An objective-function-free algorithm for general smooth constrained optimization

Stefania Bellavia[*], Serge Gratton[†], Benedetta Morini[‡], Philippe L. Toint[§]

February 12, 2026

Abstract

A new algorithm for smooth constrained optimization is proposed that never computes the value of the problem's objective function and that handles both equality and inequality constraints. The algorithm uses an adaptive switching strategy between a normal step aiming at reducing constraint's infeasibility and a tangential step improving dual optimality, the latter being inspired by the AdaGrad-norm method. Its worst-case iteration complexity is analyzed, showing that the norm of the gradients generated converges to zero like $\mathcal{O}(1/\sqrt{k+1})$ for problems with full-rank Jacobians. Numerical experiments show that the algorithm's performance is remarkably insensitive to noise in the objective function's gradient.

Keywords: Objective-function-free optimization (OFFO), general constraints, nonconvex problems, reliability in the presence of noise, complexity.

1 Introduction

The design and analysis of deterministic algorithms for solving constrained continuous optimization problems have a long history and have produced well-assessed techniques such as penalty methods, SQP methods, interior-point methods or filter methods (see [17, 6, 10, 24] for example). These techniques all require the computation of both the function and derivative evaluation of the objective and the constraints. By contrast, this paper addresses the solution of the problem

$$\min_{x\in\mathcal{F}} f(x) \quad \text{where} \quad \mathcal{F}=\{x\in\mathbb{R}^n,\ \ c(x)=0\ \ x\geq 0\}, \tag{1}$$

using a first-oder objective function-free (OFFO) method. Here f is a smooth (possibly nonconvex) function from an open set containing the feasible region $\mathcal{F}\subseteq\mathbb{R}^n$ into $\mathbb{R}$, $c(x):\mathbb{R}^n\to\mathbb{R}^m$ with $m\leq n$ and equalities and inequalities are meant componentwise. Problem (1) includes general constrained optimization since all problems in this class can be cast into this form by using slack variables. We assume that, given x, we can compute both the gradient $g(x)=\nabla_x f(x)$ of f and the value of the constraints $c(x)$ as well as their Jacobian $J(x)=\nabla_x c(x)\in\mathbb{R}^{m\times n}$, which we will assume (for the purpose of our analysis) is full rank for any $x\geq 0$.

[*]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italia. Member of the INdAM Research Group GNCS. Email: stefania.bellavia@unifi.it. Work partially supported by Progetti di Ricerca INDAM-GNCS.

[†]Université de Toulouse, INP, IRIT, Toulouse, France. Work partially supported by 3IA Artificial and Natural Intelligence Toulouse Institute (ANITI), French "Investing for the Future - PIA3" program under the Grant agreement ANR-19-PI3A-0004. Email: serge.gratton@toulouse-inp.fr.

[‡]Dipartimento di Ingegneria Industriale, Università degli Studi di Firenze, Firenze, Italia. Member of the INdAM Research Group GNCS. Email: benedetta.morini@unifi.it. Work partially supported by Progetti di Ricerca INDAM-GNCS.

[§]Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium.Work partially supported by 3IA Artificial and Natural Intelligence Toulouse Institute (ANITI) and DIEF (Florence). Email: philippe.toint@unamur.be.

First-order OFFO procedures do not employ the value of the objective function but rely on gradients. They are known to be suitable for the solution of problems in which the function is approximated or subject to noise, and have exhibited remarkable robustness in the presence of noisy gradients [2, 7, 13, 15, 23].

Our approach is inspired by the "trust-funnel" approach [12, 20], which, as [25], has roots in the much older Himmelblau's "flexible tolerance " method [16]. As in these references, the new method uses an adaptive switching strategy to select a normal or a tangential steps, the first aiming at reducing the violation of the constraints and the second at improving the objective-function value. The stepsize in the latter is reminiscent of the stepsize formula used in Adagrad-norm method [8, 22] for unconstrained optimization. Our method can therefore be seen as a AdaGrad-like method for solving equality and inequality constrained problems. Other OFFO methods for constrained optimization using AdaGrad stepsizes have been proposed by the authors in the papers [1, 15]. In [1] bound constrained optimization problems are considered; stochastic estimators of the gradient are allowed and second-order information are used when available. The paper [15] considers nonlinear equality constrained problems with full-rank Jacobians and proposes a first-order algorithm that adaptively selects steps in the plane tangent to the constraints or steps that reduce infeasibility. The evaluation complexity is analyzed, in both cases yielding a global convergence rate in $O(1/\sqrt{k+1})$, identical in order to that of steepest-descent and Newton's methods for unconstrained problems [5].

Our present proposal builds on these contributions and extends [15] to handle inequality constraints and thus to cover general smooth constrained optimization. To accommodate such constraints, we revisited the procedure from [15] by introducing suitable primal and dual criticality measures and redefining both tangential and normal steps, while avoiding a technical assumption on the first iteration. Three different techniques are provided for the computation of the tangential step. We analyze the worst-case iteration complexity of our procedures and show that the norm of the gradients generated converges to zero like $\mathcal{O}(1/\sqrt{k+1})$. Numerical experiments show that, in line with what happens on simpler problems, the algorithm's performance is remarkably insensitive to noise in the objective function's gradient.

The authors are aware of four other papers on OFFO procedures [2, 7, 9, 21] for constrained problems. The paper [2] presents objective function-free Sequential Quadratic Programming (SQP) algorithms to solve smooth optimization problems with stochastic objective and deterministic nonlinear equality constraints. It employs a stepsize selection scheme based on Lipschitz constants (or adaptively estimated Lipschitz constants) in place of the linesearch. This approach has been extended in [7] to handle deterministic inequality constraints. A convergence analysis in expectation is carried out, but the worst-case complexity has not been analyzed. The method introduced in [9] is designed to solve nonlinear optimization problems with stochastic objectives and deterministic equality constraints. It again employs normal and tangential steps, the latter being computed using a standard trust-region technique; an explicit penalty parameter is used and dynamically updated throughout the process, without requiring the objective function's computation. Global almost-sure convergence is proved. [21] proposes a variant of the SQP approach of [2] for equality-constrained problems with full-rank Jacobian using first-order methods with momentum and analyzes its rate of convergence.

Our paper is organized as follows. The ADIC (ADagrad with Inequality Constraints) class of algorithms is introduced in Section 2 with its algorithmic options. Section 3 analyzes its worst-case complexity. Results obtained from the numerical validation of the algorithms are described in Section4. Section 5 finally summarizes our contributions and discusses perspectives for further research.

Notations: In what follows, $\|\cdot\|$ denotes the Euclidean norm unless otherwise specified, and $\sigma_{\min}[A]$ denotes the smallest singular value of the matrix A.

2 The ADIC class of algorithms

In the new class of iterative methods that we are going to describe, a new iterate is formed using either a tangential step (i.e. a step in the plane tangent to the manifold of equality constraints) or a normal step (mostly orthogonal to that manifold), the choice between the two being based on a comparison of the primal and dual criticality measures. One of the interesting features of this algorithmic framework is that it allows the use of fairly general non-negative bounded dual and primal criticality measures, denoted $\omega_T(x)$ and $\omega_N(x)$ respectively.

In the algorithm's description on the following page, the successive iterates are denoted by x_k and we let $g_k = g(x_k)$, $c_k = c(x_k)$, $J_k = J(x_k)$, $\omega_T(x_k) = \omega_{T,k}$, $\omega_N(x) = \omega_{N,k}$. These criticality measures are computed in Step 1, together with a tangential stepsize whose form is, as we will detail later, directly inspired by the AdaGrad [8, 22] algorithm. Whether the step taken is tangential or normal is decided by comparing their sizes, each of these steps being designed to provide a first-order improvement (of a carefully chosen Lyapunov function) comparable to the relevant criticality measure while being of a size ensuring that first-order effects dominate (as we will prove below).

In our subsequent analysis, we need to distinguish between iterates using tangential or normal steps. We denote by $\{k_\tau\} \subseteq \{k\}$ the index subsequence of iterations such that a tangential step $s_{T,k}$ was computed (implying that (3) holds), while $\{k_\nu\}$ is the index subsequence of iterations where a normal step $s_{N,k}$ was computed. Note that $\{k_\tau\}$ and $\{k_\nu\}$ need not be disjoint, but that $\{k\} = \{k_\tau\} \cup \{k_\nu\}$. By convention, we will define $s_{T,k} = 0$ for $k \not\in \{k_\tau\}$ and $s_{N,k} = 0$ for $k \not\in \{k_\nu\}$.

We will also consider the Lyapounov function (whose value is hopefully decreased as the iterations progress) given by

$$\psi(x,\lambda) \stackrel{\rm def}{=} L(x,\lambda) + \rho\|c(x)\|, \tag{11}$$

where ρ is a fixed constant (to be determined below) and $L(x,\lambda)$ is the standard Lagrangian

$$L(x,\lambda) = f(x) + \lambda^T c(x), \tag{12}$$

for some multiplier $\lambda \in \mathbb{R}^m$. The function $\psi(x,\lambda)$ is sometimes called the "sharp augmented Lagrangian" (see [3, 4, 19] for instance). Of particular interest in our argument is the least-squares Lagrange multiplier $\widehat{\lambda}(x)$ defined by

$$\big(J(x)J(x)^T\big)\,\widehat{\lambda}(x) \;=\; -\,J(x)\,g(x) \tag{13}$$

when the Jacobian $J(x)$ has full rank.

