Inertial Newton Algorithms Avoiding Strict Saddle Points

Camille Castera

Inertial Newton Algorithms Avoiding Strict Saddle Points

Camille Castera
CNRS - IRIT, Université de Toulouse,
Toulouse, France

Abstract

We study the asymptotic behavior of second-order algorithms mixing Newton's method and inertial gradient descent in non-convex landscapes. We show that, despite the Newtonian behavior of these methods, they almost always escape strict saddle points. We also evidence the role played by the hyper-parameters of these methods in their qualitative behavior near critical points. The theoretical results are supported by numerical illustrations.

1 Introduction

With the ever-growing size of machine learning problems, the design of algorithms for solving large-scale optimization problems remains a major challenge. In particular, many of these problems amount to the unconstrained minimization of a so-called *loss functions*,

$$\min_{\theta\in\mathbb{R}^P} \mathcal{J}(\theta), \tag{1}$$

where $P \in \mathbb{N}_{>0}$ and $\mathcal{J}\colon \mathbb{R}^P \to \mathbb{R}$. Throughout this paper, we consider a loss functions $\mathcal{J}$ that is twice continuously differentiable and lower bounded on $\mathbb{R}^P$. Lots of efforts are put into building algorithms exploiting second-order derivatives in frameworks where both storage and computational cost are limited. In this context, Castera et al. (2021) recently proposed an inertial Newton algorithm (INNA) relying only on the computation of first-order derivatives, making it practical for large-scale applications such as the training of deep neural networks. One of the assets of this algorithm is that it is built upon the following ordinary differential equation (ODE), introduced by Alvarez et al. (2002),

$$\frac{\mathrm{d}^2\theta}{\mathrm{d}t^2}(t) + \alpha\frac{\mathrm{d}\theta}{\mathrm{d}t}(t) + \beta\nabla^2\mathcal{J}(\theta(t))\frac{\mathrm{d}\theta}{\mathrm{d}t}(t) + \nabla\mathcal{J}(\theta(t)) = 0, \quad \text{for all } t > 0. \tag{2}$$

In (2), $\nabla\mathcal{J}$ and $\nabla^2\mathcal{J}$ denote the gradient and the Hessian of $\mathcal{J}$ respectively, $\theta : \mathbb{R}_+ \to \mathbb{R}^P$ is a twice continuously differentiable function called *solution* of the ODE, or *trajectory*, and α and β are non-negative parameters. This ODE, called DIN for *dynamical inertial Newton-like system*, echos famous optimization algorithms. Indeed, taking $\beta = 0$, (2) models the heavy-ball with friction (HBF) method (Polyak, 1964) and is linked to the famous Nesterov's accelerated gradient (Nesterov, 1983; Su et al., 2014). Similarly, when $\alpha = 0$ the ODE represents an inertial Newton method (Attouch and Redont, 2001) so overall (2) models a mix between inertial gradient descent and Newton's algorithm. Additionally, (2) is relevant to tackle (1). Indeed, when the solutions of (2) converge (i.e., when they reach a limit point as $t \to \infty$), Alvarez

et al. (2002) proved that they converge to critical points of $\mathcal{J}$ (points where the gradient of $\mathcal{J}$ vanishes), similarly, the accumulation points (the limit of sub-sequences of iterations) of INNA with vanishing step-sizes yield critical points of $\mathcal{J}$ (Castera et al., 2021).

Yet, in many applications—for example in deep learning—the function $\mathcal{J}$ is non-convex, and may thus posses spurious critical points (critical points that are not local minima). While it has been proved that gradient descent and HBF are likely to avoid *strict* saddles (critical points where $\nabla^2\mathcal{J}$ has a negative eigenvalue, Goudou and Munier 2009; Lee et al. 2016; O'Neill and Wright 2019), vanilla Newton's method (with unit step-sizes) is however attracted by any type of critical points, not only minima (see e.g., Dauphin et al. 2014), which is problematic when solving minimization problems like (1). Since (2) mixes Newton's method and HBF, the following question remains open: *are the solutions of DIN—and the INNA algorithm—likely to avoid strict saddle points?* The main contribution of this paper is to answer positively to this question both for DIN and INNA with fixed step-sizes, regardless the choice of the hyper-parameters $\alpha > 0$ and $\beta > 0$. Additionally, we shed light on the link between the choice of α and β and the asymptotic behavior of the solutions of DIN, with in particular the emergence of spirals when $\alpha\beta < 1$, this gives a better understanding of the role played by these hyper-parameters. We also provide numerical experiments illustrating our theoretical results.

Main contributions. To summarize, our main contributions are the following:

- Prove that the solutions of DIN avoid strict saddle points for almost any initialization.
- Prove the convergence of INNA with fixed step-sizes to critical points for loss functions with Lipschitz continuous gradient.
- Provide sufficient conditions such that the limit of INNA is not a strict saddle point, for almost any initialization.
- Connect the choice of the hyper-parameters α and β and the qualitative behavior of the solutions of DIN.
- Produce numerical experiments illustrating the theoretical results.

Organization. The rest of the paper is organized as follows. We recall essential notions and optimality conditions in Section 2. Section 3 states the main results regarding DIN and introduces key theorems for proving the avoidance of strict saddles (Section 3.1). A study of the qualitative asymptotic behavior of DIN is carried out in Section 3.2. The reader only interested in convergence results for the algorithm INNA may skip Section 3 and go directly to Section 4 where we provide convergence guarantees to local minimizers for INNA. Some conclusions are finally drawn. We first review the literature related to this work.

Related work. The DIN system was first introduced by Alvarez et al. (2002) and was then studied by many, in particular Attouch et al. (2014, 2016); Shi et al. (2018) considered extensions of DIN where the hyper-parameters α and β vary over time. As we shall see, a powerful feature of DIN is that it can be written as a first-order system where the Hessian does not appear explicitly, this feature was exploited for example by Castera et al. (2021); Chen and Luo (2019); Attouch et al. (2020), these authors also used discretization techniques to build discrete optimization algorithms from DIN, among which INNA, IPAHD, or HNAG to name a few. Regarding the effect of the parameters α and β, Attouch et al. (2020) recently provided a global understanding of the link between these parameters and quantitative properties such

as asymptotic rates of convergence, in particular for convex and strongly-convex settings. In this work we rather study qualitative properties (such as the existence of spiraling solutions) and consider non-convex landscapes. This analysis is based on the Hartman-Grobman theorem (Grobman, 1959; Hartman, 1960).

Our analysis mainly relies on results from the theory of dynamical systems, and in particular on the stable manifold theorem (Pliss, 1964; Kelley, 1966). This theorem can be used to prove that optimization algorithms are likely to avoid strict saddle points, it has been used for example by Goudou and Munier (2009); Lee et al. (2016); O'Neill and Wright (2019) for gradient descent and HBF. Finally, the convergence of INNA with vanishing step-sizes was proved by Castera et al. (2021) for stochastic optimization, whereas here, we prove the convergence of INNA with fixed step-sizes for deterministic optimization.

2 Preliminary discussions and definitions

Before analyzing asymptotic behaviors of optimization methods, we recall some fundamental notions which will be important in what follows. Consider a function $g\colon \mathbb{R}^P \to \mathbb{R}$ twice continuously differentiable, a point $\theta^\star \in \mathbb{R}^P$ is a called local minimizer of g if there exists a neighborhood of $\theta^\star$ such that $g(\theta^\star)$ is the smallest value achieved by g on this neighborhood. It is a global minimizer if $g(\theta^\star)$ is the smallest value achieved by g on $\mathbb{R}^P$. Local and global maximizers are defined similarly considering the largest values of g. We now recall the following optimality conditions, see for example Nocedal and Wright (2006).

Proposition 1 (Optimality conditions)**.** *Let $g\colon \mathbb{R}^P \to \mathbb{R}$ be a twice continuously differentiable function and let $\theta^\star \in \mathbb{R}^P$. If $\theta^\star$ is a local minimizer of g, then the following holds:*

- *First-order condition: $\theta^\star$ is a critical point of g, i.e., $\nabla g(\theta^\star) = 0$.*
- *Second-order condition: The Hessian matrix $\nabla^2 g(\theta^\star)$ is positive semidefinite. Equivalently, all the eigenvalues of $\nabla^2 g(\theta^\star)$ are* non-negative.

Similarly, for any $\theta^\star \in \mathbb{R}^P$, if $\nabla g(\theta^\star) = 0$ and $\nabla^2 g(\theta^\star)$ is positive definite (or equivalently, $\nabla^2 g(\theta^\star)$ has only positive *eigenvalues), then $\theta^\star$ is a local minimizer of g.*

Similar results hold for maximizers but with negativity conditions for the Hessian matrix. The link between the eigenvalues of the Hessian matrix of the loss function $\mathcal{J}$ and the nature of its critical points plays a crucial role in the sequel. We distinguish three types of critical points:

- Those where the Hessian of $\mathcal{J}$ has only positive eigenvalues. From Proposition 1, these points are local minima.
- The points where the Hessian matrix has at least one negative eigenvalue, which are referred to as *strict saddle points*. Such a point cannot be a local minimum, it is either a maximum or not an extremum.
- The points where the Hessian matrix has only non-negative eigenvalues and at least one zero eigenvalue, we call them *non-strict* saddle points. Such points may be maximizers, minimizers, or neither of them. For example, consider the functions $(\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \frac{1}{2}\theta_1^2 + \frac{1}{2}\theta_2^2 + \theta_1\theta_2$ and $(\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \theta_1^3 + \theta_2^2$. For both functions, $(0, 0)$ is a critical point and the eigenvalues of their Hessian matrices at $(0, 0)$ are 0 and 2. Yet, one can easily check that $(0, 0)$ is a minimizer for the first function and is not an extremum for the second one. These considerations are illustrated on Figure 1.

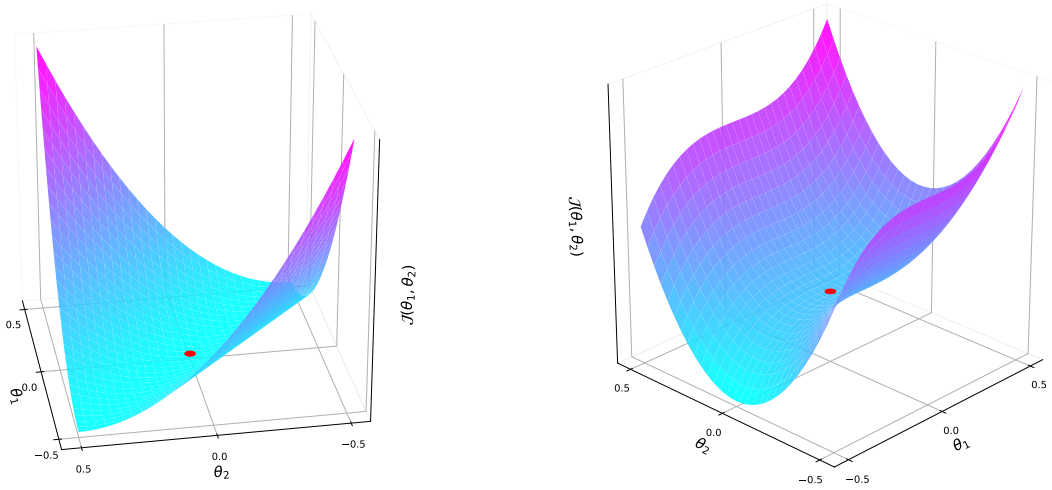


Figure 1: Example of two functions whose Hessian matrices are singular at $(0,0)$. For the function $(\theta_1,\theta_2) \in \mathbb{R}^2 \mapsto \frac{1}{2}\theta_1^2 + \frac{1}{2}\theta_2^2 + \theta_1\theta_2$ (on the left), the critical point $(0,0)$ is a minimum. For the function $(\theta_1,\theta_2) \in \mathbb{R}^2 \mapsto \theta_1^3 + \theta_2^2$ (on the right), the critical point $(0,0)$ is neither a minimum nor a maximum.

Due to the difficulties raised by the existence of non-strict saddle points, some results of this paper hold only for Morse functions, defined next.

Definition 1. *A twice continuously differentiable function $g\colon \mathbb{R}^P \to \mathbb{R}$ is a Morse function if for any $\theta^\star \in \mathbb{R}^P$ such that $\nabla g(\theta^\star) = 0$, the Hessian $\nabla^2 g(\theta^\star)$ has no zero eigenvalues.*

Morse functions are functions for which all saddles are strict and other critical points are minima. Some of the following results are restricted to Morse functions, others are more general, yet, in every case we will need the following assumption.

Assumption 1. *The loss function $\mathcal{J}$ in (1) has isolated critical points: for any $\theta^\star \in \mathbb{R}^P$ such that $\nabla\mathcal{J}(\theta^\star) = 0$, there exists a neighborhood $\Omega \subset \mathbb{R}^P$ of $\theta^\star$ such that $\theta^\star$ is the only critical point inside Ω.*

This assumption guarantees in particular that $\mathcal{J}$ has at most a countable (possibly infinite) number of critical points. Note additionally that Assumption 1 holds for Morse functions. Let us now move on to the analysis of DIN.

3 Asymptotic behavior of the solutions of DIN

We recall that $\mathcal{J} : \mathbb{R}^P \to \mathbb{R}$ is a twice continuously differentiable function and that we denote by $\nabla\mathcal{J}$ and $\nabla^2\mathcal{J}$ its gradient and its Hessian respectively. Let $\alpha \geq 0$ and $\beta > 0$, as mentioned in the introduction, a powerful property of (2) is that it is equivalent to the following first-order system (Alvarez et al., 2002),

$$\begin{cases} \frac{\mathrm{d}\theta}{\mathrm{d}t}(t) &= -\left(\alpha - \frac{1}{\beta}\right)\theta(t) - \frac{1}{\beta}\psi(t) - \beta\nabla\mathcal{J}(\theta(t)) \\ \frac{\mathrm{d}\psi}{\mathrm{d}t}(t) &= -\left(\alpha - \frac{1}{\beta}\right)\theta(t) - \frac{1}{\beta}\psi(t) \end{cases}, \quad \text{for all } t > 0, \tag{3}$$

where $(\theta, \psi)\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}^P \times \mathbb{R}^P$ is differentiable for all $t > 0$. For a given initial condition $(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P$, we say that a continuously differentiable function (θ, ψ) is a solution of (3) if $(\theta(0), \psi(0)) = (\theta_0, \psi_0)$ and (3) holds for all $t > 0$. We say that (θ, ψ) converges if there exists $(\theta^\star, \psi^\star) \in \mathbb{R}^P \times \mathbb{R}^P$ such that $(\theta(t), \psi(t))$ converges to $(\theta^\star, \psi^\star)$ as $t \to \infty$. Since $\mathcal{J}$ is twice continuously differentiable, the existence and uniqueness (with respect to initial conditions) of the solutions of (3) are granted by the Cauchy-Lipschitz theorem, see Alvarez et al. (2002). In this work, we focus on the asymptotic behavior of the solutions with respect to the initial conditions.

