On Invariance and Linear Convergence of Evolution Strategies with Augmented Lagrangian Constraint Handling

Asma Atamna, Anne Auger, Nikolaus Hansen

On Invariance and Linear Convergence of Evolution Strategies with Augmented Lagrangian Constraint Handling

Asma Atamna[a,b,*], Anne Auger[a,b], Nikolaus Hansen[a,b]

[a] *RandOpt Team, Inria Saclay–Île-de-France*
[b] *CMAP UMR 7641 École Polytechnique CNRS, Route de Saclay, 91128 Palaiseau Cedex France*

Abstract

In the context of numerical constrained optimization, we investigate stochastic algorithms, in particular evolution strategies, handling constraints via augmented Lagrangian approaches. In those approaches, the original constrained problem is turned into an unconstrained one and the function optimized is an augmented Lagrangian whose parameters are adapted during the optimization. The use of an augmented Lagrangian however breaks a central invariance property of evolution strategies, namely invariance to strictly increasing transformations of the objective function. We formalize nevertheless that an evolution strategy with augmented Lagrangian constraint handling should preserve invariance to strictly increasing affine transformations of the objective function and the scaling of the constraints—a subclass of strictly increasing transformations. We show that this invariance property is important for the linear convergence of these algorithms and show how both properties are connected.

Keywords: Augmented Lagrangian, constrained optimization, evolution strategies, Markov chain, adaptive randomized algorithms, invariance

*Corresponding author
Email addresses: asma.atamna@inria.fr (Asma Atamna), anne.auger@inria.fr (Anne Auger), nikolaus.hansen@inria.fr (Nikolaus Hansen)

1. Introduction

Evolution strategies (ESs) are randomized (or stochastic) algorithms that are widely used in industry for solving real-wold continuous optimization problems. Their success is due to their robustness and their ability to deal with a wide range of difficulties encountered in practice such as non-separability, ill-conditioning, and multi-modality. They are also well-suited for *black-box* optimization, a common scenario in industry where the mathematical expression of the objective function—or the source code that computes it—is not available. The covariance matrix adaptation evolution strategy (CMA-ES) [15] is nowadays considered the state-of-the-art method and is able to achieves linear convergence on a large class of functions when solving unconstrained optimization problems.

Linear convergence is a desirable property for an ES; it represents the fastest possible rate of convergence for a randomized algorithm. It has been widely investigated in the unconstrained case on comparison-based adaptive randomized algorithms [8, 7, 11, 6, 9], where the connection between linear convergence and invariance of the studied algorithms has been established.

On ESs for unconstrained optimization, linear convergence is commonly analyzed using a Markov chain approach that consists in finding an underlying homogeneous Markov chain with some "stability" properties, generally positivity and Harris-recurrence. If such a Markov chain exists, linear convergence can be deduced by applying a law of large numbers (LLN) for Markov chains. In [8], it is shown that the existence of a homogeneous Markov chain of interest stems from the invariance of the algorithm, namely invariance to strictly increasing transformations of the objective function, translation-invariance, and scale-invariance.

In this work, we study ESs for *constrained* optimization where the constraints are handled using an *augmented Lagrangian* approach. A general constrained optimization[1] problem can be written as[2]

$$\begin{aligned} &\arg\min_{\mathrm{x}} \ f(\mathrm{x}) \\ &\text{subject to } g(\mathrm{x}) \leq \mathbf{0} \ , \end{aligned} \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function and $g : \mathbb{R}^n \to \mathbb{R}^m$ is the constraint

[1] We consider minimization in this work. Therefore, by "optimization", we implicitly refer to minimization.

[2] An equality constraint can be written as two inequality constraints, hence the absence of equality constraints in (1).

function. The notation $g(\mathrm{x}) \leq \mathbf{0}$ in this case is equivalent to

$$g_i(\mathrm{x}) \leq 0,\ i = 1, \cdots, m\ ,$$

where $g(\mathrm{x}) = (g_1(\mathrm{x}), \cdots, g_m(\mathrm{x}))^\intercal$ and $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, \cdots, m$. Augmented Lagrangian methods transform the initial constrained problem (1) into one or many unconstrained problems by defining a new function to minimize, the augmented Lagrangian. The use of an augmented Lagrangian, however, results in the loss of invariance to strictly increasing transformations of f, as well as g. Yet, invariance to a subset of strictly increasing transformations can be achieved: namely invariance to strictly increasing affine transformations of the objective function f and to the scaling of the constraint function g. We formulate that this invariance should be satisfied for an augmented Lagrangian ES. We explain how this property, along with translation-invariance and scale-invariance, is related to linear convergence of the algorithm by exhibiting a homogeneous Markov chain whose stability implies linear convergence.

This paper is organized as follows: first, we give an overview of augmented Lagrangian methods in Section 2. Then, we present our algorithmic setting in Section 3: we describe a general framework for building augmented Lagrangian randomized algorithms from adaptive randomized algorithms for unconstrained optimization in Section 3.1, then we use this framework to instantiate a practical ES with adaptive augmented Lagrangian in Section 3.2 and a more general step-size adaptive algorithm with augmented Lagrangian in Section 3.3. In Section 4, we discuss important invariance properties for augmented Lagrangian methods. Section 5 is dedicated to the analysis: we start by showing that our general augmented Lagrangian step-size adaptive algorithm satisfies the previously defined invariance properties in Section 5.1. In Section 5.2, we give an overview of the Markov chain approach for analyzing linear convergence in the unconstrained case, then we apply the same approach to investigate linear convergence of our general algorithm. We show in particular how invariance allows to achieve linear convergence on problems with linear constraints. We present our numerical results in Section 6 and provide a discussion in Section 7.

A preliminary version of this work was published in [5]. The focus was on identifying a homogeneous Markov chain for the general augmented Lagrangian algorithm we study, then deducing its linear convergence under sufficient stability conditions.

Notations

We denote $\mathbb{Z}_{\geq 0}$ the set of non-negative integers $\{0, 1, \cdots\}$ and $\mathbb{Z}_{>0}$ the set of positive integers $\{1, 2, \cdots\}$. We denote $\mathbb{R}_{\geq 0}$ the set of non-negative real numbers and $\mathbb{R}_{>0}$ the set of positive real numbers. We denote $[\mathrm{x}]_i$ the ith entry of a vector x. For a matrix M, $[\mathrm{M}]_{ij}$ denotes the entry in its ith row and jth column. We denote $\mathbf{0}$ the vector $(0, \cdots, 0)^\intercal \in \mathbb{R}^n$ and $\mathrm{I}_{n\times n} \in \mathbb{R}^{n\times n}$ the identity matrix. We denote $\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})$ the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathrm{I}_{n\times n}$. We refer to a multivariate normal variable with mean $\mathbf{0}$ and covariance matrix $\mathrm{I}_{n\times n}$ as *standard multivariate normal variable* in the remainder of the paper. We denote $\mathrm{Im}(f)$ the image of a function f and $\circ$ the function composition operator. We denote $\odot$ the entrywise (Hadamard) product. For a vector $\mathrm{x} = (\mathrm{x}_1, \cdots, \mathrm{x}_k)^\intercal$, x^2 denotes the vector $(\mathrm{x}_1^2, \cdots, \mathrm{x}_k^2)^\intercal$.

2. Augmented Lagrangian Methods: Overview and Related Work

Augmented Lagrangian (AL) methods are a family of constraint handling approaches. They were first introduced in [16, 20] as an alternative to penalty function methods, in particular quadratic penalty methods, whose convergence necessitates the penalty parameters to grow to infinity as the optimization progresses, thereby causing ill-conditioning [19].

Analogously to penalty function methods, AL methods proceed by transforming the constrained problem into one or many unconstrained optimization problems by constructing a new objective function, the AL, as a combination of a Lagrangian and a penalty function part. Consider the constrained optimization problem in (1). The Lagrangian is a function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ defined as

$$\mathcal{L}(\mathrm{x}, \gamma) = f(\mathrm{x}) + \gamma^\intercal g(\mathrm{x}) \ , \tag{2}$$

where $\gamma = (\gamma^1, \cdots, \gamma^m)^\intercal$ is a vector of Lagrange factors. The Lagrangian $\mathcal{L}$ is used in the literature to formulate the so-called *Karush-Kuhn-Tucker* (KKT) first-order necessary conditions of optimality, one of which is the KKT stationarity condition that states the following: if a point $\mathrm{x}^* \in \mathbb{R}^n$ is a local minimum for the constrained problem (1), then, assuming the derivatives of f and g at x^* exist and some constraint qualifications[3] hold at x^*, there exists a vector $\gamma^* \in \mathbb{R}^m_{\geq 0}$ of

[3]Constraint qualifications are regularity conditions that the constraint functions need to satisfy in order for the KKT conditions to apply. The *linear independence constraint qualification* (LICQ) and the *Mangasarian-Fromovitz constraint qualification* (MFCQ) [19] are commonly used in the literature.

Lagrange multipliers such that

$$\nabla \mathcal{L}(\mathrm{x}^*, \gamma^*) = \nabla f(\mathrm{x}^*) + \gamma^{*\intercal} \nabla g(\mathrm{x}^*) = \mathbf{0}^\intercal \ , \tag{3}$$

where $\nabla f(\mathrm{x}^*)$ is the gradient of f at x^* and $\nabla g(\mathrm{x}^*)$ is a $m \times n$ Jacobian matrix whose ith row is the gradient $\nabla g_i(\mathrm{x}^*)$. We say that x^* is a *KKT point*. The KKT conditions ensure the *existence* of a vector γ^* of Lagrange multipliers. If the constraints satisfy the linear independence constraint qualification (LICQ) [19] however, i.e. if the constraint gradients are linearly independent, the vector γ^* of Lagrange multipliers is unique [19].

The AL, denoted h, is constructed by "augmenting" the Lagrangian $\mathcal{L}$ in (2) by a penalty term. The resulting function is of the form:

$$h(\mathrm{x}, \phi) = f(\mathrm{x}) + \varphi(g(\mathrm{x}), \phi) \ , \tag{4}$$

where the vector γ of Lagrange factors is part of the parameter ϕ and the function φ combines the constraints and the parameter vector ϕ. We consider ALs that are parametrized by γ and a vector $\omega = (\omega^1, \cdots, \omega^m)^\intercal \in \mathbb{R}^m_{>0}$ of penalty factors, i.e. $\phi = (\gamma, \omega)$, such as the practical AL for (1) defined as

$$h(\mathrm{x}, \gamma, \omega) = f(\mathrm{x}) + \underbrace{\sum_{i=1}^{m} \begin{cases} \gamma^i g_i(\mathrm{x}) + \frac{1}{2}\omega^i g_i(\mathrm{x})^2 & \text{if } \gamma^i + \omega^i g_i(\mathrm{x}) \geq 0 \\ -\frac{{\gamma^i}^2}{2\omega^i} & \text{otherwise} \end{cases}}_{\varphi_1(g(\mathrm{x}), \gamma, \omega)} \ . \tag{5}$$

The quality of a solution x is determined by adding $f(\mathrm{x})$ and (i) $\gamma^i g_i(\mathrm{x}) + \frac{\omega^i}{2} g_i(\mathrm{x})^2$ if $g_i(\mathrm{x})$ is larger than $-\frac{\gamma^i}{\omega^i}$ or (ii) $-\frac{{\gamma^i}^2}{2\omega^i}$ otherwise, for each constraint g_i. Therefore, when "far" in the feasible domain, the objective function is only altered by a constant. In this work, however, we study a particular case of (1) where all the constraints are active at the optimum[4] (see Section 5 for details on the considered constrained problem). Therefore, we use the following (simpler) AL for our analysis:

$$h(\mathrm{x}, \gamma, \omega) = f(\mathrm{x}) + \underbrace{\gamma^\intercal g(\mathrm{x}) + \frac{1}{2}\omega^\intercal g(\mathrm{x})^2}_{\varphi_2(g(\mathrm{x}), \gamma, \omega)} \ . \tag{6}$$

In comparison to the AL in (5), a penalization is applied even when a point is far in the feasible domain. In practice, the function φ in (4) is designed such that a KKT point x^* is a stationary point for h, that is, for all $\omega \in \mathbb{R}^m_{>0}$, $\nabla_{\mathrm{x}} h(\mathrm{x}^*, \gamma^*, \omega) = \mathbf{0}^\intercal$.

[4] We say that a constraint g_i is *active* at a point $\bar{\mathrm{x}}$ if $g_i(\bar{\mathrm{x}}) = 0$.

In *adaptive* AL approaches, γ is adapted to approach the Lagrange multipliers and ω is generally increased to favour feasible solutions. A good adaptation mechanism for ω should only increase ω when necessary to prevent ill-conditioning. Indeed, with AL approaches, penalty factors do not need to tend to infinity in order to converge [19].

AL approaches have gained a large interest since their introduction in the 1960s and their convergence has been widely investigated in the mathematical nonlinear programming community [12, 17, 10]. In evolutionary computation, there exist some examples of evolutionary algorithms using ALs to handle constraints, as in [21] where the authors present an AL coevolutionary method for constrained optimization, where a population of Lagrange factors is evolved in parallel with a population of candidate solutions using an evolution strategy with self-adaptation.

In [13], the authors present a genetic algorithm for constrained optimization with an AL approach. Their algorithm requires a local search procedure to improve the current best solution in order to converge to the optimal solution and to Lagrange multipliers.

More recently, an AL approach was implemented for a $(1 + 1)$-ES to handle one constraint in [2]. The authors present an adaptation rule for the penalty parameter and observe the linear convergence of their approach on a linearly constrained sphere function and a linealy constrained moderately ill-conditioned ellipsoid function. This algorithm was analyzed in [3] using a Markov chain approach. The authors construct a homogeneous Markov chain and deduce linear convergence to the optimum and the associated Lagrange multipliers under the stability of this Markov chain. In [4], a general framework for building an adaptive AL randomized algorithm is presented for the case of one constraint. The authors use this general framework to implement the AL approach presented in [2] for CMA-ES.

3. Algorithmic Framework

Given an adaptive randomized algorithm for unconstrained optimization, it is possible to build an AL algorithm for constrained optimization by applying the general framework described in [4]. In the following, we extend this framework to the case of multiple constraints and use it to construct a practical $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES with adaptive AL, as well as a general AL adaptive randomized algorithm that includes the previous $(\mu/\mu_{\mathrm{w}}, \lambda)$-ES as a particular case.

3.1. Methodology for Building AL Adaptive Randomized Algorithms

Let consider a general adaptive randomized algorithm minimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ and whose state at time t is given by the state variable $\theta_t^{\mathrm{u}} \in \Omega^{\mathrm{u}}$, where Ω^{u} is the state space and where the superscript "u" stands for "unconstrained". The algorithm can be viewed as a sequence $\{\theta_t^{\mathrm{u}} : t \in \mathbb{Z}_{\geq 0}\}$ of its states, where each state is defined recursively via a deterministic transition function $\mathcal{F}$ parameterized by f, according to

$$\theta_{t+1}^{\mathrm{u}} = \mathcal{F}^f(\theta_t^{\mathrm{u}}, \mathrm{U}_{t+1}) \ , \tag{7}$$

where in our case $\mathrm{U}_{t+1} = (\mathrm{U}_{t+1}^1, \cdots, \mathrm{U}_{t+1}^\lambda) \in \mathbb{R}^{n \times \lambda}$ is a vector of λ independent identically distributed (i.i.d.) random vectors U_{t+1}^i, $i = 1, \cdots, \lambda$. The update of the state variable θ_t^{u} typically includes the generation of λ candidate solutions $\{\mathrm{X}_{t+1}^i : i = 1, \cdots, \lambda\}$ whose f-values are used to compute the new state $\theta_{t+1}^{\mathrm{u}}$. In the case of ESs, this update is *comparison-based*, that is, the f-values of the candidate solutions are only used through comparison. Therefore, the update in (7) can be rewritten as a deterministic function $\mathcal{G}$ of the current state and the ordered vector $\mathrm{U}_{t+1}^{\varsigma^{\mathrm{u}}} = (\mathrm{U}_{t+1}^{\varsigma^{\mathrm{u}}(1)}, \cdots, \mathrm{U}_{t+1}^{\varsigma^{\mathrm{u}}(\lambda)})$ as follows:

$$\theta_{t+1}^{\mathrm{u}} = \mathcal{G}(\theta_t^{\mathrm{u}}, \mathrm{U}_{t+1}^{\varsigma^{\mathrm{u}}}) \ , \tag{8}$$

where ς^{u} is the permutation of indices obtained from ranking the candidate solutions according to their f-values. More formally, ς^{u} satisfies

$$f(\mathrm{X}_{t+1}^{\varsigma^{\mathrm{u}}(1)}) \leq \cdots \leq f(\mathrm{X}_{t+1}^{\varsigma^{\mathrm{u}}(\lambda)}) \ . \tag{9}$$

This formalism was first introduced in [8] as part of a general methodology to analyze linear convergence on comparison-based adaptive randomized algorithms for unconstrained optimization with a Markov chain approach.

In the presence of constraints handled with an AL, the objective function f is replaced by the AL, h, defined by the general equation (4). In adaptive AL approaches, the parameter ϕ is updated. This leads to a *dynamic* optimization problem where the objective function changes possibly at each iteration. Therefore, starting from a comparison-based adaptive randomized algorithm for unconstrained optimization with a state θ_t^{u} and the update function $\mathcal{G}$ in (8), one can build a randomized algorithm with adaptive AL constraint handling as follows:

- The state θ_t of the new algorithm is defined as the state of the original algorithm to which we add the vector ϕ_t of the AL parameters. Formally,

  $$\theta_t = (\theta_t^{\mathrm{u}}, \phi_t) \ , \tag{10}$$

  and we denote Ω the state space of the new algorithm.

