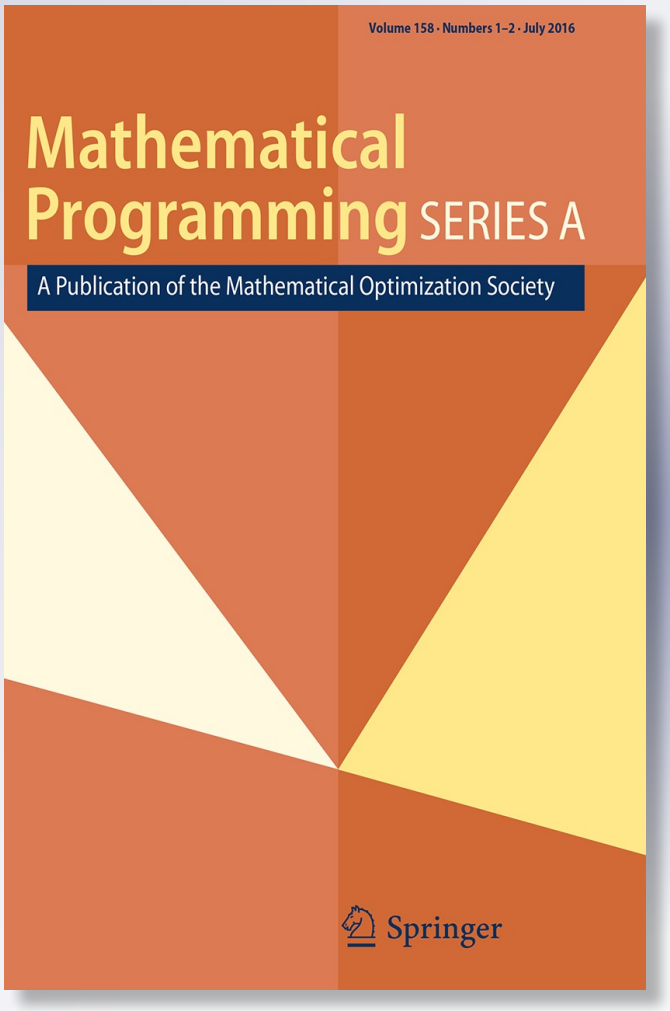
A proximal method for composite minimization

A. S. Lewis & S. J. Wright

Mathematical Programming
A Publication of the Mathematical Optimization Society

ISSN 0025-5610
Volume 158
Combined 1-2

Math. Program. (2016) 158:501-546
DOI 10.1007/s10107-015-0943-9



Springer

Math. Program., Ser. A (2016) 158:501–546
DOI 10.1007/s10107-015-0943-9

FULL LENGTH PAPER

A proximal method for composite minimization

A. S. Lewis[1] · **S. J. Wright**[2]

Received: 2 December 2008 / Accepted: 13 August 2015 / Published online: 29 August 2015

Abstract We consider minimization of functions that are compositions of convex or prox-regular functions (possibly extended-valued) with smooth vector functions. A wide variety of important optimization problems fall into this framework. We describe an algorithmic framework based on a subproblem constructed from a linearized approximation to the objective and a regularization term. Properties of local solutions of this subproblem underlie both a global convergence result and an identification property of the active manifold containing the solution of the original problem. Preliminary computational results on both convex and nonconvex examples are promising.

Keywords Prox-regular functions · Polyhedral convex functions · Sparse optimization · Global convergence · Active constraint identification

Mathematics Subject Classfication 49M37 · 90C30

A.S. Lewis's research supported in part by NSF Award DMS-1208338.

S.J. Wright's research supported in part by NSF Awards DMS-1216318 and IIS-1447449, ONR Award N00014-13-1-0129, AFOSR Award FA9550-13-1-0138, and Subcontract 3F-30222 from Argonne National Laboratory.

✉ A. S. Lewis
adrian.lewis@cornell.edu

S. J. Wright
swright@cs.wisc.edu

[1] School of ORIE, Cornell University, Ithaca, NY 14853, USA

[2] Computer Sciences Department, University of Wisconsin, 1210 W. Dayton Street, Madison, WI 53706, USA

1 Introduction

We consider optimization problems of the form

$$\min_x \; h\big(c(x)\big), \tag{1.1}$$

where the inner function $c : \Re^n \to \Re^m$ is smooth. The outer function $h : \Re^m \to [-\infty, +\infty]$ may be nonsmooth, but is usually convex (even polyhedral), and sufficiently well-structured to allow us to solve, relatively easily, subproblems of the form

$$\min_d \; h\big(\Phi(d)\big) + \frac{\mu}{2}|d|^2, \tag{1.2}$$

for *affine* maps Φ and scalars $\mu > 0$ (where $|\cdot|$ denotes the Euclidean norm). We analyze a "proximal" method for the problem (1.1). In its simplest form, for a finite convex function h, the method is shown below as Algorithm 1.

Algorithm 1 ProxDescent for Finite Convex h

Define constants $\tau > 1$, $\sigma \in (0, 1)$, and $\mu_{\min} > 0$;
Choose $x_0 \in \Re^n$, $\mu_0 \geq \mu_{\min}$;
for $k = 0, 1, 2, \ldots$ **do**
  Set accept $\leftarrow$ false;
  while not accept **do**
    Find the minimizer d of the function $h\big(c(x) + \nabla c(x)d\big) + \frac{1}{2}\mu|d|^2$
    (terminating if $d = 0$);
    if $h\big(c(x)\big) - h\big(c(x+d)\big) \geq \sigma\left[h\big(c(x)\big) - h\big(c(x) + \nabla c(x)d\big)\right]$ **then**
      $\mu \leftarrow \max(\mu_{\min}, \mu/\tau)$;
      accept $\leftarrow$ true;
    else
      $\mu \leftarrow \tau\mu$;
    end if
  end while
  $x \leftarrow x + d$;
end for

The method repeatedly solves a *proximal linearized subproblem* of the form

$$\min_d \; h_{x,\mu}(d) := h\big(c(x) + \nabla c(x)d\big) + \frac{\mu}{2}|d|^2, \tag{1.3}$$

to find a trial step d, where the linear map $\nabla c(x) : \Re^n \to \Re^m$ is the derivative of the map c at x (representable by the $m \times n$ Jacobian matrix). In the algorithmic framework that we discuss later, where the function h is not restricted to being polyhedral or convex, the subproblem solution d is just a first approximation to the step. If h is sufficiently well-structured—an assumption we make concrete using "partial smoothness," a generalization of the idea of an active set in nonlinear programming—we may then be able to enhance the step, possibly with the use of higher-order derivative information.

Although many important problems of the form (1.1) involve finite convex functions h, we explore extensions to broader classes of functions h. Specifically, we allow that

- h may be extended-valued, allowing it to incorporate constraints that must be enforced;
- h is "prox-regular" rather than convex.

(We note in passing that our analysis extends easily to the case where the function c is defined only locally.) This broader framework requires additional technical overhead, but we point out throughout the simplifications that are available in the case of continuous convex h, and in particular polyhedral h.

1.1 Outline

In the next subsection, we discuss the building blocks from variational analysis that are used in later sections, focusing on key ideas that may be unfamiliar to many readers—"prox-regularity" and "partial smoothness"—and deferring more standard formal definitions to an "Appendix". In Sect. 2, we describe how a wide variety of examples can be posed as composite optimization problems of the form (1.1). We include problems from approximation, nonlinear programming, and regularized minimization, including nonconvex examples. Given its prevalence, historical importance, and significance in building intuition and terminology, we pay particular attention to the case in which the function h is finite and polyhedral. Next, in Sect. 3, we survey the extensive body of related work.

Section 4 contains our main theoretical tools, pertaining to the subproblem (1.3). We note that, since any local solution $\bar{x}$ for the problem (1.1) is *critical* (meaning that $0 \in \partial(h \circ c)(\bar{x})$, where ∂ denotes the subdifferential), a chain rule typically implies the existence of a vector $\bar{v}$ such that

$$\bar{v} \in \partial h(\bar{c}) \cap \text{Null}(\nabla c(\bar{x})^*), \tag{1.4}$$

where $\bar{c} := c(\bar{x})$ and * denotes the adjoint map. In examples, we can interpret the vector $\bar{v}$ as a Lagrange multiplier. We begin by showing that, when the current point x is near the critical point $\bar{x}$, the proximal linearized subproblem (1.3) has a local solution d of size $O(|x - \bar{x}|)$. To illuminate this idea, consider first a function h that is convex, lower semicontinuous, and never $-\infty$. Assuming that the vector $c(x) + \nabla c(x)d$ lies in the domain of h for some step $d \in \Re^n$, the subproblem (1.3) involves minimizing a strictly convex function with nonempty compact level sets, and thus has a unique solution $d = d(x)$. If we assume slightly more—that $c(x) + \nabla c(x)d$ lies in the relative interior of the domain of h for some d (as holds obviously if h is continuous at $c(x)$), a standard chain rule from convex analysis implies that $d = d(x)$ is the unique solution of the following inclusion:

$$\nabla c(x)^* v + \mu d = 0, \quad \text{for some } v \in \partial h\big(c(x) + \nabla c(x)d\big). \tag{1.5}$$

When h is prox-regular rather than convex, reasonable conditions ensure that the subproblem (1.3) still has a unique local solution close to zero, for μ sufficiently

large, also characterized by property (1.5). Then, by projecting the point $x + d$ onto the inverse image under the map c of the domain of the function h, we can obtain a step that reduces the objective.

The final part of Sect. 4 focuses on the common situation in which the function h is partly smooth at the point $\bar{c}$ relative to a certain manifold $\mathcal{M}$—a generalization of the surface defined by the active constraints in classical nonlinear programming. We give conditions guaranteeing that, when x is close to $\bar{x}$, the algorithm "identifies" $\mathcal{M}$, in the sense that the solution d of the subproblem (1.2) has $\Phi(d) \in \mathcal{M}$.

Section 5 presents the ProxDescent algorithm in full generality and proves a global convergence result. Finally, Sect. 6 describes some promising preliminary computational experiments, on convex and nonconvex regularized linear least-squares problems, together with a polyhedral penalty function arising from a nonlinear programming application.

1.2 Variational analysis tools

We begin with some important basic ideas and notation. We denote by $P_S(v)$ the usual Euclidean projection of a vector $v \in \Re^m$ onto a closed set $S \subset \Re^m$. The *distance* between x and the set S is

$$\operatorname{dist}(x, S) = \inf_{y \in S} |x - y|.$$

We use $B_\epsilon(x)$ to denote the closed Euclidean ball of radius ϵ around a point x.

We write $\bar{\Re}$ for the extended reals $[-\infty, +\infty]$, and consider a function $h : \Re^m \to \bar{\Re}$. The notion of the subdifferential of h at a point $\bar{c} \in \Re^m$, denoted $\partial h(\bar{c})$, provides a powerful unification of the classical gradient of a smooth function and the subdifferential from convex analysis. It is a set of generalized gradient vectors, coinciding exactly with the classical convex subdifferential [39] when h is lower semicontinuous and convex, and equaling $\{\nabla h(\bar{c})\}$ when h is $\mathcal{C}^1$ around $\bar{c}$. For formal definitions from variational analysis, we refer the reader to standard texts such as [40] and [35]. For ease of reading, we collect such definitions (along with brief discussions) in an "Appendix".

Since the notion of "prox-regularity" is crucial for our development, we quote a full definition here, from [40, Definition 13.27].

Definition 1.1 A function $h : \Re^m \to \bar{\Re}$ is *prox-regular at a point* $\bar{c} \in \Re^m$ *for a subgradient* $\bar{v} \in \partial h(\bar{c})$ if h is finite at $\bar{c}$, locally lower semicontinuous around $\bar{c}$, and there exists $\rho > 0$ such that

$$h(c') \geq h(c) + \langle v, c' - c \rangle - \frac{\rho}{2}|c' - c|^2$$

whenever points $c, c' \in \Re^m$ are near $\bar{c}$ with the value $h(c)$ near the value $h(\bar{c})$ and for every subgradient $v \in \partial h(c)$ near $\bar{v}$. Further, h is *prox-regular at* $\bar{c}$ if it is prox-regular at $\bar{c}$ for every $\bar{v} \in \partial h(\bar{c})$.

While this definition appears formidably technical in its generality, it holds commonly in practice. Prevalent examples include continuous functions h with the property that

the function $h + \kappa|\cdot|^2$ is convex for some constant κ. For further discussion, see the "Appendix".

A weaker property than the prox-regularity of a function h is "subdifferential regularity." Formal definitions and discussion can be found in standard texts and in the "Appendix". Here, we simply note that both $\mathcal{C}^1$ functions and lower semicontinuous convex functions are subdifferentially regular, as are sums of such functions.

We next turn to the idea of "partial smoothness" introduced by Lewis [30], a variational-analytic formalization of the notion of the active set in classical nonlinear programming: see also Hare and Lewis [22, Definition 2.3]. A set $\mathcal{M} \subset \Re^m$ is a *manifold about* a point $\bar{c} \in \mathcal{M}$ if it can be described locally by a collection of smooth equations with linearly independent gradients. More precisely, there exists a map $F : \Re^m \to \Re^k$ that is $\mathcal{C}^2$ around $\bar{c}$, with $\nabla F(\bar{c})$ surjective, such that points $c \in \Re^m$ near $\bar{c}$ lie in $\mathcal{M}$ if and only if $F(c) = 0$. The *normal space* to $\mathcal{M}$ at $\bar{c}$, denoted $N_{\mathcal{M}}(\bar{c})$ is then just the range of $\nabla F(\bar{c})^*$.

Definition 1.2 Given a manifold $\mathcal{M} \subset \Re^m$ about a point $\bar{c}$, a function $h : \Re^m \to \bar{\Re}$ is *partly smooth* at $\bar{c}$ relative to $\mathcal{M}$ if h is subdifferentially regular at all points $c \in \mathcal{M}$ near $\bar{c}$, the dependence of the value $h(c)$ and the (nonempty) subdifferential $\partial h(c)$ on the point $c \in \mathcal{M}$ are $\mathcal{C}^2$ and continuous respectively, and furthermore the affine span of $\partial h(\bar{c})$ is a translate of the normal space $N_{\mathcal{M}}(\bar{c})$. We refer to $\mathcal{M}$ as the *active manifold*.

As with prox-regularity, this definition appears technical. To illustrate, consider again the example of continuous function h such that $h + \kappa|\cdot|^2$ is convex for some κ. Since such functions are always subdifferentially regular, partial smoothness amounts to smoothness of the restriction $h|_{\mathcal{M}}$, continuity with respect to the point $c \in \mathcal{M}$ of the classical directional derivative $h'(c; d)$ for all fixed directions d, and the property $h'(\bar{c}; d) > -h'(\bar{c}; -d)$ for all nonzero $d \in N_{\mathcal{M}}(\bar{c})$.

2 Examples

Our basic framework admits a wide variety of interesting problems, as we show in this section.

2.1 Approximation problems

Example 2.1 (Least squares, ℓ_1, and Huber approximation) The formulation (1.1) encompasses both the usual (nonlinear) least squares problem if we define $h(\cdot) = |\cdot|^2$, and the ℓ_1 approximation problem if we define $h(\cdot) = |\cdot|_1$, the ℓ_1 norm. Another popular robust loss function is the Huber function defined by $h(c) = \sum_{i=1}^m \phi(c_i)$, where

$$\phi(c_i) = \begin{cases} \frac{1}{2}c_i^2 & (|c_i| \le T) \\ Tc_i - \frac{1}{2}T^2 & (|c_i| > T). \end{cases}$$

Example 2.2 (Sum of Euclidean norms) Given a collection of smooth vector functions $g_i : \Re^n \to \Re^{m_i}$, for $i = 1, 2, \ldots, t$, consider the problem

$$\min_x \sum_{i=1}^{t} |g_i(x)|.$$

We can place such problems in the form (1.1) by defining the smooth vector function $c : \Re^n \to \Re^{m_1} \times \Re^{m_2} \times \cdots \times \Re^{m_t}$ by $c = (g_1, g_2, \ldots, g_t)$, and the nonsmooth function $h : \Re^{m_1} \times \cdots \times \Re^{m_t} \to \Re$ by

$$h(g_1, g_2, \ldots, g_t) = \sum_{i=1}^{t} |g_i|.$$

This form is seen in facility location problems and in regularized optimization problems with group-sparse regularizers.

2.2 Nonlinear programming penalty functions

Next, we consider examples motivated by penalty functions for nonlinear programming.

Example 2.3 (ℓ_1 penalty function). Consider the following nonlinear program:

$$\begin{aligned} &\min\ f(x) \\ \text{subject to}\ \ &g_i(x) = 0 \quad (1 \le i \le j), \\ &g_i(x) \le 0 \quad (j+1 \le i \le k), \\ &x \in X, \end{aligned} \tag{2.1}$$

where the polyhedron $X \subset \Re^n$ describes constraints on the variable x that are easy to handle directly. The ℓ_1 penalty function formulation is

$$\min_{x \in X} f(x) + \nu \sum_{i=1}^{j} |g_i(x)| + \nu \sum_{i=j+1}^{k} \max\big(0, g_i(x)\big), \tag{2.2}$$

where $\nu > 0$ is a scalar parameter. We can express this problem in the form (1.1) by defining the smooth vector function

$$c(x) = \Big(f(x)\,,\ \big(g_i(x)\big)_{i=1}^{k}\,,\ x\Big) \in \Re \times \Re^k \times \Re^n$$

and the extended polyhedral convex function $h : \Re \times \Re^k \times \Re^n \to \bar{\Re}$ by

$$h(f, g, x) = \begin{cases} f + \nu \sum_{i=1}^{j} |g_i| + \nu \sum_{i=j+1}^{k} \max(0, g_i) & (x \in X) \\ +\infty & (x \notin X). \end{cases}$$

2.3 The finite polyhedral case

A generalization of the polyhedral convex function of the previous subsection is obtained by defining

$$h(c) = \max_{i \in I} \{\langle h_i, c\rangle + \beta_i\}, \tag{2.3}$$

where I is a finite set of indices, with $h_i \in \Re^m$ and $\beta_i \in \Re$ for all $i \in I$. We return to this case below to illustrate much of our theory (in Sects. 4.4, 4.5, for example).

