Chapter 2

The Iterative Method seen as an Ordinary Differential Equation

2.1 Motivation

Chapter 1 introduced the gradient descent method for unconstrained optimization (1.18), as well as a number of methods for constrained optimization, such as the penalty method of (1.31), the multiplier method in (1.40), the Arrow-Hurwicz method in (1.42) and the projection method of (1.39). All of these algorithms are in general of the form:

$$\theta_{n+1} = \theta_n + \epsilon_n \, d(\theta_n). \tag{2.1}$$

This type of recursive algorithm is very useful in numerical analysis, computer science and adaptive learning algorithms. In the sequel, we will be studying algorithms that have a stochastic direction $d(\theta_n)$, and to analyse them we will use some of the concepts that we will introduce in this Chapter. We start with constant step size sequence, for ease of presentation.

To set the stage, we illustrate the issues with a simple example. Suppose that we wish to solve

$$\min_{\theta \in \mathbb{R}^2} J(\theta) = \min_{\theta^\top = (\theta_1, \theta_2)} \left\{ 2\theta_1^2 + \theta_2^2 \right\}. \tag{2.2}$$

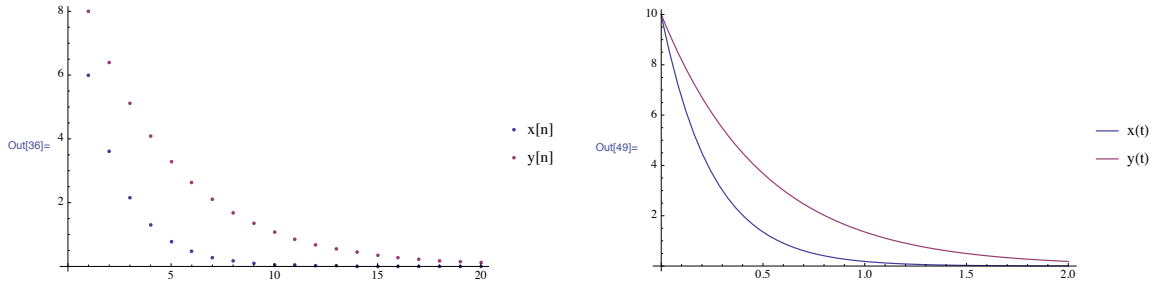
According to the results in Chapter 1 the gradient-based algorithm

$$\theta_{n+1} = \theta_n - \epsilon \nabla J(\theta_n)^\top = \theta_n - \epsilon \begin{pmatrix} 4\theta_{n,1} \\ 2\theta_{n,2} \end{pmatrix}, \tag{2.3}$$

with $\theta_n^\top = (\theta_{n,1}, \theta_{n,2})$, will approximate the solution.

In the following we will analyze the properties of the sequence of points generated by recursions such as (2.3). That (2.3) is motivated by an optimization problem is not of importance for this analysis and we will change the notation of the variables to x and y, respectively, when studying the convergence properties. We now let $x_n = \theta_{n,1}$ and $y_n = \theta_{n,2}$. Plotting consecutive values of the sequence (2.3):

$$\begin{aligned} x_{n+1} &= x_n - \epsilon \, 4x_n \\ y_{n+1} &= y_n - \epsilon \, 2y_n \end{aligned} \tag{2.4}$$



(a) The sequences $x(n), y(n)$. (b) Limit functions $x(t), y(t)$.

Figure 2.1: Visualizing the algorithm.

we obtain Figure 2.7(a), using $\epsilon = 0.1$ for initial values $x(0) = y(0) = 10$. The count number on the x-axis is the iteration number. If we plot the same figure using closer points (i.e. smaller ϵ), our mind will interpolate the plots and interpret the graph as a *function* rather than a sequence of points.

Figure 2.1(b) shows the plots of the continuous functions

$$x(t) = x(0)e^{-4t}$$
$$y(t) = y(0)e^{-2t}$$

and they are "very close" to the interpolation of the dots. How are these two processes related?

Differentiating w.r.t. t, the functions $x(t), y(y)$ satisfy the ordinary differential equations:

$$\frac{dx(t)}{dt} = -4x(t)$$
$$\frac{dy(t)}{dt} = -2y(t).$$

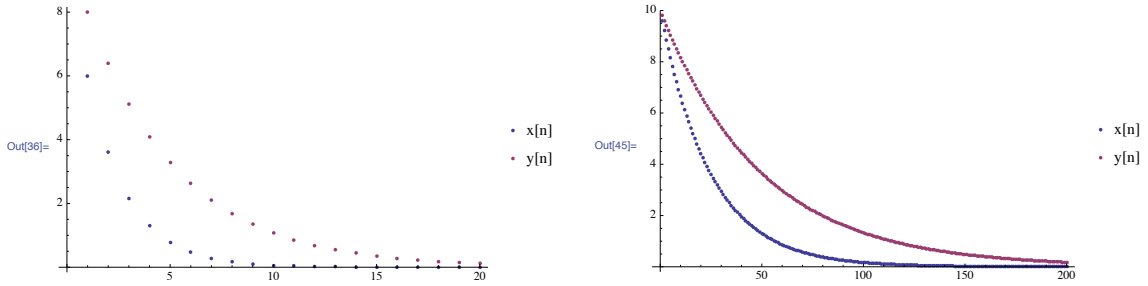
To "solve" numerically an ODE over the interval $(0, T)$, Euler proposed the following method. First choose a grid of equally spaced points with spacing $\epsilon > 0$. Call the points $t_n = \epsilon\, n$, for $0 \leq n \leq T/\epsilon$. The total number of points in the interval is inversely proportional to the sub-interval size ϵ. Next, use the finite difference approximation of a derivative at each point t_n to express:

$$\frac{x(t_{n+1}) - x(t_n)}{\epsilon} \approx -4x(t_n), \quad \text{and} \quad \frac{y(t_{n+1}) - y(t_n)}{\epsilon} \approx -2y(t_n).$$

The resulting sequences approximate the continuous function and satisfy exactly the recursions (2.4). Figure 2.2 compares the plots of (2.4) using $\epsilon = 0.1$ and a smaller $\epsilon = .01$. As ϵ decreases the plots look more and more similar to Figure 2.1(b).

Iteration count is related to the computing time required to execute the sequential algorithm (2.4) and therefore it is very important in analyzing the performance of algorithms. If we are only interested in the limit as $n \to \infty$, then we can use the results of Chapter 1. In this chapter we explore further and we will provide a framework to identify not just the limit, but the behavior of the algorithm as a dynamic process. The corresponding ODE is called the *limit process* of the algorithm, where the limit is taken with respect to ϵ.

REMARK. Any function $d(\cdot)$ such that the recursion (2.1) converges to the solution in the original optimization problem in (2.2) is a possible choice, with the negative of the gradient being a natural



(a) The sequences $x(n), y(n)$ for $\epsilon = 0.1$.

(b) The sequences $x(n), y(n)$ for $\epsilon = 0.01$.

Figure 2.2: Visualizing convergence.

choice for $d(\cdot)$. However, inspecting the behavior of the algorithm in the above example it is apparent that the first component approaches zero (which is the solution to (2.2)) faster than the second (actually exactly twice as fast). On occasion it is possible to first analyse the ODE and tweak the drift $d(\cdot)$ to achieve faster convergence. In this simple illustration, if instead of $-\nabla J^\top$ we use the scaled vector $d(x, y) = (-4x, -4y)^\top$ then both equations will follow the same trajectory and it is in principle possible to achieve accuracy in both components faster. Observe that such scaling preserves the descent direction of the algorithm, but sometimes it is more convenient to use a "surrogate" gradient than the actual one. Naturally the behavior of the ODE (and in consequence of the algorithm) will also depend on the initial condition. When we study the problem in this manner, we call the desired trajectory the *target* ODE and then build the algorithm from it. These are concepts that will help us establish the *relative efficiency* of algorithms, a topic that we will discuss in Chapter 5 for the more general stochastic approximation algorithms.

This chapter contains a summary of the theoretical tools required to correctly define "close to" for sequences and functions, and the notion of "convergence". In addition we present the full proofs of convergence and the methodology for analyzing the behavior of the (ϵ-) limit ODEs.

2.2 Stability of ODE's

We are concerned with the characterisation of the *limit points* of dynamic systems governed by ODE's, because these will often describe the convergence of numerical recursive algorithms of interest. The following theorem is a standard result in ODE literature and we present it without proof.

