Lyapunov Functions for First-Order Methods: Tight Automated Convergence Guarantees

Adrien Taylor [*1] **Bryan Van Scoy** [*2] **Laurent Lessard** [*2 3]

Abstract

We present a novel way of generating Lyapunov functions for proving linear convergence rates of first-order optimization methods. Our approach provably obtains the *fastest* linear convergence rate that can be verified by a quadratic Lyapunov function (with given states), and only relies on solving a small-sized semidefinite program. Our approach combines the advantages of performance estimation problems (PEP, due to Drori & Teboulle (2014)) and integral quadratic constraints (IQC, due to Lessard et al. (2016)), and relies on convex interpolation (due to Taylor et al. (2017c;b)).

1. Introduction

In this work, we study first-order methods for solving the (unconstrained) minimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) \tag{$\mathcal{P}$}$$

where $f:\mathbb{R}^d\to\mathbb{R}$. In the sequel, we focus on the case where f is L-smooth and μ-strongly convex, though our methodology can be adapted to a broader class of problems.

To solve ($\mathcal{P}$), we consider methods that iteratively update their estimate of the optimizer using only gradient evaluations. One method for proving convergence of such methods is by finding a *Lyapunov function*.

A Lyapunov function can be interpreted as defining an "energy" that decreases geometrically with each iteration of the method, with an energy of zero corresponding to reaching the optimal solution of ($\mathcal{P}$). The existence of such an energy function thus provides a straightforward certificate of linear convergence for the iterative method.

In this paper, we present an automated way of generating quadratic Lyapunov functions for certifying linear convergence of first-order iterative methods to solve ($\mathcal{P}$). The procedure relies on solving a small-sized semidefinite program (SDP) so it is computationally efficient. Moreover, the procedure is *tight*, meaning that if the SDP is infeasible, then no such quadratic Lyapunov function exists.

Our results unify recent SDP-based works for certifying convergence of first-order methods, namely: performance estimation problems (Drori & Teboulle, 2014; Taylor et al., 2017c) and integral quadratic constraints from robust control (Lessard et al., 2016), using smooth strongly convex interpolation (Taylor et al., 2017c). These connections are further discussed in Section 4.3.

1.1. Organization

The paper is organized as follows. We describe the class of methods under consideration and basic properties of Lyapunov functions in Sections 2 and 3 respectively. Our main results are then presented in Section 4, which also features numerical examples and comparisons to other approaches. The corresponding proof is presented in Section 5. Finally, we explore extensions of our approach in Section 6, and conclude in Section 7.

1.2. Preliminaries

A function $f:\mathbb{R}^d\to\mathbb{R}$ is called L-smooth if its gradient is Lipschitz continuous with parameter L, i.e.,

$$\|\nabla f(x)-\nabla f(y)\|\leq L\|x-y\| \quad \text{for all } x,y\in\mathbb{R}^d. \tag{1}$$

Furthermore, f is called convex if

$$f(x)\geq f(y)+\nabla f(y)^{\mathsf{T}}(x-y) \quad \text{for all } x,y\in\mathbb{R}^d, \tag{2}$$

and μ-strongly convex if $f(x)-\frac{\mu}{2}\|x\|^2$ is convex. The set of L-smooth and μ-strongly convex functions is denoted $\mathcal{F}_{\mu,L}$, and we define $\kappa:=\frac{L}{\mu}$, the corresponding condition number.

[*]Equal contribution [1]INRIA, Département d'informatique de l'ENS, École normale supérieure, CNRS, PSL Research University, Paris, France [2]Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, Wisconsin, USA [3]Department of Elecrical and Computer Engineering, University of Wisconsin–Madison, Madison, Wisconsin, USA. Correspondence to: Adrien Taylor <adrien.taylor@inria.fr>, Bryan Van Scoy <vanscoy@wisc.edu>, Laurent Lessard <laurent.lessard@wisc.edu>.

When $f \in \mathcal{F}_{\mu,L}$ with $0 < \mu \leq L$, optimization problem ($\mathcal{P}$) has a unique minimizer denoted $x_\star := \arg\min_x f(x)$. The function and gradient values at optimality are denoted $f_\star := f(x_\star)$ and $g_\star := \nabla f(x_\star) = \mathbf{0}_d$, respectively.

2. First-order iterative fixed-step methods

To solve the optimization problem ($\mathcal{P}$), we consider *first-order iterative fixed-step methods* of the form

$$\begin{aligned} y_k &= \sum_{j=0}^{N} \gamma_j \, x_{k-j} \\ x_{k+1} &= \sum_{j=0}^{N} \beta_j \, x_{k-j} - \alpha \, \nabla f(y_k) \end{aligned} \qquad (\mathcal{M})$$

for $k \geq 0$ where α, β_j, γ_j are the (fixed) step-sizes and $x_j \in \mathbb{R}^d$ for $j = -N, \ldots, 0$ are the initial conditions. We call the constant $N \geq 0$ the *degree* of the method.

Many first-order optimization methods are of the form ($\mathcal{M}$), including: the Gradient Method, Heavy Ball Method (Polyak, 1964), Fast Gradient Method for smooth strongly convex minimization (Nesterov, 2004), Triple Momentum Method (Van Scoy et al., 2018), and Robust Momentum Method (Cyrus et al., 2018).

For method ($\mathcal{M}$) to solve ($\mathcal{P}$), it must have a fixed-point at the optimizer $x_\star$. Hence, we require the step-sizes to satisfy

$$\sum_{j=0}^{N} \beta_j = 1 \qquad \text{and} \qquad \sum_{j=0}^{N} \gamma_j = 1.$$

For simplicity, let us define the concatenated error vectors at iteration k as $\mathbf{x}_k, \mathbf{g}_k \in \mathbb{R}^{(N+1)d}$ and $\mathbf{f}_k \in \mathbb{R}^{N+1}$ with

$$\mathbf{x}_k := \begin{bmatrix} (x_k - x_\star)^\mathsf{T} & \ldots & (x_{k-N} - x_\star)^\mathsf{T} \end{bmatrix}^\mathsf{T} \qquad (3a)$$

$$\mathbf{g}_k := \begin{bmatrix} (g_k - g_\star)^\mathsf{T} & \ldots & (g_{k-N} - g_\star)^\mathsf{T} \end{bmatrix}^\mathsf{T} \qquad (3b)$$

$$\mathbf{f}_k := \begin{bmatrix} (f_k - f_\star) & \ldots & (f_{k-N} - f_\star) \end{bmatrix}^\mathsf{T} \qquad (3c)$$

where $x_k \in \mathbb{R}^d$ are the iterates, $f_k := f(y_k) \in \mathbb{R}$ are the function values, and $g_k := \nabla f(y_k) \in \mathbb{R}^d$ are the gradient values. Note that we shifted $(\mathbf{x}_k, \mathbf{g}_k, \mathbf{f}_k)$ so that the optimal solution corresponds to $(\mathbf{x}_\star, \mathbf{g}_\star, \mathbf{f}_\star) = (\mathbf{0}, \mathbf{0}, \mathbf{0})$.

3. What is a Lyapunov function?

Lyapunov functions are one of the fundamental tools in control theory that can be used to verify stability of a dynamical system (Kalman & Bertram, 1960a;b).

Consider applying method ($\mathcal{M}$) to solve problem ($\mathcal{P}$). Our goal is to find the smallest possible $0 \leq \rho < 1$ such that $\{x_k\}$ converges linearly to the optimizer $x_\star$ with rate ρ. A *Lyapunov function* $\mathcal{V}$ is a continuous function $\mathcal{V} : \mathbb{R}^n \to \mathbb{R}$ that satisfies the following properties:

1. (nonnegative) $\mathcal{V}(\xi) \geq 0$ for all ξ,
2. (zero at fixed-point) $\mathcal{V}(\xi) = 0$ if and only if $\xi = \xi_\star$,
3. (radially unbounded) $\mathcal{V}(\xi) \to \infty$ as $\|\xi\| \to \infty$,
4. (decreasing) $\mathcal{V}(\xi_{k+1}) \leq \rho^2 \, \mathcal{V}(\xi_k)$ for $k \geq N$,

where $\xi_k := (\mathbf{x}_k, \mathbf{g}_k, \mathbf{f}_k)$ is the *state* of the system at iteration k. The state at iteration k includes past iterates, function values, and gradient values from iterations $k - N$ up to k. If we can find such a $\mathcal{V}$, then it can be used to show that the state converges linearly to the fixed-point from any initial condition (the rate of convergence depends on both ρ and the structure of $\mathcal{V}$).

Lyapunov functions are typically found by searching over a parameterized family of functions (called Lyapunov function candidates). In the simple case where the state $\{\xi_k\}$ is generated by a linear dynamical system, one can search over quadratic Lyapunov function candidates by solving a semidefinite program, as illustrated in Example 1 below.

Example 1 (Quadratic Lyapunov function)**.** *Consider the linear dynamical system described by*

$$\xi_{k+1} = A\xi_k, \qquad \xi_0 \in \mathbb{R}^n$$

with fixed-point $\xi_\star \in \mathbb{R}^n$ (i.e., $\xi_\star = A\xi_\star$). Suppose that

$$\underset{P \in \mathbb{S}^n}{\text{feasible}} \quad 0 \succeq A^\mathsf{T} P A - \rho^2 P, \quad P \succ 0 \qquad (4)$$

has solution $P_\star$. Then a Lyapunov function for the system is

$$\mathcal{V}(\xi) = (\xi - \xi_\star)^\mathsf{T} P_\star (\xi - \xi_\star)$$

which can be used to show that $\xi_k \to \xi_\star$ linearly with rate ρ. Specifically, we have the bound

$$\|\xi_k - \xi_\star\|_{P_\star} \leq \rho^k \, \|\xi_0 - \xi_\star\|_{P_\star} \quad \text{for } k \geq 0.$$

To find the best bound, we can perform a bisection search on ρ to find the smallest ρ such that (4) is feasible.