Assume that the map $c\colon \Re^n \to \Re^m$ is $\mathcal{C}^1$ around a critical point $\bar{x} \in \Re^n$ for the composite function $h \circ c$, and let $\bar{c} = c(\bar{x})$. Define the set of "active" indices

$$\bar{I} = \operatorname{argmax}\big\{\langle h_i, \bar{c}\rangle + \beta_i : i \in I\big\}.$$

Then, denoting convex hulls by conv, we have $\partial h(\bar{c}) = \operatorname{conv}\{h_i : i \in \bar{I}\}$. The basic criticality condition (1.4) becomes existence of a vector $\lambda \in \Re^{\bar{I}}$ satisfying

$$\lambda \geq 0 \quad \text{and} \quad \sum_{i \in \bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{2.4}$$

The subgradient $\bar{v}$ is then $\sum_{i \in \bar{I}} \lambda_i h_i$.

Compare this condition with the one obtained from the standard nonlinear programming framework, which is

$$\min_{(x,t) \in \Re^n \times \Re} -t \quad \text{subject to} \quad \langle h_i, c(x)\rangle + \beta_i + t \leq 0 \quad (i \in I). \tag{2.5}$$

At the point $\big(\bar{x}, -h(\bar{c})\big)$, the conditions (2.4) are just the standard first-order optimality conditions, with Lagrange multipliers λ_i. The fact that the vector $\bar{v}$ in the criticality condition (1.4) is closely identified with λ via the relationship $\bar{v} = \sum_{i \in \bar{I}} \lambda_i h_i$ motivates our terminology "multiplier vector".

2.4 Regularized minimization problems

A large family of instances of (1.1) arises in the area of regularized minimization, where the minimization problem has the following general form:

$$\min_x \; f(x) + \tau |x|_* \tag{2.6}$$

where $f : \Re^n \to \Re$ is a smooth objective, while $|x|_*$ is a continuous, nonnegative, usually nonsmooth function, and τ is a nonnegative *regularization parameter*. Such formulations arise when we seek an approximate minimizer of f that is "simple" in some sense; the purpose of the second term $|x|_*$ is to promote this simplicity property.

Larger values of τ tend to produce solutions x that are simpler, but less accurate as minimizers of f. The problem (2.6) can be put into the framework (1.1) by defining

$$c(x) = \begin{bmatrix} f(x) \\ x \end{bmatrix} \in \Re^{n+1}, \qquad h(f, x) = f + \tau|x|_*. \tag{2.7}$$

We list now some interesting cases of (2.6).

Example 2.4 (ℓ_1-Regularized minimization) The choice $|\cdot|_* = |\cdot|_1$ in (2.6) tends to produce solutions x that are *sparse*, in the sense of having relatively few nonzero components. Larger values of τ tend to produce sparser solutions. Compressed sensing is a particular area of interest, in which the objective f is typically a least-squares function $f(x) = (1/2)|Ax - b|^2$; see [9] for a survey. Regularized least-squares problems (or equivalent constrained-optimization formulations) are also encountered in statistics; see for example the LASSO [47] and LARS [16] procedures, and basis pursuit [10].

A related application is regularized logistic regression, where again $|\cdot|_* = |\cdot|_1$, but f is (the negative of) an a posteriori log likelihood function [45]. Here, the components of x are weights applied to the features in a data vector. We aim to identify those features (corresponding to the nonzero locations in x) that are most effective in predicting a binary outcome.

Another interesting class of regularized minimization problems arises in *matrix completion*, where we seek an $m \times n$ matrix X of smallest rank that is consistent with given knowledge of various linear combinations of the elements of X; see [7,8,37]. Much as the ℓ_1 norm of a vector x is used as a surrogate for cardinality of x in the formulations of Example 2.4, the *nuclear norm* is used as a surrogate for the rank of X in formulations of the matrix completion problem. The nuclear norm $|X|_*$ is defined as the sum of singular values of X, and we have the following specialization of (2.6):

$$\min_{X \in \Re^{m \times n}} \frac{1}{2}|\mathcal{A}(X) - b|^2 + \tau|X|_*, \tag{2.8}$$

where $\mathcal{A}$ denotes a linear operator from $\Re^{m \times n}$ to $\Re^p$, and $b \in \Re^p$ is the observation vector. Note that the nuclear norm is a continuous and convex function of X.

Finally, we mention image denoising and deblurring problems, which are often posed in the form (2.6), where $|\cdot|_*$ is a total-variation regularizer [41] that induces "natural" qualities in the solution images. Specifically, the recovered images contain large areas of near-constant color or shade, separated by sharp edges.

For regularized minimization problems of the form (2.6), the subproblem (1.3) has the form

$$\min_d \ f(x) + \langle \nabla f(x), d \rangle + \frac{\mu}{2}|d|^2 + \tau|x + d|_*. \tag{2.9}$$

An equivalent formulation can be obtained by shifting the objective and making the change of variable $z := x + d$:

$$\min_z \frac{\mu}{2}|z-y|^2 + \tau|z|_*, \qquad \text{where} \quad y = x - \frac{1}{\mu}\nabla f(x). \tag{2.10}$$

When the regularization function $|\cdot|_*$ is separable in the components of x, as when $|\cdot|_* = |\cdot|_1$ or $|\cdot| = |\cdot|_2^2$, this problem can be solved in $O(n)$ time. (This fact is key to the practical efficiency of methods based on these subproblems in compressed sensing; see [51].) For the case $|\cdot|_* = |\cdot|_1$, if we set $\alpha = \tau/\mu$, the solution of (2.10) is

$$z_i = \begin{cases} 0 & (|y_i| \le \alpha) \\ y_i - \alpha & (y_i > \alpha) \\ y_i + \alpha & (y_i < -\alpha). \end{cases} \tag{2.11}$$

This operation is known commonly as the "shrink operator."

For matrix completion (2.8), the formulation (2.10) of the subproblem becomes

$$\min_{Z \in \Re^{m\times n}} \frac{\mu}{2}|Z-Y|_F^2 + \tau|Z|_*, \tag{2.12}$$

where $|\cdot|_F$ denotes the Frobenius norm of a matrix and

$$Y = X - \frac{1}{\mu}\mathcal{A}^*[\mathcal{A}(X) - b]. \tag{2.13}$$

It is known (see for example [7]) that (2.12) can be solved by using the singular-value decomposition of Y. Writing $Y = U\Sigma V^T$, where U and V are orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{\min(m,n)})$, we have $Z = U\Sigma_{\tau/\mu}V^T$, where the diagonals of $\Sigma_{\tau/\mu}$ are $\max(\sigma_i - \tau/\mu, 0)$ for $i = 1, 2, \ldots, \min(m, n)$. In essence, we apply the shrink operator to the singular values of Y, and reconstruct Z by using the orthogonal matrices U and V from the decomposition of Y.

2.5 Nonconvex problems

Each of the examples above involves a convex outer function h. In principle, however, the techniques we develop here also apply to a variety of nonconvex functions. This section discusses some applications in which h is nonconvex.

Example 2.5 (problems involving quadratics) Given a general quadratic function $f : \Re^p \to \Re$ (possibly nonconvex) and a smooth function $c_1 : \Re^n \to \Re^p$, consider the problem $\min_x f\big(c_1(x)\big)$. This problem trivially fits into the framework (1.1), and the function f, being $\mathcal{C}^2$, is everywhere prox-regular. The subproblems (1.2), for sufficiently large values of the parameter μ, simply amount to solving a linear system.

More generally, given another general quadratic function $g : \Re^q \to \Re$, and another smooth function $c_2 : \Re^n \to \Re^q$, consider the problem

$$\min_{x\in\Re^n} f\big(c_1(x)\big) \quad \text{subject to} \quad g\big(c_2(x)\big) \le 0.$$

We can express this problem in the form (1.1) by defining the smooth vector function $c = (c_1, c_2)$ and defining an extended-valued nonconvex function

$$h(c_1, c_2) = \begin{cases} f(c_1) & (g(c_2) \le 0) \\ +\infty & (g(c_2) > 0). \end{cases}$$

The epigraph of h is

$$\big\{(c_1, c_2, t) : g(c_2) \le 0,\ t \ge f(c_1)\big\},$$

a set defined by two smooth inequality constraints: hence h is prox-regular at any point (c_1, c_2) satisfying $g(c_2) \le 0$ and $\nabla g(c_2) \ne 0$. The resulting subproblems (1.2) are all in the form of the standard trust-region subproblem, and hence relatively straightforward to solve quickly.

As one more example of this type, we consider the case in which the outer function h is defined as the maximum of a finite collection of quadratic functions (possibly nonconvex): $h(x) = \max\{f_i(x) : i = 1, 2, \ldots, k\}$. The subproblems (1.2) are as follows:

$$\min \Big\{ t : t \ge f_i\big(\Phi(d)\big) + \frac{\mu}{2}|d|^2,\ d \in \Re^m,\ t \in \Re,\ i = 1, 2, \ldots, k \Big\}.$$

where the map Φ is affine. For sufficiently large values of the parameter μ, this is a quadratically-constrained convex quadratic program, which can in principle be solved efficiently by an interior point method.

To conclude, we consider three more nonconvex examples. The first, due to Mangasarian [31], is used by Jokar and Pfetsch [24] to find sparse solutions of underdetermined linear equations. The formulation of [24] can be stated in the form (2.6) where the regularization function $|\cdot|_*$ has the form

$$|x|_* = \sum_{i=1}^{n} (1 - e^{-\alpha |x_i|})$$

for some parameter $\alpha > 0$. It is easy to see that this function is nonconvex but prox-regular, and nonsmooth only at $x_i = 0$.

Fan and Li [17] propose the smoothly clipped absolute deviation (SCAD) regularizer. This problem has the form (2.6), and behaves like the ℓ_1 norm near the origin, transitioning (via a concave quadratic) to a constant for large loss values. Specifically, we have $|\cdot|_* = \sum_{i=1}^{n} \phi(x_i)$, where

$$\phi(x_i) = \begin{cases} \lambda |x_i| & (|x_i| \le \lambda) \\ -(|x_i|^2 - 2a\lambda|x_i| + \lambda^2)/\big(2(a-1)\big) & (\lambda < |x_i| \le a\lambda) \\ (a+1)\lambda^2/2 & (|x_i| > a\lambda). \end{cases}$$

Here $\lambda > 0$ and $a > 1$ are tuning parameters. The minimum concave penalty (MCP) regularizer of Zhang [54] has a similar form, with

$$\phi(x_i) = \begin{cases} \lambda |x_i| - |x_i|^2/(2a) & (|x_i| \le a\lambda) \\ a\lambda^2/2 & (|x_i| > a\lambda). \end{cases} \tag{2.14}$$

SCAD and MCP have been shown to avoid the bias property associated with the ℓ_1 penalty function, in which nonzero values of x are skewed toward zero.

3 Related work

We discuss here some connections of our approach with existing literature.

3.1 Convex h

Burke [3] uses a similar composite function to the one analyzed here, and a subproblem like (1.2) to calculate the search direction d. In contrast to our approach, the analysis in [3] is restricted to finite convex h, and the algorithm uses a backtracking line search to ensure descent in the composite objective at each iteration. In place of the prox term $|d|^2/2$ of (1.2), Burke uses "casting functions" that serve a similar purpose of ensuring well posedness of the subproblem. Sagastizábal [42] considers the problem (1.1) in which h is finite, convex, and positively homogeneous. Her algorithm is based on a subproblem like (1.3), differing mainly in that h is replaced by a lower-bounding bundle approximation. Lan [27, Section 4] discusses (1.1) in which h and the components of $c(x)$ are all Lipschitz continuous and convex. Under certain assumptions on the smoothness of c, a subproblem is defined that makes use of an approximation like the $h\big(c(x) + \nabla c(x)d\big)$ of (1.3), but taking the maximum of such approximations over all previous iterates, not just the one from the latest iterate. Global convergence is proved [27, Corollary 1] at rates that are optimal among first-order schemes.

3.2 Polyhedral h

Various approaches have been proposed for the case of h finite and polyhedral. One work closely related to ours is by Fletcher and Sainz de la Maza [18], who discuss an algorithm for minimization of the ℓ_1 penalty function (2.2) for the nonlinear optimization problem (2.1). At each iteration, their method solves a linearized trust-region problem that can be expressed in our general notation as follows:

$$\min_d \; h\big(c(x) + \nabla c(x)d\big) \quad \text{subject to} \quad |d| \le \rho, \tag{3.1}$$

where ρ is some trust-region radius. Note that this subproblem is closely related to our linearized subproblem (1.3) when the Euclidean norm is used to define the trust region. However, the ℓ_∞ norm is preferred in [18], as it allows the subproblem (3.1) to be expressed as a linear program. The algorithm in [18] uses the solution of (3.1) to estimate the active constraint manifold, then computes a step that minimizes a model of the Lagrangian function for (2.1) while fixing the identified constraints as equalities. An active-constraint identification result is proved [18, Theorem 2.3]; this result is related to our Theorems 4.11 and 5.5 below.

Byrd et al. [6] describe a successive linear-quadratic programming method, based on [18], which starts with solution of the linear program (3.1) (with ℓ_∞ trust region) and

uses it to define an approximate Cauchy point, then approximately solves an equality-constrained quadratic program (EQP) over a different trust region to enhance the step. This algorithm is implemented in the KNITRO package for nonlinear optimization as the KNITRO-ACTIVE option.

Friedlander et al. [19] solve a problem of the form (1.3) for the case of nonlinear programming, where h is the sum of the objective function f and the indicator function for the equalities and the inequalities defining the feasible region. The resulting step can be enhanced by solving an EQP.

Other related literature on composite nonsmooth optimization problems with general finite polyhedral convex functions (Sect. 2.3) includes the papers of Yuan [52,53] and Wright [49]. The approaches in [49,53] solve a linearized subproblem like (3.1), from which an analog of the "Cauchy point" for trust-region methods in smooth unconstrained optimization can be calculated. This calculation involves a line search along a piecewise quadratic function and is therefore more complicated than the calculation in [18], but serves a similar purpose, namely as the basis of an acceptability test for a step obtained from a higher-order model.

3.3 Regularized form (2.6)

For general outer functions h, the theory is more complex. An early approach to regularized minimization problems of the form (2.6) for a lower semicontinuous convex function $|\cdot|_*$ is due to Fukushima and Mine [20]. They calculate a trial step at each iteration by solving the linearized problem (2.9).

Subproblems of the form (2.9) were used in compressed sensing algorithms by Wright, Nowak, and Figueiredo [51] and Hale et al. [21], in conjunction with an adaptive strategy for choosing μ. (Indeed, this application provided the motivation for the current study.)

Combettes and Wajs [12] study formulations similar to (2.6) and algorithms that use subproblems like (2.9). Apart from assuming convexity, their setting is more general. Convergence is proved for algorithms that use values of μ in (2.9) that are large enough to guarantee descent in the objective at every iteration, regardless of iterate x. This assumption contrasts with the adaptive approach used in [51] and in Sect. 5 below.

3.4 $c(x) = x$: Proximal-point methods

The case when the map c is simply the identity has a long history. The iteration $x_{k+1} = x_k + d_k$, where d_k minimizes the function $d \mapsto h(x_k + d) + \frac{\mu}{2}|d|^2$, is the well-known *proximal point method*. For lower semicontinuous convex functions h, convergence was proved by Martinet [32] and generalized by Rockafellar [38]. For nonconvex h, a good survey up to 1998 is by Kaplan and Tichatschke [25]. Pennanen [36] took an important step forward, showing in particular that if the graph of the subdifferential ∂h agrees locally with the graph of the inverse of a Lipschitz function (a condition verifiable using second-order properties including prox-regularity—see Levy [29, Cor. 3.2]), then the proximal point method converges linearly if started nearby and with regularization parameter μ bounded away from zero. This result was

foreshadowed in much earlier work of Spingarn [46], who gave conditions guaranteeing local linear convergence of the proximal point method for a function h that is the sum of lower semicontinuous convex function and a $\mathcal{C}^2$ function, conditions which furthermore hold "generically" under perturbation by a linear function. Inexact variants of Pennanen's approach are discussed by Iusem, Pennanen, and Svaiter [23] and Combettes and Pennanen [11]. In this current work, we make no attempt to build on this more sophisticated theory, preferring a more direct and self-contained approach.

3.5 Manifold identification

The issue of identification of the face of a constraint set on which the solution of a constrained optimization problem lies has been the focus of numerous works. For the problem $\min_{x\in X} f(x)$, for a closed set $X \subset \Re^n$, some papers show that the projection of the point $x - (1/\mu)\nabla f(x)$ onto the feasible set X (for some fixed $\mu > 0$) lies on the same face as the solution $\bar{x}$, under certain nondegeneracy assumptions on the problem and geometric assumptions on X. Identification of so-called quasi-polyhedral faces of convex X was described by Burke and Moré [5]. An extension to the nonconvex case is provided by Burke [4], who considers algorithms that work with linearizations of the constraints describing X. Wright [50] considers surfaces of a convex set X that can be parametrized by a smooth algebraic mapping, and shows how algorithms of gradient projection type can identify such surfaces once the iterates are sufficiently close to a solution. Lewis [30] and Hare and Lewis [22] extend these identification results to the nonconvex, nonsmooth case by using concepts from nonsmooth analysis, including partly smooth functions and prox-regularity. In their setting, the concept of an identifiable face of a feasible set is extended to a certain type of manifold with respect to which the function h in (1.1) is partly smooth (see Definition 1.2 above).

A rich class of convex composite functions with partly smooth structure was discussed in detail by Bonnans and Shapiro [2] and Shapiro [44]. For a detailed discussion of the relationship between that class and partial smoothness, see [1].

3.6 Alternative subproblems

Another line of relevant work is associated with the $\mathcal{VU}$ theory introduced by Lemaréchal, et al. [28] and subsequently elaborated by these and other authors. The focus is on minimizing convex functions $f(x)$ that, again, are partly smooth—smooth ("U-shaped") along a certain manifold through the solution $\bar{x}$, but nonsmooth ("V-shaped") in the transverse directions. Sagastizábal and Mifflin [43] discuss the "fast track," which is essentially the manifold containing the solution $\bar{x}$ along which the objective is smooth. Similarly to [18], they are interested in algorithms that identify the fast track and then take a minimization step for a certain Lagrangian function along this track. It is proved in [43, Theorem 5.2] that under certain assumptions, when x is near $\bar{x}$, the proximal point $x + d$ obtained by solving the problem

$$\min_d \; f(x+d) + \frac{\mu}{2}|d|^2 \tag{3.2}$$

lies on the fast track. This identification result is similar to the one we prove in Sect. 4.5, but the calculation of d is different. In our case of $f = h \circ c$, (3.2) becomes

$$\min_d \; h\big(c(x+d)\big) + \frac{\mu}{2}|d|^2. \tag{3.3}$$

In many applications of interest, c is nonlinear, so the subproblem (3.3) is generally harder to solve for the step d than our subproblem (1.3).

