A PROXIMAL QUASI-NEWTON TRUST-REGION METHOD FOR NONSMOOTH REGULARIZED OPTIMIZATION*

ALEKSANDR Y. ARAVKIN†, ROBERT BARALDI‡, AND DOMINIQUE ORBAN§

Abstract. We develop a trust-region method for minimizing the sum of a smooth term f and a nonsmooth term h, both of which can be nonconvex. Each iteration of our method minimizes a possibly nonconvex model of $f + h$ in a trust region. The model coincides with $f + h$ in value and subdifferential at the center. We establish global convergence to a first-order stationary point when f satisfies a smoothness condition that holds, in particular, when it has Lipschitz-continuous gradient, and h is proper and lower semi-continuous. The model of h is required to be proper, lower-semi-continuous and prox-bounded. Under these weak assumptions, we establish a worst-case $O(1/\epsilon^2)$ iteration complexity bound that matches the best known complexity bound of standard trust-region methods for smooth optimization. We detail a special instance, named TR-PG, in which we use a limited-memory quasi-Newton model of f and compute a step with the proximal gradient method, resulting in a practical proximal quasi-Newton method. We establish similar convergence properties and complexity bound for a quadratic regularization variant, named R2, and provide an interpretation as a proximal gradient method with adaptive step size for nonconvex problems. R2 may also be used to compute steps inside the trust-region method, resulting in an implementation named TR-R2. We describe our Julia implementations and report numerical results on inverse problems from sparse optimization and signal processing. Both TR-PG and TR-R2 exhibit promising performance and compare favorably with two linesearch proximal quasi-Newton methods based on convex models.

Key words. Nonsmooth optimization, nonconvex optimization, composite optimization, trust-region methods, quasi-Newton methods, proximal gradient method, proximal quasi-Newton method.

AMS subject classifications. 49J52, 65K10, 90C53, 90C56

1. Introduction.

We consider the problem class

$$\underset{x}{\text{minimize}} \;\; f(x) + h(x), \tag{1.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is proper and lower semi-continuous, and both may be nonconvex. Smooth and nonsmooth optimization problems are special cases corresponding to $h := 0$ and $f := 0$, respectively. Certain authors [9, 23] refer to (1.1) as a *composite* problem. We use instead the term *nonsmooth regularized* to differentiate with problems where $f = 0$ and $h(x) = g(c(x))$, where g is nonsmooth and c is smooth, which is indeed the composition of two functions.

*Disclaimer: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

†Department of Applied Mathematics, University of Washington, Seattle WA., USA. E-mail: saravkin@uw.edu. Research partially supported by the Washington Research Foundation.

‡Department of Applied Mathematics, University of Washington, Seattle WA., USA. E-mail: rbaraldi@uw.edu. This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-FG02-97ER25308.

§GERAD and Department of Mathematics and Industrial Engineering, Polytechnique Montréal, QC, Canada. E-mail: dominique.orban@gerad.ca. Research partially supported by an NSERC Discovery Grant.

In practice, h is often a regularizer designed to promote desirable properties in solutions, such as sparsity. The class (1.1) captures the natural structure of a wide range of problems; problems with simple constraints, exact penalty formulations, basis selection problems with both convex [40, 41] and nonconvex [2, 6, 46] regularization, and more general inverse and learning problems [1, 7, 10].

We describe a trust-region method for (1.1) in which steps are computed by approximately minimizing simpler nonsmooth iteration-dependent models inside a trust region defined by an arbitrary norm. In practice, the norm is chosen based on the nonsmooth term in the model and the tractability of the step-finding subproblem, which is not required to be convex. Our analysis hinges on the observation that in the nonsmooth context, the first step of the proximal gradient method is the right generalization of the gradient in smooth optimization. We establish global convergence in terms of an optimality measure describing the decrease achievable in the model by a single step of the proximal gradient method inside the trust-region. We also establish a worst-case complexity bound of $O(1/\epsilon^2)$ iterations to bring this optimality measure below a tolerance $0 < \epsilon < 1$. Others [9, 20] have observed that it is possible to devise trust-region methods for regularized optimization with complexity equivalent to that for smooth optimization. However, past research typically assumes that h is either globally Lipschitz continuous and/or convex.

We also revisit a quadratic regularization method, and establish similar convergence properties and same worst-case compexity under the same assumptions. Our description highlights the connection between the quadratic regularization method and the standard proximal gradient method. The former may be seen as an implementation of the latter with adaptive step size.

We provide implementation details and illustrate the performance of an instance where the trust-region model is the sum of a limited-memory quasi-Newton, possibly nonconvex, approximation of f with a nonsmooth model of h and various choices of the trust-region norm. Our trust-region algorithm exhibits promising performance and compares favorably with linesearch proximal quasi-Newton methods based on convex models [38, 39]. Our open source implementations are available from `github.com/UW-AMO/TRNC` as packages in the emerging Julia programing language [5].

As far as we can tell from the literature, the method described in the present paper is the first trust-region method for the fully nonconvex nonsmooth regularized problem. Our approach offers flexibility in the choice of the norm used to define the trust-region, provided an efficient procedure is known to solve the subproblem. We show that such procedures are easily obtained in a number of applied scenarios.

Related research. We focus on (1.1) and do not provide an extensive review of approaches for smooth optimization. Conn et al. [11] cover trust-region methods for smooth optimization thoroughly, as well as a number of select generalizations, and we refer the reader to their comprehensive treatment for background.

Yuan [43] formulates conditions for convergence of trust-region methods for convex-composite objectives, i.e., $g(c(x))$ where c is continuously differentiable and g is convex. In particular, he considers models of the form $s \mapsto g(c(x) + \nabla c(x)s)$, that are relevant to exact penalty methods for constrained optimization, and that are a special case of the models we consider.

Dennis et al. [14] develop convergence properties of trust-region methods for the case where $f = 0$ and h is Lipschitz continuous. Their analysis is based on a generalization of the concept of Cauchy point in terms of Clarke directional derivatives, but they do not provide an approach to solve the typically nonsmooth subproblem.

Kim et al. [22] analyze a trust-region method for (1.1) when f is convex and h is continuous and convex with assumptions based on those of Dennis et al. [14]. Their model around a current x has the form $f(x) + \nabla f(x)^T s + \frac{1}{2}\alpha\|s\|^2 + h(x+s)$, where α is a Barzilai-Borwein step length safeguarded to stay sufficiently positive and bounded. By contrast, our approach allows general quadratic models, possibly indefinite, and explicitly accounts for the trust-region constraint in the subproblem by devising specialized proximal operators.

Qi and Sun [33] propose a trust-region method inspired by that of Dennis et al. [14] for the case where $f = 0$ and h is locally Lipschitz continuous with bounded level sets. They establish convergence under the further assumption that the models are [0, 1]-subhomogeneous. Martínez and Moretti [27] employ similar assumptions to generalize the approach to problems with linear constraints.

Cartis et al. [9] consider (1.1) where h is convex and globally Lipschitz continuous. They analyze both a trust-region algorithm and a quadratic regularization variant, develop convergence and iteration complexity results, but do not provide guidance on how to compute steps in practice. Their analysis revolves around properties of a stationarity measure that are strongly anchored to the convexity assumption. The algorithms that we develop below are most similar to theirs but rest upon significantly weaker assumptions and concrete subproblem solvers. Grapiglia et al. [20] detail a unified convergence theory for smooth optimization that has trust-region methods as a special case. They also generalize the results of [9] but focus on objectives of the form $f(x) + g(c(x))$ where f and c are smooth and g is convex and globally Lipschitz.

Lee et al. [23] fully explore the global and fast local convergence properties of exact and inexact proximal Newton and quasi-Newton methods for the case where both f and h are convex. They show that those methods inherit all the desired properties of their counterparts in smooth optimization.

Bolte et al. [7] present a proximal alternating method for objectives of the form $g(x) + Q(x, y) + h(y)$ where g and h are proper and lower semi-continuous and the coupling function Q is continuously differentiable. Their setting has (1.1) as a special case. They establish convergence under the Kurdyka-Łojasiewicz assumption and provide a general recipe for algorithmic convergence under such an assumption.

Li and Lin [24] consider monotone and non-monotone accelerations of the proximal gradient method for possibly nonconvex f and h. They establish global convergence under the assumptions that f has a Lipschitz continuous gradient, h is proper and lower semi-continuous, and that $f + h$ is coercive. This leads to a sublinear iteration complexity bound when a Kurdyka-Łojasiewicz condition holds. Boţ et al. [8] employ an inertial acceleration strategy which converges under the assumptions that h is bounded below and possesses a Kurdyka-Łojasiewicz condition.

Stella et al. [38] initially devised PANOC, a linesearch quasi-Newton method for (1.1) with limited-memory BFGS Hessian approximations, for model predictive control. PANOC assumes that the objective has the form $f(x)+h_1(x)+h_2(c(x))$, where f and c are smooth, h_1 is nonsmooth and may be nonconvex, and h_2 is nonsmooth and convex. Themelis et al. [39] develop ZeroFPR, a nonmonotone linesearch proximal quasi-Newton method for (1.1) based on the concept of forward-backward envelope. ZeroFPR converges under a Kurdyka-Łojasiewicz assumption and enjoys the fast local convergence properties of quasi-Newton methods for smooth optimization when a Dennis-Moré condition holds.

Notation. Sets are represented by calligraphic letters. The cardinality of set $\mathcal{S}$ is represented by $|\mathcal{S}|$. We use $\|\cdot\|$ to denote a generic norm on $\mathbb{R}^n$. The symbols ν,

λ, σ and Δ are scalars. $\mathbb{B}(0,\Delta)$ is the ball centered at 0 with radius $\Delta > 0$ defined by a norm that should be clear from the context. We use the shorthands $\mathbb{B} = \mathbb{B}(0,1)$ and $\Delta\mathbb{B} = \mathbb{B}(0,\Delta)$. When necessary, we write $\mathbb{B}_p$ to indicate that the ℓ_p-norm is used. Functional symbols f, g, h, as well as ϕ, φ and ψ are used for functions. $\chi(\cdot;A)$ represents the indicator function of $A \subseteq \mathbb{R}^n$. In particular, the indicator of $\mathbb{B}(0,\Delta)$ is denoted $\chi(\cdot;\Delta\mathbb{B})$ or just $\chi(\cdot;\Delta)$ when the norm is clear from the context. We use the alternative notation $\chi(\cdot;\Delta\mathbb{B}_p)$ to emphasize that the ℓ_p-norm is used to define the ball. If $A \subseteq \mathbb{R}^n$ and $x \in \mathbb{R}^n$, $\operatorname{dist}(x;A) = \inf\{\|a - x\| \mid a \in A\}$ is the Euclidean distance from x to A. If A is closed and convex, $\operatorname{proj}_A(x)$ denotes the unique projection of x into A, i.e., $\{\operatorname{proj}_A(x)\} = \operatorname{argmin}\{\|a - x\| \mid a \in A\}$. Finally, j and k are iteration counters.

Roadmap. The paper proceeds as follows. In section 2, we gather preliminary concepts for trust-region methods and variational analysis used in the theory. section 3 develops the general trust-region method for (1.1), including the new Algorithm 3.1, and introduces several innovations that yield the main results. In section 4, we explain how to compute a trust-region step based on a proximal quasi-Newton model. New relevant proximal operators needed to implement the trust-region method are studied in section 5. A quadratic regularization variant of the trust-region algorithm together with its convergence analysis are presented in section 6. Numerical results and experiments are in section 7. We end with a brief discussion in section 8.

2. Preliminaries.

2.1. Smooth context. When $f \in \mathcal{C}^1$ and $h = 0$ in (1.1), trust-region methods are known for strong convergence properties and favorable numerical performance on both small and large-scale problems. At an iterate x_k, they compute a step s_k as an approximate solution of

$$\underset{s}{\text{minimize}}\ m_k(s;x_k) \quad \text{subject to}\ \|s\| \le \Delta_k,$$

where $m_k(\cdot;x_k)$ is a model of f about x_k, $\|\cdot\|$ is a norm and $\Delta_k > 0$ is the trust-region radius. The predicted decrease $m_k(0;x_k) - m_k(s_k;x_k)$ is compared to the actual decrease $(f+h)(x_k) - (f+h)(x_k+s_k)$ to decide whether s_k should be accepted or rejected. If s_k is accepted, the iteration is *successful*; otherwise it is *unsuccessful*. Typically, $m_k(\cdot;x_k)$ is a quadratic expansion of f about x_k and the Euclidean norm is used in the trust region. The Euclidean norm is favored because efficient numerical schemes are known for the quadratic subproblem, which can be solved either exactly by way of the method of Moré and Sorensen [28] or approximately by way of the truncated conjugate gradient method of Steihaug [37]. See [11] for more information.

2.2. Nonsmooth context. We denote $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$. We call $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ *proper* if $h(x) > -\infty$ for all x and $h(x) < \infty$ for at least one x, and *lower semi-continuous*, or *lsc*, at $\bar{x}$ if $\liminf_{x\to\bar{x}} h(x) = h(\bar{x})$. We say that h is *(lower-)level bounded* if all its level sets are bounded. If h is proper, lsc and level bounded, then $\operatorname{argmin} h$ is nonempty and compact [36, Theorem 1.9].

DEFINITION 2.1. *For a proper lsc function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ and a parameter $\nu > 0$, the* Moreau envelope $e_{\nu h}$ *and the* proximal mapping $\operatorname{prox}_{\nu h}$ *are defined by*

$$e_{\nu h}(x) := \inf_w \tfrac{1}{2}\nu^{-1}\|w - x\|^2 + h(w) = \nu^{-1} \inf_w \tfrac{1}{2}\|w - x\|^2 + \nu h(w), \tag{2.1a}$$

$$\operatorname*{prox}_{\nu h}(x) := \operatorname*{argmin}_w \tfrac{1}{2}\nu^{-1}\|w - x\|^2 + h(w) = \operatorname*{argmin}_w \tfrac{1}{2}\|w - x\|^2 + \nu h(w). \tag{2.1b}$$

Under certain assumptions, including strong convexity of the objective of (2.1b), the set $\text{prox}_{\nu h}(x)$ is a singleton. However, in general, the set-valued mapping $\text{prox}_{\nu h}$ may be empty or contain multiple elements. For a given h, the range of parameter values for which the Moreau envelope assumes a finite value is given by the following definition.

