Convergence and robustness guarantees for numerical optimization algorithms

Giuseppe Giorgio Colabufo
giuseppe.colabufo@polytechnique.edu

September 6, 2019

Contents

Notations

Gradient and Hessian matrix		
∇f	gradient of the function f	
H	Hessian matrix of the function f	
F_k	approximation of the Hessian matrix at a point x_k	(SUBSECTION 4.5)
B_k	approximation of the inverse Hessian matrix at a point x_k	(SUBSECTION 4.6)
Functions		
V	an (ISS / iISS / δISS) Lyapunov function	Definitions (5.3), (2.11), (2.23), (2.26)
α	a $\mathcal{K}$ or $\mathcal{K}_\infty$ function	(SUBSECTION 2.1)
β	a $\mathcal{KL}$ function	(Definition 2.5)
ρ	a comparison function	(SUBSECTION 2.1)
σ	a $\mathcal{K}$ function	(Definition 2.2)
Errors		
ε_k *or* r_k	error in the evaluation of $\nabla f(x_k)$	(SUBSECTION 4.4)
ε_H	error in the evaluation of $H(x)$	(SUBSECTION 4.4)
ζ_k	error in the inversion of the Hessian H	(SUBSECTION 4.4)
Δ_H *or* s_k	error in the evaluation of $H^{-1}(x_k)$	(SUBSECTION 4.4)
Δ_k	error in the Newton's Method step	(SUBSECTION 1.1)
Vectors and matrices		
$\|x\|$	euclidean norm of a vector $x \in \mathbb{R}^n$	
$\|A\|$	induced matrix norm for $A \in \mathbb{R}^{n\times m}$	
x^T *or* A^T	denotes the transpose of a vector x or a matrix A	
$A \otimes B$	Kronecker product between matrices A and B	(Definition 5.5)
Other notations		
$\|u\|_\infty$	ℓ_∞ norm of a bounded control $u\colon \mathbb{N} \to \mathbb{R}^n$	(Remark 2.5.3)
$u_{[k]}$	truncation of u	(Remark 2.5.7)
$x(\cdot, \xi, u)$	trajectory of system (12) with initial state $x(0) = \xi \in \mathbb{R}^n$ and input u	(Remark 2.5.3)

Summary of the work done

What I learnt. I have learnt the formalism of the comparison functions, I've come to know stability definitions and properties like input-to-state-stability, integral ISS and incremental ISS. I've looked at some examples and learnt the relationships among the different definitions of stability, as well as their characterization via Lyapunov functions. I understood the importance of having consistent and intuitive notation throughout the document.

About Lyapunov functions. I investigated the changes in gain and transient bounds after a scaling (linear or nonlinear) of a given ISS and δISS Lyapunov function.

Results on Newton's Method. Given some results on convergence of Newton's Method, I found some examples to illustrate the value of the different assumptions required to use the algorithm, implemented it in Matlab and plot some domains of attraction. I extended the results on convergence, both to the exact point and relaxing the convergence to a ball, giving sufficient conditions for the cases in which an error is made in the evaluation of the gradient or the Hessian matrix.

Results on approximated Newton's Method. Given a result of convergence for an approximated version of Newton's Method, I made explicit the error term to obtain a sufficient condition of practical convergence to a ball. Studying the Quasi-Newton Method, in particular the BFGS and the DFP, I compared these algorithm with the original one, trying to find conditions for incremental stability and input-to-state-stability. Considering that a Lyapunov function might be easier to find for a continuous time problem, I tried to state the equivalent dynamics of BFGS and DFP in a continuous time frame.

Results on stability for Newton's Method. I looked for a Lyapunov function for the generic iteration step: for the classical Lyapunov definition there exists an easy one; for the ISS case, I provided sufficient conditions for the input-to-state-stability of Newton's Method using comparison functions and investigated the class of functions that satisfy these conditions; a weak result on incremental stability (sufficient conditions assuming the Lyapunov function is the 2-norm) was obtained; two results give sufficient conditions for the incremental input-to-state-stability of Newton's Method.

Results on stability for Quasi-Newton Method. After having rewritten the updating step for the matrix B^{BFGS} in a vectorial form using Kronecker product, I studied the eigenvalues of the iteration matrices for this new linear dynamic.

In the Appendix besides the many examples, used to clarify the concepts, I wrote a scheme to switch from CT systems to DT and vice versa.

Plan

- In (SECTION 1) we provide context for the problem of guarantee convergence and robustness guarantees for numerical optimization algorithms.

- In (SECTION 2) different types of stability are presented (Lyapunov, ISS, Input-output, incremental stability ...) with the ensemble of definitions and main results.
- In (SECTION 3) we briefly discuss about the modifications on gains and transient bounds that occur if the Lyapunov function (ISS or δISS) is scaled by a linear factor or by a nonlinear $\mathcal{K}_\infty$ function.
- In (SECTION 4) Newton's method is described. After a brief overview, where the role of each assumption is emphasized, and some remarks we provide a proof for convergence and its generalization when the update is noisy. A Newton-like method (in which the inverse of the Hessian matrix is replaced by a local approximation) and Quasi-Newton methods are briefly described. Finally, we translate the method in the formalism of ISS and provide a Lyapunov function that guarantees convergence.
- (SECTION 5) makes the link between Newton's Method, Quasi-Newton Method and all the stability properties presented in (SECTION 2). Lyapunov functions and sufficient conditions are provided for input-to-state-stability, incremental-input-to-state-stability, integral-input-to-state-stability...
- (APPENDIX A) contains: some tools that might be useful for the proofs of this document (APPENDIX A.2); some extra examples of ISS and incremental ISS systems are presented in (APPENDIX A.3) and (APPENDIX A.5); a link between continuous time systems and discrete time systems sharing stability properties is made in (APPENDIX A.5).

References for the different sections

List of the main references used for each section of the document.

(SECTION 2): **Stability.** [1–6] And, in detail for each subsection:

(SUBSECTION 2.1): Khalil [1], Kellett [2]
(SUBSECTION 2.2): Khalil [1], Sontag [3]
(SUBSECTION 2.3): Sontag [3], Jiang and Wang [4], Tran et al. [5], Jiang et al. [7]
(SUBSECTION 2.4): Tran et al. [5]
(SUBSECTION 2.5): Sontag [3], Tran et al. [5]
(SUBSECTION 2.6): Sontag [3], Angeli [6], Tran et al. [8], Bayer et al. [9], Tran et al. [10]
(SUBSECTION 2.7) and (SUBSECTION 2.8): Sontag [3], Tran et al. [5], Angeli [6]

(SECTION 3): **Lyapunov functions and gains.** Jiang and Wang [4], Grüne and Kellett [11], Jiang and Wang [12]

(SECTION 4): **Newton's Method.** [13–16] And in detail for each subsection:

(SUBSECTION 4.1), (SUBSECTION 4.2) and (SUBSECTION 4.3): Luenberger and Ye [13], NMF [14], Bertsekas [15]
(SUBSECTION 4.4) and (SUBSECTION 4.5): Kelley [17] and notes by Dr Iman Shames
(SUBSECTION 4.6): Bertsekas [15], Nocedal and Wright [16]

(SECTION 5): **Newton's Method in the ISS formalism.** Jiang and Wang [12]
And in detail for each subsection:

(SUBSECTION 5.6): Magnus and Neudecker [18]
(SUBSECTION 5.7): Nocedal and Wright [16]

(APPENDIX A): **Appendix.** Tran et al. [5], Angeli [6], Li et al. [19], Sontag [20]
And in detail for each subsection:

(APPENDIX A.3): Tran et al. [5], Li et al. [19]
(APPENDIX A.5): Angeli [6], Sontag [20]

1 About the problem

We have a dynamical system $\dot{x} = f(x, u)$ that we can solve to get the trajectories $x \equiv x(t)$. These trajectories are used to build a control feedback $u = k(x)$ that will be injected into the dynamic as input for the following step. Figure (Figure 1) illustrates this cycle. We might have "disturbances" (i.e. uncertainty or noise in the calculation of u or in solving for the trajectory) and we want a guarantee that our dynamic will be stable. In other words we want to prove some robustness results under hypothesis of "small" perturbations of the input. We can observe and measure the trajectory given by the solver of the dynamical system (possibly with some noise). In this report we are interested in discrete time dynamics, that is system (12): $x_{k+1} = f(x_k, u_k)$.

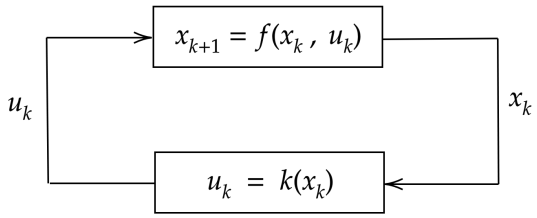


Figure 1: Feedback stabilization, closed-loop system $x_{k+1} = f(x_k, k(x_k))$.

We have essentially 3 cases:

(i) noisy feedback u: $\hat{u} = u + \eta_1(x)$;

(ii) noisy dynamic x: $\hat{x} = x + \eta_2(x)$;

(iii) noisy feedback u and dynamic x: $\hat{u} = u + \eta_1(x)$ and $\hat{x} = x + \eta_2(x)$.

Remark 1.0.1. We observe that even if x is given exactly as output of the dynamic solver, it will contain some noise due to the noisy input injected in the function f. ◀

We can model the noisy results as $\hat{u} = u + \eta(x)$ where $\eta(x)$ is a random variable with a distribution independent from the point x (and the notation simply means that we compute a realization of the random variable at any x when computing $\hat{u}$.

Remark 1.0.2. This model also include the case where the noise is multiplicative:

$$\hat{u} = u(1 + \widetilde{\eta}(x)) = u + u\widetilde{\eta}(x) = u + \eta(x).$$

◀

In order to have a chance of convergence we suppose that the noise $\eta(x)$ is bounded.

Remark 1.0.3. The noise can be stochastic or deterministic (i.e. round-off errors), but in this report we will stick to the deterministic case, that is will make no assumptions on the nature of the errors and we suppose that we don't know the explicit form of u, that is $k(x)$, but only the class of functions to which it belongs. For example, we may suppose that u is bounded in L^∞ norm. ◀

1.1 Link with the Newton's method

Iterative algorithms can be represented as dynamical systems. Indeed, consider an iterative algorithm and a dynamical system with state vector x_k and dynamics $x_{k+1} = f(x_k)$. This dynamical system represents the algorithm when the state x_k of the dynamical system is equal to the k-th iterate of the algorithm for all k. In particular we are interested in Newton's method algorithm for finding stationary points of a function f, as a discrete time dynamical system. The iteration step of the Newton's method (see (SECTION 4)) is given by

$$x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x_k)$$

where ∇f is the gradient of f and H is its Hessian matrix. This iteration step is an example of closed loop dynamics, as shown in diagram of figure (Figure 2).

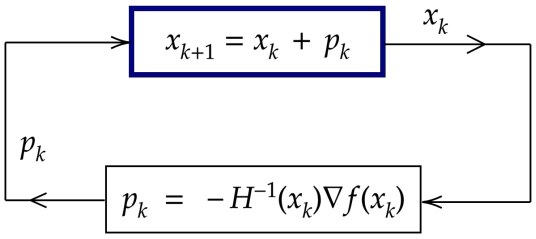


Figure 2: Closed-loop system for Newton's Method: $x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k)$.

In real world application an algorithm has finite precision so we have to add finite precision errors in the model. This could be done by considering the finite precision errors as disturbance inputs in the dynamical model. In this case there are several possibilities in which an error may occur:

i) error in evaluation: this includes the evaluation of ∇f at points x_k or evaluation of H^{-1} at point x_k. In figure (Figure 3) we call the first error ε_k and the second one ε_H. One can suppose that the evaluation error is

the same in both cases (the two error are of the same nature, so their size should be comparable). Notice that the same error appears in the evaluation of f but is not directly seen in this equation, however we should take it into account because in the evaluation of ∇f we will probably do an approximation that requires $f(x_k)$.

ii) error in the inversion of the Hessian matrix $H^{-1}(x)$ (which is noted ζ_k in figure (Figure 3));

In general both types of errors will be present. Even if we have already made the remark that the form of the noise can be always considered as additive, keep the structure of errors may actually be important to prove convergence and to give more accurate bounds on the norm of the error that one should assume to guarantee convergence. Figure (Figure 3) schematize the general situation: an error ε_k in evaluation of ∇f, an error ε_H in evaluation of H^{-1} at point x_k and an error $\zeta_k$in inverting the hessian matrix are combined together. We can estimate the error due to inversion as follows: we consider the hessian matrix $H(x)$ but we only know its noisy value $H(x)+\varepsilon_H$. When the matrix is inverted, we are actually computing $(H(x)+\varepsilon_H)^{-1}$. We can use the expansion of the geometric series to estimate this quantity:

$$\begin{aligned}(H(x)+\varepsilon_H)^{-1} &= (H(x)(I+H^{-1}(x)\varepsilon_H))^{-1}\\ &= (I+H^{-1}(x)\varepsilon_H)^{-1}H^{-1}(x)\\ &= \left(I-H^{-1}(x)\varepsilon_H+\varepsilon_H^T\left(H^{-1}(x)\right)^2\varepsilon_H-\dots\right)H^{-1}(x)\\ &\approx H^{-1}(x)-H^{-1}(x)\varepsilon_H H^{-1}(x).\end{aligned}$$

When we evaluate this quantity at x_k another error ε_k adds to it. So, the computed values are

- $\nabla f(x_k)+\varepsilon_k$ for the gradient of f at x_k;
- $H^{-1}(x_k)-H^{-1}(x)\varepsilon_H H^{-1}(x)+\zeta_k$ for the inverse of the Hessian matrix H at x_k;

these two quantities are multiplied together, so that at the end the update $p_k = x_{k+1}-x_k$ for the iteration step becomes:

$$\begin{aligned}p_k &:= -\left(H^{-1}(x_k)-H^{-1}(x)\varepsilon_H H^{-1}(x)+\zeta_k\right)\left(\nabla f(x_k)+\varepsilon_k\right)\\ p_k &= -H^{-1}(x_k)\nabla f(x_k)-H^{-1}(x_k)\varepsilon_k-\zeta_k\nabla f(x_k)\\ &\quad +H^{-1}(x)\varepsilon_H H^{-1}(x)\nabla f(x_k)+H^{-1}(x)\varepsilon_H H^{-1}(x)\varepsilon_k-\zeta_k\varepsilon_k.\end{aligned}$$

Remark 1.0.4. In order to prove the stability of this dynamics, we can arbitrarily decompose the iteration step, making x_{k+1} a more general function of x_k and an input u_k. Indeed we notice that taking $u_k = -H(x_k)^{-1}\nabla f(x_k)$ makes the

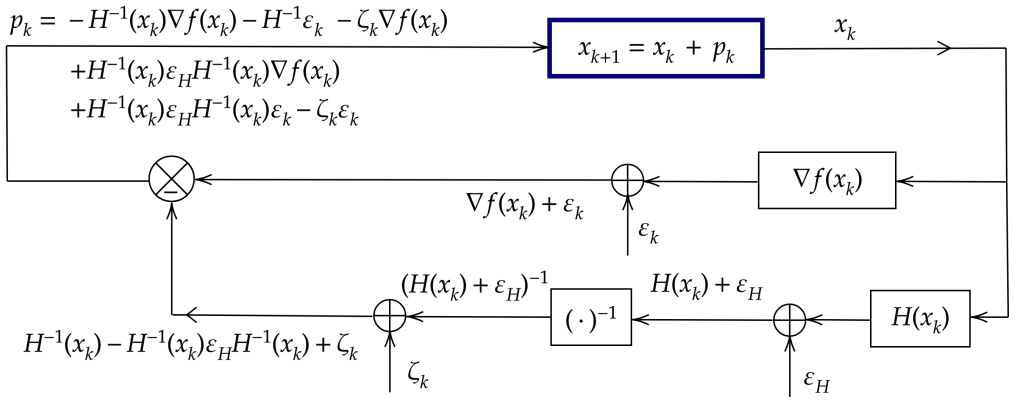


Figure 3: A scheme for the general case of noisy Newton's method.

system not to be ISS[1] (with respect to u_k). For example a possible (linear) decomposition would be $x_{k+1} = 0.9x_k + (0.1x_k - H(x_k)^{-1}\nabla f(x_k))$. In this case, the system is ISS with input $u_k = 0.1x_k - H(x_k)^{-1}\nabla f(x_k)$. ◀

Remark 1.0.5 (Two possible main questions.). In this setting we have two possible main directions to explore:

I. Assuming that there exists a unique point x^* stationary point and minimum point for a given function f, what assumptions are needed to guarantee the convergence of Newton's method to that point?

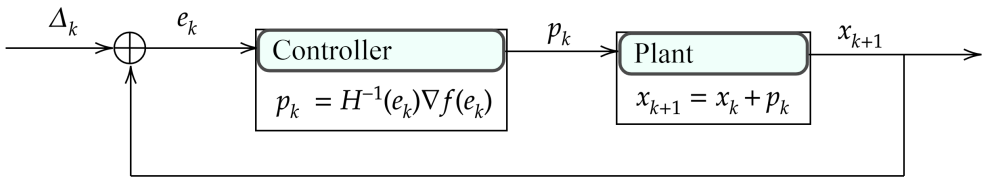
II. What other assumptions one need to make to show that such a point exists?

These two directions contains each multiple sub-questions. For example, in the first case a natural generalization would be the case of multiple stationary points, and in this case one may just be interested in convergence to any point or one in particular. ◀

Remark 1.0.6. In the literature there are many results for the convergence of Newton's method (and other iterative algorithms) to the exact solution x^* when no kind of disturbances are considered in the iteration step. An example of these results are (Theorem 4.1) or (Theorem 4.2). In (SUBSECTION 4.4) and (SUBSECTION 4.4) are presented similar results that take into account disturbances in the iteration step and still guarantee a convergence to the exact point. However, in general, for many difference equations a solution is considered a stable

[1] See (Definition 2.10).

solution if it enters and remains in a sufficiently small set. For example, under the proper conditions all solutions of the Newton's equation (28) approach the desired solution as $k \to \infty$. This is what proved for instance in (Fact 1), (Fact 2), (Proposition 4.4) and (Proposition 4.9). In some cases, if all the solutions become and remain close to the desired solution, then the method is judged to be satisfactory. This type of stability is called **practical stability**. An example of exact definition is (Definition 2.25) of incremental stability in (SUBSECTION 2.6). ◀



$$\Delta_k = -H^{-1}(x_k)\varepsilon_k - \zeta_k \nabla f(x_k) + H^{-1}(x_k)\varepsilon_H H^{-1}(x_k)\nabla f(x_k) + H^{-1}(x_k)\varepsilon_H H^{-1}(x_k)\varepsilon_k - \zeta_k \varepsilon_k.$$

Figure 4: Is it OK? Newton's Method in the basic control system scheme.

1.2 Conclusion of this section

In this first section we provided an introduction to the general scheme of dynamical system with control feedback. In the setting of discrete time systems, we highlights the way that disturbances can be introduced in the system so that noise arises when measuring the output. Motivated by the goal of guaranteeing convergence and robustness for numerical optimization algorithms, we translated Newton's Method in this framework and presented a diagram with all the errors introduced at each iteration. We concluded the section with some remarks about the different directions of study and research and a window on the following sections.

2 Stability

Various types of stability may be discussed for the solutions of differential equations or difference equations describing dynamical systems. The most important type is that concerning the stability of solutions near to a point of equilibrium. This may be discussed by the theory of Lyapunov.

2.1 Comparison functions formalism

These classes of functions are used in stability theory to characterize the stability properties of control systems.

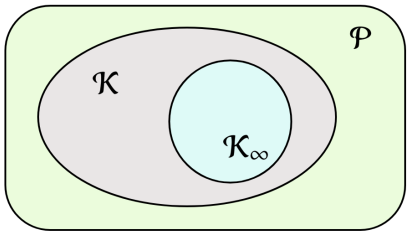


Figure 5: Subsets of comparison functions.

Definition 2.1 (Positive definite functions)**.** The class of positive definite functions is the class

$$\mathcal{P} = \{\gamma \in C(\mathbb{R}_+; \mathbb{R}_+) \mid \gamma(0) = 0 \text{ and } \forall r > 0\ \gamma(r) > 0\}.$$

Definition 2.2 (Class $\mathcal{K}$)**.** A continuous function $\gamma\colon [0, a) \to \mathbb{R}_+$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and $\gamma(0) = 0$. In other words

$$\mathcal{K} = \{\gamma \in \mathcal{P} \mid \gamma \nearrow \text{ (strictly increasing)}\}.$$

Definition 2.3 (Class $\mathcal{K}_\infty$)**.** A continuous function $\gamma\colon \mathbb{R}_+ \to \mathbb{R}_+$ is said to belong to class $\mathcal{K}_\infty$ if it belongs to class $\mathcal{K}$ with $a = \infty$ and it is unbounded. In other words

$$\mathcal{K}_\infty = \{\gamma \in \mathcal{K} \mid \lim_{r\to\infty} \gamma(r) = \infty\}.$$

Definition 2.4 (Class $\mathcal{L}$)**.** A continuous function $\gamma\colon \mathbb{R}_+ \to \mathbb{R}_+$ is said to belong to class $\mathcal{L}$ if it is strictly decreasing:

$$\mathcal{L} = \{\gamma \in C(\mathbb{R}_+; \mathbb{R}_+) \mid \gamma \searrow \text{ and } \lim_{r\to\infty} \gamma(r) = 0\}.$$

Definition 2.5 (Class $\mathcal{KL}$)**.** A continuous function $\beta\colon [0, a) \times \mathbb{R}_+ \to \mathbb{R}_+$ is said to belong to class $\mathcal{KL}$ if $\beta(\cdot, t) \in \mathcal{K}$ for each fixed t and $\beta(r, \cdot) \in \mathcal{L}$ for each fixed r:

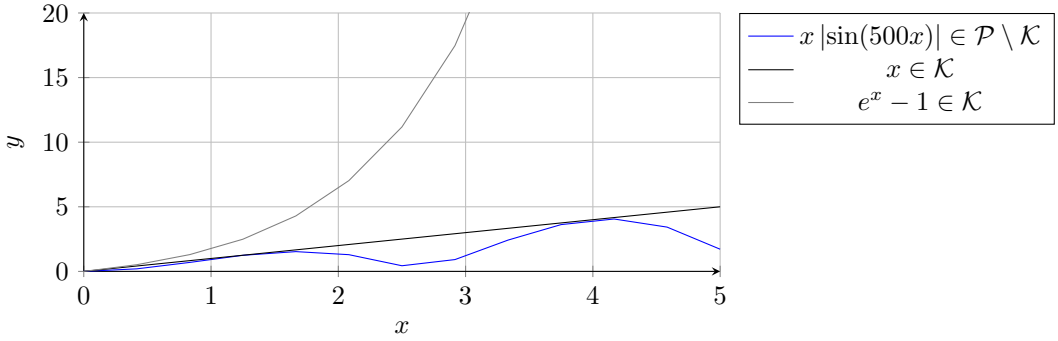
$$\mathcal{KL} = \{\beta \in C([0, a) \times \mathbb{R}_+; \mathbb{R}_+) \mid \forall t \geq 0\ \beta(\cdot, t) \in \mathcal{K} \text{ and } \forall r \in (0, a)\ \beta(r, \cdot) \in \mathcal{L}\}.$$

In other words β is strictly increasing in the first variable and decreasing to 0 in the second one and $\beta(0, 0) = 0$.

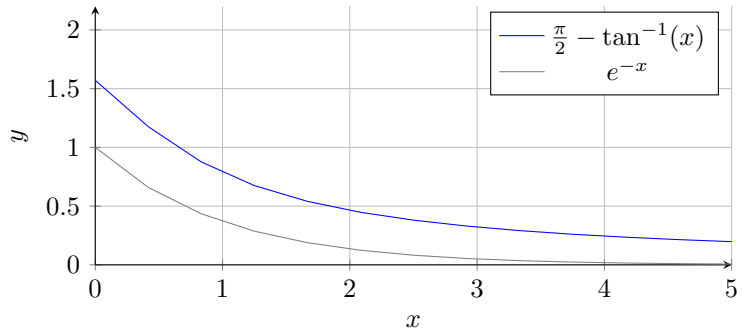
Lemma 2.1 (Comparison function properties)**.** *[1, Lemma 4.2, p. 145] Let α_1 and α_2 be class $\mathcal{K}$ functions on $[0, a)$, α_3 and α_4 be class $\mathcal{K}_\infty$ functions, and β be a class $\mathcal{KL}$ function. Then,*

- α_1^{-1} *is defined on* $[0, \alpha_1(a))$ *and belongs to class* $\mathcal{K}$*;*
- α_3^{-1} *is defined on* $[0, +\infty)$ *and belongs to class* $\mathcal{K}_\infty$*;*
- $\alpha_1 \circ \alpha_2 \in \mathcal{K}$*;*
- $\alpha_3 \circ \alpha_4 \in \mathcal{K}_\infty$*;*
- $\sigma(r, s) = \alpha_1(\beta(\alpha_2(r), s)) \in \mathcal{KL}$*.*

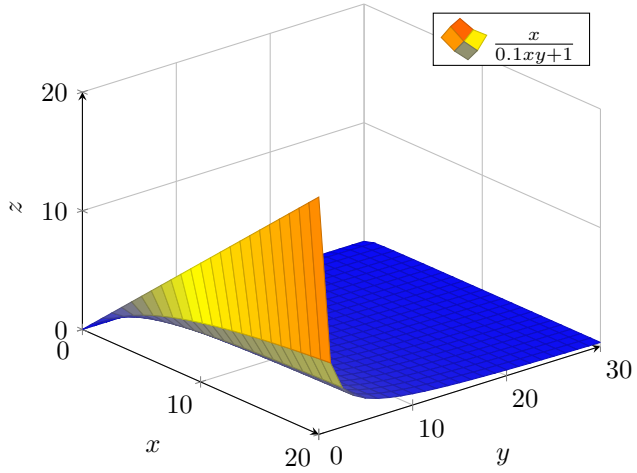
Comparison functions: examples in classes $\mathcal{P} \supset \mathcal{K}$



Comparison functions: examples in class $\mathcal{L}$



Comparison functions: an example in class $\mathcal{KL}$



Remark 2.1.1 (Other properties of comparison functions). In addition to the basic properties listed by the previous (Lemma 2.1), comparison functions have also other interesting properties:

- If $\alpha \in \mathcal{K}$ and $\sigma \in \mathcal{L}$ then $\alpha \circ \sigma \in \mathcal{L}$.
- For any $\sigma_1, \sigma_2 \in \mathcal{L}$, $\sigma_1 \circ \sigma_2 - \sigma_1(\sigma_2(0)) \in \mathcal{K}$.
- We can always bound comparison functions from above and below by smooth functions on $\mathbb{R}_+$; we may additionally control how close the smooth function is to the given function. [2, Lemmas 1-4, pp. 345-346]
- (Sontag's Lemma on $\mathcal{KL}$-Estimates) For any given $\beta \in \mathcal{KL}$ and any constant $\lambda > 0$

$$\exists \rho_1,\ \rho_2 \in \mathcal{K}_\infty \quad \beta(s,r) \leq \rho_1(\rho_2(s)e^{-\lambda r})\ \forall s, r \geq 0 \tag{1}$$

- (A lower bound for $\mathcal{KL}$ functions) [2, Lemma 19, p. 351] For any $\beta \in \mathcal{KL}$ there exist $\alpha_1 \in \mathcal{K}$, $\alpha_2 \in \mathcal{K}_\infty$ such that

$$\beta(s,t) \geq \frac{\alpha_1(s)}{1+\alpha_2(t)} \quad \forall s,t \in \mathbb{R}_+.$$

- (A triangle inequality) for any function $\alpha \in \mathcal{K}$ and any $a, b \in \mathbb{R} \geq 0$,

$$\alpha(a+b) \leq \alpha(2a) + \alpha(2b). \tag{2}$$

This is a special case of a more general inequality involving a $\mathcal{K}_\infty$ function φ (see [2, Lemma 10, p. 347]):

Lemma 2.2 (Triangle inequality for comparison functions). *Given $\alpha \in \mathcal{K}$ and any function $\varphi \in \mathcal{K}_\infty$ such that $\varphi - id \in \mathcal{K}_\infty$, then for any $a, b \geq 0$:*

$$\alpha(a+b) \leq \alpha(\varphi(a)) + \alpha(\varphi \circ (\varphi - id)^{-1}(b)). \tag{3}$$

- The integral of a class-$\mathcal{K}$ function is convex.

Proof. Let $\alpha \in \mathcal{K}$ and call $\varphi(s) = \int_0^s \alpha(\tau)d\tau$. We need to prove that $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for every $\lambda \in [0,1]$. It is just a change of variables:

$$\begin{aligned}
\varphi(\lambda x + (1-\lambda)y) &= \int_0^{\lambda x} \alpha(\tau)d\tau + \int_0^{(1-\lambda)y} \alpha(\tau)d\tau \\
&= \lambda \int_0^x \alpha(\lambda\tau)d\tau + (1-\lambda)\int_0^y \alpha(\lambda\tau)d\tau \\
&\leq \lambda \int_0^x \alpha(\tau)d\tau + (1-\lambda)\int_0^y \alpha(\tau)d\tau \\
&= \lambda\varphi(x) + (1-\lambda)\varphi(y).
\end{aligned}$$

Where we used that $\lambda \leq 1$ and the fact that α is increasing to obtain the inequality. □

- There is a *comparison principle* (or *comparison lemma*) which makes use of a scalar differential inequality to make statements about the nature of solutions to a scalar differential equation (that is give upper or lower bounds by $\mathcal{K}$ or $\mathcal{KL}$ function on them).
- Given a $\mathcal{K}$-function, it is possible to find another $\mathcal{K}$ function that upper bounds the given function away from the origin and is linear near the origin. (see [2, Lemma 26, p. 354])
- For any $\alpha \in \mathcal{K}_\infty$ there is $\hat{\alpha} \in \mathcal{K}_\infty$ satisfying:

$$\hat{\alpha}(s) \leq \alpha(s) \quad \forall s \geq 0 \qquad \text{and} \qquad id - \hat{\alpha} \in \mathcal{K}.$$

◀

2.2 Input-output stability

An input-output model relates the otput of the system directly to the input (with no knowledge of the interior structure that is represented by the state equation, so the system is like a black box).

The input-output stability is useful in studying interconnected systems: the gain allows us to track how the norm of a signal increases or decreases as it passes through the system.

We consider $y = Hu$ where $u\colon \mathbb{R}^+ \to \mathbb{R}^m$ is a (piece-wise constant / bounded / piece-wise continuous / square integrable) input. We set, for $1 \le p < \infty$,

$$\mathcal{L}_p^m := \left\{ u \text{ piece-wise continuous with } \|u\|_{\mathcal{L}} := \left(\int_0^\infty \|u(t)\|_p \, dt \right)^{(1/p)} < \infty \right\}.$$

We extend this space:

$$\mathcal{L}_e^m := \left\{ u \in \mathcal{L}^m \mid u_\tau := u(t)\mathbb{1}_{[0,\tau]}(t) \in \mathcal{L}^m \ \forall \tau \in [0,\infty) \right\}.$$

We say that H is **causal** if the value of the output at any time t depends only on vales of the input up to time t.

Definition 2.6 ($\mathcal{L}$-stable)**.** [1, Definition 5.1, p. 197] $H\colon \mathcal{L}_e^m \to \mathcal{L}_e^q$ is **$\mathcal{L}$-stable** if $\exists \alpha \in \mathcal{K}\ \ \exists \beta \ge 0$ such that

$$\|(Hu)_\tau\|_{\mathcal{L}} \le \alpha\left(\|u_\tau\|_{\mathcal{L}}\right) + \beta \qquad \forall u \in \mathcal{L}_e^m \ \forall \tau \in \mathbb{R}^+$$

Definition 2.7 (finite gain $\mathcal{L}$-stable)**.** [1, Definition 5.1, p. 197] $H\colon \mathcal{L}_e^m \to \mathcal{L}_e^q$ is **finite gain $\mathcal{L}$-stable** if $\exists \gamma, \beta \ge 0$ such that

$$\|(Hu)_\tau\|_{\mathcal{L}} \le \gamma \|u_\tau\|_{\mathcal{L}} + \beta \qquad \forall u \in \mathcal{L}_e^m \ \forall \tau \in \mathbb{R}^+$$

Remark 2.2.1. The notion of $\mathcal{L}^\infty$ stability is the same of BIBO: for every bounded input the output is bounded. ◀

Remark 2.2.2. One could give a nonlinear version of this (Definition 2.7), requiring the upper bound to be $\gamma \in \mathcal{K}_\infty$ a function of $\|u_\tau\|$. ◀

Definition 2.8 (small-signal $\mathcal{L}$-stable)**.** [1, Definition 5.2, p. 201] $H\colon \mathcal{L}_e^m \to \mathcal{L}_e^q$ is **small-signal $\mathcal{L}$-stable** if $\exists r \ge 0$ such that

$$\forall u \in \mathcal{L}_e^m \ \ \sup_{[0,\tau]} \|u(t)\| \le r \ \Rightarrow \|(Hu)_\tau\|_{\mathcal{L}} \le \alpha\left(\|u_\tau\|_{\mathcal{L}}\right) + \beta$$

Definition 2.9 (small-signal finite gain $\mathcal{L}$-stable)**.** [1, Definition 5.2, p. 201] $H\colon \mathcal{L}_e^m \to \mathcal{L}_e^q$ is **small-signal finite gain $\mathcal{L}$-stable** if $\exists r \ge 0$ such that

$$\forall u \in \mathcal{L}_e^m \ \ \sup_{[0,\tau]} \|u(t)\| \le r \ \Rightarrow \|(Hu)_\tau\|_{\mathcal{L}} \le \gamma \|u_\tau\|_{\mathcal{L}} + \beta$$

Lyapunov stability tools can be used to establish $\mathcal{L}$ stability of nonlinear system represented by state models.

2.2.1 Summary of results

We consider the system

$$\begin{cases} \dot{x} = f(t,x,u) \quad x(0) = x_0 \\ y = h(t,x,u) \end{cases} \tag{4}$$

where f is piece-wise continuous in t and locally Lipschitz in (x,u) and h is piece-wise continuous in t and continuous in (x,u). Then the following theorems hold.

Theorem 2.3 (Small-signal finite-gain $\mathcal{L}_p$ stable)**.** *[1, Theorem 5.1, p. 202] Consider the system (4) and take $r > 0$ and $r_u > 0$ such that $\{\|x\| \leq r\} \subset D \subset \mathbb{R}^n$ the x-domain for f and h and $\{\|u\| \leq r_u\} \subset D_u \subset \mathbb{R}^m$ the u-domain for f and h. Suppose that*

- *$x = 0$ is an exponentially stable equilibrium point for $\dot{x} = f(t,x,0)$ with Lyapunov function V that satisfies*

$$c_1 \|x\|^2 \leq V(t,x) \leq c_2 \|x\|^2 \tag{5a}$$
$$\partial_t V + \partial_x V f(t,x,0) \leq -c_3 \|x\|^2 \tag{5b}$$
$$\|\partial_x V\| \leq c_4 \|x\| \tag{5c}$$

  for some positive constants c_i;

- *f L-Lipschitz in u:*

$$\|f(t,x,u) - f(t,x,0)\| \leq L \|u\|$$

- *h satisfies inequality*

$$\|h(t,x,u)\| \leq \eta_1 \|x\| + \eta_2 \|u\| \tag{6}$$

then the system (4) is small-signal finite-gain $\mathcal{L}_p$ stable for each $\|x_0\| \leq r\sqrt{c_1/c_2}$.

Remark 2.3.1. We can precise the result a bit:

- it holds for each $p \in [1, \infty]$;
- it holds locally or globally (depending on where the assumptions hold) if $D = \mathbb{R}^n$ and $D_u = \mathbb{R}^m$;
- we have an explicit form for the constants γ and β:

$$\gamma = \eta_2 + \frac{\eta_1 c_2 c_4 L}{c_1 c_3} \qquad \beta = \eta_1 \|x_0\| \sqrt{\frac{c_2}{c_1}} \rho \quad \text{where} \quad \rho = \begin{cases} 1 & p = \infty \\ \left(\frac{2c_2}{c_3 p}\right)^{1/p} & p \in [1, \infty) \end{cases}.$$

- As a corollary we can apply the theorem if f is continuously differentiable in a neighborhood of $(x = 0, u = 0)$ with $\partial_x f$ and $\partial_u f$ uniformly bounded in t.
- As a corollary, the linear time-invariant system

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx + Du \end{cases}$$

is finite-gain $\mathcal{L}_p$ stable if A is Hurwitz.

◀

Theorem 2.4 (small-signal $\mathcal{L}_\infty$ stable)**.** *[1, Theorem 5.2, p. 206] Consider the system (4) and take $r > 0$ and $r_u > 0$ such that $\{\|x\| \leq r\} \subset D \subset \mathbb{R}^n$ the x-domain for f and h. Suppose that*

- *$x = 0$ is an uniformly asymptotically stable equilibrium point of $\dot{x} = f(t, x, 0)$ with a Lyapunov function V that satisfies*

$$\alpha_1(\|x\|) \leq V(t, x) \leq \alpha_2(\|x\|) \tag{7a}$$
$$\partial_t V + \partial_x V f(t, x, 0) \leq -\alpha_3(\|x\|) \tag{7b}$$
$$\|\partial_x V\| \leq \alpha_4(\|x\|) \tag{7c}$$

  for some $\mathcal{K}$ functions α_i;

- *f and h satisfy inequalities*

$$\|f(t, x, u) - f(t, x, 0)\| \leq \alpha_5(\|u\|) \tag{8a}$$
$$\|h(t, x, u)\| \leq \alpha_6(\|x\|) + \alpha_7(\|u\|) + \eta \tag{8b}$$

  for some $\mathcal{K}$ functions $\alpha_5, \alpha_6, \alpha_7$ and a constant $\eta \geq 0$.

then the system is small-signal $\mathcal{L}_\infty$ stable for each $\|x_0\| \leq \alpha_2^{-1}(\alpha_1(r))$.

Remark 2.4.1. In this case we cannot generalize to obtain a global result. ◀

Theorem 2.5 ($\mathcal{L}_\infty$ stable)**.** *[1, Theorem 5.3, p. 208] Consider the system (4) with $D = \mathbb{R}^n$ and $D_u = \mathbb{R}^m$ and suppose that $\dot{x} = f(t, x, u)$ is ISS stable and*

$$\|h(t, x, u)\| \leq \alpha_1(\|x\|) + \alpha_2(\|u\|) + \eta \tag{9}$$

for some $\mathcal{K}$ functions α_1, α_2 and a constant $\eta \geq 0$. Then, for each $x_0 \in \mathbb{R}^n$ the system is $\mathcal{L}_\infty$ stable.

2.2.2 Calculation of $\mathcal{L}_2$ gain

- [1, Theorem 5.4, p. 210] gives an explicit form for a $\mathcal{L}_2$ gain of a linear time-independent system

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx + Du \end{cases}$$

  which is given by $\sup_{\omega \in \mathbb{R}} \|G(i\omega)\|$ where $G(s) = C(sI - A)^{-1}B + D$.

- [1, Theorem 5.5, p. 211] If we consider the system

$$\begin{cases} \dot{x} = f(x) + G(x)u \quad x(0) = x_0 \\ y = h(x) \end{cases} \tag{10}$$

  where $f(0) = 0 = h(0)$ and f locally Lipschitz, h, G continuous. If V satisfies the Hamilton-Jacobi inequality

$$\partial_x V f(x) + \frac{1}{2\gamma^2} \partial_x V G(x) G(x)^T (\partial_x V)^T + \frac{1}{2} h(x)^T h(x) \leq 0 \tag{11}$$

  with $\gamma > 0$ then the system is finite-gain $\mathcal{L}_2$ stable with $\mathcal{L}_2$ gain $\leq \gamma$.

  - the theorem hold globally or locally as long as that the solution $x(t)$ of system (10) remains in the same domain;
  - we can use asymptotic stability of the origin of $\dot{x} = f(x)$ when $\|x_0\|$ and $\sup_{[0,\tau]} \|u(t)\|$ are sufficiently small to get the same result (see [1, Lemma 5.1, p. 215]).
  - we can check the asymptotic stability of the origin with linearization or by a Lyapunov function. Under certain conditions we can use V satisfying Hamilton-Jacobi inequality as a Lyapunov function for showing asymptotic stability.

- [1, Theorem 5.6, p. 218] If we have also that $f \in C^1$ and no solution of $\dot{x} = f(x)$ identically stays in $\{h(x) = 0\}$ other than $x \equiv 0$ then the origin is asymptotically stable and $\exists k \ \|x_0\| \leq k \Rightarrow$ the system is finite-gain $\mathcal{L}_2$ stable with $\mathcal{L}_2$ gain $\leq \gamma$.

  - this is the same result of the previous theorem but the asymptotic stability is a consequence of stronger hypothesis on the solutions.

Remark 2.5.1 (Hamilton-Jacobi in the linear case.). Suppose that the system (10) is linear:

$$\begin{cases} \dot{x} = Ax + Bu \\ y = Cx \end{cases}$$

and call $V(x) = \frac{1}{2}x^T Px$. Then the Hamilton-Jacobi equation becomes

$$x^T PAx + \frac{1}{2\gamma^2}x^T PB^T BPx + \frac{1}{2}x^T C^T Cx = 0$$

which lead to the Riccati equation

$$A^T P + PA + \frac{1}{\gamma^2}PB^T BP + C^T C = 0$$

for the positive definite matrix P. ◀

Remark 2.5.2 (Fundamental Relationship Among ISS, IOS, and IOSS). Input-output stability combined with input-output-to-state stability is equivalent input-to-state stability. ◀

2.3 Input-to-state stability

The notion of input to state stability (ISS) (Definition 2.10) plays a central role in nonlinear systems. In particular this notion has many good properties, such as: that the states are bounded for bounded inputs, and they tend to the equilibrium of the systems when the inputs tend to zero.

- ISS applies Lyapunov notions to systems with inputs.
- The property concerns with the continuity of state trajectories on the initial state and input.
- Every state trajectory corresponding to a bounded control remains bounded (the trajectory becomes small if the input signal is small no matter what the initial state is).
- ISS employed the stability analysis and control synthesis of nonlinear systems with complex structure.
- the discrete time system can be rendered ISS iff it is globally stabilizable via state-feedback.
- ISS is particularly useful as a tool for the robust stability analysis of nonlinear system and interconnected systems.

Remark 2.5.3 (Notations.). [2] If $x \in \mathbb{R}^n$ then $\|x\|$ is its euclidean norm. For a matrix $A \in \mathbb{R}^{n\times m}$ $\|A\|$ stands for its induced matrix norm. The same notation for the bounded controls $u\colon \mathbb{N} \to \mathbb{R}^n$ stands for their ℓ_∞ norm. For each $\xi \in \mathbb{R}^n$ and each input u, $x(\cdot, \xi, u)$ denotes the trajectory of system (12) with initial state $x(0) = \xi$ and input u. ◀

[2] I used essentially the same notations used in [4] but for the euclidean norm.

We consider the system

$$x_{k+1} = f(x_k, u_k). \tag{12}$$

Remark 2.5.4. The system (12) is an autonomous system since f does not explicitly depend on time k. One can always turn a system into autonomous: given a non-autonomous system $\dot{x}(t) = f(x,t)$, one can introduce a new vector function $X(t) = (x(t), t)$ which satisfies the autonomous system $\dot{X}(t) = g(X) := (f(x), 1)$. Doing this, the dimension of the system goes up. However, while autonomous systems are often easier to understand by analysis of their equilibria, the new equivalent system $\dot{X}(t) = g(X)$ has no equilibria (because g never vanishes). ◀

Definition 2.10 (ISS)**.** The system (12) is (globally) **ISS** if $\exists \beta \in \mathcal{KL}$ $\exists \gamma \in \mathcal{K}$ such that $\forall u \in l_\infty^m$ $\forall \xi \in \mathbb{R}^n$

$$\|x(k, \xi, u)\| \leq \beta(\|\xi\|, k) + \gamma(\|u\|) \quad \forall k \in \mathbb{N}$$

The function β in this definition describes the decaying effect of the initial condition ξ, while the function γ describes the effect of the input signal u.

