An objective-function-free algorithm for nonconvex stochastic optimization with deterministic equality and inequality constraints

S. Gratton* and Ph. L. Toint†

31 III 2026

Abstract

An algorithm is proposed for solving optimization problems with stochastic objective and deterministic equality and inequality constraints. This algorithm is objective-function-free in the sense that it only uses the objective's gradient and never evaluates the function value. It is based on an adaptive selection of function-decreasing and constraint-improving iterations, the first ones using an Adagrad-type stepsize. When applied to problems with full-rank Jacobian, the combined primal-dual optimality measure is shown to decrease at the rate of $\mathcal{O}(1/\sqrt{k})$, which is identical to the convergence rate of first-order methods in the unconstrained case.

Keywords: General constrained optimization, objective-function-free optimization (OFFO), first-order methods, AdaGrad, evaluation complexity, stochastic analysis.

1 Introduction

This paper proposes an "objective-function-free" algorithm for finding a first-order critical point of the problem

$$\min_{x\in\mathbb{R}^n} F(x) = \mathbb{E}[f(x,\zeta)] \qquad \text{such that} \qquad c(x)=0 \;\text{ and }\; \ell \le x \le u, \tag{1.1}$$

where ζ is a suitably defined random variable. We assume that f is a smooth function of its first argument $x \in \mathbb{R}^n$ and c is a smooth function from $\mathbb{R}^n$ into $\mathbb{R}^m$ $(m \le n)$. The inequality $\ell \le x \le u$ must be understood componentwise and the components of the vectors of lower and upper bounds on the variables ℓ and u need not be finite.

The use of "objective-function-free" (OFFO) algorithms, that is algorithms avoiding the evaluation of the objective function completely, has been central to many applications where noise is present, such as problems arising in deep learning where f is the sum of a very large number of terms and it is only realistic to compute a sample approximation of $f(x)$ and its gradient $\nabla_x f(x)$. The unconstrained case has been extensively studied, starting with the famous stochastic gradient descent [37] and all its modern variants with or without adaptive stepsize ([39, 19, 33, 44] for example). Among them the AdaGrad method [16] has been of particular interest because of its solid and extensive convergence theory (e.g. [44, 30, 15, 3, 26, 41]). Extensions of stochastic gradient methods to the constrained case have also been investigated, so far focused mainly on the case of convex constraints, in particular covering the most common case of bound constraints [4]. Several approaches have been used, such as projected [2, 1] or conditional gradients [18, 7].

[*] Université de Toulouse, INP, IRIT, Toulouse, France. Email: serge.gratton@toulouse-inp.fr. 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"

[†] Université de Toulouse, INP, IRIT, Toulouse, France, and NAXYS, University of Namur, Namur, Belgium. mail: philippe.toint@unamur.be

When constraints are nonconvex, as is necessarily the case if they include nonlinear equalities, things are more complicated. Common practice in deep-learning is to use a simple penalty approach, where a suitably large penalization of the constraint violation is added to the objective function, resulting in an unconstrained problem. This is the approach followed, for instance, in the growing literature on Physics Informed Neural Networks (PINNs) (see [32, 28, 23] and the references therein for instance) but also in other contexts [42]. Its drawback is that a large penalization parameter is sometimes required to obtain a reasonably small constraint violation, which leads to ill-conditioning of the objective function, in turn causing (possibly very) slow convergence of first-order methods.

A second approach is to consider augmented-Lagrangian-based algorithms. Such methods rely on an estimates of the dual variables (Lagrange multipliers) to compute, at each iterate, a step which is approximately tangent to the active constraints. Stochastic adaptations have mostly been inspired by the deterministic iALM method of [45], itself a recent instantiation of the classical Hestenes-Powell augmented Lagrangian technique (see [25, 36, 6, 11] or [12, Section 14.4] and the references therein), and have been proposed and analyzed for instance in [31, 35, 38, 27]. As it turns out, the sensitivity of the multipliers to small variations of the gradients is quite high (we comment on this below). In our view, this makes their practical use in numerical algorithms possibly problematic.

A third approach extends the classical Sequential Quadratic Optimization[1] [34], in which, at each iteration, a local (quadratic) model of the objective function is minimized in the tangent plane to the active constraints. In our noisy context, the main advantage of this technique is that the objective function's gradient's randomness does not affect this plane. A trust-region algorithm of this type is proposed in [13], where a positive-definite approximation is built and minimized in the intersection of the tangent plane and a trust region (and other inequalities). The progress of the algorithm is monitored using a merit function involving a parameter with a somewhat complex update mechanism and behaviour. Almost sure convergence to first-order critical points is proved assuming bounded and Lipschitz continuous gradients, bounded constraints and bounded Lipschitz continuous Jacobians which are required to be full-rank, as well as unbiased gradients with vanishing variance. No complexity estimate is presented. The TR-StoSQP algorithm proposed in [17] is similar in spirit, in that it uses a decomposition of the step within a single trust-region and relies on a parameter-dependent merit function for monitoring progress. Again, almost sure convergence is proved (no complexity) under assumptions very similar to those of [13]. The algorithms proposed in [24, 5] are deterministic and use a trust-funnel approach [21, 14] with linear models. In this method, "constraint improving" and "function decreasing" iterations alternate according to an adaptive switching condition. This technique avoids using any merit function and thus the complexities linked to its parameter. The recent paper [43] follows up on [13] but is also cast in the framework of trust-funnel techniques. It is restricted to equality constraints (as [24]) and uses a first-order adaptive stepsize strategy with momentum based on the "isotropic" version[2] of the Adam method [29]. Under assumptions again similar to those of [13], the authors establish a "perturbed global rate" of convergence where the square of the optimality measure is bounded above by the inverse of the iteration number plus constants which may be made as small as desired by suitable user choices. Although this falls short of a proper complexity analysis [40], in particular because the stepsize has to be chosen proportional to the inverse of the (unknown) gradient's Lipschitz constant, this type of result is standard for Adam-like methods (e.g. [15]).

The method proposed in this theoretical paper is also based on the trust-funnel approach and elaborates on the numerically successful proposal of [5]. Its main contributions are the following.

1. Its fully "componentwise" approach (where each component of the vector of variables is considered with its own stepsize) distinguishes it from [13, 17, 24, 5, 43]. This feature can be important as it is often considered that componentwise first-order methods outperform full-space ones.

[1] Formerly Sequential Quadratic Programming (SQP).

[2] Meaning that the same (adaptive) stepsize is used for all components of the vector of variables.

2. The new method allows for both equality and inequality nonlinear constraints, at variance with [24, 5, 43].
3. It also provides a mechanism to take (possibly approximate and nonconvex) second-order information into account, which is not the case for [13, 5, 43], the first of these references nevertheless allowing strictly convex Hessian approximations.
4. Unlike [13, 17, 24, 43], its complexity is fully analyzed in the stochastic setting, yielding an optimal rate of convergence where the average optimality measure is bounded above by the inverse of the square root of the iteration number, under a "root-mean-square" stochastic assumption.

Section 2 describes the new algorithmic framework and states two of its basic properties. Its global rate of convergence (and hence its worst-case evaluation complexity) is analyzed in Section 3, including a discussion of the stochastic assumptions and how to weaken them. Some perspectives are finally outlined in Section 4.

Notations: The symbol $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$. Expectations and probabilities are denoted using the symbols $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$, respectively. If M is a matrix, $\sigma_{\min}(M)$ denotes its smallest singular value.

2 An algorithm

For a given vector x, we assume that we can compute a random approximation $g(x)$ of the gradient $G(x) = \nabla_x F(x)$, as well use the (exact) Jacobian of the constraints at x

$$J(x) = \nabla c(x) \in \mathbb{R}^{m \times n}.$$

We will use the following "projected gradient" dual first-order optimality measure

$$\Omega_T(x) = \|\mathcal{P}_{x+\mathcal{F}(x)}[x - G(x)] - x\| = \|\mathcal{P}_{\mathcal{F}(x)}[-G(x)]\| \tag{2.1}$$

where

$$\mathcal{F}(x) = \mathcal{T}(x) \cap \mathcal{X}(x) \quad \text{with} \quad \mathcal{T}(x) = \{y \in \mathbb{R}^n \mid J(x)y = 0\} \text{ and } \mathcal{X}(x) = \{y \in \mathbb{R}^n \mid \ell \leq x+y \leq u\},$$

$\mathcal{F}(x)$ being the (convex) tangent feasibility set at x shifted to the origin and $\mathcal{P}_{\mathcal{F}(x)}$ is the orthogonal projection onto this set. The corresponding measure using the approximate gradient $g(x)$ is then given by

$$\omega_T(x) = \|\mathcal{P}_{\mathcal{F}(x)}[-g(x)]\|. \tag{2.2}$$

Note that $\Omega_T(x)$ and $\omega_T(x)$ use projections onto the same set $\mathcal{F}(x)$, and that $\Omega_T(x)$ is a continuous function of x. We also choose a primal (first-order) optimality measure $\omega_N(x)$, here $\omega_N(x) = \|J(x)^T c(x)\|$, but other choices are acceptable. (Since the constraints are deterministic, there is no need to define $\Omega_N(x)$.) As a consequence, x is a true first-order critical point for problem (1.1) if and only if

$$\Omega_T(x) = 0 \quad \text{and} \quad \omega_N(x) = 0, \tag{2.3}$$

while it is an "approximate" first-order critical point if

$$\omega_T(x) = 0 \quad \text{and} \quad \omega_N(x) = 0. \tag{2.4}$$

Our algorithmic framework, dubbed STRADIC for Stochastic Trust-Region AdaGrad with Inequality Constraints, is presented on the following page.

We now briefly comment on some aspects of the STRADIC framework.

Algorithm 2.1: STRADIC

Step 0: Initialization. A starting point x_0 is given, together with constants $\theta_N, \theta_T > 1$, $\beta, \eta, \varsigma, \tau \in (0,1]$. Set $\Gamma_{-1,i} = 0$ for $i \in \{1,\ldots,n\}$ and $k = 0$.

