Hansknecht, Christoph; Kirches, Christian; Manns, Paul

Convergence of successive linear programming algorithms for noisy functions

Computational Optimization and Applications

Suggested Citation: Hansknecht, Christoph; Kirches, Christian; Manns, Paul (2024) : Convergence of successive linear programming algorithms for noisy functions, Computational Optimization and Applications, ISSN 1573-2894, Springer US, New York, Vol. 88, Iss. 2, pp. 567-601, https://doi.org/10.1007/s10589-024-00564-w

Computational Optimization and Applications (2024) 88:567–601
https://doi.org/10.1007/s10589-024-00564-w

Convergence of successive linear programming algorithms for noisy functions

Christoph Hansknecht[1] · **Christian Kirches**[1] · **Paul Manns**[2]

Received: 16 February 2023 / Accepted: 30 January 2024 / Published online: 26 February 2024

Abstract

Gradient-based methods have been highly successful for solving a variety of both unconstrained and constrained nonlinear optimization problems. In real-world applications, such as optimal control or machine learning, the necessary function and derivative information may be corrupted by noise, however. Sun and Nocedal have recently proposed a remedy for smooth unconstrained problems by means of a stabilization of the acceptance criterion for computed iterates, which leads to convergence of the iterates of a trust-region method to a region of criticality (Sun and Nocedal in Math Program 66:1–28, 2023. https://doi.org/10.1007/s10107-023-01941-9). We extend their analysis to the successive linear programming algorithm (Byrd et al. in Math Program 100(1):27–48, 2003. https://doi.org/10.1007/s10107-003-0485-4, SIAM J Optim 16(2):471–489, 2005. https://doi.org/10.1137/S1052623403426532) for unconstrained optimization problems with objectives that can be characterized as the composition of a polyhedral function with a smooth function, where the latter and its gradient may be corrupted by noise. This gives the flexibility to cover, for example, (sub)problems arising in image reconstruction or constrained optimization algorithms. We provide computational examples that illustrate the findings and point to possible strategies for practical determination of the stabilization parameter that balances the size of the critical region with a relaxation of the acceptance criterion (or descent property) of the algorithm.

Keywords Nonsmooth optimization · Global convergence · Noisy optimization

✉ Christoph Hansknecht
c.hansknecht@tu-braunschweig.de

Christian Kirches
c.kirches@tu-braunschweig.de

Paul Manns
paul.manns@tu-dortmund.de

[1] Institute for Mathematical Optimization, TU Braunschweig, Braunschweig, Germany

[2] Chair of Numerical Analysis and Optimization, TU Dortmund University, Dortmund, Germany

1 Introduction

Handling non-smoothness is an ubiquitous research question in nonlinear optimization because it arises naturally in different areas, for example, penalty functions for constrained optimization [4], statistical data analysis and signal processing [5, 6], and neural network architectures [7]. In this work we study the convergence properties of successive linear programming algorithms to solve the optimization problem

$$\min_{x\in\mathbb{R}^n} \phi(x) := \omega(F(x)), \tag{P}$$

where $\omega : \mathbb{R}^p \to \mathbb{R}$ is convex and Lipschitz continuous with polyhedral epigraph, and $F : \mathbb{R}^n \to \mathbb{R}^p$ is twice continuously differentiable. Moreover, we assume that F and its Jacobian can only be accessed inexactly so that their evaluations are corrupted by noise. This and similar problems have been studied in the literature, see, for example, [8–12] and the references therein.

Many optimization problems can be formulated in terms of problem (P) such as the Lagrangian form

$$\min_{x\in\mathbb{R}^n} \|y - Ax\|_2^2 + \beta\|x\|_1$$

of the famous LASSO problem [5, 13] with $A \in \mathbb{R}^{m\times n}$, $\beta > 0$, and $y \in \mathbb{R}^m$ that is particularly popular among data scientists for sparse parameter identification in over-parameterized models. More broadly speaking, a wide class of nonlinear optimization problems fall under (P) as well. As an example, the unconstrained minimization of a smooth objective $f : \mathbb{R}^n \to \mathbb{R}$ can be modeled by setting $p = 1$ and $\omega(x) = x$ for $x \in \mathbb{R}$. We can also examine nonlinear programs constrained by smooth functions $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^k$ yielding problems of the form

$$\min_{x\in\mathbb{R}^n} f(x) \text{ s.t. } g(x) \le 0,\ h(x) = 0. \tag{NLP}$$

These problems may be solved by minimizing a non-smooth exact penalty function of the form

$$\phi(x, \nu) := f(x) + \nu \left\| \left(g(x)_+^{\mathrm{T}}, h(x)^{\mathrm{T}} \right)^{\mathrm{T}} \right\|_1, \tag{1}$$

where $y_+ := \max(y, 0)$ and $\nu > 0$. In fact, strict local solutions of (NLP) are local minimizers of $\phi(x, \nu)$ for a sufficiently large value of ν if g and h are smooth and satisfy the Mangasarian–Fromovitz constraint qualification [14, Theorem 4.4], [4, Theorem 17.3] at the respective points. The penalty function ϕ can be expressed as $\omega(F(x))$, where $F(x) := (f(x), g(x)^{\mathrm{T}}, h(x)^{\mathrm{T}})^{\mathrm{T}}$ is smooth and $\omega(x, y, z) := x + \nu\|({y_+}^{\mathrm{T}}, z^{\mathrm{T}})^{\mathrm{T}}\|_1$ is convex and polyhedral. Besides problems of type (NLP), a variety of other problems, such as linear or nonlinear fitting problems can be formulated in terms of (P) as well.

Noisy functions

The combination of unconstrained optimization with noisy observations has recently been examined [1, 15–21]. The authors consider the minimization of a smooth function $\phi : \mathbb{R}^n \to \mathbb{R}$ while only having access to

$$f(x) = \phi(x) + \varepsilon(x) \text{ and } g(x) = \nabla\phi(x) + e(x), \tag{2}$$

where the only assumptions are that both $|\varepsilon|$ and $\|e\|$ are uniformly bounded. Consequently, it is not generally possible to generate a sequence $\{x_k\}$ of iterates converging to a local optimum or stationary point of ϕ. Intuitively, while the gradient noise e is small compared to $\nabla\phi$, the direction g is a suitable search direction with respect to ϕ. This allows for the use of an Armijo-like globalization strategy [16, 20] or, in case of [1, 17], a trust-region method, where the noise is handled by stabilizing the reduction ratio, which is of course closely related to the Armijo condition. As soon as a region is reached where the noise produced by ε and e becomes too large relative to ϕ and $\nabla\phi$ respectively, no further progress can be expected and the algorithm may stall. However, this critical region is visited infinitely often and once reaching it, the algorithm does not produce objective values much larger than the objective values attained in the critical region. The authors also study the problem of adapting quasi-Newton methods to the noisy setting.

Contribution

We build on the ideas in [1, 15, 21] and consider the non-smooth problem (P) in a setting, where function and derivative evaluations are only available as *noisy* observations. As the authors in [1, 15, 21], we assume the following noise model: Rather than being able to evaluate F and its derivative F' directly, we only have access to

$$\tilde{F}(x) := F(x) + \delta_F(x) \text{ and } G(x) := F'(x) + \delta_{F'}(x).$$

These proxies consist of the original functions F and F' as well as error functions $\delta_F : \mathbb{R}^n \to \mathbb{R}^p$ and $\delta_{F'} : \mathbb{R}^n \to \mathbb{R}^{p\times n}$. In terms of the problem (NLP), this is tantamount to noise in the objective f, the constraints g, h, and their respective derivatives. Contrary to this, we assume that the function ω does not suffer from any noise. What is more, we presume that the structure of ω is well understood in the sense that, for example, its Lipschitz constant is known, which is certainly the case for the penalty function in (1).

In order to solve optimization problems of the form (P), we propose a trust-region algorithm leaning on the successive linear programming template proposed in [2] and a convergence analysis that builds on the ideas in [3, 15, 21]. Specifically, we use a stabilization of the iterate acceptance test in order to assert that a neighborhood of a stationary point is visited infinitely often by the iterates produced by the algorithm. The polyhedral structure of ω is handled by first solving a linear program in order to determine a direction for a subsequent Cauchy point determination. This can also be interpreted as an active set determination for the corresponding kinks of the polyhedral epigraph of ω.

We also provide computational examples that illustrate the theoretical results and the practical behavior of the algorithm. Moreover, the results point to open questions and possible approaches regarding the choice of the correct stabilization parameter in the acceptance test.

Structure of the Remainder

We introduce the successive linear programming algorithm and the modified acceptance test in Sect. 2. The asymptotics of the algorithm are analyzed in Sect. 3. We

provide computational examples and the corresponding results in Sect. 4. We draw a conclusion in Sect. 5.

2 A noise-tolerant successive linear programming algorithm

In the noisy setting, we cannot expect to find the true optimum or stationary points of ϕ, since we do not have access to F and F'. Specifically, in a small region around the true optimum x^*, $\tilde{F}$ and G may oscillate, thereby making their evaluations unreliable. This impairs globalization strategies in nonlinear programming because their acceptance tests require reliable evaluations of F and a model function involving F'.

In the non-noisy regime, a trust-region method produces a sequence $\{x_k\}$ of iterates by assembling and subsequently optimizing model functions $q_k : \mathbb{R}^n \to \mathbb{R}$, yielding a step d_k. The quality of d_k is determined according to the reduction ratio

$$\rho_k := \frac{\phi(x_k) - \phi(x_k + d_k)}{\phi(x_k) - q_k(d_k)},$$

which is used to determine whether or not the step will be accepted. However, in the noisy setting, we only have access to $\tilde{F}$ leading to a noisy composite function $\tilde{\phi}(x) := \omega(\tilde{F}(x))$. While we can build a model $\tilde{q}_k : \mathbb{R}^n \to \mathbb{R}^p$ which coincides with $\tilde{F}$ at x_k, we cannot control the numerator $\tilde{\phi}(x_k) - \tilde{\phi}(x_k + d_k)$. Indeed, if we reduce the trust region, sending d_k to zero, the denominator of ρ_k will tend to zero while the numerator will oscillate, making the ratio unreliable. To alleviate this problem, we turn towards a recent adaptation [1] of trust-region methods in order to solve the noisy counterpart of (P). The authors of [1] add a correction term, that is a positive constant $\vartheta > 0$, to both the numerator and denominator of the reduction ratio ρ_k to mitigate the effect of noisy evaluations, yielding a modified ratio

$$\tilde{\rho}_k := \frac{\tilde{\phi}(x_k) - \tilde{\phi}(x_k + d_k) + \vartheta}{\tilde{\phi}(x_k) - \tilde{q}_k(d_k) + \vartheta}.$$

The parameter ϑ can then be chosen appropriately in order to stabilize the ratio. As we will see, this means that for ϑ large enough, the iterates of the successive linear programming algorithm converge to a critical region around a stationary point. The downside is that this region grows with ϑ and the algorithm also accepts steps that do not improve the objective.

Apart from this adjustment, we follow the algorithmic approach in [3]. Specifically, we use the following *partially linearized* and *quadratic* models at $x_k \in \mathbb{R}^n$

$$\tilde{\ell}(x; d) := \omega\left(\tilde{F}(x) + G(x)d\right), \tag{3}$$

$$\tilde{\ell}_k(d) := \tilde{\ell}(x_k; d), \text{ and} \tag{4}$$

$$\tilde{q}_k(d) := \tilde{\ell}_k(d) + \tfrac{1}{2}\langle d, B_k d\rangle, \tag{5}$$

where the $B_k \in \mathbb{R}^{n\times n}$ are symmetric (not necessarily positive definite) approximations of the curvature of $\omega \circ F$. Here and throughout we let $\langle\cdot,\cdot\rangle$ be the standard scalar product in $\mathbb{R}^n$, inducing the 2-norm, which we simply denote by $\|\cdot\|$. For matrices, we let $\|\cdot\|$ be the corresponding operator norm, given by its largest singular value.

Algorithm

Based on the models above, our noise-tolerant approach to solving (P) is laid out in Algorithm 1. In each iteration, an initial step d^{LP} is computed in Line 3 by solving the problem

$$\min_{\|d\|_{\mathrm{LP}}\le\Delta_k^{\mathrm{LP}}} \tilde{\ell}_k(d),$$

where $\|\cdot\|_{\mathrm{LP}}$ is a norm on $\mathbb{R}^n$ defining the trust region associated with $\tilde{\ell}_k$. While we make no further assumptions regarding this norm, it is in practice advantageous to cast this subproblem as a linear program to be solved using state-of-the-art LP solvers [22, 23]. To this end, two conditions should be met. First, the epigraph of ω should be polyhedral. Second, the feasible region should be polyhedral as well, or, equivalently, $\|\cdot\|_{\mathrm{LP}}$ should be a polyhedral norm, such as $\|\cdot\|_1$ or $\|\cdot\|_\infty$. In any case, due to the equivalence of norms in $\mathbb{R}^n$ there exists a constant $\gamma > 0$ such that for each $d \in \mathbb{R}^n$ it holds that

$$\|d\| \le \gamma \|d\|_{\mathrm{LP}}. \tag{6}$$

The algorithm proceeds to compute a Cauchy step d_k^C in Lines 4–7. To this end, it employs a line search initialized with a step size sufficiently small to ensure that the Cauchy step falls into the trust region bounded by Δ_k^{LP}. During the line search the step size is shortened by a factor of $0 < \tau < 1$ until the quadratic reduction achieved by the Cauchy point is within a factor of $0 < \eta < 1$ of its linear reduction.