- The objective function on which the candidate solutions $\{\mathrm{X}_{t+1}^i : i = 1, \cdots, \lambda\}$ are evaluated is the AL, $h(\mathrm{x}, \phi_t)$.

- The update of the state θ_t of the new algorithm takes place in two stages: first, θ_t^{u} is updated via
$$\theta_{t+1}^{\mathrm{u}} = \mathcal{G}(\theta_t^{\mathrm{u}}, \mathrm{U}_{t+1}^{\varsigma}) \ , \tag{11}$$
where the permutation ς extracts the indices of the ordered candidate solutions on h, i.e.
$$h(\mathrm{X}_{t+1}^{\varsigma(1)}, \phi_t) \leq \cdots \leq h(\mathrm{X}_{t+1}^{\varsigma(\lambda)}, \phi_t) \ .$$
Then, the vector ϕ_t of the AL parameters is updated via
$$\phi_{t+1} = \mathcal{H}^{(f,g)}(\phi_t, \theta_t^{\mathrm{u}}, \mathrm{U}_{t+1}^{\varsigma}) \ , \tag{12}$$
where we assume the update function $\mathcal{H}$ to depend also on θ_t^{u}, on the vector U_{t+1}, and possibly on the f-values and g-values of the candidate solutions $\{\mathrm{X}_{t+1}^i : i = 1, \cdots, \lambda\}$.

We now use this framework to instantiate a practical adaptive randomized algorithm with AL constraint handling.

3.2. Practical ES with Adaptive AL Constraint Handling

We present a non-elitist, multi-parent, step-size adaptive ES with an adaptive AL constraint handling. We refer to our algorithm as the $(\mu/\mu_{\mathrm{w}}, \lambda)$-CSA$_{\mathrm{off}}$-AL, where "CSA" denotes the cumulative step-size adaptation rule [15], and the subscript "off" indicates a variant of CSA where the cumulation is deactivated. This ES generalizes the adaptive AL approach presented in [2] to non-elitist selection and multiple constraints; we extend in particular the adaptation of the penalty parameters to the case of multiple constraints.

The pseudocode is given in Algorithm 1. Lines 0–3 define the input of the algorithm, its constants, the AL under consideration, and the initial parameters values. First, λ candidate solutions $\{\mathrm{X}_{t+1}^i : i = 1, \cdots, \lambda\}$ are sampled in Line 4 according to
$$\mathrm{X}_{t+1}^i = \mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i \ , \tag{13}$$
where X_t is the current estimate of the optimum, also referred to as the *mean vector*, and where

A1 each $\mathrm{U}_{t+1} = (\mathrm{U}_{t+1}^1, \cdots, \mathrm{U}_{t+1}^\lambda) \in \mathbb{R}^{n \times \lambda}$ for $t \in \mathbb{Z}_{\geq 0}$ satisfies $\{\mathrm{U}_{t+1}^i : i = 1, \cdots, \lambda\}$ are i.i.d. standard multivariate normal variables. The sequence $\{\mathrm{U}_{t+1} : t \in \mathbb{Z}_{\geq 0}\}$ is i.i.d.

The factor $\sigma_t > 0$ is the *step-size* and determines the "width" of the sampling distribution. The candidate solutions are then ranked according to their h-values in Line 5, where ς is the permutation that contains the indices of the ranked candidate solutions and is defined according to

$$h(\mathrm{X}_{t+1}^{\varsigma(1)}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}_{t+1}^{\varsigma(\lambda)}, \gamma_t, \omega_t) \ . \tag{14}$$

The new solution X_{t+1} is computed in Line 6 by recombining the μ best candidate solutions (or parents) in a weighted sum according to

$$\mathrm{X}_{t+1} = \mathrm{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)} \ . \tag{15}$$

The constants w_i, $i = 1, \cdots, \mu$, are the weights associated to the parents, where larger weights are attributed to better parents. This recombination scheme is called *weighted recombination* and is denoted by "w" in $(\mu/\mu_{\mathrm{w}}, \lambda)$.

The step-size is adapted using a simplified variant of the CSA rule [15], where the cumulation is deactivated. The update is given in Line 7 according to

$$\sigma_{t+1} = \sigma_t \exp^{\frac{1}{d_\sigma}} \left(\frac{\sqrt{\mu_{\mathrm{eff}}} \| \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)} \|}{E\|\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})\|} - 1 \right) \ . \tag{16}$$

The algorithm only uses the weighted sum $\sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)}$ of the best steps in the current iteration, as opposed to the sum of successive steps over the iterations in the original CSA. The norm of this weighted sum is compared to the expected norm of a standard multivariate normal variable by computing the ratio $\frac{\sqrt{\mu_{\mathrm{eff}}} \| \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)} \|}{E\|\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})\|}$, where μ_{eff} is a normalization factor, and the step-size σ_t is updated depending on the result of the comparison: if $\frac{\sqrt{\mu_{\mathrm{eff}}} \| \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)} \|}{E\|\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})\|} \geq 1$, suggesting that the progress is too slow, σ_t is increased. Otherwise, σ_t is decreased. A damping factor d_σ is used to attenuate the changes in σ_t values.

The vector γ_t of Lagrange factors is adapted in Line 8 according to

$$\gamma_{t+1} = \gamma_t + \frac{1}{d_\gamma} \omega_t \odot g(\mathrm{X}_{t+1}) \ . \tag{17}$$

A Lagrange factor γ_t^i is increased if the new estimate of the optimum X_{t+1} violates the corresponding constraint g_i, and decreased if X_{t+1} satisfies the constraint. A damping factor d_γ is used to control the changes of γ_t.

The vector ω_t of penalty factors is adapted in Line 9 as follows:

$$\omega_{t+1} = \omega_t \odot \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathrm{X}_{t+1})^2 < k_1 \frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t) - h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < |g_i(\mathrm{X}_t)| \\ \chi^{-1/d_\omega} & \text{otherwise} \end{cases} \right)_{i=1,\cdots,m} \tag{18}$$

This update extends the original update presented in [2] to the case of multiple constraints. A penalty factor ω_t^i is increased if the influence of the penalty part, $\omega_t^i g_i(\mathrm{X}_{t+1})^2$, in the h-value of the new mean vector X_{t+1} is too small. This is expressed by the first inequality in Line 9 where the penalty part is compared to the difference between h-values of X_t and X_{t+1}. A penalty factor is also increased if the difference in the corresponding constraint value $|g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)|$ is significantly smaller than $|g_i(\mathrm{X}_t)|$ (second inequality in Line 9). Increasing the penalty factor in this case favors the selection of solutions with smaller constraint values and, hence, solutions on the boundary of the feasible domain. For the sake of readability, we introduce the function $\mathcal{W}^{(f,g)} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^m$ defined as follows:

$$\mathcal{W}^{(f,g)}(\gamma, \omega, \mathrm{x}, \mathrm{y}) = \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega^i g_i(\mathrm{y})^2 < k_1 \frac{|h(\mathrm{y},\gamma,\omega) - h(\mathrm{x},\gamma,\omega)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{y}) - g_i(\mathrm{x})| < |g_i(\mathrm{x})| \\ \chi^{-1/d_\omega} & \text{otherwise} \end{cases} \right)_{i=1,\cdots,m} , \tag{19}$$

to define the update of ω_t in the remainder of this paper. The superscript (f, g) indicates the objective and the constraint functions that are used in h. Therefore, Line 9 in Algorithm 1 will simply read:

$$\omega_{t+1} = \omega_t \odot \mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) \ . \tag{20}$$

The $(\mu/\mu_\mathrm{w}, \lambda)$-CSA$_\mathrm{off}$-AL is an adaptive randomized algorithm with state $\theta_t = (\mathrm{X}_t, \sigma_t, \gamma_t, \omega_t)$. It is built from a comparison-based adaptive randomized algorithm for unconstrained optimization following the framework introduced in Section 3.1 (see (11) and (12) in particular). Given the current state θ_t and the vector $\mathrm{U}_{t+1} = \left(\mathrm{U}_{t+1}^1, \cdots, \mathrm{U}_{t+1}^\lambda\right)$ of i.i.d. normal vectors, the new state θ_{t+1} is given by

$$\theta_{t+1} = \mathcal{F}^{(f,g)}(\theta_t, \mathrm{U}_{t+1}) = \begin{pmatrix} \mathcal{G}((\mathrm{X}_t, \sigma_t), \mathrm{U}_{t+1}^\varsigma) \\ \mathcal{H}^{(f,g)}((\gamma_t, \omega_t), (\mathrm{X}_t, \sigma_t), \mathrm{U}_{t+1}^\varsigma) \end{pmatrix} , \tag{21}$$

Algorithm 1 The $(\mu/\mu_{\mathrm{w}}, \lambda)$-CSA$_{\mathrm{off}}$-AL

0 **input:** $n \in \mathbb{Z}_{>0}$, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$

1 **set** $\chi, k_1, k_2, d_\gamma, d_\omega, d_\sigma \in \mathbb{R}_{>0},\ \lambda, \mu \in \mathbb{Z}_{>0},\ 0 \leq w_i < 1,\ i = 1, \cdots, \mu,$
$\sum_{i=1}^{\mu} w_i = 1,\ \mu_{\mathrm{eff}} = 1/\sum_{i=1}^{\mu} w_i^2$ // constants

2 **define** $h(\mathrm{x}, \gamma, \omega) = f(\mathrm{x}) + \gamma^{\intercal} g(\mathrm{x}) + \frac{1}{2}\omega^{\intercal} g(\mathrm{x})^2$ // augmented Lagrangian

3 **initialize** $\mathrm{X}_0 \in \mathbb{R}^n,\ \sigma_0 \in \mathbb{R}_{>0},\ \gamma_0 \in \mathbb{R}^m,\ \omega_0 \in \mathbb{R}^m_{>0},\ t = 0$

4 **while** stopping criterion not met

$\mathrm{X}_{t+1}^i = \mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i$, where $\mathrm{U}_{t+1}^i \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n}),\ i = 1, \cdots, \lambda$
// sample λ i.i.d. candidate solutions

5 Extract the permutation ς of the indices $\{1, \cdots, \lambda\}$ such that:
$h(\mathrm{X}_{t+1}^{\varsigma(1)}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}_{t+1}^{\varsigma(\lambda)}, \gamma_t, \omega_t)$

6 $\mathrm{X}_{t+1} = \mathrm{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)}$ // update mean vector

7 $\sigma_{t+1} = \sigma_t \exp^{\frac{1}{d_\sigma}} \left(\frac{\sqrt{\mu_{\mathrm{eff}}} \| \sum_{i=1}^{\mu} w_i \mathrm{U}_{t+1}^{\varsigma(i)} \|}{E\|\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})\|} - 1 \right)$ // update step-size

8 $\gamma_{t+1} = \gamma_t + \dfrac{1}{d_\gamma} \omega_t \odot g(\mathrm{X}_{t+1})$ // update Lagrange factors

9 $\omega_{t+1} = \omega_t \odot \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathrm{X}_{t+1})^2 < k_1 \frac{|h(\mathrm{X}_{t+1}, \gamma_t, \omega_t) - h(\mathrm{X}_t, \gamma_t, \omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < |g_i(\mathrm{X}_t)| \\ \chi^{-1/d_\omega} & \text{otherwise} \end{cases} \right)_{i=1,\cdots,m}$
// update penalty factors

10 $t = t + 1$

where $\mathrm{U}_{t+1}^{\varsigma} = (\mathrm{U}_{t+1}^{\varsigma(1)}, \cdots, \mathrm{U}_{t+1}^{\varsigma(\lambda)})$ and ς satisfies (14). The function $\mathcal{G}$ updates the state variables X_t and σ_t while the function $\mathcal{H}$ (parameterized by the the objective function f and the constraint function g) updates the state variables γ_t and ω_t of the AL. Given the particular updates of the state variables used in the $(\mu/\mu_{\mathrm{w}}, \lambda)$-

CSA$_{\text{off}}$-AL, we can further write $\mathcal{G}$ and $\mathcal{H}$ as:

$$\mathcal{G}((\mathrm{X}_t, \sigma_t), \mathrm{U}^{\varsigma}_{t+1}) = \begin{pmatrix} \mathrm{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i \mathrm{U}^{\varsigma(i)}_{t+1} \\ \sigma_t \exp^{\frac{c_\sigma}{d_\sigma}} \left(\frac{\sqrt{\mu_{\text{eff}}} \| \sum_{i=1}^{\mu} w_i \mathrm{U}^{\varsigma(i)}_{t+1} \|}{E\|\mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})\|} - 1 \right) \end{pmatrix} ,$$

$$\mathcal{H}^{(f,g)}((\gamma_t, \omega_t), (\mathrm{X}_t, \sigma_t), \mathrm{U}^{\varsigma}_{t+1}) = \begin{pmatrix} \gamma_t + \frac{1}{d_\gamma} \omega_t \odot g(\underbrace{\mathrm{X}_t + \sigma_t \sum_{i=1}^{\mu} w_i \mathrm{U}^{\varsigma(i)}_{t+1}}_{\mathrm{X}_{t+1}}) \\ \omega_t \odot \mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) \end{pmatrix} .$$

3.3. Case Study: General Algorithm with Adaptive AL Constraint Handling

The $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL (Algorithm 1) is a particular case of a more general algorithm where the update rules for X_t and σ_t are given by deterministic functions $\mathcal{G}_{\text{x}}$ and $\mathcal{G}_\sigma$ according to

$$\mathrm{X}_{t+1} = \mathcal{G}_{\text{x}}((\mathrm{X}_t, \sigma_t), \mathrm{U}^{\varsigma}_{t+1}) \ , \tag{22}$$

$$\sigma_{t+1} = \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^{\varsigma}_{t+1}) \ , \tag{23}$$

and such that $\mathcal{G}_{\text{x}}$ and $\mathcal{G}_\sigma$ satisfy the following conditions [8]:

A2 For all $\mathrm{x}, \mathrm{x}_0 \in \mathbb{R}^n$, for all $\sigma > 0$, for all $\mathrm{y} \in \mathbb{R}^{n\times\lambda}$,

$$\mathcal{G}_{\text{x}}((\mathrm{x} + \mathrm{x}_0, \sigma), \mathrm{y}) = \mathcal{G}_{\text{x}}((\mathrm{x}, \sigma), \mathrm{y}) + \mathrm{x}_0 \ .$$

A3 For all $\mathrm{x} \in \mathbb{R}^n$, for all $\alpha, \sigma > 0$, for all $\mathrm{y} \in \mathbb{R}^{n\times\lambda}$,

$$\mathcal{G}_{\text{x}}((\mathrm{x}, \sigma), \mathrm{y}) = \alpha \mathcal{G}_{\text{x}}\left(\left(\frac{\mathrm{x}}{\alpha}, \frac{\sigma}{\alpha}\right), \mathrm{y}\right) \ .$$

A4 For all $\alpha, \sigma > 0$, for all $\mathrm{y} \in \mathbb{R}^{n\times\lambda}$,

$$\mathcal{G}_\sigma(\sigma, \mathrm{y}) = \alpha \mathcal{G}_\sigma\left(\frac{\sigma}{\alpha}, \mathrm{y}\right) \ .$$

In [8], the authors show that comparison-based adaptive algorithms for unconstrained optimization with update functions $\mathcal{G}_{\text{x}}$ and $\mathcal{G}_\sigma$ of the form (22) and (23) are translation-invariant and scale-invariant if conditions A2–A4 are satisfied.

This general algorithm—denoted GSAR-AL for General Step-size Adaptive Randomized algorithm with Augmented Lagrangian constraint handling—is defined by replacing Lines 6 and 7 in Algorithm 1 with the general update functions

$\mathcal{G}_{\mathrm{x}}$ (22) and $\mathcal{G}_\sigma$ (23) respectively, with the assumption that conditions A2–A4 are satisfied. The pseudocode is presented in Algorithm 2, where the main changes in comparison to Algorithm 1 are highlighted in grey. We conduct the analysis of invariance and linear convergence on the GSAR-AL. Therefore, our theoretical results are applicable to any AL-ES covered by the definition of the GSAR-AL, including the $(\mu/\mu_{\mathrm{w}}, \lambda)$-CSA$_{\mathrm{off}}$-AL.