Step 1: Evaluations. Evaluate $c_k = c(x_k)$, $J_k = J(x_k)$ and $g_k = \nabla_x f(x_k, \zeta_k)$. Then compute $\omega_{T,k} = \|d_k\|$, where

$$d_k = \mathcal{P}_{\mathcal{F}_k}[-g_k]. \tag{2.5}$$

where $\mathcal{F}_k = \mathcal{T}(x_k) \cap \mathcal{X}(x_k) \stackrel{\text{def}}{=} \mathcal{T}_k \cap \mathcal{X}_k$. If $\|d_k\| \le \epsilon_D$ and $\omega_{N,k} \le \epsilon_C$, terminate. Otherwise, set

$$\alpha_{k,i} = \frac{\eta}{\sqrt{\Gamma_{k,i} + d_{k,i}^2 + \varsigma}} \quad \text{for } \; i \in \{1,\ldots,n\} \quad \text{and} \quad \Delta_k = \text{diag}\left(\frac{1}{\alpha_{k,i}|d_{k,i}|}\right), \tag{2.6}$$

and compute

$$s_k^L = \mathcal{P}_{\mathcal{F}_k \cap \mathcal{S}_k}[-g_k] \quad \text{where} \quad \mathcal{S}_k = \{y \in \mathbb{R}^n \mid \|\Delta_k y\|_\infty \le 1\}. \tag{2.7}$$

Step 2: Normal step. Except possibly if

$$\omega_{N,k} \le \beta \|s_k^L\|_\infty, \tag{2.8}$$

set $x_k^+ = x_k + s_{N,k}$ where the step $s_{N,k}$ is such that

$$s_{N,k} \in \mathcal{X}_k, \quad \text{and} \quad \|s_{N,k}\|_\infty \le \theta_N \, \omega_{N,k}, \tag{2.9}$$

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

$$\tfrac{1}{2}\|c(x_k + s_{N,k})\|^2 \le \tfrac{1}{2}\|c_k\|^2 - \kappa_n \, \omega_{N,k}^2. \tag{2.10}$$

If (2.8) holds and $s_{N,k}$ was not computed, set $x_k^+ = x_k$.

Step 3: Tangential step. If (2.8) holds, select B_k a symmetric approximation of $\nabla_x^2 f(x_k)$, compute a "Cauchy step"

$$s_k^C = \left(\operatorname*{argmin}_{t \in [0,1]} m_k(t s_k^L)\right) s_k^L \quad \text{where} \quad m_k(s) = g_k^T s + \frac{1}{2} s^T B_k s, \tag{2.11}$$

choose a step $s_{T,k}$ such that

$$s_{T,k} \in \mathcal{F}_k \cap \mathcal{S}_k, \quad \|s_{T,k}\| \ge \|s_k^C\| \quad \text{and} \quad m_k(s_{T,k}) \le \tau m_k(s_k^C). \tag{2.12}$$

and set $x_{k+1} = x_k^+ + s_{T,k}$ and

$$\Gamma_{k+1,i} = \Gamma_{k,i} + d_{k,i}^2 \quad (i \in \{1,\ldots,n\}) \tag{2.13}$$

Otherwise (that is if (2.8) fails), set $x_{k+1} = x_k^+$ and $\Gamma_{k+1} = \Gamma_k$.

Step 4: Loop. Increment k by one and go to Step 2.

1. We have not covered possible implementations of the normal step in detail, but many are possible. As a matter of fact, most algorithms for bound-constrained nonlinear least-squares are acceptable. It is for instance argued in detail in [5] that a few (usually one) steps of a trust-region based Gauss-Newton method are enough to guarantee (2.10), but this is not limitative. Also note that the computation of the normal step is totally independent of the rest of the algorithm. As a consequence, constraint-based preconditioners may be applied without having to take the objective function into account.

2. For notational consistency, we define $s_{N,k} = 0$ at iteration where it is not computed, so that $x_k^+ = x_k + s_{N,k}$ whether or not $s_{N,k}$ was computed.

3. As written above, the definition of Δ_k assumes that none of the $d_{k,i}$ is zero. Should $d_{k,i}$ be zero, we define $\Delta_{k,i,i} = 1$ (it turns out that the chosen constant is irrelevant). Also note that the second part of (2.6), the definition of $\mathcal{S}_k$ and (2.12) ensure that, in all cases,

$$|s_{T,k,i}| \leq \alpha_{k,i}|d_{k,i}| \quad (i \in \{1,\ldots,n\}). \tag{2.14}$$

4. Our formulation remains deliberately vague about how to choose the approximate Hessian B_k and covers a number of possible choices such as Barzilai-Borwein, finite differences or (limited-memory) quasi-Newton approximations. If available, using the true Hessian $\nabla_x^2 f(x_k)$ is also possible.

5. Observe that the choice $s_{T,k} = s_k^L$ is always acceptable. This is useful if computing B_k is deemed too expensive and $B_k = 0$ is chosen at each iteration. When $s_{T,k} = s_k^L$, STRADIC reduces to a strictly first-order method for constrained problems, in the spirit of [5] albeit using a componentwise approach.

6. In the same vein, the choice $s_{T,k} = s_k^C$ as defined by (2.11) is also fully acceptable for (2.12), and has the advantage of limiting the computation involved to take second-order information into account to a single matrix-vector product (for computing γ_k in (2.24), at tangential iterations only). Also note that s_k^C is the first iterate of a Krylov method for minimizing $m_k(s)$ in $\mathcal{F}_k \cap \mathcal{S}_k$, should further minimization of this quadratic model be desired.

For our subsequent analysis, we need to identify, at iteration k, the index of the smallest stepsize $\alpha_{k,i}$. we therefore denote

$$\mu(k) = \operatorname*{argmin}_{i\in\{1,\ldots,n\}} \alpha_{k,i} = \operatorname*{argmax}_{i\in\{1,\ldots,n\}} \sqrt{\Gamma_{k,i} + d_{k,i}^2 + \varsigma}. \tag{2.15}$$

Because of (2.6), we also have that, for $k \geq 0$ and $i \in \{1,\ldots,n\}$,

$$\alpha_{k,i}|d_{k,i}| < 1 \quad \text{and thus} \quad \Delta_{k,i,i} > 1. \tag{2.16}$$

Lemma 2.1 Suppose that a tangential step $s_{T,k}$ is computed at iteration k. Then

$$g_k^T s_k^L \leq -\frac{3}{4}\min[\alpha_{k,\mu(k)}, 1]\,\|d_k\|^2, \tag{2.17}$$

and

$$|g_k^T s_k^L| \geq \|s_k^L\|^2. \tag{2.18}$$

Proof. Since $0 \in \mathcal{F}_k \cap \mathcal{S}_k$, the definition of s_k^L in (2.7) as a projection implies that

$$\big(-g_k - s_k^L\big)^T\big(0 - s_k^L\big) \leq 0,$$

and thus

$$g_k^T s_k^L \le -\|s_k^L\|^2, \tag{2.19}$$

yielding (2.18). Similarly, the definition of d_k in (2.5) gives that

$$g_k^T d_k \le -\|d_k\|^2. \tag{2.20}$$

Consider now the vector

$$y_k = \min[\alpha_{k,\mu(k)}, 1]\, d_k \tag{2.21}$$

Then, for each $i \in \{1, \ldots, n\}$,

$$\left| \frac{\min[\alpha_{k,\mu(k)}, 1]\, d_{k,i}}{\alpha_{k,i}\, d_{k,i}} \right| \le 1.$$

and $y_k \in \mathcal{S}_k$. Moreover, since both d_k and the origin belong to the convex set $\mathcal{F}_k$, so does y_k. Thus $y_k \in \mathcal{F}_k \cap \mathcal{S}_k$. Since $s_k^L = \mathcal{P}_{\mathcal{F}_k \cap \mathcal{S}_k}[-g_k]$ by (2.7), we then derive that

$$(-g_k^T - s_k^L)^T(y_k - s_k^L) \le 0,$$

from which we deduce that

$$-g_k^T(y_k - s_{T,k}) \le (s_k^L)^T(y_k - s_{T,k}) = (s_k^L)^T y_k - \|s_k^L\|^2 = -\|s_k^L - \tfrac{1}{2} y_k\|^2 + \tfrac{1}{4}\|y_k\|^2 \le \tfrac{1}{4}\|y_k\|^2.$$

Substituting (2.21) in this inequality then gives that

$$-g_k^T(y_k - s_k^L) \le \tfrac{1}{4} \min[\alpha_{k,\mu(k)}, 1]^2\, \|d_k\|^2 \le \tfrac{1}{4} \min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2$$

Combining this inequality with (2.20) and (2.21), we finally obtain that

$$g^T s_k^L = g_k^T y_k - g_k^T(y_k - s_{T,k}) \le -\min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2 + \tfrac{1}{4} \min[\alpha_{k,\mu(k)}, 1])\|d_k\|^2$$

yielding (2.17). □

We conclude our brief description of the algorithm's basic properties by showing that our requirement to compute a "Cauchy step" in (2.11) implies two simple but useful bounds.

Lemma 2.2 Suppose that a tangential step $s_{T,k}$ is computed at iteration k. Then

$$\|s_{T,k}\| \ge \|s_k^C\| \ge \frac{\beta}{\max[1, \|B_k\|]} \|s_k^L\| \tag{2.22}$$

and

$$g_k^T s_{T,k} \le -\frac{3\tau}{4 \max[1, 2\|B_k\|]} \min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2 + \frac{1}{2}\|B_k\| \sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2. \tag{2.23}$$

Proof. One verifies that the Cauchy step (2.11) is given by

$$s_k^C = \gamma_k s_k^L \quad \text{where} \quad \gamma_k = \begin{cases} \min\left[1, \dfrac{-g_k^T s_k^L}{(s_k^L)^T B_k s_k^L}\right] & \text{if} \quad (s_k^L)^T B_k s_k^L > 0 \\ 1 & \text{otherwise,} \end{cases} \tag{2.24}$$

and one first observes that, when $(s_k^L)^T B_k s_k^L > 0$, (2.17) and (2.18) imply that

$$\gamma_k = \frac{-g_k^T s_k^L}{\|s_k^L\|^2} \frac{\|s_k^L\|^2}{(s_k^L)^T B_k s_k^L} = \frac{|g_k^T s_k^L|}{\|s_k^L\|^2} \frac{\|s_k^L\|^2}{(s_k^L)^T B_k s_k^L} \ge \frac{1}{\|B_k\|},$$

while $\gamma_k = 1$ otherwise. Thus (2.22) follows from the middle part of (2.12). Now note that $s_k^C \in \mathcal{F}_k \cap \mathcal{S}_k$ because both s_k^L and the origin belong to this convex set. Suppose first that $\gamma_k < 1$, implying that $(s_k^L)^T B_k s_k^L > 0$. Then, using Lemma 2.1,

$$\begin{aligned} g_k^T s_k^C + \frac{1}{2}(s_k^C)^T B_k s_k^C &= -\frac{(g_k^T s_k^L)^2}{2(s_k^L)^T B_k s_k^L} \\ &\leq -\frac{3}{8\|B_k\|} \min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2 \\ &\leq -\frac{3}{4\max[1, 2\|B_k\|]} \min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2 \end{aligned}$$