It is important to note to this point that, because all norms are equivalent in $\mathbb{R}^n$, our theoretically convenient choice of expressing (10) and (5) in Euclidean norm is by no means crucial. Should other norms be used, as we will see below, the relevant equivalence constants may be absorbed in θ_T and θ_N. It is also useful to notice (2) implies that

$$\alpha_{T,k} \le \frac{\eta}{\sqrt{\varsigma}} \quad \text{and} \quad \alpha_{T,k}\omega_{T,k} < \eta. \tag{14}$$

Clearly, much else remains to be specified in our algorithmic outline: details of which criticality measures are considered together with which norm and methods to compute the tangential step $s_{T,k}$ itself as well as the normal step $s_{N,k}$ must be clarified. In order to simplify exposition, we focus in our theory on a single technique for computing the normal step $s_{N,k}$, and propose to define it by one (or more) step(s) of a trust-region algorithm applied on the constrained violation $\frac{1}{2}\|c(x)\|^2$ using a linear model. Lemma 3.2 below will show that such a step satisfies our requirements of Step 2 with

$$\omega_{N,k} = \chi_{N,k} = |c_k^T J_k d_{N,k}|. \tag{15}$$

Algorithm 2.1: $\mathsf{ADIC}(x_0)$

Step 0: Initialization: The constants $\beta, \eta > 0$, $\theta_T, \theta_N \geq 1$, $0 < \eta_{\min} \leq \eta_{\max}$ and $\kappa_t, \kappa_n, \varsigma \in (0, \frac{1}{2}]$ are given.
Project x_0 onto the positive orthant.
Set $k = 0$ and $\Gamma_0 = 0$.

Step 1: Evaluations: Evaluate $c_k = c(x_k)$, $J_k = J(x_k)$, $g_k = \nabla f(x_k)$. Then compute the dual measure $\omega_{T,k}$, the primal measure $\omega_{N,k}$ and the stepsize

$$\alpha_{T,k} = \frac{\eta}{\sqrt{\Gamma_k + \omega_{T,k}^2 + \varsigma}}. \tag{2}$$

Step 2: Normal step: Except possibly if

$$\omega_{N,k} \leq \beta\, \alpha_{T,k}\omega_{T,k}, \tag{3}$$

compute $s_{N,k}$ such that

$$x_k + s_{N,k} \geq 0, \tag{4}$$
$$\|s_{N,k}\| \leq \theta_N\, \omega_{N,k}, \tag{5}$$

and there exists a contant $\kappa_n \in (0, \frac{1}{2})$ independent of k such that

$$\tfrac{1}{2}\|c(x_k + s_{N,k})\|^2 \leq \tfrac{1}{2}\|c_k\|^2 - \kappa_n\, \omega_{N,k}^2. \tag{6}$$

Then set $x_k^+ = x_k + s_{N,k}$. If (3) holds and $s_{N,k}$ was not computed, set $x_k^+ = x_k$.

Step 3: Tangential step: If (3) holds, compute a step $s_{T,k}$ such that

$$x_k + s_{T,k} \geq 0 \tag{7}$$
$$J_k s_{T,k} = 0, \tag{8}$$
$$g_k^T s_{T,k} \leq -\kappa_t\, \alpha_{T,k}\, \omega_{T,k}^2, \tag{9}$$
$$\|s_{T,k}\| \leq \theta_T\, \alpha_{T,k}\, \omega_{T,k}, \tag{10}$$

and set $x_{k+1} = x_k^+ + s_{T,k}$ and $\Gamma_{k+1} = \Gamma_k + \omega_{T,k}^2$.
Otherwise (i.e. if (3) fails), set $x_{k+1} = x_k^+$ and $\Gamma_{k+1} = \Gamma_k$.

Step 4: Loop: Increment k by one and go to Step 1.

where $d_{N,k}$ solves the problem

$$\begin{array}{ll} \min_d & c_k^T J_k d \\ & x_k + d \geq 0 \\ & \|d\|_\infty \leq 1. \end{array} \tag{16}$$

By contrast, we will exploit the freedom in our model to introduce a few variants for the computation of the tangential step.

2.1 ADIC-LP: two variants based on linear optimization

We start by describing a variant based on the dual criticality measure given by

$$\omega_{T,k} = \chi_{T,k} = |g_k^T d_{T,k}|, \tag{17}$$

where $d_{T,k}$ is the solution of the linear optimization[1] problem

$$\begin{array}{ll} \min_d & g_k^T d \\ & J_k d = 0 \\ & x_k + d \geq 0 \\ & \|d\|_\infty \leq 1. \end{array} \tag{18}$$

Also observe that

$$\chi_{T,k} \leq \|g_k\| \, \|d_{T,k}\| \leq \sqrt{n}\|g_k\|. \tag{19}$$

The tangential step $s_{T,k}$ can then be computed in two ways. The first is to define $s_{T,k}$ as the solution of the linear programming problem

$$\begin{array}{ll} \min_s & g_k^T s \\ & J_k s = 0 \\ & x_k + s \geq 0 \\ & \|s\|_\infty \leq \alpha_{T,k}\omega_{T,k}. \end{array} \tag{20}$$

(Note that (20) only differs from (18) in the definition of its bounds, and that we have used the liberty in the choice of norms to express the bound on the step in $\|\cdot\|_\infty$). A second, simpler, possibility is to choose a multiple of $d_{T,k}$ and simply set

$$s_{T,k} = \frac{\alpha_{T,k}\omega_{T,k}}{\|d_{T,k}\|_\infty} \, d_{T,k}. \tag{21}$$

Defined in either of these ways, $s_{T,k}$ clearly satisfies (7), (8) and (10) (with $\theta_T \geq \sqrt{n}$), and it is not difficult to verify that it also satisfies (9) with $\kappa_t = 1/\max[\eta, 1]$.

Lemma 2.1 Suppose that, at tangential iteration k_τ, s_{T,k_τ} is defined by either (20) or (21). Then we have that

$$|g_{k_\tau}^T s_{T,k_\tau}| \geq \alpha_{T,k_\tau}\chi_{T,k_\tau}^2 = \frac{\alpha_{T,k_\tau}}{\max[\eta, 1]}\omega_{T,k_\tau}^2. \tag{22}$$

Proof. Suppose first that $\|s_{T,k_\tau}\|_\infty \geq \|d_{T,k_\tau}\|_\infty$. Then d_{T,k_τ} is feasible for problem (20) and thus

$$|g_{k_\tau}^T s_{T,k_\tau}| \geq |g_{k_\tau}^T d_{T,k_\tau}| = \omega_{T,k_\tau} \geq \frac{\alpha_{T,k_\tau}}{\eta}\omega_{T,k_\tau}^2,$$

where we used (14) to deduce the last inequality. Suppose now that $\|s_{T,k_\tau}\|_\infty < \|d_{T,k_\tau}\|_\infty$. Then we must have that $\|s_{T,k_\tau}\|_\infty = \alpha_{T,k_\tau}\omega_{T,k_\tau}$. The vector $y = (\|s_{T,k_\tau}\|_\infty/\|d_{T,k_\tau}\|_\infty)d_{T,k_\tau}$ is therefore feasible for problem (18) and thus

$$|g_{k_\tau}^T y| = \frac{\|s_{T,k_\tau}\|_\infty}{\|d_{T,k_\tau}\|_\infty}|g_{k_\tau}^T d_{T,k_\tau}| = \frac{\alpha_{T,k_\tau}\omega_{T,k_\tau}}{\|d_{T,k_\tau}\|_\infty}\,\omega_{T,k_\tau} \geq \alpha_{T,k_\tau}\omega_{T,k_\tau}^2 \tag{23}$$

[1] Formerly known as "linear programming".

If $s_{T,k}$ is defined by (21), then $s_{T,k} = y$ and (22) follows. Otherwise, we obtain from (20) that $|g_{k_\tau}^T s_{T,k_\tau}| \geq |g_{k_\tau}^T y|$ and (22) also follows from (23). □

When the measure (17) is used, it is also useful to note that, for x such that $c(x) = 0$,

$$\chi_{Tk} = \chi_T(x_k) = \min_{\|d\|_\infty \leq 1} \{\nabla_x L(x_k, \widehat{\lambda}(x_k))^T d \mid J(x_k)d = 0 \text{ and } x_k + d \geq 0\} \tag{24}$$

where the Lagrangian $L(x, \lambda)$ is defined in (12) and $\widehat{\lambda}(x)$ is given by (13).

2.2 ADIC-P1: a projection-based variant

We next consider a variant based on the dual criticality measure given by

$$\omega_{T,k} = \pi_T(x_k) \quad \text{with} \quad \pi_T(x) = \|\Pi_{\mathcal{F}}(x)\Big(x - g(x)\Big) - x\| \stackrel{\text{def}}{=} \|p_1(x)\|. \tag{25}$$

where $\Pi_{\mathcal{F}}(x)$ is the orthogonal projection onto $\mathcal{F}(x) \stackrel{\text{def}}{=} \{x + y \in \mathbb{R}^n \mid J(x)y = 0 \text{ and } x + y \geq 0\}$ (see [6, Section 12.1.4], for instance). In this setting, one still defines $\alpha_{T,k}$ by (2) and one simply chooses

$$s_{T,k} = \min[\alpha_{T,k}, 1]\, p_1(x_k). \tag{26}$$

Again, we note that

$$\pi_T(x_k) \leq \|g_k\|. \tag{27}$$

The minimum in (26) ensures that, by construction, $x_k + s_{T,k} \in \mathcal{F}(x_k)$ and thus that (7) holds. The definition (26) also implies that (8) holds, while (10) with $\theta_T = 1$ directly results from (25). The nature of the orthogonal projection also ensures the following result.

Lemma 2.2 Suppose that, at a tangential iteration k_τ, s_{T,k_τ} is defined by (26). Then

$$|g_{k_\tau}^T s_{T,k_\tau}| \geq \frac{\sqrt{\varsigma}}{\max[\eta, 1]} \alpha_{T,k_\tau} \pi_T(x_{k_\tau})^2 = \frac{\sqrt{\varsigma}}{\max[\eta, 1]} \alpha_{T,k_\tau} \omega_{T,k_\tau}^2. \tag{28}$$

Proof. The optimal nature of the projection implies that

$$\Big([x_{k_\tau} - g_{k_\tau}] - [x_{k_\tau} + p_1(x_{k_\tau})]\Big)^T \Big([x_{k_\tau} + p_1(x_{k_\tau})] - x_{k_\tau}\Big) \geq 0$$

and thus

$$g_{k_\tau}^T p_1(x_{k_\tau}) \leq -\|p_1(x_{k_\tau})\|^2 = -\pi_T(x_{k_\tau})^2.$$

Suppose first that $s_{T,k} = \alpha_{T,k} p_1(x_k)$. Then,

$$|g_{k_\tau}^T s_{T,k}| \geq \alpha_{T,k} \pi_T(x_{k_\tau})^2. \tag{29}$$

Alternatively, if $s_{T,k} = p_1(x_k)$, this implies that $\alpha_{T,k} \geq 1$. Now, (2) gives that $\sqrt{\varsigma}\alpha_{T,k} \leq \max[\eta, 1]$ and hence

$$|g_{k_\tau}^T s_{T,k}| \geq \pi_T(x_{k_\tau})^2 \geq \frac{\sqrt{\varsigma}}{\max[\eta, 1]} \alpha_{T,k} \pi_T(x_{k_\tau})^2. \tag{30}$$

Combining (29) and (30) yields (28). □

Thus the step (26) also satisfies (9) with $\kappa_t = \sqrt{\varsigma}/\max[\eta, 1]$.