DEFINITION 2.2. *The proper lsc function $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ is* prox-bounded *if there exists $\nu > 0$ and at least one $x \in \mathbb{R}^n$ such that $e_{\nu h}(x) > -\infty$. The* threshold of prox-boundedness *ν_h of h is the supremum of all such $\nu > 0$.*

If h is level bounded, then so is $w \mapsto \frac{1}{2}\nu^{-1}\|w - x\|^2 + h(w)$ for all $x \in \mathbb{R}^n$ and all $\nu > 0$, so $e_{\nu h}(x) > -\infty$ [36, Theorem 1.9] and h is prox-bounded. The following result summarizes some properties of (2.1a)–(2.1b). Further properties appear in [36, Theorem 1.25].

PROPOSITION 2.3. *Let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be proper lsc and prox-bounded with threshold $\nu_h > 0$. For every $\nu \in (0, \nu_h)$ and all $x \in \mathbb{R}^n$,*

1. *$\text{prox}_{\nu h}(x)$ is nonempty and compact;*
2. *$e_{\nu h}(x)$ depends continuously on (ν, x) and $e_{\nu h}(x) \nearrow h(x)$ as $\nu \searrow 0$.*

2.3. Optimality conditions. We use the following notions of subgradient and subdifferential [36, Definition 8.3].

DEFINITION 2.4 (Limiting subdifferential). *Consider $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ and $\bar{x} \in \mathbb{R}^n$ with $\phi(\bar{x}) < \infty$. We say that $v \in \mathbb{R}^n$ is a* regular subgradient *of ϕ at $\bar{x}$, and we write $v \in \hat{\partial}\phi(\bar{x})$ if*

$$\liminf_{x \to \bar{x}} \frac{\phi(x) - \phi(\bar{x}) - v^T(x - \bar{x})}{\|x - \bar{x}\|} \geq 0.$$

The set of regular subgradients is also called the Fréchet subdifferential. *We say that v is a* general subgradient *of ϕ at $\bar{x}$, and we write $v \in \partial\phi(\bar{x})$, if there are sequences $\{x_k\}$ and $\{v_k\}$ such that $x_k \to \bar{x}$, $\phi(x_k) \to \phi(\bar{x})$, $v_k \in \hat{\partial}\phi(x^k)$ and $v^k \to v$. The set of general subgradients is called the* limiting subdifferential.

If ϕ is convex, the Fréchet and limiting subdifferentials coincide with the subdifferential of convex analysis. If ϕ is differentiable at x, $\partial\phi(x) = \{\nabla\phi(x)\}$ and if ϕ is continuously differentiable at x, $\hat{\partial}\phi(x) = \{\nabla\phi(x)\}$ [36, Section 8.8].

In the following, we do not make use of the precise definition of the relevant subdifferential, but merely rely on the following criticality property.

PROPOSITION 2.5 (36, Theorem 10.1). *If $\phi : \mathbb{R}^n \to \overline{\mathbb{R}}$ is proper and has a local minimum at $\bar{x}$, then $0 \in \hat{\partial}\phi(\bar{x}) \subseteq \partial\phi(\bar{x})$. If ϕ is convex, the latter condition is also sufficient for $\bar{x}$ to be a global minimum. If $\phi = f + h$ where f is continuously differentiable on a neighborhood of $\bar{x}$ and h is finite at $\bar{x}$, then $\partial\phi(\bar{x}) = \nabla f(\bar{x}) + \partial h(\bar{x})$.*

2.4. The proximal gradient method. Consider the generic nonsmooth regularized problem

$$\underset{s}{\text{minimize}} \;\; \varphi(s) + \psi(s), \tag{2.2}$$

where φ is continuously differentiable and ψ is proper, lower semi-continuous and prox-bounded. The notation φ and ψ is intentionally different from (1.1) and will be reused to denote models of f and h in section 3.

A natural method to solve (2.2) that generalizes the gradient method of smooth optimization is the *proximal gradient method* [3, 25]. When initialized from $s_0 \in \mathbb{R}^n$

where ψ is finite, it generates iterates according to

$$s_{j+1} \in \operatorname*{prox}_{\nu\psi}(s_j - \nu\nabla\varphi(s_j)), \quad j \geq 0, \tag{2.3}$$

where $\nu > 0$ is a step size. If ψ is the indicator of a closed convex set, the proximal gradient method reduces to the projected gradient method.

The first-order optimality conditions of (2.3) are

$$0 \in s_{j+1} - s_j + \nu\nabla\varphi(s_j) + \nu\partial\psi(s_{j+1}). \tag{2.4}$$

The proximal literature primarily focuses on the generalized gradient

$$G_\nu(s) := \nu^{-1}(s - \operatorname*{prox}_{\nu\psi}(s - \nu\nabla\varphi(s))), \tag{2.5}$$

with $G_\nu(0) = \nabla\varphi(0)$ in the case of smooth optimization. The following result gives conditions under which the proximal gradient method is monotonic.

PROPOSITION 2.6 (7, Lemma 2). *Let φ be continuously differentiable, $\nabla\varphi$ be Lipschitz continuous with constant $L > 0$ and ψ be proper, lsc and bounded below. For any $0 < \nu < 1/L$, any s_0 where ψ is finite, the iteration (2.3) is such that*

$$(\varphi + \psi)(s_{j+1}) \leq (\varphi + \psi)(s_j) - \tfrac{1}{2}(\nu^{-1} - L)\|s_{j+1} - s_j\|^2, \quad j \geq 0.$$

It is possible to remove the assumption that ψ is bounded below from Proposition 2.6 and replace it with the weaker assumption that ψ is prox-bounded and that ν is chosen smaller than the threshold of prox-boundedness of ψ.

In the smooth case, where $\psi = 0$, we have $s_1 = -\nu\nabla\varphi(s_0)$ and the decrease is

$$\varphi(s_1; x) \leq \varphi(s_0) - \tfrac{1}{2}\nu^2(\nu^{-1} - L)\|\nabla\varphi(s_0)\|^2. \tag{2.6}$$

3. Trust-region methods for nonsmooth regularized optimization. In this section, we develop and analyze a general trust-region method for (1.1). Section 3.1 examines properties of trust-region subproblems. Section 3.2 discusses optimality measures, and highlights the role of the prox-gradient step in quantifying descent in the general context of (1.1). In Section 3.3, we present the trust-region approach, and highlight key innovations that make it possible to obtain the convergence results and complexity analysis presented in Section 3.4.

3.1. Properties of trust-region subproblems. For fixed $x \in \mathbb{R}^n$, consider the parametric problem and its optimal set

$$p(\Delta; x) := \operatorname*{minimize}_s \ \varphi(s; x) + \psi(s; x) + \chi(s; \Delta), \tag{3.1a}$$

$$P(\Delta; x) := \operatorname*{argmin}_s \ \varphi(s; x) + \psi(s; x) + \chi(s; \Delta), \tag{3.1b}$$

where $\varphi(s; x) \approx f(x + s)$, $\psi(s; x) \approx h(x + s)$, $\chi(s; \Delta)$ is the indicator function of the trust region $\Delta\mathbb{B}$ and $\Delta > 0$. The form of (3.1) is representative of a trust-region subproblem for (1.1) in which f and h are modeled separately and the trust-region constraint appears implicitly via an indicator function.

We make the following additional assumption.

MODEL ASSUMPTION 3.1. *For any $x \in \mathbb{R}^n$, $\varphi(\cdot; x)$ is continuously differentiable, $\psi(\cdot; x)$ is proper and lsc.*

By Proposition 2.5,

$$s \in P(\Delta; x) \quad \Longrightarrow \quad 0 \in \nabla\varphi(s; x) + \partial(\psi(\cdot; x) + \chi(\cdot; \Delta))(s).$$

The following result summarizes properties of (3.1).

PROPOSITION 3.1. *Let Model Assumption 3.1 be satisfied. If we define* $p(0; x) := \varphi(0; x) + \psi(0; x)$ *and* $P(0; x) = \{0\}$*, the domain of* $p(\cdot; x)$ *and* $P(\cdot; x)$ *is* $\{\Delta \mid \Delta \geq 0\}$*. In addition,*

1. $p(\cdot; x)$ *is proper lsc and for each* $\Delta \geq 0$*,* $P(\Delta; x)$ *is nonempty and compact;*
2. *if* $\{\Delta_k\} \to \bar{\Delta} \geq 0$ *in such a way that* $\{p(\Delta_k; x)\} \to p(\bar{\Delta}; x)$*, and for each* k*,* $s_k \in P(\Delta_k; x)$*, then* $\{s_k\}$ *is bounded and all its limit points are in* $P(\bar{\Delta}; x)$*;*
3. *if* $\varphi(\cdot; x) + \psi(\cdot; x)$ *is strictly convex,* $P(\Delta; x)$ *is single-valued;*
4. *if* $\bar{\Delta} > 0$ *and there exists* $\bar{s} \in P(\bar{\Delta}; x)$ *such that* $\|\bar{s}\| < \bar{\Delta}$*, then* $p(\cdot; x)$ *is continuous at* $\bar{\Delta}$ *and* $\{p(\Delta_k; x)\} \to p(\bar{\Delta}; x)$ *holds in part 2.*

Proof. Model Assumption 3.1 and compactness of the trust region ensure that the objective of (3.1a) is always level-bounded in s locally uniformly in Δ [36, Definition 1.16] because for any $\bar{\Delta} > 0$ and $\epsilon > 0$, and for any $\Delta \in (\bar{\Delta} - \epsilon, \bar{\Delta} + \epsilon)$ with $\Delta \geq 0$, the level sets of $\varphi(\cdot; x) + \psi(\cdot; x) + \chi(\cdot; \Delta)$ are contained in $\Delta\mathbb{B} \subseteq (\bar{\Delta} + \epsilon)\mathbb{B}$. Parts 1–2 follow by Rockafellar and Wets [36], Theorems 1.17 and 7.41. Part 3 follows from Rockafellar and Wets [36, Exercice 7.45]. Part 4 follows by noting that if $\|\bar{s}\| < \bar{\Delta}$, then $\varphi(\bar{s}; x) + \psi(\bar{s}; x) + \chi(\bar{s}; \Delta)$ is continuous in Δ in a neighborhood of $\bar{\Delta}$; the rest follows from Rockafellar and Wets [36, Theorem 1.17c]. ∎

It is not necessary to assume that $\psi(\cdot; x)$ is prox-bounded in Model Assumption 3.1 because under the assumptions stated and compactness of the trust region, the objective of (3.1a) is necessarily bounded below, and therefore prox-bounded. Proposition 3.1 allows us to think of how approximate solutions "truncated" by a trust-region constraint approach $\bar{s}$ as the trust-region radius increases. Indeed, we may choose any $\bar{\Delta} > \|\bar{s}\|$ in parts 2 and 4. When $\psi(\cdot; x) = 0$ and $\varphi(\cdot; x)$ is quadratic and strictly convex, the graph of $P(\cdot; x)$ is known to be a smooth curve such that $P(0; x) = \{x_k\}$, that is tangential to $-\nabla f(x_k)$ at $\Delta = 0$ and such that $\lim_{\Delta \to \infty} P(\Delta; x)$ contains the Newton step as its only element. This observation gives rise to several numerical methods to approximate the solution of (3.1), including the dogleg [32] and double dogleg methods [15].

3.2. Optimality measures. In this section, we seek a convenient way of assessing whether a given x is first-order critical for (1.1) based on the trust-region subproblem (3.1). We begin with the following result.

PROPOSITION 3.2. *Let Model Assumption 3.1 be satisfied. Assume in addition that* $\nabla_s\varphi(0; x) = \nabla f(x)$*,* $\partial\psi(0; x) = \partial h(x)$*, and let* $\Delta > 0$*. Then* $0 \in P(\Delta; x) \Longrightarrow s = 0$ *is first-order stationary for (3.1)* $\Longleftrightarrow$ x *is first-order stationary for (1.1).*

Proof. By definition, x is first-order stationary if and only if $0 \in \nabla f(x) + \partial h(x) = \nabla_s\varphi(0; x) + \partial\psi(0; x)$. But $\psi(0; x) = \psi(0; x) + \chi(0; \Delta)$ and $\partial(\psi(\cdot; x) + \chi(\cdot; \Delta))(0) = \partial\psi(0; x) + \partial\chi(0; \Delta)$ because $\partial\chi(0; \Delta) = \{0\}$. Thus we obtain $0 \in \nabla_s\varphi(0; x) + \partial(\psi(\cdot; x) + \chi(\cdot; \Delta))(0)$, i.e., $s = 0$ is first-order stationary for (3.1). ∎

Proposition 3.2 suggests we may use an element of $P(\Delta; x)$ as first-order optimality measure for any $\Delta > 0$, such as for example $\|g(\Delta; x)\|$, where $g(\Delta; x)$ is the least-norm element of $P(\Delta; x)$. However, the dependency on Δ is inconvenient. In order to circumvent this difficulty, we focus our attention temporarily on the choice

$$\begin{aligned} \varphi(s; x) &= f(x) + \nabla f(x)^T s + \tfrac{1}{2}\nu^{-1}\|s\|^2 \\ &= \tfrac{1}{2}\nu^{-1}\|s + \nu\nabla f(x)\|^2 + f(x) - \tfrac{1}{2}\nu\|\nabla f(x)\|^2, \end{aligned} \tag{3.2}$$

where $\nu > 0$ is fixed, so that for any $x \in \mathbb{R}^n$,

$$p(\Delta; x, \nu) = e_{\nu\psi(\cdot;x)+\chi(\cdot;\Delta)}(-\nu\nabla f(x)) + f(x) - \tfrac{1}{2}\nu\|\nabla f(x)\|^2, \tag{3.3a}$$

$$P(\Delta; x, \nu) = \operatorname*{prox}_{\nu\psi(\cdot;x)+\chi(\cdot;\Delta)}(-\nu\nabla f(x)), \tag{3.3b}$$

and p only differs from a Moreau envelope by a constant. The above choice of $\varphi(\cdot; x)$ allows us to derive a convenient, computable optimality measure, and to generalize the concept of decrease along the steepest descent direction, also known as Cauchy decrease, which is so fundamental to the convergence analysis of computational methods for smooth optimization.

In the special case where $\psi(\cdot; x) = 0$, Proposition 3.1 part 3 indicates that $P(\Delta; x, \nu)$ is single valued, and its only element is the projection of $-\nu\nabla f(x)$ into the trust region. On the other hand, $p(\Delta; x, \nu)$ measures the decrease of (3.2) in the direction of the projected gradient. Cartis et al. [9] study the special case where $h(x) = g(c(x))$ with g convex and globally Lipschitz continuous, and c smooth. In lieu of (3.3a), they minimize $f(x) + \nabla f(x)^T s + g(c(x) + \nabla c(x)^T s)$ in the trust region, which is analogous.