Theorem 2.1 *Let $T \subset \mathbb{R}^+$, with $0 \in T$, be a time interval of interest. Suppose that $G\colon T \times \mathbb{R}^N \to \mathbb{R}^N$ is piecewise continuous and satisfies the Lipschitz condition:*

$$\|G(t, x) - G(t, y)\| \leq L\|x - y\|,$$

for $L < \infty$, and that $\sup_{t\in T} \|G(t, x_0)\| < \infty$, for any initial condition x_0. Then

$$\frac{dx(t)}{dt} = G(t, x(t)); \quad x(0) = x_0$$

has a unique solution over $t \in T$. Moreover, the solution $x(t)$ is Lipschitz continuous on T.

For each initial condition $x(0) \in \mathbb{R}^N$ the function $x(t)$ that solves the ODE is called a *trajectory* of the ODE. The notion of a *vector field* can be used to understand the dynamic behaviour of the successive iterations by visualising the trajectories as well as their speed. A *vector field* on $\mathbb{R}^d$ is a mapping that assigns a vector $v(x) \in \mathbb{R}^d$ to each point $x \in \mathbb{R}^d$. Figure 2.3 shows an example of a vector field: the vector $v(\theta)$ is shown at each point $\theta \in \mathbb{R}^2$. The size of the arrow is proportional to the magnitude of $v(\theta)$. Such representations are useful to visualize the "drift" of recursive equations. The vector field associated with recursion (2.1) is defined by the "drift" mapping $d(\theta)$. For example, in the recursive description (2.3) we have $d(\theta) = (4\theta_1, \theta_2)^\top$. Successive values of numerical algorithms can be studied as a *dynamical system* that evolves in "time" (iteration number).

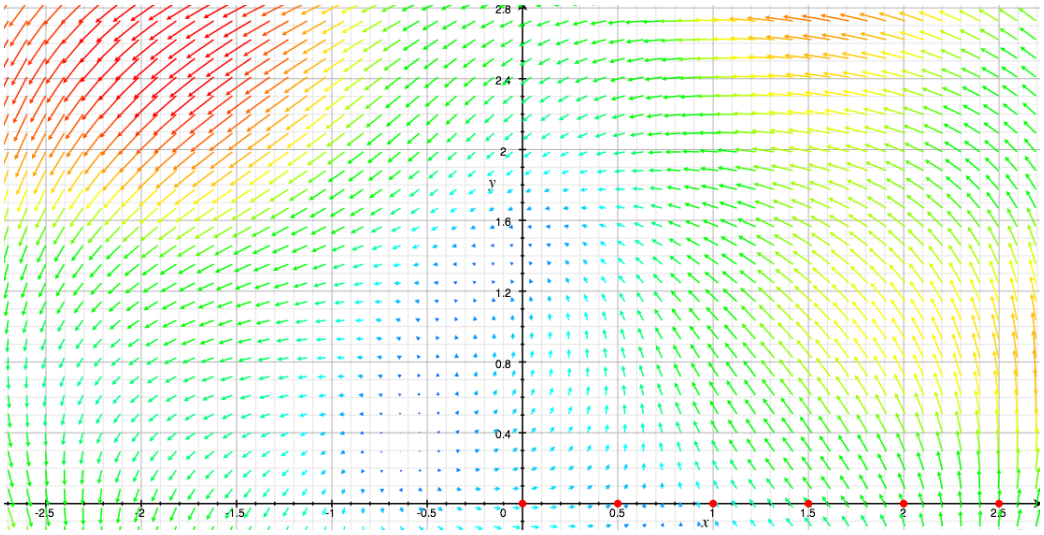


Figure 2.3: Example of a vector field

The function $G(\cdot)$ in Theorem 2.1 is a *vector field*, called the "drift" of the dynamical system. When possible, it can be plotted to help visualize the dynamics of trajectories, as illustrated in Figure 2.3. Note that the vector field displayed in Figure 2.3 is not the one given in (2.3).

For the following definition, recall that a mapping $f : \mathbb{R}^N \to \mathbb{R}^N$ is said to be locally Lipschitz continuous if for any x there exits a neighborhood U_x so that f is Lipschitz continuous on U_x.

Definition 2.1 *Let $G\colon \mathbb{R}^N \to \mathbb{R}^N$ be locally Lipschitz continuous. A point $\bar{x}$ is called an* equilibrium (or stationary) point *of the* autonomous *ODE:*

$$\frac{dx(t)}{dt} = G(x(t)) \tag{2.5}$$

if $G(\bar{x}) = 0$. The interpretation is that when $x(0)$ is an equilibrium point, then the dynamical system does not move: it remains at the equilibrium.

Definition 2.2 *An equilibrium point $\bar{x}$ of (2.5) is said to be:*

- stable *if $\forall \epsilon > 0\ \exists\, \delta > 0$ such that $\|x(0) - \bar{x}\| < \delta \Rightarrow \|x(t) - \bar{x}\| < \epsilon$ for all $t \geq 0$,*
- asymptotically stable *if it is stable and, in addition, $\exists \delta > 0$ such that if $\|x(0) - \bar{x}\| < \delta$ then $\lim_{t\to\infty} x(t) = \bar{x}$,*

- unstable *if it is not stable.*

Asymptotic stability is a desired property. Once the ODE comes close to $\bar{x}$ it will eventually reach $\bar{x}$. The domain of attraction of $\bar{x}$ is then at least $\{x : ||x - \bar{x}|| \leq \delta\}$.

EXAMPLE 2.1. Consider a linear drift with a symmetric $(d \times d)$ matrix $\mathbb{A}$ of full rank, that is,

$$\frac{dx(t)}{dt} = \mathbb{A}x(t). \tag{2.6}$$

Because the matrix is full rank, the only solution to $\mathbb{A}x = 0$ is the zero vector, which shows that the origin is the unique equilibrium point of the above ODE. Recall that all eigenvalues, $\rho_k, 1 \leq k \leq d$, of a symmetric matrix are real-valued. If, in addition, the matrix is full rank, then eigenvectors v_k corresponding the eigenvalues ρ_k are orthogonal and span $\mathbb{R}^d$. Specifically, $\mathbb{A}v_k = \rho_k v_k$, and for each point $\bar{x} \in \mathbb{R}^d$ there are unique coefficients α_i such that $\bar{x} = \sum_{k=1}^d \alpha_k \, v_k$. The coefficients $(\alpha_k, k = 1, \ldots, d)$ are also called the *coordinates* of x in the orthogonal basis $\{v_k\}$.

Express $x(t)$ in this coordinate system we set $x(t) = \sum_{k=1}^d \alpha_k(t) \, v_k$, so that

$$\frac{dx(t)}{dt} = \frac{d}{dt} \sum_{k=1}^d \alpha_k(t) \, v_k = \sum_{k=1}^d \frac{d}{dt} \alpha_k(t) \, v_k$$

and

$$\mathbb{A}x(t) = \mathbb{A} \left(\sum_{k=1}^d \alpha_k(t) \, v_k \right) = \sum_{k=1}^d \alpha_k(t) \rho_k \, v_k.$$

By (2.6) this yields

$$\sum_{k=1}^d \left(\frac{d\alpha_k(t)}{dt} - \alpha_k(t) \rho_k \right) \, v_k = 0.$$

Because the eigenvectors are orthonormal they are linearly independent, which implies that for each component k it holds

$$\frac{d}{dt} \alpha_k(t) = \alpha_k(t) \rho_k,$$

which is a one-dimensional ODE with solution $\alpha_k(t) = \alpha_k(0) \, e^{\rho_k t}$. Therefore, the (unique) equilibrium point $\bar{x} = 0$ is

- stable $\iff \max_k \rho_k \leq 0$. Indeed, assume $\max \rho_k = 0$, then at least one eigenvalue has real part null, in which case for each k either $\alpha_k(t) = \alpha_k(0)$, or $\alpha_k(t) = \alpha_k(0) e^{-|\rho_k| \, t}$. Thus if $\|x(0) - \bar{x}\| \leq \epsilon$, then $\|x(t) - \bar{x}\| \leq \epsilon$;
- asymptotically stable $\iff \max_k \rho_k < 0$, so all coordinate processes satisfy $\alpha_k(t) = \alpha_k(0) e^{-|\rho_k| t}$. Therefore $x(t) \to \bar{x} = 0$;
- unstable $\iff \max_k \rho_k > 0$. In this final case, all coordinate processes satisfy $\alpha_k(t) = \alpha_k(0) \, e^{\rho_k \, t}$, which means that they diverge to infinity.

⁂

As detailed in the above example for a linear problem, the eigenvalues of the matrix defining the drift characterize the type of stability. This motivates the following definition.