2.3 Comments

Some observations are in order at this stage.

1. Three types of iterations may occur in the course of the execution of the algorithm.
   - The first is when the constraint violation is large, in which case condition (3) typically fails. A normal step $s_{N,k}$ is then computed but a tangential step is not, which is probably reasonable because the meaning of a move in the tangent plane far away from the constraint is debatable, as it could result in very large steps which take forever to recover from.
   - The second is when the constraint violation is moderate and (3) holds. Both normal and tangential step may then be computed.
   - The third is when the constraint violation is small. Condition (3) holds so that a tangential step is computed, but a normal step is not.

   What actually happens in a run depends on the choice of the constant β in (3) and the user's decision to avoid or force a normal step when possible.

2. The tangential stepsize formula (2) is of course reminiscent of the stepsize formula used in AdaGrad for unconstrained problems. Note that the running sum of squares of dual measures (Γ_k) is only updated at tangential iterations.

3. We have chosen to use the $\|.\|_\infty$ norm in (16), (18) and (20) so that these problems are standard linear programs, but, as we noted above, this is not necessary. In particular, variants using the (isotropic) Euclidean norm or preconditioned version of these norms may also be considered. One reason to consider Euclidean or other ellipsoidal norms is that n inequality constraints created by the box constraints in the linear programs are replaced by a single constraint.

4. In our statement of the ADIC framework, we have assumed that subproblems ((18), (20) or the projection problem in (25)) are solved exactly. This is not necessary and it is sufficient that approximate solution are accurate enough to produce a decrease in the Lagrangian at least a fraction of the optimal one (as suggested by the introduction of the constant κ_t).

3 Worst-case complexity analysis

This section is devoted to the theoretical study of the ADIC method(s). Its main result is that, under suitable conditions, the average value of $(\omega_{T,k} + \|c_k\|)$ tends to zero like $1/\sqrt{k+1}$. We need the following assumptions to derive it.

AS.0: f and c are continuously differentiable on on open set containing the positive orthant of $\mathbb{R}^n$.

AS.1: For all $x \geq 0$, $f(x) \geq f_{\rm low}$.

AS.2: For all $x \geq 0$, $\|g(x)\| \leq \kappa_g$ where $\kappa_g \geq \eta\beta$.

AS.3: For all $x \geq 0$, $\|c(x)\| \leq \kappa_c$, where $\kappa_c > 1$.

AS.4: For all $x \geq 0$, $\|J(x)\| \leq \kappa_J$

AS.5: For all $x \geq 0$, $\sigma_{\min}[J(x)] \geq \sigma_0 \in (0, 1]$,

AS.6: The gradient $g(x)$ is globally Lipschitz continuous on the positive orthant (with constant L_g).

AS.7: The Jacobian $J(x)$ is globally Lipschitz continuous on the positive orthant (with constant L_J).

AS.8: There exists a constant $\xi \in (0,1]$ such that, for all $k \geq 0$, $\omega_{N,k} \geq \xi \|c_k\|$.

Assumptions AS.1–AS.4 hold if the iterates remain, as is often the case, in a closed bounded set. Using the fact that the product of bounded and Lipschitz functions is Lipschitz, we deduce the following properties, whose detailed proofs can be found in appendix.

Lemma 3.1 Suppose that AS.0 and AS.2–AS.7 hold. Then we have that

1. $c(x)$ is Lipschitz continuous on the positive orthant (with constant L_c),
2. $\nabla_x(\frac{1}{2}\|c(x)\|^2) = J(x)^T c(x)$ is Lipschitz continuous on the positive orthant (with constant $L_{JTc} \geq 1$),
3. $\widehat{\lambda}(x)$ is well-defined on the positive orthant,
4. $\widehat{\lambda}(x)$ is bounded (by the constant κ_λ) and Lipschitz continuous (with constant L_λ) on the positive orthant,
5. $\nabla_x L(x,\lambda)$ is Lipschitz continuous on the positive orthant (with constant L_L).

We also observe that (16), (15), AS.3 and AS.4 ensure that

$$\chi_{N,k} \leq \sqrt{n}\,\|J_k^T c_k\| \leq \sqrt{n}\,\kappa_J\,\kappa_c, \tag{31}$$

Finally, AS.8 assumes that there exists a "sufficient-descent" direction for the problem (16). Specifically, the normal step is designed to reduce $\chi_{N,k}$ but it does not guarantee that $\{\|c_{k_\nu}\|\}$ also converges to zero. In fact, without further assumption, the minimization of $\frac{1}{2}\|c(x)\|^2$ may end up at a local minimizer $x_{\rm loc}$ of this function which is infeasible for the original problem because $c(x_{\rm loc}) \neq 0$. The existence of such local minimizers may be caused by a singular Jacobian $J(x_{\rm loc})$ (in which case $J(x_{\rm loc})^T c(x_{\rm loc}) = 0$ does not imply $c(x_{\rm loc}) = 0$), or by the presence of bounds since $-J(x_{\rm loc})^T c(x_{\rm loc})$ may then belong to the normal cone of the bound constraints at $x_{\rm loc}$. Unfortunately, convergence to such an $x_{\rm loc}$ cannot be avoided without either applying a global optimization method to minimize $\frac{1}{2}\|c(x)\|^2$, or restricting the class of problems under consideration. Here we follow the second approach and first note that AS.5 already ensures that $J(x_{\rm loc})^T c(x_{\rm loc}) = 0$ implies $c(x_{\rm loc}) = 0$. Making AS.8 is motivated by the observation that descent along any direction for problem (16) not hitting the non-negativity constraints must be limited by the bound $\|d\|_\infty \leq 1$. Thus at least one component of $d_{N,k}$, say components $i \in \mathcal{I} \subseteq \{1,\ldots,n\}$, must be equal to one in absolute value, which implies that $\chi_{N,k} = |c_k^T J_k d_{N,k}| \geq \|[J_k^T c_k]_\mathcal{I}\|_1$. AS.8 then guarantees that $\|[J_k^T c_k]_\mathcal{I}\|_1$ is not negligible with respect to $\|J_k^T c_k\| \geq \sigma_0 \|c_k\|$.

To maintain generality, we finally assume that the considered criticality measures are bounded.

AS.9: There exists a constant $\kappa_\omega > 0$ such that, for all $x \in \mathbb{R}^n$, $\omega_T(x) \leq \kappa_\omega$ and $\omega_N(x) \leq \kappa_\omega$.

For the special cases discussed in Sections 2.1 and 2.2, AS.9 automatically results from AS.2–AS.4, as can be seen from (19), (27) and (31).

The next result shows our requirements on the normal step in Step 3 of the ADIC are not excessive. This is achieved by exhibiting one particular computational scheme (a trust-region method) which satisfies the conditions (4)–(6).

Lemma 3.2 The normal step $s_{N,k}$ (in Step 2) can be computed using a trust-region algorithm applied to minimizing $\frac{1}{2}\|c(x)\|^2$ subject to (4) and (5) with $\|.\| = \|\cdot\|_\infty$ and $\omega_{N,k} = \chi_{N,k}$ defined by (15), starting with the radius $\theta_N \chi_{N,k}$.

Proof. For any $\Delta \leq \min[1, \theta_N \chi_{N,k}]$, let $s_{TR}(\Delta)$ be the solution of the problem of minimizing $c_k^T J_k s$ over the constraints $x + s_{TR}(\Delta) \geq 0$ and $\|s_{TR}(\Delta)\|_\infty \leq \Delta$. Using the Lipschitz continuity of $J(x)^T c(x)$ (with constant L_{JTc}), we obtain that

$$\tfrac{1}{2}\Big(\|c(x_k + s_{TR}(\Delta))\|^2 - \|c_k\|^2\Big) \leq c_k^T J_k s_{TR}(\Delta) + \frac{L_{JTc}}{2}\Delta^2 \leq c_k^T J_k s_{TR}(\Delta) + \frac{\max[\kappa_\omega, L_{JTc}]}{2}\Delta^2, \tag{32}$$

where κ_ω is defined in AS.9. Now, since $\Delta \leq \min[1, \theta_N \chi_{N,k}]$, given the vector $d_{N,k}$ solution to (16), the vector $\Delta d_{N,k}$ is feasible for (4)-(5), and thus, from (15),

$$c_k^T J_k s_{TR}(\Delta) \leq c_k^T J_k (\Delta d_{N,k}) = -\chi_{N,k}\Delta.$$

Hence

$$c_k^T J_k s_{TR}(\Delta) + \frac{\max[\kappa_\omega, L_{JTc}]}{2}\Delta^2 \leq -\chi_{N,k}\Delta + \frac{\max[\kappa_\omega, L_{JTc}]}{2}\Delta^2. \tag{33}$$

It is then easy to verify that, if $\Delta \leq \chi_{N,k}/\max[\kappa_\omega, L_{JTc}]$ then (33) gives that

$$c_k^T J_k s_{TR}(\Delta) + \frac{\max[\kappa_\omega, L_{JTc}]}{2}\Delta^2 \leq -\frac{1}{2}\chi_{N,k}\Delta. \tag{34}$$