Remark 2.5.5. ISS implies 0-GAS and Converging Input Converging Output (bounded input $\Rightarrow$ trajectories in a ball).
However, 0-GAS + CICO don't imply ISS (see counterexample below). ◀

Example 2.1: ISS is stronger than 0-GAS + CICO. Consider the system

$$x_{k+1} = \frac{1}{2}(1 + \sin(u_k))x_k$$

which is clearly 0-GAS and CICO simply using the definitions. It is not ISS because when taking input $u_k \equiv \frac{\pi}{2}$ the dynamic is just $x_{k+1} = x_k$ so that the trajectories are constant and cannot go to 0 as $k \to \infty$). ◁

Remark 2.5.6. [3] Since, in general, $\max\{a, b\} \leq a + b \leq \max\{2a, 2b\}$, one can restate the ISS condition in a slightly different manner, namely, asking for the existence of some $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}_\infty$ (in general different from the ones in the ISS definition) such that

$$\|x(t)\| \leq \max\{\beta(\|x_0\|, t), \gamma(\|u\|_\infty)\}$$

holds for all solutions. ◀

Remark 2.5.7. The truncation of u is defined as

$$u_{[k]}(j) = \begin{cases} u_j & j \le k \\ 0 & j > k \end{cases}.$$

By causality, the same definition of ISS would result if one would replace $\|u\|$ by the norm of the truncation $\|u_{[k]}\|$. ◀

Lemma 2.6 (ISS characterization)**.** *System (12) is ISS if and only if there exists α, η, $\sigma \in \mathcal{K}_\infty$ such that*

$$\sum_{j=0}^{k} \alpha(\|x_j\|) \le \eta(\|x_0\|) + \sum_{j=0}^{k-1} \sigma(\|u_j\|) \quad \forall x_0 \in \mathbb{R}^n \; \forall k \ge 0. \tag{13}$$

Definition 2.11 (ISS-Lyapunov function)**.** $V\colon \mathbb{R}^n \to \mathbb{R}^+$ is a **ISS-Lyapunov function** for system (12) if it is continuous and

$$\exists \alpha_1,\ \alpha_2 \in \mathcal{K}_\infty \quad \alpha_1(\|\xi\|) \le V(\xi) \le \alpha_2(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \tag{14a}$$

$$\exists \alpha_3 \in \mathcal{K}_\infty \;\; \exists \sigma \in \mathcal{K} \quad V(f(\xi,\mu)) - V(\xi) \le \sigma(\|\mu\|) - \alpha_3(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \; \forall \mu \in \mathbb{R}^m \tag{14b}$$

Remark 2.6.1. It is obvious by the (Definition 2.11) that, if $\lambda > 0$ is a constant and V an ISS-Lyapunov function, then λV is still an ISS-Lyapunov function, and all the comparison functions used for bounds are just multiplied by the same constant λ. ◀

Remark 2.6.2. Observe that if V is an ISS-Lyapunov function for (12), then V is a Lyapunov function for the 0-input system $x_{k+1} = f(x_k, 0)$. Indeed the first condition (14a) gives $V(0) = 0$ and $V(x) > 0$ for $x \neq 0$, while the second condition (14b) gives $V(f(x,0)) - V(x) < 0$. ◀

Remark 2.6.3. The first inequality in (14a) states that V is positive definite and radially unbounded. The second property (14b) of (Definition 2.11) is equivalent to

$$\exists \alpha_4 \in \mathcal{K}_\infty \; \exists \chi \in \mathcal{K} \quad \|\xi\| \ge \chi(\|\mu\|) \Rightarrow V(f(\xi,\mu)) - V(\xi) \le -\alpha_4(\|\xi\|). \tag{15}$$

◀

Proof. Clearly (14b) ⇒ (15):

$$\begin{aligned} V(f(\xi,\mu)) - V(\xi) &\le \sigma(\|\mu\|) - \alpha_3(\|\xi\|) \\ &\le -\frac{1}{2}\alpha_3(\|\xi\|) - \frac{1}{2}\alpha_3(\|\xi\|) + \sigma(\|\mu\|) \\ &\le \underbrace{-\frac{1}{2}\alpha_3(\|\xi\|)}_{=:\alpha_4(\|\xi\|)} \end{aligned}$$

whenever $-\frac{1}{2}\alpha_3(\|\xi\|) + \sigma(\|\mu\|) \leq 0$ that is $\|\xi\| \geq \alpha_3^{-1}(2\sigma(\|\mu\|)) =: \chi(\|\mu\|)$. For the implication (15) $\Rightarrow$ (14b) one can take $\alpha_3 = \alpha_4$ and consider

$$\tilde{\sigma}(r) := \max\{V(f(\xi,\mu)) - V(\xi) + \alpha_4(\chi(\|\mu\|)) \mid \|\mu\| \leq r \ \ \|\xi\| \leq \chi(r)\}.$$

Then define $\sigma(r) := \max\{0, \tilde{\sigma}(r)\}$. We can assume $\sigma \in \mathcal{K}$ (otherwise we can always majorize it by a $\mathcal{K}$ function) and show that it verifies

$$\sigma(r) \geq \sup_{\|\mu\|=r} V(f(\xi,\mu)) - V(\xi) + \alpha_4(\chi(\|\mu\|))$$

Indeed there are two cases: when $\|\xi\| \geq \chi(\|\mu\|)$ then the RHS is non-positive and the LHS $\sigma(r)$ is non-negative by definition. When $\|\xi\| \leq \chi(\|\mu\|)$ it follows from the definition of $\tilde{\sigma}(r) \leq \sigma(r)$. Then the inequality (14b)

$$V(f(\xi,\mu)) - V(\xi) \leq \sigma(\|\mu\|) - \alpha_3(\|\xi\|)$$

yields. □

Example 2.2: An example of ISS-Lyapunov function. We can illustrate the notion of ISS-Lyapunov function in the case of a linear discrete system

$$x_{k+1} = Ax_k + Bu_k$$

where the matrix A has all its eigenvalues strictly inside the unit disk. We can then choose two constants $c > 0$ and $0 \leq \sigma < 1$ such that $\left\|A^k\right\| \leq c\sigma^k$. Since the system is linear, we know the exact form of the trajectories:

$$x_{k+1} = A^{k+1}x_0 + \sum_{j=0}^{k} A^{k-j}Bu_j$$

and the ISS property as defined in (Definition 2.10) follows with

$$\beta(r,k) = c\sigma^k r \qquad \gamma(r) = \sum_{j=0}^{\infty} c\sigma^j \|B\| r = \frac{c\|B\| r}{1-\sigma}$$

(it is obvious that $\beta \in \mathcal{KL}$ as it is linear in r and exponentially decreasing to 0 in k and $\gamma \in \mathcal{K}$ because it is affine and the coefficient is positive since $c > 0$ and $1 - \sigma > 0$). One can show that this linear system has a quadratic ISS-Lyapunov function given by $V(x) = x^T Px$ where $P > 0$ is the unique solution to the matrix equation

$$A^T PA - P = -Q$$

for Q a symmetric positive-definite matrix. Clearly the first property of (Definition 2.11) is satisfied with:

$$\alpha_1(\|x\|) := \lambda_{min}(P)\|x\|^2 \leq V(x) \leq \lambda_{max}(P)\|x\|^2 =: \alpha_2(\|x\|).$$

From a direct computation follows

$$V(x_{k+1}) - V(x_k) \leq \underbrace{-\frac{1}{2}\lambda_{min}(Q)\left\|x_k\right\|^2}_{=:\,-\alpha_3(\|x_k\|)} + \underbrace{\left(\frac{2\left\|A^TPB\right\|^2}{\lambda_{min}(Q)} + \left\|B^TPB\right\|^2\right)\left\|u_k\right\|^2}_{=:\,\sigma(\|u_k\|)}$$

that is the second requirement of (Definition 2.11). Therefore the quadratic function $V(x) = x^TPx$ is an ISS-Lyapunov function for the linear system $x_{k+1} = Ax_k + Bu_k$. ◁

Definition 2.12 ($\mathcal{K}$-asymptotic gain)**.** The system $x_{k+1} = f(x_k, u_k)$ has a **$\mathcal{K}$-asymptotic gain** if $\exists\gamma_a \in \mathcal{K}$ such that

$$\limsup_{k\to\infty} \|x(k,\xi,u)\| \leq \gamma_a(\limsup_{k\to\infty} \|u_k\|) \quad \forall\xi \in \mathbb{R}^n.$$

Definition 2.13 (LIM)**.** The system (12) $x_{k+1} = f(x_k, u_k)$ satisfies the limit property (LIM) if $\exists\theta \in \mathcal{K}_\infty$ such that

$$\inf_{k\geq 0} \|x_k\| \leq \theta(\|u\|_\infty) \quad \forall\xi \in \mathbb{R}^n.$$

Definition 2.14 (UBIBS)**.** The system (12) $x_{k+1} = f(x_k, u_k)$ is uniformly bounded input bounded state (**UBIBS**) if bounded initial states and controls produce uniformly bounded trajectories:

$$\exists\sigma_1,\ \sigma_2 \in \mathcal{K} \quad \forall\xi \in \mathbb{R}^n\ \forall u \in l_\infty^m \quad \sup_k \|x(k,\xi,u)\| \leq \max\{\sigma_1(\|\xi\|), \sigma_2(\|u\|)\}$$

Definition 2.15 (Robustly stable)**.** The system (12) $x_{k+1} = f(x_k, u_k)$ is **robustly stable** if

$$\exists\rho \in \mathcal{K}_\infty \quad x_{k+1} = f(x_k, d_k\rho(\|x_k\|)) =: g(x_k, d_k) \text{ is UGAS.}^3$$

Definition 2.16 (continuously stabilizable)**.** The system (12) $x_{k+1} = f(x_k, u_k)$ is **continuously stabilizable** if $\exists w\colon \mathbb{R}^n \to \mathbb{R}^m$ continuous, $w(0) = 0$ such that under control $u = w(x)$ the system $x_{k+1} = f(x_k, w(x_k))$ is GAS.

Definition 2.17 (continuously ISS stabilizable)**.** The system (12) $x_{k+1} = f(x_k, u_k)$ is **continuousy ISS stabilizable** if $\exists w\colon \mathbb{R}^n \to \mathbb{R}^m$ continuous, $w(0) = 0$ and $\exists\,\Gamma$ $n \times n$ matrix of continuous functions invertible such that under $u = w(x) + \Gamma(x)v$, the system $x_{k+1} = f(x_k, w(x_k) + \Gamma(x_k) + v_k)$ is ISS.

[3] Uniformly Globally Asymptotically Stable.

2.3.1 Summary of results.

Theorem 2.7 (Equivalent formulations of ISS property.)**.** *Consider the system (12). Then the following equivalences yield: ISS* $\iff$ *UBIBS + admits a* $\mathcal{K}$*-asymptotic gain* $\iff$ *robustly stable* $\iff$ *smooth ISS-Lyapunov function.*

Sketch of the proof. The proof consists in the following implications steps:

- ISS-Lyapunov function $\Rightarrow$ ISS;
- ISS $\Rightarrow$ UBIBS + admits a $\mathcal{K}$-asymptotic gain;
- GAS $\iff$ UGAS $\iff$ smooth ISS-Lyapunov function;
- robustly stable $\Rightarrow$ smooth ISS-Lyapunov function;
- UBIBS + admits a $\mathcal{K}$-asymptotic gain $\Rightarrow$ robustly stable.

$\diamondsuit$

Theorem 2.8 (Explicit the gain from a Lyapunov function)**.** *We can obtain an ISS gain function from an ISS-Lyapunov function* V *(if the bounds for* V *are explicit). In particular the* $\mathcal{K}$*-asymptotic gain would be* $\gamma_a(s) := \alpha_1^{-1} \circ \alpha_3^{-1} \circ \sigma(s)$ *for an ISS-Lyapunov function* V *as in (Definition 2.11).*

Theorem 2.9 (Explicit the estimate from a Lyapunov function)**.** *We can obtain an explicit estimate (13) for ISS characterization of (Lemma 2.6) gain function from an ISS-Lyapunov function* V *(if the bounds for* V *are explicit). In particular the functions would be* $\sigma(s) := \sigma(s)$*,* $\eta(s) := \alpha_2(s)$ *and* $\alpha(s) := \tilde{\alpha} \circ \alpha_1(s)$ *where* $\tilde{\alpha}(s) := \min\{s, \alpha_3 \circ \alpha_2^{-1}(s)\}$ *for an ISS-Lyapunov function* V *as in (Definition 2.11).*

Remark 2.9.1. Using (Theorem 2.8) and (Theorem 2.9), together with (Remark 2.6.1), we can see how the asymptotic gain and the transient bounds change when scaling an ISS-Lyapunov function V. It is easy to see that if $\hat{\alpha}(s) = \lambda\alpha(s)$ then $\hat{\alpha}^{-1}(s) = \alpha^{-1}\left(\frac{s}{\lambda}\right)$, so that the bounds given by the preceding theorems become:

$$\begin{aligned}
\hat{\gamma}_a(s) &= \hat{\alpha}_1^{-1} \circ \hat{\alpha}_3^{-1} \circ \hat{\sigma}(s) = \alpha_1^{-1} \circ \frac{1}{\lambda}\alpha_3^{-1} \circ \sigma(s) \\
\hat{\alpha}(s) &= \hat{\bar{\alpha}} \circ \hat{\alpha}_1(s) = \lambda\bar{\alpha} \circ \alpha_1(s) \\
\hat{\eta}(s) &= \hat{\alpha}_2(s) = \lambda\alpha_2(s) \\
\hat{\sigma}(s) &= \lambda\sigma(s).
\end{aligned}$$

Notice that the only nonlinear change appears in the $\mathcal{K}$-asymptotic gain $\hat{\gamma}_a$. ◀

Corollary 2.10 (ISS and l_p gain)**.** *If in the previous (Theorem 2.9) there exist two constant c_α and c_σ so that $\alpha(s) \geq c_\alpha s^p$ and $\sigma(s) \leq c_\sigma s^p$ then the system (12) satisfies the linear l_p gain property of (Definition 2.20) with transient and gain bounds $\kappa(s) = \frac{1}{c_\alpha}\eta(s)$ and $\lambda = \left(\frac{c_\sigma}{c_\alpha}\right)^{1/p}$.*

Other results.

- if two subsystems are ISS then the whole interconnected system is ISS (Theorem 2.11);
- if we have ISS-Lyapunov function for two subsystem, then the whole interconnected system is ISS (Theorem 2.12);
- system (12) is ISS-stabilizable $\iff$ continuously stabilizable.

Theorem 2.11 (ISS for interconnected systems)**.** *Consider the interconnected and nonlinear discrete-time systems*

$$\begin{cases} x_1(k+1) = f_1(x_1(k), v_1(k), u_1(k)) \\ x_2(k+1) = f_2(x_2(k), v_2(k), u_2(k)) \end{cases} \tag{16}$$

subject to the interconnection constraints

$$v_1(k) = x_2(k) \qquad v_2(k) = x_1(k). \tag{17}$$

Suppose that both subsystems in (16) are ISS:

$$\begin{aligned} \|x_1(k, \xi, v_1, u_1)\| &\leq \max\{\beta_1(\|\xi_1\|, k), \gamma_1^x(\|v_1\|), \gamma_1^u(\|u_1\|)\} \\ \|x_2(k, \xi, v_2, u_2)\| &\leq \max\{\beta_2(\|\xi_2\|, k), \gamma_2^x(\|v_2\|), \gamma_2^u(\|u_2\|)\} \end{aligned}$$

and $\gamma_1^x \circ \gamma_2^x(s) < s$ for all $s > 0$. Then the interconnected system (16) and (17) is ISS with input (u_1, u_2).

Remark 2.11.1. One would like to apply this theorem to the sub-dynamics of Newton's method (for x_k and for s_k), however, as already remarked in (1.0.4) we need a smarter decomposition since the dynamic $x_{k+1} = x_k$ is unstable and therefor not ISS. ◀

Theorem 2.12 (ISS-Lyapunov for interconnected systems)**.** *Suppose that both subsystems in (16) admit ISS-Lyapunov functions V_1 and V_2 respectively that satisfy*

$$V_i(f_i(\xi_i, v_i, \mu_i)) - V_i(\xi_i) \leq -\sigma_i(V_i(\xi_i)) + \rho_i^x(V_j(v_i)) + \rho_i^u(\|\mu_i\|) \quad i \neq j \in \{1, 2\} \tag{18}$$

with $id - \sigma_i \in \mathcal{K}$ for $i = 1, 2$. If there exists $\rho \in \mathcal{K}_\infty$ such that

$$\sigma_1^{-1} \circ (id + \rho) \circ \rho_1^x \circ \sigma_2^{-1} \circ (id + \rho) \circ \rho_2^x < id\,,$$

then the interconnected system (16) and (17) is ISS with input (u_1, u_2).

ISS concept was modified in many directions, providing different relationships between input, output, and state. Some of the different concepts derived from ISS will be explored in next sections ((SUBSECTION 2.2), integral ISS (iISS) (SUBSECTION 2.5), incremental ISS (SUBSECTION 2.6)) but others exist (Input-to-Output Stability (IOS), Input-Output-to-State Stability (IOSS)).

2.4 l_p stability

Most of this section comes from [5].

There exist many other forms of stability for discrete time systems. Here we list a series of definition that were proved to be qualitatively equivalent to ISS (see (SUBSECTION 2.3) and (Definition 2.10)) or iISS (see (SUBSECTION 2.5) and (Definition 2.22)).

We still consider the system (12): $x_{k+1} = f(x_k, u_k)$.

Definition 2.18 (α-summable)**.** System (12) is 0-input α-summable if $\exists \alpha,\, \eta \in \mathcal{K}_\infty$ such that for input $u \equiv 0$

$$\sum_{j=0}^{k} \alpha(\|x_j\|) \leq \eta(\|x_0\|) \quad \forall x_0 \in \mathbb{R}^n \; \forall k \geq 0.$$

Remark 2.12.1. It can be proved that ([5, Theorem 1, p.362]) for a system with no input $x_{k+1} = f(x_k)$ then the origin is α-summable if and only if it is GAS. ◀

Definition 2.19 (l_p-stable)**.** For a fixed $p > 1$, system (12) is 0-input l_p-stable if $\exists \kappa \in \mathcal{K}_\infty$ such that for input $u \equiv 0$

$$\|x\|^p_{l_p[0,k]} \leq \kappa(\|x_0\|) \quad \forall x_0 \in \mathbb{R}^n \; \forall k \geq 0.$$

Example 2.3: A 0-input l_2-stable system. Consider the scalar system

$$x_{k+1} = f(x_k) := \frac{1}{2} x_k + \frac{3}{2} x_k^2 u_k.$$

For $u_k \equiv 0$ the solution is simply given by

$$x_k = \frac{x_0}{2^k},$$

therefore

$$\|x\|^2_{l_2[0,k]} = \sum_{j=0}^{k} \left| \frac{x_0}{2^j} \right|^2 \leq 2x_0^2$$

so that the system is 0-input l_2-stable. ◁

Definition 2.20 (linear l_p gain)**.** For a fixed $p > 1$, system (12) has linear l_p gain property with transient bound $\kappa \in \mathcal{K}_\infty$ sand gain bound $\gamma \geq 0$ if

$$\|x\|^p_{l_p[0,k]} \leq \kappa(\|x_0\|) + \gamma^p \|u\|^p_{l_p[0,k-1]} \quad \forall x_0 \in \mathbb{R}^n \; \forall k \geq 0.$$

Definition 2.21 (nonlinear l_p gain)**.** For a fixed $p > 1$, system (12) has nonlinear l_p gain property with transient bound $\kappa \in \mathcal{K}_\infty$ sand gain bound $\mu \in \mathcal{K}_\infty$ if

$$\|x\|^p_{l_p[0,k]} \leq \kappa(\|x_0\|) + \mu\left(\|u\|^p_{l_p[0,k-1]}\right) \quad \forall x_0 \in \mathbb{R}^n \; \forall k \geq 0.$$

Example 2.4: A system without l_2 gain. The system in the previous (Example 2.3) does not satisfy the nonlinear l_2 property. Indeed, consider the input $u_k = 2^{-k}$, which has finite l_2 norm: $\|u\|^2_{l_2} = \sum_{k\geq 0} \frac{1}{2^2 k} < \sum_{k\geq 0} \frac{1}{2^k} = 2$. The solution of the system is then given by $x_k = 2^k$ And taking the norm, for every κ, $\mu \in \mathcal{K}_\infty$, there exist a time k such that

$$\|x\|^2_{l_2[0,k]} > \kappa(1) + \mu(2) \geq \kappa(|x_0|) + \mu(\|u\|^2_{l_2[0,k-1]}).$$

This means that the system does not satisfy the nonlinear l_2 property. ◁

Theorem 2.13 (l_p stable and α-summable)**.** *[5, Theorem 10, p. 363] For any fixed $p \geq 1$, if system $x_{k+1} = f(x_k)$ is l_p-stable then it is α-summable. Conversely, if system $x_{k+1} = f(x_k)$ is α-summable then there exists a change of coordinates such that the system in the new coordinates is l_p-stable.*

Example 2.5: l_2 stable and α-summable system. The system

$$x_{k+1} = \frac{x_k}{\sqrt{x_k^2 + 1}} \qquad \text{with } x_0 = \xi \in \mathbb{R}^n$$

has explicit solution

$$x(k, \xi) = \frac{\xi}{\sqrt{k\xi^2 + 1}} \quad \forall k \geq 0.$$

Notice that this implies that the origin is GAS. For the Lyapunov function $V(x) = |x|^2$ and $x \neq 0$ yields

$$V(x_{k+1}) - V(x_k) = -\frac{x_k^4}{x_k^2 + 1} < 0$$

thus, defining $\alpha(s) = \frac{s^4}{s^2+1} \leq s^2 = V(s)$ for all $s \in \mathbb{R}_+$. Summing along any trajectory:

$$\begin{aligned}\sum_{j=0}^{k-1} V(x_{j+1}) - V(x_j) &= |x_k|^2 - |x_0|^2 \\ &= -\sum_{j=0}^{k-1} \alpha(|x_j|)\end{aligned}$$

that gives

$$\sum_{j=0}^{k} \alpha(|x_j|) \leq |x_k|^2 + \sum_{j=0}^{k-1} \alpha(|x_j|) = |x_0|^2 .$$

Thus, the system $x_{k+1} = \frac{x_k}{\sqrt{x_k^2+1}}$ is α-summable by (Definition 2.18) with $\eta(|x_0|) = V(x_0) = |x_0|^2$. Now, for $p = 2$:

$$\|x\|^2_{l_2[0,k]} = \sum_{j=0}^{k} \frac{x_0^2}{jx_0^2 + 1}$$

which corresponds to the harmonic series when $x_0 = 1$, so in particular the system is not l_2-stable. However, as stated by (Theorem 2.13), there exists a change of coordinates such that the system is l_2-stable. This change of coordinates is

$$z = \frac{x\,|x|}{\sqrt{x^2+1}}.$$

Indeed, define $\kappa \in \mathcal{KL}$ such that $\kappa^{-1}(s) = \frac{s}{\sqrt{s+1}}$, so that $|z_0| = \kappa^{-1}(|x_0|^2)$ and

$$\begin{aligned}
\|z\|_{l_2[0,k]} &= \sum_{j=0}^{k} z_j^2 = \sum_{j=0}^{k} \frac{|x_j|^4}{|x_j|^2 + 1} \\
&= \sum_{j=0}^{k} \alpha(|x_j|) \leq |x_0|^2 \\
&= \kappa(|z_0|)
\end{aligned}$$

therefore by (Definition 2.19) the system in the new coordinates is l_2-stable. ◁

Similar to the (Theorem 2.13) we have the following result (see also (Theorem 2.19)).

Theorem 2.14 (ISS and l_p-gain)**.** *[5, Theorem 11, p. 364] For a fixed $p \geq 1$, if system (12) satisfies the linear l_p-gain property then it is ISS. Conversely, if system (12) is ISS then there exists a change of coordinates for state and input such that the system in the new coordinates satisfies the linear l_p-gain property.*

Example 2.6: ISS and l_2-gain. Consider the preceding (Example 2.5) with a perturbation:

$$x_{k+1} = \frac{x_k}{\sqrt{x_k^2+1}} + u_k \qquad \text{with } x_0 = \xi \in \mathbb{R}^n.$$

The Lyapunov function $V(x) = x^2$ is an ISS-Lyapunov function:

$$\begin{aligned} V(x_{k+1}) - V(x_k) &= \left(\frac{x_k}{\sqrt{x_k^2+1}} + u_k \right)^2 - x_k^2 \\ &= \frac{x_k^2}{x_k^2+1} + 2\frac{x_k}{\sqrt{x_k^2+1}} u_k + u_k^2 - x_k^2 \\ &\leq -\frac{x_k^4}{x_k^2+1} + 2\frac{|x_k|}{\sqrt{x_k^2+1}} |u_k| + |u_k|^2 \\ &\leq -\frac{x_k^4}{x_k^2+1} + (2|u_k| + u_k^2) \end{aligned}$$

so it satisfies (Definition 2.11) with $\alpha(s) = \frac{s^4}{s^2+1}$ and $\sigma(s) = 2s + s^2$. Thus, the system is ISS by (Theorem 2.7). We already observed that for $u_k \equiv 0$ the system cannot be l_2-stable, neither it can have l_2-gain. With the same change of coordinates for the state $z = \frac{x|x|}{\sqrt{x^2+1}}$ and with $v = \operatorname{sign}(u)\sqrt{2|u| + u^2}$, we can prove that the system has l_2-gain in the new coordinates. Indeed, with the same kind of manipulations of the preceding example yields

$$\sum_{j=0}^{k} \frac{|x_j|^4}{|x_j|^2 + 1} \leq x_0^2 + \sum_{j=0}^{k-1} (2|u_j| + |u_j|^2).$$

Therefore, with $\kappa \in \mathcal{KL}$ such that $\kappa^{-1}(s) = \frac{s}{\sqrt{s+1}}$,

$$\begin{aligned} \|z\|_{l_2[0,k]} = \sum_{j=0}^{k} z_j^2 &\leq \kappa(|z_0|) + \sum_{j=0}^{k-1} v_j^2 \\ &\leq \kappa(|z_0|) + \|v\|_{l_2[0,k-1]}^2 \end{aligned}$$

which is the (Definition 2.20) with transient bound $\kappa \in \mathcal{K}_\infty$ and gain bound $\gamma = 1$. Note that this gain can be arbitrarly chosen, via the change of coordinates: $v = \frac{1}{q}\operatorname{sign}(u)\sqrt{2|u| + u^2}$ would give a gain bound $\gamma = q$. ◁

2.5 Integral Input-to-state stability

Definition 2.22 (iISS)**.** System (12) is integral input-to-state stable (iISS) if $\exists \alpha, \sigma \in \mathcal{K}_\infty$ and $\exists \beta \in \mathcal{KL}$ such that

$$\alpha(\|x_k\|) \leq \beta(\|x_0\|, k) + \sum_{j=0}^{k-1} \sigma(\|u_j\|) \quad \forall x_0 \in \mathbb{R}^n \; \forall k \geq 0.$$

Lemma 2.15 (iISS characterization)**.** *System (12) is iISS if and only if there exists α, η, γ, $\sigma \in \mathcal{K}_\infty$ such that*

$$\sum_{j=0}^{k} \alpha(\|x_j\|) \leq \eta(\|x_0\|) + \gamma\left(\sum_{j=0}^{k-1} \sigma(\|u_j\|)\right) \quad \forall x_0 \in \mathbb{R}^n \ \forall k \geq 0. \tag{19}$$

As for the ISS property (Definition 2.10) one can define a Lypunov function (Definition 2.11) to characterize the iISS property (Definition 2.22):

Definition 2.23 (iISS-Lyapunov function)**.** $V : \mathbb{R}^n \to \mathbb{R}^+$ is a **iISS-Lyapunov function** for system (12) if it is continuous and

$$\exists \alpha_1,\ \alpha_2 \in \mathcal{K}_\infty \quad \alpha_1(\|\xi\|) \leq V(\xi) \leq \alpha_2(\|\xi\|) \ \ \forall \xi \in \mathbb{R}^n \tag{20a}$$

$$\exists \rho \in \mathcal{P} \ \ \exists \hat{\sigma} \in \mathcal{K}_\infty \quad V(f(\xi,\mu)) - V(\xi) \leq \hat{\sigma}(\|\mu\|) - \rho(\|\xi\|) \ \ \forall \xi \in \mathbb{R}^n \ \forall \mu \in \mathbb{R}^m \tag{20b}$$

Remark 2.15.1. Notice the similarity with an ISS-Lyapunov function (Definition 2.11). The estimate has the same form but where $\alpha_3 \in \mathcal{K}_\infty$ in (14b). Since $\mathcal{K}_\infty \subset \mathcal{P}$, an ISS-Lyapunov function is also an iISS-Lyapunov function. ◀

With this definition there is an analogous of (a part of) (Theorem 2.7) that is the following:

Theorem 2.16 (iISS and iISS Lyapunov function)**.** *System (12) is iISS if and only if there exists an iISS Lyapunov function for the system.*

As in (Theorem 2.9), we can obtain an explicit estimate (19) for ISS characterization of (Lemma 2.15) from an ISS-Lyapunov function V (if the bounds for V are explicit).

Theorem 2.17 (Explicit the estimate from a Lyapunov function)**.** *Given an iISS-Lyapunov function V as in (Definition 2.23), let $\rho_1 \in \mathcal{K}_\infty$ and $\rho_2 \in \mathcal{L}$ such that $\rho(s) \geq \rho_1(s)\rho_2(s)$ and $\rho \circ \alpha_1^{-1}(s) \leq 1$ for all $s \geq 0$. Define*

$$\chi(s) := \frac{1}{\rho_2 \circ \alpha_2^{-1}} - 1 \qquad \tilde{\alpha}(s) := \min\{s, \rho \circ \alpha_1^{-1}(s)\}$$

Then the system satisfies the iISS estimation (19) with

$$\begin{cases} \alpha(s) := \tilde{\alpha} \circ \alpha_1(s) \\ \gamma(s) := \chi(2s)s + \frac{1}{2}\chi(2s)^2 + \frac{1}{2}s^2 + s \\ \eta(s) := \gamma \circ \alpha_2(s) \\ \sigma(s) := \sigma(s). \end{cases}$$

Corollary 2.18 (iISS and l_p gain)**.** *If in the previous (Theorem 2.17) there exist two constant c_α and c_σ so that $\alpha(s) \geq c_\alpha s^p$ and $\sigma(s) \leq c_\sigma s^p$ then the system (12) satisfies the nonlinear l_p gain property (2.21) with transient and gain bounds $\kappa(s) = \frac{1}{c_\alpha}\eta(s)$ and $\mu(s) = \frac{1}{c_\alpha}\gamma(c_\sigma s)$.*

Similar to the theorems (2.13) and (2.14) we have the following:

Theorem 2.19 (iISS and l_p-gain)**.** *[5, Theorem 12, p. 364] For a fixed $p \geq 1$, if system (12) satisfies the nonlinear l_p-gain property then it is iISS. Conversely, if system (12) is iISS then there exists a change of coordinates for state and input such that the system in the new coordinates satisfies the nonlinear l_p-gain property.*

Example 2.7: iISS and l_2-gain. Consider the (Example 2.6) with an additional linear term:

$$x_{k+1} = f(x_k, u_k) := \frac{x_k}{\sqrt{x_k^2 + 1}} + u_k x_k + u_k \qquad \text{with } x_0 = \xi \in \mathbb{R}^n.$$

For $u_k \equiv 1$ and initial condition $x_0 > 0$, the solution will always be positive and strictly increasing since

$$f(x_k, u_k) - x_k = \frac{x_k}{\sqrt{x_k^2 + 1}} + 1 > 1 \quad \forall k \geq 0 \; \forall x_0 \in \mathbb{R}.$$

This means that the system is not ISS. The Lyapunov function candidate will be $V(x) = V_1(x) + V_2(x)$ with

$$V_1(x) = \arctan(|x|) \qquad V_2(x) = \log(|x| + 1).$$

It is possible to prove that

$$\begin{aligned} V_1(x_{k+1}) - V_1(x_k) &\leq \frac{|x_{k+1}| - |x_k|}{x_k^2 + 1} \\ &\leq -\frac{|x_k|}{x_k^2 + 1}\left(1 - \frac{1}{\sqrt{x_k^2 + 1}}\right) + \frac{|u_k x_k|}{x_k^2 + 1} + |u_k| \\ &\leq -\frac{|x_k|}{x_k^2 + 1}\left(1 - \frac{1}{\sqrt{x_k^2 + 1}}\right) + 2\,|u_k| \\ &=: -\rho(|x_k|) + \sigma_1(|u_k|) \end{aligned}$$

where in particular ρ is positive definite and $\sigma_1 \in \mathcal{K}_\infty$. In addition

$$\begin{aligned} V_2(x_{k+1}) - V_2(x_k) &\leq \log\left(|x_k - |x_k|\,|u_k| + |u_k| + 1|\right) - \log(|x_k| + 1) \\ &\leq \log\left((|x_k| + 1)(|u_k| + 1)\right) - \log(|x_k| + 1) \\ &\leq \log(|u_k| + 1) \\ &=: \sigma_2(|u_k|). \end{aligned}$$

Define $\sigma(s) := \sigma_1(s) + \sigma_2(s)$ and hence,

$$V(x_{k+1}) - V(x_k) \leq -\rho(|x_k|) + \sigma(|u_k|)$$

which guarantee that the system is iISS by (Theorem 2.16). In the same way as in (Example 2.5), taking $u_k \equiv 0$ the system cannot have nonlinear l_2-gain.

However, (Theorem 2.19) states that there exists a change of coordinates for which the system has a nonlinear l_2-gain. This change of coordinates is given by

$$z = \operatorname{sign}(x)\kappa(|x|) = \operatorname{sign}(x)\sqrt{\alpha(|x|)}$$
$$v = \operatorname{sign}(u)\sqrt{\sigma(|u|)} = \operatorname{sign}(u)\sqrt{2\,|u| + \log(|u|+1)}$$

where

$$\alpha(s) := \min\left\{\log(s+1), s(\sqrt{s^2+1}-1)\right\}.$$

Using the following definitions:

$$\varphi(s) = (e^s - 1)^2 \qquad \mu(s) = \varphi(2s)^4 + 2\varphi(2s)^2 s + 2\varphi(2s)s + \varphi(2s)^2 + 2s^2 + s$$
$$\alpha_2(s) = \log(s+1) + \arctan(s) \qquad \hat{\kappa}(s) = \mu \circ \alpha_2 \circ \kappa^{-1}(s)$$

one obtains:

$$\begin{aligned}
\|z\|^2_{l_2[0,k]} &= \sum_{j=0}^{k} |z_j|^2 = \sum_{j=0}^{k} \alpha(|x_j|) \\
&\leq \mu(\alpha_2(|x_0|)) + \mu\left(\sum_{j=0}^{k-1} \sigma(|u_j|)\right) \\
&= \mu(\alpha_2(\kappa^{-1}(|z_0|))) + \mu\left(\sum_{j=0}^{k-1} |v_j|^2\right) \\
&= \hat{\kappa}(|z_0|) + \mu(\|v\|^2_{l_2[0,k-1]}).
\end{aligned}$$

Thus, in the new coordinates, the system satisfies the nonlinear l_2-gain property. ◁

2.6 Incremental stability

Incremental stability extends the classical notion of asymptotic stability of an equilibrium of a nonlinear system to consider the asymptotic behavior of any solution with respect to any other solution. Specifically, any two solutions must eventually asymptotically converge to each other regardless of their initial conditions.

Consider the system (12):

$$x_{k+1} = f(x_k, u_k) \qquad x(0) = \xi$$

Definition 2.24 (δGAS)**.** The system (12) is **incremental globally asymptotically stable (δGAS)** if $\exists \beta \in \mathcal{KL}$ such that for every sequence of disturbances u and every initial states ξ_1, $\xi_2 \in \mathbb{R}^n$

$$\|x(k, \xi_1, u) - x(k, \xi_2, u)\| \leq \beta(\|\xi_1 - \xi_2\|, k) \quad \forall k \geq 0.$$

Remark 2.19.1. Incremental stability describes the convergence of trajectories with respect to themselves, rather than with respect to an equilibrium point or a particular trajectory. ◀

Example 2.8: A δGAS system. The system

$$x_{k+1} = -\frac{k}{2} - 1 + \frac{x_k}{2} \qquad \text{with } x_0 = \xi \in \mathbb{R}^n$$

has an explicit solution

$$x(k, \xi) = -k + \frac{\xi}{2^k} \qquad \forall k \geq 0.$$

For any two initial conditions $\xi_1, \xi_2 \in \mathbb{R}^n$ yields

$$\|x(k, \xi_1) - x(k, \xi_2)\| = \frac{\|\xi_1 - \xi_2\|}{2^k} =: \beta(\|\xi_1 - \xi_2\|, k)$$

and $\beta(s, r) = \frac{s}{2^r}$ is a $\mathcal{KL}$ function. Hence, the system $x_{k+1} = -\frac{k}{2} - 1 + \frac{x_k}{2}$ is globally asymptotically incrementally stable. However, notice that for the specific initial condition $\xi = 0$, the solution $x(k, \xi) = -k$ from this initial condition is unbounded. ◁

The following theorem is a discrete-time Lyapunov function characterization of incremental stability.

Theorem 2.20 (δGAS Lyapunov function.)**.** *[8, Theorem 9, p. 7] System (12) is δGAS if and only if there exists a smooth function $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ which satisfies*

$$\alpha_1(\|x_1 - x_2\|) \leq V(x_1, x_2) \leq \alpha_2(\|x_1 - x_2\|)$$
$$V(f(x_1), f(x_2)) - V(x_1, x_2) \leq -\alpha_3(\|x_1 - x_2\|)$$

for some $\alpha_1, \alpha_2, \in \mathcal{K}_\infty$ and $\alpha_3 \in \mathcal{P}$ for all $x_1, x_2 \in \mathbb{R}^n$ and all $k \geq 0$.

Incremental ISS is the notion that estimates differences $\|x_1(t) - x_2(t)\|$ in terms of $\mathcal{KL}$ decay of differences of initial states, and differences of norms of inputs. Consider system (12) where $u \in U$ a closed and convex set of $\mathbb{R}^m$. Also suppose $f(0, 0) = 0$.

Under these assumptions we define:

Definition 2.25 (δISS)**.** The system (12) is **incrementally input-to-state stable (δISS)** if there exists a function $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}_\infty$ such that for any $k \geq 0$, any initial states ξ_1, $\xi_2 \in \mathbb{R}^n$ and any couple of disturbances u_1, u_2 the following is true

$$\|x(k, \xi_1, u_1) - x(k, \xi_2, u_2)\| \leq \beta(\|\xi_1 - \xi_2\|, k) + \gamma(\|u_1 - u_2\|_\infty).$$

This kind of stability definition provides a way to formulate notions of sensitivity to initial conditions and controls. Notice that in particular when there are no inputs one obtains incremental GAS property of (Definition 2.24). Indeed, note that the difference between the above inequality and the one defining incremental global asymptotic stability is that the latter property holds for arbitrary identical input signals $u_1 = u_2 = u$, and arbitrary pairs of initial states.

Remark 2.20.1. Again, in the previous definition, the summation on the RHS may be replaced by $\max\{\beta(\|\xi_1 - \xi_2\|, k), \gamma(\|u_1 - u_2\|_\infty)\}$. ◀

Remark 2.20.2. Since $f(0,0) = 0$ it is easy to check that δISS implies ISS just comparing an arbitrary trajectory with $x_k \equiv 0$. ◀

A Lyapunov characterization exists for incremental input-to-state stability as well. In the following definition, x^i is an abbreviation for $x(k, \xi_i, u_i)$.

Definition 2.26 (δISS Lyapunov function)**.** A function $V\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called a **δISS Lyapunov function** if for any u_1, $u_2 \in U$ and any x^1, $x^2 \in \mathbb{R}^n$

$$\alpha_1(\|x^1 - x^2\|) \leq V(x^1, x^2) \leq \alpha_2(\|x^1 - x^2\|) \tag{21a}$$

$$V(f(x^1, u_1), f(x^2, u_2)) - V(x^1, x^2) \leq -\alpha_4(\|x^1 - x^2\|) + \sigma(\|u_1 - u_2\|) \tag{21b}$$

for some α_1, α_2, $\alpha_4 \in \mathcal{K}_\infty$ and $\sigma \in \mathcal{K}$.

Remark 2.20.3. The second condition (21b) can be restated in an implication form: there exists $\kappa \in \mathcal{K}_\infty$ such that for every x^1, x^2 and $u_1, u_2 \in U$

$$\begin{aligned} &\|u_1 - u_2\| \leq \kappa(\|x^1 - x^2\|) \Rightarrow \\ &V(f(x^1, u_1), f(x^2, u_2)) - V(x^1, x^2) \leq -\rho(\|x^1 - x^2\|) \quad \forall k \geq 0 \end{aligned} \tag{22}$$

where $\rho \in \mathcal{K}_\infty$ ◀

Proof. Indeed, if condition (21b) holds and $\|u_1 - u_2\| \leq \kappa(\|x^1 - x^2\|)$ then

$$\begin{aligned} V(f(x^1, u_1), f(x^2, u_2)) - V(x^1, x^2) &\leq -\alpha_4(\|x^1 - x^2\|) + \sigma(\|u_1 - u_2\|) \\ &\leq \underbrace{-\alpha_4(\|x^1 - x^2\|) + \sigma(\kappa(\|x^1 - x^2\|))}_{-\rho(\|x^1 - x^2\|)}. \end{aligned}$$

□

Remark 2.20.4. The reverse of the preceding remark requires additional assumptions: suppose $V\colon G \times G \to \mathbb{R}$ with $G \subset \mathbb{R}^n$ compact and $U \subset \mathbb{R}^m$ compact. Then, if system (12) admits an implication form incremental ISS Lyapunov function (22) it admits a dissipation form incremental ISS-Lyapunov function (21b). ◀

Theorem 2.21 (δISS and δISS Lyapunov function.)**.** *If the system (12) admits a time-invariant δISS Lyapunov function, then it is δISS. Moreover, if the set U is compact the two conditions are equivalent.*

Remark 2.21.1. As stated in [9], the following classes of systems are δISS:

- linear time-invariant systems which are asymptotically stable;
- globally Lipschitz systems with a Lipschitz constant $0 < L < 1$;

◀

2.7 Relationships between the different concepts

We might be interested in which relationships are (if any) between the different notions of stability given in (SECTION 2).

We have already remarked the Fundamental Relationship Among ISS, IOS, and IOSS in (Remark 2.5.2): see figure (Figure 6).

Figure 6: Fundamental Relationship Among ISS, IOS, and IOSS (Remark 2.5.2).

(Example 2.1) shows that ISS property is stronger than Converging Input Converging Output property combined with 0-Global Asymptotic Stability.

Figure 7: ISS is stronger than 0-GAS + CICO: (Example 2.1).

In (Remark 2.20.2) we observed that a δISS system is clearly ISS when $f(0,0) = 0$. In the continuous case, [6] suggest a counter-example to the implication ISS $\Rightarrow$ δISS.

Example 2.9. The system $\dot{x} = -x + u^3$ is ISS with respect to the equilibrium point $x_u = \bar{u}^3$ and the input signal $u - \bar{u}$ for all $\bar{u} \in \mathbb{R}$. For this, is enough to choose $V_{\bar{u}}(x) = (x - \bar{u}^3)^2$ as ISS Lyapunov function. Moreover it is GAS (with respect to the "disturbance" u in any compact subset of $\mathbb{R}$):

$$\frac{d}{dt}\left(x(t,\xi_1,u) - x(t,\xi_2,u)\right) = -\left(x(t,\xi_1,u) - x(t,\xi_2,u)\right).$$

Nevertheless, it is not δISS: we pick u_1 and u_2 as given by the closed-loop feedback that makes the system into $\dot{x} = 1$. That means $u_i^3 = x(t, \xi_i, u_i) + 1$ and thus $x(t, \xi_i, u_i) = t + \xi_i$ for $i = 1, 2$ so that

$$u_1(t) = (t + \xi_1 + 1)^{1/3} \qquad u_2(t) = (t + \xi_2 + 1)^{1/3}$$

and in particular $x(t, \xi_1, u_1) - x(t, \xi_2, u_2) = \xi_1 - \xi_2$ is constant whereas $u_1(t) - u_2(t) \to 0$. This contradicts the converging-inputs–converging-states property of δISS.

◁

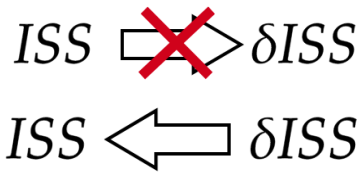


Figure 8: A δISS system is ISS but there are ISS systems that are not δISS: (Example 2.9) and (Example 2.10).

We can transpose the same example in a discrete time setting (see (Appendix A.5)).

Example 2.10. Consider the discrete time system

$$x_{k+1} = (1 - \delta)x + \delta u^3.$$

This is ISS with respect to the equilibrium point $x_u = \bar{u}^3$ and the input signal $u - \bar{u}$. Indeed, $V(x) = (x - \bar{u}^3)^2$ is a ISS Lyapunov function. We show that the system is not δISS, by contradiction with the converging-inputs-converging-states properties. Take two different input u, v corresponding to the solutions x and y respectively. Choose them in order to have $u_k^3 = x_k + 1$ and $v_k^3 = y_k + 1$ so that the solutions are

$$x_k = x_0 + k\delta = \xi_1 + k\delta \qquad y_k = y_0 + k\delta = \xi_2 + k\delta$$

and the corresponding inputs become

$$u_k = (k\delta + \xi_1 + 1)^{1/3} \qquad v_k = (k\delta + \xi_2 + 1)^{1/3}.$$

Now the difference $x - y = \xi_1 - \xi_2$ is constant even though $u - v \to 0$. ◁

As noticed in (Remark 2.15.1), an ISS-Lyapunov function is an iISS-Lyapunov function as well, so that ISS $\Rightarrow$ iISS; the converse is not true, as shown in (Example 2.7).

2.8 Summary of implications

In this section we propose some diagrams to summarize the relationship among all the different stability properties for discrete-time systems.

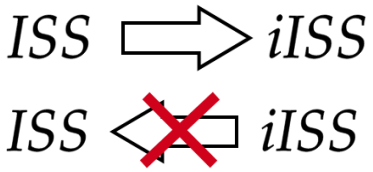


Figure 9: An ISS system is iISS - using the characterization via Lyapunov functions - but an iISS system is not necessarily ISS.

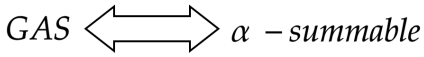


Figure 10: Equivalence from [5, Theorem 1, p. 362]

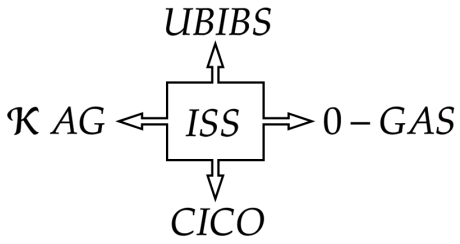


Figure 11: Implications of ISS property as showed in (SUBSECTION 2.3).

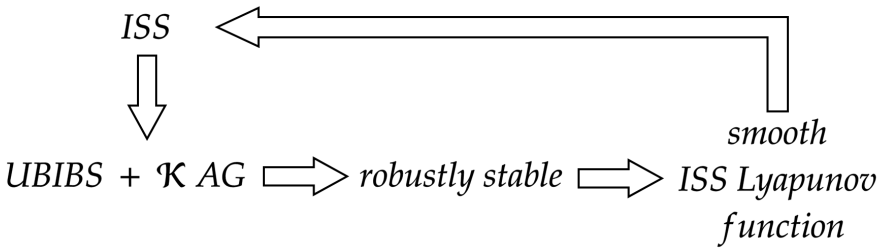


Figure 12: Equivalence of ISS property from (Theorem 2.7).

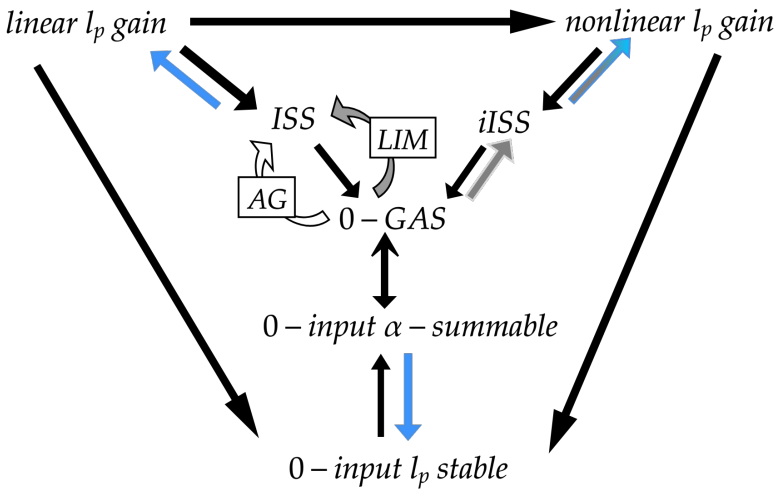


Figure 13: Implication diagram showing the relationship between stability properties for discrete-time systems. The gray arrows are for implications that require f in system (12) to be continuous. Blue arrows are for implications that require change of coordinates. Where indicated more assumptions are needed (asymptotic gain (AG) (Definition 2.12), and limit (LIM) (Definition 2.13)).

2.9 Conclusion of this section

This section contains the definitions and main results of different types of stability (Lyapunov, ISS, Input-output, incremental stability ...) for discrete-time systems. In particular, we presented the comparison functions formalism, which is a powerful and useful tool to prove stability results. For each notion some basic examples are provided as well as some proofs (which I made as an exercise). At the end, different diagrams help to understand the different relationships between the different concepts and the major implications among the several types of stability. Hereafter, the most important tools will be from (SUBSECTION 2.1) (comparison functions), (SUBSECTION 2.3) (input-to-state-stability), and (SUBSECTION 2.5) (integral-input-to-state-stability).

3 Lyapunov functions and gains

In this section we will investigate how scaling the Lyapunov function affects the asymptotic gain and the transient bound for a stable system (ISS or δISS).