Crucially, (3.3) describes the first step of the proximal gradient method with step size ν applied to (3.1a) where $\varphi(\cdot; x)$ is as in (3.2) from $s = 0$ with a trust region of radius Δ. In the notation of section 2.4, φ is $\varphi(\cdot; x)$ and ψ is $\psi(\cdot; x) + \chi(\cdot; \Delta)$. If $\psi(\cdot; x)$ is finite at $s_0 = 0$, the first step of the proximal gradient method is

$$\begin{aligned} s_1 &\in \operatorname*{argmin}_s \ \tfrac{1}{2}\nu^{-1}\|s + \nu\nabla f(x)\|^2 + \psi(s; x) + \chi(s; \Delta) \\ &= \operatorname*{argmin}_s \ f(x) + \nabla f(x)^T s + \tfrac{1}{2}\nu^{-1}\|s\|^2 + \psi(s; x) + \chi(s; \Delta), \end{aligned} \tag{3.4}$$

and yields the decrease

$$(\varphi + \psi)(s_1; x) \leq (f + h)(x) - \tfrac{1}{2}(\nu^{-1} - L)\|s_1\|^2 \tag{3.5}$$

Moreover, s_1 is also the first step of the proximal-gradient method applied to (3.1a) where $\varphi(\cdot; x)$ is *any* model of f about x that is differentiable at $s = 0$ with $\nabla_s\varphi(0; x) = \nabla f(x)$, and, in particular, any quadratic expansion of f about x. In the sequel, we use s_1 as the appropriate generalization to the nonsmooth context of the projected gradient step, which allows us to derive an adequate optimality measure.

Let

$$\xi(\Delta; x, \nu) := f(x) + h(x) - p(\Delta; x, \nu), \tag{3.6}$$

where $p(\Delta; x, \nu)$ is defined in (3.3a). In view of the above, $\xi(\Delta; x, \nu)$ measures the decrease predicted by the first step of the proximal gradient method applied to (3.1a) from $s = 0$ with trust-region radius Δ and step length $\nu > 0$, where $\varphi(\cdot; x)$ is any model of f about x that is differentiable at $s = 0$ with $\nabla_s\varphi(0; x) = \nabla f(x)$.

Assume from now on that $\varphi(0; x) = f(x)$ and $\psi(0; x) = h(x)$. Because $p(\Delta; x, \nu) \leq \varphi(0; x) + \psi(0; x) + \chi(0; \Delta) = f(x) + h(x)$, we necessarily have $\xi(\Delta; x, \nu) \geq 0$.

Examples of models of f satisfying the above assumptions include Taylor expansions of f about x, and in particular quadratic models $f(x) + \nabla f(x)^T s + \frac{1}{2}s^T B s$ where $B = B^T$. The most straightforward example of a model of h satisfying the above is $\psi(s; x) = h(x + s)$. If $h(x) = g(c(x))$, where $g : \mathbb{R}^m \to \overline{R}$ is proper, lsc and level-bounded, and $c : \mathbb{R}^n \to \mathbb{R}^m$ is continuously differentiable, other possible models

include $\psi(s;x) = g(c(x) + \nabla c(x)^T s)$ and $\psi(s;x) = g(c(x) + \nabla c(x)^T s + \sum_{i=1}^m s^T B_i s)$, where each $B_i = B_i^T$.

The following result allows us to rely on the computable values $p(\Delta;x,\nu)$ and $\xi(\Delta;x,\nu)$ to assess stationarity.

PROPOSITION 3.3. *Let Model Assumption 3.1 be satisfied where $\varphi(0;x) = f(x)$ and $\nabla_s\varphi(0;x) = \nabla f(x)$. Assume furthermore that $\psi(0;x) = h(x)$ and $\partial\psi(0;x) = \partial h(x)$, and let $\Delta > 0$. Then, $\xi(\Delta;x,\nu) = 0 \iff 0 \in P(\Delta;x,\nu) \implies x$ is first-order stationary for (1.1).*

Proof. $\xi(\Delta;x,\nu) = 0$ if and only if $p(\Delta;x,\nu) = f(x) + h(x) = \varphi(0;x) + \psi(0;x) + \chi(0;\Delta)$, which occurs if and only if $0 \in P(\Delta;x,\nu)$. Proposition 3.2 then implies that x is first-order stationary for (1.1). ∎

3.3. A trust-region algorithm. We focus on the solution of (1.1) under Problem Assumption 3.1.

PROBLEM ASSUMPTION 3.1. *In (1.1), $f \in \mathcal{C}^1(\mathbb{R}^n)$, and h is proper and lsc.*

At iteration k, we construct a model $m_k(s;x_k) := \varphi(s;x_k) + \psi(s;x_k) \approx f(x_k + s) + h(x_k + s)$ and we approximately solve

$$\underset{s}{\text{minimize}}\ m_k(s;x_k) \quad \text{subject to } \|s\| \le \Delta_k \tag{3.7}$$

by computing a step s_k required to result in at least a fraction of the decrease achieved with one step of the proximal gradient method. Step Assumption 3.1 formalizes our requirement.

STEP ASSUMPTION 3.1. *There exists $\kappa_{\rm m} > 0$ and $\kappa_{\rm mdc} \in (0,\,1)$ such that for all k, $\|s_k\| \le \Delta_k$ and*

$$|f(x_k + s_k) + h(x_k + s_k) - m_k(s_k;x_k)| \le \kappa_{\rm m}\|s_k\|^2, \tag{3.8a}$$

$$m_k(0;x_k) - m_k(s_k;x_k) \ge \kappa_{\rm mdc}\xi(\Delta_k;x_k,\nu_k), \tag{3.8b}$$

where m_k is defined above and $\xi(\Delta_k;x_k,\nu_k)$ is defined in (3.6).

Condition (3.8a) is certainly satisfied if both f and φ are twice continuously differentiable with bounded second derivatives, and $\psi(s;x_k) := h(x_k + s)$. It also holds when $h(x) = g(c(x))$ where $c : \mathbb{R}^n \to \mathbb{R}^m$ has Lipschitz-continuous Jacobian and $g : \mathbb{R}^m \to \mathbb{R}^n$ is Lipschitz continuous. Such a situation arises when (1.1) results from penalizing infeasibility in the process of solving a smooth constrained problem. A useful model is then $\psi(s;x_k) := g(c(x_k) + \nabla c(x_k)^T s)$. If $L > 0$ is the Lipschitz constant of g and $M > 0$ that of the Jacobian of c, we have

$$|h(x_k + s) - \psi(s;x_k)| \le L\|c(x_k + s) - c(x_k) - \nabla c(x_k)^T s\| \le \tfrac{1}{2}LM\|s\|^2,$$

for all s, and (3.8a) is satisfied.

In order to develop a convergence analysis, we further assume that the gradient of $\varphi(\cdot;x_k)$ is Lipschitz continuous, which is satisfied, for instance, in the case of a quadratic model. It is not necessary to assume at this point that those Lipschitz constants are uniformly bounded; we will make such an assumption when needed. We gather the assumptions on the model from sections 3.1 and 3.2 in Model Assumption 3.2.

MODEL ASSUMPTION 3.2. *For any $x \in \mathbb{R}^n$, $\varphi(\cdot;x)$ is continuously differentiable with $\varphi(0;x) = f(x)$ and $\nabla_s\varphi(0;x) = \nabla f(x)$. In addition, $\nabla_s\varphi(\cdot;x)$ is Lipschitz continuous with constant $L(x)$ for all $x \in \mathbb{R}^n$. Finally, $\psi(\cdot;x)$ is proper, lsc, and satisfies $\psi(0;x) = h(x)$ and $\partial\psi(0;x) = \partial h(x)$.*

The complete process is formalized in Algorithm 3.1, which differs from a traditional trust-region algorithm in a few respects. First, each iteration begins with the choice of a steplength $\nu_k > 0$ for the proximal-gradient method. Steplength ν_k must be below $1/L(x_k)$ to ensure descent; in addition, we connect ν_k explicitly to Δ_k for a reason that becomes apparent in Theorem 3.4. Second, a step computation occurs in two phases. In the first phase, we compute the first step $s_{k,1}$ of the proximal-gradient method applied to our model with trust-region radius Δ_k. Step $s_{k,1}$ is an analog of the scaled projected gradient for nonsmooth regularized problems. In the second phase, we continue the proximal-gradient iterations from $s_{k,1}$ but possibly modify the trust-region radius so it does not exceed $\beta\|s_{k,1}\|$ for a prescribed $\beta \geq 1$. This choice is similar in spirit to the analysis of Curtis et al. [12] for smooth problems, who set the radius to be proportional to the gradient norm. More precisely, if $\|s_{k,1}\| < \Delta_k$, we explore a trust region of radius $\beta\|s_{k,1}\| \geq \|s_{k,1}\|$. Because the constraint $\|s\| \leq \Delta_k$ is inactive at $s_{k,1}$, the first step of the proximal gradient method computed in the updated trust region remains $s_{k,1}$, so that subsequent proximal gradient iterations will result in further decrease and the ultimate step s_k will satisfy (3.8b). If, on the other hand, $\|s_{k,1}\| = \Delta_k$, the first step of the proximal gradient method computed in a larger trust region might differ from $s_{k,1}$, which would jeopardize satisfaction of (3.8b). In order to preserve (3.8b), we leave Δ_k unchanged.

A final difference with traditional trust-region methods is that h and/or $\psi(\cdot; x)$ can take the value $+\infty$. In accordance with [36], we employ extended arithmetic rules in which $+\infty \cdot 0 = 0 \cdot (+\infty) = +\infty/(+\infty) := 0$. Thus if $h(x_k + s_k) = \psi(s_k; x_k) = +\infty$, we set $\rho_k := 0$ at step 9 of Algorithm 3.1.

Algorithm 3.1 Nonsmooth Regularized Trust-Region Algorithm.

1: Choose constants

$$0 < \eta_1 \leq \eta_2 < 1, \quad 0 < \gamma_1 \leq \gamma_2 < 1 < \gamma_3 \leq \gamma_4 \quad \text{and} \quad \alpha > 0,\ \beta \geq 1.$$

2: Choose $x_0 \in \mathbb{R}^n$ where h is finite, $\Delta_0 > 0$, compute $f(x_0) + h(x_0)$.
3: **for** $k = 0, 1, \ldots$ **do**
4: Choose $0 < \nu_k \leq 1/(L(x_k) + \alpha^{-1}\Delta_k^{-1})$.
5: Define $m_k(s; x_k) := \varphi(s; x_k) + \psi(s; x_k)$ satisfying Model Assumption 3.2.
6: Define $m_k^\nu(s; x_k) := \varphi^\nu(s; x_k) + \psi(s; x_k)$ where $\varphi^\nu(\cdot; x_k)$ is as in (3.2).
7: Compute $s_{k,1}$ as the solution of (3.7) with model $m_k^\nu(s; x_k)$.
8: Compute an approximate solution s_k of (3.7) with model $m_k(s; x_k)$ satisfying Step Assumption 3.1 and such that $\|s_k\| \leq \min(\Delta_k, \beta\|s_{k,1}\|)$.
9: Compute the ratio

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{m_k(0; x_k) - m_k(s_k; x_k)}.$$

10: If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise, set $x_{k+1} = x_k$.
11: Update the trust-region radius according to

$$\Delta_{k+1} \in \begin{cases} [\gamma_3\Delta_k, \gamma_4\Delta_k] & \text{if } \rho_k \geq \eta_2, \quad \text{(very successful iteration)} \\ [\gamma_2\Delta_k, \Delta_k] & \text{if } \eta_1 \leq \rho_k < \eta_2, \quad \text{(successful iteration)} \\ [\gamma_1\Delta_k, \gamma_2\Delta_k] & \text{if } \rho_k < \eta_1 \quad \text{(unsuccessful iteration).} \end{cases}$$

12: **end for**

3.4. Convergence analysis and iteration complexity. Our first result states that a successful step is guaranteed provided the trust-region radius is small enough.

THEOREM 3.4. *Let Model Assumption 3.2 and Step Assumption 3.1 be satisfied and let*

$$\Delta_{\text{succ}} := \frac{\kappa_{\text{mdc}}(1-\eta_2)}{2\kappa_{\text{m}}\alpha\beta^2} > 0. \tag{3.9}$$

If x_k is not first-order stationary and $\Delta_k \leq \Delta_{\text{succ}}$, then iteration k is very successful and $\Delta_{k+1} \geq \Delta_k$.

Proof. Because x_k is not first-order stationary, $s_{k,1} \neq 0$ and $s_k \neq 0$. Note first that (3.5), (3.6) and Model Assumption 3.2 give

$$\xi(\Delta_k; x_k, \nu_k) \geq (f+h)(x_k) - (\varphi+\psi)(s_1; x_k) \geq \tfrac{1}{2}(\nu_k^{-1} - L(x_k))\|s_{k,1}\|^2.$$

Line 4 of Algorithm 3.1 implies in turn that $\nu_k^{-1} - L(x_k) \geq \alpha^{-1}\Delta_k^{-1}$, so that

$$\xi(\Delta_k; x_k, \nu_k) \geq \tfrac{1}{2}\alpha^{-1}\Delta_k^{-1}\|s_{k,1}\|^2.$$

Model Assumption 3.2 and Step Assumption 3.1 together with the bound $\|s_k\| \leq \beta\|s_{k,1}\|$ yield

$$\begin{aligned}
|\rho_k - 1| &= \left| \frac{f(x_k+s_k) + h(x_k+s_k) - m_k(s_k; x_k)}{m_k(0; x_k) - m_k(s_k; x_k)} \right| \\
&\leq \frac{\kappa_{\text{m}}\|s_k\|^2}{\kappa_{\text{mdc}}\xi(\Delta_k; x_k, \nu_k)} \\
&\leq \frac{\kappa_{\text{m}}\beta^2\|s_{k,1}\|^2}{\frac{1}{2}\alpha^{-1}\Delta_k^{-1}\|s_{k,1}\|^2} \\
&= \frac{2\kappa_{\text{m}}\alpha\beta^2}{\kappa_{\text{mdc}}}\Delta_k.
\end{aligned}$$

Therefore, $\Delta_k \leq \Delta_{\text{succ}}$ implies $\rho_k \geq \eta_2$ and iteration k is very successful. The trust-region update of Algorithm 3.1 ensures that $\Delta_{k+1} \geq \Delta_k$. ∎

A careful examination of the proof of Theorem 3.4 reveals that the model adequacy condition (3.8a) could be replaced with the weaker condition

$$|f(x_k+s_k) + h(x_k+s_k) - m_k(s_k; x_k)| \leq \kappa_{\text{m}}\beta^2\|s_{k,1}\|^2, \tag{3.10}$$

which encapsulates the step size and the trust-region radius simultaneously, and suggests that $s_{k,1}$ is the appropriate generalization of the projected gradient for nonsmooth regularized optimization.

