SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Analysis of first order optimization methods

av

Fredrik Krypta

2019 - No K20

Analysis of first order optimization methods

Fredrik Krypta

Självständigt arbete i matematik 15 högskolepoäng, grundnivå

Handledare: Yishao Zhou

2019

Analysis of first order optimization methods

Fredrik Kypta

June 2019

Abstract

In this thesis we deal with optimization algorithms that are commonly used in machine learning, the Gradient descent method and variants of it. These are called first order algorithms because they depend on first order derivative information. We do not work on the computational aspects, but rather do mathematical and structural analysis of the algorithms.

Acknowledgements

I would like to thank my supervisor Yishao Zhou for her encouragement, guidance and all the interesting discussions we have had. I also would like to thank Martin Tamm for proof-reading this thesis and giving lots of constructive feedback.

1 Introduction

The content of this text is essentially split into three parts. In the first part we look at the basic form of the Gradient descent algorithm, and prove its convergence. The Algorithms and Convergence section in the preliminaries contains definitions and theorems related to the convergence proof of the basic Gradient descent algorithm. In the second part we switch to the framework of control theory and dynamical systems to prove upper bounds of convergence rates in some special cases, in particular the objective function to minimize will be strongly convex and smooth. In the last part we take a look at the concept of Lyapunov stability.

2 Preliminaries

2.1 Algorithms and Convergence

2.1.1 Algorithmic Maps

Consider the problem of minimizing $f(x)$ subject to S, where f is the objective function and S is the feasible region. An algorithm for solving this problem is an iterative process that generates a sequence of points according to a set of instructions, including a termination criterion.

Given a vector x_k and applying the instructions, we get a new point x_{k+1}. This process can be described by an *algorithmic map* A. This map is generally a point-to-set map and assigns to each point in the domain X a subset of X. So given an initial point x_1, the algorithmic map generates the sequence $x_1, x_2, \ldots$, where $x_{k+1} \in A(x_k)$ for each k.

2.1.2 Closedness of the Line Search Algorithmic Map

Consider the line search problem to minimize $\theta(\lambda)$ subject to $\lambda \in L$, where $\theta(\lambda) = f(x + \lambda d)$ and L is a closed interval in $\mathbb{R}$ and $d \in \mathbb{R}^n$. This line search problem can be described by the algorithmic map $M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$, defined by $\mathbf{M}(x, d) = \{y : y = x + \overline{\lambda}d$ for some $\overline{\lambda} \in L$ and $f(y) \leq f(x + \lambda d)$ for each $\lambda \in L\}$. Note that there might be more than one minimizing point y. The following theorem shows that the map M is closed.

2.1.3 Theorem

Let $f : \mathbb{R}^n \to \mathbb{R}$, and let L be a closed interval in $\mathbb{R}$. Consider the line search map $M : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ defined by

$M(x, d) = \{y : y = x + \overline{\lambda}d$ for some $\overline{\lambda} \in L$ and $f(\mathrm{y}) \leq f(x + \lambda d)$ for each $\lambda \in L\}$.

If f is continuous at x and $d \neq 0$, then M is closed at (x, d).

2.1.4 Definition

Let X, Y, and Z be nonempty sets in $\mathbb{R}^n$, $\mathbb{R}^p$, and $\mathbb{R}^q$, respectively. Let $B : X \to Y$ and $C : Y \to Z$ be point-to-set maps. The *composite map* $A = CB$ is defined as the point-to-set map $A : X \to Z$ with

$$A(x) = \cup\{C(y) : y \in B(x)\}.$$

2.1.5 Theorem

Let X, Y, and Z be nonempty sets in $\mathbb{R}^n$, $\mathbb{R}^p$, and $\mathbb{R}^q$, respectively. Let $B : X \to Y$ and $C : Y \to Z$ be point-to-set maps. Consider the composite map $A = CB$. Suppose that B is closed at x and that C is closed on $B(x)$. Furthermore, suppose that if $x_k \to x$ and $y_k \in B(x_k)$, then there is a convergent subsequence of $\{y_k\}$. Then A is closed at x.

2.1.6 Corollary

Let X, Y, and Z be nonempty sets in $\mathbb{R}^n$, $\mathbb{R}^p$, and $\mathbb{R}^q$, respectively. Let $B : X \to Y$ be a function, and let $C : Y \to Z$ be a point-to-set map. If B is continuous at x, and C is closed on $B(x)$, then $A = CB$ is closed at x.

Note that without the assumption that a convergent subsequence $\{y_k\}$ exists in the theorem, even if the maps B and C are closed, the composite map $A = CB$ is not necessarily closed.

2.1.7 Theorem

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\overline{x}$. If there is a vector d such that $\nabla f(\overline{x})^\top d < 0$, there exists a $\delta > 0$ such that $f(\overline{x} + \lambda d) < f(\overline{x})$ for each $\lambda \in (0, \delta)$, so that d is a *descent direction* of f at $\overline{x}$.

2.1.8 Zangwill's Convergence Theorem

Let X be a nonempty closed set in $\mathbb{R}^n$, and let the nonempty set $\Omega \subseteq X$ be the solution set. Let $A : X \to X$ be a point-to-set map. Given $x_1 \in X$, the sequence $\{x_k\}$ is generated iteratively as follows: If $x_k \in \Omega$, then stop; otherwise, let $x_{k+1} \in A(x_k)$, replace k by $k + 1$, and repeat.

Suppose that the sequence $x_1, x_2, \ldots$ produced by the algorithm is contained in a compact subset of X, and suppose that there exists a continuous function ϕ, called the *descent function*, such that $\phi(y) < \phi(x)$ if $x \notin \Omega$ and $y \in A(x)$. If the map A is closed over the complement of Ω, then either the algorithm stops in a finite number of steps with a point in Ω or it generates an infinite sequence $\{x_k\}$ such that:

1. Every convergent subsequence of $\{x_k\}$ has a limit in Ω; that is, all accumulation points $\{x_k\}$ belong to Ω.
2. $\phi(x_k) \to \phi(x)$ for some $x \in \Omega$.

2.2 Linear algebra

We will use two notations for the scalar product, that is both $\langle v_1, v_2 \rangle$ and $v_1^\top v_2$ mean the scalar product of the vectors v_1 and v_2. Also $\|v\|$ denotes the Euclidean norm of the vector v.

Definition. [Spectral radius] Let $\lambda_1, ..., \lambda_n$ be the (real or complex) eigenvalues of an $n \times n$ matrix A. Then its spectral radius $\rho(A)$ is defined as:

$$\rho(A) = \max_{1 \le j \le n} |\lambda_j|.$$

Definition. [Positive semi-definite symmetric real matrices]
A $n \times n$ symmetric real matrix A is said to be positive semi-definite if $x^\top A x \ge 0$ for all non-zero x in $\mathbb{R}^n$.

The notation

$$A \succeq B$$

means that $A - B$ is a positive semi-definite matrix.

2.3 Convex functions

Definition. A set in S in $\mathbb{R}^n$ is said to be *convex* if the line segment joining any points of the set also belongs to the set. I.e. if x_1 and x_2 are in S then $tx_1 + (1-t)x_2$ must also belong to S for each $t \in [0,1]$.

Definition. A function f is convex if

$$f(tx + (1-t)y) \leq f(x) + (1-t)f(y) \quad \forall x, y \in \mathbb{R}^n, t \in [0,1].$$

Note that to check this condition we need three points and thus we refer this as a three-point criterion of convexity. In general it is difficult to use this criterion. However, the situation will be improved if we assume smoothness on the function f.

If f is differentiable and convex then every tangent line to the graph of f bounds the function values from below, that is

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x), \quad \forall x, y \in \mathbb{R}^n$$

which can be obtained by first dividing by t in the definition and rearranging

$$\frac{f(y + t(x-y)) - f(y)}{t} \leq f(x) - f(y)$$

and then taking the limit $t \to 0$. We call this the two-point criterion of convexity. For convenience we will use the scalar product form

$$f(y) \geq f(x) + \langle \nabla f(x), (y - x) \rangle \quad \forall x, y \in \mathbb{R}^n.$$

If f is twice differentiable, then taking a directional derivative in the v direction on the point x in the two-point criterion gives

$$0 \geq \langle \nabla f(x), v \rangle + \langle \nabla^2 f(x) v, y - x \rangle - \langle \nabla f(x), v \rangle = \langle \nabla^2 f(x) v, y - x \rangle \quad \forall x, y, v \in \mathbb{R}^n$$

which is equivalent to saying that the Hessian is positive semi-definite

$$\nabla^2 f(x) \succeq 0, \quad \forall x \in \mathbb{R}^n.$$

We call this one-point criterion of convexity.

Definition. [Quasiconvex function] Let $f\colon\ S \to \mathbb{R}$, where S is a nonempty convex set in $\mathbb{R}^n$. The function f is said to be *quasiconvex* if for each x_1 and $x_2 \in S$, the following inequality is true:

$$f(tx_1 + (1-t)x_2) \leq \max\{f(x_1), f(x_2)\} \quad \text{for each} \ \ t \in (0,1).$$

Definition. [Pseudoconvex function] Let S be a nonempty open set in $\mathbb{R}^n$, and let $f : S \to \mathbb{R}$ be differentiable on S. The function f is said to be *pseudoconvex* if for each $x_1, x_2 \in S$ with $\nabla f(x_1)^\top (x_2 - x_1) \geq 0$, we have $f(x_2) \geq f(x_1)$, or equivalently, if for each distinct $x_1, x_2 \in S$, $f(x_1) \leq f(x_2)$ implies that $\nabla f(x_1)^\top (x_2 - x_1) < 0$.