Necessary condition for being a stationary point. As previously said, one of the main interests of (2) and hence of (3) in optimization, is that stationary points of the solutions (where the solutions stabilize) yield critical points of $\mathcal{J}$. Indeed, a solution (θ, ψ) of (3) reaches a *stationary point* of (3) if there exists $t_0 > 0$, such that $\frac{\mathrm{d}\theta}{\mathrm{d}t}(t_0) = 0$ and $\frac{\mathrm{d}\psi}{\mathrm{d}t}(t_0) = 0$. This is equivalent to $\psi(t_0) = (1 - \alpha\beta)\theta(t_0)$ and $\nabla\mathcal{J}(\theta(t_0)) = 0$. Hence, the set of stationary points of the solutions of (3) is,

$$\mathsf{S} = \left\{(\theta^\star, \psi^\star) \in \mathbb{R}^P \times \mathbb{R}^P \middle| \nabla\mathcal{J}(\theta^\star) = 0,\ \psi^\star = (1 - \alpha\beta)\theta^\star\right\}.$$

Bounded solutions (θ, ψ) of (3) converge to points of S, hence the first coordinate θ of a bounded solution converges to a critical point of $\mathcal{J}$, see Alvarez et al. (2002); Castera et al. (2021). We will thus study the type of stationary points which the solutions of (3) are likely to converge to, and then study the qualitative asymptotic behavior of these solutions.

3.1 DIN is likely to avoid strict saddle points

We start with our main result regarding the limit (as $t \to +\infty$) of the solutions of DIN.

3.1.1 Main convergence results

For convenience, we denote by $\mathsf{S}_{<0} \subset \mathsf{S}$ the set of stationary points $(\theta^\star, \psi^\star)$ such that $\theta^\star$ is a strict saddle point of $\mathcal{J}$, namely,

$$\mathsf{S}_{<0} = \left\{(\theta^\star, \psi^\star) \in \mathsf{S} \middle| \nabla^2\mathcal{J}(\theta^\star) \text{ has at least one negative eigenvalue}\right\}.$$

Theorem 2. *Suppose that Assumption 1 holds for $\mathcal{J}$, then for almost any initialization, the corresponding solution of* (3) *does not converge to a point in $\mathsf{S}_{<0}$.*

Before proving the theorem, the following corollary is an immediate consequence suited for practical applications.

Corollary 3. *Assume that $\mathcal{J}$ is a twice continuously differentiable Morse function. Assume also that $\mathcal{J}$ is coercive (i.e., that $\lim_{\|\theta\|\to\infty} \mathcal{J}(\theta) = +\infty$). Then for any initialization the associated solution of* (3) *converges. Moreover, let (θ_0, ψ_0) be a non-degenerate random variable on $\mathbb{R}^P \times \mathbb{R}^P$, and let (θ, ψ) be the solution of* (3) *initialized at (θ_0, ψ_0) and converging to $(\theta^\star, \psi^\star) \in \mathbb{R}^P \times \mathbb{R}^P$. Then with probability one with respect to the draw of (θ_0, ψ_0), $\theta^\star$ is a local minimizer of $\mathcal{J}$.*

This corollary states in particular that for a coercive Morse function, we can pick an initialization sampled from a non-degenerate distribution on $\mathbb{R}^P \times \mathbb{R}^P$, for example a Gaussian or uniform distribution, and with probability one, the first coordinate of the limit of the solution (with respect to the initialization) is a local minimizer of $\mathcal{J}$.

Proof of Corollary 3. From Castera et al. (2021, Section 3.2), the coercivity of $\mathcal{J}$ guarantees that any solution of (3) remains bounded, and thus from Alvarez et al. (2002), any solution is converging. Then, the limit of any solution belongs to S. Let (θ_0, ψ_0) be a random variable sampled from a non-degenerate distribution on $\mathbb{R}^P \times \mathbb{R}^P$, by definition the support of the distribution has non-zero measure. In addition, according to Theorem 2, the set of initializations such that the solutions of (3) converge to $\mathsf{S}_{<0}$ has zero measure. So, almost surely with respect to the random variable (θ_0, ψ_0), the solution of (3) initialized at (θ_0, ψ_0) converges toward $\mathsf{S} \backslash \mathsf{S}_{<0}$. Finally, since $\mathcal{J}$ is a Morse function, $\mathsf{S} \setminus \mathsf{S}_{<0}$ is exactly the set of local minimizers. □

Remark 1. *We could state a more general (but more abstruse) result than Corollary 3 which would not require the coercivity assumption but only that the set of initializations such that the associated solutions of* (3) *converge has positive Lebesgue measure.*

We now introduce the main tool to prove Theorem 2: the stable manifold theorem.

3.1.2 The stable manifold theorem

To simplify the notations we introduce the following mapping,

$$G : (\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P \mapsto \begin{pmatrix} -\left(\alpha - \frac{1}{\beta}\right)\theta - \frac{1}{\beta}\psi - \beta \nabla \mathcal{J}(\theta) \\ -\left(\alpha - \frac{1}{\beta}\right)\theta - \frac{1}{\beta}\psi \end{pmatrix},$$

so that (3) can be re-written,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\theta(t)\\ \psi(t)\end{pmatrix} = G(\theta(t), \psi(t)), \quad \text{for all } t > 0. \tag{4}$$

For any $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$, we also denote by $DG(\theta, \psi) \in \mathbb{R}^{2P \times 2P}$ the Jacobian matrix of G at (θ, ψ). Remark that for any $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$, $(\theta, \psi) \in \mathsf{S} \iff G(\theta, \psi) = 0$, so the stationary points of (3) are exactly the zeros of G. We now state the stable manifold theorem which is the keystone to prove Theorem 2.

Theorem 4 (Stable manifold theorem (Haragus and Iooss, 2010; Perko, 2013))**.** *Let $F \colon \mathbb{R}^{2P} \to \mathbb{R}^{2P}$ be a C^1 mapping and denote by DF the Jacobian of F, consider the autonomous ODE,*

$$\frac{\mathrm{d}\Theta}{\mathrm{d}t}(t) = F(\Theta(t)), \quad \textit{for all } t > 0. \tag{5}$$

Let $\Theta^\star \in \mathbb{R}^{2P}$ such that $F(\Theta^\star) = 0$. Let $E^{sc}(\Theta^\star)$ be the linear subspace of $\mathbb{R}^{2P}$ spanned by the eigenvalues of $DF(\Theta^\star)$ with non-positive real part. There exists a neighborhood Ω of $\Theta^\star$ and a C^1 manifold $\mathsf{W}^{sc}(\Theta^\star)$ tangent to $E^{sc}(\Theta^\star)$ at $\Theta^\star$—whose dimension is the number of eigenvalues of $DF(\Theta^\star)$ with non-positive real part—such that, for any solution Θ of (5)*,*

(i) If $\Theta(0) \in \mathsf{W}^{sc}(\Theta^\star) \cap \Omega$ and for $T \geq 0$, $\Theta([0, T]) \subset \Omega$, then $\Theta([0, T]) \subset \mathsf{W}^{sc}(\Theta^\star)$. (Invariance)

(ii) If $\forall t \geq 0,\ \Theta(t) \in \Omega$, then $\Theta(0) \in \mathsf{W}^{sc}(\Theta^\star)$.

We see why this theorem plays an important role in the proof of Theorem 2. It states in particular that all the solutions of (5) converging to $\Theta^\star$ must enter inside $\mathsf{W}^{sc}(\Theta^\star)$ after some time, and that $\mathsf{W}^{sc}(\Theta^\star)$ has zero measure as soon as $DF(\Theta^\star)$ has at least one positive eigenvalue. We will show that this holds true for G and for any point in $\mathsf{S}_{<0}$. Now that we introduced our main tool, we can prove Theorem 2.

3.1.3 Proof of Theorem 2

We state an elementary lemma that will be useful.

Lemma 5. *Let $\alpha \geq 0$, $\beta > 0$ and $\lambda \in \mathbb{R}$. The quantity $(\alpha + \beta\lambda)^2 - 4\lambda$ is non-positive if and only if $\alpha\beta \leq 1$ and $\lambda \in \left[\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right]$.*

Proof of Lemma 5. Let $\alpha \geq 0$, and $\beta > 0$, the function $\lambda \in \mathbb{R} \mapsto (\alpha+\beta\lambda)^2-4\lambda = \beta^2\lambda^2+2(\alpha\beta-2)\lambda + \alpha^2$ is a second-order polynomial in λ whose discriminant is $16(1-\alpha\beta)$. If $\alpha\beta > 1$ this discriminant is negative thus the polynomial has no real roots and hence is always positive. If $\alpha\beta \leq 1$, then the discriminant is positive and the roots of the polynomial are $\frac{(2-\alpha\beta)}{\beta^2} \pm \frac{2\sqrt{1-\alpha\beta}}{\beta^2}$. □

We now prove Theorem 2.

Proof of Theorem 2. Let $(\theta^\star, \psi^\star) \in \mathbb{R}^P \times \mathbb{R}^P$ such that $G(\theta^\star, \psi^\star) = 0$. We first compute the eigenvalues of $DG(\theta^\star, \psi^\star)$, in order to apply Theorem 4 to (4) around $(\theta^\star, \psi^\star)$. By differentiating G we obtain the following Jacobian matrix, displayed in four blocks,

$$DG(\theta^\star, \psi^\star) = \begin{pmatrix} -\beta\nabla^2\mathcal{J}(\theta^\star) - \left(\alpha - \frac{1}{\beta}\right) I_P & -\frac{1}{\beta} I_P \\ -\left(\alpha - \frac{1}{\beta}\right) I_P & -\frac{1}{\beta} I_P \end{pmatrix},$$

where I_P denotes the identity matrix of $\mathbb{R}^{P\times P}$. We need to compute the eigenvalues of $DG(\theta^\star, \psi^\star)$ and study the sign of their real parts. In particular, we want to show that if $\nabla^2\mathcal{J}(\theta)$ has a (strictly) negative eigenvalue (i.e., $(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}$), then $DG(\theta^\star, \psi^\star)$ has at least one (strictly) positive eigenvalue, and thus according to Theorem 4, the stable manifold associated to $(\theta^\star, \psi^\star)$ has zero measure.

First, $\nabla^2\mathcal{J}(\theta^\star)$ is real and symmetric, so the spectral theorem states that there exists an orthogonal matrix V such that $V^T\nabla^2\mathcal{J}(\theta^\star)V$ is a diagonal matrix. Thus the matrix,

$$\begin{pmatrix} V^T & 0 \\ 0 & V^T \end{pmatrix} DG(\theta^\star, \psi^\star) \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} -\beta V^T\nabla^2\mathcal{J}(\theta^\star)V - \left(\alpha - \frac{1}{\beta}\right) I_P & -\frac{1}{\beta} I_P \\ -\left(\alpha - \frac{1}{\beta}\right) I_P & -\frac{1}{\beta} I_P \end{pmatrix} \tag{6}$$

is a sparse matrix with only 3 non-zero diagonals and whose eigenvalues are the same as those of $DG(\theta^\star, \psi^\star)$. Exploiting the tridiagonal structure, there exists a symmetric permutation $U \in \mathbb{R}^{2P\times 2P}$—specified in (19) in Section A of the appendix—such that we can transform (6) into a block diagonal matrix,

$$U^T \begin{pmatrix} V^T & 0 \\ 0 & V^T \end{pmatrix} DG(\theta^\star, \psi^\star) \begin{pmatrix} V & 0 \\ 0 & V \end{pmatrix} U = \begin{pmatrix} M_1 & & \\ & \ddots & \\ & & M_P \end{pmatrix}, \tag{7}$$

where for each $p \in \{1, \dots, P\}$, M_p is a 2×2 matrix defined as follows. Denote by $(\lambda_p)_{p\in\{1,\dots,P\}}$ the eigenvalues of $\nabla^2\mathcal{J}(\theta^\star)$, for all $p \in \{1, \dots, P\}$, $M_p = \begin{pmatrix} -\left(\alpha - \frac{1}{\beta}\right) - \beta\lambda_p & -\frac{1}{\beta} \\ -\left(\alpha - \frac{1}{\beta}\right) & -\frac{1}{\beta} \end{pmatrix}$, up to a symmetric permutation.

The eigenvalues of $DG(\theta^\star, \psi^\star)$ are obtained by computing those of the matrices M_p. Let $p \in \{1, \dots, P\}$, the eigenvalues of M_p are the roots of its characteristic polynomial: $\chi_{M_p} : X \in \mathbb{R} \mapsto X^2 - \mathrm{trace}(M_p)X + \det(M_p)$, which gives for any $X \in \mathbb{R}$,

$$\chi_{M_p}(X) = X^2 + (\alpha + \beta\lambda_p)X + \lambda_p.$$

This is a second-order polynomial, whose discriminant is,

$$\Delta_{M_p} = (\alpha + \beta\lambda_p)^2 - 4\lambda_p. \tag{8}$$

The eigenvalues of M_p depend on the sign of Δ_{M_p} which is given by Lemma 5. We now show that if $\lambda_p < 0$, then M_p has a positive eigenvalue.