Note that although $\mathcal{V}$ depends explicitly on the fixed point $\xi_\star$, we do not need to know $\xi_\star$ to solve the SDP (4).

The linear dynamical system of Example 1 converges linearly if and only if a quadratic Lyapunov function exists, which happens if and only if the SDP (4) is feasible (Lyapunov & Fuller, 1992; Vidyasagar, 2002).

4. Main results

Similar to Example 1, we now show how to use quadratic Lyapunov functions to prove linear convergence of a first-order iterative fixed-step method applied to the minimization of a smooth strongly convex function. Furthermore, we show that such Lyapunov function exists if and only if a small-sized semidefinite program is feasible (whose optimal solution produces the Lyapunov function).

4.1. Quadratic Lyapunov functions

We begin with sufficiency: if we can find a quadratic Lyapunov function, we can use it to prove linear convergence.

Lemma 2 (Quadratic Lyapunov function)**.** *Consider applying the first-order iterative fixed-step method ($\mathcal{M}$) of degree N to a smooth strongly convex function $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ with $0 < \mu \leq L$. Define the state $\xi_k := (\mathbf{x}_k, \mathbf{g}_k, \mathbf{f}_k)$ as in (3). Consider the quadratic function*

$$\mathcal{V}(\xi_k) = \begin{bmatrix}\mathbf{x}_k \\ \mathbf{g}_k\end{bmatrix}^\mathsf{T} (P \otimes I_d) \begin{bmatrix}\mathbf{x}_k \\ \mathbf{g}_k\end{bmatrix} + p^\mathsf{T}\mathbf{f}_k \quad \text{for } k \geq N \tag{5}$$

with parameters $P \in \mathbb{S}^{2(N+1)}$ and $p \in \mathbb{R}^{N+1}$, and where $\otimes$ denotes the Kronecker product. Suppose $\mathcal{V}$ is a Lyapunov function for the system with rate ρ. Then, the following bound is satisfied:

$$\mathcal{V}(\xi_k) \leq \rho^{2(k-N)}\, \mathcal{V}(\xi_N) \quad \text{for } k \geq N. \tag{6}$$

Proof. Suppose $\mathcal{V}$ is a Lyapunov function for method ($\mathcal{M}$) with $f \in \mathcal{F}_{\mu,L}$. Then $0 \geq \mathcal{V}(\xi_{i+1}) - \rho^2\, \mathcal{V}(\xi_i)$ for $i \geq N$. Multiplying this inequality by $\rho^{2(k-i-1)}$ and summing over $i = N, \ldots, k-1$ gives a telescoping sum that yields (6). ■

As a consequence to Lemma 2, we have the relations:

$$\|x_k - x_\star\| = \mathcal{O}(\rho^k) \tag{7a}$$
$$\|\nabla f(y_k)\| = \mathcal{O}(\rho^k) \tag{7b}$$
$$f(y_k) - f_\star = \mathcal{O}(\rho^{2k}) \tag{7c}$$

where $x_\star \in \mathbb{R}^d$ is the optimizer of ($\mathcal{P}$) and $f_\star := f(x_\star)$.

Remark 3. *The Lyapunov function (5) is only defined for $k \geq N$ since the state ξ_k is a function of the previous N function and gradient values. This is why the bound (6) is expressed in terms of $\mathcal{V}(\xi_N)$.*

Remark 4. *The states used in the Lyapunov function (5) can be modified to include other iterates (such as y_k) in the quadratic term as well as the function and gradient values evaluated at iterates other than y_k. We chose the form in (5) because it contains all necessary ingredients while also being straightforward to generalize to other cases.*

In addition, note that the structure of (5) makes it permutation-invariant *(i.e., it does not depend on the ordering of the coordinate set). This is largely motivated by the fact that there is no reason to favor any coordinate among $\mathbb{R}^d$.*

Lemma 2 shows that if we can find a quadratic Lyapunov function, then we can use this to prove linear convergence of method ($\mathcal{M}$) when $f \in \mathcal{F}_{\mu,L}$. In the following section, we construct an SDP whose feasibility is necessary and sufficient for the existence of such a Lyapunov function.

4.2. SDP for quadratic Lyapunov functions

Given parameters α, β_j, and γ_j for a method ($\mathcal{M}$) of degree N and a rate ρ to be verified, we construct the semidefinite program as follows.

Step 1: Initialization. First, we initialize the row vectors $\bar{x}_k^{(K)}, \bar{g}_k^{(K)} \in \mathbb{R}^{N+K+2}$ and $\bar{f}_k^{(K)} \in \mathbb{R}^{K+1}$, corresponding to the initial conditions, gradient values, and function values, respectively, as

$$\bar{x}_k^{(K)} := \mathbf{e}_{k+N+1}^\mathsf{T} \quad \text{for } k \in \{-N, \ldots, 0\} \tag{8a}$$
$$\bar{g}_k^{(K)} := \mathbf{e}_{k+N+2}^\mathsf{T} \quad \text{for } k \in \{0, \ldots, K\} \tag{8b}$$
$$\bar{f}_k^{(K)} := \mathbf{e}_{k+1}^\mathsf{T} \quad \text{for } k \in \{0, \ldots, K\} \tag{8c}$$

for $K \in \{N, N+1\}$ ($\mathbf{e}_i$ denotes the i^{th} unit vector with appropriate dimension). These form a basis for all iterates, function values, and gradient values up to iteration K. Also, define the row vectors corresponding to the fixed-point as

$$\bar{y}_\star^{(K)} := \mathbf{0}_{N+K+2}^\mathsf{T}, \quad \bar{g}_\star^{(K)} := \mathbf{0}_{N+K+2}^\mathsf{T}, \quad \bar{f}_\star^{(K)} := \mathbf{0}_{K+1}^\mathsf{T}.$$

We also introduce the following SDP variables:

$$P \in \mathbb{S}^{2(N+1)}, \qquad \lambda_{ij} \in \mathbb{R} \quad \text{for } i,j \in \mathcal{I}_N,$$
$$p \in \mathbb{R}^{N+1}, \qquad \eta_{ij} \in \mathbb{R} \quad \text{for } i,j \in \mathcal{I}_{N+1},$$

where $\mathcal{I}_K := \{0, 1, \ldots, K, \star\}$ is an index set.

Step 2: Method. Next, we iterate the method for $k = 0, \ldots, K$ using the row vectors we previously defined.

$$\bar{y}_k^{(K)} = \sum_{j=0}^{N} \gamma_j\, \bar{x}_{k-j}^{(K)} \tag{9a}$$
$$\bar{x}_{k+1}^{(K)} = \sum_{j=0}^{N} \beta_j\, \bar{x}_{k-j}^{(K)} - \alpha\, \bar{g}_k^{(K)}. \tag{9b}$$

Step 3: Interpolation conditions[1]**.** Using the computed vectors, define $m_{ij}^{(K)} \in \mathbb{R}^{K+1}$ and $M_{ij}^{(K)} \in \mathbb{S}^{N+K+2}$ as

$$m_{ij}^{(K)} := (L-\mu)\big(\bar{f}_i^{(K)} - \bar{f}_j^{(K)}\big)^\mathsf{T} \tag{10a}$$
$$M_{ij}^{(K)} := \frac{1}{2} \begin{bmatrix}\bar{y}_i^{(K)} \\ \bar{y}_j^{(K)} \\ \bar{g}_i^{(K)} \\ \bar{g}_j^{(K)}\end{bmatrix}^\mathsf{T} M \begin{bmatrix}\bar{y}_i^{(K)} \\ \bar{y}_j^{(K)} \\ \bar{g}_i^{(K)} \\ \bar{g}_j^{(K)}\end{bmatrix} \tag{10b}$$

[1]The terms $M_{ij}^{(K)}$ and $m_{ij}^{(K)}$ are related to *interpolation* by smooth strongly convex functions as discussed in Section 5.1

for $i, j \in \mathcal{I}_K$ where

$$M := \begin{bmatrix} -\mu L & \mu L & \mu & -L \\ \mu L & -\mu L & -\mu & L \\ \mu & -\mu & -1 & 1 \\ -L & L & 1 & -1 \end{bmatrix}. \tag{11}$$

Step 4: Lyapunov function. We now construct the linear and quadratic terms in the Lyapunov function, denoted $v_k^{(K)} \in \mathbb{R}^{K+1}$ and $V_k^{(K)} \in \mathbb{S}^{N+K+2}$, respectively, as

$$v_k^{(K)} := p^\mathsf{T} \bar{\mathbf{f}}_k^{(K)} \tag{12a}$$

$$V_k^{(K)} := \begin{bmatrix} \bar{\mathbf{x}}_k^{(K)} \\ \bar{\mathbf{g}}_k^{(K)} \end{bmatrix}^\mathsf{T} P \begin{bmatrix} \bar{\mathbf{x}}_k^{(K)} \\ \bar{\mathbf{g}}_k^{(K)} \end{bmatrix} \tag{12b}$$

where the matrices $\bar{\mathbf{x}}_k^{(K)}, \bar{\mathbf{g}}_k^{(K)} \in \mathbb{R}^{(N+1)\times(N+K+2)}$ and $\bar{\mathbf{f}}_k^{(K)} \in \mathbb{R}^{(N+1)\times(K+1)}$ are defined as

$$\bar{\mathbf{x}}_k^{(K)} := \begin{bmatrix} \bar{x}_k^{(K)} \\ \vdots \\ \bar{x}_{k-N}^{(K)} \end{bmatrix} \quad \bar{\mathbf{g}}_k^{(K)} := \begin{bmatrix} \bar{g}_k^{(K)} \\ \vdots \\ \bar{g}_{k-N}^{(K)} \end{bmatrix} \quad \bar{\mathbf{f}}_k^{(K)} := \begin{bmatrix} \bar{f}_k^{(K)} \\ \vdots \\ \bar{f}_{k-N}^{(K)} \end{bmatrix}.$$

Also, define the decrease in the linear and quadratic terms of the Lyapunov function as

$$\Delta v_k^{(K)} := v_{k+1}^{(K)} - \rho^2 v_k^{(K)} \tag{13a}$$

$$\Delta V_k^{(K)} := V_{k+1}^{(K)} - \rho^2 V_k^{(K)} \tag{13b}$$

where ρ is the convergence rate to be verified.