We are now in position to show that Algorithm 3.1 identifies a first-order critical point. We first consider the case where there are finitely many successful iterations.

THEOREM 3.5. *Let Model Assumption 3.2 and Step Assumption 3.1 be satisfied. If Algorithm 3.1 only generates finitely many successful iterations, then $x_k = x^*$ for all sufficiently large k and x^* is first-order critical.*

Proof. The proof mirrors that of Conn et al. [11, Theorem 6.4.4]. Under the assumptions given, there exists $k_0 \in \mathbb{N}$ such that all iterations $k \geq k_0$ are unsuccessful and $x_k = x_{k_0} = x^*$. Assume by contradiction that x^* is not first-order critical. The

mechanism of Algorithm 3.1 ensures that Δ_k decreases on unsuccessful iterations. Thus, there must be $k_1 \geq k_0$ such that $\Delta_k \leq \Delta_{\text{succ}}$, where Δ_{succ} is defined in Theorem 3.4, which ensures that iteration k_1 is successful and contradicts our assumption. ∎

We now turn to the case where there are infinitely many successful iterations and show that the objective is either unbounded below or a measure of criticality converges to zero. The mechanism of Algorithm 3.1 and Theorem 3.4 together ensure that

$$\Delta_k \geq \Delta_{\min} \quad \text{for all } k \in \mathbb{N} \text{ where } \quad \Delta_{\min} := \min(\Delta_0, \gamma_1 \Delta_{\text{succ}}) > 0. \tag{3.11}$$

Thus, by definition of $\xi(\cdot; x_k, \nu_k)$ and (3.11), we have

$$\xi(\Delta_k; x_k, \nu_k) \geq \xi(\Delta_{\min}; x_k, \nu_k) \quad \text{for all } k \in \mathbb{N}. \tag{3.12}$$

Following this last observation and in view of Proposition 3.3 and (2.6), we define $\nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}}$ as our measure of criticality. Observe the similarity between this measure and $\|G_{\nu_k}(0)\|$ defined in (2.5).

Our objective is to establish that $\liminf \nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k) = 0$ provided $f + h$ is bounded below. While doing so, we also establish a complexity result.

Let $\epsilon > 0$ be a stopping tolerance set by the user. We are interested in determining the smallest iteration number $k(\epsilon)$ at which we achieve the first-order optimality condition

$$\nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} \leq \epsilon \quad (0 < \epsilon < 1). \tag{3.13}$$

We denote

$$\mathcal{S} := \{k \in \mathbb{N} \mid \rho_k \geq \eta_1\}, \tag{3.14a}$$

$$\mathcal{S}(\epsilon) := \{k \in \mathcal{S} \mid k < k(\epsilon)\}, \tag{3.14b}$$

$$\mathcal{U}(\epsilon) := \{k \in \mathbb{N} \mid k \notin \mathcal{S} \text{ and } k < k(\epsilon)\}, \tag{3.14c}$$

respectively the set of all successful iterations, the set of successful iterations for which (3.13) has not yet been attained, and the set of unsuccessful iterations before (3.13) is first attained.

We make the following additional assumption on the model.

Model Assumption 3.3. *In Model Assumption 3.2, there exists $L > 0$ such that $0 < L(x_k) \leq L$ for all $k \in \mathbb{N}$. In addition, we select ν_k at line 4 of Algorithm 3.1 in a way that there exists $\nu_{\min} > 0$ such that $\nu_k \geq \nu_{\min}$ for all $k \in \mathbb{N}$.*

We stress that it is not necessary to know the value of or estimate L; only to ensure that such a constant exists, which may be achieved either by controling the norm of quasi-Newton approximations [26] or employing exact Hessians and substituting one for a bounded approximation when its norm is too large. Finally, in view of (3.11), there exists $\nu_{\min} > 0$ satisfying the assumption. For instance, choosing $\nu_k := 1/(L(x_k) + \alpha^{-1}\Delta_k^{-1})$ at each iteration ensures that $\nu_k \geq \nu_{\min} := 1/(L + \alpha^{-1}\Delta_{\min}^{-1}) > 0$.

The following two results parallel the now-classic complexity analysis of Cartis et al. [9] and references therein.

Lemma 3.6. *Let Model Assumptions 3.2 and 3.3 and Step Assumption 3.1 be satisfied. Assume there are infinitely many successful iterations and that $f(x_k) + h(x_k) \geq (f + h)_{\text{low}}$ for all $k \in \mathbb{N}$. Then, for all $\epsilon \in (0, 1)$,*

$$|\mathcal{S}(\epsilon)| \leq \frac{(f + h)(x_0) - (f + h)_{\text{low}}}{\eta_1 \kappa_{\text{mdc}} \nu_{\min}^2 \epsilon^2} = O(\epsilon^{-2}). \tag{3.15}$$

Proof. If $k \in \mathcal{S}(\epsilon)$, Model Assumption 3.3 and Step Assumption 3.1 and (3.12) imply

$$\begin{aligned} f(x_k) + h(x_k) - f(x_k + s_k) - h(x_k + s_k) &\geq \eta_1(m_k(0; x_k) - m_k(s_k; x_k)) \\ &\geq \eta_1 \kappa_{\text{mdc}} \xi(\Delta_k; x_k, \nu_k) \\ &\geq \eta_1 \kappa_{\text{mdc}} \xi(\Delta_{\min}; x_k, \nu_k) \\ &\geq \eta_1 \kappa_{\text{mdc}} \nu_k^2 \epsilon^2 \\ &\geq \eta_1 \kappa_{\text{mdc}} \nu_{\min}^2 \epsilon^2. \end{aligned}$$

Because $f + h$ is bounded below by $(f + h)_{\text{low}}$, summing the above inequalities over all $k \in \mathcal{S}(\epsilon)$ yields

$$(f + h)(x_0) - (f + h)_{\text{low}} \geq \sum_{k \in \mathcal{S}(\epsilon)} (f + h)(x_k) - (f + h)(x_{k+1}) \geq |\mathcal{S}(\epsilon)| \eta_1 \kappa_{\text{mdc}} \nu_{\min}^2 \epsilon^2,$$

which establishes (3.15). ∎

In order to derive a similar bound on the total number of iterations before (3.13) is first attained, we need to bound the number of unsuccessful iterations.

LEMMA 3.7. *Under the assumptions of Lemma 3.6,*

$$|\mathcal{U}(\epsilon)| \leq \log_{\gamma_2}(\Delta_{\min}/\Delta_0) + |\mathcal{S}(\epsilon)||\log_{\gamma_2}(\gamma_4)| = O(\epsilon^{-2}). \tag{3.16}$$

Proof. Each unsuccessful iteration reduces the trust-region radius by a factor at least γ_2, while at each successful iteration, $\Delta_{k+1} \leq \gamma_4 \Delta_k$. Thus if $k(\epsilon) - 1$ is the iteration index just before (3.13) occurs for the first time,

$$\Delta_{\min} \leq \Delta_{k(\epsilon)-1} \leq \Delta_0 \gamma_2^{|\mathcal{U}(\epsilon)|} \gamma_4^{|\mathcal{S}(\epsilon)|}.$$

Taking logarithms on both sides and remembering that $0 < \gamma_2 < 1$ gives

$$|\mathcal{U}(\epsilon)| \log(\gamma_2) + |\mathcal{S}(\epsilon)| \log(\gamma_4) \geq \log(\Delta_{\min}/\Delta_0),$$

and establishes (3.16). ∎

Finally, the total number of iteration until (3.13) is attained is given in the next result, which simply combines Lemma 3.6 and Lemma 3.7.

THEOREM 3.8. *Under the assumptions of Lemma 3.6,*

$$|\mathcal{S}(\epsilon)| + |\mathcal{U}(\epsilon)| = O(\epsilon^{-2}). \tag{3.17}$$

We use the update $\Delta_{k+1} \in [\gamma_3 \Delta_k, \gamma_4 \Delta_k]$ on very successful iterations but other possibilities exist. For instance, it is common to set $\Delta_{k+1} = \max(\gamma_3 \|s_k\|, \Delta_k)$ instead. Lemma 3.7 continues to hold because on successful iterations, $\Delta_{k+1} \leq \max(\gamma_3 \Delta_k, \Delta_k) = \gamma_3 \Delta_k$.

Curtis et al. [12] establish a complexity bound of $O(\epsilon^{-2})$ by making Δ_k proportional to an optimality measure—in their context of smooth optimization, they choose the gradient norm. Grapiglia et al. [20] study the convergence and complexity of a generic algorithm that has trust-region methods as a special case and obtain the $O(\epsilon^{-2})$ complexity bound under stronger smoothness assumptions than ours. Among others, they establish a bound for regularized optimization but also require h to be convex and

globally Lipschitz continuous. Curtis et al. [13] describe a nonstandard trust-region algorithm with a stronger $O(\epsilon^{-3/2})$ complexity bound.

A straightforward consequence of Theorem 3.8 is that if $f + h$ is bounded below, a subsequence of the criticality measure converges to zero.

COROLLARY 3.9. *Let Model Assumptions 3.2 and 3.3, and Step Assumption 3.1 be satisfied. If there are infinitely many successful iterations, then, either*

$$\lim_{k\to\infty} f(x_k) + h(x_k) \to -\infty \quad \text{or} \quad \liminf_{k\to\infty} \nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} = 0.$$

Proof. Follows directly from Theorem 3.8. ∎

In order to give an interpretation of Corollary 3.9, consider (3.1) with $\Delta = \Delta_{\min} > 0$ along with its value function $p(\Delta_{\min}; x, \nu)$, optimal set $P(\Delta_{\min}; x, \nu)$ and the optimality measure $\xi(\Delta_{\min}; x, \nu)$, where (x, ν) now plays the role of the parameter. Similar to Proposition 3.1, though with slightly stronger assumptions than Model Assumption 3.1, we have the following result.

PROPOSITION 3.10. *Let Problem Assumption 3.1 be satisfied and consider (3.1) with φ as in (3.2). Assume ψ is proper and lsc in the joint variables (s, x) and $\psi(s; x) + \chi(s; \Delta_{\min})$ is level-bounded in s locally uniformly in x. Then, the domain of $p(\Delta_{\min}; \cdot, \cdot)$ and $P(\Delta_{\min}; \cdot, \cdot)$ is $\mathbb{R}^n \times \{\nu \mid \nu > 0\}$. In addition,*

1. *$p(\Delta_{\min}; \cdot, \cdot)$ is proper continuous and for all $x \in \mathbb{R}^n$ and $\nu > 0$, $P(\Delta_{\min}; x, \nu)$ is nonempty and compact. In addition, $\xi(\Delta_{\min}; \cdot, \cdot)$ is proper lsc;*
2. *if $\{x_k\} \to \bar{x}$ and $\{\nu_k\} \to \bar{\nu} > 0$, and for each k, $s_k \in P(\Delta_{\min}; x_k, \nu_k)$, then $\{s_k\}$ is bounded and all its limit points are in $P(\Delta_{\min}; \bar{x}, \bar{\nu})$.*

Proof. Because h is proper lsc, (3.6) implies that $\xi(\Delta_{\min}; \cdot, \cdot)$ is proper whenever $p(\Delta_{\min}; \cdot, \cdot)$ is proper and is lsc whenever $p(\Delta_{\min}; \cdot, \cdot)$ is continuous. The latter holds because $p(\Delta_{\min}; \cdot, \cdot)$ is the composition of ∇f, which is continuous, with the Moreau envelope of $\psi(\cdot; x) + \chi(\cdot; \Delta)$, and such Moreau envelope is continuous in (x, ν)—see, [36, Theorem 1.25]. The rest follows by [36, Theorems 1.17 and 7.41]. ∎

By Corollary 3.9, if $f + h$ is bounded below, there is an index set $\mathcal{K}$ such that $\{\nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}}\}_{\mathcal{K}} \to 0$. Assume that $\{(x_k, \nu_k)\}_{\mathcal{K}}$ possesses a limit point and, without loss of generality, that $\{(x_k, \nu_k)\}_{\mathcal{K}} \to (\bar{x}, \bar{\nu})$ with $\bar{\nu} > 0$. That implies that $\{\xi(\Delta_{\min}; x_k, \nu_k)\}_{\mathcal{K}} \to 0$ because for all sufficiently large k,

$$\nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} \geq \tfrac{1}{2}\bar{\nu}^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} \geq 0.$$

Under the assumptions of Proposition 3.10, $\xi(\Delta_{\min}; \cdot, \cdot)$ is lsc, which means exactly that

$$0 = \liminf_{k\in\mathcal{K}} \xi(\Delta_{\min}; x_k, \nu_k) = \xi(\Delta_{\min}; \bar{x}, \bar{\nu}),$$

so that $\bar{x}$ is first-order critical.

It turns out that a stronger conclusion holds without further assumptions; the following result implies that *every* limit point of $\{(x_k, \nu_k)\}$ determines a first-order critical point. The proof follows the logic of [11, Theorem 6.4.6] but is significantly simpler due to the form of Step Assumption 3.1 and (3.12).

THEOREM 3.11. *Let Model Assumptions 3.2 and 3.3 and Step Assumption 3.1 be satisfied. If there are infinitely many successful iterations,*

$$\lim_{k\to\infty} f(x_k) + h(x_k) \to -\infty \quad \text{or} \quad \lim_{k\to\infty} \nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{\frac{1}{2}} = 0.$$

Proof. If $\{\nu_k^{-1}\xi(\Delta_{\min};x_k,\nu_k)^{\frac{1}{2}}\} \not\to 0$, there exist $\epsilon > 0$ and an infinite set $\mathcal{K} \subset \mathcal{S}$ such that $\nu_k^{-1}\xi(\Delta_{\min};x_k,\nu_k)^{\frac{1}{2}} \geq \epsilon$ for all $k \in \mathcal{K}$. Because each $k \in \mathcal{K}$ is a successful iteration, Step Assumption 3.1 and (3.12) yield

$$\begin{aligned}(f+h)(x_k) - (f+h)(x_{k+1}) &\geq \eta_1\kappa_{\text{mdc}}\xi(\Delta_k;x_k,\nu_k)\\ &\geq \eta_1\kappa_{\text{mdc}}\xi(\Delta_{\min};x_k,\nu_k)\\ &\geq \eta_1\kappa_{\text{mdc}}\nu_{\min}^2\epsilon^2\end{aligned}$$

for all $k \in \mathcal{K}$, which is a contradiction if $\{f(x_k)+h(x_k)\}$ is not bounded below. ∎

4. Proximal-quasi-Newton trust-region method. In this section, we consider the computation of a trust-region step and develop a special case of Proposition 2.6 in which

$$\varphi(s;x) := f(x) + \nabla f(x)^T s + \tfrac{1}{2}s^T B s, \tag{4.1}$$

where $B = B^T$. We assume that $\Delta > 0$ is fixed. For conciseness, we use the notation $\varphi(s) := \varphi(s;x)$ and $\psi(s) := \psi(s;x) + \chi(s;\Delta)$. We work under Model Assumption 3.2, i.e., we assume that ψ is proper and lsc with prox-boundedness coming from $\chi(\cdot;\Delta)$.

