SIAM J. OPTIM.
Vol. 21, No. 4, pp. 1721–1739

ON THE EVALUATION COMPLEXITY OF COMPOSITE FUNCTION MINIMIZATION WITH APPLICATIONS TO NONCONVEX NONLINEAR PROGRAMMING[*]

CORALIA CARTIS[†], NICHOLAS I. M. GOULD[‡], AND PHILIPPE L. TOINT[§]

Abstract. We estimate the worst-case complexity of minimizing an unconstrained, nonconvex composite objective with a structured nonsmooth term by means of some first-order methods. We find that it is unaffected by the nonsmoothness of the objective in that a first-order trust-region or quadratic regularization method applied to it takes at most $\mathcal{O}(\epsilon^{-2})$ function evaluations to reduce the size of a first-order criticality measure below ϵ. Specializing this result to the case when the composite objective is an exact penalty function allows us to consider the objective- and constraint-evaluation worst-case complexity of nonconvex equality-constrained optimization when the solution is computed using a first-order exact penalty method. We obtain that in the reasonable case when the penalty parameters are bounded, the complexity of reaching within ϵ of a KKT point is at most $\mathcal{O}(\epsilon^{-2})$ problem evaluations, which is the same in order as the function-evaluation complexity of steepest-descent methods applied to unconstrained, nonconvex smooth optimization.

Key words. nonlinear programming, nonsmooth optimization, steepest descent methods, trust region methods, quadratic regularization methods, exact penalty methods, global complexity bounds, global rate of convergence

AMS subject classifications. 90C30, 90C26, 65K05, 49M05, 49M37, 58C15, 90C60, 68Q25

DOI. 10.1137/11082381X

1. Introduction.

We consider the unconstrained minimization of the composite function

$$\Phi_h(x) := f(x) + h(c(x)), \tag{1.1}$$

where $h : \mathbb{R}^m \to \mathbb{R}$ is convex but may be nonsmooth and where $f : \mathbb{R}^n \to \mathbb{R}$ and $c : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable throughout the domain of interest but may be nonconvex. We shall be concerned with estimating the function-evaluation worst-case complexity of solving (1.1) to approximate first-order optimality from an arbitrary initial guess. We will investigate two approaches, namely, (first-order) trust-region and quadratic regularization, the latter mindful of Levenberg–Morrison–Marquardt techniques [14]. If Φ_h were differentiable, generating an iterate within ϵ of a first-order criticality measure for Φ_h could be achieved in $\mathcal{O}(\epsilon^{-2})$ function evaluations by steepest-descent [11, p. 29], by trust-region [4, 8, 9], and by quadratic-regularization techniques [11, p. 29], [2]. We show that the order of this bound stays the same for (first-order) trust-region and quadratic regularization when Φ_h has a nonsmooth component. The worst-case complexity of minimizing a composite function with a nonsmooth term by gradient methods has been addressed in [12], but there,

[*] Received by the editors February 9, 2011; accepted for publication (in revised form) September 12, 2011; published electronically December 22, 2011. This work was supported by the Royal Society through the International Joint Project 14265.
http://www.siam.org/journals/siopt/21-4/82381.html

[†] School of Mathematics, University of Edinburgh, The King's Buildings, Edinburgh, EH9 3JZ, Scotland, UK (coralia.cartis@ed.ac.uk).

[‡] Computational Science and Engineering Department, Rutherford Appleton Laboratory, Chilton, Oxfordshire, OX11 0QX, England, UK (nick.gould@stfc.ac.uk). This author's work was supported by EPSRC grant EP/E053351/1.

[§] Department of Mathematics, FUNDP - University of Namur, 61, rue de Bruxelles, B-5000, Namur, Belgium (philippe.toint@fundp.ac.be).

the nonsmooth term is assumed to be convex. By contrast, the nonsmooth term of Φ_h is here a composition of the convex nonsmooth function h with the nonconvex smooth vector-valued function $c(x)$. Similarly, the global rate of convergence of solving a system of nonlinear equations by means of a sharp (potentially nonsmooth) merit function and quadratic regularization has been investigated in [13]. There, a worst-case bound of order $\mathcal{O}(\epsilon^{-2})$ was obtained for the general nonconvex case and then further improved to reflect fast local convergence in the case of zero-residual problems and uniformly nondegenerate Jacobians. These results and the proposed quadratic regularization techniques apply directly to instances (1.1) when $f = 0$ and h is the Euclidean or some other norm; here, we address a more general framework by imposing fewer requirements on h than in [13] and allowing the addition of the smooth objective term f.

An illustrative example of (1.1) is the exact penalty function

$$\Phi(x, \rho) = f(x) + \rho\|c(x)\|, \tag{1.2}$$

with the penalty parameter $\rho > 0$ and associated to the equality-constrained optimization problem

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} f(x) \text{ subject to } c(x) = 0, \tag{1.3}$$

where $m \leq n$. We can now make use of the above-mentioned algorithms and their complexity bounds when applied to (1.2) so as to estimate the worst-case problem—that is, objective and constraints—evaluation complexity of generating an approximate solution of (1.3) by means of an exact penalty method, noting that each function evaluation of the penalty function $\Phi(\cdot, \rho)$ requires one evaluation of the objective and constraints of (1.3). To the best of our knowledge, the results presented here are the first worst-case global evaluation bounds for constrained optimization when both the objective and the constraints are allowed to be nonconvex.

For approximate optimality for problem (1.3), we are content with getting sufficiently close to a KKT point of our problem (1.3), namely, to any x_* satisfying

$$g(x_*) + J(x_*)^T y_* = 0 \text{ and } c(x_*) = 0 \tag{1.4}$$

for some Lagrange multiplier $y_* \in \mathbb{R}^m$, where g denotes the gradient of f, and J the Jacobian of the constraints c. Recall that the KKT points (1.4) of (1.3) correspond to critical points of (1.2) for sufficiently large ρ, provided usual constraint qualifications hold [1,6,14]. The exact penalty algorithm for solving (1.3) proceeds by sequentially minimizing the penalty function (1.2) using the trust-region or quadratic-regularization approach, and then adaptively increasing the penalty parameter ρ through a steering procedure [1]. We obtain that when the penalty parameter is bounded—which is a reasonable assumption since the penalty is exact—the exact penalty algorithm takes at most $\mathcal{O}(\epsilon^{-2})$ total problem evaluations to satisfy the KKT conditions (1.4) within ϵ or reach within ϵ of an infeasible (first-order) critical point of the feasibility measure $\|c(x)\|$. Otherwise, when the penalty parameter grows unbounded, the algorithm takes at most $\mathcal{O}(\epsilon^{-5})$ total problem evaluations to satisfy the same approximate optimality conditions.

The above exact penalty approach can be extended to problems that also have finitely many inequality constraints by choosing h in (1.1) appropriately. In particular, if the constraints $\tilde{c}(x) \geq 0$ are added to (1.3), then it is suitable to replace the penalty term in (1.2) by $h((c, \tilde{c})(x)) = \rho\|c(x)\| + \rho\|\tilde{c}^-(x)\|$, where $\tilde{c}^-(x)$ is defined componentwise as $\tilde{c}^-(x) \overset{\text{def}}{=} \min\{\tilde{c}_i(x), 0\}$.

The structure of the paper is as follows. Sections 2.1 and 2.2 address the global evaluation complexity of minimizing a composite nonconvex function that may have a nonsmooth term, by employing a first-order trust-region and quadratic regularization method, respectively. Then by letting the composite function be the exact penalty function (1.2), section 3.1 connects the approximate critical points of (1.2) to approximate KKT points of (1.3), while section 3.2 applies the complexity results in section 2 in the context of an exact penalty algorithm for problem (1.3), to deduce a bound on the worst-case complexity of the latter. We draw our conclusions in section 4.

2. Function-evaluation complexity for composite nonsmooth unconstrained minimization.

Let us consider the unconstrained minimization of the general function (1.1), where h may be nonsmooth. The following assumptions will be required throughout, namely,

AF.1 $$f,\, c_i \in \mathcal{C}^1(\mathbb{R}^n),\ i \in \{1, \ldots, m\},$$

and, letting g denote the gradient of f, and $J(x)$ the Jacobian of c at x,

AF.2 g and J are Lipschitz continuous on the path of the iterates and trial points $[x_k, x_k + s_k]$ for all k, with constants $L_g \geq 1$ and L_J, respectively, independent of k.

Similarly, for h, we assume that

AH.1 h is convex and globally Lipschitz continuous, with Lipschitz constant L_h.

Note that h being convex implies that h is globally Lipschitz continuous at all required points (in the results that follow), provided the iterates lie in a bounded set or h is bounded above and below on $\mathbb{R}^n$ [10, pp. 173–174]. In the case of (1.2), $h \overset{\text{def}}{=} \rho\|\cdot\|$ and so AH.1 holds with $L_h = \rho$.