Mifflin and Sagastizábal [33] describe an algorithm in which an approximate solution of (3.2) is obtained, again for the case of a convex objective, by making use of a piecewise linear underapproximation to their objective f, usually constructed from a bundle of subgradients gathered at earlier iterations. Approximations to the manifold of smoothness for f are constructed, and a Newton-like step for the Lagrangian is taken along this manifold. Daniilidis et al. [13] use the terminology "predictor-corrector" to describe algorithms of this type. Miller and Malick [34] show how algorithms of this type are related to Newton-like methods that have been proposed earlier in various contexts.

Various of the algorithms discussed above make use of curvature information for the objective on the active manifold to accelerate local convergence. The algorithmic framework that we describe in Sect. 5 can be modified to incorporate similar techniques, while retaining its global convergence and manifold identification properties. Algorithms with this flavor have been described in [45] for the case of ℓ_1-regularized logistic regression, and [48] for ℓ_1-regularized least squares.

4 Properties of the proximal linearized subproblem

We show in this section that when h is prox-regular at $\bar{c}$, under a mild additional assumption, the subproblem (1.3) has a local solution d with norm $O(|x - \bar{x}|)$, when the parameter μ is sufficiently large. When h is convex, this solution is the unique global solution of the subproblem. We show too that a point x^+ near $x + d$ can be found such that the objective value $h\big(c(x^+)\big)$ is close to the prediction of the model function $h(c(x) + \nabla c(x)d)$ from (1.3). Further, we describe conditions under which the subproblem correctly identifies the manifold $\mathcal{M}$ with respect to which h is partly smooth at the solution of (1.1).

4.1 Lipschitz properties

We start with technical preliminaries. Allowing non-Lipschitz or extended-valued outer functions h in our problem (1.1) is conceptually appealing, since it allows us to model constraints that must be enforced. However, this flexibility presents certain technical challenges, which we now address. We begin with a simple example, to illustrate some of the difficulties.

Example 4.1 Define a $\mathcal{C}^2$ function $c : \Re \to \Re^2$ by $c(x) = (x, x^2)$, and a lower semicontinuous convex function $h : \Re^2 \to \bar{\Re}$ by

$$h(y,z) = \begin{cases} y & (z \geq 2y^2) \\ +\infty & (z < 2y^2). \end{cases}$$

The composite function $h \circ c$ is simply $\delta_{\{0\}}$, the indicator function of $\{0\}$. This function has a global minimum value zero, attained uniquely by $\bar{x} = 0$.

At any point $x \in \Re$, the derivative map $\nabla c(x) : \Re \to \Re^2$ is given by $\nabla c(x)d = (d, 2xd)$ for $d \in \Re$. Then, for all nonzero x, it is easy to check that

$$h\big(c(x) + \nabla c(x)d\big) = +\infty \quad \text{for all } d \in \Re,$$

so the corresponding proximal linearized subproblem (1.3) has no feasible solutions: its objective value is identically $+\infty$.

The adjoint map $\nabla c(0)^* : \Re^2 \to \Re$ is given by $\nabla c(0)^* v = v_1$ for $v \in \Re^2$, and

$$\partial h(0,0) = \big\{v \in \Re^2 : v_1 = 1,\ v_2 \leq 0\big\}.$$

Hence the criticality condition (1.4) has no solution $\bar{v} \in \Re^2$.

This example illustrates two fundamental difficulties. The first is theoretical: the basic criticality condition (1.4) may be unsolvable, essentially because the chain rule fails. The second is computational: if, implicit in the function h, are constraints on acceptable values for $c(x)$, then curvature in these constraints can cause infeasibility in linearizations. As we see below, resolving both difficulties requires a kind of "transversality" condition common in variational analysis.

In this section we make use of the *normal cone* to a set S at a point $s \in S$, denoted by $N_S(s)$, defined in the "Appendix". When S is convex, it coincides exactly with the classical normal cone from convex analysis, while for smooth manifolds it coincides with the classical normal space.

The transversality condition we need involves the "horizon subdifferential" of the function $h : \Re^m \to \bar{\Re}$ at the point $\bar{c} \in \Re^m$, denoted $\partial^\infty h(\bar{c})$. This object, which recurs throughout our analysis, consists of a set of "horizon subgradients", capturing information about directions in which h grows faster than linearly near $\bar{c}$. (See the "Appendix" for a formal definition.) This idea simplifies in important special cases. If h is convex, finite, and lower semicontinuous at $\bar{c}$, we have the following relationship between the subdifferential and the classical normal cone to the domain (see [40, Proposition 8.12]): $\partial^\infty h(\bar{c}) = N_{\mathrm{dom}\, h}(\bar{c})$. We have further that $\partial^\infty h(\bar{c}) = \{0\}$ if h is locally Lipschitz around $\bar{c}$. This condition holds in particular for a convex function h that is continuous at $\bar{c}$.

We seek conditions guaranteeing a reasonable step in the proximal linearized subproblem (1.3). Our key tool is the following technical result.

Theorem 4.1 *Consider a lower semicontinuous function $h : \Re^m \to \bar{\Re}$, a point $\bar{z} \in \Re^m$ where $h(\bar{z})$ is finite, and a linear map $\bar{G} : \Re^n \to \Re^m$ satisfying*

$$\partial^\infty h(\bar{z}) \cap \mathrm{Null}(\bar{G}^*) = \{0\}.$$

Then there exists a constant $\gamma > 0$ such that, for all vectors $z \in \Re^m$ and linear maps $G : \Re^n \to \Re^m$ with (z, G) near $(\bar{z}, \bar{G})$, there exists a vector $w \in \Re^n$ satisfying

$$|w| \leq \gamma |z - \bar{z}| \quad \text{and} \quad h(z + Gw) \leq h(\bar{z}) + \gamma |z - \bar{z}|.$$

Notice that this result is trivial if h is locally Lipschitz (or in particular continuous and convex) around $\bar{z}$, since we can simply choose $w = 0$. The non-Lipschitz case is harder; our proof appears below following the introduction of a variety of ideas from variational analysis whose use is confined to this subsection. We refer the reader to Rockafellar and Wets [40] or Mordukhovich [35] for further details. First, we need a "metric regularity" result, which is proved below by means of a result from Dontchev, Lewis, and Rockafellar [15]. An alternative proof, which sets the result in a broader context, appears in the "Appendix".

Theorem 4.2 (Uniform metric regularity under perturbation) *Suppose that the closed set-valued mapping $F : \Re^p \rightrightarrows \Re^q$ is metrically regular at a point $\bar{u} \in \Re^p$ for a point $\bar{v} \in F(\bar{u})$: in other words, there exist positive constants κ and a such that all points $u \in B_a(\bar{u})$ and $v \in B_a(\bar{v})$ satisfy*

$$\operatorname{dist}\big(u, F^{-1}(v)\big) \leq \kappa \operatorname{dist}\big(v, F(u)\big). \tag{4.1}$$

Then there exist constants $\delta, \gamma > 0$ such that all linear maps $H : \Re^p \to \Re^q$ with $\|H\| < \delta$ and all points $u \in B_\delta(\bar{u})$ and $v \in B_\delta(\bar{v})$ satisfy

$$\operatorname{dist}\big(u, (F + H)^{-1}(v)\big) \leq \gamma \operatorname{dist}\big(v, (F + H)(u)\big). \tag{4.2}$$

Proof We follow the notation of the proof of [15, Theorem 3.3]. Fix any constants

$$\lambda \in (0, \kappa^{-1}), \qquad \alpha \in \left(0, \frac{a}{4}(1 - \kappa\lambda) \min\{1, \kappa\}\right), \qquad \delta \in \left(0, \min\left\{\frac{\alpha}{4}, \frac{\alpha}{4\kappa}, \lambda\right\}\right).$$

Then the proof shows inequality (4.2), if we define $\gamma = \kappa/(1 - \kappa\lambda)$. □

Using this result, and given a closed set S containing 0, we identify a condition under which any vector v can be projected to S along the range space of a given matrix, with the difference between v and its projection being bounded in terms of $|v|$. We prove this result in the "Appendix".

Corollary 4.3 *Consider a closed set $S \subset \Re^q$ with $0 \in S$, and a linear map $\bar{A} : \Re^p \to \Re^q$ satisfying*

$$N_S(0) \cap \operatorname{Null}(\bar{A}^*) = \{0\}.$$

Then there exists a constant $\gamma > 0$ such that, for all vectors $v \in \Re^q$ and linear maps $A : \Re^p \to \Re^q$ with (v, A) near $(0, \bar{A})$, the inclusion

$$v + Au \in S$$

has a solution $u \in \Re^p$ satisfying $|u| \leq \gamma |v|$.

We are now ready to prove the main result of this subsection.

Proof of Theorem 4.1 Let $S \subset \Re^m \times \Re$ be the epigraph of h, and define a map $\bar{A}\colon \Re^n \times \Re \to \Re^m \times \Re$ by $\bar{A}(z,\tau) = (\bar{G}z, \tau)$. From $\partial^\infty h(\bar{z}) \cap \mathrm{Null}(\bar{G}^*) = \{0\}$, we have $\mathrm{Null}(\bar{A}^*) = \mathrm{Null}(\bar{G}^*) \times \{0\}$, so [40, Theorem 8.9] shows that

$$N_S(\bar{z}, h(\bar{z})) \cap \mathrm{Null}(\bar{A}^*) = \{(0,0)\}.$$

For any vector z and linear map G with (z, G) near $(\bar{z}, \bar{G})$, the vector $(z, 0) \in \Re^m \times \Re$ is near the vector $(\bar{z}, 0)$ and the map $(w, \tau) \mapsto (Gw, \tau)$ is near the map $(w, \tau) \mapsto (\bar{G}w, \tau)$. The previous corollary shows the existence of a constant $\gamma > 0$ such that, for all such z and G, the inclusion

$$(z, 0) + (Gw, \tau) \in S$$

has a solution satisfying $|(w, \tau)| \le \gamma |(z - \bar{z}, 0)|$, and the result follows. □

We end this subsection with another tool to be used later, whose proof (in the "Appendix") is a straightforward application of standard ideas from variational analysis. Like Theorem 4.2, this tool concerns metric regularity, this time for a constraint system of the form $F(z) \in S$ for an unknown vector z, where the map F is smooth, and S is a closed set.

Theorem 4.4 (Metric regularity of constraint systems) *Consider a $\mathcal{C}^1$ map $F\colon \Re^p \to \Re^q$, a point $\bar{z} \in \Re^p$, and a closed set $S \subset \Re^q$ containing the vector $F(\bar{z})$. Suppose the condition*

$$N_S\big(F(\bar{z})\big) \cap \mathrm{Null}(\nabla F(\bar{z})^*) = \{0\}$$

holds. Then there exists a constant $\kappa > 0$ such that all points $z \in \Re^p$ near $\bar{z}$ satisfy the inequality

$$\mathrm{dist}\big(z, F^{-1}(S)\big) \le \kappa\, \mathrm{dist}(F(z), S).$$

4.2 The proximal step

We now prove a key result. Under a standard transversality condition, and assuming the proximal parameter μ is sufficiently large (if the function h is nonconvex), we show the existence of a step $d = O(|x - \bar{x}|)$ in the proximal linearized subproblem (1.3) with corresponding objective value close to the critical value $h(\bar{c})$.

When the outer function h is locally Lipschitz (or, in particular, continuous and convex), this result and its proof simplify considerably. First, the transversality condition is automatic. Second, while the proof of the result appeals to the technical tool we developed in the previous subsection (Theorem 4.1), this tool is trivial in the Lipschitz case, as we noted earlier. We state the theorem in a form that encompasses both the general case and the specialization to convex h.

Theorem 4.5 (Proximal step) *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$. Assume the transversality condition*

$$\partial^\infty h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*) = \{0\}. \tag{4.3}$$

Then there exist numbers $\bar{\mu} \geq 0$, $\delta > 0$, and $\bar{\rho} \geq 0$, and a mapping $d : B_\delta(\bar{x}) \times (\bar{\mu}, \infty) \to \Re^n$ such that the following properties hold.

(a) For all points $x \in B_\delta(\bar{x})$ and all parameter values $\mu > \bar{\mu}$, the step $d(x, \mu)$ is a local minimizer of the proximal linearized subproblem (1.3) with

$$h\big(c(x) + \nabla c(x) d(x, \mu)\big) + \frac{\mu}{2}|d(x, \mu)|^2 \leq h\big(c(x)\big),$$

and moreover $|d(x, \mu)| \leq \bar{\rho}|x - \bar{x}|$.

(b) Given any sequences $x_r \to \bar{x}$ and $\mu_r > \bar{\mu}$, then if either $\mu_r |x_r - \bar{x}|^2 \to 0$ or $h\big(c(x_r)\big) \to h(\bar{c})$, we have

$$h\big(c(x_r) + \nabla c(x_r) d(x_r, \mu_r)\big) \to h(\bar{c}). \tag{4.4}$$

(c) When h is convex and lower semicontinuous, the results of parts (a) and (b) hold with $\bar{\mu} = 0$.

Proof Without loss of generality, suppose $\bar{x} = 0$ and $\bar{c} = c(0) = 0$, and furthermore $h(0) = 0$. By assumption, $0 \in \partial(h \circ c)(0) \subset \nabla c(0)^* \partial h(0)$, using the chain rule [40, Thm 10.6], so there exists a vector $v \in \partial h(0) \cap \mathrm{Null}(\nabla c(0)^*)$.

We first prove part (a). By prox-regularity, there exists a constant $\rho \geq 0$ such that

$$h(z) \geq \langle v, z \rangle - \frac{\rho}{2}|z|^2 \tag{4.5}$$

for all small vectors $z \in \Re^m$. Hence, there exists a constant $\delta_1 > 0$ such that ∇c is continuous on $B_{\delta_1}(0)$ and

$$h_{x,\mu}(d) \geq \langle v, c(x) + \nabla c(x) d \rangle - \frac{\rho}{2}|c(x) + \nabla c(x) d|^2 + \frac{\mu}{2}|d|^2$$

for all vectors $x, d \in B_{\delta_1}(0)$. As a consequence, we have that

$$h_{x,\mu}(d) \geq \min_{|x| \leq \delta_1,\, |d| = \delta_1} \left\{ \langle v, c(x) + \nabla c(x) d \rangle - \frac{\rho}{2}|c(x) + \nabla c(x) d|^2 \right\} + \frac{\mu}{2}|d|^2,$$

and the term in braces is finite by continuity of c and ∇c on $B_{\delta_1}(0)$. Hence by choosing $\bar{\mu}$ sufficiently large (certainly greater than $\rho \|\nabla c(0)\|^2$) we can ensure that $h_{x,\bar{\mu}}(d) \geq 1$ whenever $|x| \leq \delta_1$, $|d| = \delta_1$. Then for $x \in B_{\delta_1}(0)$, $|d| = \delta_1$, and $\mu \geq \bar{\mu}$, we have

$$h_{x,\mu}(d) = h_{x,\bar{\mu}}(d) + \frac{1}{2}(\mu - \bar{\mu})|d|^2 \geq 1 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2. \tag{4.6}$$

Since c is $\mathcal{C}^2$ at 0, there exist constants $\beta > 0$ and $\delta_2 \in (0, \delta_1)$ such that, for all $x \in B_{\delta_2}(0)$, the vector

$$z(x) = c(x) - \nabla c(x)x \tag{4.7}$$

satisfies $|z(x)| \leq \beta|x|^2$. Setting $G = \nabla c(x)$, $\bar{G} = \nabla c(0)$, $\bar{z} = 0$, and $z = z(x)$ in Theorem 4.1, we obtain the following result. For some constants $\gamma > 0$ and $\delta_3 \in (0, \delta_2)$, given any vector $x \in B_{\delta_3}(0)$, there exists a vector $\hat{d}(x) \in \Re^n$ (defined by $\hat{d}(x) := w - x$, in the notation of the theorem) satisfying

$$\begin{aligned} |x + \hat{d}(x)| &\leq \gamma|z(x)| \leq \gamma\beta|x|^2 \\ h\big(c(x) + \nabla c(x)\hat{d}(x)\big) &\leq \gamma|z(x)| \leq \gamma\beta|x|^2. \end{aligned}$$

We deduce the existence of a constant $\delta_4 \in (0, \delta_3)$ such that, for all $x \in B_{\delta_4}(0)$, the corresponding $\hat{d}(x)$ satisfies $|\hat{d}(x)| \leq |x| + \gamma\beta|x|^2 < \delta_1$, and

$$\begin{aligned} h_{x,\mu}(\hat{d}(x)) &= h(c(x) + \nabla c(x)\hat{d}(x)) + \frac{\mu}{2}|\hat{d}(x)|^2 \\ &\leq \gamma\beta|x|^2 + \frac{\bar{\mu}}{2}\big(|x| + \gamma\beta|x|^2\big)^2 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2 \\ &< 1 + \frac{1}{2}(\mu - \bar{\mu})\delta_1^2. \end{aligned}$$

The lower semicontinuous function $h_{x,\mu}$ must have a minimizer (which we denote $d(x, \mu)$) over the compact set $B_{\delta_1}(0)$. Since $d = 0$ is feasible for $B_{\delta_1}(0)$, we must have $h_{x,\mu}\big(d(x, \mu)\big) \leq h_{x,\mu}(0) = h\big(c(x)\big)$. Moreover, the inequality above implies that the corresponding minimum value is majorized by $h_{x,\mu}\big(\hat{d}(x)\big)$, and thus is strictly less than $1 + (1/2)(\mu - \bar{\mu})\delta_1^2$. But inequality (4.6) implies that this minimizer must lie in the *interior* of the ball $B_{\delta_1}(0)$; in particular, it must be an unconstrained local minimizer of $h_{x,\mu}$. By setting $\delta = \delta_4$, we complete the proof of the first part of (a). Notice further that for $x \in B_{\delta_4}(0)$, we have

$$\begin{aligned} &h\big(c(x) + \nabla c(x)d(x, \mu)\big) \\ &\quad \leq h_{x,\mu}\big(d(x, \mu)\big) \leq h_{x,\mu}\big(\hat{d}(x)\big) \leq \gamma\beta|x|^2 + \frac{\mu}{2}\big(|x| + \gamma\beta|x|^2\big)^2. \end{aligned} \tag{4.8}$$

We now prove the remainder of part (a), that is, uniform boundedness of the ratio $|d(x, \mu)|/|x|$. Suppose there are sequences $x_r \in B_\delta(\bar{x})$ and $\mu_r > \bar{\mu}$ such that $|d_r|/|x_r| \to \infty$, where we use notation $d_r := d(x_r, \mu_r)$ for brevity. Since $|d_r| \leq \delta_1$ by the arguments above, we must have $x_r \to 0$. By the arguments above, for all large r we have the following inequalities:

$$\begin{aligned} &\gamma\beta|x_r|^2 + \frac{\mu_r}{2}\big(|x_r| + \gamma\beta|x_r|^2\big)^2 \\ &\quad \geq h_{x_r,\mu_r}(d_r) \end{aligned}$$

$$\geq \langle v, c(x_r) + \nabla c(x_r)d_r\rangle - \frac{\rho}{2}|c(x_r) + \nabla c(x_r)d_r|^2 + \frac{\mu_r}{2}|d_r|^2.$$

Dividing each side by $(1/2)\mu_r|x_r|^2$ and letting $r \to \infty$, we recall the inequalities $\mu_r > \bar{\mu} > \rho\|\nabla c(0)\|^2 \geq 0$ and observe that the left-hand side remains finite, while the right-hand side is eventually dominated by $(1-\rho\|\nabla c(0)\|^2/\mu_r)|d_r|^2/|x_r|^2$, which approaches ∞, yielding a contradiction.