First assume that $\Delta_{M_p} \leq 0$. Lemma 5 states that in this case, we have $\alpha\beta \leq 1$ and $\lambda_p \in \left[\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right]$. An elementary study of the function $x \in [0,1] \mapsto 2 - x - 2\sqrt{1-x}$ shows however that for $0 \leq \alpha\beta \leq 1$, $\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2} \geq 0$, so $\Delta_{M_p} \leq 0$ implies $\lambda_p \geq 0$. Thus, we do not need to further investigate the case $\Delta_{M_p} \leq 0$ since this case never occurs when $\lambda_p < 0$.

Suppose now that $\Delta_{M_p} > 0$, then M_p has two real eigenvalues,

$$\begin{cases} \sigma_{p,+} = \frac{-(\alpha+\beta\lambda_p)}{2} + \frac{\sqrt{\Delta_{M_p}}}{2} = \frac{-(\alpha+\beta\lambda_p)}{2} + \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2} \\ \sigma_{p,-} = \frac{-(\alpha+\beta\lambda_p)}{2} - \frac{\sqrt{\Delta_{M_p}}}{2} = \frac{-(\alpha+\beta\lambda_p)}{2} - \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2} \end{cases}.$$

In this case, assume that $\lambda_p < 0$. If $\alpha + \beta\lambda_p \leq 0$, then $\sigma_{p,+}$ is a sum of a non-negative and a positive term, so $\sigma_{p,+} > 0$. If $\alpha+\beta\lambda_p \geq 0$, then $2\sigma_{p,+} = -(\alpha+\beta\lambda_p)+\sqrt{(\alpha+\beta\lambda_p)^2 + 4(-\lambda_p)} > 0$ since $4(-\lambda_p) > 0$. Overall, we showed that in every case, $\lambda_p < 0 \implies \sigma_{p,+} > 0$. So whenever there exists $p \in \{1, \dots, P\}$ such that $\lambda_p < 0$, $DG(\theta^\star, \psi^\star)$ has at least one positive eigenvalue.

We can now apply the stable manifold theorem. Let $(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}$. Let an initialization (θ_0, ψ_0) such that the corresponding solution (θ, ψ) of (4) converges to $(\theta^\star, \psi^\star)$. Denote $\Phi : \mathbb{R}^P \times \mathbb{R}^P \times \mathbb{R} \to \mathbb{R}^P \times \mathbb{R}^P \times \mathbb{R}$ the flow of the solutions of (4), so that we have in particular for all $t \geq 0$, $(\theta(t), \psi(t)) = \Phi((\theta_0, \psi_0), t)$ and $(\theta_0, \psi_0) = \Phi((\theta(t), \psi(t), -t)$. Consider the manifold $\mathsf{W}^{sc}(\theta^\star, \psi^\star)$ and the neighborhood Ω as defined in Theorem 4. The convergence of (θ, ψ) implies that there exists $t_0 \geq 0$ such that for all $t \geq t_0$, $(\theta(t), \psi(t)) \in \Omega$, so according to Theorem 4, $\forall t \geq t_0$, $(\theta(t), \psi(t)) \in \Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$. Expressing this in terms of flows, $\forall t \geq t_0$, $\Phi((\theta_0, \psi_0), t) \in \Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$ and hence $\forall t \geq t_0$,

$$\Phi((\theta_0, \psi_0), t) \in \bigcup_{k \in \mathbb{N}} \Phi\left(\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star), -k\right), \tag{9}$$

where the right-hand side in (9) corresponds to the union over $k \in \mathbb{N}$ of initial conditions such that the associated solution has reached $\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$ at time k. Let

$$\mathsf{W}(\theta^\star, \psi^\star) = \left\{ (\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P \middle| \Phi((\theta_0, \psi_0), t) \xrightarrow[t \to +\infty]{} (\theta^\star, \psi^\star) \right\},$$

the set of all initial conditions such that the associated solution converges to $(\theta^\star, \psi^\star)$. According to (9), we have proved that,

$$\mathsf{W}(\theta^\star, \psi^\star) \subset \bigcup_{k \in \mathbb{N}} \Phi\left(\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star), -k\right). \tag{10}$$

Now, we previously showed that since $(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}$, $DG(\theta^\star, \psi^\star)$ has one or more positive eigenvalues, so according to the stable manifold theorem, the dimension of $\mathsf{W}^{sc}(\theta^\star, \psi^\star)$ is strictly less than $2P$, hence this manifold has zero measure. Due to the uniqueness of the solutions of (4), for any $k \in \mathbb{N}$, $\Phi(\cdot, -k)$ is a local diffeomorphism, hence it maps zero-measure sets to zero-measure sets. Consequently, the right-hand side in (10) is a countable union of zero-measure sets, so it has zero measure as well, and the same goes for $\mathsf{W}(\theta^\star, \psi^\star)$.

To conclude the proof of the theorem, by Assumption 1, the critical points are isolated so their number is countable. So $\bigcup_{(\theta^\star,\psi^\star)\in\mathsf{S}_{<0}} \mathsf{W}(\theta^\star,\psi^\star)$ is a countable union of zero-measure sets so it has zero measure. □

Remark 2. *Keeping the notations of the proof of Theorem 2, we proved that for $p \in \{1,\dots,P\}$, if $\lambda_p < 0$ then $\sigma_{p,+} > 0$. Looking at the proof, note that we could also show quite easily that $\lambda_p > 0 \implies \sigma_{p,+} < 0$, so for any local minimizer with non-singular Hessian, the associated stable manifold does not have zero measure. In particular, the stable manifold associated to any local minima of a twice differentiable Morse functions does not have zero measure.*

3.1.4 On the complex eigenvalues of DG

In the proof of Theorem 2 we showed that for any $(\theta^\star,\psi^\star) \in \mathsf{S}$, if an eigenvalue λ_p of the Hessian $\nabla\mathcal{J}(\theta^\star)$ is negative, then the associated discriminant Δ_{M_p} is positive and thus the eigenvalues of $DG(\theta^\star,\psi^\star)$ are real. Note however that when there exists $p \in \{1,\dots,P\}$ such that $\lambda_p \geq 0$ (and hence in particular around local minima), we may have $\Delta_{M_p} \leq 0$. This is the case whenever $\alpha\beta \leq 1$, and $\lambda_p \in \left[\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right]$. When the aforementioned conditions hold, $DG(\theta^\star,\psi^\star)$ has complex eigenvalues,

$$\begin{cases} \sigma_{p,+} = \frac{-(\alpha+\beta\lambda_p)}{2} + i\frac{\sqrt{-\Delta_{M_p}}}{2} = \frac{-(\alpha+\beta\lambda_p)}{2} + i\frac{\sqrt{4\lambda_p-(\alpha+\beta\lambda_p)^2}}{2} \\ \sigma_{p,-} = \frac{-(\alpha+\beta\lambda_p)}{2} - i\frac{\sqrt{-\Delta_{M_p}}}{2} = \frac{-(\alpha+\beta\lambda_p)}{2} - i\frac{\sqrt{4\lambda_p-(\alpha+\beta\lambda_p)^2}}{2} \end{cases} . \tag{11}$$

Overall, we proved that DIN is likely to avoid strict saddle points, and we additionally observed that around local minima, the Jacobian DG may or may not have eigenvalues with non-zero imaginary part. The existence of complex eigenvalues may change the qualitative behavior of the solutions around local minima. Next section is devoted to studying this matter.

3.2 Behavior of the solutions of DIN around stationary points

Accordingly to what we just discussed in Section 3.1.4, we wish to characterize the qualitative asymptotic behavior of the convergent trajectories of DIN.

3.2.1 The Hartman-Grobman theorem

To this aim, we introduce the Hartman-Grobman theorem.

Theorem 6 (Hartman–Grobman (Perko, 2013))**.** *Consider the following dynamical system,*

$$\frac{\mathrm{d}\Theta}{\mathrm{d}t}(t) = F(\Theta(t)), \quad t \in \mathbb{R} \tag{12}$$

where $\Theta : \mathbb{R} \to \mathbb{R}^{2P}$, $F : \mathbb{R}^{2P} \to \mathbb{R}^{2P}$ is C^1 and denote by DF the Jacobian matrix of F. Assume that there exists $\Theta^\star \in \mathbb{R}^{2P}$ such that $F(\Theta^\star) = 0$ and $DF(\Theta^\star)$ has only non-zero eigenvalues. Then, there exists a neighborhood Ω of $\Theta^\star$ and a homeomorphism H (a bijective continuous function whose inverse is continuous) such that, for any $\Theta_0 \in \Omega$, if Θ is a solution of (12) with $\Theta(0) = \Theta_0$, there exists an open interval of time $\mathsf{T} \subset \mathbb{R}$ containing 0 such that the function $\Phi = H \circ \Theta$ is the solution of

$$\frac{\mathrm{d}\Phi}{\mathrm{d}t}(t) = DF(\Theta^\star)\Phi(t), \quad t \in \mathsf{T}, \tag{13}$$

with initial condition $\Phi(0) = H(\Theta_0)$. The homeomorphism H preserves the parameterization by time (it does not reverse time).

This theorem essentially states that, in a neighborhood of a stationary point $\Theta^\star$ where the Jacobian matrix $DF(\Theta^\star)$ is non-singular the solutions of (12) have a qualitative behavior similar to those of the linearized system (13).

3.2.2 Application to DIN

Application of the theorem. Let $(\theta^\star, \psi^\star) \in \mathsf{S}$ be a stationary point of (3) such that $\theta^\star$ is a local minimizer of $\mathcal{J}$ and such that $\nabla^2\mathcal{J}(\theta^\star)$ has only non-zero eigenvalues (this is guaranteed for all local minima if $\mathcal{J}$ is a Morse function). Consider the differential equation,

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}\widetilde{\theta}(t)\\ \widetilde{\psi}(t)\end{pmatrix} = DG(\theta^\star, \psi^\star)\begin{pmatrix}\widetilde{\theta}(t)\\ \widetilde{\psi}(t)\end{pmatrix}, \quad t \in \mathbb{R}. \tag{14}$$

According to Remark 2 and the proof of Theorem 2, all the eigenvalues of $DG(\theta^\star, \psi^\star)$ have strictly negative real parts. So there exists a homeomorphism H and a neighborhood Ω of $(\theta^\star, \psi^\star)$ on which Theorem 6 holds. In particular, for any initial condition $(\theta_0, \psi_0) \in \Omega$, the associated solution of (4) converges to $(\theta^\star, \psi^\star)$ (because the corresponding solution of (14) converges to $(\theta^\star, \psi^\star)$ and H preserves the parameterization by time).

So for any initialization in $(\theta_0, \psi_0) \in \Omega$, the corresponding solution (θ, ψ) of (4) remains in Ω, i.e., $\forall t \geq 0, (\theta(t), \psi(t)) \in \Omega$. Thus we can use the Hartman-Grobman around $(\theta(t), \psi(t))$ for any $t \geq 0$. As a result, for any initialization in $(\theta_0, \psi_0) \in \Omega$, and for all $t \geq 0$, the associated solution (θ, ψ) of (4) reads, $(\theta(t), \psi(t)) = H^{-1}(\widetilde{\theta}(t), \widetilde{\psi}(t))$ where $(\widetilde{\theta}, \widetilde{\psi})$ is the solution of (14) with initial condition $(\widetilde{\theta}(t_0), \widetilde{\psi}(t_0)) = H(\theta(0), \psi(0))$.

We give a more precise expression for this solution. As done in the proof of Theorem 2 we can diagonalize $DG(\theta^\star, \psi^\star)$: there exists a matrix $Q \in \mathbb{R}^{2P\times 2P}$ such that $DG(\theta^\star, \psi^\star) = QDQ^{-1}$, where $D = \mathrm{diag}(\sigma_1, \ldots, \sigma_{2P})$, and $(\sigma_p)_{\{1,\ldots,2P\}}$ are the eigenvalues of $DG(\theta^\star, \psi^\star)$. Using the diagonalization, the solution of (14) is given for all $t \in \mathbb{R}$ by $\begin{pmatrix}\widetilde{\theta}(t)\\ \widetilde{\psi}(t)\end{pmatrix} = Qe^{tD}Q^{-1}\begin{pmatrix}\widetilde{\theta}(0)\\ \widetilde{\psi}(0)\end{pmatrix}$. So going back to (θ, ψ), we have,

$$\big(\theta(t), \psi(t)\big) = H^{-1}\left(Qe^{tD}Q^{-1}H(\theta(0), \psi(0))\right), \quad \text{for all } t \geq 0. \tag{15}$$

Finally, let any initialization $(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P$ (not necessarily belonging to Ω), such that the corresponding solution (θ, ψ) of (4), converges to $(\theta^\star, \psi^\star)$. Then there exists $t_0 \geq 0$ such that for all $t \geq t_0$, $(\theta(t), \psi(t)) \in \Omega$, and the arguments above apply after t_0.

Form of the solutions. We proved that after some time, a solution (θ, ψ) of (4) that converges to $(\theta^\star, \psi^\star) \in \mathsf{S}$ can be expressed with formula (15) (up to a time shift). If $\alpha\beta \leq 1$ and all the eigenvalues of $\nabla^2\mathcal{J}(\theta^\star)$ are not in $\left(\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right)$, or if $\alpha\beta > 1$, then all the eigenvalues of $DG(\theta^\star, \psi^\star)$ are real so the coordinates of $Q^{-1}e^{tD}Q$ in (15) are sums of exponential functions decreasing in time.

However, if $\alpha\beta \leq 1$ and there exists eigenvalues of $\nabla^2\mathcal{J}(\theta^\star)$ belonging to the interval mentioned above, then there exists eigenvalues of $DG(\theta^\star, \psi^\star)$ with non-zero imaginary part. Let $p \in \{1, \ldots, P\}$ such that λ_p is an eigenvalue of $\nabla^2\mathcal{J}(\theta^\star)$ belonging to the aforementioned interval. From (11), there exists two complex eigenvalues: $\frac{-(\alpha+\beta\lambda_p)}{2} \pm i\frac{\sqrt{4\lambda_p - (\alpha+\beta\lambda_p)^2}}{2}$, and thus the coordinates of the matrix e^{tD} in (15) contain terms of the form,

$$e^{\frac{-(\alpha+\beta\lambda_p)}{2}t}\left(\cos(\omega_p t) \pm i\sin(\omega_p t)\right),$$

where $\omega_p = \frac{\sqrt{4\lambda_p - (\alpha+\beta\lambda_p)^2}}{2}$. So in this setting, and in this setting only, the imaginary parts of the eigenvalues of $DG(\theta^\star, \psi^\star)$ generate oscillating terms and the solution of the linearized model $t \mapsto Qe^{tD}Q^{-1}$ (plus initial condition) spirals around $(\theta^\star, \psi^\star)$ as it converges toward it. This is illustrated on Figure 2 where such a behavior is indeed observed for a (numerically approximated) solution of (3). Note finally that ω_p is a decreasing function of β, hence increasing the parameter β reduces the oscillations.

