Orthogonal Directions Constrained Gradient Method: from non-linear equality constraints to Stiefel manifold

Sholom Schechtman SHOLOM.SCHECHTMAN@TELECOM-SUDPARIS.EU
SAMOVAR, Télécom Sudparis, Institut Polytechnique de Paris, 91120 Palaiseau, France

Daniil Tiapkin DTYAPKIN@HSE.RU
HSE University, Pokrovsky Blvd, 11, Moscow, Russia, 109028

Michael Muehlebach MICHAELM@TUEBINGEN.MPG.DE
Max Planck Ring 4, 72076 Tuebingen, Germany

Éric Moulines ERIC.MOULINES@POLYTECHNIQUE.EDU
CMAP, École Polytechnique, Route de Saclay, 91128, Palaiseau

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We consider the problem of minimizing a non-convex function over a smooth manifold $\mathcal{M}$. We propose a novel algorithm, the *Orthogonal Directions Constrained Gradient Method* (`ODCGM`), which only requires computing a projection onto a vector space. `ODCGM` is infeasible but the iterates are constantly pulled towards the manifold, ensuring the convergence of `ODCGM` towards $\mathcal{M}$. `ODCGM` is much simpler to implement than the classical methods, which require the computation of a retraction. Moreover, we show that `ODCGM` exhibits the near-optimal oracle complexities $\mathcal{O}(1/\varepsilon^2)$ and $\mathcal{O}(1/\varepsilon^4)$ in the deterministic and stochastic cases, respectively. Furthermore, we establish that, under an appropriate choice of the projection metric, our method recovers the `landing` algorithm of Ablin and Peyré (2022), a recently introduced algorithm for optimization over the Stiefel manifold. As a result, we significantly extend the analysis of Ablin and Peyré (2022), establishing near-optimal rates both in deterministic and stochastic frameworks. Finally, we perform numerical experiments, which shows the efficiency of `ODCGM` in a high-dimensional setting.

Keywords: constrained optimization, non-convex optimization, Riemannian optimization, stochastic optimization, Stiefel manifold

1. Introduction

Given a continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$, we consider the following optimization problem:

$$\min_{x\in\mathcal{M}} f(x), \qquad \text{with } \mathcal{M} := \{x \in \mathbb{R}^n : h(x) = 0\}, \tag{1}$$

where $h : \mathbb{R}^n \to \mathbb{R}^{n_h}$ is continuously differentiable, non-convex, $n_h > 0$ represents the number of constraints and $\mathcal{M}$ denotes the feasible set. Optimization problems with nonlinear constraints naturally arise in a number of areas in machine learning, with a specific emphasis on matrix manifold optimization (see Li et al. (2019); Yang (2007); Sato (2021)). Examples include independent component analysis (Hyvärinen et al., 2009; Ablin et al., 2018), Procrustes estimation (Bojanczyk and Lutoborski, 1999; Turaga et al., 2008, 2011) and the orthogonally normalized neural networks in deep learning (Arjovsky et al., 2016; Li et al., 2019; Bansal et al., 2018; Qi et al., 2020).

When the projection to $\mathcal{M}$ is computationally tractable, the projected gradient method – in which a gradient descent step on f is combined with the projection to $\mathcal{M}$ – is often the preferred

option. The convergence guarantees for projected gradient methods are similar to those for unconstrained gradient descent. Moreover, projected gradients are a first-order procedure and efficiently handle the stochastic case, where only an unbiased estimate of ∇f is known (see, e.g., Ghadimi and Lan (2013, 2016)). When $\mathcal{M}$ is a submanifold, a typical approach is to determine a search direction in the tangent space and then apply a retraction (see e.g. Absil and Malick (2012); Bonnabel (2013); Boumal et al. (2019); Boumal (2020); Sato (2021)). Similarly, retraction-based gradient algorithms have optimal convergence rates in both deterministic and stochastic settings (see Zhang and Sra (2016); Sato et al. (2019)). These methods are feasible, i.e., the iterates always belong to $\mathcal{M}$. In most cases, however, computing the retraction is expensive and requires solving a nontrivial optimization problem.

Infeasible methods (i.e. the iterates do not remain on $\mathcal{M}$), such as augmented Lagrangian and proximally guided methods, seek a solution to (1) by solving a sequence of optimization problems (see Li et al. (2021); Lin et al. (2022); Xie and Wright (2021); Hong et al. (2017)). Here, the iterates are not feasible but are gradually pushed towards $\mathcal{M}$. Nevertheless, each of the optimization problems in the inner loop might be computationally involved. Moreover, these methods are sensitive to the choice of hyperparameters both in theory (often the sub-problems in the inner loop are required to be convex) and in practice.

In this work, we propose `ODCGM`, which stands for *Orthogonal Directions Constrained Gradient Method*, a new class of algorithms that are both easy to implement and computationally inexpensive while retaining the good convergence properties of gradient descent. `ODCGM` realizes a trade-off between two opposite goals: minimizing f and guaranteeing feasibility of solutions. In order to set up the stage, we define for each $x \in \mathbb{R}^n$ *i)* $\nabla H(x) := (1/2)\nabla(\|h(x)\|^2)$, and *ii)* $V(x)$ the vector space orthogonal to $\operatorname{span}(\{\nabla h_i(x)\}_{i=1}^{n_h})$. Then, a vanilla version of `ODCGM` produces iterates as follows:

$$x_{k+1} = x_k - \gamma_k \nabla H(x_k) - \gamma_k \nabla_V f(x_k)\,, \tag{2}$$

where $\gamma_k > 0$ is a step size and $\nabla_V f$ denotes the orthogonal projection of ∇f onto $V(x)$. Since $\nabla H(x)$ is orthogonal to $V(x)$ by construction, the iterates, even if allowed to be infeasible, are constantly shifted in the direction of $\mathcal{M}$. Moreover, $-\nabla_V f$ strives to be as close as possible to $-\nabla f$, which is the direction of descent for f, and thus tends to minimize f; see Figure 1.

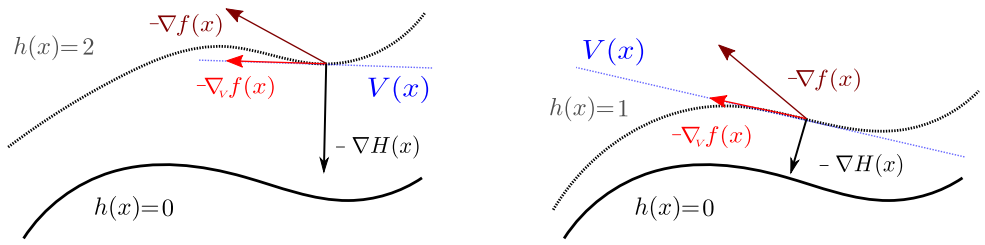

Figure 1: Construction of the orthogonal directions

`ODCGM` is a first-order algorithm that only requires a projection onto the vector space $V(x)$ at each iteration. It is scalable, simple to implement, and can be easily generalized to the stochastic setting. In addition, we provide `ODCGM` with strong theoretical guarantees: we establish convergence bounds that are equivalent to those of (unconstrained) gradient descent in both deterministic and stochastic settings: $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-4})$, respectively. We also present `RODCGM` a computationally cheaper version of `ODCGM` where $V(x)$ is replaced by a hyperplane orthogonal to $\nabla H(x)$.

The advantage RODCGM is that the projection (i.e. the computation of $\nabla_V f(x)$) now comes essentially for free, which has the potential to efficiently solve high-dimensional problems, $n - n_h \gg 1$. This version of ODCGM is inherently non-smooth and we obtain a $\mathcal{O}(\varepsilon^{-3})$ convergence rate in the deterministic setting and $\mathcal{O}(\varepsilon^{-4})$ in the stochastic setting.

ODCGM is closely related to two recently proposed methods. First, the algorithm developed in Muehlebach and Jordan (2022), when applied to equality constraints, is a special instance of ODCGM. Our finite-time complexity analysis extends Muehlebach and Jordan (2022) to the non-convex setting (see also Schechtman et al. (2022); Leconte et al. (2023)). Second, ODCGM is closely related to the landing algorithm proposed by Ablin and Peyré (2022); Gao et al. (2022). The landing algorithm deals with the case where $\mathcal{M}$ is the Stiefel (or orthogonal) manifold: it avoids retractions and requires only a few matrix multiplications at each iteration. For the orthogonal manifold case (and in the deterministic setting), Ablin and Peyré (2022) provides convergence guarantees, however, with a suboptimal convergence rate. Following Gao et al. (2022), we show that by choosing an appropriate metric for the projection on $V(x)$, we obtain a closed-form solution for $\nabla_V f$, and we recover landing as a specific instance of ODCGM. As a consequence, when $\mathcal{M}$ is the Stiefel manifold, we significantly extend the analysis of Ablin and Peyré (2022) by establishing near-optimal rates both in the deterministic and stochastic framework. In particular, we show that landing indeed converges to $\mathcal{M}$, which was only conjectured by Ablin and Peyré (2022).

Main contributions.

- We propose ODCGM, a novel family of algorithms that do not require projections or retractions to the feasible set $\mathcal{M}$.
- We establish convergence rates that coincide with the one of gradient descent in the non-convex setting: $\mathcal{O}(\varepsilon^{-2})$ in the deterministic and $\mathcal{O}(\varepsilon^{-4})$ in the stochastic cases; see Section 3.
- We propose RODCGM which significantly decreases the computational cost per iteration. The cost of this computational reduction is a slightly degraded convergence rate: $\mathcal{O}(\varepsilon^{-3})$ in the deterministic case and $\mathcal{O}(\varepsilon^{-4})$ in the stochastic case; see Section 4.
- We introduce ODRGM, a geometry-aware version of ODCGM, which is applicable when an underlying geometrical structure of the problem is available. In particular, the landing method of Ablin and Peyré (2022) is a particular version of ODRGM. Convergence guarantees of ODRGM are identical to the one ODCGM; see Section 5.
- We perform various numerical experiments on high-dimensional problems that highlight the claim on efficiency of our method; see Section 6.

Notations. For a smooth function $f : \mathbb{R}^n \to \mathbb{R}$, $\nabla f(x) \in \mathbb{R}^n$ denotes its gradient. For a smooth function $h : \mathbb{R}^n \to \mathbb{R}^{n_h}$, we denote $\nabla h(x) \in \mathbb{R}^{n \times n_h}$ the matrix in which the i-th column is $\nabla h_i(x)$. Given a matrix A, $\ker A$ denotes its kernel. Given a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a filtration $(\mathscr{F}_k)$, $\mathbb{E}[\cdot|\mathscr{F}_k]$ is denoted as $\mathbb{E}_k[\cdot]$. The orthogonal projector on the linear subspace V is denoted P_V and the Euclidean norm will be denoted $\|\cdot\|$.

Submanifolds. A set $\mathcal{M} \subset \mathbb{R}^n$ is called a submanifold of dimension $n - n_h$, with $n_h \leqslant n$, if for every point $x \in \mathcal{M}$, there is a neighborhood $U \subset \mathbb{R}^n$ of x and a smooth function $h : U \to \mathbb{R}^{n_h}$ such that $h^{-1}(0) = U \cap \mathcal{M}$ and ∇h is of full rank on U. The tangent plane of $\mathcal{M}$ at x is $\mathcal{T}_x\mathcal{M} = \ker(\nabla h(x)^\top)$. For a smooth function $f : \mathbb{R}^n \to \mathbb{R}$ and $x \in \mathcal{M}$, $\operatorname{Grad} f(x) = P_{\mathcal{T}_x\mathcal{M}} \nabla f(x)$ denotes the Riemannian gradient of f at x in the case when the Riemannian metric is inherited from

the ambient space. More generally, for $(\mathcal{M}, g)$ a manifold equipped with a Riemannian metric g, $\mathrm{Grad}_{\mathcal{M}} f(x) \in \mathcal{T}_x\mathcal{M}$ denotes the Riemannian gradient: a vector in the tangent plane such that for any $\xi \in \mathcal{T}_x\mathcal{M}$, $g_x(\mathrm{Grad}_{\mathcal{M}} f(x), \xi) = \nabla f(x)^\top \xi$, where g_x is a scalar product induced by g on $\mathcal{T}_x\mathcal{M}$.

2. Problem formulation and preliminaries

We consider submanifolds of $\mathbb{R}^n$ defined by a single function $h\colon \mathbb{R}^n \to \mathbb{R}^{n_h}$. A point $x \in \mathcal{M}$ is a *critical point* of (1) if:

$$\mathrm{Grad}\, f(x) := P_{\mathcal{T}_x\mathcal{M}} \nabla f(x) = 0\,. \tag{3}$$

In particular, any local minimum of (1) is a critical point. To each $x \in \mathbb{R}^n$ we associate a vector space $V(x) = \{v \in \mathbb{R}^n : \nabla h(x)^\top v = 0\}$. Note that, for any $x \in \mathcal{M}$, $V(x) = \mathcal{T}_x\mathcal{M}$. If x is such that $\nabla h(x)$ has full rank, then $V(x)$ is the tangent plane of the manifold $\{y \in \mathbb{R}^n : h(y) = h(x)\}$ (perhaps restricted to some neighborhood of x). Therefore, $V(x)$ extends the tangent plane outside of $\mathcal{M}$. The *orthogonal directions field* is defined as:

$$\mathrm{O}_D(x) = -\nabla h(x) A(x) h(x) - \nabla_V f(x)\,, \tag{4}$$

where for all $x \in \mathbb{R}^n$, $\nabla_V f(x)$ is the orthogonal projection of $\nabla f(x)$ onto $V(x)$ and $A(x) \in \mathbb{R}^{n_h \times n_h}$ is chosen such that $\nabla h(x)^\top \nabla h(x) A(x)$ is a symmetric positive definite matrix. As we will see in the next sections, under this assumption, the directions along $\mathrm{O}_D(x)$ tend to decrease $\|h\|$. Note also that the first term, $\nabla h(x) A(x) h(x)$, is orthogonal to $\nabla_V f(x)$ by construction. Before discussing the possible choices of $x \mapsto A(x)$, we show in the following lemma, that $\mathrm{O}_D(x)$ is a meaningful way to measure the closeness of x to a critical point. In particular, it is consistent with the notions of ε-1o point of Xie and Wright (2021) and ε-KKT point of Birgin et al. (2018); Haeser et al. (2019).