Remembering (31), we may then define

$$s_{N,k} = s_{TR}(\Delta_*) \quad \text{with} \quad \Delta_* = \min\left[1, \frac{\chi_{N,k}}{\max[\kappa_\omega, L_{JTc}]}\right] = \frac{\chi_{N,k}}{\max[\kappa_\omega, L_{JTc}]} \leq \theta_N \chi_{N,k}$$

where the last inequality, which shows that (5) holds, is derived using the bounds $L_{Tc} \geq 1$ and $\theta_N > 1$. Substituting this value in (34) and using (32) then yields that

$$\tfrac{1}{2}\Big(\|c(x_k + s_{N,k})\|^2 - \|c_k\|^2\Big) \leq -\frac{\chi_{N,k}^2}{2\max[\kappa_\omega, L_{JTc}]}$$

which proves the desired conclusion (with $\kappa_n = 1/(2\max[\kappa_\omega, L_{JTc}]) \in (0, \frac{1}{2})$), because the radius Δ can then be reduced (if necessary) starting from $\theta_N \chi_{N,k}$ until (6) holds. □

The previous result implies that the normal step can be computed using a trust-region algorithm for minimizing $\frac{1}{2}\|c(x)\|^2$ subject to (4) and (5), $\|.\| = \|\cdot\|_\infty$ and imposing that the trust-region solution $s_{TR}(\Delta)$ satisfies

$$\tfrac{1}{2}\|c(x_k + s_{TR}(\Delta))\|^2 \leq \tfrac{1}{2}\|c_k\|^2 - \tfrac{1}{2}\chi_{N,k}(\Delta).$$

Our analysis now proceeds by studying the behaviour of the Lyapunov function (11) for iterations using normal and tangential steps (indexed by k_ν and k_τ, respectively), before combining the results and deriving global rates of convergence of $(\omega_{T,k} + \|c_k\|)$ to zero along the sequences $\{k_\tau\}$, $\{k_\nu\}$ and, finally, $\{k\}$. For brevity, we define the abbreviated notations

$$\psi(x) \stackrel{\text{def}}{=} \psi\big(x, \widehat{\lambda}(x)\big) \quad \text{and} \quad \widehat{\lambda}_k = \widehat{\lambda}(x_k). \tag{35}$$

We also observe that (12) and (13) ensure that, for $\lambda^\dagger = \widehat{\lambda}(x)$,

$$\nabla_x L(x, \lambda^\dagger) = g(x) + J(x)^T \widehat{\lambda}(x) = g_T(x), \tag{36}$$

where $g_T(x)$ is the orthogonal projection of $g(x)$ onto the nullspace of $J(x)$, and consequently, using AS.2,

$$\|\nabla_x L(x, \lambda^\dagger)\| \leq \|g(x)\| \leq \kappa_g. \tag{37}$$

3.1 Descent at normal steps

We first consider the effect of normal steps on the value of the Lyapunov function ψ. We start by a very simple observation.

Lemma 3.3 Suppose that AS.5 and AS.8 hold and that a normal step is used at iteration k_ν. Then $c_{k_\nu}^+ = c(x_{k_\nu} + s_{N,k_\nu})$ satisfies

$$\|c_{k_\nu}^+\| - \|c_{k_\nu}\| \le -\kappa_n \xi \, \omega_{N,k_\nu}. \tag{38}$$

Proof. We have from (6) that $\|c_{k_\nu}^+\| < \|c_{k_\nu}\|$. Then,

$$2\|c_{k_\nu}\|(\|c_{k_\nu}\| - \|c_{k_\nu}^+\|) \ge (\|c_{k_\nu}\| + \|c_{k_\nu}^+\|)(\|c_{k_\nu}\| - \|c_{k_\nu}^+\|) = \|c_{k_\nu}\|^2 - \|c_{k_\nu}^+\|^2,$$

and therefore, using (6) and AS.8, that

$$\|c_{k_\nu}^+\| - \|c_{k_\nu}\| \le -\frac{\kappa_n \omega_{N,k_\nu}^2}{\|c_{k_\nu}\|} \le -\kappa_n \xi \, \omega_{N,k_\nu}$$

□

We then use this observation to deduce the following result.

Lemma 3.4 Suppose that AS.3–AS.9 hold and that a normal step is used at iteration k_ν. Define

$$\rho = \frac{1}{\kappa_n \xi}, \left[(\kappa_g + \kappa_c \, L_\lambda)\theta_N + \left(\frac{L_L}{2} + L_\lambda L_c \right) \theta_N^2 \kappa_\omega + \eta \right] \tag{39}$$

Then $x_{k_\nu}^+ = x_{k_\nu} + s_{N,k_\nu}$ satisfies

$$\psi(x_{k_\nu}^+) - \psi(x_{k_\nu}) \le -\eta \, \omega_{N,k_\nu}. \tag{40}$$

Proof. We have that

$$\psi(x_{k_\nu}^+) - \psi(x_{k_\nu}) = \underbrace{\psi(x_{k_\nu}^+, \widehat{\lambda}_{k_\nu}) - \psi(x_{k_\nu}, \widehat{\lambda}_{k_\nu})}_{\Delta_x} + \underbrace{\psi(x_{k_\nu}^+, \widehat{\lambda}_{k_\nu}^+) - \psi(x_{k_\nu}^+, \widehat{\lambda}_{k_\nu})}_{\Delta_\lambda}. \tag{41}$$

Now consider Δ_x and Δ_λ separately. Using the Lipschitz continuity of $\nabla_x \psi(x, \widehat{\lambda})$ (ρ is fixed in (39)) and (38), we obtain that

$$\begin{aligned} \Delta_x &= \psi(x_{k_\nu}^+, \widehat{\lambda}_{k_\nu}) - \psi(x_{k_\nu}, \widehat{\lambda}_{k_\nu}) \\ &= L(x_{k_\nu}^+, \widehat{\lambda}_{k_\nu}) - L(x_{k_\nu}, \widehat{\lambda}_{k_\nu}) + \rho\big(\|c_{k_\nu}^+\| - \|c_{k_\nu}\|\big) \\ &\le (\nabla_x L(x_{k_\nu}, \widehat{\lambda}_{k_\nu})^T s_{N,k_\nu} + r_3 - \rho \kappa_n \xi \, \omega_{N,k_\nu} \end{aligned} \tag{42}$$

with

$$|r_3| \le \frac{L_L}{2} \|s_{N,k_\nu}\|^2.$$

We now invoke the Cauchy-Schwartz inequality, (37) and (5) to deduce that

$$\begin{aligned}\Delta_x &\le \|\nabla_x L(x_{k_\nu}, \widehat{\lambda}_{k_\nu})\| \, \|s_{N,k_\nu}\| - \rho \frac{\kappa_n \xi}{2} \, \omega_{N,k_\nu} + \frac{L_L}{2}\|s_{N,k_\nu}\|^2 \\ &\le \kappa_g \|s_{N,k_\nu}\| - \rho\kappa_n\xi \, \omega_{N,k_\nu} + \frac{L_L}{2}\|s_{N,k_\nu}\|^2 \\ &\le \kappa_g \theta_N \omega_{N,k_\nu} - \rho\kappa_n\xi \, \omega_{N,k_\nu} + \frac{L_L}{2}\theta_N^2 \omega_{N,k_\nu}^2 .\end{aligned} \tag{43}$$

Using now the definition of Δ_λ in (41), AS.6, the Lipschitz continuity of $\widehat{\lambda}$ and c and AS.3 then yields that

$$\begin{aligned}\Delta_\lambda &= \psi(x_{k_\nu}^+, \widehat{\lambda}(x_{k_\nu}^+)) - \psi(x_{k_\nu}^+, \widehat{\lambda}(x_{k_\nu})) \\ &\le (\|c_{k_\nu}\| + \|c_{k_\nu}^+ - c_{k_\nu}\|) \, \|\widehat{\lambda}_{k_\nu}^+ - \widehat{\lambda}_{k_\nu}\| \\ &\le L_\lambda \, \|s_{N,k_\nu}\| \, \|c_{k_\nu}\| + L_\lambda L_c \|s_{N,k_\nu}\|^2 \\ &\le L_\lambda \, \theta_N \kappa_c \omega_{N,k_\nu} + L_\lambda L_c \theta_N^2 \omega_{N,k_\nu}^2\end{aligned} \tag{44}$$

and thus, summing (43) and (44), that

$$\begin{aligned}&\psi(x_{k_\nu}^+) - \psi(x_{k_\nu}) \\ &\quad \le -\rho\kappa_n\xi \, \omega_{N,k_\nu} + \kappa_g\theta_N\omega_{N,k_\nu} + L_\lambda \, \kappa_c \, \theta_N \omega_{N,k_\nu} + \left(\frac{\theta_N^2 L_L}{2} + \theta_N^2 L_\lambda L_c\right) \omega_{N,k_\nu}^2 \\ &\quad \le -\rho\kappa_n\xi\omega_{N,k_\nu} + (\kappa_g\theta_N + L_\lambda \, \kappa_c \, \theta_N) \, \omega_{N,k_\nu} + \left(\frac{\theta_N^2 L_L}{2} + \theta_N^2 L_\lambda L_c\right) \kappa_\omega \, \omega_{N,k_\nu},\end{aligned}$$

where we have used AS.9 to deduce the second inequality. The bound (40) then follows from (39).
□

Note that, should $s_{N,k}$ belong to the range space of J_k, the first term in the last right-hand side of (42) vanishes and κ_g disappears from (43) and, consequently, from (39).

3.2 Descent at tangential steps

We now turn to considering the effect of tangential steps.