For part (b), suppose first that $\mu_r|x_r|^2 \to 0$. By substituting $(x, \mu) = (x_r, \mu_r)$ into (4.8), we have that

$$\limsup\ h\big(c(x_r) + \nabla c(x_r)d_r\big) \leq 0. \tag{4.9}$$

From part (a), we have that $|d_r|/|x_r|$ is uniformly bounded, hence $d_r \to 0$ and thus $c(x_r) + \nabla c(x_r)d_r \to 0$. Being prox-regular, h is lower semicontinuous at 0, so

$$\liminf\ h\big(c(x_r) + \nabla c(x_r)d_r\big) \geq 0.$$

Combining these last two inequalities gives $h\big(c(x_r) + \nabla c(x_r)d_r\big) \to 0$, as required.

Now suppose instead that $h\big(c(x_r)\big) \to h(\bar{c}) = 0$. We have from (4.8) that

$$h\big(c(x_r) + \nabla c(x_r)d_r\big) \leq h_{x_r,\mu_r}(d_r) \leq h_{x_r,\mu_r}(0) = h\big(c(x_r)\big).$$

Taking the lim sup, we again obtain (4.9), and the result follows as before.

For part (c), when h is lower semicontinuous and convex, the argument simplifies. We set $\rho = 0$ in (4.5) and choose the constant $\delta > 0$ so the map ∇c is continuous on $B_\delta(0)$. For constants β and γ as before, Theorem 4.1 again guarantees the existence, for all small points x, of a step $\hat{d}(x)$ satisfying $h\big(c(x) + \nabla c(x)\hat{d}(x)\big) \leq \gamma\beta|x|^2$. It follows that the proximal linearized objective $h_{x,\mu}$ is somewhere finite, so has compact level sets, by coercivity. Thus it has a global minimizer $d(x, \mu)$ (unique, by strict convexity), which must satisfy the inequality

$$h\big(c(x) + \nabla c(x)d(x, \mu)\big) \leq h\big(c(x) + \nabla c(x)\hat{d}(x)\big) \leq \gamma\beta|x|^2.$$

The remainder of the argument proceeds as before. □

We elaborate on Theorem 4.5(b) by giving a simple example of a function prox-regular at $c(\bar{x})$ such that for sequences $x_r \to \bar{x}$ and $\mu_r \to \infty$ that satisfy neither $\mu_r|x_r - \bar{x}|^2 \to 0$ nor $h\big(c(x_r)\big) \to h\big(c(\bar{x})\big)$, there exists a sequence of *global* minimizers $d_r := d(x_r, \mu_r)$ of the subproblem (1.3) for which (4.4) is not satisfied. For a scalar x, take $c(x) = x$ and

$$h(c) = \begin{cases} -c & (c \leq 0) \\ 1 + c & (c > 0). \end{cases}$$

The unique critical point is clearly $\bar{x} = 0$ with $c(\bar{x}) = 0$ and $h\big(c(\bar{x})\big) = 0$, and this problem satisfies the assumptions of the theorem. Consider $x > 0$, for which the subproblem (1.3) is

$$\min_d h_{x,\mu}(d) = h(x+d) + \frac{\mu}{2}d^2 = \begin{cases} -x - d + \frac{\mu}{2}d^2 & (x+d \le 0) \\ 1 + x + d + \frac{\mu}{2}d^2 & (x+d > 0). \end{cases}$$

When $\mu_r x_r \in (0, 1]$, then $d_r = -x_r$ is the only local minimizer of h_{x_r,μ_r}. When $\mu_r x_r > 1$, the situation is more interesting. The value $d_r = -\mu_r^{-1}$ minimizes the "positive" branch of h_{x_r,μ_r}, with function value $1 + x_r - (2\mu_r)^{-1}$, and there is a second local minimizer at $d_r = -x_r$, with function value $(\mu_r/2)x_r^2$. (In both cases, these minimizers satisfy the estimate $|d_r| = O(|x_r - \bar{x}|)$ proved in part (a).) Comparison of the function values show that in fact the global minimum is achieved at the former point ($d_r = -\mu_r^{-1}$) when $x_r > \mu_r^{-1} + \sqrt{2}\mu_r^{-1/2}$. If this step is taken, we have $x_r + d_r > 0$, so the new iterate remains on the upper branch of h. For sequences $x_r = \mu_r^{-1} + 2\mu_r^{-1/2}$ and $\mu_r \to \infty$, we thus have for the global minimizer $d_r = -\mu_r^{-1}$ of h_{x_r,μ_r} that $h(c(x_r) + \nabla c(x_r)d_r) > 1$ for all r, while $h\big(c(\bar{x})\big) = 0$, so that (4.4) does not hold. The alternative sequence of local minimizers $d_r = -x_r$ of does, however, satisfy the limit (4.4).

4.3 Restoring feasibility

In the algorithmic framework to be discussed below, the basic iteration starts at a current point $x \in \Re^n$ such that the function h is finite at the vector $c(x)$. We then solve the proximal linearized subproblem (1.3) to obtain the step $d = d(x, \mu) \in \Re^n$. Under reasonable conditions we have shown that, for x near the critical point $\bar{x}$, we have $d = O(|x - \bar{x}|)$ and furthermore we know that the value of h at the vector $c(x) + \nabla c(x)d$ is close to the critical value $h\big(c(\bar{x})\big)$.

The algorithmic idea is now to update the point x to a new point $x + d$. When the function h is Lipschitz, this update is motivated by the fact that, since the map c is $\mathcal{C}^2$, we have, uniformly for x near the critical point $\bar{x}$,

$$c(x+d) - (c(x) + \nabla c(x)d) = O(|d|^2)$$

and hence

$$h\big(c(x+d)\big) - h\big((c(x) + \nabla c(x)d)\big) = O(|d|^2).$$

However, if h is not Lipschitz, it may not be appropriate to update x to $x + d$: the value $h\big(c(x+d)\big)$ may even be infinite.

In order to take another step, we need somehow to restore the point $x + d$ to feasibility, or more generally to find a nearby point with objective value not much worse than our linearized estimate $h(c(x) + \nabla c(x)d)$. Depending on the form of the function h, this may or may not be easy computationally. However, as we now discuss, our fundamental transversality condition (4.3), guarantees that such a restoration is always possible in theory. In the next section, we refer to this restoration process as an "efficient projection."

Theorem 4.6 (Linear estimator improvement) *Consider a map $c\colon \Re^n \to \Re^m$ that is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$, and a lower semicontinuous function $h\colon \Re^m \to \bar{\Re}$ that is finite at the vector $\bar{c} = c(\bar{x})$. Assume that the transversality condition (4.3) holds. Then there exist constants γ and $\delta > 0$ such that, for any point $x \in B_\delta(\bar{x})$ and any step $d \in B_\delta(0) \subset \Re^n$ for which $|h(c(x) + \nabla c(x)d) - h(\bar{c})| < \delta$, there exists a point $x^+ \in \Re^n$ satisfying*

$$|x^+ - (x+d)| \le \gamma |d|^2 \quad \text{and} \quad h\big(c(x^+)\big) \le h(c(x) + \nabla c(x)d) + \gamma |d|^2. \tag{4.10}$$

Proof Define a $\mathcal{C}^2$ map $F\colon \Re^n \times \Re \to \Re^m \times \Re$ by $F(x,t) = (c(x), t)$. Notice that the epigraph $\operatorname{epi} h$ is a closed set containing the vector $F\big(\bar{c}, h(\bar{c})\big)$. Clearly we have

$$\operatorname{Null}\Big(\nabla F\big(\bar{x}, h(\bar{c})\big)^*\Big) = \operatorname{Null}\big(\nabla c(\bar{x})^*\big) \times \{0\}.$$

Recalling the relationship (6.5) between $\partial^\infty h$ and $\operatorname{epi} h$ at $\bar{c}$, we have

$$(y, 0) \in N_{\operatorname{epi} h}\big(\bar{c}, h(\bar{c})\big) \Leftrightarrow y \in \partial^\infty h(\bar{c}).$$

Hence the transversality condition is equivalent to

$$N_{\operatorname{epi} h}\big(\bar{c}, h(\bar{c})\big) \cap \operatorname{Null}\Big(\nabla F\big(\bar{x}, h(\bar{c})\big)^*\Big) = \{0\}.$$

We next apply Theorem 4.4 to deduce the existence of a constant $\kappa > 0$ such that, for all vectors (u,t) near the vector $\big(\bar{c}, h(\bar{c})\big)$ we have

$$\operatorname{dist}\big((u,t), F^{-1}(\operatorname{epi} h)\big) \le \kappa \operatorname{dist}(F(u,t), \operatorname{epi} h).$$

Thus there exists a constant $\delta > 0$ such that, for any point $x \in B_\delta(\bar{x})$ and any step $d \in \Re^n$ satisfying $|d| \le \delta$ and $|h(c(x) + \nabla c(x)d) - h(\bar{c})| \le \delta$, we have

$$\begin{aligned}
&\operatorname{dist}\Big(\big(x+d, h(c(x) + \nabla c(x)d)\big), F^{-1}(\operatorname{epi} h)\Big) \\
&\quad \le \kappa \operatorname{dist}\big(F\big(x+d, h(c(x) + \nabla c(x)d)\big), \operatorname{epi} h\big) \\
&\quad = \kappa \operatorname{dist}\Big(\big(c(x+d), h(c(x) + \nabla c(x)d)\big), \operatorname{epi} h\Big) \\
&\quad \le \kappa\, |c(x+d) - (c(x) + \nabla c(x)d)|,
\end{aligned}$$

since

$$\big(c(x) + \nabla c(x)d, h(c(x) + \nabla c(x)d)\big) \in \operatorname{epi} h.$$

Since the map c is $\mathcal{C}^2$, by reducing δ if necessary we can ensure the existence of a constant $\gamma > 0$ such that the right-hand side of the above chain of inequalities is bounded above by $\gamma |d|^2$.

We have therefore shown the existence of a vector $(x^+, t) \in F^{-1}(\text{epi}\, h)$ satisfying the inequalities $|x^+ - (x + d)| \le \gamma |d|^2$ and $|t - h(c(x) + \nabla c(x)d)| \le \gamma |d|^2$. Since $t \ge h\big(c(x^+)\big)$, the result follows. □

4.4 Uniqueness of the proximal step and convergence of multipliers

Our focus in this subsection is on uniqueness of the local solution of (1.3) near $d = 0$, uniqueness of the corresponding multiplier vector, and on showing that the solution $d(x, \mu)$ of (1.3) has a strictly lower subproblem objective value than $d = 0$. For the uniqueness results, we strengthen the transversality condition (4.3) to a constraint qualification that we now introduce.

Throughout this subsection we assume that the function h is prox-regular at the point $\bar{c}$. Since prox-regular functions are subdifferentially regular, the subdifferential $\partial h(\bar{c})$ is a closed and convex set in $\Re^m$, and its recession cone is exactly the horizon subdifferential $\partial^\infty h(\bar{c})$ (see [40, Corollary 8.11]). Denoting the subspace parallel to the affine span of the subdifferential by $\text{par}\, \partial h(\bar{c})$, we deduce that $\partial^\infty h(\bar{c}) \subset \text{par}\, \partial h(\bar{c})$. Hence the "constraint qualification" that we next consider, namely

$$\text{par}\, \partial h(\bar{c}) \cap \text{Null}(\nabla c(\bar{x})^*) = \{0\}, \tag{4.11}$$

implies the transversality condition (4.3).

Condition (4.11) is related to the linear independence constraint qualification in nonlinear programming. To illustrate, consider again the case of Sect. 2.3, where the function h is finite and polyhedral:

$$h(c) = \max_{i \in I} \{\langle h_i, c\rangle + \beta_i\}$$

for given vectors $h_i \in \Re^m$ and scalars β_i. Then, as we noted, $\partial h(\bar{c}) = \text{conv}\{h_i : i \in \bar{I}\}$, where $\bar{I}$ is the set of active indices, so

$$\text{par}\, \partial h(\bar{c}) = \Big\{ \sum_{i \in \bar{I}} \lambda_i h_i : \sum_{i \in \bar{I}} \lambda_i = 0 \Big\}.$$

Thus condition (4.11) states

$$\sum_{i \in \bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \sum_{i \in \bar{I}} \lambda_i h_i = 0. \tag{4.12}$$

By contrast, the linear independence constraint qualification for the corresponding nonlinear program (2.5) at the point $\big(\bar{x}, -h(\bar{c})\big)$ is

$$\sum_{i \in \bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \lambda_i = 0 \quad (i \in \bar{I}),$$

which is a stronger assumption than condition (4.12).

We now prove a straightforward technical result that addresses two issues: existence and boundedness of multipliers for the proximal subproblem (1.3), and the convergence of these multipliers to a unique multiplier that satisfies criticality conditions for (1.1), when the constraint qualification (4.11) is satisfied. The argument is routine but, as usual, it simplifies considerably in the case of h locally Lipschitz (or in particular convex and continuous) around the point $\bar{c}$, since then the horizon subdifferential $\partial^\infty h$ is identically $\{0\}$ near $\bar{c}$.

Lemma 4.7 *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$.*

When the transversality condition (4.3) holds, then for any sequences $\mu_r > 0$ and $x_r \to \bar{x}$ such that $\mu_r |x_r - \bar{x}| \to 0$, and any sequence of critical points $d_r \in \Re^n$ for the corresponding proximal linearized subproblems (1.3) satisfying the conditions

$$d_r = O(|x_r - \bar{x}|) \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r) d_r\big) \to h(\bar{c}),$$

there exists a bounded sequence of vectors $v_r \in \Re^m$ that satisfy

$$0 = \nabla c(x_r)^* v_r + \mu_r d_r, \tag{4.13a}$$

$$v_r \in \partial h\big(c(x_r) + \nabla c(x_r) d_r\big). \tag{4.13b}$$

When the stronger constraint qualification (4.11) holds, in place of (4.3), the set of multipliers $v \in \Re^m$ solving the criticality condition (1.4), namely

$$\partial h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*) \tag{4.14}$$

is in fact a singleton $\{\bar{v}\}$. Furthermore, any sequence of multipliers $\{v_r\}$ satisfying the conditions above converges to $\bar{v}$.

Proof We first assume (4.3), and claim that

$$\partial^\infty h\big(c(x_r) + \nabla c(x_r) d_r\big) \cap \mathrm{Null}(\nabla c(x_r)^*) = \{0\} \tag{4.15}$$

for all large r. Indeed, if this property should fail, then for infinitely many r there would exist a unit vector v_r lying in the intersection on the left-hand side, and any accumulation point of these unit vectors must lie in the set

$$\partial^\infty h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*), \tag{4.16}$$

by outer semicontinuity of the set-valued mapping $\partial^\infty h$ at the point $\bar{c}$ [40, Proposition 8.7], contradicting the transversality condition (4.3). As a consequence, we can apply the chain rule [40, Theorem 10.6] to deduce the existence of vectors $v_r \in \Re^n$ satisfying (4.13). This sequence must be bounded, since otherwise, after taking a subsequence, we could suppose $|v_r| \to \infty$ and then any accumulation point of the

unit vectors $|v_r|^{-1}v_r$ would lie in the set (4.16), again contradicting the transversality condition. The first claim of the theorem is proved.

For the remaining claims, note first that the chain rule implies that the set (4.14) is nonempty. The constraint qualification (4.11) then implies that this set is a singleton $\{\bar{v}\}$. Using boundedness of $\{v_r\}$, and the fact that $\mu_r d_r \to 0$, we have by taking limits in (4.13) that any accumulation point of $\{v_r\}$ lies in (4.14) (by h-attentive outer semicontinuity of ∂h at $\bar{c}$), and therefore $v_r \to \bar{v}$. □

Using Theorem 4.5, we show that the local minimizers of h_{x_r,μ_r} satisfy the desired properties, and in addition give a strict improvement over 0 in the subproblem (1.3).

Lemma 4.8 *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, that the composite function $h \circ c$ is critical at $\bar{x}$, and that the transversality condition (4.3) holds. Then there is a constant $\bar{\mu} \geq 0$ with the following property. If $\mu_r > \bar{\mu}$ and $x_r \to \bar{x}$ are sequences such that $\mu_r|x_r - \bar{x}| \to 0$, then for all r sufficiently large, we have the following.*

(a) There is a local minimizer d_r of h_{x_r,μ_r} such that

$$d_r = O(|x_r - \bar{x}|) \quad \text{and} \quad h(c(x_r) + \nabla c(x_r)d_r) \to h(\bar{c}). \tag{4.17}$$

(b) If $0 \notin \partial(h \circ c)(x_r)$ for all r, then $d_r \neq 0$ and

$$h_{x_r,\mu_r}(d_r) < h_{x_r,\mu_r}(0) \tag{4.18}$$

for all r sufficiently large.