Algorithm 2 The GSAR-AL

0 **input:** $n \in \mathbb{Z}_{>0}$, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^n \to \mathbb{R}^m$

1 **set** $\chi, k_1, k_2, d_\gamma, d_\omega \in \mathbb{R}_{>0},\ \lambda, \mu \in \mathbb{Z}_{>0}$ // constants

2 **define** $h(\mathrm{x}, \gamma, \omega) = f(\mathrm{x}) + \gamma^\intercal g(\mathrm{x}) + \frac{1}{2}\omega^\intercal g(\mathrm{x})^2$ // augmented Lagrangian

3 **initialize** $\mathrm{X}_0 \in \mathbb{R}^n,\ \sigma_0 \in \mathbb{R}_{>0},\ \gamma_0 \in \mathbb{R}^m,\ \omega_0 \in \mathbb{R}^m_{>0},\ t = 0$

4 **while** stopping criterion not met

$\mathrm{X}^i_{t+1} = \mathrm{X}_t + \sigma_t \mathrm{U}^i_{t+1}$, where $\mathrm{U}^i_{t+1} \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n}),\ i = 1, \cdots, \lambda$

// sample λ i.i.d. candidate solutions

5 Extract the permutation ς of the indices $\{1, \cdots, \lambda\}$ such that:

$h(\mathrm{X}^{\varsigma(1)}_{t+1}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}^{\varsigma(\lambda)}_{t+1}, \gamma_t, \omega_t)$

6 $\mathrm{X}_{t+1} = \mathcal{G}_{\mathrm{x}}((\mathrm{X}_t, \sigma_t), \mathrm{U}^\varsigma_{t+1})$ // update mean vector

7 $\sigma_{t+1} = \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^\varsigma_{t+1})$ // update step-size

8 $\gamma_{t+1} = \gamma_t + \dfrac{1}{d_\gamma}\omega_t \odot g(\mathrm{X}_{t+1})$ // update Lagrange factors

9 $\omega_{t+1} = \omega_t \odot \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega^i_t g_i(\mathrm{X}_{t+1})^2 < k_1 \frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t) - h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < |g_i(\mathrm{X}_t)| \\ \chi^{-1/d_\omega} & \text{otherwise} \end{cases} \right)_{i=1,\cdots,m}$

// update penalty factors

10 $t = t + 1$

4. Invariance and AL Methods

Invariance is an important notion in science. From a mathematical optimization perspective, when an algorithm is invariant, its performance on a particular

function can generalize to a whole class of functions. Comparison-based adaptive randomized algorithms for unconstrained optimization (see definition in (8) and (9)) are inherently invariant to strictly increasing transformations of the objective function f [8]. This is a direct consequence of their definition, as these algorithms use the objective function f only through the ranking of the candidate solutions according to their f-values. Therefore, if the objective function is $c \circ f$, where $c : \mathrm{Im}(f) \to \mathbb{R}$ is a strictly increasing function, the ranking remains unchanged compared to the ranking on f. This invariance is *unconditional* in that the optimization of f or $c \circ f$ results in the exact same sequence of states. The use of an AL to handle the constraints, however, breaks this invariance as we show in Section 4.1.

We discuss here invariance properties that an adaptive randomized algorithm with AL constraint handling should satisfy and formally define them. We distinguish between two types of invariances: invariance to transformations of the objective function f and the constraint function g, and invariance to transformations of the search space.

4.1. Invariance to Transformations of Objective and Constraint Functions

Let consider an AL adaptive randomized algorithm built from a comparison-based adaptive randomized algorithm for unconstrained optimization, as presented in Section 3.1. Although the algorithm for unconstrained optimization is invariant to strictly increasing transformations of the objective function f, the new algorithm does not preserve invariance to strictly increasing transformations of the objective and constraint functions, due to the use of an AL as the new objective function. Indeed, taking the AL in (6) as an example and assuming the number of constraints $m = 1$ (i.e. $g : \mathbb{R} \to \mathbb{R}$), it is easy to see that the ranking of two candidate solutions X^i_{t+1} and X^j_{t+1} on $h^{(f,g)}(\mathrm{x}, \gamma, \omega)$ may be different from their ranking on $h^{(c_1 \circ f, c_2 \circ g)}(\mathrm{x}, \gamma, \omega)$, where $c_1 : \mathrm{Im}(f) \to \mathbb{R}$ and $c_2 : \mathrm{Im}(g) \to \mathbb{R}$ are two strictly increasing functions and the upper script in h is used here to explicitly indicate the functions used by the AL. We illustrate this situation with the following example.

Example 1. Let $\gamma_t = \omega_t = 1$, $f(\mathrm{X}^i_{t+1}) = 5$, $g(\mathrm{X}^i_{t+1}) = 0.5$, $f(\mathrm{X}^j_{t+1}) = 6.5$, and $g(\mathrm{X}^j_{t+1}) = -0.5$. Let $c_1 : x \mapsto \frac{1}{2}x$ and $c_2 : x \mapsto x + 0.5$. Although $h(\mathrm{X}^i_{t+1}, \gamma_t, \omega_t) = 5.75 \leq h(\mathrm{X}^j_{t+1}, \gamma_t, \omega_t) = 6.25$, the ranking is inverted when considering $c_1 \circ f$ and $c_2 \circ g$, and we have: $h^{(c_1 \circ f, c_2 \circ g)}(\mathrm{X}^i_{t+1}, \gamma_t, \omega_t) = 4.5 > h^{(c_1 \circ f, c_2 \circ g)}(\mathrm{X}^j_{t+1}, \gamma_t, \omega_t) = 3.25$.

By looking at the ALs (5) and (6), however, we observe that the same ranking is obtained when the candidate solutions are evaluated on $h^{(f,g)}(\mathrm{x}, \gamma, \omega)$ than on $h^{(\alpha f+a_0,\beta g)}(\mathrm{x}, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega)$[5], for all $\alpha > 0$, $\beta > 0$, $a_0 \in \mathbb{R}$. In particular, we have

$$h^{(\alpha f+a_0,\beta g)}(\mathrm{x}, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega) = \alpha h^{(f,g)}(\mathrm{x}, \gamma, \omega) + a_0 \ .$$

Therefore, if $h^{(f,g)}(\mathrm{X}^i_{t+1}, \gamma, \omega) \leq h^{(f,g)}(\mathrm{X}^j_{t+1}, \gamma, \omega)$ for some candidate solutions X^i_{t+1}, X^j_{t+1}, then

$$h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^i_{t+1}, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega) \leq h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^j_{t+1}, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega) \ .$$

This latter property suggests that AL algorithms can be invariant to strictly increasing affine transformations of f and scaling of g—a subclass of strictly increasing transformations. We postulate this invariance as an important feature AL adaptive algorithms should have that is particularly important for linear convergence. A natural property to demand from an AL adaptive algorithm in this case is that the algorithm exhibits the same behavior when dealing with f and g as with $\alpha f + a_0$ (strictly increasing affine transformation) and βg (scaling), for all $\alpha > 0$, $\beta > 0$, $a_0 \in \mathbb{R}$. Formally, this consists in finding a bijective state-space transformation such that we obtain the same state when performing one step of the algorithm on (f, g) in the original state space as when performing one step on $(\alpha f + a_0, \beta g)$ in the transformed state space, then applying the inverse transformation to the resulting state. More formally we define:

Definition 1. An adaptive randomized algorithm with transition function $\mathcal{F}^{(f,g)} : \Omega \times \mathbb{R}^{n\times\lambda} \to \Omega$, where f is the objective function to minimize and g is the constraint function, is invariant to strictly increasing affine transformations of f and scaling of g if for all $\alpha > 0$, for all $\beta > 0$, for all $a_0 \in \mathbb{R}$, there exists a bijective state-space transformation $T_{\alpha,\beta,a_0} : \Omega \to \Omega$ such that for any objective function f, for any constraint function g, for any state $\theta \in \Omega$, and for any $\mathrm{u} \in \mathbb{R}^{n\times\lambda}$,

$$\mathcal{F}^{(f(\mathrm{x}),g(\mathrm{x}))}(\theta, \mathrm{u}) = T^{-1}_{\alpha,\beta,a_0}\left(\mathcal{F}^{(\alpha f(\mathrm{x})+a_0,\beta g(\mathrm{x}))}(T_{\alpha,\beta,a_0}(\theta), \mathrm{u})\right) \ .$$

[5] $\alpha f + a_0$ and βg are the functions defined as $\mathrm{x} \mapsto \alpha f(\mathrm{x}) + a_0$ and $\mathrm{x} \mapsto \beta g(\mathrm{x})$ respectively.

Invariance to strictly increasing affine transformations of f and scaling of g appears as an important property when we aim at linear convergence as it implies invariance on a scaled problem. We will see in Section 5 that indeed, together with other invariance properties, it allows to conclude on the existence of a homogeneous Markov chain whose stability implies the linear convergence of the algorithm.

4.2. Invariance to Transformations of the Search Space

We discuss here invariance to two particular transformations of the search space, namely translation-invariance and scale-invariance. In the case of comparison-based adaptive randomized algorithms for unconstrained optimization, translation-invariance and scale-invariance are tightly connected to linear convergence. In [8, 7], the authors show that if an algorithm is translation-invariant and scale-invariant, there exits a homogeneous Markov chain which can be used to prove the linear convergence of the algorithm.

4.2.1. Translation-Invariance

Translation-invariance is usually satisfied by most algorithms for unconstrained optimization. It expresses the non-sensitivity of an algorithm to the choice of its initial solution, that is, the algorithm's behavior when optimizing $\mathrm{x} \mapsto f(\mathrm{x})$ or $\mathrm{x} \mapsto f(\mathrm{x} - \mathrm{x}_0)$ is the same for any x_0. More formally, an algorithm is translation-invariant if there exists a bijective state-space transformation such that optimizing $\mathrm{x} \mapsto f(\mathrm{x})$ or $\mathrm{x} \mapsto f(\mathrm{x} - \mathrm{x}_0)$ is the same up to the state-space transformation. The following definition of translation-invariance is a generalization to the constrained case of the one in [8].

Definition 2. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,g)} : \Omega \times \mathbb{R}^{n\times\lambda} \to \Omega$, where f is the objective function to minimize and g is the constraint function, is translation-invariant if for all $\mathrm{x}_0 \in \mathbb{R}^n$, there exists a bijective state-space transformation $T_{\mathrm{x}_0} : \Omega \to \Omega$ such that for any objective function f, for any constraint function g, for any state $\theta \in \Omega$, and for any $\mathrm{u} \in \mathbb{R}^{n\times\lambda}$,

$$\mathcal{F}^{(f(\mathrm{x}),g(\mathrm{x}))}(\theta, \mathrm{u}) = T_{\mathrm{x}_0}^{-1}\left(\mathcal{F}^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(T_{\mathrm{x}_0}(\theta), \mathrm{u})\right) \quad .$$

4.2.2. Scale-Invariance

Scale-invariance informally translates the non-sensitivity of an algorithm to the scaling of the search space. When an algorithm is scale-invariant, there exists

a bijective state-space transformation such that optimizing $\mathrm{x} \mapsto f(\alpha \mathrm{x})$ is the same as optimizing $\mathrm{x} \mapsto f(\mathrm{x})$ up to the state-space transformation. The following definition of scale-invariance is a generalization to the constrained case to the one in [8].

Definition 3. A randomized adaptive algorithm with transition function $\mathcal{F}^{(f,g)} : \Omega \times \mathbb{R}^{n \times \lambda} \to \Omega$, where f is the objective function to minimize and g is the constraint function, is scale-invariant if for all $\alpha > 0$, there exists a bijective state-space transformation $T_\alpha : \Omega \to \Omega$ such that for any objective function f, for any constraint function g, for any state $\theta \in \Omega$, and for any $\mathrm{u} \in \mathbb{R}^{n \times \lambda}$,

$$\mathcal{F}^{(f(\mathrm{x}),g(\mathrm{x}))}(\theta, \mathrm{u}) = T_\alpha^{-1}\left(\mathcal{F}^{(f(\alpha \mathrm{x}),g(\alpha \mathrm{x}))}(T_\alpha(\theta), \mathrm{u})\right) \quad .$$

In the next section, we analyze the invariance properties we have defined for the specific case of the GSAR-AL and show how invariance is connected to linear convergence of the algorithm.

5. Invariance and Linear Convergence

We illustrate here the connection between invariance and linear convergence via the analysis of the GSAR-AL. We first analyze the invariance of the algorithm, then we show how this invariance can be used to define a homogeneous Markov chain whose stability implies the linear convergence of the algorithm.

We conduct the Markov chain analysis on a particular case of problem (1) where the constraints are linear, i.e.

A5 the constraint function $g : \mathbb{R}^n \to \mathbb{R}^m$ is defined as $g(\mathrm{x}) = \mathrm{A}\mathrm{x} + \mathrm{b}$, where $\mathrm{A} \in \mathbb{R}^{m \times n}$ is the matrix whose rows a_i represent the gradients of the linear constraints g_i, $i = 1, \cdots, m$, and $\mathrm{b} = (b_1, \cdots, b_m)^\intercal \in \mathbb{R}^m$.

We further assume that

A6 the problem has a unique global optimum x_{opt},

A7 the constraints are active at x_{opt}, i.e. $g(\mathrm{x}_{\text{opt}}) = \mathbf{0}$.

We argue that the case of active constraints is the most difficult in practice. Indeed, non-active constraints will typically not be "seen" by the algorithm when close enough to the optimum and can theoretically be ignored. Therefore, in the absence

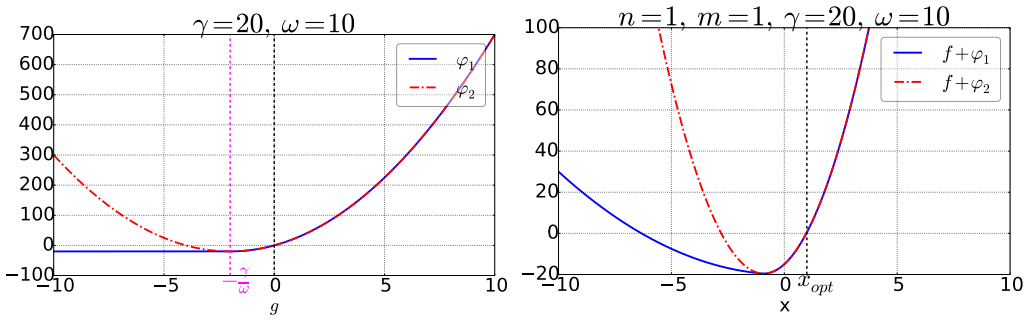

Figure 1: Left: $\varphi_j(g(\mathrm{x}), \gamma, \omega)$ for $j = 1$ (blue) and $j = 2$ (red), as a function of g (see (5) and (6)). Right: Augmented Lagrangians, $f(x) + \varphi_j(g(x), \gamma, \omega)$, for $j = 1$ (blue) and $j = 2$ (red), in $n = 1$ with $g : \mathbb{R}^n \to \mathbb{R}$ (one constraint). $f(x) = \frac{1}{2}x^2$, $g(x) = x - 1$, and $x_{\text{opt}} = 1$.

of active constraints, the algorithm will behave similarly to the unconstrained case, which is well-understood theoretically for step-size adaptive algorithms in the case of scaling-invariant functions [8]. We also assume that

A8 the gradient at x_{opt}, $\nabla f(\mathrm{x}_{\text{opt}})$, and the Jacobian $\nabla g(\mathrm{x}_{\text{opt}})$ exist,

A9 the constraints are linearly independent, that is, they satisfy the linear independence constraint qualification (LICQ). In this case, the vector γ_{opt} of Lagrange multipliers is unique [19].

For the analysis, we use the AL in (6) so that we can construct a Markov chain candidate for stability. Indeed, when the constraints are active at the optimum, ALs (5) and (6) are equivalent in the vicinity of the optimum, as illustrated in Figure 1 for $m = 1$. The left graph shows the functions φ_1 and φ_2 defined in (5) and (6) respectively plotted as functions of g while the right graph shows the ALs in $n = 1$ with $f(x) = \frac{1}{2}x^2$ and $g(x) = x - 1$.

5.1. Analysis of Invariance

We show that the GSAR-AL is invariant to strictly increasing affine transformations of the objective function f and scaling of the constraint function g, and is also translation-invariance and scale-invariance. For each invariance property, we show the existence of a bijective state-space transformation such that Definitions 1–3 are satisfied.

The following result states the algorithm's invariance to strictly increasing affine transformations of the objective function f and scaling of the constraint function g.

Proposition 1. *The GSAR-AL is invariant to affine transformations of the objective function f, $x \mapsto \alpha f(x) + a_0$, and to the scaling of the constraint function g, $x \mapsto \beta g(x)$, and for all $\alpha, \beta > 0$, for all $a_0 \in \mathbb{R}$, the associated state-space transformation T_{α,β,a_0} is defined as*

$$T_{\alpha,\beta,a_0}(x, \sigma, \gamma, \omega) = (x, \sigma, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega) \ , \tag{24}$$

for all $x \in \mathbb{R}^n$, for all $\sigma \in \mathbb{R}$, and for all $\gamma, \omega \in \mathbb{R}^m$.

This invariance stems from the updates of the AL parameters γ_t and ω_t used in the GSAR-AL, as shown in the proof of Proposition 1 in Appendix A. The next result states that the GSAR-AL is translation-invariant.

Proposition 2. *Assume that condition A2 holds. The GSAR-AL is translation-invariant and for all $x_0 \in \mathbb{R}^n$, the associated state-space transformation T_{x_0} is given by*

$$T_{x_0}(x, \sigma, \gamma, \omega) = (x + x_0, \sigma, \gamma, \omega) \ , \tag{25}$$

for all $x \in \mathbb{R}^n$, for all $\sigma \in \mathbb{R}$, and for all $\gamma, \omega \in \mathbb{R}^m$.

Translation-invariance stems from property A2 of the update function $\mathcal{G}_{\mathrm{x}}$, as shown in the proof of Proposition 2 presented in Appendix A. The next result states scale-invariance of the GSAR-AL.

Proposition 3. *Assume that conditions A3–A4 hold. The GSAR-AL is scale-invariant and for all $\alpha > 0$, the associated state-space transformation T_α is defined as*

$$T_\alpha(x, \sigma, \gamma, \omega) = (x/\alpha, \sigma/\alpha, \gamma, \omega) \ , \tag{26}$$

for all $x \in \mathbb{R}^n$, for all $\sigma \in \mathbb{R}$, and for all $\gamma, \omega \in \mathbb{R}^m$.