3.1 Lyapunov function and ISS bounds

Recall (14a) and (14b) from (Definition 2.11):

$$\exists \alpha_1,\ \alpha_2 \in \mathcal{K}_\infty \quad \alpha_1(\|\xi\|) \leq V(\xi) \leq \alpha_2(\|\xi\|) \ \ \forall \xi \in \mathbb{R}^n \tag{14a}$$

$$\exists \alpha_3 \in \mathcal{K}_\infty \ \ \exists \sigma \in \mathcal{K} \quad V(f(\xi,\mu)) - V(\xi) \leq \sigma(\|\mu\|) - \alpha_3(\|\xi\|) \ \ \forall \xi \in \mathbb{R}^n \ \forall \mu \in \mathbb{R}^m \tag{14b}$$

Concerning the ISS property, we already observed in (Remark 2.6.1) that if V is an ISS-Lyapunov function, then for every positive λ, $\hat{V} = \lambda V$ is still an ISS-Lyapunov function. Then using (Theorem 2.8) and the $\mathcal{K}$-asymptotic gain becomes $\hat{\gamma}_a(s) = \alpha_1^{-1} \circ \frac{1}{\lambda}\alpha_3^{-1} \circ \sigma(s)$ while from (Theorem 2.9) the $\mathcal{K}_\infty$ functions for the characterization (13) of (Lemma 2.6) become

$$\begin{cases} \hat{\alpha}(s) &= \hat{\bar{\alpha}} \circ \hat{\alpha}_1(s) = \lambda \bar{\alpha} \circ \alpha_1(s) \\ \hat{\eta}(s) &= \hat{\alpha}_2(s) = \lambda \alpha_2(s) \\ \hat{\sigma}(s) &= \lambda \sigma(s). \end{cases}$$

Notice that the only nonlinear change appears in the $\mathcal{K}$-asymptotic gain $\hat{\gamma}_a$. From the proof of (Theorem 2.7), we can explicit the $\mathcal{KL}$ function β and the $\mathcal{K}_\infty$ function γ of the (Definition 2.10) of input-to-state stability. In fact, the proof is "quasi" constructive and the bounds on V are used to build this functions as:

$$\begin{aligned} \beta(s,t) &= \alpha_1^{-1}(\tilde{\beta}(\alpha_2(s),t)) \\ \gamma(r) &= \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1} \circ \rho^{-1} \circ \sigma(r) \end{aligned} \tag{23}$$

where $\tilde{\beta}$ is given by the comparison (Lemma 3.1) and $\rho \in \mathcal{K}_\infty$ is such that $id - \rho \in \mathcal{K}_\infty$. The proof of this construction can be found in [4, Lemma 3.5, p. 860] and shows that if system (12) admits a continuous ISS-Lyapunov function, then it is ISS. If we decide to scale $\hat{V} = \lambda V$, then all the functions α_i and σ from (Definition 2.11) are also scaled and the previous equations for β and γ change as follows:

$$\begin{aligned} \hat{\beta}(s,t) &= \alpha_1^{-1}\left(\frac{1}{\lambda}\tilde{\beta}(\lambda\alpha_2(s),t)\right) \\ \hat{\gamma}(r) &= \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1}\left(\frac{1}{\lambda}\hat{\rho}^{-1}(\lambda\sigma(r))\right) \end{aligned} \tag{24}$$

where $\hat{\rho}$ is a $\mathcal{K}_\infty$ function such that $id - \rho$ is again $\mathcal{K}_\infty$. According to the choice of $\hat{\rho}$ (which may depend on λ) we can decide to simplify a little the expression

for $\hat{\gamma}$:

$$\hat{\rho}(s) = \lambda\rho(s) \Rightarrow \hat{\gamma}(r) = \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1} \left(\frac{1}{\lambda} \rho^{-1}(\sigma(r)) \right)$$

or

$$\hat{\rho}(s) = \rho\left(\frac{s}{\lambda}\right) \Rightarrow \hat{\gamma}(r) = \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1} \left(\rho^{-1}(\lambda\sigma(r)) \right).$$

Notice that in this way the effect of scaling appears only once. Of course, choices like $\hat{\rho} = \rho$ or $\hat{\rho}(s) = \rho(\lambda s)$ are possible as well. Another interesting choice is

$$\hat{\rho}(s) = \lambda\rho\left(\frac{s}{\lambda}\right) \quad \text{so that} \quad \hat{\rho}^{-1}(s) = \lambda\rho^{-1}\left(\frac{s}{\lambda}\right)$$

and the gain remains unchanged:

$$\hat{\gamma} = \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1} \circ \rho^{-1} \circ \sigma(s) = \gamma.$$

Remark that this choise is always possible for any $\lambda > 0$. Indeed, if $\rho \in \mathcal{K}_\infty$ and such that $id - \rho \in \mathcal{K}_\infty$, such a $\hat{\rho}$ is still a $\mathcal{K}_\infty$ function and for every $s > 0$

$$s - \hat{\rho}(s) = s - \lambda\rho\left(\frac{s}{\lambda}\right) > 0$$

so that $id - \hat{\rho}$ is again $\mathcal{K}_\infty$.

Remark 3.0.1. In the implication form (15) when V is scaled by λ, then α_4 is too, while χ stays unchanged:

$$\exists \alpha_4 \in \mathcal{K}_\infty \; \exists \chi \in \mathcal{K} \quad \|\xi\| \geq \chi(\|\mu\|) \Rightarrow \lambda V(f(\xi, \mu)) - \lambda V(\xi) \leq -\lambda\alpha_4(\|\xi\|).$$

◀

Let us state the comparison lemma mentioned before.

Lemma 3.1 (Comparison lemma)**.** *[12, Lemma 4.3, p. 55] For each $\mathcal{K}$-function α there exists a $\mathcal{KL}$-function $\beta_\alpha(s,t)$ with the following property: if $y\colon \mathbb{N} \to [0,\infty)$ is a function satisfying*

$$y(k+1) - y(k) \leq -\alpha(y(k)) \quad \forall\, 0 \leq k \leq k_1$$

for some $k_1 \leq \infty$, then

$$y(k) \leq \beta_\alpha(y(0), k) \quad \forall\, k < k_1.$$

In the proof given in [12], the β_α is defined from $id - \alpha$ through iteration and max operations. In the contest of ISS-Lyapunov function, y corresponds to V and α is defined as being $\alpha := (id - \rho) \circ \alpha_3 \circ \alpha_2^{-1}$, so that with the last mentioned choice of $\hat{\rho}$, the new $\mathcal{K}$ function would just be a scaled version: $\hat{\alpha} = \lambda\alpha$.

We focused on a linear scaling for the ISS-Lyapunov function, but nonlinear scaling are possible. Let $\phi \in \mathcal{K}_\infty$ and pose $W(x) = \phi(V(x))$. We want to found the conditions for which, if V is an ISS-Lyapunov function, then W is still an ISS-Lyapunov function for the same system.

Proposition 3.2 (Nonlinear scaling of ISS-Lyapunov function)**.** *Given an ISS-Lyapunov function V for the system (12) that satisfies*

$$\begin{aligned}\alpha_1(\|\xi\|) \leq V(\xi) \leq \alpha_2(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \\ V(f(\xi,\mu)) - V(\xi) \leq \sigma(\|\mu\|) - \alpha_3(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \; \forall \mu \in \mathbb{R}^m\end{aligned}$$

and applying a non-linear scaling: $W(x) = \phi(V(x))$ for some smooth $\phi \in \mathcal{K}_\infty$ such that $\phi' \in \mathcal{K}$. Then $W(x)$ is an ISS-Lyapunov function as it satisfies the conditions in the implication forms:

$$\begin{aligned}\hat{\alpha}_1(\|\xi\|) \leq W(\xi) \leq \hat{\alpha}_2(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \\ \|x\| \geq \chi(\|u\|) \Rightarrow \; W(f(\xi,\mu)) - W(\xi) \leq -\hat{\alpha}_4(\|\xi\|) \;\; \forall \xi \in \mathbb{R}^n \; \forall \mu \in \mathbb{R}^m\end{aligned}$$

Proof. The first condition (14a) is trivially satisfied:

$$\hat{\alpha}_1(\|x\|) = \phi(\alpha_1(x)) \leq \underbrace{\phi(V(x))}_{W(x)} \leq \phi(\alpha(\|x\|)) = \hat{\alpha}_2(\|x\|)$$

while condition (14b) becomes

$$\begin{aligned}W(x_+) - W(x) &= \phi(V(x_+)) - \phi(V(x)) \\ &\leq \phi(V(x_+) - V(x) + V(x)) - \phi(V(x)) \\ &\leq \phi(\sigma(\|u\|) - \alpha_4(\|x\|) + V(x)) - \phi(\alpha_1(\|x\|)) \\ &\leq \phi(\sigma(\|u\|) + (\alpha_2 - \alpha_4)(\|x\|)) - \hat{\alpha}_1(\|x\|) \\ &\leq \underbrace{\phi(\sigma(\|u\|))}_{\hat{\sigma}(\|u\|)} - \hat{\alpha}_1(\|x\|)\end{aligned}$$

where the last inequality yields if $\alpha_2 \leq \alpha_4$. A better condition on α_4 is $\|x\| \geq \alpha_4^{-1}(V(x))$:

$$\begin{aligned}W(x_+) - W(x) &\leq \phi(\sigma(\|u\|) - \alpha_4(\|x\|) + V(x)) - \phi(\alpha_1(\|x\|) \\ &\leq \underbrace{\phi(\sigma(\|u\|))}_{\hat{\sigma}(\|u\|)} - \hat{\alpha}_1(\|x\|)\end{aligned}$$

Supposing ϕ smooth and $\phi' \in \mathcal{K}$:

$$\begin{aligned}W(x_+) - W(x) &= \phi(V(x_+)) - \phi(V(x)) \\ &= \phi'(\tau)(V(x_+) - V(x)) \\ &\leq \phi'(V(x_+) + V(x))(-\alpha_4(\|x\|) + \sigma(\|u\|)) \\ &\leq \phi'(-\alpha_4(\|x\|) + \sigma(\|u\|) + 2V(x))(-\alpha_4(\|x\|) + \sigma(\|u\|)) \\ &\leq \phi'(2\alpha_2(\|x\|) + \sigma(\|u\|))(-\alpha_4(\|x\|) + \sigma(\|u\|)) \\ &\leq (\phi'(4\alpha_2(\|x\|)) + \phi'(2\sigma(\|u\|)))\,(-\alpha_4(\|x\|) + \sigma(\|u\|))\end{aligned}$$

and if we use the implication form (15) with $\|x\| \geq \alpha_4^{-1}(2\sigma(\|u\|)) =: \chi(\|u\|)$:

$$\begin{aligned} W(x_+) - W(x) &\leq \phi'(4\alpha_2(\|x\|))\,(\sigma(\|u\|) - \alpha_4(\|x\|)) + \phi'(2\sigma(\|u\|))\,(\sigma(\|u\|) - \alpha_4(\|x\|)) \\ &\leq -\frac{1}{2}\alpha_4(\|x\|)\,(\phi'(4\alpha_2(\|x\|)) + \phi'(\alpha_4(\|x\|))) \\ &\leq -\hat{\alpha}_4(\|x\|) \end{aligned}$$

□

Let now $\lambda \in (0,1)$. If α_3 is replace by λV in (14b) of (Definition 2.11), one obtain a so called **exponential ISS-Lyapunov function**, for which

$$\exists \alpha_1,\ \alpha_2 \in \mathcal{K}_\infty \quad \alpha_1(\|\xi\|) \leq V(\xi) \leq \alpha_2(\|\xi\|) \quad \forall \xi \in \mathbb{R}^n \tag{14a}$$

$$\exists \sigma \in \mathcal{K} \quad V(f(\xi,\mu)) \leq \sigma(\|\mu\|) + \lambda V(\|\xi\|)\ \ \forall \xi \in \mathbb{R}^n \quad \forall \mu \in \mathbb{R}^m \tag{25}$$

hold. In [11] it is proved that if there exists an ISS-Lyapunov function V, then there exists an exponential ISS-Lyapunov function $\hat{V}$ as well, given by a nonlinear scaling $\hat{V} = \hat{\alpha}(V)$ for some $\hat{\alpha} \in \mathcal{K}_\infty$. The corresponding implication form of condition (25) is

$$\exists \chi \in \mathcal{K} \quad \|\xi\| \geq \chi(\|\mu\|) \Rightarrow V(f(\xi,\mu)) \leq \lambda V(\xi). \tag{26}$$

3.2 Lyapunov function and δISS bounds

Recall (21a) and (21b) from (Definition 2.26):

$$\alpha_1(\|x^1 - x^2\|) \leq V(x^1,x^2) \leq \alpha_2(\|x^1 - x^2\|) \tag{21a}$$

$$V(f(x^1,u_1),f(x^2,u_2)) - V(x^1,x^2) \leq -\alpha_4(\|x^1 - x^2\|) + \sigma(\|u_1 - u_2\|) \tag{21b}$$

for some α_1, α_2, $\alpha_4 \in \mathcal{K}_\infty$ and $\sigma \in \mathcal{K}$.

In this case as well, if V is a δISS-Lyapunov function, then for every positive λ, $\hat{V} = \lambda V$ is still a δISS-Lyapunov function for the system.

The reasoning applied to the ISS property still holds: from the proof of [8, Theorem 8, p.482], which still uses (Lemma 3.1), the gain and the transient bound are given by (23)

$$\begin{aligned} \beta(s,t) &= \alpha_1^{-1}(\tilde{\beta}(\alpha_2(s),t)) \\ \gamma(r) &= \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1} \circ \rho^{-1} \circ \sigma(r) \end{aligned} \tag{23}$$

so that a scaling $\hat{V} = \lambda V$ produces again (24)

$$\begin{aligned} \hat{\beta}(s,t) &= \alpha_1^{-1}\left(\frac{1}{\lambda}\tilde{\beta}(\lambda\alpha_2(s),t)\right) \\ \hat{\lambda}(r) &= \alpha_1^{-1} \circ \alpha_2 \circ \alpha_3^{-1}\left(\frac{1}{\lambda}\hat{\rho}^{-1}(\lambda\sigma(r))\right). \end{aligned} \tag{24}$$

Again for a nonlinear scaling by a smooth $\mathcal{K}_\infty$ function, one still has a δISS-Lyapunov function.

3.3 Conclusion of this section

In this section some effects after linear and nonlinear scaling of an ISS or δ-ISS Lyapunov function are depicted. In particular we were interested in the changes on the asymptotic gain and the transient bound for a stable system.

4 Newton's method

4.1 An introduction to Newton's method in dimension 1

As a method to find the root of a function $f\colon \mathbb{R} \to \mathbb{R}$: it uses an approximation of the function by its tangent line and the x-intercept will be a better approximation of the root, then we can iterate. We define the Newton transform $N_f(x) = -f'(x)^{-1}f(x)$ and we have the iteration step $x_{n+1} = x_n + N_f(x_n)$ solving the equation $f'(x_n)(x - x_n) + f(x_n) = 0$ for x. We can justify a quadratic convergence by a Taylor expansion of the function f around its root α:

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2}f''(\xi_n)(\alpha - x_n)^2$$

that leads to a recurrence formula on the error $|\varepsilon_{n+1}| := |\alpha - x_{n+1}| = \frac{|f''(\xi_n)|}{2|f'(x_n)|}\,|\varepsilon_n|^2$. In particular we need the following three hypothesis:

i) $f'(x) \neq 0$ in a neighbourhood of the root α;

ii) f'' to be continuous on this neighbourhood;

iii) the initial guess x_0 close enough to the root so that we can justify the Taylor expansion and neglect the higher terms: $\frac{1}{2}\frac{|f''(\xi_n)|}{|f'(x_n)|} \leq C\frac{|f''(\alpha)|}{|f'(\alpha)|} < \frac{1}{\varepsilon_n}$

4.2 General case

We can easily generalize this method in dimension $k > 1$, and in this case one just need to use the gradient and the Hessian matrix of f instead of its first and second derivative. The iteration step is then of the form:

$$x_{k+1} = x_k - \nabla f^{-1}(x_k)f(x_k). \tag{27}$$

Remark 4.0.1. Another generalization of Newton's method can be done: take f a functional on Banach space saying Df its Frechet derivative. In this case we could use a damping strategy aiming to avoid the appearance of possibly large updates in the iterations: $\begin{cases} x_{n+1} = x_n + \eta\delta_n \\ Df(x_n)\delta_n = -f(x_n) \end{cases}$ for $0 < \eta < 1$. ◀

Remark 4.0.2 (Newton's method for stationary points.). We can see Newton's method as a descent method (an iterative algorithm for minimizing a function g) where the descent direction is $d_k = -H_g^{-1}(x_k)\nabla g(x_k)$ so that the iteration takes the form $x_{k+1} = x_k - \alpha_k - H_g^{-1}(x_k)\nabla g(x_k)$. The idea in Newton's method is to minimize at each iteration the quadratic approximation off around the current point x_k. The price for the fast convergence of Newton's method is

the overhead required to calculate the Hessian matrix, and to solve the linear system of equations

$$H_g(x_k)d_k = \nabla g(x_k).$$

The stationary point of g correspond to the roots of $f = \nabla g$ in the previous form. ◀

In the rest of this document, we will always consider the Newtons method in this form, that is with the following iteration step:

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) \tag{28}$$

Remark 4.0.3. For a special class of function, Newton's method converge to the exact solution in only one step! This corresponds to the quadratic case. Consider

$$f(x) = x^T Ax + Bx + c$$

where A is symmetric positive definite. Its gradient and Hessian matrix are

$$\nabla f(x) = Ax + B \qquad H(x) = A$$

so that an iteration step is

$$x_+ = x - A^{-1}(Ax + B) = -A^{-1}B$$

which is a stationary point for f: $\nabla f(x_+) = -AA^{-1}B + B = 0$. ◀

4.3 Importance of the hypothesis

All the simulations and plots in this section have been realized in MATLAB.

A few problems may arise if one or more of the hypothesis are not verified: we can see cycles appearing, divergence or oscillation when the first derivative is not defined or unbounded at roots...

Remark 4.0.4. Many factor can slow the convergence down: if we cannot calculate f' analytically (*secant method*), if a root has multiplicity bigger than 1, in case the algorithm encounters stationary points (we could be dividing by 0), if the second derivative does not exist at the root... ◀

Table (1) presents different situations that may occur in the case in which Newton's method is used to find roots of a real valued function $f\colon \mathbb{R} \to \mathbb{R}$. The starting point is always $x_0 = 0$ and the number of iteration was chosen to be $N = 10$. Here is the explanation of the different examples:

(a) the method works well because $f(x)$ is smooth.

(b) In this case the problem is that during the iteration step one try to divide by 0 which is the value of the derivative at x_0. The problem is only due to the particular starting point: $x_0 > 0$ would make the algorithm converge to 1 and $x_0 < 0$ would make the algorithm converge to -1.

(c) Again, starting point enters a cycle and the algorithm does not converge; choosing $x_0 = 2$ would make it converge to the real root $x = -\frac{2}{(27-3\sqrt{57})^{1/3}} - \frac{(9-\sqrt{57})^{1/3}}{3^{2/3}} \approx -1.7693$. Figure (Figure 14) is a graphical view of this case.

(d) As in case (b), the problem is that the derivative does not exist at root, however in this case a different starting point would not avoid the problem (as choosing a point close to 0 would make the algorithm diverge).

(e) The derivative is discontinuous at 0 and because of the particular starting point the method fails; with $x_0 = 0.01$ a value of $3.5410e-06$ is found after 1000 iterations.

(f) In this case the algorithm has no problem in converging to the exact solution, but the convergence rate is not quadratic since there is no second derivative at the root.

(g) Function $f(x) = x^2 + 1$ has no real roots, so Newton's method will chaotically move around the x axis.

Remark 4.0.5 (What is really needed.). In order to guarantee the convergence of Newton's method, we need *(i)* a sufficient regularity of the function (∇f continuously differentiable with some bounds linking $\min |\nabla f(x)|$ and $\max |f'(x)|$);

(ii) we want to be able to inverse the Hessian matrix; *(iii)* iterations should start close enough to the stationary point of f. However if we consider Newton's method in the complex field, we can see the convergence to $\pm i$ according to the starting point in the upper or lower semi-plane. ◀

	$f(x)$	$f'(x)$	exact root	found root	error
(a)	$e^x + x$	$e^x + 1$	≈ -0.56714	-0.5671	$1.2537e-07$
(b)	$1 - x^2$	$-2x$	± 1	`NaN`	`NaN`
(c)	$x^3 - 2x + 2$	$3x^2 - 2$	$-\frac{2}{(27-3\sqrt{57})^{1/3}} - \frac{(9-\sqrt{57})^{1/3}}{3^{2/3}}$	cycle of 0 and 1	1
(d)	$x^{1/3}$	$\frac{1}{3}x^{-2/3}$	0	*failed: derivative is* 0	
(e)	$x + x^2 \sin\left(\frac{2}{x}\right)$	$1 + 2x \sin\left(\frac{2}{x}\right) - 2\cos\left(\frac{2}{x}\right)$	0	*failed:* x_1 *is* `NaN`	
(f)	$x + x^{4/3}$	$1 + \frac{4}{3}x^{1/3}$	0	$1.6797e-10$	$1.0669e-07$
(g)	$x^2 + 1$	$2x$	$\pm i \notin \mathbb{R}$	changes slightly	≥ 1

Table 1: Different behaviours of Newton's method in dimension 1. The starting point was chosen $x_0 = 0$ (but $x_0 = 0.1$ in case (g)) and the number of iteration fixed to $N = 10$.

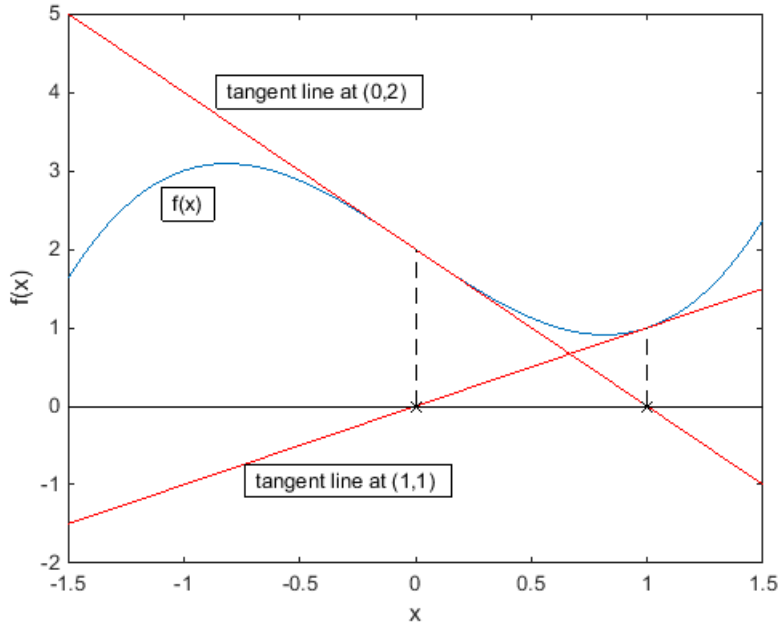


Figure 14: The function $f(x) = x^3 - 2x + 2$ of example (c) (in blue). The tangent lines at points corresponding to $x = 0$ and $x = 1$ are marked in red. Following these lines we understand why the method enters a cycle when starting with $x_0 = 0$.

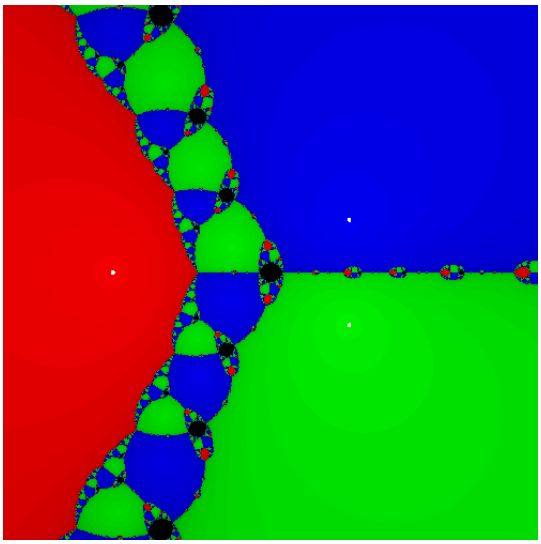


Figure 15: The domains of attraction for the three complex roots of $f(x) = x^3 - 2x + 2$ of example (c). Roots are marked with a white spot. The black zones correspond to where NM fails to converge (note that the origin is in black and the point $(1, 0)$ as well).

Remark 4.0.6 (Interesting questions). what happens if we use pseudo-inverse H^+ or regularize $(H + \varepsilon I)$; what if we have many near roots (domains of attraction); in which cases the algorithm does not converge; where the numerical computations introduce noises (approximation of ∇f, evaluation, ...); ◀

4.3.1 Multiple roots

An interesting question is: what happens if the function $f(x)$ has more than one root? Which of these will Newton's method converge to? This is actually asking for the domain of attraction of each root, that is the ensemble of starting points x_0 that will converge to the root. There are two ways to determine the domain of attraction of a root. The first one is to apply Newton's method to each different possible starting point and see which is the root it converges to. This method gives the entire domain of attraction for each root and it is used for the simulations below. A second way is to identify the region using the sufficient conditions given by the theorem that guarantee convergence (like (Theorem 4.1)). Notice that this approach could leave some points out of the different domains of attraction. A rough estimation of the domain of attraction based on (Theorem 4.1) is given in figure (Figure 17), where the corresponding domains are drawn as circles.Each root has a domain of attraction that contains a circle of radius 1/3. To estimate the radius, we noticed that (computations...) $L = 6,\ h = 3\ \beta = 2$ for $|x| \leq 1$.

Remark 4.0.7 (Third roots of unity.). A common example is the function $f(x) = x^3 - 1$, which has a real root $\alpha_{(1)} = 1$ and two complex roots $\alpha_{(2)} = e^{i\frac{2\pi}{3}} = -\frac{1}{2} + i\frac{\sqrt{3}}{2}$ and $\alpha_{(3)} = e^{i\frac{4\pi}{3}} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$. We can use Newtons formula on each point on a $2D$ complex plane. The calculated root is searched in array of roots (previously calculate with the command `roots([1 0 0 -1])`. Each root is associated to a different colour (blue, green or red in figure (Figure 17)), so that the index of array (that tells to which one of the three roots the algorithm has converged) is the used to decide the colour for the starting point.

◀

Figure (Figure 18) shows the domain of attraction of the four roots of the polynomial $f(x) = x^4 + 1$. We can see that it is very similar to the previous figure (Figure 17). Both this figure are in fact fractal figures, known as Newton fractals. This web page provides a canvas for the roots of $f(x) = x^3 - 1$ with the possibility to click anywhere in the canvas to zoom in. A static version is shown in figure (Figure 16) below (from Wikipedia).

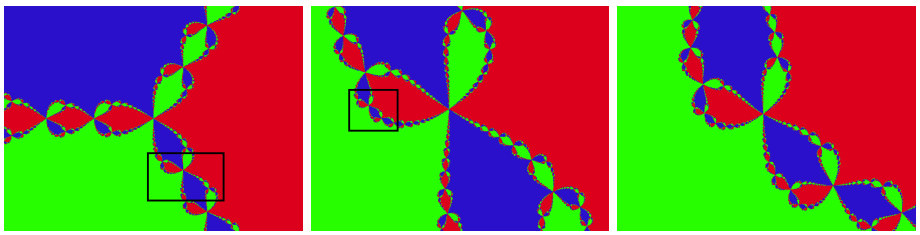


Figure 16: Successive zoom on figure showing the three domains of attraction for roots of polynomial $f(x) = x^3 - 1$.

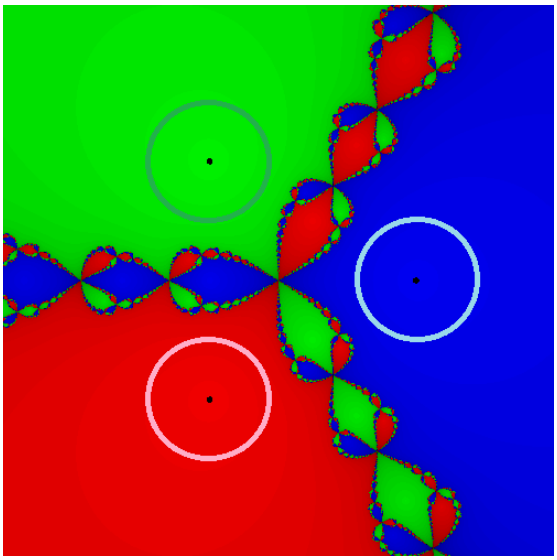


Figure 17: The three domains of attraction for the (complex) roots of the function $f(x) = x^3 - 1$. To each root is associated a colour: blue for $\alpha_{(1)} = 1$, green for $\alpha_{(2)} = e^{i\frac{2\pi}{3}}$ and red for $\alpha_{(3)} = e^{i\frac{4\pi}{3}}$. Roots are marked by black spots, the circles correspond to the estimation of the domain of attraction based on (Theorem 4.1).

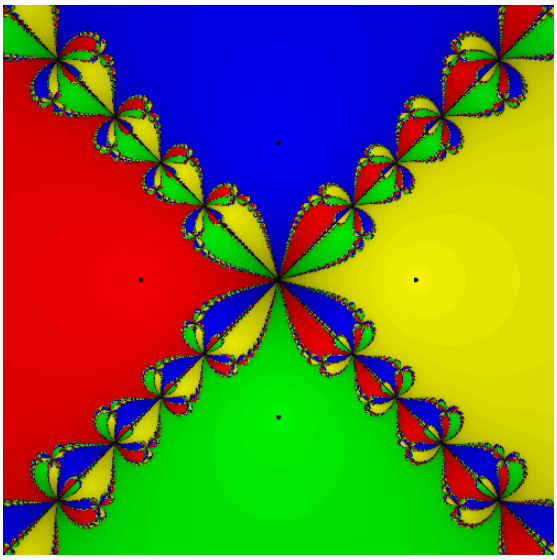


Figure 18: The four domains of attraction for the (complex) roots of the function $f(x) = x^4 + 1$. To each root is associated a colour: blue, green, red and yellow. Roots are marked by black spots.

4.4 A proof of convergence

I own the following (Theorem 4.1) and its proof to Dr Iman Shames. The following theorem gives sufficient conditions for the convergence of Newton's method (as a method for searching stationary points: $x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k)$ where $H(x)$ is the Hessian matrix of $f(x)$).

Theorem 4.1 (Newton's method convergence)**.** *Suppose:*

(i) x^* *is a stationary point for* f*:* $\nabla f(x^*) = 0$*;*

(ii) $\exists h > 0 \ \left\|H^{-1}(x^*)\right\| \leq \frac{1}{h}$

(iii) $\exists \beta,\, L > 0 \ \ \|x - x^*\| \leq \beta \Rightarrow \|H(x) - H(x^*)\| \leq L\,\|x - x^*\|$*;*

(iv) $\|x_0 - x^*\| < \gamma := \min\left(\beta, \frac{2h}{3L}\right)$

Then the iteration $x_+ = x - H^{-1}(x)\nabla f(x)$ *is such that:*

1. $\|x_+ - x^*\| \leq \|x - x^*\|^2 \frac{L}{2(h - L\|x - x^*\|)}$
2. $\|x_+ - x^*\| < \|x - x^*\| < \gamma$
3. $\|x_+ - x^*\| \leq \|x - x^*\|^2 \frac{3L}{2h}$.

Proof. We will need the two following lemmas: for $h > 0$

$$M \in \mathbb{R}^{n\times n}\ M = M^T \Rightarrow \left\|M^{-1}\right\| \leq \frac{1}{h} \iff \|Mv\| \geq h\,\|v\| \ \ \forall v \in \mathbb{R}^n \tag{29}$$

and (for a continually differentiable function)

$$F(z) - F(x) = \int_0^1 \nabla f(x + t(z - x))(z - x)dt. \tag{30}$$

Let's compute

$$\begin{aligned} x_+ - x^* &= x - x^* - H^{-1}(x)\left(\nabla f(x^*) - \nabla f(x)\right) \\ &= x - x^* - H^{-1}(x)\int_0^1 H(x + t(x^* - x))(x^* - x)dt \quad \text{using (30)} \\ &= H^{-1}(x)\left[\int_0^1 \left(H(x + t(x^* - x)) - H(x)\right)(x^* - x)dt\right]. \end{aligned}$$

So, taking norms and using the hypothesis:

$$\begin{aligned} \|x_+ - x^*\| &\leq \left\|H^{-1}(x)\right\| \int_0^1 \|H(x + t(x^* - x)) - H(x)\|\,\|x^* - x\|\,dt \\ &\leq \left\|H^{-1}(x)\right\| L\,\|x^* - x\|^2 \int_0^1 t\,dt \\ &\leq \left\|H^{-1}(x)\right\| \frac{L}{2}\,\|x^* - x\|^2. \end{aligned}$$

Now, using (29) and the properties of matrix norm:

$$\|H(x)v\| = \|H(x^*)v + (H(x)v - H(x^*)v)\| \geq h\|v\| - L\|x - x^*\|\|v\| \quad \forall v$$

so that (again by (29))

$$\|x_+ - x^*\| \leq \frac{L}{2(h - L\|x - x^*\|)}\|x - x^*\|^2$$

holds. The other inequalities follows noticing that $L\|x - x^*\| < \frac{2h}{3}$ because of hypothesis (iv). □

Remark 4.1.1. This theorem tells us that if x^* is a stationary point in which $H^{-1}(x^*)$ is well defined and H is locally Lipschitz near x^*, if we start sufficiently close to x^* then at each iteration we get closer, the convergence rate is quadratic and the constant mostly depends on H. ◀

Remark 4.1.2. The second hypothesis (ii) of the (Theorem 4.1) corresponds to the assumption that H is coercive (see property (29)). ◀

Remark 4.1.3. In the previous theorem (4.1), the convergence result is obtained with an argument fixed point-like. Indeed we show that the iteration map is a contraction. ◀

Remark 4.1.4 (What if?). One could ask what would happen if one or more assumptions are not satisfied. Consider the examples in (SUBSECTION 4.3).

(a) We saw that the method do not present particular problems in convergence. However, notice that the Lipschitz condition on $H(x) = e^x + 1$ is not satisfied globally, but only in a neighborhood of x^* (i.e. the β is a finite constant) when one can for example use the approximation $H(x) - H(x^*) = e^x - e^{x^*} \approx x - x^*$.

(b) Does not present any problem (unless one takes $x_0 = 0$), as $H(x) = -2x$ is globally 2-Lipschitz. However, condition (ii) in (Theorem 4.1) is not satisfied: $\left\|H(x^*)^{-1}\right\| = \left|\frac{1}{2x^*}\right|\Big|_{x^*=0}$ is unbounded; moreover, $\gamma = \infty$ but actually the situation is asymmetric: one cannot start farther than 1 on one side of the root.

(c) In this case $H(x) = 3x^2 - 2$ so that $H^{-1}(x^*) \approx 0.1353$ and one can take $h = 7.3$. Noticing that

$$\|H(x) - H(x^*)\| = 3|x + x^*||x - x^*| \leq 3(|x| + |x^*|)|x - x^*|$$

the constant L for the Lipschitz condition depends on $|x|$ and so we can actually fix β big as much as we want and $L = 3(\beta + |x^*|)$. Then $\gamma := \min\left(\beta, \frac{2h}{3L}\right) = \frac{2h}{3L} \approx \frac{14}{9\beta}$ becomes very small. The theorem suggests a trade-off to find between the two conditions.

(d) The coercivity (ii) and Lipschitz condition (iii) are not satisfied (as $H(x) = \frac{1}{3}x^{-2/3}$) and indeed the algorithm does not converge.

(e) $H(x)$ oscillates heavily near the origin (conditions (ii) and (iii) are not satisfied).

(f) In this case $h = 1$ satisfies condition (ii) but the Lipschitz condition is not satisfied as $x^{1/3} > x$ near the origin. The method converges despite this fact.

(g) The first condition is not satisfied if we consider only real functions. Otherwise the conditions are satisfied with $x^* = \pm i$, $h = 2 = L$ and $\gamma = \frac{2}{3}$. We can see that the domain of attraction is only a circle of radius $\frac{2}{3}$ but the maximal circle just need the radius to be less than 1 and the whole domain of attraction is a semiplane.

◀

(Theorem 4.1) can be generalized using comparison functions (to replace the Lipschitz condition (iii)).

Theorem 4.2 (Newton's method convergence 2)**.** *Suppose:*

(i) x^* *is a stationary point for* f*:* $\nabla f(x^*) = 0$*;*

(ii) H *is coercive:* $\exists h > 0 \ \left\|H^{-1}(x^*)\right\| \leq \frac{1}{h}$

(iii) $\exists \beta \in \mathcal{KL} \ \|H(x) - H(x^*)\| \leq \beta(\|x - x^*\|, k) \ \forall k \geq 0$*;*

(iv) $\|x_0 - x^*\| < \frac{2h}{3}$

Then the iteration $x_+ = x - H^{-1}(x)\nabla f(x)$ *is such that* $\|x_+ - x^*\| \leq \frac{2}{3}\frac{h}{h-\beta(\frac{2h}{3},k)}\beta(\|x - x^*\|, k) \leq \frac{4}{3}\beta(\frac{2h}{3}, k)$.

Proof. As in the previous proof, compute

$$\begin{aligned} x_+ - x^* &= x - x^* - H^{-1}(x)\left(\nabla f(x^*) - \nabla f(x)\right) \\ &= x - x^* - H^{-1}(x)\int_0^1 H(x + t(x^* - x))(x^* - x)dt \quad \text{using (30)} \\ &= H^{-1}(x)\left[\int_0^1 \left(H(x + t(x^* - x)) - H(x)\right)(x^* - x)dt\right]. \end{aligned}$$

So, taking norms and using the hypothesis, (29) and the properties of matrix norm:

$$\begin{aligned}
\|x_+ - x^*\| &\leq \left\|H^{-1}(x)\right\| \int_0^1 \|H(x + t(x^* - x)) - H(x)\| \, \|x^* - x\| \, dt \\
&\leq \left\|H^{-1}(x)\right\| \|x^* - x\| \, \beta(\|x^* - x\|, k) \int_0^1 dt \\
&\leq \frac{\|x - x^*\| \, \beta(\|x - x^*\|, k)}{h - \beta(\|x - x^*\|, k)} \\
&\leq \frac{2h}{3} \frac{1}{h - \beta(\frac{2h}{3}, k)} \beta\left(\frac{2h}{3}, k\right) \\
&\leq \frac{4}{3} \beta\left(\frac{2h}{3}, k\right).
\end{aligned}$$

In the last two lines we used that, by induction, if $\|x - x^*\| < \frac{2h}{3}$ and $\beta(\frac{2h}{3}, 0) < \frac{h}{2}$ ($\beta(\frac{2h}{3}, 0) < h$ already needed to use (29)), then $\|x_+ - x^*\| < \frac{2h}{3}$. The inequality between β and h also leads to $\frac{h}{h - \beta(\frac{2h}{3}, k)} < 2$. This guarantees convergence as $k \to \infty$. □

4.4.1 Errors in Functions, Gradients, and Hessians

In the presence of errors in functions and gradients, the problem of convergence becomes a bit more difficult. In this section we discus this briefly and we'll study the different cases more in detail in the next sections. A first significant example in which errors appears is if only functions are available and gradients and Hessians must be computed with differences. We Let consider a simple one-dimensional analysis to better understand the size of the this errors. Assume that we can only compute the function f approximately, say we compute $\hat{f} = f + \varepsilon_f$ rather than f. Then an approximation of the gradient with forward difference is

$$D_h f(x) = \frac{\hat{f}(x + h) - \hat{f}(x)}{h} = \nabla f(x) + O(h + \varepsilon_f / h)$$

and the error on the gradient is at least $\varepsilon_g = O(\sqrt{\varepsilon_f})$. Doing the same for the Hessian, its error will be $\varepsilon_H = O(\sqrt{\varepsilon_g}) = O(\varepsilon_f^{1/4})$.
From a numerical point of view, when ε_f is larger than machine round-off, this implies that using a second numerical differentiation of f to compute the Hessians will not be very accurate, and in addition it will be very expensive to compute if the Hessian is dense. Alternatively one can obtain better results with centered differences, but at a cost of twice the number of function evaluations. An approximation of $\nabla f(x)$ will be:

$$D_h f(x) = \frac{\hat{f}(x + h) - \hat{f}(x - h)}{2h} = \nabla f(x) + O(h^2 + \varepsilon_f / h)$$

so that the error in this case is $\varepsilon_g = O(\varepsilon_f^{2/3})$. This leads to an error in the Hessian matrix of $\varepsilon_H = O(\varepsilon_f^{4/9})$.

4.4.2 Error in the evaluation of the gradient

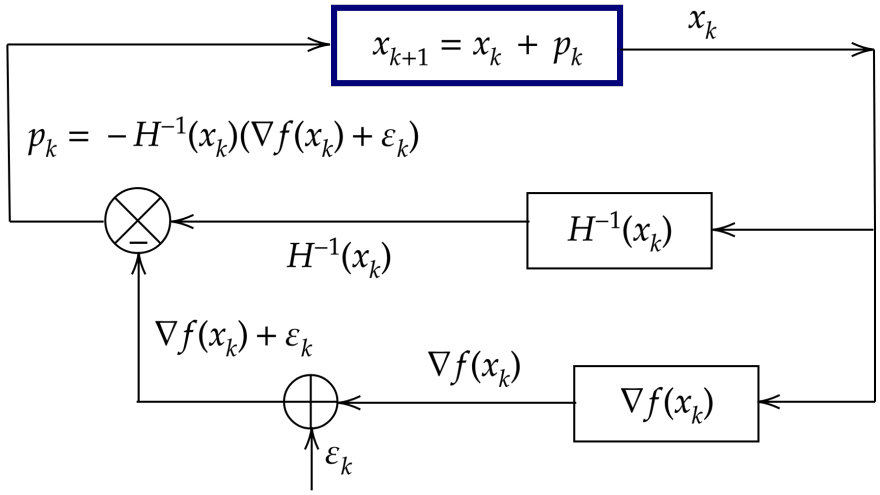


Figure 19: Diagram for the case the error is only in the evaluation of the gradient. As observed in (Remark 1.0.4), in order to prove the stability of this dynamic, we can arbitrarily decompose the iteration step, making x_{k+1} a more general function of x_k and an input u_k. This means that we can change this diagram and make it more general.

We try to understand what happens in the case that for every iteration we have some noise Δ_k due to the computation of the step update p_k such that $H(x_k)p_k = -\nabla f(x_k)$:

$$x_{k+1} = x_k + p_k + \Delta_k = x_k - H^{-1}(x_k)\nabla f(x_k) + \Delta_k. \tag{31}$$

This corresponds to the case presented in (SECTION 1) where the evaluation of the gradient is noisy (i.e. $\Delta_k = -H^{-1}(x_k)\varepsilon_k$ where ε_k is the noise on the evaluation of ∇f at x_k). Using the proof of the previous (Theorem 4.1) one obtains

$$\|x_{k+1} - x^*\| \leq c\,\|x_k - x^*\|^2 + \|\Delta_k\| \leq c^k\,\|x_0 - x^*\|^{2k} + \sum_{i=0}^{k} c^{k-i}\,\|\Delta_i\|$$

for a positive constant $c < 1$. If one requires that the method converges to a compact set (say a ball), then this result is sufficient whenever the series at the

RHS converges.
This is true, for example, if the norms $\|\Delta_i\|$ are uniformly bounded (since the constant $c < 1$): for every $i \geq 0$ $\|\Delta_i\| \leq M$. One can more precisely find the required bound on the error imposing the radius of the ball, say δ. Then in order to have $\|x_\infty - x^*\| \leq \delta$ we need:

$$\lim_{k\to\infty} \sum_{i=0}^{k} c^{k-i} \|\Delta_i\| \leq M \lim_{k\to\infty} \sum_{i=0}^{k} c^{k-i} \leq \delta$$

that is

$$\lim_{k\to\infty} c^k \frac{1-c^{-k}}{1-1/c} = \lim_{k\to\infty} \frac{c^{k+1}-c}{c-1} = \frac{c}{1-c} \leq \frac{\delta}{M}$$

which gives an upper bound for M:

$$M \leq \delta \frac{1-c}{c}.$$

We just proved the following fact:

Fact 1. Suppose that the evaluation of the gradient in the iteration step (28) of the Newton's method is noisy and call ε_k the error of this evaluation. The iteration step becomes then (31) with $\Delta_k = -H^{-1}(x_k)\varepsilon_k$. Suppose

$$\|\Delta_k\| = \left\|H^{-1}(x_k)\varepsilon_k\right\| \leq \delta \frac{1-c}{c} \quad \forall k \geq 0$$

where $c = \frac{3L}{2h} < 1$ is a constant depending on the Hessian matrix H. Then for every starting point x_0 sufficiently close to the stationary point x^* of f, [4] the algorithm convergence to a point that lies in a ball of radius δ. ♢

Remark 4.2.1. Applying the version of (Theorem 4.1) with the comparison functions (Theorem 4.2), the upper bound on the error is just the size of the desired ball. In other words, the previous fact can be restated as follows:

Fact 2. Suppose that the evaluation of the gradient in the iteration step (28) of the Newton's method is noisy and call ε_k the error of this evaluation. The iteration step becomes then (31) with $\Delta_k = -H^{-1}(x_k)\varepsilon_k$. Suppose that the assumptions of (Theorem 4.2) and

$$\|\Delta_k\| = \left\|H^{-1}(x_k)\varepsilon_k\right\| \leq \delta \quad \forall k \geq 0$$

Then for every starting point x_0 sufficiently close to the stationary point x^* of f, the algorithm convergence to a point that lies in a ball of radius δ. ♢ ◀

[4] such that it satisfies the last hypothesis (iv) of (Theorem 4.1).

Suppose otherwise that we are interested in the convergence of the method to the exact point x^*. In this case the previous argument is not enough and we need to add some additional hypothesis on the norm of the error $\|\Delta_k\|$ and on the Hessian matrix $H(x)$. Again we suppose that the uncertainty is in the evaluation of the gradient $\nabla f(x_k)$.