Lemma. [Maximum of convex functions] Assume that $\{f_\lambda\}_{\lambda \in \Lambda}$ are convex functions, where Λ is an index set. Then $f = \max\{f_1, ..., f_n\}$ is convex.

3 Gradient descent

Here we partly follow the exposition from [1]. Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. We are going to look at the optimization problem of minimizing $f(x)$, and in particular how it can be done with the method of gradient descent. Note that for a point x^* to be optimal, a necessary and sufficient condition is that $\nabla f(x^*) = \mathbf{0}$ (this is sufficient when f is convex).

A vector d is called a direction of descent of the function f at x if there exists $\delta > 0$ such that $f(x + \alpha d) < f(x)$ for all $\alpha \in (0, \delta)$. If

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} < 0,$$

then d is a direction of descent. The idea of the method is to move along the direction d (with $\|d\| = 1$) which minimizes the above limit, i.e. move in the direction of steepest descent. The following lemma hints at the reason why the method is called *gradient descent*.

3.0.1 Lemma.

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at x and suppose $\nabla f(x) \neq \mathbf{0}$. Then the optimal solution to the problem to minimize $f'(x; d)$ subject to $\|d\| \leq 1$ is given by $\overline{d} = -\nabla f(x)/\|\nabla f(x)\|$.

Proof.

From the differentiability of f at x we have that

$$f'(x; d) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha d) - f(x)}{\alpha} = \nabla f(x)^\top d.$$

By the Cauchy-Schwarz inequality, with $\|d\| \leq 1$, we have

$$\nabla f(x)^\top d \geq -\|\nabla f(x)\| \cdot \|d\| \geq -\|\nabla f(x)\|,$$

so the rate of change in a direction cannot be smaller than $-\|\nabla f(x)\|$.
The equalities above hold if and only if $d = -\nabla f(x)/\|\nabla f(x)\|$ and that is the optimal solution.

With the previous lemma in mind we will describe a natural algorithm for solving the problem of minimizing $f(x)$. An algorithm in this context is an iterative process that generates a sequence of points according to a set of instructions, together with a criterion for terminating the process. We now have an idea of which direction to step in for each iteration, but how large should a step be? When designing an algorithm it is possible to choose a constant stepsize. We will deal with that case in much more detail in section 4.

3.1 Line search

An algorithm for minimizing f might proceed as follows: Given a point x_k, find a directon vector d_k and a suitable step size α_k. Take a step to the new point $x_{k+1} = x_k + \alpha_k d_k$. Repeat this process until a termination criterion is fulfilled. To find the step size α_k we solve the subproblem of minimizing $f(x_k + \alpha d_k)$, this is a one-dimensional search problem in the variable α. Let $\theta(\alpha) = f(x + \alpha d)$. One option is to minimize θ exactly, this is called *exact line search*. An exact line search is used when the cost of the minimization problem is low compared to the cost of computing the search direction itself. In some special cases the minimizer along the line can be found analytically, and in others it can be computed efficiently. Many line searches in practice are *inexact*, the step length is chosen to approximately minimize f along the line, or even to just reduce f "enough" [8]. It is possible to do line search without derivates, but we will look at an example of a method that needs (first order) derivative information.

3.2 Bisection search method

Suppose we want to minimize θ over a closed and bounded interval. Furthermore suppose θ is pseudoconvex and differentiable. Let $[a_k, b_k]$ be the interval of uncertainty at iteration k. Suppose the derivative $\theta'(\alpha_k)$ is known, then there are three possibilities:

1. $\theta'(\alpha_k) = 0$, then by the pseudoconvexity of θ, α_k is the minimum.
2. If $\theta'(\alpha_k) > 0$, then for $\alpha > \alpha_k$ we have $\theta'(\alpha_k)(\alpha - \alpha_k) > 0$, and by the pseudoconvexity of θ it follows that $\theta(\alpha) > \theta(\alpha_k)$. So the minimum has to occur on the left of α_k. The new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[a_k, \alpha_k]$.
3. If $\theta'(\alpha_k) < 0$, then for $\alpha < \alpha_k$, $\theta'(\alpha_k)(\alpha - \alpha_k) > 0$, so that $\theta(\alpha) > \theta(\alpha_k)$. In this case the minimum occurs to right of α_k, and the new interval of uncertainty $[a_{k+1}, b_{k+1}]$ is given by $[\alpha_k, b_k]$.

We want to place α_k in the interval $[a_k, b_k]$ so that the maximum possible length of the new interval of uncertainty is minimized. In other words α_k must be chosen to minimize the maximum of $\alpha_k - a_k$ and $b_k - \alpha_k$. Clearly the optimal position of α_k is the midpoint $(1/2)(a_k + b_k)$.

Observe that the length of the interval of uncertainty after k iterations is $(1/2)^k(b_1 - a_1)$, i.e. the interval size gets bisected every iteration so the method will converge to a minimum within a desired degree of accuracy.

3.3 Gradient Descent Algorithm

Initialization Let $\epsilon > 0$ be the termination scalar. Choose a starting point x_1 (guess), let $k = 1$ and go to the Main Step.

Main Step If $\|\nabla f(x_k)\| < \epsilon$, stop; otherwise, let $d_k = -\nabla f(x_k)$, and let α_k be an optimal solution to the problem to minimize $f(x_k + \alpha d_k)$ subject to $\alpha \geq 0$. Let $x_{k+1} = x_k + \alpha_k d_k$, replace k by $k + 1$, and repeat Main Step.

3.4 Convergence of the Gradient Descent Method

Let $\Omega = \{\overline{x} : \nabla f(\overline{x}) = 0\}$, and let f be the descent function. The algorithmic map is $A = MD$, where $D(x) = [x, \nabla f(x)]$ and M is the line search map over the closed interval $[0, \infty)$. Under the assumption that f is continuously differentiable, D is continuous. Furthermore, M is closed by Theorem 2.3. Therefore, the algorithmic map A is closed by the Corollary to Theorem 2.5.1. Finally, if $x \notin \Omega$, then $\nabla f(x)^\top d < 0$, where $d = -\nabla f(x)$. By Theorem 2.6, d is a descent direction, so $f(y) < f(x)$ for $y \in A(x)$. Assuming that the sequence generated by the algorithm is contained in a compact set, then by Theorem 2.7, the gradient descent algorithm converges to a point with zero gradient.

3.5 Zig-Zagging of the Gradient Descent Method and an Example

This instructional example is taken from [6]. Consider $f(x, y) = \frac{1}{2}(x^2 + by^2)$ with $0 < b \leq 1$. The gradient ∇f has two components $\partial f/\partial x = x$ and $\partial f/\partial y = by$. If we use gradient descent with exact line search it turns out there is a formula for each point (x_k, x_y) in the descent down towards the minimum $(0, 0)$. If we start from $(x_0, y_0) = (b, 1)$ the formulas are:

$$x_k = b\left(\frac{b-1}{b+1}\right)^k, \quad y_k = \left(\frac{1-b}{b+1}\right)^k, \quad f(x_k, y_k) = \left(\frac{1-b}{b+1}\right)^{2k} f(x_0, y_0).$$

Note that in this particular example exact line search results in a stepsize $\alpha_k = \frac{2}{b+1}$ for all k, so the stepsize is constant. In the case of $b = 1$ the point (x_1, y_1) is $(0, 0)$ - success after just one iteration. In this case the graph of the

function is a symmetrically shaped bowl and the gradient goes exactly through $(0, 0)$. However, the point of this example is to see what happens when b is small. If we look at the ratio $(b-1)/(b+1)$ in the equations above, we can see that as b gets smaller it approaches -1. If b is very small the progress towards $(0, 0)$ becomes painfully slow. The path takes on a zig-zag pattern and looks something like:

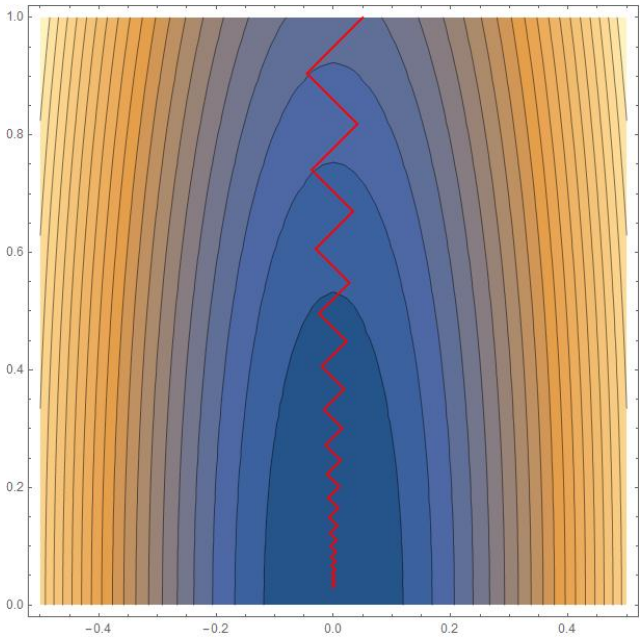

Figure 1: Example with $b = \frac{1}{20}$.