4.1. Computing a trust-region step. The following result states a fundamental relationship between G_ν and $\partial\psi$.

LEMMA 4.1.
Let s_{j+1} be given by (2.3) and $G_\nu(s_j)$ be defined by (2.5). Then,

$$G_\nu(s_j) - \nabla\varphi(s_j) \in \partial\psi(s_{j+1}). \tag{4.2a}$$

$$(B - \nu^{-1}I)(s_{j+1} - s_j) \in \nabla\varphi(s_{j+1}) + \partial\psi(s_{j+1}). \tag{4.2b}$$

Proof. (4.2a) is a simple restatement of (2.4) and (4.2b) results from adding $\nabla\varphi(s_{j+1})$to both sides of (2.4) and substituting the gradient of φ using (4.1). ∎

The next result shows that (2.3) is a descent method when φ is a quadratic.

LEMMA 4.2. *Let $\{s_j\}$ be generated according to (2.3). For all $j \geq 0$,*

$$\psi(s_{j+1}) + \nabla\varphi(s_j)^T(s_{j+1} - s_j) \leq \psi(s_j) - \tfrac{1}{2}\nu^{-1}\|s_{j+1} - s_j\|^2, \tag{4.3a}$$

$$(\varphi+\psi)(s_{j+1}) \leq (\varphi+\psi)(s_j) + \tfrac{1}{2}(s_{j+1} - s_j)^T(B - \nu^{-1}I)(s_{j+1} - s_j). \tag{4.3b}$$

Proof. Because s_{j+1} solves (2.3),

$$\tfrac{1}{2}\nu^{-1}\|s_{j+1} - (s_j - \nu\nabla\varphi(s_j))\|^2 + \psi(s_{j+1}) \leq \tfrac{1}{2}\nu^{-1}\|\nu\nabla\varphi(s_j)\|^2 + \psi(s_j).$$

By expanding the squared norm in the left-hand-side of the above and cancelling the common term $\|\nu\nabla\varphi(s_j)\|^2$, we obtain (4.3a). Because φ is quadratic,

$$\varphi(s_{j+1}) = \varphi(s_j) + \nabla\varphi(s_j)^T(s_{j+1} - s_j) + \tfrac{1}{2}(s_{j+1} - s_j)^T B(s_{j+1} - s_j).$$

We now add $\psi(s_{j+1})$ to both sides and use (4.3a) and obtain

$$\begin{aligned}(\varphi+\psi)(s_{j+1}) &\leq \varphi(s_j) + \psi(s_j) - \tfrac{1}{2}\nu^{-1}\|s_{j+1} - s_j\|^2 + \tfrac{1}{2}(s_{j+1} - s_j)^T B(s_{j+1} - s_j)\\ &= (\varphi+\psi)(s_j) + \tfrac{1}{2}(s_{j+1} - s_j)^T(B - \nu^{-1}I)(s_{j+1} - s_j).\end{aligned}$$ ∎

We now examine two choices of $\nu > 0$ that result in two decrease behaviors.

COROLLARY 4.3. *Under the assumptions of Lemma 4.2, assume* $0 < \nu \leq (1 - \theta)/\|B\|$ *for some* $\theta \in (0, 1)$*, or simply that* $\nu > 0$ *if* $B = 0$*, in which case* $\theta = 1$*. Then,*

$$(\varphi + \psi)(s_{j+1}) \leq (\varphi + \psi)(s_j) - \tfrac{1}{2}\theta\nu^{-1}\|s_j - s_{j+1}\|^2, \quad (j \geq 0). \tag{4.4}$$

Proof. If $B = 0$, (4.4) with $\theta = 1$ follows directly from (4.3a). If $B \neq 0$, we have by assumption $(1 - \theta)\nu^{-1} \geq \|B\|$, so that $\lambda_{\max}(B - \nu^{-1}I) \leq -\theta\nu^{-1} < 0$, and therefore,

$$(s_{j+1} - s_j)^T(B - \nu^{-1}I)(s_{j+1} - s_j) \leq -\theta\nu^{-1}\|s_{j+1} - s_j\|^2,$$

which combines with (4.3b) to complete the proof. ∎

COROLLARY 4.4. *Under the assumptions of Lemma 4.2, assume* $B \neq 0$*, let* $0 < \theta < 1/(4\|B\|)$ *and* $\nu^{\min} \leq \nu \leq \nu^{\max}$*, where*

$$\nu^{\min} := \frac{1 - \sqrt{1 - 4\theta\|B\|}}{2\|B\|}, \qquad \nu^{\max} := \frac{1 + \sqrt{1 - 4\theta\|B\|}}{2\|B\|}.$$

Then, for all $j \geq 0$*,*

$$(\varphi+\psi)(s_{j+1}) \leq (\varphi+\psi)(s_j) - \tfrac{1}{2}\theta\nu^{-2}\|s_j - s_{j+1}\|^2 = (\varphi+\psi)(s_j) - \tfrac{1}{2}\theta\|G_\nu(s_j)\|^2. \tag{4.5}$$

Proof. Under our assumptions, the quadratic $p(\nu) := \|B\|\nu^2 - \nu + \theta$ has the two positive real roots $\nu^{\min}$ and $\nu^{\max}$. Moreover, for all $\nu \in [\nu^{\min}, \nu^{\max}]$, $p(\nu) \leq 0$, which can also be written $\|B\| - \nu^{-1} \leq -\theta\nu^{-2}$. Therefore, if $\nu \in [\nu^{\min}, \nu^{\max}]$, then for all j,

$$(s_{j+1} - s_j)^T(B - \nu^{-1}I)(s_{j+1} - s_j) \leq -\theta\nu^{-2}\|s_{j+1} - s_j\|^2 = -\theta\|G_\nu(s_j)\|^2,$$

which combines with (4.3b) to complete the proof. ∎

Because $s_0 = 0$ and $(\varphi + \psi)(s_0) = f(x) + h(x) < +\infty$, if ν is chosen as in Corollary 4.3 or Corollary 4.4, (2.3) generates iterates $\{s_j\}$ such that $\{(\varphi + \psi)(s_j)\}$ is monotonically decreasing and all its terms are finite. Finiteness implies that $\|s_j\| \leq \Delta$ for all $j \geq 0$, i.e., all iterates lie in the trust region. In particular, for any $j \geq 1$,

$$m_k(s_{j+1}; x_k) \leq m_k(s_j; x_k) \leq m_k(s_1; x_k) = m_k^\nu(s_1; x_k), \tag{4.6}$$

where $m_k^\nu(s_1; x_k) = \frac{1}{2}\nu^{-1}\|s_1 + \nu\nabla f(x_k)\|_2^2 + (\psi + \chi)(s_1)$ and hence s_j satisfies the sufficient decrease condition (3.8b), and the final equality results from the fact that s_1 is the same for any model of the form (4.1).

With regards to proximal gradient convergence, two situations may occur. In the first, (2.3) results in $s_{j_0+1} = s_{j_0}$ for a smallest index $j_0 > 0$. In that case, (2.4) yields

$$0 \in \partial(\varphi + \psi)(s_{j_0}),$$

i.e., we have identified a stationary point of (3.7) in a finite number of iterations, while decreasing the value of m_k at each iteration. Otherwise, $s_{j+1} \neq s_j$ for all $j \geq 0$, and the next result establishes sub-linear convergence of the proximal gradient method (2.3).

THEOREM 4.5. *Let* $\{s_j\}$ *be generated according to* (2.3) *with* ν *as in Corollary 4.3. Denote* $(\varphi + \psi)_{\text{low}} := \inf(\varphi + \psi) > -\infty$*. Let* v_{j+1} *denote the left-hand side of* (4.2b)*. For any* $N \geq 1$*,*

$$\min_{j=0,\ldots,N-1} \|v_{j+1}\| \leq \sqrt{\frac{2}{N\theta}(\nu^{-1} - \lambda_{\min}(B))\,((\varphi + \psi)(s_0) - (\varphi + \psi)_{\text{low}})}.$$

Proof. We rearrange (4.4) and sum from iteration $j = 0$ to iteration $j = N - 1$:

$$\sum_{j=0}^{N-1} \|s_j - s_{j+1}\|^2 \leq \frac{2\nu}{\theta}((\varphi+\psi)(s_0) - (\varphi+\psi)(s_N)) \leq \frac{2\nu}{\theta}((\varphi+\psi)(s_0) - (\varphi+\psi)_{\text{low}}).$$

For any positive sequence $\{c_j\}$,

$$\min_{0\leq j\leq N-1} c_j = \sqrt{\min_{0\leq j\leq N-1} c_j^2} \leq \sqrt{\frac{1}{N}\sum_{j=0}^{N-1} c_j^2}.$$

Therefore,

$$\min_{0\leq j\leq N-1} \|s_j - s_{j+1}\| \leq \sqrt{\frac{2\nu}{N\theta}((\varphi+\psi)(s_0) - (\varphi+\psi)_{\text{low}})}.$$

Because $\|v_{j+1}\| \leq \|B - \nu^{-1}I\|\,\|s_j - s_{j+1}\| = (\nu^{-1} - \lambda_{\min}(B))\,\|s_j - s_{j+1}\| \leq \nu^{-1}\|s_j - s_{j+1}\|$, we obtain the desired result. ∎

When solving (3.7), a reasonable stopping condition would be $\|v_{j+1}\| \leq \epsilon$ for a user-chosen tolerance $\epsilon > 0$. Theorem 4.5 indicates that such stopping condition is attained after $N(\epsilon)$ iterations, where

$$N(\epsilon) = \left\lceil \frac{2}{\epsilon^2\theta}(\nu^{-1} - \lambda_{\min}(B))\,((\varphi+\psi)(s_0) - (\varphi+\psi)_{\text{low}}) \right\rceil.$$

A result similar to Theorem 4.5 can be established under the step size rule of Corollary 4.4, with nearly identical proof.

THEOREM 4.6. *Let $\{s_j\}$ be generated according to (2.3) with ν as in Corollary 4.4 with $0 < \theta < 1/(4\|B\|)$. Assume ψ, and therefore $\varphi + \psi$, is bounded below and denote $(\varphi+\psi)_{\text{low}} := \inf(\varphi+\psi) > -\infty$. For any $N \geq 1$,*

$$\min_{j=0,\ldots,N-1} \|G_\nu(s_{j+1})\| \leq \sqrt{\frac{2}{N\theta}((\varphi+\psi)(s_0) - (\varphi+\psi)_{\text{low}})}.$$

5. Proximal Operators for Trust-Region Subproblems. In this section, we develop techniques for computing (2.3) for use in Steps 7 and 8 of Algorithm 3.1. Many standard proximal operators for both convex and nonconvex prox-bounded functions ψ have been worked out [4, 10], and new examples for nonconvex problems continuously appear. Well-known examples include the firm-thresholding penalty [19], the SCAD penalty [17], MCP penalty [44], lower C^2 functions [21], any ℓ_p^p-seminorm for $0 < p < 1$ [46, Appendix A], and other exotic operators, see e.g. [45, Table 1]. We refer to such functions ψ as *prox-friendly*. However, Algorithm 3.1 requires evaluating proximal operators for modified functions that combine a shift and a summation with an indicator function. By Model Assumption 3.2, our model $\psi(s;x) \approx h(x+s)$ must coincide with h in value and subdifferential at $s = 0$. In particular, the choice $\psi(s;x) = h(x+s)$ seems natural when h itself is prox-friendly. Here we consider

$$\psi(s;x) := h(x+s) + \chi(s;\Delta\mathbb{B}_p), \tag{5.1}$$

where h is prox-friendly, x is a shift, and $p \in \{1, 2, \infty\}$. Below, we provide closed form solutions and/or efficient routines for (5.1) with focus on the following cases:

1. for an arbitrary separable prox-friendly h, we evaluate $\text{prox}_{\nu\psi(\cdot;x)}$ by leveraging $\text{prox}_{\nu h}$, but we restrict our attention to $p = \infty$. This allows us to consider (5.1) with $h(x) = \lambda\|x\|_1$ and $h(x) = \lambda\|x\|_0$;
2. we consider $h(x) = \lambda\|x\|_1$ in (5.1) for $p = 2$.