Theorem 4.3 (Convergence of NM with noise in ∇f)**.** *Consider the Newton iteration:*

$$x_+ = x - H^{-1}(x)\nabla f(x) + \Delta = x - H^{-1}(x)\nabla f(x) + H^{-1}(x)r. \tag{31}$$

Suppose that there exists a sequence $\{\eta_k\}$ *with* $\sup_k \eta_k < \eta \leq c < 1$ *such that*

$$\frac{\|r_k\|}{\|\nabla f(x_k)\|} \leq \eta_k < \eta. \tag{32}$$

Also suppose that there exists $\gamma > 0$ *such that*

$$m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma) \leq c < 1.$$

Then there exists $\varepsilon > 0$ *such that if* $\|x_0 - x^*\| \leq \varepsilon$ *the sequence of iterates* $\{x_k\}$ *converges to* x^**.*

Proof. Call

$$M := \|H(x^*)\| \qquad m := \left\|H^{-1}(x^*)\right\| \tag{33}$$

and observe that, by definition

$$\forall v \in \mathbb{R}^n \quad \frac{1}{m}\|v\| \leq \|H(x^*)v\| \leq M\|v\|. \tag{34}$$

Now we can choose $\varepsilon > 0$ such that for all $\|x - x^*\| \leq Mm\varepsilon$ the hypothesis

$$\|H(x) - H(x^*)\| \leq \gamma \tag{35a}$$

$$\left\|H^{-1}(x) - H^{-1}(x^*)\right\| \leq \gamma \tag{35b}$$

$$\|\nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*)\| \leq \gamma\|x - x^*\| \tag{35c}$$

yield. By induction, if $\|x_0 - x^*\| \leq \varepsilon$ then the $\|x_k - x^*\| \leq Mm\varepsilon$ and we can apply (35a), (35b) and (35c). We rewrite the iteration step:

$$x_+ = x - H^{-1}(x)\nabla f(x) + \Delta = x - H^{-1}(x)\nabla f(x) + H^{-1}(x)r \tag{36}$$

Some tricks in rewriting:

$$\begin{aligned}
x_+ - x^* &= (x - x^*) - H^{-1}(x)\nabla f(x) + H^{-1}(x)r \\
&\quad \text{we put } H^{-1}(x^*) \text{ in factor} \\
&= H^{-1}(x^*)\left(H(x^*)H^{-1}(x)r + H(x^*)H^{-1}(x)H(x)(x - x^*) - H(x^*)H^{-1}(x)\nabla f(x)\right) \\
&\quad \text{we put } H(x^*)H^{-1}(x) \text{ in factor} \\
&= H^{-1}(x^*)\left(H(x^*)H^{-1}(x)(r + H(x)(x - x^*) - \nabla f(x))\right) \\
&\quad \text{since } \nabla f(x^*) = 0 \text{ and } a = a + b - b: \\
&= H^{-1}(x^*)\big[(1 + H(x^*)H^{-1}(x) - 1)\cdot \\
&\qquad \cdot (r + H(x)(x - x^*) - \nabla f(x) + H(x^*)(x - x^*) - H(x^*)(x - x^*) + \nabla f(x^*))\big]
\end{aligned}$$

And in conclusion we obtain

$$x_+ - x^* = H^{-1}(x^*)\left[\left(1 + H(x^*)(H^{-1}(x) - H^{-1}(x^*))\right)(A - B)\right] \tag{37}$$

where

$$A = r + (H(x) - H(x^*))(x - x^*) \qquad B = \nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*).$$

Taking the norms:

$$\begin{aligned}\|x_+ - x^*\| \leq \left\|H^{-1}(x^*)\right\| \Big[&\left(1 + \|H(x^*)\| \left\|H^{-1}(x) - H^{-1}(x^*)\right\|\right) \\ &\cdot \left(\|r\| + \|H(x) - H(x^*)\| \|x - x^*\| + \|\nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*)\| \right)\Big]\end{aligned}$$

which gives the inequality

$$\|x_+ - x^*\| \leq \left\|H^{-1}(x^*)\right\| (1 + M\gamma)(\|r\| + 2\gamma \|x - x^*\|). \tag{38}$$

Now we notice that

$$\begin{aligned}\|\nabla f(x)\| &= \|H(x^*)(x - x^*) + \nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*)\| \\ &\leq \|H(x^*)(x - x^*)\| + \|\nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*)\|\end{aligned}$$

that is

$$\|\nabla f(x)\| \leq \|H(x^*)(x - x^*)\| + \gamma \|x - x^*\|. \tag{39}$$

Putting (32) and (39) into (38) one obtains

$$\begin{aligned}\|x_+ - x^*\| &\leq \left\|H^{-1}(x^*)\right\| (1 + M\gamma)(\eta \|\nabla f(x)\| + 2\gamma \|x - x^*\|) \\ &\leq \left\|H^{-1}(x^*)\right\| (1 + M\gamma)(M\eta + \eta\gamma + 2\gamma) \|x - x^*\| \\ &\leq m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma) \|x - x^*\|. \end{aligned} \tag{40}$$

The result then follows from the choice of γ. □

Remark 4.3.1. As said for (Theorem 4.1), this last result tells us that if x^* is a stationary point in which $H^{-1}(x^*)$ is well defined and we start sufficiently close to x^* so that H and H^{-1} have little variations near x^* (hypothesis (35a) and (35b)) and we have a Lipschitz-like condition (hypothesis (35c)), if we can bound the norm of the error as in (32), then the algorithm converges. Notice that in this case, the convergence rate isn't quadratic anymore. ◀

Remark 4.3.2. One could relax the assumptions on γ: choosing three different parameters in (35a), (35b), (35c) so that $m(1 + M\gamma_2)(M\eta + \eta\gamma_3 + \gamma_1 + \gamma_2) \leq c < 1$.

◀

A simple generalization of previous (Theorem 4.3) can guarantee the convergence only to a ball of radius δ rather than to the exact point x^*.

Proposition 4.4 (Practical convergence of NM with noise in ∇f)**.** *Consider the Newton iteration:*

$$x_+ = x - H^{-1}(x)\nabla f(x) + \Delta = x - H^{-1}(x)\nabla f(x) + H^{-1}(x)r. \tag{31}$$

Suppose that there exists a sequence $\{\eta_k\}$ *with* $\sup_k \eta_k < \eta \leq c < 1$ *such that*

$$\|r_k\| \leq \eta_k \|\nabla f(x_k)\| < \eta \|\nabla f(x_k)\| + \tilde{\delta}.$$

Also suppose that there exists $\gamma > 0$ *such that*

$$m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma) \leq c < 1.$$

Then there exists $\varepsilon > 0$ *such that if* $\|x_0 - x^*\| \leq \varepsilon$ *the sequence of iterates* $\{x_k\}$ *converges to a ball of center* x^* *and radius* $\delta = \frac{m(1+M\gamma)}{1-m(1+M\gamma)(M\eta+\eta\gamma+2\gamma)}\tilde{\delta}$.

Proof. Since the estimation on the norm of the error is

$$\|r_k\| \leq \eta_k \|\nabla f(x_k)\| < \eta \|\nabla f(x_k)\| + \tilde{\delta},$$

then from inequality (38) follows

$$\begin{aligned}
\|x_+ - x^*\| &\leq \left\|H^{-1}(x^*)\right\| (1 + M\gamma)(\|r\| + 2\gamma \|x - x^*\|) \\
&\leq \left\|H^{-1}(x^*)\right\| (1 + M\gamma)(\eta \|\nabla f(x)\| + \tilde{\delta} + 2\gamma \|x - x^*\|) \\
&\leq m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma) \|x - x^*\| + m(1 + M\gamma)\tilde{\delta} \\
&=: c_1 \|x - x^*\| + c_2\tilde{\delta}.
\end{aligned}$$

Now taking $\tilde{\delta} = \dfrac{1 - c_1}{c_2}\delta = \dfrac{1 - m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma)}{m(1 + M\gamma)}\delta$, when $k \to \infty$, since $c_1 < 1$:

$$\|x_{k+1} - x^*\| \leq c_1^{k+1} \|x_0 - x^*\| + \frac{c_2}{1 - c_1}\tilde{\delta} \Rightarrow \|x_\infty - x^*\| \leq \delta.$$

□

4.4.3 Error in the evaluation of the gradient and of the inverse Hessian matrix

We will try to do the same thing, that is proving the convergence of Newton's method to the stationary point x^* of f in the case the update $p_k = -H^{-1}(x_k)\nabla f(x_k)$ is noisy both because of error in the evaluation of the gradient $\nabla f(x_k)$ and in the inversion of the Hessian matrix H. For the sake of simplicity, we change the notation of (SECTION 1) as follows: we call the error on the gradient ε_k and we write the error on the Hessian matrix as

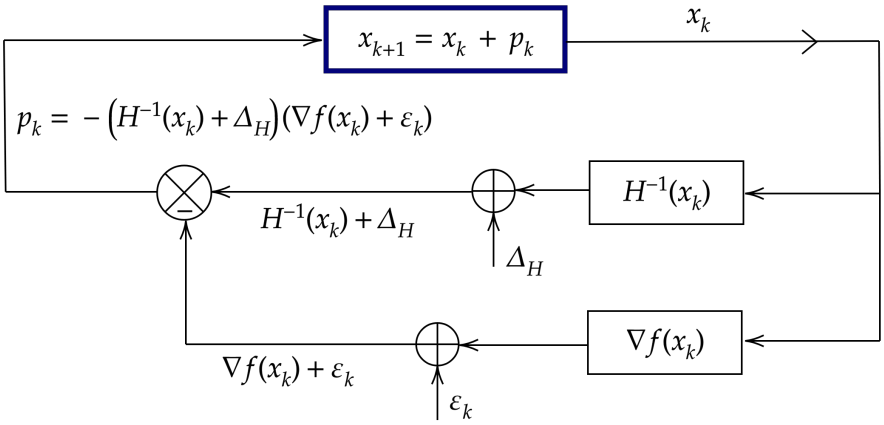


Figure 20: Diagram for the case the error is in the evaluation of the gradient and of the inverse of Hessian matrix. As observed in (Remark 1.0.4), in order to prove the stability of this dynamic, we can arbitrarily decompose the iteration step, making x_{k+1} a more general function of x_k and an input u_k. This means that we can change this diagram and make it more general

$\Delta_H = -H^{-1}(x_k)\varepsilon_H H^{-1}(x_k) + \zeta_k$. We suppose then that it takes the form $p_k = -(H^{-1}(x_k) + \Delta_H)(\nabla f(x_k) + \varepsilon_k)$.

The idea is again to write

$$x_+ - x^* = x - x^* - H^{-1}(x)\nabla f(x) - H^{-1}(x)\varepsilon - \Delta_H \nabla f(x) - \Delta_H \varepsilon$$

and try to control the mixed terms with stronger hypothesis on H and ∇f. The term $H^{-1}(x)\varepsilon$ should be treated as the term $H^{-1}(x)r$ in the previous (Theorem 4.3). Using (39) if we suppose Δ_H small enough, we should be able to control $\Delta_H \nabla f(x)$ too. Then $\|\Delta_H \varepsilon\| \leq \|\Delta_H\| \, \|\nabla f(x)\| \, \eta$ and with a smaller γ the constant should still be < 1. Renaming

$$r = \varepsilon \qquad s = \Delta_H \tag{41}$$

for the sake of simplicity, we can state the following (Theorem 4.5).

Theorem 4.5 (Convergence of NM with error in ∇f and H)**.** *Consider Newton's method iteration step*

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) - H^{-1}(x_k)r_k - s_k \nabla f(x_k) - s_k r_k. \tag{42}$$

Suppose that there exists a sequence $\{\eta_k\}$ *with* $\sup_k \eta_k < \eta \leq c < 1$ *such that*

$$\frac{\|r_k\|}{\|\nabla f(x_k)\|} \leq \eta_k < \eta. \tag{32}$$

Also suppose that there exists $\gamma > 0$ such that

$$m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma + M\eta(1 + \eta)(M + \gamma)) \leq c < 1.$$

If in addition

$$\|s_k\| \leq \eta \leq c < 1 \tag{43}$$

then there exists $\varepsilon > 0$ such that if $\|x_0 - x^*\| \leq \varepsilon$ the sequence of iterates $\{x_k\}$ converges to x^*.

Proof. Rewrite the iteration step and obtain something similar to (37):

$$x_+ - x^* = H^{-1}(x^*)\left[\left(1 + H(x^*)(H^{-1}(x) - H^{-1}(x^*))\right)(A - B - C)\right] \tag{44}$$

where

$$\begin{aligned} A &= (H(x) - H(x^*))(x - x^*) - r \\ B &= \nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*) \\ C &= H(x)s(\nabla f(x) + r). \end{aligned}$$

Taking the norms, after some computations one obtains something similar to (40):

$$\|x_+ - x^*\| \leq m(1+M\gamma)(\eta(M+\gamma)+2\gamma+M(1+\eta)\|s\|(M+\gamma))\|x - x^*\|. \tag{45}$$

Thanks to the hypothesis (43) this becomes

$$\|x_+ - x^*\| \leq m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma + M\eta(1 + \eta)(M + \gamma))\|x - x^*\|. \tag{46}$$

Now, since $\eta \leq c < 1$, if we can choose γ small enough so that

$$m(1 + M\gamma)(M\eta + \eta\gamma + 2\gamma + M\eta(1 + \eta)(M + \gamma)) \leq c < 1$$

the result follows. □

We can try to find conditions on r and s in order to obtain practical convergence to a ball in the general case. Suppose that

$$\|r\| \leq \eta_1 \|\nabla f(x)\| + \delta_1 \qquad \|s\| \leq \eta_2 \left\|H^{-1}(x)\right\| + \delta_2$$

and that the point x is close enough to x^* so that the three conditions

$$\begin{aligned} \|H(x) - H(x^*)\| &\leq \gamma_1 \\ \left\|H^{-1}(x) - H^{-1}(x^*)\right\| &\leq \gamma_2 \\ \|\nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*)\| &\leq \gamma_3 \|x - x^*\| \end{aligned}$$

hold. As above, call

$$M := \|H(x^*)\| \qquad m := \left\|H^{-1}(x^*)\right\|$$

and notice that

$$\|\nabla f(x)\| \leq (M + \gamma_3) \|x - x^*\|.$$

Then one can compute:

$$\begin{aligned} x_+ - x^* &= H^{-1}(x^*) \left[\left(1 + H(x^*)(H^{-1}(x) - H^{-1}(x^*))\right)(A - B) - C \right] \\ &\quad \text{with} \\ A &= (H(x) - H(x^*))(x - x^*) - r \\ B &= \nabla f(x) - \nabla f(x^*) - H(x^*)(x - x^*) \\ C &= H(x^*)s(\nabla f(x) + r). \end{aligned}$$

Taking the norms one obtains

$$\begin{aligned} \|x_+ - x^*\| \leq\ & m(1 + M\gamma_2)(M\eta_1 + \eta_1\gamma_3 + \gamma_1 + \gamma_3 + mM^2\eta_2 + M^2\eta_2\gamma_2 + M^2\delta_2 + \\ & + mM\eta_2\gamma_3 + M\eta_2\gamma_2\gamma_3 + M\gamma_3\delta_2 + mM^2\eta_1\eta_2 + M^2\eta_1\eta_2\gamma_2 + \\ & + M^2\eta_1\delta_2 + mM\eta_1\eta_2\gamma_3 + M\eta_1\eta_2\gamma_2\gamma_3 + M\eta_1\gamma_3\delta_2) \|x - x^*\| \\ & + m(1 + M\gamma_2)(1 + mM\eta_2 + M\eta_2\gamma_2 + M\delta_2)\delta_1 \\ =:\ & c \|x - x^*\| + d \end{aligned}$$

So that, for $\eta_1, \eta_2 < 1$ and x close enough to x^* one can find positive γ_1, γ_2 and γ_3 such that $c < 1$ and

$$\|x_{k+1} - x^*\| \leq c^{k+1} \|x_0 - x^*\| + d \sum_{i=0}^{k} c^i$$

converges

$$\|x_\infty - x^*\| \leq \frac{d}{1 - c}$$

that is practical convergence of the sequence $\{x_k\}_k$ to a ball of radius $\delta := \dfrac{d}{1 - c}$.

Remark 4.5.1. In order to look for conditions that yield $c < 1$, one could study c as function of $\gamma_1, \gamma_2, \gamma_3$ under the constraint that the three gammas are positive. However, writing the Lagrangian and imposing the derivative with respect to γ_1 to vanish, one obtains $m(1 + M\gamma_2) = 0$ that means γ_2 should be negative, thus there are not admissible solutions. ◀

Remark 4.5.2. An easy lower bound $\underline{c}$ for c is obtained imposing the three gammas to be 0:

$$c > \underline{c} := m(M\eta_1 + mM^2\eta_2 + M^2\delta_2 + mM^2\eta_1\eta_2 + M^2\eta_1\delta_2).$$

This can never happen because of their definition, so the inequality is strict. ◀

Remark 4.5.3. If we change the two definitions of γ_1 and γ_2 to be some sort of Lipschitz constants:

$$\|H(x) - H(x^*)\| \leq \gamma_1 \|x - x^*\|$$
$$\left\|H^{-1}(x) - H^{-1}(x^*)\right\| \leq \gamma_2 \|x - x^*\|$$

then in the end we find

$$\|x_+ - x^*\| \leq c_1 \|x - x^*\| + c_2 \|x - x^*\|^2 + c_3 \|x - x^*\|^3 + d$$

where

$$\begin{aligned}
c_1 &= m\left(M\left(M((1+\delta_1)(\delta_2 + m\eta_2) + \eta_1(\eta_2 + \delta_2)) + \gamma_3(1+\eta_1)(\delta_2 + m\eta_2) + \eta_1 + \gamma_2\delta_1(1+\eta_2)\right) + \gamma_3(1+\eta_1)\right)\\
c_2 &= m\left(\gamma_1 + M\gamma_2\left((\eta_1 + \eta_2 + \eta_1\eta_2 + M\delta_2)(\gamma_3 + M) + mM(\eta_1(1+\delta_1) + \eta_2(M\eta_1 + \gamma_3)) + \gamma_3\right)\right)\\
c_3 &= mM\gamma_2(\gamma_1 + M\eta_2\gamma_2(1+\eta_1)(M+\gamma_3)))\\
d &= m(1 + M(m\eta_2 + \delta_2))\delta_1
\end{aligned}$$

are positive constants. If we rewrite the inequality as

$$\|x_+ - x^*\| \leq \underbrace{\left(c_1 + c_2 \|x - x^*\| + c_3 \|x - x^*\|^2\right)}_{c} \|x - x^*\| + d$$

Sufficient conditions for a convergence to a ball of radius $\dfrac{d}{1-c}$ are

$$\begin{cases} c_2^2 > 4(c_1 - 1)c_3 \\ \max(0, \underline{z}) < \|x_k - x^*\| < \overline{z} \quad \forall k \\ \overline{z} > 0 \end{cases}$$

with

$$\underline{z} := \frac{-c_2 - \sqrt{c_2^2 - 4(c_1-1)}}{2c_3} < \frac{-c_2 + \sqrt{c_2^2 - 4(c_1-1)}}{2c_3} =: \overline{z}.$$

The first condition guarantees that $c < 1$ when each step is close enough to the stationary point x^* the sequence will converge; the second condition control how much the steps should be close; the third condition just ensures that the interval on the second condition is not empty.
Another sufficient condition would be imposing that all the three constants are smaller than 1:

$$c_1 < 1 \quad c_2 < 1 \quad c_3 < 1$$

and the radius of the ball would be $\delta = \delta(c_1, c_2, c_3, d)$.
A third approach that leads to sufficient conditions makes use of a general theorem of stability for nonlinear first-order recurrences. Let $e_k = x_k - x^*$. From the theory of nonlinear first-order recurrences, we know that

$$e_{k+1} = g(e_k) := c_1 e_k + c_2 e_k^2 + c_3 e_k^3 + d$$

is locally stable, meaning that it converges to the fixed point $e = \overline{e}$ (that corresponds to $x = x^*$) from points sufficiently close to it (thus x sufficiently close to x^*), if the slope of g in the neighborhood of $\overline{e}$ is smaller than unity in absolute value: that is,

$$|g'(\overline{e})| = |3c_3\overline{e} + 2c_2\overline{e} + c_1| < 1.$$

This condition leads to three possible cases:

$$\text{(a)} \quad \begin{cases} \dfrac{-c_2 - \sqrt{c_2^2 - 3c_3(c_1 - 1)}}{3c_3} < \overline{e} < \dfrac{-c_2 + \sqrt{c_2^2 - 3c_3(c_1 - 1)}}{3c_3} \\ \left| \frac{c_2^2}{3c_3} - c_1 \right| < 1 \end{cases}$$

$$\text{(b)} \quad \begin{cases} \dfrac{-c_2 - \sqrt{c_2^2 - 3c_3(c_1 - 1)}}{3c_3} < \overline{e} < \dfrac{-c_2 - \sqrt{c_2^2 - 3c_3(c_1 + 1)}}{3c_3} \\ c_2^2 \geq 3c_3(c_1 + 1) \end{cases}$$

$$\text{(c)} \quad \begin{cases} \dfrac{-c_2 + \sqrt{c_2^2 - 3c_3(c_1 + 1)}}{3c_3} < \overline{e} < \dfrac{-c_2 + \sqrt{c_2^2 - 3c_3(c_1 - 1)}}{3c_3} \\ c_2^2 \geq 3c_3(c_1 + 1) \end{cases}$$

and we recall that a zero of a cubic function can be written as an expression in its coefficients as:

$$\overline{e} = -\frac{1}{3c_3}\left(c_2 + C + \frac{\Delta_0}{C}\right)$$

with

$$\Delta_0 = c_2 - 3c_3c_1 \qquad \Delta_1 = 2c_2^3 - 9c_1c_2c_3 + 27c_1^2d$$

$$C = \sqrt[3]{\frac{\Delta_1 \pm \sqrt{\Delta_1^2 - 4\Delta_0^3}}{2}}.$$

Condition $|g'(\overline{e})| < 1$ could be satisfied for either one, two or three roots of the function g. This means that local convergence is guaranteed but not to a specific root (domains of attraction could be chaotic: see figure (Figure 17)). ◀

There is another result that guarantees practical convergence of the Newton's Method to a ball: in [17, Theorem 2.3.4, p. 18] we find the following theorem.

Theorem 4.6 (Estimation on the error for NM)**.** *Let x^* be a stationary point for f: $\nabla f(x^*) = 0$. Suppose that the Hessian is γ-Lipschitz and $H(x^*)$ is positive definite. Then there are $K > 0, \delta > 0$, and $\delta_1 > 0$ such that if $\|x - x^*\| \leq \delta$ and $\|\varepsilon_H\| < \delta_1$ then $H(x_+) + \varepsilon_H$ is non-singular and after one step*

$$x_+ = x - (H(x) + \varepsilon_H)^{-1}(\nabla f(x) + \varepsilon)$$

the error satisfies

$$\|x_+ - x^*\| \leq K\left(\|x - x^*\|^2 + \|\varepsilon_H\|\,\|x - x^*\| + \|\varepsilon\|\right) \tag{47}$$

with

$$\begin{aligned} K &= (4+\gamma)\left\|H^{-1}(x^*)\right\| + 16\left\|H^{-1}(x^*)\right\|^2 \|H(x^*)\| \\ &= \left(4+\gamma+16\left\|H^{-1}(x^*)\right\| \|H(x^*)\|\right)\left\|H^{-1}(x^*)\right\|. \end{aligned}$$

As a consequence, one cannot hope to find a minimizer with more accuracy that one can evaluate ∇f and in most cases the iteration will stagnate once $\|x - x^*\|$ is (roughly) the same size as ε. The speed of convergence will be governed by the accuracy in the Hessian.

Using this theorem we can find sufficient conditions on the coefficient of the recurrence equation (47) in order to keep the norm of the error smaller than a fixed value ζ. The equation

$$e_{k+1} = g(e_k) := Ke_k^2 + K\|\varepsilon_H\| e_k + K\|\varepsilon\| \tag{47}$$

has two fix points $\overline{e}_\pm = \dfrac{1 - K\|\varepsilon_H\| \pm \sqrt{(K\|\varepsilon_H\| - 1)^2 - 4K^2\|\varepsilon\|}}{2K}$. The stability conditions are given by $\|g'(\overline{e})\| < 1$ which is possible if and only if the discriminant is positive and the fix point is $\overline{e} = \overline{e}_-$, that is:

$$\begin{cases} \|g'(\overline{e}_-)\| = \left\|1 - \sqrt{(K\|\varepsilon_H\| - 1)^2 - 4K^2\|\varepsilon\|}\right\| < 1 \\ (K\|\varepsilon_H\| - 1)^2 - 4K^2\|\varepsilon\| > 0 \end{cases}$$

that means

$$0 < (K\|\varepsilon_H\| - 1)^2 - 4K^2\|\varepsilon\| < 1$$

$$\Rightarrow \begin{cases} 0 < K < 2\frac{\|\varepsilon_H\| + 2\|\varepsilon\|}{\|\varepsilon_H\|^2} \\ K < \frac{\|\varepsilon_H\| + 2\|\varepsilon\| - 2\sqrt{\|\varepsilon\|(\|\varepsilon\| + \|\varepsilon_H\|}}{\|\varepsilon_H\|^2} \end{cases} \quad \vee \quad \begin{cases} 0 < K < 2\frac{\|\varepsilon_H\| + 2\|\varepsilon\|}{\|\varepsilon_H\|^2} \\ K > \frac{\|\varepsilon_H\| + 2\|\varepsilon\| + 2\sqrt{\|\varepsilon\|(\|\varepsilon\| + \|\varepsilon_H\|}}{\|\varepsilon_H\|^2} \end{cases}$$

and since K is always positive the conditions become

$$K < \frac{\|\varepsilon_H\| + 2\|\varepsilon\| - 2\sqrt{\|\varepsilon\|(\|\varepsilon\| + \|\varepsilon_H\|}}{\|\varepsilon_H\|^2} \tag{♣1}$$

or

$$2\frac{\|\varepsilon_H\| + 2\|\varepsilon\|}{\|\varepsilon_H\|^2} > K > \frac{\|\varepsilon_H\| + 2\|\varepsilon\| + 2\sqrt{\|\varepsilon\|(\|\varepsilon\| + \|\varepsilon_H\|}}{\|\varepsilon_H\|^2}. \tag{♣2}$$

The following proposition summarizes the considerations above.

Proposition 4.7 (Practical convergence of NM with noise in ∇f & H)**.** *Suppose that the Hessian is γ-Lipschitz and $H(x^*)$ is positive definite. Let δ, δ_1 and K given by (Theorem 4.6). If K satisfies one of the two conditions (♣1) or (♣2) then for every $\eta > 0$ the Newton's Method with iteration step*

$$x_{k+1} = x_k - (H(x_k) + \varepsilon_H)^{-1}(\nabla f(x_k) + \varepsilon_k)$$

stays in a ball of radius $\zeta := \eta + \overline{e}_-$ for k sufficiently large.

4.5 An approximation of Newton's method

The following (Theorem 4.8) and its proof are a reformulation of a theorem showed me by Dr Iman Shames.
Instead of considering the Newton's method iteration (28), we consider

$$x_{k+1} = x_k - F_k^{-1} \nabla f(x_k) \tag{48}$$

where F_k^{-1} is a local approximation of the inverse Hessian matrix $H^{-1}(x_k)$. Then the following theorem yields.

Theorem 4.8 (Local convergence of Newton type updates)**.** *Let x^* a stationary point of f and consider the iteration (48). Suppose there exist bounded positive scalars ω and δ where $\delta \leq 1$ and a sequence $\{\delta_k\}$ where $\sup_k\{\delta_k\} < \delta$ such that for all x and x_k*

(i) $\|x_0 - x^*\| \leq \frac{2(1-\delta)}{\omega}$; *(proximity of the initial guess)*

(ii) $\left\|F_k^{-1}(H(x_k) - H(x))\right\| \leq \omega \|x_k - x\|$; *(Lipschitz condition)*

(iii) $\left\|F_k^{-1}(H(x_k) - F_k)\right\| \leq \delta_k < \delta$. *(quality of approximation)*

Then the sequence $\{x_k\}$ converges to x^*.

Proof. Compute

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - F_k^{-1}\nabla f(x_k) \\ &= F_k^{-1}\left(F_k(x_k - x^*) - \int_0^1 H(x^* + t(x_k - x^*))(x_k - x^*)dt\right) \\ &= F_k^{-1}\left((F_k - H(x_k))(x_k - x^*) - \int_0^1 (H(x^* + t(x_k - x^*)) - H(x_k))(x_k - x^*)dt\right). \end{aligned}$$

In the light of conditions (ii) and (iii), one obtains

$$\begin{aligned} \|x_{k+1} - x^*\| &\leq \delta_k \|x_k - x^*\| + \int_0^1 \omega \|x^* + t(x_k - x^*) - x_k\| \|x_k - x^*\| dt \\ &\leq \left(\delta_k + \frac{\omega}{2}\|x_k - x^*\|\right)\|x_k - x^*\| \\ &\leq (\delta_k + 1 - \delta)\|x_k - x^*\| \end{aligned}$$

Due to (i) and the fact that $\sup_k\{\delta_k\} < \delta \leq 1$ this results in convergence. □

Remark 4.8.1. We could "explicit" the approximation, that is write $F_k^{-1} = H^{-1}(x_k) + s_k$. The three assumptions become:

(i) $\|x_0 - x^*\| \leq \frac{2(1-\delta)}{\omega}$; (proximity of the initial guess)

(*ii*) $\left\|1 - H^{-1}(x_k)H(x) + s_k(H(x_k) - H(x))\right\| \leq \omega \|x_k - x\|$; (Lipschitz condition)

(*iii*) $\|s_k H(x_k)\| \leq \delta_k < \delta$. (quality of approximation)

The iteration step of Newton's method corresponding to this case is

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) - s_k \nabla f(x_k)$$

which is slightly different to the hypothesis of (Fact 1): there is no error in the evaluation of the gradient, but there is some error in the evaluation of the Hessian matrix. In particular this is a special case of (Theorem 4.5) when taking $r_k = 0$. ◀

As done before for (Theorem 4.1), the previous (Theorem 4.8) can be easily generalized to the case of an extra error ε_k in the iteration step, that is every iteration step is in the form

$$x_{k+1} = x_k - F_k^{-1}\nabla f(x_k) + \varepsilon_k,$$

in order to guarantee convergence to a ball of radius ζ. This is more precisely stated in the following propostion.

Proposition 4.9 (Practical convergence of NM approximation)**.** *Consider the approximation (48) of Newton's method iteration step and suppose that it is noisy. Call this noise ε_k so that the iteration step is in the form*

$$x_{k+1} = x_k - F_k^{-1}\nabla f(x_k) + \varepsilon_k$$

Suppose there exists bounded positive scalars ω and $\delta \leq 1$ and a sequence $\sup_k\{\delta_k\} < \delta$ such that for all x and x_k the three assumptions (i), (ii), (iii) of (Theorem 4.8) are satisfied. Call $c := \delta - \sup_k\{\delta_k\} < 1$ and suppose also that the noise is bounded: $\sup_k \|\varepsilon_k\| \leq \zeta$. Then the algorithm converges to a ball of radius $\frac{\zeta}{c}$ when $k \to \infty$.

Proof. Compute as in the proof of (Theorem 4.8):

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq (1 - \delta + \delta_k)\|x_k - x^*\| + \|\varepsilon_k\| \\
&\leq (1-c)^{k+1}\|x_0 - x^*\| + \zeta \sum_{i=0}^{k}(1-c)^i \\
&\leq (1-c)^{k+1}\|x_0 - x^*\| + \zeta \frac{1}{1-(1-c)} \\
&\overset{k\to\infty}{\longrightarrow} \frac{\zeta}{c}.
\end{aligned}$$

□

4.6 Quasi-Newton methods

Quasi-Newton methods are generalizations of the secant method to find the root of the first derivative for multidimensional problems. In quasi-Newton methods, the Hessian matrix of second derivatives is not computed. Instead, only first order information is used and the Hessian matrix is approximated using updates specified by gradient evaluations (or approximate gradient evaluations). So, quasi-Newton methods differ in how they update the approximate Hessian. According to [13] «the earliest, and certainly one of the most clever schemes for constructing the inverse Hessian, was originally proposed by Davidon and later developed by Fletcher and Powell.»

The Davidon–Fletcher–Powell (**DFP**) and the Broyden–Fletcher–Goldfarb–Shanno (**BFGS**) algorithms are iterative methods for solving unconstrained nonlinear optimization problems.

DFP Algorithm. From an initial guess x_0 and an approximate inverse of the Hessian matrix B_k^{DFP} the following steps are repeated as x_k converges to the solution:

DFP Algorithm

1. Obtain a direction d_k by solving $d_k = -B_k^{DFP}\nabla f(x_k)$.
2. Find an acceptable stepsize α_k in the direction found in the first step, so $\alpha_k = \arg\min_{\alpha} f(x_k + \alpha d_k)$.
3. Set $p_k = \alpha_k d_k$ and update $x_{k+1} = x_k + p_k$.
4. $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
5. $B_{k+1}^{DFP} = B_k^{DFP} - \dfrac{B_k^{DFP} q_k q_k^T B_k^{DFP}}{q_k^T B_k^{DFP} q_k} + \dfrac{p_k p_k^T}{q_k^T p_k}$.

BFGS Algorithm. The BFGS formula is the *dual* or *complementary* of the DFP formula (that means one only needs to interchanging the roles of q and p in the two updating formulas for Hessian approximation). This comes from the fact that the two equations

$$H^{-1}(x)\nabla f(x) = d$$

and

$$\nabla f(x) = H(x)d$$

have exactly the same form. So one can approximate the Hessian itself rather than its inverse. The corresponding update to the Hessian approximation $F_k = (B_k^{DFP})^{-1}$ is given by

$$F_{k+1} = F_k + \frac{q_k q_k^T}{q_k^T p_k} - \frac{F_k p_k p_k^T F_k}{p_k^T F_k p_k} \tag{49}$$

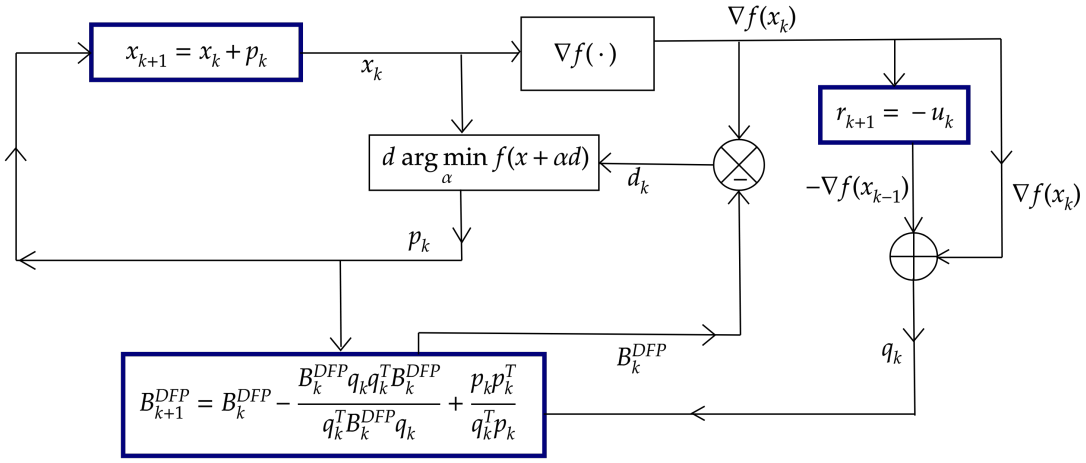


Figure 21: A diagram summarizing the DFP algorithm.

where again $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. The BFGS algorithm uses the inverse of the matrix F_k, which can be obtained efficiently by applying the Sherman–Morrison formula (65d) to the previous updating step (49), giving (with $B_k^{BFGS} = F_k^{-1}$)

$$B_{k+1}^{BFGS} = \left(I - \frac{p_k q_k^T}{q_k^T p_k}\right) B_k^{BFGS} \left(I - \frac{q_k p_k^T}{q_k^T p_k}\right) + \frac{p_k p_k^T}{q_k^T p_k}. \tag{50}$$

recognizing that B_k^{BFGS} is symmetric, and that $q_k^T B_k^{BFGS} q_k$ and $p_k^T q_k$ are scalars, we can expand this update to compute B_k^{BFGS} efficiently. The explicit derivation of (50) from (49) can be found in (Appendix A.2). The entire algorithm is summarized in the following steps: from an initial guess x_0 and an approximate Hessian matrix B_0 the following steps are repeated as x_k converges to the solution:

BFGS Algorithm

1. Obtain a direction d_k by $d_k = -B_k^{BFGS} \nabla f(x_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable step-size α_k in the direction found in the first step, so $\alpha_k = \arg\min_\alpha f(x_k + \alpha d_k)$.
3. Set $p_k = \alpha_k d_k$ and update $x_{k+1} = x_k + p_k$.
4. $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
5. $B_{k+1}^{BFGS} = B_k^{BFGS} + \frac{(p_k^T q_k + q_k^T B_k^{BFGS} q_k)(p_k p_k^T)}{(p_k^T q_k)^2} - \frac{B_k^{BFGS} q_k p_k^T + p_k q_k^T B_k^{BFGS}}{p_k^T q_k}$.

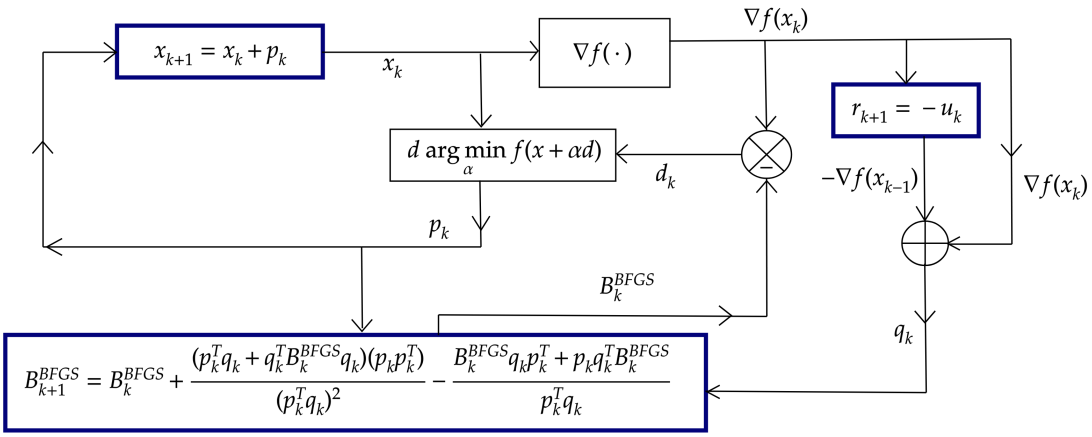


Figure 22: A diagram summarizing the BFGS algorithm.

Remark 4.9.1. The last step of (BFGS Algorithm) - which is the expansion of (50) - can be also rewritten as

$$B_{k+1}^{BFGS} = B_k^{BFGS} + \left(1 + \frac{q_k^T B_k^{BFGS} q_k}{p_k^T q_k}\right) \frac{p_k p_k^T}{p_k^T q_k} - \frac{B_k^{BFGS} q_k p_k^T + p_k q_k^T B_k^{BFGS}}{p_k^T q_k}$$

◀

One of the advantages of the BFGS algorithm is that the matrix B_k does not appear in the denominator of its updating step. A proof of the convergence in a finite number of step for the DFP method can be find in [13, p. 292]. This proof is given assuming f quadratic with positive definite Hessian and seeing the DFP as a particular conjugate direction method. The same proof holds for BFGS under the same assumptions.

However, even for general nonquadratic functions, these two methods offer several advantages: *(i)* they require only that first-order information be available (the Hessian matrix does not need to be computed); *(ii)* the directions generated can always be guaranteed to be directions of descent by arranging for B_k^{DFP} to be positive definite throughout the process; *(iii)* since for a quadratic problem the matrices B_k^{DFP} converge to the inverse Hessian in at most n steps, convergence will be superlinear.

Remark 4.9.2. The (Theorem 4.8) can also be applied to both DFP and BFGS method to prove their convergence. Indeed, if we suppose $\alpha_k = 1$ in the prevoius algorithms, their iteration steps become

$$x_{k+1}^{DFP} = x_k^{DFP} - B_k^{DFP} \nabla f(x_k^{DFP})$$

and

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

B_k (for both DFP and BFGS algorithms) plays the role of F_k^{-1} in (Theorem 4.8), so if it satisfies the assumptions (some Lipschitz condition and a good quality of approximation) the theorem can be directly applied. However, the assumption $\alpha_k = 1$ is quite strong and one should then find the neighborhood of trajectories where this stays true. ◀

Attempt to apply (Theorem 4.8) to DFP algorithm. In this case, $B_k \equiv B_k^{DFP}$ plays the role of F_k^{-1}, so we do not need to express the inverse in a clever way from the update step. To make things simpler: suppose that $H(x) \equiv H$ is constant. Then the Lipschitz condition (ii) is always satisfied, since it becomes

$$\left\| F_k^{-1}(H(x_k) - H(x)) \right\| = \left\| F_k^{-1}(H - H) \right\| = 0 \leq \omega \left\| x_k - x \right\|.$$

Now, try to use induction to see if B_k satisfies condition (iii). Suppose that $\|B_k H - I\| \leq \delta_k$. Using the updating step, the condition to prove becomes

$$\begin{aligned} \|B_{k+1}H - I\| &= \left\| B_k H - \frac{1}{a_k} B_k q_k q_k^T B_k H + \frac{1}{b_k} p_k p_k^T H - I \right\| \\ &= \left\| \frac{1}{a_k b_k} \left(a_k b_k - b_k B_k q_k q_k^T - a_k B_k \nabla f(x_k) \nabla f(x_k)^T \right) B_k H - I \right\| \\ &\leq \delta_{k+1} \end{aligned}$$

where we supposed $\alpha_k = 1$ so that $p_k = d_k = -B_k \nabla f(x_k)$ and we have set

$$a_k = q_k^T q_k \qquad \text{and} \qquad b_k = -q_k^T B_k \nabla f(x_k).$$

The inequality is true as long as $B_k H$ gets closer to the identity matrix, i.e. if the quality of the approximation doesn't get worse.

Remark 4.9.3. As emphasized in figure (Figure 22), the BFGS algorithm is composed by two different dynamics, one involving the sequence x_k and the other involving B_k. This two dynamics are interconnected so one may think about proving the ISS property for each dynamic (with respect to an input and respectively) and then apply (Theorem 2.11) to obtain ISS property for the whole system. However, it is clear than the dynamic

$$x_{k+1} = x_k + p_k$$

is not ISS with input p_k (because if we take a vanishing input the dynamic is unstable). So to have a chance of applying a result like (Theorem 2.11) one need to split out the system in a different and less obvious way. ◀

We can summarize the two algorithms in a more general frame of plant, controller and estimator. This is precisely done in the following figure (Figure 23):

- the plant corresponds to the integrator that establish the dynamic of x_k plus a static map that computes $\nabla f(x_k)$;
- the estimator uses this two information to build a matrix B_k that approximate the inverse of the Hessian matrix, thus it contains the dynamic for B_k and the memory of $\nabla f(x_{k-1})$;
- the controller uses all the information from the plant and the estimator to compute α_k (with an exact or inexact line search) and multiplies it by the direction obtained by the gradient to provide a feedback p_k to the plant and the estimator.

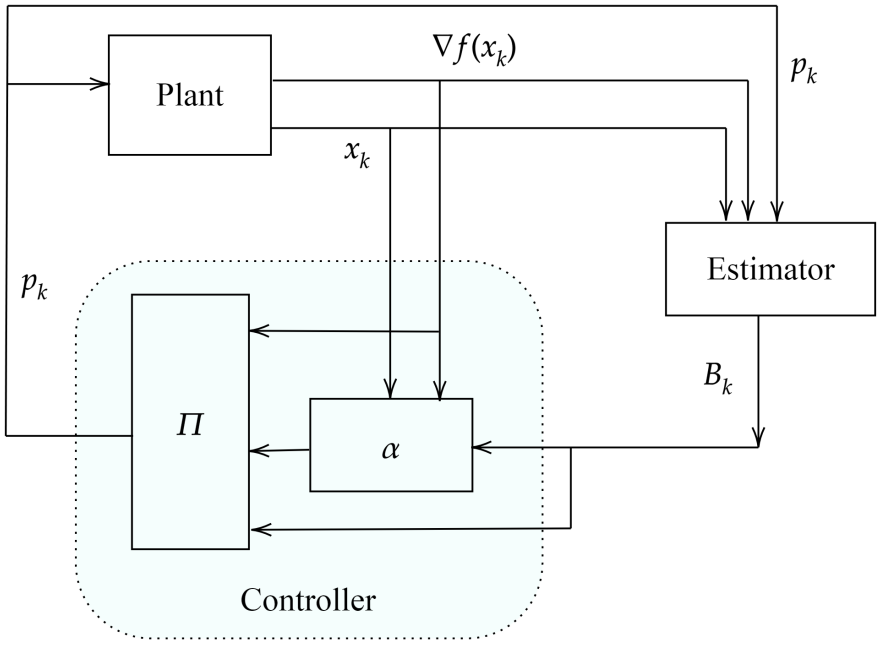


Figure 23: A general diagram to summarize DFP and BFGS algorithms.

Explicit error analysis of BFGS. In all that follows the bars indicate the exact value, that is the value that the variables would have theoretically without any error. Suppose in a first approximation that $B_k \equiv B_k^{BFGS} = H(x_k)^{-1} + \varepsilon_H$. Then following the steps of the (BFGS Algorithm) and considering the evaluation errors we have that

$$\begin{aligned} d &= -(H(x_k) + \varepsilon_H)^{-1}(\nabla f(x_k) + \varepsilon_g) \\ &= \underbrace{-H^{-1}(x_k)\nabla f(x_k)}_{\overline{d_k}} \underbrace{- (H^{-1}(x_k)\varepsilon_g + \varepsilon_H \nabla f(x_k) + \varepsilon_H \varepsilon_g)}_{\varepsilon_d} \end{aligned}$$

and so

$$\alpha_k = \arg\min f(x_k + \alpha d_k) = \overline{\alpha_k} + \varepsilon_\alpha$$

and

$$\begin{aligned} p_k = \alpha_k d_k &= (\overline{\alpha_k} + \varepsilon_\alpha)(\overline{d_k} + \varepsilon_d) \\ &= \underbrace{\overline{\alpha_k}\overline{d_k}}_{\overline{p_k}} + \underbrace{\alpha_k \varepsilon_d + \varepsilon_\alpha d_k}_{\varepsilon_p}. \end{aligned}$$

The main equation on x is then given by

$$x_{k+1} = x_k + p_k = \underbrace{x_k + \overline{p_k}}_{\overline{x_{k+1}}} + \varepsilon_p.$$

When computing the gradient direction:

$$\begin{aligned} q_k = \nabla f(x_{k+1}) - \nabla f(x_k) &= \nabla f(\overline{x_{k+1}} + \varepsilon_p) - \nabla f(x_k) \\ &\approx \underbrace{\nabla f(\overline{x_{k+1}}) - \nabla f(x_k)}_{\overline{q_k}} + \underbrace{H\varepsilon_p}_{\varepsilon_q} \end{aligned}$$

That leads, after some computations, to the equation on B:

$$\begin{aligned} B_{k+1} &= B_k \frac{(p_k^T q_k + q_k^T B_k q_k)(p_k p_k^T)}{(p_k^T q_k)^2} - \frac{B_k q_k p_k^T + p_k q_k^T B_k}{p_k^T q_k} \\ &= \overline{B_{k+1}} + \varepsilon_B. \end{aligned}$$

This should be re-injected into the equation for d_k (the computation is the same with ε_H that becomes $\hat{\varepsilon}_H = \varepsilon_H + \varepsilon_B$ at the next step). Similarly to (Proposition 5.2), we can provide sufficient condition for the two dynamics on x and B to be input-to-state-stability, with inputs $u_x = \varepsilon_p$ and $u_B = \varepsilon_B$. The problem is to explicit the construction of the $\mathcal{K}_\infty$ functions (or at least prove their existence) that guarantees the stability.

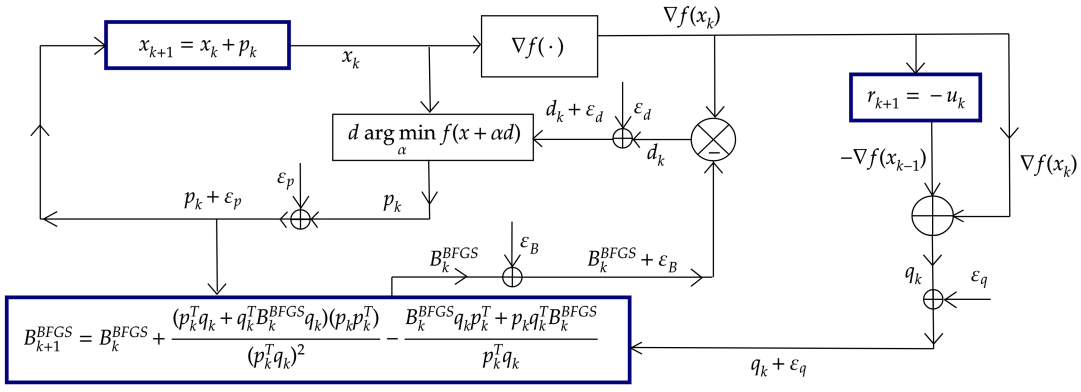


Figure 24: Error diagram for (BFGS Algorithm).