The reason that the progress is so slow is that at every iteration the stepsize α_k was chosen to minimize f along a line. But the direction of $-\nabla f$ even if it is the steepest is pointing far from $(x^*, y^*) = (0, 0)$. When b is small the graph of f looks like a narrow valley, and the path needlessly crosses the valley instead of moving further down the valley towards the bottom.

3.6 Momentum and the Path of a Heavy Ball

We want to improve the performance of the gradient descent method. The key idea here is that zig-zagging would not happen for a heavy ball rolling

downhill. Its momentum would result in a smoother path, bumping the sides but moving forwards for the most part. Mathematically this translates to adding a momentum term with coefficient β to the gradient. The new step with direction d_k "remembers" the previous direction d_{k-1}. The next step is calculated by

$$x_{k+1} = x_k - \alpha d_k \quad \text{with} \quad d_k = \nabla f(x_k) + \beta d_{k-1}.$$

Now there are two coefficients to be determined, the stepsize α and β. Note that the expression for x_{k+1} in the equation above involves d_{k-1}. The addition of momentum has turned a one-step method into a two-step method. To remedy this we rewrite the equation as two coupled equations for the state at time $k+1$ (one vector equation).

$$\begin{aligned} x_{k+1} &= x_k - \alpha d_k \\ d_{k+1} - \nabla f(x_{k+1}) &= \beta d_k. \end{aligned}$$

This is like reducing a single second order differential equation to a system of two first order equations. The heavy ball method can be applied to the previous example [6], with a choice of constant parameters

$$\alpha = \left(\frac{2}{1+\sqrt{b}}\right)^2 \quad \text{and} \quad \beta = \left(\frac{1-\sqrt{b}}{1+\sqrt{b}}\right)^2,$$

we will return to why these choices make sense later. These choices of stepsize and momentum give a convergence rate that looks like the rate for ordinary gradient descent, but with one difference b is replaced with $\sqrt{b}$.

$$\text{Ordinary descent factor: } \left(\frac{1-b}{1+b}\right)^2$$

$$\text{Accelerated descent factor: } \left(\frac{1-\sqrt{b}}{1+\sqrt{b}}\right)^2$$

When b is very small the descent factor is essentially $1-4b$, very close to 1. The accelerated descent factor is essentially $1-4\sqrt{b}$, much further from 1. To emphasize this suppose $b = \frac{1}{100}$, then $\sqrt{b} = \frac{1}{10}$ and the convergence factors become

$$\text{Ordinary descent factor: } \left(\frac{.99}{1.01}\right)^2 = .96$$

$$\text{Accelerated descent factor: } \left(\frac{0.9}{1.1}\right)^2 = .67$$

3.7 Nesterov Acceleration

This method is due to Yuri Nesterov. Instead of evaluating the gradient ∇f at x_k the idea is to evaluate it at the point $x_k + \gamma_k(x_k - x_{k-1})$, so this is also a way of utilizing x_{k-1} in the formula for x_{k+1}. By choosing $\gamma = \beta$ (momentum coefficient) both ideas are combined. Accelerated descent involves three parameters α, β, γ:

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(x_k + \gamma(x_k - x_{k-1})).$$

The following table illustrates how the parameters are related to the three methods:

Gradient descent	Stepsize α	$\beta = 0$	$\gamma = 0$
Heavy ball	Stepsize α	Momentum β	$\gamma = 0$
Nesterov acceleration	Stepsize α	Momentum β	shift ∇f by $\gamma\Delta\mathbf{x}$

We will do some analysis of convergence and convergence rates of these algorithms with fixed parameters α and $\beta = \gamma$, but first we need to introduce another class of functions. The main source for the next section is [3] if not otherwise stated.

4 Strongly convex and smooth functions

This special class of objective functions $f(x)$ has a benefit for fast first order methods, because one can implicitly make use of information on second derivatives in the error estimations. We give definitions and describe properties of the following in this section: functions whose gradient satisfies a Lipschitz condition, β-smooth functions, strongly convex functions and the combination of smoothness and convexity. The section is closed off by an investigation of two commonly used functions in machine learning.

Definition. [L-Lipschitz] A differentiable function f is said to be *L-Lipschitz* if its gradients are Lipschitz continuous, that is

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \ \forall x, y \in \mathbb{R}^n.$$

Lemma. [Descent lemma] *If f is twice differentiable and L-Lipschitz then*

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|x - y\|^2.$$

Proof. If f is twice differentiable then we have, by using first order expansion

$$\nabla f(x) - \nabla f(x + \alpha d) = \int_{t=0}^{\alpha} \nabla^2 f(x + td) d \, dt \quad d \neq 0.$$

Taking the norm gives

$$\left\| \int_{t=0}^{\alpha} \nabla^2 f(x + td) d \, dt \right\| \leq L\alpha \|d\|.$$

Dividing by α

$$\frac{\left\| \int_{t=0}^{\alpha} \nabla^2 f(x + td) d \, dt \right\|}{\alpha} \leq L\|d\|,$$

then dividing through by $\|d\|$ and taking the limit as $\alpha \to 0$ we have that

$$\frac{\left\| \int_{t=0}^{\alpha} \nabla^2 f(x + td) d \, dt \right\|}{\alpha \|d\|} = \frac{\|\alpha \nabla^2 f(x) d\|}{\alpha \|d\|} + O(\alpha) \underset{\alpha \to 0}{\Rightarrow} \frac{\|\nabla^2 f(x) d\|}{d} \leq L.$$

Taking the supremum over $0 \neq d \in \mathbb{R}^n$ we get the Hessian

$$\nabla^2 f(x) \preceq LI.$$

Furthermore, using the Taylor expansion of $f(x)$ and the uniform bound over Hessian we have that

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|x - y\|^2$$

□.

Motivated by this lemma we introduce a new terminology often used in the literature on the gradient descent method.

Definition. [β-smooth function] The function f is called *β-smooth* if

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{\beta}{2} \|x - y\|^2.$$

Clearly f is β-smooth if its gradient is β-Lipschitz. Now we "strengthen" the convexity notion by defining μ-strong convexity based on the two-point criterion:

Definition. [Strong convexity] A function f is said to be *μ-strongly convex* if

$$f(y) \geq f(x) + \langle \nabla f(x), (y - x) \rangle + \frac{\mu}{2} \|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

Minimizing both sides of this equation in y we get

$$f(x) - f(x^*) \leq \langle \nabla f(x), (x - x^*) \rangle - \frac{\mu}{2} \|y - x\|^2 =$$

$$= -\frac{1}{2}\|\sqrt{\mu}(x - x^*) - \frac{1}{\sqrt{\mu}}\nabla f(x)\|^2 + \frac{1}{2\mu}\|\nabla f(x)\|^2 \leq \frac{1}{2\mu}\|\nabla f(x)\|^2,$$

proving the following lemma:

Lemma [Polyak-Lojasiewicz condition] *If f is μ-strongly convex then it satisfies the following inequality*

$$\|\nabla f(x)\|^2 \geq 2\mu(f(x) - f(x^*))$$

where x^* is the minimum of f.

It can also be verified that a function f is μ-strongly convex if and only if $f(x) - \frac{\mu}{2}\|x\|^2$ is convex.

There are many problems in optimization where the function is both smooth and convex. Furthermore, such a combination results in some interesting consequences and Lemmas - that we will use to prove convergence of the Gradient descent method.

Lemma [Smooth and convex] *If $f(x)$ is convex and L-smooth, then*

$$f(y) - f(x) \leq \langle \nabla f(y), y - x\rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2$$

and

$$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geq \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Proof. By using the two-point criterion of convexity and the descent lemma we obtain

$$f(y)-f(x) = (f(y)-f(z))+(f(z)-f(x)) \leq \langle \nabla f(y), y-z\rangle+\langle \nabla f(x), z-x\rangle+\frac{L}{2}\|z-x\|^2.$$

Minimizing the right hand side (a quadratic function of z) over z yields

$$z = -\frac{1}{L}(\nabla f(x) - \nabla f(y)).$$

Substituting this in the previous inequality yields

$$\begin{aligned}
&f(y) - f(x)\\
\leq &\left\langle \nabla f(y), y - x + \frac{1}{L}(\nabla f(x) - \nabla f(y))\right\rangle - \frac{1}{L}\langle \nabla f(x), \nabla f(x) - \nabla f(y)\rangle + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2\\
= &\langle \nabla f(y), y - x\rangle - \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2\\
= &\langle \nabla f(y), y - x\rangle - \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|^2.
\end{aligned}$$

This proves the first inequality.

Changing the roles of x and y in the first inequality gives

$$f(x) - f(y) \leq \langle \nabla f(x), x - y \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Adding this to the first inequality results in

$$0 \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle - \frac{1}{2L} \|\nabla f(x) - \nabla f(y)\|^2.$$

Now we give an equivalent statement of μ-strongly convex functions.

Theorem [Equivalence of strong convexity and smoothness] *That $f(x)$ is μ-strongly convex and L-smooth is equivalent to*

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2, \quad \forall x, y.$$

Proof. Recall the definition of μ-strong convexity of f: for any x, y f satisfies

$$f(y) \geq f(x) + \langle \nabla f(x), x - y \rangle + \frac{\mu}{2} \|x - y\|^2.$$

Exchanging the role of x and y we get

$$f(x) \geq f(y) + \langle \nabla f(y), y - x \rangle + \frac{\mu}{2} \|x - y\|^2.$$

Adding this to the previous inequality yields

$$f(x) + f(y) \geq f(x) + f(y) + \langle \nabla f(x) - \nabla f(y), x - y \rangle + \mu \|x - y\|^2,$$

equivalently

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu \|x - y\|^2.$$

This theorem motivates the following definition.