Lemma 1 *For $x \in \mathbb{R}^n$ and $\varepsilon > 0$, let λ denote the minimal singular value of $\nabla h(x) A(x)$. If $\|\mathrm{O}_D(x)\| \leqslant \varepsilon$, then $\|\nabla_V f(x)\| \leqslant \varepsilon$ and $\|h(x)\| \leqslant \lambda^{-1}\varepsilon$. In particular, if $\mathrm{O}_D(x) = 0$, then x is a critical point of Problem 1.*

Example 1 (Vanilla orthogonal directions field) *The first natural example is to choose $A(x) \equiv \alpha\,\mathrm{Id}$, where $\mathrm{Id} \in \mathbb{R}^{n_h \times n_h}$ is the identity matrix. In this case, denoting $H(x) = 1/2\,\|h(x)\|^2$, it holds that $\mathrm{O}_D(x) = -\alpha \nabla H(x) - \nabla_V f(x)$. An adaptive version of the method is obtained by choosing $A(x) = \alpha(x)\,\mathrm{Id}$, with $\alpha : \mathbb{R}^n \to \mathbb{R}_+$ a strictly positive function.*

Example 2 (MJ orthogonal directions field) *For x such that $\nabla h(x)$ is of full rank, another natural example is obtained by setting $A(x) = \alpha(\nabla h(x)^\top \nabla h(x))^{-1}$, where $\alpha > 0$. In this case, it turns out that O_D is an instance of Muehlebach and Jordan (2022). Denote $V_\alpha(x) := \{v \in \mathbb{R}^n : \nabla h(x)^\top v = -\alpha h(x)\}$. Note that for $x \in \mathcal{M}$, $V_\alpha(x) = V(x) = \mathcal{T}_x\mathcal{M}$ and that $V_\alpha(x)$ is non-empty as soon as $\nabla h(x)$ is of full rank. A direct calculation (see Lemma 7) shows that*

$$\mathrm{O}_D(x) = \operatorname*{arg\,min}_{v \in V_\alpha(x)} \frac{1}{2}\,\|v + \nabla f(x)\|^2\,. \tag{5}$$

Since the computational cost of the projection on V_α and V is similar, it might be interesting to compute this vector field by directly solving (5). However, we will see in Section 5 that for important examples of Stiefel and orthogonal manifolds we can modify the geometry of the ambient space to obtain a computationally tractable projection onto V.

3. Main results

3.1. Continuous-time flow

In this section, we analyze the ordinary differential equation $\dot{\mathsf{x}}(t) = \mathrm{O}_D(\mathsf{x}(t))$. In all the remainder, we fix $r_1 > 0$ and $K \subset \mathbb{R}^n$ with $K = \{x \in \mathbb{R}^n : \|h(x)\| \leqslant r_1\}$. Consider the following assumption:

A1 *i) The set K is compact and ∇h is of full rank on K.*
ii) It holds that $\nabla h^\top \nabla h A \in \mathbb{R}^{n_h \times n_h}$ is symmetric positive definite on K.
iii) The function $A : K \to \mathbb{R}^{n_h \times n_h}$ can be extended to a locally Lipschitz continuous function on some neighborhood of K.
iv) There is $\alpha_m > 0$ such that $\inf_{x \in K} \lambda_m(x) > \alpha_m$, where $\lambda_m(x)$ is the minimal eigenvalue of $\nabla h^\top(x) \nabla h(x) A(x)$

Note that as soon as $\mathcal{M}$ is compact, there is always some $r_1 > 0$ such that **A**1-i) holds. Moreover, **A** 1-ii)–iii) are satisfied for the matrices A given in Examples 1 and 2. As is often the case, to analyze the trajectory of an ordinary differential equation we need to find an energy (or Lyapunov) function. For $M > 0$, we define $\Lambda_M : \mathbb{R}^n \to \mathbb{R}$ as:

$$\Lambda_M = f + M \|h\| \ . \tag{6}$$

The following theorem is our first main result. It shows that for M large enough, Λ_M decreases along any trajectory. This observation immediately implies the convergence of any bounded trajectory to the set of critical points.

Theorem 2 *Assume **A**1. For any x_0 such that $\|h(x_0)\| \leqslant r_1$, there is a unique solution to*

$$\dot{\mathsf{x}}(t) = \mathrm{O}_D(\mathsf{x}(t)) \tag{7}$$

starting at x_0. In addition, denoting this solution by $\mathsf{x} : \mathbb{R}_+ \to \mathbb{R}^n$, it holds that:

1. *For any $t \geqslant 0$, $\|h(\mathsf{x}(t))\| \leqslant \mathrm{e}^{-\alpha_m t} \|h(x_0)\|$, where α_m is defined in **A**1-iv).*
2. *For all $M \geqslant \overline{M} = M_1/\alpha_m$, with $M_1 = \sup_{x \in K} \left\|A^\top \nabla h^\top (\nabla f - \nabla h A h)\right\|$, we get*

$$\inf_{0 \leqslant t \leqslant T} \|\mathrm{O}_D(\mathsf{x}(t))\|^2 = \inf_{0 \leqslant t \leqslant T} \|\dot{\mathsf{x}}(t)\|^2 \leqslant \frac{1}{T} \int_0^T \|\dot{\mathsf{x}}(t)\|^2 \, \mathrm{d}t \leqslant \frac{\Lambda_M(\mathsf{x}(0)) - \Lambda_M(\mathsf{x}(T))}{T} \, .$$

3. *Let x^* be in the limit set of x, i.e. there is $t_n \to +\infty$ such that $\mathsf{x}(t_n) \to x^*$. Then x^* is a critical point of (1).*

Proof The existence and uniqueness of a local solution of (7) follows from the fact that O_D is locally Lipschitz continuous. As we shall see, such a solution must lie in K, which is compact by **A**1. This implies that the domain of a local solution can be extended to $\mathbb{R}_+$. Indeed, let x be such a solution. Since for all $v \in V$, it holds that $\nabla h^\top v = 0$, we get using **A**1-iv):

$$\frac{\mathrm{d}}{\mathrm{d}t} \|h(\mathsf{x})\|^2 = -2 h^\top(\mathsf{x}) \nabla h^\top(\mathsf{x}) \nabla h(\mathsf{x}) A(\mathsf{x}) h(\mathsf{x}) \leqslant -2\alpha_m \|h(\mathsf{x})\|^2 \ , \tag{8}$$

and Grönwall's lemma implies that $\|h(\mathsf{x}(t))\| \leqslant \mathrm{e}^{-\alpha_m t} \|h(\mathsf{x}(0))\|$, for $t \geqslant 0$. Therefore, any local solution stays away from the boundary of K and can be extended to a global solution for which the

first claim holds. We now prove the second claim. Denote $D_h = (\nabla h^\top \nabla h)^{-1}$. In order to simplify the notations we omit the dependence on x (see Lemma 7), and get

$$\mathrm{O}_D = -\nabla f + \nabla h \left(D_h \nabla h^\top \nabla f - Ah \right) , \tag{9}$$

where $D_h := (\nabla h^\top \nabla h)^{-1}$. This implies $\nabla h^\top \mathrm{O}_D = -\nabla h^\top \nabla h A h$. Therefore, we have

$$\begin{aligned} \left\| (\mathrm{O}_D + \nabla f)^\top \mathrm{O}_D \right\| &= \left\| \left(D_h \nabla h^\top \nabla f - Ah \right)^\top \nabla h^\top \mathrm{O}_D \right\| \\ &\leqslant \left\| h^\top A^\top \nabla h^\top \nabla h A h - \nabla f^\top \nabla h A h \right\| \leqslant M_1 \left\| h \right\| . \end{aligned} \tag{10}$$

Finally, if $\mathsf{x} \notin \mathcal{M}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t} f(\mathsf{x}) = \nabla f(\mathsf{x})^\top \dot{\mathsf{x}} = - \left\| \dot{\mathsf{x}} \right\|^2 + \left(\dot{\mathsf{x}} + \nabla f(\mathsf{x}) \right)^\top \dot{\mathsf{x}}(t) \leqslant - \left\| \dot{\mathsf{x}} \right\|^2 + M_1 \left\| h(\mathsf{x}) \right\| , \tag{11}$$

and from (8)

$$\frac{\mathrm{d}}{\mathrm{d}t} \left\| h(\mathsf{x}) \right\| = \frac{1}{2 \left\| h(\mathsf{x}) \right\|} \frac{\mathrm{d}}{\mathrm{d}t} \left\| h(\mathsf{x}) \right\|^2 \leqslant -\alpha_m \left\| h(\mathsf{x}) \right\| . \tag{12}$$

Therefore, using (12) and (11) we obtain, for $x \notin \mathcal{M}$,

$$\frac{\mathrm{d}}{\mathrm{d}t} \Lambda_M(\mathsf{x}) \leqslant - \left\| \dot{\mathsf{x}} \right\|^2 \leqslant - \left\| \nabla_V f(\mathsf{x}) \right\|^2 , \tag{13}$$

where the last inequality comes from the fact that the projection of $\dot{\mathsf{x}}(t)$ onto V is $\nabla_V f$. Furthermore, as soon as there is $t_1 > 0$ such that $\mathsf{x}(t_1) \in \mathcal{M}$, x remains in $\mathcal{M}$ by (8). Thus, for $t \geqslant t_1$, $\mathrm{O}_D(\mathsf{x}(t)) = -\operatorname{Grad}_{\mathcal{M}}(f(\mathsf{x}(t)) = -\nabla_V f(\mathsf{x}(t))$ and the flow reduces to the Riemannian gradient flow. Since, for such t, $f(\mathsf{x}(t)) = \Lambda_M(\mathsf{x}(t))$, (13) still holds. We obtain the second claim by integrating the latter.

To establish the third claim, we notice that $\mathrm{O}_D \neq 0$ as soon as $x \notin \mathcal{M}$ or $x \in \mathcal{M}$ and $\operatorname{Grad}(f) \neq 0$. Equation (13) then shows that Λ_M is a strict Lyapunov function for the ODE (7) and the set of critical points of (1). In particular, LaSalle's invariance principle (see e.g. (Haraux, 1991, Theorem 2.17)) then implies that any limit point of x must be contained in the set of critical points of (1). ■

3.2. Algorithm

In this section we analyze the algorithms provided by the discretization of the ODE (7) both in the deterministic and stochastic settings. Consider a filtered probability space $(\Omega, \mathscr{F}, \{\mathscr{F}_k, k > 0\}, \mathbb{P})$. Fix $x_0 \in K$ and let $(\eta_k)_{k \geqslant 1}$ be a sequence of random variables adapted to $(\mathscr{F}_k)$. Our method, `ODCGM`, produces iterates as follows:

$$x_{k+1} = x_k + \gamma_k v_k + \gamma_k \eta_{k+1}, \qquad \text{with } v_k = \mathrm{O}_D(x_k) \tag{14}$$

and with (γ_k) a sequence of positive step sizes. The perturbation (η_k) allows to capture the case where $\nabla f(x)$ (and hence $\nabla_V f(x)$) is unknown. This covers both streaming data and finite-sum problems in machine learning; see (Lan, 2020). Recall that $\mathbb{E}_k$ denotes the conditional expectation given $\mathscr{F}_k$ and consider the following assumptions.

A2 *i) The function f (respectively h) has L_f (respectively L_h) Lipschitz gradients on K.*

ii) The iterates (x_k) remain in K, $\mathbb{P}$-almost surely.
iii) For every $k \in \mathbb{N}$, it holds that $\eta_{k+1} \in V(x_k)$ and $\mathbb{E}_k[\eta_{k+1}] = 0$.
iv) There is a constant $\sigma \geqslant 0$ such that for all $k \in \mathbb{N}$, $\mathbb{E}_k[\|\eta_{k+1}\|^2] \leqslant \sigma^2$.

Example 3 *In the stochastic approximation framework, it is assumed that there is a probability space $(\Xi, \mathscr{T}, \mu)$ and a μ-integrable function $g : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ such that for each $x \in \mathbb{R}^n$, $\int g(x,s)\mu(\mathrm{d}s) = \nabla f(x)$. Let $(\xi_k)_{k\geqslant 1}$ be a sequence of i.i.d random variables defined on $(\Omega, \mathscr{F}, \mathbb{P})$, taking values in Ξ and such that the distribution of ξ_k is μ. We consider the following recursion*

$$x_{k+1} = x_k - \gamma_k \nabla h(x_k) A(x_k) h(x_k) - \gamma_k g_V(x_k, \xi_{k+1}) \,,$$

where $g_V(x,\xi)$ denotes the orthogonal projection of $g(x,\xi)$ onto $V(x)$. Thus, if we denote $\eta_{k+1} := \nabla_V f(x_k) - g_V(x_k, \xi_{k+1})$ and $\mathscr{F}_k := \sigma(\xi_1, \dots, \xi_k)$, we obtain (14). Note also that in this case $\eta_{k+1} \in V(x_k)$, $\mathbb{E}_k[\eta_{k+1}] = 0$, and if for some $\sigma > 0$, it holds that $\sup_{x\in\mathbb{R}^n} \mathbb{E}[\|g(x,\xi) - \nabla f(x)\|^2] \leqslant \sigma^2$, then $\mathbb{E}_k[\|\eta_{k+1}\|^2] \leqslant \sigma^2$.

The deterministic setting is recovered by setting $\sigma = 0$. If A is defined only on K (see Example 2), then **A**2-ii) is required for the recursions to be properly defined. However, for A as in Example 1, this assumption is not needed. Nevertheless, it is necessary for our convergence analysis. The following proposition shows that if the sequence (η_{k+1}) is bounded (which is the case in both the deterministic and finite-sum settings), a sufficiently small step size forces the algorithm to stay in K. Its proof is given in Appendix B.2.

To state the result, we denote C_h, C_f as the Lipschitz constants of h, f on K and define $C_A := \sup_{x\in K} \|\nabla h A\|$.

Proposition 3 (safe step-size) *Assume **A**1 and **A**2-i) and that there is a constant $b > 0$ such that $\sup_k \|\eta_{k+1}\| \leqslant b$. Consider $\delta > 0$ and let γ be defined as*

$$\gamma = \min\left(\frac{1}{\alpha_m}, \frac{\delta}{C_h\sqrt{2\left(C_A^2 r_1^2 + C_f^2 + b^2\right)}}, \frac{\alpha_m}{2L_h C_A^2 r_1}, \frac{2\alpha_m(r_1 - \delta)}{3L_h(C_f^2 + b^2)} \right). \tag{15}$$

If (γ_k) is bounded by γ and $\|h(x_0)\| \leqslant r_1 - \delta$, then the sequence (x_k) produced by `ODCGM` remains in K.

As explained in Remark 11, although (15) might be intractable to compute explicitly, a simple doubling trick can enforce the algorithm to stay in K.

The following theorem is the discrete counterpart of Theorem 2. It shows that `ODCGM` converges to the set of the critical points essentially at the same rate than (unconstrained) gradient descent.