Lemma 3.5 Suppose that AS.4–AS.8 hold. Then

$$\psi(x_{k_\tau+1}) - \psi(x_{k_\tau}^+) \le -\kappa_t \alpha_{T,k_\tau} \omega_{T,k_\tau}^2 + \kappa_{\tan} \, \alpha_{T,k_\tau}^2 \omega_{T,k_\tau}^2 . \tag{45}$$

where

$$\kappa_{\tan} = \left[\frac{\theta_T^2}{2}\Big(L_L + \rho L_c\Big) + \beta\theta_N\theta_T\Big(L_L + \kappa_J L_\lambda + \rho L_J\Big)\right] + \frac{\beta\theta_T L_\lambda}{\xi} + \theta_T^2 L_c L_\lambda. \tag{46}$$

Proof. As in (41), we now have that

$$\psi(x_{k_\tau+1}) - \psi(x_{k_\tau}^+) = \underbrace{\psi(x_{k_\tau+1}, \widehat{\lambda}_{k_\tau}^+) - \psi(x_{k_\tau}^+, \widehat{\lambda}_{k_\tau}^+)}_{\Delta_x} + \underbrace{\psi(x_{k_\tau+1}, \widehat{\lambda}_{k_\tau+1}) - \psi(x_{k_\tau+1}, \widehat{\lambda}_{k_\tau}^+)}_{\Delta_\lambda} . \tag{47}$$

The Lipschitz continuity of $\nabla_x \psi(x, \widehat{\lambda})$, (11) and (35) give that

$$\Delta_x = \nabla_x L(x_{k_\tau}^+, \widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau} + r_0 + \rho(\|c_{k_\tau+1}\| - \|c_{k_\tau}^+\|) \quad \text{with} \quad |r_0| \le \frac{L_L}{2}\|s_{T,k_\tau}\|^2. \tag{48}$$

Equation (8) gives

$$\begin{aligned}\|c(x_{k_\tau+1})\| &= \|c(x_{k_\tau}^+) - J_{k_\tau}s_{T,k_\tau} + (J_{k_\tau}^+ - J_{k_\tau})s_{T,k_\tau} + r_1\| \\ &\leq \|c_{k_\tau}^+\| + \|r_1\| + \|J_{k_\tau}^+ - J_{k_\tau}\|\,\|s_{T,k_\tau}\| \\ &\leq \|c_{k_\tau}^+\| + \|r_1\| + L_J\|s_{N,k_\tau}\|\,\|s_{T,k_\tau}\|\end{aligned}$$

with $\|r_1\| \leq \frac{L_c}{2}\|s_{T,k_\tau}\|^2$. Now $\|s_{N,k_\tau}\|$ is either zero (if $k_\tau \notin \{k_\nu\}$) or, using (3) for k_τ,

$$\omega_{N,k_\tau} \leq \beta\alpha_{T,k_\tau}\omega_{T,k_\tau}$$

and thus

$$\|s_{N,k_\tau}\| \leq \theta_N\omega_{N,k_\tau} \leq \beta\theta_N\alpha_{T,k_\tau}\omega_{T,k_\tau}, \tag{49}$$

so that, whether $k_\tau \in \{k_\nu\}$ or not, using (10),

$$\|c(x_{k_\tau+1})\| - \|c(x_{k_\tau}^+)\| \leq \left(\frac{\theta_T^2 L_c}{2} + \beta\theta_N\theta_T L_J\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2. \tag{50}$$

Now differentiating L with respect to its first argument and using the Lipschitz continuity of $\nabla_x L$ with respect to this first argument, AS.4, (8) and the Lipschitz continuity of $\widehat{\lambda}$ gives that

$$\begin{aligned}\nabla_x L(x_{k_\tau}^+,\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau} &= \Big(\nabla_x L(x_{k_\tau}^+,\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau} - \nabla_x L(x_{k_\tau},\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau}\Big) \\ &\quad + \Big(\nabla_x L(x_{k_\tau},\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau} - \nabla_x L(x_{k_\tau},\widehat{\lambda}_{k_\tau})^T s_{T,k_\tau}\Big) \\ &\quad + g_{k_\tau}^T s_{T,k_\tau} + \widehat{\lambda}_{k_\tau}^T J_{k_\tau}s_{T,k_\tau} \\ &= \Big(\nabla_x L(x_{k_\tau}^+,\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau} - \nabla_x L(x_{k_\tau},\widehat{\lambda}_{k_\tau}^+)^T s_{T,k_\tau}\Big) \\ &\quad + \Big((\widehat{\lambda}_{k_\tau}^+)^T J_{k_\tau} - \widehat{\lambda}_{k_\tau}^T J_{k_\tau}\Big)^T s_{T,k_\tau} + g_{k_\tau}^T s_{T,k_\tau} + \widehat{\lambda}_{k_\tau}^T J_{k_\tau}s_{T,k_\tau} \\ &\leq \Big(L_L + \kappa_J L_\lambda\Big)\|s_{N,k_\tau}\|\,\|s_{T,k_\tau}\| + g_{k_\tau}^T s_{T,k_\tau} + \widehat{\lambda}_{k_\tau}^T J_{k_\tau}s_{T,k_\tau} \\ &\leq g_{k_\tau}^T s_{T,k_\tau} + \beta\Big(L_L + \kappa_J L_\lambda\Big)\theta_N\theta_T\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2\end{aligned}$$

where the last inequality results from (8) and (49). Hence we obtain from (48), (9), (10) and (50) that

$$\begin{aligned}\Delta_x &\leq g_{k_\tau}^T s_{T,k_\tau} + r_0 + \rho\left(\frac{\theta_T^2 L_c}{2} + \beta\theta_N\theta_T L_J\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2 + \beta\theta_N\theta_T\Big(L_L + \kappa_J L_\lambda\Big)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2 \\ &\leq g_{k_\tau}^T s_{T,k_\tau} + \left(\frac{\theta_T^2}{2}(L_L + \rho L_c) + \beta\theta_N\theta_T\Big(L_L + \kappa_J L_\lambda + \rho L_J\Big)\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2 \\ &\leq -\kappa_t\alpha_{T,k_\tau}\omega_{T,k_\tau}^2 + \left(\frac{\theta_T^2}{2}(L_L + \rho L_c) + \beta\theta_N\theta_T\Big(L_L + \kappa_J L_\lambda + \rho L_J\Big)\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2.\end{aligned} \tag{51}$$

Now, we may use the Lipschitz continuity of $\widehat{\lambda}$ and c, inequality (6), the Cauchy-Schwartz inequality, and AS.8 to deduce that

$$\begin{aligned}\Delta_\lambda &= c_{k_\tau+1}^T\big(\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\big) \\ &= (c_{k_\tau+1} - c_{k_\tau}^+)^T\big(\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\big) + (c_{k_\tau}^+)^T\big(\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\big) \\ &\leq \|c_{k_\tau}^+\|\,\|\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\| + \|c_{k_\tau+1} - c_{k_\tau}^+\|\,\|\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\| \\ &\leq \|c_{k_\tau}\|\,\|\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\| + \|c_{k_\tau+1} - c_{k_\tau}^+\|\,\|\widehat{\lambda}_{k_\tau+1} - \widehat{\lambda}_{k_\tau}^+\| \\ &\leq \frac{L_\lambda}{\xi}\omega_{N,k_\tau}\|s_{T,k_\tau}\| + L_cL_\lambda\|s_{T,k_\tau}\|^2.\end{aligned} \tag{52}$$

Again using (3) for $k \in \{k_\tau\}$, (10) and (49), we obtain that

$$\Delta_\lambda \le \frac{\beta\theta_T L_\lambda}{\xi}\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2 + \theta_T^2 L_c L_\lambda \alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2.$$

Thus, summing Δ_x and Δ_λ, we deduce that

$$\begin{aligned}\psi(x_{k_\tau+1}) - \psi(x_{k_\tau}^+) \le -\kappa_t\alpha_{T,k_\tau}\omega_{T,k_\tau}^2 + \left(\frac{\theta_T^2}{2}\Big(L_L+\rho L_c\Big) + \beta\theta_N\theta_T\Big(L_L+\kappa_J L_\lambda+\rho L_J\Big)\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2 \\ + \left(\frac{\beta\theta_T L_\lambda}{\xi} + \theta_T^2 L_c L_\lambda\right)\alpha_{T,k_\tau}^2\omega_{T,k_\tau}^2\end{aligned}$$

and (45) follows. □

Observe that the second term in the bracket of (46) only appears when $k_\tau \in \{k_\nu\}$. The bound (45) quantifies the effect of tangential steps on the Lyapunov function, and its right-hand side involves a first-order (descent) term and a second-order perturbation term. We now derive crucial bounds on these terms, using the fact that Γ_k is not updated at normal iterations.

Lemma 3.6 Suppose that AS.2 and AS.5 hold. If we denote

$$\Gamma_{k_{\tau_0}} = 0, \quad \Gamma_{k_{\tau+1}} = \Gamma_{k_\tau} + \omega_{T,k_\tau}^2, \quad \alpha_{T,k_\tau} = \frac{\eta}{\sqrt{\varsigma + \Gamma_{k_{\tau+1}}}},$$

then, for all $\tau_0 \le \tau_1$,

$$\sum_{\tau=\tau_0}^{\tau_1} \alpha_{T,k_\tau}\,\omega_{T,k_\tau}^2 > \eta\sqrt{\varsigma}\sqrt{1+\frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} - \eta\sqrt{\varsigma} \tag{53}$$

$$\sum_{\tau=\tau_0}^{\tau_1} \alpha_{T,k_\tau}^2\,\omega_{T,k_\tau}^2 \le \eta^2 \log\left(1+\frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right). \tag{54}$$

Proof. Let $w_{k_\tau+1} = \sqrt{\Gamma_{k_{\tau+1}}+\varsigma}$. The definition of α_{T,k_τ} in (2) implies that

$$\begin{aligned}\sum_{\tau=\tau_0}^{\tau_1}\alpha_{T,k_\tau}\,\omega_{T,k_\tau}^2 &= \eta\sum_{\tau=\tau_0}^{\tau_1}\frac{\omega_{T,k_\tau}^2}{\sqrt{\varsigma+\Gamma_{k_{\tau+1}}}} \\ &> \eta\sum_{\tau=\tau_0}^{\tau_1}\frac{\omega_{T,k_\tau}^2}{w_{k_{\tau+1}}+w_{k_\tau}} \\ &= \eta\sum_{\tau=\tau_0}^{\tau_1}\frac{w_{k_{\tau+1}}^2 - w_{k_\tau}^2}{w_{k_{\tau+1}}+w_{k_\tau}} \\ &= \eta\sum_{\tau=\tau_0}^{\tau_1}(w_{k_{\tau+1}} - w_{k_\tau}) \\ &= \eta\big(w_{k_{\tau_1+1}} - w_{k_{\tau_0}}\big).\end{aligned}$$

Now observe that, using $\Gamma_{k_{\tau_0}} = 0$,

$$w_{k_{\tau_1+1}} - w_{k_{\tau_0}} = \sqrt{\varsigma+\Gamma_{k_{\tau_1+1}}} - \sqrt{\varsigma+\Gamma_{k_{\tau_0}}} = \sqrt{\varsigma+\Gamma_{k_{\tau_1+1}}} - \sqrt{\varsigma},$$

which then gives (53). Using the concavity and the increasing nature of the logarithm, we also have from (2) that

$$\alpha_{T,k_\tau}^2 \omega_{T,k_\tau}^2 = \eta^2 \frac{\omega_{T,k_\tau}^2}{\varsigma + \Gamma_{k_\tau+1}} = \eta^2 \frac{\Gamma_{k_{\tau+1}} - \Gamma_{k_\tau}}{\varsigma + \Gamma_{k_{\tau+1}}} \le \eta^2 \left[\log(\varsigma + \Gamma_{k_{\tau+1}}) - \log(\varsigma + \Gamma_{k_\tau})\right].$$

Summing for $\tau \in \{\tau_0, \ldots, \tau_1\}$ then yields that

$$\sum_{\tau=\tau_0}^{\tau_1} \alpha_{T,k_\tau}^2 \, \omega_{T,k_\tau}^2 \le \eta^2 \left[\log(\varsigma + \Gamma_{k_{\tau_1+1}}) - \log(\varsigma + \Gamma_{k_{\tau_0}})\right],$$

and (54) follows, again using $\Gamma_{k_{\tau_0}} = 0$. □

3.3 Telescoping sum

Having considered the impacts of tangential and normal steps separately, we now combine them to derive a crucial inequality.