5.1. $p = \infty$, h separable. For the special case of $\mathbb{B}_\infty$, (2.1b) and (5.1) yield

$$\operatorname*{prox}_{\nu\psi}(q) := \operatorname*{argmin}_s \tfrac{1}{2}\nu^{-1}\|s-q\|^2 + h(x+s) + \chi(s;\Delta\mathbb{B}_\infty). \tag{5.2}$$

If h is separable, i.e., $h(x) = \sum_i h_i(x_i)$, (5.2) decouples in each coordinate:

$$\operatorname*{prox}_{\nu\psi}(q)_i = \operatorname*{argmin}_{s_i} \tfrac{1}{2}\nu^{-1}(s_i-q_i)^2 + h_i(x_i+s_i) + \chi(s_i;[-\Delta,\Delta]).$$

Using the change of variable $v_i = x_i + s_i$, we may rewrite

$$\operatorname*{prox}_{\nu\psi}(q)_i = \operatorname*{argmin}_{v_i}\{\tfrac{1}{2}\nu^{-1}(v_i-x_i-q_i)^2 + h_i(v_i) + \chi(v_i;[x_i-\Delta,x_i+\Delta])\} - x_i.$$

If h is convex, we may work backwards from the form of the solution. For any $p_i \in \operatorname{prox}_{\nu\psi}(q)_i$, either

1. $|p_i| < \Delta$, in which case $p_i \in \operatorname{prox}_{\nu h_i}(q+x)_i - x_i$;
2. otherwise, $|p_i| = \Delta$ by construction, and

$$\begin{aligned}\operatorname*{prox}_{\nu\psi}(q)_i &= \operatorname*{argmin}_{v_i = x_i \pm \Delta} (\tfrac{1}{2}\nu^{-1}(v_i-(x_i+q_i))^2 + h_i(v_i)) - x_i \\ &= \operatorname*{argmin}_{s_i=\pm\Delta} \tfrac{1}{2}\nu^{-1}(s_i-q_i)^2 + h_i(x_i+s_i) \subseteq \{-\Delta,\,\Delta\}.\end{aligned}$$

In such cases, the definition of convexity implies that set of bound-constrained solutions includes the projection of the unconstrained solutions into the bounds. Because the objective of (5.2) is strictly convex, equality holds:

$$\operatorname*{prox}_{\nu\psi}(q)_i = \{\operatorname*{proj}_{[x_i-\Delta,x_i+\Delta]}(\operatorname*{prox}_{\nu h_i}(q+x)_i)\} - x_i = \operatorname*{proj}_{[-\Delta,\Delta]}(\operatorname*{prox}_{\nu h_i}(q+x)_i - x_i),$$

For example, let $h(x) = \lambda\|x\|_1$. Then,

$$\begin{aligned}\operatorname*{prox}_{\nu\psi}(q)_i &= \operatorname*{proj}_{[-\Delta,\Delta]}(\operatorname*{prox}_{\nu\lambda|\cdot|}(q+x)_i - x_i) = \operatorname*{proj}_{[-\Delta,\Delta]}\left(\begin{cases} q_i-\nu\lambda & x_i+q_i > \nu\lambda \\ -x_i & |x_i+q_i| \le \nu\lambda \\ q_i+\nu\lambda & x_i+q_i < -\nu\lambda \end{cases}\right) \\ &= \operatorname*{proj}_{[-\Delta,\Delta]}\left(\operatorname*{proj}_{[q_i-\nu\lambda,q_i+\nu\lambda]}(-x_i)\right).\end{aligned}$$

When h is nonconvex, there may be a greater variety of cases. For instance, if $h(x) = \lambda\|x\|_0$, a global solution of (5.2) may be one of the bounds, or either of the unconstrained local minimizers q and $-x$ if they lie inside the bounds. A simple strategy consists in evaluating the objective of (5.2) at those four points and choosing one with lowest objective value.

5.2. $p = 2$, $h(x) = \lambda\|x\|_1$. When using other norms to define the trust region, additional computations are required. For certain norms, we can dualize h to solve (5.2). We focus on $h(x) = \lambda\|x\|_1$ with an ℓ_2-norm trust-region throughout because the ℓ_2-norm is standard in the literature, and is used in section 7.1.

First, we rewrite the scaled ℓ_1-norm using its conjugate:

$$\lambda\|x+s\|_1 = \sup_{w\in\lambda\mathbb{B}_\infty} w^T(x+s),$$

recharacterizing (2.1b) and (5.1) as

$$\min_{s} \sup_{w \in \lambda \mathbb{B}_\infty} \tfrac{1}{2}\nu^{-1}\|s-q\|^2 + w^T(x+s) + \chi(s; \Delta\mathbb{B}_2). \tag{5.3}$$

Strong duality holds in this case since the objective is convex, piecewise linear-quadratic, and the primal solution is attained. We interchange the order of minimization and maximization and complete squares in s and in w to obtain

$$\sup_{w \in \lambda \mathbb{B}_\infty} \min_{s} \tfrac{1}{2}\nu^{-1}\|s-q+\nu w\|^2 + \chi(s; \Delta\mathbb{B}_2) - \tfrac{1}{2}\nu^{-1}\left\|x+q-\nu w\right\|^2 + \tfrac{1}{2}\nu^{-1}\|x+q\|^2. \tag{5.4}$$

The solution of the inner problem is

$$s(w) := \operatorname*{proj}_{\Delta\mathbb{B}_2}(q-\nu w). \tag{5.5}$$

We substitute (5.5) back into (5.4) to rewrite the dual objective as

$$\sup_{w \in \lambda \mathbb{B}_\infty} \tfrac{1}{2}\nu^{-1}\operatorname{dist}\left(q-\nu w; \Delta\mathbb{B}_2\right)^2 - \tfrac{1}{2}\nu^{-1}\left\|x+q-\nu w\right\|^2 + \tfrac{1}{2}\nu^{-1}\|x+q\|^2. \tag{5.6}$$

The change of variable

$$y = q - \nu w, \tag{5.7}$$

transforms (5.6) into

$$\min_{q-\nu\lambda\mathbf{1} \le y \le q+\nu\lambda\mathbf{1}} \tfrac{1}{2}\nu^{-1}\left(\|y+x\|^2 - \operatorname{dist}\left(y; \Delta\mathbb{B}_2\right)^2\right), \tag{5.8}$$

where $\mathbf{1}$ is a vector of all ones. As the value function of (5.3) with respect to s, the objective of (5.8) is convex [36, Proposition 2.22]. The first-order optimality conditions of (5.8) are

$$0 \in x + \frac{y}{\max\{1, \|y\|/\Delta\}} + \nu\partial\chi(y; [q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]). \tag{5.9}$$

Once we have an optimal solution of (5.8) , denoted y^+, we can evaluate (5.5) at the corresponding w^+ to obtain

$$s = \operatorname*{proj}_{\Delta\mathbb{B}_2}(y^+).$$

which solves (5.2). To characterize y^+ more explicitly, we work backwards from properties of the solution. There are only two possibilities to consider: y^+ is in the trust region, and y^+ is outside of the trust region.

1. if $\|y^+\| < \Delta$, $\operatorname{dist}(y^+; \Delta\mathbb{B}_2) = 0$, and (5.8) and (5.9) simplify:

$$s = y^+ = \operatorname*{proj}_{[q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]}(-x),$$

   where we used (5.5) and (5.7);
2. if $\|y^+\| \ge \Delta$, (5.9) becomes

$$0 \in x + \frac{\Delta}{\|y\|}y + \nu\partial\chi(y; [q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]).$$

Multiplying through by $\|y\|/\Delta$ yields

$$0 \in y + \frac{\|y\|}{\Delta} x + \frac{\nu\|y\|}{\Delta} \partial\chi(y; [q - \nu\lambda\mathbf{1}, q + \nu\lambda\mathbf{1}]). \tag{5.10}$$

Suppose first that $\eta := \|y^+\|$ is known. A solution y^+ to (5.10) can be obtained by solving

$$\min_{y \in [q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]} \tfrac{1}{2}\|y + \frac{\eta}{\Delta} x\|^2$$

which can be written in closed form as

$$y = \operatorname*{proj}_{[q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]} \left(-\frac{\eta}{\Delta} x\right). \tag{5.11}$$

Taking the norm of each side of (5.11) gives a scalar root finding equation that characterizes η:

$$\eta = \left\| \operatorname*{proj}_{[q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]} \left(-\frac{\eta}{\Delta} x\right) \right\|.$$

Once we have solved for $\eta = \|y^+\|$, we obtain y^+ from (5.11), and, using (5.5),

$$s = \operatorname*{proj}_{\Delta\mathbb{B}_2} \left(\operatorname*{proj}_{[q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]} \left(-\frac{\eta}{\Delta} x\right) \right) = \left(\operatorname*{proj}_{[q-\nu\lambda\mathbf{1}, q+\nu\lambda\mathbf{1}]} \left(-\frac{\eta}{\Delta} x\right) \right) \frac{\Delta}{\eta}.$$

6. A quadratic regularization variant. We now describe a variant of the trust-region algorithm of the previous sections inspired by the modified Gauss-Newton scheme proposed by Nesterov [30] in the context of nonlinear least-squares problems. Here again, Cartis et al. [9] establish a complexity of $O(\epsilon^{-2})$ iterations to attain a near-optimality condition under the assumption that h is convex and globally Lipschitz continuous. In the sequel, we obtain the same complexity bound under Problem Assumption 3.1. The quadratic regularization method decribed below is closely related to the standard proximal gradient method with the exception that it employs an adaptive steplength. It may be used as an alternative to a linesearch-based proximal gradient method such as those of Li and Lin [24] and Boţ et al. [8].

In the quadratic regularization method, we use the linear model

$$\varphi(s; x) = f(x) + \nabla f(x)^T s \approx f(x + s) \tag{6.1}$$

together with a model of $\psi(s; x)$ that satisfies Model Assumption 3.2. The first difference is that in the present setting, the Lipschitz constant of $\nabla\varphi(\cdot; x)$ is $L(x) = 0$ for all $x \in \mathbb{R}^n$. The second difference is that we must now assume that $\psi(\cdot; x)$ is prox-bounded. In particular, there exists $\lambda_x \in \mathbb{R} \cup \{+\infty\}$ such that $\psi(\cdot; x) + \frac{1}{2}\lambda_x^{-1}\|\cdot\|^2$ is bounded below [36, Exercise 1.24c]. We refer to the supremum of all such λ_x as the *threshold of boundedness* of $\psi(\cdot; x)$. If $\psi(\cdot; x)$ is itself bounded below, we may choose $\lambda_x = +\infty$. At x, we define

$$p(\sigma; x) := \operatorname*{minimize}_s \; m(s; x, \sigma), \tag{6.2a}$$

$$P(\sigma; x) := \operatorname*{argmin}_s \; m(s; x, \sigma), \tag{6.2b}$$

where

$$m(s;x,\sigma) := \varphi(s;x) + \psi(s;x) + \tfrac{1}{2}\sigma\|s\|^2, \tag{6.3}$$

and $\sigma > 0$ is a regularization parameter. From x, the method computes a step $s \in P(\sigma;x)$. As earlier, let us also define

$$\xi(\sigma;x) := f(x) + h(x) - p(\sigma;x) \geq 0. \tag{6.4}$$

If we combine (6.1) with (6.3), we may write

$$m(s;x,\sigma) = \tfrac{1}{2}\sigma\|s + \sigma^{-1}\nabla f(x)\|^2 + \psi(s;x) + f(x) - \tfrac{1}{2}\sigma^{-1}\|\nabla f(x)\|^2, \tag{6.5}$$

where the last two terms are independent of s. In (6.5), we recognize a model of the form (3.4), so that minimizing (6.3) amounts to performing a single step of the proximal gradient method with step size $1/\sigma$ and Lipschitz constant $L = 0$. The decrease guaranteed by the proximal gradient method is given by (3.5), i.e.,

$$\xi(x;\sigma) = f(x) + h(x) - m(s;x,\sigma) \geq \tfrac{1}{2}\sigma\|s\|^2, \tag{6.6}$$

so that

$$f(x) + h(x) - (\varphi(s;x) + \psi(s;x)) \geq \sigma\|s\|^2, \tag{6.7}$$

provided that $\sigma > 1/\lambda_x$. Because of (6.7), there is no need for a sufficient decrease assumption such as (3.8b) in the quadratic regularization method.

In view of (6.5), Proposition 2.3 applies to (6.2). In particular, $p(\sigma;x)$ is continuous in (σ,x), and $P(\sigma;x)$ is nonempty and compact for all $\sigma > 1/\lambda_x$.

By Proposition 2.5, for any $\sigma > 1/\lambda_x$, if $s \in P(\sigma;x)$, then $0 \in \nabla f(x) + \partial\psi(s;x) + \sigma s$. Thus, we have the following optimality result.

LEMMA 6.1. *Let Model Assumption 3.2 be satisfied, $\psi(\cdot;x)$ be prox-bounded, and let $\sigma > 1/\lambda_x$. Then $\xi(\sigma;x) = 0 \iff 0 \in P(\sigma;x) \implies x$ is first-order stationary for (1.1).*

As in the trust-region context, we require that the difference between the model and the actual objective be bounded by a multiple of $\|s_k\|^2$:

STEP ASSUMPTION 6.1. *There exists $\kappa_{\mathrm{m}} > 0$ such that for all k,*

$$|f(x_k + s_k) + h(x_k + s_k) - \varphi_k(s_k;x_k) - \psi(s_k;x_k)| \leq \kappa_{\mathrm{m}}\|s_k\|^2. \tag{6.8}$$

Once a step s has been computed, its quality is assessed by comparing the decrease in $\varphi(\cdot;x) + \psi(\cdot;x)$ with that in the objective $f + h$, similarly to Algorithm 3.1. If both are in strong agreement, σ decreases. Otherwise, σ increases. We state the overall algorithm as Algorithm 6.1. Contrary to Algorithm 3.1, it is possible that, at certain iterations, $\sigma_k \leq 1/\lambda_{x_k}$, and in such a situation, it is possible that $m(s_k;x_k,\sigma_k) = -\infty$, and therefore that $\varphi(s_k;x_k) + \psi(s_k;x_k) \leq m(s_k;x_k,\sigma_k) = -\infty$. Because h is proper, $h(x_k + s_k)$ is either finite or $+\infty$. In such a case, thanks to extended arithmetic rules, $\rho_k = 0$, s_k is rejected, and σ_k is increased. After a finite number of such increases, we obtain $\sigma_k > 1/\lambda_{x_k}$.

We now combine (6.7) with Step Assumption 6.1 into the following result.

THEOREM 6.2. *Let Model Assumption 3.2 and Step Assumption 6.1 be satisfied, $\psi(\cdot;x_k)$ be prox-bounded for each $k \in \mathbb{N}$ with threshold of boundedness λ_{x_k}, and let*

$$\sigma_{\mathrm{succ}} := \kappa_{\mathrm{m}}/(1 - \eta_2) > 0. \tag{6.9}$$

Algorithm 6.1 Nonsmooth quadratic regularization algorithm.

Choose constants $0 < \eta_1 \leq \eta_2 < 1$ and $0 < \gamma_3 \leq 1 < \gamma_1 \leq \gamma_2$.
Choose $x_0 \in \mathbb{R}^n$ where h is finite, $\sigma_0 > 0$, compute $f(x_0) + h(x_0)$.
for $k = 0, 1, \ldots$ **do**
  Define $m(s; x_k, \sigma_k)$ as in (6.3) satisfying Model Assumption 3.2 with $L = 0$.
  Compute a solution s_k of (6.2) such that Step Assumption 6.1 holds.
  Compute the ratio

$$\rho_k := \frac{f(x_k) + h(x_k) - (f(x_k + s_k) + h(x_k + s_k))}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))}.$$

  If $\rho_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$. Otherwise, set $x_{k+1} = x_k$.
  Update the regularization parameter according to

$$\sigma_{k+1} \in \begin{cases} [\gamma_3 \sigma_k, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \eta_1 \leq \rho_k < \eta_2, \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k < \eta_1. \end{cases}$$

end for

If x_k is not first-order stationary and $\sigma_k \geq \max(1/\lambda_{x_k}, \sigma_{\text{succ}})$, then iteration k is very successful and $\sigma_{k+1} \leq \sigma_k$.