Definition. [Class $S(m, L)$ convex function, [2]] Assume that the function $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable, convex. Assume that f has Lipschitz gradients with parameter L, i.e., f satisfies

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \leq L \|x - y\|^2 \ \forall x, y \in \mathbb{R}^n.$$

Let m be given such that $0 < m < L$ and

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geq m \|x - y\|^2, \ \forall x, y \in \mathbb{R}^n.$$

In other words, the continuously differentiable and convex function with parameters m and L satisfies the inequalities

$$m\|x-y\|^2 \leq (\nabla f(x) - \nabla f(y))^\top (x-y) \leq L\|x-y\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

We call such a function f a *strongly convex* function with L-smoothness. The set of all such functions with parameters m and L is denoted as $S(m, L)$. We call $\kappa := L/m$ the condition ratio of $f \in S(m, L)$. We adopt this terminology to distinguish the condition ratio of a function from the related concept of condition number of a matrix. The connection is that if f is twice differentiable, we have the bound: $\text{cond}(\nabla^2 f(x)) \leq \kappa \ \forall x \in \mathbb{R}^n$, where $\text{cond}(\cdot)$ is the common notion of the condition number.

Theorem [Class $S(m, L)$-functions] *Assume that f μ-strongly convex and L-smooth. Then for any $x, y \in \mathbb{R}^n$*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\mu L}{\mu + L}\|x-y\|^2 + \frac{1}{\mu + L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Proof. Note that when $\mu = L$, we have, by the the Smooth and convex lemma and the equivalence Theorem of strong convexity and smoothness,

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{\mu}\|\nabla f(y) - \nabla f(x)\|$$

and

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \mu\|x-y\|^2$$

respectively. Adding them yields

$$2\langle f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\mu}\|\nabla f(y) - \nabla f(x)\| + \mu\|x-y\|^2.$$

Dividing by 2 on both sides gives the desired inequality

$$\langle f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{2\mu}\|\nabla f(y) - \nabla f(x)\| + \frac{\mu}{2}\|x-y\|^2.$$

Now we assume that $L > \mu$ we show that the convex function $\phi(x) = f(x) - \frac{\mu}{2}\|x\|^2$ is $(L-\mu)$-smooth. That is we have to show that

$$\phi(y) \leq \phi(x) + \langle \nabla \phi(x), y - x \rangle + \frac{L-\mu}{2}\|y-x\|^2, \quad \forall x, y.$$

Since f is L-smooth f satisfiles

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y-x\|^2, \quad \forall x, y.$$

Now $\nabla\phi(x) = \nabla f(x) - \mu x$. Then

$$\begin{aligned}
\phi(y) &= f(y) - \frac{\mu}{2}\|y\|^2 \\
&\leq f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2, \ (f \text{ is } L\text{-smooth}) \\
&= \phi(x) + \frac{\mu}{2}\|x\|^2 + \langle \nabla\phi(x), y - x\rangle + \mu\langle x, y - x\rangle + \frac{L}{2}\|y - x\|^2 - \frac{\mu}{2}\|y\|^2 \\
&= \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 - \mu\left(-\langle x, y - x\rangle - \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2\right) \\
&= \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 - \mu\left(\langle y - x, y - x\rangle - \langle y, y - x\rangle - \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2\right) \\
&= \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 - \mu\left(\langle y - x, y - x\rangle - \frac{1}{2}\|y\|^2 + \langle x, y\rangle - \frac{1}{2}\|x\|^2\right) \\
&= \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 - \mu\left(\langle y - x, y - x\rangle - \frac{1}{2}\|x - y\|^2\right) \\
&= \phi(x) + \langle \nabla\phi(x), y - x\rangle + \frac{L}{2}\|x - y\|^2 - \frac{\mu}{2}\|x - y\|^2
\end{aligned}$$

thus proving that ϕ is $L - \mu$-smooth. Next we invoke the Smooth and convex lemma

$$\langle \nabla\phi(x) - \nabla\phi(y), x - y\rangle \geq \frac{1}{L - \mu}\|\nabla\phi(x) - \nabla\phi(y)\|^2.$$

Equivalently

$$\begin{aligned}
&\langle \nabla f(x) - \nabla f(y), x - y\rangle - \mu\|x - y\|^2 \\
\geq &\frac{1}{L - \mu}\left(\|\nabla f(x) - \nabla f(y)\|^2 - 2\mu\langle \nabla f(x) - \nabla f(y), x - y\rangle + \mu^2\|x - y\|^2\right) \\
\Leftrightarrow &
\end{aligned}$$

$$(1 - \frac{2\mu}{L - \mu})\langle \nabla f(x) - \nabla f(y), x - y\rangle \geq \frac{1}{L - \mu}\|\nabla f(x) - \nabla f(y)\|^2 + \frac{L\mu}{L - \mu}\|x - y\|^2.$$

And a last re-arrangment yields

$$\langle \nabla\phi(y) - \nabla\phi(x), y - x\rangle \geq \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Now we study the above properties for some functions.

Example 1. The function $f(x) = x^\top Ax$ is a μ strongly convex function when A is a symmetric positive definite matrix whose eigenvalues are all greater than or equal to $\frac{1}{2}\mu$. It follows from the fact that $\nabla f(x) = 2Ax$ and the Hessian $\nabla^2 f(x) = 2A \succeq \mu I$ since the eigenvalues of A are all greater than or equal to $\frac{1}{2}\mu$. Together with the Taylor expansion, we obtain

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2}(y - x)^\top(2A)(y - x)$$

$$\geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}(y-x)^\top (y-x).$$

This automatically means that the function is convex, as can be seen by checking the Hessian matrix.

This function is in the class $S(m, L)$ since $(\nabla f(x) - \nabla f(y))^\top (x-y) = 2(x-y)^\top A(x-y)$ which is greater than or equal to the minimal eigenvalue of A and less than or equal to the maximal eigenvalue of A. So the the condition number of f, κ is also equal to the condition number of A, i.e. the ratio between the maximum and minimum eigenvalues of A. In geometric terms, when κ is close to 1, it means that the level sets of f are nearly round, while if κ is large it means that the level sets of f may be quite elongated.

Example 2. The function $f(x) = (a^\top x)^+ := \max\{a^\top x, 0\}$, where a is any nonzero vector in $\mathbb{R}^d$. The function is convex since for all $x, y \in \mathbb{R}^d$ and any $0 \leq t \leq 1$ we have

$$f(tx + (1-t)y) = (a^\top (tx + (1-t)y))^+ = \max\{a^\top (tx + (1-t)y), 0\}$$

$$= \max\{ta^\top x + (1-t)a^\top y, 0\} \leq t \max\{a^\top x, 0\} + (1-t)\max\{a^\top y, 0\} = tf(x) + (1-t)f(y)$$

Nevertheless this function is neither linear nor strongly convex because

$$\nabla f(x) = \begin{cases} 0 & \text{if } a^\top x < 0 \\ a & \text{if } a^\top x > 0 \end{cases}$$

Clearly the function $(a^\top x) - \frac{\mu}{2}\|x\|^2$ is non-convex, thus $f(x)$ is not strongly convex.

4.1 Basics on linear finite dimensional control theory: discrete-time systems

We want to study convergence and convergence rates of optimization methods by using stability analysis of linear control systems. The following material can be found in any introductory books on this topic. Here we use [5]. We use a state space description, the system of first order difference equations

$$\begin{aligned} x_i(k+1) &= f_i(x_1(k), ..., x_n(k), u_1(k), ..., u_m(k)),\ i = 1, ..., n \\ y_j(k) &= g_j(x_1(k), ..., x_n(k), u_1(k), ..., u_m(k)),\ j = 1, ..., p \end{aligned}$$

is called an autonomous control system with state $x_1, ..., x_n$, *inputs* $u_1, ..., u_m$ and *outputs* $y_1, ..., y_p$. Often we write this in matrix form:

$$x(k+1) = f(x(k), u(k)), \quad y(k) = g(x(k), u(k))$$

where $x(k)^\top = (x_1(k), ..., x_n(k))$, $u(k)^\top = (u_1(k), ..., u_m(k))$, and $y(k)^\top = (y_1(k), ..., y_p(k))$, and $f(\cdot)^\top = (f_1(\cdot), ..., f_n(\cdot))$, and $g(\cdot)^\top = (g_1(\cdot), ..., g_p(\cdot))$,

and m, n, p are positive integers. Here we assume that $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$. If

$$f(x, u) = Ax + Bu, \quad g(x, u) = Cx + Du,$$

then we call the resulting system a linear control system,

$$x(k + 1) = A(k)x(k) + B(k)u(k), \quad y(k) = C(k)x(k) + D(k)u(k)$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$. If these matrices are constant then the system is called a linear time-invariant system. If there is a matrix $K \in \mathbb{R}^{m\times n}$ such that $u(x) = Kx$ we call K the state feedback and the matrix $A_{\text{cl}} := A - BK$ is called the feedback matrix or closed loop matrix.