Step 5: Semidefinite program. Finally, we compute the quadratic Lyapunov function (if one exists) for a given rate ρ by solving the following semidefinite program:

SDP for quadratic Lyapunov function (ρ-SDP)

$$\begin{aligned} \underset{\substack{P\in\mathbb{S}^{2(N+1)} \\ p\in\mathbb{R}^{N+1} \\ \{\lambda_{ij}\} \\ \{\eta_{ij}\}}}{\text{feasible}} \quad & 0 \prec V_N^{(N)} - \sum_{i,j\in\mathcal{I}_N} \lambda_{ij} M_{ij}^{(N)} \\ & 0 < v_N^{(N)} - \sum_{i,j\in\mathcal{I}_N} \lambda_{ij} m_{ij}^{(N)} \\ & 0 \succeq \Delta V_N^{(N+1)} + \sum_{i,j\in\mathcal{I}_{N+1}} \eta_{ij} M_{ij}^{(N+1)} \\ & 0 \geq \Delta v_N^{(N+1)} + \sum_{i,j\in\mathcal{I}_{N+1}} \eta_{ij} m_{ij}^{(N+1)} \\ & 0 \leq \lambda_{ij} \quad \text{for } i,j \in \mathcal{I}_N \\ & 0 \leq \eta_{ij} \quad \text{for } i,j \in \mathcal{I}_{N+1} \end{aligned}$$

Theorem 5 (Main Result)**.** *Consider applying the first-order iterative fixed-step method ($\mathcal{M}$) of degree N to a smooth strongly convex function $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ with $0 < \mu \leq L$. Let the step-sizes α, β_j, and γ_j be such that $\alpha \neq 0$, $\gamma_0 \neq 0$, and*

$$\sum_{j=0}^{N} \beta_j = \sum_{j=0}^{N} \gamma_j = 1.$$

Then there exists a quadratic Lyapunov function of the form (5) with rate ρ that is valid for all $d \in \mathbb{N}$ if and only if (ρ-SDP) is feasible.

From Theorem 5, we can perform bisection on ρ to find the minimum ρ such that (ρ-SDP) is feasible to produce the *fastest* linear convergence rate that is able to be verified using a quadratic Lyapunov function with states $(\mathbf{x}, \mathbf{g}, \mathbf{f})$.

4.3. Comparison to PEP and IQC

Our results are closely related to several other recent approaches utilizing semidefinite programs for studying convergence of first-order methods, which we discuss now.

Performance Estimation Problem (PEP). The performance estimation approach was introduced by Drori & Teboulle (2014) as a systematic way to obtain worst-case performance guarantees of a given method. In the context of fixed-step first-order methods, performance estimation problems (PEP) can be formulated as semidefinite programs.

The key idea in PEP is to look for a tuple $(x_{-N}, \ldots, x_0, f)$ such that the given algorithm behaves in the worst possible way, according to a given performance measure. The dual of PEP corresponding to the performance measure $\mathcal{V}(\xi_{N+1})/\mathcal{V}(\xi_N)$ for some fixed P and p is exactly the same as solving $\min_\rho \rho^2$ subject to (ρ-SDP) being feasible.

The difference between PEP and our approach is that in PEP, the optimization is for a fixed performance measure carried out over multiple timesteps. This yields exact worst-case bounds, but at the cost of solving an SDP whose size is proportional to the number of timesteps (this allows, among others, dealing with time-varying methods and sublinear convergence rates). In our approach, for a fixed ρ, we optimize *the performance measure itself*. This yields a Lyapunov function with a guaranteed decrease at every iteration while (1) maintaining tightness and (2) solving a small SDP of fixed size. Both approaches ensure tightness via *(smooth) convex interpolation*, developed by Taylor et al. (2017c).

Integral Quadratic Constraints (IQCs). Integral quadratic constraints are an analysis method for bounding the worst-case performance of dynamical systems in feedback with nonlinearities (Megretski & Rantzer, 1997). This approach was recently adapted for use in analyzing

first-order optimization algorithms (Lessard et al., 2016). In the optimization context, the nonlinear component is the gradient of the objective function, while the dynamical system is the iterative method being analyzed.

The key idea with IQCs is to replace the nonlinearity (∇f) by quadratic constraints that it must satisfy. This is precisely the idea behind *interpolation* (discussed in Section 5.1), which is a foundational concept in our methodology.

The difference between IQCs and our approach is that the interpolation conditions are necessary and sufficient to characterize ∇f when $f \in \mathcal{F}_{\mu,L}$. However, the sector IQC and weighted off-by-one IQC used by Lessard et al. (2016) are a strict subset of the interpolation conditions; they are only sufficient for describing ∇f when $f \in \mathcal{F}_{\mu,L}$. In particular, the IQC framework does not use any constraints on ∇f that explicitly involve function values. This amounts to solving (ρ-SDP) with additional constraints on λ_{ij} and η_{ij} such that all the function values cancel out in the SDP.

4.4. Numerical comparisons

To illustrate our results, we consider the Gradient Method (GM), Heavy Ball Method (HBM), Fast Gradient Method (FGM), and Triple Momentum Method (TMM). Each of these methods can be parametrized as

$$y_k = x_k + \gamma\,(x_k - x_{k-1}) \tag{14a}$$
$$x_{k+1} = x_k + \beta\,(x_k - x_{k-1}) - \alpha\,\nabla f(y_k) \tag{14b}$$

for $k \geq 0$ where $x_{-1}, x_0 \in \mathbb{R}^d$ are the initial conditions, and the parameters for each method are:

Method	α	β	γ
GM	$\frac{1}{L}$	0	0
HBM	$\frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$	$\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$	0
FGM	$\frac{1}{L}$	$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$	$\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$
TMM	$\frac{2\sqrt{L}-\sqrt{\mu}}{L\sqrt{L}}$	$\frac{(\sqrt{\kappa}-1)^2}{\kappa+\sqrt{\kappa}}$	$\frac{(\sqrt{\kappa}-1)^2}{2\kappa+\sqrt{\kappa}-1}$

We use (ρ-SDP) to find corresponding Lyapunov functions. The corresponding convergence rates are provided in Figure 1; the results match those obtained using IQCs (Lessard et al., 2016) for GM, HBM, and FGM, and those for TMM provided in (Van Scoy et al., 2018). For more complicated cases, the performance estimation toolbox PESTO (Taylor et al., 2017a) can be used to perform numerical validations.

For illustrative purposes, we present results obtained using a restricted class of Lyapunov functions. We fixed $\lambda_{ij} = 0$ in (ρ-SDP) and plotted the best achievable ρ in Figure 2. We observe that this restricted class is not sufficient to recover the rates obtained in Figure 1.

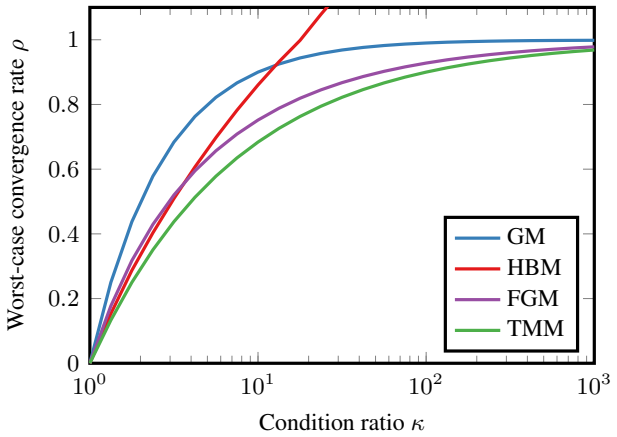

Figure 1. Worst-case linear convergence rates from (ρ-SDP).

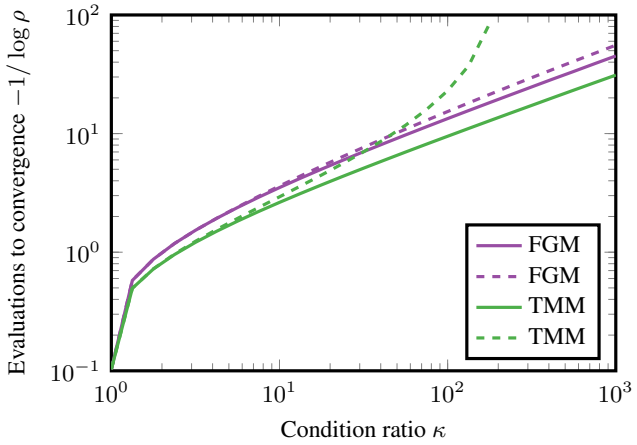

Figure 2. Order of magnitude of the worst-case number of iterations, which is $\mathcal{O}(-1/\log\rho)$, to solve problem ($\mathcal{P}$). The bounds are obtained by searching for Lyapunov functions of the form (5) in two cases: (i) (P, p) found using (ρ-SDP) (solid), and (ii) restricting $P \succ 0$ and $p > 0$ (dashed).

5. Proof of Theorem 5

5.1. Sampled smooth strongly convex functions

To prove Theorem 5, we first need a result on smooth strongly convex functions that are sampled at discrete points. Indeed, inequalities (1) and (2) completely characterize functions that are smooth and strongly convex. However, these inequalities are defined on an infinite set of points, and it was shown in Section 2.2 of (Taylor et al., 2017c) that using them to prove convergence may introduce conservatism. Therefore, in order to completely characterize points which are sampled from smooth strongly convex functions, we need the concept of *interpolation*.