Proof. Let s_k be the step computed at iteration k of Algorithm 6.1. Because x_k is not first-order stationary, $s_k \neq 0$. Step Assumption 6.1 and (6.7) combine to yield

$$|\rho_k - 1| = \frac{|f(x_k + s_k) + h(x_k + s_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))|}{\varphi(0; x_k) + \psi(0; x_k) - (\varphi(s_k; x_k) + \psi(s_k; x_k))} \leq \frac{\kappa_{\text{m}} \|s_k\|^2}{\sigma_k \|s_k\|^2}.$$

After simplifying by $\|s_k\|^2$, we obtain $\sigma_k \geq \sigma_{\text{succ}} \implies \rho_k \geq \eta_2$. ∎

If there exists $\lambda > 0$ such that $\lambda_{x_k} \geq \lambda$ for all $k \in \mathbb{N}$, Theorem 6.2 ensures existence of a constant $\sigma_{\max} > 0$ such that

$$\sigma_k \leq \sigma_{\max} := \min(\sigma_0, \gamma_2 \max(1/\lambda, \sigma_{\text{succ}})) > 0 \quad \text{for all } k \in \mathbb{N}. \tag{6.10}$$

We paraphrase the above assumption by stating that $\psi(\cdot; x_k)$ is *uniformly prox-bounded*.

A result analogous to Theorem 3.5 holds for Algorithm 6.1. We omit the proof, as it is nearly identical.

THEOREM 6.3. *Let Model Assumption 3.2 and Step Assumption 6.1 be satisfied, and $\psi(\cdot; x_k)$ be uniformly prox-bounded. If Algorithm 6.1 only generates finitely many successful iterations, $x_k = x^*$ for sufficiently large k and x^* is first-order critical.*

According to Proposition 2.3 part 2, and the identification $\nu = \sigma^{-1}$, $p(\sigma; x)$ increases as σ increases, so that $\xi(\sigma; x)$ decreases as σ increases, and (6.10) yields

$$\xi(\sigma_k; x_k) \geq \xi(\sigma_{\max}; x_k) \quad \text{for all } k \in \mathbb{N}. \tag{6.11}$$

Lemma 6.1, (6.7) and (6.11) suggest using $\xi(\sigma_{\max}; x_k)^{\frac{1}{2}}$ as stationarity measure.

Let $\epsilon > 0$ be a tolerance set by the user and consider the sets (3.14). We are now in position to establish complexity results analogous to those obtained for Algorithm 3.1. The proof is nearly identical and is omitted.

THEOREM 6.4. *Let Model Assumption 3.2 and Step Assumption 6.1 be satisfied, and $\psi(\cdot; x_k)$ be uniformly prox-bounded. Assume there are infinitely many successful iterations and that $f(x_k) + h(x_k) \geq (f+h)_{\text{low}}$ for all $k \in \mathbb{N}$. Then, for all $\epsilon \in (0, 1)$,*

$$|\mathcal{S}(\epsilon)| = O(\epsilon^{-2}), \quad |\mathcal{U}(\epsilon)| = O(\epsilon^{-2}), \quad |\mathcal{S}(\epsilon)| + |\mathcal{U}(\epsilon)| = O(\epsilon^{-2}). \tag{6.12}$$

7. Implementation and numerical results. Algorithms 3.1 and 6.1 are implemented in Julia [5] and are available at `github.com/UW-AMO/TRNC`, along with scripts to reproduce our experiments. Our design allows the user to choose a method to compute a step, an important feature given the nonstandard $\psi_k + \chi_k$ operator.

We compare the performance of Algorithm 3.1 (TR) to other proximal quasi-Newton routines: PANOC [38] and ZeroFPR [39]. PANOC can be viewed as a proximal gradient descent scheme accelerated by limited-memory BFGS steps. It performs proximal gradient iterations with a backtracking linesearch, and then 20 quasi-Newton steps computed using the proximal gradient method. ZeroFPR is similar, but takes a fixed number of quasi-Newton steps between each proximal gradient step; it defaults to proximal gradient descent if no progress is made during the inner quasi-Newton steps. To compare, we count gradient evaluations as well as proximal operator evaluations, but in our example problems, proximal evaluations are far cheaper than gradients.

In the following experiments, we set $\psi(s; x_k) := h(x_k + s)$. Our stopping criteria for Algorithm 3.1 is $\xi(\Delta_k; x_k, \nu_k)^{1/2}$, which we use as a proxy for the first-order error measure $\nu_k^{-1}\xi(\Delta_{\min}; x_k, \nu_k)^{1/2}$ defined in (3.6). We set $\Delta_0 := 1.0$. We compute trust-region steps using the proximal-gradient (PG) method with step length chosen as in Corollary 4.3, denoted TR-PG in figures and tables. The user could choose accelerated variants for the subproblem, including our quadratic regularization procedure Algorithm 6.1 (R2), signified by TR-R2. In our experiments, the latter performed similarly to the proximal gradient method, although it typically required fewer inner iterations. We use proximal operators that include both $\psi(\cdot; x_k)$ and the indicator of the trust region as described in section 5. The criticality measure used in the inner PG iterations is the norm of the subgradient (4.2b), while that used in the R2 inner iterations is $\xi(\sigma_k; x_k)^{1/2}$, which is a proxy for $\xi(\sigma_{\max}; x_k)^{1/2}$. We set the inner tolerance to

$$\min(0.01, \xi(\Delta_k; x_k + s_{k,1}, \nu_k)^{\frac{1}{2}})\, \xi(\Delta_k; x_k + s_{k,1}, \nu_k),$$

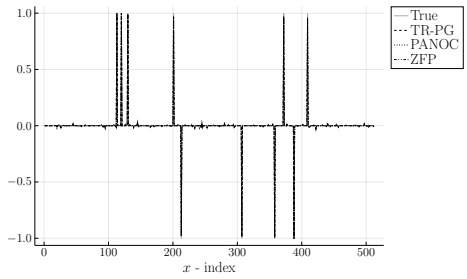
which is inspired from inexact Newton methods to encourage fast local convergence. Note that ξ is computed with the first step $s_{k,1}$ from Line 7 of Algorithm 3.1.

We use automatic differentiation as implemented in the ForwardDiff package [35] to obtain $\nabla f(x)$ and construct limited-memory quasi-Newton approximations by way of the LinearOperators package [31]. Below, we use LSR1 and LBFGS approximations with memory 5 for the BPDN and ODE examples, respectively.

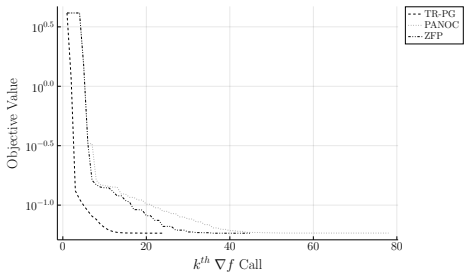
7.1. LASSO/BPDN. The first set of experiments concerns LASSO/basis pursuit de-noise (BPDN) problems, which arise in statistical [40] and compressed sensing [16] applications. We seek to recover a sparse signal $x_{\text{true}} \in \mathbb{R}^n$ given observed noisy data $b \in \mathbb{R}^m$. x_{true} is a sparse vector containing mostly zeros and 10 values of ± 1 where both the index of the nonzero entry and $\pm$ are randomly generated.

We set $m = 200$, $n = 512$, $b := Ax_{\text{true}} + \varepsilon$ where $\varepsilon \sim \mathcal{N}(0, .01)$ and A to have orthonormal rows—A^T is generated by taking the Q factor in the thin QR decomposition of a random $n \times m$ matrix. To recover x, we solve

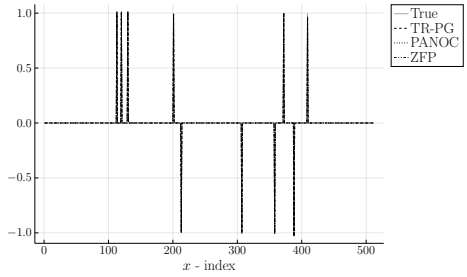
$$\underset{x}{\text{minimize}}\ \tfrac{1}{2}\|Ax - b\|_2^2 + h(x). \tag{7.1}$$



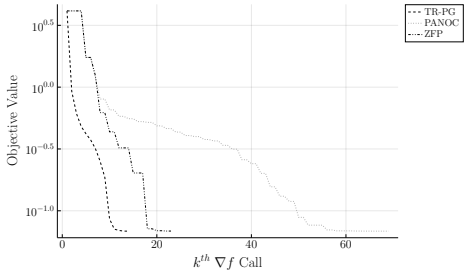
(a) Signal: $h = \lambda\|\cdot\|_1$, $\Delta\mathbb{B}_2$



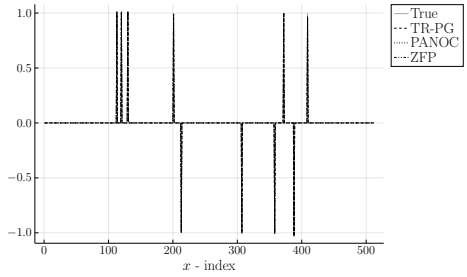
(b) History: $h = \lambda\|\cdot\|_1$, $\Delta\mathbb{B}_2$



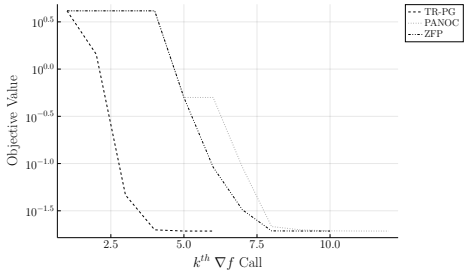
(c) Signal: $h = \lambda\|\cdot\|_0$, $\Delta\mathbb{B}_\infty$



(d) History: $h = \lambda\|\cdot\|_0$, $\Delta\mathbb{B}_\infty$



(e) Signal: $h = \chi(\cdot;\lambda\mathbb{B}_0)$, $\Delta\mathbb{B}_\infty$



(f) History: $h = \chi(\cdot;\lambda\mathbb{B}_0)$, $\Delta\mathbb{B}_\infty$

FIG. 7.1. *BPDN results* (7.1) *with Algorithm* 3.1 *and a proximal gradient subsolver (TR-PG), PANOC and ZeroFPR (ZFP): Signal plots (left) and objective value history (right).* $\Delta\mathbb{B}_p$ *indicates the norm used to define the trust region.*

We first consider $h(x) = \lambda\|x\|_p$ for $p \in \{0, 1\}$ with $\lambda = 0.1\|A^T b\|_\infty$ in the vein of [41], and employ both the ℓ_2 and ℓ_∞ norms to define the trust region. We also consider $h(x) = \chi(x; \lambda\mathbb{B}_0)$ with $\lambda = 10$ and an ℓ_∞-norm trust region. We set the maximum number of inner iterations to 5000 and $\epsilon = 10^{-3}$. The quasi-Newton model is defined by a limited-memory SR1 approximation with memory 5. All algorithms use $x_0 = 0$.

Table 7.1 and Figure 7.1 summarize our results. Table 7.1 shows that Algorithm 3.1 performs comparably to PANOC and ZeroFPR in terms of parameter fit, it performs significantly fewer gradient evaluations and significantly more proximal operator evaluations. Thus there is an advantage when proximal evaluations are cheap relative to gradient evaluations, especially in situations where the proximal operator of ψ is simpler or cheaper than that of h. All algorithms yield nearly identical solution

quality. The objective value history in Figure 7.1 shows a steeper initial decrease for Algorithm 3.1 with shorter tails in all cases. Results with R2 as subproblem solver are nearly identical though R2 performed fewer inner iterations than PG.

TABLE 7.1
BPDN results (7.1) with Algorithm 3.1 and a proximal gradient subsolver (TR-PG), PANOC and ZeroFPR (ZFP). $\Delta\mathbb{B}_p$ indicates the norm used in the trust region. The true value of $h(\cdot)/\lambda$ is 10 for $\|\cdot\|_1$ and $\|\cdot\|_0$, but 0 for $\chi(\cdot;\lambda\mathbb{B}_0)$.

		$h=\lambda\|\cdot\|_1$, $\Delta\mathbb{B}_2$			$h=\lambda\|\cdot\|_0$, $\Delta\mathbb{B}_\infty$			$h=\chi(\cdot;\lambda\mathbb{B}_0)$, $\Delta\mathbb{B}_\infty$		
	True	TR-PG	PANOC	ZFP	TR-PG	PANOC	ZFP	TR-PG	PANOC	ZFP
$f(x)$	0.020	0.005	0.005	0.005	0.019	0.019	0.019	0.019	0.019	0.019
$h(x)/\lambda$	10/0	10.750	10.767	10.750	10	10	10	0	0	0
$\|x-x_{\text{true}}\|_2/\|A\|$	0	0.134	0.141	0.133	0.055	0.055	0.056	0.054	0.056	0.055
∇f evals		24	78	45	14	69	23	6	12	10
$\text{prox}_{\nu\psi}$ calls		270	52	95	90	36	57	32	6	14

7.2. A nonlinear inverse problem. We next consider an inverse problem consisting in recovering the regularized solution to a system of nonlinear ODEs. We seek parameters $x_{\text{true}} \in \mathbb{R}^n$ given observed noisy data $b = F(x_{\text{true}}) + \varepsilon$ where $F : \mathbb{R}^n \to \mathbb{R}^m$ and $\varepsilon \sim \mathcal{N}(0, 0.1)$. The data generating mechanism F is given by the FitzHugh [18] and Nagumo et al. [29] model for neuron activation

$$\frac{\mathrm{d}V}{\mathrm{d}t} = (V - V^3/3 - W + x_1)x_2^{-1}, \quad \frac{\mathrm{d}W}{\mathrm{d}t} = x_2(x_3V - x_4W + x_5), \tag{7.2}$$

which, if $x_1 = x_4 = x_5 = 0$, becomes the Van der Pol [42] oscillator

$$\frac{\mathrm{d}V}{\mathrm{d}t} = (V - V^3/3 - W)x_2^{-1}, \quad \frac{\mathrm{d}W}{\mathrm{d}t} = x_2(x_3V). \tag{7.3}$$

Both models are highly nonlinear and ill-conditioned.