We consider linearizing the argument of Φ_h around (any) x to obtain the approximation

$$l(x, s) \overset{\text{def}}{=} f(x) + g(x)^T s + h(c(x) + J(x)s), \quad s \in \mathbb{R}^n. \tag{2.1}$$

An appropriate *criticality measure* for Φ_h is the quantity

$$\Psi(x) \overset{\text{def}}{=} l(x, 0) - \min_{\|s\| \leq 1} l(x, s). \tag{2.2}$$

In particular, following [1, 15], $\Psi(x)$ is continuous for all x, and we say that x_* is a critical point of Φ_h if

$$\Psi(x_*) = 0. \tag{2.3}$$

Note that other first-order necessary optimality conditions for Φ_h such as [6, pp. 369] can be shown to be equivalent to (2.3) [15, Lemma 2.1]. Note also the connection of (2.2) to the criticality measure for smooth constrained optimization $\chi(x)$ in [5, section 12.1.4] that we employed for the complexity analysis of cubic regularization variants for convex-constrained problems [3].

We will investigate two techniques, namely, first-order trust region and quadratic regularization for minimizing Φ_h. These algorithms generate a sequence of iterates $\{x_k\}$ and trial steps $\{s_k\}$ from a given initial point x_0. At each iterate x_k, we let

$$f_k \overset{\text{def}}{=} f(x_k), \quad g_k \overset{\text{def}}{=} g(x_k), \quad J_k \overset{\text{def}}{=} J(x_k), \text{ and } \Psi_k \overset{\text{def}}{=} \Psi(x_k).$$

On the basis of (2.3), we will terminate each method as soon as we find an iterate for which $\Psi_k \leq \epsilon$, where $\epsilon > 0$ is a(ny) user-defined accuracy tolerance. We will address the global function-evaluation complexity of these methods until termination is achieved. Note that each of the algorithms applied to Φ_h requires one evaluation of Φ_h per each iteration, or, equivalently, one objective and constraints evaluation of problem (1.3), while only the so-called (very) successful iterations, when the trial step s_k is employed in forming the new iterate, evaluate the gradients of f and c.

2.1. A trust-region approach.

Let us now apply a (first-order) trust-region method to minimizing Φ_h, which is summarized in Algorithm 2.1. At each iterate k, the trial step s_k is computed as the solution of the trust-region subproblem

$$\min_{s \in \mathbb{R}^n} l(x_k, s) \text{ subject to } \|s\| \le \Delta_k, \tag{2.4}$$

where $l(x_k, s)$ is defined in (2.1). Since h is convex, (2.1) implies that the subproblem (2.4) is also convex. Thus, provided that h is computationally inexpensive to minimize, the cost of computing s_k is acceptable. In particular, if $h = \|\cdot\|$ is a polyhedral norm, then (2.4) can be solved as a linear programming problem. Note also that the solution of (2.4) does not require additional problem evaluations to those already computed for constructing the model (2.1) of Φ.

The radius Δ_k is adjusted, and the new iterate constructed, according to standard trust-region rules based on the value of the ratio r_k of the actual function decrease $\Phi_h(x_k) - \Phi_h(x_k + s_k)$ to the optimal model decrease, namely,

$$\Psi(x_k, \Delta_k) \stackrel{\text{def}}{=} l(x_k, 0) - \min_{\|s\| \le \Delta_k} l(x_k, s) = l(x_k, 0) - l(x_k, s_k); \tag{2.5}$$

note that $l(x_k, 0) = \Phi_h(x_k)$. Note also the connection between the optimality measure (2.2) and the model decrease (2.5), namely, $\Psi_k = \Psi(x_k, 1)$.

```
Algorithm 2.1: A trust-region algorithm for minimizing $\Phi_h$.

Step 0: Initialization. Initial data: $x_0$, $\Delta_0$, $0 < \eta_1 \le \eta_2 < 1$, $0 < \gamma_1 \le \gamma_2 < 1$,
        $\epsilon > 0$. Set $k = 0$.
While $\Psi_k > \epsilon$, do:
Step 1: Step calculation. Compute the step $s_k$ as the solution of (2.4).
Step 2: Acceptance of trial point. Compute $\Phi_h(x_k + s_k)$ and define

        (2.6)   $r_k = \frac{\Phi_h(x_k) - \Phi_h(x_k + s_k)}{\Psi(x_k, \Delta_k)}$,

        where $\Psi(x_k, \Delta_k)$ is defined in (2.5).
        If $r_k \ge \eta_1$, then $x_{k+1} = x_k + s_k$; else, $x_{k+1} = x_k$.
Step 3: Trust-region radius update. Set
        (2.7)
        $\Delta_{k+1} \in \begin{cases} [\Delta_k, \infty) & \text{if } r_k \ge \eta_2, & [k \text{ very successful}] \\ [\gamma_2 \Delta_k, \Delta_k] & \text{if } r_k \in [\eta_1, \eta_2), & [k \text{ successful}] \\ [\gamma_1 \Delta_k, \gamma_2 \Delta_k] & \text{if } r_k < \eta_1. & [k \text{ unsuccessful}] \end{cases}$

        Increment $k$ by one and return to Step 1.
```

Now, we investigate the function-evaluation complexity of Algorithm 2.1 generating $\Psi_k \le \epsilon$. Our results follow similarly to those in [1, section 3].

LEMMA 2.1 (see [1, Lemma 3.1]). *Let AF.1 and AH.1 hold. Then*

$$\Psi(x_k, \Delta_k) \geq \min\{\Delta_k, 1\}\Psi_k. \tag{2.8}$$

Proof. Assume first that $\Delta_k \geq 1$. Then

$$\min_{\|s\|\leq 1} l(x_k, s) \geq \min_{\|s\|\leq \Delta_k} l(x_k, s),$$

and so $\Psi_k \leq \Psi(x_k, \Delta_k)$, which proves (2.8) in this case since $\min\{\Delta_k, 1\} = 1$.

Now let $\Delta_k < 1$ and $s_k^* \overset{\text{def}}{=} \arg\min_{\|s\|\leq 1} l(x_k, s)$. Then $\|\Delta_k s_k^*\| \leq \Delta_k$ and so $l(x_k, s_k) \leq l(x_k, \Delta_k s_k^*)$, implying

$$\Psi(x_k, \Delta_k) \geq l(x_k, 0) - l(x_k, \Delta_k s_k^*) \geq \Delta_k[l(x_k, 0) - l(x_k, s_k^*)] = \Delta_k \Psi_k,$$

where the second inequality follows from $\Delta_k \leq 1$ and l in (2.1) being convex due to AH.1. □

The next lemmas deduce a lower bound on Δ_k.

LEMMA 2.2. *Let AF.1, AF.2, and AH.1 hold. Then, provided $\Psi_k \neq 0$, we have that*

$$\Delta_k \leq \kappa_L \sqrt{\Psi_k} \min\{1, \sqrt{\Psi_k}\} \quad \Longrightarrow \quad k \textit{ is very successful in the sense of } (2.7), \tag{2.9}$$

where

$$\kappa_L \overset{\text{def}}{=} \frac{1-\eta_2}{L_g + \frac{1}{2}L_h L_J}. \tag{2.10}$$

Proof. From (1.1), (2.6), (2.5), and (2.1), we have

$$\begin{aligned}
|r_k - 1| &= \frac{1}{\Psi(x_k, \Delta_k)} \left|\Phi_h(x_k + s_k) - l(x_k, s_k)\right| \\
&= \frac{1}{\Psi(x_k, \Delta_k)} \left|f(x_k + s_k) - f_k - g_k^T s_k + h(c(x_k + s_k)) - h(c_k + J_k s_k)\right| \\
&\leq \frac{1}{\Psi(x_k, \Delta_k)} \left\{\left|f(x_k + s_k) - f_k - g_k^T s_k\right| + \left|h(c(x_k + s_k)) - h(c_k + J_k s_k)\right|\right\}.
\end{aligned}$$

The Taylor expansions $f(x_k + s_k) = f_k + g(\xi_k)^T s_k$ for some $\xi_k \in [x_k, x_k + s_k]$ and $c(x_k + s_k) = c_k + \int_0^1 J(x_k + ts_k)s_k dt$ imply, together with AF.2 and AH.1, that

$$\begin{gathered}
\left|f(x_k + s_k) - f_k - g_k^T s_k\right| \leq L_g \|s_k\|^2 \text{ and} \\
|h(c(x_k + s_k)) - h(c_k + J_k s_k)| \leq \tfrac{1}{2} L_h L_J \|s_k\|^2.
\end{gathered}$$

From (2.10), it follows that

$$|r_k - 1| \leq \frac{(1-\eta_2)\|s_k\|^2}{\kappa_{\mathrm{L}} \Psi(x_k, \Delta_k)} \leq \frac{1-\eta_2}{\kappa_{\mathrm{L}}} \cdot \frac{\Delta_k^2}{\min\{\Delta_k, 1\}\Psi_k},$$

where in the second inequality, we used $\|s_k\| \leq \Delta_k$ and (2.8). The implication (2.9) now follows from (2.7) and $\kappa_{\mathrm{L}} \leq 1$, the latter being provided by $L_g \geq 1$ and $\eta_2 \in (0, 1)$. □

LEMMA 2.3. *Let AF.1, AF.2, and AH.1 hold. Also, let $\epsilon \in (0,1]$ such that*

$$\Psi_k > \epsilon \text{ for all } k = 0, \dots, j, \tag{2.11}$$

where $j \leq \infty$. Then

$$\Delta_k \geq \min\{\Delta_0, \gamma_1 \kappa_L \epsilon\} \text{ for all } k = 0, \dots, j, \tag{2.12}$$

where κ_L is defined in (2.10).