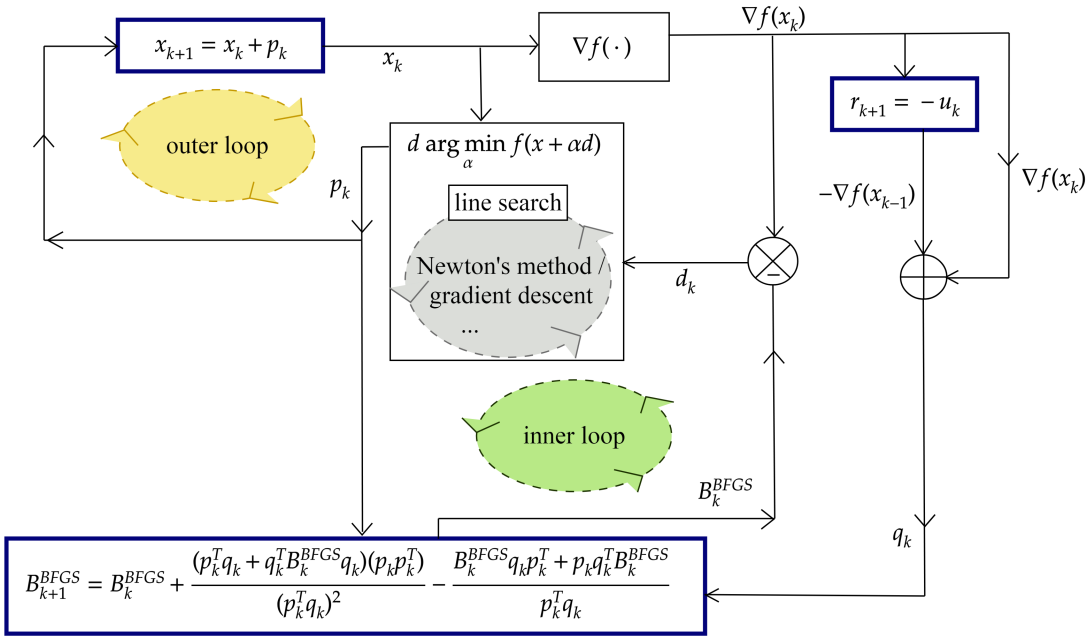


Figure 25: The three main loops of (BFGS Algorithm).

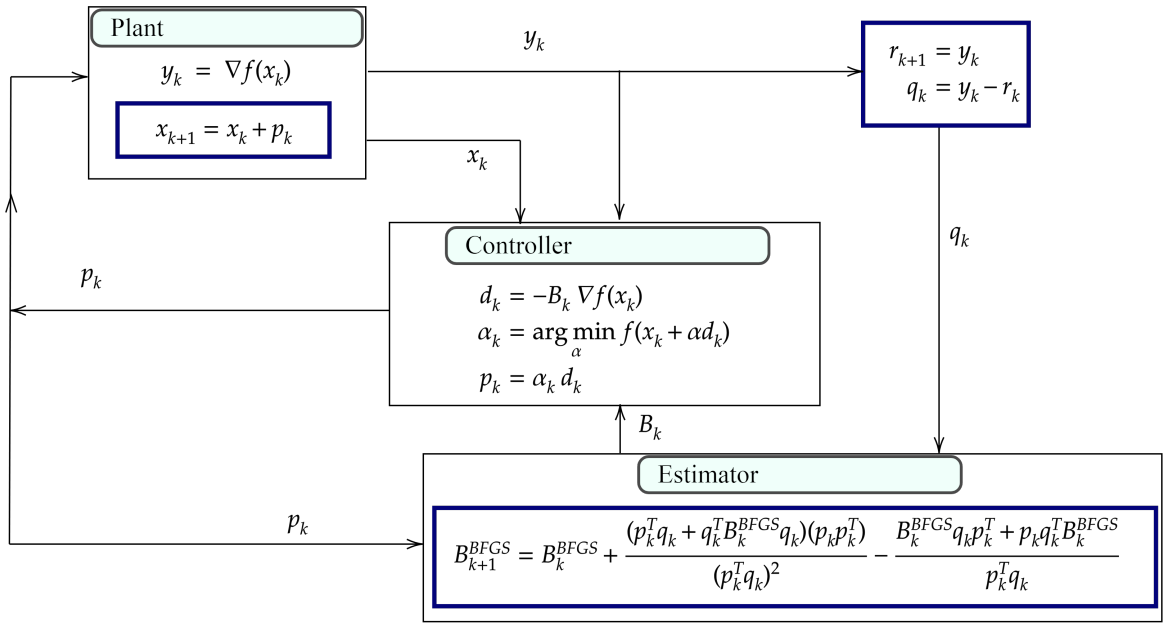


Figure 26: Diagram for (BFGS Algorithm) in the general form of (Figure 23).

4.7 Conclusion of this section

This was the first "big section" of this report. It began with the introduction and the description of Newton's method as a tool to compute zeros or a function (or equivalently stationary points) in (SUBSECTION 4.1) and (SUBSECTION 4.2). The role of each assumption was emphasized in (SUBSECTION 4.3). In (SUBSECTION 4.4) we stated and gave a proof for a usual theorem of convergence (Theorem 4.1), together with extensive comments on the use of its assumptions and its application to the examples of the preceding section. The same result is then restated using comparison functions (Theorem 4.2). Follows a general a brief presentation about the presence of errors in the evaluations of function, gradient and hessian matrix (SUBSECTION 4.4). Additional assumptions on the size of the error let us apply the previous theorems even when some noise is added at each step (Fact 1) and (Fact 2). After fixing the notations, corresponding theorems for both exact (Theorem 4.3) and practical (Proposition 4.4) convergence are enunciated and proved in the case of noisy gradient. The same was done for the general case in which the hessian is also noisy: (Theorem 4.5) and (Proposition 4.7). In this more complicated context, a long discussion follows (Theorem 4.5) on how modifying the sufficient conditions in order to guarantee practical stability. Different solutions are proposed, but none of them seems to be easy to check.

Quasi-Newton methods - in which the inverse of the Hessian matrix is replaced by a local approximation - are subject of (SUBSECTION 4.5) and (SUBSECTION 4.6). After a brief description, two general result for local convergence - again exact (Theorem 4.8) and practical (Proposition 4.9) - are stated and proved. The principal Quasi-Newton methods we dealt with were (DFP Algorithm) and (BFGS Algorithm). Since it appears difficult to directly apply the local result (its requirements are quite strong), we tried different way of looking at them as dynamical systems (see diagrams in (Figure 25) and (Figure 26)).

5 Newton's method in the ISS formalism

Recall the result of (Theorem 4.1): $\|x_{k+1} - x^*\| \leq c \|x_k - x^*\|$ for some constant $c < 1$, then

$$\|x_{k+1} - x^*\| \leq c^{k+1} \|x_0 - x^*\| \quad \forall k \; \forall x_0 \; \forall u$$

that is Newton's method has asymptotic stability.

Having in mind the (Definition 2.10), we would like to prove that the ISS property holds for Newton's method. As already observed, if we consider the iteration step $x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) = x_k + u_k$ then the ISS property does not hold: if one takes $u_k = 0$ the dynamic is unstable. Consider instead the decomposition

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) = \frac{1}{2}x_k + \frac{1}{2}x_k - H^{-1}(x_k)\nabla f(x_k) = \frac{1}{2}x_k + u_k.$$

It is easy to estimate

$$\begin{aligned}
\|x_{k+1}\| &= \left\| \frac{1}{2}x_k + u_k \right\| \\
&\leq \frac{1}{2}\|x_k\| + \|u_k\| \\
&\leq \underbrace{\frac{1}{2^{k+1}}\|x_0\|}_{=:\beta(\|x_0\|,k)} + \sum_{i=0}^{k} \frac{1}{2^i}\|u_{k-i}\| \\
&\leq \beta(\|x_0\|, k) + \underbrace{2\|u\|_\infty}_{=:\gamma(\|u\|_\infty)}.
\end{aligned}$$

That means that the dynamic is ISS with input $u_k = \frac{1}{2}x_k - H^{-1}(x_k)\nabla f(x_k)$. Notice that the this stays true with any other decomposition $x_k = \lambda x_k + (1-\lambda)x_k$ for $0 < \lambda < 1$, as one would obtain a term in $\lambda^k \overset{k\to\infty}{\longrightarrow} 0$ and a convergent series $\sum_{i\geq 0}(1-\lambda)^i = \frac{1}{\lambda}$. We are now interested in introducing an error in the iteration step and try to obtain the ISS property with input containing the error term. For this purpose, consider the dynamic corresponds to $x_{k+1} = g(x_k, u_k) = x_k - H^{-1}(x_k)\nabla f(x_k) + u_k$.

Remark 5.0.1. Recall that if the perturbation is in the form $u_k = H^{-1}(x_k)\eta_k$, from (Theorem 4.3) ensure asymptotic stability when starting close enough to x^*, and from the discussion and the fact stated after the same (Theorem 4.3) practical stability to a ball of radius δ is ensured with weaker and slightly different assumptions. ◀

Let us define the solution error as $e_k := x_k - x^*$, where x_k is the approximate solution at iteration k and x^* is the exact solution of a numerical problem. Given a dynamical system representation of an algorithm in state space form, $x_{k+1} = g(x_k, u_k)$, we can find the dynamics for the solution error as

$e_{k+1} = g_e(e_k, u_k) := g(e_k + x^*, u_k) + x^*$. Note that this dynamical system for the solution error has an equilibrium at the origin for $w_k = 0$.

Proposition 5.1 (Error bound for an ISS system)**.** *Assume that x^* is an equilibrium point of the dynamical system when there is no disturbance, i.e. $u_k = 0$. If the dynamical system for the solution error, $e_{k+1} = g_e(e_k, u_k)$ is ISS, then the norm of the error, is bounded:*

$$\|e_k\| \leq \beta(\|e_0\|, k) + \gamma(\|u\|_\infty)$$

for each $k \geq 0$, where $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$.

Proof. It is exactly the (Definition 2.10) of the ISS property for a dynamical system. □

Remark 5.1.1. As a corollary from (Proposition 5.1) we have the following bound:

$$\lim_{k\to\infty} \|e_k\| \leq \gamma(\|u\|_\infty).$$

Also remark that (Proposition 5.1) is true for a general dynamical system, i.e. g does not have to correspond to Newton's method, since the proof only is based on the assumption of the system for the error being ISS. ◀

We recall the definitions of a GAS and UGAS system (as given in [12]).

Definition 5.1 (GAS)**.** A discrete time system

$$x_{k+1} = f(x_k, u_k)$$

is said to be globally asymptotically stable (**GAS**) if

- for each $\varepsilon > 0$ there is a $\delta > 0$ so that for every $k \geq 0$ and u taking values in a compact set $\Omega \subset \mathbb{R}^m$, if $\|\xi\| < \delta$ then $\|x(k, \xi, u)\| < \varepsilon$; (local uniform stability) [5]
- $\lim_{k\to\infty} \|x(k, \xi, u)\| = 0$ for each $\xi \in \mathbb{R}^n$ and u taking values in a compact set $\Omega \subset \mathbb{R}^m$. (global attraction)

Alternatively, we can say that the origin is GAS if $\exists \beta \in \mathcal{KL}$ such that

$$\|x_k\| \leq \beta(\|x_0\|, k) \quad \forall x_0 \in \mathbb{R}^n \ \forall k \geq 0. \tag{51}$$

If $u \equiv 0$ and the previous condition (51) holds, we say that the origin is 0-GAS.

Example 5.1. The system (without disturbances)

$$x_{k+1} = x_k \left(1 - \frac{1}{k+2}\right)$$

[5] We can state this as "the trajectory is a map from $\mathbb{R}^n \ni \xi \mapsto x(\cdot, \xi) \in C(\mathbb{R}_+; \mathbb{R}^n)$ continuous at 0".

is GAS. Indeed, the solution in terms of $x_0 = \xi$ is given by

$$x_k = \frac{1}{k+1}\xi,$$

which clearly satisfies the global attraction property from (Definition 5.1). Local uniform stability also holds: for every ε, we can choose $\delta = \varepsilon$ and whenever $\|\xi\| < \delta$, then $\|x(k,\xi,u)\| = \left\|\frac{1}{k+1}\xi\right\| < \delta = \varepsilon$. ◁

Definition 5.2 (UGAS)**.** The system $x_{k+1} = g(k, x_k, u_k)$ is called uniformly globally asymptotically stable (**UGAS**) if

- $\exists \delta \in \mathcal{K}_\infty$ such that for each $\varepsilon > 0$ and $\|\xi\| < \delta(\varepsilon)$ yields $\|x(k, k_0, \xi, u)\| \leq \varepsilon$ for each u taking values in a compact set $\Omega \subset \mathbb{R}^m$; (uniform stability)
- for all $r, \varepsilon > 0$ there is a time $T \in \mathbb{N}$ so that for every $k \geq T + k_0$ and $\|\xi\| \leq r$ the norm of the trajectory is $\|x(k, k_0, \xi, u)\| \leq \varepsilon$ for every $k_0 \in \mathbb{N}$ and for each u taking values in a compact set $\Omega \subset \mathbb{R}^m$. (uniform global attraction)

Remark 5.1.2. In fact, we know that a system is UGAS if and only if there exists a $\mathcal{KL}$ function β which is an upper bound for the norm of the trajectory:

$$\|x(k, k_0, \xi, u)\| \leq \beta(\|\xi\|, k - k_0) \quad \forall \xi \in \mathbb{R}^n \ \forall k, k_0 \in \mathbb{N} \ \text{ s.t. } \ k \geq k_0.$$

◀

It is clear that UGAS implies GAS, but the converse is not generally true (it is for periodic systems, see [12, Proposition 3.2, p. 52])..

5.1 A Lyapunov function for Newton's method

We can apply the different characterizations of ISS (Theorem 2.7) and try to give an ISS-Lyapunov function for the Newton's method. In the following we suppose that a stationary point x^* exists.

Recall the definition of a classical Lyapunov function for a dynamical system:

Definition 5.3 (Lyapunov function)**.** A **Lyapunov function** for an autonomous dynamical system $\dot{x} = f(x)$ with an equilibrium point x^* is a continuous function V which is locally positive definite in a neighborhood of the point and for which $\nabla V \cdot f$ is negative definite. In other words we require:

- $V(x^*) = 0$ and $V(x) > 0$ for $x \neq x^*$;
- $\nabla V(x) \cdot f(x) = \dfrac{\partial}{\partial x_1} V(x) f_1(x) + \cdots + \dfrac{\partial}{\partial x_n} V(x) f_n(x) \leq 0.$

Example 5.2: A Lyapunov function. Suppose that x^* is a stationary point for f, i.e. $\nabla f(x^*) = 0$ which is a minimum for f. We look for a function V such that $\dot{V}(x) = \nabla V(x)^T(x - H(x)^{-1}\nabla f(x))$ is negative definite. Since we need $\dot{V}(x^*) = 0$ one must have $\nabla V(x^*) = 0$. Let

$$V(x) = f(x) - f(x^*)$$

and let us prove that V is a Lyapunov function for the system

$$x_{k+1} = g(x_k) = x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k).$$

It is clear that $V(x^*) = 0$ and locally $V(x) > 0$ for $x \neq x^*$. The derivative of V is $\nabla V(x) = \nabla f(x)$ and $-\nabla V \cdot g(x) = -\nabla f(x)^T x + \nabla f(x)^T(\nabla^2 f(x))^{-1}\nabla f(x)$ vanishes in $x = x^*$ and is otherwise strictly positive in any point $x = x_k$ of the sequence of the algorithm. To see this, we consider Taylor's expansion of f:

$$f(x + h) = f(x) + \nabla f(x)^T h + \frac{1}{2}h^T H(x)h + O(\|h\|^3)$$

that can be rewritten as (using $y = x + h$):

$$\begin{aligned} f(y) &= f(x) + \nabla f(x)^T(x - y) + \frac{1}{2}(y - x)^T H(x)(y - x) + O(\|y - x\|^3) \\ &= constants + (\nabla f(x) - H(x)x)^T y + \frac{1}{2}y^T H(x)y + O(\|y - x\|^3). \end{aligned}$$

Now,

$$\begin{aligned} dV(y)^T(x_+ - x) &= (\nabla f(x) - H(x)x + H(x)y)^T(-H^{-1}(x)\nabla f(x)) \\ &= -\nabla f(x)^T H(x)\nabla f(x) + x^T H(x) - y^T H(x) \end{aligned}$$

which in $y = x$ becomes $dV(x)^T(x_+ - x) = -\nabla f(x)^T H(x)\nabla f(x)$ that is quadratic and thus always negative. So, V is a continuously differentiable function that vanishes in x^*, is locally positive-definite, and for which the differential is negative-definite. This means that V is a Lyapunov function for the system. ◁

Example 5.3: An easier attempt. Another possible Lyapunov function for Newton's method may be

$$V(x_k) = \min_{i \leq k}\{f(x_i)\} - f(x^*).$$

Indeed, if x^* is a minimum for the function f, then $V(x^*) = 0$ and for each k is clear that $V(x_k) \geq 0$. Moreover, the sequence $\{V(x_k)\}_k$ is non-increasing

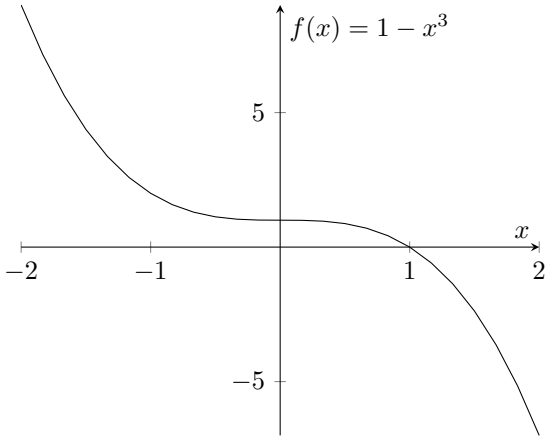
because for each k holds

$$\begin{aligned} V(x_{k+1}) - V(x_k) &= \min_{i\le k+1}\{f(x_i)\} - f(x^*) - \min_{i\le k}\{f(x_i)\} + f(x^*) \\ &= \begin{cases} 0 & \text{if } f(x_{k+1}) \ge f(x_k) \\ f(x_{k+1}) - \min_{i\le k}\{f(x_i)\} < 0 \end{cases} \\ &\le 0. \end{aligned}$$

Notice that this function V looks very similar to the previous one but it was easier to prove that is a Lyapunov function for the Newton's method. ◁

5.1.1 Lyapunov functions for the examples of (SUBSECTION 4.3)

Let's take for example the dynamic $\dot{x} = f(x)$ with $f(x) = 1 - x^3$.



It is really easy to find a Lyapunov function for this dynamic: just take $V(x) = \frac{1}{2}(x-1)^2$ which is obviously positive definite and vanishes only for $x = 1$; in addiction $V'(x)f(x) = (x-1)(1-x^3) \le 0$ and the equality holds only for $x = 1$. That means that such a function V is a Lyapunov function for the dynamical system $\dot{x} = 1 - x^3$.

Remark 5.1.3. In general, if $f(x)$ is a decreasing function with a simple root in $x = x^*$, then $V(x) = \frac{1}{2}(x - x^*)^2$ is a Lyapunov function for the dynamic $\dot{x} = f(x)$, as $V'(x)f(x) \le 0$ and $V(x) \ge 0$ with equalities only at point x^*. ◀

In case of multiple roots, one can only obtain local Lyapunov function. For example, consider $f(x) = 1 - x^2$ which has two roots $x_1 = -1$ and $x_2 = 1$. Then the two functions

$$V_1(x) = \frac{1}{2}(x+1)^2 \quad \text{for } x < 0 \qquad V_2(x) = \frac{1}{2}(x-1)^2 \quad \text{for } x > 0$$

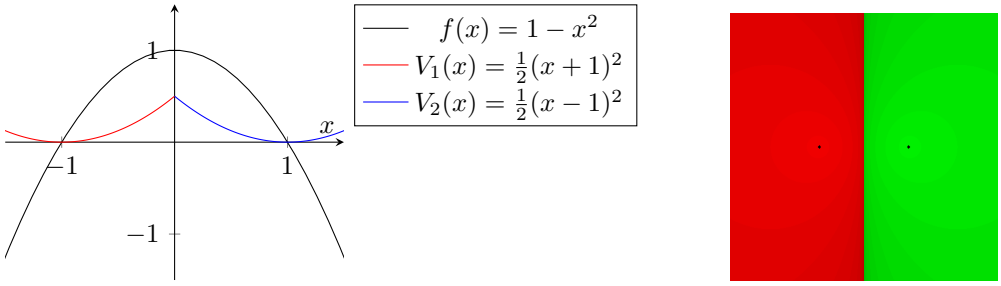


Figure 27: *Left*: $f(x) = 1 - x^2$ and the two Lyapunov functions for the roots. *Right*: The domains of attraction for the two roots $x_1 = -1$ and $x_2 = 1$ of $f(x)$.

are Lyapunov function for the two roots respectively. The domains of attraction are simply $\{x < 0\}$ and $\{x > 0\}$ (as observed in subsection (SUBSECTION 4.3) the point $x = 0$ lead to a division by zero in the first step) as shown in figure (Figure 27).

5.2 An ISS-Lyapunov function for Newton's method

Example 5.4: An ISS-Lyapunov function. We found a "classical" Lyapunov function, but we are now interested in an ISS-Lyapunov function for Newton's method. Using (Theorem 2.8) we look for an ISS-Lyapunov function $V(x)$ that satisfies (Definition 2.11) with $\sigma \equiv 0$. Since x^* is a stationary point for

$$g(x, u) = x - H^{-1}(x)\nabla f(x) + u$$

as $g(x^*, u) = x^* + u$, from the second property in (Definition 2.11) one obtain that $\alpha_3(\|x^*\|)$ must be 0.

We try again with the same function $V(x_k) = \min_{i \le k}\{f(x_i)\} - f(x^*)$. If we use Taylor expansion on $f(x)$ we obtain:

$$\begin{aligned}
V(x_+) - V(x) &= f(x_+) - f(x) \\
&= f\left(x - H^{-1}(x)\nabla f(x) + u\right) - f(x) \\
&= -\nabla f(x)^T \left(H^{-1}(x)\nabla f(x) + u\right) + \frac{1}{2}\left(\nabla f(x)^T H^{-1}(x) + u^T\right) H(x) \left(H^{-1}(x)\nabla f(x) + u\right) + \dots \\
&= -\nabla f(x)^T H^{-1}(x)\nabla f(x) - \nabla f(x)u \\
&\qquad + \frac{1}{2}\left[\nabla f(x)^T H^{-1}(x)\nabla f(x) + \nabla f(x)u + u^T\nabla f(x) + u^T H(x)u\right] + \dots \\
&= -\frac{1}{2}\nabla f(x)^T H^{-1}(x)\nabla f(x) + \frac{1}{2}u^T H(x)u + O(\left\|H^{-1}(x)\nabla f(x) + u\right\|^3) \\
&\overset{?}{\le} -\alpha_3(\|x\|) + \sigma(\|u\|)
\end{aligned}$$

It is not obvious that one can always find a $\mathcal{K}_\infty$ function $\alpha_3(\|x\|)$ which is smaller than the quadratic function in $\nabla f(x)$. Moreover because of the first property (14a) of (Definition 2.11), $V(x)$ should be unbounded in every direction, but is not the case for our choice. The same thing is true for $V(x_k) = \min_{i \leq k}\{f(x_i)\} - f(x^*)$. This is an example showing that a "classical" Lyapunov function is not necessarily an ISS-Lyapunov function for the same system. On the other hand, observe that if V is an ISS-Lyapunov function for the system $\dot{x} = f(x, u)$, then V is a Lyapunov function, in the usual sense, for the autonomous system $\dot{x} = f(x, 0)$ obtained when no controls are applied. The inequality can still be true for some functions. One condition on f to satisfy the inequality is

$$-\frac{1}{2}\nabla f(x)^T H^{-1}(x)\nabla f(x) \leq -\frac{1}{2}x^T H^{-1}x \Rightarrow \|\nabla f(x)\| \geq \|x\|\,.$$

This condition implies that the function growth is at least superlinear.

◁

In order to find an ISS-Lyapunov function, I tried to retrace the steps of the proof of (Theorem 2.7). So, we know that NM is ISS, which implies UBIBS and AG with 0-gain (as we remarked the result of (Theorem 2.8)). Next step is to see that is robustly stable, and finally that this gives us a smooth ISS-function. The main idea is to recall the following property:

$$\forall \beta \in \mathcal{KL}\ \exists \rho_1,\ \rho_2 \in \mathcal{K}_\infty \quad \beta(s, r) \leq \rho_1(\rho_2(s)e^{-r})\ \forall s, r \geq 0$$

Then we can use the characterization of a UGAS system, so that NM is UGAS if and only if

$$\exists \rho_1,\ \rho_2 \in \mathcal{K}_\infty \quad \|x(k, \xi, u)\| \leq \rho_1(\rho_2(\|\xi\|)e^{-k})$$

Once we have found such a $\mathcal{K}_\infty$ function ρ_1 the work is done since, as in the proof of [12, Theorem 1] we can define $\omega = \rho_1^{-1}$ and $V(\xi) := \sup_d \sum_{k \geq 0} \omega(\|x(k, \xi, d)\|)$ which is shown to be an ISS-Lyapunov function that satisfies the property (14a) and (14b) of (Definition 2.11) with:

$$\omega(\|\xi\|) \leq V(\xi) \leq \frac{e}{e-1}\rho_2(\|\xi\|)$$
$$V(f(\xi, \mu)) - V(\xi) \leq -\omega(\|\xi\|)$$

5.2.1 Using comparison functions

One obtains better results using comparison functions. The following statement gives sufficient conditions for the Newton's method to be input-to-state stable.

Proposition 5.2 (Sufficient condition for ISS of NM)**.** *Suppose there exist $\psi \in \mathcal{K}_\infty$ such that $\psi - id \in \mathcal{K}_\infty$ and $\varphi \in \mathcal{K}_\infty$ such that the following conditions hold:*

(i) $\varphi(\lambda \|x\|) \leq \lambda \varphi(\|x\|)$ *for all* $0 < \lambda < 1$ *and* $x \in \mathbb{R}^n$;

(ii) *there exists* $0 < c < 1$ *such that* $\psi\left(\left\|x - H^{-1}(x)\nabla f(x)\right\|\right) \leq c \|x\|$ *for all* $x \in \mathbb{R}^n$.

Then $V(x) = \varphi(\|x\|)$ *is an ISS-Lyapunov function for the system :*

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) \underbrace{-H^{-1}(x_k)r_k - s_k\nabla f(x_k) - s_k r_k}_{u_k}. \tag{42}$$

Proof. Applying the hypothesis we obtain that

$$\begin{aligned} V(x_{k+1}) - V(x_k) &= \varphi\left(\left\|x_k - H^{-1}(x_k)\nabla f(x_k) + u_k\right\|\right) - \varphi(\|x_k\|) \\ &\leq \varphi\left(\psi\left(\left\|x - H^{-1}(x)\nabla f(x)\right\|\right)\right) + \underbrace{\varphi\left(\psi \circ (\psi - id)^{-1}(\|u_k\|)\right)}_{\sigma(\|u_k\|)} - \varphi(\|x_k\|) \\ &\leq \varphi\left(\psi\left(\left\|x - H^{-1}(x)\nabla f(x)\right\|\right)\right) - \varphi(\|x_k\|) + \sigma(\|u_k\|) \\ &\leq \underbrace{(c-1)\,\varphi(\|x_k\|)}_{-\alpha(\|x_k\|)} + \sigma(\|u_k\|) \end{aligned}$$

□

Remark 5.2.1. The same result with the same computations yields for $V(x) = \varphi(\|x\|^2)$ (or other powers of $\|x\|$), as we only use the fact that $\|x\|$ is positive.

◀

Remark 5.2.2. If we change the definition of the disturbance u_k in the iteration step (42). In (Proposition 5.2) we assumed $u_k = -H^{-1}(x_k)r_k - s_k\nabla f(x_k) - s_k r_k$. Considering instead

$$u_k = -H^{-1}(x_k)r_k - s_k r_k$$

the result still holds replacing condition (ii) by

(ii') $\psi\left(\left\|x - \left(H^{-1}(x) + s_k\right)\nabla f(x)\right\|\right) \leq c \|x\|$ for all $x \in \mathbb{R}^n$.

◀

Remark 5.2.3. There are specific conditions that f must satisfy in order to be "eligible" to the application of this result. Since $\psi - id \in \mathcal{K}_\infty$, $\psi > id$ and we must have

$$\left\|x - H^{-1}(x)\nabla f(x)\right\| < \psi\left(\left\|x - H^{-1}(x)\nabla f(x)\right\|\right) \leq c \|x\| \tag{52}$$

for all $x \in \mathbb{R}^n$. This is a necessary condition on f required to apply (Proposition 5.2). A natural question is then to characterize the classes of functions that satisfies this condition. ◀

Let us start with a few simple examples to answer to the preceding remark. Consider the monomial $f(x) = kx^a$. The condition (52) is written as

$$\left\| x - \frac{x}{a-1} \right\| = \left\| \frac{a-2}{a-1} x \right\| \leq \underbrace{\left\| \frac{a-2}{a-1} \right\|}_{c \in [0,1)} \|x\|$$

which is satisfied for $a \neq 1$ (for $a = 1$ there's no point in applying Newton's Method because we know there are no stationary points). For a general polynomial the condition does not hold: if $f(x) = \sum_{i=0}^{n} a_i x^i$ then

$$c\,\|x\| > \left\| x - \frac{\sum_{i=1}^{n} i a_i x^{i-1}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} \right\| = \left\| \frac{\sum_{i=1}^{n} i(i-2) a_i x^{i-1}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} \right\|$$

is not satisfied in $x = 0$ unless $a_1 = 0$ (the condition gives $0 > \|a_1/2a_2\|$), and in this case

$$\begin{aligned} \left\| x - \frac{\sum_{i=1}^{n} i a_i x^{i-1}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} \right\| &= \left\| x \left(1 - \frac{\sum_{i=2}^{n} i a_i x^{i-2}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} \right) \right\| \\ &< c\,\|x\| \end{aligned}$$

for some $c \in (0,1)$, whenever

$$-1 < 1 - \frac{\sum_{i=2}^{n} i a_i x^{i-2}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} = \frac{\sum_{i=2}^{n} i(i-2) a_i x^{i-2}}{\sum_{i=2}^{n} i(i-1) a_i x^{i-2}} < 1.$$

This can be rewritten as

$$\begin{cases} \sum_{i=2}^{n} i a_i x^{i-2} > 0 \\ \sum_{i=2}^{n} i(2i-3) a_i x^{i-2} > 0 \end{cases}.$$

The summations are positive in particular if each term is positive, and that leads to

$$\begin{cases} a_i > 0 & \text{if } i \text{ is even} \\ a_i = 0 & \text{if } i \text{ is odd} \end{cases}. \tag{53}$$

Remark 5.2.4. Since the condition is on the gradient and Hessian matrix of f, adding a constant to the function f do not change the truth value of the condition. ◀

Example 5.5. We can apply the preceding result of (Proposition 5.2) to the polynomial

$$f(x) = x^4 + x^2 + 3$$

with the ISS-Lyapunov function $V(x) = |x|^2$ and the $\mathcal{K}_\infty$ function $\psi(s) = \frac{11}{10} s > s$. Such a ψ satisfies the condition $(\psi - id)(s) = s/10 \in \mathcal{K}_\infty$ and

$$\psi\left(\left| x - H^{-1}(x) \nabla f(x) \right| \right) = \frac{11}{10} \left| \left(1 - \frac{2x^2 + 1}{6x^2 + 1} \right) x \right| < \underbrace{\frac{11}{15}}_{c<1} |x|\,.$$

Indeed, the iteration step of Newton's Method for this function

$$x_{k+1} = \left(1 - \frac{2x_k^2 + 1}{6x_k^2 + 1}\right) x_k + u_k$$

is ISS with respect to a disturbance u_k:

$$\begin{aligned} V(x_{k+1}) - V(x_k) &= \left| \left(1 - \frac{2x_k^2 + 1}{6x_k^2 + 1}\right) x_k + u_k \right|^2 - |x_k|^2 \\ &\leq \left| \frac{11}{10} \left(1 - \frac{2x_k^2 + 1}{6x_k^2 + 1}\right) x_k \right|^2 + |11u_k|^2 - |x_k|^2 \\ &\leq \left(\frac{484}{900} - 1\right) |x_k|^2 + \underbrace{11\,|u_k|^2}_{\sigma(|u_k|)} \\ &\leq \underbrace{-0.46\,|x_k|^2}_{-\alpha(|x_k|)} + \sigma(|u_k|) \end{aligned}$$

◁

The following example shows that the condition (53) on the coefficients of a polynomial is not a necessary condition.

Example 5.6. Consider the polynomial

$$f(x) = x^4 + 4x^3 + 9x^2 + 1$$

and observe that $a_3 = 4 \neq 0$ so that (53) does not hold. However, take $V(x) = |x|^2$ and $\psi(s) = \frac{11}{10}s$ as in the previous example. Since

$$\psi\left(\left|x - H^{-1}(x)\nabla f(x)\right|\right) = \frac{11}{10} \left| \left(1 - \frac{2x^2 + 6x + 9}{6x^2 + 12x + 9}\right) x \right| < \underbrace{\frac{88}{100}}_{c<1} |x|$$

the (Proposition 5.2) can be applied, thus $V(x)$ is an ISS-Lyapunov function for the Newton's Method dynamic

$$x_{k+1} = \left(1 - \frac{2x_k^2 + 6x_k + 9}{6x_k^2 + 12x_k + 9}\right) x_k + u_k$$

satisfying the inequality

$$V(x_{k+1}) - V(x_k) \leq \underbrace{-0.2256\,|x_k|^2}_{-\alpha(|x_k|)} + \underbrace{11\,|u_k|}_{\sigma(|u_k|)}.$$

◁

5.2.2 Another form of V

Suppose $\varphi \in \mathcal{K}_\infty$ and define

$$V(x) = \int_0^{\|x\|} \varphi(s)ds.$$

Two cases are possible: $\|x\| \leq \|x_+\|$ and in this case

$$0 \leq V(x_+) - V(x) = \int_{\|x\|}^{\|x_+\|} \varphi(s)ds \leq -\alpha(\|x\|) + \sigma(\|u\|) \quad \forall x,\, u$$

leads to an impossible condition (take $u = 0$ and a "big" x); then $\|x\| > \|x_+\|$ and

$$V(x_+) - V(x) = -\int_{\|x_+\|}^{\|x\|} \varphi(s)ds = -\varphi(\tau)(\|x\| - \|x_+\|) < 0$$

for some $\tau = \|x_+\| + t\|x\|$ with $t \in (0, 1)$. If for every x the difference $\left\|x - H^{-1}(x)\nabla f(x)\right\| - \|x\|$ is negative, the condition (14b) is satisfied:

$$-\varphi(\tau)(\|x\| - \|x_+\|) \leq \underbrace{\varphi(\tau)\|u\|}_{\sigma\|u\|} + \underbrace{\varphi(\tau)\left(\left\|x - H^{-1}(x)\nabla f(x)\right\| - \|x\|\right)}_{\leq -\alpha(\|x\|)}.$$

However, the condition $\|x\| > \|x_+\| = \left\|x - H^{-1}(x)\nabla f(x) + u\right\|$ for all x, u is inconsistent (take $u = H^{-1}(x)\nabla f(x)$). So this form for an ISS-Lyapunov function is possible only within a certain class of functions f and disturbances u, as for the case $V(x) = \varphi(\|x\|)$.

Recap. Different form for a Lyapunov function were tested.

(a) $V(x) = \|x\|$;

(b) $V(x) = \|x\|^2$;

(c) $V(x) = \|x\|^r$;

(d) $V(x) = \varphi(\|x\|)$ for some $\varphi \in \mathcal{K}_\infty$;

(e) $V(x) = \int_0^{\|x\|} \varphi(s)ds$ for some $\varphi \in \mathcal{K}_\infty$;

For the form (a) the condition

$$\exists \alpha \in \mathcal{K}_\infty \quad \left\|x - H^{-1}(x)\nabla f(x)\right\| - \|x\| \leq -\alpha(\|x\|)\ \forall x$$

is needed in order to let V satisfy the (Definition 2.11). For the function of the form (d) we obtained the result of (Proposition 5.2). Form (e) is possible only for some specific class of functions f and disturbances u.

5.3 Incremental stability for NM

Proposition 5.3 (NM (without noise) is δISS)**.** *Suppose that the function $x \mapsto H^{-1}(x)\nabla f(x)$ is 1-Lipschitz. Then, Newton's Method is incrementally ISS with input $u = H^{-1}(x)\nabla f(x)$.*

Proof. Consider Newton's method iteration step

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) = x_k + u_k \tag{28}$$

and $V(x, y) = \|x - y\|^2$. Then the δISS Lyapunov condition (22) is given by

$$\begin{aligned} V(x_k + u_k, y_k + v_k) - V(x_k, y_k) &= \|x_k + u_k - y_k - v_k\|^2 - \|x_k - y_k\|^2 \\ &\leq \|u_k - v_k\|^2 - 2\|x_k - y_k\| \|u_k - v_k\| \\ &\leq \|x_k - v_k\|^2 - 2\|x_k - y_k\| \|x_k - v_k\| \\ &= \underbrace{-\|x_k - y_k\|^2}_{-\rho(\|x_k - y_k\|)} \end{aligned}$$

whenever $\|u_k - v_k\| \leq \|x_k - y_k\| = \kappa(\|x_k - y_k\|)$. This correspond to the condition

$$\left\|H^{-1}(x)\nabla f(x) - H^{-1}(y)\nabla f(y)\right\| \leq \|x - y\| \quad \forall x, y \tag{54}$$

which is a Lipschitz condition for the function $H^{-1}(\cdot)\nabla f(\cdot)$. □

Remark 5.3.1. The preceding (Proposition 5.3) means that Newton's method without noise is δISS with input $u = H^{-1}(x)\nabla f(x)$. Notice that the same reasoning would apply to the more general case (42) in which

$$x_{k+1} = x_k \underbrace{-H^{-1}(x_k)\nabla f(x_k) - H^{-1}(x_k)r_k - s_k\nabla f(x_k) - s_k r_k}_{u_k}$$

but with a stronger condition

$$\begin{aligned} &\left\|H^{-1}(x)\nabla f(x) - H^{-1}(x)r_1 - s_1\nabla f(x) - s_1 r_1 + \right. \\ &\quad \left. -H^{-1}(y)\nabla f(y) + H^{-1}(y)r_2 + s_2\nabla f(y) + s_2 r_2\right\| \leq \|x - y\| \quad \forall x, y. \end{aligned}$$

◀

Remark 5.3.2. If

$$\sup_x \left\|H^{-1}(x)\right\| \leq M_H < \infty \qquad \sup_x \|\nabla f(x)\| \leq M_\nabla < \infty$$

then if we require H^{-1} to be L_H-Lipschitz and ∇f to be L_∇-Lipschitz, the previous 1-Lipschitz condition on $H^{-1}.\nabla f$ is satisfied whenever $M_H L_\nabla + M_\nabla L_H \leq 1$.

◀

Remark 5.3.3. If we prove that some stability holds for the general dynamical system (42), then it is proved for the Quasi Newton Methods (DFP Algorithm) and (BFGS Algorithm) which are particular cases ($r_k = 0$ in iteration step (42)). ◀

Proposition 5.4 (NM is δISS)**.** *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ a function of class C^3. Consider two different trajectories given by Newton's Method iteration*

$$\begin{cases} x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) - H^{-1}(x_k)r_k^x - s_k^x \nabla f(x_k) - s_k^x r_k^x \\ y_{k+1} = y_k - H^{-1}(y_k)\nabla f(y_k) - H^{-1}(y_k)r_k^y - s_k^y \nabla f(y_k) - s_k^y r_k^y \end{cases} \tag{55}$$

where we can suppose that the error in the evaluation of gradient and hessian for the trajectories x and y are linked by the following relationships:

$$r_k^y = r_k^x + \delta_r \qquad s_k^y = s_k^x + \delta_s.$$

Suppose to have the following uniform bounds:

$$\|H(z)\| \le M \qquad \left\|H^{-1}(z)\right\| \le m \qquad \|\nabla f(z)\| \le g$$

for all $z \in \mathbb{R}^n$. Also suppose that H is locally L-Lipschitz, i.e.

$$\exists \zeta > 0 \quad \|x - y\| \le \zeta \Rightarrow \|H(x) - H(y)\| \le L\,\|x - y\|\,.$$

If the initial points are close, namely

$$\|x_0 - y_0\| \le \frac{c - m^2 Lg}{mL(1 + mM)}$$

for some positive constant $1 > c > m^2 Lg$, then the system (55) is δISS with inputs

$$\begin{aligned} u_k &= -H^{-1}(x_k)r_k^x - s_k^x \nabla f(x_k) - s_k^x r_k^x \\ v_k &= -H^{-1}(y_k)r_k^y - s_k^y \nabla f(y_k) - s_k^y r_k^y. \end{aligned}$$

Proof. For the sake of clarity we write simply $s_k \equiv s_k^x$. Then we can write the

second equation as function of the difference $\Delta_k = y_k - x_k$:

$$\begin{aligned}
y_{k+1} &= x_k + \Delta_k - H^{-1}(x_k + \Delta_k)\nabla f(x_k + \Delta_k) - H^{-1}(x_k + \Delta_k)r_k - s_k\nabla f(x_k + \Delta_k) \\
&\quad - H^{-1}(x_k + \Delta_k)\delta_r - \delta_s\nabla f(x_k + \Delta_k) - (s_k + \delta_s)(r_k + \delta_r) \\
&= x_k + \Delta_k - \left(H^{-1}(x_k) - H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\left(\nabla f(x_k) + H(\zeta)\Delta_k\right) \\
&\quad - \left(H^{-1}(x_k) - H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)r_k - s_k\left(\nabla f(x_k) + H(\zeta)\Delta_k\right) \\
&\quad - \left(H^{-1}(x_k) - H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\delta_r - \delta_s\left(\nabla f(x_k) + H(\zeta)\Delta_k\right) \\
&\quad - (s_k + \delta_s)(r_k + \delta_r) \\
&= x_k - H^{-1}(x_k)\nabla f(x_k) - H^{-1}(x_k)r_k - s_k\nabla f(x_k) - s_k r_k \\
&\quad + \left(I - H^{-1}(x_k)H(\zeta) + H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta) - (s_k + \delta_s)H(\zeta)\right)\Delta_k \\
&\quad + \left(H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\left(\nabla f(x_k) + r_k + \delta_r\right) \\
&\quad - H^{-1}(x_k)\delta_r - \delta_s\nabla f(x_k) - s\delta_r - \delta_s r - \delta_s\delta_r \\
&= x_{k+1} + \left(I - H^{-1}(x_k)H(\zeta) + H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta) - (s_k + \delta_s)H(\zeta)\right)\Delta_k \\
&\quad + \left(H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\left(\nabla f(x_k) + r_k + \delta_r\right) \\
&\quad - H^{-1}(x_k)\delta_r - \delta_s\nabla f(x_k) - s\delta_r - \delta_s r_k - \delta_s\delta_r
\end{aligned}$$

that is

$$\begin{aligned}
\Delta_{k+1} &= \left(I - H^{-1}(x_k)H(\zeta) + H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta) - (s_k + \delta_s)H(\zeta)\right)\Delta_k \\
&\quad + \left(H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\left(\nabla f(x_k) + r_k + \delta_r\right) \\
&\quad - H^{-1}(x_k)\delta_r - \delta_s\nabla f(x_k) - s\delta_r - \delta_s r_k - \delta_s\delta_r.
\end{aligned}$$

Above we called τ and ζ two points between x_k and y_k. Call

$$\begin{aligned}
u_k &= -H^{-1}(x_k)r_k^x - s_k^x\nabla f(x_k) - s_k^x r_k^x \\
v_k &= -H^{-1}(y_k)r_k^y - s_k^y\nabla f(y_k) - s_k^y r_k^y
\end{aligned}$$

so that we can write

$$\begin{aligned}
\Delta_{k+1} &= \left(H^{-1}(x_k)\left(H(x_k) - H(\zeta)\right) + H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta)\right)\Delta_k \\
&\quad + \left(H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right)\nabla f(x_k) + v_k - u_k
\end{aligned}$$

and taking the norms

$$\begin{aligned}\|\Delta_{k+1}\| &\leq \left\|H^{-1}(x_k)\left(H(x_k)-H(\zeta)\right)+H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta)\right\|\|\Delta_k\| \\ &\quad+\left\|H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right\|\|\nabla f(x_k)\|+\|u_k-v_k\| \\ &\leq mL(1+mM)\|\Delta_k\|^2+m^2Lg\|\Delta_k\|+\|u_k-v_k\| \\ &\leq c\|\Delta_k\|+\|u_k-v_k\| \\ &\leq c^k\|\Delta_0\|+\|u_k-v_k\|\sum_{j=0}^{k}c^j \\ &\leq \underbrace{c^k\|\Delta_0\|}_{\beta(\|\Delta_0\|,k)}+\underbrace{\frac{\|u-v\|_\infty}{1-c}}_{\gamma(\|u-v\|_\infty)}.\end{aligned}$$

We used the uniform bounds for the second inequality and the hypothesis of closeness in the third one (by induction). The resulting inequality

$$\|\Delta_{k+1}\|\leq\beta\left(\|\Delta_0\|,k\right)+\gamma\left(\|u-v\|_\infty\right)$$

corresponds to (Definition 2.25) of incremental-input-to-state-stability. □

We also supposed that for each k the norms

$$\begin{cases}\left\|I-H^{-1}(x_k)H(\zeta)+H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta)-(s_k+\delta_s)H(\zeta)\right\|\leq c \\ \left\|H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right\|\|\nabla f(x_k)+r_k+\delta_r\| \\ \quad+\left\|(H^{-1}(x_k)+s)\delta_r\right\|+\|\delta_s(\nabla f(x_k)+r_k+\delta_r)\|\leq d\end{cases}$$

are uniformly bounded by two constants $c,d>0$. This is reasonable because:

- when the Newton's Method converges, the gradient $\nabla f(x_k)\to 0$;
- when x_k and y_k are close to each other, ζ and τ are both close to x_k and to each other; we expect the first coefficient to be uniformly bounded by $\sup_k\|s_k+\delta_s\|\|H(x_k)\|$.

Now, if we can say that $c<1$, the previous inequality proves the incremental input-to-state-stability of noisy Newton's Method. We remark that, for a quadratic function f, the two conditions are easily restated as

$$\begin{cases}\|s_k^yH\|\leq c<1 \\ \left\|(H^{-1}+s)\delta_r+\delta_s(\nabla f(x_k)+r_k+\delta_r)\right\|\leq d\end{cases}$$

that is, the first one simply relates the norm of the error $\|s\|_\infty$ to the norm of the hessian matrix $\|H\|$. If instead, we consider the case in which the two errors $r=s=0$, then

$$\begin{cases}\left\|I-H^{-1}(x_k)H(\zeta)+H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)H(\zeta)\right\|\leq c \\ \left\|H^{-1}(\tau)(DH(\tau)\Delta_k)H^{-1}(\tau)\right\|\|\nabla f(x_k)\|\leq d\end{cases}$$

the second condition is reasonable when the Newton's Method converges to a stationary points and H is locally Lipschitz (the LHS is a product of something bounded and something that goes to 0). For the first condition, we can observe that, when $s = 0$ the inequality becomes

$$\|y_{k+1} - x_{k+1}\| \leq \left\|H^{-1}(\tau)\right\|^2 \|DH(\tau)\| \|H(\zeta)\| \|y_k - x_k\|^2 + \left(\left\|H^{-1}(x_k)\left(H(x_k) - H(\zeta)\right)\right\| + \left\|H^{-1}(\tau)\right\|^2 \|DH(\tau)\| \|\nabla f(x_k)\|\right) \|y_k - x_k\|.$$

and the stability condition for this latter is obtained imposing the RHS $\leq \kappa \|y_k - x_k\|$ for some $\kappa < 1$:

$$\begin{cases} \left\|H^{-1}(x_k)\left(H(x_k) - H(\zeta)\right)\right\| + \left\|H^{-1}(\tau)\right\|^2 \|DH(\tau)\| \|\nabla f(x_k)\| \leq \kappa \\ \|y_k - x_k\| \leq \frac{\kappa - \left\|H^{-1}(x_k)(H(x_k) - H(\zeta))\right\| + \left\|H^{-1}(\tau)\right\|^2 \|DH(\tau)\| \|\nabla f(x_k)\|}{\|H^{-1}(\tau)\|^2 \|DH(\tau)\| \|H(\zeta)\|} \end{cases}$$

which, again, seems reasonable because the first term is close to 0 and the second one is bounded (same as for the other condition).