Theorem 4 *Assume **A**1–2. For any $M \geqslant \overline{M}$, where $\overline{M}$ is defined in Theorem 2, denote $D_M := \Lambda_M(x_0) - \inf_{x\in K} \Lambda_M(x)$ and let $\gamma \leqslant \gamma_{\max} = \min\left(\alpha_m^{-1}, (L_f + ML_h)^{-1}\right)$. Then, the following holds.*

1. If $\sigma = 0$, and for all k, $\gamma_k \equiv \gamma$, then:

$$\inf_{0\leqslant k\leqslant N-1} \|\mathrm{O}_D(x_k)\|^2 = \inf_{0\leqslant k\leqslant N-1} \|v_k\|^2 \leqslant \frac{2D_M}{N\gamma} \,. \tag{16}$$

Furthermore, it holds that $\mathrm{O}_D(x_k) \to 0$ and any accumulation point x^* of (x_k) is a critical point of Problem (1).

2. *Otherwise, fix some constant* $\bar{D} > 0$, $N > 0$ *and* $\gamma := \min(\gamma_{\max}, \bar{D}(\sigma\sqrt{N})^{-1})$. *If* $\gamma_k \equiv \gamma$, *and* $\hat{k}$ *is uniformly sampled in* $\{0, \dots, N-1\}$, *then:*

$$\mathbb{E}\left[\left\|\mathrm{O}_D(x_{\hat{k}})\right\|^2\right] \leqslant \frac{2D_M(L_f + ML_h + \alpha_m)}{N} + \frac{\sigma}{\sqrt{N}}\left(\bar{D}(L_f + ML_h) + \frac{2D_M}{\bar{D}}\right). \tag{17}$$

Proof Using a Taylor expansion of Λ_M and using the upper-bound on γ_k, we obtain

$$2\left(\mathbb{E}_k[\Lambda_M(x_{k+1})] - \Lambda_M(x_k)\right) \leqslant -\gamma\left\|v_k\right\|^2 + (L_f + ML_h)\sigma^2\gamma^2. \tag{18}$$

Our claims then follow by telescoping this inequality and applying a standard proof technique (see e.g. Lan (2020, Chapter 6)) both in the deterministic and stochastic framework. Further details are given in Appendix B.1. ■

The preceding theorem shows that the rate of convergence of our algorithm, measured through O_D, is identical to the one obtained by gradient descent in a non-convex framework: $\mathcal{O}(\varepsilon^{-2})$ in the deterministic setting and $\mathcal{O}(\varepsilon^{-4})$ in the stochastic setting. As recently shown in Carmon et al. (2020); Arjevani et al. (2023), these rates are tight, which makes our algorithm near-optimal in both cases.

The term $(L_f + ML_h)$ in the definition of $\gamma_{\max}$ is the Lipschitz constant of $\nabla f + M\nabla h$, hence our bound on the step sizes is reminiscent of the L_f^{-1} bound required for convergence of standard gradient descent. Note also that only an upper bound on $\overline{M}$ is required to achieve such rates. Indeed, in the deterministic setting, we can combine our method with line search; see Remark 11. In the stochastic framework, performing line search is not an option, but we note that the discussion of Ghadimi and Lan (2013, Corollary 2.2.) applies here as well. In particular, we can make an error of the order of $\sqrt{N}$ in estimating $(L_f + ML_h)$ while maintaining our rate of convergence of $\mathcal{O}(\varepsilon^{-4})$. The constant $\overline{D}$ is presented here to deal with misspecified step sizes, i.e. a lack of knowledge of required constants. Nevertheless, if all constants are known, then the optimal $\overline{D}$ in Equation (17) is $\sqrt{2D_M/(L_f + ML_h)}$. Finally, a nonconstant choice of step sizes is possible without affecting the final results; see (Lan, 2020, Chapter 6). The choice of step size is further discussed in Appendix B.2.

4. Reducing the computational cost: reduced ODCGM

While ODCGM provides optimal theoretical guarantees, it does so by computing, at every iteration, a projection onto a $n - n_h$-dimensional vector space. For $n - n_h \gg 1$, such a projection might be computational expensive. In this section, we therefore propose a modification of ODCGM that only projects onto a hyperplane, which comes essentially for free. The main idea is to reparametrize our problem by noting that $\mathcal{M} = \{x \in \mathbb{R}^n : H(x) := \|h(x)\|^2/2 = 0\}$. Introducing the vector spaces $\tilde{V}(x) := \{v \in \mathbb{R}^n : \nabla H(x)^\top v = 0\}$, the iterates of RODCGM are defined as follows:

$$x_{k+1} = x_k - \alpha(x_k)\gamma_k\nabla H(x_k) - \gamma_k\nabla_{\tilde{V}} f(x_k) + \gamma_k\eta_{k+1}, \tag{19}$$

where, as previously, $\nabla_{\tilde{V}} f(x)$ denotes the projection of $\nabla f(x)$ onto $\tilde{V}(x)$, (η_{k+1}) is a perturbation sequence, and $\alpha(x)$ corresponds to the choice $A(x) = \alpha(x)\,\mathrm{Id}$ and is specified in Theorem 5 below.

Note that, as soon as $\nabla H(x) \neq 0$, $\tilde{V}(x)$ is a hyperplane. Therefore, the computation of $\nabla_{\tilde{V}} f(x)$ is straightforward, preserving, at the same time, its orthogonality to $\nabla H(x)$. Thus, RODCGM follows

the same idea as ODCGM while significantly reducing its computational cost. Unfortunately, this construction damages the continuity of $\tilde{V}(x)$ near $\mathcal{M}$. Indeed, since $\nabla H(x) = \nabla h(x) \cdot h(x)$, we obtain $\nabla H(x) = 0$ and $\tilde{V}(x) = \mathbb{R}^n$ on $\mathcal{M}$. This observation shows that the field associated with RODCGM is non-smooth. The inherent non-smoothness of RODCGM deteriorates its convergence properties, but we can still derive a $\mathcal{O}(\epsilon^{-3})$ rate of convergence in deterministic environments and a $\mathcal{O}(\varepsilon^{-4})$ rate of convergence in stochastic environments. The latter is reminiscent of the convergence rate of subgradient methods in non-smooth environments (see Davis and Drusvyatskiy (2019)).

To properly analyze RODCGM, and due to a non-smooth choice of $\alpha(x)$, we consider assumptions that are slightly different from **A**1. More precisely, we assume **A**1 for $A(x) = \mathrm{Id}$. We will call this set of assumptions **A**1', and we denote the smallest eigenvalue of $\nabla h(x)^\top \nabla h(x)$ as μ_h^2.

We note that the compactness of K and Lipchitz-continuity of ∇f and ∇h (**A**2-i)) implies that f, h, and $\nabla H = \nabla h \cdot h$ are Lipschitz-continuous with Lipchitz constants C_f, C_h, and L_H respectively. Moreover, since ∇h is continuous and K is compact, we have $\sup_{x \in K} \|\nabla h(x)\|_2 \leqslant M_h$.

Theorem 5 *Assume **A** 1'-2. Let $\bar{D}, \alpha > 0$ and $\alpha(x) = \alpha \cdot H(x) / \|\nabla H(x)\|^2$. Denote $D_0 = f(x_0) - \inf_{x \in K} f(x)$, $\gamma_{\max} = \min(\alpha^{-1}, (L_f + \alpha \mu_h^{-2} L_H)^{-1})$ and $\tilde{C} = B_f M_h \mu_h^{-2}$. Finally, assume that $x_0 \in \mathcal{M}$ and fix $N > 0$ the number of iterations. The following holds:*

1. *If $\sigma^2 = 0$ and for all $k \in \mathbb{N} : \gamma_k \equiv \gamma$ for $\gamma = \min(\gamma_{\max}, \bar{D} \cdot N^{-1/3})$, then choosing $\alpha = \gamma$, we obtain*

$$\inf_{k=0,\dots,N-1} \left\{ \left\| \nabla_{\tilde{V}} f(x_k) \right\|^2 + \frac{1}{2} \|h(x_k)\|^2 \right\} \leqslant \frac{8 D_0 (L_f + L_h \mu_h^{-2})}{N} + \frac{\left(\frac{8 D_0}{\bar{D}} + 8 \tilde{C} L_H \cdot \bar{D} \right)}{N^{2/3}}.$$

2. *Otherwise, if for all $k \in \mathbb{N} : \gamma_k \equiv \gamma$, with $\gamma = \min(\gamma_{\max}, \bar{D} \cdot N^{-1/2})$, we obtain by choosing $\alpha = \gamma$ and $\hat{k}$ uniformly sampled in $\{0, \dots, N-1\}$*

$$\mathbb{E}\left[\left\| \nabla_{\tilde{V}} f(x_{\hat{k}}) \right\|^2 + \frac{1}{2} \left\| h(x_{\hat{k}}) \right\|^2 \right] \leqslant \frac{4 D_0 (L_f + L_h \mu_h^{-2})}{N} + \frac{4 D_0}{\bar{D} \cdot \sqrt{N}} + \frac{4 \tilde{C}^2 \bar{D}^2 \cdot L_H}{N} + \frac{\bar{D}}{\sqrt{N}} \left(2 \left(L_f + \gamma L_H \mu_h^{-2} \right) \cdot \sigma^2 + 2 \tilde{C} \cdot \sqrt{\frac{L_H \sigma^2}{2}} \right).$$

The main difficulty in establishing this result relies in the lack of a suitable Lyapunov function for RODCGM. The latter comes from its inherent non-smoothness and the fact that the Lagrange multipliers that arise in the problem of projection on $\tilde{V}$ are unbounded. A complete proof of this theorem is provided in Appendix B.3.

In the deterministic setting, RODCGM outputs $\hat{x}$ such that $\|h(\hat{x})\| \leqslant \varepsilon$ and $\left\| \nabla_{\tilde{V}} f(\hat{x}) \right\| \leqslant \varepsilon$ in $\mathcal{O}(\varepsilon^{-3})$ iterations. In the stochastic setting, RODCGM outputs a point $\hat{x} = x_{\hat{k}}$ such that $\mathbb{E}[\|h(\hat{x})\|] \leqslant \varepsilon$ and $\mathbb{E}[\left\| \nabla_{\tilde{V}} f(\hat{x}) \right\|] \leqslant \varepsilon$ in $\mathcal{O}(\varepsilon^{-4})$ iterations. One drawback of such a method is that we are no longer guaranteed to converge towards the feasible set. Nevertheless, the condition $\left\| \nabla_{\tilde{V}} f(\hat{x}) \right\| \leqslant \varepsilon$, could be rewritten as ε-1o point with appropriate Lagrange multipliers proportional to $h(\hat{x})$ (see Xie and Wright (2021) for the definition of an ε-1o point).

5. A geometry aware version of ODCGM

As mentioned earlier, a drawback of ODCGM lies in the fact that at each iteration the method evaluates a projection on $V(x)$. RODCGM requires only one projection onto a hyperplane but does not

exhibit optimal convergence guarantees. In fact, since the main feature of our analysis was to exploit the orthogonality of $\nabla h(x)$ and $V(x)$, one might think that the projection onto $V(x)$ (and thus $\nabla_V f(x)$) is not necessarily defined through the canonical metric. This observation is the main idea behind our *Orthogonal Directions Riemannian Gradient Method* (`ODRGM`), where the type of projection might depend on x. This implicitly provides the ambient space with a Riemannian metric and turns out to be particularly interesting for optimization over the Stiefel manifold. In fact, by a specific choice of metric, the projection has a closed form, which recovers the `landing` algorithm of Ablin and Peyré (2022). In particular, our results imply near-optimal rates of `landing`, significantly improving the ones presented in Ablin and Peyré (2022).

Before proceeding, let us introduce some notations. Let $Q : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ be such that for all $x \in \mathbb{R}^n$, $Q(x)$ is a positive definite matrix. Given $v, u \in \mathbb{R}^n$, we set $q_x(u, v) = u^\top Q(x)v$. As a result, we change the geometry of $\mathbb{R}^n$ and transform it into a Riemannian manifold with q_x as the Riemannian inner product. For $v \in \mathbb{R}^n$, we will denote $\|v\|_{q_x} = \sqrt{q_x(v, v)}$. We are now ready to present the *geometry-aware orthogonal directions field*

$$\mathrm{O}_G(x) = -\nabla h(x)A(x)h(x) + \operatorname*{arg\,min}_{v\in V(x)} \frac{1}{2} \left\|v + Q^{-1}(x)\nabla f(x)\right\|_{q_x}^2 . \tag{20}$$

By replacing O_D with O_G in the algorithms of Section 3, we obtain geometry-aware deterministic and stochastic algorithms.

To describe a more geometric viewpoint on this algorithm, let us define a family of manifolds $\mathcal{M}_{h(x)} = \{y \in \mathbb{R}^n : h(y) = h(x)\}$ with Riemannian metric $g^{h(x)}$ such that $g_x^{h(x)} = q_x$ parameterized by $x \in \mathbb{R}^n$. In this case, we can prove that the projection in (20) exactly corresponds to the negative Riemannian gradient (see Lemma 19 in Appendix B.4):

$$-\operatorname{Grad}_{\mathcal{M}_{h(x)}} f(x) = \operatorname*{arg\,min}_{v\in V(x)} \frac{1}{2} \left\|v + Q^{-1}(x)\nabla f(x)\right\|_{q_x}^2 . \tag{21}$$

In particular, if the problem at hand has a geometrical structure, one might hope that a particular choice of Q might reduce the computational costs (or even exhibit a closed form solution) of the right-hand side of (21). This idea explains the "geometry-aware" nature of the algorithm.

The main motivation for `ODRGM` is the example of the orthogonal, or, more generally, Stiefel manifold. In this case, for $X \in \mathbb{R}^{n\times p}$, following the recent work of Gao et al. (2022), the constraints are defined by $h(X) = X^\top X - \mathrm{Id}$ and the manifolds $\mathcal{M}_{h(X)}$ correspond to $\mathrm{St}_{X^\top X}(p, n)$. For any of such $\mathcal{M}_{h(X)}$, we obtain a natural Riemannian metric inherited from the Stiefel manifold $\mathrm{St}(p, n)$ through a family of diffeomorphisms. This provides us with a natural way of defining Q and we obtain (see Gao et al. (2022) for a detailed discussion):

$$\operatorname{Grad}_{\mathcal{M}_{h(X)}} f(X) = \psi(X)X, \quad \text{where } \psi(X) = \left(\nabla f(X)X^\top - X(\nabla f(X))^\top\right).$$

In particular, by setting $A(x) = \lambda\,\mathrm{Id}$, our algorithm exactly recovers the `landing` algorithm (Ablin and Peyré, 2022; Gao et al., 2022)

$$X_{k+1} = X_k - \lambda\gamma_k\nabla H(X_k) - \gamma_k\psi(X_k)X_k.$$

In other words, our approach is a generalization of the `landing` algorithm beyond the orthogonal and Stiefel manifolds.