Proof. For any $k \in \{0, \dots, j\}$, $\epsilon \in (0,1]$ and (2.11) give

$$\kappa_{\mathrm{L}}\epsilon = \kappa_{\mathrm{L}}\sqrt{\epsilon}\min\{1, \sqrt{\epsilon}\} \leq \kappa_{\mathrm{L}}\sqrt{\Psi_k}\min\{1, \sqrt{\Psi_k}\}$$

and so Lemma 2.2 and (2.7) provide the implication

$$\Delta_k \leq \kappa_{\mathrm{L}}\epsilon \quad \Longrightarrow \quad \Delta_{k+1} \geq \Delta_k. \tag{2.13}$$

Thus when $\Delta_0 \geq \gamma_1\kappa_{\mathrm{L}}\epsilon$, (2.13) implies that $\Delta_k \geq \gamma_1\kappa_{\mathrm{L}}\epsilon$ for all $k \in \{0, \dots, j\}$, where the factor γ_1 is introduced for the case when Δ_k is greater than $\kappa_{\mathrm{L}}\epsilon$ and iteration k is not very successful. Letting $k = 0$ in (2.13) gives (2.12) when $\Delta_0 < \gamma_1\kappa_{\mathrm{L}}\epsilon$ since $\gamma_1 \in (0,1)$. □

We are now ready to give the main result of this section.

THEOREM 2.4. *Let AF.1, AF.2, and AH.1 hold, and let $\{\Phi_h(x_k)\}$ be bounded below by Φ_h^{low}. Given any $\epsilon \in (0,1]$, assume that $\Psi_0 > \epsilon$, and let $j_1 \leq \infty$ be the first iteration such that $\Psi_{j_1+1} \leq \epsilon$. Then the trust-region algorithm, Algorithm* 2.1, *takes at most*

$$J_1^s \stackrel{\text{def}}{=} \lceil \kappa_{TR}^s \epsilon^{-2} \rceil$$

successful iterations, or, equivalently, problem-gradient evaluations, to generate $\Psi_{j_1+1} \leq \epsilon$, where

$$\kappa_{TR}^s \stackrel{\text{def}}{=} \frac{\Phi_h(x_0) - \Phi_h^{\mathrm{low}}}{\eta_1 \min\{\Delta_0, \gamma_1\kappa_L\}}, \tag{2.14}$$

where κ_L is defined in (2.10).

Additionally, assume that on each very successful iteration k, Δ_{k+1} is chosen such that

$$\Delta_{k+1} \leq \gamma_3 \Delta_k \tag{2.15}$$

for some $\gamma_3 > 1$. Then

$$j_1 \leq \lceil \kappa_{TR}\epsilon^{-2} \rceil \stackrel{\text{def}}{=} J_1, \tag{2.16}$$

and so Algorithm 2.1 *takes at most J_1 (successful and unsuccessful) iterations, or, equivalently, problem evaluations, to generate $\Psi_{j_1+1} \leq \epsilon$, where*

$$\kappa_{TR} \stackrel{\text{def}}{=} \kappa_{TR}^s\left(1 - \frac{\log \gamma_3}{\log \gamma_2}\right) + \frac{1}{|\log \gamma_2|} \cdot \frac{\Delta_0}{\gamma_1 \kappa_L}. \tag{2.17}$$

Proof. The definition of j_1 in the statement of the theorem is equivalent to

$$\Psi_k > \epsilon \text{ for all } k = 0, \dots, j_1, \quad \text{and} \quad \Psi_{j_1+1} \leq \epsilon. \tag{2.18}$$

Thus Lemma 2.3 applies with $j = j_1$. It follows from (2.8) and (2.12) that

$$\Psi(x_k, \Delta_k) \geq \min\{1, \Delta_0, \gamma_1 \kappa_{\text{L}} \epsilon\} \Psi_k, \quad k = 0, \ldots, j_1,$$

which further becomes, due to ϵ, $\kappa_{\text{L}} \in (0, 1]$, $\gamma_1 \in (0, 1)$, and again (2.18),

$$\Psi(x_k, \Delta_k) \geq \min\{\Delta_0, \gamma_1 \kappa_{\text{L}}\} \epsilon^2, \quad k = 0, \ldots, j_1. \tag{2.19}$$

Now let $k \in \mathcal{S} \cap \{0, \ldots, j_1\}$, where $\mathcal{S}$ denotes the set of all successful or very successful iterations in the sense of (2.7). Then (2.6), (2.7), and (2.19) imply

$$\Phi_h(x_k) - \Phi_h(x_k + s_k) \geq \eta_1 \Psi(x_k, \Delta_k) \geq \eta_1 \min\{\Delta_0, \gamma_1 \kappa_{\text{L}}\} \epsilon^2. \tag{2.20}$$

Summing up (2.20) over $k \in \{0, \ldots, j_1\}$, recalling that function values remain unchanged on unsuccessful iterations and that $\Phi_h(x_{j_1}) \geq \Phi_h^{\text{low}}$, we get

$$\Phi_h(x_0) - \Phi_h^{\text{low}} \geq k_{j_1}^s \eta_1 \min\{\Delta_0, \gamma_1 \kappa_{\text{L}}\} \epsilon^2,$$

where $k_{j_1}^s$ denotes the number of successful iterations that occur up to iteration j_1. The latter gives the iteration upper bound J_1^s. To prove the bound J_1, we need to bound the number of unsuccessful iterations up to j_1. First, (2.15) implies

$$\Delta_{k+1} \leq \gamma_3 \Delta_k, \quad k \in \{0, \ldots, j_1\} \cap \mathcal{S},$$

and (2.7) gives

$$\Delta_{i+1} \leq \gamma_2 \Delta_i, \quad i \in \{0, \ldots, j_1\} \setminus \mathcal{S}.$$

Thus we deduce inductively that

$$\Delta_{j_1} \leq \Delta_0 \gamma_3^{k_{j_1}^s} \gamma_2^{k_{j_1}^u},$$

where $k_{j_1}^u$ denotes the number of unsuccessful iterations up to j_1; from (2.12), this further becomes

$$\min\left\{1, \frac{\gamma_1 \kappa_{\text{L}} \epsilon}{\Delta_0}\right\} \leq \gamma_3^{k_{j_1}^s} \gamma_2^{k_{j_1}^u},$$

and so, taking the logarithm on both sides and recalling that $\gamma_2 \in (0, 1)$, we get

$$k_{j_1}^u \leq -k_{j_1}^s \frac{\log \gamma_3}{\log \gamma_2} - \frac{1}{\log \gamma_2} \log \frac{\Delta_0}{\gamma_1 \kappa_{\text{L}} \epsilon}.$$

Hence, using also that $\log(\Delta_0/(\gamma_1 \kappa_{\text{L}} \epsilon)) \leq \Delta_0/(\gamma_1 \kappa_{\text{L}} \epsilon)$,

$$j_1 = k_{j_1}^s + k_{j_1}^u \leq k_{j_1}^s \left(1 - \frac{\log \gamma_3}{\log \gamma_2}\right) + \frac{\Delta_0}{\gamma_1 \kappa_{\text{L}} |\log \gamma_2|} \cdot \frac{1}{\epsilon},$$

which together with the bound J_1^s on $k_{j_1}^s$ and $\epsilon \in (0, 1]$ yields (2.16). □

When applying Algorithm 2.1 to (1.2) in place of Φ_h, Theorem 2.4 applies and the value of every constant stays the same in the bounds except L_h in expression (2.10), which is replaced by ρ; thus for ρ sufficiently large, $\kappa_{\text{L}} = \mathcal{O}(\rho^{-1})$ and $\Phi_h(x_0) - \Phi_h^{\text{low}} = \mathcal{O}(\rho)$, and so κ_{TR}^s and κ_{TR} are both $\mathcal{O}(\rho^2)$. Note also that to ensure that $\Phi(\cdot, \rho)$ is bounded below it is sufficient to require that f be bounded below on $\mathbb{R}^n$; both of these, however, are restrictive assumptions when related to problem (1.3), as discussed in greater detail in section 4.