Lemma 3.7 Suppose that AS.0–AS.8 hold. Then, for any $\tau_1 > 0$ and any $\nu_1 \ge 0$,

$$\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} + \sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} \le \kappa_{\rm gap} + \frac{\kappa_{\tan}}{\kappa_t\sqrt{\varsigma}} \log\left(1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right), \tag{55}$$

where

$$\kappa_{\rm gap} = \frac{1}{\eta\kappa_t\sqrt{\varsigma}} \left(1 + \psi(x_0) + \kappa_c\kappa_\lambda + \rho\kappa_c - f_{\rm low}\right).$$

Proof. Consider $k \ge 0$. Then, defining $\min[k_{\nu_0}, k_{\tau_0}] = 0$ and $\max[k_{\nu_1}, k_{\tau_1}] = k$, we have that $\Gamma_{k_{\tau_0}} = 0$ and we may apply Lemma 3.6. Combining (45), (53) and (54), we obtain that

$$\sum_{\tau=\tau_0}^{\tau_1} \left(\psi(x_{k_\tau+1}) - \psi(x_{k_\tau}^+)\right) \le \eta\kappa_t\sqrt{\varsigma} - \eta\kappa_t\sqrt{\varsigma}\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} + \eta^2\kappa_{\tan}\log\left(1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right). \tag{56}$$

Also considering (40) and observing that $x_k^+ = x_{k+1}$ when $k \in \{k_\nu\} \setminus \{k_\tau\}$ and $x_k^+ = x_k$ when $k \in \{k_\tau\} \setminus \{k_\nu\}$ therefore yields that

$$\begin{aligned}
\psi(x_{k+1}) - \psi(x_0) &= \sum_{\tau=\tau_0}^{\tau_1} \left(\psi(x_{k_\tau+1}) - \psi(x_{k_\tau}^+)\right) + \sum_{\nu=\nu_0}^{\nu_1} \left(\psi(x_{k_\nu}^+) - \psi(x_{k_\nu})\right) \\
&\le \eta\kappa_t\sqrt{\varsigma} - \eta\kappa_t\sqrt{\varsigma}\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} - \eta\sum_{\nu=\nu_0}^{\nu_1}\omega_{N,k_\nu} + \eta^2\kappa_{\tan}\log\left(1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right) \\
&\le \eta\kappa_t\sqrt{\varsigma} - \eta\kappa_t\sqrt{\varsigma}\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} - \eta\kappa_t\sqrt{\varsigma}\sum_{\nu=\nu_0}^{\nu_1}\omega_{N,k_\nu} + \eta\kappa_{\tan}\log\left(1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right),
\end{aligned} \tag{57}$$

where we used the facts that $\kappa_t\sqrt{\varsigma} < 1$ and $\eta \le 1$. Using now (11), (35), the Cauchy-Schwartz inequality, the boundedness of $\widehat{\lambda}(x)$, AS.1 and AS.3, we have that

$$\psi(x_{k+1}) - \psi(x_0) - \eta\kappa_t\sqrt{\varsigma} \ge f_{\rm low} - \kappa_c\kappa_\lambda - \rho\kappa_c - \psi(x_0) - \eta\kappa_t\sqrt{\varsigma} \stackrel{\rm def}{=} -\eta\kappa_t\sqrt{\varsigma}\,\kappa_{\rm gap},$$

so that (57) implies (55). □

3.4 Tangential complexity

Lemma 3.7 implies upper bounds for both $\Gamma_{k\tau}$ and ω_{N,k_ν}. We now exploit the first of these to derive the rate of convergence for tangential steps proper, after establishing a useful technical result.

Lemma 3.8 Suppose that $at \leq b + c\,\log(t)$ for $t \geq 1$ and $a, c > 0$. Then,

$$t \leq \frac{2b}{a} + \frac{2c}{a}\left[\log\left(\frac{2c}{a}\right) - 1\right].$$

Proof. See [1, Lemma 3.2]. □

Lemma 3.9 Suppose that AS.0–AS.9 hold. Then, for any $\tau_1 > 0$,

$$\sqrt{\varsigma + \Gamma_{k_{\tau_1+1}}} \leq \kappa_T \stackrel{\rm def}{=} 2\kappa_{\rm gap}\sqrt{\varsigma} + \frac{4\kappa_{\rm tan}}{\kappa_t}\left[\log\left(\frac{4\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}}\right) - 1\right] \tag{58}$$

and

$$\xi\sum_{\tau=\tau_0}^{\tau_1}\Big(\omega_{T,k_\tau} + \|c_{k_\tau}\|\Big) \leq \sum_{\tau=\tau_0}^{\tau_1}\Big(\omega_{T,k_\tau} + \omega_{N,k_\tau}\Big) \leq \kappa_T\sqrt{\tau_1+1}\left(1 + \frac{\beta\eta}{\sqrt{\varsigma}}\right). \tag{59}$$

Proof. The bound (55) implies that

$$\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} \leq \kappa_{\rm gap} + \frac{\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}}\log\left(1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}\right) = \kappa_{\rm gap} + \frac{2\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}}\log\left(\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}}\right).$$

Using Lemma 3.8 with

$$t = \sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}}, \qquad a = 1, \qquad b = \kappa_{\rm gap} \qquad \text{and} \qquad c = \frac{2\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}},$$

we then obtain that

$$\sqrt{\varsigma + \Gamma_{k_{\tau_1+1}}} = \sqrt{\varsigma}\sqrt{1 + \frac{\Gamma_{k_{\tau_1+1}}}{\varsigma}} \leq \sqrt{\varsigma}\left\{2\kappa_{\rm gap} + \frac{4\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}}\left[\log\left(\frac{4\kappa_{\rm tan}}{\kappa_t\sqrt{\varsigma}}\right) - 1\right]\right\}.$$

This is (58). We may now invoke the inequality

$$\sum_{j=0}^{k} a_j \leq \sqrt{k+1}\sqrt{\sum_{j=0}^{k} a_j^2}$$

for nonnegative $\{a_j\}_{j=0}^k$ to deduce from the definition of Γ_{k_τ} and (58) that

$$\sum_{\tau=\tau_0}^{\tau_1}\omega_{T,k_\tau} \leq \sqrt{\tau_1+1}\sqrt{\sum_{\tau=\tau_0}^{\tau_1}\omega_{T,k_\tau}^2} = \sqrt{\tau_1+1}\sqrt{\Gamma_{k_{\tau_1}+1}} < \sqrt{\tau_1+1}\sqrt{\varsigma + \Gamma_{k_{\tau_1}+1}} \leq \sqrt{\tau_1+1}\,\kappa_T. \tag{60}$$

Using the switching condition (3) and the first part of (14), we then deduce that, whether k_τ belongs to $\{k_\nu\}$ or not,

$$\sum_{\tau=\tau_0}^{\tau_1} \omega_{N,k_\tau} \leq \sum_{\tau=\tau_0}^{\tau_1} \beta\alpha_{T,k_\tau}\omega_{T,k_\tau} \leq \frac{\beta\eta}{\sqrt{\varsigma}} \sum_{\tau=\tau_0}^{\tau_1} \omega_{T,k_\tau} < \frac{\beta\eta\kappa_T\sqrt{\tau_1+1}}{\sqrt{\varsigma}}.$$

Summing this bound with (60) then gives the second inequality of (59). The first results from AS.8. □

3.5 Normal complexity

We now exploit the bound on ω_{N,k_ν} stated in Lemma 3.7 to analyze the complexity of the subsequence of normal iterations. We first show that the sum of the norms of constraint violations is bounded.

Lemma 3.10 Suppose that AS.0–AS.9 hold. Then, for any $\nu_1 > 0$,

$$\xi \sum_{\nu=\nu_0}^{\nu_1} \|c_{k_\nu}\| \leq \sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} < \kappa_N, \tag{61}$$

where

$$\kappa_N = \kappa_{\rm gap} + \kappa_{\rm tan} \log\left(1 + \frac{\kappa_T^2}{\varsigma}\right). \tag{62}$$

Proof. The bound (55) ensures that

$$\sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} \leq \kappa_{\rm gap} + \kappa_{\rm tan} \log\left(1 + \frac{\Gamma_{k_{\tau_1}+1}}{\varsigma}\right), \tag{63}$$

where k_{τ_1} is the index of the last tangential iteration before k_{ν_1}. Substituting the bound (58) in this inequality then gives the second inequality of (61), the first resulting again from AS.8. □

This allows us to derive boundedness of a combined primal and dual criticality measure.