Proof Part (a) follows from parts (a) and (b) of Theorem 4.5 when we choose $\bar{\mu}$ as in that theorem and set $d_r = d(x_r, \mu_r)$.

For part (b), we have from (4.17) and Lemma 4.7 that there exists v_r satisfying (4.13). If we were to have $d_r = 0$, these conditions would reduce to $\nabla c(x_r)^* v_r = 0$ and $v_r \in \partial h\big(c(x_r)\big)$, so that $0 \in \partial(h \circ c)(x_r)$, by subdifferential regularity of h. Hence we must have $d_r \neq 0$. To prove (4.18), suppose for contradiction that there are sequences μ_r, x_r with the assumed properties such that this inequality does not hold for all r sufficiently large. Without losing generality, we can assume that (4.18) fails to hold for *every* r. By taking limits in (4.13) and from boundedness of $\{v_r\}$, we can assume without loss of generality that $v_r \to \bar{v}$, for some $\bar{v}$ with $\nabla c(\bar{x})^*\bar{v} = 0$, $\bar{v} \in \partial h(\bar{c})$, where we have used h-attentive outer semicontinuity of $\partial h(\cdot)$ to obtain the latter inclusion. Let ρ be the constant from Definition 1.1 associated with $\bar{c}$ and $\bar{v}$, and choose $\bar{\mu}$ such that $\bar{\mu} > \rho\|\nabla c(\bar{x})\|^2$. By prox-regularity, we have

$$\begin{aligned} h\big(c(x_r)\big) &\geq h(c(x_r) + \nabla c(x_r)d_r) + \langle v_r, -\nabla c(x_r)d_r\rangle - \frac{\rho}{2}|\nabla c(x_r)d_r|^2 \\ &= h(c(x_r) + \nabla c(x_r)d_r) + \mu_r|d_r|^2 - \frac{\rho}{2}|\nabla c(x_r)d_r|^2 \qquad \text{by (4.13a)} \\ &\geq h(c(x_r) + \nabla c(x_r)d_r) + \frac{\mu_r}{2}|d_r|^2 + \frac{\mu_r - \rho\|\nabla c(x_r)\|^2}{2}|d_r|^2 \end{aligned}$$

$$
\begin{aligned}
&= h_{x_r,\mu_r}(d_r) + \frac{\mu_r - \rho\|\nabla c(x_r)\|^2}{2}|d_r|^2 \qquad \text{by (1.3)}\\
&> h_{x_r,\mu_r}(d_r),
\end{aligned}
$$

where the final inequality holds because of our choice of $\bar{\mu}$. Since $h_{x_r,\mu_r}(0) = h\big(c(x_r)\big)$, we have a contradiction, and the proof is complete. □

Returning to the assumptions of Theorem 4.5, but now with the constraint qualification (4.11) replacing the weaker transversality condition (4.3), we can derive local uniqueness results about critical points for the proximal linearized subproblem. When the outer function h is convex, uniqueness is obvious, since then the proximal linearized objective $h_{\mu,x}$ is strictly convex for any $\mu > 0$. For lower $\mathcal{C}^2$ functions, the argument is much the same: such functions have the form $g - \kappa|\cdot|^2$, locally, for some continuous convex function g, so again $h_{\mu,x}$ is locally strictly convex for large μ. For general prox-regular functions, the argument requires slightly more care.

Theorem 4.9 (Unique step) *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Suppose that c is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$, that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and that the composite function $h \circ c$ is critical at $\bar{x}$. Suppose further that the constraint qualification (4.11) holds. Then there exists $\bar{\mu} \geq 0$ such that the following properties hold. Given any sequence $\{\mu_r\}$ with $\mu_r > \bar{\mu}$ for all r and any sequence $x_r \to \bar{x}$ such that $\mu_r|x_r - \bar{x}| \to 0$, there exists a sequence of local minimizers d_r of h_{x_r,μ_r} and a corresponding sequence of multipliers v_r with the following properties:*

$$
0 \in \partial h_{x_r,\mu_r}(d_r), \quad d_r = O(|x_r - \bar{x}|), \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r)d_r\big) \to h(\bar{c}), \tag{4.19}
$$

as $r \to \infty$, and satisfying (4.13), with $v_r \to \bar{v}$, where $\bar{v}$ is the unique vector that solves the criticality condition (1.4). Moreover, d_r is uniquely defined for all r sufficiently large.

In the case of a convex, lower semicontinuous function $h : \Re^m \to (-\infty, +\infty]$, the result holds with $\bar{\mu} = 0$.

Proof Existence of sequences $\{d_r\}$ and $\{v_r\}$ with the claimed properties follows from Theorem 4.5 and Lemma 4.7, where we select $\bar{\mu}$ in the same way as in Theorem 4.5. We need only prove the claim about uniqueness of the vectors d_r, and the final claim about the special case of h convex and lower semicontinuous.

We first show the uniqueness of d_r in the general case. Since the function h is prox-regular at $c(\bar{x})$, its subdifferential ∂h has a hypomonotone localization T around the point $(c(\bar{x}), \bar{v})$ with constant $\rho > 0$ (see the "Appendix"). If the uniqueness claim does not hold, we have by taking a subsequence if necessary that there is a sequence $x_r \to \bar{x}$ and distinct sequences of $d_r^1 \neq d_r^2$ in $\Re^n$ satisfying the conditions

$$
0 \in \partial h_{x_r,\mu_r}(d_r^i), \quad d_r^i = O(|x_r - \bar{x}|) \to 0, \quad \text{and} \quad h\big(c(x_r) + \nabla c(x_r)d_r^i\big) \to h\big(c(\bar{x})\big),
$$

as $r \to \infty$, for $i = 1, 2$. Lemma 4.7 shows the existence of sequences of vectors $v_r^i \in \Re^n$ satisfying

$$0 = \nabla c(x_r)^* v_r^i + \mu_r d_r^i$$
$$v_r^i \in \partial h\big(c(x_r) + \nabla c(x_r) d_r^i\big),$$

for all large r, and furthermore $v_r^i \to \bar{v}$ for each $i = 1, 2$. Consequently, for all large r we have

$$v_r^i \in T\big(c(x_r) + \nabla c(x_r) d_r^i\big) \quad \text{for } i = 1, 2,$$

so that

$$-\mu_r |d_r^1 - d_r^2|^2 = \langle v_r^1 - v_r^2, \nabla c(x_r)\,(d_r^1 - d_r^2)\rangle \geq -\rho \left|\nabla c(x_r)\left(d_r^1 - d_r^2\right)\right|^2.$$

Since $\bar{\mu} > \rho\|\nabla c(\bar{x})\|^2$, we have the contradiction $\rho\|\nabla c(x_r)\|^2 \geq \mu_r > \bar{\mu} > \rho\|\nabla c(\bar{x})\|^2$ for all large r.

For the special case of h convex and lower semicontinuous, we have from Theorem 4.5(c) that unique d_r with the properties (4.19) exists, for $\bar{\mu} = 0$. □

4.5 Manifold identification

We next work toward the identification result. Consider a sequence of points $\{x_r\}$ in $\Re^n$ converging to the critical point $\bar{x}$ of the composite function $h \circ c$, and let μ_r be a sequence of positive proximality parameters. Suppose now that the outer function h is partly smooth at the point $\bar{c} = c(\bar{x}) \in \Re^m$ relative to some manifold $\mathcal{M} \subset \Re^m$. Our aim is to find conditions guaranteeing that the update to the point $c(x_r)$ predicted by minimizing the proximal linearized objective h_{x_r,μ_r} lies on $\mathcal{M}$: in other words,

$$c(x_r) + \nabla c(x_r) d_r \in \mathcal{M} \quad \text{for all large } r,$$

where d_r is the unique small critical point of h_{x_r,μ_r}. We would furthermore like to ensure that the "efficient projection" x^+ resulting from this prediction, guaranteed by Theorem 4.6 (linear estimator improvement), satisfies $c(x^+) \in \mathcal{M}$.

To illustrate, we return to our ongoing example from Sect. 2.3, the finite polyhedral function (2.3). If $\bar{I}$ is the active index set corresponding to the point $\bar{c}$, then it is easy to check that h is partly smooth relative to the manifold

$$\mathcal{M} = \big\{c : \langle h_i, c\rangle + \beta_i = \langle h_j, c\rangle + \beta_j \text{ for all } i, j \in \bar{I}\big\}.$$

Our analysis requires one more assumption, in addition to those of Theorem 4.9. The basic criticality condition (1.4) requires the existence of a multiplier vector:

$$\partial h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*) \neq \emptyset.$$

We now strengthen this assumption slightly, to a "strict" criticality condition:

$$\mathrm{ri}\big(\partial h(\bar{c})\big) \cap \mathrm{Null}(\nabla c(\bar{x})^*) \neq \emptyset, \tag{4.20}$$

where ri denotes the relative interior of a convex set. The condition (4.20) is related to the strict complementarity assumption in nonlinear programming. For finite polyhedral h (2.3), since $\partial h(\bar{c}) = \mathrm{conv}\{h_i : i \in \bar{I}\}$, we have

$$\mathrm{ri}\big(\partial h(\bar{c})\big) = \Big\{ \sum_{i\in\bar{I}} \lambda_i h_i : \sum_{i\in\bar{I}} \lambda_i = 1,\ \lambda > 0 \Big\}.$$

Hence, the strict criticality condition (4.20) becomes the existence of a vector $\lambda \in \Re^{\bar{I}}$ satisfying

$$\lambda > 0 \quad \text{and} \quad \sum_{i\in\bar{I}} \lambda_i \begin{bmatrix} \nabla c(\bar{x})^* h_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}. \tag{4.21}$$

The only change from the corresponding basic criticality condition (2.4) is that the condition $\lambda \geq 0$ has been strengthened to $\lambda > 0$, corresponding exactly to the extra requirement of strict complementarity in the nonlinear programming formulation (2.1).

Recall that the constraint qualification (4.11) implies the uniqueness of the multiplier vector $\bar{v}$, by Lemma 4.7. Assuming in addition the strict criticality condition (4.20), we then have

$$\bar{v} \in \mathrm{ri}\big(\partial h(\bar{c})\big) \cap \mathrm{Null}(\nabla c(\bar{x})^*).$$

We now prove a trivial modification of [22, Theorem 5.3].

Theorem 4.10 *Suppose the function $h : \Re^m \to \bar{\Re}$ is partly smooth at the point $\bar{c} \in \Re^m$ relative to the manifold $\mathcal{M} \subset \Re^m$, and is prox-regular there. Consider a subgradient $\bar{v} \in \mathrm{ri}\,\partial h(\bar{c})$. Suppose the sequence $\{\hat{c}_r\} \subset \Re^m$ satisfies $\hat{c}_r \to \bar{c}$ and $h(\hat{c}_r) \to h(\bar{c})$. Then $\hat{c}_r \in \mathcal{M}$ for all large r if and only if* $\mathrm{dist}\big(\bar{v}, \partial h(\hat{c}_r)\big) \to 0$.

Proof The proof proceeds exactly as in [22, Theorem 5.3], except that instead of defining a function $g : \Re^m \times \Re \to \Re$ by $g(c, r) = r$, we set $g(c, r) = r - c^T \bar{v}$. □

We can now prove our main identification result.

Theorem 4.11 *Consider a function $h : \Re^m \to \bar{\Re}$, and a map $c : \Re^n \to \Re^m$ that is $\mathcal{C}^2$ around the point $\bar{x} \in \Re^n$. Suppose that h is prox-regular at the point $\bar{c} = c(\bar{x})$, and partly smooth there relative to the manifold $\mathcal{M}$. Suppose further that the constraint qualification* (4.11) *and the strict criticality condition* (4.20) *both hold for the composite function $h \circ c$ at $\bar{x}$. Then there exist nonnegative constants $\hat{\mu}$ and γ with the following property. Given any sequence $\{\mu_r\}$ with $\mu_r > \hat{\mu}$ for all r, and any sequence $x_r \to \bar{x}$ such that $\mu_r |x_r - \bar{x}| \to 0$, the local minimizer d_r of h_{x_r,μ_r} defined in* Theorem 4.9 *satisfies, for all large r, the condition*

$$c(x_r) + \nabla c(x_r) d_r \in \mathcal{M}, \tag{4.22}$$

and also the inequalities

$$|x_r^{\text{new}} - (x_r + d_r)| \leq \gamma |d_r|^2 \quad \text{and} \quad h\big(c(x_r^{\text{new}})\big) \leq h\big(c(x_r) + \nabla c(x_r) d_r\big) + \gamma |d_r|^2, \tag{4.23}$$

hold for some point x_r^{new} *with* $c(x_r^{\text{new}}) \in \mathcal{M}$.

In the special case when $h : \Re^m \to (-\infty, +\infty]$ *is convex and lower semicontinuous, the result holds with* $\hat{\mu} = 0$.

Proof Theorem 4.9 implies $d_r \to 0$, so $\hat{c}_r = c(x_r) + \nabla c(x_r) d_r \to \bar{c}$. The theorem also shows $h(\hat{c}_r) \to h(\bar{c})$, and that there exist multiplier vectors $v_r \in \partial h(\hat{c}_r)$ satisfying $v_r \to \bar{v} \in \text{ri}\, \partial h(\bar{c})$. Since $\text{dist}\big(\bar{v}, \partial h(\hat{c}_r)\big) \leq |\bar{v} - v_r| \to 0$, we can apply Theorem 4.10 to obtain property (4.22).

Let us now define a function $h_{\mathcal{M}} \colon \Re^m \to \bar{\Re}$, agreeing with h on the manifold $\mathcal{M}$ and taking the value $+\infty$ elsewhere. By partial smoothness, $h_{\mathcal{M}}$ is the sum of a smooth function and the indicator function of $\mathcal{M}$, and hence $\partial^\infty h_{\mathcal{M}}(\bar{c}) = N_{\mathcal{M}}(\bar{c})$. Partial smoothness also implies $\text{par}\big(\partial h(\bar{c})\big) = N_{\mathcal{M}}(\bar{c})$. We can therefore rewrite the constraint qualification (4.11) in the form $\partial^\infty h_{\mathcal{M}}(\bar{c}) \cap \text{Null}\big(\nabla c(\bar{x})^*\big) = \{0\}$. This condition allows us to apply Theorem 4.6 (linear estimator improvement), with the function $h_{\mathcal{M}}$ replacing the function h, to deduce the existence of the point x_r^{new}, as required. □

5 A proximal descent algorithm

We now describe Algorithm ProxDescent, a simple first-order algorithm that manipulates the proximality parameter μ in (1.3) to achieve a "sufficient decrease" in h at each iteration. (This algorithm is shown in the figure below as Algorithm 2.) We follow our description with results concerning the global convergence behavior of this method and its ability to identify the manifold $\mathcal{M}$ discussed in Sect. 4.5.

A few remarks about Algorithm ProxDescent are in order. First, we are not specific about the derivation of x^+ from $x_k + d$, but we assume that the "efficient projection" technique that is the basis of Theorem 4.6 is used when possible. Lemma 4.8 indicates that for μ sufficiently large and x near a critical point $\bar{x}$ of $h \circ c$, it is indeed possible to find a local solution d of (1.3) which satisfies $h_{x,\mu}(d) < h_{x,\mu}(0)$ as required by the algorithm, and which also satisfies the conditions of Theorem 4.6. Lemma 5.2 below shows further that the new point x^+ satisfies the acceptance tests in the algorithm. However, Lemma 5.2 is more general in that it also gives conditions for acceptance of the step when x_k is *not* in a neighborhood of a critical point of $h \circ c$.

Second, we note that the framework allows x^+ to be improved further. For example, we could use higher-order derivatives of c to take a further step along the manifold of h identified by the subproblem (1.3) (analogous to an "EQP step" in nonlinear programming) and reset x^+ accordingly if this step produces a reduction in $h \circ c$. We discuss this point further at the end of the section.

We start our convergence analysis with a technical result showing that in the neighborhood of a non-critical point $\bar{x}$, and for bounded μ, the steps d do not become too short.

Algorithm 2 ProxDescent

Define constants $\tau > 1$, $\sigma \in (0, 1)$, and $\mu_{\min} > 0$;
Choose $x_0 \in \Re^n$, $\mu_0 \geq \mu_{\min}$;
Set $\mu \leftarrow \mu_0$;
for $k = 0, 1, 2, \ldots$ **do**
  Set accept $\leftarrow$ false;
  while not accept **do**
    if $d = 0$ is a local minimizer of (1.3) **then**
      Terminate with $\bar{x} = x_k$;
    end if
    Find a local minimizer d of (1.3) with $x = x_k$, that is

$$\min_d h_{x_k,\mu}(d) := h\big(c(x_k) + \nabla c(x_k)d\big) + \frac{\mu}{2}|d|^2,$$

    such that $h_{x_k,\mu}(d) < h_{x_k,\mu}(0)$;
    if no such d exists **then**
      $\mu \leftarrow \tau\mu$;
    else
      Derive x^+ from $x_k + d$ (by an efficient projection and/or other enhancements);
      if $h\big(c(x_k)\big) - h\big(c(x^+)\big) \geq \sigma\left[h\big(c(x_k)\big) - h(c(x_k) + \nabla c(x_k)d)\right]$
          and $|x^+ - (x_k + d)| \leq \frac{1}{2}|d|$ **then**
        $x_{k+1} \leftarrow x^+$;
        $d_k \leftarrow d$;
        $\mu_k \leftarrow \mu$;
        $\mu \leftarrow \max(\mu_{\min}, \mu/\tau)$;
        accept $\leftarrow$ true;
      else
        $\mu \leftarrow \tau\mu$;
      end if
    end if
  end while
end for

Lemma 5.1 *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Let $\bar{x}$ be such that: c is $\mathcal{C}^1$ near $\bar{x}$; h is finite at the point $\bar{c} = c(\bar{x})$ and subdifferentially regular there; the transversality condition (4.3) holds; but the criticality condition (1.4) is* not *satisfied. Then there exists a quantity $\epsilon > 0$ such that for any sequence $x_r \to \bar{x}$ with $h\big(c(x_r)\big) \to h(\bar{c})$, and any sequence $\{\mu_r\}$ with $\mu_r \geq \mu_{\min}$, any sequence of critical points d_r of h_{x_r,μ_r} satisfying $h_{x_r,\mu_r}(d_r) \leq h_{x_r,\mu_r}(0)$ must also satisfy* $\liminf_r \mu_r|d_r| \geq \epsilon$.