Definition. (Stability of the linear system) A time-invariant discrete-time linear system $x(k+1) = Ax(k)$ is (asymptotically) stable if all eigenvalues of A lie inside the unite circle.

This implies that the trajectory converges to its fixed point $x = Ax$. Note that we have simplified our exposition on stability a little bit in order to avoid technical details. Also note that if the system is not linear the stability can be local or global. The convergence of an iterative method can be studied by stability theory, however it does not always provide the convergence rate.

For later use we state the following characterization of the locations of eigenvalues.

Proposition. (Root locations of a second degree polynomial) *Let* $p(z) := z^2 + az + b$ *with real numbers* a, b. *Both roots of* $p(z) = 0$ *lie inside the unit circle if and only if* $b < 1$, *and* $1 - a + b > 0$, *and* $1 + a + b > 0$.

Proof. We show first that the polynomial equation $s^2 + as^2 + b = 0$ has all roots on the open left half complex plane, $\mathbb{C}^-$ if and only if $a > 0$ and $b > 0$. Assume s_1, s_2 are the roots. Then we have

$$a = -(s_1 + s_2), \ b = s_1 s_2.$$

It is obvious that $a > 0$ and $b > 0$ if $s_1$and s_2 have negative real parts. On the other hand, let $b > 0$ and $a > 0$. First we consider the real roots. It is clear that $b > 0$ implies that s_1 and s_2 must have the same signs. If they are both negative then $a > 0$. If they were both positive then a would by negative, contradicting the condition that $a > 0$. Now consider a pair of complex conjugate roots since this is a real polynomial. Let $s_1 = \sigma + i\omega$, and $s_2 = \sigma - i\omega$ and we want to show that $\sigma < 0$. Now $b = s_1 s_2 = \sigma^2 + \omega^2 > 0$, but $0 < a = -2\sigma$. So σ must be negative.

Next we show that p has all zeros inside the unit circle if and only if $q(s) =$

$p\left(\frac{1+s}{1-s}\right)(1-s)^2$ has all zeros in $\mathbb{C}^-$. This is true because the mapping $z = \frac{1+s}{1-s}$ maps $\mathbb{C}^-$ to the open unit disk and the inverse of this maps the unit disc to $\mathbb{C}^-$ by elementary complex analysis.

Now

$$q(s) = p\left(\frac{1+s}{1-s}\right)(1-s)^2 = (1-a+b)s^2 + 2(1-b)s + 1 + a + b.$$

Then, by the first step of this proof we get p has both zeros inside the unit circle if and only if $1-a+b > 0, 1-b > 0$ and $1+a+b > 0$.

5 Analysis of the class of momentum methods

Previously we mentioned the Heavy ball method and Nesterov's accelerated method as ways of speeding up the convergence of the Gradient descent method, without showing how - so now we give a unified treatment using dynamical systems. This is based on the framework in [2]. The paper has an interesting topic because it relates the optimization methods to robust control theory. As pointed out in [2], *Convex optimization algorithms provide a powerful toolkit for robust, efficient, large-scale optimization algorithms. They provide not only effective tools for solving optimization problems, but are guaranteed to converge to accurate solutions in provided time budgets, are robust to errors and time delays, and are amendable to declarative modeling that decouples the algorithm design from the problem formulation. However, as we push up against the boundaries of the convex analysis framework, try to build more complicated models, and aim to deploy optimization systems in highly complex environments, the mathematical guarantees of convexity start to break. The standard proof techniques for analyzing convex optimization rely on deep insights by experts and are devised on an algorithm-by-algorithm basis. It is thus not clear how to extend the toolkit to more diverse scenarios where multiple objectives – such as robustness, accuracy, and speed – need to be delicately balanced.*

The research in [2] makes an attempt at providing a systematized approach to the design and analysis of optimization algorithms using techniques from control theory. Since that topic is well beyond the scope of this thesis we will only show that the first order optimization methods introduced up to now can be cast in dynamical system form, and provide the basic ideas for convergence analysis.

We want to understand the algorithms designed to solve the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

as a dynamical system with control in the input-output form

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ y_k &= C\xi_k + Du_k \\ u_k &= \phi(y_k). \end{aligned}$$

The linear system (two first equations) is connected in *feedback* with a non-linearity ϕ. The output y is transformed by the map $\phi : \mathbb{R}^d \to \mathbb{R}^d$ and is used as the input to the linear system. In our case the interconnected non-linearity will have the form $\phi(y) = \nabla f(y)$ with $f \in S(m, L)$. We will be content to limit our study to the special case of a quadratic objective function f.

5.1 The well-known first order methods as control systems

In this subsection we prove that Gradient descent, Nesterov's method and Polyak's Heavy-ball method can all be cast in the dynamical system setting.

Proposition. (Gradient descent method) *The gradient descent method is equivalent to the dynamical system described above with*

$$A = I_d, B = -\alpha I_d, C = I_d, D = 0_d.$$

Proof. Writing the dynamical system in its explicit form we have

$$\begin{aligned} \xi_{k+1} &= \xi_k - \alpha u_k \\ y_k &= \xi_k \\ u_k &= \nabla f(y_k). \end{aligned}$$

Eliminating y_k and u_k we get

$$\xi_{k+1} = \xi_k - \alpha \nabla f(\xi_k).$$

Renaming ξ to x yields

$$x_{k+1} = x_k - \alpha \nabla f(x_k).$$

This is the Gradient descent with constant stepsize.

Proposition. (Nesterov's method) *Nesterov's method is equivalent to the dynamical system described above with*

$$A = \begin{pmatrix} (1+\beta)I_d & -\beta I_d \\ I_d & 0_d \end{pmatrix}, B = \begin{pmatrix} -\alpha I_d \\ 0_d \end{pmatrix}, C = \begin{pmatrix} (1+\beta)I_d & -\beta I_d \end{pmatrix}, D = 0_d.$$

Proof. From the form of A we see that there are two block components in the vector ξ. So let $\xi_k^\top = (\xi_k^{(1)}, \xi_k^{(2)})$ and note that the decomposition of ξ should be in accordance with the decomposition of A. Now writing the dynamical system

explicitly we get

$$\begin{aligned}\xi_{k+1}^{(1)} &= (1+\beta)\xi_k^{(1)} - \beta\xi_k^{(2)} - \alpha u_k\\ \xi_{k+1}^{(2)} &= \xi_k^{(1)}\\ y_k &= (1+\beta)\xi_k^{(1)} - \beta\xi_k^{(2)}\\ u_k &= \nabla f(y_k).\end{aligned}$$

Note that the second equation above is equivalent to $\xi_k^{(2)} = \xi_{k-1}^{(1)}$, i.e. the partial state $\xi^{(2)}$ is a delayed version of the state $\xi^{(1)}$. Substituting this into the preceding equations yields

$$\begin{aligned}\xi_{k+1}^{(1)} &= (1+\beta)\xi_k^{(1)} - \beta\xi_{k-1}^{(1)} - \alpha u_k\\ y_k &= (1+\beta)\xi_k^{(1)} - \beta\xi_{k-1}^{(1)}\\ u_k &= \nabla f(y_k).\end{aligned}$$

By eliminating u_k, and by renaming $\xi^{(1)}$ to x we obtain the common form of Nesterov's method

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

Proposition. (Heavy-ball method) *Heavy-ball method method is equivalent to the dynamical system described above with*

$$A = \begin{pmatrix}(1+\beta)I_d & -\beta I_d\\ I_d & 0_d\end{pmatrix}, B = \begin{pmatrix}-\alpha I_d\\ 0_d\end{pmatrix}, C = \begin{pmatrix}I_d & 0_d\end{pmatrix}, D = 0_d.$$

Proof. As in proving the previous proposition we have the following dynamical system

$$\begin{aligned}\xi_{k+1}^{(1)} &= (1+\beta)\xi_k^{(1)} - \beta\xi_k^{(2)} - \alpha u_k\\ \xi_{k+1}^{(2)} &= \xi_k^{(1)}\\ y_k &= \xi_k^{(1)}\\ u_k &= \nabla f(y_k).\end{aligned}$$

Substituting the second equation into the first yields

$$\begin{aligned}\xi_{k+1}^{(1)} &= (1+\beta)\xi_k^{(1)} - \beta\xi_{k-1}^{(1)} - \alpha u_k\\ y_k &= \xi_k^{(1)}\\ u_k &= \nabla f(y_k).\end{aligned}$$

Eliminating u and renaming ξ to x we get

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}).$$

This is the heavy-ball method.

5.2 Proof of convergence: the quadratic objective function

The standard two-step procedure in convergence analysis of a convex optimization algorithm is:
1. We first show that the algorithm has a fixed point that solves the optimization problem at hand.
2. Then we prove that the algorithm converges at a specified rate to its optimal solution for a suitable choice of the initial value.

Such analysis is called stability analysis in the dynamical system setting. By writing a first order algorithm as a dynamical system, we can unify the stability analysis. If we know that the minimum is at y^*, a necessary condition for optimality is that $u^* = \nabla f(y^*) = 0$. Substituting into the dynamical system the fixed point satisfies

$$y^* = C\xi^* \text{ and } \xi^* = A\xi^*.$$

This means in particular that A must have an eigenvalue of value 1. If the block matrices of A are diagonal as in the cases of Gradient descent or Heavy-ball or Nesterov's method shown above then the eigenvalues 1 will have a geometric multiplicity of at least d.