Scale-invariance results from properties A3–A4 of the update functions $\mathcal{G}_{\mathrm{x}}$ and $\mathcal{G}_\sigma$. The proof of Proposition 3 is available in Appendix A.

5.2. Analysis of Linear Convergence

Linear convergence of an algorithm roughly speaking states that $\|\mathrm{X}_t - \mathrm{x}_{\mathrm{opt}}\|$ decreases to zero geometrically. There are, however, several (non fully equivalent ways) to formulate linear convergence for a stochastic algorithm. We consider in the sequel almost sure convergence and, therefore, formulate asymptotic linear convergence of X_t towards $\mathrm{x}_{\mathrm{opt}}$ as

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathrm{X}_t - \mathrm{x}_{\mathrm{opt}}\|}{\|\mathrm{X}_0 - \mathrm{x}_{\mathrm{opt}}\|} = -\mathrm{CR} \ \text{ almost surely} \ ,$$

where CR $\in \mathbb{R}$ (positive when convergence occurs) is called the *convergence rate*.

Taking the GSAR-AL whose invariance is established in Propositions 1–3, we demonstrate how linear convergence to the optimal solution $\mathrm{x}_{\mathrm{opt}}$ and to the corresponding vector of Lagrange multipliers γ_{opt} can be deduced by exploiting translation-invariance, scale-invariance, and invariance to strictly increasing affine transformations of the objective function and scaling of the constraint function.

To analyze linear convergence, we adopt the Markov chain approach described in [8, 7] for the unconstrained case. Informally, a discrete-time Markov chain is a sequence $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ of (multivariate) random variables that satisfies the Markov property, that is, the conditional distribution of Φ_{t+1} given the past states, $\Phi_0, \cdots, \Phi_t$, depends only on Φ_t. We consider *time homogeneous* Markov chains where the conditional distribution of Φ_{t+1} given Φ_t does not depend on t. In our context, we consider homogeneous Markov chains that follow the so-called nonlinear state-space model [11], where

$$\Phi_{t+1} = F(\Phi_t, \mathrm{U}_{t+1}) \ ,$$

with F being a measurable function and $\{\mathrm{U}_{t+1} : t \in \mathbb{Z}_{>0}\}$ a sequence of i.i.d. random vectors. More definitions related to Markov chains are provided in Appendix B for completeness.

Markov chain theory [18] provides powerful tools to prove linear convergence of randomized algorithms whose state is a Markov chain. The general approach consists in finding a class of objective functions for which an underlying homogeneous Markov chain candidate to be "stable"[6] exists, then proving the stability of the identified Markov chain. The convergence rate can then be expressed as a function of the stable Markov chain and linear convergence follows from a LLN for Markov chains[7] [8, 7, 6, 3, 9].

In the following section, we illustrate the Markov chain approach to analyze linear convergence on a general comparison-based adaptive algorithm for unconstrained optimization.

5.2.1. Linear Convergence via Markov Chain Analysis: Unconstrained Case

We consider the comparison-based adaptive randomized algorithm for unconstrained optimization defined in (8) and (9), with state $\theta_t = (\mathrm{X}_t, \sigma_t)$ and update

[6] In our context, we define stability as positivity and Harris-recurrence.

[7] The LLN generalizes to Markov chains if stability (i.e. positivity and Harris-recurrence) holds. See Theorem 3 in Appendix B.

functions $\mathcal{G}_{\mathrm{x}}$ (22) and $\mathcal{G}_\sigma$ (23) that satisfy conditions A2–A4. We assume, for the sake of simplicity, the minimization of a convex quadratic function

$$f(\mathrm{x}) = \frac{1}{2}\mathrm{x}^\intercal \mathrm{H}\mathrm{x} \ , \tag{27}$$

with optimum in zero, without loss of generality. This algorithm, which is the unconstrained version of the GSAR-AL, is translation-invariant and scale-invariant [8] as a result of properties A2–A4 of its update functions. Consequently, the sequence $\{\mathrm{Y}_t = \frac{\mathrm{X}_t}{\sigma_t} : t \in \mathbb{Z}_{\geq 0}\}$ is a homogeneous Markov chain that can be defined independently of (X_t, σ_t), given $\mathrm{Y}_0 = \frac{\mathrm{X}_0}{\sigma_0}$, as

$$\mathrm{Y}_{t+1} = \frac{\mathcal{G}_{\mathrm{x}}((\mathrm{Y}_t, 1), \mathrm{U}_{t+1}^\varsigma)}{\mathcal{G}_\sigma(1, \mathrm{U}_{t+1}^\varsigma)} \ ,$$

where the permutation ς results from the ranking of $\{\mathrm{Y}_t + \mathrm{U}_{t+1}^i : i = 1, \cdots, \lambda\}$ on f [8, Proposition 4.1] (this result is true more generally for scaling-invariant objective functions).

Using the property of the logarithm, the decrease of the log-distance to the optimum (zero here) normalized by t, that is $\frac{1}{t} \ln \frac{\|\mathrm{X}_t\|}{\|\mathrm{X}_0\|}$, can be expressed as a function of Y_t as follows:

$$\begin{aligned}
\frac{1}{t} \ln \frac{\|\mathrm{X}_t\|}{\|\mathrm{X}_0\|} &= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathrm{X}_{k+1}\|}{\|\mathrm{X}_k\|} = \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathrm{X}_{k+1}\|}{\|\mathrm{X}_k\|} \frac{\sigma_k \mathcal{G}_\sigma(1, \mathrm{U}_{k+1}^\varsigma)}{\sigma_{k+1}} \\
&= \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathrm{Y}_{k+1}\|}{\|\mathrm{Y}_k\|} \mathcal{G}_\sigma(1, \mathrm{U}_{k+1}^\varsigma) \ ,
\end{aligned} \tag{28}$$

where we have successively artificially introduced $\sigma_{k+1} = \sigma_k \mathcal{G}_\sigma(1, \mathrm{U}_{k+1}^\varsigma)$ then used that $\mathrm{Y}_k = \mathrm{X}_k / \sigma_k$ and $\mathrm{Y}_{k+1} = \mathrm{X}_{k+1} / \sigma_{k+1}$. In (28), we have expressed the term whose limit we are interested in as the empirical average of a function of a Markov chain. However, we know from Markov chain theory that if some sufficient stability conditions—given for instance in Theorem 17.0.1 from [18]—are satisfied by $\{\mathrm{Y}_t : t \in \mathbb{Z}_{\geq 0}\}$, then a LLN for Markov chains can be applied to the right-hand side of the previous equation. Consequently,

$$\begin{aligned}
\lim_{t \to \infty} \frac{1}{t} \ln \frac{\|\mathrm{X}_t\|}{\|\mathrm{X}_0\|} &= \lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t-1} \ln \frac{\|\mathrm{Y}_{k+1}\|}{\|\mathrm{Y}_k\|} \mathcal{G}_\sigma(1, \mathrm{U}_{k+1}^\varsigma) = \int \ln \|\mathrm{y}\| \pi(d\mathrm{y}) \\
&\quad - \int \ln \|\mathrm{y}\| \pi(d\mathrm{y}) + \underbrace{\int E(\ln(\mathcal{G}_\sigma(1, \mathrm{U}_{t+1}^\varsigma)) | \mathrm{Y}_t = \mathrm{y}) \pi(d\mathrm{y})}_{-\mathrm{CR}} \ ,
\end{aligned}$$

where π is the invariant probability measure of the Markov chain $\{\mathrm{Y}_t : t \in \mathbb{Z}_{\geq 0}\}$. Hence, assuming that a LLN holds for the Markov chain $\{\mathrm{Y}_t : t \in \mathbb{Z}_{\geq 0}\}$, the algorithm described by the iterative sequence $\{(\mathrm{X}_t, \sigma_t) : t \in \mathbb{Z}_{\geq 0}\}$ will converge linearly at a rate CR expressed as minus the expected log step-size change (where the expectation is taken with respect to the invariant probability measure of $\{\mathrm{Y}_t : t \in \mathbb{Z}_{\geq 0}\}$).

This methodology to prove the linear convergence of adaptive algorithms (including many ESs) in the unconstrained case holds on scaling-invariant functions (that include particularly functions that write $c \circ f$, where c is a 1-D strictly increasing function and f is positive homogeneous). Translation-invariance and scale-invariance are key elements in the analysis. Indeed, the existence of a homogeneous Markov chain that is candidate to be stable stems from both translation-invariance and scale-invariance.

In the following section, we apply the Markov chain approach described above to investigate linear convergence of the GSAR-AL.

5.2.2. Generalization to Constrained Optimization

We show here the existence of an underlying homogeneous Markov chain for the GSAR-AL. For the sake of clarity, we present the analysis on convex quadratic objective functions with optimum in zero (see definition in (27)) before presenting more general results. Our study consists in identifying the homogeneous Markov chain by exploiting the invariance of the algorithm. Then, we *assume* the stability of the Markov chain to deduce linear convergence. The stability of the identified Markov chain is investigated numerically in Section 6.

When the objective function is convex quadratic and the constraint functions are linear, KKT conditions are sufficient conditions for optimality, that is, a KKT point is also the global optimum $\mathrm{x}_{\mathrm{opt}}$ of the constrained problem [19]. Since we assume the constraint functions to be linearly independent, the associated Lagrange multiplier to $\mathrm{x}_{\mathrm{opt}}$ is unique, and we denote it γ_{opt}.

The following theorem defines the homogeneous Markov chain.

Theorem 1. *Consider the GSAR-AL defined in Algorithm 2 solving the constrained problem (1), where f is convex quadratic with optimum in zero, i.e. $f(x) = \frac{1}{2} x^{\intercal} H x$. Let $\{(X_t, \sigma_t, \gamma_t, \omega_t) : t \in \mathbb{Z}_{\geq 0}\}$ be the Markov chain associated to the algorithm and assume conditions A1–A4, A5, A7, and A9 are satisfied. Let*

$$Y_t = \frac{X_t - x_{opt}}{\sigma_t} \quad \textit{and} \quad \Gamma_t = \frac{\gamma_t - \gamma_{\mathrm{opt}}}{\sigma_t} \quad , \tag{29}$$

where x_{opt} is the global optimum of the constrained problem and γ_{opt} is the associated vector of Lagrange multipliers. Then, the sequence $\{\Phi_t = (Y_t, \Gamma_t, \omega_t) : t \in \mathbb{Z}_{\geq 0}\}$ is a homogeneous Markov chain that can be defined independently of $(X_t, \sigma_t, \gamma_t, \omega_t)$ as $Y_0 = (X_0 - x_{opt})/\sigma_0$, $\Gamma_0 = (\gamma_0 - \gamma_{\text{opt}})/\sigma_0$, and for all t

$$Y_{t+1} = \mathcal{G}_x((Y_t, 1), U^{\varsigma}_{t+1}) / \mathcal{G}_\sigma(1, U^{\varsigma}_{t+1}) \ , \tag{30}$$

$$\Gamma_{t+1} = \frac{\Gamma_t + \frac{1}{d_\gamma}\omega_t \odot g(\tilde{Y}_{t+1} + x_{opt})}{\mathcal{G}_\sigma(1, U^{\varsigma}_{t+1})} \ , \tag{31}$$

$$\omega_{t+1} = \omega_t \odot \mathcal{W}^{(f(x+x_{opt}), g(x+x_{opt}))}(\Gamma_t + \gamma_{\text{opt}}, \omega_t, Y_t, \tilde{Y}_{t+1}) \ , \tag{32}$$

where

$$\tilde{Y}_{t+1} = Y_{t+1}\mathcal{G}_\sigma(1, U^{\varsigma}_{t+1}) \ , \tag{33}$$

and the permutation ς extracts the indices of the ordered vectors $\{Y_t + U^i_{t+1} : i = 1, \cdots, \lambda\}$ on $h(x + x_{opt}, \Gamma_t + \gamma_{\text{opt}}, \omega_t)$, i.e. ς satisfies

$$h(Y_t + U^{\varsigma(1)}_{t+1} + x_{opt}, \Gamma_t + \gamma_{\text{opt}}, \omega_t) \leq \cdots \leq h(Y_t + U^{\varsigma(\lambda)}_{t+1} + x_{opt}, \Gamma_t + \gamma_{\text{opt}}, \omega_t) \ . \tag{34}$$

The proof of Theorem 1 is given in Appendix A. A key idea of the proof is that on a convex quadratic objective function f, the same permutation ς is obtained when ranking candidate solutions $\{\mathrm{X}_t + \sigma_t \mathrm{U}^i_{t+1} : i = 1, \cdots, \lambda\}$ on $h(\mathrm{x}, \gamma_t, \omega_t)$ than when ranking $\{\mathrm{Y}_t + \mathrm{U}^i_{t+1} : i = 1, \cdots, \lambda\}$ on $h(\mathrm{x} + \mathrm{x}_{\text{opt}}, \Gamma_t + \gamma_{\text{opt}}, \omega_t)$. The existence of the homogeneous Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ is due to translation-invariance and scale-invariance of the GSAR-AL, as well as to the particular updates of the parameters γ_t and ω_t of the AL. While translation-invariance and scale-invariance are explicitly exploited to build the Markov chain (to compute Y_{t+1}), the effect of invariance to strictly increasing affine transformations of f and scaling of g is implicit through the update rules for γ_t and ω_t.

The next theorem gives sufficient stability conditions on the Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ under which the GSAR-AL converges linearly to the optimum x_{opt} and to the associated vector of Lagrange multipliers γ_{opt}. Following the same reasoning as in Section 5.2.1, we show that if stability conditions hold for the constructed Markov chain, the sequence $\{\mathrm{X}_t : t \in \mathbb{Z}_{\geq 0}\}$ converges linearly to x_{opt} at the same speed the sequence $\{\sigma_t : t \in \mathbb{Z}_{\geq 0}\}$ converges to zero. Additionally, the sequence $\{\gamma_t : t \in \mathbb{Z}_{\geq 0}\}$ of Lagrange factors converges linearly, and also at the same speed, to γ_{opt}.

Theorem 2. *Let $\{(X_t, \sigma_t, \gamma_t, \omega_t) : t \in \mathbb{Z}_{\geq 0}\}$ be the Markov chain associated to the GSAR-AL defined in Algorithm 2, optimizing the augmented Lagrangian h defined in (6) associated to the constrained problem (1), where the objective function is convex quadratic with optimum in zero, i.e. $f(x) = \frac{1}{2}x^\intercal Hx$, and where x_{opt} is the global optimum of the problem and γ_{opt} is the associated vector of Lagrange multipliers. We assume that the algorithm satisfies conditions A1–A4 and that the optimization problem satisfies conditions A5, A7, and A9. Let $\{\Phi_t = (Y_t, \Gamma_t, \omega_t) : t \in \mathbb{Z}_{\geq 0}\}$ be the Markov chain defined in Theorem 1 and assume that it is positive Harris-recurrent with invariant probability measure π, that $E_\pi(|\ln \|[\phi]_1\||) < \infty$, $E_\pi(|\ln \|[\phi]_2\||) < \infty$, and $E_\pi(|\mathcal{R}(\phi)|) < \infty$, where*

$$\mathcal{R}(\phi) = E\big(\ln(\mathcal{G}_\sigma(1, U_{t+1}^\varsigma))|\Phi_t = \phi\big) \ . \tag{35}$$

Then for all X_0, for all σ_0, for all γ_0, and for all ω_0,

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|X_t - x_{opt}\|}{\|X_0 - x_{opt}\|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\gamma_t - \gamma_{\text{opt}}\|}{\|\gamma_0 - \gamma_{\text{opt}}\|} = \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$$
$$= -CR \ \textit{almost surely} \ ,$$

where $-CR = \int \mathcal{R}(\phi)\pi(d\phi)$.

The proof idea for Theorem 2 is similar to the one discussed in Section 5.2.1 for the unconstrained case, where the quantities $\frac{1}{t} \ln \frac{\|\mathrm{X}_t - \mathrm{x}_{\text{opt}}\|}{\|\mathrm{X}_0 - \mathrm{x}_{\text{opt}}\|}$, $\frac{1}{t} \ln \frac{\|\gamma_t - \gamma_{\text{opt}}\|}{\|\gamma_0 - \gamma_{\text{opt}}\|}$, and $\frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$ are expressed as a function of the Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$. The detailed proof is given in Appendix A.

5.2.3. More General Results

The result stated in Theorem 1 for convex quadratic objective functions is a particular case of a more general result. In fact, the sequence $\{\Phi_t = (\mathrm{Y}_t, \Gamma_t, \omega_t) : t \in \mathbb{Z}_{\geq 0}\}$, where

$$\mathrm{Y}_t = \frac{\mathrm{X}_t - \bar{\mathrm{x}}}{\sigma_t} \quad \text{and} \quad \Gamma_t = \frac{\gamma_t - \bar{\gamma}}{\sigma_t} \ ,$$

and where $\bar{\mathrm{x}}$ is any vector in $\mathbb{R}^n$ that satisfies $g(\bar{\mathrm{x}}) = \mathbf{0}$ and $\bar{\gamma}$ is any vector in $\mathbb{R}^m$, is a homogeneous Markov chain on the class of objective functions f such that the following condition holds:

A10 The function $\mathcal{D}h_{\bar{\mathrm{x}},\bar{\gamma},\omega} : (\mathrm{x}, \gamma) \mapsto h(\mathrm{x}, \gamma, \omega) - h(\bar{\mathrm{x}}, \bar{\gamma}, \omega)$ is *positive homogeneous* of degree 2 with respect to $(\bar{\mathrm{x}}, \bar{\gamma})$, for all $\omega \in \mathbb{R}^m_{>0}$,

where a function $p : X \to Y$ is positive homogeneous of degree $k > 0$ with respect to $\mathrm{x}^* \in X$ if for all $\alpha > 0$ and for all $\mathrm{x} \in X$, $p(\mathrm{x}^* + \alpha\mathrm{x}) = \alpha^k p(\mathrm{x}^* + \mathrm{x})$.