The actual step d_k to be taken in Line 8 can be different from the Cauchy step d_k^C, provided that it improves upon the quadratic reduction of d_k^C. This gives some algorithmic flexibility, allowing for the computation of Newton-type steps in order to achieve local quadratic convergence. Based on the stabilized reduction ratio $\tilde{\rho}_k$ computed in Line 9, the step is either accepted (Lines 11–12) or rejected (Lines 14–15) according to an acceptance threshold of $\rho_u > 0$. Additionally, the trust-region radii Δ_{k+1} and $\Delta_{k+1}^{\mathrm{LP}}$ are adjusted based on $\tilde{\rho}_k$:

1. The value of Δ_{k+1} is increased or decreased based on whether $\tilde{\rho}_k$ achieves a value of at least ρ_s. The decrease is such that the new trust-region radius is at most $\kappa_u < 1$ times as large as the previous one, thereby ensuring a true reduction, while being at least $\kappa_l\|d_k\|$ (with $0 < \kappa_l \le \kappa_u$) in order to prevent an immediate collapse of the trust region.
2. If $\tilde{\rho}_k$ achieves at least ρ_u (note that $\rho_u \le \rho_s$), the LP trust-region radius $\Delta_{k+1}^{\mathrm{LP}}$ is increased beyond $\|d_k^C\|_{\mathrm{LP}}$, as long as it does not exceed the upper bound of $\Delta_{\max}^{\mathrm{LP}} \ge 1$. The new LP trust-region radius is also only increased beyond $\Delta_{k+1}^{\mathrm{LP}}$ if the full LP step d^{LP} was accepted (i.e., $\alpha_k = 1$), indicating that the partially linearized model $\tilde{\ell}_k$ is a good approximation of $\tilde{\phi}_k$ across the entire LP trust region. If $\tilde{\rho}_k$ falls short of ρ_u, $\Delta_{k+1}^{\mathrm{LP}}$ is decreased while being kept within a factor of $\theta > 0$ of $\|d_k\|_{\mathrm{LP}}$.

Remark 2.1 When applied to problem (NLP), Algorithm 1 uses the strategies introduced in [2], which form the basis of the active set method in the highly successful KNITRO code [24], which combines sequential linear programming with equality constrained quadratic programming approaches in order to achieve robust performance over a range of large-scale nonlinear programming problems.

Algorithm 1: A noise-tolerant algorithm to minimize $\phi(x) = \omega(F(x))$

Input : Functions ω, $\tilde{F}$, G,
Initial point $x_0 \in \mathbb{R}^n$,
Initial trust region radii $0 < \Delta_0^{\mathrm{LP}} \leq \Delta_{\max}^{\mathrm{LP}}$, $0 < \Delta_0$

Parameters: Acceptance thresholds $0 < \rho_u \leq \rho_s < 1$,
Step adjustments $0 < \kappa_l \leq \kappa_u < 1$, $\theta > 0$,
Cauchy line search parameters $0 < \eta < 1$, $0 < \tau < 1$,
Ratio stabilizer $\vartheta > 0$,
Maximum LP trust region radius $\Delta_{\max}^{\mathrm{LP}} \geq 1$

Output : Primal point $x^* \in \mathbb{R}^n$

$k \leftarrow 0$
until *Some termination criterion is satisfied*
  Compute LP step

$$d_k^{\mathrm{LP}} \leftarrow \arg\min_{\|d\|_{\mathrm{LP}} \leq \Delta_k^{\mathrm{LP}}} \tilde{\ell}_k(d)$$

  $\alpha_k \leftarrow \min\left(1, \Delta_k / \|d_k^{\mathrm{LP}}\|\right)$
  while $\tilde{\phi}(x_k) - \tilde{q}_k(\alpha_k d_k^{\mathrm{LP}}) < \eta\left[\tilde{\phi}(x_k) - \tilde{\ell}_k(\alpha_k d_k^{\mathrm{LP}})\right]$ **do**
    $\alpha_k \leftarrow \tau \alpha_k$
  $d_k^C \leftarrow \alpha_k d_k^{\mathrm{LP}}$
  $d_k \leftarrow$ Step d such that $\|d\| \leq \Delta_k$ and $\tilde{q}_k(d) \leq \tilde{q}_k(d_k^C)$
  Compute stabilized reduction ratio

$$\tilde{\rho}_k \leftarrow \frac{\tilde{\phi}(x_k) - \tilde{\phi}(x_k + d_k) + \vartheta}{\tilde{\phi}(x_k) - \tilde{q}_k(d_k) + \vartheta}$$

  if $\tilde{\rho}_k \geq \rho_u$ **then** ▷ *Accept step*
    Set $x_{k+1} \leftarrow x_k + d_k$
    Pick $\Delta_{k+1}^{\mathrm{LP}} \in \left[\|d_k^C\|_{\mathrm{LP}}, \Delta_{\max}^{\mathrm{LP}}\right]$ such that $\Delta_{k+1}^{\mathrm{LP}} \leq \Delta_k^{\mathrm{LP}}$ if $\alpha_k < 1$
  else ▷ *Reject step*
    Set $x_{k+1} \leftarrow x_k$
    Pick $\Delta_{k+1}^{\mathrm{LP}} \in \left[\min(\theta\|d_k\|_{\mathrm{LP}}, \Delta_k^{\mathrm{LP}}), \Delta_k^{\mathrm{LP}}\right]$
  if $\tilde{\rho}_k \geq \rho_s$ **then**
    Set $\Delta_{k+1} \geq \Delta_k$
  else
    Choose $\Delta_{k+1} \in [\kappa_l\|d_k\|, \kappa_u \Delta_k]$
  $k \leftarrow k + 1$
return $x^* = x_k$

3 Convergence analysis of Algorithm 1

We begin our convergence analysis with the introduction of the standing assumptions and a recap of the relevant stationarity concept for (P) in Sect. 3.1. We analyze the criticality measure for this notion of stationarity in the noisy setting in Sect. 3.2. We use these results to prove lower bounds on the trust-region radii that occur in Algorithm 1 in Sect. 3.3, which are then used to obtain sufficient decrease and, as a consequence, convergence of the produced iterates to critical regions in Sect. 3.4.

3.1 Standing assumptions and stationarity

In order to study the convergence properties of Algorithm 1, we make several assumptions regarding the amount of noise, the functions ω, F, and the matrices B_k used in the quadratic models $\tilde{q}_k$.

Assumption 1 We assume that the noise is uniformly bounded via $\|\delta_F(x)\| \leq \varepsilon_F$ and $\|\delta_{F'}(x)\| \leq \varepsilon_{F'}$ for all $x \in \mathbb{R}^n$. We refer to ε_F and $\varepsilon_{F'}$ as the *noise levels* of the functions $\tilde{F}$ and G respectively.

Assumption 2 ω is Lipschitz-continuous with constant L^ω, i.e., it holds for all $x, y \in \mathbb{R}^n$ that

$$|\omega(x) - \omega(y)| \leq L^\omega \|x - y\|.$$

Assumption 3 F and F' are Lipschitz-continuous with constants L^F and $L^{F'}$, i.e., it holds for all $x, y \in \mathbb{R}^n$ that

$$\begin{aligned} \|F(x) - F(y)\| &\leq L^F \|x - y\| \text{ and} \\ \|F'(x) - F'(y)\| &\leq L^{F'} \|x - y\|. \end{aligned}$$

Assumption 4 The Hessian approximations B_k are bounded, i.e., there exists $\beta > 0$ such that for all $d \in \mathbb{R}^n$, $k > 0$ it holds that

$$|\langle d, B_k d \rangle| \leq \beta \|d\|^2.$$

Our aim in the following is to find a local optimum of ϕ. A first-order necessary condition (see [25, p. 184]) of optimality for (P) states that $x^* \in \mathbb{R}^n$ can only be a local optimum if

$$\max_{\lambda \in \partial\omega(F(x^*))} \langle \lambda, F'(x^*)d \rangle \geq 0 \quad \text{for all } d \in \mathbb{R}^n, \tag{7}$$

where $\partial\omega(z)$ denotes the subdifferential of ω at $z \in \mathbb{R}^p$. To measure how close a point $x \in \mathbb{R}^n$ is to satisfying these conditions, we use the *criticality* measure introduced in [26]. Specifically, we let

$$\Psi(x; \Delta) := \phi(x) - \min_{\|d\|_{\mathrm{LP}} \leq \Delta} \ell(x; d)$$

and set $\Psi_k(\Delta) := \Psi(x_k; \Delta)$ as before. Clearly, since $d = 0$ is a feasible solution of the inner optimization problem, the value $\Psi(x; \Delta)$ is always non-negative. On the other hand, the following result establishes that a vanishing reduction over a nontrivial trust region is tantamount to reaching a point satisfying first-order conditions, which we call a *critical point*:

Lemma 3.1 ([26], Lemma 2.1) *A point $x^* \in \mathbb{R}^n$ satisfies conditions (7) iff there exists $\Delta > 0$ such that*

$$\Psi(x^*; \Delta) = 0.$$

As a consequence of this result, an algorithm solving (P) should aim at generating a sequence of iterates such that $\liminf_{k\to\infty} \Psi_k(\Delta) = 0$ for some fixed $\Delta > 0$ (assumed to be 1 in the following). This ensures the existence of an accumulation point x^* of the iterates satisfying first-order conditions.

3.2 Analysis of model function and criticality measure in the presence of noise

Since we do not have access to the values of F and F' required to compute Ψ_k, we define a noisy measure of criticality via

$$\tilde{\Psi}(x; \Delta) := \tilde{\phi}(x) - \min_{\|d\|_{\mathrm{LP}} \le \Delta} \tilde{\ell}(x; d)$$

and set $\tilde{\Psi}_k(\Delta) := \tilde{\Psi}(x_k; \Delta)$ as before. This function is also non-negative if the same realization of the function noise $\delta_F(x_k)$ is used when computing $\tilde{\phi}(x_k)$ and constructing the linear approximation $\tilde{\ell}_k(d)$. We analyze its properties and relationship to Ψ_k below. Since δ_F cannot be assumed to be continuous, neither can $\tilde{\phi}$. This differs from the analysis in [3], where the Lipschitz-continuity of ϕ is used to argue that the reduction ratio approaches one if the trust-region radius is driven to zero. We can, however, state that the criticality measures Ψ_k and $\tilde{\Psi}_k$ are related by the following approximation result: when considering a fixed x_k, we claim that $\tilde{\Psi}_k(1) \to \Psi_k(1)$ for $\varepsilon_F \to 0$ and $\varepsilon_{F'} \to 0$ and that we also have convergence of the minimizers of the convex programs in the definitions of $\tilde{\Psi}_k(1)$ and $\Psi_k(1)$. This follows from the epi-convergence of the functionals

$$\mathcal{L}_{\varepsilon_F, \varepsilon_{F'}}(d) := \tilde{\ell}_k(d) + i_{\|d\|_{\mathrm{LP}} \le \Delta}(d) \quad \text{and} \quad \mathcal{L}_{0,0}(d) := \ell_k(d) + i_{\|d\|_{\mathrm{LP}} \le \Delta}(d),$$

where $i_A : \mathbb{R}^d \to \{0, \infty\}$ is the indicator function of $A \subset \mathbb{R}^d$, that is $i_A(x) = \infty$ if $x \notin A$ and $i_A(x) = 0$ else. We recall that the functionals $\mathcal{L}_{\varepsilon_F, \varepsilon_{F'}}$ epi-converge to $\mathcal{L}_{0,0}$ if and only if for all $d \in \mathbb{R}^n$ the inequalities

$$\mathcal{L}_{0,0}(d) \le \liminf_{\varepsilon_F, \varepsilon_{F'} \to 0} \mathcal{L}_{\varepsilon_F, \varepsilon_{F'}}(d_\varepsilon) \text{ for all sequences } d_\varepsilon \to d \text{ and} \tag{8}$$

$$\mathcal{L}_{0,0}(d) \ge \limsup_{\varepsilon_F, \varepsilon_{F'} \to 0} \mathcal{L}_{\varepsilon_F, \varepsilon_{F'}}(d_\varepsilon) \text{ for some sequence } d_\varepsilon \to d \tag{9}$$

hold, see, for example, [27, § 7], which is shown below.

Proposition 3.2 *Let $\Delta > 0$. Then the functionals $\mathcal{L}_{\varepsilon_F,\varepsilon_{F'}}$ epi-converge to $\mathcal{L}_{0,0}$ for $\varepsilon_F \to 0$ and $\varepsilon_{F'} \to 0$. In particular, $\tilde{\Psi}_k(1) \to \Psi_k(1)$ for $\varepsilon_F \to 0$ and $\varepsilon_{F'} \to 0$ in case of a fixed $x_k \in \mathbb{R}^n$.*

Proof We begin by showing the first inequality and consider $d_\varepsilon \to d$. W.l.o.g. we assume that $\liminf_{\varepsilon_F,\varepsilon_{F'}\to 0} \mathcal{L}_{\varepsilon_F,\varepsilon_{F'}}(d_\varepsilon) < C$ for some $C > 0$, which implies that there is a subsequence (denoted by d'_ε) such that $\|d'_\varepsilon\|_{\mathrm{LP}} \leq \Delta$ for all elements d'_ε. The continuity of the norm $\|\cdot\|_{\mathrm{LP}}$ yields $0 = i_{\|d'_\varepsilon\|_{\mathrm{LP}}\leq\Delta} = i_{\|d\|_{\mathrm{LP}}\leq\Delta}$ for all elements of the subsequence inside the trust region with radius Δ. Moreover, $\tilde{F}(x_k) \to F(x_k)$ and $G(x_k) \to F'(x_k)$ for $\varepsilon_F \to 0$ and $\varepsilon_{F'} \to 0$ and thus the continuity of ω gives $\tilde{\ell}_k(d_\varepsilon) \to \ell_k(d)$ and in turn the first inequality.