Next we analyze `ODRGM` under the following assumption.

A3 *There is a constant $C_q > 0$ such that*

$$\sup_{x \in K} \max(\left\|Q^{-1}(x)\right\|, \|Q(x)\|) \leqslant C_q \,.$$

The following theorem shows that ODRGM exhibits the same type of rates than ODCGM. We emphasize that all our analysis automatically holds for the landing algorithms as a special case. In particular, we obtain new and better rates for landing, where only an $\mathcal{O}(\epsilon^{-6})$ rate was previously proven for the deterministic (and with decreasing step-sizes) version of the algorithm. Furthermore, we establish the convergence of the deterministic version of landing to the Stiefel manifold, which was only conjectured in Ablin and Peyré (2022). A full proof is provided in Appendix B.4.

Theorem 6 *Let A1–3 hold. Then, there exists $\overline{M}_q$ (detailed in the proof) such that for all $M \geqslant \overline{M}_q$, with $\gamma_{\max} = \min(\alpha_m^{-1}, C_q^{-1}(L_f + ML_h)^{-1})$, the following holds:*

1. *If $\sigma = 0$ and $\gamma_k \equiv \gamma$, with $\gamma \leqslant \gamma_{\max}$, then:*

$$\inf_{0 \leqslant k \leqslant N-1} \|\mathrm{O}_G(x_k)\|^2 \leqslant \inf_{0 \leqslant k \leqslant N-1} \|v_k\|^2 \leqslant 2C_q \frac{\Lambda_M(x_0) - \Lambda_M(x_n)}{N\gamma} \,.$$

   Furthermore, $\|\mathrm{O}_G(x_k)\| \to 0$ and any accumulation point of (x_k) is a critical point of Problem (1).

2. *Otherwise, fix some constant $\bar{D}$, $N > 0$ and $\gamma = \min(\gamma_{\max}, \bar{D}(\sigma\sqrt{N})^{-1})$. If $\gamma_k \equiv \gamma$ and $\hat{k}$ is uniformly sampled in $\{0, \dots, N-1\}$, then*

$$\mathbb{E}_k\left[\left\|\mathrm{O}_G(x_{\hat{k}})\right\|^2\right] \leqslant \frac{2C_q D_M(L_f + ML_h + \alpha_m)}{N} + \frac{C_q\sigma}{\sqrt{N}}\left(\bar{D}(L_f + ML_h) + \frac{2D_M}{\bar{D}}\right) \,.$$

6. Numerical experiments

We showcase the efficiency of the proposed algorithms on different optimization problems.

Procrustes problem Let A, B be matrices with $A \in \mathbb{R}^{q\times q}$ and $B \in \mathbb{R}^{p\times q}$, where $p \geqslant q$. We consider the orthogonal Procrustes problem of finding a matrix $X \in \mathbb{R}^{p\times q}$ with orthonormal columns solving the minimization problem $\min_{X^\top X = \mathrm{Id}_q} \|AX - B\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm. This is referred to as the Procrustes problem on the Stiefel manifold; see (Eldén and Park, 1999). We compare ODCGM, RODCGM, ODRGM with Riemannian gradient descent with two different choices of Riemannian metric: Euclidean and canonical. The results are shown in Figure 2 in log-log scale for $p = 60, q = 40$. The results are averaged over $n = 100$ draws for the matrices A and B [the entries of the matrices are sampled from a standard normal distributions]. For this experiment, we choose $A(x) = 5\,\mathrm{Id}$; we use a constant step size $\gamma_k = 10^{-2}$ for ODCGM and ODRGM, and decreasing step size for RODCGM $\gamma_k = 10^{-2} \cdot k^{-1/3}$. In particular, we find that ODRGM outperforms the Riemannian gradient descent methods for both the Euclidean and canonical Riemannian metrics, and achieves the orthogonality error at the level of machine accuracy. We also see numerical confirmation of the $\mathcal{O}(\varepsilon^{-2})$ convergence of ODCGM and ODRGM and the slower convergence of RODCGM. Additional experiments on a large instance of the problem are presented in Appendix A.

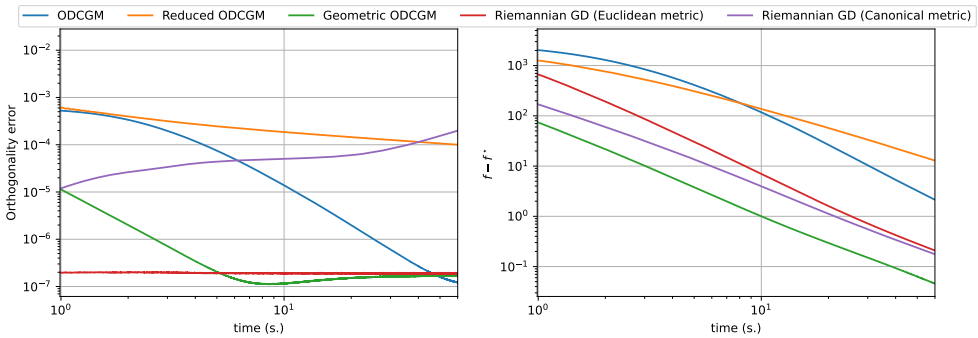

Figure 2: Comparison of ODCGM (blue), RODCGM (orange), and ODRGM (green) with Riemannian gradient descent with two different Riemannian metrics (red and purple). The upper plot shows the orthogonality error for X, the lower plot shows the convergence of the objective function (averages over 100 seeds).

Hanging chain As a second non-convex and nonlinear example, we compute the shape of a hanging chain. The problem can be formulated as follows:

$$\min_{(\xi_1,\ldots,\xi_N)\in\mathbb{R}^{2N}} \frac{1}{N^3}\sum_{i=1}^{N}\left(\frac{k_s}{r^4}(\xi_{i-1}-\xi_i)^\mathsf{T}(\xi_{i+1}-\xi_i)+y_i\right)$$
$$\text{s.t.} \quad \sqrt{(\xi_{k-1}-\xi_k)^\mathsf{T}(\xi_{k-1}-\xi_k)} \leqslant r,\ k=1,2,\ldots,N+1, \quad (22)$$

where $\xi_k = (x_k, y_k)$ denotes the xy-position of the k-th element, $k = 1,\ldots N$, and $\xi_0 = (0,0)$ and $\xi_{N+1} = (9,0)$ are the two endpoints. Further details are given in Appendix A. We compare the results of ODCGM with $A(x) = \alpha(\nabla h(x)^\top \nabla h(x))^{-1}$ (hereafter abbreviated as ODCGM), RODCGM, and an augmented Lagrangian method. The results are summarized in Figure 3 for the case $N = 10,000$, which leads to $20,000$ decision variables and $10,001$ nonlinear constraints. We note that RODCGM and augmented Lagrangian converge much more slower than ODCGM. We also find that fine-tuning the augmented Lagrangian method is quite difficult, while the time steps of ODCGM and RODCGM are easy to set (see Appendix A for details). The execution time per iteration of ODCGM is about five times that of RODCGM and the augmented Lagrangian. To demonstrate the potential of RODCGM, we run the same example for $N = 2\times10^5$, resulting in a large optimization problem with 4×10^5 decision variables and 2×10^5 nonlinear constraints. Under these conditions, solving the Karush-Kuhn-Tucker system becomes challenging at each iteration, which is required for ODCGM. However, the RODCGM still performs well, requiring only about 0.85 seconds to execute a single iteration.

7. Conclusion

In this paper, we propose ODCGM a novel infeasible method for optimization on an immersed manifold $\mathcal{M}$. An attractive property of ODCGM is that it avoids retractions and only projections on a vector space need to be computed. ODCGM achieve near-optimal oracle complexities $\mathcal{O}(1/\varepsilon^2)$ and $\mathcal{O}(1/\varepsilon^4)$ in the deterministic and stochastic cases, respectively. Various extensions of ODCGM are

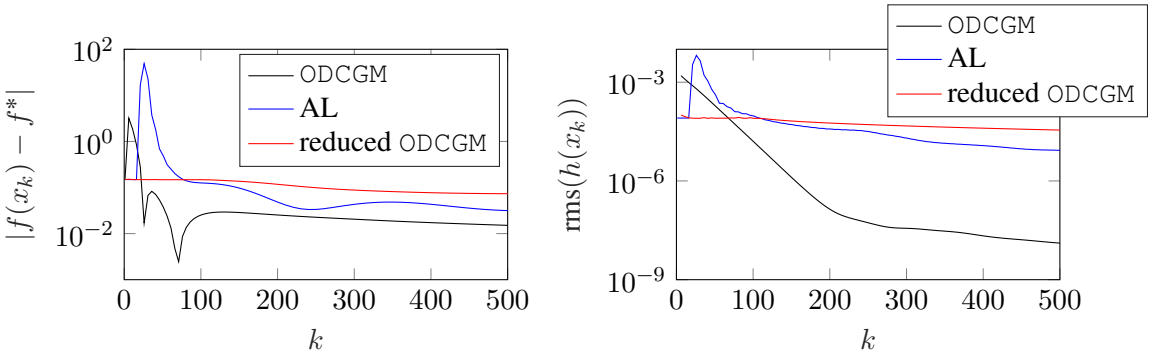

Figure 3: The figure compares the results of ODCGM (black) with RODCGM (red) and an augmented Lagrangian method (blue). RODCGM is performed with a decreasing step size of $\mathcal{O}(k^{-1/2})$. The left plot shows the convergence in objective function, while the right plot gives the mean square error (denoted by rms) of the constraint violations. We note that ODCGM converges much faster than RODCGM and the augmented Lagrangian.

presented. First, we introduce RODCGM, a computationally friendly version of ODCGM. Here we only need to compute one projection onto a hyperplane, but at the price of a slightly worse complexity bound. Second, we introduce ODRGM a geometry-based version of ODCGM, where the projections account for the local Riemannian metric. When specialized to the Stiefel manifold, ODRGM generalizes the landing algorithm (Ablin and Peyré, 2022). We show that ODRGM enjoys the same oracle complexity as ODCGM . As a result, we also establish oracle complexity bounds for landing on the Stiefel manifold. Numerical experiments illustrate the performance of ODCGM and its variants.

Acknowledgments

E.M. and S.S. received support from the grant ANR-19-CHIA-0002 SCAI. Part of this work was conducted under the auspices of the Lagrange Center for Mathematics and Computing. The work of D.T. was prepared within the framework of the HSE University Basic Research Program. M.M. thanks the German Research Foundation and the Branco Weiss Fellowship, administered by ETH Zurich, for the support.

References

Pierre Ablin and Gabriel Peyré. Fast and accurate optimization on the orthogonal manifold without retraction. In *International Conference on Artificial Intelligence and Statistics*, pages 5636–5657, 2022.

Pierre Ablin, Jean-François Cardoso, and Alexandre Gramfort. Faster independent component analysis by preconditioning with Hessian approximations. *IEEE Transactions on Signal Processing*, 66(15):4040–4049, 2018.

P-A Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2): 165–214, 2023.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *Proceedings of the International Conference on Machine Learning*, pages 1120–1128, 2016.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep CNNs? In *Proceedings of the International Conference on Neural Information Processing Systems*, page 4266–4276, 2018.

Ernesto G Birgin, Gabriel Haeser, and Alberto Ramos. Augmented Lagrangians with constrained subproblems and convergence to second-order stationary points. *Computational Optimization and Applications*, 69:51–75, 2018.

Adam W Bojanczyk and Adam Lutoborski. The Procrustes problem for orthogonal Stiefel matrices. *SIAM Journal on Scientific Computing*, 21(4):1291–1304, 1999.

Silvère Bonnabel. Stochastic gradient descent on Riemannian manifolds. *IEEE Transactions on Automatic Control*, 58(9):2217–2229, 2013.

Nicolas Boumal. An introduction to optimization on smooth manifolds. Available online, Nov 2020. URL http://www.nicolasboumal.net/book.

Nicolas Boumal, Pierre-Antoine Absil, and Coralia Cartis. Global rates of convergence for nonconvex optimization on manifolds. *IMA Journal of Numerical Analysis*, 39(1):1–33, 2019.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184(1-2):71–120, 2020.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Lars Eldén and Haesun Park. A Procrustes problem on the Stiefel manifold. *Numerische Mathematik*, 82(4):599–619, 1999.

Bin Gao, Simon Vary, Pierre Ablin, and P.-A. Absil. Optimization flows landing on the Stiefel manifold. *Proceedings of the International Symposium on Mathematical Theory of Networks and Systems*, 55(30):25–30, 2022.

Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming*, 156(1–2):59–99, 2016.

Gabriel Haeser, Hongcheng Liu, and Yinyu Ye. Optimality condition and complexity analysis for linearly-constrained optimization without differentiability on the boundary. *Mathematical Programming*, 178(1-2):263–299, 2019.

Alain Haraux. *Systèmes Dynamiques Dissipatifs et Applications*, volume 17. Elsevier Masson, 1991.

Mingyi Hong, Davood Hajinezhad, and Ming-Min Zhao. Prox-PDA: The proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pages 1529–1538, 2017.

Aapo Hyvärinen, Jarmo Hurri, Patrik O Hoyer, Aapo Hyvärinen, Jarmo Hurri, and Patrik O Hoyer. Independent component analysis. *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, pages 151–175, 2009.

Guanghui Lan. *First-order and stochastic optimization methods for machine learning*, volume 1. Springer, 2020.

Louis Leconte, Sholom Schechtman, and Eric Moulines. AskewSGD: an annealed interval-constrained optimisation method to train quantized neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 3644–3663, 2023.

Shuai Li, Kui Jia, Yuxin Wen, Tongliang Liu, and Dacheng Tao. Orthogonal deep neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(4):1352–1368, 2019.

Zichong Li, Pin-Yu Chen, Sijia Liu, Songtao Lu, and Yangyang Xu. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *Proceeding of the International Conference on Artificial Intelligence and Statistics*, pages 2170–2178, 2021.

Qihang Lin, Runchao Ma, and Yangyang Xu. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. *Computational Optimization and Applications*, 82(1):175–224, 2022.

Michael Muehlebach and Michael I Jordan. On constraints in first-order optimization: A view from non-smooth dynamical systems. *Journal of Machine Learning Research*, 23(256):1–47, 2022.

Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, second edition, 2006.

Haozhi Qi, Chong You, Xiaolong Wang, Yi Ma, and Jitendra Malik. Deep isometric learning for visual recognition. In *Proceedings of the International Conference on Machine Learning*, pages 7824–7835, 2020.

Hiroyuki Sato. *Riemannian optimization and its applications*. Springer, 2021.

Hiroyuki Sato, Hiroyuki Kasai, and Bamdev Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019.