We generalize the results presented in Section 5.2.2 by replacing the assumption that the objective function f is convex quadratic with condition A10 in Theorems 1–2, and x_{opt} and γ_{opt} with $\bar{\mathrm{x}}$ and $\bar{\gamma}$ respectively in Theorem 1. We also assume that conditions A6 and A8 hold in Theorems 1–2. In this case, the proof of Theorem 1 generalizes by exploiting the positive homogeneity of the newly defined function $\mathcal{D}h_{\bar{\mathrm{x}},\bar{\gamma},\omega}$ (instead of the explicit expression of f) to write the permutation ς and the vector of penalty factors ω_t, in particular, as a function of the Markov chain. As for Theorem 2, the proof remains unchanged.

In the next section, we numerically verify the linear convergence of the $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL—an instance of the GSAR-AL—and the stability of the Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ on the linearly constrained sphere and ellipsoid functions.

6. Numerical Results

We evaluate the $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL (Algorithm 1) on two linearly constrained convex quadratic functions: the sphere, f_{sphere}, and the ellipsoid, $f_{\text{ellipsoid}}$, with a moderate condition number. These functions are defined according to (27) by taking $\mathrm{H} = \mathrm{I}_{n\times n}$ for f_{sphere} and H diagonal with diagonal elements $[\mathrm{H}]_{ii} = \alpha^{\frac{i-1}{n-1}}$, $i = 1, \cdots, n$, for $f_{\text{ellipsoid}}$ and with a condition number $\alpha = 10$.

We choose x_{opt} to be at $(10, \cdots, 10)^\intercal$ and construct the (active) linear constraints following the steps below:

1. For the first constraint, the normal $\mathrm{a}_1 = -\nabla f(\mathrm{x}_{\text{opt}})^\intercal$ and $b_1 = -\mathrm{a}_1^\intercal \mathrm{x}_{\text{opt}}$,
2. For the $m-1$ remaining constraints, we choose the constraint normal a_i as a standard multivariate normal variable ($\mathrm{a}_i \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_{n\times n})$) and $b_i = -\mathrm{a}_i^\intercal \mathrm{x}_{\text{opt}}$. We choose the point $\nabla f(\mathrm{x}_{\text{opt}})^\intercal = -\mathrm{a}_1$ to be feasible along with x_{opt}. Therefore, if $g_i(\nabla f(\mathrm{x}_{\text{opt}})^\intercal) > 0$, we modify a_i and b_i according to: $\mathrm{a}_i = -\mathrm{a}_i$ and $b_i = -b_i$.

With the construction above, the constraints are linearly independent with probability one and the unique vector of Lagrange multipliers associated to x_{opt} is $\gamma_{\text{opt}} = (1, 0, \cdots, 0)^\intercal$.

As for the parameters of the $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL, we choose the default values in [14] for both λ and μ. We set the weights w_i, $i = 1, \cdots, \mu$, according to [1], where they are chosen to be optimal on the sphere function in infinite dimension.

We set $d_\sigma = 2 + 2\max\left(0, \sqrt{\frac{1/\sum_{k=1}^{\mu} w_k^2 - 1}{n+1}} - 1\right)$ as recommended in [14]. We take $d_\gamma = d_\omega = 5$, $\chi = 2^{1/n}$, $k_1 = 3$, and $k_2 = 5$.

We run the $(\mu/\mu_\mathrm{w}, \lambda)$-CSA$_\mathrm{off}$-AL and simulate the Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ defined in Theorem 1 in $n = 10$ on f_sphere and $f_\mathrm{ellipsoid}$ with $m \in \{1, 2, 5, 9\}$ constraints. For each problem, we test three different initial values of the penalty vector $\omega_0 \in \{(1, \cdots, 1)^\intercal, (10^3, \cdots, 10^3)^\intercal, (10^{-3}, \cdots, 10^{-3})^\intercal\}$. In all the tests, X_0 and Y_0 are sampled uniformly in $[-5, 5]^n$, $\sigma_0 = 1$, and $\gamma_0 = \Gamma_0 = (5, \cdots, 5)^\intercal$. We discuss in this section results for $m \in \{1, 9\}$ and $\omega_0 = (1, \cdots, 1)^\intercal$. The remaining results are given in Appendix C.

Figure 2 shows simulations of the Markov chain on f_sphere (left column) and $f_\mathrm{ellipsoid}$ (right column) subject to 1 constraint (top row) and 9 constraints (bottom row). Displayed are the normalized distance to x_opt, $\|\mathrm{Y}_t\|$ (red), the normalized distance to γ_opt, $\|\Gamma_t\|$ (green), and the norm of the vector of penalty factors, $\|\omega_t\|$ (blue) in log-scale for $\omega_0 = (1, \cdots, 1)^\intercal$. In both cases, we observe an overall convergence to a stationary distribution after a certain number of iterations. This adaptation phase before reaching the stationary state is overall longer for larger values of ω_0 on both f_sphere and $f_\mathrm{ellipsoid}$. It also increases with increasing m (compare Figures C.4 and C.7 for example).

Figure 3 shows single runs of the $(\mu/\mu_\mathrm{w}, \lambda)$-CSA$_\mathrm{off}$-AL on the same constrained problems described previously. Results on constrained f_sphere and constrained $f_\mathrm{ellipsoid}$ are displayed in left and right columns respectively, for $m = 1$ (top row) and $m = 9$ (bottom row). The displayed quantities are the distance to the optimum, $\|\mathrm{X}_t - \mathrm{x}_\mathrm{opt}\|$ (red), the distance to the Lagrange multipliers, $\|\gamma_t - \gamma_\mathrm{opt}\|$ (green), the norm of the penalty vector, $\|\omega_t\|$ (blue), and the step-size, σ_t (purple), in log-scale. Linear convergence occurs after an adaptation phase whose length depends on the accuracy of the choice of the initial parameters. We observe that the number of iterations needed to reach a given precision increases with the number of constraints: it takes more than twice longer to reach a distance to the optimum of 10^{-4} on both f_sphere and $f_\mathrm{ellipsoid}$ with $m = 9$ than with $m = 1$. These results are consistent with the simulations of the Markov chain in that the observed stability of the Markov chain leads to linear convergence of the algorithm, as stated in Theorem 2.

7. Discussion

We discussed throughout this work the connection between invariance and linear convergence of randomized adaptive algorithms for constrained optimization

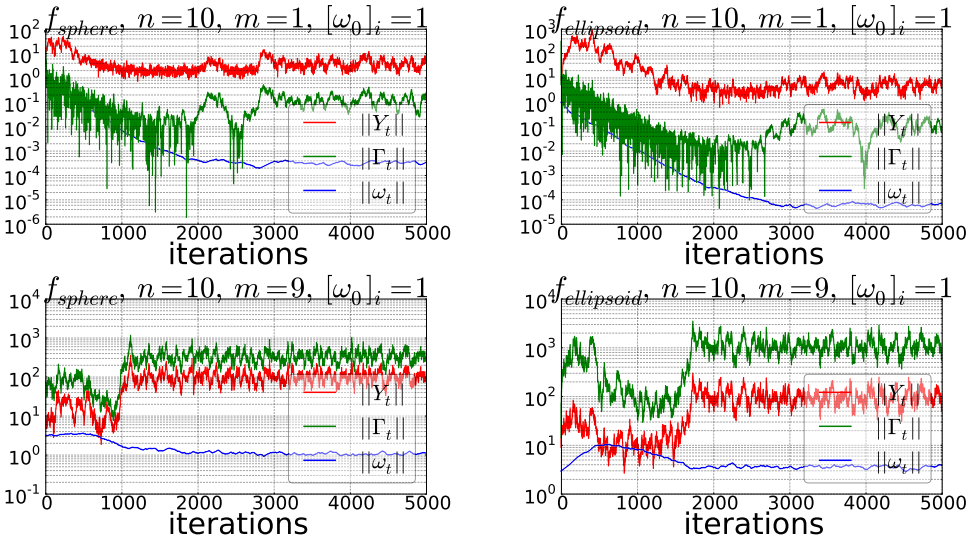

Figure 2: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ (top) and $m = 9$ (bottom) in $n = 10$.

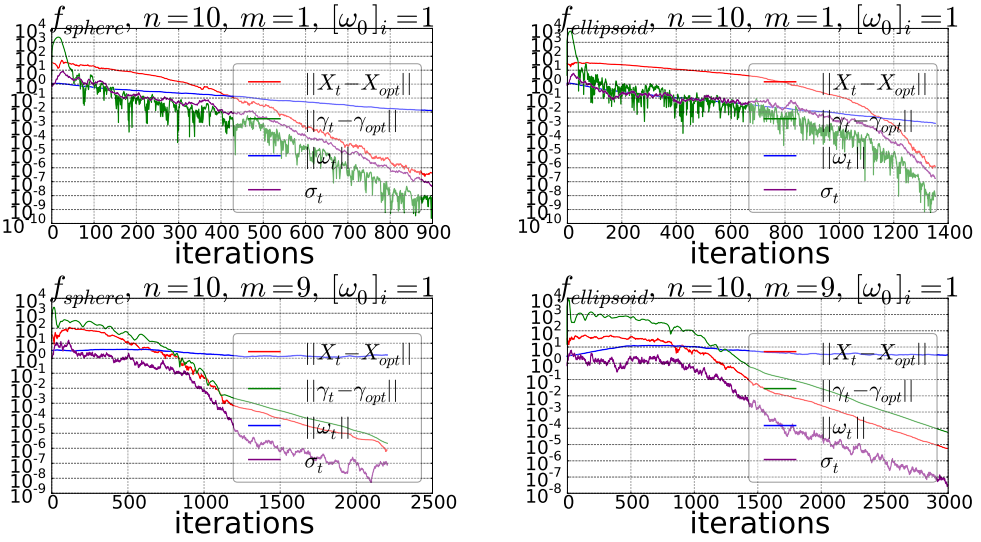

Figure 3: Single runs of the $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ (top) and $m = 9$ (bottom) in $n = 10$.

when the constraints are handled with an augmented Lagrangian approach.

We formalized invariance properties for augmented Lagrangian algorithms that are important to achieve linear convergence. We showed that although un-

conditional invariance to strictly increasing transformations of the objective and the constraint functions does no longer hold due to the use of an augmented Lagrangian, invariance to a subclass of these transformations—namely strictly increasing affine transformations of the objective function and scaling of the constraints—is still achievable.

We presented a general framework for building augmented Lagrangian algorithms from adaptive randomized algorithms for unconstrained optimization, then used this framework to instantiate a practical evolution strategy, the $(\mu/\mu_{\mathrm{w}}, \lambda)$-$\mathrm{CSA}_{\mathrm{off}}$-AL, as well as a more general step-size adaptive algorithm, the GSAR-AL, which we analyzed.

We showed that the GSAR-AL is invariant to strictly increasing affine transformations of the objective function and scaling of the constraints, and is also translation-invariant and scale-invariant. Following a Markov chain approach, we illustrated how these invariance properties can lead to linear convergence of the algorithm in the case of linear inequality constraints. In this case, the existence of a homogeneous Markov that can be used to deduce linear convergence under sufficient stability conditions, is a consequence of the algorithm's invariance. We exhibited a class of objective functions on which such a Markov chain exists. This class includes convex quadratic functions and is defined such that the augmented Lagrangian, centered at the optimum $\mathrm{x}_{\mathrm{opt}}$ and the corresponding vector of Lagrange multipliers γ_{opt}, is positive homogeneous of degree two. The stability of the constructed Markov chain, as well as linear convergence of the practical $(\mu/\mu_{\mathrm{w}}, \lambda)$-$\mathrm{CSA}_{\mathrm{off}}$-AL, were validated numerically on the linearly constrained sphere and ellipsoid functions.

Acknowledgments

This work was supported by the PGMO Numerical Black-Box Optimization for Energy Applications (NumBER) project and by the grant ANR-2012-MONU-0009 (NumBBO) from the French National Research Agency.

[1] D. V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms*, pages 215–237. Springer, 2005.

[2] D. V. Arnold and J. Porter. Towards an Augmented Lagrangian Constraint Handling Approach for the $(1 + 1)$-ES. In *Genetic and Evolutionary Computation Conference*, pages 249–256. ACM Press, 2015.

[3] A. Atamna, A. Auger, and N. Hansen. Analysis of Linear Convergence of a $(1+1)$-ES with Augmented Lagrangian Constraint Handling. In *Genetic and Evolutionary Computation Conference*, pages 213–220. ACM Press, 2016.

[4] A. Atamna, A. Auger, and N. Hansen. Augmented Lagrangian Constraint Handling for CMA-ES—Case of a Single Linear Constraint. In *Parallel Problem Solving from Nature*, pages 181–191. Springer, 2016.

[5] A. Atamna, A. Auger, and N. Hansen. Linearly Convergent Evolution Strategies via Augmented Lagrangian Constraint Handling. In *Proceedings of the 14th ACM/SIGEVO Conference on Foundations of Genetic Algorithms*, pages 149–161. ACM, 2017.

[6] A. Auger. Convergence Results for the $(1, \lambda)$-SA-ES Using the Theory of φ-Irreducible Markov Chains. *Theoretical Computer Science*, 334(1-3):35–69, 2005.

[7] A. Auger and N. Hansen. Linear Convergence on Positively Homogeneous Functions of a Comparison Based Step-Size Adaptive Randomized Search: the $(1 + 1)$ ES with Generalized One-Fifth Success Rule. arXiv:1310.8397, 2013.

[8] A. Auger and N. Hansen. Linear Convergence of Comparison-Based Step-Size Adaptive Randomized Search via Stability of Markov Chains. *SIAM Journal on Optimization*, 26(3):1589–1624, 2016.

[9] A. Bienvenüe and O. François. Global Convergence of Evolution Strategies in Spherical Problems: Some Simple Proofs and Difficulties. *Theoretical Computer Science*, 306(1–3):269–289, 2003.

[10] E. G. Birgin, C. A. Floudas, and J. M. Martínez. Global Minimization Using an Augmented Lagrangian Method with Variable Lower-Level Constraints. *Mathematical Programming*, 125(1):139–162, 2010.

[11] A. Chotard and A. Auger. Verifiable Conditions for Irreducibility, Aperiodicity and T-chain Property of a General Markov Chain. Accepted for publication in Bernoulli, 2015.

[12] A. R. Conn, N. I. M. Gould, and P. 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.

[13] K. Deb and S. Srivastava. A Genetic Algorithm Based Augmented Lagrangian Method for Constrained Optimization. *Computational Optimization and Applications*, 53(3):869–902, 2012.

[14] N. Hansen. The CMA Evolution Strategy: A Tutorial. http://arxiv.org/pdf/1604.00772v1.pdf, 2016.

[15] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[16] M. R. Hestenes. Multiplier and Gradient Methods. *Journal of Optimization Theory and Applications*, 4(5):303–320, 1969.

[17] R. M. Lewis and V. Torczon. A Globally Convergent Augmented Lagrangian Pattern Search Algorithm for Optimization with General Constraints and Simple Bounds. *SIAM Journal on Optimization*, 12(4):1075–1089, 2002.

[18] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, 1993.

[19] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, 2nd edition, 2006.

[20] M. J. D. Powell. A Method for Nonlinear Constraints in Minimization Problems. In R. Fletcher, editor, *Optimization*, pages 283–298. Academic Press, 1969.

[21] M.-J. Tahk and B.-C. Sun. Coevolutionary Augmented Lagrangian Methods for Constrained Optimization. *IEE Transactions on Evolutionary Computation*, 4(2):114–124, 2000.