2.2. A quadratic-regularization approach.

Let us now apply instead a (first-order) quadratic-regularization method to minimizing Φ_h, which is mindful of Levenberg–Morrison–Marquardt techniques; see Algorithm 2.2. Our approach and results here mirror those in [13, section 2], while employing a more general merit function due to the choice of h and the addition of the smooth objective term f.

At iteration k, the step s_k is now computed as the solution of the regularized subproblem

$$\min_{s\in\mathbb{R}^n} l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2, \tag{2.21}$$

where $l(x_k, s)$ is defined in (2.1). The cost of computing s_k is manageable for some h as (2.21) is a convex unconstrained problem with simple quadratic terms; furthermore, it does not require additional problem evaluations to those already computed for constructing the model (2.1) of Φ. The regularization weight $\sigma_k > 0$ and the new iterate are chosen adaptively, based on the value of the ratio r_k^r of the actual function decrease $\Phi_h(x_k) - \Phi_h(x_k + s_k)$ to the optimal model decrease, namely,

$$\Psi^r(x_k, \sigma_k) \overset{\text{def}}{=} l(x_k, 0) - \min_{s\in\mathbb{R}^n}\left[l(x_k, s) + \frac{\sigma_k}{2}\|s\|^2\right] = l(x_k, 0) - l(x_k, s_k) - \frac{\sigma_k}{2}\|s_k\|^2. \tag{2.22}$$

As termination criterion in Algorithm 2.2, we use the same optimality measure Ψ_k as for the trust-region approach in the previous section, namely, (2.2). Note that (2.22) with $\sigma_k = 1$ is also an optimality measure for Φ_h, but it is not scaled appropriately in that when $c = 0$, it is of order $\|g_k\|^2$ rather than $\|g_k\|$. As a result of this, using (2.22) with $\sigma_k = 1$ in the termination condition of Algorithm 2.2 worsens its complexity bound.

Algorithm 2.2: A quadratic-regularization algorithm for minimizing Φ_h.

Step 0: Initialization. Initial data: x_0, σ_0, $0 < \eta_1 \le \eta_2 < 1$, $1 < \gamma_1 \le \gamma_2$, $\epsilon > 0$. Set $k = 0$.

While $\Psi_k > \epsilon$, do:

Step 1: Step calculation. Compute the step s_k as the solution of (2.21).

Step 2: Acceptance of trial point. Compute $\Phi_h(x_k + s_k)$ and define

$$r_k^r = \frac{\Phi_h(x_k) - \Phi_h(x_k + s_k)}{\Psi^r(x_k, \sigma_k)}, \tag{2.23}$$

where $\Psi^r(x_k, \sigma_k)$ is defined in (2.22).

If $r_k^r \ge \eta_1$, then $x_{k+1} = x_k + s_k$; else, $x_{k+1} = x_k$.

Step 3: Updating the regularization weight. Set

$$\sigma_{k+1} \in \begin{cases} (0,\, \sigma_k] & \text{if } r_k^r \ge \eta_2, & [k \text{ very successful}] \\ [\sigma_k, \gamma_1\sigma_k] & \text{if } r_k^r \in [\eta_1, \eta_2), & [k \text{ successful}] \\ [\gamma_1\sigma_k, \gamma_2\sigma_k] & \text{if } r_k^r < \eta_1. & [k \text{ unsuccessful}] \end{cases} \tag{2.24}$$

Increment k by one and return to Step 1.

Now, we investigate the problem-evaluation complexity of Algorithm 2.2 generating $\Psi_k \le \epsilon$. First, we relate the model decrease $\Psi^r(x_k, \sigma_k)$ to the optimality measure Ψ_k in (2.2).

LEMMA 2.5. *Let AF.1 and AH.1 hold. Then*

$$\Psi^r(x_k,\sigma_k) \geq \frac{1}{2}\min\left\{1, \frac{\Psi_k}{\sigma_k}\right\}\Psi_k. \tag{2.25}$$

Proof. Assume first that $\sigma_k \leq \Psi_k$. Then clearly

$$\begin{aligned}\min_{s\in\mathbb{R}^n}\left[l(x_k,s)+\frac{\sigma_k}{2}\|s\|^2\right] &\leq \min_{\|s\|\leq 1}\left[l(x_k,s)+\frac{\sigma_k}{2}\|s\|^2\right]\\ &\leq \min_{\|s\|\leq 1} l(x_k,s)+\frac{\sigma_k}{2} \leq \min_{\|s\|\leq 1} l(x_k,s)+\frac{\Psi_k}{2},\end{aligned}$$

and so, from (2.22) and (2.2),

$$\Psi^r(x_k,\sigma_k) \geq l(x_k,0) - \min_{\|s\|\leq 1} l(x_k,s) - \frac{\Psi_k}{2} = \Psi_k - \frac{\Psi_k}{2} = \frac{\Psi_k}{2},$$

which proves (2.25) in the case when $\sigma_k \leq \Psi_k$.

Now let $\sigma_k > \Psi_k$ and $s_k^* \overset{\text{def}}{=} \arg\min_{\|s\|\leq 1} l(x_k,s)$. Then the definition of s_k as the solution of (2.21) implies

$$l(x_k,s_k)+\frac{\sigma_k}{2}\|s_k\|^2 \leq l\left(x_k,\frac{\Psi_k}{\sigma_k}s_k^*\right)+\frac{\sigma_k}{2}\left\|\frac{\Psi_k}{\sigma_k}s_k^*\right\|^2 \leq l\left(x_k,\frac{\Psi_k}{\sigma_k}s_k^*\right)+\frac{\Psi_k^2}{2\sigma_k},$$

where, to obtain the second inequality, we used $\|s_k^*\| \leq 1$. This and (2.22) give

$$\Psi^r(x_k,\sigma_k) \geq l(x_k,0) - l\left(x_k,\frac{\Psi_k}{\sigma_k}s_k^*\right) - \frac{\Psi_k^2}{2\sigma_k}. \tag{2.26}$$

Using $0 < \Psi_k/\sigma_k < 1$ and l in (2.1) being convex due to AH.1, we deduce

$$l\left(x_k,\frac{\Psi_k}{\sigma_k}s_k^*\right) \leq \left(1-\frac{\Psi_k}{\sigma_k}\right)l(x_k,0)+\frac{\Psi_k}{\sigma_k}l(x_k,s_k^*),$$

which substituted into (2.26) gives

$$\Psi^r(x_k,\sigma_k) \geq \frac{\Psi_k}{\sigma_k}\left[l(x_k,0)-l(x_k,s_k^*)\right]-\frac{\Psi_k^2}{2\sigma_k} = \frac{\Psi_k^2}{\sigma_k}-\frac{\Psi_k^2}{2\sigma_k} = \frac{\Psi_k^2}{2\sigma_k},$$

where we also used (2.2) and the choice of s_k^*. □

Lemma 2.5 implies that r_k^r in (2.23) is well-defined whenever the current iterate is not first-order critical, namely, $\Psi_k \neq 0$. The next lemma deduces an upper bound on σ_k.

LEMMA 2.6. *Let AF.1, AF.2, and AH.1 hold. Then*

$$\sigma_k \leq \max\{\sigma_0, \gamma_2(2L_g+L_hL_J)\} \overset{\text{def}}{=} \kappa_\sigma, \quad \textit{for all } k \geq 0. \tag{2.27}$$

Proof. Let $\kappa_{\sigma,1} \overset{\text{def}}{=} 2L_g + L_hL_J$. To prove (2.27), it is sufficient to show the implication

$$\sigma_k \geq \kappa_{\sigma,1} \quad\Longrightarrow\quad k \text{ is very successful in the sense of (2.24)}, \tag{2.28}$$

and so $\sigma_{k+1} \leq \sigma_k$. We allow the factor γ_2 in κ_σ for the case when σ_k is only slightly less than $\kappa_{\sigma,1}$ and k is not very successful, while the term σ_0 in (2.27) accounts for the choice at start-up.