Definition 2.3 *A matrix $\mathbb{A}$ is called a* Hurwitz matrix *if the maximum of the real parts of its eigenvalues is strictly negative.*

In the following we will study the stability of an ODE around a stationary point via a linearisation of the vector field around the stationary point. Let w^* be a stationary point of the vector field G, so that $G(w^*) = 0$, and consider the ODE

$$\frac{dw(t)}{dt} = G(w(t)).$$

Provided that G is twice continuously differentiable, using a Taylor expansion of G at w^* yields,

$$G(w) = G(w^*) + \nabla G(w^*)(w - w^*) + \mathcal{O}(\|w - w^*\|^2).$$

Close to the stationary point, we approximate the behaviour of $w(t)$ by the behaviour of the linearised ODE:

$$\frac{dw(t)}{dt} = \mathbb{A}(w(t) - w^*), \tag{2.7}$$

where $\mathbb{A} = \nabla G(w^*)^T$ is the gradient of the vector field, evaluated at the stationary point, for which $G(w^*) = 0$. The linear system has the solution $w(t) = w^* + e^{\mathbb{A}t}(w(0) - w^*)$, so that the solution $w(t)$ converges to w^* if all the eigenvalues of $\mathbb{A}$ have strictly negative real part; that is, if $\mathbb{A}$ is Hurwitz then w^* is asymptotically stable.

EXAMPLE 2.2. Consider the following problem:

$$\min_{\theta=(\theta_1,\theta_2)^\top \in \mathbb{R}^2} J(\theta) \stackrel{\text{def}}{=} 2\theta_1^2 + \theta_2^2,$$

with solution at the origin. Let us analyse the behaviour of the gradient-driven ODE:

$$\frac{dx(t)}{dt} = -(\nabla J(x(t)))^\top.$$

The gradient and Hessian at a point $\theta = (\theta_1, \theta_2)^\top \in \mathbb{R}^2$ are given by:

$$\nabla J(\theta) = \begin{pmatrix} 4\,\theta_1, 2\,\theta_2 \end{pmatrix}$$

and

$$\nabla^2 J(\theta) = \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}.$$

We can verify that the vector field $G = -\nabla J$ here has a unique equilibrium point at the origin (the same as the stationary point for the optimisation problem). The eigenvalues ρ of $\mathbb{A}$, with $\mathbb{A} = -\nabla^2 J(\theta)$, satisfy $(-4 - \rho)(-2 - \rho) = 0$, so that $\rho_1 = -2$ and $\rho_2 = -4$ are negative, which implies that the origin is asymptotically stable. Moreover, as the ODE is linear, all the trajectories $x(t)$ move exponentially fast to the origin, regardless of the initial value $x(0)$. The vector field for the gradient search algorithm is shown in Figure 2.4. The corresponding trajectories are those shown in Figure 2.1(b).

⁂

EXAMPLE 2.3. Consider now the problem:

$$\min_{\theta=(\theta_1,\theta_2)^\top \in \mathbb{R}^2} J(\theta) \stackrel{\text{def}}{=} 2\theta_1^2 + \theta_1^2\theta_2^2,$$

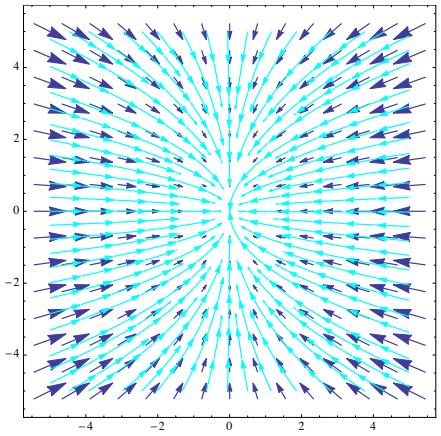


Figure 2.4: Stability for unconstrained optimisation, $\mathbb{A}$ Hurwitz, Example 2.2

with solution at the origin. Let us analyze the behavior of $dx(t)/dt = -(\nabla J(x(t)))^\top$. In this case the gradient and Hessian at $\theta \in \mathbb{R}^2$ are given by:

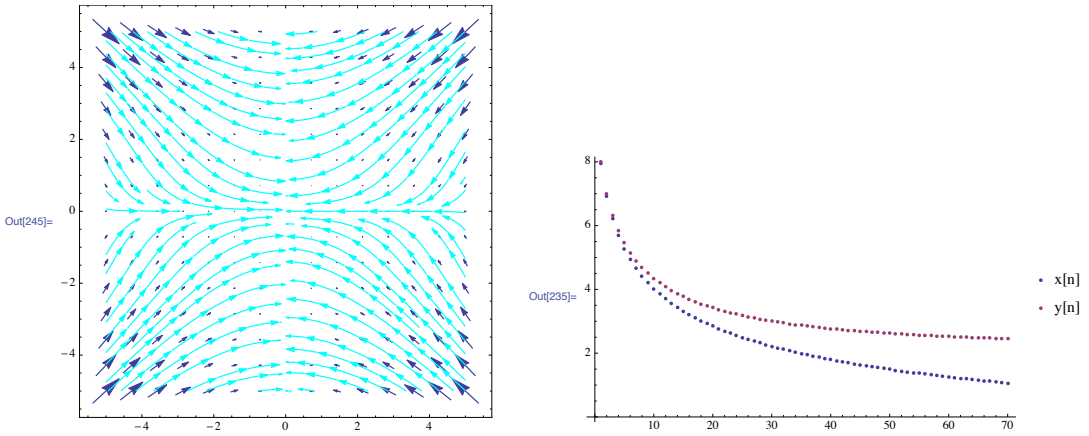
$$\nabla J(\theta) = \begin{pmatrix} 4\theta_1 + 2\theta_1\theta_2^2 \\ 2\theta_1^2\,\theta_2 \end{pmatrix}, \quad \mathbb{A} = -\nabla^2 J(0) = \begin{pmatrix} -4 & 0 \\ 0 & 0 \end{pmatrix}.$$

The eigenvalues ρ of $\mathbb{A}$ satisfy $(-4-\rho)(-\rho) = 0$, so that $\rho_1 = -4, \rho_2 = 0$ and $\mathbb{A}$ fails to be Hurwitz. The vector field is shown in Figure 2.5(a) with some trajectories streamlined for easy visualization. Every point with $\theta_1 = 0$ is a stationary point of the ODE. Therefore, the limit point of a trajectory $x(t)$ may depend on the initial condition. Notice that although there is a unique value for the minimum $J(\theta^*) = 0$, the set of optimal values $\{\theta \in \mathbb{R}^2 \colon J(\theta) = 0\}$ is not unique: all stationary points are minimisers. Thus the ODE will drive the trajectories towards optimality, but the limit points depend on the initial condition $x(0)$. Put differently, different initial conditions will yield different limit points. In addition, the vector field is relatively much weaker as it approaches the stability region. Figure 2.5(b) plots the trajectory of the corresponding approximation for the initial point $(10, 10)$. It shows that many more iterations would be required to get reasonably close to the limit, as expected.

⁂

The two examples above illustrate cases of the limit behaviour associated with a gradient-driven ODE, and they illustrate how the ODE trajectories can provide insight into the iterative gradient descent algorithm. We will present an example later on where the trajectories exhibit a more complex behaviour around the stable point. Specifically, we will show an example where the stable point is reached via spiral trajectories.

REMARK. To link this material with Chapter 1, consider $J \in C^2$ as a function that we wish to minimize (without constraints), and let x^* be such that the Hessian of J at x^* is positive definite. Then, the function J behaves in a neighborhood of x^* like a convex function and it is conceivable that any descent direction algorithm, once entering the neighborhood of x^*, will be attracted to the stationary point x^*. Observe that in the special case that the descent direction is given by the negative gradient we have $\mathbb{A} = -\nabla^2 J(x^*)$ in (2.7), so that the positive definiteness condition on the Hessian $\nabla^2 J(x^*)$ (introduced in Chapter 1 as the second order condition for optimality) is equivalent to $\mathbb{A}$ being Hurwitz.



(a) Vector field with trajectories.

(b) A given trajectory with $\epsilon = 0.001$.

Figure 2.5: Vector field and approximation, Example 2.3

2.3 ODE limit of recursive algorithms

The behaviour of algorithms will be studied by establishing convergence in the functional space of algorithm trajectories to solutions of ODE's. Consider a recursive equation of the form:

$$\theta_{n+1} = \theta_n + \epsilon_n \, G(\theta_n), \quad \theta \in \mathbb{R}^d, \tag{2.8}$$

where the function $G\colon \mathbb{R}^d \to \mathbb{R}^d$ is bounded and continuous. This formulation covers all the gradient based optimisation methods for unconstrained as well as constrained problems that we mentioned in Chapter 1. Although our notation is suggestive of a "gradient G", we will see that there is no need to restrict the analysis to the (simplest) case of unconstrained optimisation of a convex function. Here we will establish the limiting behaviour of the algorithm in general settings, which include algorithms for constrained optimisation, notably the Arrow-Hurwicz method (1.42) of Chapter 1. Importantly, we will give particular attention to the *constant gain* case, where $\epsilon_n \equiv \epsilon$ is a constant.