We continue with the second inequality and consider the constant sequence $d_\varepsilon := d$. Then $i_{\|d_\varepsilon\|_{\mathrm{LP}}\leq\Delta} = i_{\|d\|_{\mathrm{LP}}\leq\Delta}$ and $\tilde{F}(x_k) \to F(x_k)$ and $G(x_k) \to F'(x_k)$ for $\varepsilon_F \to 0$ and $\varepsilon_{F'} \to 0$. Again, the continuity of ω gives $\tilde{\ell}_k(d_\varepsilon) \to \ell_k(d)$ and in turn the second inequality.

The functionals $\mathcal{L}_{\varepsilon_F,\varepsilon_{F'}}$ always admit a minimizer because the feasible set $\{d \mid \|d\|_{\mathrm{LP}} \leq \Delta\}$, on which $\mathcal{L}_{\varepsilon_F,\varepsilon_{F'}}$ is finite, is compact. A standard argument yields that all accumulation points of a sequence of minimizers of the functionals $\mathcal{L}_{\varepsilon_F,\varepsilon_{F'}}$ minimize the limit functional $\mathcal{L}_{0,0}$. □

Consequently, if we drive $\tilde{\Psi}_k$ to zero over the iterations, we have an upper bound on Ψ_k, defining a critical region (sublevel set) into which the iterates converge.

Lemma 3.3 *Under Assumption 3 it holds for all $x, d \in \mathbb{R}^n$ that*

$$\|F(x+d) - F(x) - F'(x)d\| \leq L^{F'}\|d\|^2.$$

Proof This follows directly from Assumption 3 and the mean value theorem. □

Lemma 3.4 *Under Assumptions 1–4, it holds that*

$$|\tilde{\phi}(x_k + d_k) - \tilde{q}_k(d_k)| \leq M_0^\varepsilon + M_1^\varepsilon\|d_k\| + M_2^\varepsilon\|d_k\|^2, \tag{10}$$

where $M_0^\varepsilon := 2L^\omega\varepsilon_F$, $M_1^\varepsilon := L^\omega\varepsilon_{F'}$, and $M_2^\varepsilon := L^\omega L^{F'} + \frac{1}{2}\beta$.

Proof The assumed Lipschitz continuity of ω, F, F', the representations $\tilde{F} = F + \delta_F$, $G = F' + \delta_{F'}$, Lemma 3.3, and the bounds on δ_F, $\delta_{F'}$ yield the claim by repeated application of the triangle inequality. □

Remark 3.5 If Assumptions 2 to 4 hold in the noiseless case, the achievable reduction is related to the trial value via $|q_k(d_k) - \phi(x_k + d_k)| \leq M\|d_k\|^2$ for some $M > 0$. This is no longer the case in the noisy model.

Several of the following results are due to [3] and are largely unaffected by moving from the noiseless to the noisy regime. We refer to their counterparts in [3] and prove them in the appendix. We begin by establishing that the linearized model $\tilde{\ell}$ is still Lipschitz-continuous, albeit with a Lipschitz-constant affected by the noise level $\varepsilon_{F'}$:

Lemma 3.6 *Under Assumptions 1–3 it holds for all* $d \in \mathbb{R}^n$ *that*

$$|\tilde{\ell}_k(d) - \tilde{\ell}_k(0)| \leq L_\varepsilon^\ell \|d\|_{\mathrm{LP}},$$

where $L_\varepsilon^\ell := \gamma L^\omega (L^{F'} + \varepsilon_{F'})$.

Proof The assumed Lipschitz continuity of ω, F', the representation $G + \delta_{F'}$, Lemma 3.3. and the bound on $\delta_{F'}$ and the norm estimation (6) yield the claim by repeated application of the triangle inequality. □

We proceed to examine the reduction according to the partially linearized model as a function of the size of an improvement step. The following result establishes that the reduction is well behaved in the step size in the following sense: the model reduction that is achieved for a reduced step size is bounded from below by the model reduction achieved without step reduction multiplied by the step reduction.

Lemma 3.7 *It holds for all* $\alpha \in [0, 1]$ *that*

$$\tilde{\phi}(x_k) - \tilde{\ell}_k(\alpha d) \geq \alpha \left[\tilde{\phi}(x_k) - \tilde{\ell}_k(d) \right].$$

Proof This follows directly from the fact that $\tilde{\ell}_k$ is convex, where we note that $\tilde{\ell}_k(0) = \tilde{\phi}(x_k)$ holds for $d = 0$. □

Next, we establish that the criticality $\tilde{\Psi}_k(\Delta)$ for a given trust-region radius $\Delta > 0$ is bounded below by $\tilde{\Psi}_k(1)$ multiplied by Δ if the latter is less than one. This holds in particular during the computation of the LP step in Algorithm 1. The proof requires the relationship established in Lemma 3.7.

Lemma 3.8 (Lemma 3.1 in [3]) *It holds for any* $\Delta > 0$ *that*

$$\tilde{\Psi}_k(\Delta) \geq \min(\Delta, 1) \tilde{\Psi}_k(1).$$

Proof The proof is in Appendix A. □

The next result states that when progress is possible with respect to the criticality $\tilde{\Psi}_k(1)$, the LP step either lies on the trust-region boundary or has a norm proportional to $\tilde{\Psi}_k(1)$.

Lemma 3.9 (Lemma 3.2 in [3]) *Suppose that Assumptions 1–3 hold and that* $\tilde{\Psi}_k(1) \neq 0$. *Let* d_Δ *be a minimizer achieving* $\tilde{\Psi}_k(\Delta)$ *for some* $\Delta > 0$. *Then it follows that*

$$\|d_\Delta\|_{\mathrm{LP}} \geq \min\left(\Delta, \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell} \right).$$

Proof The proof is in Appendix A. □

We are now ready to examine the step d_k computed by Algorithm 1 with respect to the reduction achieved by the model $\tilde{q}_k$. Specifically, if progress can be made with respect to the criticality $\tilde{\Psi}_k(1)$, then we can expect a positive reduction in $\tilde{q}_k$. We use this result to prove that the objective $\tilde{\phi}$ decreases as well as long as $\tilde{\Psi}_k(1)$ is sufficiently large.

Lemma 3.10 (Lemma 3.3 in [3]) *The model decrease satisfies*

$$\tilde{\phi}(x_k) - \tilde{q}_k(d_k) \geq \tilde{\phi}(x_k) - \tilde{q}_k(d_k^C) \geq \eta\alpha_k \tilde{\Psi}_k(\Delta_k^{\mathrm{LP}}) \geq \eta\alpha_k \min(\Delta_k^{\mathrm{LP}}, 1)\tilde{\Psi}_k(1).$$

Proof The proof is in Appendix A. □

The following technical lemma shows that if $\tilde{\Psi}_k(1) \neq 0$, then $\|d_k^C\|_{\mathrm{LP}}$ is bounded below, which we will need to ensure that the updated trust region radii do not collapse while progress in the objective can still be made.

Lemma 3.11 (Lemma 3.5 in [3]) *Under Assumptions 1–4 it holds that*

$$\begin{aligned}\alpha_k \Delta_k^{\mathrm{LP}} &\geq \|d_k^C\|_{\mathrm{LP}} \\ &\geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell}, \min\left(1, \frac{1}{\Delta_k^{\mathrm{LP}}}\right)\frac{2(1-\eta)\tau\tilde{\Psi}_k(1)}{\beta\gamma^2}\right).\end{aligned}$$

Proof The proof is in Appendix A. □

3.3 Lower bounds on the trust-region radii

We are now able to state a key result that provides lower bounds on both the trust-region radius Δ for the quadratic model and the LP trust-region radius Δ^{LP}. It ensures that the algorithm does not stall while progress can be made with respect to the noisy criticality $\tilde{\Psi}_k$. The proof strategy follows Lemma 3.6 in [3] for the noiseless case. In order to compensate for the noise, we need to assume a sufficiently large stabilization parameter ϑ, which in turn depends on the constants introduced by the noise.

Lemma 3.12 *Consider an application of Algorithm 1 to the noisy variant of problem (P). Suppose that Assumptions 1–4 hold, $\tilde{\Psi}_k(1) \geq \delta > 0$ for all k, and that*

$$\vartheta \geq \vartheta_\varepsilon^* := \frac{M_0^\varepsilon + M_1^\varepsilon}{1 - \rho_s}, \tag{11}$$

with M_0^ε, M_1^ε from Lemma 3.4. Then it follows that

$$\Delta_k \geq \Delta_{\min} \quad \text{and} \quad \alpha_k \Delta_k^{\mathrm{LP}} \geq \frac{\Delta_{\min}}{\gamma},$$

where $\Delta_{\min} = \min(A, B\delta)$ *with*

$$A := \min\left(\theta, \kappa_l, \Delta_0, \Delta_0^{\mathrm{LP}}\gamma\right), \text{ and}$$
$$B := \min\left(\frac{\theta^2(1-\rho_u)\eta}{\gamma M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}, \frac{\kappa_l^2(1-\rho_s)\eta}{\gamma M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}, \frac{\gamma}{L_{\varepsilon}^{\ell}}, \frac{2(1-\eta)\tau}{\beta\gamma\Delta_{\max}^{\mathrm{LP}}}\right).$$

Proof Using $\tilde{\Psi}_k(1) \geq \delta$ and $\Delta_{\max}^{\mathrm{LP}} \geq 1$ the bound in Lemma 3.11 becomes

$$\begin{aligned}\|d_k^C\|_{\mathrm{LP}} &\geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \frac{\tilde{\Psi}_k(1)}{L_{\varepsilon}^{\ell}}, \min\left(1, \frac{1}{\Delta_k^{\mathrm{LP}}}\right)\frac{2(1-\eta)\tau\tilde{\Psi}_k(1)}{\beta\gamma^2}\right)\\ &\geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \frac{\delta}{L_{\varepsilon}^{\ell}}, \frac{2(1-\eta)\tau\delta}{\Delta_{\max}^{\mathrm{LP}}\beta\gamma^2}\right) = \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right), \quad (12)\end{aligned}$$

where

$$\Delta_{\mathrm{crit}} := \min\left(\frac{\delta}{L_{\varepsilon}^{\ell}}, \frac{2(1-\eta)\tau\delta}{\Delta_{\max}^{\mathrm{LP}}\beta\gamma^2}\right).$$

If a step is accepted in the k-th iteration (that is $\tilde{\rho}_k \geq \rho_u$), it follows that

$$\Delta_{k+1}^{\mathrm{LP}} \geq \|d_k^C\|_{\mathrm{LP}} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right). \quad (13)$$

If, on the other hand, the step is rejected, we can deduce the inequalities

$$1-\rho_u < 1-\tilde{\rho}_k = \frac{\tilde{\phi}(x_k+d_k) - \tilde{q}_k(d_k)}{\tilde{\phi}(x_k) - \tilde{q}_k(d_k) + \vartheta} \leq \frac{M_0^{\varepsilon} + M_1^{\varepsilon}\|d_k\| + M_2^{\varepsilon}\|d_k\|^2}{\eta\alpha_k \min\left(\Delta_k^{\mathrm{LP}}, 1\right)\tilde{\Psi}_k(1) + \vartheta}$$

from Lemmas 3.4 and 3.10. Based on the bounds $\tilde{\Psi}_k(1) \geq \delta$, $\Delta_k^{\mathrm{LP}} \leq \Delta_{\max}^{\mathrm{LP}}$, and $\|d_k^C\|_{\mathrm{LP}} = \|\alpha_k d_k^{\mathrm{LP}}\|_{\mathrm{LP}} \leq \alpha_k\Delta_k^{\mathrm{LP}}$, we can estimate the denominator via

$$\begin{aligned}\eta\alpha_k \min\left(\Delta_k^{\mathrm{LP}}, 1\right)\tilde{\Psi}_k(1) + \vartheta &\geq \eta\alpha_k\Delta_k^{\mathrm{LP}}\min\left(1, \frac{1}{\Delta_k^{\mathrm{LP}}}\right)\delta + \vartheta\\ &\geq \eta\|d_k^C\|_{\mathrm{LP}}\frac{\delta}{\Delta_{\max}^{\mathrm{LP}}} + \vartheta\\ &\underset{(12)}{\geq} \eta\min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right)\frac{\delta}{\Delta_{\max}^{\mathrm{LP}}} + \vartheta.\end{aligned}$$

Consequently, we obtain the relationship

$$(1-\rho_u)\eta\min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right)\frac{\delta}{\Delta_{\max}^{\mathrm{LP}}} + (1-\rho_u)\vartheta < M_0^{\varepsilon} + M_1^{\varepsilon}\|d_k\| + M_2^{\varepsilon}\|d_k\|^2. \quad (14)$$

Furthermore, the choice of ϑ together with $0 < \rho_u \leq \rho_s < 1$ implies that

$$\vartheta \geq \frac{M_0^\varepsilon + M_1^\varepsilon}{1-\rho_s} \geq \frac{M_0^\varepsilon + M_1^\varepsilon}{1-\rho_u}. \tag{15}$$

To obtain a lower bound on $\Delta_{k+1}^{\mathrm{LP}}$, we distinguish two cases with respect to $\|d_k\|$:

1. If $\|d_k\| \geq 1$, it follows that $\theta\|d_k\|_{\mathrm{LP}} \geq \frac{\theta}{\gamma}\|d_k\| \geq \frac{\theta}{\gamma}$. Using that $\Delta_{k+1}^{\mathrm{LP}} \in [\min(\theta\|d_k\|_{\mathrm{LP}}, \Delta_k^{\mathrm{LP}}), \Delta_k^{\mathrm{LP}}]$ implies that $\Delta_{k+1}^{\mathrm{LP}} \geq \min\left(\frac{\theta}{\gamma}, \Delta_k^{\mathrm{LP}}\right)$.
2. If $\|d_k\| < 1$, (14) and the lower bound on ϑ given by (15) imply that

$$\frac{1-\rho_u}{M_2^\varepsilon}\eta \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \frac{\delta}{\Delta_{\max}^{\mathrm{LP}}} < \|d_k\|^2. \tag{16}$$

We now make a case distinction on the minimizer resulting from Line 15 in Algorithm 1. If $\Delta_{k+1}^{\mathrm{LP}}$ has a value of at least $\theta\|d_k\|_{\mathrm{LP}}$, we obtain

$$\begin{aligned}
\left(\Delta_{k+1}^{\mathrm{LP}}\right)^2 &\geq \theta^2\|d_k\|_{\mathrm{LP}}^2 \geq \theta^2/\gamma^2\|d_k\|^2 && (17)\\
&> \frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^\varepsilon \Delta_{\max}^{\mathrm{LP}}} \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) && (18)\\
&\geq \min\left(\frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^\varepsilon \Delta_{\max}^{\mathrm{LP}}}, \frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right)^2 && (19)
\end{aligned}$$

If $\theta\|d_k\|_{\mathrm{LP}} > \Delta_{k+1}^{\mathrm{LP}}$, then

$$\Delta_{k+1}^{\mathrm{LP}} \geq \min\left(\frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^\varepsilon \Delta_{\max}^{\mathrm{LP}}}, \frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \tag{20}$$

holds true.