Proof If the result were not true, there would exist sequences x_r, μ_r, and d_r as above except that $\mu_r d_r \to 0$. We would then have $0 \leq |d_r| \leq \mu_r|d_r|/\mu_{\min} \to 0$. Noting that $h(c(x_r) + \nabla c(x_r)d_r) \to h(\bar{c})$ (using lower semicontinuity and the fact that the left-hand side is dominated by $h\big(c(x_r)\big)$, which converges to $h(\bar{c})$), we have that

$$\partial^\infty h\big(c(x_r) + \nabla c(x_r)d_r\big) \cap \mathrm{Null}(\nabla c(x_r)^*) = \{0\},$$

for all r sufficiently large. (If this were not true, we could use an h-attentive outer semicontinuity argument based on [40, Proposition 8.7] to deduce that $\partial^\infty h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*)$ contains a nonzero vector, thus violating the transversality condition

(4.3).) Hence, we can apply the chain rule and deduce that there are multiplier vectors v_r such that (4.13) is satisfied, that is,

$$0 = \nabla c(x_r)^* v_r + \mu_r d_r,$$
$$v_r \in \partial h\big(c(x_r) + \nabla c(x_r) d_r\big),$$

for all sufficiently large r. If the sequence $\{v_r\}$ is unbounded, we can assume without loss of generality that $|v_r| \to \infty$. Any accumulation point of the sequence $v_r/|v_r|$ would be a unit vector in the set $\partial^\infty h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*)$, contradicting (4.3). Hence, the sequence $\{v_r\}$ is bounded, so by taking limits in the conditions above and using $\mu_r d_r \to 0$ and outer semicontinuity of $\partial h(c)$ at $\bar{c}$, we can identify a vector $\bar{v}$ such that $\bar{v} \in \partial h(\bar{c}) \cap \mathrm{Null}(\nabla c(\bar{x})^*)$. Using the chain rule and subdifferential regularity, this contradicts non-criticality of $\bar{x}$. □

The next result makes use of the efficient projection mechanism of Theorem 4.6. When the conditions of this theorem are satisfied, we show that Algorithm ProxDescent can perform the projection to obtain the point x_k^+ in such a way that (4.10) is satisfied.

Lemma 5.2 *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$ that is $\mathcal{C}^2$ around a point $\bar{x} \in \Re^n$. Assume that h lower semicontinuous and finite at $\bar{c} = c(\bar{x})$ and that the transversality condition (4.3) holds at $\bar{x}$ and $\bar{c}$. Then there exist constants $\tilde{\mu} > 0$ and $\tilde{\delta} > 0$ with the following property: For any $x \in B_{\tilde{\delta}}(\bar{x})$, $d \in B_{\tilde{\delta}}(0)$, and $\mu \geq \tilde{\mu}$ such that*

$$h_{x,\mu}(d) \leq h_{x,\mu}(0), \qquad |h(c(x) + \nabla c(x)d) - h\big(c(\bar{x})\big)| < \tilde{\delta}, \tag{5.1}$$

there is a point $x^+ \in \Re^n$ such that

$$h\big(c(x)\big) - h\big(c(x^+)\big) \geq \sigma\big[h\big(c(x)\big) - h\big(c(x) + \nabla c(x)d\big)\big], \tag{5.2a}$$
$$|x^+ - (x+d)| \leq \frac{1}{2}|d|. \tag{5.2b}$$

Proof Define δ and γ as in Theorem 4.6 and set $\tilde{\delta} = \min\big(\delta, 1/(2\gamma)\big)$. By applying Theorem 4.6, we obtain a point x^+ for which $|x^+ - (x+d)| \leq \gamma|d|^2 \leq \frac{1}{2}|d|$ (thus satisfying (5.2b)) and $h\big(c(x^+)\big) \leq h(c(x) + \nabla c(x)d) + \gamma|d|^2$. Also note that because of $h_{x,\mu}(d) \leq h_{x,\mu}(0)$, we have

$$h(c(x) + \nabla c(x)d) + \frac{\mu}{2}|d|^2 \leq h\big(c(x)\big)$$

and hence

$$|d|^2 \leq \frac{2}{\mu}\big[h\big(c(x)\big) - h(c(x) + \nabla c(x)d)\big].$$

We therefore have

$$\begin{aligned} h\big(c(x)\big) - h\big(c(x^+)\big) &\geq h\big(c(x)\big) - h(c(x) + \nabla c(x)d) - \gamma|d|^2 \\ &\geq \big[h\big(c(x)\big) - h(c(x) + \nabla c(x)d)\big]\left(1 - \frac{2\gamma}{\mu}\right). \end{aligned}$$

By choosing $\tilde{\mu}$ large enough that $1 - 2\gamma/\tilde{\mu} > \sigma$, we obtain (5.2a). □

We also need the following elementary lemma.

Lemma 5.3 *For any constants $\tau > 1$ and $\rho > 0$ and any positive integer t, we have*

$$\min\left\{\sum_{i=1}^{t} \alpha_i^2 \tau^i : \sum_{i=1}^{t} \alpha_i \geq \rho,\ \alpha \in \Re_+^t\right\} > \rho^2(\tau - 1).$$

Proof By scaling, we can suppose $\rho = 1$. Clearly the optimal solution of this problem must lie on the hyperplane $H = \{\alpha : \sum_i \alpha_i = 1\}$. The objective function is convex, and its gradient at the point $\bar{\alpha} \in H$ defined by

$$\bar{\alpha}_i = \frac{\tau^{1-i} - \tau^{-i}}{1 - \tau^{-t}} > 0$$

is easily checked to be orthogonal to H. Hence $\bar{\alpha}$ is optimal, and the corresponding optimal value is easily checked to be strictly larger than $\tau - 1$. □

For the main convergence result, we make the additional assumption that h can be bounded below, globally, by a (concave) quadratic function, that is,

$$h(c) \geq h_0 - q_0|c|^2 \quad \text{for all } c \in \Re^m, \tag{5.3}$$

for some scalars h_0 and $q_0 \geq 0$. Such functions are called *prox-bounded* [40]. This assumption holds for all h considered in the examples of Sect. 2. The other assumptions made on h, c, and $\bar{x}$ in the theorem below allow us to apply both Lemmas 5.1 and 5.2.

Theorem 5.4 (Global convergence) *Consider a function $h\colon \Re^m \to \bar{\Re}$ and a map $c\colon \Re^n \to \Re^m$. Suppose that the sequence $\big(x_k, h\big(c(x_k)\big)\big)$ generated by Algorithm ProxDescent has an accumulation point at $\big(\bar{x}, h(\bar{c})\big)$, where $\bar{c} := c(\bar{x})$. Suppose that c is $\mathcal{C}^2$ near $\bar{x}$, that h is subdifferentially regular (thus lower semicontinuous) at $\bar{c}$ and is prox-bounded, and that the transversality condition (4.3) holds at $\bar{x}$. Then the criticality condition (1.4) is satisfied at $\bar{x}$.*

Proof Suppose for contradiction that $\big(\bar{x}, h(\bar{c})\big)$ is an accumulation point but is not critical. Since the sequence $\{h\big(c(x_r)\big)\}$ generated by the algorithm is monotonically decreasing, we have $h\big(c(x_r)\big) \downarrow h(\bar{c})$. By the acceptance test in the algorithm and the definition of $h_{x,\mu}$ in (1.3), we have that

$$\begin{aligned} h\big(c(x_{r+1})\big) &\leq h\big(c(x_r)\big) - \sigma\big[h\big(c(x_r)\big) - h\big(c(x_r) + \nabla c(x_r)d_r\big)\big] \\ &\leq h\big(c(x_r)\big) - \sigma\frac{\mu_r}{2}|d_r|^2. \end{aligned} \tag{5.4}$$

We thus have

$$h\big(c(x_0)\big) - h\big(c(\bar{x})\big) \geq \sum_{r=0}^{\infty} h\big(c(x_r)\big) - h\big(c(x_{r+1})\big)$$
$$\geq \frac{\sigma}{2} \sum_{r=1}^{\infty} \mu_r |d_r|^2 \geq \frac{\sigma}{2} \mu_{\min} \sum_{r=1}^{\infty} |d_r|^2,$$

which implies that $d_r \to 0$. Further, we have that

$$\begin{aligned} &|h(c(x_r) + \nabla c(x_r) d_r) - h(\bar{c})| \\ &\quad \leq \big[h\big(c(x_r)\big) - h(c(x_r) + \nabla c(x_r) d_r)\big] + \big[h\big(c(x_r)\big) - h(\bar{c})\big] \\ &\quad \leq \sigma^{-1} \big[h\big(c(x_r)\big) - h\big(c(x_{r+1})\big)\big] + \big[h\big(c(x_r)\big) - h(\bar{c})\big] \to 0. \end{aligned} \tag{5.5}$$

Because $\bar{x}$ is an accumulation point, we can define a subsequence of indices r_j, $j = 0, 1, 2, \ldots$ such that $\lim_{j\to\infty} x_{r_j} = \bar{x}$. The corresponding sequence of regularization parameters μ_{r_j} must be unbounded, since Lemma 5.1 indicates that $\liminf_j \mu_{r_j} |d_{r_j}| \geq \epsilon > 0$. Defining $\tilde{\mu}$ and $\tilde{\delta}$ as in Lemma 5.2, we can assume without loss of generality that $\mu_{r_j} > \tau\tilde{\mu}$ and $\mu_{r_{j+1}} > \mu_{r_j}$ for all j. Moreover, since $x_{r_j} \to \bar{x}$ and $d_r \to 0$, and using (5.5), we can assume that

$$x_{r_j} \in B_{\tilde{\delta}/2}(\bar{x}), \qquad \text{for } j = 0, 1, 2, \ldots, \tag{5.6a}$$
$$d_r \in B_{\tilde{\delta}}(0), \qquad \text{for all } r > r_0, \tag{5.6b}$$
$$\big|h(c(x_r) + \nabla c(x_r) d_r) - h(\bar{c})\big| \leq \tilde{\delta}, \qquad \text{for all } r > r_0. \tag{5.6c}$$

Suppose first that there are infinitely many r_j, $j = 0, 1, 2, \ldots$, such that μ is increased in an inner iteration of iteration r_j. Without loss of generality, we can assume that this behavior happens for *all* r_j, $j = 0, 1, 2, \ldots$. We consider reasons why the previously tried value μ_{r_j}/τ would have been rejected. The first possible reason for rejection is that (1.3) does not have a local minimizer for $x = x_{r_j}$ and $\mu = \mu_{r_j}/\tau$. Because of (5.3), we have

$$\begin{aligned} &h\big(c(x_{r_j}) + \nabla c(x_{r_j}) d\big) + \frac{\mu_{r_j}}{2\tau} |d|^2 \\ &\quad \geq h_0 - q_0 \big|c(x_{r_j}) + \nabla c(x_{r_j}) d\big|^2 + \frac{\mu_{r_j}}{2\tau} |d|^2 \\ &\quad \geq \Big[h_0 - 2q_0 |c(x_{r_j})|^2\Big] + \Big[\frac{\mu_{r_j}}{2\tau} - 2q_0 |\nabla c(x_{r_j})|^2\Big] |d|^2 \\ &\quad \geq \bar{h}_0 + \Big(\frac{\mu_{r_j}}{2\tau} - \bar{q}_0\Big) |d|^2, \end{aligned}$$

where the last inequality follows from (5.6a) and smoothness of c. We conclude that $h_{x_{r_j}, \mu_{r_j}/\tau}$ has bounded level sets for all μ_{r_j} sufficiently large, thus by lower semicontinuity of h, it attains a minimizer [40, Theorem 1.9]. Note that $d = 0$ is not

the minimizer (otherwise, ProxDescent would have terminated), so at least one of the local minimizers that exists has $h_{x_{r_j},\mu_{r_j}/\tau}(d) < h_{x_{r_j},\mu_{r_j}/\tau}(0)$,

We can thus assume that a local minimizer $\hat{d}_{r_j}$ is found at $x = x_{r_j}$ and $\mu = \mu_{r_j}/\tau$. If $\lim_j \hat{d}_{r_j} = 0$, we have from $h(c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j}) \le h(c(x_{r_j})) \downarrow h(c(\bar{x}))$, the fact that $x_{r_j} \to \bar{x}$, and lower semicontinuity of h that $h(c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j}) \to h(c(\bar{x}))$. Thus, all conditions of Lemma 5.2 are satisfied, by $x = x_{r_j}$, $d = \hat{d}_{r_j}$, $\mu = \mu_{r_j}/\tau$, so it follows from this lemma that a step would have been taken and μ would *not* have been increased above μ_{r_j}/τ, a contradiction. Hence, we must have $\hat{d}_{r_j} \not\to 0$. We can therefore identify a constant $\hat{\epsilon} > 0$ and assume without loss that $|\hat{d}_{r_j}| \ge \hat{\epsilon}$ for all $j = 0, 1, 2, \ldots$. Since $h_{x_{r_j},\mu_{r_j}/\tau}(\hat{d}_{r_j}) < h_{x_{r_j},\mu_{r_j}/\tau}(0)$, we have

$$h(c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j}) < h(c(x_{r_j})) - \frac{\mu_{r_j}}{2\tau}|\hat{d}_{r_j}|^2. \tag{5.7}$$

Since $\mu_{r_j} \uparrow \infty$, this inequality contradicts prox-boundedness for all j sufficiently large. To see this, we have from (5.3) and $|\hat{d}_{r_j}| \ge \hat{\epsilon} > 0$ (for large j) that

$$\begin{aligned} h(c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j}) &\ge h_0 - q_0 \left| c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j} \right|^2 \\ &\ge \left[h_0 - 2q_0|c(x_{r_j})|^2 \right] - \left[2q_0|\nabla c(x_{r_j})|^2 \right] |\hat{d}_{r_j}|^2 \\ &> h(c(x_{r_j})) - \frac{\mu_{r_j}}{2\tau}|\hat{d}_{r_j}|^2 \quad \text{for all } j \text{ sufficiently large} \\ &> h(c(x_{r_j}) + \nabla c(x_{r_j})\hat{d}_{r_j}), \end{aligned}$$

giving a contradiction. We conclude that the case of $\hat{d}_{r_j} \not\to 0$ also cannot hold, so there are *not* infinitely many r_j, $j = 0, 1, 2, \ldots$ such that μ is increased during an internal iteration of major iteration r_j. In fact, we can claim, with no loss of generality, that μ is not increased in iteration r_j, so that the first value of μ tried in each iteration r_j, $j = 0, 1, 2, \ldots$ is accepted.

Let k_j, $j = 1, 2, \ldots$ denote the latest iteration prior to iteration r_j for which μ is increased in an internal iteration. Since $\mu_{r_j} > \mu_{r_{j-1}}$, the index k_j is well defined, and from the discussion above, we have $r_{j-1} < k_j < r_j$. Since no increases are performed internally during iterations $k_j + 1, \ldots, r_j$, the value μ_k at each of these steps is the first value tried, that is,

$$\tau\tilde{\mu} < \mu_{r_j} = \tau^{-1}\mu_{r_j-1} = \tau^{-2}\mu_{r_j-2} = \cdots = \tau^{k_j - r_j}\mu_{k_j}. \tag{5.8}$$

We show first that

$$x_{k_j} - x_{r_j} \to 0. \tag{5.9}$$

If this limit did not hold, there would be a value $\hat{\delta} > 0$ such that $|x_{k_j} - x_{r_j}| \ge \hat{\delta}$ for infinitely many j—without loss of generality for *all* j. From the acceptance criteria in Algorithm ProxDescent, we have

$$|x_{k+1} - x_k| \le |x_{k+1} - (x_k + d_k)| + |d_k| \le \frac{3}{2}|d_k|, \tag{5.10}$$

and so

$$\hat{\delta} \le |x_{r_j} - x_{k_j}| \le \sum_{k=k_j}^{r_j-1} |x_{k+1} - x_k| \le \frac{3}{2} \sum_{k=k_j}^{r_j-1} |d_k|.$$

To bound the decrease in objective function over the steps from x_{k_j} to x_{r_j}, we have from (5.4) and (5.8) that

$$\begin{aligned} h\big(c(x_{k_j})\big) - h\big(c(x_{r_j})\big) &= \sum_{k=k_j}^{r_j-1} h\big(c(x_k)\big) - h\big(c(x_{k+1})\big) \\ &\ge \frac{\sigma}{2} \sum_{k=k_j}^{r_j-1} \mu_k |d_k|^2 = \frac{\sigma}{2} \mu_{r_j} \sum_{k=k_j}^{r_j-1} \tau^{r_j-k} |d_k|^2. \end{aligned}$$

To obtain a lower bound on the final summation, we apply Lemma 5.3 with $\rho = 2\hat{\delta}/3$ (from (5.10)) and $t = r_j - k_j \ge 1$ to obtain

$$h\big(c(x_{k_j})\big) - h\big(c(x_{r_j})\big) \ge \frac{\sigma}{2} \mu_{r_j} (2\hat{\delta}/3)^2 (\tau - 1) \ge \frac{2}{9} \tau \tilde{\mu} \hat{\delta}^2 (\tau - 1) > 0,$$

where we have used $\mu_{r_j} > \tau\tilde{\mu}$. This inequality contradicts $h(c(x_r)) \downarrow h(c(\bar{x}))$, so we conclude that (5.9) holds. It thus follows from the definition of $\{r_j\}$ that

$$\lim_{j\to\infty} x_{k_j} = \lim_{j\to\infty} x_{r_j} + \lim_{j\to\infty} (x_{k_j} - x_{r_j}) = \bar{x}. \tag{5.11}$$

An identical argument to the one we used to show that μ cannot be increased at iteration r_j can now be applied to the sequence $\{k_j\}$, $j = 0, 1, 2, \ldots$, to show that the second-to-last value μ_{k_j}/τ tried at iteration k_j would be been accepted for all j sufficiently large. This contradicts the definition of k_j. Summarizing these arguments, we conclude that the sequence $\{r_j\}$ does not exist, so the desired contradiction is obtained, and $\bar{x}$ must be a critical point. □

We note that this global convergence result (stationarity of accumulation points) is typical of algorithms for nonlinear programming and composite nonsmooth optimization; see for example [18, Theorem 2.1], [52, Theorem 3.1].

To illustrate the idea of identification, we state a simple manifold identification result for the case when the function h is convex and finite.