REMARK. Interestingly, in his book *Institutionum calculi integralis* (published between 1768 and 1770) Euler proposed to use (2.8) with constant step size ϵ to solve the differential equation (2.5). We proceed in the inverse way here, starting with a difference equation and establishing conditions under which a "limit" ODE can be used to assess the behaviour of successive iterations.

We wish to study the limit of the sequence of points θ_n given by (2.8) and compare it to the solution of an ODE. In order to do, so will first interpret the sequence θ_n as mapping of t and thus obtaining an interpolation process to be defined formally in the following.

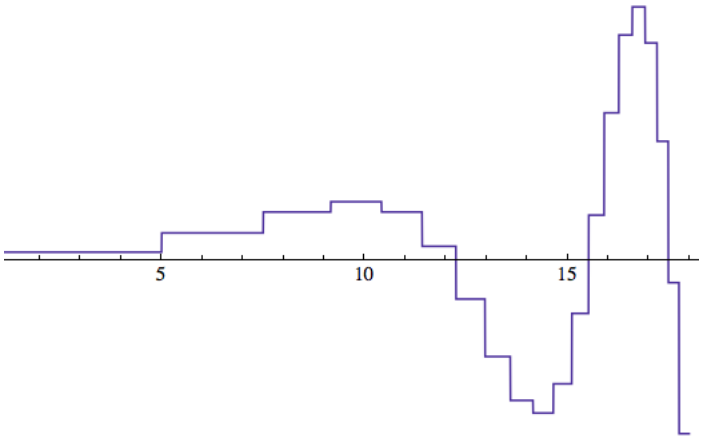
Definition 2.4 *For any sequence* $\boldsymbol{\epsilon} = \{\epsilon_n\}$, *with* $\epsilon_n > 0$, *the* interpolation process $\vartheta^{\boldsymbol{\epsilon}}(\cdot)$ of the recursion put forward in (2.8) *is defined for* $t \in \mathbb{R}^+, \vartheta(0) = \theta_0$ *by*

$$\vartheta^{\boldsymbol{\epsilon}}(t) = \theta_{m(t)},$$

where

$$m(t) = \min\left\{ n\colon t_n \stackrel{\text{def}}{=} \sum_{i=0}^{n} \epsilon_i \geq t \right\}. \tag{2.9}$$

In case $\epsilon_n \equiv \epsilon$ we simply write $\vartheta^\epsilon(\cdot)$ for the interpolation process. Moreover, to simplify the notation we will suppress indexation of the stepsize sequence and write $\vartheta(\cdot)$ whenever this causes no confusion.



The step size here is $\epsilon_n = 5/(n+1)$ and first 20 values of the sequence $\{\theta_n\}$ are: 0.479, 1.682, 2.992, 3.637, 2.992, 0.846, -2.455, -6.054, -8.797, -9.589, -7.760, -3.352, 2.796, 9.197, 14.07, 15.829, 13.574, 7.418, -1.427, -10.880.

Figure 2.6: The interpolated process $\vartheta^{\boldsymbol{\epsilon}}(t) = \theta_{m(t)}$.

To study convergence of the interpolation process, we introduce in the following the concept of the limit of functions as elements of a sequence. For any $N \in \mathbb{N}$ and $a, b \in \mathbb{R}$, with a, b, call $F_N(a,b)$ the space of functions $f\colon [a,b] \to \mathbb{R}^N$. The sup-norm of a mapping $f \in F_N(a,b)$ is given by

$$||f||_\infty = \sup_{x\in[a,b]} \|f(x)\|.$$

EXAMPLE 2.4. This example illustrates that the limit of a pointwise convergent sequence of continuous functions may fail to be continuous. Consider the sequence of functions $f_n : [0,T] \to \mathbb{R}$:

$$f_n(x) = \begin{cases} 0 & 0 \le x < 1 - \frac{1}{n}, \\ nx + (1-n) & 1 - \frac{1}{n} \le x < 1, \\ 1 & 1 \le x \le T, \end{cases}$$

so each f_n is a continuous and piecewise linear function. For each $x < 1$, and all $n > 1/(1-x)$, $f_n(x) = 0$, so that $\lim_{n\to\infty} f_n(x) = 0$. As well, if $x \ge 1$ then $f_n(x) = 1$ for all n. This implies that $f_n(x)$ converges point-wise to the function:

$$f(x) = \begin{cases} 0 & 0 \le x < 1, \\ 1 & 1 \le x \le T, \end{cases}$$

which is **not** a continuous function. That is, f_n does not converge in the sup-norm.

⁂

In the above example, the "limiting" function contains a jump, so it's discontinuous. For our project we are faced with an even harder problem as we are studying the limit of interpolation processes which are piecewise constant functions and therefore not even continuous. The concept of equicontinuity in the extended sense, as introduced presently, is a generalization of uniform continuity and will provide sufficient conditions for convergence of our interpolation processes to a continuous limit, which is a technical result required for proving the main result of this chapter.

Definition 2.5 *The set $\mathcal{F} = \{f_n\} \subset F_N(a,b)$ is called* equicontinuous in the extended sense *if for each $\eta > 0$ there exists $\delta_\eta > 0$ such that for $r, q \in (a,b)$*

$$|r-q| < \delta_\eta \quad \textit{implies} \quad \limsup_{n\to\infty} \|f_n(r) - f_n(q)\|_\infty < \eta.$$

For a proof of the following theorem we refer to [10, 28].

Theorem 2.2 (Ascoli-Arzelà) *Let $\{f_n\} \in F_N(a,b)$ be a sequence of functions and assume that it is equicontinuous in the extended sense and uniformly bounded, that is, there exists $M < \infty$ such that $\|f_n(x)\|_\infty \le M$ for all n and $x \in (a,b)$. Then every sequence has a convergent subsequence, and all accumulation points are continuous functions on (a,b).*

The equicontinuity condition puts a bound on the modulus of continuity of the whole family of functions, effectively bounding the slopes or growth rate across the whole family. In our example above, the slope of the functions is n, which is not bounded, thus creating a discontinuity for the limit function, even when each of the functions in the set fails to be continuous. For more details on equicontinuity in the extended sense we refer to [21].

We now establish the main result of this chapter, which proves the existence of a unique solution of the ODE on finite intervals $[0,T]$ for arbitrary T. In the light of the above discussion this is a surprising result as, under the conditions put forward in the following theorem, the limit of a sequence of discontinuous functions (in this case the interpolation processes) is a continuous function.

Theorem 2.3 *Let $G\colon \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz continuous function and consider recursion (2.8):*

$$\theta_{n+1} = \theta_n + \epsilon \left(G(\theta_n) + \beta_\epsilon(\theta_n)\right),$$

with constant step size $\epsilon > 0$. Let $x_\epsilon(t) = \vartheta^\epsilon(t), 0 \le t \le T$, denote the interpolation process of $\{\theta_n\}$ on $[0,T]$, for $T > 0$, with $x_\epsilon(0) = \theta_0$.

If $\sup_\theta \|\beta_\epsilon(\theta)\| = \mathcal{O}(\epsilon)$, then the ϵ-indexed sequence of processes $\{(x_\epsilon(t); 0 \le t \le T) : \epsilon > 0\}$ converges as $\epsilon \to 0$ (in the sup norm) to the solution of the ODE:

$$\frac{dx(t)}{dt} = G(x(t)), \tag{2.10}$$

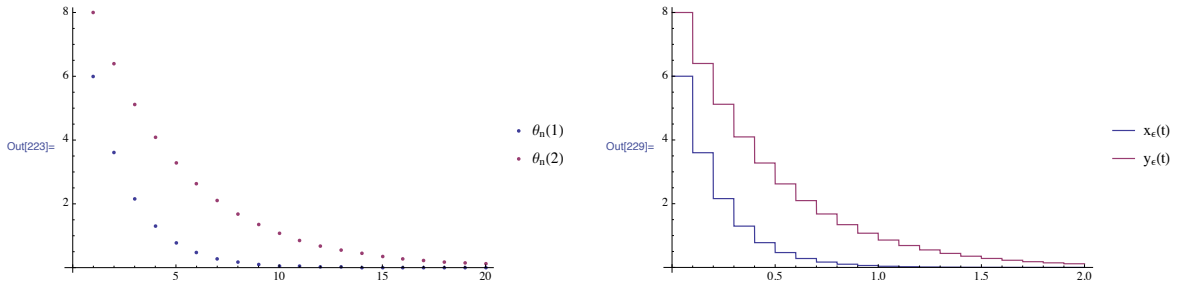
for $0 \le t \le T$.

The following proof is neither the simplest nor the most elegant one. The method of proof, however, will become an important tool when we analyse the stochastic version of the recursions (2.8). In the case of constant stepsize ϵ we have $t_n = n\epsilon$ and $m(t) = \lfloor t/\epsilon \rfloor$ is the integer part of t/ϵ.