Appendix A. Proofs

Appendix A.1. Proof of Proposition 1

Starting from the transformed state

$$(\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t) = T_{\alpha,\beta,a_0}(\mathrm{X}_t, \sigma_t, \gamma_t, \omega_t) \ ,$$

where T_{α,β,a_0} is defined in (24), we compute the new state

$$(\mathrm{X}'_{t+1}, \sigma'_{t+1}, \gamma'_{t+1}, \omega'_{t+1}) = \mathcal{F}^{(\alpha f+a_0,\beta g)}((\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t), \mathrm{U}_{t+1}) \ ,$$

considering the augmented Lagrangian in (6) where we replace f and g with $\alpha f + a_0$ and βg respectively, that is, we consider

$$h^{(\alpha f+a_0,\beta g)}(\mathrm{x}, \gamma, \omega) := \alpha f(\mathrm{x}) + a_0 + \beta\gamma^\intercal g(\mathrm{x}) + \frac{\beta^2}{2}\omega^\intercal g(\mathrm{x})^2 \ .$$

We have

$$\mathrm{X}'_{t+1} = \mathcal{G}_\mathrm{x}((\mathrm{X}'_t, \sigma'_t), \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\mathrm{x}((\mathrm{X}_t, \sigma_t), \mathrm{U}^\varsigma_{t+1}) \ , \tag{A.1}$$

$$\sigma'_{t+1} = \mathcal{G}_\sigma(\sigma'_t, \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^\varsigma_{t+1}) \ , \tag{A.2}$$

where the permutation ς extracts the indices of the candidate solutions ranked on $h^{(\alpha f+a_0,\beta g)}$, i.e. ς satisfies

$$\begin{gathered} h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^{\varsigma(1)}_{t+1}, \gamma'_t, \omega'_t) \leq \cdots \leq h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^{\varsigma(\lambda)}_{t+1}, \gamma'_t, \omega'_t) \\ \Leftrightarrow \\ h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^{\varsigma(1)}_{t+1}, \frac{\alpha}{\beta}\gamma_t, \frac{\alpha}{\beta^2}\omega_t) \leq \cdots \leq h^{(\alpha f+a_0,\beta g)}(\mathrm{X}^{\varsigma(\lambda)}_{t+1}, \frac{\alpha}{\beta}\gamma_t, \frac{\alpha}{\beta^2}\omega_t) \ . \end{gathered} \tag{A.3}$$

Notice however that

$$h^{(\alpha f+a_0,\beta g)}(\mathrm{x}, \frac{\alpha}{\beta}\gamma, \frac{\alpha}{\beta^2}\omega) = \alpha h(\mathrm{x}, \gamma, \omega) + a_0 \ , \tag{A.4}$$

for all $\mathrm{x} \in \mathbb{R}^n$, for all $\gamma, \omega \in \mathbb{R}^m$, for all $\alpha, \beta > 0$, and for all $a_0 \in \mathbb{R}$. Therefore, (A.3) is equivalent to

$$h(\mathrm{X}^{\varsigma(1)}_{t+1}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}^{\varsigma(\lambda)}_{t+1}, \gamma_t, \omega_t) \ . \tag{A.5}$$

Consequently, (A.1) and (A.2) become

$$\mathrm{X}'_{t+1} = \mathcal{G}_\mathrm{x}((\mathrm{X}_t, \sigma_t), \mathrm{U}^\varsigma_{t+1}) = \mathrm{X}_{t+1} \ , \tag{A.6}$$

$$\sigma'_{t+1} = \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^\varsigma_{t+1}) = \sigma_{t+1} \ , \tag{A.7}$$

where the permutation ς satisfies (A.5).

Using the definition of γ_{t+1} in (17), we have

$$\gamma'_{t+1} = \gamma'_t + \frac{1}{d_\gamma}\omega'_t \odot g(\mathrm{X}'_{t+1}) = \frac{\alpha}{\beta}\gamma_t + \frac{\alpha}{\beta d_\gamma}\omega_t \odot g(\mathrm{X}_{t+1}) = \frac{\alpha}{\beta}\gamma_{t+1} \ . \tag{A.8}$$

Using the definition of ω_{t+1} in (19) and (20) and exploiting (A.4), we obtain

$$\begin{aligned}
\omega'_{t+1} &= \omega'_t \odot \mathcal{W}^{(\alpha f + a_0, \beta g)}(\gamma'_t, \omega'_t, \mathrm{X}'_t, \mathrm{X}'_{t+1}) \\
&= \frac{\alpha}{\beta^2}\omega_t \odot \mathcal{W}^{(\alpha f + a_0, \beta g)}(\frac{\alpha}{\beta}\gamma_t, \frac{\alpha}{\beta^2}\omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) \\
&= \frac{\alpha}{\beta^2}\omega_t \odot \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \alpha\omega_t^i g_i(\mathrm{X}_{t+1})^2 < k_1 \times \\ & \qquad \alpha \frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t) - h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } \beta k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < \beta |g_i(\mathrm{X}_t)| \\ \chi^{-1/d_\omega} & \text{otherwise,} \end{cases} \right)_{i=1,\cdots,m} \\
&= \frac{\alpha}{\beta^2}\omega_t \odot \mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) = \frac{\alpha}{\beta^2}\omega_{t+1} \ .
\end{aligned} \tag{A.9}$$

By applying the inverse transformation $T^{-1}_{\alpha,\beta,a_0} : (\mathrm{x}, \sigma, \gamma, \omega) \mapsto (\mathrm{x}, \sigma, \frac{\beta}{\alpha}\gamma, \frac{\beta^2}{\alpha}\omega)$ to (A.6), (A.7), (A.8), and (A.9), we recover the new state in the initial state space, $(\mathrm{X}_{t+1}, \sigma_{t+1}, \gamma_{t+1}, \omega_{t+1})$.

Appendix A.2. Proof of Proposition 2

Starting from $(\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t) = T_{\mathrm{x}_0}(\mathrm{X}, \sigma_t, \gamma_t, \omega_t)$, where T_{x_0} is defined in (25), and considering $f(\mathrm{x} - \mathrm{x}_0)$ and $g(\mathrm{x} - \mathrm{x}_0)$, we have

$$(\mathrm{X}'_{t+1}, \sigma'_{t+1}, \gamma'_{t+1}, \omega'_{t+1}) = \mathcal{F}^{(f(\mathrm{x}-\mathrm{x}_0), g(\mathrm{x}-\mathrm{x}_0))}((\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t), \mathrm{U}_{t+1}) \ ,$$

where X'_{t+1}, σ'_{t+1}, γ'_{t+1}, and ω'_{t+1} are defined according to (22), (23), (17), and (20) and (19) respectively. Using (25) and the translation property of $\mathcal{G}_\mathrm{x}$ in A2, we have

$$\begin{aligned}
\mathrm{X}'_{t+1} &= \mathcal{G}_\mathrm{x}((\mathrm{X}'_t, \sigma'_t), \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\mathrm{x}((\mathrm{X}_t + \mathrm{x}_0, \sigma_t), \mathrm{U}^\varsigma_{t+1}) \\
&= \mathcal{G}_\mathrm{x}((\mathrm{X}_t, \sigma_t), \mathrm{U}^\varsigma_{t+1}) + \mathrm{x}_0 \ , \\
\sigma'_{t+1} &= \mathcal{G}_\sigma(\sigma'_t, \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^\varsigma_{t+1}) \ ,
\end{aligned}$$

where the permutation ς is extracted by ranking the candidate solutions $\{\mathrm{X}_t + \mathrm{x}_0 + \sigma_t \mathrm{U}^i_{t+1} : i = 1, \cdots, \lambda\}$ on $h^{(f(\mathrm{x}-\mathrm{x}_0), g(\mathrm{x}-\mathrm{x}_0))}(\mathrm{x}, \gamma'_t, \omega'_t)$, where $\gamma'_t = \gamma_t$ and $\omega'_t = \omega_t$,

i.e. ς satisfies the following:

$$h^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(\mathrm{X}_t+\mathrm{x}_0+\sigma_t\mathrm{U}_{t+1}^{\varsigma(1)},\gamma_t,\omega_t)\leq\cdots \\ \leq h^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(\mathrm{X}_t+\mathrm{x}_0+\sigma_t\mathrm{U}_{t+1}^{\varsigma(\lambda)},\gamma_t,\omega_t)\ . \quad (A.10)$$

On the other hand, we have by definition of h in (6)

$$h^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(\mathrm{x},\gamma,\omega)=h(\mathrm{x}-\mathrm{x}_0,\gamma,\omega)\ ,$$

for all $\mathrm{x}\in\mathbb{R}^n$, for all $\gamma,\omega\in\mathbb{R}^m$. Therefore, (A.10) is equivalent to:

$$h(\mathrm{X}_t+\sigma_t\mathrm{U}_{t+1}^{\varsigma(1)},\gamma_t,\omega_t)\leq\cdots\leq h(\mathrm{X}_t+\sigma_t\mathrm{U}_{t+1}^{\varsigma(\lambda)},\gamma_t,\omega_t)\ ,$$

meaning that the same permutation ς is extracted as when ranking candidate solutions $\{\mathrm{X}_t+\sigma_t\mathrm{U}_{t+1}^i : i=1,\cdots,\lambda\}$ on the original augmented Lagrangian h. It follows that

$$\mathrm{X}'_{t+1}=\mathrm{X}_{t+1}+\mathrm{x}_0\ , \quad (A.11)$$
$$\sigma'_{t+1}=\sigma_{t+1}\ . \quad (A.12)$$

Using the definition of γ_{t+1} in (17), as well as (25) and (A.11), we have

$$\gamma'_{t+1}=\gamma'_t+\frac{1}{d_\gamma}\omega'_t\odot g(\mathrm{X}'_{t+1}-\mathrm{x}_0)=\gamma_{t+1}\ . \quad (A.13)$$

As for ω'_{t+1}, we have from (19), (20), (25), and (A.11)

$$\begin{aligned}\omega'_{t+1}&=\omega'_t\odot\mathcal{W}^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(\gamma'_t,\omega'_t,\mathrm{X}'_t,\mathrm{X}'_{t+1})\\&=\omega_t\odot\mathcal{W}^{(f(\mathrm{x}-\mathrm{x}_0),g(\mathrm{x}-\mathrm{x}_0))}(\gamma_t,\omega_t,\mathrm{X}_t+\mathrm{x}_0,\mathrm{X}_{t+1}+\mathrm{x}_0)\\&=\omega_t\odot\left(\begin{cases}\chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathrm{X}_{t+1})^2<k_1\times\frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t)-h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n}\\ & \text{or } k_2|g_i(\mathrm{X}_{t+1})-g_i(\mathrm{X}_t)|<|g_i(\mathrm{X}_t)|\\ \omega_t^i\chi^{-1/d_\omega} & \text{otherwise, } i=1,\cdots,m\end{cases}\right)_{i=1,\cdots,m}\\&=\omega_t\odot\mathcal{W}^{(f,g)}(\gamma_t,\omega_t,\mathrm{X}_t,\mathrm{X}_{t+1})=\omega_{t+1}\ . \quad (A.14)\end{aligned}$$

By applying the inverse transformation $T_{\mathrm{x}_0}^{-1}:(\mathrm{x},\sigma,\gamma,\omega)\mapsto(\mathrm{x}-\mathrm{x}_0,\sigma,\gamma,\omega)$ to (A.11), (A.12), (A.13), and (A.14), we recover the new state in the initial state space, $(\mathrm{X}_{t+1},\sigma_{t+1},\gamma_{t+1},\omega_{t+1})$.

Appendix A.3. Proof of Proposition 3

Starting from $(\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t) = T_\alpha(\mathrm{X}, \sigma_t, \gamma_t, \omega_t)$, where T_α is defined in (26), and considering $f(\alpha \mathrm{x})$ and $g(\alpha \mathrm{x})$, we have

$$(\mathrm{X}'_{t+1}, \sigma'_{t+1}, \gamma'_{t+1}, \omega'_{t+1}) = \mathcal{F}^{(f(\alpha \mathrm{x}), g(\alpha \mathrm{x}))}((\mathrm{X}'_t, \sigma'_t, \gamma'_t, \omega'_t), \mathrm{U}_{t+1}) \ ,$$

where X'_{t+1}, σ'_{t+1}, γ'_{t+1}, and ω'_{t+1} are defined according to (22), (23), (17), and (20) and (19) respectively. Using the definition of T_α in (26) and the properties of $\mathcal{G}_\mathrm{x}$ and $\mathcal{G}_\sigma$ in A3 and A4 respectively, it follows:

$$\begin{aligned}
\mathrm{X}'_{t+1} &= \mathcal{G}_\mathrm{x}((\mathrm{X}'_t, \sigma'_t), \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\mathrm{x}((\mathrm{X}_t/\alpha, \sigma_t/\alpha), \mathrm{U}^\varsigma_{t+1}) \\
&= \frac{1}{\alpha} \mathcal{G}_\mathrm{x}((\mathrm{X}_t, \sigma_t), \mathrm{U}^\varsigma_{t+1}) \ , \\
\sigma'_{t+1} &= \mathcal{G}_\sigma(\sigma'_t, \mathrm{U}^\varsigma_{t+1}) = \mathcal{G}_\sigma(\sigma_t/\alpha, \mathrm{U}^\varsigma_{t+1}) = \frac{1}{\alpha} \mathcal{G}_\sigma(\sigma_t, \mathrm{U}^\varsigma_{t+1}) \ .
\end{aligned}$$

The permutation ς is extracted by ranking the candidate solutions $\{\frac{\mathrm{X}_t}{\alpha} + \frac{\sigma_t}{\alpha} \mathrm{U}^i_{t+1} : i = 1, \cdots, \lambda\}$ on $h^{(f(\alpha \mathrm{x}), g(\alpha \mathrm{x}))}(\mathrm{x}, \gamma'_t, \omega'_t)$, where $\gamma'_t = \gamma_t$ and $\omega'_t = \omega_t$, and we have

$$\begin{aligned}
h^{(f(\alpha \mathrm{x}), g(\alpha \mathrm{x}))} \left(\frac{\mathrm{X}_t}{\alpha} + \frac{\sigma_t}{\alpha} \mathrm{U}^{\varsigma(1)}_{t+1}, \gamma_t, \omega_t \right) \leq \cdots \\
\leq h^{(f(\alpha \mathrm{x}), g(\alpha \mathrm{x}))} \left(\frac{\mathrm{X}_t}{\alpha} + \frac{\sigma_t}{\alpha} \mathrm{U}^{\varsigma(\lambda)}_{t+1}, \gamma_t, \omega_t \right) \ . \quad \text{(A.15)}
\end{aligned}$$

From (6), however,

$$h^{(f(\alpha \mathrm{x}), g(\alpha \mathrm{x}))}(\mathrm{x}, \gamma, \omega) = h(\alpha \mathrm{x}, \gamma, \omega) \ ,$$

for all $\mathrm{x} \in \mathbb{R}^n$, for all $\gamma, \omega \in \mathbb{R}^m$. Therefore, (A.15) is equivalent to:

$$h(\mathrm{X}_t + \sigma_t \mathrm{U}^{\varsigma(1)}_{t+1}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}_t + \sigma_t \mathrm{U}^{\varsigma(\lambda)}_{t+1}, \gamma_t, \omega_t) \ .$$

This means that the same permutation ς is obtained as when ranking the candidate solutions $\{\mathrm{X}_t + \sigma_t \mathrm{U}^i_{t+1} : i = 1, \cdots, \lambda\}$ on h and consequently

$$\mathrm{X}'_{t+1} = \mathrm{X}_{t+1}/\alpha \ , \quad \text{(A.16)}$$

$$\sigma'_{t+1} = \sigma_{t+1}/\alpha \ . \quad \text{(A.17)}$$

Using (17), (26), and (A.16), we have

$$\gamma'_{t+1} = \gamma'_t + \frac{1}{d_\gamma} \omega'_t \odot g(\alpha \mathrm{X}'_{t+1}) = \gamma_{t+1} \ . \quad \text{(A.18)}$$

Finally, we have from (19), (20), (26), and (A.16)

$$\begin{aligned}
\omega'_{t+1} &= \omega'_t \odot \mathcal{W}^{(f(\alpha \mathrm{x}),g(\alpha \mathrm{x}))}(\gamma'_t, \omega'_t, \mathrm{X}'_t, \mathrm{X}'_{t+1}) \\
&= \omega_t \odot \mathcal{W}^{(f(\alpha \mathrm{x}),g(\alpha \mathrm{x}))}(\gamma_t, \omega_t, \mathrm{X}_t/\alpha, \mathrm{X}_{t+1}/\alpha) \\
&= \omega_t \odot \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathrm{X}_{t+1})^2 < k_1 \times \frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t) - h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < |g_i(\mathrm{X}_t)| \\ \omega_t^i \chi^{-1/d_\omega} & \text{otherwise, } i = 1, \cdots, m \end{cases} \right)_{i=1,\cdots,m} \\
&= \omega_t \odot \mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) = \omega_{t+1} \ . \qquad \text{(A.19)}
\end{aligned}$$

By applying the inverse transformation $T_\alpha^{-1} : (\mathrm{x}, \sigma, \gamma, \omega) \mapsto (\alpha \mathrm{x}, \alpha\sigma, \gamma, \omega)$ to (A.16), (A.17), (A.18), and (A.19), we obtain the new state in the original search space, $(\mathrm{X}_{t+1}, \sigma_{t+1}, \gamma_{t+1}, \omega_{t+1})$.

Appendix A.4. Proof of Theorem 1

We express $\Phi_{t+1} = (\mathrm{Y}_{t+1}, \Gamma_{t+1}, \omega_{t+1})$ as a function of $\Phi_t = (\mathrm{Y}_t, \Gamma_t, \omega_t)$ and the i.i.d. random vectors $\mathrm{U}_{t+1} = (\mathrm{U}_{t+1}^1, \cdots, \mathrm{U}_{t+1}^\lambda)$.

According to (29), we have

$$\mathrm{Y}_{t+1} = \frac{\mathrm{X}_{t+1} - \mathrm{x}_{\mathrm{opt}}}{\sigma_{t+1}} = \frac{\mathcal{G}_{\mathrm{x}}((\mathrm{X}_t, \sigma_t), \mathrm{U}_{t+1}^\varsigma) - \mathrm{x}_{\mathrm{opt}}}{\mathcal{G}_\sigma(\sigma_t, \mathrm{U}_{t+1}^\varsigma)} = \frac{\mathcal{G}_{\mathrm{x}}((\mathrm{Y}_t, 1), \mathrm{U}_{t+1}^\varsigma)}{\mathcal{G}_\sigma(1, \mathrm{U}_{t+1}^\varsigma)} \ , \qquad \text{(A.20)}$$

where we used translation-invariance and scale-invariance of Algorithm 2.