Theorem 5.5 *Consider a function $h : \Re^m \to \Re$, a map $c : \Re^n \to \Re^m$, and a point $\bar{x} \in \Re^n$ such that c is C^2 near $\bar{x}$ and that the constraint qualification (4.11) and the strict criticality condition (4.20) both hold for the composite function $h \circ c$ at $\bar{x}$.*

Suppose too that h is convex, and continuous on $\operatorname{dom} h$ *near* $\bar{c} := c(\bar{x})$. *Suppose in addition that h is partly smooth at $\bar{c}$ relative to the manifold $\mathcal{M}$. Then if Algorithm ProxDescent generates a sequence $x_r \to \bar{x}$, we have that $c(x_r) + \nabla c(x_r) d_r \in \mathcal{M}$ for all r sufficiently large.*

Proof Note that h, c, and $\bar{x}$ satisfy the assumptions of Theorem 4.11, with $\hat{\mu} = 0$. To apply Theorem 4.11 and thus prove the result, we need to show only that $\mu_r |x_r - \bar{x}| \to 0$. In fact, we show that $\{\mu_r\}$ is bounded, so that this estimate is satisfied trivially.

Suppose for contradiction that $\{\mu_r\}$ is unbounded, so without loss of generality we can choose an infinite subsequence $\{r_j\}_{j=0,1,2,\ldots}$ with the following properties:

$$\mu_{r_j} \uparrow \infty, \tag{5.12a}$$

$$\lim_{j\to\infty} x_{r_j} = \bar{x}, \tag{5.12b}$$

$$\mu \text{ was increased at an internal iteration of iteration } r_j. \tag{5.12c}$$

Similarly to the proof of Theorem 5.4, we consider the reasons why the value μ_{r_j}/τ was rejected as a possible value for μ at iteration r_j. Let $\hat{d}_{r_j}$ be the value of d obtained by solving (1.3) with $x = x_{r_j}$ and $\mu = \mu_{r_j}/\tau$. If $\lim_{j\to\infty} \hat{d}_{r_j} = 0$, we have from (5.12a) and (5.12b) and continuity of h that the conditions of Lemma 5.2 are satisfied by $x = x_{r_j}$, $d = \hat{d}_{r_j}$, and $\mu = \mu_{r_j}/\tau$, for all j sufficiently large. This lemma implies that μ_{r_j}/τ would have been accepted at iteration r_j, a contradiction. We must therefore have $\hat{d}_{r_j} \not\to 0$, so may as well assume that we can identify $\hat{\epsilon} > 0$ such that $|d_{r_j}| \geq \hat{\epsilon}$ for all j sufficiently large. Since $h_{x_{r_j},\mu_{r_j}/\tau}(\hat{d}_{r_j}) < h_{x_{r_j},\mu_{r_j}/\tau}(0)$, inequality (5.7) holds. The assumptions on h imply that h is globally bounded below by a *linear* function (the supporting hyperplane at $c(\bar{x})$, for example), so as in the proof of Theorem 5.4, inequality (5.7) also leads to a contradiction. We conclude that $\{\mu_r\}$ is bounded, as claimed. □

To enhance the step d obtained from (1.3), we might try to incorporate second-order information inherent in the structure of the subdifferential ∂h at the new value of c predicted by the linearized subproblem. Knowledge of the subdifferential $\partial h\big(c(x) + \nabla c(x) d\big)$ allows us in principle to compute the tangent space to $\mathcal{M}$. We could then try to "track" $\mathcal{M}$ using second-order information, since both the map c and the restriction of the function h to $\mathcal{M}$ are $\mathcal{C}^2$.

6 Computational results

We present results for Algorithm ProxDescent applied to three problems drawn from the examples of Sect. 2. Our results are far from exhaustive; the wide range of applications of our framework make a comprehensive study impossible. Moreover, the algorithmic framework that we present and analyze is of a bare-bones nature. Significant improvements in efficiency could be gained by enhancing it in various ways (for example by making the strategy to increase and decrease μ more adaptive) and by customizing it to the various applications. Our goal here is to show that even the basic

ProxDescent algorithm gives good performance on a diverse set of applications. Two of our applications are regularized linear least-squares problems, one with a nonconvex regularizer. The other is a nonsmooth penalty function from a nonlinear programming application in power systems.

We start with the following ℓ_1-regularized least-squares problem:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \nu\|x\|_1, \tag{6.1}$$

where $\nu > 0$ is a regularization parameter. This problem has been widely studied in recent years in the context of compressed sensing [9] (where $m < n$) and LASSO [47] (where typically $m > n$). As mentioned earlier, Algorithm ProxDescent applied to this problem is closely related to the SpaRSA algorithm for compressed sensing; we refer to [51] for more detailed numerical testing.

We use ProxDescent to solve a compressed sensing signal recovery problem in which 51 components of a $n = 4096$-dimensional vector $\hat{x}$ were chosen to have nonzero values, and 256 random linear observations $A\hat{x}$ were made, where each entry of A is drawn i.i.d. from a normal distribution with mean zero and standard deviation $1/(2n)$. Random normal noise of mean 0 and standard deviation $10^{-4}/(2n)$ is added to each observation, to yield the vector b in (6.1). The nonzero values of $\hat{x}$ have a wide range of magnitudes. We choose the regularization parameter ν to be $.02\|A^T b\|_\infty$, which gives good recovery accuracy, and use $x = 0$ as the starting point. For the parameters in ProxDescent, we used $\tau = 1.25$, $\sigma = .01$, and $\mu_{\min} = 10^{-4}$. Termination was declared when the relative change in function value between two successive iterations dropped below 10^{-4}. (We note that because of the sufficient decrease condition in ProxDescent, this quantity dominates a multiple of the first-order predicted decrease in h, which quantity is zero at a stationary point.)

Results are shown in Fig. 1. ProxDescent runs for 92 iterations before declaring convergence. The top subfigure illustrates the solution $\hat{x}$ (with nonzero components shown as vertical bars) and the recovered solution x^* (indicated by circles), which has 25 nonzero components. Note that x^* appears to have captured all larger-magnitude components of $\hat{x}$ accurately. The middle figure plots log of objective function value against iteration number, showing apparent linear convergence. The bottom plot shows the log of μ_k plotted against iteration number k. This value shows a slow downward trend and is not constrained by the minimum value $\mu_{\min}$.

Consider now the linear least-squares problem with a component-wise MCP regularizer $\phi(\cdot)$ defined in (2.14):

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \nu \sum_{i=1}^{n} \phi(x_i), \tag{6.2}$$

where $\nu > 0$ is again the regularization parameter. In replacing the regularizer $\|\cdot\|_1$ of (6.1) with the nonconvex regularizer of (6.2), we reduce bias in the solution at the cost of introducing nonconvexity and thus the possibility of local minima.

To test ProxDescent on this problem, we used a different random instance of the same problem as in (6.1), with the same parameter settings. We define the parameters

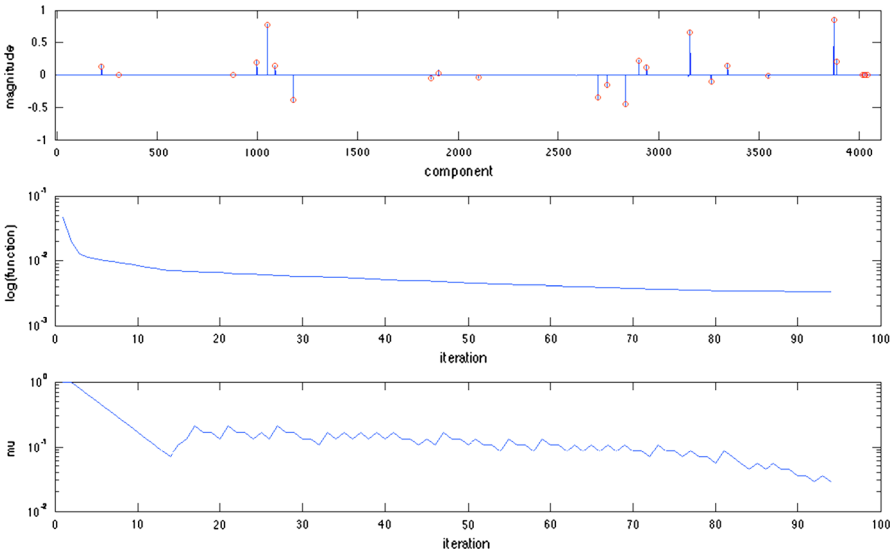


Fig. 1 Results for formulation (6.1). *Top figure* shows spikes in true signal (*bars*) and recovered spikes (*circles*), showing accurate recovery of the true signal. *Middle figure* shows the function values at each iteration, while the *bottom figure* shows the values of μ_k at each iteration k

of the MCP regularizer (2.14) to be $\lambda = 1$ and $a = \|\hat{x}\|_\infty/3$. These choices ensure that the MCP function has similar slope to $\|\cdot\|_1$ near zero and that it achieves its maximum value of $a\lambda^2/2$ for the larger spikes. Results are shown in Fig. 2, using the same format as Fig. 1 for the subfigures. Despite the nonconvexity, ProxDescent appears to have no trouble finding the global minimum of (6.2), and in a similar number of iterations as for (6.1) (84, in this particular instance). In fact, a close comparison of the top subfigures in Figs. 1 and 2 indicates that the recovered spikes in Fig. 1 have slightly lower magnitudes in general than the true spikes, an effect that is not present in Fig. 2. This effect is subtle for the choice of ν used here (detectable only in high-precision versions of the plots), but it illustrates nicely the unbiasedness property of the MCP regularizer. Once again, we see a faint downward trend in the value of μ_k, and a convergence rate that is clearly linear, until the solution is identified to high accuracy at about iteration 80.

Finally, we consider the following nonlinear optimization problem:

$$\min\ p^T x \quad \text{subject to} \quad c(x) = 0, \quad \underline{x} \le x \le \overline{x}, \tag{6.3}$$

where $c : \Re^n \to \Re^q$ is a smooth nonlinear vector function. A nonsmooth penalty formulation of this problem, stated in a form consistent with (1.1), is as follows:

$$\min\ p^T x + \nu\|c(x)\|_1 + I_{[\underline{x},\overline{x}]}(x), \tag{6.4}$$

where the indicator function $I_{[\underline{x},\overline{x}]}(x)$ takes the value 0 if the bound constraints $\underline{x} \le x \le \overline{x}$ are satisfied, and ∞ otherwise. This problem was considered in [26],

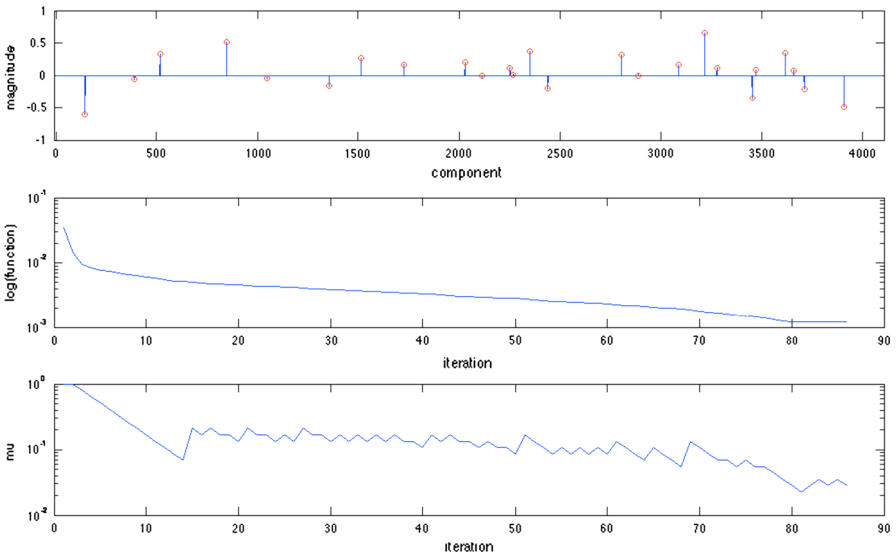


Fig. 2 Results for formulation (6.2). *Top figure* shows spikes in true signal (*bars*) and recovered spikes (*circles*), showing accurate recovery of the true signal. *Middle figure* shows the function values at each iteration, while the *bottom figure* shows the values of μ_k at each iteration k

where a sequential ℓ_1-linear programming algorithm was proposed to solve it. When ProxDescent is applied to (6.4), the only essential difference between it and the algorithm of [26] is that the latter uses an ℓ_∞ ("box-shaped") trust region on the step d in its subproblem, in place of the quadratic prox-term $(\mu/2)|d|^2$ of (1.3), which is equivalent to an ℓ_2-norm (circular) trust region. (In fact, the code used to obtain numerical results in [26] was easily modified to produce the results shown here.)

We use ProxDescent on the framework (6.4) to solve two problems from [26], arising from the restoration of stable operation of a power grid following a disruption, such as loss of a transmission line. In this application, the variables x represent voltage phasors at each node of the grid and various slacks in the formulation, while $c(x)$ is derived from the (nonlinear) model of AC power flow. The bounds on x represent acceptable deviations of voltage magnitude from 1, and acceptable values of the amount of load to be shed from the nodes of the grid. The first problem is of a type that commonly arises in the power grid application, where the number of constraints active at the solution of (6.3) equals the number of variables, so that methods that use linearization of the constraints (including ProxDescent and the algorithm of [26]) reduce to Newton's method on the system of nonlinear equations represented by the active constraints, and rapid convergence is observed once the active set has been determined correctly. In the second problem, the number of active constraints is fewer than the number of variables, so rapid convergence cannot be expected from a first-order method. Here, as in [26], convergence is considerably slower.

For both datasets, we set $\tau = 1.5$, $\sigma = 10^{-3}$, and $\mu_{\min} = 10^{-3}$ in ProxDescent, and terminate when the relative change in objective falls below 10^{-5}. Results for the first problem are shown in Fig. 3. This problem has 143 variables and 143 active

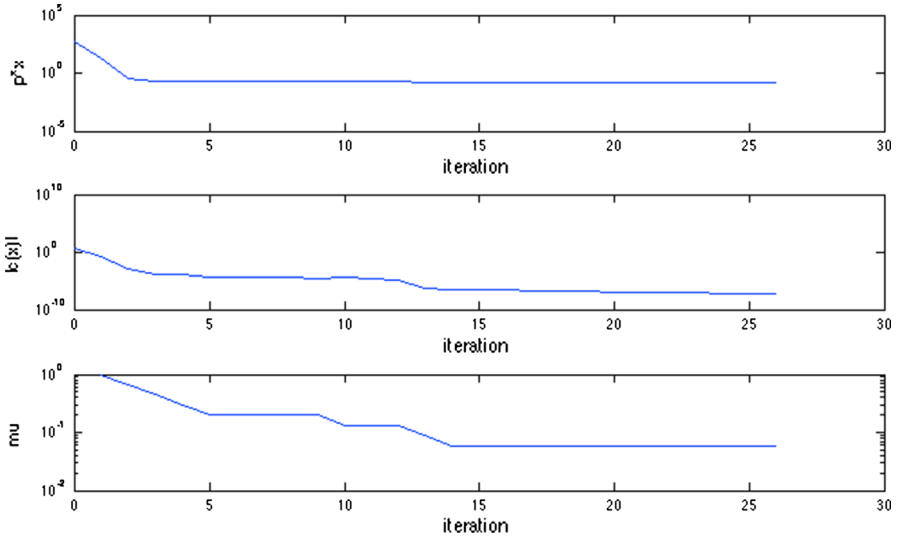


Fig. 3 Results for formulation (6.4), derived from a 57-bus power grid in which the number of active constraints at the solution equals the number of variables. *Top figure* shows $p^T x_k$ plotted against iteration k; *middle figure* show $\|c(x_k)\|_1$; *bottom figure* shows μ_k

constraints at the solution. Convergence occurred in 26 iterations, with a total of 44 subproblems solved. In Fig. 3, we consider separately the contributions from the $p^T x$ term and the penalty term $\|c(x)\|_1$. Both exhibit steady linear convergence to their optimal values. Two subproblems are solved on most iterations, because we try to decrease the value of μ then increase it again when the smaller value fails to satisfy the sufficient decrease test. Note that μ_k stabilizes at .058 on later iterations. Less than one second of execution time was required on a MacBook Pro (2 GHz Intel i7 with 8GB RAM), using Matlab, the MATPOWER package [55] for modeling and solving power grid problems, and CPLEX. The number of major iterations required was similar to the Sℓ_1LP algorithm described in [26]. We also coded a version of the algorithm that attempts to determine the set of active constraints manually once the active set appears to have settled down, solving a system of nonlinear equations based on the KKT conditions and making small heuristic adjustments to the active set in search of a stationary point. This version takes 21 iterations, and about half the run time.

The second data set for (6.4) is derived from a 118-bus system, and has 262 variables, and 260 constraints active at the solution. Convergence behavior is quite similar to the first case, featuring Q-linear convergence at a rate of about .5 for the constraint violation measure $\|c(x_k)\|_1$. μ_k stabilizes at the same value .058 as for the first data set, and convergence is declared after 21 iterations, with 34 subproblems solved, in about .6 seconds of CPU time (Fig. 4). An active-set version of the approach requires only 11 iterations and about .18 seconds of CPU time.

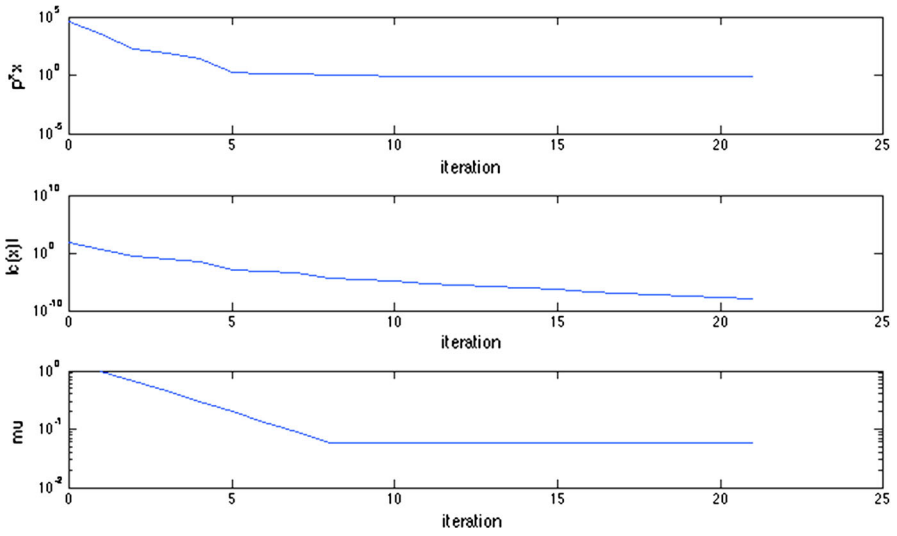


Fig. 4 Results for formulation (6.4), derived from a 118-bus power grid in which the number of active constraints at the solution is smaller than the number of variables. *Top figure* shows $p^T x_k$ plotted against iteration k; *middle figure* show $\|c(x_k)\|_1$; *bottom figure* shows μ_k

Acknowledgments We acknowledge the support of NSF Grants 0430504 and DMS-0806057. We are grateful for the comments of two referees, which were most helpful in revising earlier versions. We thank Mr. Taedong Kim for obtaining computational results for the formulation (6.4).