3.2.3 Numerical illustration of the spiraling phenomenon

Setting. To illustrate the spiraling phenomenon, we consider a simple quadratic function $\mathcal{J}\colon (\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \theta_1^2 + 2\theta_2^2$. This loss function is $\mathcal{C}^2(\mathbb{R}^2)$, and for all $(\theta_1, \theta_2) \in \mathbb{R}^2$, it has a constant diagonal Hessian: $\nabla^2\mathcal{J}(\theta_1, \theta_2) = \begin{pmatrix} 2 & 0 \\ 0 & 4 \end{pmatrix}$. This is a convex function whose unique global minimizer is $(\theta^\star, \psi^\star) = (0, 0)$. Instead of solving exactly (3), we find an approximate solution via the discrete algorithm INNA derived from (3) and presented in next section. To do so, we ran the algorithm with very small step-sizes. The algorithm is initialized at $(1, 1)$. We consider two choices of parameters: $(\alpha, \beta) = (2, 0.1)$ and $(\alpha, \beta) = (2, 1)$. The former illustrates the case $\alpha\beta < 1$ while the second corresponds to the case where $\alpha\beta > 1$. According to Section 3.1.4, with the configuration $(\alpha, \beta) = (2, 0.1)$, the range of eigenvalues for which we should observe spirals is approximately $[1, 359]$ so both eigenvalues of the Hessian of $\mathcal{J}$ lie in this interval.

Results. The expected behavior (discussed on Section 3.2.2) can be observed on the left of Figure 2. When $\alpha\beta < 1$ (red curve), the trajectory spirals around the critical point $(0, 0)$. On the contrary, the phenomenon does not occur when $\alpha\beta > 1$ (orange curve). Remark also that when zooming very close to $(0, 0)$, the oscillating behavior is still present. Note however that this qualitative results says nothing about the speed of convergence, as evidenced on the right of Figure 2. Despite the presence of spirals, the setting where $\alpha\beta < 1$ yields a faster algorithm both in terms of loss function values and distance to the minimizer. From a theoretical point of view, the Hartman-Grobman theorem connects the solutions of (3) and those of its linearized approximation through a mapping which is homeomorphic (hence continuous) but not necessarily differentiable. As a consequence, the theorem does not guarantee that the speed of convergence is preserved. Regarding the speed of convergence, we refer to Attouch et al. (2020, 2021) in the convex setting and to Castera et al. (2021) for the non-convex case.

Vanishing viscous damping. To finish this section, we empirically investigate the oscillating phenomenon when using an asymptotically vanishing damping. More precisely, we consider a viscous damping $\alpha(t)$ that may vary over time, and in particular that progressively decreases to zero as $t \to \infty$. Such damping has been given a lot of attention after the work of Su et al. (2014) who made a connection between Nesterov's accelerated gradient (Nesterov, 1983) and a differential equation with a damping proportional to $1/t$. As for DIN, if such a damping is used while keeping β fixed, we eventually have $\alpha(t)\beta \leq 1$ after some time. Our approach is however purely empirical since the Hartman-Grobman theorem holds only for autonomous ODEs[1] (hence with α not decreasing with t). In this setting we do observe spirals (see Figure 2), actually, we see on the blue curve that for $(\alpha(t), \beta) = (2/t, 0.1)$, the spirals are so large that the algorithm is much slower than it was for fixed values of (α, β). However when taking a

[1]The theorem can be extended to some non-autonomous ODEs (Palmer, 1973), which we do not consider for the sake of simplicity.

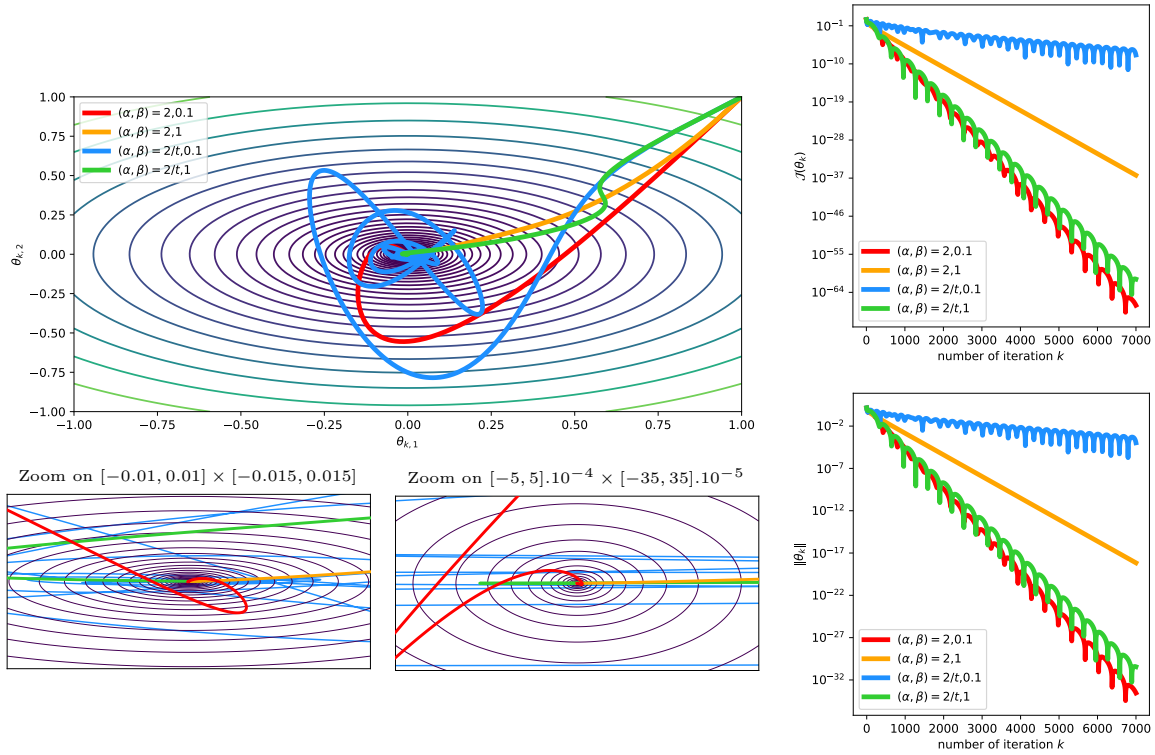


Figure 2: Illustration of the spiral phenomenon discussed in Section 3.2 on the quadratic function $\mathcal{J}\colon (\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \theta_1^2 + 2\theta_2^2$. Top-left figure displays the evolution of the iterates on the landscape of the loss function with two zooms on bottom-left figures. Right figures show the value of the loss function and the distance to the global minimizer $(0, 0)$ as a function of the iterations.

larger β (green curve), these oscillations are damped (although still noticeable), yielding better performances in terms of speed.

4 Asymptotic behavior of INNA

We now turn our attention to the asymptotic behavior of the algorithm INNA introduced by Castera et al. (2021). Throughout this section we fix $\alpha \geq 0$ and $\beta > 0$ the hyper-parameters of INNA. The INNA algorithm is originally designed for non-smooth and stochastic applications, however, in order to study its asymptotic behavior, we consider a simpler framework. We analyze the algorithm for a loss function $\mathcal{J}$ that is twice continuously differentiable and consider a deterministic version of the algorithm. In this framework, we may use fixed step-sizes: let $\gamma > 0$ be a step-size, in this setting, the INNA algorithm reads,

$$\begin{cases} \theta_{k+1} &= \theta_k + \gamma \left[-(\alpha - \frac{1}{\beta})\theta_k - \frac{1}{\beta}\psi_k - \beta \nabla \mathcal{J}(\theta_k) \right] \\ \psi_{k+1} &= \psi_k + \gamma \left[-(\alpha - \frac{1}{\beta})\theta_k - \frac{1}{\beta}\psi_k \right] \end{cases}. \tag{16}$$

This algorithm is obtained via an explicit Euler discretization of (3), thus, we may hope the asymptotic behavior of INNA to be similar to that of the solutions of DIN. Actually, the set of stationary points of INNA is the same as that of DIN (see Section B.1 of the appendix):

$$\mathsf{S} = \left\{ (\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P \middle| \nabla \mathcal{J}(\theta) = 0,\ \psi = (1 - \alpha\beta)\theta \right\}.$$

In this section, we prove that for almost any initialization, INNA does not converge to strict saddle points. To this aim we recall the definition of the set $\mathsf{S}_{<0}$,

$$\mathsf{S}_{<0} = \left\{(\theta, \psi) \in \mathsf{S} \,\middle|\, \nabla^2 \mathcal{J}(\theta) \text{ has at least one negative eigenvalue}\right\}.$$

4.1 INNA generically avoids strict saddles

In order to derive results for INNA similar to those for DIN, we will have to carefully choose the step-size $\gamma > 0$. To this aim, we need the following assumption.

Assumption 2. *There exists $L_{\nabla\mathcal{J}} > 0$ such that the gradient $\nabla\mathcal{J}$ of the loss function $\mathcal{J}$ is $L_{\nabla\mathcal{J}}$-Lipschitz continuous on $\mathbb{R}^P$ (with respect to a given norm $\|\cdot\|$ on $\mathbb{R}^P$). In other words, for any $\theta_1 \in \mathbb{R}^P$ and $\theta_2 \in \mathbb{R}^P$,*

$$\|\nabla\mathcal{J}(\theta_1) - \nabla\mathcal{J}(\theta_2)\| \leq L_{\nabla\mathcal{J}}\|\theta_1 - \theta_2\|$$

This assumption implies that at any point of $\mathbb{R}^P$, the eigenvalues of $\nabla^2\mathcal{J}$ are bounded by the constant $L_{\nabla\mathcal{J}}$. Under this assumption our main result regarding INNA follows.

Theorem 7. *Under Assumption 1 and 2, if $\alpha > 0$ and the step-size γ is such that,*

$$0 < \gamma < \min\left(\frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}, \beta\right), \tag{17}$$

—where the right-hand side in (17) is always positive—then for almost any initialization, INNA does not converge to a point in $\mathsf{S}_{<0}$.

The proof relies on arguments similar to those of Theorem 2 and it thus postponed to Section B of the appendix for the sake of readability. Theorem 7 is particularly relevant if INNA converges (otherwise the statement is trivial). Next result provides sufficient conditions so that convergence holds.

Theorem 8. *Assume that $\alpha > 0$. Let $(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P$ and let $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ be the sequence generated by INNA initialized at (θ_0, ψ_0). Under Assumption 2, if the step-size γ is such that,*

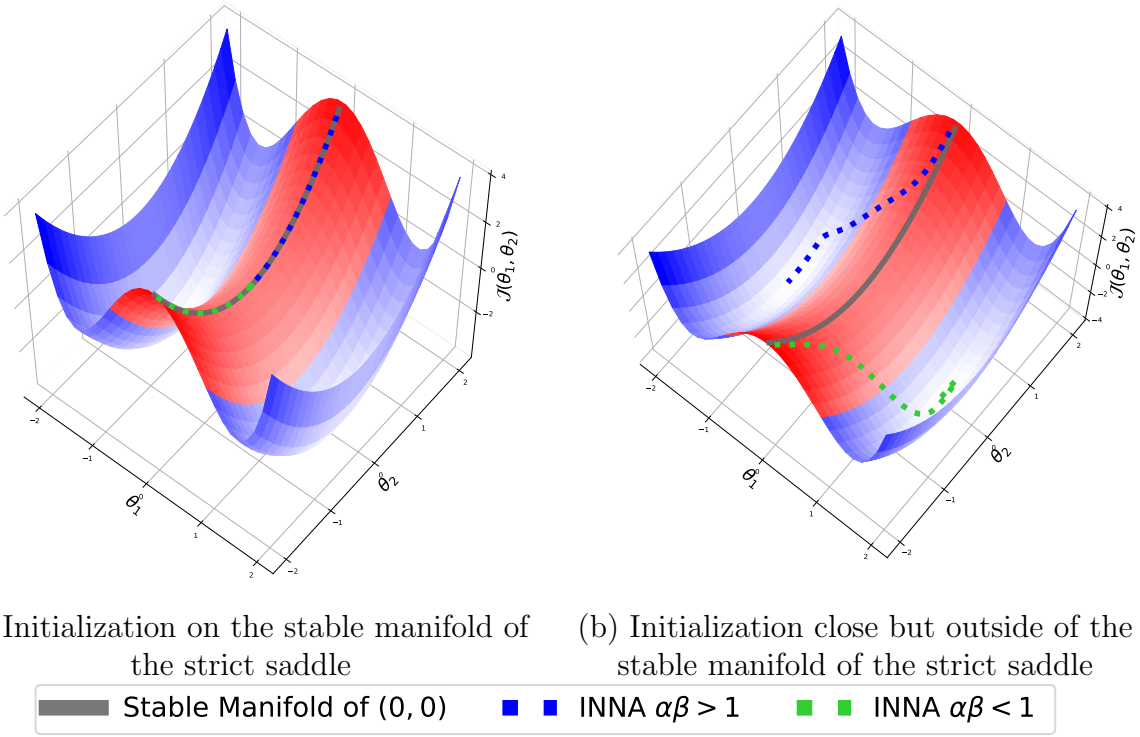
$$0 < \gamma < \min\left(\frac{2\alpha}{(1+\alpha\beta)L_{\nabla\mathcal{J}} + \alpha^2}, \frac{1}{\alpha} + \beta, 2\beta\right), \tag{18}$$

then the sequence of values $(\mathcal{J}(\theta_k))_{k\in\mathbb{N}}$ converges and $\lim_{k\to+\infty} \|\nabla\mathcal{J}(\theta_k)\|^2 = 0$.
If in addition the sequence $(\theta_k)_{k\in\mathbb{N}}$ is bounded and Assumption 1 holds, then $(\theta_k)_{k\in\mathbb{N}}$ converges to a critical point of $\mathcal{J}$.