To prove (2.28), note that (2.24) provides that $r_k^r \geq 1$ implies k is very successful. It follows from (2.23), $\Psi^r(x_k,\sigma_k) > 0$, (2.22), and $\Phi_h(x_k) = l(x_k,0)$ that $r_k^r \geq 1$, provided

$$D_k \stackrel{\text{def}}{=} \Phi_h(x_k+s_k) - \left[l(x_k,s_k) + \frac{\sigma_k}{2}\|s_k\|^2\right] \leq 0. \tag{2.29}$$

From (1.1) and (2.1), and Taylor expansions for f and c, we have

$$\begin{aligned} D_k &= \left[f(x_k+s_k) - f_k - g_k^T s_k\right] + \left[h(c(x_k+s_k)) - h(c_k+J_k s_k)\right] - \frac{\sigma_k}{2}\|s_k\|^2 \\ &\leq [g(\xi_k^1) - g_k]^T s_k + L_h\|c(x_k+s_k) - c_k - J_k s_k\| - \frac{\sigma_k}{2}\|s_k\|^2 \\ &\leq [g(\xi_k) - g_k]^T s_k + L_h\left\|\int_0^1 J(x_k+ts_k)s_k dt - J_k s_k\right\| - \frac{\sigma_k}{2}\|s_k\|^2, \end{aligned}$$

where $\xi_k \in (x_k, x_k+s_k)$, and where we also used AH.1 in the second inequality. Now using AF.1 and $\|\xi_k - x_k\| \leq \|s_k\|$, the last displayed inequality further becomes

$$D_k \leq \left(L_g + \tfrac{1}{2}L_h L_J - \tfrac{1}{2}\sigma_k\right)\|s_k\|^2.$$

Thus (2.29) holds whenever $\sigma_k \geq \kappa_{\sigma,1}$. □

The main result of this section follows.

THEOREM 2.7. *Let AF.1, AF.2, and AH.1 hold, and let $\{\Phi_h(x_k)\}$ be bounded below by Φ_h^{low}. Then, given any $\epsilon \in (0,1]$, the total number of successful iterations and problem-gradient evaluations with*

$$\Psi_k > \epsilon \tag{2.30}$$

that occur when applying the quadratic-regularization Algorithm 2.2 to Φ_h is at most

$$J_1^{s,r} \stackrel{\text{def}}{=} \lceil \kappa_{QR}^s \epsilon^{-2} \rceil,$$

where

$$\kappa_{QR}^s \stackrel{\text{def}}{=} 2\kappa_\sigma \eta_1^{-1}\left(\Phi_h(x_0) - \Phi_h^{\text{low}}\right), \tag{2.31}$$

with κ_σ defined in (2.27). Assuming (2.30) holds at $k=0$, Algorithm 2.2 takes at most $J_1^{s,r}+1$ successful iterations and problem-gradient evaluations to generate a first iterate, say j_1, such that $\Psi_{j_1+1} \leq \epsilon$.

Additionally, assume that on each very successful iteration k, σ_{k+1} is chosen such that

$$\sigma_{k+1} \geq \gamma_3 \sigma_k \tag{2.32}$$

for some $\gamma_3 \in (0,1)$ independent of k. Then

$$j_1 \leq \lceil \kappa_{QR}\epsilon^{-2} \rceil \stackrel{\text{def}}{=} J_1^r, \tag{2.33}$$

and so Algorithm 2.2 takes at most J_1^r (successful and unsuccessful) iterations, or, equivalently, problem evaluations, to generate $\Psi_{j_1+1} \leq \epsilon$, where

$$\kappa_{QR} \stackrel{\text{def}}{=} \kappa_{QR}^s\left(1 - \frac{\log\gamma_3}{\log\gamma_1}\right) + \frac{1}{\log\gamma_1}\log\frac{\kappa_\sigma}{\sigma_0}.$$

Proof. It follows from (2.25) and (2.27) that

$$\Psi^r(x_k, \sigma_k) \geq \frac{1}{2} \min\left\{1, \frac{\Psi_k}{\kappa_\sigma}\right\} \Psi_k, \quad k \geq 0.$$

Thus, while Algorithm 2.2 does not terminate, (2.30) and $\epsilon \leq 1$ provide

$$\Psi^r(x_k, \sigma_k) \geq \frac{1}{2} \min\left\{1, \frac{1}{\kappa_\sigma}\right\} \epsilon^2 = \frac{\epsilon^2}{2\kappa_\sigma} \text{ for all } k \text{ with (2.30)}, \tag{2.34}$$

where the equality follows from κ_σ in (2.27) satisfying $\kappa_\sigma \geq 1$ due to $\gamma_2 \geq 1$ and $L_g \geq 1$. Let $\mathcal{S}$ denote the set of all successful or very successful iterations in the sense of (2.24). Now (2.23), (2.24), and (2.34) imply

$$\Phi_h(x_k) - \Phi_h(x_{k+1}) \geq \eta_1 \Psi^r(x_k, \sigma_k) \geq \frac{\eta_1}{2\kappa_\sigma} \epsilon^2 \tag{2.35}$$

for all $k \in \mathcal{S}$ satisfying (2.30); assume there are k_ϵ such iterations. Summing up (2.35) over all such k, and recalling that function values remain unchanged on unsuccessful iterations and that $\Phi_h(k) \geq \Phi_h^{\text{low}}$, we get

$$\Phi_h(x_0) - \Phi_h^{\text{low}} \geq \sum_k \left[\Phi_h(x_k) - \Phi_h(x_{k+1})\right] \geq \sum_{k=0, k\in\mathcal{S}} \left[\Phi_h(x_k) - \Phi_h(x_{k+1})\right] \geq k_\epsilon \frac{\eta_1 \epsilon^2}{2\kappa_\sigma},$$

and so $k_\epsilon \leq 2\kappa_\sigma \left[\Phi_h(x_0) - \Phi_h^{\text{low}}\right] / (\eta_1 \epsilon^2)$, which is the bound $J_1^{s,r}$. To prove the bound J_1^r, we need to bound the number of unsuccessful iterations up to j_1. First, (2.32) implies

$$\sigma_{k+1} \geq \gamma_3 \sigma_k, \quad k \in \{0, \ldots, j_1\} \cap \mathcal{S},$$

and (2.24) gives

$$\sigma_{i+1} \geq \gamma_1 \sigma_i, \quad i \in \{0, \ldots, j_1\} \setminus \mathcal{S}.$$

Thus we deduce inductively that

$$\sigma_{j_1} \geq \sigma_0 \gamma_3^{k_{j_1}^s} \gamma_1^{k_{j_1}^u},$$

where $k_{j_1}^u$ denotes the number of unsuccessful iterations up to j_1; from (2.27), this further becomes

$$\frac{\kappa_\sigma}{\sigma_0} \geq \gamma_3^{k_{j_1}^s} \gamma_1^{k_{j_1}^u},$$

and so, taking the logarithm on both sides and recalling that $\gamma_1 > 1$, we get

$$k_{j_1}^u \leq -k_{j_1}^s \frac{\log \gamma_3}{\log \gamma_1} + \frac{1}{\log \gamma_1} \log \frac{\kappa_\sigma}{\sigma_0}.$$

Hence, since $\epsilon \in (0, 1]$, we deduce

$$j_1 = k_{j_1}^s + k_{j_1}^u \leq k_{j_1}^s \left(1 - \frac{\log \gamma_3}{\log \gamma_1}\right) + \frac{\epsilon^{-2}}{\log \gamma_1} \log \frac{\kappa_\sigma}{\sigma_0},$$

which together with the bound $J_1^{s,r}$ on $k_{j_1}^s$ yields (2.33). ∎

When applying Algorithm 2.2 to (1.2) in place of Φ_h, Theorem 2.7 applies and the constants remain the same in the bounds, except L_h in expression (2.27) is replaced by ρ; thus for ρ sufficiently large, $\kappa_\sigma = \mathcal{O}(\rho)$ and $\Phi_h(x_0) - \Phi_h^{\text{low}} = \mathcal{O}(\rho)$, and so κ_{QR}^s and κ_{QR} are both $\mathcal{O}(\rho^2)$, hence the same in order as for the (first-order) trust-region approach in the previous section. Note also that to ensure $\Phi(\cdot, \rho)$ is bounded below, it again suffices to require that f be bounded below on $\mathbb{R}^n$; again, both of these are restrictive assumptions when related to problem (1.3), as we discuss in greater detail in section 4.

3. An exact penalty-function algorithm for problem (1.3).

We now return to the problem-evaluation complexity of solving (1.3). In what follows, we let $\Phi_h = \Phi(\cdot, \rho)$, where $\Phi(\cdot, \rho)$ is defined in (1.2) for a(ny) $\rho > 0$, and so the criticality measure (2.2) becomes in this case

$$\Psi_\rho(x) \overset{\text{def}}{=} l^\rho(x, 0) - \min_{\|s\| \leq 1} l^\rho(x, s), \tag{3.1}$$

where

$$l^\rho(x, s) = f(x) + g(x)^T s + \rho \|c(x) + J(x)s\| \text{ for any } x \text{ and } s,$$

is the approximation (2.1) when $\Phi_h = \Phi(\cdot, \rho)$.