The permutation ς is extracted by ranking the candidate solutions $\{\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i : i = 1, \cdots, \lambda\}$ on h, that is, the following inequality holds:

$$h(\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^{\varsigma(1)}, \gamma_t, \omega_t) \leq \cdots \leq h(\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^{\varsigma(\lambda)}, \gamma_t, \omega_t) \ . \qquad \text{(A.21)}$$

We observe, however, that the h-value of a candidate solution $\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i$ satisfies:

$$\begin{aligned}
h(\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i, \gamma_t, \omega_t) &= h(\sigma_t(\mathrm{Y}_t + \mathrm{U}_{t+1}^i) + \mathrm{x}_{\mathrm{opt}}, \sigma_t \Gamma_t + \gamma_{\mathrm{opt}}, \omega_t) \\
&= \underbrace{f(\sigma_t(\mathrm{Y}_t + \mathrm{U}_{t+1}^i) + \mathrm{x}_{\mathrm{opt}})}_{A} \\
&\quad + \underbrace{(\sigma_t \Gamma_t + \gamma_{\mathrm{opt}})^\intercal g(\sigma_t(\mathrm{Y}_t + \mathrm{U}_{t+1}^i) + \mathrm{x}_{\mathrm{opt}})}_{B} \\
&\quad + \underbrace{\frac{1}{2} \omega_t^\intercal g(\sigma_t(\mathrm{Y}_t + \mathrm{U}_{t+1}^i) + \mathrm{x}_{\mathrm{opt}})^2}_{C} \ ,
\end{aligned}$$

where we used the definitions of Y_t and Γ_t in (29). By developing A, B, and C, and using the definitions of f and $g(\mathrm{x}) = \mathrm{Ax} + \mathrm{b}$, the first KKT condition in (3), and the fact that $\nabla f(\mathrm{x}) = \mathrm{x}^\intercal \mathrm{H}$ and $\nabla g(\mathrm{x}) = \mathrm{A}$, we obtain:

$$A = \sigma_t^2 f(\mathrm{Y}_t + \mathrm{U}_{t+1}^i + \mathrm{x}_{\text{opt}}) + (1-\sigma_t^2) f(\mathrm{x}_{\text{opt}}) + \sigma_t(1-\sigma_t) \underbrace{\mathrm{x}_{\text{opt}}^\intercal \mathrm{H}}_{\nabla f(\mathrm{x}_{\text{opt}})} (\mathrm{Y}_{t+1} + \mathrm{U}_{t+1}^i) \ ,$$

$$B = \sigma_t^2 (\Gamma_t + \gamma_{\text{opt}})^\intercal g(\mathrm{Y}_{t+1} + \mathrm{U}_{t+1}^i + \mathrm{x}_{\text{opt}}) + \sigma_t(1-\sigma_t)\gamma_{\text{opt}}^\intercal \underbrace{\mathrm{A}}_{\nabla g(\mathrm{x}_{\text{opt}})} (\mathrm{Y}_{t+1} + \mathrm{U}_{t+1}^i) \ ,$$

$$C = \frac{\sigma_t^2}{2} \omega_t^\intercal g(\mathrm{Y}_{t+1} + \mathrm{U}_{t+1}^i + \mathrm{x}_{\text{opt}})^2 \ .$$

Therefore

$$\begin{aligned} h(\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i, \gamma_t, \omega_t) &= h(\sigma_t(\mathrm{Y}_t + \mathrm{U}_{t+1}^i) + \mathrm{x}_{\text{opt}}, \sigma_t \Gamma_t + \gamma_{\text{opt}}, \omega_t) \\ &= \sigma_t^2 h(\mathrm{Y}_t + \mathrm{U}_{t+1}^i + \mathrm{x}_{\text{opt}}, \Gamma_t + \gamma_{\text{opt}}, \omega_t) \\ &+ \underbrace{(1-\sigma_t^2) f(\mathrm{x}_{\text{opt}})}_{\text{constant}} \ , \end{aligned} \tag{A.22}$$

meaning that the same permutation ς is obtained when ranking the vectors $\{\mathrm{Y}_t + \mathrm{U}_{t+1}^i : i = 1, \cdots, \lambda\}$ on $h(\mathrm{x} + \mathrm{x}_{\text{opt}}, \Gamma_t + \gamma_{\text{opt}}, \omega_t)$ as when ranking the candidate solutions $\{\mathrm{X}_t + \sigma_t \mathrm{U}_{t+1}^i : i = 1, \cdots, \lambda\}$ on $h(\mathrm{x}, \gamma_t, \omega_t)$. That is, ς also satisfies

$$h(\mathrm{Y}_t + \mathrm{U}_{t+1}^{\varsigma(1)} + \mathrm{x}_{\text{opt}}, \Gamma_t + \gamma_{\text{opt}}, \omega_t) \leq \cdots \leq h(\mathrm{Y}_t + \mathrm{U}_{t+1}^{\varsigma(\lambda)} + \mathrm{x}_{\text{opt}}, \Gamma_t + \gamma_{\text{opt}}, \omega_t) \ .$$

According to (29), we also have

$$\Gamma_{t+1} = \frac{\gamma_{t+1} - \gamma_{\text{opt}}}{\sigma_{t+1}} = \frac{\gamma_t + \frac{1}{d_\omega} \omega_t \odot g(\mathrm{X}_{t+1}) - \gamma_{\text{opt}}}{\mathcal{G}_\sigma(\sigma_t, \mathrm{U}_{t+1}^\varsigma)} \ .$$

Using the definitions of X_{t+1}, g, and $\tilde{\mathrm{Y}}_{t+1}$ in (22), A5, and (33) respectively, along with translation-invariance and scale-invariance of Algorithm 2, we obtain

$$\Gamma_{t+1} = \frac{\Gamma_t + \frac{1}{d_\gamma} \omega_t \odot g(\tilde{\mathrm{Y}}_{t+1} + \mathrm{x}_{\text{opt}})}{\mathcal{G}_\sigma(1, \mathrm{U}_{t+1}^\varsigma)} \ . \tag{A.23}$$

Finally, we have by definition:

$$\omega_{t+1} = \omega_t \odot \mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) \ ,$$

with

$$\mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) = \left(\begin{cases} \chi^{1/(4d_\omega)} & \text{if } \omega_t^i g_i(\mathrm{X}_{t+1})^2 < k_1 \times \frac{|h(\mathrm{X}_{t+1},\gamma_t,\omega_t) - h(\mathrm{X}_t,\gamma_t,\omega_t)|}{n} \\ & \text{or } k_2 |g_i(\mathrm{X}_{t+1}) - g_i(\mathrm{X}_t)| < |g_i(\mathrm{X}_t)| \\ \chi^{-1/d_\omega} & \text{otherwise,} \end{cases} \right)_{i=1,\cdots,m} . \tag{A.24}$$

Using the definitions of X_{t+1}, Y_t, Γ_t, and $\tilde{\mathrm{Y}}_{t+1}$, along with translation-invariance and scale-invariance of the algorithm, we have

$$\begin{aligned} h(\mathrm{X}_{t+1}, \gamma_t, \omega_t) &= h(\mathcal{G}_{\mathrm{x}}((\mathrm{X}_t, \sigma_t), \mathrm{U}_{t+1}^{\varsigma}), \gamma_t, \omega_t) \\ &= h(\sigma_t \tilde{\mathrm{Y}}_{t+1} + \mathrm{x}_{\mathrm{opt}}, \sigma_t \Gamma_t + \gamma_{\mathrm{opt}}, \omega_t) \ . \end{aligned} \tag{A.25}$$

Using (A.22), we deduce that

$$\begin{aligned} h(\mathrm{X}_{t+1}, \gamma_t, \omega_t) - h(\mathrm{X}_t, \gamma_t, \omega_t) = \sigma_t^2 (& h(\tilde{\mathrm{Y}}_{t+1} + \mathrm{x}_{\mathrm{opt}}, \Gamma_t + \gamma_{\mathrm{opt}}, \omega_t) \\ & - h(\mathrm{Y}_t + \mathrm{x}_{\mathrm{opt}}, \Gamma_t + \gamma_{\mathrm{opt}}, \omega_t)) \ . \end{aligned}$$

On the other hand, we have by the definition of g

$$\begin{aligned} g_i(\mathrm{X}_t) &= \sigma_t g_i(\mathrm{Y}_t + \mathrm{x}_{\mathrm{opt}}) \\ g_i(\mathrm{X}_{t+1}) &= \sigma_t g_i(\tilde{\mathrm{Y}}_{t+1} + \mathrm{x}_{\mathrm{opt}}) \ . \end{aligned}$$

Replacing in (A.24), we obtain

$$\mathcal{W}^{(f,g)}(\gamma_t, \omega_t, \mathrm{X}_t, \mathrm{X}_{t+1}) = \mathcal{W}^{(f(\mathrm{x}+\mathrm{x}_{\mathrm{opt}}), g(\mathrm{x}+\mathrm{x}_{\mathrm{opt}}))}(\Gamma_t + \gamma_{\mathrm{opt}}, \omega_t, \mathrm{Y}_t, \tilde{\mathrm{Y}}_{t+1}) \ ,$$

and therefore

$$\omega_{t+1} = \omega_t \odot \mathcal{W}^{(f(\mathrm{x}+\mathrm{x}_{\mathrm{opt}}), g(\mathrm{x}+\mathrm{x}_{\mathrm{opt}}))}(\Gamma_t + \gamma_{\mathrm{opt}}, \omega_t, \mathrm{Y}_t, \tilde{\mathrm{Y}}_{t+1}) \ .$$

$\Phi_{t+1} = (\mathrm{Y}_{t+1}, \Gamma_{t+1}, \omega_{t+1})$ depends only on $\Phi_t = (\mathrm{Y}_t, \Gamma_t, \omega_t)$ and the i.i.d. random vectors U_{t+1}. Therefore, $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ is a homogeneous Markov chain.

Appendix A.5. Proof of Theorem 2

We express $\frac{1}{t} \ln \frac{\|\mathrm{X}_t - \mathrm{x}_{\mathrm{opt}}\|}{\|\mathrm{X}_0 - \mathrm{x}_{\mathrm{opt}}\|}$, $\frac{1}{t} \ln \frac{\|\gamma_t - \gamma_{\mathrm{opt}}\|}{\|\gamma_0 - \gamma_{\mathrm{opt}}\|}$, and $\frac{1}{t} \ln \frac{\sigma_t}{\sigma_0}$ as a function of the homogeneous Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ defined in Theorem 1. Using the property

of the logarithm, we have

$$\begin{aligned}\frac{1}{t}\ln\frac{\|\mathrm{X}_t-\mathrm{x}_{\mathrm{opt}}\|}{\|\mathrm{X}_0-\mathrm{x}_{\mathrm{opt}}\|}&=\frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathrm{X}_{k+1}-\mathrm{x}_{\mathrm{opt}}\|}{\|\mathrm{X}_k-\mathrm{x}_{\mathrm{opt}}\|}=\frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathrm{Y}_{k+1}\|}{\|\mathrm{Y}_k\|}\mathcal{G}_\sigma(1,\mathrm{U}_{k+1}^{\varsigma})\\&=\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathrm{Y}_{k+1}\|-\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathrm{Y}_k\|\\&\quad+\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\mathrm{U}_{t+1}^{\varsigma})\ .\end{aligned}\tag{A.26}$$

$\{\Phi_t : t\in\mathbb{Z}_{\geq 0}\}$ is positive Harris-recurrent with an invariant probability measure π and $E_\pi(|\ln\|[\phi]_1\||)<\infty$, $E_\pi(|\ln\|[\phi]_2\||)<\infty$, and $E_\pi(|\mathcal{R}(\phi)|)<\infty$, where $\mathcal{R}(\phi)$ is defined in (35). Therefore, we can apply a LLN to the right-hand side of (A.26). We obtain

$$\begin{aligned}\lim_{t\to\infty}\frac{1}{t}\ln\frac{\|\mathrm{X}_t-\mathrm{x}_{\mathrm{opt}}\|}{\|\mathrm{X}_0-\mathrm{x}_{\mathrm{opt}}\|}&=\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathrm{Y}_{k+1}\|-\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\mathrm{Y}_k\|\\&\quad+\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\mathrm{U}_{t+1}^{\varsigma})\\&=\int\ln\|[\phi]_1\|\pi(d\phi)-\int\ln\|[\phi]_1\|\pi(d\phi)\\&\quad+\int\mathcal{R}(\phi)\pi(d\phi)=-\mathrm{CR}\ .\end{aligned}$$

We proceed similarly with $\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\mathrm{opt}}\|}{\|\gamma_0-\gamma_{\mathrm{opt}}\|}$ and $\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}$.

$$\begin{aligned}\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\mathrm{opt}}\|}{\|\gamma_0-\gamma_{\mathrm{opt}}\|}&=\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\Gamma_{k+1}\|-\frac{1}{t}\sum_{k=0}^{t-1}\ln\|\Gamma_k\|\\&\quad+\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\mathrm{U}_{t+1}^{\varsigma})\ ,\end{aligned}\tag{A.27}$$

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}=\frac{1}{t}\sum_{k=0}^{t-1}\frac{\sigma_{k+1}}{\sigma_k}=\lim_{t\to\infty}\frac{1}{t}\sum_{k=0}^{t-1}\ln\mathcal{G}_\sigma(1,\mathrm{U}_{t+1}^{\varsigma})\ .\tag{A.28}$$

By applying a LLN to the right-hand side of (A.27) and (A.28), we obtain

$$\lim_{t\to\infty}\frac{1}{t}\ln\frac{\|\gamma_t-\gamma_{\mathrm{opt}}\|}{\|\gamma_0-\gamma_{\mathrm{opt}}\|}=\lim_{t\to\infty}\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0}=-\mathrm{CR}\ .$$

Appendix B. Definitions Related to Markov Chains

This appendix contains complementary notions on Markov chains. We define in particular notions related to the "stability", such as irreducibility, positivity, and Harris-recurrence. For further reading on the Markov chain theory, see [18].

Let consider a Markov chain $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ that takes its values in a set $\mathcal{S}$ equipped with its Borel σ-algebra $\mathcal{B}(\mathcal{S})$. The transition probabilities are given by the transition probability kernel P such that for $\phi \in \mathcal{S}$ and $B \in \mathcal{B}(\mathcal{S})$,

$$P(\phi, B) = \Pr(\Phi_{t+1} \in B \mid \Phi_t = \phi) \ .$$

We say that a Markov chain is φ-irreducible if there exists a nonzero measure φ on the state space such that for any $\phi \in \mathcal{S}$ and for any $B \in \mathcal{B}(\mathcal{S})$ such that $\varphi(B) > 0$,

$$\sum_{k \in \mathbb{Z}_{>0}} P^k(\phi, B) > 0 \ .$$

In such a case, there exists a maximal irreducibility measure ψ that dominates other irreducibility measures [18].

Let now π be a probability measure on $\mathcal{S}$. We say that π is invariant if

$$\pi(B) = \int_{\mathcal{S}} \pi(d\phi) P(\phi, B) \ .$$

We say that a Markov chain is *positive* if there exists an invariant probability measure for this Markov chain.

Harris-recurrence [18] is related to the notion of irreducibility. Let $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ be a ψ-irreducible Markov chain. A measurable set $B \in \mathcal{B}(\mathcal{S})$ is Harris-recurrent if

$$\Pr\Big(\sum_{t \in \mathbb{Z}_{>0}} 1_B(\Phi_t) = \infty \mid \Phi_0 = \phi\Big) = 1 \ ,$$

for all $\phi \in B$, where 1_B is the indicator function. By extension, we say that $\{\Phi_t : t \in \mathbb{Z}_{\geq 0}\}$ is Harris-recurrent if all ψ-positive sets are Harris-recurrent.

We can now recall Theorem 17.0.1 from [18] that gives sufficient conditions for the application of a LLN for Markov chains.

Theorem 3 (Theorem 17.0.1 from [18])**.** *Let Z be a positive Harris-recurrent chain with invariant probability π. Then, the LLN holds for any function q such that $\pi(|q|) = \int |q(\boldsymbol{z})|\pi(d\boldsymbol{z}) < \infty$. That is, for any initial state Z_0,*

$$\lim_{t \to \infty} \frac{1}{t} \sum_{k=0}^{t-1} q(Z_k) = \pi(q) \ \textit{almost surely} \ .$$

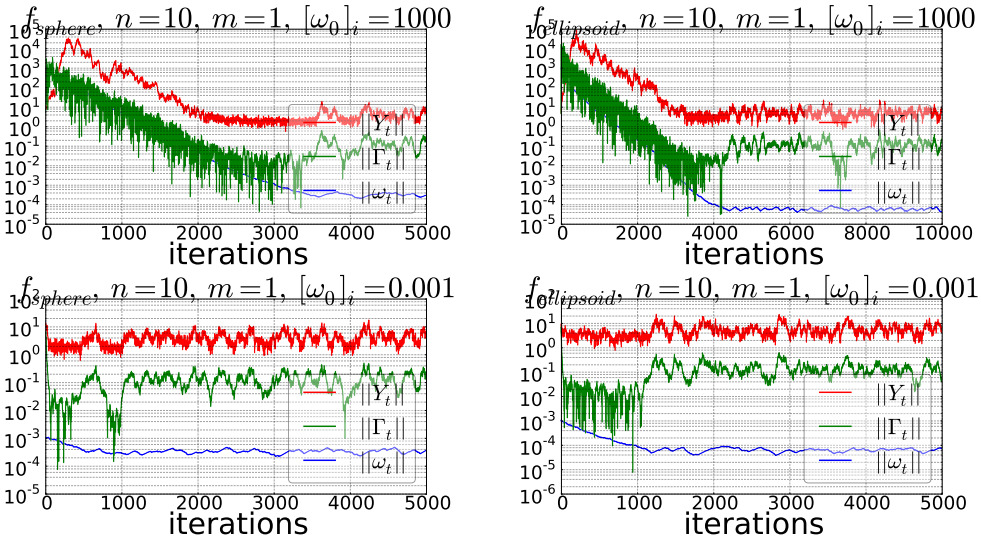

Figure C.4: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ in $n = 10$.