Remark 3 (Comments on Theorem 8)**.**

– *Note that in (18), only the condition $\gamma < 2\alpha/\left((1+\alpha\beta)L_{\nabla\mathcal{J}} + \alpha^2\right)$ depends on the function $\mathcal{J}$ (through the Lipschitz constant $L_{\nabla\mathcal{J}}$), the other two being only related to the hyper-parameters of INNA. This is similar to the well-known condition $\gamma < 2/L_{\nabla\mathcal{J}}$ used for gradient descent, see e.g., Bertsekas et al. (1998, Proposition 2.3.2).*

– *The first part of the theorem states in particular that any accumulation point (the limit of any sub-sequence) is a critical point of $\mathcal{J}$. The additional boundedness assumption is necessary only to guarantee the convergence of $(\theta_k)_{k\in\mathbb{N}}$ to such critical points. Due to the convergence of $(\mathcal{J}(\theta_k))_{k\in\mathbb{N}}$, this assumption is valid, for example, when $\mathcal{J}$ is coercive, which holds in many problems met in practice.*



(a) Initialization on the stable manifold of the strict saddle

(b) Initialization close but outside of the stable manifold of the strict saddle

Figure 3: Evolution of the iterates of INNA on the landscape of the toy function $\mathcal{J} : (\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \theta_1^4 - 4\theta_1^2 + \theta_2^2$. This function has two minimizers $(-\sqrt{2}, 0)$ and $(\sqrt{2}, 0)$ and one strict saddle point $(0, 0)$. The red and blue surfaces represent the parts where $\mathcal{J}$ is locally concave and convex respectively. The stable manifold of $\mathcal{J}$ around $(0, 0)$ is represented by the grey curve. Left figure shows the behavior of INNA for two choices of hyper-parameters and for two initializations belonging to the stable manifold of $(0, 0)$. In this setting the algorithm does converge to the strict saddle $(0, 0)$. When initialized near but outside the manifold (right figure), the algorithm avoids the strict saddle and converges to a local minimizer for both choices of hyper-parameters.

The proof relies on a Lyapunov argument and is also postponed to the appendix (see Section C). Combining Theorem 7 and 8, we can formulate a corollary suited for practical applications.

Corollary 9. *Assume that $\mathcal{J}$ is a twice continuously differentiable coercive Morse function and that Assumption 2 holds. Assume that $\alpha > 0$, and that the step-size $\gamma > 0$ is such that both (17) and (18) hold. Let (θ_0, ψ_0) be a non-degenerate random variable on $\mathbb{R}^P \times \mathbb{R}^P$, and let $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ be the iterations of INNA initialized at (θ_0, ψ_0). Then $(\theta_k)_{k\in\mathbb{N}}$ converges, and with probability one, the limit $\theta^\star \in \mathbb{R}^P$ is a local minimizer of $\mathcal{J}$.*

The proof follows similar lines as those of Corollary 3 and the practical consequences of Corollary 9 are the same as those discussed for DIN. We conclude the study with numerical illustrations.

4.2 Numerical Illustration

We finish the study of INNA with a short empirical illustration of Theorem 7 on a toy example. To this aim we consider the function $\mathcal{J} : (\theta_1, \theta_2) \in \mathbb{R}^2 \mapsto \theta_1^4 - 4\theta_1^2 + \theta_2^2$. This function is twice continuously differentiable on $\mathbb{R}^2$, non-convex, has a diagonal Hessian and three critical points:

Table 1: Empirical validation of the results of Theorem 7

		Percentage of convergence to each critical point			Average number of iterations to escape a saddle point
		$(-\sqrt{2},0)$	$(\sqrt{2},0)$	$(0,0)$	
Initialization outside the stable manifold, very close to $(0,0)$	INNA $\alpha\beta<1$	48,8%	51,2%	0%	36
	INNA $\alpha\beta>1$	50,7%	49,3%	0%	35
	Gradient Descent	50,2%	49,8%	0%	37
Initialization on the stable manifold	INNA $\alpha\beta<1$	0%	0%	100%	-
	INNA $\alpha\beta>1$	0%	0%	100%	-
	Gradient Descent	0%	0%	100%	-

two local minimizers $(-\sqrt{2},0)$ and $(\sqrt{2},0)$ and one strict saddle point $(0,0)$. The landscape of $\mathcal{J}$ is displayed on Figure 3. The set of initializations such that INNA converges to the strict saddle point $(0,0)$ is the manifold $\theta_2 \in \mathbb{R} \mapsto (0,\theta_2)$ which has indeed zero measure on $\mathbb{R}^2$. Figure 3 shows that when initialized on this manifold, the algorithm does converge to $(0,0)$ but when initialized anywhere else, it avoids the strict saddle.

In addition to this illustration, we ran INNA and gradient descent —which is also known for almost surely escaping strict saddle points (Lee et al., 2016)—for 1000 random Gaussian initializations sampled from $\mathcal{N}_2(0,10^{-24})$, hence extremely close to the saddle point $(0,0)$. We also perform the same experiment but with a random initialization on the stable manifold. The results reported on Table 1 demonstrate that the algorithm always escapes the saddle point and converges to one of the two local minimizers. This empirically illustrates Theorem 7 and Corollary 9.

5 Conclusion

In this work, we provided a better understanding of the role played by the hyper-parameters α and β. This could help users of INNA to choose these parameters in practical applications. More importantly, we proved that the asymptotic behaviors of INNA and DIN make them relevant to tackle non-convex minimization problems. In particular, we provided conditions so that INNA converges and so that it is likely to avoid strict saddle points for almost all initializations.

Acknoledgments

The author acknowledges the support of the European Research Council (ERC FACTORY-CoG-6681839) and the Air Force Office of Scientific Research (FA9550-18-1-0226). The author deeply thanks Jérôme Bolte, Cédric Févotte, and Edouard Pauwels for their valuable comments and suggestions which helped in the process of writing this article.
The numerical experiments were made thanks to the development teams of the following libraries: Python (Rossum, 1995), Numpy (Walt et al., 2011) and Matplotlib (Hunter, 2007).

References

Alvarez, F., Attouch, H., Bolte, J., and Redont, P. (2002). A second-order gradient-like dissipative dynamical system with Hessian-driven damping: Application to optimization and mechanics. *Journal de Mathématiques Pures et Appliquées*, 81(8):747–779.

Attouch, H., Chbani, Z., Fadili, J., and Riahi, H. (2020). First-order optimization algorithms via inertial systems with Hessian driven damping. *Mathematical Programming*.

Attouch, H., Chbani, Z., Fadili, J., and Riahi, H. (2021). Convergence of iterates for first-order optimization algorithms with inertia and hessian driven damping. *arXiv preprint:2107.05943*.

Attouch, H., Peypouquet, J., and Redont, P. (2014). A dynamical approach to an inertial forward-backward algorithm for convex minimization. *SIAM Journal on Optimization*, 24(1):232–256.

Attouch, H., Peypouquet, J., and Redont, P. (2016). Fast convex optimization via inertial dynamics with Hessian driven damping. *Journal of Differential Equations*, 261(10):5734–5783.

Attouch, H. and Redont, P. (2001). The second-order in time continuous newton method. In *Approximation, optimization and mathematical economics*, pages 25–36. Springer.

Bertsekas, D. P., Hager, W., and Mangasarian, O. (1998). *Nonlinear programming*. Athena Scientific Belmont, MA.

Castera, C., Bolte, J., Févotte, C., and Pauwels, E. (2021). An inertial newton algorithm for deep learning. *Journal of Machine Learning Research*, 22(134):1–31.

Chen, L. and Luo, H. (2019). First order optimization methods based on Hessian-driven Nesterov accelerated gradient flow. *arXiv:1912.09276*.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S., and Bengio, Y. (2014). Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27.

Goudou, X. and Munier, J. (2009). The gradient and heavy ball with friction dynamical systems: the quasiconvex case. *Mathematical Programming*, 116(1):173–191.

Grobman, D. M. (1959). Homeomorphism of systems of differential equations. *Doklady Akademii Nauk SSSR*, 128(5):880–881.

Haragus, M. and Iooss, G. (2010). *Local bifurcations, center manifolds, and normal forms in infinite-dimensional dynamical systems*. Springer Science & Business Media.

Hartman, P. (1960). A lemma in the theory of structural stability of differential equations. *Proceedings of the American Mathematical Society*, 11(4):610–620.

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3):90–95.

Kelley, A. (1966). The stable, center-stable, center, center-unstable, unstable manifolds. *Journal of Differential Equations*.

Lange, K. (2013). *Optimization*, volume 2. Springer texts in statistics.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In *Conference on learning theory (COLT)*, pages 1246–1257. PMLR.

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

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

O'Neill, M. and Wright, S. J. (2019). Behavior of accelerated gradient methods near critical points of nonconvex functions. *Mathematical Programming*, 176(1):403–427.

Palmer, K. J. (1973). A generalization of Hartman's linearization theorem. *Journal of Mathematical Analysis and Applications*, 41(3):753–758.

Perko, L. (2013). *Differential equations and dynamical systems*, volume 7. Springer Science & Business Media.

Pliss, V. A. (1964). A reduction principle in the theory of stability of motion. *Izvestiya Akademii Nauk SSSR. Seriya Matematicheskaya*, 28.

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

Rossum, G. (1995). *Python reference manual*. CWI (Centre for Mathematics and Computer Science).

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. (2018). Understanding the acceleration phenomenon via high-resolution differential equations. *arXiv:1810.08907*.

Shub, M. (2013). *Global stability of dynamical systems*. Springer Science & Business Media.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2510–2518.

Walt, S. v. d., Colbert, S. C., and Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. *Computing in science & engineering*, 13(2):22–30.

A Permutation matrices

In this section, we specify the permutations matrices necessary to obtain the block diagonalization in (7) and (21). Denote by mod the modulo operator and let $P \in \mathbb{N}_{>0}$. We can choose the permutation matrix $U \in \mathbb{R}^{2P \times 2P}$ as the matrix whose coefficients are all zero except the

following, for all $p \in \{1, \dots, P\}$,

$$\text{if } P \text{ is odd: } \begin{cases} U_{P-p+1,p} & = 1 - \mathrm{mod}(p,2) \\ U_{P+p,2P-p+1} & = \mathrm{mod}(p,2) \\ U_{p,2P-p} & = \mathrm{mod}(p,2) \end{cases},$$

$$\text{if } P \text{ is even: } \begin{cases} U_{p,p} & = \mathrm{mod}(p,2) \\ U_{P+p,P+p} & = 1 - \mathrm{mod}(p,2) \\ U_{P+p,p} & = \mathrm{mod}(p,2) \\ U_{p,P+p} & = \mathrm{mod}(p,2) \end{cases}. \tag{19}$$

For example, for $P = 3$ and $P = 4$ this yields the following matrices (where the zero coefficients are omitted for the sake of readability),

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 \\ 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 \end{pmatrix}.$$

B Proof of Theorem 7

In order to prove Theorem 7, we will use a version of the stable manifold theorem suited to the analysis of discrete processes.

B.1 Stable manifold theorem for discrete processes

We introduce a different version of the stable manifold theorem. This version was used by Lee et al. (2016) and O'Neill and Wright (2019) to analyze gradient descent and the HBF methods respectively. For a function $F : \mathbb{R}^P \to \mathbb{R}^P$ and for all $k \in \mathbb{N}_{>0}$, we introduce the following notation: $F^k = \underbrace{F \circ \ldots \circ F}_{k \text{ compisitions}}$. The result is the following.

Theorem 10 (III.7 (Shub, 2013))**.** *Let $\Theta^\star \in \mathbb{R}^{2P}$ be a fixed point for the C^1 local diffeomorphism $F : \mathsf{U} \to \mathbb{R}^{2P}$ where $\mathsf{U} \subset \mathbb{R}^{2P}$ is a neighborhood of $\Theta^\star$. Let $\mathsf{E}^{sc}(\Theta^\star)$ be the linear subspace spanned by the (complex) eigenvalues of $DF(\Theta^\star)$ with magnitude less than one. There exists a neighborhood Ω of $\Theta^\star$ and a C^1 manifold $\mathsf{W}^{sc}(\Theta^\star)$ tangent to $\mathsf{E}^{sc}(\Theta^\star)$ at $\Theta^\star$—whose dimension is the number of eigenvalues of $DF(\Theta^\star)$ with magnitude less than one—such that, for $\Theta_0 \in \mathbb{R}^{2P}$,*

(i) If $\Theta_0 \in \mathsf{W}^{sc}(\Theta^\star)$ and $F(\Theta_0) \in \Omega$ then $F(\Theta_0) \in \mathsf{W}^{sc}(\Theta^\star)$ (Invariance).

(ii) If $\forall k \in \mathbb{N}_{>0},\ F^k(\Theta_0) \in \Omega$, then $\Theta_0 \in \mathsf{W}^{sc}(\Theta^\star)$.

Although we study an iterative algorithm and not the solutions of an ODE, the results stated in Theorem 10 are very similar to those of Theorem 4, thus, the proof of Theorem 7 is close to the proof of Theorem 2.

Formulating INNA to use Theorem 10. Proceeding similarly to Section 3.1, for any $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$, we redefine the mapping G as,

$$G\begin{pmatrix}\theta\\ \psi\end{pmatrix} = \begin{pmatrix}\theta + \gamma\left[-(\alpha - \frac{1}{\beta})\theta - \frac{1}{\beta}\psi - \beta\nabla\mathcal{J}(\theta)\right]\\ \psi + \gamma\left[-(\alpha - \frac{1}{\beta})\theta - \frac{1}{\beta}\psi\right]\end{pmatrix}, \tag{20}$$

so that an iteration $k \in \mathbb{N}$ of INNA reads $(\theta_{k+1}, \psi_{k+1}) = G(\theta_k, \psi_k)$. Remark that unlike for (3), we now study the fixed points of G and not its zeros. Indeed, the iterative process INNA consists in successive compositions of the operator G and the set of fixed points of G is exactly S. To prove this statement, let $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$,

$$G(\theta,\psi) = (\theta,\psi) \iff \begin{cases}-(\alpha - \frac{1}{\beta})\theta - \frac{1}{\beta}\psi - \beta\nabla\mathcal{J}(\theta) = 0\\ -(\alpha - \frac{1}{\beta})\theta - \frac{1}{\beta}\psi = 0\end{cases} \iff \begin{cases}\nabla\mathcal{J}(\theta) = 0\\ \psi = (1-\alpha\beta)\theta\end{cases}.$$

So (θ, ψ) is a fixed point of G if and only if $\nabla\mathcal{J}(\theta) = 0$ and $\psi = (1 - \alpha\beta)\theta$.