Sholom Schechtman, Daniil Tiapkin, Eric Moulines, Michael I. Jordan, and Michael Muehlebach. First-order constrained optimization: Non-smooth dynamical system viewpoint. *Proceedings of the IFAC Workshop on Control Applications of Optimization*, 55(16):236–241, 2022.

Pavan Turaga, Ashok Veeraraghavan, and Rama Chellappa. Statistical analysis on Stiefel and Grassmann manifolds with applications in computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

Pavan Turaga, Ashok Veeraraghavan, Anuj Srivastava, and Rama Chellappa. Statistical computations on Grassmann and Stiefel manifolds for image and video-based recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2273–2286, 2011.

Yue Xie and Stephen J. Wright. Complexity of proximal augmented Lagrangian for nonconvex optimization with nonlinear equality constraints. *Journal of Scientific Computing*, 86:1–30, 2021.

Yaguang Yang. Globally convergent optimization algorithms on Riemannian manifolds: Uniform framework for unconstrained and constrained optimization. *Journal of Optimization Theory and Applications*, 132(2):245–265, 2007.

Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Proceedings of the Annual Conference on Learning Theory*, pages 1617–1638, 2016.

Appendix A. Numerical experiments

A.1. Procrustes problem

We provide additional numerical experiment on matrices $A \in \mathbb{R}^{q\times q}, B \in \mathbb{R}^{p\times q}$ for $p = 1000$ and $q = 500$. Plots are presented in Figure 4. The large scale of the problem adds a lot of challenges for all algorithms and notably it affects RODCGM's convergence for constraints. However, we see that ODRGM again outperforms all the baselines. All experiments for the Procrustes problem are performed in PyTorch on a CPU with an Intel Core-i7 processor.

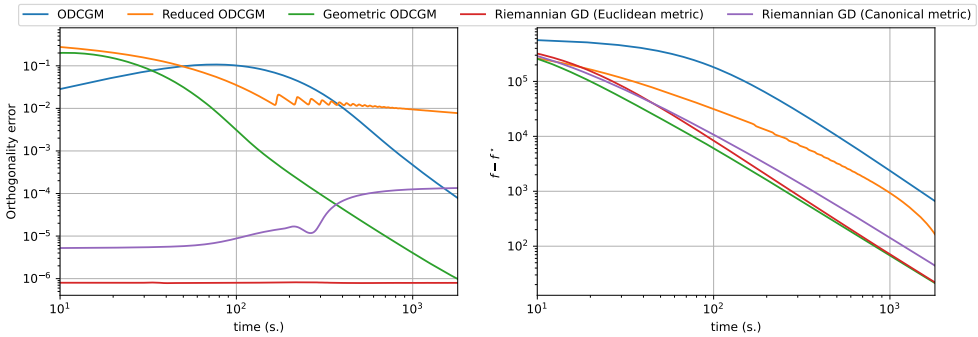

Figure 4: Comparison of ODCGM (blue), RODCGM (orange), and ODRGM (green) with Riemannian gradient descent with two different Riemannian metrics (red and purple). The left plot shows the orthogonality error for X, the right plot shows the convergence of the objective function (averages over 128 seeds).

A set K that satisfies A1. Even though the existence of $K = \{x : \|h(x)\| \leqslant r_1\}$ satisfying **A** 1 is guaranteed as soon as $\mathcal{M}$ is a compact manifold, finding the right constant $r_1 > 0$ is problem dependent and might be nontrivial. Nevertheless, in the case of Stiefel's manifold we can show that for $r_1 < 1$ and $A = \mathrm{Id}$, the set $K := \{x : \|h(x)\| \leqslant r_1\}$ satisfies **A**1. Furthermore, in that case it holds that $\alpha_m \geqslant 4(1 - \sqrt{r_1})$.

First, recall that the Stiefel's manifold is defined as $\{X \in \mathbb{R}^{p\times q} : X^\top X = \mathrm{Id}\}$. Thus, (see e.g. Boumal (2020)) we have $\nabla h(x)^\top : \mathbb{R}^{p\times q} \to S^q$, with $S^q \subset \mathbb{R}^{q\times q}$ denoting the set of symmetric matrices and

$$\nabla h(X)^\top H = X^\top H + H^\top X\,.$$

To compute $\lambda_{\min}(\nabla h(X)^\top \nabla h(X))$, we need to minimize $\langle H', \nabla h(X)^\top \nabla h(X) H'\rangle$ over the set of $H' \in S^q$ of unitary norm. We have:

$$\begin{aligned}\left\|\nabla h(X)H'\right\|_2^2 := \langle H', \nabla h(X)^\top(\nabla h(X)H')\rangle &= \langle H', X^\top \nabla h(X)H' + H'\nabla h(X)^\top X\rangle \\ &= \langle H', (\nabla h(X)^\top X)^\top H' + \nabla h(X)^\top X\rangle\end{aligned}$$

and $\nabla h(X)^\top X = 2X^\top X$. Thus, we obtain:

$$2\,\mathrm{tr}(H'X^\top X H' + H'H'X^\top X) = 4\,\mathrm{tr}(H'X^\top X H')\,.$$

Without losing generality, we can consider the case where $X^\top X$ is diagonal with, $\lambda_1 \geqslant \cdots \geqslant \lambda_q$, being its eigenvalues. If we denote $h_1, \ldots, h_q \in \mathbb{R}^q$ the lines of H', we have $\|H'\|^2 = \sum_{i=1}^q \|h_i\|^2$ and

$$\operatorname{tr}(H' X^\top X H') = \sum_{i=1}^q \lambda_i \|h_i\|^2 .$$

Thus, the best way to minimize this sum is to take $h_1, \ldots, h_{q-1} = 0$ and $h_q = (0, \ldots, 1)^\top$. We obtain:

$$\min_{H' \in S^q} \left\|\nabla h(X) H'\right\|_2^2 = \min_{H' \in S^q} (\left\|H' \nabla h(X)^\top \nabla h(X) H'\right\|) = 4 \min_{H' \in S^q} \operatorname{tr}(H' X^\top X H') = 4\lambda_q .$$

Therefore, $\lambda_{\min}(\nabla h(X)^\top \nabla h(X)) = 4\lambda_{\min}(X^\top X)$. Furthermore, if $\|h(X)\| \leqslant r_1 < 1$, then:

$$\left\|X^\top X - \mathrm{Id}\right\|^2 = \sum_{i=1}^q (1 - \lambda_i)^2 \leqslant r_1 ,$$

and therefore $(1 - \lambda_q) \leqslant \sqrt{r_1} \implies \lambda_{\min}(X^\top X) = \lambda_q \geqslant 1 - \sqrt{r_1} > 0$. This shows that for $r_1 < 1$ and $A = \mathrm{Id}$, the set $K := \{x : \|h(x)\| \leqslant r_1\}$ satisfies **A**1. Furthermore, in that case it holds that $\alpha_m \geqslant 4(1 - \sqrt{r_1})$.

Euclidean projection on tangent space. Additionally, in the case of optimization over the Stiefel manifold $\mathcal{M} = \{X \in \mathbb{R}^{p \times q} : X^\top X = \mathrm{Id}\}$, we discuss a way of projecting onto $V(x)$ (necessary for `ODCGM`) which is more efficient than solving a linear system of size pq.

First, we notice that the tangent space can be described as follows (see Gao et al. (2022)):

$$V(X) = \{Y \in \mathbb{R}^{p \times q} : Y^\top X + X^\top Y = 0\} .$$

To optimize over this set, we apply the Lagrange multipliers method

$$\min_{Y \in \mathbb{R}^{p \times q}} \frac{1}{2} \|Y - U\|_2^2 , \quad \text{s.t. } Y^\top X + X^\top Y = 0. \tag{23}$$

To solve this problem, we start from the reparametrization $Y = (X^+)^\top Z$, with Z a skew-symmetric matrix, and where X^+ denotes the pseudoinverse. In this way, we obtain:

$$\min_{Z \in \mathbb{R}^{n \times p}} \frac{1}{2} \left\|(X^+)^\top Z - U\right\|_2^2 , \quad \text{s.t. } Z^\top + Z = 0 . \tag{24}$$

The first order optimality condition implies $X^+((X^+)^\top Z - U) - (\Lambda + \Lambda^\top) = 0$, where Λ are Lagrange multipliers. By the properties of the pseudoinverse matrices with full column rank we have $X^+((X^+)^\top = (X^\top X)^{-1}$ and thus

$$Z = (X^\top X)(\Lambda + \Lambda^\top) + X^\top U.$$

Next, we have to choose M to satisfy $Z^\top + Z = 0$:

$$(X^\top X)(\Lambda + \Lambda^\top) + (\Lambda + \Lambda^\top)(X^\top X) = -(X^\top U + U^\top X) .$$

Since the right-hand side is symmetric, we only need to compute, over P, any solution to the following *Sylvester's equation*

$$(X^\top X)P + P(X^\top X) = -2(X^\top U + U^\top X)$$

and symmetrize it, that is $\Lambda = (P + P^\top)/2$. The solution to this system can be easily found using the SVD of the matrix X. Notice that since $X^\top X \to I$ we may expect that all operations will be numerically stable. The total complexity of the Euclidian projection is therefore equal to $\mathcal{O}(pq^2)$.

A.2. Hanging chain

We compute the shape of a hanging chain as a numerical example. The chain has length $l = 10$ and is divided into $N+1$ segments of equal length $r = 10/(N+1)$. Each two segments are connected by a joint and a torsion spring, which models the stiffness of the chain. The torsion spring has a spring constant of $k_s = 100$. The chain is suspended at positions $(0, 0)$ and $(9, 0)$, and an example with three nodes ($N = 3$) is shown in Figure 5. The optimization variables are given by the coordinates $\xi_i = (x_i, y_i)$ of the nodes, $i = 1, \ldots, N$, and a non-convex distance constraint restricts the length of each segment to r. We compute the shape of the chain by minimizing its potential energy, i.e.,

$$\begin{aligned} \min_{(\xi_1,\ldots,\xi_N)\in\mathbb{R}^{2N}} \frac{1}{N^3} \sum_{i=1}^{N} \left(\frac{k_s}{r^4} (\xi_{i-1} - \xi_i)^\mathsf{T} (\xi_{i+1} - \xi_i) + y_i \right) \\ \text{s.t.} \quad \sqrt{(\xi_{k-1} - \xi_k)^\mathsf{T} (\xi_{k-1} - \xi_k)} \leqslant r, \ k = 1, 2, \ldots, N+1, \end{aligned} \tag{25}$$

where $\xi_0 = (0, 0)$ and $\xi_{N+1} = (9, 0)$ are the two endpoints. We note that the first term of the objective function contains a discrete approximation to the curvature of the chain that models the spring potential, while the second term corresponds to the gravitational potential. This example is motivated by the fact that it leads to a simple problem formulation that includes non-convex distance constraints, but also allows us to scale N to values of 10^5 or more. Finally, Euler-Bernoulli beam theory gives us a reasonable initial estimate for the start of the optimization. All calculations are performed in MATLAB on a standard laptop (Dell XPS 15 with an Intel Core-i7 processor, 32 gigabytes of RAM, and a Windows operating system).

We start with a chain of $10,001$ segments, leading to an optimization problem with $20,000$ decision variables and $10,001$ nonlinear constraints. We compare the three algorithms: `ODCGM` with $A(x) = \alpha(\nabla h(x)^\top \nabla h(x))^{-1}$ (denoted simply by `ODCGM`), `RODCGM` , and an augmented Lagrangian approach. Figure 5 (right) shows the initial estimate and the final result as computed by `ODCGM` (the result of the other algorithms is similar). The step size for `ODCGM` is set to $\gamma_k = T$, where $T = 0.1/k_s$ and $\alpha = 0.05/T$; the step size for `RODCGM` is set to $\gamma_k = T$ for $k \leqslant 100$ and $\gamma_k = T/\sqrt{k - 100}$ for $k > 100$ (the scaling with $1/k_s$ results from the Hessian of (25)). Figure 3 (main text) shows the value of the objective function and the root mean square error of the constraint violation over the course of the optimization. We find that `ODCGM` leads to fast convergence in terms of constraint violations and function value, while the convergence of the augmented Lagrangian approach and `RODCGM` is much slower. Moreover, the performance of the augmented Lagrangian is quite sensitive to the initial value of the dual variable, which may even lead to divergence. In contrast, setting the step size of `ODCGM` and `RODCGM` is very simple. Figure 6 illustrates that `RODCGM` must be executed with decreasing step size; if a constant step size is used, the constraint violations remain as shown in the left panel, which is also consistent with our theoretical analysis. The right panel shows the execution time per iteration of the different algorithms by computing a moving average over past iterations. We conclude that `RODCGM` and the augmented Lagrangian require only about one-fifth of the time of `ODCGM` for a single iteration. This can be explained by the fact that `ODCGM` requires the solution of a linear system of size $30,0001 \times 30,0001$ at each iteration (we have exploited parsimony but have not taken into account the special structure of the equality constraints in (25)). Although `RODCGM` and the augmented Lagrangian approach have lower execution time per iteration, they also converges much more slowly.

In order to highlight the potential of `RODCGM` , we run the same example for $N = 200,000$, which results in a large-scale optimization problem with $400,000$ decision variables and $200,001$

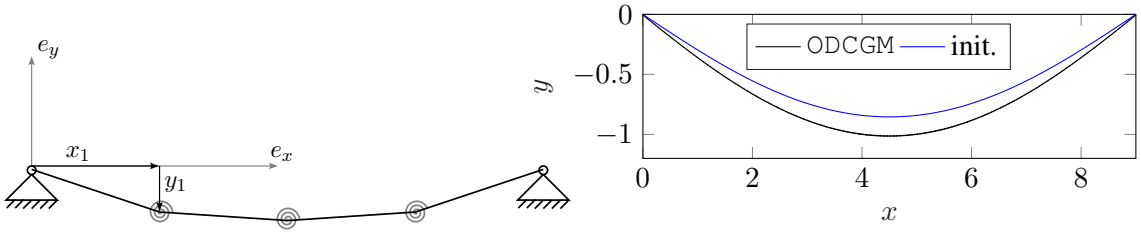

Figure 5: The figure shows a sketch of the hanging chain (left), the results arising from optimizing (25) (black, right) and the results predicted by the Euler-Bernoulli beam theory (red, right). The predictions from the Euler-Bernoulli beam theory are used as initial guess.