Proof: The proof has the following main parts: (i) the integral representation, (ii) the 'equicontinuity in the extended sense' argument, and (iii) the compactness argument. For ease of presentation, we first prove the statement for the unbiased case, i.e., $\beta_n(\theta_n) = 0$ for all n.

We now turn to part (i) of the proof. First, from (2.8), we obtain

$$\theta_{n+m} - \theta_n = \sum_{i=n}^{n+m-1} (\theta_{i+1} - \theta_i) = \sum_{i=n}^{n+m-1} \epsilon\, G(\theta_i), \tag{2.11}$$



(a) The sequences θ_n. (b) The interpolations with $\epsilon = 0.1$.

Figure 2.7: The interpolation process $x_\epsilon(t)$.

which, using $x_\epsilon(t) = \theta_{m(t)}$ for storing the latest θ_n value just after t, is equivalent to:

$$x_\epsilon(t+s) - x_\epsilon(t) = \sum_{i=m(t)}^{m(t+s)-1} \epsilon\, G(\theta_i). \tag{2.12}$$

Because $x_\epsilon(\cdot)$ is piecewise constant, $G(x_\epsilon(\cdot))$ is also piecewise constant and its jump times are given by $\{t_n = n\epsilon, n \in \mathbb{N}\}$. Thus the definite integral on $[t, t+s]$ of $G(x_\epsilon(\cdot))$ is a sum that can be expressed as

$$\int_t^{t+s} G(x_\epsilon(u))\, du = \sum_{i=m(t)}^{m(t+s)-1} \epsilon\, G(\theta_i) + \rho(\epsilon),$$

where $\rho(\epsilon)$ is the error in the approximation, due to the discretisation at the end points.

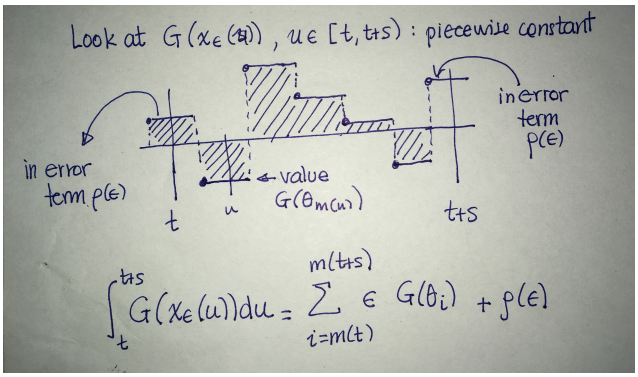


Figure 2.8: The sum and its approximation by an integral.

When both t and $t+s$ are multiples of ϵ, this approximation error is null (see Figure 2.8. Therefore,

$$x_\epsilon(t+s) - x_\epsilon(t) = \int_t^{t+s} G(x_\epsilon(u))\, du \; - \rho(\epsilon). \tag{2.13}$$

We now turn to part (ii): the proof of establishing equicontinuity of x_ϵ in the extended sense. Let

L denote the Lipschitz constant for G, then

$$\|G(\theta_{n+1}) - G(\theta_n)\|_\infty \le L\|\theta_{n+1} - \theta_n\|_\infty \overset{(2.8)}{\le} \epsilon\, L\|G(\theta_n)\|_\infty. \tag{2.14}$$

Thus by the triangle inequality $\|G(\theta_{n+1})\|_\infty \le \|G(\theta_n)\|_\infty(1+\epsilon\, L)$ and by induction, $\|G(\theta_{n+m})\|_\infty \le \|G(\theta_n)\|_\infty(1+\epsilon\, L)^m$. Therefore using $m(t) = \lfloor t/\epsilon \rfloor$, we have:

$$\|G(\theta_{m(T)-1})\|_\infty \le \|G(\theta_0)\|_\infty(1+\epsilon\, L)^{m(T)-1} \le \|G(\theta_0)\|_\infty(1+\epsilon\, L)^{T/\epsilon}. \tag{2.15}$$

Use now the fact that $\lim_{\epsilon\to 0}(1+\epsilon\, L)^{T/\epsilon} = e^{LT}$ to bound the terms with constant $\bar{G} = \|G(\theta_0)\|_\infty\, e^{LT}$, for sufficiently small ϵ.

Notice that the sum in (2.12) contains $m(q) - m(r) - 1$ terms. For ϵ sufficiently small, $m(r) \ge r/\epsilon$ and $m(q) \le q/\epsilon$, so that the number of terms is bounded by $(q-r)/\epsilon$. This yields, for small ϵ,

$$\|x_\epsilon(q) - x_\epsilon(r)\|_\infty = \left\| \sum_{i=m(r)}^{m(q)-1} \epsilon\, G(\theta_i) \right\|_\infty \le \epsilon\, \bar{G}\, \frac{(q-r)}{\epsilon} = \bar{G}\,(q-r), \tag{2.16}$$

To summarize, for ϵ sufficiently small, we have shown that for any $\eta > 0$, me may let $\delta_\eta = \eta/\bar{G}$, so that it follows that $\|x_\epsilon(q) - x_\epsilon(r)\|_\infty \le \eta$ whenever $|q-r| \le \delta_\eta$. This establishes equicontinuity in the extended sense. Indeed, for any family $\{\epsilon_k\}$ such that $0 \le \epsilon_k \le \epsilon$, for ϵ sufficiently small, we have $\limsup_{k\to\infty} \|x_{\epsilon_k}(q) - x_{\epsilon_k}(r)\|_\infty \le \eta$ whenever $|q-r| \le \delta_\eta$.

The remainder of the proof is devoted to part (iii): the compactness argument that will establish the validity of (2.10). The sequence $\{x_{\epsilon_k}(\cdot)\}$, as introduced above, is equicontinuous in the extended sense. Let $a < t$ and $b > t+s$ and consider $\{x_{\epsilon_k}(\cdot)\}$ on (a,b). Note that $x_{\epsilon_k}(0) = \theta_0$ for all k. Therefore, for ϵ sufficiently small, by (2.16)

$$\|x_{\epsilon_k}(r)\|_\infty \le \|\theta_0\|_\infty + r\,\bar{G},$$

for all $r > 0$, which suffices to show that x_ϵ is uniformly bounded in (a,b). This together with equicontinuity of $\{x_{\epsilon_k}(\cdot)\}$ implies by the Ascoli-Arzelà Theorem 2.2 that any infinite subsequence has a convergent subsequence with a continuous limit on (a,b). Consider a convergent subsequence along $\epsilon_r \to 0$, so that $\hat{x}(\cdot) = \lim_{r\to\infty} x_{\epsilon_r}(\cdot)$ (in the sup norm). Then,

$$\begin{aligned}
\lim_{r\to\infty}(x_{\epsilon_r}(t+s) - x_{\epsilon_r}(t)) &\overset{\text{(a)}}{=} \lim_{r\to\infty}\int_t^{t+s} G(x_{\epsilon_r}(u))du \\
&\overset{\text{(b)}}{=} \int_t^{t+s} \lim_{r\to\infty} G(x_{\epsilon_r}(u))du \\
&\overset{\text{(c)}}{=} \int_t^{t+s} G(\hat{x}(u))du,
\end{aligned}$$

where (a) follows from the fact that $\rho(\epsilon)$ in (2.13) is bounded by $\|\rho(\epsilon)\|_\infty \le 2\epsilon\bar{G}$ and thus of order $\mathcal{O}(\epsilon)$, (b) follows from Lebesgue Dominated Convergence Theorem, and (c) is a consequence of the continuity of $G(\hat{x}(\cdot))$ on (a,b). We arrive for $s > 0$ at

$$\frac{\hat{x}(t+s) - \hat{x}(t)}{s} = \frac{1}{s}\int_t^{t+s} G(\hat{x}(u))du.$$

By continuity of $G(\hat{x}(\cdot))$ on $[t, t+s]$, taking the limit as s goes to zero, the above right-hand side converges to $G(\hat{x}(t))$, which establishes (2.10) for $\hat{x}(\cdot)$. Because G is continuous and bounded on the trajectory $\hat{x}$, it follows from Theorem 2.1 that (2.10) has a unique solution for each initial condition, establishing that all accumulation points have the same limit, proving the claim for the unbiased case.

We now turn to the proof in case the bias term is present. The proof follows by noticing that the perturbations will appear in the part (i) of the above proof via

$$\int_t^{t+s} G(x_\epsilon(u))\, du = \sum_{i=m(t)}^{m(t+s)-1} \epsilon\, G(\theta_i) + \sum_{i=m(t)}^{m(t+s)-1} \epsilon\, \beta_\epsilon(\theta_i) + \rho(\epsilon),$$

and using the bound on the perturbations, for any $r < q$

$$\sum_{i=m(r)}^{m(q)-1} \epsilon\, \beta_\epsilon(\theta_i) = (q-r)\mathcal{O}(\epsilon)$$

so this term can be added to the approximation error $\rho(\epsilon)$ in (2.13) and the proof follows directly.