B.2 Proof of Theorem 7

Block-diagonal transformation. Throughout the proof we use a block-diagonal transformation. Let $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$. Since $\mathcal{J}$ is C^2 on $\mathbb{R}^P$ then G is C^1 on $\mathbb{R}^P \times \mathbb{R}^P$ and the Jacobian matrix of G at (θ, ψ) (displayed by block) reads,

$$DG(\theta,\psi) = \begin{pmatrix}(1 - \gamma(\alpha - \frac{1}{\beta}))I_P - \gamma\beta\nabla^2\mathcal{J}(\theta) & -\frac{\gamma}{\beta}I_P\\ -\gamma(\alpha - \frac{1}{\beta})I_P & (1 - \frac{\gamma}{\beta})I_P\end{pmatrix}.$$

Proceeding like in Section 3.1, their exists an orthogonal matrix $V \in \mathbb{R}^{P\times P}$ and a permutation $U \in \mathbb{R}^{2P\times 2P}$ (defined in Section A) such that,

$$U^T\begin{pmatrix}V^T & 0\\ 0 & V^T\end{pmatrix} DG(\theta,\psi)\begin{pmatrix}V & 0\\ 0 & V\end{pmatrix}U = \begin{pmatrix}M_1 & & \\ & \ddots & \\ & & M_P\end{pmatrix}, \tag{21}$$

where for each $p \in \{1, \dots, P\}$, $M_p = \begin{pmatrix}1 - \gamma(\alpha - \frac{1}{\beta}) - \gamma\beta\lambda_p & -\frac{\gamma}{\beta}\\ -\gamma(\alpha - \frac{1}{\beta}) & 1 - \frac{\gamma}{\beta}\end{pmatrix}$—up to a symmetric permutation—and $(\lambda_p)_{p\in\{1,\dots,P\}}$ are the eigenvalues of $\nabla^2\mathcal{J}(\theta)$. To apply the stable manifold theorem and prove Theorem 7 we need G to be a local diffeomorphism. This result is non-straightforward to obtain, so we state it as a theorem before proving Theorem 7.

Theorem 11. *Under Assumption 2, for any $\alpha > 0$, $\beta > 0$ and*

$$0 < \gamma < \min\left(\frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha + \beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}, \beta\right),$$

the mapping G defined in (20) *is a local diffeomorphism from $\mathbb{R}^P \times \mathbb{R}^P$ to $\mathbb{R}^P \times \mathbb{R}^P$.*

We finish the proof of Theorem 7 first, and then prove this theorem in Section B.3.

Application of Theorem 10 to prove Theorem 7. We can now prove Theorem 7.

Proof of Theorem 7. Let $\alpha > 0$, $\beta > 0$ and $0 < \gamma < \min\left(\frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}, \beta\right)$. Consider the mapping G defined in (20) with these parameters. By direct application of Theorem 11, G is a local diffeomorphism. Let $(\theta^\star, \psi^\star) \in \mathbb{R}^P \times \mathbb{R}^P$ be a fixed point of G. Our goal is to apply the stable manifold theorem in a neighborhood of this point.

To this aim, we study under which conditions on the eigenvalues of $\nabla^2\mathcal{J}(\theta^\star)$ the eigenvalues of $DG(\theta^\star, \psi^\star)$ have magnitude less than one. Throughout the proof we consider the same block-diagonal transformation of $DG(\theta^\star, \psi^\star)$ as in (21), and we keep the same notations. Let $p \in \{1, \dots, P\}$, the eigenvalues of M_p are the roots of the following polynomial,

$$\chi_{M_p}(X) = X^2 - \mathrm{trace}(M_p)X + \det(M_p) = X^2 - (2 - \gamma(\alpha + \beta\lambda_p))X + 1 - \gamma(\alpha + \beta\lambda_p) + \gamma^2\lambda_p.$$

The discriminant of χ_{M_p} is,

$$\begin{aligned}\Delta_{M_p} &= (2 - \gamma(\alpha + \beta\lambda_p))^2 - 4(1 - \gamma(\alpha + \beta\lambda_p) + \gamma^2\lambda_p) \\ &= 4 + \gamma^2(\alpha + \beta\lambda_p)^2 - 4\gamma(\alpha + \beta\lambda_p) - 4 + 4\gamma(\alpha + \beta\lambda_p) - 4\gamma^2\lambda_p \\ &= \gamma^2\left((\alpha + \beta\lambda_p)^2 - 4\lambda_p\right).\end{aligned}$$

Remark that up to a factor $\gamma^2 > 0$, this is the same discriminant as in (8) from Section 3.1. Therefore, we can once again use Lemma 5 to deduce that Δ_{M_p} is non-positive if and only if $\alpha\beta \leq 1$ and $\lambda_p \in \left[\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right]$. We split the study with respect to the sign of Δ_{M_p}.

- If $\Delta_{M_p} > 0$, then M_p has two real eigenvalues,

$$\begin{cases}\sigma_{p,+} &= 1 - \frac{1}{2}\gamma(\alpha + \beta\lambda_p) + \frac{1}{2}\gamma\sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p} \\ \sigma_{p,-} &= 1 - \frac{1}{2}\gamma(\alpha + \beta\lambda_p) - \frac{1}{2}\gamma\sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p}\end{cases}.$$

  We then study whether the magnitudes of the eigenvalues are smaller or larger than 1, the computations are very similar to those of Section 3.1. If $\lambda_p < 0$, then $|(\alpha + \beta\lambda_p)| < \sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p}$, so $\sigma_{p,+} > 1$ and $\sigma_{p,-} < 1$, so we have at least one eigenvalue with magnitude larger than one. If $\lambda_p = 0$, then $\sigma_{p,+} = 1$ and $|\sigma_{p,-}| = |1 - \gamma\alpha| \leq 1$.[2]

  In order to be exhaustive, remark that if $\lambda_p > 0$, then $(\alpha + \beta\lambda_p) > \sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p}$, so $0 \leq \sigma_{p,+} < 1$ and $-\infty < \sigma_{p,-} < 1$, so the bounds enforced on γ ensure that both eigenvalues have magnitude less than 1. This is indeed the case whenever $-\sigma_{p,-} < 1$, which is equivalent to $\gamma\left[\alpha + \beta\lambda_p + \sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p}\right] < 4$ and the latter always holds true. Indeed, when $\alpha\beta > 1$, one can show that the function $x > 0 \mapsto \alpha + \beta x + \sqrt{(\alpha + \beta x)^2 - 4x}$ is increasing (by differentiating it). Then, using Assumption 2 and the upper bound enforced on γ it holds,

$$\begin{aligned}&\gamma\left[\alpha + \beta\lambda_p + \sqrt{(\alpha + \beta\lambda_p)^2 - 4\lambda_p}\right] \\ &< \frac{(\alpha + \beta L_{\nabla\mathcal{J}}) - \sqrt{(\alpha + \beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}\left[(\alpha + \beta L_{\nabla\mathcal{J}}) + \sqrt{(\alpha + \beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}\right] \\ &< \frac{4L_{\nabla\mathcal{J}}}{2L_{\nabla\mathcal{J}}} = 2 < 4. \quad (22)\end{aligned}$$

[2] This is ensured by the boundaries enforced on γ as shown in the proof of Theorem 11 in Section B.3.

On the other hand, when $\alpha\beta \leq 1$, by studying again the function $x > 0 \mapsto \alpha + \beta x + \sqrt{(\alpha+\beta x)^2 - 4x}$, one can show[3] that,

$$\alpha + \beta\lambda_p + \sqrt{(\alpha+\beta\lambda_p)^2 - 4\lambda_p} \leq \max\left(2\alpha, \alpha + \beta L_{\nabla\mathcal{J}} + \sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}\right). \quad (23)$$

Then if the maximum in the right-hand side of (23) is 2α, it holds,

$$\gamma\left[\alpha + \beta\lambda_p + \sqrt{(\alpha+\beta\lambda_p)^2 - 4\lambda_p}\right] \leq 2\alpha\gamma \leq 2\alpha\beta \leq 2 < 4,$$

and if the maximum is the other value, we use (22) again. To summarize, when $\Delta_{M_p} > 0$, $\lambda_p \geq 0 \iff |\sigma_{p,+}| \leq 1$ and $|\sigma_{p,-}| \leq 1$.

- If $\Delta_{M_p} \leq 0$ then this implies that $\lambda_p \geq 0$ so $(\theta^\star, \psi^\star) \notin \mathsf{S}_{<0}$ and we do not need additional arguments. However, for the sake of completeness, we check whether the manifold in Theorem 10 may have positive measure around any local minimizer with non-singular Hessian (in particular around any minimizer of a Morse function). The eigenvalues of M_p are,

$$\begin{cases} \sigma_{p,+} &= 1 - \frac{1}{2}\gamma(\alpha+\beta\lambda_p) + \frac{i}{2}\gamma\sqrt{4\lambda_p - (\alpha+\beta\lambda_p)^2} \\ \sigma_{p,-} &= 1 - \frac{1}{2}\gamma(\alpha+\beta\lambda_p) - \frac{i}{2}\gamma\sqrt{4\lambda_p - (\alpha+\beta\lambda_p)^2} \end{cases}.$$

Both eigenvalues have the same magnitude,

$$|\sigma_{p,+}|^2 = |\sigma_{p,-}|^2 = \left(1 - \frac{1}{2}\gamma(\alpha+\beta\lambda_p)\right)^2 + \frac{1}{4}\gamma^2(4\lambda_p - (\alpha+\beta\lambda_p)^2) = 1 - \gamma(\alpha+\beta\lambda_p) + \gamma^2\lambda_p,$$

so,

$$|\sigma_{p,+}|^2 < 1 \iff -\gamma(\alpha+\beta\lambda_p) + \gamma^2\lambda_p < 0 \iff (\gamma-\beta)\lambda_p < \alpha \iff \gamma < \beta + \frac{\alpha}{\lambda_p}.$$

This is always true since,

$$\gamma < \frac{1}{2}(\beta + \frac{\alpha}{L_{\nabla\mathcal{J}}}) - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}} \leq \frac{1}{2}(\beta + \frac{\alpha}{L_{\nabla\mathcal{J}}}) \leq \beta + \frac{\alpha}{\lambda_p}.$$

We just proved that the eigenvalues of $DG(\theta^\star, \psi^\star)$ have magnitude less than one if and only if $(\theta^\star, \psi^\star) \in \mathsf{S} \setminus \mathsf{S}_{<0}$.

We can now use the stable manifold theorem. Let $(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}$. Let an initialization (θ_0, ψ_0) such that the associated realization $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ of INNA converges to $(\theta^\star, \psi^\star)$. Consider the manifold $\mathsf{W}^{sc}(\theta^\star, \psi^\star)$ and the neighborhood Ω as defined in Theorem 10. Since $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ converges, there exists $k_0 \in \mathbb{N}$ such that for all $k \geq k_0$, $(\theta_k, \psi_k) \in \Omega$, so according to Theorem 10, $\forall k \geq k_0$, $(\theta_k, \psi_k) \in \Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$. Rewriting this with the operator G, $\forall k \geq k_0$, $G^k(\theta_0, \psi_0) \in \Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$, and hence $\forall k \geq k_0$,

$$G^k(\theta_0, \psi_0) \in \bigcup_{j\in\mathbb{N}} G^{-j}\left(\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)\right), \quad (24)$$

where $G^{-j}\left(\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)\right)$ corresponds to all the initial conditions such that INNA has reached $\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star)$ after j iterations. Let

$$\mathsf{W}(\theta^\star, \psi^\star) = \left\{(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P \middle| G^k(\theta_0, \psi_0) \xrightarrow[k\to+\infty]{} (\theta^\star, \psi^\star)\right\},$$

[3]The proof is similar to the one of Lemma 14 proved in Section B.3.

the set of all initial conditions such that INNA converges to $(\theta^\star, \psi^\star)$. From (24), it holds that,

$$\mathsf{W}(\theta^\star, \psi^\star) \subset \bigcup_{j \in \mathbb{N}} G^{-j} \left(\Omega \cap \mathsf{W}^{sc}(\theta^\star, \psi^\star) \right). \tag{25}$$

Then, we showed that since $(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}$, then $DG(\theta^\star, \psi^\star)$ has at least one eigenvalue with magnitude strictly larger than one, so according to the stable manifold theorem, the dimension of $\mathsf{W}^{sc}(\theta^\star, \psi^\star)$ is strictly less than $2P$, hence this manifold has zero measure. By assumption the step-size γ is chosen such that G is a local diffeomorphism (from Theorem 11), so $\forall j \in \mathbb{N}$, G^{-j} is also a local diffeomorphism, hence it maps zero-measure sets to zero-measure sets. As a result, the right-hand side in (25) is a countable union of zero-measure sets, so it has zero measure, as well as $\mathsf{W}(\theta^\star, \psi^\star)$.

This proves the theorem since Assumption 1 guarantees that there is at most a countable number of critical points. So $\bigcup_{(\theta^\star, \psi^\star) \in \mathsf{S}_{<0}} \mathsf{W}(\theta^\star, \psi^\star)$ is a countable union of zero-measure sets so it has zero measure. □

B.3 Proof of Theorem 11

To prove Theorem 11, we introduce three technical lemmas.