If $\gamma_k = 1$, $s_k^C = s_k^L$ and the last inequality directly results from (2.17). Hence,

$$\begin{aligned} g_k^T s_{T,k} &= g_k^T s_{T,k} + \frac{1}{2} s_{T,k}^T B_k s_{T,k} - \frac{1}{2} s_{T,k}^T B_k s_{T,k} \\ &\leq \tau(g_k^T s_k^L + \frac{1}{2}(s_k^L)^T B_k s_k^L) + \frac{1}{2}\|B_k\|\, \|s_{T,k}\|^2 \\ &\leq -\frac{3\tau}{4\max[1, 2\|B_k\|]} \min[\alpha_{k,\mu(k)}, 1]\, \|d_k\|^2 + \frac{1}{2}\|B_k\| \sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2, \end{aligned}$$

proving (2.23). □

3 Complexity analysis

In our subsequent analysis, we need to distinguish between "normal" and "tangential" iterations. We denote by $\{k_\tau\} \subseteq \{k\}$ the index subsequence of iterations at which (2.8) holds, while $\{k_\nu\} \subseteq \{k\}$ is the index subsequence of iterations at (2.8) fails. Thus

$$\{0, \ldots, k\} = \{k_{\tau_0}, \ldots, k_{\tau_1}\} \cup \{k_{\nu_0}, \ldots, k_{\nu_1}\} \quad \text{with} \quad \min[k_{\tau_0}, k_{\nu_0}] = 0 \;\text{ and }\; \max[k_{\tau_1}, k_{\nu_1}] = k.$$

The definition of the algorithm implies that a tangential step and (possibly) a normal step are computed for $k \in \{k_\tau\}$ and a normal step (but no tangential step) is computed for $k \in \{k_\nu\}$. We will also consider the "sharp augmented Lagrangian" Lyapunov function introduced in [5] (whose value is hopefully decreasing as the iterations progress), which is given by

$$\psi(x, \lambda) \stackrel{\text{def}}{=} L(x, \lambda) + \rho\|c(x)\|, \tag{3.1}$$

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

$$L(x, \lambda) = f(x) + \lambda^T c(x), \tag{3.2}$$

for some multiplier $\lambda \in \mathbb{R}^m$. Of particular interest in our argument is the least-squares Lagrange multiplier $\widehat{\lambda}(x)$ defined by

$$\big(J(x)J(x)^T\big)\, \widehat{\lambda}(x) \;=\; -\,J(x)\, g(x) \tag{3.3}$$

when the Jacobian $J(x)$ has full rank. We also use the abbreviation

$$\psi(x) \stackrel{\text{def}}{=} \psi_\rho\big(x, \widehat{\lambda}(x)\big). \tag{3.4}$$

3.1 Assumptions

Our analysis uses the following assumptions.

AS.1: f and c are continuously differentiable on $\mathbb{R}^n$.

AS.2: There exists a constant $f_{\rm low}$ such that, for all $\ell \le x \le u$ and all ζ, $f(x,\zeta) \ge f_{\rm low}$.

AS.3: There exists a constant $\kappa_g \ge 1$ such that $\|g(x)\| \le \kappa_g$ for all $\ell \le x \le u$.

AS.4: There exists a constant $\kappa_c > 1$ such that $\|c(x)\| \le \kappa_c$ for all $\ell \le x \le u$.

AS.5: There exists a constant $\kappa_J > 1$ such that $\|J(x)\| \le \kappa_J$ for all $\ell \le x \le u$.

AS.6: There exists a constant $\sigma_0 \in (0,1]$ such that $\sigma_{\min}(J(x)) \ge \sigma_0$ for all $\ell \le x \le u$.

AS.7: The gradient $g(x)$ is globally Lipschitz continuous (with constant L_g).

AS.8: The Jacobian $J(x)$ is globally Lipschitz continuous (with constant L_J).

AS.9: There exists a constant $\kappa_B \ge 1$ such that $\|B_k\| \le \kappa_B$ for all $k \ge 0$.

AS.10: There exists a constant $\xi \in (0,1]$ such that, for all $k \ge 0$, $\omega_{N,k} \ge \xi \|c_k\|$.

AS.1–AS.9 are identical to those used in [4, 13] and other papers mentioned in the introduction., and AS.2–AS.5 automatically hold in the common occurrence where the iterates remain in a bounded set. We immediately note that, for all $k \ge 0$,

$$\|g_k + J_k^T \widehat{\lambda}_k\| = \|P_{\mathcal{T}(x_k)}(g_k)\| \le \kappa_g. \tag{3.5}$$

because of AS.3 and the contractive nature of the projection.

Although the normal step is designed to reduce constraint violation, it does not guarantee that the sequence $\{\|c_k\|\}$ converges to zero. Without further assumptions, the iterates may end up at an infeasible local minimizer $x_{\rm loc}$ of $\frac{1}{2}\|c(x)\|^2$. Such a situation 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 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 constrained global optimization method to minimize $\frac{1}{2}\|c(x)\|^2$ subject to the bounds, or restricting the class of problems under consideration. As in [13, 17, 5, 43] , we follow here the second approach: AS.6 precludes the first cause of the problem, ensuring that $J(x_{\rm loc})^T c(x_{\rm loc}) = 0$ implies $c(x_{\rm loc}) = 0$, while AS.10 prevents the criticality measure $\omega_{N,(}x_{\rm loc})$ to vanish at an infeasible local minimizer.

Because of the random nature of the gradient estimator, the STRADIC algorithm generates a random process where, for a given iterate x_k, the oracle computes the gradient approximation $g_k = \nabla_x f(x_k, \zeta_k)$ where ζ_k is a random variable (whose distribution may depend on x_k), with probability space $(\Sigma, \mathcal{W}, \mathbb{P})$. The expectation conditioned to knowing $g_0, \ldots, g_{k-1}$ will be denoted by the symbol $\mathbb{E}_k[\cdot]$. Note that $\widehat{\lambda}(x_k)$ is measurable with respect to the past. We will use the abbreviations

$$\mathbb{E}_k^\tau[\cdot] \stackrel{\rm def}{=} \mathbb{E}_k[\cdot | k \in \{k_\tau\}] \quad \text{and} \quad \mathbb{E}_k^\nu[\cdot] \stackrel{\rm def}{=} \mathbb{E}_k[\cdot | k \in \{k_\nu\}] \, .$$

We next require a "root-mean-square error" condition [4] along the tangential step $s_{T,k}$ given by

AS.11: There exists a constant $\kappa_{\rm dir} > 0$ such that, for all $k \ge 0$,

$$\mathbb{E}_k^\tau\left[|(G_{T,k} - g_{T,k})^T s_{T,k}|\right] \le \frac{\kappa_{\rm dir}}{2} \mathbb{E}_k^\tau\left[\|s_{T,k}\|^2\right] .$$

This condition only considers the gradient error along the step, which is the minimum that can be required given that (2.12) only enforces a very loose relation between $s_{T,k}$ and the gradient. As we will see below, AS.11 is strong enough to ensure an optimal rate of convergence of our algorithm for generally constrained optimization, comparable to that of standard first-order methods (like steepest descent) on unconstrained problems. Note that it is weaker than requiring the maybe more natural total variance condition

$$\mathbb{E}_k^\tau\left[\|G_{T,k} - g_{T,k}\|^2\right] \le \kappa_{\rm dir2}^2 \mathbb{E}_k^\tau\left[\|s_{T,k}\|^2\right] \tag{3.6}$$

since the latter and Cauchy-Schartz inequality ensure that

$$\mathbb{E}_k^\tau\left[|(G_{T,k}-g_{T,k})^T s_{T,k}|\right] \leq \sqrt{\mathbb{E}_k^\tau[\|G_{T,k}-g_{T,k}\|^2]}\sqrt{\mathbb{E}_k^\tau[\|s_{T,k}\|^2]} \leq \kappa_{\rm dir2}\mathbb{E}_k^\tau\left[\|s_{T,k}\|^2\right].$$

Note that, (3.6) is itself akin to the "strong growth condition"

$$\mathbb{E}_k^\tau\left[|(G_k-g_k)^T s_{T,k}|\right] \leq \kappa_{\rm dir2}\mathbb{E}_k^\tau\left[\|g_k\|^2\right]. \tag{3.7}$$

used in [41] in their so far sharpest analysis of (unconstrained) AdaGrad. Observe also that AS.11 only applies to tangential iterations. Additional discussion about AS.11, its relation to and how to weaken it is provided in Section 3.8.
Before moving on, we note that

- AS.1 and AS.5 imply that c is Lipschitz continuous (with constant L_c),
- AS.1, AS.7 and AS.8 imply that $\nabla_x L(x,\lambda)$ is Lipschitz continuous (with constant L_L),
- AS.6 implies that $\widehat{\lambda}(x)$ is well-defined for all x, and thus that the norm of the projected gradient in the nullspace of J is also bounded (that is $\|g_k + J_k^T\widehat{\lambda}(x_k)\| \leq \kappa_g$) and that $\|d_k\| \leq \kappa_g$,
- AS.1, AS.3, AS.6, AS.7 and AS.8 ensure that $\widehat{\lambda}(x)$ is bounded (by κ_λ) and Lipschitz continuous (with constant L_λ).

Detailed proofs of these statements are available in [5, Lemma 3.1]. As turns out, Ł$_\lambda$ is proportional to $\kappa_g\kappa_J^2/\sigma_{\min}^2$, which justifies the comment made in Section 1 regarding the sensitivity of the Lagrange multipliers.

3.2 Normal steps

Our analysis hinges on the fact that first-order descent can be shown on the Lyapounov $\psi(x)$, both for tangential and normal steps, despite the fact that neither $\widehat{\lambda}(x_k)$ or ρ (which we still need to define) appears in the algorithm. We start by considering normal steps.