Assume that f is a convex quadratic function

$$f(y) = \frac{1}{2}y^\top Qy - p^\top y + r$$

where

$$mI_d \preceq Q \preceq LI_d$$

in the positive definite ordering. The gradient of f is

$$\nabla f(y) = Qy - p$$

and the optimal solution is at $y^* = Q^{-1}p$. Now substituting these conditions into the dynamical system we have the following specific form of dynamical system

$$\begin{aligned} \xi_{k+1} &= A\xi_k + Bu_k \\ y_k &= C\xi_k \\ u_k &= \nabla f(y_k) = Qy_k - p = Q(y_k - y^*). \end{aligned}$$

Subtracting ξ^* from both sides of the first equation and using the equations ξ^* and y^*: $y^* = C\xi^*$, and $\xi^* = A\xi^*$ yield the following feedback system for $\xi - \xi^*$:

$$\xi_{k+1} - \xi^* = (A + BQC)(\xi_k - \xi^*).$$

Let the feedback matrix $A + BQC$ be A_{cl}. This system is (asymptotically) stable, i.e. the trajectory generated by this system will converge to its fixed point

if and only if all eigenvalues of A_{cl} lie inside the unit circle. Or equivalently its spectral radius, $\rho(A_{\text{cl}})$ is less than 1.

We have that

$$\rho(A) \leq \|A^k\|^{1/k} \quad \text{for all } k \text{ and} \quad \rho(A) = \lim_{k\to\infty} \|A\|^{1/k},$$

here $\|\cdot\|$ denotes the matrix norm induced by the vector 2-norm. So for any $\epsilon > 0$, and for all k sufficiently large, we have that $\rho(A_{\text{cl}})^k \leq \|A_{\text{cl}}^k\| \leq (\rho(A_{\text{cl}}) + \epsilon)^k$. Hence the convergence rate can be bounded:

$$\|\xi_k - \xi_*\| = \|A_{\text{cl}}^k(\xi_0 - \xi_*)\| \leq \|A_{\text{cl}}^k\|\|\xi_0 - \xi_*\| \leq (\rho(A_{\text{cl}}) + \epsilon)^k\|\xi_0 - \xi_*\|.$$

So the spectral radius determines the rate of convergence of the algorithm.Note that the spectral radius of a positive definite matrix is its largest eigenvalue. Now we will give the convergence rates.

Theorem. *Assume that $f : \mathbb{R}^n \to \mathbb{R}$ is defined as $f(x) = x^\top Qx - p^\top x + r$ and Q is any matrix that satisfies $mI_n \preceq Q \preceq LI_n$. Let $\kappa := \frac{L}{m}$. Then we have the following convergence rate bound ρ:*

1. *Gradient descent method: $\rho = 1 - \frac{1}{\kappa}$ if $\alpha = \frac{1}{L}$, and $\rho = \frac{\kappa-1}{\kappa+1}$ if $\alpha = \frac{2}{L+m}$.*
2. *Nesterov's method: $\rho = 1 - \frac{1}{\sqrt{\kappa}}$ if $\alpha = \frac{1}{L}$, $\beta = \frac{\kappa-1}{\kappa+1}$, and $\rho = 1 - \frac{2}{\sqrt{3\kappa+1}}$ if $\alpha = \frac{4}{3L+m}$, $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$.*
3. *Heavy-ball method: $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ if $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$.*

Proof. To find the worst-case convergence rate is equivalent to solving the following maximization problem

$$\rho = \max_{mI_n \preceq Q \preceq LI_n} \rho(A_{\text{cl}}).$$

Assume that eigenvalues of Q are $0 < m \leq \lambda_n \leq \lambda_{n-1} \leq \cdots \leq \lambda_1 \leq L$. Then Q can be factorized as

$$Q = U\Lambda U^\top, \text{ where } \Lambda = \text{diag}(\lambda_1, ..., \lambda_n) \text{ and } UU^\top = I_n.$$

(1) The dynamical system of the gradient descent method is $A = I_n$,$B = -\alpha I_n$ and $C = I_n$. Then

$$A_{\text{cl}} = I - \alpha Q = U = U(I_n - \alpha\Lambda)U^\top.$$

Since $\rho(A_{\text{cl}}) = \rho((I_n - \alpha\Lambda))$ the problem is reduced to

$$\rho(A_{\text{cl}}) = \max_{m\leq\lambda\leq L} |1 - \alpha\lambda|.$$

Now the functions $|1-\alpha\lambda|$ are convex so $\max_{m\le\lambda\le L}|1-\alpha\lambda|$ is also convex by the Lemma on maximum of convex functions. Then the maximum must occur at the boundary, i.e., $\lambda = m$ and/or $\lambda = L$. Now we have

$$\rho(A_{\rm cl}) = \max\{|1-\alpha m|, |1-\alpha L|\}.$$

Clearly, when $\alpha = 1/L$, $\rho(A_{\rm cl}) = 1 - m/L = 1 - 1/\kappa$. If $\alpha = 2/(L+m)$ we have

$$\rho(A_{\rm cl}) = \max\left\{\frac{L-m}{L+m}, \frac{m-L}{L+m}\right\} = \frac{L-m}{L+m} = \frac{1-\kappa}{1+\kappa}.$$

(2) The dynamical system of Nesterov's method is

$$A = \begin{pmatrix}(1+\beta)I_n & -\beta I_n \\ I_n & 0_d\end{pmatrix}, B = \begin{pmatrix}-\alpha I_n \\ 0_n\end{pmatrix}, C = \begin{pmatrix}(1+\beta)I_n & -\beta I_n\end{pmatrix}.$$

Then

$$A_{\rm cl} = \begin{pmatrix}(1+\beta)I_n - \alpha(1+\beta)Q & -\beta I_n + \alpha\beta I_n \\ I_n & 0\end{pmatrix}.$$

$$= \begin{pmatrix}U & 0_n \\ 0_n & U\end{pmatrix}\begin{pmatrix}(1+\beta)(I_n - \alpha\Lambda) & -\beta(1-\alpha\Lambda) \\ I_n & 0_n\end{pmatrix}\begin{pmatrix}U & 0_n \\ 0_n & U\end{pmatrix}^\top$$

By permuting rows and columns the matrix

$$\begin{pmatrix}(1+\beta)(I_n - \alpha\Lambda) & -\beta(1-\alpha\Lambda) \\ I_n & 0_n\end{pmatrix}$$

can be transformed into a block diagonal matrix, that is similar to $A_{\rm cl}$, where the main diagonal blocks consists of matrices of the form

$$\begin{pmatrix}(1+\beta)(1-\alpha\lambda_i) & -\beta(1-\alpha\lambda_i) \\ 1 & 0\end{pmatrix}, i = 1, ..., n$$

Therefore the eigenvalues of $A_{\rm cl}$ are all the eigenvalues of these submatrices.

Thus the optimization problem $\rho = \max_{mI_n \preceq Q \preceq LI_n}\rho(A_{\rm cl})$ is reduced to

$$\max_{m\le\lambda\le L}\max\{|z_1(\lambda)|, |z_2(\lambda)|\}$$

where $z_1(\lambda)$ and $z_2(\lambda)$ are eigenvalues of the matrix $\begin{pmatrix}(1+\beta)(1-\alpha\lambda) & -\beta(1-\alpha\lambda) \\ 1 & 0\end{pmatrix}$, i.e., they are roots of the following equation

$$z^2 - (1+\beta)(1-\alpha\lambda)z + \beta(1-\alpha\lambda) = 0.$$

The magnitudes of the roots satisfy:

$$\max\{|z_1(\lambda)|, |z_2(\lambda)|\} = \begin{cases} \frac{1}{2}|(1+\beta)(1-\alpha\lambda)| + \frac{1}{2}\sqrt{\Delta} & \text{if } \Delta \geq 0 \\ \sqrt{\beta(1-\alpha\lambda)} & \text{if } \Delta < 0 \end{cases}$$

where $\Delta := (1+\beta)^2(1-\alpha\lambda)^2 - 4\beta(1-\alpha\lambda)$. If α, β are fixed, then $h(\lambda) = \max\{|z_1(\lambda)|, |z_2(\lambda)|\}$ is a function of λ. We are going to show that $h(\lambda)$ is continuous and quasiconvex, because that would imply that the maximum over λ occurs at a boundary point.

Let's first consider when $\Delta \geq 0$. So

$$(1-\alpha\lambda)((1+\beta)^2(1-\alpha\lambda) - 4\beta) \geq 0$$

which is equivalent to

$$\lambda \leq \frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2 \text{ or } \lambda \geq \frac{1}{\alpha}.$$

So we get that

$$h(\lambda) = \begin{cases} \frac{1}{2}(1+\beta)(\alpha\lambda - 1) + \frac{1}{2}\sqrt{\Delta} & \text{if } \frac{1}{\alpha} \leq \lambda \leq L \\ \sqrt{\beta(1-\alpha\lambda)} & \text{if } \frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2 < \lambda < \frac{1}{\alpha} \\ \frac{1}{2}(1+\beta)(1-\alpha\lambda) + \frac{1}{2}\sqrt{\Delta} & \text{if } m \leq \lambda \leq \frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2 \end{cases}$$

The left and right limits agrees at the point $\frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2$:

$$h\left(\frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2\right)^- = \frac{2\beta}{1+\beta} = \lim_{\lambda \to \frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^{2+}} \sqrt{\beta(1-\alpha\lambda)},$$

and likewise for the point $\frac{1}{\alpha}$:

$$h(\frac{1}{\alpha})^+ = 0 = \lim_{\lambda \to \frac{1}{\alpha}^-} \sqrt{\beta(1-\alpha\lambda)},$$

so h is indeed continuous.