3.1. Approximate solutions.

Let us relate the minimizers of (1.2) to the solutions of our original problem (1.3). It is well known that the penalty function (1.2) is exact in that for sufficiently large ρ, strict local minimizers of (1.3) satisfying the Mangasarian–Fromovitz constraint qualification (MFCQ) are minimizers of $\Phi(\cdot, \rho)$ [1,14]. Conversely, very similarly to the proof of [1, Theorem 4.1], we can show that if x_* is a critical point of $\Phi(\cdot, \rho)$ for some $\rho > 0$ and it is feasible for (1.3), then x_* is a KKT point of (1.3); if x_* is a critical point of $\Phi(\cdot, \rho)$ for all sufficiently large ρ that is infeasible for (1.3), then x_* is an (infeasible) critical point of the measure $\|c(x)\|$ of constraint violation. In the next theorem, we prove a similar result for when we have an approximate critical point of $\Phi(\cdot, \rho)$ in the sense that the optimality measure (3.1) is sufficiently small.

THEOREM 3.1. *Let AF.1 hold and let $\rho > 0$. Consider minimizing $\Phi(\cdot, \rho)$ by some algorithm and obtaining an approximate solution x such that*

$$\Psi_\rho(x) \leq \epsilon \tag{3.2}$$

for a given tolerance $\epsilon > 0$. Then there exists $y_*(\rho)$ such that

$$\|g(x) + J(x)^T y_*(\rho)\| \leq \epsilon. \tag{3.3}$$

Additionally, if $\|c(x)\| \leq \kappa_c \epsilon$ for some $\kappa_c > 0$, then x is an approximate KKT point of problem (1.3)*, within ϵ.*

Proof. Note that it is straightforward that if (3.3) and $\|c(x)\| \leq \kappa_c \epsilon$ hold, then the KKT conditions (1.4) for (1.3) hold with a residual norm error of order ϵ, so that x is an approximate KKT point of (1.3). Thus it remains to show that (3.2) implies (3.3). Let

$$s_* = \arg \min_{\|s\| \leq 1} l^\rho(x, s) = \arg \min_{\|s\| \leq 1} f(x) + g(x)^T s + \rho \|J(x)s + c(x)\|. \tag{3.4}$$

Let us first assume that we are in the case $\|s_*\| < 1$. Then (3.4) is essentially unconstrained and convex, and first-order conditions [10, Theorem 2.2.1] provide that $(0 \in \partial l^\rho(x, s_*))$, and so there exists $y_* \in \partial(\|J(x)s_* + c(x)\|)$ such that $g(x) + \rho J(x)^T y_* = 0$, which implies that (3.3) trivially holds with $y_*(\rho) \overset{\text{def}}{=} \rho y_*$. It remains to consider $\|s_*\| = 1$. Then first-order conditions for (3.4) imply that there exist $y_* \in \partial(\|J(x)s_* + c(x)\|)$, $\lambda_* \geq 0$, and $z_* \in \partial(\|s_*\|)$ such that

$$g(x) + \rho J(x)^T y_* + \lambda_* z_* = 0. \tag{3.5}$$

It follows from the definition (3.1) of $\Psi_\rho(x)$ that

$$\Psi_\rho(x) = l^\rho(x,0) - l^\rho(x,s_*) = -g(x)^T s_* + \rho\left\{\|c(x)\| - \|J(x)s_* + c(x)\|\right\},$$

and replacing $g(x)$ from (3.5) into the above, we deduce

$$\begin{aligned}\Psi_\rho(x) &= \rho\left\{\|c(x)\| - \|J(x)s_* + c(x)\| + s_*^T J(x)^T y_*\right\} + \lambda_* s_*^T z_* \\ &= \rho\left\{\|c(x)\| - \|J(x)s_* + c(x)\| + s_*^T J(x)^T y_*\right\} + \lambda_*,\end{aligned} \tag{3.6}$$

where we also used that $s_*^T z_* = 1$. Let $p(s) = \|J(x)s + c(x)\|$, which is convex; then $p(0) - p(s_*) \geq (-s_*)^T J(x)^T y$ for any $y \in \partial(\|J(x)s_* + c(x)\|)$. Letting $y = y_*$, we deduce

$$\|c(x)\| - \|J(x)s_* + c(x)\| + (s_*)^T J(x)^T y_* \geq 0,$$

and so, from (3.2) and (3.6), we have that

$$\epsilon \geq \Psi_\rho(x) \geq \lambda_*. \tag{3.7}$$

From (3.5) and $\|z_*\| = 1$, we deduce

$$\lambda_* = \lambda_*\|z_*\| = \|g(x) + \rho J(x)^T y_*\|. \tag{3.8}$$

Finally, (3.7) and (3.8) yield (3.3) with $y_*(\rho) \overset{\text{def}}{=} \rho y_*$. □

Let us introduce the following function as a measure of constraint violation:

$$v(x) = \|c(x)\|. \tag{3.9}$$

Clearly, this is a special case of Φ_h and $\Phi(\cdot,\rho)$, obtained by letting $f = 0$ in (1.1) and in (1.2), as well as $h = \|\cdot\|$ in the former and $\rho = 1$ in the latter. Hence the criticality measure and results in the previous section apply to v. We let

$$l^v(x,s) = \|c(x) + J(x)s\| \text{ for any } x \text{ and } s$$

be the value of the approximation (2.1) for $\Phi_h = v$, and

$$\theta(x) \overset{\text{def}}{=} l^v(x,0) - \min_{\|s\|\leq 1} l^v(x,s) \tag{3.10}$$

the criticality measure (2.2) for $\Phi_h = v$ at some point x.

By letting $f = 0$, $g = 0$, and $\rho = 1$ in Theorem 3.1, we deduce the implication

$$\theta(x) \leq \epsilon \quad \Longrightarrow \quad \|J(x)^T y_*\| \leq \epsilon \tag{3.11}$$

for some $y_* \in \mathbb{R}^m$ and $\epsilon > 0$, where $\theta(x)$ is defined in (3.10). Note that $\|y_*\| \leq 1$; further, if $\|y_*\| < 1$ in (3.11), including the case when $y_* = 0$, then we are approximately feasible, namely, $\|c(x)\| \leq \epsilon$. Indeed, it follows from (3.4) and (3.5) that $y_* \in \partial(\|J(x)s_* + c(x)\|)$, where now $s_* = \arg\min_{\|s\|\leq 1} l^v(x,s)$. This and the definition of the subdifferential of the norm [10, section VI.3] imply that $\|y_*\| \leq 1$, and that $J(x)s_* + c(x) = 0$ whenever $\|y_*\| < 1$. In the latter case, (3.10) becomes $\theta(x) = \|c(x)\|$, and so (3.11) implies that $\|c(x)\| \leq \epsilon$.

It follows from (3.11) that when $\theta(x)$ is small, we are within ϵ of a KKT point of the feasibility problem

$$\min_x 0 \text{ subject to } c(x) = 0.$$

Note, however, that this may not imply that x is close to being feasible for the constraints c as required at the end of Theorem 3.1. Indeed, as we shall see in what follows, the exact penalty algorithm below may terminate at an infeasible critical point of v.

3.2. The outer penalty algorithm with a steering procedure. The algorithm for solving (1.3) that we analyze next is a standard exact penalty method [14], apart from the inclusion of a steering procedure [1] that we use when updating the penalty parameter ρ; see Step 1 of Algorithm 3.1. This heuristic ensures that the (main) iterates x_k generated by this Algorithm satisfy

$$\Psi_{\rho_k}(x_k) \geq \xi \rho_k \theta(x_k) \quad \text{for all } k \geq 0, \tag{3.12}$$

and that if ρ is increased on the kth iteration, it is because

$$\Psi_{\rho_{k-1}}(x_k) < \xi \rho_{k-1} \theta(x_k). \tag{3.13}$$

Steering helps ensure that we cannot be close to a critical point of $\Phi(\cdot, \rho)$ without being near a critical point of the feasibility measure v. Note that steering does not involve any additional problem evaluations of (1.3), only additional computations of the optimality measure (3.1) whenever ρ is increased.

Algorithm 3.1: Exact penalty-function algorithm for solving (1.3).

Step 0: Initialization. An initial point x_0, a steering parameter $\xi \in (0,1)$, an initial penalty parameter $\rho_{-1} \geq 1/\xi$, and a minimal increase factor $\tau > 0$, as well as a tolerance $\epsilon \in (0,1]$, are given. Set $k = 0$.

Step 1: Update the penalty parameter. If $\rho = \rho_{k-1}$ satisfies

$$\Psi_\rho(x_k) \geq \xi \rho \theta(x_k), \tag{3.14}$$

then set $\rho_k = \rho_{k-1}$. Else, choose any ρ_k such that $\rho_k \geq \rho_{k-1} + \tau$ and that (3.14) holds with $\rho = \rho_k$.