Lemma 3.1 Suppose that AS.6 and AS.10 hold. Then, if $c_k^+ = c(x_k^+)$ and $s_{N,k} \neq 0$,

$$\|c_k^+\| - \|c_k\| \leq -\frac{\kappa_n\xi}{2}\,\omega_{N,k}. \tag{3.8}$$

Proof. We have from (2.10) that $\|c_k^+\| \leq \|c_k\|$. Then,

$$2\|c_k\|(\|c_k\| - \|c_k^+\|) \geq (\|c_k\| + \|c_k^+\|)(\|c_k\| - \|c_{k+1}\|) = \|c_k\|^2 - \|c_k^+\|^2,$$

and therefore, using (2.10) and AS.10, that

$$\|c_k^+\| - \|c_k\| \leq -\frac{\kappa_n\omega_{N,k}^2}{2\|c_k\|} \leq -\frac{\kappa_n\xi}{2}\,\omega_{N,k}$$

□

Lemma 3.2 Suppose that AS.4–AS.10 hold and that a normal step is used at iteration. Define

$$\rho = \frac{2}{\kappa_n \xi}\left[(\kappa_g + \kappa_c\, L_\lambda)\,\theta_N\,\sqrt{n} + \kappa_J\kappa_c\left(\frac{L_L}{2} + L_\lambda L_c\right)\theta_N^2\, n + \eta\right] \tag{3.9}$$

Then we have that

$$\psi(x_k^+) - \psi(x_k) \le -\eta\,\omega_{N,k}. \tag{3.10}$$

Proof. We have that

$$\psi(x_k^+) - \psi(x_k) = \underbrace{\psi(x_k^+,\widehat{\lambda}_k) - \psi(x_k,\widehat{\lambda}_k)}_{\Delta_x} + \underbrace{\psi(x_k^+,\widehat{\lambda}_k^+) - \psi(x_k^+,\widehat{\lambda}_k)}_{\Delta_\lambda}. \tag{3.11}$$

where $\widehat{\lambda}_k = \widehat{\lambda}(x_k)$ and $\widehat{\lambda}_k^+ = \widehat{\lambda}(x_k^+)$. Now consider Δ_x and Δ_λ separately.

Using (3.4), the Lipschitz continuity of $\nabla_x\psi(x,\widehat{\lambda})$ (ρ is fixed in (3.9)) and (3.8), we obtain that

$$\begin{aligned}\Delta_x &= \psi(x_k^+,\widehat{\lambda}_k) - \psi(x_k,\widehat{\lambda}_k)\\ &= L(x_k^+,\widehat{\lambda}_k) - L(x_k,\widehat{\lambda}_k) + \rho\big(\|c_k^+\| - \|c_k\|\big)\\ &\le (\nabla_x L(x_k,\widehat{\lambda}_k)^T s_{N,k} + \frac{L_L}{2}\|s_{N,k}\|^2 - \tfrac{1}{2}\rho\kappa_n\xi\,\omega_{N,k}.\end{aligned}$$

We now invoke the Cauchy-Schwarz inequality, (3.5) and (3.14) to deduce that

$$\begin{aligned}\Delta_x &\le \|\nabla_x L(x_k,\widehat{\lambda}_k)\|\,\|s_{N,k}\| - \rho\frac{\kappa_n\xi}{2}\,\omega_{N,k} + \frac{L_L}{2}\|s_{N,k}\|^2\\ &\le \|g_k + J_k^T\widehat{\lambda}_k\|\,\|s_{N,k}\| - \rho\frac{\kappa_n\xi}{2}\,\omega_{N,k} + \frac{L_L}{2}\|s_{N,k}\|^2\\ &\le \kappa_g\|s_{N,k}\| - \rho\frac{\kappa_n\xi}{2}\,\omega_{N,k} + \frac{L_L}{2}\|s_{N,k}\|^2.\end{aligned} \tag{3.12}$$

Using now the definition of Δ_λ in (3.11), the Lipschitz continuity of $\widehat{\lambda}$ and c and AS.4 then yields that

$$\begin{aligned}\Delta_\lambda &= \psi(x_k^+,\widehat{\lambda}_k^+) - \psi(x_k^+,\widehat{\lambda}_k)\\ &\le (\|c_k\| + \|c_k^+ - c_k\|)\,\|\widehat{\lambda}_k^+ - \widehat{\lambda}_k\|\\ &\le L_\lambda\,\|s_{N,k}\|\,\|c_k\| + L_\lambda L_c\|s_{N,k}\|^2\\ &\le L_\lambda\,\kappa_c\|s_{N,k}\| + L_\lambda L_c\|s_{N,k}\|^2.\end{aligned} \tag{3.13}$$

We also observe that, because of (2.12),

$$\|s_{N,k}\| \le \sqrt{n}\|s_{N,k}\|_\infty \le \theta_N\sqrt{n}\,\omega_{N,k}. \tag{3.14}$$

and thus, summing (3.12) and (3.13) and taking into account that $\omega_{N,k} \le \kappa_J\kappa_c$ because of AS.4 and AS.5, that

$$\begin{aligned}&\psi(x_k^+) - \psi(x_k)\\ &\qquad\le -\rho\frac{\kappa_n\xi}{2}\,\omega_{N,k} + \kappa_g\|s_{N,k}\| + L_\lambda\,\kappa_c\|s_{N,k}\| + \left(\frac{L_L}{2} + L_\lambda L_c\right)\|s_{N,k}\|^2\\ &\qquad\le -\rho\frac{\kappa_n\xi}{2}\omega_{N,k} + (\kappa_g + L_\lambda\,\kappa_c)\,\theta_N\sqrt{n}\,\omega_{N,k} + \kappa_J\kappa_c\left(\frac{L_L}{2} + L_\lambda L_c\right)\theta_N^2\,n\,\omega_{N,k}.\end{aligned}$$

The bound (3.10) then follows from (3.9). □

3.3 Tangential steps

Lemma 3.3 Suppose that AS.5–AS.11 hold and that a tangential step is taken at iteration k (i.e. $k \in \{k_\tau\}$). Then

$$\mathbb{E}_k^\tau[\psi(x_{k+1})] - \psi(x_k^+) \le -\frac{3}{8\kappa_B}\mathbb{E}_k^\tau\big[\min[\alpha_{k,\mu(k)},1]\,\|d_k\|^2\big] + \kappa_{\tan}\,\mathbb{E}_k^\tau\left[\sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2\right]. \qquad (3.15)$$

where

$$\kappa_{\tan} = \kappa_{\rm dir} + \frac{\beta\theta_N}{\kappa_B}(L_L+\kappa_J L_\lambda) + \rho\left(\frac{\beta\theta_N}{\kappa_B}\,L_J + \frac{L_c}{2}\right) + \frac{\kappa_B}{2} + \frac{\beta L_\lambda}{\xi} + L_\lambda L_c. \qquad (3.16)$$

Proof. We again use the decomposition

$$\begin{aligned}\mathbb{E}_k^\tau[\psi(x_{k+1})] - \psi(x_k^+) &= \underbrace{\mathbb{E}_k^\tau\Big[\psi_\rho(x_{k+1},\widehat{\lambda}_k)\Big] - \psi_\rho(x_k^+,\widehat{\lambda}_k^+)}_{\Delta_x} \\ &\quad + \underbrace{\mathbb{E}_k^\tau\Big[\psi_\rho(x_{k+1},\widehat{\lambda}_{k+1}) - \psi_\rho(x_{k+1},\widehat{\lambda}_k^+))\Big]}_{\Delta_\lambda}\end{aligned} \qquad (3.17)$$

and consider Δ_x and Δ_λ separately. The Lipschitz continuity of $\nabla_x L(x,\lambda)$ gives that

$$\Delta_x = -\mathbb{E}_k^\tau\left[\nabla_x\psi(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k} + \frac{L_L}{2}\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big]\right] + \rho(\mathbb{E}_k^\tau[\|c_{k+1}\|] - \|c_k^+\|).$$

Successively using the Lipschitz continuity of $\nabla_x L(x,\lambda)$, that of $\widehat{\lambda}$ and the identity $J_k s_{T,k} = 0$, we now verify that

$$\begin{aligned}\nabla_x L(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k} &= \Big(\nabla_x L(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k} - \nabla_x L(x_k,\widehat{\lambda}_k^+)^T s_{T,k}\Big) \\ &\quad + \Big(\nabla_x L(x_k,\widehat{\lambda}_k^+)^T s_{T,k} - \nabla_x L(x_k,\widehat{\lambda}_k)^T s_{T,k}\Big) + g_k^T s_{T,k} + \widehat{\lambda}_k^T J_k s_{T,k} \\ &= \Big(\nabla_x L(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k} - \nabla_x L(x_k,\widehat{\lambda}_k^+)^T s_{T,k}\Big) \\ &\quad + \Big((\widehat{\lambda}_k^+)^T J_k - \widehat{\lambda}_k^T J_k\Big)^T s_{T,k} + g_k^T s_{T,k} + \widehat{\lambda}_k^T J_k s_{T,k} \\ &\le \Big(L_L + \kappa_J L_\lambda\Big)\|s_{N,k}\|\,\|s_{T,k}\| + g_k^T s_{T,k}\end{aligned}$$

But (2.8) must hold on iteration where the tangential step is computed, and hence, using also (2.9), (2.22) and AS.9,

$$\|s_{N,k}\|_\infty \le \theta_N\omega_{N,k} \le \beta\theta_N\|s_k^L\|_\infty \le \beta\theta_N\|s_k^L\| \le \frac{\beta\theta_N}{\kappa_B}\|s_{T,k}\|. \qquad (3.18)$$

As a consequence,

$$\nabla_x L(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k} \le g_k^T s_{T,k} + \frac{\beta\,\theta_N}{\kappa_B}\Big(L_L + \kappa_J L_\lambda\Big)\|s_{T,k}\|^2.$$

Hence, using (2.23), AS.9, AS.11 and (2.14),

$$\begin{aligned}
\mathbb{E}_k^\tau\Big[\nabla_x\psi(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k}\Big] &\le \mathbb{E}_k^\tau\big[g_k^T s_{T,k}\big] + \mathbb{E}_k^\tau\big[|(G_k^T - g_k^T)s_{T,k}|\big] + \frac{\beta\,\theta_N}{\kappa_B}(L_L+\kappa_J L_\lambda)\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big] \\
&\le -\frac{3}{8\kappa_B}\mathbb{E}_k^\tau\big[\min[\alpha_{k,\mu(k),1]}\,\|d_k\|^2\big] + \big(\kappa_{\rm dir} + \frac{\beta\,\theta_N}{\kappa_B}(L_L+\kappa_J L_\lambda)\big)\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big] + \frac{\kappa_B}{2}\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big] \\
&\le -\frac{3}{8\kappa_B}\mathbb{E}_k^\tau\big[\min[\alpha_{k,\mu(k)},1]\,\|d_k\|^2\big] + \left(\kappa_{\rm dir} + \frac{\beta\,\theta_N}{\kappa_B}(L_L+\kappa_J L_\lambda) + \frac{\kappa_B}{2}\right)\mathbb{E}_k^\tau\left[\sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2\right]
\end{aligned} \tag{3.19}$$