The following theorem is borrowed from (Taylor et al., 2017c) and forms the basic building block for our analysis.

Theorem 6 ($\mathcal{F}_{\mu,L}$-interpolation)**.** *Let $\mathcal{I}$ be an index set, and consider the set of triples $S = \{(y_i, g_i, f_i)\}_{i\in\mathcal{I}}$ where $y_i, g_i \in \mathbb{R}^d$ and $f_i \in \mathbb{R}$ for all $i \in \mathcal{I}$. There exists a function[2] $f \in \mathcal{F}_{\mu,L}$ such that $f(y_i) = f_i$ and $\nabla f(y_i) = g_i$ for all $i \in \mathcal{I}$ if and only if $\phi_{ij} \geq 0$ for all $i, j \in \mathcal{I}$ where*

$$\phi_{ij} := (L-\mu)(f_i - f_j) + \begin{bmatrix} y_i \\ y_j \\ g_i \\ g_j \end{bmatrix}^\mathsf{T} (M \otimes I_d) \begin{bmatrix} y_i \\ y_j \\ g_i \\ g_j \end{bmatrix} \tag{15}$$

with $M \in \mathbb{S}^4$ defined in (11).

5.2. Positive definite quadratics from sampling

Recall from Section 3 that the Lyapunov function must satisfy two conditions: (i) $\mathcal{V}$ must be positive definite (i.e, nonnegative, zero at the fixed-point, and radially unbounded), and (ii) $\Delta\mathcal{V} := \mathcal{V}_{k+1} - \rho^2\,\mathcal{V}_k$ must be negative semidefinite (i.e., $\mathcal{V}$ must satisfy the decrease condition). To prove both (i) and (ii), we use the following theorem, which provides necessary and sufficient conditions for a quadratic form to be positive (semi-)definite when the iterates are generated by method ($\mathcal{M}$) applied to $f \in \mathcal{F}_{\mu,L}$.

Theorem 7 (Sampled positive definite quadratics)**.** *Consider applying the first-order iterative fixed-step method ($\mathcal{M}$) of degree N to a smooth strongly convex function $f \in \mathcal{F}_{\mu,L}(\mathbb{R}^d)$ for K iterations. Suppose the step-sizes α, β_j, and γ_j are such that $\alpha \neq 0$, $\gamma_0 \neq 0$, and*

$$\sum_{j=0}^{N} \beta_j = \sum_{j=0}^{N} \gamma_j = 1.$$

Define the vectors $\mathbf{x} \in \mathbb{R}^{(N+1)d}$, $\mathbf{g} \in \mathbb{R}^{(K+1)d}$, and $\mathbf{f} \in \mathbb{R}^{K+1}$ as

$$\mathbf{x} := \begin{bmatrix} (x_{-N} - x_\star)^\mathsf{T} & \dots & (x_0 - x_\star)^\mathsf{T} \end{bmatrix}^\mathsf{T} \tag{16a}$$

$$\mathbf{g} := \begin{bmatrix} (g_0 - g_\star)^\mathsf{T} & \dots & (g_K - g_\star)^\mathsf{T} \end{bmatrix}^\mathsf{T} \tag{16b}$$

$$\mathbf{f} := \begin{bmatrix} f_0 - f_\star & \dots & f_K - f_\star \end{bmatrix}^\mathsf{T} \tag{16c}$$

and denote the triple $\xi := (\mathbf{x}, \mathbf{g}, \mathbf{f})$. Define $m_{ij}^{(K)} \in \mathbb{R}^{K+1}$ and $M_{ij}^{(K)} \in \mathbb{S}^{N+K+2}$ such that

$$\phi_{ij}(\xi) = \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix}^\mathsf{T} (M_{ij}^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} + \left(m_{ij}^{(K)}\right)^\mathsf{T} \mathbf{f} \tag{17}$$

for $i, j \in \mathcal{I}_K := \{0, \dots, K, \star\}$ where ϕ_{ij} is defined in (15). Consider the quadratic function

$$\sigma(\xi) = \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix}^\mathsf{T} (Q \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} + q^\mathsf{T} \mathbf{f}$$

where $Q \in \mathbb{S}^{N+K+2}$ and $q \in \mathbb{R}^{N+1}$. Suppose the dimension d satisfies[3] $d \geq N + K + 2$.

Then σ is positive semidefinite (i.e., nonnegative) if and only if there exists $\tau_{ij} \geq 0$ for $i, j \in \mathcal{I}_K$ such that

$$0 \preceq Q - \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, M_{ij}^{(K)}$$

$$0 \leq q - \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, m_{ij}^{(K)}.$$

Furthermore, σ is positive definite if and only if there exists $\tau_{ij} \geq 0$ for $i, j \in \mathcal{I}_K$ such that

$$0 \prec Q - \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, M_{ij}^{(K)} \tag{19a}$$

$$0 < q - \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, m_{ij}^{(K)}. \tag{19b}$$

Proof. We prove the second statement that σ is positive definite if and only if there exists $\tau_{ij} \geq 0$ for $i, j \in \mathcal{I}_K$ such that (19) holds; the proof of the first statement is similar.

Here we prove that the conditions are sufficient for σ to be positive definite; necessity is more involved and can be found in the supplementary material.

(Sufficiency). Suppose there exists $\tau_{ij} \geq 0$ for $i, j \in \mathcal{I}_K$ such that (19) holds. Clearly, we have $\sigma(0) = 0$. Now assume that $\xi \neq 0$. Sum the following two inequalities: (i) take the Kronecker product of (19a) with I_d and multiply the result on the left and right by $\begin{bmatrix} \mathbf{x}^\mathsf{T} & \mathbf{g}^\mathsf{T} \end{bmatrix}$ and its transpose, respectively, and (ii) multiply the tranpose of (19b) on the right by $\mathbf{f}$. Doing so gives the inequality

$$0 < \sigma(\xi) - \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, \phi_{ij}(\xi) \tag{20}$$

which is strict due to the strict inequalities in (19) and since $\xi \neq 0$. Since $f \in \mathcal{F}_{\mu,L}$, we have $\phi_{ij} \geq 0$ from Thm. 6, so

$$0 \leq \sum_{i,j\in\mathcal{I}_K} \tau_{ij}\, \phi_{ij}(\xi) < \sigma(\xi).$$

Then $\sigma(\xi) \geq 0$, and $\sigma(\xi) = 0$ if and only if $\xi = 0$. Finally, note that the strict inequalities in (19) imply that the right side of (20) grows arbitrarily large as $\|\xi\| \to \infty$, so σ is radially unbounded. Thus, σ is positive definite. ∎

Remark 8. *Theorem 7 can be seen as a specialized application of the S-procedure (Boyd et al., 1994; Megretski & Treil, 1993) where the points in ξ are generated by method ($\mathcal{M}$), and the positive semidefinite quadratic terms come from the interpolation conditions in Theorem 6. While*

[2] In other words, we say that the set S is $\mathcal{F}_{\mu,L}$*-interpolable*.

[3] This requirement is only used for necessity.

the S-procedure is known to be lossy in certain cases (i.e., the conditions are sufficient but not necessary for σ to be positive (semi)definite), Theorem 7 shows that it is in fact lossless *under the large-scale assumption $d \geq N + K + 2$.*

We now apply Theorem 7 to obtain necessary and sufficient conditions for both $\mathcal{V}$ to be positive definite and $\Delta\mathcal{V} := \mathcal{V}(\xi_{k+1}) - \rho^2\,\mathcal{V}(\xi_k)$ to be negative semidefinite. To that end, note that the basis vectors $\bar{x}_k^{(K)}$, $\bar{g}_k^{(K)}$, and $\bar{f}_k^{(K)}$ in (8) are such that

$$\begin{aligned}
x_k - x_\star &= (\bar{x}_k^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} && \text{for } k \in \{-N, \dots, K\} \\
g_k - g_\star &= (\bar{g}_k^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} && \text{for } k \in \{0, \dots, K\} \\
f_k - f_\star &= \bar{f}_k^{(K)} \mathbf{f} && \text{for } k \in \{0, \dots, K\} \\
y_k - x_\star &= (\bar{y}_k^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} && \text{for } k \in \{0, \dots, K\}.
\end{aligned}$$

where $\mathbf{x}$, $\mathbf{g}$, and $\mathbf{f}$ are defined in (16) and we used the iterations in (9). We can then sum the following: (i) take the Kronecker product of $M_{ij}^{(K)}$ in (10) with I_d and multiply the result on the left and right by $\begin{bmatrix} \mathbf{x}^\mathsf{T} & \mathbf{g}^\mathsf{T} \end{bmatrix}$ and its transpose, respectively, and (ii) multiply the transpose of $m_{ij}^{(K)}$ on the right by $\mathbf{f}$. Adding these two quantities gives (17). Similarly, the Lyapunov function in (5) is given by

$$\mathcal{V}(\xi_k) = \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix}^\mathsf{T} (V_k^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} + \left(v_k^{(K)}\right)^\mathsf{T} \mathbf{f} \qquad (21)$$

and the decrease in the Lyapunov function is given by

$$\Delta V(\xi_k) = \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix}^\mathsf{T} (\Delta V_k^{(K)} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} + \left(\Delta v_k^{(K)}\right)^\mathsf{T} \mathbf{f} \quad (22)$$

using the definitions in (12) and (13). This leads to the following results.