We combine all of the above cases by taking their minimum, resulting in

$$\Delta_{k+1}^{\mathrm{LP}} \geq \min\left(\frac{\theta}{\gamma}, \frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^\varepsilon \Delta_{\max}^{\mathrm{LP}}}, \frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right). \tag{21}$$

This lower bound on $\Delta_{k+1}^{\mathrm{LP}}$ dominates the previously shown lower bound (13) for accepted iterates. In order to derive a uniform lower bound on Δ_k^{LP} (that is independent of k), we may assume the worst case, i.e. all steps are rejected, and resort to only (21).

Regarding trust-region radius Δ_{k+1} for the quadratic model, we can follow a similar chain of reasoning as for $\Delta_{k+1}^{\mathrm{LP}}$: If the reduction ratio $\tilde{\rho}_k$ achieved by the step is at least ρ_s, then we know that $\Delta_{k+1} \geq \Delta_k$. Otherwise, it holds that $\Delta_{k+1} \geq \kappa_l\|d_k\|$. We can use the same estimations as before to derive a bound analogous to (14):

$$(1-\rho_s)\eta \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \frac{\delta}{\Delta_{\max}^{\mathrm{LP}}} + (1-\rho_s)\vartheta < M_0^\varepsilon + M_1^\varepsilon\|d_k\| + M_2^\varepsilon\|d_k\|^2.$$

The case distinction regarding $\|d_k\|$ is similar: If $\|d_k\| \geq 1$, then it must hold that $\Delta_{k+1} \geq \kappa_l$. Otherwise, we can deduce the inequality

$$\frac{\Delta_{k+1}}{\gamma} \geq \min\left(\frac{\kappa_l^2(1-\rho_s)\eta\delta}{\gamma^2 M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}, \frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right).$$

using the same steps as before, where κ_l has the role of θ in the argument before (after dividing Δ_{k+1} by γ). We can combine these two cases to derive a lower bound on Δ_{k+1}, which again holds uniformly regardless of whether the step is accepted or not.

Combining the lower bounds on $\Delta_{k+1}^{\mathrm{LP}}$ and $\Delta_{k+1}^{\mathrm{LP}}/\gamma$, we obtain the lower bound

$$\min\left(\frac{\Delta_{k+1}}{\gamma}, \Delta_{k+1}^{\mathrm{LP}}\right) \geq \min\left(\frac{\theta}{\gamma}, \frac{\kappa_l}{\gamma}, \frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}, \frac{\kappa_l^2(1-\rho_s)\eta\delta}{\gamma^2 M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}\right).$$

Starting from some k, we apply the inequality above recursively while decrementing k and arrive at

$$\min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}\right) \geq \min\left(\frac{\theta}{\gamma}, \frac{\kappa_l}{\gamma}, \frac{\Delta_0}{\gamma}, \Delta_0^{\mathrm{LP}}, \Delta_{\mathrm{crit}}, \frac{\theta^2(1-\rho_u)\eta\delta}{\gamma^2 M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}, \frac{\kappa_l^2(1-\rho_s)\eta\delta}{\gamma^2 M_2^{\varepsilon}\Delta_{\max}^{\mathrm{LP}}}\right) =: \Delta_{\mathrm{low}}.$$

It must therefore hold that $\Delta_k \geq \gamma\Delta_{\mathrm{low}}$. Moreover, it holds that

$$\alpha_k \Delta_k^{\mathrm{LP}} \geq \|d_k^C\|_{\mathrm{LP}} \geq \min\left(\frac{\Delta_k}{\gamma}, \Delta_k^{\mathrm{LP}}, \Delta_{\mathrm{crit}}\right) \geq \min\left(\Delta_{\mathrm{low}}, \Delta_{\mathrm{crit}}\right) = \Delta_{\mathrm{low}}.$$

The result follows from grouping the terms in $\gamma\Delta_{\mathrm{low}}$, according to whether or not they contain δ. □

3.4 Global convergence theorem

We are now ready to establish the convergence of Algorithm 1. In order to simplify the proof of the main theorem, we separately handle the special case in which Algorithm 1 converges in a finite number of iterations to a point x^*, which is critical with respect to the noisy function $\tilde{\phi}$:

Lemma 3.13 (Corollary 3.7 in [3]) *Consider an application of Algorithm 1 to the noisy variant of problem (P). Suppose that Assumptions 1 to 4 and (11) hold. If there are finitely many successful iterations (that is $\tilde{\rho}_k \geq \rho_u$) during the execution of Algorithm 1, then it holds that x_k is invariant and critical for all sufficiently large k. In particular,*

$$x_k \equiv x^* \quad \text{and} \quad \tilde{\Psi}_k(1) = 0$$

for some x^* and all sufficiently large k.

Proof The proof is in Appendix A. □

The following convergence theorem states that when the objective of (P) is bounded below, an application of Algorithm 1 will produce one of two possible mutually exclusive outcomes: the algorithm may stop at a critical point after a finite number of iterations as described in Lemma 3.13 or, alternatively, Algorithm 1 visits a critical region infinitely often. In terms of the functions ω, $\tilde{F}$, and G, the *critical region* is defined as

$$C(\delta) := \left\{ x \;\middle|\; \tilde{\Psi}(x; 1) \leq \delta \right\}.$$

By definition, an iterate x_k produced during the execution of Algorithm 1 is contained in $C(\delta)$ if $\tilde{\Psi}_k(1) \leq \delta$. What is more, Proposition 3.2 establishes that $\tilde{\Psi}_k(1)$ tends to $\Psi_k(1)$ as the errors ε_F and $\varepsilon_{F'}$ approach zero. These results therefore suggest that the iterate is *close* to being optimal in the sense of Lemma 3.1.

Theorem 3.14 *Consider an application of Algorithm 1 to the noisy variant of problem (P). Suppose that Assumptions 1 to 4 and (11) hold. Then either*

$$\tilde{\Psi}_k(1) = 0 \text{ for some } k \geq 0,$$

or

$$\lim_{k \to \infty} \tilde{\phi}(x_k) = -\infty,$$

or there are infinitely many $k \in \mathbb{N}$ such that $x_k \in C(\delta_{\max})$, where

$$\delta_{\max} := \max\left(\sqrt{\frac{\vartheta(1-\rho_u)\gamma\,\Delta_{\max}^{\mathrm{LP}}}{\rho_u \eta B}}, \frac{\vartheta(1-\rho_u)\gamma\,\Delta_{\max}^{\mathrm{LP}}}{\rho_u \eta A} \right)$$

is given in terms of the constants A, B from Lemma 3.12.

Proof If there are only finitely many accepted steps, the result follows from Lemma 3.13, yielding the first possibility. Otherwise, we can assume that during the algorithm, an infinite number of accepted steps occurs. If $\tilde{\phi}(x_k)$ tends to $-\infty$, the second possibility occurs, so we can assume in the following that $\tilde{\phi}(x_k)$ (and hence $\phi(x_k)$) is bounded below.

Let $\mathcal{K}$ be the sequence of accepted steps, i.e., consisting of those k where $x_{k+1} \neq x_k$. Clearly, if $\liminf_{k \to \infty} \tilde{\Psi}_k(1) = 0$, then the result follows. So we can assume that there exists a $\delta > 0$ such that $\tilde{\Psi}_k(1) \geq \delta$ for all $k \geq k_0$. The claim stating that the region $C(\delta_{\max})$ is visited infinitely often is tantamount to ensuring that $\delta \leq \delta_{\max}$, which will be the aim of the remainder of this proof. For each $k \in \mathcal{K}$, $k \geq k_0$ we have that

$$\frac{\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) + \vartheta}{\tilde{\phi}(x_k) - \tilde{q}_k(d_k) + \vartheta} \geq \rho_u > 0.$$

We deduce using Lemma 3.10 that

$$\begin{aligned}\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) &\geq \rho_u \left[\tilde{\phi}(x_k) - \tilde{q}_k(d_k)\right] + (\rho_u - 1)\vartheta \\ &\geq \rho_u \left[\eta\alpha_k \min(\Delta_k^{\mathrm{LP}}, 1)\tilde{\Psi}_k(1)\right] + (\rho_u - 1)\vartheta.\end{aligned}$$

It follows that

$$\begin{aligned}\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) &\geq \rho_u \left[\eta\alpha_k \Delta_k^{\mathrm{LP}} \min\left(\frac{1}{\Delta_k^{\mathrm{LP}}}, 1\right)\tilde{\Psi}_k(1)\right] + (\rho_u - 1)\vartheta \\ &\geq \rho_u \left[\eta\alpha_k \Delta_k^{\mathrm{LP}} \frac{1}{\Delta_{\max}^{\mathrm{LP}}}\tilde{\Psi}_k(1)\right] + (\rho_u - 1)\vartheta.\end{aligned}$$

We can now apply Lemma 3.12 to bound $\alpha_k \Delta_k^{\mathrm{LP}}$ below based on $\Delta_{\min}$ and the constants A and B:

$$\begin{aligned}\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) &\geq \rho_u \left[\eta \frac{\Delta_{\min}}{\gamma \Delta_{\max}^{\mathrm{LP}}}\tilde{\Psi}_k(1)\right] + (\rho_u - 1)\vartheta \\ &= \rho_u \left[\eta \frac{\min(A, B\delta)}{\gamma \Delta_{\max}^{\mathrm{LP}}}\tilde{\Psi}_k(1)\right] + (\rho_u - 1)\vartheta \\ &\geq \rho_u \left[\eta \frac{\min(A, B\delta)}{\gamma \Delta_{\max}^{\mathrm{LP}}}\delta\right] + (\rho_u - 1)\vartheta.\end{aligned}$$

Let us assume towards a contradiction that $\delta > \delta_{\max}$. We distinguish two cases with respect to the minimum $\min(A, B\delta)$:

1. The minimum is attained at A, implying that

$$\begin{aligned}\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) &\geq \rho_u \left[\eta \frac{A}{\gamma \Delta_{\max}^{\mathrm{LP}}}\delta\right] + (\rho_u - 1)\vartheta \\ &\geq \rho_u \left[\eta \frac{A}{\gamma \Delta_{\max}^{\mathrm{LP}}}\delta_{\max}\right] + (\rho_u - 1)\vartheta + C_1.\end{aligned}$$

   for a constant $C_1 > 0$. Using the fact that

$$\delta_{\max} \geq \frac{\vartheta(1 - \rho_u)\gamma \Delta_{\max}^{\mathrm{LP}}}{\rho_u \eta A}$$

   by definition of $\delta_{\max}$, this implies that $\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) \geq C_1 > 0$.
2. The minimum is attained at $B\delta$, implying that

$$\begin{aligned}\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) &\geq \rho_u \left[\eta \frac{B\delta^2}{\gamma \Delta_{\max}^{\mathrm{LP}}}\right] + (\rho_u - 1)\vartheta \\ &\geq \rho_u \left[\eta \frac{B\delta_{\max}^2}{\gamma \Delta_{\max}^{\mathrm{LP}}}\right] + (\rho_u - 1)\vartheta + C_2\end{aligned}$$

for a constant $C_2 > 0$. We now use the fact that

$$\delta_{\max} \geq \sqrt{\frac{\vartheta(1-\rho_u)\gamma\,\Delta_{\max}^{\mathrm{LP}}}{\rho_u \eta B}}$$

to deduce that $\tilde{\phi}(x_k) - \tilde{\phi}(x_{k+1}) \geq C_2 > 0$.

In either case $\tilde{\phi}$ decreases by $\min(C_1, C_2) > 0$ from x_k to x_{k+1}. Since this decrease is strictly positive and there are infinitely many accepted steps in the sequence $\mathcal{K}$, it follows that $\tilde{\phi}(x_k)$ tends to $-\infty$, which is a contradiction. It must therefore hold that $\delta \leq \delta_{\max}$ as desired. □

Interpretation of Theorem 3.14

In theoretical terms, the result in Theorem 3.14 is as expected: the size of the critical region C depends on the stabilization parameter ϑ, If we increase ϑ, Algorithm 1 can tolerate a larger amount of noise at the cost of a decreased accuracy with respect to the criticality measure $\tilde{\Psi}$. Of course, problem (P) can generally also be unbounded. The remaining case, where $\tilde{\Psi}_k(1) = 0$ for some $k \in \mathbb{N}$, can involve different scenarios. If $\Psi_k(1) = 0$ holds as well, the iterate x_k is a critical point of (P), which is the ideal situation. Otherwise, the noises $\delta_F(x_k)$ and $\delta_{F'}(x_k)$ attain values such that x_k *appears to be critical* in the noisy model. In case of an unconstrained version of (NLP), this is tantamount to a non-zero gradient that is canceled out by noise.