Moreover, the Lipschitz continuity of c and J and the identity $J_k s_{T,k} = 0$ ensure that

$$\begin{aligned}
\mathbb{E}_k^\tau[\|c(x_{k+1})\|] &= \mathbb{E}_k^\tau\big[\|c(x_k^+) + J_k s_{T,k} + (J_k^+ - J_k)s_{T,k} + r_1\|\big] \\
&= \mathbb{E}_k^\tau[\|c(x_k)+r_1\|] + L_J\mathbb{E}_k^\tau[\|s_{N,k}\|\,\|s_{T,k}\|] \\
&\le \|c(x_k)\| + L_J\mathbb{E}_k^\tau[\|s_{N,k}\|\,\|s_{T,k}\|] + \frac{L_c}{2}\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big]\,.
\end{aligned} \tag{3.20}$$

Hence, combining (2.14), (3.19), (3.20) and (3.18), we deduce that

$$\begin{aligned}
\Delta_x \le & -\frac{3}{8\kappa_B}\mathbb{E}_k^\tau\big[\min[\alpha_{k,\mu(k),1]}\,\|d_k\|^2\big] \\
& + \left[\kappa_{\rm dir} + \frac{\beta\,\theta_N}{\kappa_B}(L_L+\kappa_J L_\lambda) + \frac{\kappa_B}{2} + \rho\left(\frac{\beta\,\theta_N}{\kappa_B}\,L_J + \frac{L_c}{2}\right)\right]\mathbb{E}_k^\tau\left[\sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2\right].
\end{aligned} \tag{3.21}$$

We may now use the Lipschitz continuity of $\widehat{\lambda}$ and c to deduce that

$$\begin{aligned}
\Delta_\lambda &= \mathbb{E}_k^\tau\Big[c_{k+1}^T\big(\widehat{\lambda}_{k+1} - \widehat{\lambda}_k^+\big)\Big] \\
&= \mathbb{E}_k^\tau\Big[(c_{k+1} - c_k^+)^T\big(\widehat{\lambda}_{k+1} - \widehat{\lambda}_k^+)\big)\Big] + \mathbb{E}_k^\tau\Big[c_k^{+T}\big(\widehat{\lambda}_{k+1} - \widehat{\lambda}_k^+\big)\Big] \\
&\le \mathbb{E}_k^\tau\Big[\|c_{k+1} - c_k^+\|\,\|\widehat{\lambda}_{k+1} - \widehat{\lambda}_k^+\|\Big] + \mathbb{E}_k^\tau\Big[\|c_k\|\,\|\widehat{\lambda}_{k+1} - \widehat{\lambda}_k^+\|\Big] \\
&\le L_\lambda\mathbb{E}_k^\tau[\|c_k\|\|s_{T,k}\|] + L_\lambda L_c\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big]
\end{aligned} \tag{3.22}$$

Taking into account the fact that, for $k\in\{k_\tau\}$, (2.8) and AS.10 give that $\|c_k^+\| \le \omega_{N,k}/\xi \le \beta\|s_{T,k}\|_\infty/\xi$, we obtain that

$$\Delta_\lambda \le \left(\frac{\beta L_\lambda}{\xi} + L_\lambda L_c\right)\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big] \le \left(\frac{\beta L_\lambda}{\xi} + L_\lambda L_c\right)\mathbb{E}_k^\tau\left[\sum_{i=1}^n \alpha_{k,i}^2 d_{k,i}^2\right], \tag{3.23}$$

where we used (2.14) to obtain the last inequality. The bound (3.15) the follows by summing the bounds on Δ_x (in (3.21) and Δ_λ (in (3.23)) and using the definition of $\kappa_{\rm tan}$ in (3.16). □

Note that (3.16) reduces to

$$\kappa_{\rm tan} = \kappa_{\rm dir} + \rho\frac{L_c}{2} + \frac{\beta L_\lambda}{\xi} + L_\lambda L_c + \frac{\kappa_B}{2}$$

if no normal step is computed.

Lemma 3.4 If we denote, for $i \in \{1,\ldots,n\}$,

$$\Gamma_{k_{\tau_0},i} = 0, \quad \Gamma_{k_{\tau+1},i} = \Gamma_{k_\tau,i} + |d_{k_\tau,i}|^2 \text{ and } \alpha_{k_\tau,i} = \frac{\eta}{\sqrt{\Gamma_{k_\tau,i} + |d_{k_\tau,i}|^2 + \varsigma}},$$

then, for all $0 \le \tau_0 \le \tau_1$, all $i \in \{1,\ldots,n\}$ and all realizations of the algorithm,

$$\sum_{\tau=\tau_0}^{\tau_1} \min[\alpha_{k_\tau,\mu(k_\tau)},1]\,\|d_{k_\tau}\|^2 > \eta\sqrt{\varsigma}\sqrt{1+\frac{1}{\varsigma}\max_{i\in\{1,\ldots,n\}}\Gamma_{k_{\tau_1+1},i}} - \eta\max[\eta,\sqrt{\varsigma}], \tag{3.24}$$

$$\sum_{\tau=\tau_0}^{\tau_1} \alpha_{k_\tau,i}^2\, d_{k_\tau,i}^2 \le \eta^2 \log\left(1+\frac{1}{\varsigma}\max_{i\in\{1,\ldots,n\}}\Gamma_{k_{\tau_1+1},i}\right). \tag{3.25}$$

Proof. Let $\gamma_{k_\tau,i} = \sqrt{\varsigma+\Gamma_{k_\tau,i}}$ and note that $\mu(k_\tau) = \text{argmax}_{i\in\{1,\ldots,n\}}\gamma_{k_{\tau+1},i}$. Since $\Gamma_{\tau_0,i} = 0$ for all $i \in \{1,\ldots,n\}$, we have also that $\gamma_{\tau_0,\mu(\tau_0)-1} = \sqrt{\varsigma}$. Define τ_α the smallest $\tau \ge \tau_0$ such that $\alpha_{k_\tau,\mu(k_\tau)} < 1$ (or $\tau_\alpha = \infty$ if such an $\alpha_{k_\tau,\mu(k_\tau)}$ does not exist). Then

$$\begin{aligned}
\sum_{\tau=\tau_0}^{\tau_1} \min[\alpha_{k_\tau,\mu(k_\tau)},1]\,\|d_{k_\tau}\|^2 &= \sum_{\tau=\tau_0}^{\tau_{\alpha-1}} \|d_{k_\tau}\|^2 + \sum_{\tau=\tau_\alpha}^{\tau_1} \alpha_{k_\tau,\mu(k_\tau)} \sum_{i\in\{1,\ldots,n\}} |d_{k_\tau,i}|^2 \\
&\ge \sum_{\tau=\tau_\alpha}^{\tau_1} \alpha_{k_\tau,\mu(k_\tau)}\,|d_{k_\tau,\mu(k_\tau)}|^2 \\
&= \eta \sum_{\tau=\tau_\alpha}^{\tau_1} \frac{|d_{k_\tau,\mu(k_\tau)}|^2}{\gamma_{k_{\tau+1},\mu(k_\tau)}} \\
&> \eta \sum_{\tau=\tau_\alpha}^{\tau_1} \frac{|d_{k_\tau,\mu(k_\tau)}|^2}{\gamma_{k_{\tau+1},\mu(k_\tau)} + \gamma_{k_\tau,\mu(k_\tau)}} \\
&= \eta \sum_{\tau=\tau_\alpha}^{\tau_1} \frac{\gamma_{k_{\tau+1},\mu(k_\tau)}^2 - \gamma_{k_\tau,\mu(k_\tau)}^2}{\gamma_{k_{\tau+1},\mu(k_\tau)} + \gamma_{k_\tau,\mu(k_\tau)}} \\
&= \eta \sum_{\tau=\tau_\alpha}^{\tau_1} \left(\gamma_{k_{\tau+1},\mu(k_\tau)} - \gamma_{k_\tau,\mu(k_\tau)}\right)
\end{aligned}$$

But $\gamma_{k_\tau,\mu(k_\tau)} \le \gamma_{k_\tau,\mu(k_{\tau-1})}$ by definition of $\mu(k_{\tau-1})$, and therefore

$$\sum_{\tau=\tau_0}^{\tau_1} \min[\alpha_{k_\tau,\mu(k_\tau)},1]\,\|d_{k_\tau}\|^2 \ge \eta \sum_{\tau=\tau_0}^{\tau_1} \left(\gamma_{k_{\tau+1},\mu(k_\tau)} - \gamma_{k_\tau,\mu(k_{\tau-1})}\right) > \eta\left(\gamma_{k_{\tau_1+1},\mu(k_{\tau_1})} - \gamma_{k_{\tau_\alpha},\mu(k_{\tau_\alpha-1})}\right). \tag{3.26}$$

Moreover, if $\tau_\alpha = \tau_0$, then $\gamma_{k_{\tau_\alpha},\mu(k_{\tau_\alpha}-1)} = \sqrt{\varsigma}$, while, because of (2.6), we must have that

$$\gamma_{k_{\tau_\alpha},\mu(k_{\tau_\alpha}-1)} = \sqrt{\varsigma + \Gamma_{k_{\tau_\alpha},\mu(k_{\tau_\alpha-1})}} \le \eta$$

when $\tau_\alpha > \tau_0$. Thus (3.26) ensures that

$$\sum_{\tau=\tau_0}^{\tau_1} \min[\alpha_{k_\tau,\mu(k_\tau)},1]\,\|d_{k_\tau}\|^2 \ge \eta\left(\sqrt{\varsigma + \Gamma_{k_{\tau_1+1},\mu(k_{\tau_1})}} - \max[\eta,\sqrt{\varsigma}]\right),$$

yielding (3.24). Finally, using the concavity and the increasing nature of the logarithm, we also have from (2.6) that, for each $i \in \{1,\ldots,n\}$,

$$\begin{aligned}
\sum_{\tau=\tau_0}^{\tau_1} \alpha_{k,i}^2 |d_{k,i}|^2 &= \eta^2 \sum_{\tau=\tau_0}^{\tau_1} \frac{|d_{k,i}|^2}{\varsigma + \Gamma_{k_{\tau+1},i}} \\
&= \eta^2 \sum_{\tau=\tau_0}^{\tau_1} \frac{\Gamma_{k_{\tau+1},i} - \Gamma_{k_\tau,i}}{\varsigma + \Gamma_{k_{\tau+1},i}} \\
&\leq \eta^2 \sum_{\tau=\tau_0}^{\tau_1} \Big(\log(\varsigma + \Gamma_{k_{\tau+1},i}) - \log(\varsigma + \Gamma_{k_\tau,i}) \Big) \\
&= \eta^2 \Big(\log(\varsigma + \Gamma_{k_{\tau_1+1},i}) - \log(\varsigma + \Gamma_{k_{\tau_0},i}) \Big)
\end{aligned}$$

giving (3.25) because $\Gamma_{k_{\tau_0},i} = 0$. □

3.4 Telescoping sum for values of the Lyapunov function

We now use the bounds developed for both types of step to consider the global decrease of the Lyapunov function ψ, from which a bound on

$$\Theta_k = 1 + \frac{1}{\varsigma} \max_{i \in \{1,\ldots,n\}} \Gamma_{k_{\tau_1+1},i} \tag{3.27}$$

can be deduced.