As for (Proposition 5.2), we can try a more general approach with comparison functions.

Recap. Different form for a Lyapunov function were tested.

(a) $V(x, y) = \|x - y\|$;

(b) $V(x, y) = \|x - y\|^2$;

For the form (a) the condition

$$\exists \alpha \in \mathcal{K}_\infty \quad \left\|x - y + H^{-1}(y)\nabla f(y) - H^{-1}(x)\nabla f(x)\right\| - \|x - y\| \leq -\alpha(\|x - y\|) \; \forall x \, y$$

is needed in order to let V satisfy the (Definition 2.11). For the case (b), a sufficient condition is the Lipschitz continuity (54).

5.4 An iISS-Lyapunov function for Newton's method

We consider the iteration step

$$x_+ = x - H^{-1}(x)\nabla f(x) + u \tag{42}$$

and try to find an iISS-Lyapunov function for this system. The following result holds:

Proposition 5.5 (Newton's Method iISS: sufficient condition)**.** *The Newton's Method*

$$x_+ = x - H^{-1}(x)\nabla f(x) + u \tag{42}$$

is iISS with respect to an input u under the assumption

$$\left\|x - H^{-1}(x)\nabla f(x)\right\| \leq \frac{1}{2}\|x\| \quad \forall x \in \mathbb{R}^n.$$

Proof. $V(x) = \log(\|x\| + 1) \in \mathcal{K}_\infty$ is an iISS-Lyapunov function for the system according to (Definition 2.23) with

$$\begin{cases} \alpha_1(s) = V(s) = \alpha_2(s) \\ \rho(s) = \frac{1}{2}\log(s+1) \qquad \hat{\sigma}(s) = \frac{1}{2}\log(2s+1) \end{cases}.$$

Indeed one can compute

$$\begin{aligned} V(x_+) - V(x) &= \log(\|x_+\| + 1) - \log(\|x\| + 1) \\ &\leq \log\left(\left\|x - H^{-1}(x)\nabla f(x)\right\| + \|u\| + 1\right) - \log(\|x\| + 1) \\ &\leq \log\left(\left\|x - H^{-1}(x)\nabla f(x)\right\| + \frac{1}{2} + \|u\| + \frac{1}{2}\right) - \log(\|x\| + 1) \\ &\leq \frac{1}{2}\log\left(2\left\|x - H^{-1}(x)\nabla f(x)\right\| + 1\right) + \underbrace{\frac{1}{2}\log(2\|u\| + 1)}_{\sigma(\|u\|)} - \log(\|x\| + 1) \end{aligned} \tag{$\dagger$}$$

and the condition $2\left\|x - H^{-1}(x)\nabla f(x)\right\| \leq \|x\|$ gives:

$$\begin{aligned} V(x_+) - V(x) &\leq \underbrace{-\frac{1}{2}\log(\|x\| + 1)}_{-\rho(\|x\|)} + \sigma(\|u\|) \\ &= -\rho(\|x\|) + \sigma(\|u\|). \end{aligned} \tag{$\dagger$}$$

The result follows by (Theorem 2.16). Since $\rho \in \mathcal{K}_\infty$ the system is also ISS. □

Example 5.7. Consider the function $f(x) = \frac{3}{5}x^{5/3}$. The assumption of (Proposition 5.5) is satisfied:

$$\left\|x - H^{-1}(x)\nabla f(x)\right\| = \left\|x - \frac{1}{2}\frac{3}{2}x^{1/3}x^{2/3}\right\| = \frac{1}{4}\|x\| \leq \frac{1}{2}\|x\|$$

and then the system

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) + u_k = \frac{1}{4}x_k + u_k$$

is (i)ISS with input u_k. Notice that this was clear already from the explicit writing of the function that rules the dynamic. ◁

Another multidimensional example:

Example 5.8. Consider $f(x) = \frac{3}{5}x_1^{5/3} + \frac{1}{2}x_2^2 + \frac{4}{7}x_3^{7/4}$. The gradient and hessian matrix are given by

$$\nabla f(x) = \begin{pmatrix} x_1^{2/3} \\ x_2 \\ x_3^{3/4} \end{pmatrix} \text{ and } H(x) = \begin{pmatrix} \frac{2}{3}x_1^{-1/3} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{3}{4}x_3^{-1/4} \end{pmatrix}$$

We can verify the condition

$$\begin{aligned}\left\|x - H^{-1}(x)\nabla f(x)\right\| &= \left\| \begin{pmatrix} -\frac{1}{2}x_1 \\ 0 \\ -\frac{1}{3}x_3 \end{pmatrix} \right\|_2 = \sqrt{\frac{1}{4}x_1^2 + \frac{1}{9}x_3^2} \\ &\leq \frac{1}{2}\sqrt{x_1^2 + x_3^2} = \frac{1}{2}\|x\|_2\end{aligned}$$

that guarantees that the system is (i)ISS. ◁

The condition from previous (Proposition 5.5) can be relaxed:

Proposition 5.6 (Newton's Method iISS: another sufficient condition)**.** *If there exists $\rho \in \mathcal{P}$ such that*

$$\|x\| - \left\|x - H^{-1}(x)\nabla f(x)\right\| \geq \rho(\|x\|) \quad \forall x \in \mathbb{R}^n.$$

then, the Newton's Method

$$x_+ = x - H^{-1}(x)\nabla f(x) + u \tag{42}$$

is iISS with respect to an input u.

Proof. Consider $V(x) = \|x\| + \arctan(\|x\|)$. From the assumption we can deduce ($\rho(s) \geq 0$)

$$\|x_+\| - \|x\| \leq \|u\| + \left\|x - H^{-1}(x)\nabla f(x)\right\| - \|x\| \leq \|u\|$$

and therefore

$$\|x\| =: \alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|) := \|x\| + \arctan(\|x\|).$$

Moreover, for $\tau_x = \|x_+\| + t(\|x_+\| - \|x\|)$ with $t \in (0, 1)$

$$\begin{aligned}V(x_+) - V(x) &= \|x_+\| - \|x\| + \arctan(\|x_+\|) - \arctan(\|x\|) \\ &\leq \|u\| + \frac{1}{1+\tau_x^2}(\|x_+\| - \|x\|) \\ &\leq 2\|u\| + \|x\| - \left\|x - H^{-1}(x)\nabla f(x)\right\| \\ &\leq -\rho(\|x\|) + \hat{\sigma}(\|u\|).\end{aligned}$$

□

Remark 5.6.1. The condition in (Proposition 5.5) is a particular case of (Proposition 5.6). Indeed, it corresponds to $\rho(\|x\|) = \frac{1}{2}\|x\|$:

$$\|x\| - \left\|x - H^{-1}(x)\nabla f(x)\right\| \geq \rho(\|x\|) = \frac{1}{2}\|x\| \iff \left\|x - H^{-1}(x)\nabla f(x)\right\| \leq \frac{1}{2}\|x\|.$$

The particular form of $\rho(\|x\|) = \frac{1}{2}\|x\|$, makes it not only positive definite, but also $\mathcal{K}_\infty$ so that one obtain the stronger result of the system being ISS (not only integral ISS). ◀

Example 5.9. Suppose to have a function f such that its gradient is given by

$$\nabla f(x) = \exp\left(-\frac{\sqrt{x^2+1} - 3x^2\log(x) + x^2\log(\sqrt{x^2+1}+1)+1)}{2x^2}\right)$$

so that the hessian is

$$H(x) = \left(\frac{-\frac{x}{\sqrt{x^2+1}} - 2x\log(\sqrt{x^2+1}+1) - \frac{x^3}{\sqrt{x^2+1}(\sqrt{x^2+1}+1)} + 3x + 6x\log(x)}{2x^2} - \frac{-\sqrt{x^2+1} + 3x^2\log(x) - x^2\log(\sqrt{x^2+1}+1) - 1)}{x^3}\right)$$
$$\exp\left(-\frac{\sqrt{x^2+1} - 3x^2\log(x) + x^2\log(\sqrt{x^2+1}+1)+1)}{2x^2}\right).$$

It easy to compute

$$\begin{aligned}\|x\| - \left\|x - H^{-1}(x)\nabla f(x)\right\| &= \|x\| - \left\|x - x + \frac{x}{\sqrt{x^2+1}}\right\| \\ &\geq \underbrace{\frac{\|x\|}{x^2+1}\left(1 - \frac{1}{\sqrt{x^2+1}}\right)}_{\rho(\|x\|)}\end{aligned}$$

which is the condition of (5.6) that guarantees that the corresponding dynamics for Newton's Method

$$x_+ = x - H^{-1}(x)\nabla f(x) + u$$

is iISS with input u. Notice that $\rho \in \mathcal{P} \setminus \mathcal{K}$ so that this system is not ISS. A plot of the function ρ is given in the following (Figure 28).

◁

5.5 Stability for Quasi Newton Methods

ISS-Lyapunov function. We can try to apply the same ideas of (Proposition 5.2) to the (DFP Algorithm). Call $B \equiv B_k^{DFP}$ and $B_+ \equiv B_{k+1}^{DFP}$. Suppose $V(x) = \varphi(\|x\|)$ and compute

$$\begin{aligned}V(B_+) - V(B) &= \varphi(\|B_+\|) - \varphi(\|B\|) \\ &\leq \varphi\left(\left\|B - \frac{Bqq^TB}{q^TBq}\right\| + \|u\|\right) - \varphi(\|B\|) \\ &\leq \varphi\circ\psi\left(\|B\|\left\|I - \frac{qq^TB}{q^TBq}\right\|\right) - \varphi(\|B\|) + \sigma(\|u\|) \qquad (\bigstar)\end{aligned}$$

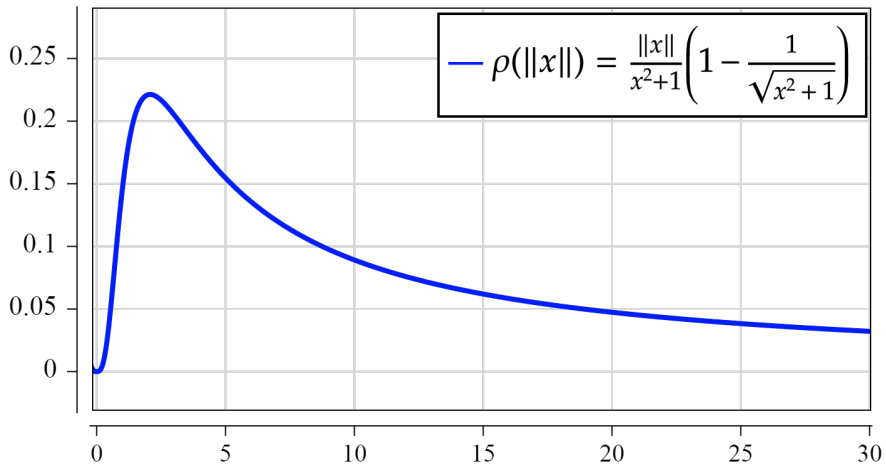


Figure 28: The function $\rho(x) \in \mathcal{P} \setminus \mathcal{K}$.

Now consider

$$\lambda := \min_k \lambda_{min}(B_k) > 0$$

and define $\psi(\|x\|) = \frac{1}{2}\sqrt{\frac{\lambda\|x\|}{2}}$ for $\|x\| \leq \frac{\lambda}{8}$ (ψ must be $> id$). Now we have

$$\begin{aligned}\psi\left(\|B\|\left\|I - \frac{qq^TB}{q^TBq}\right\|\right) &= \frac{1}{2}\sqrt{\frac{\lambda\|B\|\left\|I - \frac{qq^TB}{q^TBq}\right\|}{2}} \\ &\leq \frac{1}{2}\|B\|\end{aligned}$$

because

$$\begin{aligned}\|B\| = \frac{\|x\|}{\left\|I - \frac{qq^TB}{q^TBq}\right\|} &\geq \frac{\|x\|}{1 + \frac{\|qq^TB\|}{q^TBq}} \\ &\geq \frac{\|x\|\left\|q^TBq\right\|}{2\|B\|\|q\|^2} \geq \|x\|\frac{\lambda_{min}(B)}{2\|B\|} \\ \Rightarrow \|B\|^2 &\geq \frac{\lambda}{2}\|x\| \\ \Rightarrow \frac{1}{2}\|B\| &\geq \frac{1}{2}\sqrt{\frac{\lambda\|x\|}{2}}.\end{aligned}$$

Then we have

$$V(B_+) - V(B) \leq \varphi \circ \psi \left(\|B\| \left\| I - \frac{qq^T B}{q^T B q} \right\| \right) - \varphi(\|B\|) + \sigma(\|u\|) \qquad (\bigstar)$$
$$\leq \varphi\left(\frac{1}{2}\|B\|\right) - \varphi(\|B\|) + \sigma(\|u\|)$$
$$\leq -\alpha(\|B\|) + \sigma(\|u\|).$$

Comparing Newton's Method and BFGS. Since trying to look for incremental stability leads to many variables to handle, we can try to compare a Quasi Newton's Method with the original Newton's Method. In other words, consider a first system which follows the (BFGS Algorithm) and a second one who consists in the classical iteration step with an additional scaling:

$$y_+ = y - \alpha_2 H^{-1}(y)\nabla f(y)$$

(one can think of it as a BFGS method in which the hessian is computed exactly at each step, so that the updating of matrix B_k is just the evaluation of H^{-1} at the point y_k). In this case, the differences between trajectories in the two systems depends on the quality of the approximation in the update of B_k. The idea is to measure how the BFGS mimic the dynamic of the original Newton's Method. Fix the notation for the first system:

$$\begin{cases} d_1 = -B\nabla f(x) & \alpha_1 = \arg\min_\alpha f(x + \alpha d_1) \\ p = \alpha_1 d_1 & x_+ = x + p \\ q = \nabla f(x_+) - \nabla f(x) & B_+ = B + \left(1 + \frac{q^T B q}{p^T q}\right)\frac{pp^T}{p^T q} - \frac{Bqp^T + pq^T B}{p^T q} \end{cases} \qquad (\spadesuit_1)$$

and for the second system:

$$\begin{cases} d_2 &= -H^{-1}(y)\nabla f(y) \\ \alpha_2 &= \arg\min_\alpha f(x + \alpha d_2) \\ y_+ &= y - \alpha_2 H^{-1}(y)\nabla f(y) \end{cases} . \qquad (\spadesuit_2)$$

There are two different dynamics to compare:

$$\|x_+ - y_+\| = \|x + \alpha_1 d_1 - y - \alpha_2 d_2\|$$
$$= \left\|x - y + \alpha_2 H^{-1}(y)\nabla f(y) - \alpha_1 B^{-1}\nabla f(x)\right\| \qquad (56)$$
$$\left\|B_+ - H_+^{-1}\right\| = \left\| B + \left(1 + \frac{q^T B q}{p^T q}\right)\frac{pp^T}{p^T q} - \frac{Bqp^T + pq^T B}{p^T q} - H^{-1}(y_+) \right\| \qquad (57)$$

and we can write $H(y_+)$ in function of $H(y)$ to compare it with the matrix B:

$$H^{-1}(y_+) = H^{-1}(y) + \alpha_2 \nabla H^{-1}(\tau_y) d_2$$

for some $\tau_y = y + t(y_+ - y)$ with $t \in (0,1)$ and ∇H^{-1} a rank 3 tensor, so that

$$\left\|B_+ - H_+^{-1}\right\| = \left\| B - H^{-1}(y) + \left(1 + \frac{q^T B q}{p^T q}\right)\frac{pp^T}{p^T q} - \frac{Bqp^T + pq^T B}{p^T q} - \alpha_2 \nabla H^{-1}(y) H^{-1}(y)\nabla f(y) \right\|$$

We can use mean value theorem to express q as a function of H:

$$q = \nabla f(x + \alpha_1 d_1) - \nabla f(x) = \alpha_1 H(\tau_x) d_1$$

for some $\tau_x = x + t(x_+ - x)$ with $t \in (0, 1)$. In this way we can express the difference

$$\begin{aligned}
\left\|B_+ - H_+^{-1}\right\| = \Big\| & B - H^{-1}(y) - \alpha_2 \nabla H^{-1}(\tau_y) H^{-1}(y) \nabla f(y) \\
& + \left(1 + \frac{\alpha_1^2 \nabla f(x)^T (BH(\tau_x))^2 B \nabla f(x)}{\alpha_1^2 \nabla f(x)^T BH(\tau_x) B \nabla f(x)}\right) \frac{\alpha_1^2 B \nabla f(x) \nabla f(x)^T B}{\alpha_1^2 \nabla f(x)^T BH(\tau_x) B \nabla f(x)} \\
& - \frac{\alpha_1^2 BH(\tau_x) B \nabla f(x) \nabla f(x)^T B + \alpha_1^2 B \nabla f(x) \nabla f(x)^T BH(\tau_x) B}{\alpha_1^2 \nabla f(x)^T BH(\tau_x) B \nabla f(x)} \Big\| \\
= \Big\| & B - H^{-1}(y) - \alpha_2 \nabla H^{-1}(\tau_y) H^{-1}(y) \nabla f(y) \\
& + \left(1 + \frac{\nabla f(x)^T (BH(\tau_x))^2 B \nabla f(x)}{\nabla f(x)^T BH(\tau_x) B \nabla f(x)}\right) \frac{B \nabla f(x) \nabla f(x)^T B}{\nabla f(x)^T BH(\tau_x) B \nabla f(x)} \\
& - \frac{BH(\tau_x) B \nabla f(x) \nabla f(x)^T B + B \nabla f(x) \nabla f(x)^T BH(\tau_x) B}{\nabla f(x)^T BH(\tau_x) B \nabla f(x)} \Big\|.
\end{aligned}$$

Supposing $\alpha_1 = \alpha_2 \equiv 1$, the equation (56) can be rewritten

$$\begin{aligned}
\|x_+ - y_+\| &= \left\|x - y + (H^{-1}(y) - B^{-1}) \nabla f(y) + B^{-1}(\nabla f(y) - \nabla f(x))\right\| \\
&= \left\|x - y + (H^{-1}(y) - B^{-1}) \nabla f(y) + B^{-1} H(\tau_x^y)(y - x)\right\| \\
&= \left\|(H^{-1}(y) - B^{-1}) \nabla f(y) + (I - B^{-1} H(\tau_x^y))(x - y)\right\|
\end{aligned}$$

where $\tau_x^y = x + t(y - x)$ for some $t \in (0, 1)$. Or, using $\nabla f(y) - \nabla f(x) \approx H(y)(x - y)$,

$$\begin{aligned}
\|x_+ - y_+\| &= \left\|x - y + (H^{-1}(y) - B^{-1}) \nabla f(x) + H^{-1}(y)(\nabla f(y) - \nabla f(x))\right\| \\
&\approx \left\|x - y + (H^{-1}(y) - B^{-1}) \nabla f(x) + H^{-1}(y) H(y)(x - y)\right\| \\
&= \left\|(H^{-1}(y) - B^{-1}) \nabla f(x) + 2(x - y)\right\|.
\end{aligned}$$

We could try to express $H^{-1}(y) - B^{-1}$ as function of $H^{-1}(y) - B$ and inject (57) in (56). It can be verified that

$$\begin{aligned}
H^{-1}(y) - B^{-1} &= (H^{-1}(y) - B)^{-1}(I - H^{-1}(y) B^{-1} - BH^{-1}(y) + H^{-2}) \\
&= (H^{-1}(y) - B)^{-1}[(H^{-1}(y) - B) H^{-1}(y) - H^{-1}(y) B^{-1} + I].
\end{aligned}$$

Instead of comparing B_k and $H^{-1}(y)$ we could compare $F_k = B_k^{-1}$ and $H(y)$, knowing that the updating step for F_k is

$$F_+ = F + \frac{qq^T}{q^T p} - \frac{Fpp^T F}{p^T F p} \tag{49}$$

An easy computation shows that, if $FB = I = BF$, then $F_+B_+ = I = B_+F_+$.

$$\begin{aligned}
\|H_+ - F_+\| &= \left\| H - F + \eta - \frac{qq^T}{q^Tp} + \frac{Fpp^TF}{p^TFp} \right\| \\
&= \left\| H - F + \eta - \frac{H(\tau_x)B\nabla f(x)\nabla f(x)^TBH(\tau_x)}{\nabla f(x)^TBH(\tau_x)B\nabla f(x)} + \frac{\nabla f(x)\nabla f(x)^T}{\nabla f(x)^TB\nabla f(x)} \right\| \\
&= \left\| H - F + \eta - (1+\varepsilon)^2\frac{\nabla f(x)\nabla f(x)^T}{(1+\varepsilon)\nabla f(x)^TB\nabla f(x)} + \frac{\nabla f(x)\nabla f(x)^T}{\nabla f(x)^TB\nabla f(x)} \right\| \\
&= \left\| H - F + \eta - \varepsilon\frac{\nabla f(x)\nabla f(x)^T}{\nabla f(x)^TB\nabla f(x)} \right\|
\end{aligned}$$

where we used

$$H_+ = H+\eta, \qquad \tau_x = x+t(x_+-x) \text{ for some } t \in (0,1) \quad \text{and} \qquad H(\tau_x)B = BH(\tau_x) = I+\varepsilon.$$

Comparing the dynamics on x and y:

$$\begin{aligned}
\|x_+ - y_+\| &= \|x + \alpha_1 d_1 - y - \alpha_2 d_2\| \\
&= \left\|x - y + \alpha_2 H^{-1}(y)\nabla f(y) - \alpha_1 B\nabla f(x)\right\|.
\end{aligned}$$

Supposing $\alpha_1 = \alpha_2 \equiv 1$, this equation can be rewritten

$$\begin{aligned}
\|x_+ - y_+\| &= \left\|x - y + (H^{-1}(y) - B)\nabla f(y) + B(\nabla f(y) - \nabla f(x))\right\| \\
&= \left\|x - y + (H^{-1}(y) - B)\nabla f(y) + BH(\tau_x^y)(y - x)\right\| \\
&= \left\|(H^{-1}(y) - B)\nabla f(y) + (I - H(\tau_x^y))(x - y)\right\|
\end{aligned}$$

where $\tau_x^y = x+t(y-x)$ for some $t \in (0,1)$. Or, using $\nabla f(y) - \nabla f(x) \approx H(y)(x-y)$,

$$\begin{aligned}
\|x_+ - y_+\| &= \left\|x - y + (H^{-1}(y) - B)\nabla f(x) + H^{-1}(y)(\nabla f(y) - \nabla f(x))\right\| \\
&\approx \left\|x - y + (H^{-1}(y) - B)\nabla f(x) + H^{-1}(y)H(y)(x - y)\right\| \\
&= \left\|(H^{-1}(y) - B)\nabla f(x) + 2(x - y)\right\|.
\end{aligned}$$

We could try to express $H^{-1}(y) - B$ as function of $H(y) - F$ and combine the two dynamics...

Trying with SR1. SR1 is a Quasi Newton's Method that uses a rank one approximation of the Hessian matrix. The updating matrix step is given by

$$B_+ = B + \frac{(p - Bq)(p - Bq)^T}{q^T(p - Bq)}$$

using the same notations of BFGS and DFP. This update maintains the symmetry of the matrix but does not guarantee that the update be positive definite. Similar computations...

5.6 Updating B^{BFGS} with a Kronecker product

The vectorization of a matrix is a linear transformation which converts the matrix into a column vector.

Definition 5.4 (Vectorization)**.** The vectorization of an $m \times n$ matrix A, denoted $\mathrm{vec}(A)$, is the $mn \times 1$ column vector obtained by stacking the columns of the matrix A on top of one another:

$$\mathrm{vec}(A) = [a_{1,1}, \ldots, a_{m,1}, a_{1,2}, \ldots, a_{m,2}, \ldots, a_{1,n}, \ldots, a_{m,n}]^T.$$

A brief recap of the definition and some properties of a Kronecker product:

Definition 5.5 (Kronecker product)**.** If $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, then the Kronecker product $A \otimes B$ is the $mp \times nq$ block matrix:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{58}$$

Proposition 5.7 (Properties of the Kronecker product)**.** *The Kronecker product has the following properties:*

- *Bilinearity and associativity:*

$$\begin{aligned} A \otimes (B + C) &= A \otimes B + A \otimes C, \\ (A + B) \otimes C &= A \otimes C + B \otimes C, \\ (kB) \otimes B &= A \otimes (kB) = k(A \otimes B), \\ (A \otimes B) \otimes C &= A \otimes (B \otimes C); \end{aligned}$$

- *it is non-commutative, however, $A \otimes B$ and $B \otimes A$ are permutation equivalent, meaning that there exist permutation matrices P and Q (so called commutation matrices) such that*

$$A \otimes B = P\,(B \otimes A)\,Q;$$

- *mixed-product property:*

$$(A \otimes B)(C \otimes D) = (AC) \otimes (BD);$$

- *inverse of a Kronecker product:*

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1};$$

- *transposition and conjugate transposition are distributive over the Kronecker product:*

$$(A \otimes B)^T = A^T \otimes B^T \quad \textit{and} \quad (A \otimes B)^* = A^* \otimes B^*;$$

- *trace and determinant behave as follows for A and B square matrices of size n and m respectively:*

$$\operatorname{tr}(A \otimes B) = \operatorname{tr} A \operatorname{tr} B \quad \textit{and} \quad \det(A \otimes B) = (\det A)^m (\det B)^n;$$

- *the rank of the Kronecker product is the product of the ranks:*

$$\operatorname{rank}(A \otimes B) = \operatorname{rank} A \operatorname{rank} B.$$

The following result link the Kronecker product with the vectorization of a matrix:

Proposition 5.8 (Kronecker product and vectorization)**.** *For matrices of correct sizes yields*

$$\operatorname{vec}(AXB) = (B^T \otimes A) \operatorname{vec}(X)$$

As a consequence of (Proposition 5.8), one has:

Corollary 5.9 (Kronecker product and vectorization)**.** *If $X = I$ in (Proposition 5.8), then*

$$\operatorname{vec}(AB) = (I \otimes A) \operatorname{vec}(B) = (B^T \otimes I) \operatorname{vec}(A).$$

Moreover, for a vector $a \in \mathbb{R}^n$

$$\operatorname{vec}(aa^T) = a \otimes a.$$

Now we have all the instruments to write the updating step for the $B \equiv B_k^{BFGS}$ matrix

$$B_+ = B + \frac{(p^T q + q^T B_k q)(pp^T)}{(p^T q)^2} - \frac{Bqp^T + pq^T B}{p^T q}$$

in a vectorization form:

$$\begin{aligned}
\operatorname{vec}(B_+) &= \operatorname{vec}(B) - \frac{1}{p^T q}\left(\operatorname{vec}(Bqp^T) + \operatorname{vec}(pq^T B)\right) + \left(1 + \frac{q^T Bq}{p^T q}\right)\frac{1}{p^T q}\operatorname{vec}(pp^T) \\
&= \operatorname{vec}(B) - \frac{1}{p^T q}\left(\operatorname{vec}(Bqp^T) + K^{(n,n)}\operatorname{vec}(Bqp^T)\right) + \left(1 + \frac{q^T Bq}{p^T q}\right)\frac{1}{p^T q} p \otimes p \\
&= \left(I_{n^2} - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)\right)\operatorname{vec}(B) + \left(1 + \frac{q^T Bq}{p^T q}\right)\frac{1}{p^T q} p \otimes p \\
&=: \hat{M}_{p,q,n} \operatorname{vec}(B) + N_{p,q,B}.
\end{aligned}$$

Where $K^{(n,n)}$ is the commutation matrix such that $K^{(n,n)} \operatorname{vec}(A) = \operatorname{vec}(A^T)$ (in this case $A = Bqp^T$). In the second line we used (Corollary 5.9) and in the third one (Proposition 5.8). In this way the dynamic on B can be studied in this linear system, and sufficient conditions for its stability will lead to conditions on vectors p and q. The problem in this formulation is that B appears in the drift term $N_{p,q,B}$. This can be solved writing $q^T Bqpp^T$ in a clevver way:

$$q^T Bqpp^T = pq^T Bqp^T \Rightarrow \operatorname{vec}(q^T Bqpp^T) = \operatorname{vec}(pq^T Bqp^T) = (pq^T \otimes pq^T) \operatorname{vec}(B)$$

thus

$$\begin{aligned}\operatorname{vec}(B_+) &= \left(I_{n^2} - \frac{1}{p^Tq}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n) + \frac{1}{(p^Tq)^2}(pq^T \otimes pq^T)\right)\operatorname{vec}(B) + \frac{1}{p^Tq}p \otimes p \\ &=: M_{p,q,n}\operatorname{vec}(B) + N_{p,q}. \end{aligned} \tag{59}$$

Notice that the matrices $M_{p,q,n}$ and $N_{p,q}$, depending on p and q, are not constant at each step. Since p and q depend on k and in order to simplify the notation we rename them M_k and N_k respectively. Thus, the general term $\operatorname{vec}(B_{k+1})$ is written as a function of the initial guess $\operatorname{vec}(B_0)$ as

$$\operatorname{vec}(B_{k+1}) = \left(\prod_{j=0}^{k} M_j\right)\operatorname{vec}(B_0) + \sum_{i=0}^{k}\left(\prod_{j=i+i}^{k} M_j\right)N_i \tag{60}$$

where the notation $\prod_{j=0}^{k} M_j$ stands for the matrix product $M_kM_{k-1}\dots M_0$ and the product $\prod_{j=i+1}^{k} M_j$ is the identity matrix (empty product) if $i+1 \geq k$. For a more compact notation, we could include the first term in the summation, defining $N_{-1} := \operatorname{vec}(B_0)$:

$$\operatorname{vec}(B_{k+1}) = \sum_{i=-1}^{k}\left(\prod_{j=i+i}^{k} M_j\right)N_i. \tag{61}$$

<u>*Remark*</u> 5.9.1. Starting from the alternative equation (50) to update B_k, one obtains an alternative form for the matrix M_k:

$$M_k = \left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right) \otimes \left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right)$$

as

$$\begin{aligned} B_{k+1} &= \left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right)B_k\left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right)^T + \frac{p_kp_k^T}{p_k^Tq_k} \\ \Rightarrow \operatorname{vec}(B_{k+1}) &= \underbrace{\left[\left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right) \otimes \left(I - \frac{p_kq_k^T}{p_k^Tq_k}\right)\right]}_{M_k}\operatorname{vec}(B_k) + \underbrace{\frac{1}{p_k^Tq_k}\operatorname{vec}(p_kp_k^T)}_{N_k}. \end{aligned}$$

◀

We want to study the convergece of B_k to $\beta := \operatorname{vec}(H^{-1}(x^*))$.

Fact 1. Suppose that there is an uniform bound on the matrices M_k and the sequence N_k is bounded by a summable sequence, that is:

- $\|M_k\| \leq m < 1 \ \forall k$;
- $\forall k \ \exists n_k \ \ \|N_k\| \leq n_k$;

- $\sum_{k=0}^{\infty} n_k < +\infty$

Then the sequence B_k converges to a ball of center $\beta = \text{vec}(H^{-1}(x^*))$ and radius $\|\beta\|$. ♢

Proof. The equation (60) becomes

$$\begin{aligned}\|\text{vec}(B_{k+1}) - \beta\| &= \|M_k(\text{vec}(B_k) - \beta) + N_k + M_k\beta - \beta\| \\ &= \left\| \left(\prod_{j=0}^{k} M_j\right)(\text{vec}(B_0) - \beta) + \sum_{i=0}^{k}\left(\prod_{j=i+i}^{k} M_j\right) N_i + \left(\prod_{j=0}^{k} M_j\right)\beta - \beta \right\| \\ &\leq m^{k+1}\|\text{vec}(B_0) - \beta\| + \sum_{i=0}^{k} m^{k-i-1} n_i + (m^{k+1}+1)\|\beta\|.\end{aligned}$$

Now, the sum can be seen as the term from Cauchy product of the two series of n_i and m^{k-i-1}. Indeed, define $b_j := m^{j+1}$ so that

$$\left(\sum_{i=0}^{\infty} n_i\right)\cdot\left(\sum_{j=0}^{\infty} b_j\right) = \sum_{k=0}^{\infty} c_k \text{ where } c_k = \sum_{i=0}^{k} n_i b_{k-i} = \sum_{i=0}^{k} n_i m^{k-i-1}.$$

This product is well defined because we supposed $\sum_{k=0}^{\infty} n_k < +\infty$ and $m < 1$ so that $\left(\sum_{j=0}^{\infty} b_j\right) = \frac{1}{1-m}$. That means that the product series is convergent, which implies $c_k \to 0$. In other words

$$\begin{aligned}\|\text{vec}(B_{k+1}) - \beta\| &\leq \underbrace{m^{k+1}\|\text{vec}(B_0) - \beta\|}_{\searrow 0} + \underbrace{\sum_{i=0}^{k} m^{k-i-1} n_i}_{\searrow 0} + \underbrace{(m^{k+1}+1)\|\beta\|}_{\searrow \|\beta\|} \\ &\longrightarrow \|\beta\|.\end{aligned}$$

□

We study the eigenvalues of M_k to study the stability of this system. There are a few remarks that can be stated:

Remark 5.9.2.

- The spectral radius of $\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)$ verifies

$$\rho\left(\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)\right) = 1.$$

Proof. Unless $p = 0$,[6] it is an eigenvector for the rank one matrix pq^T with eigenvalue $q^T p$:

$$(pq^T)p = p(q^T p) = (q^T p)p$$

[6] If $p = 0$ at some step, then the gradient $\nabla f(x)$ would be zero and the algorithm has found a stationary point. So while the algorithm is running one can assume $p \neq 0$.

and therefore, from (Proposition 5.7) we know that $(pq^T \otimes pq^T)$ has rank 1 and the only eigenvalue which is non-zero is given by

$$\mathrm{tr}(pq^T \otimes pq^T) = \mathrm{tr}(pq^T)^2 = (q^T p)^2.$$

This means that $\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)$ has eigenvalues 0 with multiplicity $n^2 - 1$ and $\frac{1}{(q^T p)^2}(q^T p)^2 = 1$ with multiplicity 1. □

- The eigenvector of $\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)$ associated with the non-zero eigenvalue 1 is given by $\mathrm{vec}(pp^T)$:

$$\begin{aligned}\left[\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)\right]\mathrm{vec}(pp^T) &= \frac{1}{(p^T q)^2}\,\mathrm{vec}((pq^T)(pp^T)qp^T) \\ &= \frac{1}{(p^T q)^2}(q^T p)^2\,\mathrm{vec}(pp^T) \\ &= \mathrm{vec}(pp^T).\end{aligned}$$

- The commutation matrix $K^{(n,n)}$ has eigenvalues 1 and -1 with multiplicities $\frac{1}{2}n(n+1)$ and $\frac{1}{2}n(n-1)$. In particular this means that $\rho(K^{(n,n)}) = 1$.

Proof. The proof is given in [18, Theorem 3.1 (vi), p. 383]. □

- The spectral radius of $\frac{1}{p^T q}(I_{n^2} + K^{(n,n)})$ verifies

$$\rho\left(\frac{1}{p^T q}(I_{n^2} + K^{(n,n)})\right) \leq \frac{2}{|p^T q|}.$$

Proof. It is straightforward from the previous bullet point. □

- The spectral radius of $pq^T \otimes I_n$ verifies

$$\rho\left(pq^T \otimes I_n\right) = \left|p^T q\right|.$$

Proof. From (Proposition 5.7) we know that $pq^T \otimes I_n$ has rank n and trace $(q^T p)n$. Its eigenvalues are thus 0 with multiplicity $n^2 - n$ and we show that $q^T p$ is eigenvalue with multiplicity n. Let $v \in \mathbb{R}^{n^2} \setminus \{0\}$ an eigenvector of $pq^T \otimes I_n$ with eigenvalue λ. We claim that v is of the form

$$v \in \left\{ \begin{pmatrix} p_1 \\ \mathbf{0}_{n-1} \\ p_2 \\ \mathbf{0}_{n-1} \\ \vdots \\ p_n \\ \mathbf{0}_{n-1} \end{pmatrix}, \begin{pmatrix} 0 \\ p_1 \\ \mathbf{0}_{n-1} \\ p_2 \\ \mathbf{0}_{n-1} \\ \vdots \\ p_n \\ \mathbf{0}_{n-2} \end{pmatrix}, \ldots, \begin{pmatrix} \mathbf{0}_{n-1} \\ p_1 \\ \mathbf{0}_{n-1} \\ p_2 \\ \vdots \\ \mathbf{0}_{n-1} \\ p_n \end{pmatrix} \right\}$$

and a simple computation shows that

$$((pq^T \otimes I_n)v)_i = \begin{cases} (q^T p)p_k & v_i = p_k \\ 0 & v_i = 0 \end{cases} = (q^T p)v_i$$

that is v is an eigenvector with eigenvalue $q^T p$. □

- The two preceding bullet points imply that

$$\rho\left(\frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)\right) \leq \frac{2}{|p^T q|}\left|p^T q\right| = 2.$$

◀

From all that precedes the eigenvalues of M_k are between

$$0 \leq \lambda(M_k) \leq 2.$$

Moreover, we cannot bound the spectral radius of M_k with some constant $m < 1$: let $w \in \mathbb{R}^n \setminus \{0\}$ such that $pq^T w = 0$. Using w, one can construct $v \in \mathbb{R}^{n^2} \setminus \{0\}$ such that $(pq^T \otimes I)v = 0$ in a similar way as done above for the eigenvector. But then $(pq^T \otimes pq^T)v = 0$ as well and $M_k v = v$, that is $\rho(M_k) \geq 1$. This means in particular that $\|M_k\| \geq \rho(M_k) \geq 1$ for every induced matrix norm $\|\cdot\|$.

Remark 5.9.3. We have the following useful implications:

- $(pq^T \otimes I_n)v = 0 \Rightarrow (pq^T \otimes pq^T)v = 0$ and the converse implication is not true: take $p \neq 0$, $q = (0\ 1)^T$ and $v = (0\ 1\ 1\ 0)^T$.
- $(pq^T \otimes I_n)v = 0 \iff q \in \mathrm{Ker}(V)$ and $v = \mathrm{vec}(V)$.
- $(pq^T \otimes pq^T)v = 0 \iff q^T V q = 0$ and $v = \mathrm{vec}(V)$.

◀

Remark 5.9.4. In the (BFGS Algorithm) we suppose to initialize B_0 with a positive definite matrix. It can be shown that this property is preserved by the updating step. Unfortunately, this means that $x^T Bx > 0$ for every $x \in \mathbb{R}^n \setminus \{0\}$ and in particularly $q^T Bq > 0$, meaning that the third term that compose the matrix M_k is not vanishing:

$$\frac{1}{(p^T q)^2}(pq^T \otimes pq^T)\,\mathrm{vec}(B) = q^T Bq\ \mathrm{vec}(pp^T) \neq 0.$$

Using the equivalence $(pq^T \otimes I_n)v = 0 \iff q \in \mathrm{Ker}(V)$ and $v = \mathrm{vec}(V)$ from the remark above, B being positive definite means that

$$(pq^T \otimes I_n)\,\mathrm{vec}(B) \neq 0.$$

Thus, the second term composing the matrix M_k can be zero if and only if

$$(pq^T \otimes I_n)\,\mathrm{vec}(B) \in \mathrm{Ker}(I_{n^2} + K^{(n,n)}).$$

If this is the case, one cannot have $M_k\,\mathrm{vec}(B) < \mathrm{vec}(B)$. Otherwise, if we may prove that for the sequence $\{B_k\}_k$ holds $(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)\,\mathrm{vec}(B) \neq 0$, then there may be a chance that

$$\|M_k\,\mathrm{vec}(B)\| < \|\mathrm{vec}(B)\|$$

whereas p and q satisfy some conditions, and this ensures the stability of the dynamical system given by equation (60). ◀

Lemma 5.10 (Kernel of identity + commutation matrix)**.** $\mathrm{vec}(A) \in \mathrm{Ker}(I_{n^2} + K^{(n,n)})$ *if and only if* A *is skew-symmetric, i.e.* $A^T = -A$.

Proof. It is a simple computation:

$$\begin{aligned}(I_{n^2} + K^{(n,n)})\,\mathrm{vec}(A) = 0 &\iff K^{(n,n)}\,\mathrm{vec}(A) = -\,\mathrm{vec}(A)\\ &\iff \mathrm{vec}(A^T) = -\,\mathrm{vec}(A)\\ &\iff A^T = -A\end{aligned}$$

□

The (Lemma 5.10) above is good news in the light of preceding (Remark 5.9.4). Indeed B_k is a symmetric matrix at every step k so that $(pq^T \otimes I_n)\,\mathrm{vec}(B)$ is not a vectorization of a skew-symmetric matrix and therefore it is not in $\mathrm{Ker}(I_{n^2} + K^{(n,n)})$:

$$(p \neq 0 \text{ and } Bq \neq 0) \Rightarrow (pq^T \otimes I_n)\,\mathrm{vec}(B) = \mathrm{vec}(Bqp^T) \neq 0$$

so that there exists some index i, j for which $(Bqp^T)_{ij} \neq 0$, which means

$$\begin{aligned}&(Bqp^T)_{ij} \neq -(Bqp^T)_{ij} = -(pq^TB)^T_{ji}\\ &\Rightarrow\ Bqp^T \neq -(pq^TB)^T\\ \Rightarrow\ &(pq^T \otimes I_n)\,\mathrm{vec}(B) \notin \mathrm{Ker}(I_{n^2} + K^{(n,n)}).\end{aligned}$$

Now, we consider $v_1 = \mathrm{vec}(pp^T)$ and complete it to a basis of $\mathbb{R}^{n^2}$. Then we can decompose $\mathrm{vec}(B) = \gamma v_1 + w$ where $w \in \mathrm{Span}(v_1)^\perp$ so that $(pq^T \otimes pq^T)w = 0$. An explicit form for γ is given by

$$\gamma = \frac{\mathrm{vec}(pp^T)^T\,\mathrm{vec}(B)}{p^Tp} = \frac{\mathrm{tr}(pp^TB)}{p^Tp} = \frac{p^TBp}{p^Tp}.$$

Then we can compute

$$
\begin{aligned}
M_k \operatorname{vec}(B) &= \operatorname{vec}(B) - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)\operatorname{vec}(B) + \frac{1}{(p^T q)^2}(pq^T \otimes pq^T)\operatorname{vec}(B) \\
&= \operatorname{vec}(B) - \frac{\gamma}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)\operatorname{vec}(pp^T) - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n)w \\
&\quad + \gamma \operatorname{vec}(pp^T) \\
&= \operatorname{vec}(B) - \frac{\gamma}{p^T q}(I_{n^2} + K^{(n,n)})\operatorname{vec}(pp^T qp^T) - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})\operatorname{vec}(Wqp^T) \\
&\quad + \gamma \operatorname{vec}(pp^T) \\
&= \operatorname{vec}(B) - \gamma \operatorname{vec}(pp^T) - \gamma K^{(n,n)} \operatorname{vec}(pp^T) - \frac{1}{p^T q}\operatorname{vec}(Wqp^T) - \frac{1}{p^T q}K^{(n,n)} \operatorname{vec}(Wqp^T) \\
&\quad + \gamma \operatorname{vec}(pp^T) \\
&= \operatorname{vec}(B) - 2\gamma \operatorname{vec}(pp^T) - \frac{1}{p^T q}\operatorname{vec}(Wqp^T) - \frac{1}{p^T q}\operatorname{vec}(pq^T W) + \gamma \operatorname{vec}(pp^T) \\
&= w - \frac{1}{p^T q}\operatorname{vec}(Wqp^T) - \frac{1}{p^T q}\operatorname{vec}(pq^T W)
\end{aligned}
$$

where W is a symmetric matrix such that $w = \operatorname{vec}(W)$. This also shows that $v_1 = \operatorname{vec}(pp^T) \in \operatorname{Ker}(M_k)$. Recall that for any real matrix A and $p \geq 1$ the *entrywise* p-norm of A is equal to the p-norm of $\operatorname{vec}(A)$, so that in the particular case of $p = 2$ (i.e. the Frobenius norm) one has

$$\|A\|_F = \|A\|_2 = \|\operatorname{vec}(A)\|_2 = \sqrt{\operatorname{tr}(AA^T)}.$$

A (too) rough estimation on $\|M_k \operatorname{vec}(B)\|_2$ is then (Frobenius norm is not induced but it is submultiplicative)

$$\|M_k \operatorname{vec}(B)\|_2 \leq \|w\|_2 \left(1 + \frac{2}{|p^T q|}\|p\|_2 \|q\|_2\right)$$

so that the condition we wish for is

$$\|w\|_2 < \frac{|p^T q|}{|p^T q| + 2\|p\|_2 \|q\|_2}.$$

What we know about W:

- $q^T W q = 0$ because of previous remark and $(pq^T \otimes pq^T)w = 0$;
- $p^T W p = 0$ because

$$w \perp \operatorname{vec}(pp^T) \iff w^T \operatorname{vec}(pp^T) = \operatorname{tr}(W^T pp^T) = p^T W p = 0.$$

We can try to estimate the distance of the sequence from the desired limit

$\beta = \mathrm{vec}(H^{-1}(x^*))$:

$$\begin{aligned}\|\mathrm{vec}(B_{k+1}) - \beta\| &= \|M_k\,\mathrm{vec}(B_k) + N_k - \beta\| \\ &= \|(\mathrm{vec}(B_k) - \beta) + N_k + (M_k - I_{n^2})\,\mathrm{vec}(B_k)\| \\ &= \Big\|(\mathrm{vec}(B_k) - \beta) + \frac{1}{p_k^T q_k}\,\mathrm{vec}(p_k p_k^T) - \\ &\qquad (p_k^T B_k p_k)\,\mathrm{vec}(p_k p_K^T) - \frac{1}{p_k^T q_k}(\mathrm{vec}(W_k q_k p_k^T + p_k q_k^T W_k))\Big\|\end{aligned}$$

and decomposing

$$\mathrm{vec}(W_k q_k p_k^T) = (p_k^T p_k)(q_k^T W_k p_k)\,\mathrm{vec}(p_k p_k^T) + z_1$$

and

$$\mathrm{vec}(p_k q_k^T W_k) = (p_k^T p_k)(q_k^T W_k p_k)\,\mathrm{vec}(p_k p_k^T) + z_2$$

with $z_1, z_2 \in \mathrm{Span}(\mathrm{vec}(p_k p_k^T))^\perp$:

$$\|\mathrm{vec}(B_{k+1}) - \beta\| = \Big\|(\mathrm{vec}(B_k) - \beta) + \left(\frac{1 - 2(p_k^T p_k)(q_k^T W_k p_k)}{p_k^T q_k} - (p_k^T B_k p_k)\right)\mathrm{vec}(p_k p_k^T) - \frac{1}{p_k^T q_k}(z_1 + z_2)\Big\|.$$

Notice that z_1 and z_2 are relied by the following relation:

$$z_2 = \mathrm{vec}(Z^T) = K^{(n,n)}\,\mathrm{vec}(Z) = K^{(n,n)} z_1.$$

For numerical simulations, see (Appendix A.6).