If $\lambda \in \left[m, \frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2\right]$ we have that

$$h'(\lambda) = -\frac{1}{2}(1+\beta)\alpha + \frac{\alpha}{2\sqrt{\Delta}}((1+\beta)^2(\alpha\lambda - 1) + 2\beta) < 0.$$

If $\lambda \in \left(\frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2, \frac{1}{\alpha}\right)$ then

$$h'(\lambda) = \frac{\alpha\beta}{2\sqrt{\beta(1-\alpha\lambda)}} < 0.$$

If $\lambda \in \left[\frac{1}{\alpha}, L\right]$ then

$$h'(\lambda) = \frac{1}{2}(1+\beta)\alpha + \frac{\alpha}{2\sqrt{\Delta}}((1+\beta)^2(\alpha\lambda - 1) + 2\beta) > 0.$$

The function h attains its minimum at $\frac{1}{\alpha}$, it is non-increasing on $[m, \frac{1}{\alpha}]$ and non-decreasing on $[\frac{1}{\alpha}, L]$, so it is quasiconvex. Thus h attains it maximum at $\lambda = m$ or $\lambda = L$.

For the case when $\alpha = \frac{1}{L}$ and $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, the choice of $\lambda = L$ yields zero, so the maximum must be achieved at $\lambda = m$, which yields:

$$\rho = \sqrt{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\left(1 - \frac{1}{\kappa}\right)} = \sqrt{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\frac{(\sqrt{\kappa}+1)(\sqrt{\kappa}-1)}{\kappa}} = 1 - \frac{1}{\sqrt{\kappa}}.$$

For the case when $\alpha = \frac{4}{3L+m}$ and $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$ the disciminant Δ is zero. Also

$$\frac{1}{\alpha}\left(\frac{1-\beta}{1+\beta}\right)^2 = \frac{3L+m}{4}\left(\frac{2}{\sqrt{3\kappa+1}}\right)^2 = \frac{m(3\kappa+1)}{4}\left(\frac{2}{\sqrt{3\kappa+1}}\right)^2 = m,$$

and

$$\frac{1}{\alpha} = \frac{3L+m}{4} < \frac{3L+L}{4} = L.$$

Since $h(m) = 1 - \frac{2}{\sqrt{3\kappa+1}} > \frac{k-1}{\sqrt{3\kappa+2}\sqrt{3\kappa+1}} = h(L)$ for all κ we get that

$$\rho = 1 - \frac{2}{\sqrt{3\kappa+1}},$$

as desired.

(3) The dynamical system of the Heavy-ball method is :

$$A = \begin{pmatrix}(1+\beta)I_n & -\beta I_n \\ I_d & 0_n\end{pmatrix}, B = \begin{pmatrix}-\alpha I_n \\ 0_n\end{pmatrix}, C = \begin{pmatrix}I_n & 0_n\end{pmatrix}.$$

Then

$$A_{\text{cl}} = \begin{pmatrix} (1+\beta)I_n - \alpha Q & -\beta I_n \\ I_n & 0 \end{pmatrix} = \begin{pmatrix} U & 0_n \\ 0_n & U \end{pmatrix} \begin{pmatrix} (1+\beta)I_n - \alpha\Lambda & -\beta I_n \\ I_n & 0_n \end{pmatrix} \begin{pmatrix} U & 0_n \\ 0_n & U \end{pmatrix}^\top$$

The eigenvalues of A_{cl} are all the eigenvalues of the submatrices

$$\begin{pmatrix} (1+\beta) - \alpha\lambda_i & -\beta \\ 1 & 0 \end{pmatrix}, i = 1, ..., n.$$

As in the case of Nesterov's method the eigenvalues of these matrices $z_1(\lambda)$ and $z_2(\lambda)$ satisfy

$$z^2 - (1 + \beta - \alpha\lambda)z + \beta = 0.$$

Hence we have to find the solution to the following optimization problem

$$\rho = \max_{m \le \lambda \le L} \max\{|z_1(\lambda)|, |z_2(\lambda)|\}.$$

The magnitudes of the roots satisfy:

$$\max\{|z_1(\lambda)|, |z_2(\lambda)|\} = \begin{cases} \frac{1}{2}|1 + \beta - \alpha\lambda| + \frac{1}{2}\sqrt{\Delta} & \text{if } \Delta \ge 0 \\ \sqrt{\beta} & \text{if } \Delta < 0 \end{cases}$$

where $\Delta := (1 + \beta - \alpha\lambda)^2 - 4\beta$. Let $h(\lambda) = \max\{|z_1(\lambda)|, |z_2(\lambda)|\}$ for fixed α, β. We want to show that $h(\lambda)$ is continuous and quasiconvex and thus attains it maximum at a boundary point. By calculating for which λ (in terms of α,β) the discriminant Δ is non-negative we get that:

$$h(\lambda) = \begin{cases} \frac{1}{2}(\alpha\lambda - 1 - \beta) + \frac{1}{2}\sqrt{\Delta} & \text{if } \frac{1}{\alpha}(1 + 2\sqrt{\beta} + \beta) \le \lambda \le L \\ \sqrt{\beta} & \text{if } \frac{1}{\alpha}(1 - 2\sqrt{\beta} + \beta) < \lambda < \frac{1}{\alpha}(1 + 2\sqrt{\beta} + \beta) \\ \frac{1}{2}(1 + \beta - \alpha\lambda) + \frac{1}{2}\sqrt{\Delta} & \text{if } m \le \lambda \le \frac{1}{\alpha}(1 - 2\sqrt{\beta} + \beta) \end{cases}$$

For brevity denote $\underline{\lambda} := \frac{1}{\alpha}(1 - 2\sqrt{\beta} + \beta)$ and $\overline{\lambda} := \frac{1}{\alpha}(1 + 2\sqrt{\beta} + \beta)$. We have that

$$h(\underline{\lambda})^- = \sqrt{\beta} = h(\underline{\lambda})^+$$

and

$$(\overline{\lambda})^- = \sqrt{\beta} = h(\overline{\lambda})^+,$$

so h is continuous.

If $\lambda \in [m, \underline{\lambda}]$ then

$$h'(\lambda) = -\frac{\alpha}{2} - \frac{\alpha(1 + \beta - \alpha\lambda)}{\sqrt{\Delta}} < 0.$$

If $\lambda \in (\underline{\lambda}, \overline{\lambda})$ then

$$h'(\lambda) = 0,$$

so h is constant on the interval.

If $\lambda \in [\overline{\lambda}, L]$ then

$$h'(\lambda) = \frac{\alpha}{2} - \frac{\alpha(1+\beta-\alpha\lambda)}{\sqrt{\Delta}} > 0.$$

Given a point $\lambda^* \in [\underline{\lambda}, \overline{\lambda}]$, h attains its minimum. It is non-increasing on $[m, \lambda^*]$ and non-decreasing on $[\lambda^*, L]$ and thus it is quasiconvex. So h attains its maximum at $\lambda = m$ or $\lambda = L$.

Now with $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$ and $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ we get that

$$\underline{\lambda} = \frac{1}{\alpha}(1-\sqrt{\beta})^2 = \frac{(\sqrt{L}+\sqrt{m})^2}{4}\left(\frac{2}{\sqrt{\kappa}+1}\right)^2 = m\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}+1}\right)^2 = m,$$

and

$$\overline{\lambda} = \frac{1}{\alpha}(1+\sqrt{\beta})^2 = \frac{(\sqrt{L}+\sqrt{m})^2}{4}\left(\frac{2\sqrt{\kappa}}{\sqrt{\kappa}+1}\right)^2 = L\left(1+\sqrt{\frac{m}{L}}\right)^2\left(\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1}\right)^2 = L.$$

Finally

$$h(m) = h(L) = \sqrt{\beta} = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} = \rho,$$

as desired.

Remarks.
1. Note that the spectral radius of Q for the positive definite matrix is the largest eigenvalue of Q so we in fact get the worst-case rates.

2. We have different choices of the parameters α, β in the algorithms.

- In gradient descent we chose $\alpha = \frac{1}{L}$ in the first case. Note that this is just a choice but a popular choice among the users. However, $\alpha = \frac{2}{L+m}$ is an optimal choice. To see this, we look for the intersection of the curve $y = |1-\alpha m|$ and the curve $y = |1-\alpha L|$, as functions of α. It is clear that the intersection whose value is the minimum is the intersection between the line $y = L\alpha - 1$ and $y = 1 - m\alpha$. Thus the worst case spectral radius is attained at $\alpha = \frac{2}{L+m}$.
- Similarly the choice of $\alpha = \frac{1}{L}$ and $\beta = \frac{\kappa-1}{\kappa+1}$, is a standard choice for the Nesterov method. However the choice of $\alpha = \frac{4}{3L+m}$, $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$ is an

optimal tuning. In this case the optimum is reached when the discriminant equals 0. This is true for the Heavy-ball algorithm as well thus $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$ is an optimal choice.
With optimal tuning of the parameters the rate bounds obtained for the two momentum methods are better than for the fixed step Gradient descent method.