Lemma 3.5 Suppose that AS.1–AS.11 hold and define Θ_k as in (3.27). Then

$$\mathbb{E}\Big[\sqrt{\Theta_k}\Big] + \sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} \leq \kappa_{\rm gap} + n\,\kappa_{\rm tan}\,\eta\,\mathbb{E}[\log(\Theta_k)] \tag{3.28}$$

where

$$\kappa_{\rm gap} = \frac{8\kappa_B}{3\eta\sqrt{\varsigma}} \Big(\eta \max[\eta, \sqrt{\varsigma}] + \psi(x_0) + \kappa_\lambda \kappa_c + \rho\kappa_c - f_{\rm low} \Big).$$

Proof. Using the law of total expectation, we have from (3.10) that

$$\sum_{\nu=\nu_0}^{\nu_1} \mathbb{E}[\psi(x_j) - \psi(x_{j+1})] \geq \eta \sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu}.$$

But (3.10) also implies, when combined with (3.15), that

$$
\begin{aligned}
\sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\psi(x_{k_\tau}) - \psi(x_{k_\tau+1})] &= \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\psi(x_{k_\tau}) - \psi(x_{k_\tau}^+)\big] + \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\psi(x_{k_\tau}^+) - \psi(x_{k_\tau+1})\big] \\
&\geq \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\psi(x_{k_\tau}^+) - \psi(x_{k_\tau+1})\big] \\
&\geq \frac{3}{8\kappa_B} \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\min[\alpha_{k_\tau,\mu(k_\tau),1}] \, \|d_{k_\tau}\|^2\big] - \kappa_{\rm tan}\mathbb{E}\big[\|s_{T,k_\tau}\|^2\big] \\
&\geq \frac{3}{8\kappa_B} \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\min[\alpha_{k_\tau,\mu(k_\tau),1}] \, \|d_{k_\tau}\|^2\big] - \kappa_{\rm tan}\mathbb{E}\left[\sum_{i=1}^{n} \alpha_{k_\tau,i}^2 d_{k_\tau,i}^2\right], \\
&\geq \frac{3}{8\kappa_B}\eta\sqrt{\varsigma}\,\mathbb{E}\left[\sqrt{\Theta_k}\right] - \eta\max[\eta,\sqrt{\varsigma}] - n\kappa_{\rm tan}\eta^2\mathbb{E}[\log(\Theta_k)]\,,
\end{aligned}
\tag{3.29}
$$

where we have used the inequality $\|\Delta_{k_\tau} s_{T,k_\tau}\|^2 \leq 1$ and the second part of (2.6) to deduce the penultimate inequality, and (3.24) and (3.25) (which are both true for every realization) to derive the last. Thus, for $\min[\tau_0,\nu_0] = 0$ and $\max[\tau_1,\nu_1] = k$,

$$
\begin{aligned}
\mathbb{E}[\psi(x_0) - \psi(x_{k+1})] &= \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\psi(x_{k_\tau}) - \psi(x_{k_\tau+1})] + \sum_{\nu=\nu_0}^{\nu_1} \mathbb{E}[\psi(x_{k_\nu}) - \psi(x_{k_\nu+1})] \\
&\geq \frac{3}{4}\eta\sqrt{\varsigma}\mathbb{E}\left[\sqrt{\Theta_k}\right] - \eta\max[\eta,\sqrt{\varsigma}] - n\kappa_{\rm tan}\eta^2\mathbb{E}[\log(\Theta_k)] + \eta\sum_{\nu=\nu_0}^{\nu_1}\omega_{N,k_\nu}.
\end{aligned}
\tag{3.30}
$$

Now, using (3.1), and the bound $\|\widehat{\lambda}(x)\| \leq \kappa_\lambda$, we have that, for all realizations,

$$
\mathbb{E}[\psi(x_0) - \psi(x_{k+1})] \leq \psi(x_0) + \kappa_\lambda\kappa_c + \rho\kappa_c - f_{\rm low} \stackrel{\rm def}{=} \eta\sqrt{\varsigma}\kappa_{\rm gap} - \eta\max[\eta,\sqrt{\varsigma}],
$$

Substituting this inequality in (3.30) and using $\varsigma \in (0,1]$ gives that

$$
\eta\sqrt{\varsigma}\kappa_{\rm gap} \geq \frac{3}{4}\eta\sqrt{\varsigma}\mathbb{E}\left[\sqrt{\Theta_k}\right] + \eta\sqrt{\varsigma}\sum_{\nu=\nu_0}^{\nu_1}\omega_{N,k_\nu} - n\kappa_{\rm tan}\eta^2\mathbb{E}[\log(\Theta_k)]\,,
$$

from which (3.28) may be deduced. □

3.5 Tangential complexity

Our next step is to use (3.28) to derive a global rate of convergence over tangential steps. We start by quoting a useful technical result.

Lemma 3.6 Suppose that $t \leq b + c\,\log(t)$ for some $t \geq 1$. Then $t \leq 2b + 2c\,[\log(2c) - 1]$.

Proof. See [4, Lemma 3.6] for $a = 1$. □

We now consider the rate of convergence along the subsequence of tangential steps.

Lemma 3.7 Suppose that AS.1–AS.11 hold. Then

$$\mathbb{E}\left[\sqrt{\varsigma + \max_{i\in\{1,\ldots,n\}} \Gamma_{k,i}}\right] \le \kappa_T \stackrel{\rm def}{=} \sqrt{\varsigma}\,\Big[2\kappa_{\rm gap} + 4n\kappa_{\rm tan}\eta\Big(\log(4n\kappa_{\rm tan}\eta) - 1\Big)\Big] \tag{3.31}$$

and

$$\sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\omega_{T,k_\tau} + \omega_{N,k_\tau}] \le \kappa_T\sqrt{n}\sqrt{\tau_1+1}\left(1+\frac{\beta\eta}{\sqrt{\varsigma}}\right). \tag{3.32}$$

Proof. The inequality (3.28) yields that

$$\mathbb{E}\Big[\sqrt{\Theta_k}\Big] \le \kappa_{\rm gap} + 2n\kappa_{\rm tan}\eta\mathbb{E}\left[\log\left(\sqrt{1+\frac{1}{\varsigma}\Gamma_{\max}}\right)\right].$$

Jensen's inequality and the concavity of the logarithm then imply that

$$\mathbb{E}\Big[\sqrt{\Theta_k}\Big] \le \kappa_{\rm gap} + 2n\kappa_{\rm tan}\eta\log\left(\mathbb{E}\left[\sqrt{1+\frac{1}{\varsigma}\Gamma_{\max}}\right]\right),$$

so that we obtain from Lemma 3.6 that

$$\mathbb{E}\left[\sqrt{\varsigma + \max_{i\in\{1,\ldots,n\}} \Gamma_{k,i}}\right] \le \sqrt{\varsigma}\,\Big[2\kappa_{\rm gap} + 4n\kappa_{\rm tan}\eta\Big(\log(4n\kappa_{\rm tan}\eta) - 1\Big)\Big].$$

This is (3.31). We now invoke the inequality $\sum_{j=0}^k a_j \le \sqrt{k+1}\sqrt{\sum_{j=0}^k a_j^2}$ for nonnegative $\{a_j\}_{j=0}^k$ to deduce, from (2.13) and (3.31), that

$$\begin{aligned}
\sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\|d_{k_\tau}\|] &= \mathbb{E}\left[\sum_{\tau=\tau_0}^{\tau_1} \|d_{k_\tau}\|\right] \\
&\le \mathbb{E}\left[\sqrt{\tau_1+1}\,\sqrt{\sum_{\tau=\tau_0}^{\tau_1} \|d_{k_\tau}\|^2}\right] \\
&\le \sqrt{\tau_1+1}\,\mathbb{E}\left[\sqrt{n\max_{i\in\{1,\ldots,n\}} \Gamma_{k,i}}\right] \\
&\le \sqrt{n(\tau_1+1)}\,\mathbb{E}\left[\sqrt{\varsigma + \max_{i\in\{1,\ldots,n\}} \Gamma_{k,i}}\right] \\
&\le \sqrt{n(\tau_1+1)}\,\kappa_T.
\end{aligned} \tag{3.33}$$

Using the switching condition (2.8) and the fact that $\alpha_{k_\tau,i} \le \eta/\sqrt{\varsigma}$ for all $i\in\{1,\ldots,n\}$, we also deduce that

$$\sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\omega_{N,k_\nu}] \le \sum_{\tau=\tau_0}^{\tau_1} \beta\,\mathbb{E}\big[\|s_k^L\|\big] \le \frac{\beta\,\eta}{\sqrt{\varsigma}} \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\|d_{k_\tau}\|] \le \frac{\beta\,\eta\,\kappa_T\,\sqrt{n}\,\sqrt{\tau_1+1}}{\sqrt{\varsigma}}.$$

Summing this bound with (3.33) then gives (3.32). □

3.6 Normal complexity

The analysis of the complexity of normal step also uses the switching condition, but in the other direction.