Step 2: Inner minimization. (Approximately) solve the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \Phi(x, \rho_k) \tag{3.15}$$

by applying some algorithm (e.g., Algorithm 2.1/2.2), starting from some x_k^{s} and stopping at an (approximate) solution x_{k+1} for which

$$\Psi_{\rho_k}(x_{k+1}) \leq \epsilon, \tag{3.16}$$

where $\Psi_{\rho_k}(x_{k+1})$ is defined in (3.1) with $\rho = \rho_k$ and $x = x_{k+1}$.

Step 3: Termination. If the value of v's criticality measure θ at x_{k+1} satisfies

$$\theta(x_{k+1}) \leq \epsilon, \tag{3.17}$$

where $\theta(x_{k+1})$ is (3.10) with $x = x_{k+1}$, then terminate. Else, increment k by 1 and go to Step 1.

Let us argue that Step 1 of Algorithm 3.1 is well-defined for any $\xi \in (0,1)$; namely, condition (3.14) can be ensured for sufficiently large ρ; see also [1]. From (3.1), we have

$$\begin{aligned} \Psi_\rho(x_k) &= \rho\|c(x_k)\| - \min_{\|s\|\leq 1} \left\{ g(x_k)^T s + \rho\|c(x_k) + J(x_k)s\| \right\} \\ &\geq - \min_{\|s\|\leq 1} \left\{ \|g(x_k)\| \cdot \|s\| \right\} + \rho \left\{ \|c(x_k)\| - \min_{\|s\|\leq 1} \|c(x_k) + J(x_k)s\| \right\} \\ &\geq -\|g(x_k)\| + \rho\theta(x_k), \end{aligned} \tag{3.18}$$

where we also used the Cauchy–Schwarz inequality and (3.10). Thus (3.14) holds, provided

$$\rho \geq \frac{\|g(x_k)\|}{(1-\xi)\theta(x_k)}. \tag{3.19}$$

In practice, the value (3.19) is considerably larger than necessary. In particular, notice that as x_k approaches feasibility, $\theta(x_k)$ approaches zero and so the right-hand side of (3.19) blows up; thus, (3.10) should not be used for choosing ρ in Step 1 of Algorithm 3.1 [1].

Note the termination condition in Step 3 of Algorithm 3.1. The condition (3.16) ensures (3.3), due to Theorem 3.1, but to be close to a KKT point of (1.3), we still need to ensure that $\|c(x_k)\| \leq \kappa_c\epsilon$ for some $\kappa_c > 0$. We will show, however, in the theorem below, that only the weaker termination condition (3.17) can be ensured by Algorithm 3.1; see also our remarks following (3.11).

Let us now investigate the problem- (namely, function- and constraint-)evaluation worst-case complexity of Algorithm 3.1. We need to show that (3.17) will hold after ρ_k has been finitely or infinitely increased, so that Algorithm 3.1 terminates either with an approximate KKT point of (1.3) or an approximate (infeasible) critical point of v.

THEOREM 3.2. *Let AF.*1 *and AF.*2 *hold, and assume that f is bounded below over $\mathbb{R}^n$. Let either Algorithm* 2.1 *or* 2.2 *be applied on each major iteration k of Algorithm* 3.1 *for solving the subproblem* (3.15).

(i) *Assume that there exists $\overline{\rho} > 0$ such that $\rho_k \leq \overline{\rho}$ for all k. Then Algorithm* 3.1 *will terminate either with an approximate KKT point of* (1.3) *or an infeasible critical point of the feasibility measure* (3.9) *in at most*

$$\left\lceil \frac{\kappa_{ep}\overline{\rho}^3}{\epsilon^2} \right\rceil \tag{3.20}$$

problem evaluations, where κ_{ep} is a positive problem-dependent constant.

(ii) *Alternatively, assume that ρ_k grows unboundedly as k increases. Assume also that the sequence of (major) iterates $\{x_k\}$ is bounded. Then Algorithm* 3.1 *will terminate either with an approximate KKT point of* (1.3) *or an infeasible critical point of the feasibility measure* (3.9) *in at most*

$$\left\lceil \frac{\kappa_{ep,inf}}{\epsilon^5} \right\rceil \tag{3.21}$$

problem evaluations, where $\kappa_{ep,inf}$ is positive problem-dependent constant.

Proof. First, note that on any iteration $k \geq 1$, either Algorithm 3.1 terminates or ρ must be increased to satisfy (3.14). Indeed, if (3.14) holds with $\rho = \rho_{k-1}$, then (3.16) on iteration $k-1$ implies that

$$\theta(x_k) \leq \frac{\Psi_{\rho_{k-1}}(x_k)}{\xi\rho_{k-1}} \leq \frac{\epsilon}{\xi\rho_{k-1}} \leq \frac{\epsilon}{\xi\rho_{-1}} \leq \epsilon,$$

where we also used that the penalty parameters are monotonically increasing and the assumption $\rho_{-1}\xi \geq 1$. Thus except for maybe the first iteration $k = 0$, the penalty parameter ρ_k will be increased in each iteration until termination.

(i) Since ρ_k will be increased at most $\lceil(\overline{\rho}-\rho_0)/\tau\rceil$ times before reaching its upper bound $\overline{\rho}$, there will be at most $\lceil\overline{\rho}/\tau\rceil$ subproblems (3.15) solved. Let us assume Algorithm 2.1 is employed to solve (3.15). Then Theorem 2.4 applies (since $\Phi(\cdot, \rho_k)$ satisfies AH.1 with $L_h = \rho_k$), yielding that the subproblem solution set (3.15) takes at most $\lceil\kappa_{\rm TR}\epsilon^{-2}\rceil$ problem evaluations, where $\kappa_{\rm TR}$ is defined in (2.17). Note that since $f(x) \geq f_{\text{low}}$ for all x, $\rho_k \leq \overline{\rho}$ for all k, and since there are finitely many subproblems (3.15) needing to be solved, we have that for all major iterations $k \geq 0$,

$$\begin{aligned}\Phi(x_k^{\text{s}}, \rho_k) - \Phi^{\text{low}}(x, \rho_k) &\leq \Phi(x_k^{\text{s}}, \rho_k) - f_{\text{low}} \\ &\leq \max_{1\leq k\leq \lceil \overline{\rho}/\tau \rceil} \{f(x_k^{\text{s}}) + \overline{\rho}\|c(x_k^{\text{s}})\|\} - f_{\text{low}} \\ &\leq \kappa_1\overline{\rho} + \kappa_2\end{aligned} \tag{3.22}$$

for some $\kappa_{1,2} > 0$. This, (2.10) with $L_h = \rho_k$, and (2.17) imply that $\kappa_{\text{TR}} = \mathcal{O}(\rho_k\overline{\rho}) \leq \mathcal{O}(\overline{\rho}^2)$. Thus Algorithm 3.1 will terminate after at most as many problem evaluations as given in (3.20), where $\kappa_{ep} \overset{\text{def}}{=} \kappa_{\text{TR}}/(\tau\overline{\rho}^2)$. The termination criteria (3.17) implies that we are within ϵ of a critical point of the feasibility measure v; see (3.10). If this approximate critical point x_k of v is approximately feasible with respect to the constraints so that $\|c(x_k)\| \leq \epsilon$, then (3.16) and Theorem 3.1 imply that we are near a KKT point of (1.3) in the sense of (3.3).

A similar argument can be given when applying Algorithm 2.2 to the subproblem (3.15), yielding similar problem-evaluation counts by employing Theorems 2.7 and, again, 3.1.

(ii) In this case, we must have that (3.13) holds for all k apart from possibly $k = 0$. Then using (3.18) with $\rho = \rho_{k-1}$, we deduce

$$\Psi_{\rho_{k-1}}(x_k) \geq -\|g(x_k)\| + \rho_{k-1}\theta(x_k),$$

which together with (3.13) gives

$$\xi\rho_{k-1}\theta(x_k) > -\|g(x_k)\| + \rho_{k-1}\theta(x_k),$$

or, equivalently,

$$\theta(x_k) \leq \frac{\|g(x_k)\|}{(1-\xi)\rho_{k-1}}. \tag{3.23}$$

As we have assumed that the iterates are bounded, $\{\|g(x_k)\|\}$ is also bounded above by say, M_g, and so (3.23) becomes

$$\theta(x_k) \leq \frac{M_g}{(1-\xi)\rho_{k-1}} \tag{3.24}$$

for all $k \geq 1$. We conclude that $\theta(x_k) \leq \epsilon$ once

$$\rho_{k-1} \geq \frac{M_g}{(1-\xi)\epsilon}, \tag{3.25}$$

and then Algorithm 3.1 would terminate. The remainder of the proof now follows similarly to case (i) by letting $\overline{\rho}$ be the right-hand side of (3.25). □

Note that the condition on the initial choice of penalty and steering parameters that $\rho_{-1}\xi \geq 1$ that we imposed in Algorithm 3.1 is merely for convenience, and if ignored, only a change by a constant multiple occurs in the accuracy required in either (3.16) or (3.17).