Lemma 3.11 Suppose that AS.0–AS.9 hold. Then, for any $\nu_1 \geq 0$,

$$\xi \sum_{\nu=\nu_0, k_\nu \notin \{k_\tau\}}^{\nu_1} \Big(\omega_{T,k_\nu} + \|c_{k_\nu}\|\Big) \leq \sum_{\nu=\nu_0, k_\nu \notin \{k_\tau\}}^{\nu_1} \Big(\omega_{T,k_\nu} + \omega_{N,k_\nu}\Big) < \kappa_N \left(1 + \frac{\kappa_T}{\beta\eta}\right). \tag{64}$$

Proof. Using the switching condition (3) for $k_\nu \notin \{k_\tau\}$, we obtain that, for such k_ν with $\nu \in \{\nu_0, \ldots, \nu_1\}$,

$$\omega_{N,k_\nu} > \beta\,\alpha_{T,k_\nu}\omega_{T,k_\nu}. \tag{65}$$

As in the previous lemma, let k_{τ_1} be the index of the last tangential iteration before k_{ν_1}. Thus using (58),

$$\alpha_{T,k_\nu} = \frac{\eta}{\sqrt{\varsigma + \Gamma_{k_\nu}}} \geq \frac{\eta}{\kappa_T}.$$

Substituting this bound in (65), we find that, for $\nu \in \{\nu_0, \dots, \nu_1\}$,

$$\omega_{N,k_\nu} \geq \frac{\beta\eta}{\kappa_T}\,\omega_{T,k_\nu}. \tag{66}$$

With (61), this implies that

$$\sum_{\nu=\nu_0}^{\nu_1} \omega_{T,k_\nu} \leq \frac{\kappa_T}{\beta\eta}\sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} \leq \frac{\kappa_N\,\kappa_T}{\beta\eta}.$$

Summing this bound with (61) and using AS.8 gives (64). □

3.6 Combined complexity

We finally assemble the pieces of the puzzle to derive our main result on the global rate of convergence of the ADIC algorithm.

Theorem 3.12 Suppose that AS.0-AS.9 hold. Then, for any $k \geq 0$,

$$\frac{1}{k+1}\sum_{j=0}^{k}\Big(\omega_{T,j} + \|c_j\|\Big) \leq \frac{\kappa_{\mathsf{ADIC},1}}{\sqrt{k+1}} + \frac{\kappa_{\mathsf{ADIC},2}}{k+1} = \mathcal{O}\left(\frac{1}{\sqrt{k+1}}\right), \tag{67}$$

where

$$\kappa_{\mathsf{ADIC},\,1} = \frac{\kappa_T}{\xi}\left(1 + \frac{\beta\eta}{\sqrt{\varsigma}}\right) \quad \text{and} \quad \kappa_{\mathsf{ADIC},\,2} = \frac{\kappa_N}{\xi}\left(1 + \frac{\kappa_T}{\beta\eta}\right).$$

Proof. Consider iterations of both types (tangential and normal) from 0 to k by defining $\min[k_{\nu_0}, k_{\tau_0}] = 0$ and $\max[k_{\nu_1}, k_{\tau_1}] = k$ (as in Lemma 3.7). We then obtain, by combining (59) and (64), that

$$\begin{aligned}\sum_{j=0}^{k}\Big(\omega_{T,j} + \|c_j\|\Big) &= \sum_{\tau=\tau_0}^{\tau_1}\Big(\omega_{T,k_\tau} + \|c_{k_\tau}\|\Big) + \sum_{\nu=\nu_0,k_\nu\not\in\{k_\tau\}}^{\nu_1}\Big(\omega_{T,k_\nu} + \|c_{k_\nu}\|\Big)\\ &\leq \frac{\kappa_T}{\xi}\sqrt{k+1}\left(1 + \frac{\beta\eta}{\sqrt{\varsigma}}\right) + \frac{\kappa_N}{\xi}\left(1 + \frac{\kappa_T}{\beta\eta}\right),\end{aligned}$$

where we used the inequalities $\tau_1 \leq k_{\tau_1} \leq k$ and $k_{\nu_1} \leq k$. The bound (67) is finally obtained by dividing both sides by $k+1$. □

Remarkably, Theorem 3.12 implies that obtaining an ϵ-approximate first-order critical point, that is an iterate x_k such that $\omega_{T,j} + \|c_j\| \leq \epsilon$, requires at most $\mathcal{O}(\epsilon^{-2})$ iterations of the ADIC algorithm, a complexity which is, in order, the same as that of steepest-descent and Newton's methods on unconstrained problems [5, Theorems 2.2.2 and 3.1.1].

4 Numerical illustration

We now illustrate the behaviour of three variants of the ADIC algorithm on problems from the CUTEst [11] collection as provided in Matlab by S2MPJ [14]. All nonlinear optimization problems in the collection involving general constraints and at most 200 variables were considered, leading to a test set of 312 problems. The algorithmic variants are defined as follows.

- The first variant (ADIC-LP) follows Section 2.1 and computes the dual criticality measure and tangential step using the linear optimization subproblems (18) and (20), respectively.
- The second variant (ADIC-BK) again follows Section 2.1 and computes the dual criticality measure using (18), but then uses the simple formula (21) to define the tangential step.
- The third variant (ADIC-PR) uses the projection approach of Section 2.2, in which the dual criticality is given by (25) and the tangential step is defined by (26).

All three variants have been (trivially) extended to handle general lower and upper bounds on the variables (instead of mere non-negativity constraints), thereby making them applicable to general constrained problems (after transformation of inequality constraints into equalities and the introduction of slack variables, if needed).

Because the variants use different criticality measures, a uniform (external) termination criterion was implemented in order to enforce consistency in the comparison. For all variants, a problem was considered solved as soon as

$$\chi_{T,k} \leq 10^{-4} \quad \text{and} \quad \chi_{N,k} \leq 10^{-5},$$

where $\chi_{T,k}$ and $\chi_{N,k}$ are defined in (17) and (15), respectively. Note that this accomodates the (unfortunate but unavoidable) case where an infeasible minimizer of the equality constraint's violation is found (the bound constraints are satisfied throughout the algorithms). This situation is excluded from our theoretical analysis by AS.8 but does occur in practice. A maximum number of 50000 iterations and a 3600 seconds time limit were imposed. Finally, the algorithmic parameters were chosen as

$$\varsigma = 10^{-5}, \quad \eta = 2, \quad \theta_T = 1, \quad \theta_N = 5, \quad \kappa_n = 10^{-2} \text{ and } \beta = 10^3.$$

We first report on a set of experiments in which the gradients used by the three variants were exact. To summarize the results, we computed three performance statistics: efficiency in terms of iterations, efficiency in terms of CPU time needed and reliability. The latter, which we denote by "Rel" in what follows, is simply computed as the percentage of successfully solved problems. For the two first efficiency statistics, we follow the approach of [18] and compute, for each variant, the area below the relevant curve in a performance profile comparing the three variants, truncated at a "ratio to best performance" equal to 10. The iteration-based statistic is denoted by "Iters" and the CPU-based one by "Time". Values of these statistics should be as close to one as possible. Results are presented in Table 1. The corresponding iteration and CPU performance profiles are shown in Figure 1.

Variant	Iters	Time	Rel
ADIC-LP	0.54	0.57	68.27
ADIC-BK	0.43	0.48	61.54
ADIC-PR	0.61	0.59	71.15

Table 1: Efficiency and reliability statistics for three variants of ADIC on 312 constrained CUTEst problems (noiseless gradients)

Table 1 and Figure 1 indicate that the projection-based ADIC-PR clearly outperforms both ADIC-LP and ADIC-BK, on all 3 statistics. The dominance of ADIC-PR over ADIC-LP in CPU time is however marginal, despite the fact that two linear optimization subproblems must be solved at each iteration of ADIC-LP, against a single projection subproblem for ADIC-PR. The ADIC-BK variant, which only requires the solution of a single linear optimization problem per iteration, remains slower mostly because it typically needs more iterations than other variants per successfully solved problem.

We finally show that our claim that OFFO methods are reliable in the presence of noise is vindicated in practice. To analyze this, we considered the subset of our problems by keeping those

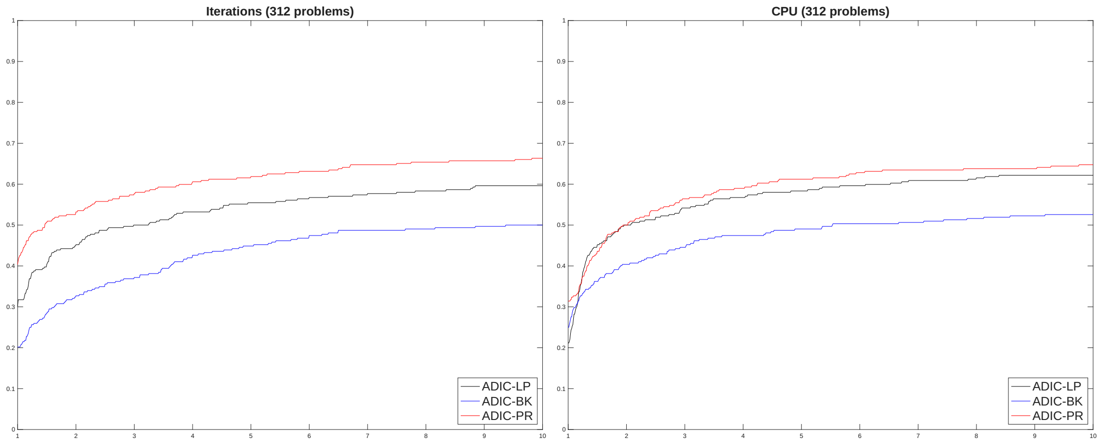

Figure 1: Iteration (left) and CPU time (right) performance profiles for three variants of ADIC on 312 constrained CUTEst problems (noiseless gradients)

Variant	0%	5%	15%	25%	50%
ADIC-LP	81.78	74.55	72.85	73.14	72.73
ADIC-BK	74.09	69.11	67.11	65.77	59.37
ADIC-PR	89.88	77.79	74.82	72.57	67.19

Table 2: Reliability statistics for three variants of ADIC on 247 constrained CUTEst problems for relative random Gaussian noise levels of 0%, 5%, 15%, 25% and 50% on the objective function's gradient

that were solved in the absence of noise by at least one variant, giving a set of 247 test problems. We then added relative random Gaussian[2] noise of increasing magnitude (5%, 15%, 25% and 50%) to the gradients of the objective function and ran each problem 20 times independently, with $\chi_{T,k} \leq 10^{-3}$ and $\chi_{N,k} \leq 10^{-3}$. We then computed the total reliability of our three variants on the resulting 4940 runs for each of the 5 noise levels. The results are presented in Table 2. They show an impressive stability for increasing noise levels, and indicate that, in our view remarkably, ADIC is capable of handling very substantial perturbations of the gradient of the objective function (50% relative noise results in barely one significant digit in the gradient) for a reasonable accuracy requirement. It is also interesting to note that ADIC-LP becomes marginally more reliable than ADIC-PR for large noise levels.