QED

Theorem 2.3 helps to establish the ϵ-limit of the algorithm as an ODE over any finite interval $[0, T]$. However for our purposes, we wish to know what is the limit as $t \to \infty$ in oder to verify that the algorithm converges to the correct limit point. The following theorem establishes the conditions under which we can extend the convergence on the whole real line $t \in [0, \infty)$. The key observation in the proof of convergence is that if G is bounded along the trajectory $x(t)$, then (2.16) holds and T needs not be finite.

Theorem 2.4 *Let $G\colon \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz continuous function and assume that the ODE (2.10):*

$$\frac{dx(t)}{dt} = G(x(t))$$

has bounded trajectories $x(t)$ for any $x_0 \in \mathbb{R}^d$. Consider recursion (2.8):

$$\theta_{n+1} = \theta_n + \epsilon \left(G(\theta_n) + \beta_\epsilon(\theta_n)\right),$$

with constant step size $\epsilon > 0$ and let $x_\epsilon(t) = \vartheta^\epsilon(t), 0 \le t$, denote the interpolation process of $\{\theta_n\}$, with $x_\epsilon(0) = \theta_0$.

If $\sup_\theta \|\beta_\epsilon(\theta)\| = \mathcal{O}(\epsilon)$, then the ϵ-indexed sequence of processes $\{(x_\epsilon(t) : t \ge 0) : \epsilon > 0\}$ converges as $\epsilon \to 0$ (in the sup norm) to the solution of the ODE (2.10) for all $t \ge 0$, and any accumulation point of θ_n is a asymptotically stable point of G.

Proof: Let $x(0) = \theta_0$ be given. By assumption, the trajectory $x(t)$ of (2.10) from this initial point is bounded, that is, there is a value C such that $|x(t)| \le C$. Define a "box" $H = \{x \in \mathbb{R}^d\colon x_i \in [-M, M]\}$, with $M = C + \sup_\theta \|\beta_\epsilon(\theta)\|$. Now consider the truncated version of $\theta_{n+1} = \theta_n + \epsilon \left(G(\theta_n) + \beta_\epsilon(\theta_n)\right)$ to H. As the update of the truncated algorithm is bounded by $\|G(\theta) + \beta_\epsilon(\theta)\|$, which in turn is is bounded on H, we may in the proof of Theorem 2.3 replace the bound (2.15) by a constant $\bar{G}$ independent of T. For ϵ sufficiently small, $x_\epsilon(t)$ stays inside the box H and the truncation becomes insubstantial. Elaborating on the argument of the proof of Theorem 2.3, it then follows that the algorithm converges to the solution of (2.10).

QED

Theorem 2.5 *For any $x \in \mathbb{R}^d$ let $\rho_k(x)$ denote the real part of the kth eigenvalue of $\nabla G(x)$. If $G \in \mathcal{C}^1$, and the following Hurwtiz condition*

$$\forall x \in \mathbb{R}^d \quad \max_k(\rho_k(x)) =: -\rho_{\max} \leq 0$$

holds, then G is bounded along trajectories $\{\theta_n\}$ as well as along the solution $\{x(t) : t \geq 0\}$ of the ODE $x'(t) = G(x(t))$.

Proof: Let θ_n be any point during the iteration of the recursion $\theta_{n+1} = \theta_n + \epsilon\, G(\theta_n)$. Using the mean value theorem,

$$G(\theta_{n+1}) = G(\theta_n) + \epsilon \nabla G(x)(\theta_{n+1} - \theta_n) = G(\theta_n) + \epsilon \nabla G(x) G(\theta_n),$$

for some x along the convex combination of θ_n and θ_{n+1}. Let v be any eigenvector of $\nabla G(x)$ with eigenvalue ρ. Projecting the difference along the direction v yields:

$$v^\top G(\theta_{n+1}) - v^\top G(\theta_n) = \epsilon\, v^\top \nabla G(x)\, G(\theta_n) = \epsilon\rho\, v^\top G(\theta_n).$$

Call $g_j = v^\top G(\theta_j) \in \mathbb{R}$. Then $g_{n+1} - g_n = \epsilon\rho\, g_n$, and as $\rho \leq \rho_{\max} < 0$, we have for ϵ sufficiently small that $0 < 1 + \epsilon\rho < 1$, which implies that $g_{n+1} = (1 + \epsilon\rho) g_n$ and thus $|g_{n+1}| \leq |g_n|$. Because this is true for all the directions that are eigenvectors of $\nabla G(x)$, it follows that the function $G(\theta_n)$ is non-increasing along the trajectory $\{\theta_n\}$ (for details see Example 2.1).

We now turn to the continuous version. Differentiation yields

$$\frac{d}{dt} G(x(t)) = \nabla G(x(t)) \frac{dx(t)}{dt} = \nabla G(x(t)) G(x(t)).$$

At $x = x(t)$, the rate of growth of each of the coordinates of $G(x)$ is bounded by $-\rho_{\max}$, so that if $y(t) = |G(x(t))|$, where $|\cdot|$ denotes the Euclidean norm, then

$$\frac{dy(t)}{dt} \leq -\rho_{\max}\, y(t),$$

and using Grönwall's lemma, $y(t) \leq y(0)\, e^{-\rho_{\max} t} \leq y(0)$. This implies that $G(x(t))$ is bounded for all $t \geq 0$, even if $G(x)$ itself is not bounded for all x.

QED

The result put forward in Theorem 2.4 can be extended to the case of decreasing step-sizes. We state below the result that can be proved using a simple adaptation of the proof of Theorem 2.3 and the details are left as exercise (see Exercise 2.3).

Theorem 2.6 *Let $G\colon \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz continuous function and consider recursion (2.8):*

$$\theta_{n+1} = \theta_n + \epsilon_n\, (G(\theta_n) + \beta_n(\theta_n)).$$

Assume $G(\theta)$ is bounded along tthe trajectory, and that $\sum_n \epsilon_n = \infty, \epsilon_n \to 0$, $\sum_n \epsilon_n |\beta_n(\theta_n)| < \infty$. Let $x_n(t) = \vartheta(t_n + t), t \geq 0$ denote the interpolation process of the shifted sequence $\{\theta_k\}$ with $x_n(0) = \theta_n$. Then, as $n \to \infty$, x_n converges (in the sup norm) to the solution of the ODE:

$$\frac{dx(t)}{dt} = G(x(t)).$$

We explain the use of Theorem 2.4 and the interpretation of the limit ODE with the following example of a constrained optimization problem. Moreover, the example illustrates the use of Lyapunov functions for establishing boundedness of the ODE and shows an alternative techique to Theorem 2.5.

EXAMPLE 2.5. Consider

$$\begin{aligned}\min_{\theta\in\mathbb{R}} J(\theta) &\stackrel{\text{def}}{=} \frac{1}{2}\theta^2\\ \text{s.t.}\quad c-\theta &\le 0,\end{aligned}$$

for $c>0$, and suppose that we use the Arrow-Hurwicz algorithm (1.42) with constant step size:

$$\begin{aligned}\theta_{n+1} &= \theta_n - \epsilon\left(\nabla_\theta \mathcal{L}(\theta_n,\lambda_n)\right)^\top\\ \lambda_{n+1} &= \max(0,\lambda_n+\epsilon g(\theta_n)).\end{aligned} \tag{2.17}$$

Here, the Lagrangian and constraint function becomes

$$\mathcal{L}(\theta,\lambda)=\frac{1}{2}\theta^2+\lambda(c-\theta);\quad g(\theta)=c-\theta.$$

The vector field is

$$G(\theta,\lambda)=\begin{pmatrix}-(\nabla_\theta\mathcal{L}(\theta,\lambda))^\top\\ g(\theta)\end{pmatrix}=\begin{pmatrix}-\theta+\lambda\\ c-\theta\end{pmatrix} \tag{2.18}$$

and by simple inspection, the solution to the constrained problem must satisfy $\theta^*=c$ and the corresponding Lagrange multiplier can be evaluated as the solution to the KKT point: $-\theta^*+\lambda^*=0$ so that $\lambda^*=c$.

In this case, the Hessian matrix related to the limit ODE in the region $\lambda>0$ is:

$$\mathbb{A}=\nabla_{(\theta,\lambda)}G(\theta,\lambda)=\begin{pmatrix}-1 & 1\\ -1 & 0\end{pmatrix}.$$

The eignevalues satisfy $(-1-\rho)(-\rho)+1=0$, with roots $\rho=-\frac{1}{2}\pm i\frac{\sqrt{3}}{2}$, and $G(\theta)$ there for satisfies the Hurwitz condition. By Theorem 2.5, the G is bounded along the trajectories.