Corollary 9 ($\mathcal{V}$ positive definite)**.** *$\mathcal{V}$ in (5) is positive definite for all values of $d \in \mathbb{N}$ if and only if there exists $\lambda_{ij} \geq 0$ for $i, j \in \mathcal{I}_N$ such that*

$$\begin{aligned}
0 &\prec V_N^{(N)} - \sum_{i,j \in \mathcal{I}_N} \lambda_{ij}\, M_{ij}^{(N)} \\
0 &< v_N^{(N)} - \sum_{i,j \in \mathcal{I}_N} \lambda_{ij}\, m_{ij}^{(N)}
\end{aligned}$$

where $M_{ij}^{(N)}$ and $m_{ij}^{(N)}$ defined in (10).

Proof. The result follows from applying Theorem 7 with $K = N$ to show that the quadratic function $\mathcal{V}$ in (21) is positive definite. ∎

Corollary 10 ($\Delta\mathcal{V}$ negative semidefinite)**.** *Consider $\mathcal{V}$ in (5) and define $\Delta\mathcal{V} := \mathcal{V}(\xi_{k+1}) - \rho^2\,\mathcal{V}(\xi_k)$. Then $\Delta\mathcal{V}$ is negative semidefinite for all values of $d \in \mathbb{N}$ if and only if there exists $\eta_{ij} \geq 0$ for $i, j \in \mathcal{I}_{N+1}$ such that*

$$\begin{aligned}
0 &\succeq \Delta V_N^{(N+1)} + \sum_{i,j \in \mathcal{I}_{N+1}} \eta_{ij}\, M_{ij}^{(N+1)} \\
0 &\geq \Delta v_N^{(N+1)} + \sum_{i,j \in \mathcal{I}_{N+1}} \eta_{ij}\, m_{ij}^{(N+1)}
\end{aligned}$$

where $\Delta V_N^{(N+1)}$ and $\Delta v_N^{(N+1)}$ are defined in (13).

Proof. The result follows from applying Theorem 7 with $K = N + 1$ to show that the quadratic function $\Delta\mathcal{V}$ in (22) is negative semidefinite. ∎

Theorem 5 then follows from combining the results in Corollaries 9 and 10. In particular, the inequalities in each corollary correspond to the constraints in the semidefinite program (ρ-SDP). If the problem is feasible, then $\mathcal{V}$ is positive definite and $\Delta\mathcal{V}$ is negative semidefinite at iteration N. Since this holds for any initial condition, we can apply the result for each $k \geq N$ to show that $\mathcal{V}$ is a valid Lyapunov function. On the other hand, if the problem is infeasible, then there exists no quadratic function of the form (5) such that $\mathcal{V}$ is positive definite and $\Delta\mathcal{V}$ is negative semidefinite, so no valid quadratic Lyapunov function with state ξ_k exists for the given rate ρ. This completes the proof of Thm. 5. ∎

6. Extensions

Our main result in Theorem 5 applies to methods of the form ($\mathcal{M}$) with fixed step-sizes applied to smooth strongly convex functions. Our framework, however, can be extended to many other scenarios, with or without tightness.

We now proceed with some examples of how our procedure of searching for Lyapunov functions can serve as a basis for the analysis of many *exotic* algorithms. We provide two such examples: (i) the analysis of variants of GM and HBM involving subspace searches, and (ii) the analysis of a fast gradient scheme with scheduled restarts.

6.1. Exact line searches

In this section, we search for quadratic Lyapunov functions when it is possible to perform an exact line search. We illustrate the procedure on steepest descent

$$\begin{aligned}
\alpha &= \arg\min_{\alpha} f(x_k - \alpha\,\nabla f(x_k)) \\
x_{k+1} &= x_k - \alpha\,\nabla f(x_k)
\end{aligned}$$

and on a variant of HBM:

$$\begin{aligned}
(\alpha, \beta) &= \arg\min_{\alpha,\beta} f(x_k + \beta\,(x_k - x_{k-1}) - \alpha\,\nabla f(x_k)) \\
x_{k+1} &= x_k + \beta\,(x_k - x_{k-1}) - \alpha\,\nabla f(x_k)
\end{aligned}$$

The detailed analyses can be found in the supplementary material, whereas the results are presented on Figure 3.

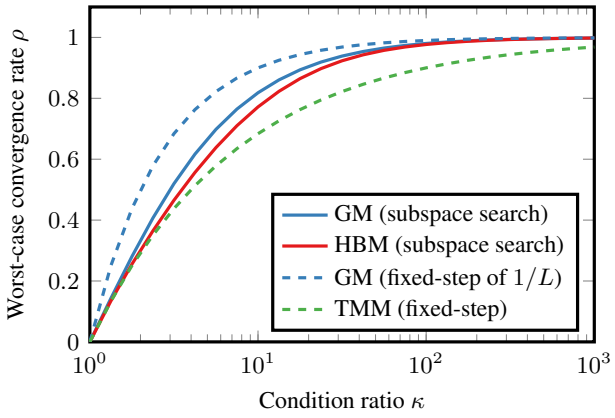

Figure 3. Convergence rates of GM and HBM with subspace searches. Note that the Gradient Method with exact line search matches the worst-case rate $\frac{\kappa-1}{\kappa+1}$ from (de Klerk et al., 2017). For comparison, the rates of the Gradient Method with step-size $1/L$ and the Triple Momentum Method are also shown.

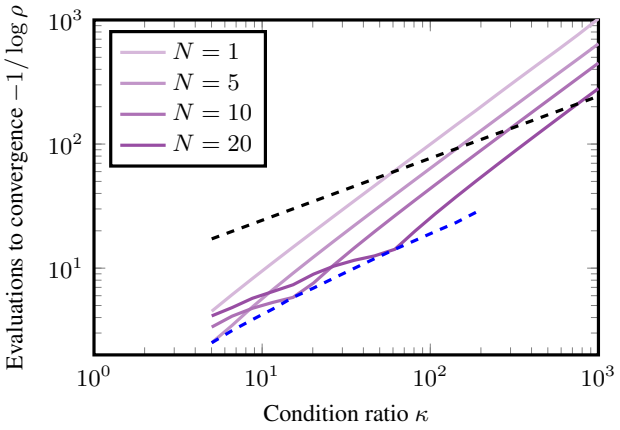

Figure 4. Worst-case number of gradient evaluations to convergence $\mathcal{O}(-1/\log\rho)$ for different restart schedules N (purple) along with the optimal restart schedule $N_\star = \arg\min_N \rho(N)$ (dashed blue). For comparison, we also plot the upper bound $\rho(N_\star) \leq \exp\left(\frac{-1}{e\sqrt{8\kappa}}\right)$ (dashed black) from (O'Donoghue & Candès, 2015). Results are not shown for small κ due to numerical limitations of the SDP solvers.

6.2. Scheduled restarts

In this section, we apply the methodology to estimate the convergence rate of FGM with scheduled restarts; motivations for this kind of techniques can be found in e.g., (O'Donoghue & Candès, 2015). We present numerical guarantees obtained when using a version of FGM tailored for smooth convex minimization, which is restarted every N iterations. This setting goes slightly beyond the fixed-step model presented in ($\mathcal{M}$), as the step-size rules depend on the iteration counter.

Define $\beta_0 := 1$ and $\beta_{i+1} := \frac{1+\sqrt{4\beta_i^2+1}}{2}$; we use the following iterative procedure

$$\begin{aligned} y_k^0, z_k^0 &\leftarrow y_{k-1}^N \\ z_k^{i+1} &= y_k^i - \frac{1}{L}\nabla f(y_k^i) \\ y_k^{i+1} &= z_k^{i+1} + \frac{\beta_i - 1}{\beta_{i+1}}(z_k^{i+1} - z_k^i) \end{aligned} \tag{23}$$

which does N steps of the standard fast gradient method (Nesterov, 1983) before restarting. We study the convergence of this scheme using quadratic Lyapunov functions with states $(y_k^N - y_\star, \nabla f(y_k^N), f(y_k^N) - f(y_\star))$. The derivations of the SDP for verifying $\mathcal{V}_{k+1} \leq \rho^{2N}\mathcal{V}_k$ (where ρ^N is the convergence rate of the inner loop) is similar to that of (ρ-SDP) (details in supplementary material). Numerical results are provided in Figure 4.

7. Conclusion

In this work, we studied first-order iterative fixed-step methods applied to smooth strongly convex functions. We presented a semidefinite formulation whose feasibility is both necessary and sufficient for the existence of a quadratic Lyapunov function.

This methodology unifies two previous approaches to worst-case analyses: performance estimation due to Drori & Teboulle (2014) and integral quadratic constraints due to Lessard et al. (2016). Moreover, this approach admits a large number of potential extensions, both in terms of classes of optimization problems and types of algorithms that can be analyzed. In particular, Lyapunov functions can be used to study sublinear convergence rates (see e.g., (Hu & Lessard, 2017)), switched systems (Lin & Antsaklis, 2009) (e.g., for adaptive methods) or noisy methods (see e.g., (Cyrus et al., 2018)).

Code

The code used to implement (ρ-SDP) and generate the figures in this paper is available at `https://github.com/QCGroup/quad-lyap-first-order`.

Acknowledgments

A. Taylor was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research

and innovation program (grant agreement 724063). Additionally, this material is based upon work supported by the National Science Foundation under Grants No. 1656951 and 1750162.

References

Boyd, S., El Ghaoui, L., Feron, E., and Balakrishnan, V. *Linear Matrix Inequalities in System and Control Theory*, volume 15 of *Studies in Applied Mathematics*. SIAM, Philadelphia, PA, June 1994.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, 2004.

Cyrus, Saman, Hu, Bin, Van Scoy, Bryan, and Lessard, Laurent. A robust accelerated optimization algorithm for strongly convex functions. In *Proceedings of the 2018 American Control Conference (to appear)*, 2018.

de Klerk, Etienne, Glineur, François, and Taylor, Adrien B. On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions. *Optimization Letters*, 11(7):1185–1199, Oct 2017.

Drori, Yoel and Teboulle, Marc. Performance of first-order methods for smooth convex minimization: A novel approach. *Mathematical Programming*, 145(1):451–482, 2014.