For a function F that is not afflicted by noise, i.e., satisfying $\varepsilon_F = \varepsilon_{F'} = 0$, it holds that $M_0^\varepsilon = M_1^\varepsilon = 0$, where M_0^ε and M_1^ε are the constants from Lemma 3.4. This allows us to set $\vartheta = 0$, whereby we recover the original algorithm discussed in [3]. If we apply Theorem 3.14 in this situation, it follows from $\vartheta = 0$ that $\delta_{\max} = 0$, implying that the region $C(\delta_{\max})$ contains precisely the points critical with respect to $\tilde{\phi}$, which itself coincides with ϕ in this particular case. Thus, the original convergence result [3, Theorem 3.8] follows for Algorithm 1 in the noiseless case.

As mentioned in the introduction, we can also apply Algorithm 1 to solve smooth unconstrained nonlinear problems affected by noise. To this end, we can set $\omega(x) = x$, achieving a Lipschitz constant of $L^\omega = 1$. Lemma 3.12, specifically (11), then suggests a stabilization of

$$\vartheta \geq \frac{2\varepsilon_F + \varepsilon_{F'}}{1-\rho_s} > 2\varepsilon_F + \varepsilon_{F'}, \tag{22}$$

which may be weaker than the stabilization of $r\varepsilon_F$ analyzed in [1] if $\varepsilon_{F'}$ becomes sufficiently large. This is particularly true for all $\varepsilon_{F'} > 0$ for the choice $r = 2/(1-\rho_s)$, which corresponds to the choice $r = 2/(1-c_2)$ in (8) in [1].

The reason is the estimate in Lemma 3.7 based on the convexity of ω. Conversely, the authors of [1] have a Lipschitz continuous derivative of the objective at hand. In that case, the criticality measure satisfies

$$x \in C(\delta) \iff \left(-\min_{\|d\|_{\mathrm{LP}} \leq 1} G(x)d\right) \leq \delta,$$

Table 1 Parameters used for numerical experiments

Symbol	Explanation	Value
Δ_0^{LP}	Initial LP trust-region radius	1
$\Delta_{\max}^{\mathrm{LP}}$	Maximum LP trust-region radius	10
Δ_0	Initial trust-region radius	1
ρ_u	Step acceptance threshold	0.1
ρ_s	Threshold for increase of Δ	0.5
κ_l	Lower bound for adjustment of Δ after failed step	0.1
κ_u	Upper bound for adjustment of Δ after failed step	0.8
θ	Lower bound for adjustment of Δ^{LP} after failed step	0.5
η	Factor of relative decrease for Cauchy step	0.1
τ	Shortening factor for Cauchy line search	0.5

which in turn is equivalent to $\|G(x)\|$ being bounded above by a constant.

4 Numerical experiments

In order to illustrate the performance and examine the behavior of Algorithm 1, we implement the algorithm[1] in `Python` (3.11.5), using `numpy` [28] (1.26.0), `scipy` [29] (1.11.3) (including `HiGHS` [23] (1.5.0) as LP solver), and `Ipopt` [30] (3.14.13) to solve the respective subproblems. We generally compare the performance of the *classical* algorithm (i.e., Algorithm 1 with a stabilization of $\vartheta = 0$), with its *stabilized* counterpart (where $\vartheta > 0$).

In terms of termination criteria, we first of all impose an iteration limit, after which the algorithm terminates. Secondly, we monitor the LP trust-region radius, Δ^{LP}. If the radius contracts to a value close to zero (1×10^{-10}), we see this as a failure of the algorithm and let it terminate. Lastly, when the noise criticality $\tilde{\Psi}_k(1)$ falls below the threshold of 1×10^{-6}, we terminate the algorithm, knowing that the current iterate is very close to being optimal. We then examine the *final* iterate x^f, i.e., the iterate x^k in Algorithm 1 of the iteration at which the termination criterion becomes satisfied. If not indicated otherwise, we choose the parameters of Algorithm 1 according to the values in Table 1 and the stabilization parameter ϑ_ε^*. In order to obtain inexact evaluations of a given function $F : \mathbb{R}^n \to \mathbb{R}^p$ and its derivative, we inject noise by setting

$$\begin{aligned} \delta_F(x) &= X_F \in \mathbb{R}^p, \ X_F \sim \mathbb{B}_n(\varepsilon_F), \text{ and} \\ \delta_{F'}(x) &= X_{F'} \in \mathbb{R}^{p\times n}, \ X_{F'} \sim \mathbb{B}_{np}(\varepsilon_{F'}), \end{aligned} \tag{23}$$

where $\mathbb{B}_n(s)$ denotes the uniform distribution on the n-dimensional Euclidean ball centered at the origin with radius $s \geq 0$. We can compute a sample from this distribution

[1] Available at https://github.com/chrhansk/noisy-nonlinear.

by first sampling n times from a standard normal distribution and scaling the resulting vector [31] by its Euclidean norm to obtain a sample from the uniform distribution on $\mathbb{S}_n(1)$. We can then use inverse transform sampling to obtain a suitable radius in $[0, s]$ and rescale the vector accordingly. The injected noise is bounded according to the noise levels ε_F and $\varepsilon_{F'}$ while being sufficiently unpredictable to significantly affect the solution process.

In Sect. 4.1, we augment an example of a quadratic test problem from [1] with a non-smooth term. We obtain qualitatively similar results in this case. Then we compare and visualize the different behaviors of the unstabilized algorithm and the stabilized algorithm for the Rosenbrock test function in Sect. 4.2. In Sect. 4.3, we apply the algorithm to an image reconstruction problem with total variation regularization and assess the impact of different choices of the stabilization parameter. Finally, in Sect. 4.4, we apply the algorithm in a penalty method for a small constrained optimization problem from `CUTest` [32] as motivated in the introduction, which points to future research directions. Sections 4.1, 4.2 and 4.4 use essentially the same type of a non-smooth objective function that includes an ℓ^1-penalty term and we provide the required estimates of its Lipschitz constant in Appendix B.

4.1 Failure of the classical algorithm

To illustrate the difference in performance between the classical algorithm and Algorithm 1, we consider the case of ℓ^1-penalized optimization problems of the form $x \mapsto f(x) + \lambda\|x\|_1$ with a smooth function $f : \mathbb{R}^n \to \mathbb{R}$. It is clear that these problems are non-smooth due to the presence of the $\|\cdot\|_1$ term, while being expressible as problems of type (P) based on suitable choices of ω and F. This problem class also enables us to minimize $\tilde{\ell}_k$ and $\tilde{q}_k$ over the trust regions defined in terms of Δ^{LP} and Δ by solving linear or quadratic programs respectively. What is more, the only curvature information in this problem class is due to f, enabling us to either use the Hessian of f (or any quasi-Newton approximation) to obtain the matrices B_k. We specifically examine the case where f is a quadratic of the form

$$f(x) = \tfrac{1}{2}\langle x, Dx\rangle,$$

where D is the matrix in $\mathbb{R}^{n\times n}$ for $n = 8$ given as

$$D = \operatorname{diag}(10^{-5}, 10^{-4.75}, 10^{-4.5}, \ldots, 10^{-3.25}),$$

taken from [1], where this optimization problem has been studied without an ℓ^1-penalty. It is apparent that the optimal solution of this instance of (P) is $x^* = 0$. We set the parameter λ to 1×10^{-2} while injecting noise according to (23) with noise levels $\varepsilon_F = 1 \times 10^{-1}$ and $\varepsilon_{F'} = 1 \times 10^{-5}$, and initialize Algorithm 1 with the initial point $x_0 = (1000, 0, \ldots, 0)^T$, limiting the number of iterations to 50, and performing quadratic steps based on the true Hessian D.

We show an example of the difference in performance in Fig. 1, where the values of the reduction ratio are clipped to ± 5 in order to properly display the results. We

see that the classical algorithm performs dramatically worse than its stabilized counterpart. Indeed, the classical algorithm stalls almost immediately, due to the reduction ratio ρ_k becoming unreliable. Consequently, the LP trust region collapses, and the classical algorithm makes no progress towards optimality. Conversely, the addition of a stabilization yields an algorithm rapidly approaching the optimum, both in terms of primal distance and objective value while maintaining a reasonably large LP trust-region radius. Similarly, noisy and noiseless criticality decrease rapidly throughout the iterations of the stabilized algorithm. Unfortunately, the criticality bound established in Theorem 3.14 attains a value of $\delta_{\max} \approx 54\,000$, limiting its use in terms of the criticality actually achieved throughout the iterations.

We would like to point out that the failure of the classical algorithm is not guaranteed in this scenario: By running the experiment with 100 different random seeds we found that the classical algorithm stalls in about half (57) of the cases, while performing well in the other half. Conversely, the stabilized algorithm consistently performs well in all cases. Its characteristics are qualitatively similar to the case that is depicted in Fig. 1. A key problem of the classical algorithm is therefore its unreliability when applied to noisy functions.

4.2 A variant of the Rosenbrock problem

Following the previous experiments conducted based on the quadratic function, we go on to examine the performance on a variant of the famous Rosenbrock function, given by

$$R(x, y) := (a - x)^2 + b(y - x^2)^2$$

with parameters of $a = 1, b = 100$. The Rosenbrock function has a unique optimum at $(x^*, y^*) = (a, a^2)$, i.e., at $(1, 1)$ for our choice of parameters. We modify the problem by adding a penalty of $\lambda\|(x, y) - (x^*, y^*)\|_1$ with a value of $\lambda = 1 \times 10^{-1}$, yielding a problem of type (P) having the same global optimum as R. We show an example of the difference in performance between the classical and stabilized algorithms in Fig. 2. The figure shows the trajectories generated by Algorithm 1 with and without stabilization starting at $(x^0, y^0) = (-1.5, 0)$, injecting noise according to (23) for different values of ε_F and a fixed value of $\varepsilon_{F'} = 1 \times 10^{-5}$, performing quadratic steps according to the true Hessian of R with an iteration limit of 50.

Examining the trajectories of the classical algorithm, shown in Fig. 2a, we find that for different values of ε_F, the trajectories are initially almost identical, until the algorithm stalls at points with a distances to the optimum increasing with ε_F. Conversely, the trajectories of the stabilized algorithm, shown in Fig. 2b, vary significantly for different noise levels. However, the stabilization yields trajectories leading significantly closer to the optimum than those of the classical algorithm even for larger noise levels. This is confirmed by the statistics shown in Fig. 3, displaying the distribution of the distance to the optimum for various noise levels for 100 different random seeds, demonstrating that the stabilized algorithm consistently outperforms the classical one, in particular for larger noise levels.

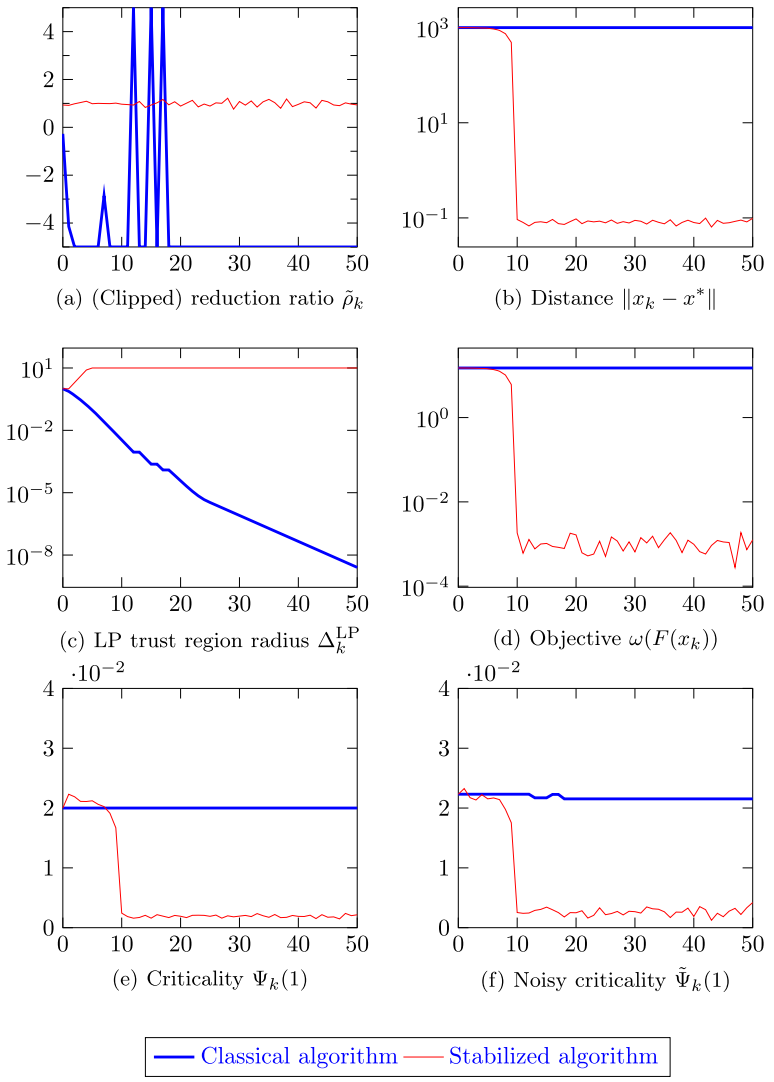
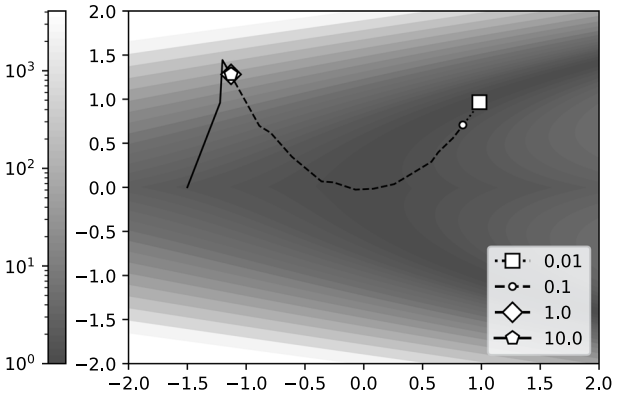


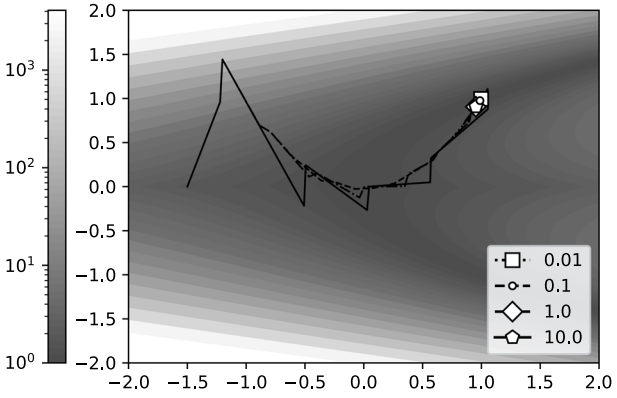
Fig. 1 Performance on an ℓ^1-penalized quadratic problem over 50 iterations of Algorithm 1 with noise levels of $\varepsilon_F = 1 \times 10^{-1}$ and $\varepsilon_{F'} = 1 \times 10^{-5}$. The x-axes always show the iteration count of Algorithm 1, the y-axes show the quantities indicated in the captions below the respective subplots

4.3 Image reconstruction

Although this is not the focus of this article, we also provide a computational example that has a meaningful problem size. Specifically, we consider an artificial task of reconstructing an image under noisy observations. That is we seek to recover a matrix $Y \in \mathbb{R}^{M \times N}$ with values normalized to be in [0, 1]. In our setting, Y is only available in the form of noisy observations. Specifically, for an input $X \in \mathbb{R}^{M \times N}$, the fidelity



(a) Trajectories of Algorithm 1 with $\vartheta = 0$ for different noise levels.