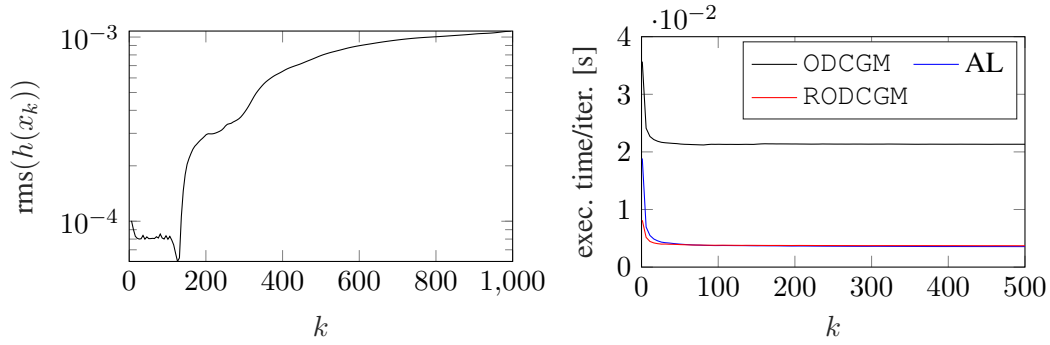

Figure 6: The figure on the left shows that RODCGM with a constant step size $\gamma_k = T$ does not converge and leads to significant constraint violations. The figure on the right shows the execution time per iteration of the different algorithms (moving average over past iterations). We note that the curve of the augmented Lagrangian and RODCGM are essentially superimposed.

non-convex equality constraints. In this case, solving the resulting Karush-Kuhn-Tucker system at every iteration, which is required for ODCGM , becomes challenging. However, RODCGM can still be applied and requires only about 0.085 seconds for executing a single iteration. The resulting function values and the evolution of the constraint violations are shown in Figure 7.

Appendix B. Supplementary proofs

B.1. Proof of Theorem 4

We preface the proof with two elementary results.

Lemma 7 *Consider $n_h \leqslant n$, $y \in \mathbb{R}^n$, $b \in \mathbb{R}^{n_h}$ and $W \in \mathbb{R}^{n \times n_h}$, with W being of full rank. It holds that:*

$$\operatorname*{arg\,min}_{v \in \mathbb{R}^n : W^\top v = b} \frac{1}{2} \|v + y\|^2 = -y + W(W^\top W)^{-1}(W^\top y + b)$$

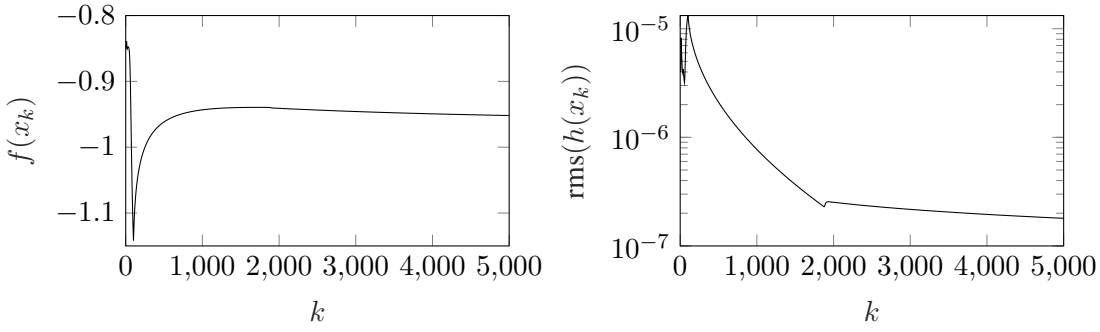

Figure 7: This figure shows the results from applying RODCGM with $\gamma_k \sim 1/\sqrt{k}$ for large k to (25) with $N = 200,000$. This leads to an optimization problem with $400,000$ decision variables and $200,001$ non-convex constraints. The evolution of the function value is shown on the left, whereas the evolution of the constraint violations is shown on the right.

Corollary 8 *If $x \in \mathbb{R}^n$ is such that $\nabla h(x)$ is of full rank, then:*

$$-\nabla_V f(x) = \underset{v \in V(x)}{\arg\min} \frac{1}{2} \|v + \nabla f(x)\|^2 = -\nabla f + \nabla h(x) D_h(x) \nabla h(x)^\top \nabla f(x)\,, \tag{26}$$

where $D_h(x) := (\nabla h(x)^\top \nabla h(x))^{-1}$.

The following proposition is the key element in our proof. It mainly follows from a Taylor expansion of Λ_M.

Proposition 9 (Discrete Lyapunov function) *Let A1–2 hold and let $\overline{M}$ be the one of Theorem 2. If for all $k \in \mathbb{N}$, $\gamma_k \leqslant \alpha_m^{-1}$, then for all $M \geqslant \overline{M}$, it holds:*

$$\mathbb{E}_k[\Lambda_M(x_{k+1})] - \Lambda_M(x_k) \leqslant -\gamma_k \|v_k\|^2 \left(1 - \frac{L_f + M L_h}{2}\gamma_k\right) + \frac{L_f + M L_h}{2}\sigma^2 \gamma_k^2\,. \tag{27}$$

Proof Since f is gradient Lipschitz continuous on K and $\mathbb{E}_k[\eta_{k+1}] = 0$, it holds that

$$\begin{aligned}
\mathbb{E}_k[f(x_{k+1})] - f(x_k) &\leqslant \gamma_k \nabla f(x_k)^\top v_k + \frac{L_f}{2}\gamma_k^2 \mathbb{E}_k[\|v_k + \eta_{k+1}\|^2] \\
&\leqslant \gamma_k \nabla f(x_k)^\top v_k + \frac{L_f}{2}\gamma_k^2 (\|v_k\|^2 + \sigma^2) \\
&\leqslant -\gamma_k \|v_k\|^2 \left(1 - \frac{L_f}{2}\gamma_k\right) + \gamma_k (\nabla f(x_k) + v_k)^\top v_k + \frac{L_f}{2}\gamma_k^2 \sigma^2 \\
&\leqslant -\gamma_k \|v_k\|^2 \left(1 - \frac{L_f}{2}\gamma_k\right) + \gamma_k M_1 \|h(x_k)\| + \frac{L_f}{2}\gamma_k^2 \sigma^2\,,
\end{aligned} \tag{28}$$

where the second inequality follows from **A**2-iv) and the last inequality follows from (10).

Similarly, since h is gradient Lipschitz on K, we obtain:

$$\begin{aligned}\mathbb{E}_k[\|h(x_{k+1})\|] &\leqslant \mathbb{E}_k[\left\|h(x_k) + \gamma_k \nabla h(x_k)^\top (v_k + \eta_{k+1})\right\|] + \frac{L_h}{2}\gamma_k^2 \mathbb{E}_k[\|v_k + \eta_{k+1}\|^2] \\ &\leqslant \left\|h(x_k) + \gamma_k \nabla h(x_k)^\top v_k\right\| + \frac{L_h}{2}\gamma_k^2 \|v_k\|^2 + \frac{L_h}{2}\gamma_k^2 \sigma^2 \\ &\leqslant \left\|h(x_k) - \gamma_k \nabla h(x_k)^\top \nabla h(x_k) A(x_k) h(x_k)\right\| + \frac{L_h}{2}\gamma_k^2 (\|v_k\|^2 + \sigma^2) \\ &\leqslant (1 - \alpha_m \gamma_k) \|h(x_k)\| + \frac{L_h}{2}\gamma_k^2 (\|v_k\|^2 + \sigma^2)\,,\end{aligned} \tag{29}$$

where in the second inequality we have used that $\eta_{k+1} \in V(x_k)$ and in the last inequality our choice of (γ_k) with **A**1-iv).

Combining Equations (28) and (29) with the fact that $M \geqslant \overline{M} = M_1/\alpha_m$ completes the proof. ■

The following corollary is obtained by telescoping (27).

Corollary 10 *Under the assumptions of Theorem 4, for $N > 0$, it holds that:*

$$\sum_{i=0}^{N-1} \mathbb{E}[\|v_i\|^2] \leqslant 2\frac{\mathbb{E}[\Lambda_M(x_0)] - \mathbb{E}[\Lambda_M(x_{N-1})]}{\gamma} + N\gamma(L_f + ML_h)\sigma^2\,. \tag{30}$$

B.1.1. DETERMINISTIC CASE: $\sigma = 0$

Fix $\sigma = 0$, from Corollary 10, we obtain:

$$\gamma \sum_{i=0}^{N-1} \|v_i\|^2 \leqslant 2(\Lambda_M(x_0) - \Lambda_M(x_{N-1}))\,.$$

This implies (16) and shows that $\|v_k\| \to 0$. Now notice that

$$\|v_k\|^2 = \|\nabla_V f(x_k)\|^2 + \|\nabla h(x_k) A(x_k) h(x_k)\|^2\,.$$

Since by **A**1 both A and ∇h are of full rank on K, this implies that $\|h(x_k)\| \to 0$. Thus, if x^* is an accumulation point of (x_k), then it must satisfy $h(x^*) = 0$, or, in other words, $x^* \in \mathcal{M}$. Finally, by continuity of O_D, we also have $0 = \lim_{k\to\infty} \|\mathrm{O}_D(x_k)\| = \|\mathrm{O}_D(x^*)\| = \|\mathrm{Grad}\, f(x^*)\|$, which completes the proof.

B.1.2. THE GENERAL CASE

Using Corollary 10, we obtain:

$$\begin{aligned}\mathbb{E}\left[\left\|v_{\hat{k}}\right\|^2\right] = \frac{1}{N}\sum_{i=0}^{N-1} \mathbb{E}[\|v_i\|^2] &\leqslant 2\frac{\mathbb{E}[\Lambda_M(x_0)] - \mathbb{E}[\Lambda_M(x_{N-1})]}{N\gamma} + \gamma(L_f + ML_h)\sigma^2 \\ &\leqslant \frac{2D_M}{N\gamma} + \gamma(L_f + ML_h)\sigma^2\,.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\left\|v_{\hat{k}}\right\|^2\right] &\leqslant \frac{2D_M}{N\gamma} + \bar{D}(L_f + ML_h)\frac{\sigma}{\sqrt{N}} \\ &\leqslant \frac{2D_M}{N}\max(\alpha_m, L_f + ML_h, \bar{D}^{-1}\sigma\sqrt{N}) + \bar{D}(L_f + ML_h)\frac{\sigma}{\sqrt{N}} \\ &\leqslant \frac{2D_M}{N}(\alpha_m + L_f + ML_h) + \frac{\sigma}{\sqrt{N}}\left(\bar{D}(L_f + ML_h) + \frac{2D_M}{\bar{D}}\right),\end{aligned}$$

which completes the proof.

B.2. Safe step size

In this section we prove Proposition 3. Recall that C_h, C_f denote the Lipschitz constants of h, f on K and $C_A = \sup_{x\in K}\|\nabla hA\|$.

Proof Let $k \in \mathbb{N}$ be such that $\|h(x_k)\| \leqslant r_1 - \delta$, we will show that $\|h(x_{k+1})\| \leqslant r_1 - \delta$, which will complete the proof by an immediate induction.

First, notice that if $\|x_{k+1} - x_k\| \leqslant \delta/C_h$, then $x_{k+1} \in K$. Indeed, assume that $x_{k+1} \notin K$ and denote for $t \in [0,1]$, $x_t = x_k + t(x_{k+1} - x_k)$. Let $u = \inf\{t \in [0,1] : x_t \notin K\}$ and note that by continuity of h, $\|h(x_u)\| = r_1$. This implies that

$$\delta \leqslant \|h(x_u)\| - \|h(x_k)\| \leqslant \|h(x_u) - h(x_k)\| \leqslant C_h\|x_u - x_k\| \leqslant u\delta\,.$$

Thus, u must be equal to 1, which is a contradiction.

Now,

$$\|x_{k+1} - x_k\|^2 \leqslant 2\gamma_k^2(\|v_k\|^2 + b^2)\,,$$

and since $\nabla_V f$ is orthogonal to ∇hAh, we obtain:

$$\|v_k\|^2 \leqslant \|\nabla h(x_k)A(x_k)h(x_k)\|^2 + \|\nabla_V f\|^2 \leqslant C_A^2\|h(x_k)\|^2 + C_f^2 \leqslant C_A^2 r_1^2 + C_f^2\,, \tag{31}$$

Hence, we get,

$$\|x_{k+1} - x_k\|^2 \leqslant 2\gamma^2(C_A^2 r_1^2 + C_f^2 + b^2) \leqslant \frac{\delta^2}{C_h^2}\,,$$

which shows that x_{k+1} remain in K. Now, since $x_k, x_{k+1} \in K$ and ∇h is L_h-Lipschitz on K, it holds that:

$$\left\|h(x_{k+1}) - h(x_k) - \gamma_k\nabla h(x_k)^\top v_k\right\| \leqslant \frac{L_h}{2}\|x_{k+1} - x_k\|^2\,,$$

where we have used the fact that $\eta_{k+1} \in V(x_k)$. Thus,

$$\|h(x_{k+1})\| \leqslant \left\|h(x_k) + \gamma_k\nabla h(x_k)^\top v_k\right\| + \frac{L_h}{2}\|x_{k+1} - x_k\|^2\,.$$

Recall that $\mathrm{Id} \in \mathbb{R}^{n_h\times n_h}$ denotes the identity matrix. It holds that:

$$\begin{aligned}\|h(x_{k+1})\| &\leqslant \left\|(\mathrm{Id} - \gamma_k\nabla h(x_k)^\top\nabla h(x_k)A(x_k))h(x_k)\right\| + \frac{L_h}{2}\|x_{k+1} - x_k\|^2 \\ &\leqslant (1 - \alpha_m\gamma_k)\|h(x_k)\| + \frac{L_h}{2}\|x_{k+1} - x_k\|^2\,.\end{aligned}$$

By examining (31) we can actually obtain a tighter upper bound on $\|x_{k+1} - x_k\|^2$

$$\|x_{k+1} - x_k\|^2 \leqslant 2b^2 + 2C_A^2 r_1 \|h(x_k)\| + 2C_f^2 \,,$$

which yields

$$\|h(x_{k+1})\| \leqslant \|h(x_k)\| + \gamma_k \|h(x_k)\| (\gamma_k L_h C_A^2 r_1 - \alpha_m) + \gamma_k^2 L_h (C_f^2 + b^2) \,.$$

Since $\gamma \leqslant \alpha_m/(2L_h C_A^2 r_1)$, it holds that:

$$\|h(x_{k+1})\| \leqslant \|h(x_k)\| - \alpha_m \gamma_k \|h(x_k)\| /2 + \gamma_k^2 L_h (C_f^2 + b^2) \,.$$

Therefore, if $\alpha_m \|h(x_k)\| \geqslant 2\gamma_k (L_h C_f^2 + b^2)$, then $\|h(x_{k+1})\| \leqslant \|h(x_k)\|$. Otherwise,

$$\|h(x_{k+1})\| \leqslant \|h(x_k)\| + \gamma_k^2 L_h (C_f^2 + b^2) \leqslant (L_h C_f^2 + b^2)\gamma_k \left(\frac{2}{\alpha_m} + \gamma_k \right) \leqslant \frac{3(L_h C_f^2 + b^2)\gamma}{2\alpha_m} \leqslant r_1 - \delta \,,$$

where the last inequality comes from our choice of γ. ■

Remark 11 *Although (15) may be intractable, it shows that the iterates remain in K for a sufficiently small γ. Therefore, we can combine our algorithm with standard line search techniques (see Nocedal and Wright (2006)). For example, if we set a threshold $\overline{\gamma}$, we check whether iterates with step sizes smaller than the threshold remain in K. If this is not the case, the threshold is divided by 2. Such a change of the threshold value can only occur finitely often, so that the convergence rates of Theorem 4 remain true.*

B.3. Proof of Theorem 5

We start with the following observation that characterizes the solution to the projection on $\tilde{V}(x) = \{v \in \mathbb{R}^n : \nabla H(x)^\top v = 0\}$, where $H(x) = \|h(x)\|^2 /2$.