Appendix C. Complementary Empirical Results

We present here additional results that correspond to simulations of the Markov chain defined in Theorem 1 (Figures C.4–C.7), as well as to single runs of the $(\mu/\mu_{\text{w}}, \lambda)$-CSA$_{\text{off}}$-AL (Figures C.8–C.11) on the linearly constrained sphere and ellipsoid functions in $n = 10$, with constraint values $m \in \{1, 2, 5, 9\}$. The parameter setting is described in Section 6. We test in particular three initial values of the penalty factors, $\omega_0 \in \{(1, \cdots, 1)^\intercal, (10^3, \cdots, 10^3)^\intercal, (10^{-3}, \cdots, 10^{-3})^\intercal\}$.

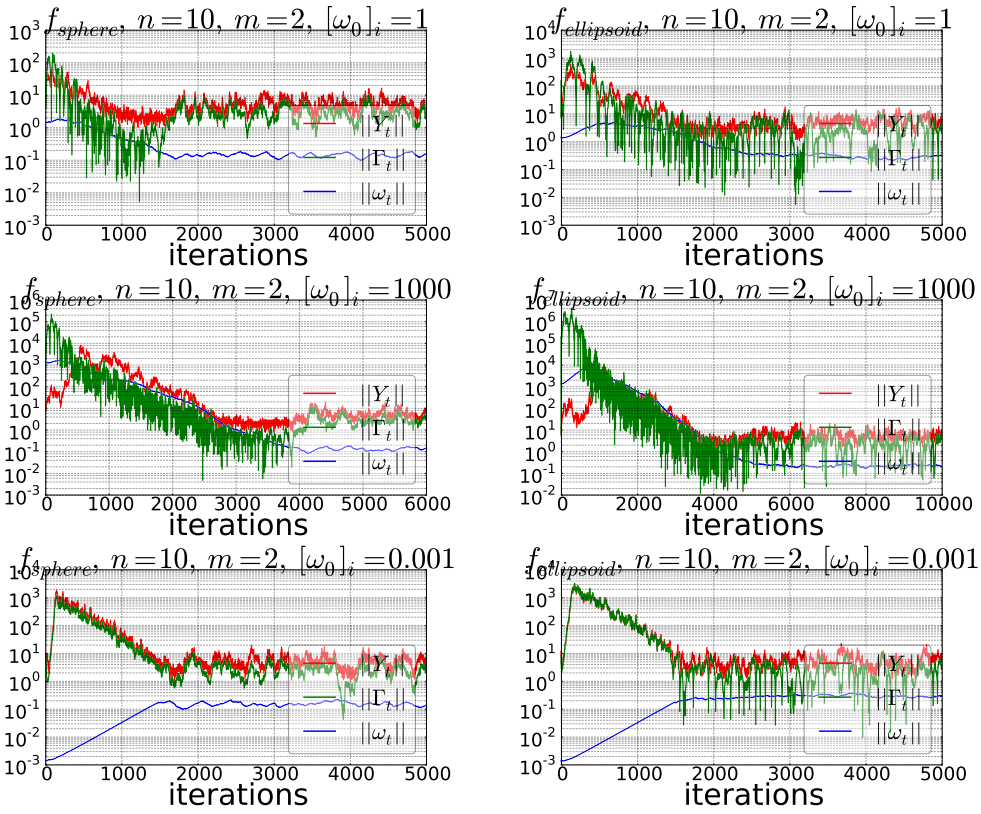

Figure C.5: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 2$ in $n = 10$.

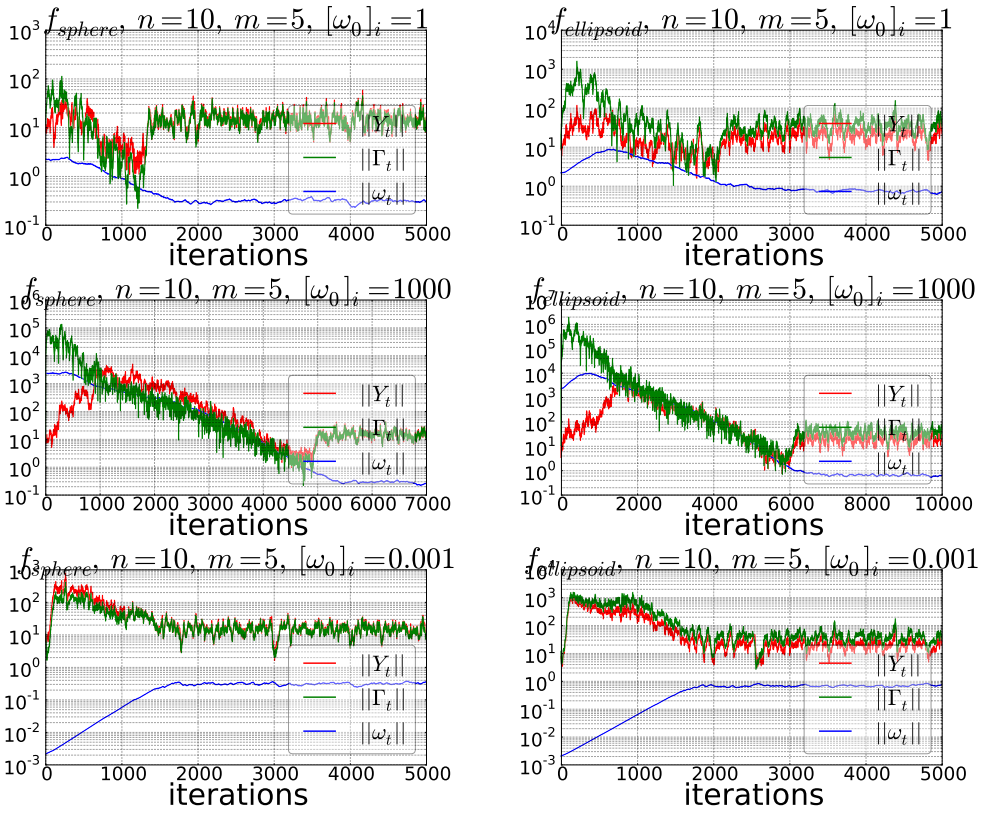

Figure C.6: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 5$ in $n = 10$.

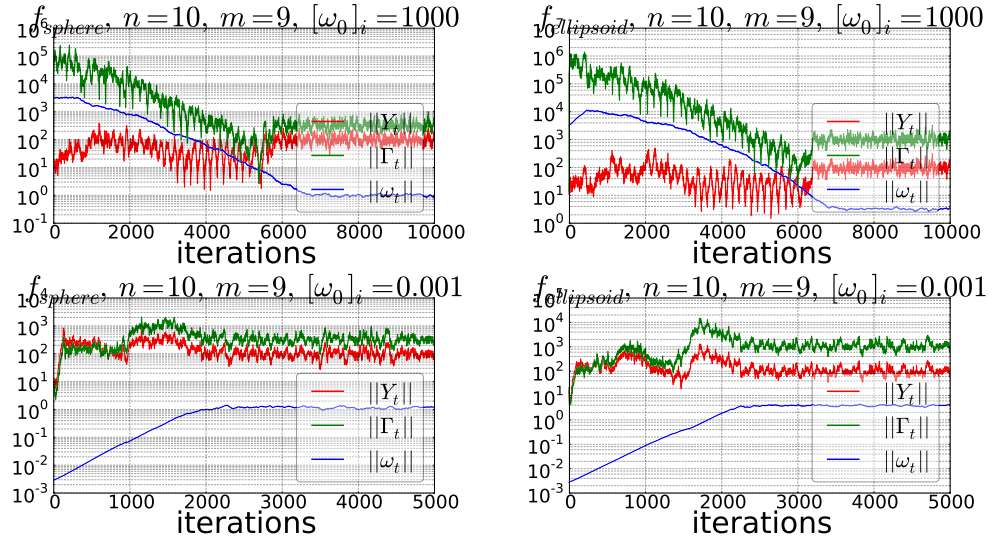

Figure C.7: Simulations of the Markov chain on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 9$ in $n = 10$.

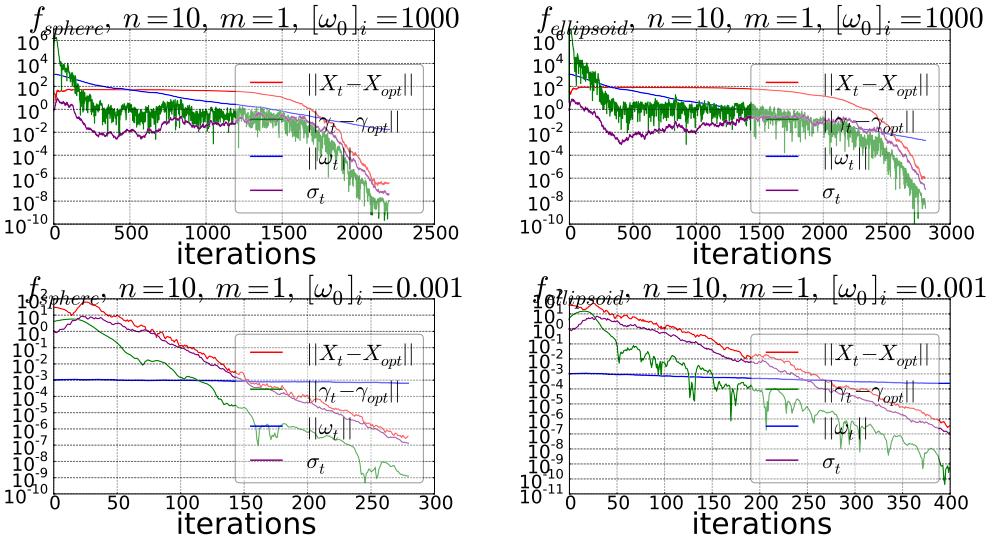

Figure C.8: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 1$ in $n = 10$, with different values of ω_0.

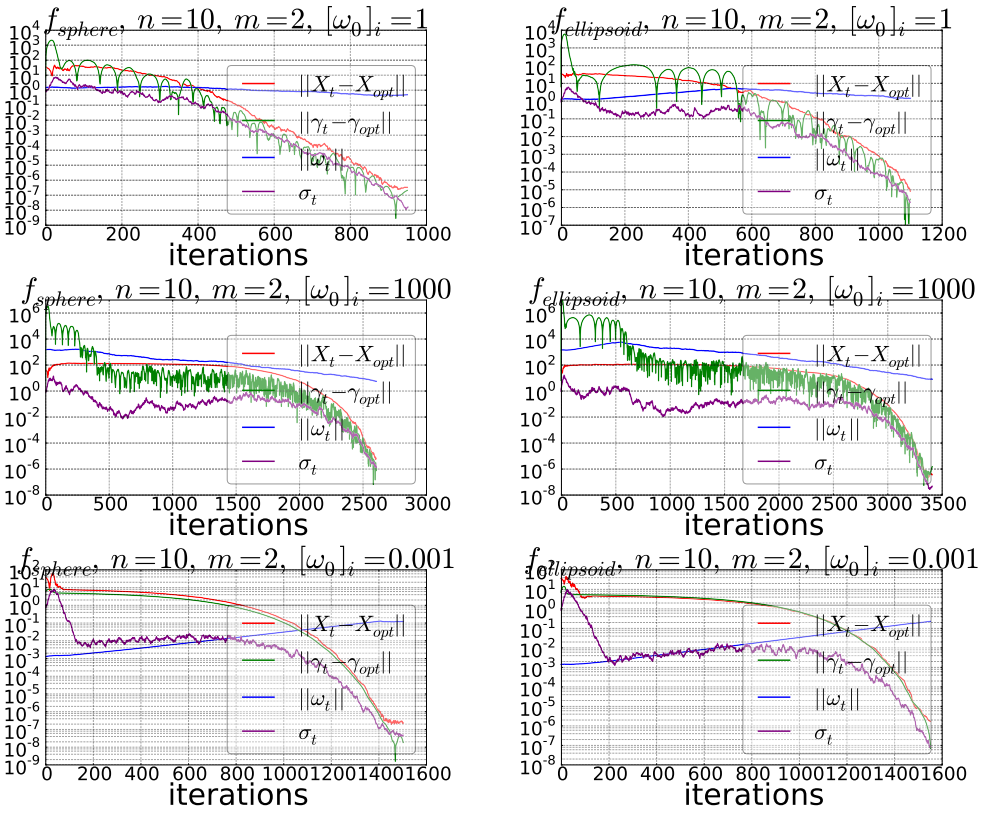

Figure C.9: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 2$ in $n = 10$, with different values of ω_0.

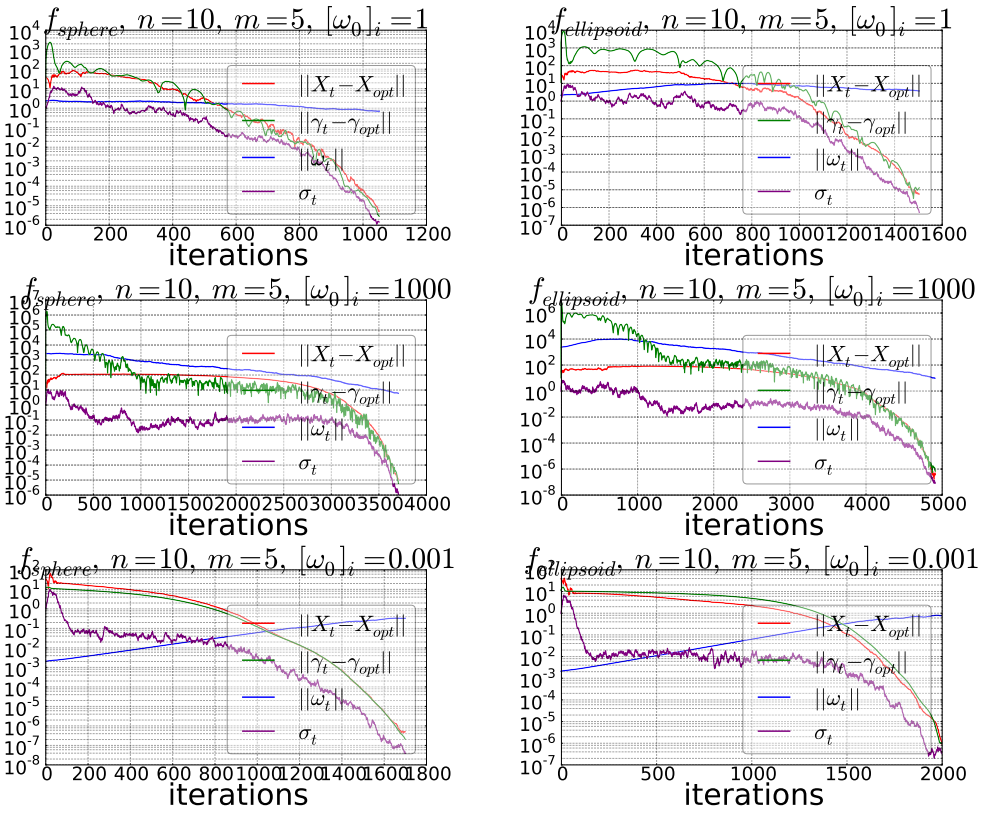

Figure C.10: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 5$ in $n = 10$, with different values of ω_0.

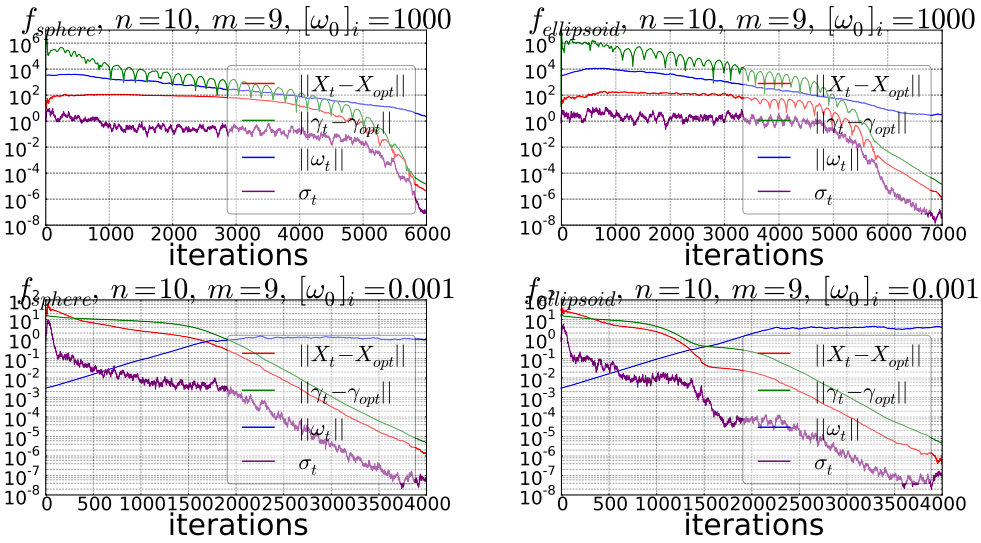

Figure C.11: Single runs on f_{sphere} (left) and $f_{\text{ellipsoid}}$ (right) with $m = 9$ in $n = 10$, with different values of ω_0.