We use initial conditions $(V, W) = (2, 0)$ and discretize the time interval $[0, 20]$ at 0.2 second increments. For given x, let $V(t; x)$ and $W(t; x)$ be solutions of (7.2). Define variables $v_i(x) \approx V(t_i; x)$, $w_i(x) \approx W(t_i; x)$, $i = 1, \ldots, n+1$ where $n = 20/0.2 = 100$. We set $F(x) := (v(x), w(x))$, where $v(x) := (v_1(x), \ldots, v_{n+1}(x))$ and $w(x) := (w_1(x), \ldots, w_{n+1}(x))$. We generate b using $x_{\text{true}} = (0, 0.2, 1, 0, 0)$, which corresponds to a solve of the Van der Pol oscillator. To recover x, we solve

$$\underset{x}{\text{minimize}}\ \tfrac{1}{2}\|F(x) - b\|_2^2 + h(x), \tag{7.4}$$

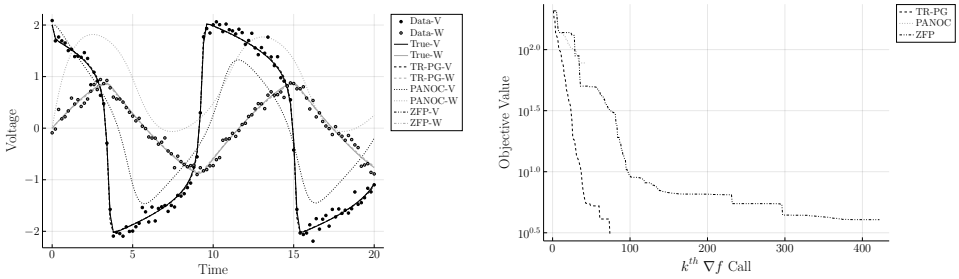
with $h(x) = \|x\|_0$. ODE solves are performed with the DifferentialEquations.jl package [34], which features an mechanism for choosing the solver, and provides $\nabla v(x)$ and $\nabla w(x)$ by way of automatic differentiation. We set $\epsilon = 10^{-3}$ in all methods, the maximum iterations to 500, and use an LBFGS approximation of the Hessian. For Algorithm 3.1, the maximum number of inner iterations is 5000.

Table 7.2 summarizes our results and Figure 7.2 shows overall data fit and objective function traces. Algorithm 3.1 with either PG or R2 as subsolver, as well as ZeroFPR, correctly identified the nonzero pattern of x with reasonable error in the nonzero elements. PANOC performs well initially, but its linesearch routine terminates prematurely as it generates a step length that is below a preset tolerance of 10^{-7}. At that point, PANOC terminates. ZeroFPR performs well, but needs many iterations to decrease the objective value to the same level as Algorithm 3.1. As

in section 7.1, Algorithm 3.1 converges with significantly fewer gradient evaluations than ZeroFPR, though with a significant number of proximal operator evaluations. However, gradient evaluations in (7.2) are far more expensive and time consuming than proximal evaluations. Figure 7.2 also reveals that the final iterate generated by Algorithm 3.1 and ZeroFPR results in trajectories that are visually indistinguishable from those associated with the exact solution. Algorithm 3.1 with Algorithm 6.1 as a subsolver reaches a similar solution as ZeroFPR, but requires much fewer proximal and gradient evaluations. The results appear in Table 7.2. Plots are nearly identical to those in Figure 7.2, and are hence omitted.

TABLE 7.2
Results for Algorithm 3.1 with proximal gradient (TR-PG) and Algorithm 6.1 (TR-R2) subsolvers, PANOC, and ZeroFPR applied to (7.2) with $h = \|\cdot\|_0$, $\Delta\mathbb{B}_\infty$ and LBFGS approximation.

Parameters										
True	TR-PG	TR-R2	PANOC	ZFP	Measure	True	TR-PG	TR-R2	PANOC	ZFP
0	0	0	0.840	0	$f(x)$	1.058	1.078	1.266	73.888	1.048
0.2	0.170	0.130	0.690	0.188	$h(x)$	2	2	3	5	3
1.0	1.136	1.408	0.952	1.048	$\|\|x - x_{\text{true}}\|\|_2$	0	0.139	0.427	1.636	0.051
0	0	0.107	0.983	0.010	∇f evals		76	61	43	422
0	0	0	0.874	0	$\text{prox}_{\nu\psi}$ calls		60143	22617	30	421



(a) Solution with data

(b) Objective Function (7.4) history

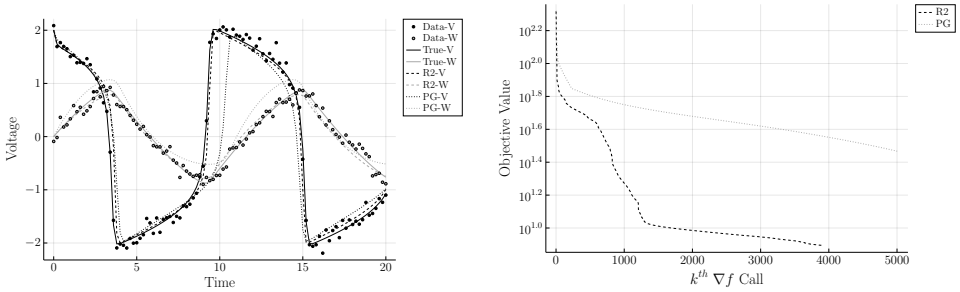
FIG. 7.2. *Solution of (7.2) with $h(x) = \|x\|_0$ in (7.4), $\Delta\mathbb{B}_\infty$ and LBFGS approximation.*

We also compare Algorithm 6.1 to our own implementation of a standard proximal gradient with linesearch on (7.4). We set the stopping tolerance for both to 10^{-3}. Table 7.3 summarizes our results and Figure 7.3 shows overall data fit and objective function traces. Both Algorithm 6.1 and proximal gradient descent converge much slower than Algorithm 3.1, where we use curvature information. Neither algorithm correctly identified the nonzero pattern of x within 5000 iterations, although Algorithm 6.1 descends considerably faster than proximal gradient descent, and attains the stopping tolerance. Figure 7.3 reveals that the final iterate generated by Algorithm 6.1 is closer to the solution than that of proximal gradient descent, though both terminated far from the correct answer.

8. Discussion and perspectives. We demonstrated the performance of trust-region methods using quasi-Newton models against two linesearch methods constrained to LBFGS models, and observed faster convergence curves with fewer gradient evaluations. Many regularizers in (1.1) have a closed-form or efficiently-computable proximal

TABLE 7.3
Results for Algorithm 6.1 (R2) and proximal gradient descent (PG) applied to (7.2) with $h = \|\cdot\|_0$.

Parameters						
True	R2	PG	Measure	True	R2	PG
0	0	0.228	$f(x)$	1.058	3.852	24.246
0.200	0.142	0.245	$h(x)$	2	4	5
1.000	1.392	1.083	$\|\|x - x_{\text{true}}\|\|_2$	0	0.737	1.045
0	0.621	0.916	∇f evals		3892	5010
0	0.022	0.440	$\text{prox}_{\nu\psi}$ calls		8891	5009



(a) Solution with data

(b) Objective Function (7.4) history

FIG. 7.3. *Solution of (7.2) for* $h = \|\cdot\|_0$ *in (7.4) with PG with linesearch and Algorithm 6.1.*

operator, whose cost is often dominated by that of a function or gradient evaluation in a large inverse problem.

The worst-case iteration complexity bound of Algorithm 3.1 matches the best known bound for trust-region methods in smooth optimization. Algorithm 6.1, a first-order method that is related to the proximal gradient method with adaptive steplength, does not require prior knowledge or estimation of a Lipschitz constant, and has a straightforward complexity analysis similar to that of Algorithm 3.1. In practice, using curvature information in Algorithm 3.1 proved useful for efficiently estimating highly nonlinear nonsmooth models. Convergence of trust-region methods for smooth optimization can be established even if Hessian approximations are unbounded, provided they do not deteriorate too fast. It may be possible to generalize our analysis along similar lines.

Interesting directions left to future work include implementation and analysis for *inexact* function, gradient, and proximal operator evaluations, and extensions of our results to cubic regularization, and more general nonlinear stepsize control-type methods, such as those of [20].

REFERENCES

[1] A. Aravkin and D. Davis. Trimmed statistical estimation via variance reduction. *Math. Oper. Res.*, 45(1):292–322, 2020.
[2] R. Baraldi, R. Kumar, and A. Aravkin. Basis pursuit denoise with nonsmooth constraints. *IEEE T. Signal Proces.*, 67(22):5811–5823, 2019.
[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer Science, 2011.
[4] A. Beck. *First Order Methods in Optimization*. SIAM, Philadelphia, USA, 2017.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1):65–98, 2017.

[6] T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. A.*, 27(3):265–274, 2009.

[7] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, (146):459—-494, 2014.

[8] R. I. Boţ, E. R. Csetnek, and S. László. An inertial forward–backward algorithm for the minimization of the sum of two nonconvex functions. *EURO J. Comput. Optim.*, (4):3–25, 2016.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. *SIAM J. Optim.*, 21(4): 1721–1739, 2011.

[10] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, pages 185–212. Springer, 2011.

[11] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. *Trust-Region Methods*. Number 1 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, 2000.

[12] F. Curtis, Z. Lubberts, and D. Robinson. Concise complexity analyses for trust region methods. *Optim. Lett.*, (12):1713—-1724, 2018.

[13] F. E. Curtis, D. P. Robinson, and M. Samadi. A trust region algorithm with a worst-case iteration complexity of $\mathcal{O}(\epsilon^{-3/2})$ for nonconvex optimization. *Math. Program., Series A*, (162): 1–32, 2017.

[14] J. Dennis, S. Li, and R. Tapia. A unified approach to global convergence of trust region methods for nonsmooth optimization. *Math. Program.*, (68):319—-346, 1995.

[15] J. E. Dennis Jr. and H. H. W. Mei. Two new unconstrained optimization algorithms which use function and gradient values. *J. Optim. Theory and Applics.*, 28:453—-482, 1979.

[16] D. L. Donoho. Compressed sensing. *IEEE T. Inform. Theory*, 52(4):1289–1306, 2006.

[17] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, 96(456):1348–1360, 2001.

[18] R. FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *B. Math. Biophys.*, 17(4):257–278, 1955.

[19] H.-Y. Gao and A. G. Bruce. Waveshrink with firm shrinkage. *Stat. Sinica*, 7:855–874, 1997.

[20] G. Grapiglia, J. Yuan, and Y. Yuan. Nonlinear stepsize control algorithms: Complexity bounds for first- and second-order optimality. *J. Optim. Theory and Applics.*, (171):980—997, 2016.

[21] W. Hare and C. Sagastizábal. Computing proximal points of nonconvex functions. *Math. Program.*, 116(1):221–258, Jan 2009.

[22] D. Kim, S. Sra, and I. S. Dhillon. A scalable trust-region algorithm with application to mixed-norm regression. In *ICML*, pages 519–526, 2010.

[23] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal Newton-type methods for minimizing composite functions. *SIAM J. Optim.*, 24(3):1420–1443, 2014.

[24] H. Li and Z. Lin. Accelerated proximal gradient methods for nonconvex programming. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 379–387, Cambridge, MA, USA, 2015. MIT Press.

[25] P. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM J. Numer. Anal.*, 16(6):964—-979, 1979.

[26] S. Lotfi, T. Bonniot de Ruisselet, D. Orban, and A. Lodi. Stochastic damped L-BFGS with controlled norm of the Hessian approximation. 2020. OPT2020 Conference on Optimization for Machine Learning.

[27] J. M. Martínez and A. C. Moretti. A trust region method for minimization of nonsmooth functions with linear constraints. *Math. Program.*, (76):431–449, 1997.

[28] J. J. Moré and D. C. Sorensen. Computing a trust region step. *SIAM J. Sci. and Statist. Comput.*, 4(3):553–572, 1983.

[29] J. Nagumo, S. Arimoto, and S. Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

[30] Y. Nesterov. Modified Gauss–Newton scheme with worst case guarantees for global performance. *Optim. Method Softw.*, 22(3):469–483, 2007.

[31] D. Orban and A. S. Siqueira. Linearoperators.jl., February 2019.

[32] M. J. D. Powell. A new algorithm for unconstrained optimization. In J. Rosen, O. Mangasarian, and K. Ritter, editors, *Nonlinear Programming*, pages 31–65. Academic Press, 1970.

[33] L. Qi and J. Sun. A trust region algorithm for minimization of locally Lipschitzian functions. *Math. Program.*, (66):25—-43, 1994.

[34] C. Rackauckas and Q. Nie. Differentialequations.jl–a performant and feature-rich ecosystem for solving differential equations in Julia. *J. Open Res. Softw.*, 5(1), 2017.

[35] J. Revels, M. Lubin, and T. Papamarkou. Forward-mode automatic differentiation in Julia, 2016. https://arxiv.org/abs/1607.07892.
[36] R. Rockafellar and R. Wets. *Variational Analysis*, volume 317. Springer Verlag, 1998.
[37] T. Steihaug. The conjugate gradient method and trust regions in large scale optimization. *SIAM J. Numer. Anal.*, 20(3):626–637, 1983.
[38] L. Stella, A. Themelis, P. Sopasakis, and P. Patrinos. A simple and efficient algorithm for nonlinear model predictive control. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 1939–1944, 2017.
[39] A. Themelis, L. Stella, and P. Patrinos. Forward-backward envelope for the sum of two nonconvex functions: Further properties and nonmonotone linesearch algorithms. *SIAM J. Optim.*, 28(3): 2274–2303, 2018.
[40] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58 (1):267–288, 1996.
[41] E. van den Berg and M. P. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, Nov. 2008. ISSN 1064-8275.
[42] B. Van der Pol. Lxxxviii. On "relaxation-oscillations". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):978–992, 1926.
[43] Y.-X. Yuan. Conditions for convergence of trust region algorithms for nonsmooth optimization. *Math. Program.*, (31):220—228, 1985.
[44] C.-H. Zhang et al. Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.*, 38(2):894–942, 2010.
[45] P. Zheng and A. Aravkin. Relax-and-split method for nonconvex inverse problems. *Inverse Problems*, 36(9):095013, 2020.
[46] P. Zheng, T. Askham, S. L. Brunton, J. N. Kutz, and A. Y. Aravkin. A unified framework for sparse relaxed regularized regression: SR3. *IEEE Access*, 7:1404–1423, 2018.