Corollary 12 *Let $f \colon \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and let $\mathcal{M} = \{x \in \mathbb{R}^n : H(x) = 0\}$. Then, for any x such that $\nabla h(x)$ is of full rank, it holds*

$$\nabla_{\tilde{V}} f(x) = \nabla f(x) + \lambda(x) \cdot \nabla H(x),$$

where

$$\lambda(x) = \begin{cases} 0, & x \in \mathcal{M} \\ -\frac{\nabla H(x)^\top \nabla f(x)}{\|\nabla H(x)\|^2}, & x \notin \mathcal{M} \end{cases}$$

Proof Apply Lemma 7 with $n_h = 1, W = \nabla H(x)$ and $b = 0$ for $x \notin \mathcal{M}$. ■

Remark 13 *Even through $\lambda(x) \to -\infty$ as $x \to \mathcal{M}$, we have that projected gradients are always bounded:*

$$\left\|\nabla_{\tilde{V}} f(x)\right\| \leqslant \|\nabla f(x)\| + \frac{|\nabla H(x)^\top \nabla f(x)|}{\|\nabla H(x)\|^2} \cdot \|\nabla H(x)\| \leqslant 2 \|\nabla f(x)\| \,.$$

Next we provide a lemma that guarantees convergence of $H(x_k)$ under the specific choice of $A(x) = \alpha(x)\,\mathrm{Id}$, where $\alpha(x) = \alpha \cdot \frac{H(x)}{\|\nabla H(x)\|^2}$ for a constant $\alpha > 0$ if $x \notin \mathcal{M}$, and $\alpha(x) = 0$ if $x \in \mathcal{M}$.

Lemma 14 *Assume A1'-2. Assume that for any $k > 0$, $\alpha\gamma_k \leqslant 1$. Define $v_k = -\alpha(x_k)\nabla H(x_k) - \nabla_{\tilde{V}} f(x_k)$. It holds*

$$\mathbb{E}[H(x_{k+1})] \leqslant H(x_0) \cdot \prod_{j=0}^{k}(1-\alpha\gamma_j) + \mathbb{E}\left[\frac{L_H}{2}\sum_{j=0}^{k}\gamma_j^2(\|v_j\|^2+\sigma^2)\prod_{\ell=j+1}^{k}(1-\alpha\gamma_\ell)\right].$$

Furthermore, if (γ_k) is a non-increasing sequence, then for all $N \in \mathbb{N}$,

$$\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k H(x_k)\right] \leqslant \frac{1}{\alpha}H(x_0) + \frac{L_H}{2\alpha}\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k^2\|v_k\|^2\right] + \frac{L_H}{2\alpha}\sum_{k=0}^{N-1}\gamma_k^2\sigma^2\,.$$

Proof Since $h(x)$ is Lispchitz and has Lipschitz-continuous gradients we have that $H(x)$ also has Lipschitz-continuous gradients with constant L_H. Thus for any $k \in \mathbb{N}$,

$$\mathbb{E}_k[H(x_{k+1})] \leqslant H(x_k) + \gamma_k \nabla H(x_k)^\top v_k + \frac{L_H\gamma_k^2}{2}\mathbb{E}_k[\|v_k+\eta_{k+1}\|^2].$$

Notice that $\nabla_{\tilde{V}} f(x_k)$ is orthogonal to $\nabla H(x_k)$, thus

$$\nabla H(x_k)^\top v_k = -\alpha(x_k)\|\nabla H(x_k)\|^2 = -\alpha H(x_k)$$

by definition of $\alpha(x_k)$. Also notice that $\mathbb{E}_k[\|v_k+\eta_{k+1}\|^2] \leqslant \|v_k\|^2 + \sigma^2$. Therefore

$$\mathbb{E}_k[H(x_{k+1})] \leqslant (1-\alpha\gamma_k)H(x_k) + \frac{L_H\gamma_k^2}{2}\left(\|v_k\|^2+\sigma^2\right).$$

Rolling out this inequality we conclude the first statement. Next we sum all inequalities for all $k = 0, \ldots, N-1$ with weights γ_k

$$\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k H(x_k)\right] \leqslant H(x_0)\cdot\sum_{k=0}^{N-1}\gamma_k\prod_{j=0}^{k-1}(1-\alpha\gamma_j) + \frac{L_H}{2}\sum_{k=0}^{N-1}\gamma_k\sum_{j=0}^{k-1}\gamma_j^2(\|v_j\|^2+\sigma^2)\prod_{\ell=j+1}^{k-1}(1-\alpha\gamma_\ell)\,.$$

First, we apply Lemma 15 to the first term. Next, we change the order of summation and apply Lemma 15 again

$$\begin{aligned}\sum_{k=0}^{N-1}\gamma_k\sum_{j=0}^{k-1}\gamma_j^2(\|v_j\|^2+\sigma^2)\prod_{\ell=j+1}^{k-1}(1-\alpha\gamma_\ell) &= \sum_{j=0}^{N-1}\gamma_j^2(\|v_j\|^2+\sigma^2)\sum_{k=j+1}^{N-1}\gamma_k\prod_{\ell=j+1}^{k-1}(1-\alpha\gamma_\ell)\\ &\leqslant \frac{1}{\alpha}\sum_{j=0}^{N-1}\gamma_j^2(\|v_j\|^2+\sigma^2).\end{aligned}$$

■

Lemma 15 *Let $b > 0$ and $(\gamma_k)_{k\geqslant 0}$ be a non-increasing sequence such that $\gamma_0 \leqslant 1/b$. Then*

$$\sum_{k=0}^{n} \gamma_k \prod_{j=0}^{k-1}(1 - b\gamma_j) = \frac{1 - \prod_{j=0}^{n}(1 - b\gamma_j)}{b}$$

Proof Introduce $u_{i:j} = \prod_{\ell=i}^{j}(1 - b\gamma_\ell)$. Notice that $u_{0:k-1} - u_{0:k} = u_{0:k-1} \cdot b\gamma_k$. Summing this equation from 0 to n concludes the statement. ■

To provide rates of convergence for the final algorithm we have to prove the following proposition.

Proposition 16 *Assume A1'-2. Let $x_0 \in \mathcal{M}$. If for all $k \in \mathbb{N}$ $\gamma_k \equiv \gamma$ where $\gamma \leqslant \min(\alpha^{-1}, (L_f + \alpha L_H \mu_h^{-2})^{-1})$, then for any $N \in \mathbb{N}$, the following holds*

$$\mathbb{E}\left[\sum_{k=0}^{N-1} \gamma \left\|v_k\right\|^2\right] \leqslant 4\Delta_N + 2\left(L_f + \alpha L_H \mu_h^{-2}\right) \cdot \sigma^2 \cdot \gamma^2 N + 4\widetilde{C}^2 \cdot \alpha L_H \cdot \gamma^2 N + 4\widetilde{C} \cdot \sqrt{\frac{\alpha\gamma N}{2}} \sqrt{\frac{L_H \sigma^2}{2} \cdot \gamma^2 N},$$

where $D_0 = f(x_0) - \inf_{x\in K} f(x)$ and $\widetilde{C} = B_f M_h \mu_h^{-2}$.

Proof First, we use the smoothness of the function f

$$\mathbb{E}_k[f(x_{k+1})] \leqslant f(x_k) + \gamma_k \nabla f(x_k)^\top v_k + \frac{L_f \gamma_k^2}{2}(\|v_k\|^2 + \sigma^2).$$

Next, we notice that $\nabla f(x_k) + (\alpha(x_k) + \lambda(x_k))\nabla H(x_k) = -v_k$, thus

$$\begin{aligned}\mathbb{E}_k[f(x_{k+1})] \leqslant f(x_k) - \gamma_k \left(1 - \frac{L_f \gamma_k}{2}\right) \|v_k\|^2 + \frac{L_f \gamma_k^2 \sigma^2}{2} \\ - \gamma_k(\alpha(x_k) + \lambda(x_k))\nabla H(x_k)^\top v_k.\end{aligned} \tag{32}$$

By the orthogonality property and the choice of $\alpha(x_k)$, we have

$$\nabla H(x_k)^\top v_k = -\alpha(x_k) \|\nabla H(x_k)\|^2 = -\alpha H(x_k),$$

and therefore, rolling out inequality for any $N \in \mathbb{N}$ results in

$$\begin{aligned}\mathbb{E}[f(x_N)] \leqslant f(x_0) + \mathbb{E}\left[\sum_{k=0}^{N-1} \left(\frac{L_f \gamma_k^2}{2} - \gamma_k\right) \|v_k\|^2 \mathbb{E}\right] + \sum_{k=0}^{N-1} \frac{L_f \gamma_k^2 \sigma^2}{2} \\ - \alpha \mathbb{E}\left[\sum_{k=0}^{N-1} \gamma_k(\alpha(x_k) + \lambda(x_k))H(x_k)\right].\end{aligned} \tag{33}$$

We will analyze the last term next. To do it, we start from the definitions of $\alpha(x_k)$, $\lambda(x_k)$ and $H(x_k)$:

$$\left|\sum_{k=0}^{N-1} \gamma_k(\alpha(x_k) + \lambda(x_k))H(x_k)\right| \leqslant \sum_{k=0}^{N-1} \gamma_k \frac{|\alpha H(x_k) - \nabla f(x_k)^\top \nabla H(x_k)|}{\|\nabla h(x) h(x_k)\|^2} \cdot \frac{1}{2} \|h(x_k)\|^2 .$$

Next, we apply the Cauchy-Schwartz inequality combined with the definition of B_f, μ_h and M_h. This yields $\|\nabla h(x)h(x_k)\|^2 \geqslant \mu_h \|h(x)\|^2$ and

$$\left|\sum_{k=0}^{N-1} \gamma_k(\alpha(x_k) + \lambda(x_k))H(x_k)\right| \leqslant \frac{\alpha}{2\mu_h^2}\sum_{k=0}^{N-1}\gamma_k H(x_k) + \frac{B_f M_h}{\sqrt{2}\cdot\mu_h^2}\sum_{k=0}^{N-1}\gamma_k\sqrt{H(x_k)}.$$

The Cauchy-Schwartz inequality implies

$$\sum_{k=0}^{N-1}\gamma_k\sqrt{H(x_k)} \leqslant \sqrt{\sum_{k=0}^{N-1}\gamma_k}\cdot\sqrt{\sum_{k=0}^{N-1}\gamma_k H(x_k)}.$$

Next, we are going to deal with the expectation term. Jensen's inequality applied to a square root concludes

$$\mathbb{E}\left[\left|\sum_{k=0}^{N-1} \gamma_k(\alpha(x_k) + \lambda(x_k))H(x_k)\right|\right] \leqslant \frac{\alpha}{2\mu_h^2}\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k H(x_k)\right] + \frac{B_f M_h}{\sqrt{2}\cdot\mu_h^2}\sqrt{\sum_{k=0}^{N-1}\gamma_k}\sqrt{\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k H(x_k)\right]}.$$

By assumption we have $\gamma_k \leqslant \alpha^{-1}$, so we can apply Lemma 14 and obtain

$$\mathbb{E}\left[\left|\sum_{k=0}^{N-1} \gamma_k(\alpha(x_k) + \lambda(x_k))H(x_k)\right|\right] \leqslant \frac{1}{2\mu_h^2}\left(H(x_0) + \frac{L_H}{2}\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k^2\|v_k\|^2\right] + \frac{L_H}{2}\sum_{k=0}^{N-1}\gamma_k^2\sigma^2\right) + \frac{B_f M_h}{\sqrt{2\alpha}\mu_h^2}\sqrt{\sum_{k=0}^{N-1}\gamma_k}\cdot\sqrt{H(x_0) + \frac{L_H}{2}\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma_k^2\|v_k\|^2\right] + \frac{L_H}{2}\sum_{k=0}^{N-1}\gamma_k^2\sigma^2}.$$

For simplicity we assume $H(x_0) = 0$ and that $\gamma_k \equiv \gamma$ satisfies the following inequality

$$\gamma \leqslant \frac{1}{L_f + \alpha L_H \mu_h^{-2}}.$$

Define $\Delta f_N = f(x_0) - f(x_N)$ and $S_N = \mathbb{E}\left[\sum_{k=0}^{N-1}\|v_k\|^2\right]$, then by rearranging term in (33) and applying inequality $\sqrt{a+b} \leqslant \sqrt{a} + \sqrt{b}$ for positive a, b

$$\frac{\gamma}{2}S_N \leqslant \Delta f_N + \frac{(L_f + \alpha L_H\mu_h^{-2})\cdot\sigma^2}{2}\gamma^2 N + \frac{B_f M_h L_H}{2\mu_h^2}\cdot(\alpha^{1/2}\gamma^{3/2}N)\cdot\sigma + \frac{B_f M_h}{\mu_h^2}\cdot\sqrt{\frac{\alpha\gamma^2 N\cdot L_H}{2}}\sqrt{\gamma S_N}.$$

This yields a quadratic inequality in $\sqrt{\gamma S_N}$ that can be easily solved. Using the fact that if $x^2 \leqslant 2ax + 2b$ then $x \leqslant a + \sqrt{a^2 + 2b} \leqslant 2a + \sqrt{2b}$ and a numeric inequality $(2a + \sqrt{2b})^2 \leqslant 8a^2 + 4b$,

we obtain

$$\mathbb{E}\left[\sum_{k=0}^{N-1}\gamma\left\|v_k\right\|^2\right] \leqslant 4\Delta f_N + 2\left(L_f + \alpha L_H \mu_h^{-2}\right)\cdot\sigma^2\cdot\gamma^2 N + \left(\frac{4B_f^2 M_h^2\cdot\alpha L_H}{\mu_h^4}\right)\cdot\gamma^2 N + \frac{2B_f M_h L_H}{\mu_h^2}\cdot(\alpha^{1/2}\gamma^{3/2}N)\cdot\sigma\,.$$

Finally, we notice that $D_0 = f(x_0) - \inf_{x\in K} f(x)$ is an upper bound on Δf_N. ■

Now we are ready to prove the main convergence results. The results will be divided into two independent propositions.