Hu, Bin and Lessard, Laurent. Dissipativity theory for Nesterov's accelerated method. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1549–1557, 2017.

Kalman, R. E. and Bertram, J. E. Control system analysis and design via the "second method" of Lyapunov: I–Continuous-time systems. *Journal of Basic Engineering*, 82(2):371–393, 1960a.

Kalman, R. E. and Bertram, J. E. Control system analysis and design via the "second method" of Lyapunov: II–Discrete-time systems. *Journal of Basic Engineering*, 82 (2):394–400, 1960b.

Lessard, Laurent, Recht, Benjamin, and Packard, Andrew. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

Lin, H. and Antsaklis, P. J. Stability and stabilizability of switched linear systems: A survey of recent results. *IEEE Transactions on Automatic Control*, 54(2):308–322, February 2009.

Lyapunov, A. M. and Fuller, A. T. *General Problem of the Stability Of Motion*. Control Theory and Applications Series. Taylor & Francis, 1992. Original text in Russian, 1892.

Megretski, A. and Rantzer, A. System analysis via integral quadratic constraints. *IEEE Transactions on Automatic Control*, 42(6):819–830, June 1997.

Megretski, A. and Treil, S. Power distribution inequalities in optimization and robustness of uncertain systems. *Journal of Mathematical Systems, Estimation, and Control*, 3 (3):301–319, 1993.

Nesterov, Yurii. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady AN USSR*, volume 269, pp. 543–547, 1983.

Nesterov, Yurii. *Introductory lectures on convex optimization: A basic course*. Springer, 2004.

O'Donoghue, Brendan and Candès, Emmanuel. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Polyak, B.T. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Taylor, Adrien B., Hendrickx, Julien, and Glineur, François. Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods. In *IEEE Conference on Decision and Control*, 2017a.

Taylor, Adrien B., Hendrickx, Julien M., and Glineur, François. Exact worst-case performance of first-order methods for composite convex optimization. *SIAM Journal on Optimization*, 27(3):1283–1313, 2017b.

Taylor, Adrien B., Hendrickx, Julien M., and Glineur, François. Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming*, 161(1):307–345, Jan 2017c.

Van Scoy, Bryan, Freeman, Randy A., and Lynch, Kevin M. The fastest known globally convergent first-order method for minimizing strongly convex functions. *IEEE Control Systems Letters*, 2(1):49–54, January 2018.

Vidyasagar, Mathukumalli. *Nonlinear systems analysis*, volume 42. SIAM, 2002.

A. Proof of Theorem 7 (sampled positive definite quadratics)

We begin by noting that σ is positive definite if and only if the optimal value of the following problem is positive for any value of $\varepsilon > 0$,

$$\inf_{\xi} \ \sigma(\xi) \quad \text{s.t.} \begin{cases} (\mathbf{x}, \mathbf{g}, \mathbf{f}) \text{ as in (16) generated by method } (\mathcal{M}) \\ \qquad \text{applied to } f \in \mathcal{F}_{\mu,L}, \\ \|\xi\| \geq \varepsilon. \end{cases}$$

Next, we discretize the problem by replacing $f \in \mathcal{F}_{\mu,L}$ with the equivalent condition that the discrete set of points $\{(y_i, g_i, f_i)\}_{i \in \mathcal{I}_K}$ is $\mathcal{F}_{\mu,L}$-interpolable (recall that $\mathcal{I}_K = \{0, \dots, K, \star\}$). Choosing a specific notion of distance for ξ, we can reformulate the previous statement as verifying that $p_\star^{(d)}(\varepsilon) > 0$ for all $\varepsilon > 0$ where

$$p_\star^{(d)}(\varepsilon) := \min_{\mathbf{x},\mathbf{g},\mathbf{f}} \ \sigma(\xi) \quad \text{s.t.} \begin{cases} \{(y_i, g_i, f_i)\}_{i \in \mathcal{I}_k} \text{ is } \mathcal{F}_{\mu,L}\text{-interpolable}, \\ (\mathbf{x}, \mathbf{g}, \mathbf{f}) \text{ as in (16) generated by } (\mathcal{M}), \\ \|\mathbf{x}\|^2 + \|\mathbf{g}\|^2 + \mathbf{1}^\mathsf{T}\mathbf{f} = \varepsilon. \end{cases}$$

Note that the last condition can equivalently be replaced by others (e.g., $\|\mathbf{x}\|^2 = \varepsilon$ or $\|\mathbf{g}\|^2 = \varepsilon$ or $\mathbf{1}^\mathsf{T}\mathbf{f} = \varepsilon$), and that the optimal value of this problem is attained (using a short homogeneity argument with respect to ε). Using the necessary and sufficient conditions for the set of points $\{(y_i, g_i, f_i)\}_{i \in \mathcal{I}_k}$ to be $\mathcal{F}_{\mu,L}$-interpolable from Theorem 6, we have

$$p_\star^{(d)}(\varepsilon) = \min_{\mathbf{x},\mathbf{g},\mathbf{f}} \ \sigma(\xi) \quad \text{s.t.} \begin{cases} \phi_{ij} \geq 0 \text{ for all } i, j \in \mathcal{I}_K, \\ (\mathbf{x}, \mathbf{g}, \mathbf{f}) \text{ as in (16) generated by } (\mathcal{M}), \\ \|\mathbf{x}\|^2 + \|\mathbf{g}\|^2 + \mathbf{1}^\mathsf{T}\mathbf{f} = \varepsilon. \end{cases}$$

Next, we define the Gram matrix $\mathbf{G} \in \mathbb{S}^{N+K+2}$ as $\mathbf{G} := \mathbf{B}^\mathsf{T}\mathbf{B}$ where

$$\mathbf{B} := \begin{bmatrix} x_{-N} - x_\star & \dots & x_0 - x_\star & g_0 & \dots & g_K \end{bmatrix}, \quad (24)$$

hence $\mathbf{G}$ is a standard Gram matrix containing all inner products between $x_i - x_\star$ for $i \in \{-N, \dots, 0\}$ and g_i for $i \in \{0, \dots, K\}$. Note that the quadratic σ can be written as a function of the Gram matrix as

$$\sigma(\mathbf{G}, \mathbf{f}) = \operatorname{tr}\big(Q\mathbf{G}\big) + q^\mathsf{T}\mathbf{f}.$$

Similarly, the interpolation conditions can also be reformulated with the Gram matrix as

$$0 \leq \phi_{ij}(\mathbf{G}, \mathbf{f}) = \operatorname{tr}\big(M_{ij}\mathbf{G}\big) + m_{ij}^\mathsf{T}\mathbf{f}$$

where M_{ij} and m_{ij} are such that

$$\phi_{ij} = \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix}^\mathsf{T} (M_{ij} \otimes I_d) \begin{bmatrix} \mathbf{x} \\ \mathbf{g} \end{bmatrix} + m_{ij}^\mathsf{T}\mathbf{f},$$

for all $i, j \in \mathcal{I}_K$ (hence also $\star$ is in the index set). Therefore, we can reformulate the previous problem as the following rank-constrained semidefinite program:

$$p_\star^{(d)}(\varepsilon) = \min_{\mathbf{G} \in \mathbb{S}^{N+K+2}, \mathbf{f} \in \mathbb{R}^{N+1}} \operatorname{tr}\big(Q\mathbf{G}\big) + q^\mathsf{T}\mathbf{f} \quad \text{s.t.} \begin{cases} 0 \leq \operatorname{tr}\big(M_{ij}\mathbf{G}\big) + m_{ij}^\mathsf{T}\mathbf{f} \quad \text{for } i, j \in \mathcal{I}, \\ \operatorname{tr}(\mathbf{G}) + \mathbf{1}^\mathsf{T}\mathbf{f} = \varepsilon, \\ \mathbf{f} \geq 0, \\ \mathbf{G} \succeq 0, \\ \operatorname{Rank}(\mathbf{G}) \leq d, \end{cases}$$

where we remind the reader that d is the dimension of the optimization problem $(\mathcal{P})$. Therefore, as discussed in (Taylor et al., 2017c;b), if we want a result that does not depend on the dimension (i.e, a σ that is positive definite whatever the value of d), we have to verify that $p^{(\infty)}(\varepsilon) > 0$ (which corresponds to assuming that $d \geq N+K+2$ since this is the dimension of $\mathbf{G}$). We then have the following semidefinite program:

$$p_\star^{(\infty)}(\varepsilon) = \min_{\mathbf{G} \in \mathbb{S}^{N+K+2}, \mathbf{f} \in \mathbb{R}^{N+1}} \operatorname{tr}\big(Q\mathbf{G}\big) + q^\mathsf{T}\mathbf{f} \quad \text{s.t.} \begin{cases} 0 \leq \operatorname{tr}\big(M_{ij}\mathbf{G}\big) + m_{ij}^\mathsf{T}\mathbf{f} \quad \text{for } i, j \in \mathcal{I}, \\ \operatorname{tr}(\mathbf{G}) + \mathbf{1}^\mathsf{T}\mathbf{f} = \varepsilon, \\ \mathbf{f} \geq 0, \\ \mathbf{G} \succeq 0. \end{cases}$$

A Slater point for this problem (i.e., a feasible point such that $\mathbf{G} \succ 0$) is obtained in the following section, so the optimal value of the primal problem is equal to the optimal value of the dual, which is given by

$$d^{(\infty)}(\varepsilon) := \max_{\{\lambda_{ij}\}, \nu} \ \nu\,\varepsilon \quad \text{s.t.} \begin{cases} \lambda_{ij} \geq 0 \text{ for all } i, j \in \mathcal{I}, \\ Q - \sum_{i,j \in \mathcal{I}} \lambda_{ij}\, M_{ij} \succeq \nu I_{N+K+2}, \\ q - \sum_{i,j \in \mathcal{I}} \lambda_{ij}\, m_{ij} \geq \nu \mathbf{1}_{N+1}. \end{cases}$$

The theorem is then proved by noting the equivalence

$$p^{(\infty)}(\varepsilon) > 0, \ \forall \varepsilon > 0 \iff d^{(\infty)}(\varepsilon) > 0, \ \forall \varepsilon > 0,$$

where the last statement amounts to verifying that

$$Q - \sum_{i,j \in \mathcal{I}} \lambda_{ij}\, M_{ij} \succ 0, \quad q - \sum_{i,j \in \mathcal{I}} \lambda_{ij}\, m_{ij} > 0. \qquad \blacksquare$$

B. Slater point for proof of Theorem 7

In this section, we show how to construct a Slater point (Boyd & Vandenberghe, 2004) for the primal semidefinite program in the proof of Theorem 7. The construction

is similar to Section 2.1.2 of (Nesterov, 2004) and the proof of Theorem 6 in (Taylor et al., 2017c).