The right-hand side of the bound (3.22)—which is used to prove both cases (i) and (ii)—may depend on the major iteration k via $f(x_k^{\rm s})$ and $c(x_k^{\rm s})$; we have shown, however, that there can only be finitely many major iterations, which is sufficient for our purposes in the above theorem. One may wish to remove this dependence on k so that the bounds depend only on initial problem data. A trivial (but likely practically inefficient) option is to set $x_k^{\rm s} = x_0$ for all $k \geq 0$ in Step 2 of Algorithm 3.1. Similarly, we may set $x_0^{\rm s} = x_0$, and only shift $x_k^{\rm s}$ (from $x_{k-1}^{\rm s}$) to be x_k if (ever) both $f(x_k) \leq f(x_{k-1}^{\rm s})$ and $\|c(x_k)\| \leq \|c(x_{k-1}^{\rm s})\|$, as then $\Phi(x_k^{\rm s}, \rho_k) = f(x_k^{\rm s}) + \rho_k\|c(x_k^{\rm s})\| \leq f(x_0) + \rho_k\|c(x_0)\|$. Alternatively, if we assume the common choice $x_k^{\rm s} = x_k$, $k \geq 0$, an inspection of the proof of Theorem 2.4 provides that the difference $\Phi(x_k^{\rm s}, \rho_k) - \Phi^{\rm low}(x, \rho_k)$—occurring in the bound (2.14) when applied to $\Phi_h = \Phi(\cdot, \rho_k)$—can be replaced by $\Phi(x_k, \rho_k) - \Phi(x_{k+1}, \rho_k)$. Thus, to obtain (3.20) and (3.21), we can sum up the latter differences with respect to k instead of summing up (3.22). Then the terms containing f cancel out and, using that $\rho_k \leq \overline{\rho}$, we are left with upper-bounding the terms $\|c_k\|$, $k \geq 0$. Due to AF.1, the latter holds if we assume, as in case (ii), that $\{x_k\}$ is uniformly bounded, independently of k.

Since the penalty function Φ is exact, some solutions of (1.3) and some feasible ones of $\Phi(\cdot, \rho)$ correspond for all $\rho \geq \rho_*$, where ρ_* is independent of ϵ, provided constraint qualifications hold [14, section 17.2]. Thus the assumptions of case (i) are reasonable, and so the bound (3.20) is of most interest and relevance, while the case (ii) may not happen too often.

Finally, note that steering can also be performed inside the main iteration k, namely, inside the subproblem solution algorithm, as in [1]. Then, depending on whether we employ Algorithm 2.1 or 2.2 for the subproblem minimization, the model decreases (2.5) or (2.22) for $\Phi(\cdot, \rho)$ and for v can be used in (3.14) instead of the criticality measures $\Psi_\rho(x_k)$ and $\theta(x_k)$, respectively. This approach yields some computational savings, as the model decreases are already readily computed; it may even decrease the worst-case problem-evaluation count, since not each subproblem (3.15) needs to be solved to ϵ accuracy. However, the loss of monotonicity in the function values $\Phi(x_k, \rho_k)$ once ρ_k is allowed to increase inside Algorithm 2.1 or 2.2 seems to prevent this approach from being amenable to our complexity analysis. Fortunately, even if a way were found to overcome the latter, the (more important) bound (3.20) is unlikely to change in the order of ϵ as it represents the worst-case function-evaluation cost of solving unconstrained nonconvex optimization problems by means of a steepest-descent-like method, which we have shown in [4] to be tight.

4. Discussion and conclusions.

The problem-evaluation complexity bounds in Theorem 3.2 rely on the assumption that f is bounded below over the whole of $\mathbb{R}^n$, which ensures that every unconstrained penalty minimization subproblem is well-defined. While such an assumption is reasonable in the context of the minimization of the (unconstrained) composite function Φ_h or $\Phi(\cdot, \rho)$, and hence in our results in section 2, it is a strong assumption when related to problem (1.3), as simple (but important) nonconvex problems such as quadratic programming fail to satisfy it. Nevertheless, convergence results for penalty methods commonly make this assumption. A way to overcome it in the quadratic programming case, for example, is to choose h in Φ_h as the "opposite"-Huber function

$$h(x) = \rho \cdot \begin{cases} \|x\| & \text{for } \|x\| \leq 1, \\ \frac{1}{2} + \frac{1}{2}\|x\|^2 & \text{for } \|x\| > 1, \end{cases}$$

which also gives an exact penalty function so that Theorem 3.2 continues to hold in this case. Crucially, the Huber function grows sufficiently at infinity to counter unboundedness of the objective for sufficiently large ρ. For the more general problem (1.3), a function h is needed that balances objective and constraint growth at infinity. Alternatively, instead of exact penalty methods, one may consider funnel techniques [7], which only require f to be bounded below in a neighborhood of the feasible set of constraints; but the latter are an SQP-based approach whose complexity appears to be more difficult to analyze.

We have analyzed the function-evaluation complexity of minimizing a composite nonlinear nonconvex function with (possibly) a nonsmooth term, when solved using a first-order trust-region and a first-order quadratic regularization method. We found that the worst-case complexity of both methods driving some first-order optimality below ϵ is of order ϵ^{-2}, the same as for smooth unconstrained nonconvex optimization. Practical examples include nonlinear fitting in polyhedral (l_1, l_∞) norms both with and without regularization. We then applied these bounds to the penalty function subproblem solution in the context of an exact penalty algorithm for the equality-constrained problem (1.3). We obtained that in the important case when the penalty parameter is bounded, the problem-evaluation complexity of reaching within ϵ of a KKT point of (1.3) is of order ϵ^{-2}, the same as for unconstrained optimization. To the best of our knowledge, this is the first worst-case problem-evaluation complexity bound for smooth constrained optimization when both the objective and constraints may be nonconvex.

Our exact penalty approach and complexity analysis can be easily extended to problems that also have finitely many inequality constraints by commonly incorporating the norm of the inequality constraint violation as an additional term of the penalty function [1, 6, 14].

REFERENCES

[1] R. H. Byrd, N. I. M. Gould, J. Nocedal, and R. A. Waltz, *On the convergence of successive linear-quadratic programming algorithms*, SIAM J. Optim., 16 (2005), pp. 471–489.

[2] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *Adaptive cubic regularisation methods for unconstrained optimization. Part* II: *Worst-case function- and derivative-evaluation complexity*, Math. Program., 130 (2011), pp. 295–319.

[3] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity*, IMA J. Numer. Anal., to appear.

[4] C. Cartis, N. I. M. Gould, and Ph. L. Toint, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization*, SIAM J. Optim., 20 (2010), pp. 2833–2852.

[5] A. R. Conn, N. I. M. Gould, and Ph. L. Toint, *Trust-Region Methods*, SIAM, Philadelphia, 2000.

[6] R. Fletcher, *Practical Methods of Optimization*, 2nd ed., John Wiley & Sons, New York, 1987.

[7] N. I. M. Gould and Ph. L. Toint, *Nonlinear programming without a penalty function or a filter*, Math. Program., 122 (2010), pp. 155–196.

[8] S. Gratton, M. Mouffe, Ph. L. Toint, and M. Weber-Mendonça, *A recursive trust-region method in infinity norm for bound-constrained nonlinear optimization*, IMA J. Numer. Anal., 28 (2008), pp. 827–861.

[9] S. Gratton, A. Sartenaer, and Ph. L. Toint, *Recursive trust-region methods for multiscale nonlinear optimization*, SIAM J. Optim., 19 (2008), pp. 414–444.

[10] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms* I, Springer-Verlag, Berlin, Heidelberg, 1993.

[11] Y. Nesterov, *Introductory Lectures on Convex Optimization*, Kluwer Academic, Dordrecht, The Netherlands, 2004.

[12] Y. Nesterov, *Gradient Methods for Minimizing Composite Objective Function*, CORE Discussion Paper 2007/76, Université Catholique de Louvain, Belgium, 2007.

[13] Y. Nesterov, *Modified Gauss-Newton scheme with worst case guarantees for global performance*, Optim. Methods Software, 22 (2007), pp. 469–483.

[14] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer-Verlag, New York, 2006.

[15] Y. Yuan, *Conditions for convergence of trust region algorithms for non-smooth optimization*, Math. Program., 31 (1985), pp. 220–228.