(b) Trajectories of Algorithm 1 with ϑ according to (11) for different noise levels.

Fig. 2 Trajectories for the modified Rosenbrock problem plotted over the *shifted* criticality $1 + \phi(x) - \min_{\|d\|_{\mathrm{LP}} \leq 1} \omega(F(x) + F'(x)d)$, which allows to show it on a logarithmic scale. The markers show the position of the final iterate

of X, given by the squared Frobenius norm of $X - Y$, $\frac{1}{2}\|X - Y\|_F^2$, as well as its derivative with respect to X cannot be evaluated. Instead, we have access to the map $X \mapsto \frac{1}{2}\|X - \tilde{Y}\|_F^2$, where $\tilde{Y}$ is a noisy version of Y, redrawn for each guess X. We obtain the term $\tilde{Y}$ by sampling from a componentwise uniform distribution

$$\delta_Y(X) = Y_{F'} \in \mathbb{R}^{M \times N}, \; Y_{F'} \sim U_{M \times N}(-\varepsilon_{\mathrm{img}}, \varepsilon_{\mathrm{img}}),$$

and setting $\tilde{Y}$ to $Y + \delta_Y(X)$ clipped back to have coefficients in $[0, 1]$. The amount of noise injected to the image is in turn governed by the parameter $\varepsilon_{\mathrm{img}} \geq 0$. This noise model translates into noise injected into the evaluations of F and F', which can be

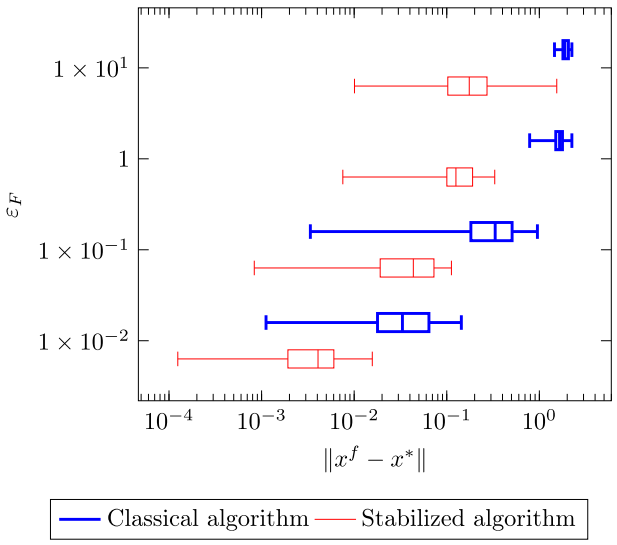


Fig. 3 Distribution of the distance between the final iterate x^f and noiseless optimum x^* for different values of the noise level ε_F

estimated in terms of ε_{img}, M, and N (see Appendix B) while not conforming to the noise model (23). We also impose an anisotropic total variation (TV) regularization penalty, defined as

$$\text{TV}(X) := \sum_{i=1}^{M-1} \sum_{j=1}^{N} |X_{i+1,j} - X_{i,j}| + \sum_{i=1}^{M} \sum_{j=1}^{N-1} |X_{i,j+1} - X_{i,j}|,$$

to our objective, turning the problem non-smooth, balancing off fidelity and regularity by a parameter $\lambda > 0$. The regularization term can be expressed as $\|A_{M,N} X\|_1$ with a suitable matrix $A_{M,N}$. Consequently, we can formulate the reconstruction problem as problem of type (P), consisting of a smooth term (the fidelity), and an ℓ^1-penalized linear function. Naturally, the regularization does not suffer from any noise.

Based on a regularization parameter of $\lambda = 5 \times 10^{-3}$ we reconstruct the image shown in Fig. 4a. To avoid having to solve large quadratic problems, we do not compute quadratic steps and opt to instead increase the number of iterations to 100 starting from $X_0 = 0$. As a baseline, Fig. 4b shows the image when we apply Algorithm 1 to the original image (i.e., setting $\varepsilon_{\text{img}} = 0$). The restored image closely resembles the original one.

We proceed to study the effect of the value of ϑ on the quality of the reconstructed image. In principle, it must hold that $\vartheta \geq \vartheta_\varepsilon^*$ in order for the criticality to provably converge. It is however unclear whether setting ϑ to ϑ_ε^* yields the best results in practice. The large value of $\delta_{\max}$ seen in Sect. 4.1 seems to suggest that (11) is rather pessimistic. We therefore examine the performance of Algorithm 1 for values of ϑ not necessarily satisfying the inequality.

(a) Original image (b) Noiseless reconstruction

Fig. 4 Sample image for the image reconstruction

To gauge performance, we record both the original and noisy objective after the iterations. The results, shown in Fig. 5, demonstrate the effect of ϑ: For small stabilization values, Algorithm 1 stalls early on, as was the case in our previous experiments. As we increase ϑ, there appears to be an optimal choice or small region, where both the noisy and the noiseless evaluation of the final objective are minimized. This effect is more pronounced for higher values of ε_{img}, where the noiseless objective for $\vartheta = 0$ is about 4 times as large as that of the optimal choice of ϑ. It is interesting to see that this sweet spot also shows in the noisy objective, suggesting that noisy observations may be sufficient to find it. Lastly, as we increase ϑ beyond the sweet spot, the final objective increases sharply. This is likely due to the case that the algorithm simply accepts too many steps, even when they are in fact disadvantageous in terms of progressing towards an optimum. Ultimately, for a sufficiently large value of ϑ, all steps are accepted, which, as the final objective suggests, leads to poor solutions. The values of ϑ^*_ε are given by 4×10^4, 2×10^5, and 4×10^5 for the respective noise levels, significantly exceeding the optimal values and beyond the point, where all steps are accepted.

We also find that the objectives are consistent with the visual appearance of the reconstructed images, shown in Fig. 6: While setting ϑ to zero yields satisfactory results, even though a grainy appearance remains for larger noise levels, a disproportionately large value of ϑ produces a distorted result with visible artifacts. For our best guess of ϑ, the restored images do not suffer from artifacts and closely resemble the original one even for larger noise levels.

4.4 Constrained optimization

As a final example and in order to demonstrate the possible use of Algorithm 1 as a subproblem solver in constrained optimization algorithms, we study a constrained optimization problem of the type (NLP). Specifically, we examine the behavior of Algorithm 1 when applied to the `HS71` benchmark problem of the `CUTest` [32]

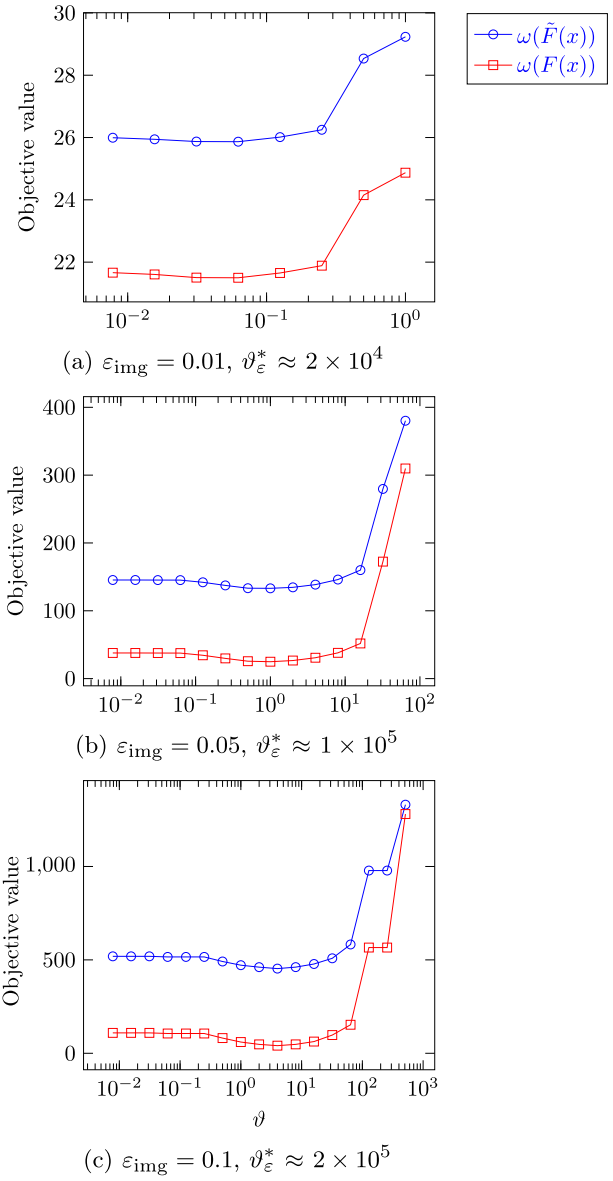


Fig. 5 Noiseless (red) and noisy evaluations (blue) of the objective values achieved by the final iterate of Algorithm 1 on an image reconstruction problem with different noise levels and stabilization parameters (Color figure online)

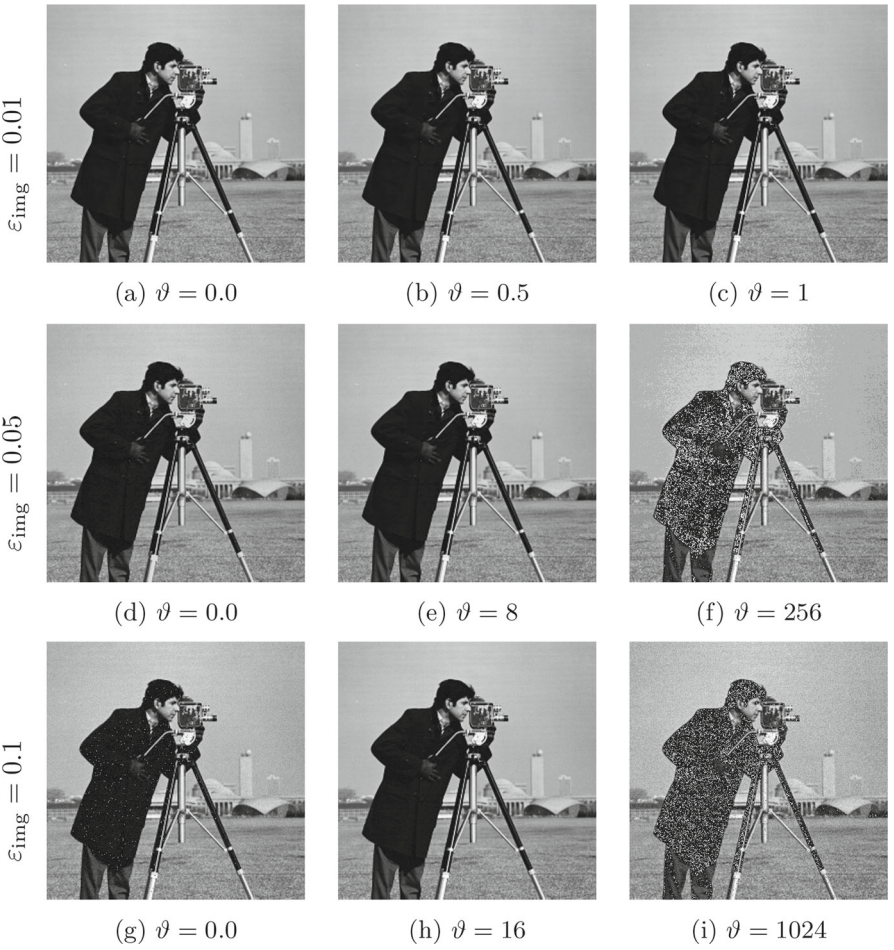


Fig. 6 Reconstructed images for different noise levels and stabilizations (left: no stabilization, center: best stabilization, right: large stabilization so that all iterates are accepted)

suite. The problem is given as

$$\begin{aligned} \min_{x} \quad & x_1 x_4 (x_1 + x_2 + x_3) + x_3 \\ \text{s.t.} \quad & x_1^2 + x_2^2 + x_3^2 + x_4^2 = 40 \\ & x_1 x_2 x_3 x_4 \geq 25 \\ & 1 \leq x \leq 5, \end{aligned} \tag{HS71}$$

leading to suitable functions g and h according to (NLP). The problem features of four bounded optimization variables, two nonlinear constraints, and a nonlinear objective with an optimum at the point $x^* \approx (1.0, 4.74, 3.82, 1.38)$ that satisfies MFCQ and in turn the conditions for the convergence of an exacty penalty method. As mentioned in