Lemma 3.8 Suppose that AS.1–AS.11 hold. Then

$$\sum_{\nu=\nu_0}^{\nu_1} \mathbb{E}[\|d_{k_\nu}\| + \omega_{N,k_\nu}] < \kappa_N \left(1 + \frac{\kappa_g\sqrt{2n(k_\nu+1)}}{\beta\eta}\right), \tag{3.34}$$

where

$$\kappa_N = \kappa_{\rm gap} + 2n\kappa_{\rm tan} \log\left(\frac{\kappa_T}{\sqrt{\varsigma}}\right). \tag{3.35}$$

Proof. We again consider the complete decrease of the Lyapunov function over all iterations and obtain, using (3.28), that, for $\min[\tau_0,\nu_0]=0$ and $\max[\tau_1,\nu_1]=k$,

$$\begin{aligned}
\eta\kappa_{\rm gap} &\ge \mathbb{E}[\psi(x_0) - \psi(x_{k+1})]\\
&= \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\psi(x_{k_\tau}) - \psi(x_{k_\tau+1})] + \sum_{\nu=\nu_0}^{\nu_1} \mathbb{E}[\psi(x_j) - \psi(x_{j+1})]\\
&\ge \eta \sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} + \eta\mathbb{E}\Big[\sqrt{\Theta_k}\Big] - \kappa_{\rm tan}\eta^2\mathbb{E}[\log(\Theta_k)]
\end{aligned}$$

Thus, using Jensen's inequality, the concavity of the logarithms, (3.28) and (3.35),

$$\begin{aligned}
\sum_{\nu=\nu_0}^{\nu_1} \omega_{N,k_\nu} &\le \kappa_{\rm gap} + n\kappa_{\rm tan}\eta\mathbb{E}[\log(\Theta_k)]\\
&\le \kappa_{\rm gap} + 2n\kappa_{\rm tan}\eta\mathbb{E}\Big[\log\Big(\sqrt{\Theta_k}\Big)\Big]\\
&\le \kappa_{\rm gap} + 2n\kappa_{\rm tan}\eta\log\Big(\mathbb{E}\Big[\sqrt{\Theta_k}\Big]\Big)\\
&\le \kappa_N.
\end{aligned} \tag{3.36}$$

Taking now the conditional expectation in the switching condition (2.8) for $k \in k_\nu$ (i.e. when (2.8) fails) obtain that, for $\nu \in \{\nu_0,\ldots,\nu_1\}$,

$$\omega_{N,k_\nu} \ge \beta\mathbb{E}_{k_\nu}\left[\|s_k^L\right]\|_\infty \ge \beta\mathbb{E}_{k_\nu}\left[\alpha_{k_\nu,\mu(k_\nu)}\|d_{k_\nu}\|_\infty\right]. \tag{3.37}$$

Now, if $k_{\tau_\diamond}$ is the index of the last tangential iteration preceding k_{ν_1}, (2.6) gives that

$$\alpha_{k_\nu,\mu(k_\nu)} = \frac{\eta}{\sqrt{\Gamma_{k_{\tau_\diamond}+1,\mu(k_\nu)} + \varsigma}} \ge \frac{\eta}{\sqrt{2\max[\Gamma_{k_{\tau_\diamond}+1,\mu(k_\nu)},\varsigma]}}.$$

But, using AS.3,

$$\Gamma_{k_{\tau_\diamond}+1,\mu(k_\nu)} \le \tau_\diamond\kappa_g^2 \le (k_\nu+1)\kappa_g^2,$$

and therefore, since $\kappa_g \ge 1 \ge \varsigma$,

$$\alpha_{k_\nu,\mu(k_\nu)} \ge \frac{\eta}{\sqrt{2\max[(k_\nu+1)\kappa_g^2,\varsigma]}} = \frac{\eta}{\kappa_g\sqrt{2(k_\nu+1)}}.$$

This inequality, (3.37) and the law of total expectation then imply that

$$\mathbb{E}[\omega_{N,k_\nu}] \geq \mathbb{E}\left[\frac{\beta\eta}{\kappa_g\sqrt{2(k_\nu+1)}}\mathbb{E}_{k_\nu}[\|d_{k_\nu}\|_\infty]\right] \geq \frac{\beta\eta}{\kappa_g\sqrt{2(k_\nu+1)}}\mathbb{E}[\|d_{k_\nu}\|_\infty]\,. \tag{3.38}$$

Using the equivalence of norms and (3.36), this implies that

$$\sum_{\nu=\nu_0}^{\nu_1}\mathbb{E}[\|d_{k_\nu}\|] \leq \sqrt{n}\sum_{\nu=\nu_0}^{\nu_1}\mathbb{E}[\|d_{k_\nu}\|_\infty] \leq \frac{\kappa_g\sqrt{2n(k_\nu+1)}}{\beta\eta}\sum_{\nu=\nu_0}^{\nu_1}\mathbb{E}[\omega_{N,k_\nu}] \leq \frac{\kappa_g\,\kappa_N\,\sqrt{2n(k_\nu+1)}}{\beta\eta}.$$

Summing this bound with (3.36) gives (3.34). □

3.7 Combined complexity

We may assemble the above results to derive a complexity result involving the expectation of the approximate dual optimality measure.

Theorem 3.9 Suppose that AS.1-AS.11 hold and that either the gradient is exact (i.e. $g_{T,k} = G_{T,k}$ for all $k \geq 0$) or AS.11 holds. Then

$$\frac{1}{k+1}\sum_{j=0}^{k}\mathbb{E}[\|d_j\| + \|c_j\|] \leq \frac{\kappa_{\rm STRAD}}{\sqrt{k+1}} + \frac{\kappa_N}{\xi(k+1)} = \mathcal{O}\left(\frac{1}{\sqrt{k+1}}\right), \tag{3.39}$$

where

$$\kappa_{\rm STRAD} = \frac{\kappa_T}{\xi}\left(1+\frac{\beta\eta}{\sqrt{\varsigma}}\right) + \frac{\kappa_N\kappa_g\sqrt{2n}}{\xi\beta\eta}.$$

Proof. Observe first that AS.11 automatically holds in the gradient is exact. Now consider iterations from 0 to k of both types (tangential and normal) by setting $\min[k_{\nu_0}, k_{\tau_0}] = 0$ and $\max[k_{\nu_1}, k_{\tau_1}] = k$ (as in Lemma 3.5). We then obtain, by combining (3.32) and (3.34) and using AS.10, that

$$\begin{aligned}\sum_{j=0}^{k}\mathbb{E}[\|d_j\| + \|c_j\|] &\leq \sum_{j=0}^{k}\mathbb{E}\left[\|d_j\| + \frac{\omega_{N,j}}{\xi}\right]\\ &\leq \frac{1}{\xi}\sum_{\tau=\tau_0}^{\tau_1}\mathbb{E}[\|d_{k_\tau}\| + \omega_{N,k_\tau}] + \frac{1}{\xi}\sum_{\nu=\nu_0}^{\nu_1}\mathbb{E}[\|d_{k_\nu}\| + \omega_{N,k_\nu}]\\ &\leq \frac{\kappa_T}{\xi}\sqrt{k+1}\left(1+\frac{\beta\eta}{\sqrt{\varsigma}}\right) + \frac{\kappa_N}{\xi}\left(1+\frac{\kappa_g\sqrt{2n(k+1)}}{\beta\eta}\right),\end{aligned}$$

where we used the inequalities $\tau_1 \leq k_{\tau_1} \leq k$ and $k_{\nu_1} \leq k$. The bound (3.39) is finally obtained dividing both sides by $k+1$. □

Our interest is now to transform this result into a result using the true dual optimality measure $\Omega_{T,k} = \Omega_T(x_k)$ given by (2.1). The reader has undoubtely noticed that we have not assumed that the gradient oracle $g(x)$ is unbiased, an assumption which is common for unconstrained problems and relates the approximate first-order optimality condition for such problems ($g(x) = 0$) to true optimality ($G(x) = 0$). When inequality constraints are present, this condition is not sufficient, as was shown in [4] for the case of bound constraints. The simplest approach is a direct analog of the unconstrained condition and is to require that the optimality measure is unbiased also for the case of constraints, that is

$$\mathbb{E}_k[\|d_k\|] = \Omega_T(x_k), \tag{3.40}$$

but that might be difficult to achieve. A slightly looser condition is to require that

$$\mathbb{E}_k[|\Omega_T(x_k) - \|d_k\||] \le \kappa_\Omega \mathbb{E}_k[\|d_k\|] \tag{3.41}$$

for some $\kappa_\Omega > 0$. Finally, the nature of the orthogonal projection in $\Omega_T(x_k)$ also implies that a condition of the form

$$\mathbb{E}_k[\|G_k - g_k\|] \le \kappa_\Omega \mathbb{E}_k[\|d_k\|] \tag{3.42}$$

is also suitable because it implies an inequality of the form (3.41) (see [4, Lemma 4.1]).

Theorem 3.10 Suppose that AS.1-AS.10 hold and that either the gradient is exact (i.e. $g_{T,k} = G_{T,k}$ for all $k \ge 0$) or AS.11 and one of (3.40), (3.41) or (3.42) hold. Then

$$\frac{1}{k+1}\sum_{j=0}^{k} \mathbb{E}[\Omega_T(x_j) + \|c_j\|] \le \frac{\kappa_{\rm STRAD}}{\sqrt{k+1}} + \frac{\kappa_N}{\xi(k+1)} = \mathcal{O}\left(\frac{1}{\sqrt{k+1}}\right), \tag{3.43}$$

where

$$\kappa_{\rm STRAD} = \frac{\kappa_\Omega \kappa_T}{\xi}\left(1 + \frac{\beta\eta}{\sqrt{\varsigma}}\right) + \frac{\kappa_N\, \kappa_g \sqrt{2}}{\xi\beta\eta}.$$

Proof. It directly results from Theorem 3.9 and either (3.40), or (3.41) and the triangle inequality. □

This result implies that, under our assumptions, the algorithm requires at most $\mathcal{O}(\min[\epsilon_D, \epsilon_C]^{-2})$ iterations to achieve ϵ-approximate first-order criticality, that is $\Omega_T(x_j) \le \epsilon_D$ and $\|c_j\| \le \epsilon_c$. It is remarkable that this complexity order, which is the best that can be achieved for deterministic unconstrained problems (see [10, 9, 8]), can also be achieved, under our assumptions, in our more general stochastic constrained context. This implies that the bound given by Theorem 3.10 is also optimal in order.

3.8 Further discussion of the stochastic conditions

.

Can AS.11 or (3.40)–(3.42) be relaxed? The theory presented above suggests two different but complementary possibilities.

The first is to relax the condition of AS.11 to admit a sufficiently small noise of the first-order descent $g_k^T s_{T,k}$ condition (in addition to the second-order noise allowed by AS.11), and transform the condition of AS.11 into

$$\mathbb{E}_k^\tau\left[|(G_{T,k} - g_{T,k})^T s_{T,k}|\right] \le \kappa_{\rm dir,*}\mathbb{E}_k^\tau\left[|g_k^T s_{T,k}|\right] + \frac{\kappa_{\rm dir}}{2}\mathbb{E}_k^\tau\left[\|s_{T,k}\|^2\right].$$

for some constants $\kappa_{\rm dir,*} < 1$ and $\kappa_{\rm dir} > 0$. In this case, (3.19) now holds for a smaller but strictly positive value of $\kappa_{\rm dir}$, the rest of the theory being unchanged.