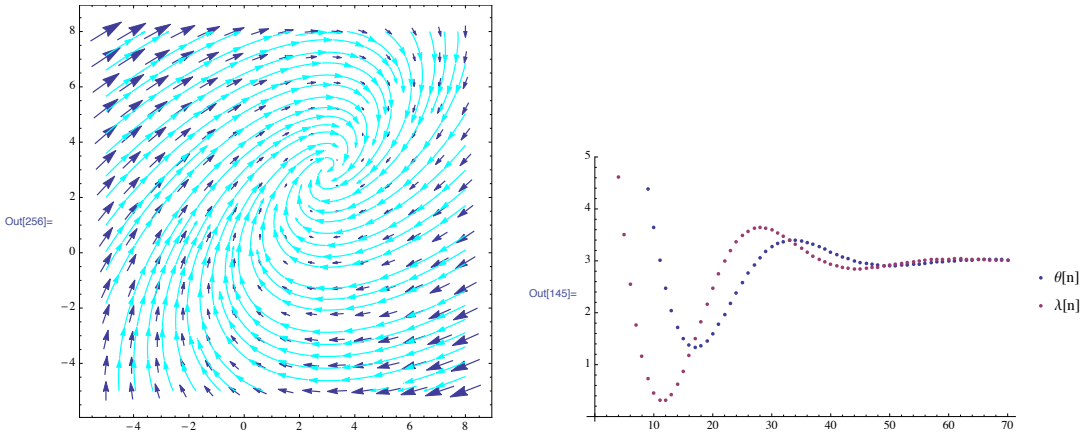
In the following we show how an "Lyapounov stability" can be applied for showing that for any finite $x(0)$, the corresponding trajectory $x(t)$ never leaves a compact set. Let $x(t)$ be a trajectory of the ODE (2.5) with vector field (2.18), i.e.,

$$\frac{dx_1(t)}{dt}=-x_1(t)+x_2(t)\quad\text{and}\quad\frac{dx_2(t)}{dt}=c-x_1(t),$$

and define the function $V(t)=\frac{1}{2}\left((x_1(t)-c)^2+(x_2(t)-c)^2\right)$ as the distance from the optimal value. Then:

$$\begin{aligned}\frac{d}{dt}V(t) &= (x_1(t)-c)\frac{dx_1}{dt}+(x_2(t)-c)\frac{dx_2}{dt}\\ &= (x_1(t)-c)(-x_1(t)+x_2(t))+(x_2(t)-c)(c-x_1(t))\\ &= -x_1^2(t)+2cx_1(t)-c^2=-(x_1(t)-c)^2\le 0\end{aligned}$$

that is, starting at a point $x(0)$, the distance to the optimal value is strictly decreasing unless $x=(\theta^*,\lambda^*)$. Following this, for all starting points inside a circle around c the entirety of the trajectory



(a) Vector field with trajectories.

(b) A given trajectory with $\epsilon = 0.2$.

Figure 2.9: To the left we show the vector field of Example 2.5 with some trajectories streamlined. To the right, a typical trajectory shows the oscillating behavior.

lies within that circle, and therefore $G(x(t))$ is bounded. Once the limit ODE has been established, its long-term behaviour can be studied around the equilibrium point (c, c), and the corresponding vector field is shown in Figure 2.9(a).

As shown in the plot, every trajectory that follows this vector field will spiral inwards towards the optimal point (c, c). The spiralling effect is due to the complex component of the eigenvalues and is typical of the Arrow Hurwicz algorithm and other differential games, where one "player" tries to minimise and the other to maximise the same economic function. Indeed, the major driving component of the algorithms depends on the exponential of the eigenvalues, and for an imaginary number $z = i\alpha$, $e^z = \sin(\alpha) + i\cos(\alpha)$ has oscillatory behaviour, as shown in Figure 2.9(b).

⁂

2.4 ODE method for Optimization and Learning

So far our limiting results establish when an interpolated difference equation converges in the sup-norm to the solution of an ODE. Whether projection, penalty or multiplier methods are used, in an optimization setting the ODE is given through a field that depends on the gradient of the cost function and of the constraints. Establishing properties of the stable points of this ODE (which are stationary points) requires the matrix $\mathbb{A} = \nabla G$ at the stable points. By Theorem 2.4, the limiting behavior of the algorithm for θ_n as $\epsilon \to 0$ is described by the ODE, and we can refer to the stability of the ODE for properties of the gradient descent algorithm for large n.

When dealing with an optimization problem, it is essential to find an appropriate model for the problem. Cost and constraint functions are not unique; one can use alternative formulations of the problem, as long as they have the same solution (i.e., the same KKT points, see Definition 1.10).

Definition 2.6 *Let $J \in \mathcal{C}^1$, and $\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0\} \subset \mathbb{R}$, with $g \in \mathcal{C}^1$, be a feasible solution set. The optimization problem*

$$\min_{\theta \in \Theta} J(\theta), \tag{2.19}$$

is said to be well posed *if the set of solutions to (2.19) is not empty, contains only KKT points, and the set of KKT points is confined to a compact set. It is called* ill-posed *otherwise.*

As seen in Chapters 1 and 2, it is possible to define a function $G(\theta)$ such that the solution θ^* of the optimization problem is an asymptotically stable point of the ODE with vector field G. In the unconstriant case $G = \nabla J$ is an obvious choice, while in the constraint case G may be choosen as the gradient of the Lagrange function (see the Arrow-Hurwicz algorithm in (1.42a) to (1.42c) in Chapter 1). Under approriate conditions, see e.g. Theorem 2.4, the vector field G will then naturally define the recursive algorithm to find the numerical solution of the problem and we call $G(\theta)$ the *target vector field* and the ODE that goes with $G(\theta)$ the *limiting ODE*. In other words, an iterative numerical method will work only if the limiting ODE tracks the solutions of the optimization problem. This leads to the following definition.

Definition 2.7 *Let $J \in \mathcal{C}^1$, and $\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0\} \subset \mathbb{R}$, with $g \in \mathcal{C}^1$, be a feasible solution set. A target vector field $G \colon \mathbb{R}^d \to \mathbb{R}^d$ that is Lipschitz continuous is said to be* coercive *for a well posed problem (2.19) if the KKT points of (2.19) are the only asymptotically stable points of the ODE*

$$\frac{dx(t)}{dt} = G(x(t)), \tag{2.20}$$

and for any initial point x_0, the vector filed G is bounded along $\{x(t) : t \geq 0\}$.

Equivalently, G is coercive for (2.19) if the KKT points of (2.19) are the only zeroes of G, and if there exists a d-box $H = \{\theta \in \mathbb{R}^d \colon \theta_i \in [-M, M]\}$, for $M \gg 0$, such that each trajectory of the truncated ODE

$$\frac{x(t)}{dt} = G(x(t))\mathbf{1}_{\{x(t)\in H\}}$$

is the same as the trajectory of (2.20), provided that $x(0) \in H$. While we use the notation $\mathbf{1}_{\{x(t)\in H\}}$ it is worth emphasizing that the ODE limit only makes sense around points which are interior to the set H, and not at the (discontinuous) boundaries.

We summarize our discussion by saying that if G is coercive for a well posed problem, then the deterministic algorithm $\theta_{n+1} = \theta_n + \epsilon\, G(\theta_n)$ will approximate the solution of the optimization problem. For proof use Theorem 2.4. In the following example we show how Lyapunov-type arguments can be used to establish boundedness along the trajectories of an ODE.