3. The bounds are tight. That is, there exists a quadratic function that achieves the worst-case ρ.

4. Note in the estimate of the convergence radii that the basic gradient descent algorithm requires $O(\kappa \log(1/\epsilon))$ iterates to reach ϵ-accuracy while Nesterov's method attains the improved complexity of $O(\sqrt{\kappa}\log(1/\epsilon))$. This is particularly relevant in Machine Learning applications since the strong convexity parameter μ can often be viewed as a regularization term, and $1/\mu$ can be as large as the sample size. Thus reducing the number of steps from "sample size" to $\sqrt{\text{sample size}}$. This is a huge deal, especially in large scale applications.

6 Discussions and further remarks

We derived and proved convergence rates for popular optimization methods when applied to a class $S(m, L)$ of quadratic functions from a control theoretical point of view. The techniques used in the last section does not extend to the case where f is a more general strongly convex function. Here we give another characterization of stability that can be useful for more general objective functions.

6.1 Lyapunov stability and the LMI approach

It is reasonably easy and intuitive to work out the convergence rates for this class of problems. We provide a reason here. First we prove:

Proposition. *Given an $n \times n$ matrix A. The following two things are equivalent.*

1. *All eigenvalues of A satisfy: $|\lambda_i(A)| < 1$ for all $i = 1, ..., n$ counted with multiplicity.*
2. *There exists a $P \succ 0$ such that $A^\top PA - P \prec 0$.*

Proof. (2) $\Rightarrow$ (1): Assume $P \succ 0$ satisfies $A^\top PA - P \prec 0$. Let $Av = \lambda v$ and $v \neq 0$. Then

$$0 > v^*(A^\top PA - P)v = (|\lambda|^2 - 1)v^*Pv.$$

This implies $|\lambda| < 1$ since $v^*Pv > 0$. So (1) holds.
(1) $\Rightarrow$ (2): Assume A has all eigenvalues inside the unit circle. We give a closed

form of the solution P.

Let $P := \sum_{k=0}^{\infty} (A^k)^\top Q A^k$, where $Q \succ 0$ is an $n \times n$ matrix. This series is well-defined, in other words, it converges, since $|\lambda_i(A)| < 1$. Then

$$A^\top P A - P = \sum_{k=1}^{\infty} (A^k)^\top Q A^k - \sum_{k=0}^{\infty} (A^k)^\top Q A^k = -Q \prec 0$$

as desired. □

In fact, we can show that the existence of P in (2) is unique. To accomplish this we consider a more general equation

$$A_1 X A_2 - X = C$$

where $A_1 \in \mathbb{R}^{m\times m}$, $A_2 \in \mathbb{R}^{n\times n}$ and $X, C \in \mathbb{R}^{m\times n}$.

Lemma *The equation $A_1 X A_2 - X = C$ has a unique solution if and only if no eigenvalue of A_1 is a reciprocal of an eigenvalue of A_2.*

Proof. We need to show that the condition on A_1 and A_2 is equivalent to the condition that $A_1 X A_2 = X$ implies $X = 0$. If we had two such solutions X_1, X_2 we would have

$$A_1(X_1 - X_2)A_2 - (X_1 - X_2) = 0 \quad \Longleftrightarrow \quad A_1 \Delta X A_2 = \Delta X$$

where $\Delta X := X_1 - X_2$. Repeating this equality we obtain

$$A_1^{k-j} \Delta X A_2^{k-j} = \Delta X \text{ and } A_1^j \Delta X = A_1^k \Delta X A_2^{k-j}, \text{ for } k \geq j \geq 0.$$

Now for a polynomial of degree k

$$p(\lambda) = \sum_{j=0}^{k} a_j \lambda^j$$

we define the polynomial of degree k as follows

$$p^*(\lambda) = \sum_{j=0}^{k} a_j \lambda^{k-j} = \lambda^k p\left(\frac{1}{\lambda}\right)$$

from which it follows that

$$p(A_1)\Delta X = A_2^k \Delta X p^*(A_2).$$

Next let $\phi_1(\lambda)$ be the characteristic polynomial of A_1 and $\phi_2(\lambda)$ be the characteristic polynomial of A_2. Since $\phi(\lambda)$ and $\phi_2^*(\lambda)$ are co-prime, there are polynomials $p(\lambda)$ and $q(\lambda)$ such that

$$p(\lambda)\phi_1(\lambda) + q(\lambda)\phi_2^*(\lambda) = 1.$$

Now define $\phi(\lambda) := q(\lambda)\phi_2^*(\lambda)$ and note that $\phi^*(\lambda) = q^*(\lambda)\phi_2(\lambda)$- It follows that $\phi^*(A_2) = 0$ and $\phi(A_1) = I$. From this it follows that $A_1\Delta X A_2 = \Delta X$ implies that $\Delta X = 0$. Thus $X_1 = X_2$.

To show the converse we assume that λ is an eigenvalue of A_1 and $1/\lambda$ is an eigenvalue of A_2. Hence it is also an eigenvalue of $A_2^\top$. let $A_1 x^1 = \lambda x^1$ and $A_2^\top x^2 = (1/\lambda)x^2$ where $x^1 \neq 0$ and $x^2 \neq 0$. Define $X = (x_1^2 x^1, x_2^2 x^1, ...x_n^2 x^1)$. Then $X \neq 0$ and $A_1 X A_2 = X$, a contradiction. □

Now in our case A has all eigenvalues inside the unit circle, then there is no reciprocal of an eigenvalue of A that is an eigenvalue. The above Lemma shows that there is a unique solution to $A^\top PA - P = -Q$.

The following is an immediate corollary of this proposition by taking $A = T/\rho$ in the Proposition.

Corollary. *Let $T \in \mathbb{R}^{d\times d}$. Then $\rho(T) < \rho$ if and only if there exists a $P \succ 0$ such that $T^\top PT - \rho^2 P \prec 0$.*

This Corollary provides a method to check the stability of a linear time invariant system. That is to say that the Linear Matrix Inequality characterizes the stability. It dates back to Lyapunov (see e.g. [5]). Now we study the dynamical system $\xi_{k+1} - \xi^* = T(\xi_k - \xi^*)$ as in the previous section. Then if there is a symmetric positive definite matrix P such that the LMI $T^\top PT - \rho^2 P \prec 0$ holds we get, on the trajectory

$$(\xi_{k+1} - \xi^*)^\top P(\xi_{k+1} - \xi^*) < \rho^2(\xi_k - \xi^*)^\top P(\xi_k - \xi^*).$$

If now $\rho < 1$, then the sequence $\{\xi_k\}_{k\geq 0}$ converges linearly to ξ^*. in this manner we get

$$(\xi_k - \xi^*)^\top P(\xi_k - \xi^*) < \rho^{2k}(\xi_0 - \xi^*)^\top P(\xi_0 - \xi^*).$$

The last inequality implies that

$$\|\xi_k - \xi^*\| < \sqrt{\mathrm{cond}(P)}\rho^k \|\xi_0 - \xi^*\|.$$

where $\mathrm{cond}(P)$ is the condition number of P. Note that the LMI can be solved by semidefinite programming, another important class of optimization problems is [4], that is to find a solution P such that

$$\begin{pmatrix} P & A^\top P \\ PA & P \end{pmatrix} \succ 0.$$

The function

$$V(\xi) = (\xi - \xi^*)^\top P(\xi - \xi^*)$$

is called a *Lyapunov function* for the dynamical system under consideration, i.e. it is a strictly decreasing function over all trajectories and hence certifies that

the algorithm is *stable*, that is the trajectory generated by the dynamical system converges to nominal values. The conventional method for proving stability of an electromechanical system is to show that some notion of *total energy* always decreases over time. A Lyapunov function is just a generalization of the total energy. From this discussion we see that if an objective function is quadratic it falls naturally in the quadratic Lyapunov function. However searching for a Lyapunov function is not a trivial task. When we have a non quadratic objective functions, the gradient does not form a linear system, so the problems are much more complicated. It amounts to finding a Lyapunov function that guarantees algorithmic convergence when f is not quadratic. An approach in control theory is to use integral quadratic constraints to capture features of the behavior of partially known components.

References

[1] M.S. Bazaraa and H.D. Sherali and C.M. Shetty, Nonlinear Programming: theory and algorithms, (3rd edition), John Wiley, cop., 2006.

[2] L. Lessard, and B. Recht and A. Packard, Analysis and Design of OptimizationAlgorithms via Integral Quadratic Constraints, https://arxiv.org/pdf/1408.3595.pdf

[3] Y. Nesterov, Introductory lectures on convex optimization: A basic course, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[4] Linear Matrix Inequalities in Control, http://www.st.ewi.tudelft.nl/roos/courses/WI4218/lmi052.pdf

[5] E. Sontag, Mathematical Control Theory: deterministic finite dimensional systems, (2nd edition) Sringer, 1998.

[6] G. Strang, Linear algebra and Learning From Data, Wellesley, Cambridge Press, 2019.

[7] A. Holst and V. Ufnarovski, Matrix Theory, Studentlitteratur, 2014

[8] L. Vandenberghe and S. Boyd, Convex Optimization, Cambridge University Press, 2009