5 Conclusions and perspectives

We have proposed a new OFFO algorithm for the solution of smooth optimization problems, with excellent stability in the presence of noise on the objective function's gradient. This "trust-funnel" algorithm uses adaptive switching between a normal step (reducing constraint violation), and tangential steps (improving dual optimality), the latter being inspired by the AdaGrad-norm algorithm [8, 22] for unconstrained problems. We have also provided a full analysis of the method's worst-case iteration complexity, showing that its global rate of convergence is, for problems with full-rank Jacobians, identical in order to the (optimal) rate of steepest-descent and Newton's method on unconstrained problems. This also provides an evaluation complexity for evaluations of the objective function's gradient, because each iteration requires a single gradient computation. Evaluation complexity for the constraint function and Jacobian is not direct and depends on the algorithm used in the normal step. We have finally conducted illustrative numerical experiments

[2]With zero mean and unit variance.

suggesting that the algorithm's performance and reliability are satisfactory (although admittedly not state-of-the-art) on noiseless problems, but that its reliability in the presence of significant noise on the objective function's gradient is very remarkable.

Many questions remain for further investigation, including the incorporation of second-order information, should it be available, a component-wise version of the algorithm (closer to AdaGrad as opposed to AdaGrad-norm) and a full stochastic complexity analysis. These topics are the subject of ongoing research. Exploiting the independent structure of normal and tangential steps to allow for specific preconditioning of the normal step and relaxing the full-rank assumption on the Jacobians are also of interest.

Acknowledgement

Philippe Toint is grateful for the continued and friendly support of the APO team at Toulouse IRIT (F) and of DIEF at the University of Florence (I).

References

[1] S. Bellavia, S. Gratton, B. Morini, and Ph. L. Toint. Fast stochastic second-order Adagrad for nonconvex bound-constrained optimization. arXiv:2505.06374, 2025.

[2] A. S. Berahas, F. E. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for nonlinear equality constrained stochastic optimization. SIAM Journal on Optimization, 31(2):1352–1379, 2021.

[3] R. S. Burachik, R. N. Gasimov, N. A. Ismayilova, and C. Y. Kaya. On a modified subgradient algorithm for dual problems via sharp augmented Lagrangian. Journal of Global Optimization, 34:55–78, 2006.

[4] R. S. Burachik, A. N. Iusem, and J. G. Melo. A primal dual modified subgradient algorithm with sharp Lagrangian. Journal of Global Optimization, 46:347–361, 2010.

[5] C. Cartis, N. I. M. Gould, and Ph. L. Toint. Evaluation complexity of algorithms for nonconvex optimization. Number 30 in MOS-SIAM Series on Optimization. SIAM, Philadelphia, USA, June 2022.

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

[7] F. Curtis, D. Robinson, and B. Zhou. Sequential quadratic optimization for stochastic optimization with deterministic nonlinear inequality and equality constraints. SIAM Journal on Optimization, 34:3592–3622, 2024.

[8] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(7):2121–2159, 2011.

[9] Y. Fang, S. Na, M. Mahoney, and M. Kolar. Fully stochastic trust-region sequential quadratic programming for equality constrained optimization problems. SIAM Journal on Optimization, 34:2007–2037, 2024.

[10] N. I. M. Gould, D. Orban, and Ph. L. Toint. Numerical methods for large-scale nonlinear optimization. Acta Numerica, 14:299–361, 2005.

[11] N. I. M. Gould, D. Orban, and Ph. L. Toint. CUTEst: a constrained and unconstrained testing environment with safe threads for mathematical optimization. Computational Optimization and Applications, 60:545–557, 2015.

[12] N. I. M. Gould and Ph. L. Toint. Nonlinear programming without a penalty function or a filter. Mathematical Programming A, 122(1):155–196, 2010. See also [?].

[13] S. Gratton, A. Kopaničáková, and Ph. L. Toint. Multilevel objective-function-free optimization with an application to neural networks training. SIAM Journal on Optimization, 33(4):2772–2800, 2023.

[14] S. Gratton and Ph. L. Toint. S2MPJ and cutest optimization problems for Matlab, Python and Julia. Optimization Methods and Software, 40(4):871–903, 2025.

[15] S. Gratton and Ph. L. Toint. A simple first-order algorithm for full-rank equality constrained optimization. arXiv:2510.16390, 2025.

[16] D. M. Himmelblau. Applied Nonlinear Programming. McGraw-Hill, New-York, 1972.

[17] J. Nocedal and S. J. Wright. Numerical Optimization. Series in Operations Research. Spinger Verlag, Heidelberg, 1999.

[18] M. Porcelli and Ph. L. Toint. BFO, a trainable derivative-free brute force optimizer for nonlinear bound-constrained optimization and equilibrium computations with continuous and discrete variables. Transactions of the ACM on Mathematical Software, 44:2007–2037, 2017.

[19] J. L. Romero, D. Fernandez, and G.A. Torres. Enhancing sharp augmented Lagrangian methods with smoothing techniques for nonlinear programming. Journal of Optimization Theory and Applications, 208(23), 2026.

[20] Ph. R. Sampaio and Ph. L. Toint. A derivative-free trust-funnel method for equality-constrained nonlinear optimization. Computational Optimization and Applications, 61(1):25–49, 2015.

[21] Q. Wang, Ch. Peirmarini, Y. Zhu, and F. E. Curtis. Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. arXiv preprint arXiv:2601.11795, 2026.

[22] R. Ward, X. Wu, and L. Bottou. Adagrad stepsizes: sharp convergence over nonconvex landscapes. Journal of Machine Learning Research, 21(1):9047–9076, 2020.

[23] X. Wu, R. Ward, and L. Bottou. WNGRAD: Learn the learning rate in gradient descent. arXiv:1803.02865, 2018.

[24] Y. Yuan. Recent advances in trust region algorithms. Mathematical Programming A, 151(1):249–281, 2015.

[25] C. Zoppke-Donaldson. A Tolerance-Tube Approach to Sequential Quadratic Programming with Applications. PhD thesis, Department of Mathematics and Computer Science, University of Dundee, Dundee, Scotland, UK, 1995.

Proof of Lemma 3.1

We prove the five statements Lemma 3.1 for arbitrary $x, y \geq 0$.

1. That $c(x)$ is Lipschitz continuous with constant $L_c = \kappa_J$ directly follows from AS.4.
2. We have, from AS.3, AS.4 and AS.7 that

$$\begin{aligned} \|J(x)^T c(x) - J(y)^T c(y)\| &= \left\| \Big(J(x) - J(y)\Big)^T c(x) + J(y)^T \Big(c(x) - c(y)\Big) \right\| \\ &\leq (\kappa_c L_J + \kappa_J L_c)\|x - y\| \\ &= (\kappa_c L_J + \kappa_J^2)\|x - y\|, \end{aligned}$$

   yielding $L_{JTc} = \max[1, \kappa_c L_J + \kappa_J^2]$.
3. Define $A(x) = J(x)J(x)^T$. AS.5 then implies that $A(x)$ is (symmetric) positive-definite with smallest eigenvalue bounded below by σ_0^2. As a consequence, $\widehat{\lambda}(x)$ is well defined by (13).
4. Moreover, AS.2 and AS.4 then imply that

$$\|\widehat{\lambda}(x)\| \leq \frac{\kappa_g \kappa_J}{\sigma_0^2},$$

   yielding $\kappa_\lambda = \kappa_g \kappa_J / \sigma_0^2$. We have also, using AS.4 and AS.7, that

$$\|A(x) - A(y)\| \leq \left\| \Big(J(x) - J(y)\Big) J(x)^T + J(y) \Big(J(x) - J(y)\Big)^T \right\| \leq 2\kappa_J L_J \|x - y\|.$$

   from which we deduce that

$$\|A(x)^{-1} - A(y)^{-1}\| = \left\| A(x)^{-1} \Big(A(x) - A(y)\Big) A(y)^{-1} \right\| \leq \frac{2\kappa_J L_J}{\sigma_0^4} \|x - y\|.$$

   We also have that

$$\|J(x)g(x) - J(y)g(y)\| = \left\| \Big(J(x) - J(y)\Big) g(x) + J(y) \Big(g(x) - g(y)\Big) \right\| \leq (\kappa_g L_J + \kappa_J L_g)\|x - y\|.$$

   where we used AS.2, AS.4, AS.6 and AS.7. Therefore, using (13),

$$\begin{aligned} \|\widehat{\lambda}(x) - \widehat{\lambda}(y)\| &= \left\| \Big(A(x)^{-1} - A(y)^{-1}\Big) J(x)g(x) + A(y)^{-1} \Big(J(x)g(x) - J(y)g(y)\Big) \right\| \\ &\leq \frac{1}{\sigma_0^2} \left(\frac{2\kappa_g \kappa_J^2 L_J}{\sigma_0^2} + \kappa_g L_J + \kappa_J L_g \right) \|x - y\|, \end{aligned}$$

   yielding $L_\lambda = \Big((2\kappa_g \kappa_J^2 L_J)/\sigma_0^2 + \kappa_g L_J + \kappa_J L_g\Big)/\sigma_0^2$.
5. Finally, we obtain that, for any λ such that $\|\lambda\| \leq \kappa_\lambda$,

$$\|\nabla_x L(x, \lambda) - \nabla_y L(y, \lambda)\| = \left\| g(x) - g(y) + \Big(J(x) - J(y)\Big)^T \lambda \right\| \leq (L_g + \kappa_\lambda L_J) \|x - y\|,$$

   yielding $L_L = L_g + \kappa_\lambda L_J = L_g + \kappa_g \kappa_J L_J / \sigma_0^2$.