Appendix

The basic building block for variational analysis (see Rockafellar and Wets [40] or Mordukhovich [35]) is the *normal cone* to a (locally) closed set S at a point $s \in S$, denoted by $N_S(s)$. It consists of all *normal vectors*: limits of sequences of vectors of the form $\lambda(u - v)$ for points $u, v \in \Re^m$ approaching s such that v is a closest point to u in S, and scalars $\lambda > 0$. On the other hand, *tangent vectors* are limits of sequences of vectors of the form $\lambda(u - s)$ for points $u \in S$ approaching s and scalars $\lambda > 0$. The set S is *Clarke regular* at s when the inner product of any normal vector with any tangent vector is always nonpositive. Closed convex sets and smooth manifolds are everywhere Clarke regular.

The *epigraph* of a function $h : \Re^m \to \bar{\Re}$ is the set

$$\operatorname{epi} h = \{(c, r) \in \Re^m \times \Re : r \geq h(c)\}.$$

If the value of h is finite at some point $\bar{c} \in \Re^m$, then h is lower semicontinuous nearby if and only if its epigraph is locally closed around the point $\big(\bar{c}, h(\bar{c})\big)$. Henceforth we focus on that case.

The *subdifferential* of h at $\bar{c}$ is the set

$$\partial h(\bar{c}) = \big\{v \in \Re^m \,:\, (v, -1) \in N_{\operatorname{epi} h}\big(\bar{c}, h(\bar{c})\big)\big\}$$

and the *horizon subdifferential* is

$$\partial^\infty h(\bar{c}) = \left\{ v \in \Re^m : (v, 0) \in N_{\operatorname{epi} h}\left(\bar{c}, h(\bar{c})\right) \right\} \tag{6.5}$$

(see [40, Theorem 8.9]). The function h is *subdifferentially regular* at $\bar{c}$ if its epigraph is Clarke regular at $\left(\bar{c}, h(\bar{c})\right)$ (as holds in particular if h is convex lower semicontinuous, or smooth). Subdifferential regularity implies that $\partial h(\bar{c})$ is a closed and convex set in $\Re^m$, and its recession cone is exactly $\partial^\infty h(\bar{c})$ (see [40, Corollary 8.11]). In the case when h is locally Lipschitz, it is almost everywhere differentiable: h is then subdifferentially regular at $\bar{c}$ if and only if its directional derivative for every direction $d \in \Re^m$ equals

$$\limsup_{c \to \bar{c}} \langle \nabla h(c), d \rangle,$$

where the lim sup is taken over points c where h is differentiable.

Consider a *subgradient* $\bar{v} \in \partial h(\bar{c})$, and a *localization* of the subdifferential mapping ∂h around the point $(\bar{c}, \bar{v})$, by which we mean a set-valued mapping $T : \Re^m \rightrightarrows \Re^m$ defined by

$$T(y) = \begin{cases} \partial h(y) \cap B_\epsilon(\bar{v}) & (|y - \bar{c}| \le \epsilon,\ |h(y) - h(\bar{c})| \le \epsilon) \\ \emptyset & (\text{otherwise}) \end{cases}$$

for some constant $\epsilon > 0$. The function h is *prox-regular at $\bar{c}$ for $\bar{v}$* if some such localization is *hypomonotone*: that is, for some constant $\rho > 0$, we have

$$z \in T(y) \quad \text{and} \quad z' \in T(y') \Rightarrow \langle z' - z, y' - y \rangle \ge -\rho |y' - y|^2.$$

This definition is equivalent to Definition 1.1 (with the same constant ρ) [40, Example 12.28 and Theorem 13.36]. Prox-regularity at $\bar{c}$ (for all subgradients v) implies subdifferential regularity.

A general class of prox-regular functions common in engineering applications is "lower $\mathcal{C}^2$" functions [40, Definition 10.29]. A function $h : \Re^m \to \Re$ is *lower $\mathcal{C}^2$* around a point $\bar{c} \in \Re^m$ if h has the local representation

$$h(c) = \max_{t \in T} f(c, t) \quad \text{for } c \in \Re^m \text{ near } \bar{c},$$

for some function $f : \Re^m \times T \to \Re$, where the space T is compact and the quantities $f(c, t)$, $\nabla_c f(c, t)$, and $\nabla^2_{cc} f(c, t)$ all depend continuously on (c, t). All lower $\mathcal{C}^2$ functions are prox-regular [40, Proposition 13.3]. A simple equivalent property, useful in theory though harder to check in practice, is that h has the form $g - \kappa |\cdot|^2$ around the point $\bar{c}$ for some continuous convex function g and some constant κ.

The normal cone is crucial to the definition of another central variational-analytic tool. Given a set-valued mapping $F : \Re^p \rightrightarrows \Re^q$ with closed graph,

$$\operatorname{gph} F = \{(u, v) : v \in F(v)\},$$

at any point $(\bar{u}, \bar{v}) \in \operatorname{gph} F$, the *coderivative* $D^*F(\bar{u}|\bar{v}) : \Re^q \rightrightarrows \Re^p$ is defined by

$$w \in D^*F(\bar{u}|\bar{v})(y) \Leftrightarrow (w, -y) \in N_{\operatorname{gph} F}(\bar{u}, \bar{v}).$$

The coderivative generalizes the adjoint of the derivative of smooth vector function: for smooth $c : \Re^n \to \Re^m$, the set-valued mapping $x \mapsto F(x) := \{c(x)\}$ has coderivative given by $D^*F(x|c(x))(y) = \{\nabla c(x)^* y\}$ for all $x \in \Re^n$ and $y \in \Re^m$. As we see next, coderivative calculations drive two of the arguments in Sect. 4.1.

Proof of Corollary 4.3 Corresponding to any linear map $A : \Re^p \to \Re^q$, define a set-valued mapping $F_A : \Re^p \rightrightarrows \Re^q$ by $F_A(u) = Au - S$. A coderivative calculation shows, for vectors $v \in \Re^p$,

$$D^*F_A(0|0)(v) = \begin{cases} \{A^*v\} & \big(v \in N_S(0)\big) \\ \emptyset & (\text{otherwise}). \end{cases}$$

Hence, by assumption, the only vector $v \in \Re^p$ satisfying $0 \in D^*F_{\bar{A}}(0|0)(v)$ is zero, so by [40, Thm 9.43], the mapping $F_{\bar{A}}$ is metrically regular at zero for zero. Applying Theorem 4.2 shows that there exist constants $\delta, \gamma > 0$ such that, if $\|A - \bar{A}\| < \delta$ and $|v| < \delta$, then we have

$$\operatorname{dist}\big(0, F_A^{-1}(-v)\big) \leq \gamma \operatorname{dist}\big(-v, F_A(0)\big),$$

or equivalently,

$$\operatorname{dist}\big(0, A^{-1}(S - v)\big) \leq \gamma \operatorname{dist}(v, S).$$

Since $0 \in S$, the right-hand side is bounded above by $\gamma|v|$, so the result follows. □

Proof of Theorem 4.4 We simply need to check that the set-valued mapping $G : \Re^p \rightrightarrows \Re^q$ defined by $G(z) = F(z) - S$ is metrically regular at $\bar{z}$ for zero. Much the same coderivative calculation as in the proof of Corollary 4.3 shows, for vectors $v \in \Re^p$, the formula

$$D^*G(\bar{z}|0)(v) = \begin{cases} \{\nabla F(\bar{z})^* v\} & \big(v \in N_S(\bar{z})\big) \\ \emptyset & (\text{otherwise}). \end{cases}$$

Hence, by assumption, the only vector $v \in \Re^p$ satisfying $0 \in D^*G(\bar{z}|0)(v)$ is zero, so metric regularity follows by [40, Thm 9.43]. □

Alternative proof of Theorem 4.2 In the text we gave a short ad hoc proof of Theorem 4.2. Here we present a more formal approach. Denote the space of linear maps from $\Re^p$ to $\Re^q$ by $L(\Re^p, \Re^q)$, and define a mapping $g : L(\Re^p, \Re^q) \times \Re^p \to \Re^q$ and a parametric mapping $g_H : \Re^p \to \Re^q$ by $g(H, u) = g_H(u) = Hu$ for maps $H \in L(\Re^p, \Re^q)$ and points $u \in \Re^p$. Using the notation of [14, Section 3], the Lipschitz constant $l[g](0; \bar{u}, 0)$, is by definition the infimum of the constants ρ for which

the inequality

$$d\big(w, g_H(u)\big) \leq \rho d\big(u, g_H^{-1}(w)\big) \tag{6.6}$$

holds for all triples (u, w, H) sufficiently near the triple $(\bar{u}, 0, 0)$. Inequality (6.6) says simply

$$|w - Hu| \leq \rho |u - z| \quad \text{for all } z \in \Re^p \text{ satisfying } Hz = w,$$

a property that holds providing $\rho \geq \|H\|$. We deduce

$$l[g](0; \bar{u}, 0) = 0. \tag{6.7}$$

We can also consider $F + g$ as a set-valued mapping from $L(\Re^p, \Re^q) \times \Re^p$ to $\Re^q$, defined by $(F + g)(H, u) = F(u) + Hu$, and then the parametric mapping $(F + g)_H \colon \Re^p \rightrightarrows \Re^q$ is defined in the obvious way: in other words, $(F + g)_H(u) = F(u) + Hu$. According to [14, Theorem 2], Equation (6.7) implies the following relationship between the "covering rates" for F and $F + g$:

$$r[F + g](0; \bar{u}, \bar{v}) = r[F](\bar{u}, \bar{v}).$$

The reciprocal of the right-hand side is, by definition, the infimum of the constants $\kappa > 0$ such that inequality (4.1) holds for all pairs (u, v) sufficiently near the pair $(\bar{u}, \bar{v})$. By metric regularity, this number is strictly positive. On the other hand, the reciprocal of the left-hand side is, by definition, the infimum of the constants $\gamma > 0$ such that inequality (4.2) holds for all triples (u, v, H) sufficiently near the pair $(\bar{u}, \bar{v}, 0)$.

References

1. Bolte, J., Daniilidis, A., Lewis, A.S.: Generic optimality conditions for semialgebraic convex problems. Math. Oper. Res. **36**, 55–70 (2011)
2. Bonnans, J.F., Shapiro, A.: Perturbation Analysis of Optimization Problems. Springer Series in Operations Research. Springer, Berlin (2000)
3. Burke, J.V.: Descent methods for composite nondifferentiable optimization problems. Math. Program. Ser. A **33**, 260–279 (1985)
4. Burke, J.V.: On the identification of active constraints II: the nonconvex case. SIAM J. Numer. Anal. **27**, 1081–1102 (1990)
5. Burke, J.V., Moré, J.J.: On the identification of active constraints. SIAM J. Numer. Anal. **25**, 1197–1211 (1988)
6. Byrd, R., Gould, N.I.M., Nocedal, J., Waltz, R.A.: On the convergence of successive linear-quadratic programming algorithms. SIAM J. Optim. **16**, 471–489 (2005)
7. Cai, J.-F., Candès, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM J. Optim. **20**, 1956–1982 (2010)
8. Candès, E., Recht, B.: Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
9. Candès, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid (2006)
10. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)

11. Combettes, P., Pennanen, T.: Proximal methods for cohypomonotone operators. SIAM J. Control Optim. **43**, 731–742 (2004)
12. Combettes, P.L., Wajs, V.R.: Signal recovery by proximal forward–backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)
13. Daniilidis, A., Hare, W., Malick, J.: Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems. Optimization **55**, 481–503 (2006)
14. Dmitruk, A.V., Kruger, A.Y.: Metric regularity and systems of generalized equations. J. Math. Anal. Appl. **342**, 864–873 (2008)
15. Dontchev, A.L., Lewis, A.S., Rockafellar, R.T.: The radius of metric regularity. Trans. Am. Math. Soc. **355**, 493–517 (2003)
16. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**, 407–499 (2004)
17. Fan, J., Li, R.: Variable selection via nonconvex penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1361 (2001)
18. Fletcher, R., Sainz de la Maza, E.: Nonlinear programming and nonsmooth optimization by successive linear programming. Math. Program. **43**, 235–256 (1989)
19. Friedlander, M.P., Gould, N.I.M., Leyffer, S., Munson, T.S.: *A filter active-set trust-region method*, Preprint ANL/MCS-P1456-0907, Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne IL 60439, September 2007
20. Fukushima, M., Mine, H.: A generalized proximal point algorithm for certain nonconvex minimization problems. Int. J. Syst. Sci. **12**, 989–1000 (1981)
21. Hale, E.T., Yin, W., Zhang, Y.: A fixed-point continuation method for ℓ_1-minimization: methodology and convergence. SIAM J. Optim. **19**, 1107–1130 (2008)
22. Hare, W., Lewis, A.: Identifying active constraints via partial smoothness and prox-regularity. J. Convex Anal. **11**, 251–266 (2004)
23. Iusem, A., Pennanen, T., Svaiter, B.: Inexact variants of the proximal point algorithm without monotonicity. SIAM J. Optim. **13**, 1080–1097 (2003)
24. Jokar, S., Pfetsch, M.E.: Exact and approximate sparse solutions of underdetermined linear equations. SIAM J. Sci. Comput. **31**, 23–44 (2008)
25. Kaplan, A., Tichatschke, R.: Proximal point methods and nonconvex optimization. J. Glob. Optim. **13**, 389–406 (1998)
26. Kim, T., Wright, S.J.: An Sℓ_1LP-active set approach for feasibility restoration in power systems, tech. rep., Computer Science Department, University of Wisconsin-Madison, May 2014. arXiv:1405.0322
27. Lan, G.: Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization. Math. Program. Ser. A **149**, 1–45 (2015)
28. Lemaréchal, C., Oustry, F., Sagastizábal, C.: The $\mathcal{U}$-Lagrangian of a convex function. Trans. Am. Math. Soc. **352**, 711–729 (2000)
29. Levy, A.: Lipschitzian multifunctions and a Lipschitzian inverse mapping theorem. Math. Oper. Res. **26**, 105–118 (2001)
30. Lewis, A.: Active sets, nonsmoothness, and sensitivity. SIAM J. Optim. **13**, 702–725 (2003)
31. Mangasarian, O.L.: Minimum-support solutions of polyhedral concave programs. Optimization **45**, 149–162 (1999)
32. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. Rev. Française Informat. Recherche Opérationnelle **4**, 154–158 (1970)
33. Mifflin, R., Sagastizábal, C.: A VU-algorithm for convex minimization. Math. Program. Ser. B **104**, 583–608 (2005)
34. Miller, S.A., Malick, J.: Newton methods for nonsmooth convex minimization: connections among $\mathcal{U}$-Lagrangian, Reimannian Newton, and SQP methods. Math. Program. Ser. B **104**, 609–633 (2005)
35. Mordukhovich, B.: Variational Analysis and Generalized Differentiation, I: Basic Theory; II: Applications. Springer, New York (2006)
36. Pennanen, T.: Local convergence of the proximal point algorithm and multiplier methods without monotonicity. Math. Oper. Res. **27**, 170–191 (2002)
37. Recht, B., Fazel, M., Parrilo, P.: Guaranteed minimum-rank solutions of matrix equations via nuclear norm minimization. SIAM Rev. **52**, 471–501 (2010)
38. Rockafellar, R.: Monotone operators and the proximal point algorithm. SIAM J. Control Optim. **14**, 877–898 (1976)
39. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)

40. Rockafellar, R.T., Wets, R.J.: Variational Analysis. Springer, Berlin (1998)
41. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Phys. D **60**, 259–268 (1992)
42. Sagastizábal, C.: Composite proximal bundle method. Math. Program. Ser. B **140**, 189–233 (2013)
43. Sagastizábal, C., Mifflin, R.: Proximal points are on the fast track. J. Convex Anal. **9**, 563–579 (2002)
44. Shapiro, A.: On a class of nonsmooth composite functions. Math. Oper. Res. **28**, 677–692 (2003)
45. Shi, W., Wahba, G., Wright, S.J., Lee, K., Klein, R., Klein, B.: LASSO-Patternsearch algorithm with application to opthalmology data. Stat. Interface **1**, 137–153 (2008)
46. Spingarn, J.: Submonotone mappings and the proximal point algorithm. Numer. Funct. Anal. Optim. **4**, 123–150 (1981/82)
47. Tibshirani, R.: Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. B **58**, 267–288 (1996)
48. Wen, Z., Yin, W., Zhang, H., Goldfarb, D.: On the convergence of an active set method for ℓ_1 minimization. SIAM J. Sci. Comput. **32**, 1832–1857 (2010)
49. Wright, S.J.: Convergence of an inexact algorithm for composite nonsmooth optimization. IMA J. Numer. Anal. **9**, 299–321 (1990)
50. Wright, S.J.: Identifiable surfaces in constrained optimization. SIAM J. Control Optim. **31**, 1063–1079 (1993)
51. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Trans. Signal Process. **57**, 2479–2493 (2009)
52. Yuan, Y.: Conditions for convergence of a trust-region method for nonsmooth optimization. Math. Program. **31**, 220–228 (1985)
53. Yuan, Y.: On the superlinear convergence of a trust region algorithm for nonsmooth optimization. Math. Program. **31**, 269–285 (1985)
54. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894–942 (2010)
55. Zimmerman, R.D., Murillo-Sánchez, C.E., Thomas, R.J.: MATPOWER: steady-state operations, planning, and analysis tools for power systems research and education. IEEE Trans. Power Syst. **26**, 12–19 (2011)