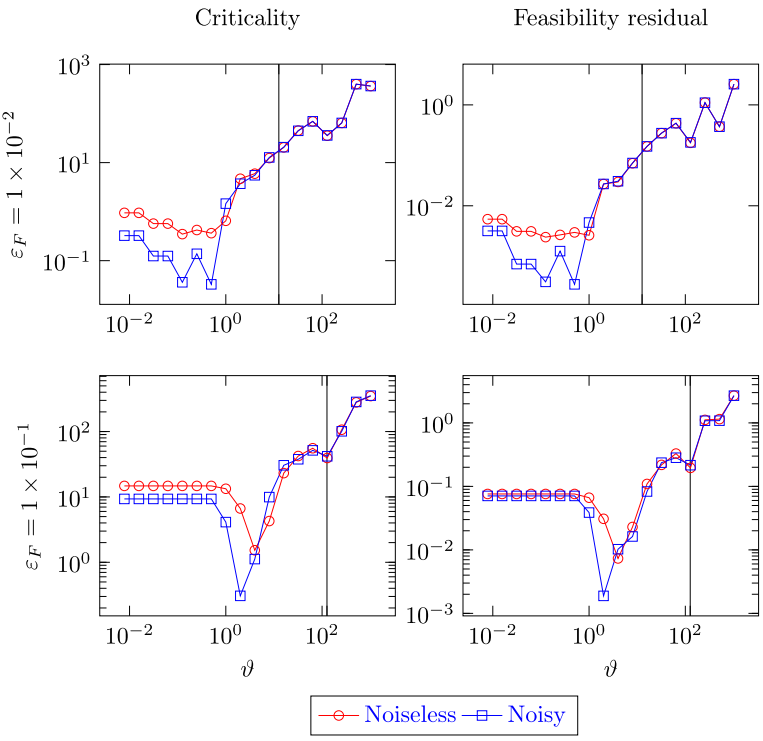


Fig. 7 Noiseless (red) and noisy (blue) performance of the criticality and feasibility residual for different values of ϑ as well as the value of ϑ_ε^* (vertical line) (Color figure online)

the introduction, we solve problems of type (NLP) by using the penalty function (1) with a suitable penalization of $\nu > 0$, knowing that convergence is guaranteed for a sufficiently large ν under mild assumptions, i.e., MFCQ. Increasing ν beyond its required value may slow down practical performance, but convergence is maintained. Consequently, ν is often set to a small initial value and increased when necessary (see for example [2]).

If the functions in (NLP) are affected by noise, the choice of ν is not as straightforward: The required stabilization (11) is dependent on M_0^ε and therefore L^ω, which increases with ν. Similarly, the value of $\delta_{\max}$ increases with L^ω and therefore with ν, so a large penalization has the adverse effect of increasing the size of the critical region $C(\delta_{\max})$, making a suitable choice of the parameter an interesting problem in and of itself. What is more, if the constraint functions g and h suffer from noise, we cannot assume the iterates x_k to tend towards feasibility in the underlying noiseless problem regardless of the value of ν.

Therefore, for our investigation, we consider a fixed value of ν, which is suitable to solve the noiseless variant of (HS71), in our case $\nu = 100$. We once again inject noise according to (23) with different values of ε_F and a fixed $\varepsilon_{F'} = 0$. Specifically, we run the algorithm with the choice $\varepsilon_F = 10^{-2}$ and $\varepsilon_F = 10^{-1}$. Since a reasonable choice of the quadratic model would likely require some dual estimation, we once again opt to skip quadratic steps and instead set the iteration limit to 100. After the algorithm

has terminated, we record the criticality $\Psi_k(1)$, the feasibility residual,

$$\max(\|g(x_k)^+\|_\infty, \|h(x^k)\|_\infty),$$

as well as their noisy counterparts for different values of ϑ (see Fig. 7).

As was the case for the image reconstruction problem, we observe pronounced minima of the quality metrics criticality and feasibility with respect to the choice of ϑ. For a given choice of ε_F, the obtained minima for both quality metrics, criticality and feasibility residual, are in close vicinity to each other. The position of these minima is also fairly consistent across the noisy and noiseless measurements of both the criticality and the feasibility residuum. Unfortunately, setting $\vartheta = \vartheta_\varepsilon^*$ does not yield optimal results, even though it appears as if ϑ_ε^* is closer to being optimal compared to the image reconstruction problem. Once again, for an informed choice of ϑ, the stabilized algorithm significantly outperforms the classical one, leading to about an order of magnitude of reduction in terms of both criticality and feasibility. The precise choice of the parameter and a systematic means to determine it do, however, remain elusive.

5 Conclusion

We have presented a noise-tolerant adaptation of a well-established trust-region method for a non-smooth optimization problem with a structured and convex non-smoothness described by a polyhedral function, which is therefore suitable to handling by linear programming techniques. The adaptation only requires knowledge of a Lipschitz constant and bounds on the noise in the objective function and its derivative. The analysis of the asymptotics of the successive linear programming algorithm can be carried out analogously to [3], where the noiseless case is handled. As we expect from the results in [1], we do not get convergence to a first-order stationary point but a critical region instead.

In a noiseless setting, both the behavior of the algorithm and its convergence properties are consistent and similar to previous analyses. The computational results show that an informed choice of the stabilization parameter ϑ may improve the quality of the obtained results significantly so that we believe it makes sense to dedicate research to improved bounds and efficient practical determination strategies.

Further analysis is also needed in order to be able to use and interpret the method as a subproblem solver for constrained optimization with noisy constraint and objective evaluations. In particular, it is necessary to study the asymptotics of the feasibility residual, identify means to control it, and classify it with respect to existing concepts from the field of uncertainty quantification like (distributional) chance constraints or expectation constraints.

Acknowledgements We thank an anonymous referee for tireless and thorough proofreading, which particularly helped to correct the proof of Lemma 3.12.

Funding Open Access funding enabled and organized by Projekt DEAL.

Data availability The authors confirm that all data generated or analysed during this study are included in this published article.

Declarations

Conflict of interest The authors declare no competing interests.

A Proofs

In the following we give the proofs of some of the result used in Sect. 3. These proofs closely follow those in [3]. We provide them here to make this article more self-contained.

Proof of Lemma 3.8 Let $d_1 \in \mathbb{R}^n$ be a minimizer of $\tilde{\Psi}_k(1)$ for $\Delta = 1$, i.e., $\tilde{\Psi}_k(1) = \tilde{\phi}(x_k) - \tilde{\ell}_k(d_1)$. If $\Delta \geq 1$, it follows that

$$\tilde{\Psi}_k(\Delta) \geq \tilde{\phi}(x_k) - \tilde{\ell}_k(d_1) = \tilde{\Psi}_k(1).$$

It remains to prove the case $\Delta < 1$. From $\|d_1\|_{\mathrm{LP}} \leq 1$ it follows that $\|\Delta d_1\|_{\mathrm{LP}} \leq \Delta$, i.e., Δd_1 is a feasible solution with respect to Δ. Therefore, it holds that

$$\tilde{\Psi}_k(\Delta) = \tilde{\phi}(x_k) - \min_{\|d\|_{\mathrm{LP}} \leq \Delta} \tilde{\ell}_k(d) \geq \Delta \left[\tilde{\phi}(x_k) - \tilde{\ell}_k(d_1)\right] = \Delta \tilde{\Psi}_k(1),$$

where the inequality is due to the feasibility of Δd_1 and Lemma 3.7. □

Proof of Lemma 3.9 Let d_1 be a minimizer for $\Delta = 1$. Assume (towards a contradiction) that $\|d_\Delta\|_{\mathrm{LP}} < \min(\Delta, \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell}) \leq \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell}$. It follows from Lemma 3.6 that

$$\tilde{\ell}_k(d_\Delta) \geq \tilde{\ell}_k(0) - L_\varepsilon^\ell \|d_\Delta\|_{\mathrm{LP}} > \tilde{\ell}_k(0) - \tilde{\Psi}_k(1) = \tilde{\ell}_k(d_1). \tag{24}$$

If $\Delta \geq 1$, (24) cannot hold because d_1 is feasible with respect to Δ and therefore cannot yield a better objective with respect to $\tilde{\ell}_k$ than the minimizer d_Δ. So it must hold that $\|d_\Delta\|_{\mathrm{LP}} \geq \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell}$ in this case $\Delta \geq 1$. If, on the other hand, $\Delta < 1$, then d_1 may not be feasible. However, since $\tilde{\ell}_k$ is convex, it holds for all $\lambda \in (0, 1]$ that

$$\tilde{\ell}_k(\lambda d_1 + (1-\lambda) d_\Delta) \leq \lambda \tilde{\ell}_k(d_1) + (1-\lambda)\tilde{\ell}_k(d_\Delta) \underset{(24)}{<} \tilde{\ell}_k(d_\Delta).$$

Therefore, any point on the line segment $(d_\Delta, d_1]$ has a strictly lower value of $\tilde{\ell}_k$ than d_Δ. Therefore, no such point can be feasible with respect to the constraint on $\|\cdot\|_{\text{LP}}$. Consequently, d_Δ must lie on the boundary of the feasible set implying that $\|d_\Delta\|_{\text{LP}} = \Delta$. The result is obtained by combining these bounds. □

Proof of Lemma 3.10 The actual step must satisfy that $\tilde{q}_k(d_k) \leq \tilde{q}_k(d_k^C)$, so the first inequality is a given. Similarly, the last inequality is an application of Lemma 3.8. To show that the remaining inequality holds, recall that the line search for the Cauchy step d_k^C terminates with an α_k such that

$$\tilde{\phi}(x_k) - \tilde{q}_k(d_k^C) \geq \eta \left[\tilde{\phi}(x_k) - \tilde{\ell}_k(d_k^C)\right] \geq \eta\alpha_k \left[\tilde{\phi}(x_k) - \tilde{\ell}_k(d_k^{\text{LP}})\right] = \eta\alpha_k \tilde{\Psi}_k(\Delta_k^{\text{LP}})$$

by means of a standard argument on Armijo line search because $0 < \eta < 1$, see also [4]. Since $d_k^C = \alpha_k d_k^{\text{LP}}$ and d_k^{LP}, the last inequality follows from Lemma 3.7, while the last equality is due to d_k^{LP} achieving $\tilde{\Psi}_k(\Delta_k^{\text{LP}})$. □

Proof of Lemma 3.11 The first inequality is due to the fact that the Cauchy step is the LP step scaled by α_k, where the LP norm of the LP step is bounded by Δ_k^{LP}. Consider two cases for the second inequality:

1. The decrease condition is immediately satisfied for the initial step size of $\alpha_k = \min(1, \Delta_k/\|d_k^{\text{LP}}\|)$. Consequently it follows that

$$\|d_k^C\|_{\text{LP}} = \|\alpha_k d_k^{\text{LP}}\|_{\text{LP}} = \min\left(1, \Delta_k/\|d_k^{\text{LP}}\|\right) \|d_k^{\text{LP}}\|_{\text{LP}}$$

   We consider two cases:

   (a) $\Delta_k/\|d_k^{\text{LP}}\| \geq 1$, which is to say that $\|d_k^C\|_{\text{LP}} = \|d_k^{\text{LP}}\|_{\text{LP}}$. It follows from Lemma 3.9 that

$$\|d_k^{\text{LP}}\|_{\text{LP}} \geq \min\left(\Delta_k^{\text{LP}}, \frac{\tilde{\Psi}_k(1)}{L_\varepsilon^\ell}\right),$$

   which implies the claimed bound.

   (b) Otherwise we know that $\|d_k^C\|_{\text{LP}} = \|d_k^{\text{LP}}\|_{\text{LP}}\Delta_k/\|d_k^{\text{LP}}\|$. We can use (6) to obtain that $\|d_k^{\text{LP}}\| \leq \gamma\|d_k^{\text{LP}}\|_{\text{LP}}$, inferring that

$$\|d_k^C\|_{\text{LP}} = \|d_k^{\text{LP}}\|_{\text{LP}}\Delta_k/\|d_k^{\text{LP}}\| \geq \|d_k^{\text{LP}}\|_{\text{LP}}\Delta_k/(\gamma\|d_k^{\text{LP}}\|_{\text{LP}}) = \Delta_k/\gamma,$$

   which implies the claimed bound.