Lemma 12. *For any $\alpha > 0$ and $\beta > 0$ such that $\alpha\beta > 1$, the function*

$$x \in \mathbb{R}_{>0} \mapsto \frac{\beta}{2} + \frac{\alpha}{2x} - \frac{\sqrt{(\alpha + \beta x)^2 - 4x}}{2x}$$

is continuous and decreasing both on $\mathbb{R}_{>0}$ and $\mathbb{R}_{<0}$.

Proof of Lemma 12. Let $\alpha > 0$ and $\beta > 0$ such that $\alpha\beta > 1$. The function $x \in \mathbb{R}_{>0} \mapsto \frac{\beta}{2} + \frac{\alpha}{2x} - \frac{\sqrt{(\alpha+\beta x)^2 - 4x}}{2x}$ is clearly $C^\infty(\mathbb{R}_{>0})$, and its first-order derivative is the function $x \in \mathbb{R}_{>0} \mapsto -\frac{\alpha\sqrt{(\beta x+\alpha)^2-4x}+(2-\alpha\beta)x-\alpha^2}{2x^2\sqrt{(\beta x+\alpha)^2-4x}}$. Since the denominator is always positive, we study the numerator of this derivative: define $h : x \in \mathbb{R}_{>0} \mapsto -\alpha\sqrt{(\beta x + \alpha)^2 - 4x} - (2 - \alpha\beta)x + \alpha^2$. We will prove that h is negative by differentiating it: for all $x \in \mathbb{R}_{>0}$,

$$\frac{\partial h}{\partial x}(x) = -\frac{\alpha(2\beta(\beta x + \alpha) - 4)}{2\sqrt{(\beta x + \alpha)^2 - 4x}} + \alpha\beta - 2. \tag{26}$$

$$\frac{\partial^2 h}{\partial x^2}(x) = -\frac{4\alpha(\alpha\beta - 1)}{((\beta x + \alpha)^2 - 4x)^{\frac{3}{2}}}. \tag{27}$$

Since $\alpha\beta > 1$, for all $x \in \mathbb{R}_{>0}$, $\frac{\partial^2 h}{\partial x^2}(x) < 0$ and hence for all $x \in \mathbb{R}_{>0}$, $\frac{\partial h}{\partial x}(x) < \lim_{t \to 0} \frac{\partial h}{\partial x}(t) = 0$. So h is also decreasing on $\mathbb{R}_{>0}$, and $\lim_{x \to 0} h(x) = 0$, so for all $x \in \mathbb{R}_{>0}$, $h(x) \leq 0$ and the claim is thus proved on $\mathbb{R}_{>0}$. The proof is very similar on $\mathbb{R}_{<0}$ except that h is increasing on $\mathbb{R}_{<0}$ but $\lim_{x \to 0} h(x) = 0$, hence the result. □

Lemma 13. *For $\alpha > 0$, $\beta > 0$ such that $\alpha\beta \leq 1$, the function*

$$x \in \mathbb{R}_{<0} \mapsto \frac{\beta}{2} + \frac{\alpha}{2x} - \frac{\sqrt{(\alpha + \beta x)^2 - 4x}}{2x}$$

is continuous and increasing on $\mathbb{R}_{<0}$.

Proof of Lemma 13. The proof follows the exact same steps as those of the proof of Lemma 12, except that in (27), $\frac{\partial^2 h}{\partial x^2}$ is always positive on $\mathbb{R}_{<0}$, and we use that to deduce that $\frac{\partial h}{\partial x}$ defined in (26) is negative on $\mathbb{R}_{<0}$. So h is decreasing on $\mathbb{R}_{<0}$ and since $h(0) = 0$ we eventually obtain the result. □

Lemma 14. *Let $\alpha > 0$, $\beta > 0$ such that $\alpha\beta \leq 1$, the function*

$$x \in \mathbb{R}_{>0} \setminus [\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}] \mapsto \frac{\beta}{2} + \frac{\alpha}{2x} - \frac{\sqrt{(\alpha+\beta x)^2 - 4x}}{2x}$$

is continuous and increasing for $x \in \left(0, \frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}\right)$ and continuous and decreasing for $x \in \left(\frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, +\infty\right)$.

Proof of Lemma 14. Let $\alpha > 0$, $\beta > 0$ such that $\alpha\beta \leq 1$. Denote by $x^- = \frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}$ and $x^+ = \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}$ The function $x \in \mathbb{R}_{>0} \setminus (x^-, x^+) \mapsto \frac{\beta}{2} + \frac{\alpha}{2x} - \frac{\sqrt{(\alpha+\beta x)^2-4x}}{2x}$ is C^∞ on $(0, x^-)$ and on $(x^+, +\infty)$. Its first-order derivative is $x \in \mathbb{R}_{>0} \setminus (x^-, x^+) \mapsto -\frac{\alpha\sqrt{(\beta x+\alpha)^2-4x}+(2-\alpha\beta)x-\alpha^2}{2x^2\sqrt{(\beta x+\alpha)^2-4x}}$. The denominator is positive, so we focus on the numerator, define $h : x \in \mathbb{R}_{>0} \setminus (x^-, x^+) \mapsto -\alpha\sqrt{(\beta x+\alpha)^2-4x} - (2-\alpha\beta)x + \alpha^2$. For all $x \in \mathbb{R}_{>0} \setminus (x^-, x^+)$, the first and second-order derivatives of h are given by (26) and (27) respectively.

Since $\alpha\beta < 1$, $\frac{\partial^2 h}{\partial x^2}$ is always positive, so $\frac{\partial h}{\partial x}$ is increasing on both intervals. First, when $x \to 0$ with $0 < x < x^-$, $\frac{\partial h}{\partial x}(x) \to 0$, so $\frac{\partial h}{\partial x}$ is positive on $(0, x^-)$ and h is increasing on $(0, x^-)$. Since $h(0) = 0$ and h is increasing, we proved the first part of the lemma. Then, when $x \to +\infty$, $\frac{\partial h}{\partial x} \to -2$, so $\frac{\partial h}{\partial x}$ is negative on $(x^+, +\infty)$ and h is decreasing on $(x^+, +\infty)$. Finally, $h(x^+) = -\frac{4(1-\alpha\beta)}{\beta^2} - \frac{2(2-\alpha\beta)\sqrt{1-\alpha\beta}}{\beta^2} \leq 0$. □

We finally use these lemmas to prove the theorem.

Proof of Theorem 11. Let $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$. Since $\mathcal{J}$ is $C^2(\mathbb{R}^P)$ then G is $C^1(\mathbb{R}^{2P})$ and the Jacobian matrix $DG(\theta, \psi)$ can be transformed into a block diagonal matrix as in (21) where for any $p \in \{1, \dots, P\}$, M_p is a 2×2 block of the diagonal and λ_p is the associated eigenvalue of $\mathcal{J}(\theta)$. To prove that G is a local diffeomorphism we prove that $DG(\theta, \psi)$ is invertible (i.e., that it has non-zero determinant) and then use the local inversion theorem. It holds that $\det(DG(\theta, \psi)) = \prod_{p=1}^P \det(M_p)$, and for each $p \in \{1, \dots, P\}$,

$$\det(M_p) = (1 - \gamma(\alpha - \frac{1}{\beta}) - \gamma\beta\lambda_p)(1 - \frac{\gamma}{\beta}) - \frac{\gamma}{\beta}\gamma(\alpha - \frac{1}{\beta}) = 1 - \gamma(\alpha + \beta\lambda_p) + \gamma^2\lambda_p. \quad (28)$$

Let $p \in \{1, \dots, P\}$, we want to choose γ such that $\det(M_p) \neq 0$ for any $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$, hence for any $\lambda_p \in [-L_{\nabla\mathcal{J}}, L_{\nabla\mathcal{J}}]$ (since the eigenvalues are bounded by $L_{\nabla\mathcal{J}}$ from Assumption 2).

First, if $\lambda_p = 0$, from (28), we must take $\gamma \neq 1/\alpha$. From now on, we assume $\lambda_p \neq 0$, so (28) is a second-order polynomial in γ and its discriminant is $\Delta_\gamma = (\alpha + \beta\lambda_p)^2 - 4\lambda_p$. Notice that Δ_γ is a polynomial in λ_p and is exactly the discriminant that we studied in Section 3.1; its sign is given by Lemma 5. When this discriminant is non-negative, there exists two real roots to (28),

$$\begin{cases} \gamma^+ = \frac{(\alpha+\beta\lambda_p)}{2\lambda_p} + \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2\lambda_p} = \frac{\beta}{2} + \frac{\alpha}{2\lambda_p} + \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2\lambda_p} \\ \gamma^- = \frac{(\alpha+\beta\lambda_p)}{2\lambda_p} - \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2\lambda_p} = \frac{\beta}{2} + \frac{\alpha}{2\lambda_p} - \frac{\sqrt{(\alpha+\beta\lambda_p)^2-4\lambda_p}}{2\lambda_p} \end{cases}. \quad (29)$$

As before, we split the study with respect to the value of $\alpha\beta$.

- If $\alpha\beta > 1$, then $\Delta_\gamma \geq 0$ and the roots of $\det(M_p)$ are given by (29).
  - First, when $\lambda_p < 0$, $\sqrt{(\alpha+\beta\lambda_p)^2 - 4\lambda_p} > |\alpha + \beta\lambda_p|$, so $\gamma^+ < 0$ and any positive choice of γ will never be equal to γ^+ in this case. Then by Lemma 12, γ^- is a decreasing function of λ_p for $\lambda_p < 0$ and when $\lambda_p \to 0$, $\gamma^- \to 1/\alpha$ (using L'Hôpital's rule), this yields a first condition $\gamma < 1/\alpha$.
  - When $\lambda_p > 0$, observe that $\gamma^+ \geq \gamma^- > 0$ so we focus the study on γ^-. Lemma 12 exactly states that γ^- is a decreasing function of $\lambda_p > 0$. Since $\mathcal{J}$ has $L_{\nabla\mathcal{J}}$-Lipschitz gradient, $\lambda_p \leq L_{\nabla\mathcal{J}}$, so, for all $\lambda_p \in (0, L_{\nabla\mathcal{J}}]$, $\gamma^- \geq \frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}$. Note in addition that when $\lambda_p \to 0$, $\gamma^- \to 1/\alpha$, this is a simple way to prove that $1/\alpha > \frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}$ when $\alpha\beta > 1$.

  To summarize, we had three conditions, $\gamma \neq 1/\alpha$, $\gamma < 1/\alpha$ and $\gamma < \frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}$ and we proved that the latter implies the first-two conditions. Remark that the condition $\gamma < \beta$ holds but is not necessary in the case $\alpha\beta > 1$, it is present in the statement of the theorem to keep it as simple as possible.

- We now assume that $\alpha\beta \leq 1$.
  - If $\lambda_p < 0$, then $\Delta_\gamma > 0$ and the roots are given by (29). As above, , it holds that $\sqrt{(\alpha+\beta\lambda_p)^2 - 4\lambda_p} > |\alpha + \beta\lambda_p|$, so $\gamma^+ < 0$. Then Lemma 13 states that γ^- is an increasing function of $\lambda_p < 0$, and when $\lambda_p \to -\infty$, $\gamma^- \to \beta$. So we need $\gamma < \beta$.
  - If $\lambda_p > 0$, then whenever $\lambda_p \in [\frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, \frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}]$, there are no real roots so $\det(M_p) \neq 0$ regardless the choice of $\gamma > 0$. If λ_p does not belong to interval previously mentioned, the roots are given by (29). Remark that $\gamma^+ > \gamma^- > 0$ so we focus on γ^-. Using Lemma 14, γ^- is increasing on $(0, \frac{2-\alpha\beta}{\beta^2} - \frac{2\sqrt{1-\alpha\beta}}{\beta^2})$ and tends to $1/\alpha$ when $\lambda_p \to 0$ (using L'Hôpital's rule). The same lemma also state that γ^- is decreasing on $(\frac{2-\alpha\beta}{\beta^2} + \frac{2\sqrt{1-\alpha\beta}}{\beta^2}, +\infty)$, so using the $L_{\nabla\mathcal{J}}$-Lipschitz gradient of $\mathcal{J}$, $\gamma^- \leq \frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}$ on this interval. Note however that we do not necessarily have $1/\alpha > \frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}$.

  Overall, for $\alpha\beta \leq 1$ we must have $0 < \gamma < \min\left(\frac{\beta}{2} + \frac{\alpha}{2L_{\nabla\mathcal{J}}} - \frac{\sqrt{(\alpha+\beta L_{\nabla\mathcal{J}})^2 - 4L_{\nabla\mathcal{J}}}}{2L_{\nabla\mathcal{J}}}, \frac{1}{\alpha}, \beta\right)$ and $\alpha\beta \leq 1 \implies \beta \leq 1/\alpha$ hence the result.

In every case we proved that the conditions mentioned in the theorem are sufficient to ensure that for all $(\theta, \psi) \in \mathbb{R}^P \times \mathbb{R}^P$, $\det(DG(\theta, \psi)) \neq 0$. So by the local inversion theorem, G is a local diffeomorphism from $\mathbb{R}^P \times \mathbb{R}^P$ to $\mathbb{R}^P \times \mathbb{R}^P$. □

C Proof of convergence of INNA

The proof of Theorem 8 relies on a Lyapunov argument stated in the following lemma.