We can simplify the computations with the following observations. Call

$$M_k = I_{n^2} - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n) + \frac{1}{(p^T q)^2}(pq^T \otimes pq^T) = I_{n^2} - \underbrace{A_1 A_2}_{A} + C$$

and remark that:

- If v is an eigenvector for A_2, there are only two possibilities:
  1. $v \in \mathrm{Ker}(A_2)$ and $Av = 0$;
  2. $v \notin \mathrm{Ker}(A_2)$ and $A_2 v = v$. In this case, the matrix $V \in \mathbb{R}^{n \times n}$ is not skew-symmetric (we know an explicit form for v from previous (Remark 5.9.2)) and $Av = A_1 v = 2v$.
- $AC = CA$, so there exists a common basis of eigenvectors for A and C.

Proof. Indeed C commutes with both A_1 and A_2:

$$AC = A_1 A_2 C = A_1 C A_2 = C A_1 A_2 = CA$$

□

- In the common basis of eigenvectors the two matrices A and C are written as

$$A = \begin{pmatrix} 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 2 & & \\ & & & & \ddots & \\ & & & & & 2 \end{pmatrix} \quad C = \begin{pmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & & 1 \end{pmatrix}$$

where the block of zeros for A has size $(n^2-n)\times(n^2-n)$ and the block of twos has size $n\times n$, while C is vanishing everywhere but for the coefficient in $C_{n^2,n^2}=1$.

Proof. Once the form for A is fixed, there are only two possibilities for C, that is the non-zero coefficient in the last entry of the diagonal or in the first one (corresponding to a 2 or a 0 for the matrix A). But the latter case is not possible: the eigenvector corresponding to the eigenvalue 1 for C is given by $\mathrm{vec}(pp^T)$: $C\,\mathrm{vec}(pp^T) = \mathrm{vec}(pp^T)$ and $A\,\mathrm{vec}(pp^T) = 2\,\mathrm{vec}(pp^T)$ so that C is written as above in this basis. □

- In this new basis the matrix M_k is diagonal

$$M = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & -1 & & & \\ & & & & \ddots & & \\ & & & & & -1 & \\ & & & & & & 0 \end{pmatrix}$$

where the ones are n^2-n, the -1s are $n-1$ and the last entry is zero. In particular the spectrum is given by $\mathrm{Spec}(M_k)=\{-1,0,1\}$, the spectral radius is $\rho(M_k)=1$ and the rank is $rk(M_k)=n^2-1$.

- We can explicit the basis of eigenvectors.
  - The first n^2-n vectors are such that $\begin{cases} Av=0 \\ Cv=0 \end{cases}$.

    Let then $w_1,\dots,w_{n-1}\in \mathrm{Span}(q)^\perp$ be linearly independent and consider

$$W_{i,j} = \begin{pmatrix} \mathbf{0} \\ w_i^T \\ \mathbf{0} \end{pmatrix} \in \mathbb{R}^{n\times n}$$

    where w_i^T is the j-th row. In this way we built $n(n-1)=n^2-n$ matrices such that $W_{i,j}q=0$, i.e. $w_{i,j}=\mathrm{vec}(W_{i,j})$ is such that

$Aw_{i,j} = 0 = Cw_{i,j}$. Moreover this vector are linearly independent by construction.

- The next $n-1$ vectors are such that $\begin{cases} Av = 2v \\ Cv = 0 \end{cases}$. Note this vectors $u_1, \dots, u_{n-1}$.
- The last one is a vector v such that $\begin{cases} Av = 2v \\ Cv = v \end{cases}$ and we already know that $v = \operatorname{vec}(pp^T)$.

- In this new basis we can write

$$\begin{aligned} \operatorname{vec}(B) &= \alpha_1 w_{1,1} + \cdots + \alpha_{n-1} w_{1,n-1} + \alpha_n w_{2,1} \cdots + \alpha_{n^2-n} w_{n,n-1} \\ &\quad + \beta_1 u_1 + \cdots + \beta_{n-1} u_{n-1} + \gamma \operatorname{vec}(pp^T) \\ M_k \operatorname{vec}(B) &= \alpha_1 w_{1,1} + \cdots + \alpha_{n-1} w_{1,n-1} + \alpha_n w_{2,1} \cdots + \alpha_{n^2-n} w_{n,n-1} \\ &\quad - \beta_1 u_1 - \cdots - \beta_{n-1} u_{n-1} \\ \operatorname{vec}(B_+) &= \alpha_1 w_{1,1} + \cdots + \alpha_{n-1} w_{1,n-1} + \alpha_n w_{2,1} \cdots + \alpha_{n^2-n} w_{n,n-1} \\ &\quad - \beta_1 u_1 - \cdots - \beta_{n-1} u_{n-1} + \tilde{\gamma} \operatorname{vec}(pp^T) \end{aligned}$$

where $\gamma = \frac{p^T B p}{p^T p}$ and $\tilde{\gamma} = 1/(p^T q)$. It is easy to see that the coefficient α_k corresponding to the vector $w_{i,j}$ is given by $\alpha_k = B_j^T w_i$, that is the scalar product between the j-th row of the matrix B and the vector w_i. This allows to explicitly compute the norms of $\operatorname{vec}(B)$ and $\operatorname{vec}(B_+)$ in terms of the coefficients α_i, β_i and γ, $\tilde{\gamma}$.

- Since the coefficient γ depends on B, we have the identity

$$\begin{aligned} \operatorname{vec}(B) - \gamma \operatorname{vec}(pp^T) &= \operatorname{vec}(B) - \frac{p^T B p}{p^T p} \operatorname{vec}(pp^T) \\ &= \operatorname{vec}(B) - \frac{1}{p^T p} \operatorname{vec}(pp^T B pp^T) \\ &= \left(I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T \right) \operatorname{vec}(B) \end{aligned}$$

and since the norm of a vector only depends on the absolute values of its coefficients:

$$\left\| \left(I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T \right) \operatorname{vec}(B) \right\|_2^2 = \sum_{j=1}^{n^2-n} |\alpha_j|^2 + \sum_{j=1}^{n-1} |\beta_j|^2 = \left\| \operatorname{vec}(B_+) - \tilde{\gamma} \operatorname{vec}(pp^T) \right\|_2^2$$

and using the sub-multiplicative property and the triangular inequality

$$\|\operatorname{vec}(B_+)\|_2 - |\tilde{\gamma}| \left\| \operatorname{vec}(pp^T) \right\|_2 \leq \left\| \operatorname{vec}(B_+) - \tilde{\gamma} \operatorname{vec}(pp^T) \right\|_2 \leq \left\| (I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T) \right\|_2 \|\operatorname{vec}(B)\|_2$$

which gives an upper bound on the norm of $\mathrm{vec}(B_+)$:

$$\|\mathrm{vec}(B_+)\|_2 \leq \left\| \left(I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T \right) \right\|_2 \|\mathrm{vec}(B)\|_2 + |\tilde{\gamma}| \left\|\mathrm{vec}(pp^T)\right\|_2 .$$

Now, the matrix $pp^T \otimes pp^T$ has rank 1, i.e. n^2-1 vanishing eigenvalues and the eigenvalue left is $\|p\|_2^4$ with corresponding eigenvector $v = \mathrm{vec}(pp^T)$:

$$(pp^T \otimes pp^T)\,\mathrm{vec}(pp^T) = \mathrm{vec}(pp^T pp^T pp^T) = (p^T p)^2\,\mathrm{vec}(pp^T) = \|p\|_2^4\,\mathrm{vec}(pp^T).$$

This tells us that

$$\left\| \left(I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T \right) \right\|_F^2 = \|p\|_2^4 - 2\|p\|_2^2 + n^2 \qquad \text{and} \qquad \left\| \left(I_{n^2} - \frac{1}{p^T p} pp^T \otimes pp^T \right) \right\|_2 = 1$$

and

$$\left\|\mathrm{vec}(pp^T)\right\|_2^2 = \mathrm{vec}(pp^T)^T\,\mathrm{vec}(pp^T) = \mathrm{tr}(pp^T pp^T) = \|p\|_2^4 .$$

In particular, using the sharpest bound of the 2-norm, the upper bound on $\|\mathrm{vec}(B_+)\|$ becomes

$$\|\mathrm{vec}(B_+)\|_2 \leq \|\mathrm{vec}(B)\|_2 + \underbrace{|\tilde{\gamma}|\,\|p\|_2^2}_{|u|} . \tag{62}$$

– The problem is that $\tilde{\gamma}$ depends on k and it blows up when $k \to +\infty$. However, assuming that $q_k \to 0$ slower than $p_k \to 0$, the drift term u_k is bounded.

This implies

$$\|\mathrm{vec}(B_{k+1})\|_2 \leq \|\mathrm{vec}(B_0)\|_2 + \sum_{j=0}^{k} |u_j| . \tag{63}$$

Summing on each side of (62) this gives

$$\sum_{j=0}^{k} \|\mathrm{vec}(B_{j+1})\|_2 \leq (k+1)\,\|\mathrm{vec}(B_0)\|_2 + \sum_{j=0}^{k} (k+1-j)\,|u_j| \qquad \text{for each } k \geq 0.$$

Proof. This can be proved by induction:

– for $k = 0$ there's nothing to prove;

– suppose that the inequality holds for k, adding $\|\mathrm{vec}(B_{k+2})\|_2$ to both sides and using the upper bound (63), we have

$$\begin{aligned}\sum_{j=0}^{k+1}\|\mathrm{vec}(B_{j+1})\|_2 &= \|\mathrm{vec}(B_{k+2})\|_2 + \sum_{j=0}^{k}\|\mathrm{vec}(B_{j+1})\|_2 \\ &\leq \|\mathrm{vec}(B_{k+2})\|_2 + (k+1)\,\|\mathrm{vec}(B_0)\|_2 + \sum_{j=0}^{k}(k+1-j)\,|u_j| \\ &\leq (k+2)\,\|\mathrm{vec}(B_0)\|_2 + \sum_{j=0}^{k}(k+2-j)\,|u_j|\end{aligned}$$

which proves the inequality.

□

- Starting from the inequality for squares

$$\|\mathrm{vec}(B_{k+1})\|_2^2 \leq \|\mathrm{vec}(B_k)\|_2^2 + \|u_k\|^2 + 2\,\|\mathrm{vec}(B_k)\|\,\|u_k\|$$

one obtains

$$\|\mathrm{vec}(B_{k+1})\|_2^2 \leq \|\mathrm{vec}(B_0)\|_2^2 + 2\,\|\mathrm{vec}(B_0)\|\sum_{j=0}^{k}\|u_j\| + \left(\sum_{j=0}^{k}\|u_j\|\right)^2.$$

Unfortunately, none of these inequalities corresponds to the characterization (13) of ISS from (Lemma 2.6).

5.7 NM and Quasi-NM in continuous time

Newton's Method in continuous time. Inspired by the reasoning of (Appendix A.5.2), we try to link Newton's iteration step, seen as a discrete time system, to a continuous time system which is hopefully easier to prove δISS via Lyapunov characterization. Then the same Lyapunov function should work in a neighborhood of the trajectories. The iteration step without noise is (28):

$$x_{k+1} = x_k - H(x_k)^{-1}\nabla f(x_k)$$

which we can rewrite as

$$x(t+\delta) - x(t) = -\delta H(x(t))^{-1}\nabla f(x(t))$$

by replacing $x_k = x(t)$, so that the continuous time system should be written

$$\dot{x} = -H(x)^{-1}\nabla f(x).$$

More precisely, suppose $f\colon \mathbb{R} \to \mathbb{R}$, we can consider the discretizations:

$$\nabla f(x) \approx \frac{f(x+h)-f(x)}{h} \qquad H(x) \approx \frac{f(x+2h)-2f(x+h)-f(x)}{h^2}$$

and

$$\begin{aligned}\frac{x(t+h)-x(t)}{h} &= -\frac{1}{h}\frac{h^2}{f(x+2h)-2f(x+h)-f(x)}\frac{f(x+h)-f(x)}{h}\\ &= -\frac{f(x+h)-f(x)}{f(x+2h)-2f(x+h)-f(x)}\end{aligned}$$

so that, for $h \to 0$ one has

$$\dot{x} = -H(x)^{-1}\nabla f(x).$$

The most general case corresponds to (42):

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla f(x_k) - H^{-1}(x_k)r_k - s_k\nabla f(x_k) - s_k r_k$$

which leads to the continuous time system

$$\dot{x} = -H^{-1}(x)\nabla f(x) + H^{-1}(x)r + s\nabla f(x) + sr.$$

The same idea could be applied to (DFP Algorithm) or (BFGS Algorithm), constructing a specific continuous dynamic for the matrices. In this case, however, one should consider the consistency of numerical method with respect to the continuous model. Let us consider the (DFP Algorithm).

BFGS in CT. The dynamics corresponding to x is given by

$$\begin{aligned}\frac{x_{k+1}-x_k}{\delta} &= \frac{p_k}{\delta} = u_k\\ \Rightarrow\ \dot{x}(t) &= u(t)\end{aligned}$$

From the definition of q_k:

$$\begin{aligned}\hat{q}_k = \frac{q_k}{\delta} &= \frac{\nabla f(x_{k+1}) - \nabla f(x_k)}{\delta}\\ &= H(x_k + \tau\delta u_k)u_k\\ \Rightarrow\ \hat{q}(t) &= H(x(t))u(t) = H(x(t))\dot{x}(t)\end{aligned}$$

So the equation for B is

$$\begin{aligned}\frac{B_+ - B}{\delta} &= \frac{1}{\delta}\frac{pp^T}{p^Tq} + \frac{1}{\delta}\frac{q^TBq}{(p^Tq)^2}pp^T - \frac{1}{\delta}\frac{Bqp^T + pq^TB}{p^Tq}\\ &= \frac{1}{\delta}\frac{uu^T}{u^T\hat{q}} + \frac{1}{\delta}\frac{\hat{q}^TB\hat{q}}{(u^T\hat{q})^2}uu^T - \frac{1}{\delta}\frac{B\hat{q}u^T + u\hat{q}^TB}{u^T\hat{q}}\end{aligned}$$

Unfortunately the first term blows up as $\delta \to 0$. Using the expression for u we may be able to avoid this: $u(t) = \frac{-\alpha(t)B(t)\nabla f(x(t))}{\delta}$ so that this last expression can be rewritten as

$$\frac{B_+ - B}{\delta} = -\frac{\alpha}{\delta^2}\frac{B\nabla f(x)\nabla f(x)^T B}{\nabla f(x)^T B\hat{q}} + \frac{1}{\delta}\frac{\hat{q}^T B\hat{q}}{(u^T\hat{q})^2}uu^T - \frac{1}{\delta}\frac{B\hat{q}u^T + u\hat{q}^T B}{u^T\hat{q}}$$

We can try to re-scale the hessian matrix. Call $\hat{B} = \eta B$ and the equation becomes:

$$\frac{\hat{B}_+ - \hat{B}}{\eta\delta} = -\frac{\alpha}{\eta\delta^2}\frac{B\nabla f(x)\nabla f(x)^T B}{\nabla f(x)^T B\hat{q}} + \frac{1}{\eta\delta}\frac{\hat{q}^T B\hat{q}}{(u^T\hat{q})^2}uu^T - \frac{1}{\eta\delta}\frac{B\hat{q}u^T + u\hat{q}^T B}{u^T\hat{q}}$$

... the problem is that p and q have the same scale size.

In order to realize a time-scale separation and make the transition to continuous time easier, one may use the following characterization of the matrix B_{k+1}^{BFGS}:

$$B_{k+1}^{BFGS} = \underset{B \ s.t. \ Bp_k = q_k}{\arg\min} \left\| W^{1/2}(B - B_k^{BFGS})W^{1/2} \right\|_F$$

where W is the average Hessian matrix

$$W = \int_0^1 H(x_k + \tau p_k)d\tau$$

and $\|\cdot\|_F$ is the Frobenius norm:

$$\|A\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\mathrm{trace}\,(A^{\mathsf{T}}A)} = \sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2(A)}.$$

Using the notation

$$\|A\|_W = \left\| W^{1/2}AW^{1/2} \right\|_F$$

we can write shortly

$$B_{k+1}^{BFGS} = \underset{B \ s.t. \ Bq_k = p_k}{\arg\min} \left\| B - B_k^{BFGS} \right\|_W .$$

In this case the diagram summarizing the (BFGS Algorithm) becomes (Figure 29).

We show that the solution to this problem is given by the matrix (50).

Proposition 5.11. *The solution of the problem*

$$\min_{B \ s.t. \ Bq_k = p_k} \left\| B - B_k^{BFGS} \right\|_W .$$

is given by the BFGS matrix

$$B_{k+1}^{BFGS} = \left(I - \frac{p_k q_k^T}{q_k^T p_k}\right) B_k^{BFGS} \left(I - \frac{q_k p_k^T}{q_k^T p_k}\right) + \frac{p_k p_k^T}{q_k^T p_k}. \tag{50}$$

Proof. Call

$$\hat{B} = W^{1/2} B W^{1/2}, \quad \hat{B}_k = W^{1/2} B_k W^{1/2}, \quad \hat{p}_k = W^{1/2} p_k, \quad \hat{q}_k = W^{-1/2} q_k.$$

Notice that since $W p_k = q_k$ then $\hat{q_k} = \hat{p_k}$. The problem becomes

$$\min \left\| \hat{B} - \hat{B}_k \right\|_F .$$

Let

$$u = \frac{\hat{p_k}}{\|\hat{p_k}\|} = \frac{\hat{q_k}}{\|\hat{q_k}\|}$$

and $U = \left[u | u_\perp\right]$. Now

$$\begin{aligned} U^T \hat{B}_k U - U^T \hat{B} U &= \begin{bmatrix} u^T \\ u_\perp^T \end{bmatrix} \hat{B}_k \begin{bmatrix} u & u_\perp \end{bmatrix} - \begin{bmatrix} u^T \\ u_\perp^T \end{bmatrix} \hat{B} \begin{bmatrix} u & u_\perp \end{bmatrix} \\ &= \begin{bmatrix} u^T \hat{B}_k u & u^T \hat{B}_k u_\perp \\ u_\perp^T \hat{B}_k u & u_\perp^T \hat{B}_k u_\perp \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & u_\perp^T \hat{B} u_\perp \end{bmatrix}. \end{aligned}$$

and the minimum for the Frobenius norm is obtained for $u_\perp^T \hat{B} u_\perp = u_\perp^T \hat{B}_k u_\perp$, that is

$$\begin{aligned} \hat{B} &= \begin{bmatrix} u & u_\perp \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & u_\perp^T \hat{B}_k u_\perp \end{bmatrix} \begin{bmatrix} u^T \\ u_\perp^T \end{bmatrix} \\ &= u u^T + u_\perp u_\perp^T \hat{B}_k u_\perp u_\perp^T \\ &= u u^T + (I - u u^T) \hat{B}_k (I - u u^T) \end{aligned}$$

using that

$$I = U U^T = \begin{bmatrix} u & u_\perp \end{bmatrix} \begin{bmatrix} u^T \\ u_\perp^T \end{bmatrix} = u u^T + u_\perp u_\perp^T \quad \Leftrightarrow \quad u_\perp u_\perp^T = I - u u^T .$$

At this point, changing variables back to the original ones gives the wished result. In fact, observe that

$$W^{-1/2} u = \frac{p_k}{\|\hat{p_k}\|} \qquad W^{1/2} u = \frac{q_k}{\|\hat{q_k}\|} \qquad \|\hat{q_k}\|^2 = p_k^T q_k$$

and compute

$$B_{k+1} = W^{-1/2} \hat{B}_k W^{-1/2} = \frac{p_k p_k^T}{p_k^T q_k} + B_k - \frac{p_k q_k^T B_k + B_k q_k p_k^T}{p_k^T q_k} + \frac{p_k q_k^T B_k q_k p_k^T}{(p_k^T q_k)^2}.$$

□

The idea is to speed up the dynamic on B with respect to the dynamic on x. Suppose to apply the following algorithm.

BFGS Algorithm with Time Scale Separation

1. Obtain a direction d_k by $d_k = -B_k \nabla f(x_k)$.
2. Perform a one-dimensional optimization (line search) to find an acceptable step-size α_k in the direction found in the first step, so $\alpha_k = \arg\min_{\alpha} f(x_k + \alpha d_k)$.
3. Set $p_k = \alpha_k d_k$ and update $x_{k+1} = x_k + p_k$.
4. $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
5. $B_{k+1}^{(1)} = \underset{B \ s.t. \ Bp_k = q_k}{\arg\min} \|B - B_k\|_W$
6. for $m = 1 : M$
   (a) $d_{k+1}^{(m)} = -B_{k+1}^{(m)} \nabla f(x_k)$;
   (b) $\alpha_{k+1}^{(m)} = \arg\min_{\alpha} f(x_k + \alpha\, d_{k+1}^{(m)})$;
   (c) Set $p_{k+1}^{(m)} = \alpha_{k+1}^{(m)} d_{k+1}^{(m)}$;
   (d) $B_{k+1}^{(m+1)} \underset{B \ s.t. \ Bp_{k+1}^{(m)} = q_k}{\arg\min} \left\| B - B_k^{(m)} \right\|_W$

   and set $\begin{cases} B_{k+1} = B_{k+1}^{(M)} \\ d_{k+1} = d_{k+1}^{(M)} \\ \alpha_{k+1} = \alpha_{k+1}^{(M)} \\ p_{k+1} = p_{k+1}^{(M)} \end{cases}$.

In this case we can see that for each k yields the following recurrence: $B_k^{(m+2)} = \frac{\alpha_k^{(m)}}{\alpha_k^{(m+1)}} B_k^{(m)}$ so that in the end (wlog M is even)

$$B_{k+1} = B_{k+1}^{(M)} = \frac{\prod_{i=0}^{M/2-2} \alpha_k^{(2i)}}{\prod_{i=0}^{M/2-1} \alpha_k^{(2i+1)}} B_k.$$

This means that we are only scaling B_k by a factor, and the approximation of the inverse of the Hessian is not improved.

What if, instead, we consider a variant of BFGS, namely L-BFGS [7]? This is equivalent to BFGS, when m is not fixed ($m = k$ at step k) and consists in the following:

[7] that stands for Limited-memory BFGS

Limited-Memory BFGS Algorithm (L-BFGS)

Fix $m > 0$ and choose a starting point x_0 and a matrix B_k^0. At iteration k:

1. $g = \nabla f(x_k)$
2. `for` $i = k-1, \ldots, k-m$
   (a) $\alpha_i = \frac{1}{q_i^T p_i} p_i^T g$
   (b) $g \leftarrow g - \alpha_i q_i$
3. $r = B_k^0 g$
4. `for` $i = k-m, \ldots, k-1$
   (a) $\beta \leftarrow \frac{1}{q_i^T p_i} q_i^T r$
   (b) $r \leftarrow r + (\alpha_i - \beta) p_i$

   now $B_k \nabla f(x_k) = r_k$.
5. $d_k = -r_k (= -B_k \nabla f(x_k))$
6. $\alpha_k = \arg\min_{\alpha} f(x_k + \alpha d_k)$.
7. Set $p_k = \alpha_k d_k$ and update $x_{k+1} = x_k + p_k$.
8. $q_k = \nabla f(x_{k+1}) - \nabla f(x_k)$.
9. If $k > m$ discard p_{k-m} and q_{k-m} from storage.

Its diagram is given in (Figure 30). In this scheme, time scale separation seems clearer and easy to operate. The idea is then: passing L-BFGS to CT, prove stability for it, obtain stability for L-BFGS, this is equivalent (with varying m) to stability for BFGS.

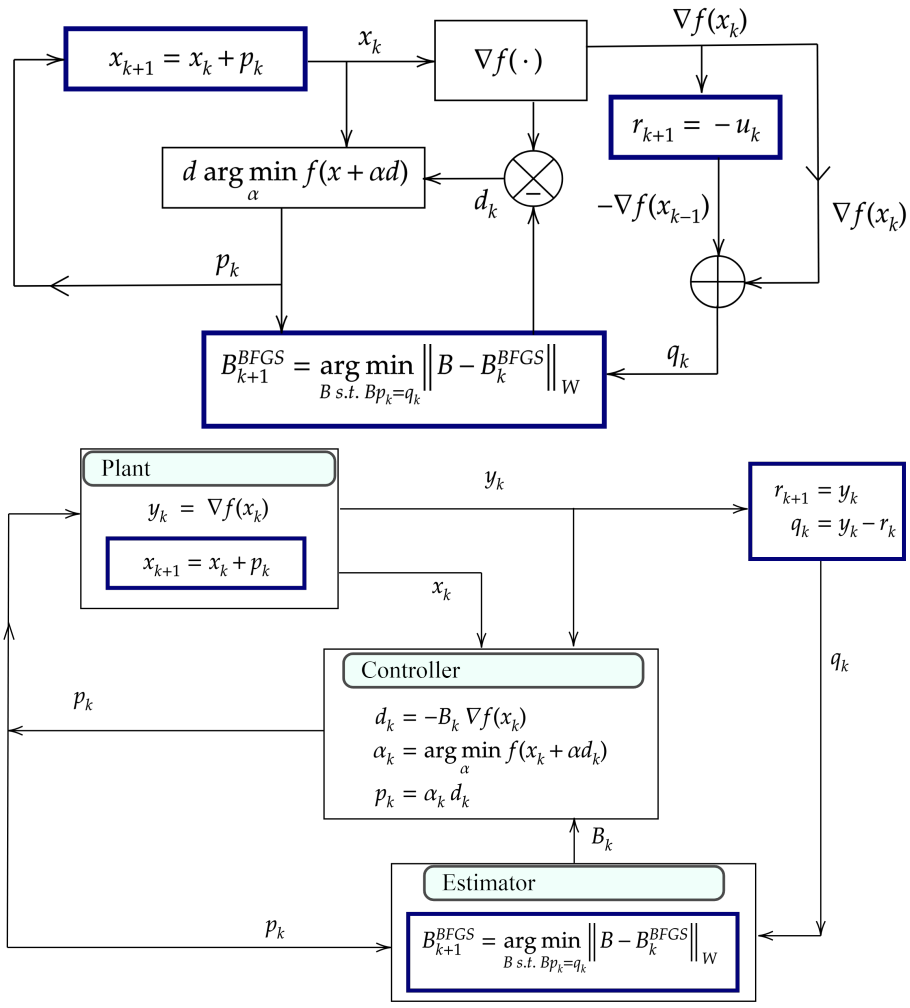


Figure 29: A diagram summarizing the BFGS algorithm, with implicit update on B_k.

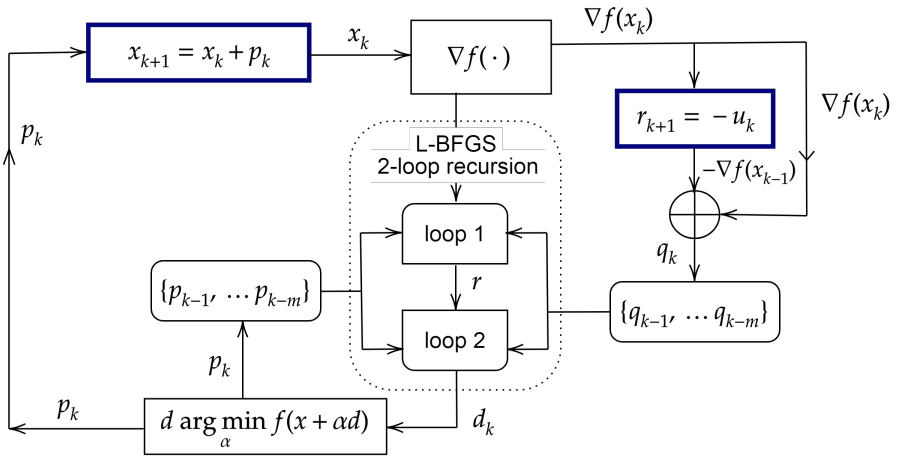


Figure 30: Diagram for L-BFGS algorithm.

6 CT

Suppose α depends on time: $\alpha = \alpha(s)$. The other parameters are defined by

$$\begin{cases} d(s) = -B(s)\nabla f(x(s)) \\ p(s) = \alpha(s)d(s) \end{cases}$$

Getting rid of the $\arg\min$: Assuming enough regularity on f, from $\alpha = \arg\min f(x + \alpha d)$ one derives

$$\dot{\alpha}(s) = -\frac{\left\langle \nabla f(x(s) + \alpha(s)d(s)), \dot{d}(s) \right\rangle}{\langle \nabla f(x(s) + \alpha(s)d(s)), d(s) \rangle}\alpha(s) - \frac{\langle \nabla f(x(s) + \alpha(s)d(s)), \dot{x}(s) \rangle}{\langle \nabla f(x(s) + \alpha(s)d(s)), d(s) \rangle}$$

Exact inverse of Hessian. Suppose $B(s) = [D\nabla f(x(s))]^{-1}$, then

$$\begin{cases} \dot{B}(s) = -\left[D\nabla f(x(s))\right]^{-1} D^2\nabla f(x(s))\dot{x}(s) \left[D\nabla f(x(s))\right]^{-1} \\ \dot{d}(s) = -\dot{B}(s)\nabla f(x(s)) - \alpha(s)d(s) \\ \qquad = \left[D\nabla f(x(s))\right]^{-1} D^2\nabla f(x(s))\dot{x}(s) \left[D\nabla f(x(s))\right]^{-1} \nabla f(x(s)) - \alpha(s)d(s) \\ \dot{p}(s) = \dot{\alpha}(s)d(s) + \alpha(s)\dot{d}(s) \\ \dot{x}(s) = p(s) = \alpha(s)d(s) \end{cases}$$

After some manipulations, one arrives to the form

$$\begin{cases} \dot{x}(s) = \alpha(s)d(s) \\ \dot{d}(s) = \alpha(s)\left[(D\nabla f(x(s)))^{-1}D^2\nabla f(x(s))d(s)(D\nabla f(x(s)))^{-1}\nabla f(x(s)) - d(s)\right] \\ \dot{p}(s) = \alpha(s)^2\Bigg[(D\nabla f(x(s)))^{-1}D^2\nabla f(x(s))d(s)(D\nabla f(x(s)))^{-1}\nabla f(x(s)) \\ \qquad - \frac{\left\langle \nabla f(x(s)+\alpha(s)d(s)), (D\nabla f(x(s)))^{-1}D^2\nabla f(x(s))d(s)(D\nabla f(x(s)))^{-1}\nabla f(x(s)) \right\rangle}{\langle \nabla f(x(s)+\alpha(s)d(s)), d(s) \rangle} d(s)\Bigg] - \alpha(s)d(s) \\ \dot{\alpha}(s) = \alpha(s)^2\left(1 - \frac{\left\langle \nabla f(x(s)+\alpha(s)d(s)), (D\nabla f(x(s)))^{-1}D^2\nabla f(x(s))d(s)(D\nabla f(x(s)))^{-1}\nabla f(x(s)) \right\rangle}{\langle \nabla f(x(s)+\alpha(s)d(s)), d(s) \rangle}\right) - \alpha(s) \\ \dot{B}(s) = -\alpha(s)(D\nabla f(x(s)))^{-1}D^2\nabla f(x(s))d(s)(D\nabla f(x(s)))^{-1} \end{cases} \tag{64}$$

Example 6.1: System (64) for a quadratic form. Let us explicit the system when

$$f(x) = \frac{1}{2}x^T H x^T + q^T x.$$

Since the Hessian is constant, B does not varies and all becomes simple:

$$\begin{cases} \dot{x}(s) = \alpha(s)d(s) \\ \dot{d}(s) = -\alpha(s)d(s) \\ \dot{p}(s) = -\alpha(s)d(s) \\ \dot{\alpha}(s) = \alpha(s)^2 - \alpha(s) \\ \dot{B}(s) = 0 \end{cases}$$

and the explicit solution is given by:

$$\begin{cases} x(s) = -e^{-s} \\ d(s) = 1 + e^{-s} \\ p(s) = e^{-s} \\ \alpha(s) = \frac{1}{e^s+1} \\ B(s) = H^{-1} \end{cases} .$$

◁

6.1 Conclusion of this section

In this section we made the link between Newton's Method, Quasi-Newton Method and all the stability properties presented in (SECTION 2). First, we provide "classical" Lyapunov functions for Newton's Method in general and for some specific functions presented as examples in (SUBSECTION 4.3). We tried to build ISS-Lyapunov functions, making use of the structure of the iteration step and of comparison functions. Sufficient conditions for the existence of ISS-Lyapunov functions are given (Proposition 5.2) and we remarked that these conditions are verified for some specific classes of functions. We tried to obtain similar conditions for incremental stability in (SUBSECTION 5.3), but without succes. In (SUBSECTION 5.4.0) two propositions guarantee integral input-to-state-stability of Newton's Method under mild assumptions. The research for stability for Quasi-Newton methods was more complex. No explicit conclusions were found for incremental stability - even after long computations. (SUBSECTION 5.6) was an attempt to prove stability for the dynamic given by the update on the matrix that approximates the inverse of the hessian in (BFGS Algorithm). In this section, we used vectorization and Kronecker product to rewrite the update equation in a linear form. We proved some results on the eigenvalues and the structures of the matrices that are present in the formula, then we proposed some numerical simulations (APPENDIX A.6) to better understand the behaviour of the dynamical system. Driven by the observation that it is usually easier to prove stability by Lyapunov functions for a continuous time system (APPENDIX A.5), in (SUBSECTION 5.7) and (SECTION 6) we tried to derive and solve the corresponding equations for Newton's Method and Quasi-Newton Method in continuous time.

A Appendix

A.1 List of Acronyms

BIBO	Bounded Input Bounded Output	
BFGS (algorithm)	Broyden–Fletcher–Goldfarb–Shanno (algorithm)	(BFGS Algorithm)
CICO	Converging Input Converging Output	
CT	Continuous Time	(Appendix A.5)
DFP (algorithm)	Davidon–Fletcher–Powel (algorithm)	(DFP Algorithm)
DT	Discrete Time	
δGAS	Incremental Globally Asymptotically Stable	(Definition 2.24)
δISS	Incremental Input to State Stability	(Definition 2.25)
GAS	Globally Asymptotically Stable	(Definition 5.1)
iISS	integral Input to State Stability	(Definition 2.22)
IOS	Input Output Stability	(subsection 2.2)
IOSS	Input Output to State Stability	(subsection 2.2)
ISS	Input to State Stability	(Definition 2.10)
$\mathcal{K}$-AG	$\mathcal{K}$-symptotic gain	(Definition 2.12)
L-BFGS (algorithm)	Limited-memory BFGS (algorithm)	(L-BFGS Algorithm)
LIM	Limit property	(Definition 2.13)
NM	Newton's Method	(section 4)
UBIBS	Uniformly Bounded Input Bounded State	(Definition 2.14)
UGAS	Uniformly Globally Asymptotically Stable	(Definition 5.2)

Table 2: All the acronyms used in the text with a reference to their definition.

A.2 Matrix inversion formulas

Trying to apply (Theorem 4.8) to DFP and BFGS algorithms in (subsection 4.6) an idea was to verify the assumptions by induction, thus expressing H_{k+1}^{-1} and B_{k+1}^{-1} as function of H_k^{-1} and B_k^{-1} respectively should help. Here's a collection of formulas for inverting sum of matrices:

Woodbury matrix identity: A, C invertible:

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1} \tag{65a}$$

when $U = V = I$

$$(A+C)^{-1} = A^{-1} - A^{-1}\left(C^{-1}+A^{-1}\right)^{-1}A^{-1} \tag{65b}$$

when $C = V = I$

$$(A+U)^{-1} = A^{-1} - A^{-1}U(A+U)^{-1} \tag{65c}$$

Sherman–Morrison formula: A invertible:

$$(A+uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u} \tag{65d}$$

No requirements on P or Q :

$$(I+P)^{-1} = I - (I+P)^{-1}P \tag{65e}$$

$$(I+PQ)^{-1}P = P(I+QP)^{-1} \tag{65f}$$

If A is invertible:

$$(A+BCD)^{-1} = A^{-1} - (I+A^{-1}BCD)^{-1}A^{-1}BCDA^{-1} \tag{65g}$$

$$= A^{-1} - A^{-1}(I+BCDA^{-1})^{-1}BCDA^{-1} \tag{65h}$$

$$= A^{-1} - A^{-1}B(I+CDA^{-1}B)^{-1}CDA^{-1} \tag{65i}$$

$$= A^{-1} - A^{-1}BC(I+DA^{-1}BC)^{-1}DA^{-1} \tag{65j}$$

$$= A^{-1} - A^{-1}BCD(I+A^{-1}BCD)^{-1}A^{-1} \tag{65k}$$

$$= A^{-1} - A^{-1}BCDA^{-1}(I+BCDA^{-1})^{-1} \tag{65l}$$

If C is also invertible we find Woodbury formula (65a) again:

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$$

A.2.1 Derivation of the update formula for B^{BFGS}

Applying (65d) twice to (49), we can derive the update formula (50) for the matrix $B \equiv B_k^{BFGS} = F_k^{-1}$. First, apply (65d) with $A = F \equiv F_k$, $u = q$ and $v = q/p^Tq$ and obtain

$$\left(F + \frac{qq^T}{p^Tq}\right)^{-1} = F^{-1} - \frac{F^{-1}qq^TF^{-1}}{p^Tq + q^TF^{-1}q} =: \hat{A}^{-1}.$$

Then with $A = \hat{A} = F + \frac{qq^T}{p^Tq}$ and $u = -Fp$, $v = Fp/(p^TFp)$:

$$\begin{aligned}
\underbrace{\left(F + \frac{qq^T}{p^Tq} - \frac{Fpp^TF}{p^TFp}\right)^{-1}}_{F_+^{-1}=B_+} &= \hat{A}^{-1} + \frac{1}{p^TFp}\frac{p^TFp}{p^TFp - p^TF\hat{A}^{-1}Fp}\hat{A}^{-1}Fpp^TF\hat{A}^{-1} \\
&= F^{-1} - \frac{F^{-1}qq^TF^{-1}}{p^Tq + q^TF^{-1}q} + \frac{pp^T}{p^TFp - p^TF\hat{A}^{-1}Fp} \\
&\quad - \frac{pp^Tqq^TF^{-1}}{(p^TFp - p^TF\hat{A}^{-1}Fp)(p^Tq + q^TF^{-1}q)} \\
&\quad - \frac{p^TqF^{-1}qp^T}{(p^TFp - p^TF\hat{A}^{-1}Fp)(p^Tq + q^TF^{-1}q)} \\
&\quad - \frac{F^{-1}qq^TF^{-1}(p^Tq)^2}{(p^TFp - p^TF\hat{A}^{-1}Fp)(p^Tq + q^TF^{-1}q)} \\
&= F^{-1} - \frac{F^{-1}qq^TF^{-1}}{p^Tq + q^TF^{-1}q} + \frac{(p^Tq + q^TF^{-1}q)}{(p^Tq)^2}pp^T \\
&\quad - \frac{pq^TF^{-1} + F^{-1}qp^T}{p^Tq} + \frac{F^{-1}qq^TF^{-1}}{p^Tq + q^TF^{-1}q} \\
&= B + \frac{(p^Tq + q^TBq)}{(p^Tq)^2}pp^T - \frac{pq^TB + Bqp^T}{p^Tq} \\
&= \left(I - \frac{pq^T}{q^Tp}\right)B\left(I - \frac{qp^T}{q^Tp}\right) + \frac{pp^T}{q^Tp}. \qquad (50)
\end{aligned}$$

A.2.2 Estimation on matrix norms

It is easy to see that, if $\|A\| < 1$ for some induced matrix norm $\|\cdot\|$, then $I - A$ is invertible and, from (65e),

$$\left\|(I-A)^{-1}\right\| \leq \frac{1}{1-\|A\|}.$$

This can be generalized as follows: suppose $\|A\| < m$. Then $I - \frac{1}{m}A$ is invertible and

$$\left\|\left(I - \frac{1}{m}A\right)^{-1}\right\| \leq \frac{1}{1 - \frac{\|A\|}{m}} = \frac{m}{m - \|A\|}$$

and this can be rewritten

$$\left\|(mI-A)^{-1}\right\| \leq \frac{1}{m-\|A\|}.$$

From this estimation we can also derive a lower bound for the inverse matrix:

$$1 + \|A\| \geq \|I - A\| \geq \frac{1}{\|(I-A)^{-1}\|} \geq 1 - \|A\|.$$

Proposition A.1 (Matrix norm and spectral radius)**.** *For every matrix A and $\varepsilon > 0$ there exists a induced matrix norm such that, call $\rho(A)$ the spectral radius,*

$$\rho(A) \leq \|A\| \leq \rho(A) + \varepsilon$$

Moreover, if all the eigenvalues μ of A such that $|\mu| = \rho(A)$ have corresponding Jordan blocks of dimension 1, then there exists an induced matrix norm such that $\|A\| = \rho(A)$.

Remark A.1.1. For every induced matrix norm, if $\rho(A)$ is the spectral radius of a matrix A,

$$\rho(A) \leq \|A\| .$$

In fact if x is an eigenvector of A with eigenvalue λ:

$$\|A\| \|x\| \geq \|Ax\| = \|\lambda x\| = |\lambda| \|x\|$$

wich implies $\|A\| \geq |\lambda|$ for all eigenvalue λ and so $\|A\| \geq \rho(A)$. ◀

A.3 Examples of ISS systems in discrete time

- As showed in (Example 2.6) by the mean of the ISS-Lyapunov function $V(x) = x^2$, the following system is ISS:

$$x_{k+1} = f(x_k, u_k) := \frac{x_k}{\sqrt{x_k^2 + 1}} + u_k \qquad \text{with } x_0 = \xi \in \mathbb{R}^n.$$

- The system

$$x_{k+1} = \frac{x_k}{2} + u_k$$

is clearly ISS. In fact using the definition:

$$\|x(k, \xi, u)\| = \left\| \frac{\xi}{2^k} + \sum_{i=0}^{k-1} \frac{u_{k-i}}{2^i} \right\| \leq \underbrace{\frac{\|\xi\|}{2^k}}_{\beta(\|\xi\|, k)} + \underbrace{2 \|u\|}_{\gamma(\|u\|)} .$$

- There exist other criteria to show that a discrete time system is ISS. Using sufficient criteria for ISS of discrete-time systems obtained with the help of indefinite difference Lyapunov functions as in [19], we can prove that the following system is also ISS:

$$x_{k+1} = \left(\frac{1}{8} + \sin^2 \left(\frac{k\pi}{4} \right) \right) x_k + u_k.$$

In order to show this, is enough to consider the indefinite Lyapunov function $V(x) = |x|$:

$$V(x_{k+1}) \leq \left(\frac{1}{8} + \sin^2 \left(\frac{k\pi}{4} \right) \right) |x_k| + |u_k|$$

which satisfies the assumptions of [19, Theorem 2, p. 73] with $M = \frac{9}{8}$, $T = 4$, $n = 1$, $\xi = \frac{255}{8^4}$.

A.4 Examples of δISS Lyapunov functions

Example A.1: Quadratic case. This is the trivial case: as observed in (Remark 4.0.3), in the quadratic case, Newton's method converges in only one step. So we can simply choose $V(x,y) = |x-y|$ as δISS Lyapunov function: if $f(x) = ax^2 + bx + c$ the iteration step gives

$$x_+ = x - \frac{1}{2a}(2ax+b) + u = -\frac{b}{2a} + u$$

and $V(x_+, y_+) = |u_1 - u_2|$ so that the condition (21b) is verified with

$$V(x_+,y_+) - V(x,y) = \underbrace{|u_1-u_2|}_{\sigma(|u_1-u_2|)} - \underbrace{|x-y|}_{\alpha(|x-y|)}$$

and $\alpha \in \mathcal{K}_\infty$ and $\sigma \in \mathcal{K}$. ◁

Example A.2: Quadratic + logarithmic case. Consider $f(x) = x^2 + \log(x^2+1)$. Easy computations show:

$$\nabla f(x) = 2x\left(\frac{x^2+2}{x^2+1}\right)$$
$$H(x) = 2\frac{x^4+x^2+2}{(x^2+1)^2}$$
$$x_+ = x - H(x)^{-1}\nabla f(x) = -\frac{2x^3}{x^4+x^2+2}.$$

Try $V(x,y) = |x-y|$ as a δISS Lyapunov function. Take any x, y:

$$\begin{aligned}
V(x_+,y_+) - V(x,y) &= \left|\frac{2x^3}{x^4+x^2+2} - \frac{2y^3}{y^4+y^2+2}\right| - |x-y| \\
&\le \left|\frac{-2x^3y^3(x-y) + 4(x^2+xy+y^2)(x-y) + 2x^2y^2(x-y)}{(x^4+x^2+2)(y^4+y^2+2)}\right| - |x-y| \\
&\le \left(2\left|\frac{-x^3y^3 + 2x^2 + 2xy + 2y^2 - x^2y^2}{(x^4+x^2+2)(y^4+y^2+2)}\right| - 1\right)|x-y| \\
&\le -0.1\,|x-y|
\end{aligned}$$

because

$$\max\left\{\left|\frac{(xy)^3 - 2x^2 - 2xy - 2y^2 - (xy)^2}{(x^4+x^2+2)(y^4+y^2+2)}\right|\right\} \approx 0.428792$$

at $(x,y) \approx (-0.804747, -0.804747)$[8]. This means that when adding distur-

[8] computed with WolframAlpha.

bances u_1 and u_2,

$$\begin{aligned} V(x_+, y_+) - V(x, y) &= \left| \frac{2x^3}{x^4 + x^2 + 2} + u_1 - \frac{2y^3}{y^4 + y^2 + 2} - u_2 \right| - |x - y| \\ &\leq \left| \frac{2x^3}{x^4 + x^2 + 2} - \frac{2y^3}{y^4 + y^2 + 2} \right| - |x - y| + |u_1 - u_2| \\ &\leq \underbrace{-0.1\,|x - y|}_{-\alpha_4(|x-y|)} + \underbrace{|u_1 - u_2|}_{\sigma(|u_1-u_2|)} \end{aligned}$$

such a V satisfies the (Definition 2.26) of δISS Lyapunov function. ◁

Remark A.1.2. In the previous example, one could have obtained a different constant, simply using the triangular inequality in a different way (e.g. applying (Lemma 2.2) to some comparison function). This underlies the following (Example A.3). ◀

In the following, we present some results that might be useful when looking for a Lyapunov function of the form $V(x, y) = |x - y|^\beta$.