Consider applying the first-order iterative fixed-step method ($\mathcal{M}$) with $\alpha \neq 0$ and $\gamma_0 \neq 0$ for K iterations to the function $f(x) = \frac{1}{2}x^T H x$ where $H \in \mathbb{S}^d$ with $d \geq N + K + 2$ is the positive definite tridiagonal matrix defined by

$$[H]_{ij} = \begin{cases} 2 & \text{if } i = j \\ 1 & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

which has maximum eigenvalue $L = 2 + 2\cos\big(\pi/(d+1)\big)$. Define the matrix $\mathbf{B}$ in (24). Using the initial condition $x_i = e_{N+1+i}$ for $i = -N, \ldots, 0$, we will show that $\mathbf{B}$ is upper triangular with nonzero diagonal elements, and hence full rank.

Since $\gamma_0 \neq 0$, y_0 has a nonzero element corresponding to e_{N+1}. Then $g_0 = Hy_0$ has a nonzero element corresponding to e_{N+2} due to the tridiagonal structure of H. Furthermore, y_0 may only have nonzero elements corresponding to e_j for $j = 1, \ldots, N+1$, so g_0 must have zero components corresponding to e_j for all $j > N+2$.

We now continue by induction. Assume that g_k has a nonzero element corresponding to e_{N+2+k} and zero elements corresponding to e_i for all $i > N+2+k$. Since $\alpha \neq 0$, x_{k+1} has a nonzero element corresponding to e_{N+2+k} and zero elements corresponding to e_i for all $i > N+2+k$. Then since $\gamma_0 \neq 0$, y_{k+1} also has the same structure. Due to the tridiagonal structure of H, g_{k+1} then has a nonzero element corresponding to e_{N+3+k} and zero elements corresponding to e_i for all $i > N+3+k$. Therefore, by induction we have shown that for all $k \geq 0$, g_k has a nonzero element corresponding to e_{N+2+k} and zero elements corresponding to e_i for all $i > N+2+k$. For P to be upper triangular, we need g_K to have dimension at least $N+2+K$. Thus, if $d \geq N+2+K$ where K is the number of iterations, then $\mathbf{B}$ is upper triangular with nonzero entries on the diagonal, and therefore has full rank.

In order to make the statement hold for general $\mu < L$, observe that the tridiagonal structure of H is preserved under the operation

$$\tilde{H} = \big(H - \lambda_{\min}(H)\, I\big)\frac{L-\mu}{\lambda_{\max}(H) - \lambda_{\min}(H)} + \mu I$$

where $\mu I \preceq \tilde{H} \preceq LI$.

Since $\mathbf{B}$ has full rank, the Gram matrix $\mathbf{G} = \mathbf{B}^T\mathbf{B} \succ 0$ is positive definite. Therefore, the primal semidefinite program satisfies Slater's condition.

C. Steepest descent

In this section, we show a similar formulation as (ρ-SDP) for steepest descent. In this case, the analysis was not *a priori* guaranteed to be tight, due to the line search conditions. In order to encode the line search, we use the corresponding optimality conditions, as in (de Klerk et al., 2017):

$$\begin{aligned} \langle x_{k+1} - x_k,\ g_{k+1}\rangle &= 0, \\ \langle g_k,\ g_{k+1}\rangle &= 0, \end{aligned} \tag{25}$$

with $g_k = \nabla f(x_k)$. For the Lyapunov function structure, we choose the following

$$V(\xi_k) = \begin{bmatrix} x_k - x_\star \\ g_k \end{bmatrix} (P \otimes I_d) \begin{bmatrix} x_k - x_\star \\ g_k \end{bmatrix}^T + p(f_k - f_\star).$$

In order to develop a SDP formulation for this problem we follow the same steps as for (ρ-SDP), starting with **Step 1** (see Section 4.2): we define the following row vectors in $\mathbb{R}^2$

$$\bar{y}_0^{(0)} := \mathbf{e}_1^T,\ \bar{x}_0^{(0)} := \mathbf{e}_1^T,\ \bar{g}_0^{(0)} := \mathbf{e}_2^T,$$

and $\bar{y}_\star^{(0)} = \bar{x}_\star^{(0)} = \bar{g}_\star^{(0)} := \mathbf{0}^T$, along with the scalars $\bar{f}_0^{(0)} := 1$ and $\bar{f}_\star^{(0)} := 0$. In addition, we use the following vectors in $\mathbb{R}^4$

$$\begin{aligned} \bar{x}_0^{(1)} &:= \mathbf{e}_1^T,\ \bar{x}_1^{(1)} := \mathbf{e}_2^T, \\ \bar{y}_0^{(1)} &:= \mathbf{e}_2^T,\ \bar{y}_1^{(1)} := \mathbf{e}_2^T, \\ \bar{g}_0^{(1)} &= \mathbf{e}_3^T,\ \bar{g}_1^{(1)} := \mathbf{e}_4^T, \end{aligned}$$

and $\bar{y}_\star^{(1)} = \bar{x}_\star^{(1)} = \bar{g}_\star^{(1)} := \mathbf{0}^T$, along with $\bar{f}_0^{(1)}, \bar{f}_1^{(1)}, \bar{f}_\star^{(1)} \in \mathbb{R}^2$ such that $\bar{f}_0^{(1)} := \mathbf{e}_1^T$, $\bar{f}_1^{(1)} := \mathbf{e}_2^T$ and $\bar{f}_\star^{(1)} := \mathbf{0}^T$.

Because of the algorithm, **Step 2** is slightly different as before; we encode the line search constraints (25) using

$$A_1 = \begin{bmatrix} \bar{x}_0^{(1)} \\ \bar{x}_1^{(1)} \\ \bar{g}_1^{(1)} \end{bmatrix}^T \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_0^{(1)} \\ \bar{x}_1^{(1)} \\ \bar{g}_1^{(1)} \end{bmatrix},$$

$$A_2 = \begin{bmatrix} \bar{g}_0^{(1)} \\ \bar{g}_1^{(1)} \end{bmatrix}^T \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{g}_0^{(1)} \\ \bar{g}_1^{(1)} \end{bmatrix}.$$

The subsequent steps (**Step 3** and **Step 4**) are exactly the same as in Section 4.2. We finally obtain a slightly modified

version of the feasibility problem (ρ-SDP):

$$\begin{array}{ll}\text{feasible} & 0 \prec V_0^{(0)} - \sum_{i,j\in\mathcal{I}_0} \lambda_{ij}\, M_{ij}^{(0)} \\ \substack{P\in\mathbb{S}^2\\ p\in\mathbb{R}^1\\ \nu_1,\nu_2\in\mathbb{R}\\ \{\lambda_{ij}\}\\ \{\eta_{ij}\}} & 0 < v_0^{(0)} - \sum_{i,j\in\mathcal{I}_0} \lambda_{ij}\, m_{ij}^{(0)} \\ & 0 \succeq \Delta V_0^{(1)} + \sum_{i,j\in\mathcal{I}_1} \eta_{ij}\, M_{ij}^{(1)} + \sum_{i=1}^{2} \nu_i\, A_i \\ & 0 \geq \Delta v_0^{(1)} + \sum_{i,j\in\mathcal{I}_1} \eta_{ij}\, m_{ij}^{(1)} \\ & 0 \leq \lambda_{ij} \quad \text{for } i,j\in\mathcal{I}_0 \\ & 0 \leq \eta_{ij} \quad \text{for } i,j\in\mathcal{I}_1 \end{array}$$

with $\mathcal{I}_0 := \{0,\star\}$ and $\mathcal{I}_1 := \{0,1,\star\}$.

D. SDP for HBM with subspace searches

We follow the steps of the previous section for steepest descent; we only make the following adaptations: (i) we look for a quadratic Lyapunov function with the states

$$V(\xi_k) = \begin{bmatrix} x_k - x_\star \\ x_{k-1} - x_\star \\ g_k \\ g_{k-1} \end{bmatrix} (P \otimes I_d) \begin{bmatrix} x_k - x_\star \\ x_{k-1} - x_\star \\ g_k \\ g_{k-1} \end{bmatrix}^\mathsf{T} + p^\mathsf{T} \begin{bmatrix} f_k - f_\star \\ f_{k-1} - f_\star \end{bmatrix},$$

(ii) we adapt the initialization (**Step 1**), (iii) adapt the line search conditions (**Step 2**) and (iv) obtain a slightly modified version of the SDP.