A second possibility is to allow further error depending on a finite number of second-order terms at past iterations, for instance by requiring

$$\mathbb{E}_k^\tau\left[|(G_{T,k} - g_{T,k})^T s_{T,k}|\right] \le \sum_{j=0}^{M} \kappa_{\rm dir,j}\mathbb{E}_{k-j}^\tau\left[\|s_{T,k-j}\|^2\right] = \kappa_{\rm dir,0}\mathbb{E}_k^\tau\left[\|s_{T,k}\|^2\right] + \sum_{j=1}^{\min[M,k]} \kappa_{\rm dir,j}\|s_{T,k-j}\|^2$$

for some fixed integer $M \geq 1$ independent of k and some $\kappa_{\rm dir,j} \geq 0$ for all $j \in \{0, \ldots, \min[M,k]\}$ with $\max_{j\in\{1,\ldots,\min[M,k]\}} \kappa_{\rm dir,j} > 0$. The inequality (3.19) then takes the form

$$\mathbb{E}_k^\tau\Big[\nabla_x\psi(x_k^+,\widehat{\lambda}_k^+)^T s_{T,k}\Big] \leq -\kappa_t\mathbb{E}_k^\tau\big[\alpha_k\omega_{T,k}^2\big] + \kappa_{\rm dir,0}\mathbb{E}_k^\tau\big[\|s_{T,k}\|^2\big] + \sum_{j=1}^{\min[M,k]} \kappa_{\rm dir,j}\|s_{T,k-j}\|^2$$

and the additional (second-order) term $\sum_{j=1}^{\min[M,k]} \kappa_{\rm dir,j}\|s_{T,k-j}\|^2$ is carried to the summation in (3.29) in Lemma 3.5, which becomes

$$\sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}[\psi(x_{k_\tau}) - \psi(x_{k_\tau+1})] = \sum_{\tau=\tau_0}^{\tau_1} \mathbb{E}\big[\alpha_{k_\tau,\mu(k_{k\tau})}\|d_{k_\tau}\|^2 - M\kappa_{\rm tan}\|s_{T,k_\tau}\|^2\big] .$$

because each second-order term appears at most M times in the sum. This merely amounts to multiplying the constant $\kappa_{\rm tan}$ by M, and the rest of the theory is again unchanged. Observe that the choice $\kappa_{\rm dir,0} = 0$ makes the right-hand side of the condition measurable with respect to the past, a potentially useful property in practice. The same technique of taking into account past iterations is also applicable to any of (3.40)–(3.42). For instance, (3.42) can be relaxed to

$$\mathbb{E}_k[\|G_k - g_k\|] \leq \kappa_{\Omega,0}\mathbb{E}_k[\|d_k\|] + \sum_{j=1}^{\min[M,k]} \kappa_{\Omega,j}\|d_{k-j}\|.$$

Of course, mixing these strategies is possible, leading to fairly relaxed conditions. When the random noise is cause by a sampling process (as is the case in many methods for the minimization of finite sums), it was for instance observed in [22] that incorporating past steps in the right-hand side of the accuracy condition covers a very much slower increase in the sample size when iterations progress.

We finally note that, because we have proved in Theorem 3.5 that Θ_k only grows very slowly, we see from (2.7), (2.12), (2.13) and (3.28) that $\|s_{T,k}\|$ cannot be much shorter that the projected approximate gradient at x_k, motivating our analogy between (3.6) and (3.7).

4 Summary and open questions

A OFFO algorithm for stochastic nonlinear optimization subject to deterministic equality and inequality constraints has been proposed, which is capable of exploiting (approximate) second-order information when available. Its global rate of convergence has been shown to be $\mathcal{O}(1/\sqrt{k+1})$ when the Jacobian of the constraints is assumed to be full-rank. This rate is, in order, identical to that of first-order methods for unconstrained problems. The proofs to obtain this result are (relatively) simple.

Of course, several questions remain. Of particular interest is weakening the assumptions of bounded gradients, full-rank Jacobians and oracle variance. Handling approximate projections would also be valuable.

Acknowledgement

The authors acknowledge the generous support of the Department of Applied Mathematics in the Hong Kong Polytechnic University during a visit in the fall of 2025. They are also pleased to acknowledge a very useful discussion about stochastic augmented Lagrangians with Xiao Wang.

References

[1] A. Alacaoglu and H. Lyu. Convergence of first-order methods for constrained nonconvex optimization with dependent data. In *International Conference on Machine Learning. PMLR*, pages 458–489, 2023.

[2] A. Alacaoglu, Y. Malitsky, and V. Cevher. Convergence od adaptive algorithms for constrained weakly-convex optimization. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NEURIPS 2021)*, 2021.

[3] A. Attia and T. Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. arxiv:2302.08783, 2023.

[4] S. Bellavia, G. Gratton, B. Morini, and Ph. L. Toint. Fast stochastic Adagrad for nonconvex bound-constrained optimization. arXiv:2505:06374, 2025.

[5] S. Bellavia, G. Gratton, B. Morini, and Ph. L. Toint. An objective-function-free algorithm for general smooth constrained optimization. arXiv:2602:11770, 2026.

[6] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods.* Academic Press, 1982.

[7] G. Braun, A. Carderera, C. Combettes, H. Hassani, A. Karbasi, A. Mokhtari, and S. Pokutta. Conditional gradient methods. arXiv;2211.14103v2, 2023.

[8] Y. Carmon, J. C. Duchi, O. Hinder, and A. Sidford. Lower bounds for finding stationary points I. *Mathematical Programming, Series A*, 184:71–120, 2020.

[9] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the complexity of the steepest-descent with exact linesearches. Technical Report naXys-16-2012, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium, 2012.

[10] C. Cartis, N. I. M. Gould, and Ph. L. Toint. On the oracle complexity of first-order and derivative-free algorithms for smooth nonconvex minimization. *SIAM Journal on Optimization*, 22(1):66–86, 2012.

[11] A. R. Conn, N. I. M. Gould, and Ph. L. Toint. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*, 28(2):545–572, 1991.

[12] 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.

[13] 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(4):3592–3622, 2024.

[14] F. E. Curtis, N. I. M. Gould, D. P. Robinson, and Ph. L. Toint. An interior-point trust-funnel algorithm for nonlinear optimization. *Mathematical Programming, Series A*, 161(1):73–134, 2017.

[15] A. Défossez, L. Bottou, F. Bach, and N. Usunier. A simple convergence proof for Adam and Adagrad. *Transactions on Machine Learning Research*, October 2022.

[16] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12, July 2011.

[17] 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(2):2007–2037, 2024.

[18] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3:95–110, 1956.

[19] S. Ghadimi and G. Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming, Series A*, 156(1-2):59–100, 2016.

[20] N. I. M. Gould, D. P. Robinson, and Ph. L. Toint. Corrigendum: Nonlinear programming without a penalty function or a filter. *Mathematical Programming, Series A*, 131(1):403–404, 2012.

[21] N. I. M. Gould and Ph. L. Toint. Nonlinear programming without a penalty function or a filter. *Mathematical Programming, Series A*, 122(1):155–196, 2010. See also [20].

[22] S. Gratton, S. Jerad, and Ph. L. Toint. A stochastic objective-function-free adaptive regularization method with optimal complexity. *Open Journal of Mathematical Optimization*, 6(5), 2025.

[23] S. Gratton, V. Mercier, E. Riccietti, and Ph. L. Toint. A block-coordinate approach of multi-level optimization with an application to physics-informed neural networks. *Computational Optimization and Applications*, 89(2):385–417, 2024.

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

[25] M. R. Hestenes. Multiplier and gradient methods. *Journal of Optimization Theory and Applications*, 4:303–320, 1969.

[26] Y. Hong and J. Lin. Revisting convergence of Adagrad with relaxed assumptions. arXiv:2403.13794v2, 2024.

[27] L. Jin and X. Wang. Stochastic nested primal-dual method for nonconvex constrained composition optimization. *Mathematics of Computation*, 94:305–358, 2025.

[28] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[29] D. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings in the International Conference on Learning Representations (ICLR)*, 2015.

[30] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, page 983–992, 2019.

[31] S. Lu, S. Zeng, X. Cui, M. Squillante, L. Horesh, B. Kingsbury, J. Liu, and M. Hong. A stochastic linearized augmented Lagrangian method for decentralized bilevel optimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 25, pages 30638–30650. Curran Associates, Inc., 2022.

[32] P. Perdikaris M. Raissi and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[33] M. C. Mukkamala and M. Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In *Proceedings of the 34th International Conference on Machine Learning*, page 2545–2553, 2017.

[34] E. O. Omojokun. *Trust region algorithms for optimization with nonlinear equality and inequality constraints*. PhD thesis, University of Colorado, Boulder, Colorado, USA, 1989.

[35] D. Papadimitriou and B. C. Vu. A stochastic Lagrangian-based method for nonconvex optimization with nonlinear constraints. Optimization Online preprint, 2025.

[36] M. J. D. Powell. A method for nonlinear constraints in minimization problems. In R. Fletcher, editor, *Optimization*, pages 283–298, London, 1969. Academic Press.

[37] H. Robbins and S. Monroe. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400—407, 1951.

[38] Q. Shi, X. Wang, and H. Wang. A momentum-based linearized augmented Lagrangian method for nonconvex constrained stochastic optimization. *Mathematics of Operations Research*, 2025.

[39] T. Tieleman and G. Hinton. Lecture 6.5-RMSPROP. COURSERA: Neural Networks for Machine Learning, 2012.

[40] Ph. L. Toint. Divergence of the ADAM algorithm with fixed-stepsize: a (very) simple example. In K. Fackeldey, A. Kannan, S. Pokutta, K. Sharma, D. Walter, A. Walther, and M. Weiser, editors, *Mathematical Optimization for Machine Learning: Proceedings of the MATH+ Thematic Einstein Semester 2023*, pages 195–199. De Gruyter Brill, 2025.

[41] B. Wang, H. Zhang, Z. Ma, and W. Chen. Convergence of Adagrad for non-convex objectives: simple proofs and relaxed assumptions. In *Proceedings of the Thirty-Sixth Conference on Learning Theory*, pages 161–190, 2023.

[42] I.-J. Wang and J. C. Spall. A constrained simultaneous perturbation stochastic approximation algorithm based on penalty functions. In *Proceedings of the 1999 Amer. Cont. Conference, vol. 1*, pages 393–399. IEEE, 1999.

[43] Q. Wang, Ch. Peirmarini, Y. Zhu, and F. E. Curtis. Projected stochastic momentum methods for nonlinear equality-constrained optimization for machine learning. arXiv:2601.11785, 2026.

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

[45] M. F. Şahin, A. Eftekhari, A. Alacaoglu, F. Latorre, and V. Cevher. An inexact augmented Lagrangian framework for nonconvex optimization with nonlinear constraints. In *NeurIPS*, 2019.