Lemma A.2 (Hölder continuity of x^α)**.** *For any $0 < \alpha \leq 1$ the function $h(x) = x^\alpha$ is α-Hölder continuous:*

$$\exists c > 0 \quad |x^\alpha - y^\alpha| \leq c\,|x - y|^\alpha\,. \tag{66}$$

Moreover, h is β-Hölder continuous for all $0 < \beta \leq \alpha$ (and it is not for $\alpha < \beta$):

$$\exists c > 0 \quad |x^\alpha - y^\alpha| \leq c\,|x - y|^\beta \qquad \text{for } \ 0 < \beta \leq \alpha \leq 1.$$

Proof. Fix $\alpha \in (0, 1]$. The function $h(x) = x^\alpha$ is concave, so for any $a, b > 0$ $\left(\frac{a}{a+b}\right)^\alpha \geq 1$ and

$$\begin{aligned} &\left(\frac{a}{a+b}\right)^\alpha + \left(\frac{b}{a+b}\right)^\alpha \geq 1 \\ &\qquad \Rightarrow \quad a^\alpha + b^\alpha \geq (a+b)^\alpha. \end{aligned}$$

Now, wlog $x > y$ and we can set $a = x - y$, $b = y$ to obtain

$$\begin{aligned} &(x - y)^\alpha + y^\alpha \geq x^\alpha \\ &\Rightarrow \quad x^\alpha - y^\alpha \leq (x - y)^\alpha \end{aligned}$$

which is exactly the Hölder continuity condition (66) with $c = 1$. To establish the second inequality it is enough to observe that for $x > 0$

$$0 < \beta \leq \alpha \leq 1 \ \Rightarrow \quad x^\alpha \leq x^\beta.$$

□

Suppose $x_+ \sim x^\alpha$ for some $0 < \alpha \leq 1$. Then choosing $V(x,y) = |x-y|^\beta$ for a strictly positive β:

$$\begin{aligned} V(x_+, y_+) - V(x,y) &= |x_+ - y_+|^\beta - |x-y|^\beta \\ &\sim |x^\alpha - y^\alpha|^\beta - |x-y|^\beta \\ &\leq c^\beta |x-y|^{\alpha\beta} - |x-y|^\beta . \end{aligned}$$

Now, if $\alpha = 1$, if $c < 1$ the RHS is smaller than $-|x-y|^\beta =: -\alpha(|x-y|)$. If instead $0 < \alpha < 1$ the inequality becomes

$$\begin{aligned} V(x_+, y_+) - V(x,y) &\leq c^\beta |x-y|^{\alpha\beta} - |x-y|^\beta \\ &\leq |x-y|^{\alpha\beta} \left(c^\beta - |x-y|^{\beta(1-\alpha)}\right) \end{aligned}$$

and this is not always negative.

Example A.3: Quadratic + perturbation. Consider the function $f(x) = x^2 + x^2\sqrt{x}$. Then we can compute

$$\nabla f(x) = 2x + \frac{5}{2}\sqrt{x^3} \qquad H(x) = 2 + \frac{15}{4}\sqrt{x}$$

and the iteration step is

$$x_+ = x\frac{4 + 10\sqrt{x}}{8 + 15\sqrt{x}} + u = \bar{x} + u$$

where u is a disturbance. After computation, we observe that

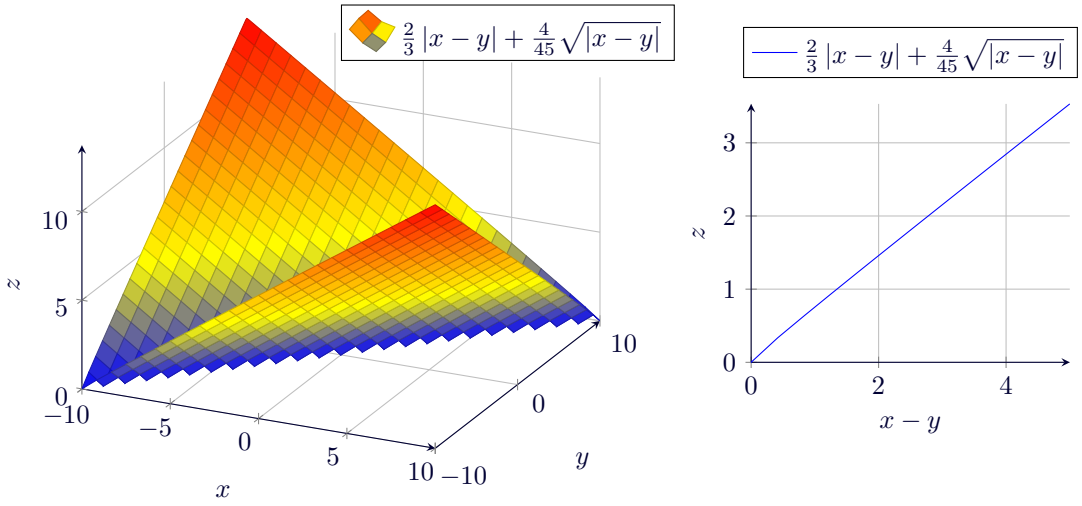
$$|\bar{x} - \bar{y}| \leq c_1 |x-y| + c_2\sqrt{|x-y|}$$

with $c_1 = \frac{2}{3}$ and $c_2 = \frac{4}{45}$:

$$\begin{aligned} |\bar{x} - \bar{y}| &= \left| x\frac{4 + 10\sqrt{x}}{8 + 15\sqrt{x}} - y\frac{4 + 10\sqrt{y}}{8 + 15\sqrt{y}} \right| \\ &= \frac{\left|(4x + 10x\sqrt{x})(8 + 15\sqrt{y}) - (4y + 10y\sqrt{y})(8 + 15\sqrt{x})\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \\ &= \frac{\left|32(x-y) - 60\sqrt{xy}(\sqrt{x} - \sqrt{y}) + 80(x\sqrt{x} - y\sqrt{y}) + 150\sqrt{xy}(x-y)\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \\ &= \frac{\left|(32 + 150\sqrt{xy} + 60\sqrt{x} + 60\sqrt{y})(x-y) - 20(y\sqrt{y} - x\sqrt{x})\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \\ &\leq \frac{\left|(32 + 150\sqrt{xy} + 80(\sqrt{x} + \sqrt{y}))(x-y)\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} + \frac{\left|20\sqrt{xy}(\sqrt{x} - \sqrt{y})\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \\ &\leq c_1 |x-y| + c_2 |x-y|^{1/2} \end{aligned}$$

with

$$c_1 = \max \frac{\left|32 + 150\sqrt{xy} + 60\sqrt{x} + 60\sqrt{y}\right|}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \leq \frac{2}{3}$$
$$c_2 = \max \frac{20\sqrt{xy}}{\left|(8 + 15\sqrt{y})(8 + 15\sqrt{x})\right|} \leq \frac{20}{225} = \frac{4}{45} \approx 0.0888888889$$



This inequality implies that there exist a K_∞ function φ, such that $\varphi - id \in \mathcal{K}_\infty$ and

$$|\bar{x} - \bar{y}| \leq c_1 |x - y| + c_2 \sqrt{|x - y|} \leq \varphi^{-1}\left(\frac{1}{2}|x - y|\right).$$

This means that

$$|\bar{x} - \bar{y}| \leq \varphi^{-1}\left(\frac{1}{2}|x - y|\right)$$
$$\varphi\left(|\bar{x} - \bar{y}|\right) \leq \frac{1}{2}|x - y|$$
$$\hat{\alpha}\left(\varphi\left(\bar{x} - \bar{y}\right)\right) \leq \hat{\alpha}\left(\frac{1}{2}|x - y|\right) \leq \frac{1}{2}\hat{\alpha}(|x - y|)$$

for any $\hat{\alpha} \in \mathcal{K}_\infty$ such that $\hat{\alpha}(\lambda s) \leq \lambda \hat{\alpha}(s)$ for all $s \geq 0$ and $0 < \lambda \leq 1$. Now,

take $V(x,y) = \hat{\alpha}(|x-y|)$ as a Lyapunov function. Then

$$\begin{aligned} V(x_+,y_+) - V(x,y) &= \hat{\alpha}(|\bar{x}+u-\bar{y}-v|) - \hat{\alpha}(|x-y|) \\ &\leq \hat{\alpha}(|\bar{x}-\bar{y}|+|u-v|) - \hat{\alpha}(|x-y|) \\ &\leq \hat{\alpha}(\varphi(|\bar{x}-\bar{y}|)) + \underbrace{\hat{\alpha}(\varphi\circ(\varphi-id)^{-1}(|u-v|))}_{\sigma(|u-v|)} - \hat{\alpha}(|x-y|) \\ &\leq \hat{\alpha}\left(\frac{1}{2}|x-y|\right) - \hat{\alpha}(|x-y|) + \sigma(|u-v|) \\ &\leq \underbrace{-\frac{1}{2}\hat{\alpha}(|x-y|)}_{-\alpha(|x-y|)} + \sigma(|u-v|) \end{aligned}$$

where we applied the triangle inequality for comparison functions (Lemma 2.2):

$$\hat{\alpha}(a+b) \leq \hat{\alpha}(\varphi(a)) + \hat{\alpha}(\varphi\circ(\varphi-id)^{-1}(b)).$$

This shows that V is a δISS Lyapunov function for the system given by the Newton's method applied to the function $f(x) = x^2 + x^2\sqrt{x}$ with additive disturbance u. The system is incrementally input to state stable thanks to the (Theorem 2.21). ◁

Remark A.2.1. Notice that the crucial step in the previous example was the inequality $|\bar{x}-\bar{y}| \leq \varphi^{-1}\left(\frac{1}{2}|x-y|\right)$. If this inequality holds, with a Lyapunov function of the form $\hat{\alpha}(|x-y|)$ for a $\mathcal{K}_\infty$ function satisfying the condition $\hat{\alpha}(\lambda s) \leq \lambda\hat{\alpha}(s)$ for all $s \geq 0$ and $0 < \lambda \leq 1$[9] and applying the triangle inequality, the last steps guarantee the incremental stability. The same reasoning can be applied to other functions as shown in the next examples. ◀

Example A.4: Variation on the previous example. Consider the function $f(x) = x^2 + x\sqrt{x}$, so that

$$\nabla f(x) = 2x + \frac{3}{2}\sqrt{x} \qquad H(x) = 2 + \frac{3}{4\sqrt{x}} = \frac{8\sqrt{x}+3}{4\sqrt{x}}$$

and the iteration step is

$$x_+ = -\frac{3}{8\sqrt{x}+3} + u = \bar{x} + u$$

where u is a disturbance. It is easy to see that

$$\begin{aligned} |\bar{x}-\bar{y}| &= \left|\frac{24(\sqrt{y}-\sqrt{x})}{64\sqrt{xy}+24(\sqrt{x}+\sqrt{y})+9}\right| \\ &\leq \frac{8}{3}\left|\sqrt{y}-\sqrt{x}\right| \leq \frac{8}{3}\sqrt{|x-y|} \\ &\leq \varphi^{-1}\left(\frac{1}{2}|x-y|\right) \end{aligned}$$

[9] For example $\hat{\alpha}(s) = s^2$ satisfies the condition.

for some $\varphi \in \mathcal{K}_\infty$. The Newton's method applied to this system is then δISS as observed in the preceding example and remark. ◁

A.5 Discrete time systems from continuous time systems

All the stability properties given in (SECTION 2) were originally for continuous time system, and (almost) all the results stated in the same section admit analogous formulations for continuous time systems. In the next subsections we will make a link between a stable (in the sense of ISS, iISS, δ ISS ...) continuous time system and its discrete time version.

A.5.1 ISS property from continuous time to discrete time systems

The ISS notion was originally proposed for continuous time systems in [20]. Similarly to (Definition 2.10), for a continuous time system

$$\dot{x} = f(x, u) \tag{67}$$

we have the following definition of the ISS property:

Definition A.1 (ISS - CT)**.** System (67) is called input-to-state stable (ISS) if for any essentially bounded input u and $x_0 \in \mathbb{R}^n$ there exist functions $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}$ such that

$$\|x(t, x_0, u)\| \leq \beta(\|x_0\|, t) + \gamma(\|u\|).$$

Remark A.2.2. As in the discrete time case, an equivalent formulation of ISS is

$$\|x(t, x_0, u)\| \leq \max\{\beta(\|x_0\|, t), \gamma(\|u\|)\}.$$

◀

Lyapunov method gives the possibility to analyze behavior of the system solutions without calculating the solutions proper as time function and the initial conditions which is a hard problem for the nonlinear systems. The analogous of (Definition 2.11) is:

Definition A.2 (ISS Lyapunov function - CT)**.** A smooth function $V\colon \mathbb{R}^n \to \mathbb{R}_+$ is called ISS Lyapunov function if for it there are the functions $\alpha_1, \alpha_2, \alpha_3, \sigma \in \mathcal{K}_\infty$ such that the conditions

$$\alpha_1(\|x\|) \leq V(x) \leq \alpha_2(\|x\|) \qquad \forall x \in \mathbb{R}^n \tag{68a}$$

$$\nabla V(x) \cdot f(x, u) \leq \sigma(\|u\|) - \alpha_3(\|x\|) \qquad \forall x \in \mathbb{R}^n \ \forall u \in \mathbb{R}^m \tag{68b}$$

Remark A.2.3. An equivalent definition by implication form is obtained replacing the second property (68b) by:

$$\exists\, \alpha_3, \chi \in \mathcal{K} \qquad \|x\| \geq \chi(\|u\|) \Rightarrow \nabla V(x) \cdot f(x,u) \leq -\alpha_3(\|x\|). \tag{69}$$

◀

Similar to (Theorem 2.7), in the continuous case the following statement yields.

Theorem A.3 (ISS Lyapunov characterization - CT)**.** *System (67) is ISS if and only if there exists for it an ISS Lyapunov function.*

Example A.5. Consider the nonlinear system

$$\dot{x} = -x^3 + u \qquad x, u \in \mathbb{R}.$$

By taking $V(x) = \frac{1}{2}x^2$, we obtain that

$$\dot{V} = x(-x^3 + u) = -x^4 + xu < -\frac{x^4}{2} =: -\alpha_3(\|x\|)$$

if $\|x\|^2 > \left(\frac{1}{\sqrt{2}}\|u\|\right)^{2/3} =: \chi(\|u\|)$. This means that $V(x) = \frac{1}{2}x^2$ can be used as a ISS Lyapunov function and the system is ISS by (Theorem A.3). ◁

We can link the continuous and the discrete time case as follows. Suppose we have a nonlinear system in continuous time (67)

$$\dot{x} = f(x,u)$$

which is proven to be ISS via a Lyapunov function $V(x)$ and (Theorem A.3). If we approximate the derivative $\dot{x} \approx \frac{x_{k+1}-x_k}{\delta t}$, we obtain a discrete time system

$$\frac{x_{k+1} - x_k}{\delta t} = f(x_k, u_k) \;\Rightarrow x_{k+1} = x_k + \delta t f(x_k, u_k).$$

Now, using Taylor expansion on V:

$$V(x_{k+1}) = V(x_k + \delta t f(x_k,u_k)) = V(x_k) + \delta t f(x_k,u_k)\nabla V(x_k) + O(\|\delta t f(x_k,u_k)\|)$$

and the condition (68b) $\nabla V(x) \cdot f(x,u) \leq \sigma(\|u\|) - \alpha_3(\|x\|)$ can be rewritten as

$$\frac{V(x_{k+1}) - V(x_k)}{\delta t} + O(\|f(x_k,u_k)\|) \leq \sigma(\|u_k\|) - \alpha_3(\|x_k\|)$$

which becomes the condition (14b)

$$V(x_{k+1}) - V(x_k) \leq \hat{\sigma}(\|u_k\|) - \hat{\alpha}_3(\|x_k\|)$$

for the discrete time system. In conclusion, when δt is small enough, if the continuous time system is ISS, the discretized system is still ISS.

Example A.6. Consider the nonlinear system of (Example A.5) and the corresponding discrete time system:

$$x_{k+1} = x_k - \delta t\, x_k^3 + \delta t u_k.$$

Take $V(x_k) = \frac{1}{2}x_k^2$ (which was a ISS Lyapunov function for the continuous time system) and compute

$$\begin{aligned} V(x_{k+1}) - V(x_k) &= \frac{1}{2}\left(x_k^2 + (\delta t)^2 x_k^6 + (\delta t)^2 u_k^2 - 2\delta t\, x_k^4 + 2\delta t\, x_k u_k - 2(\delta t)^2 x_k^3 u_k - x_k^2\right) \\ &\leq \frac{1}{2}\left((\delta t)^2 x_k^6 + (\delta t)^2 u_k^2 - 2\delta t\, x_k^4 + 2\delta t\, x_k^4 - 2(\delta t)^2 x_k^6\right) \\ &\leq \frac{(\delta t)^2}{2}(-x_k^6 + u_k^2) \\ &\leq -\hat{\alpha}_3(\|x_k\|) + \hat{\sigma}(u_k) \end{aligned}$$

for $\|x_k\|^2 > \chi(\|u_k\|) = \left(\frac{1}{\sqrt{2}}\|u_k\|\right)^{2/3}$. ◁

A.5.2 δISS property from continuous time to discrete time systems

A similar reasoning can be done for the incremental stability property. For completeness we give the analogous of (Definition 2.25) and (Definition 2.26) definitions.

Consider system (67) where $u \in U$ a closed and convex set of $\mathbb{R}^m$ containing the origin. Suppose also f locally Lipschitz and such that $f(0,0) = 0$.

Under these assumptions we define:

Definition A.3 (δISS - CT)**.** The system (67) is **incrementally input-to-state stable (δISS)** if there exists a function $\beta \in \mathcal{KL}$ and $\gamma \in \mathcal{K}_\infty$ such that for any $t \geq 0$, any ξ_1, $\xi_2 \in \mathbb{R}^n$ and any couple of input signals u_1, u_2 the following is true

$$\|x(t, \xi_1, u_1) - x(t, \xi_2, u_2)\| \leq \beta(\|\xi_1 - \xi_2\|, t) + \gamma(\|u_1 - u_2\|_\infty).$$

Remark A.3.1. Again, in the previous definition, the summation on the RHS may be replaced by $\max\{\beta(\|\xi_1 - \xi_2\|, t), \gamma(\|u_1 - u_2\|_\infty)\}$. ◀

Remark A.3.2. Since $f(0,0) = 0$ it is easy to check that δISS implies ISS just comparing an arbitrary trajectory with $x(t) \equiv 0$ (if one chooses $u_2 = 0$ and $\xi_2 = 0$ then $\dot{x} = f(0,0) = 0$ and $x(t) \equiv 0$). ◀

A necessary condition for ISS is the following: $\forall u \in U\ \exists!\, x_u$ sucht that $f(x_u, u) = 0$. This follows from the following proposition.

Proposition A.4 (Necessary condition for δISS - CT)**.** *Suppose the system (67) δISS. Then, for all constant input signals there exists a unique, globally asymptotically stable, equilibrium point.*

Remark A.4.1. Trajectories of ISS systems all converge to one another. ◀

Definition A.4 (δISS Lyapunov function - CT)**.** A smooth function $V : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ is called a **δISS Lyapunov function** if for any $u_1,\ u_2 \in U$ and any $x_1,\ x_2 \in \mathbb{R}^n$

$$\alpha_1(\|x_1 - x_2\|) \leq V(x_1, x_2) \leq \alpha_2(\|x_1 - x_2\|) \tag{70a}$$

$$\|u_1 - u_2\| \leq \kappa(\|x_1 - x_2\|) \Rightarrow \partial_{x_1} V f(x_1, u_1) + \partial_{x_2} V f(x_2, u_2) < -\rho(\|x_1 - x_2\|) \tag{70b}$$

where $\alpha_1,\ \alpha_2,\ \kappa \in \mathcal{K}_\infty$ and ρ is positive definite.

Theorem A.5 (δISS and δISS Lyapunov function - CT)**.** *If the system (67) admits a δISS Lyapunov function, then it is δISS. Moreover, if the set U is compact the two conditions are equivalent.*

Example A.7. Consider the nonlinear system

$$\dot{x} = -x + u$$

with two different solution x_1, x_2 corresponding to inputs u_1, u_2 respectively. Take the Lyapunov function $V(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$ for $\|u_1 - u_2\| \leq \frac{1}{2}\|x_1 - x_2\| = \kappa(\|x_1 - x_2\|)$. Then

$$\begin{aligned}
\partial_{x_1} V f(x_1, u_1) + \partial_{x_2} V f(x_2, u_2) &= (x_1 - x_2)(-x_1 + u_1) + (x_2 - x_1)(-x_2 + u_2) \\
&= -x_1^2 - x_2^2 + 2x_1x_2 + x_1u_1 - x_2u_1 - x_1u_2 + x_2u_2 \\
&= -(x_1 - x_2)^2 + (x_1 - x_2)(u_1 - u_2) \\
&\leq -\frac{1}{2}(x_1 - x_2)^2 =: -\rho(\|x_1 - x_2\|)
\end{aligned}$$

So the system is δISS thanks to (Theorem A.5). ◁

Now we can link the continuous and discrete time cases. Consider system

$$\dot{x} = f(x, u) \tag{67}$$

and suppose it δISS with δISS Lyapunov function $V(x, y)$. Using Taylor's expansion on V:

$$V(x + h, y + k) = V(x, y) + \partial_x V(x, y)h + \partial_y V(x, y)k + O(\|h + k\|^2)$$

in a neighborhood of (x, y). Consider now the discretization of the system:

$$\frac{x_{k+1} - x_k}{\delta t} = f(x_k, u_k) \ \Rightarrow x_{k+1} = x_k + \delta t f(x_k, u_k).$$

As in (SUBSECTION 2.6), abbreviate $x(k, \xi_i, u_i)$ in x^i for the sake of simplicity. The condition (70b) gives, for $\|u_1 - u_2\| \leq \kappa(\|x^1 - x^2\|)$

$$\begin{aligned} V(f(x^1, u_1), f(x^2, u_2)) - V(x^1, x^2) &\leq \partial_{x_1} V \, \delta t \, f(x_1, u_1) + \partial_{x_2} V \, \delta t \, f(x_2, u_2) \\ &< -\delta t \, \rho(\|x^1 - x^2\|) = \hat{\rho}(\|x^1 - x^2\|) \end{aligned}$$

which guarantees the δISS property for the discrete time system thanks to (Theorem 2.21).

Example A.8. Apply the preceding reasoning to (Example A.7), one obtains:

$$x_{k+1} = x_k + \delta(-x_k + u_k) = (1 - \delta)x_k + \delta u_k$$

and taking two different solutions $x = x(k, \xi_1, u)$ and $y = x(k, \xi_2, v)$, for the Lyapunov function $V(x, y) = \frac{1}{2}(x - y)^2$ and $\|u - v\| \leq \frac{1}{2}\|x - y\|$, yields

$$\begin{aligned} V(x_{k+1}, y_{k+1}) - V(x_k, y_k) &= \frac{1}{2}\Big((1-\delta)^2 x_k^2 + \delta^2 u_k + 2(1-\delta)\delta x_k u_k + \\ &\quad (1-\delta)^2 y_k^2 + \delta^2 v_k + 2(1-\delta)\delta y_k v_k \\ &\quad - 2(1-\delta)^2 x_k y_k - 2\delta(1-\delta) x_k v_k - 2\delta(1-\delta) y_k u_k \\ &\quad - 2\delta^2 u_k v_k - x_k^2 - y_k^2 + 2x_k y_k\Big) \\ &= \frac{1}{2}\Big((1-\delta)^2 - 1)(x_k - y_k)^2 + \delta^2(u_k - v_k)^2 \\ &\quad + 2\delta(1-\delta)(x_k - y_k)(u_k - v_k)\Big) \\ &\leq \frac{1}{2}\big((1-\delta)^2 - 1 + \frac{\delta^2}{2} + \delta(1-\delta)\big)(x_k - y_k)^2 \\ &= \frac{1}{2}\delta\left(\frac{\delta}{2} - 1\right)(x_k - y_k)^2 =: -\rho(\|x - y\|) \end{aligned}$$

with $\rho(s) = \delta\left(1 - \frac{\delta}{2}\right)s^2$ for $\delta < 2$. ◁

A.6 Results from numerical simulations on BFGS

Remind the updating formula for B_k:

$$B_{k+1}^{BFGS} = \left(I - \frac{p_k q_k^T}{q_k^T p_k}\right) B_k^{BFGS} \left(I - \frac{q_k p_k^T}{q_k^T p_k}\right) + \frac{p_k p_k^T}{q_k^T p_k}. \tag{50}$$

And its vectorization:

$$\begin{aligned} \operatorname{vec}(B_{k+1}) &= \left(I - \frac{p_k q_k^T}{p_k^T q_k}\right) \otimes \left(I - \frac{p_k q_k^T}{p_k^T q_k}\right) \operatorname{vec}(B_k) + \frac{1}{p^T q} p \otimes p \\ &= \left(I_{n^2} - \frac{1}{p^T q}(I_{n^2} + K^{(n,n)})(pq^T \otimes I_n) + \frac{1}{(p^T q)^2}(pq^T \otimes pq^T)\right) \operatorname{vec}(B_k) + \frac{1}{p^T q} p \otimes p \\ &=: M_k \operatorname{vec}(B_k) + N_k. \end{aligned} \tag{59}$$

The idea is to prove that $\{\text{vec}(B_k)\}_k$ is convergent, so we try to prove that it is a Cauchy sequence

$$\forall \varepsilon > 0 \; \exists K > 0 \text{ s.t. } \; \forall k \geq K \; \forall m > 0 \quad \|\text{vec}(B_{k+m}) - \text{vec}(B_k)\| \leq \varepsilon$$

for some small constant $c < 1$. Now, since

$$\|\text{vec}(B_{k+m}) - \text{vec}(B_k)\| \leq \sum_{i=0}^{m-1} \|\text{vec}(B_{k+i+1}) - \text{vec}(B_{k+i})\|$$

and

$$\|\text{vec}(B_{k+1}) - \text{vec}(B_k)\| \leq \|N_k\| + \|M_k - I_{n^2}\| \, \|\text{vec}(B_k)\|$$

we would study the behaviour of $\|N_k\|$, $M_k \to I_{n^2}$ and $\|\text{vec}(B_k)\|$ by numerical simulations.

Test functions I used the following test functions:

$$\begin{aligned}
f^{(1)}(x) &= \frac{1}{2} x^T \begin{pmatrix} 6 & -2 \\ -2 & 6 \end{pmatrix} x + \begin{pmatrix} 10 & 5 \end{pmatrix} x \\
f^{(2)}(x) &= \frac{1}{2} x^T \begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix} x + \begin{pmatrix} 10 & 5 \end{pmatrix} x \\
f^{(3)}(x) &= \frac{1}{2} x^T \begin{pmatrix} 101 & -99 \\ -99 & 101 \end{pmatrix} x + \begin{pmatrix} 10 & 5 \end{pmatrix} x \\
f^{(4)}(x, y) &= 100(y - x^2)^2 + (x - 1)^2 \\
f^{(5)}(x, y) &= |3x + 2y - 2|^2 + 10(|x| + |y|)
\end{aligned}$$

Remark that $f^{(2)}$ and $f^{(3)}$ are quadratic functions with less nicely conditioned Hessian than $f^{(1)}$; $f^{(4)}$ is Rosenbrock's valley-shaped function; $f^{(5)}$ is a non-differentiable lasso function. A plot of these functions can be found in (Appendix A.6).

After 1000 steps we have the parameters as in (Table 3).[10]

f	$\left\|M_\infty^{(i)} - I_4\right\|_F$	$\|\text{vec}(B_\infty)\|$	$\|\text{vec}(B_{\infty+1}) - \text{vec}(B_\infty)\|$	$\|N_\infty\|$	$\|B_\infty - \beta\|_F$
$f^{(1)}$	1.8671	0.0027	1.7618e-07	0.0020	0.2768
$f^{(2)}$	2.3534	0.0051	2.9116e-05	0.0031	0.5102
$f^{(3)}$	4.2152	0.0022	6.3568e-06	0.0021	0.4979
$f^{(4)}$	12.2826	0.0227	1.0166e-05	0.0226	2.4813
$f^{(5)}$	3.7745e+68	3.2324e+16	5.3924e+15	2.6962e+16	3.2324e+16

Table 3: Numerical values at the 1000th step of a BFGS simulation for the test functions.

[10] For the function $f^{(5)}$ the values are taken after 900 iterations, they tend to blow up after $k = 470$ and everything will be `NaN`.

From (Table 3), we remark that the term $\|\text{vec}(B_{k+1}) - \text{vec}(B_k)\|_F$ seems to converge to 0. The norm of N_k seems always to converge to a constant, which seems to be 0 as well. The norm of $\text{vec}(B_k)$ appears to be always bounded. The only problem is that the matrix M_k is not as close to the identity matrix as we were expecting. The last column shows that we have convergence of the matrices B_k to the actual inverse of the Hessian at the minimum point β for the three quadratic functions. We do not expect this to be small as we keep on iterate after convergence of the algorithm to the stationary point, re-scaling the gradient to make him always bigger than 10^{-8} (the machine epsilon is 10^{-16}). The last test function $f^{(5)}$ shows a different behaviour.[11]It needs more investigations.

In (Appendix A.6) there are some plots that shows the behaviour of the parameters in the table as a function of the iteration k.

In the first four cases, the algorithm converges exactly to the minimum, while in the last one, due to non-differentiability there's an error. I estimated the theoretical minimum by hand and find a smaller value of $f^{(5)}$ in $(1/9, 0)$. The results are summarized in (Table 4).

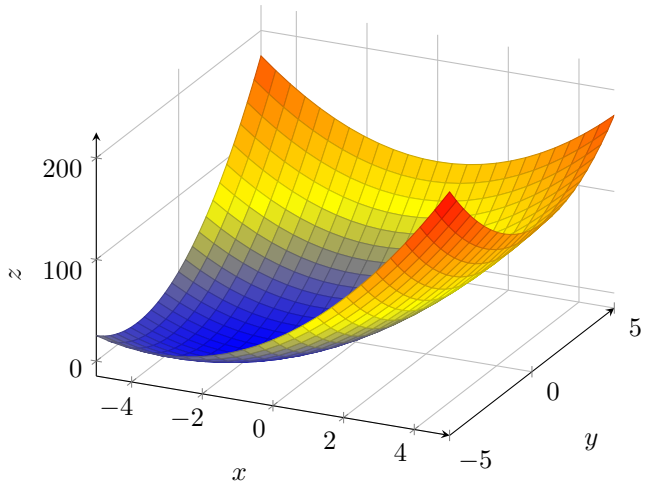
f	minimum found $\bar{x}$	theoretical point of minimum x^*	$f(\bar{x})$	$\lvert f(x^*) - f(\bar{x})\rvert$
$f^{(1)}$	(-2.1875, -1.5625)	(-2.1875, -1.5625)	-14.8438	0
$f^{(2)}$	(-4.0625, -3.4375)	(-4.0625, -3.4375)	-28.9062	0
$f^{(3)}$	(-3.7625, -3.7375)	(-3.7625, -3.7375)	-28.1562	0
$f^{(4)}$	(1,1)	(1,1)	0	0
$f^{(5)}$	(0.0433,-0.0001)	$f(x)\lvert_{x=(1/9,0)} = 3.\bar{8}$	3.9328	≤ 3.9328

Table 4: Points of minimum for the test functions: computed values and theoretical values.

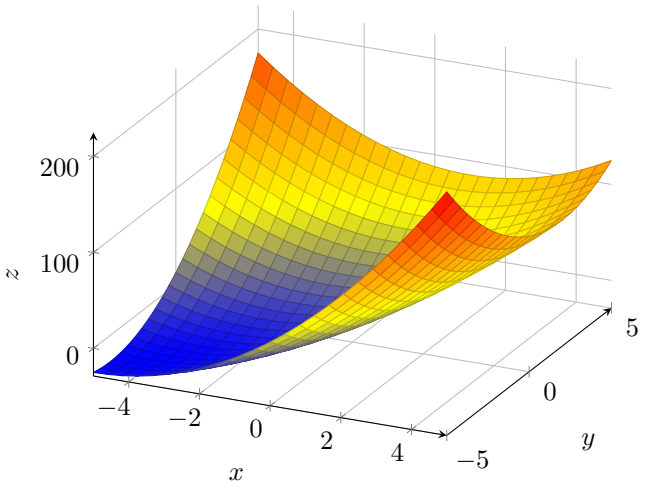
[11]See note (10).

A.6.1 Plot of the test functions

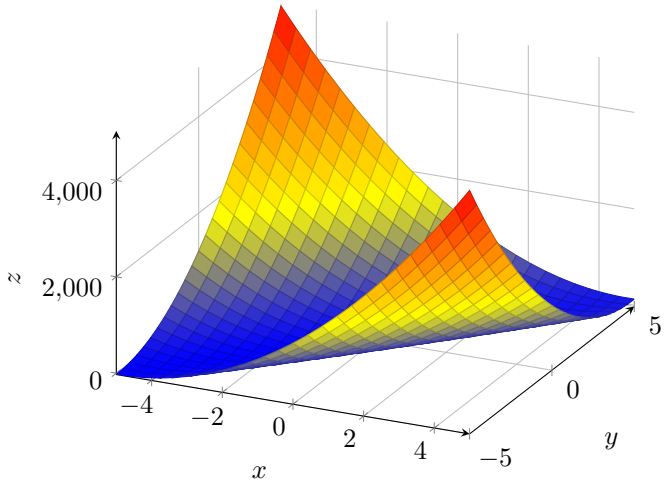
Test function $f^{(1)} = 3(x^2 + y^2) - 2xy + 10x + 5y$



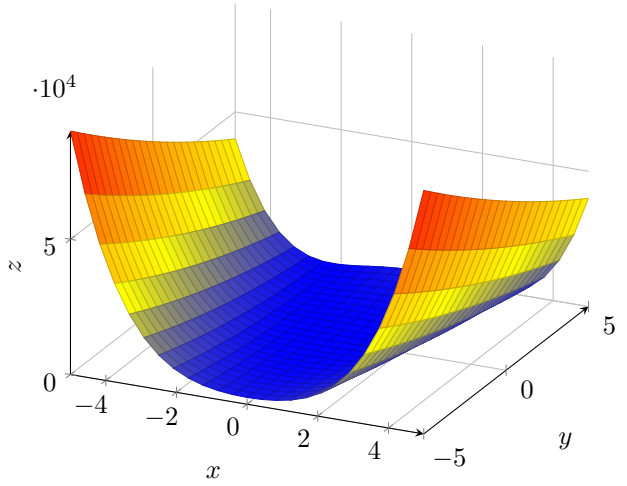
Test function $f^{(2)} = \frac{5}{2}(x^2 + y^2) - 3xy + 10x + 5y$



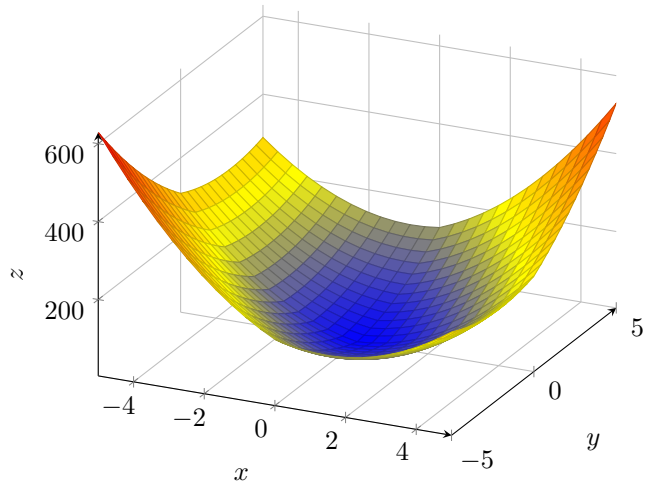
Test function $f^{(3)} = \frac{101}{2}(x^2 + y^2) - 99xy + 10x + 5y$



Test function $f^{(4)}(x, y) = 100(y - x^2)^2 + (x - 1)^2$



Test function $f^{(5)}(x, y) = |3x + 2y - 2|^2 + 10(|x| + |y|)$



A.6.2 Plot of the parameters

The following (Figure 31) shows the variations of the parameters during the time the algorithm runs. For each test function we can see the variation on $\|\text{vec}(B_{k+1}) - \text{vec}(B_k)\|$, $\|\text{vec}(B_k)\|$ and $\|N_k\|$. The behaviour of the last parameter, $\|B_k - \beta\|$ is shown in (Figure 32).

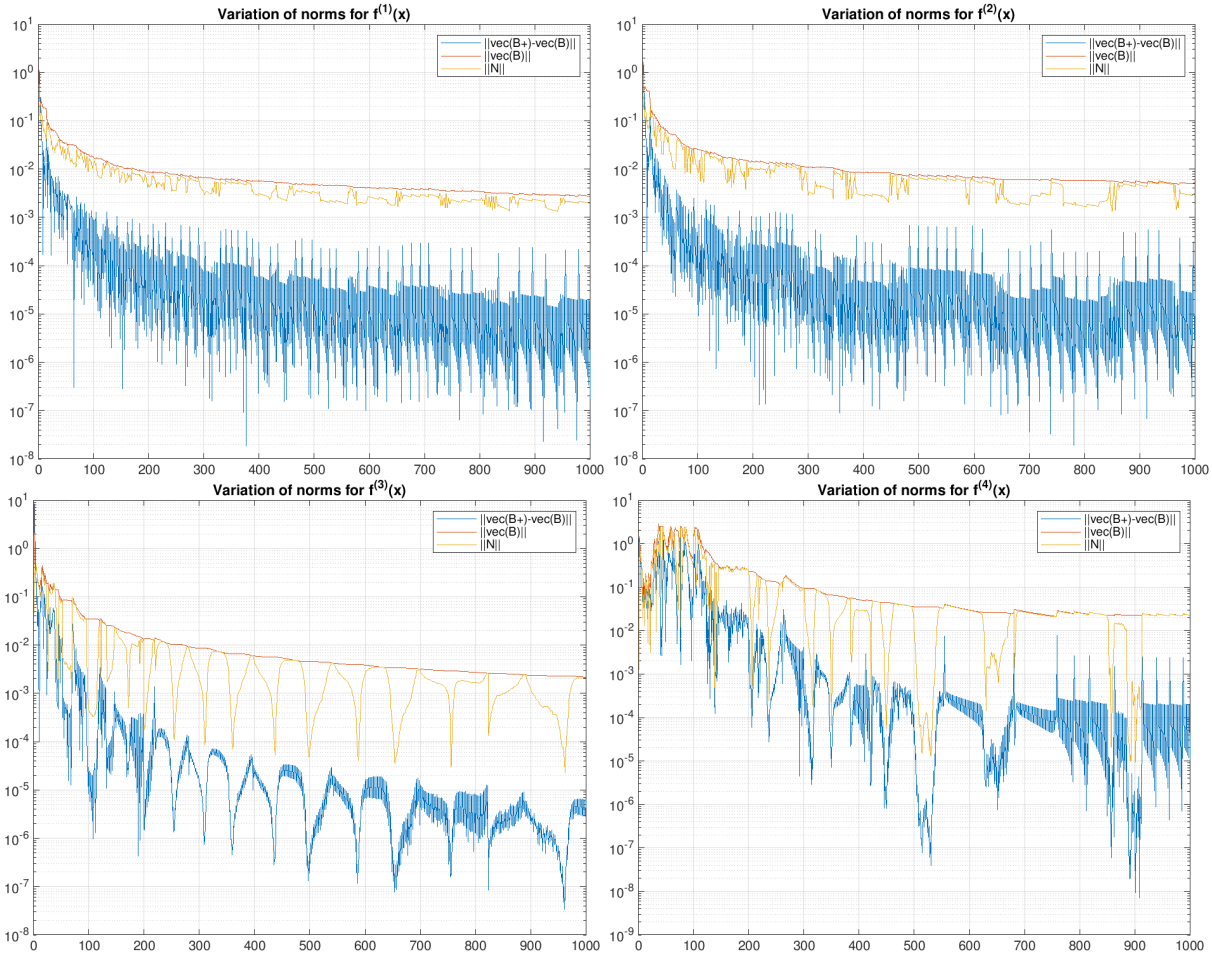


Figure 31: Variation of the parameters as function of the step k.

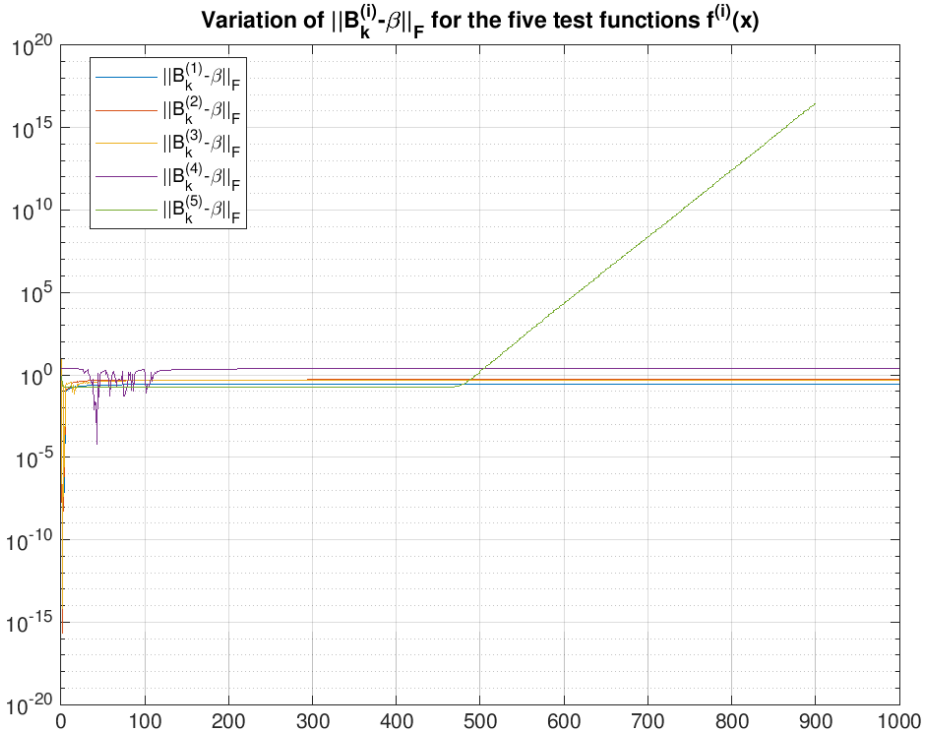


Figure 32: Variation of $\left\|B_k^{(i)} - \beta\right\|$ as function of the step k for the five test functions.

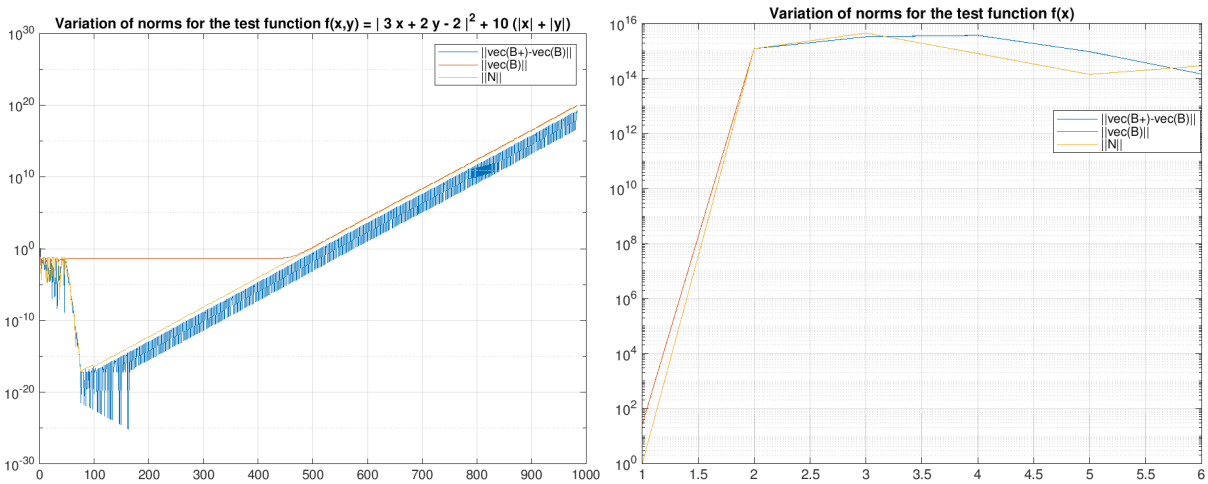


Figure 33: Behaviour of the parameter $\|\text{vec}(B_{k+1}) - \text{vec}(B_k)\|$ for the test function $f^{(5)}(x,y) = |3x + 2y - 2|^2 + 10(|x| + |y|)$ and its smooth version $\hat{f}^{(5)}(x,y) = |3x + 2y - 2|^2 + 10(x + y)$.

References

[1] Hassan K Khalil. Nonlinear systems. Prentice-Hall, 2002.

[2] Christopher M. Kellett. A compendium of comparison function results. Mathematics of Control, Signals, and Systems, 26(3):339–374, Sep 2014. ISSN 1435-568X. doi: 10.1007/s00498-014-0128-8. URL https://doi.org/10.1007/s00498-014-0128-8.

[3] Eduardo D Sontag. Input to state stability: Basic concepts and results. In Nonlinear and optimal control theory, pages 163–220. Springer, 2008.

[4] Zhong-Ping Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. Automatica, 37(6):857–869, 2001.

[5] Duc N. Tran, Christopher M. Kellett, and Peter M. Dower. Qualitative equivalences of ISS and lp-gain stability properties for discrete-time nonlinear systems. Automatica, 77(C):360–369, March 2017. ISSN 0005-1098. doi: 10.1016/j.automatica.2016.11.033. URL https://doi.org/10.1016/j.automatica.2016.11.033.

[6] David Angeli. A lyapunov approach to incremental stability properties. IEEE Transactions on Automatic Control, 47(3):410–421, 2002.

[7] Zhong-Ping Jiang, Eduardo D Sontag, and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. In Proc. 14th IFAC World Congress, pages 277–282. Citeseer, 1999.

[8] D. N. Tran, B. S. Ruffer, and C. M. Kellett. Convergence properties for discrete-time nonlinear systems. IEEE Transactions on Automatic Control, pages 1–1, 2018. ISSN 0018-9286. doi: 10.1109/TAC.2018.2879951.

[9] Florian Bayer, Mathias Bürger, and Frank Allgöwer. Discrete-time incremental ISS: A framework for robust NMPC. In 2013 European Control Conference (ECC), pages 2068–2073. IEEE, 2013.

[10] Duc N Tran, Björn S Rüffer, and Christopher M Kellett. Incremental stability properties for discrete-time systems. In 2016 IEEE 55th Conference on Decision and Control (CDC), pages 477–482. IEEE, 2016.

[11] Lars Grüne and Christopher M Kellett. ISS-lyapunov functions for discontinuous discrete-time systems. IEEE Transactions on Automatic Control, 59(11):3098–3103, 2014.

[12] Zhong-Ping Jiang and Yuan Wang. A converse lyapunov theorem for discrete-time systems with disturbances. Systems & control letters, 45 (1):49–58, 2002.

[13] David Luenberger and Yinyu Ye. Linear and nonlinear programming third edition. New York, NY: Springer, 2007. ISBN 9780387745022.

[14] Fractale de Newton kernel description. https://fr.wikipedia.org/wiki/Fractale_de_Newton, 2019. Accessed: 2019-05-01.

[15] D.P. Bertsekas. Convex Optimization Algorithms. Athena Scientific, 2015. ISBN 9781886529281. URL https://books.google.com.au/books?id=AfB5rgEACAAJ.

[16] Jorge Nocedal and Stephen Wright. Numerical optimization. Springer Science & Business Media, 2006.

[17] Carl T Kelley. Iterative methods for optimization. SIAM, 1999.

[18] Jan R. Magnus and H. Neudecker. The commutation matrix: Some properties and applications. The Annals of Statistics, 7(2):381, 1979. ISSN 00905364. URL https://ezp.lib.unimelb.edu.au/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edsjsr&AN=edsjsr.2958818&site=eds-live&scope=site.

[19] Huijuan Li, Anping Liu, and Linli Zhang. Input-to-state stability of time-varying nonlinear discrete-time systems via indefinite difference lyapunov functions. ISA transactions, 77:71–76, 2018.

[20] Eduardo D Sontag. Smooth stabilization implies coprime factorization. IEEE transactions on automatic control, 34(4):435–443, 1989.

List of Theorems

List of Definitions

List of Lemmas

List of Propositions

List of Corollaries

List of Figures

List of Tables