Proposition 17 (Convergence in the deterministic case) *Assume A1'-2 and let $x_0 \in \mathcal{M}$. Let $\sigma^2 = 0$ and also define $\bar{D}$ as a known constant. If for all $k \in \mathbb{N}$, $\gamma_k \equiv \bar{\gamma}$ where $\bar{\gamma} = \min(\alpha^{-1}, (L_f + \alpha L_H\mu_h^{-2})^{-1}, \bar{D}\cdot N^{-1/3})$, and $\alpha = \bar{\gamma}$ then for any $N \in \mathbb{N}$ the following holds*

$$\min_{k=0,\dots,N-1}\left\{\left\|\nabla_{\tilde{V}} f(x_k)\right\|^2 + \frac{1}{2}\left\|h(x_k)\right\|^2\right\} \leqslant \frac{8D_0(L_f + L_H\mu_h^{-2})}{N} + \left(\frac{8D_0}{\bar{D}} + 8\widetilde{C}L_H\cdot\bar{D}\right)\cdot N^{-2/3},$$

where $D_0 = f(x_0) - \inf_{x\in K} f(x)$ and $\widetilde{C} = B_f M_h \mu_h^{-2}$.

In particular, RODCGM outputs a point $\hat{x}$ satisfying $\|h(\hat{x})\| \leqslant \varepsilon$ and $\left\|\nabla_{\tilde{V}} f(\hat{x})\right\| \leqslant \varepsilon$ in $\mathcal{O}(\varepsilon^{-3})$ iterations.

Proof We apply Proposition 16 with $\sigma^2 = 0$ and without expectations

$$\sum_{k=0}^{N-1}\gamma\left\|v_k\right\|^2 \leqslant 4D_0 + 4\widetilde{C}^2\cdot\alpha L_H\cdot\gamma^2 N.$$

At the same time, by combining of Lemma 14 and Proposition 16, we obtain

$$\sum_{k=0}^{N-1}\gamma\left(\left\|v_k\right\|^2 + H(x_k)\right) \leqslant 4D_0\left(1+\frac{\gamma}{\alpha}\right) + 4\widetilde{C}^2\cdot\alpha L_H\cdot\gamma^2 N\left(1+\frac{\gamma}{\alpha}\right).$$

By taking $\alpha = \gamma$ and using orthongonality property of v_k we have

$$\min_{k=0,\dots,N-1}\left\{\left\|\nabla_{\tilde{V}} f(x_k)\right\|^2 + \frac{1}{2}\left\|h(x_k)\right\|^2\right\} \leqslant \frac{8D_0}{\gamma N} + 8\widetilde{C}^2 L_H\cdot\gamma^2.$$

To balance these two terms we choose $\gamma_k \equiv \bar{\gamma} = \min(1, (L_f + \bar{\gamma}L_H\mu_h^{-2})^{-1}, \bar{D}\cdot N^{1/3})$ and obtain

$$\min_{k=0,\dots,N-1}\left\{\left\|\nabla_{\tilde{V}} f(x_k)\right\|^2 + \frac{1}{2}\left\|h(x_k)\right\|^2\right\} \leqslant \frac{8D_0(L_f + L_H\mu_h^{-2})}{N} + \left(\frac{8D_0}{\bar{D}} + 8\widetilde{C}L_H\cdot\bar{D}\right)\cdot N^{-2/3}.$$

■

Proposition 18 (Convergence in the stochastic case) *Assume A1'-2 and let $x_0 \in \mathcal{M}$. Let $\sigma^2 > 0$ and also define $\bar{D}$ as a known constant. Let for all $k \in \mathbb{N}$, $\gamma_k \equiv \bar{\gamma}$, where $\bar{\gamma} = \min(\alpha^{-1}, (L_f + \alpha L_H \mu_h^{-2})^{-1}, \bar{D} \cdot N^{-1/2})$ and fix the number of steps $N > 0$. Let $\hat{k}$ be a uniform index sampled from the set $\{0, \dots, N-1\}$. Then, the following holds*

$$\mathbb{E}\left[\left\|\nabla_{\tilde{V}} f(x_{\hat{k}})\right\|^2 + \frac{1}{2}\left\|h(x_{\hat{k}})\right\|^2\right] \leqslant \frac{4D_0(L_f + L_h\mu_h^{-2})}{N} + \frac{4D_0}{\bar{D} \cdot N^{1/2}} + \frac{4\widetilde{C}^2\bar{D}^2 \cdot L_H}{N} + \frac{\bar{D}}{N^{1/2}}\left(2\left(L_f + \gamma L_H \mu_h^{-2}\right) \cdot \sigma^2 + 2\widetilde{C} \cdot \sqrt{\frac{L_H\sigma^2}{2}}\right)$$

where $D_0 = f(x_0) - \inf_{x \in K} f(x)$ and $\widetilde{C} = B_f M_h \mu_h^{-2}$.

In particular, RODCGM outputs a point $\hat{x} = x_{\hat{k}}$ such that $\mathbb{E}[\|h(\hat{x})\|] \leqslant \varepsilon$ and $\mathbb{E}[\left\|\nabla_{\tilde{V}} f(\hat{x})\right\|] \leqslant \varepsilon$ in $\mathcal{O}(\varepsilon^{-4})$ iterations.

Proof Let us start from Proposition 16 by taking $\alpha = \gamma$

$$\mathbb{E}\left[\sum_{k=0}^{N-1} \gamma \left\|v_k\right\|^2\right] \leqslant 4D_0 + 2\left(L_f + \gamma L_H \mu_h^{-2}\right) \cdot \sigma^2 \cdot \gamma^2 N + 4\widetilde{C}^2 \cdot L_H \cdot \gamma^3 N + 2\widetilde{C} \cdot \sqrt{\frac{L_H\sigma^2}{2}} \cdot \gamma^2 N.$$

Combining Lemma 14 with Proposition 16 and using the orthogonality property yields

$$\mathbb{E}\left[\frac{1}{N}\sum_{k=0}^{N-1}\left\{\left\|\nabla_{\tilde{V}} f(x_k)\right\|^2 + \frac{1}{2}\left\|h(x_k)\right\|^2\right\}\right] \leqslant \frac{4D_0}{\gamma N} + \gamma \cdot \left(2\left(L_f + \gamma L_H \mu_h^{-2}\right) \cdot \sigma^2 + 2\widetilde{C} \cdot \sqrt{\frac{L_H\sigma^2}{2}}\right) + 4\widetilde{C}^2 \cdot L_H \cdot \gamma^2.$$

Notice that the left-hand side corresponds exactly to the expectation over $\hat{k}$. Thus, by taking $\gamma_k \equiv \bar{\gamma} = \min(1, (L_f + \bar{\gamma} L_H \mu_h^{-2})^{-1}, \bar{D} \cdot N^{-1/2})$, we obtain

$$\mathbb{E}\left[\left\|\nabla_{\tilde{V}} f(x_{\hat{k}})\right\|^2 + \frac{1}{2}\left\|h(x_{\hat{k}})\right\|^2\right] \leqslant \frac{4D_0(L_f + L_h\mu_h^{-2})}{N} + \frac{4D_0}{\bar{D} \cdot N^{1/2}} + \frac{4\widetilde{C}^2\bar{D}^2 \cdot L_H}{N} + \frac{\bar{D}}{N^{1/2}}\left(2\left(L_f + \gamma L_H \mu_h^{-2}\right) \cdot \sigma^2 + 2\widetilde{C} \cdot \sqrt{\frac{L_H\sigma^2}{2}}\right).$$

■

B.4. Proof of Theorem 6

Lemma 19 *Let $(\mathcal{M}, g)$ be a Riemannian manifold with a Riemannian metric $g_x(\xi, \eta) = \langle \xi, G_x \eta \rangle$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then for any $x \in \mathcal{M}$ we have*

$$\operatorname{Grad}_{\mathcal{M}} f(x) = \operatorname*{arg\,min}_{v \in \mathcal{T}_x(\mathcal{M})} \frac{1}{2}\left\|v - G_x^{-1}\nabla f(x)\right\|_{g_x}^2. \tag{34}$$

Proof Let $v^\star$ be a solution to (34). Then it can be written as a solution to the following variational inequality

$$\forall v \in \mathcal{T}_x(\mathcal{M}) : \langle \nabla F(v^\star), v - v^\star \rangle \geqslant 0,$$

where $F(v) = \frac{1}{2} \left\| v - G_x^{-1} \nabla f(x) \right\|_{g_x}^2$. By a direct computation we have

$$\forall v \in \mathcal{T}_x(\mathcal{M}) : \langle G_x v^\star - \nabla f(x), v - v^\star \rangle \geqslant 0.$$

Fix an arbitrary $\xi \in \mathcal{T}_x(\mathcal{M})$. Since $\mathcal{T}_x(\mathcal{M})$ is a vector space, we have that $v_1 = \xi + v^\star$ and $v_2 = -\xi + v^\star$ lies in $\mathcal{T}_x(\mathcal{M})$. Thus,

$$\langle G_x v^\star - \nabla f(x), \xi \rangle \geqslant 0, \quad \langle G_x v^\star - \nabla f(x), -\xi \rangle \geqslant 0.$$

Therefore, by an arbitrary choice of ξ we have

$$\forall \xi \in \mathcal{T}_x(\mathcal{M}) : \langle G_x v^\star - \nabla f(x), \xi \rangle = 0.$$

■

Proposition 20 *Let $x \in \mathbb{R}^n$ be such that $\nabla h(x)$ is of full rank. It holds that:*

$$\mathrm{O}_G(x) = -\nabla h A h - Q^{-1} \nabla f + Q^{-1} \nabla h B \nabla f \,,$$

with $B(x) \in \mathbb{R}^{n_h \times n}$ defined as:

$$B(x) = (\nabla h(x)^\top Q^{-1}(x) \nabla h(x))^{-1} \nabla h(x)^\top Q^{-1}(x) \,.$$

Proof It holds that:

$$\operatorname*{arg\,min}_{v \in V} \left\| v + Q^{-1} \nabla f \right\|_q^2 = \operatorname*{arg\,min}_{v \in V} \left\| Q^{1/2} v + Q^{-1/2} \nabla f \right\|^2 = Q^{-1/2} \operatorname*{arg\,min}_{v \in Q^{1/2} V} \left\| v + Q^{-1/2} \nabla f \right\|^2 . \tag{35}$$

By noticing that $Q^{1/2} V := \{ v \in \mathbb{R}^n : (Q^{-1/2} \nabla h)^\top v = 0 \}$, we obtain our claim by applying Lemma 7 with $W = Q^{-1/2} \nabla h$, $y = Q^{-1/2} \nabla f$ and $b = 0$. ■

Denote M_q the following constant:

$$M_q := \sup_{x \in K} \left\| (\nabla h A h)^\top Q \nabla h A + \nabla f^\top \nabla h A - 2 (\nabla h B \nabla f)^\top \nabla h A \right\| . \tag{36}$$

The constant M_q will play the same role as M_1 (notice that $M_q = M_1$, if $Q = \mathrm{Id}_n$) in the proof of Theorem 4.

Lemma 21 *Let A1-A3. It holds that*

$$\left\| (Q^{-1} \nabla f + v)^\top Q v \right\| \leqslant M_q \left\| h \right\| \quad \text{where } v = \mathrm{O}_G(x) \,.$$

Proof Note that, as previously, $\nabla h^\top v = -\nabla h^\top \nabla h A h$. Thus, using Proposition 20, we obtain:

$$\begin{aligned}(Q^{-1}\nabla f + v)^\top Q v &= -(\nabla h A h)^\top Q v + (B\nabla f)^\top \nabla h^\top v \\ &= -(\nabla h A h)^\top Q v - (B\nabla f)^\top \nabla h^\top \nabla h A h \\ &= (\nabla h A h)^\top Q (\nabla h A h) + (\nabla h A h)^\top \nabla f - 2(\nabla h B \nabla f)^\top (\nabla h A h) \\ &= \left((\nabla h A h)^\top Q \nabla h A + \nabla f^\top \nabla h A - 2(\nabla h B \nabla f)^\top \nabla h A\right) h\,.\end{aligned}$$

This completes the proof by the definition of M_q. ■

Denote $\overline{M_q} := M_q/\alpha_m$. The following is an extension of Proposition 9 to the present case.

Proposition 22 (Geometry aware discrete Lyapunov function) *Let **A** 1–**A** 3 hold. If for all k, $\gamma_k \leqslant \alpha_m^{-1}$, then for all $M \geqslant \overline{M}_q$, it holds:*

$$\mathbb{E}_k[\Lambda_M(x_{k+1})] - \Lambda_M(x_k) \leqslant -\gamma_k \|v_k\|^2 \left(\frac{1}{C_q} - \frac{L_f + ML_h}{2}\gamma_k\right) + \frac{L_f + ML_h}{2}\sigma^2\gamma_k^2\,. \tag{37}$$

Proof Following the same path as in the proof of Proposition 9, we obtain a generalization of (28):

$$\begin{aligned}\mathbb{E}_k[f(x_{k+1})] - f(x_k) &\leqslant \gamma_k \nabla f(x_k)^\top v_k + \frac{L_f}{2}\gamma_k^2 \mathbb{E}_k[\|v_k + \eta_{k+1}\|]^2 \\ &\leqslant \gamma_k (Q^{-1}(x_k)\nabla f(x_k))^\top Q(x_k) v_k + \frac{L_f}{2}\gamma_k^2(\|v_k\|^2 + \sigma^2) \\ &\leqslant -\gamma_k v_k^\top Q(x_k) v_k + (Q^{-1}(x_k)\nabla f(x_k) + v_k)^\top Q(x_k) v_k + \frac{L_f}{2}\gamma_k^2(\|v_k\|^2 + \sigma^2) \\ &\leqslant -\gamma_k \|v_k\|^2 \left(\frac{1}{C_q} - \frac{L_f}{2}\gamma_k\right) + M_q \|h(x_k)\| + \frac{L_f}{2}\gamma_k^2\sigma^2\,,\end{aligned}$$

where we have used Lemma 21 for the third and **A**3 for the fourth inequality. Since (29) remains unchanged, we obtain the claimed inequality. ■

The end of the proof of Theorem 6 then follows, *mutatis mutatis*, the one of Theorem 4, upon replacing Proposition 9 with Proposition 22.