For (ii), we adapt the initialization procedure (**Step 1**) as follows. We define the following row vectors of $\mathbb{R}^4$:

$$\begin{aligned} \bar{x}_0^{(1)} &:= \mathbf{e}_1^\mathsf{T},\ \bar{x}_1^{(1)} := \mathbf{e}_2^\mathsf{T}, \\ \bar{y}_0^{(1)} &:= \mathbf{e}_1^\mathsf{T},\ \bar{y}_1^{(1)} := \mathbf{e}_2^\mathsf{T}, \\ \bar{g}_0^{(1)} &:= \mathbf{e}_3^\mathsf{T},\ \bar{g}_1^{(1)} := \mathbf{e}_4^\mathsf{T}, \end{aligned}$$

along with $\bar{y}_\star^{(1)} = \bar{x}_\star^{(1)} = \bar{g}_\star^{(1)} := \mathbf{0}^\mathsf{T}$, and the following in $\mathbb{R}^2$: $\bar{f}_0^{(1)} := \mathbf{e}_1^\mathsf{T}$, $\bar{f}_1^{(1)} := \mathbf{e}_2^\mathsf{T}$ and $\bar{f}_\star^{(1)} := \mathbf{0}^\mathsf{T}$. We also define the following row vectors in $\mathbb{R}^7$:

$$\begin{aligned} \bar{x}_{-1}^{(2)} :=\ \mathbf{e}_1^\mathsf{T},\ \bar{x}_0^{(2)} &:= \mathbf{e}_2^\mathsf{T},\ \bar{x}_1^{(2)} := \mathbf{e}_3^\mathsf{T},\ \bar{x}_2^{(2)} := \mathbf{e}_4^\mathsf{T}, \\ \bar{y}_0^{(2)} &:= \bar{x}_0^{(2)},\ \bar{y}_1^{(2)} := \bar{x}_1^{(2)},\ \bar{y}_2^{(2)} := \bar{x}_2^{(2)}, \\ \bar{g}_0^{(2)} &:= \mathbf{e}_5^\mathsf{T},\ \bar{g}_1^{(2)} := \mathbf{e}_6^\mathsf{T},\ \bar{g}_2^{(2)} := \mathbf{e}_7^\mathsf{T}, \end{aligned}$$

along with $y_\star^{(2)} = \bar{x}_\star^{(2)} = \bar{g}_\star^{(2)} := \mathbf{0}^\mathsf{T}$ and the vectors of $\mathbb{R}^3$: $\bar{f}_0^{(2)} := \mathbf{e}_1^\mathsf{T}$, $\bar{f}_1^{(2)} := \mathbf{e}_2^\mathsf{T}$, $\bar{f}_2^{(2)} := \mathbf{e}_3^\mathsf{T}$ and $\bar{f}_\star^{(2)} := \mathbf{0}^\mathsf{T}$.

Now for (iii) (or **Step 2**), optimality of the search conditions can be

$$\begin{aligned} \langle x_{k+1} - x_k,\ g_{k+1}\rangle &= 0, \\ \langle x_k - x_{k-1},\ g_{k+1}\rangle &= 0, \\ \langle g_k;\ g_{k+1}\rangle &= 0, \end{aligned}$$

which we can formulate in matrix form for $k \in \{0,1\}$:

$$\begin{aligned} A_{1+k} &= \begin{bmatrix} \bar{x}_k^{(2)} \\ \bar{x}_{k+1}^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_k^{(2)} \\ \bar{x}_{k+1}^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix} \\ A_{3+k} &= \begin{bmatrix} \bar{x}_{k-1}^{(2)} \\ \bar{x}_k^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{x}_{k-1}^{(2)} \\ \bar{x}_k^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix} \\ A_{5+k} &= \begin{bmatrix} \bar{g}_k^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix}^\mathsf{T} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \bar{g}_k^{(2)} \\ \bar{g}_{k+1}^{(2)} \end{bmatrix} \end{aligned}$$

The subsequent steps (**Step 3** and **Step 4**) are exactly the same as in Section 4.2. We finally obtain a slightly modified version of the feasibility problem (ρ-SDP):

$$\begin{array}{ll}\text{feasible} & 0 \prec V_1^{(1)} - \sum_{i,j\in\mathcal{I}_1} \lambda_{ij}\, M_{ij}^{(1)} \\ \substack{P\in\mathbb{S}^2\\ p\in\mathbb{R}^1\\ \nu_1,\ldots,\nu_6\in\mathbb{R}\\ \{\lambda_{ij}\}\\ \{\eta_{ij}\}} & 0 < v_1^{(1)} - \sum_{i,j\in\mathcal{I}_1} \lambda_{ij}\, m_{ij}^{(1)} \\ & 0 \succeq \Delta V_1^{(2)} + \sum_{i,j\in\mathcal{I}_2} \eta_{ij}\, M_{ij}^{(2)} + \sum_{i=1}^{6} \nu_i\, A_i \\ & 0 \geq \Delta v_1^{(2)} + \sum_{i,j\in\mathcal{I}_2} \eta_{ij}\, m_{ij}^{(2)} \\ & 0 \leq \lambda_{ij} \quad \text{for } i,j\in\mathcal{I}_1 \\ & 0 \leq \eta_{ij} \quad \text{for } i,j\in\mathcal{I}_2 \end{array}$$

with $\mathcal{I}_1 := \{0,1,\star\}$ and $\mathcal{I}_2 := \{0,1,2,\star\}$. The corresponding results are presented on Figure 3.

E. SDP for FGM with scheduled restarts

This setting goes slightly beyond the fixed-step model presented in ($\mathcal{M}$), as the step sizes depend on the iteration. We study the algorithm described by (23), which does N steps of the standard fast gradient method (Nesterov, 1983) before restarting. We study the convergence of this scheme using quadratic Lyapunov functions of the form

$$\begin{bmatrix} y_k^N - x_\star \\ \nabla f(y_k^N) \end{bmatrix}^\mathsf{T} (P \otimes I_d) \begin{bmatrix} y_k^N - x_\star \\ \nabla f(y_k^N) \end{bmatrix} + p\,[f(y_k^N) - f(x_\star)]. \tag{26}$$

Let us perform similar steps as for (ρ-SDP) for constructing the corresponding SDP. We start with the initialization

procedure **Step 1**. Let us define the following row vectors in $\mathbb{R}^2$:

$$\bar{y}_0^{(1)} = \bar{x}_0^{(1)} := \mathbf{e}_1^\mathsf{T},\ \bar{g}_0^{(1)} := \mathbf{e}_2^\mathsf{T},$$

and $\bar{x}_\star^{(1)} = \bar{g}_\star^{(1)} := \mathbf{0}^\mathsf{T}$, along with the scalars $f_0^{(1)} := 1$ and $f_\star^{(1)} := 0$. In addition, we define the row vectors of $\mathbb{R}^{N+2}$:

$$\bar{x}_0^{(N+1)} := \mathbf{e}_1^\mathsf{T},\ \bar{g}_k^{(N+1)} := \mathbf{e}_{2+k}^\mathsf{T},$$

for $k = 0, \dots, N$, along with $\bar{x}_\star^{(N+1)} = \bar{g}_\star^{(N+1)} := \mathbf{0}^\mathsf{T}$ and the row vectors $\bar{f}_k^{(N+1)} \in \mathbb{R}^{N+1}$ defined as $\bar{f}_k^{(N+1)} := \mathbf{e}_{1+k}^\mathsf{T}$ and $\bar{f}_\star^{(N+1)} := \mathbf{0}^\mathsf{T}$.

Step 2 Apply one complete loop of the algorithm as follows: for $k = 0, \dots, N-1$ define the sequence of row vectors:

$$\begin{aligned}
\bar{z}_{k+1}^{(N+1)} &= \bar{y}_k^{(N+1)} - \frac{1}{L}\bar{g}_k^{(N+1)},\\
\bar{y}_{k+1}^{(N+1)} &= \bar{z}_k^{(N+1)} + \frac{\beta_k - 1}{\beta_{k+1}}(\bar{z}_k^{(N+1)} - \bar{z}_k^{(N+1)}),
\end{aligned}$$

with $\beta_0 := 1$ and $\beta_{k+1} := \frac{1+\sqrt{4\beta_k^2+1}}{2}$. For complying with the notations of the paper, we define the sequence

$$x_k^{(K)} := y_k^{(K)}$$

for $k = 0, \dots, N$ and $K \in \{1, N+1\}$. Then, using the sets $\mathcal{I}_1 := \{0, \star\}$ and $\mathcal{I}_{N+1} := \{0, \dots, N, \star\}$, the other stages follow from the same lines as **Step 3**, and **Step 4** with the slight modification of the expression for the rate

$$\begin{aligned}
\Delta v_k^{(N+1)} &:= v_{k+1}^{(N+1)} - \rho^{2N}\, v_k^{(N+1)},\\
\Delta V_k^{(N+1)} &:= V_{k+1}^{(N+1)} - \rho^{2N}\, V_k^{(N+1)},
\end{aligned}$$

and **Step 5** follows as in Section 4.2:

$$\begin{array}{ll}
\underset{\substack{P\in\mathbb{S}^{2(N+1)}\\ p\in\mathbb{R}^{N+1}\\ \{\lambda_{ij}\}\\ \{\eta_{ij}\}}}{\text{feasible}} & 0 \prec V_1^{(1)} - \sum_{i,j\in\mathcal{I}_1} \lambda_{ij}\, M_{ij}^{(1)}\\
& 0 < v_1^{(1)} - \sum_{i,j\in\mathcal{I}_1} \lambda_{ij}\, m_{ij}^{(1)}\\
& 0 \succeq \Delta V_N^{(N+1)} + \sum_{i,j\in\mathcal{I}_{N+1}} \eta_{ij}\, M_{ij}^{(N+1)}\\
& 0 \geq \Delta v_N^{(N+1)} + \sum_{i,j\in\mathcal{I}_{N+1}} \eta_{ij}\, m_{ij}^{(N+1)}\\
& 0 \leq \lambda_{ij} \quad \text{for } i,j \in \mathcal{I}_1\\
& 0 \leq \eta_{ij} \quad \text{for } i,j \in \mathcal{I}_{N+1}
\end{array}$$

(note that the sets $\mathcal{I}_1$ and $\mathcal{I}_{N+1}$ should use the definitions of this section). Numerical results are available in Figure 4.