2. The decrease condition is only satisfied at a later iteration of the line search. Recall that the line search computes step sizes by multiplying a base length with powers of an input parameter $\tau \in (0, 1)$. We can therefore deduce that the sufficient decrease condition was not satisfied for α_k/τ in the previous iteration, i.e.,

$$\tilde{\phi}(x_k) - \tilde{q}_k(\alpha_k/\tau d_k^{\text{LP}}) < \eta\left[\tilde{\phi}(x_k) - \tilde{\ell}_k(\alpha_k/\tau d_k^{\text{LP}})\right]$$

Since the only difference between the linearized and quadratic model is the quadratic term, we have that

$$\tfrac{1}{2}(\alpha_k/\tau)^2\langle d_k^{\mathrm{LP}}, B_k d_k^{\mathrm{LP}}\rangle \geq (1-\eta)\left[\tilde{\phi}(x_k) - \tilde{\ell}_k(\alpha d_k^{\mathrm{LP}}/\tau)\right].$$

The left hand side can be bounded above by using Assumption 4 and relation (6) to yield

$$\begin{aligned}\tfrac{1}{2}(\alpha_k/\tau)^2\langle d_k^{\mathrm{LP}}, B_k d_k^{\mathrm{LP}}\rangle &\leq \tfrac{1}{2}(\alpha_k/\tau)^2\beta\gamma^2\|d_k^{\mathrm{LP}}\|_{\mathrm{LP}}^2\\ &\leq \tfrac{1}{2}(\alpha_k/\tau)^2\beta\gamma^2\|d_k^{\mathrm{LP}}\|_{\mathrm{LP}}\Delta_k^{\mathrm{LP}}.\end{aligned}$$

Similarly, for the right hand side we can use Lemmas 3.7 and 3.8 to obtain

$$\begin{aligned}\left[\tilde{\phi}(x_k) - \tilde{\ell}_k(\alpha/\tau d_k^{\mathrm{LP}})\right] &\geq \alpha_k/\tau\left[\tilde{\phi}(x_k) - \tilde{\ell}_k(d_k^{\mathrm{LP}})\right]\\ &\geq \alpha_k/\tau\,\min(1, \Delta_k^{\mathrm{LP}})\tilde{\Psi}_k(1),\end{aligned}$$

Putting these inequalities together yields the bound

$$\|d_k^C\|_{\mathrm{LP}} = \alpha_k\|d_k^{\mathrm{LP}}\|_{\mathrm{LP}} \geq \frac{2(1-\eta)\tau}{\beta\gamma^2}\min\left(1, \frac{1}{\Delta_k^{\mathrm{LP}}}\right)\tilde{\Psi}_k(1)$$

required to complete the proof.

□

Proof of Lemma 3.13 Let k_0 be the index of the last accepted step. Then, $x_{k+1} = x_k =: x^*$ for all $k > k_0$. Consequently, after finishing the k_0-the iteration, $\tilde{\Psi}_k(1)$ stays at a constant value of $\delta \geq 0$. What is more, due to the rejection of the steps following iteration k_0 it holds for all $k > k_0$ that $\Delta_{k+1} \leq \kappa_u\Delta_k < \Delta_k$ (since $\kappa_u < 1$). Therefore, Δ_k tends to zero. Recall from Lemma 3.12 that if $\delta > 0$, then Δ_k is bounded away from zero. Therefore, since Δ_k tends to zero, it must hold that $\delta = 0$. □

B Estimations

Lipschitz constant of the ℓ^1-penalty function

In the following, we give an estimation for the Lipschitz constant L^ω of the penalty function $\omega : \mathbb{R}\times\mathbb{R}^m \to \mathbb{R}$, $\omega(x, y) = x + \nu\|y\|_1$, based on the constant $\nu > 0$ and the dimension $m \in \mathbb{N}$. Since we use this function in all of the examples in Sect. 4, and since the value of ϑ_ε^* depends on the value of L^ω, we make its derivation explicit. To obtain an optimal value of L^ω, we solve the optimization problem

$$\begin{aligned}\max_{x,y,x',y'}\quad & |\omega(x', y') - \omega(x, y)|\\ \text{s.t.}\quad & (x - x')^2 + \|y - y'\|^2 \leq 1,\end{aligned}$$

i.e., we maximize the difference in values of ω while controlling the distance between the points (x, y) and (x', y'). Observe that

$$|\omega(x', y') - \omega(x, y)| = |(x - x') + \nu\|y - y'\|_1|,$$

from which it follows that both the objective and the constraint value only depend on $x - x'$ and $y - y'$. We can therefore simplify the problem by setting $y' = 0$ and $y' = 0$:

$$\begin{aligned} \max_{x,y} \quad & |x + \nu\|y\|_1| \\ \text{s.t.} \quad & x^2 + \|y\|^2 \leq 1. \end{aligned}$$

We can simplify the problem further by realizing that we can assume both x and y to be non-negative, eliminating the absolute value in the objective. The largest ratio of $\nu\|y\|_1$ over $\|y\|^2$ is achieved by setting all entries of y to the same value $y_0 \in \mathbb{R}$, yielding the problem

$$\begin{aligned} \max_{x,y_0} \quad & x + \nu m y_0 \\ \text{s.t.} \quad & x^2 + m y_0^2 \leq 1 \\ & x, y_0 \geq 0. \end{aligned}$$

By setting $z := \sqrt{m} y_0$, we obtain the problem

$$\begin{aligned} \max_{x,z} \quad & x + \nu\sqrt{m} z \\ \text{s.t.} \quad & x^2 + z^2 \leq 1 \\ & x, z \geq 0. \end{aligned}$$

The optimal solution of this problem is attained at

$$\begin{pmatrix} x^* \\ z^* \end{pmatrix} = \frac{1}{\sqrt{1 + \nu^2 m}} \begin{pmatrix} 1 \\ \nu\sqrt{m} \end{pmatrix},$$

yielding the objective $\sqrt{1 + \nu^2 m} =: L^\omega$.

Image Reconstruction

In the following, we provide estimations regarding the noise levels associated with the image fidelity map introduced in Sect. 4.3. Recall that the squared Frobenius norm of a matrix $A \in \mathbb{R}^{M \times N}$ is given by $\|A\|_F^2 := \sum_{i=1}^{M} \sum_{j=1}^{N} a_{ij}^2$. Thus, if $|a_{ij}| \leq \varepsilon$ for a given $\varepsilon > 0$, it follows that

$$\|A\|_F^2 \leq \sum_{i=1}^{M} \sum_{j=1}^{N} \varepsilon^2 = \varepsilon^2 MN,$$

and therefore, that $\|A\|_F \leq \varepsilon\sqrt{MN}$. The noisy fidelity function $\tilde{F}(X)$ satisfies the identities

$$\begin{aligned}\tilde{F}(X) &= \frac{1}{2}\|X - \tilde{Y}\|_F^2 = \frac{1}{2}\|X - (Y + \delta_Y(X))\|_F^2 \\ &= \frac{1}{2}\|X - Y\|_F^2 + \langle X - Y, \delta_Y(X)\rangle_F + \frac{1}{2}\|\delta_Y(X)\|_F^2,\end{aligned}$$

where $\langle\cdot,\cdot\rangle_F$ denotes the inner product that induces the Frobenius norm. Since the entries of X and Y are in $[0, 1]$, and therefore have absolute values bounded by 1, it follows that $\|X - Y\|_F \leq \sqrt{MN}$. The choice of distribution implies that the values in $\delta_Y(X)$ are bounded by $\pm\varepsilon_{\text{img}}$, and therefore that $\|\delta_Y(X)\|_F \leq \varepsilon_{\text{img}}\sqrt{MN}$, from which it follows that

$$|\tilde{F}(X) - F(X)| \leq \|X - Y\|_F\|\delta_Y(X)\|_F + \frac{1}{2}\|\delta_Y(X)\|_F^2 \leq \left(\varepsilon_{\text{img}} + \frac{1}{2}\varepsilon_{\text{img}}^2\right)MN,$$

by means of the Cauchy–Schwarz inequality for $\langle\cdot,\cdot\rangle_F$. Furthermore, it holds for any matrix $A \in \mathbb{R}^{m\times n}$ that $\|A\| \leq \|A\|_F$. Any estimation with respect to the Frobenius norm therefore also produces an upper bound for our standard norm $\|\cdot\|$. Specifically, the noise level ε_{img} yields a corresponding value for ε_F. Similarly, it holds that $G(X) = F'(X) - \delta_Y(X)$, and therefore $\|G(X) - F'(X)\|_F \leq \|\delta_Y(X)\|_F \leq \varepsilon_{\text{img}}\sqrt{MN}$, corresponding to a value for ε'_F.

References

1. Sun, S., Nocedal, J.: A trust region method for noisy unconstrained optimization. Math. Program. **66**, 1–28 (2023). https://doi.org/10.1007/s10107-023-01941-9
2. Byrd, R.H., Gould, N.I., Nocedal, J., Waltz, R.A.: An algorithm for nonlinear optimization using linear programming and equality constrained subproblems. Math. Program. **100**(1), 27–48 (2003). https://doi.org/10.1007/s10107-003-0485-4
3. Byrd, R.H., Gould, N.I., Nocedal, J., Waltz, R.A.: On the convergence of successive linear-quadratic programming algorithms. SIAM J. Optim. **16**(2), 471–489 (2005). https://doi.org/10.1137/S1052623403426532
4. Wright, S., Nocedal, J., et al.: Numerical Optimization, 2nd edn. Springer, Berlin (2006). https://doi.org/10.1007/978-0-387-40065-5
5. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. **58**(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
6. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. A J. Issued Courant Inst. Math. Sci. **59**(8), 1207–1223 (2006). https://doi.org/10.1002/cpa.20124
7. Fukushima, K.: Cognitron: a self-organizing multilayered neural network. Biol. Cybernet. **20**(3), 121–136 (1975). https://doi.org/10.1007/BF00342633
8. Aspelmeier, T., Charitha, C., Luke, D.R.: Local linear convergence of the admm/douglas-rachford algorithms without strong convexity and application to statistical imaging. SIAM J. Imaging Sci. **9**(2), 842–868 (2016). https://doi.org/10.1137/15M103580X
9. Cartis, C., Gould, N.I., Toint, P.L.: On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming. SIAM J. Optim. **21**(4), 1721–1739 (2011). https://doi.org/10.1137/11082381X

10. Apkarian, P., Noll, D., Ravanbod, L.: Nonsmooth bundle trust-region algorithm with applications to robust stability. Set-Valued Var. Anal. **24**(1), 115–148 (2016). https://doi.org/10.1007/s11228-015-0352-5
11. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J., et al: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends® Mach. Learn. **3**(1), 1–122 (2011). https://doi.org/10.1561/2200000016
12. Nesterov, Y.: Gradient methods for minimizing composite functions. Math. Program. **140**(1), 125–161 (2013). https://doi.org/10.1007/s10107-012-0629-5
13. Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. **7**(4), 1307–1330 (1986). https://doi.org/10.1137/0907087
14. Han, S.-P., Mangasarian, O.L.: Exact penalty functions in nonlinear programming. Math. Program. **17**(1), 251–269 (1979). https://doi.org/10.1007/BF01588250
15. Shi, H.-J.M., Xie, Y., Byrd, R., Nocedal, J.: A noise-tolerant quasi-newton algorithm for unconstrained optimization. SIAM J. Optim. **32**(1), 29–55 (2022). https://doi.org/10.1137/20M1373190
16. Irwin, B., Haber, E.: Secant penalized bfgs: a noise robust quasi-newton method via penalizing the secant condition. Comput. Optim. Appl. **84**(3), 651–702 (2023). https://doi.org/10.1007/s10589-022-00448-x
17. Cao, L., Berahas, A.S., Scheinberg, K.: First-and second-order high probability complexity bounds for trust-region methods with noisy oracles. Math. Program. **66**, 1–52 (2023). https://doi.org/10.1007/s10107-023-01999-5
18. Gal, R., Haber, E., Irwin, B., Saleh, B., Ziv, A.: How to catch a lion in the desert: on the solution of the coverage directed generation (cdg) problem. Optim. Eng. **22**(1), 217–245 (2021). https://doi.org/10.1007/s11081-020-09507-w
19. Curtis, F.E., Scheinberg, K., Shi, R.: A stochastic trust region algorithm based on careful step normalization. Informs J. Optim. **1**(3), 200–220 (2019). https://doi.org/10.1287/ijoo.2018.0010
20. Berahas, A.S., Byrd, R.H., Nocedal, J.: Derivative-free optimization of noisy functions via quasi-newton methods. SIAM J. Optim. **29**(2), 965–993 (2019). https://doi.org/10.1137/18M1177718
21. Xie, Y., Byrd, R.H., Nocedal, J.: Analysis of the BFGS method with errors. SIAM J. Optim. **30**(1), 182–209 (2020). https://doi.org/10.1137/19M1240794
22. Gurobi Optimization, LLC: Gurobi Optimizer Reference Manual (2022). https://www.gurobi.com
23. Huangfu, Q., Hall, J.J.: Parallelizing the dual revised simplex method. Math. Program. Comput. **10**(1), 119–142 (2018). https://doi.org/10.1007/s12532-017-0130-5
24. Byrd, R.H., Nocedal, J., Waltz, R.A.: KNITRO: an integrated package for nonlinear optimization. In: Large-Sscale Nonlinear Optimization, pp. 35–59. Springer, Berlin (2006). https://doi.org/10.1007/0-387-30065-1_4
25. Fletcher, R.: Practical Methods of Optimization: Vol. 2: Constrained Optimization (1981). https://doi.org/10.1002/9781118723203
26. Yuan, Y.-x: Conditions for convergence of trust region algorithms for nonsmooth optimization. Math. Program. **31**(2), 220–228 (1985). https://doi.org/10.1007/BF02591750
27. Rockafellar, R.T., Wets, R.J.-B.: Variational Analysis, vol. 317. Springer, Berlin (2009). https://doi.org/10.1007/978-3-642-02431-3
28. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E.: Array programming with NumPy. Nature **585**(7825), 357–362 (2020). https://doi.org/10.1038/s41586-020-2649-2
29. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ, Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P.: SciPy 1.0 Contributors: SciPy 1.0: fundamental algorithms for scientific computing in python. Nat. Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
30. Wächter, A., Biegler, L.T.: On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Math. Program. **106**(1), 25–57 (2006). https://doi.org/10.1007/s10107-004-0559-y

31. Marsaglia, G.: Choosing a point from the surface of a sphere. Ann. Math. Stat. **43**(2), 645–646 (1972). https://doi.org/10.1214/aoms/1177692644
32. Gould, N.I., Orban, D., Toint, P.L.: Cutest: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Comput. Optim. Appl. **60**(3), 545–557 (2015). https://doi.org/10.1007/s10589-014-9687-3