Lemma 15. *Assume that Assumption 2 holds and $\alpha > 0$. Let $(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P$, and let $\gamma > 0$ such that (18) holds. Let $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ be a sequence generated by INNA with this step-size*

and initialized at (θ_0, ψ_0). *Then the sequence* $(\mathcal{E}_k)_{k\in\mathbb{N}}$ *defined for all* $k \in \mathbb{N}$ *by*

$$\mathcal{E}_k = (1 + \alpha\beta - \gamma\alpha)\mathcal{J}(\theta_k) + \frac{1}{2}\|(\alpha - \frac{1}{\beta})\theta_k + \frac{1}{\beta}\psi_k\|^2 \tag{30}$$

is decreasing, converges, and there exists $C_1 > 0$ *and* $C_2 > 0$ *such that for all* $K \in \mathbb{N}$,

$$C_1 \sum_{k=0}^{K} \|\theta_{k+1} - \theta_k\|^2 + C_2 \sum_{k=0}^{K} \|\nabla\mathcal{J}(\theta_k)\|^2 \leq \mathcal{E}_0 - \mathcal{E}_{K+1} < +\infty. \tag{31}$$

Proof of Lemma 15. Consider INNA with the choice of step-size stated in the lemma, and let $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ be the iterates of INNA introduced in the lemma. We first introduce some notations: define $a = \alpha - \frac{1}{\beta}$ and $b = \frac{1}{\beta}$. For all $k \in \mathbb{N}$, denote also $\Delta\theta_k = \theta_{k+1} - \theta_k$ and $\Delta\psi_k = \psi_{k+1} - \psi_k$. With these notations, for all $k \in \mathbb{N}$ INNA can be rewritten as follows,

$$\begin{cases} \Delta\psi_k &= -\gamma a\theta_k - \gamma b\psi_k \\ \Delta\theta_k &= \Delta\psi_k - \gamma\beta\nabla\mathcal{J}(\theta_k) \end{cases}. \tag{32}$$

To simplify the notations denote also $\mu = 1 + \alpha\beta - \gamma\alpha$, which is positive since $\gamma < 1/\alpha + \beta$.

Now, let $k \in \mathbb{N}$, we aim to bound the difference $\mathcal{E}_{k+1} - \mathcal{E}_k$ by some non-positive quantity. First, using the $L_{\nabla\mathcal{J}}$-Lipschitz continuity of $\nabla\mathcal{J}$, we have a so-called *descent lemma* (see e.g., Bertsekas et al. 1998, Proposition A.24):

$$\mu\mathcal{J}(\theta_{k+1}) - \mu\mathcal{J}(\theta_k) \leq \mu\langle\nabla\mathcal{J}(\theta_k), \Delta\theta_k\rangle + \frac{\mu L_{\nabla\mathcal{J}}}{2}\|\Delta\theta_k\|^2,$$

which according to (32), can equivalently be rewritten as,

$$\mu\mathcal{J}(\theta_{k+1}) - \mu\mathcal{J}(\theta_k) \leq -\mu\langle\frac{\Delta\theta_k - \Delta\psi_k}{\gamma\beta}, \Delta\theta_k\rangle + \frac{\mu L_{\nabla\mathcal{J}}}{2}\|\Delta\theta_k\|^2. \tag{33}$$

We save this for later and now turn our attention to the other term in $\mathcal{E}_{k+1} - \mathcal{E}_k$,

$$\frac{1}{2}\|a\theta_{k+1} + b\psi_{k+1}\|^2 - \frac{1}{2}\|a\theta_k + b\psi_k\|^2 = \frac{1}{2}\|a\theta_k + a\Delta\theta_k + b\psi_k + b\Delta\psi_k\|^2 - \frac{1}{2}\|a\theta_k + b\psi_k\|^2.$$

We expand this and use the fact that $a\theta_k + b\psi_k = -\Delta\psi_k/\gamma$,

$$\begin{aligned} &\frac{1}{2}\|a\theta_k + a\Delta\theta_k + b\psi_k + b\Delta\psi_k\|^2 - \frac{1}{2}\|a\theta_k + b\psi_k\|^2 \\ =&\frac{1}{2}\|a\theta_k + b\psi_k\|^2 + \frac{1}{2}\|a\Delta\theta_k + b\Delta\psi_k\|^2 + \langle a\theta_k + b\psi_k, a\Delta\theta_k + b\Delta\psi_k\rangle - \frac{1}{2}\|a\theta_k + b\psi_k\|^2 \\ =&\frac{1}{2}\|a\Delta\theta_k + b\Delta\psi_k\|^2 - \frac{1}{\gamma}\langle\Delta\psi_k, a\Delta\theta_k + b\Delta\psi_k\rangle \\ =&\frac{a^2}{2}\|\Delta\theta_k\|^2 + \frac{b^2}{2}\|\Delta\psi_k\|^2 + ab\langle\Delta\theta_k, \Delta\psi_k\rangle - \frac{a}{\gamma}\langle\Delta\theta_k, \Delta\psi_k\rangle - \frac{b}{\gamma}\|\Delta\psi_k\|^2. \end{aligned} \tag{34}$$

Then, we rewrite (34) only as functions of $\Delta\theta_k$ and $\Delta\theta_k - \Delta\psi_k$ as we did in (33). To do so, we use

$$\|\Delta\psi_k\|^2 = \|\Delta\theta_k - \Delta\psi_k\|^2 + \|\Delta\theta_k\|^2 - 2\langle\Delta\theta_k, \Delta\theta_k - \Delta\psi_k\rangle,$$

and

$$\langle\Delta\theta_k, \Delta\psi_k\rangle = \|\Delta\theta_k\|^2 - \langle\Delta\theta_k, \Delta\theta_k - \Delta\psi_k\rangle.$$

Using these in (34) we obtain,

$$\frac{1}{2}\|a\theta_{k+1}+b\psi_{k+1}\|^2-\frac{1}{2}\|a\theta_k+b\psi_k\|^2=\left(\frac{a^2}{2}+\frac{b^2}{2}+ab-\frac{a}{\gamma}-\frac{b}{\gamma}\right)\|\Delta\theta_k\|^2 + \left(\frac{b^2}{2}-\frac{b}{\gamma}\right)\|\Delta\theta_k-\Delta\psi_k\|^2+\left(-b^2-ab+\frac{a}{\gamma}+\frac{2b}{\gamma}\right)\langle\Delta\theta_k,\Delta\theta_k-\Delta\psi_k\rangle.$$

We then simplify the factors using the identity $a+b=\alpha$, as well as $\frac{a^2}{2}+\frac{b^2}{2}+ab=\frac{1}{2}(a+b)^2=\frac{\alpha^2}{2}$, and $-b^2-ab=-\alpha/\beta$ which yields,

$$\begin{aligned}&\frac{1}{2}\|a\theta_{k+1}+b\psi_{k+1}\|^2-\frac{1}{2}\|a\theta_k+b\psi_k\|^2\\&=\left(\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}\right)\|\Delta\theta_k\|^2+\left(\frac{1}{2\beta^2}-\frac{1}{\gamma\beta}\right)\|\Delta\theta_k-\Delta\psi_k\|^2+\left(-\frac{\alpha}{\beta}+\frac{\alpha}{\gamma}+\frac{1}{\gamma\beta}\right)\langle\Delta\theta_k,\Delta\theta_k-\Delta\psi_k\rangle.\end{aligned} \tag{35}$$

We can finally combine (33) and (35),

$$\begin{aligned}\mathcal{E}_{k+1}-\mathcal{E}_k\leq&\left(\frac{\mu L_{\nabla\mathcal{J}}}{2}+\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}\right)\|\Delta\theta_k\|^2+\left(\frac{1}{2\beta^2}-\frac{1}{\gamma\beta}\right)\|\Delta\theta_k-\Delta\psi_k\|^2\\&+\left(-\frac{\mu}{\gamma\beta}+\frac{1+\alpha\beta-\gamma\alpha}{\gamma\beta}\right)\langle\Delta\theta_k,\Delta\theta_k-\Delta\psi_k\rangle.\end{aligned} \tag{36}$$

Then, $\mu=1+\alpha\beta-\gamma\alpha$ is specifically chosen so that the last term in (36) vanishes, so,

$$\mathcal{E}_{k+1}-\mathcal{E}_k\leq\left(\frac{\mu L_{\nabla\mathcal{J}}}{2}+\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}\right)\|\Delta\theta_k\|^2+\left(\frac{1}{2\beta^2}-\frac{1}{\gamma\beta}\right)\|\Delta\theta_k-\Delta\psi_k\|^2. \tag{37}$$

To prove the decrease of $(\mathcal{E}_k)_{k\in\mathbb{N}}$, it remains to justify that both factors in (37) are negative. First, the condition $\gamma<2\beta$ in (18) is equivalent to $\frac{1}{2\beta^2}-\frac{1}{\gamma\beta}<0$. Then,

$$\begin{aligned}&\frac{\mu L_{\nabla\mathcal{J}}}{2}+\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}<0\iff\mu L_{\nabla\mathcal{J}}\gamma+\alpha^2\gamma-2\alpha<0\\\iff&-\alpha L_{\nabla\mathcal{J}}\gamma^2+\left(\alpha^2+(1+\alpha\beta)L_{\nabla\mathcal{J}}\right)\gamma-2\alpha<0.\end{aligned}$$

A simpler sufficient condition for this to hold is $\left(\alpha^2+(1+\alpha\beta)L_{\nabla\mathcal{J}}\right)\gamma-2\alpha<0$ which is equivalent to $\gamma<2\alpha/\left(\alpha^2+(1+\alpha\beta)L_{\nabla\mathcal{J}}\right)$, and the latter holds true according to (18). So the sequence $(\mathcal{E}_k)_{k\in\mathbb{N}}$ is a decreasing. It is also lower-bounded since $\mathcal{J}$ is lower-bounded, so it converges.

It remains to prove the second part of the lemma. Let $K\in\mathbb{N}$, we sum (37) from 0 to K.

$$\sum_{k=0}^{K}\mathcal{E}_{k+1}-\mathcal{E}_k\leq\left(\frac{\mu L_{\nabla\mathcal{J}}}{2}+\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}\right)\sum_{k=0}^{K}\|\Delta\theta_k\|^2+\left(\frac{1}{2\beta^2}-\frac{1}{\gamma\beta}\right)\sum_{k=0}^{K}\|\Delta\theta_k-\Delta\psi_k\|^2.$$

The left-hand side is a telescopic series. As for the right-hand side, from (32) it holds that for all $k\in\mathbb{N}$, $\Delta\theta_k-\Delta\psi_k=-\gamma\beta\nabla\mathcal{J}(\theta_k)$. So,

$$\mathcal{E}_{K+1}-\mathcal{E}_0\leq\left(\frac{\mu L_{\nabla\mathcal{J}}}{2}+\frac{\alpha^2}{2}-\frac{\alpha}{\gamma}\right)\sum_{k=0}^{K}\|\Delta\theta_k\|^2+\left(\frac{\gamma^2}{2}-\gamma\beta\right)\sum_{k=0}^{K}\|\nabla\mathcal{J}(\theta_k)\|^2.$$

Denote finally, $C_1 = -\left(\frac{\mu L_{\nabla \mathcal{J}}}{2} + \frac{\alpha^2}{2} - \frac{\alpha}{\gamma}\right) > 0$ and $C_2 = -\left(\frac{\gamma^2}{2} - \gamma\beta\right) > 0$, it follows that,

$$C_1 \sum_{k=0}^{K} \|\theta_{k+1} - \theta_k\|^2 + C_2 \sum_{k=0}^{K} \|\nabla \mathcal{J}(\theta_k)\|^2 \leq \mathcal{E}_0 - \mathcal{E}_{K+1}.$$

And the right-hand side is finite since $(\mathcal{E}_k)_{k\in\mathbb{N}}$ converges. □

We also need the following lemma.

Lemma 16 (Lange (2013, Proposition 12.4.1))**.** *If a bounded sequence $(u_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^P$ satisfies,*

$$\lim_{k\to+\infty} \|u_{k+1} - u_k\| = 0,$$

then the set of accumulation points of $(u_k)_{k\in\mathbb{N}}$ is connected. If this set is finite then it reduces to a singleton and $(u_k)_{k\in\mathbb{N}}$ converges.

We can now prove Theorem 8.

Proof of Theorem 8. Assume that $\alpha > 0$. Let $(\theta_0, \psi_0) \in \mathbb{R}^P \times \mathbb{R}^P$, and consider INNA with step-size $\gamma > 0$ such that (18) holds. Denote $(\theta_k, \psi_k)_{k\in\mathbb{N}}$ the iterations of INNA initialized at (θ_0, ψ_0). By direct application of Lemma 15, and in particular due to (31), $\sum_{k=0}^{K} \|\nabla \mathcal{J}(\theta_k)\|^2 < +\infty$ so $\lim_{k\to+\infty} \|\nabla \mathcal{J}(\theta_k)\|^2 = 0$ and similarly, $\lim_{k\to+\infty} \|\theta_{k+1} - \theta_k\|^2 = 0$. From (16), we also have,

$$\|(\alpha - \frac{1}{\beta})\theta_k + \frac{1}{\beta}\psi_k\|^2 = \frac{1}{\gamma^2}\|\psi_{k+1} - \psi_k\|^2 \leq \frac{2}{\gamma^2}\|\theta_{k+1} - \theta_k\|^2 + 2\beta^2\|\nabla \mathcal{J}(\theta_k)\|^2 \xrightarrow[k\to\infty]{} 0. \quad (38)$$

From Lemma 15, the sequence $(\mathcal{E}_k)_{k\in\mathbb{N}}$ defined in (30) converges. Let $k \in \mathbb{N}$, since $\mathcal{E}_k = (1 + \alpha\beta - \gamma\alpha)\mathcal{J}(\theta_k) + \frac{1}{2}\|(\alpha - \frac{1}{\beta})\theta_k + \frac{1}{\beta}\psi_k\|^2$ and we just proved in (38) that the second term in $\mathcal{E}_k$ converges, this implies that $(\mathcal{J}(\theta_k))_{k\in\mathbb{N}}$ converges as well.

So far we proved the first part of the theorem. In particular since $\lim_{k\to+\infty} \|\nabla \mathcal{J}(\theta_k)\|^2 = 0$ and $\nabla \mathcal{J}$ is continuous, any accumulation point (i.e., the limit of any sub-sequence of iterates) is critical. We now assume that the critical points are isolated (Assumption 1) and that the sequence $(\theta_k)_{k\in\mathbb{N}}$ is uniformly bounded on $\mathbb{R}^P$.

According to Lemma 16, since $(\theta_k)_{k\in\mathbb{N}}$ is bounded and $\lim_{k\to+\infty} \|\theta_{k+1} - \theta_k\| = 0$, the set of accumulation points of $(\theta_k)_{k\in\mathbb{N}}$ is connected. Yet, accumulation points are critical points, which are assumed to be isolated. So the set of accumulation points of $(\theta_k)_{k\in\mathbb{N}}$ reduces to a singleton and using again Lemma 16, $(\theta_k)_{k\in\mathbb{N}}$ converges to a critical point. □