EXAMPLE 2.6. Consider a well posed unconstrained optimization problem $\min_{\theta\in\mathbb{R}^d} J(\theta)$. Then any descent direction $G(\theta)$ such that $\nabla J(\theta)G(\theta) < 0$ for all θ for $\nabla J(\theta) \neq 0$ is a coercive field for the problem if the solution $x(t)$ of the ODE

$$\frac{dx(t)}{dt} = G(x(t)),$$

converges as $t \to \infty$ towards a stable point of $G(x)$. To see this, we use a Lyapunov function. First define $x(t)$ as the solution of the ODE

$$\frac{dx(t)}{dt} = G(x(t)),$$

and let

$$V(t) = J(x(t)) - J(\theta^*), \tag{2.21}$$

where θ^* is the location of the global minimum of $J(\theta)$. Then

$$\frac{dV(t)}{dt} = \nabla J(x(t))\frac{dx(t)}{dt} = \nabla J(x(t))G(x(t)) \leq 0. \tag{2.22}$$

By construction $V(t) \geq 0$ for all t, so that necessarily $V(t)$ converges. Moreover, provided that $J, G \in \mathcal{C}^1$, $V'(t)$ is continuous. We now show that this implies $V(\bar{x}) = 0$, from which we conclude by (2.21) together with our assumption that θ^* is a location of the global minimum, that $\bar{x} = \theta^*$. We prove the claim by contradiction. Assume that $V(\bar{x}) > 0$, then $\bar{x} \neq \theta^*$, so $\nabla J(\bar{x}) \neq 0$, and from $\nabla J(\bar{x}))G(\bar{x}) \leq 0$, we conclude that $G(\bar{x}) \neq 0$. Hence, $\bar{x}$ is not a stable point of G which contradicts our assumption.
⁂

Using the results from Theorem 2.3 and Exercise 2.3, for either constant or decreasing stepsize, various algorithms for constrained optimisation can be shown to converge under appropriate conditions on the cost function $J(\cdot)$ and the constraints g, h. In their original work, Arrow and Hurwicz [3] used an ODE approach to establish optimality of (1.42). They posed the problem in terms of a competition game and they added a penalty to their objective function to ensure convexity. The following theorem assumes strict convexity, so no penalty is required. Its proof is stated within the framework of this chapter. Exercise 2.6 focuses on a simple one-dimensional case that can help to illustrate the main ideas of this proof.

Theorem 2.7 *Consider the strictly convex non linear problem (1.21), and assume that $\nabla J(\theta), \nabla g(\theta)$ are continuous and bounded on bounded intervals (for example, with uniformly bounded Lipshitz constant), and let $x_0 \in \mathbb{R}^d, y_0 > 0 (\in \mathbb{R}^p), \eta_0 \in \mathbb{R}^q$ be an initial point of the Arrow-Hurwicz algorithm (1.42) such that all inactive constraints at this point are the same as those for the optimal point θ^*. Then the limit ODE is:*

$$\begin{aligned}
\frac{dx(t)}{dt} &= -\nabla_\theta \mathcal{L}(x(t), y(t), z(t)) \\
\frac{dy(t)}{dt} &= g(x(t))\mathbf{1}_{\{y(t)\geq 0\}} \\
\frac{dz(t)}{dt} &= h(x(t)),
\end{aligned} \tag{2.23}$$

whose stable points are saddle points of the Lagrangian. In words, the above ODE is coercive for problem (1.21).

Proof: In order to apply Theorem 2.5 we study a simplification of the vector field associated to (2.23). We observe first that inactive constraints do not affect local behavior (their multipliers are zeroes) so they can be ignored from the analysis in a neighborhood of any interior point (θ, λ, η). Call $\tilde{g}(\theta)$ the corresponding vector with only those components j for which $\lambda_j > 0$. Once the total dimension has thus been reduced, the corresponding matrix ∇G is given by:

$$\mathbb{A} = \begin{pmatrix} -H(\theta, \lambda, \eta) & -B(\theta)^\top \\ B(\theta) & 0 \end{pmatrix}$$

where $H(\theta, \lambda, \eta) = \nabla^2_\theta \mathcal{L}(\theta, \lambda, \eta) \in \mathbb{R}^{d\times d}$, and $B^\top(\theta) = (\nabla \tilde{g}(\theta), \nabla h(\theta)) \in \mathbb{R}^{d\times(p+q)}$ is the gradient of the active constraints at θ. For a strictly convex non linear problem, $H(\theta, \lambda, \eta)$ is positive definite, and if $\nabla B(\theta)$ is full rank (that is, the constraint qualifications hold at θ), then the matrix $\mathbb{A}$ is a Hurwitz matrix, driving the ODE towards its asymptotically stable point, as we now show.

Let v be a eigenvector of unit norm of $\mathbb{A}$ with eigenvalue ρ. Then

$$\mathbb{A}v = \begin{pmatrix} -H(\theta,\lambda,\eta) & -B(\theta)^\top \\ B(\theta) & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -H(\theta,\lambda,\eta)\, v_1 - B(\theta)^\top v_2 \\ B(\theta) v_1 \end{pmatrix} = \rho \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

where v_1 is a d-dimensional vector and v_2 is a $(p+q)$-dimensional vector. The first claim is that v_1 cannot be the null vector. Proceeding by contradiction, if $v_1 = 0$ then the above equation would imply that $B(\theta)^\top v_2 = 0$, which has the unique solution $v_2 = 0$ because $B(\theta)$ is full rank. By assumption v is a unit norm eigenvalue and thus it cannot be the zero vector.

We use now some results from matrix algebra: (a) for a symmetric matrix the eigenvalues are real, (b) a real matrix has eigenvalues that are either real-valued, or appear in pairs of complex conjugates, and (c) $\mathbb{A} + \mathbb{A}^*$ is symmetric, where $\mathbb{A}^*$ is the conjugate transpose of $\mathbb{A}$.

The identity $\mathbb{A}v = \rho v$ implies that $v^\top \mathbb{A}^* = \rho^* v^\top$, where ρ^* is the complex conjugate of ρ. Then the real part of ρ is

$$\Re(\rho) = v^\top (\mathbb{A} + \mathbb{A}^*) v = \frac{1}{2} (v_1^\top, v_2^\top) \begin{pmatrix} -H(\theta,\lambda,\eta) & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = -v_1^\top H(\theta,\lambda,\eta) v_1$$

with $H = -[\nabla^2 J(\theta) + \lambda \nabla^2 \tilde{g}(\theta) + \eta \nabla^2 h(\theta)]$ being a positive definite matrix. Because $v_1 \neq 0$ then the real part of all eigenvalues is strictly negative, which shows that ∇G satisfies the Hurwitz condition. The claim now follows from Theorem 2.5.

QED

2.5 Exercises

EXERCISE 2.1. Let $f_n(x) = \sin(x) + \frac{x}{n}$, and $f(x) = \sin(x)$. Show that:

(a) for each compact set B and $x \in B$, $\lim_{n\to\infty} f_n(x) = f(x)$, but

(b) $f_n \not\to f$ in the sup-norm.

EXERCISE 2.2. Use boundedness of G to show that as $\epsilon \to 0$, $\{x_\epsilon(\cdot)\}$ in the proof of Theorem 2.3 is equicontinuous in the extended sense.

EXERCISE 2.3. Show Theorem 2.6 using the following steps. First consider the recursion (2.8) (bias term zero):

$$\theta_{n+1} = \theta_n + \epsilon_n G(\theta_n), \quad \theta \in \mathbb{R}^d,$$

where the function $G(\theta)$ is bounded and Lipschitz continuous. Let $\vartheta^{\boldsymbol{\epsilon}}(\cdot)$ denote its interpolation process in Definition 2.4. Assume that $\sum_n \epsilon_n = \infty, \epsilon_n \to 0$ and let $x_n(t) = \vartheta^{\boldsymbol{\epsilon}}(t_n + t), t \geq 0$, where $t_n = \sum_{k=1}^n \epsilon_n$ as before.

(a) Write the telescopic sum for $x_n(t+s) - x_n(t)$, and express this through an integral approximation.

(b) Show that $\{x_n\}$ is equicontinuous in the extended sense

(c) Use the Theorem of Ascoli-Arzelà to prove that, when $n \to \infty$, x_n converges (in the sup norm) to the solution of the ODE:

$$\frac{dx(t)}{dt} = G\big(x(t)\big). \tag{2.24}$$

(d) Argue like in the proof of Theorem 2.3 and extend your result to the biased version

$$\theta_{n+1} = \theta_n + \epsilon_n(G(\theta_n) + \beta_n(\theta_n)), \quad \theta \in \mathbb{R}^d,$$

where you assume that $\sum_n \epsilon_n^2 \|\beta_n(\theta)\| < \infty$.

EXERCISE 2.4. Consider the case $G(\theta) = -\nabla J(\theta)$ in (2.8). Discuss the differences in assumptions and conclusions between Theorem 1.3 and the result in exercise Exercise 2.3.

EXERCISE 2.5. Consider again the problem of Example 1.1 of the surfer at the beach who wishes to rescue a drowning victim. We wish to minimise $J(\theta)$, but the stationary points are only given in the implicit equation (1.6). Consider the gradient search method:

$$\theta_{n+1} = \theta_n - \epsilon J'(\theta_n).$$

(a) Show that as $\epsilon \to 0$ the interpolation processes converge to the ODE:

$$\frac{dx(t)}{dt} = \frac{\sin(\alpha_2(x(t))}{v_2} - \frac{\sin(\alpha_1(x(t))}{v_1}.$$

(b) Show that θ^* is stable and argue that a solution $x(t)$ of the above ODE must then satisfy $\lim_{t\to\infty} x(t) = \theta^*$.

(c) Program the procedure and plot the results, using $a = 2, b = 5, d = 10, v_1 = 3, v_2 = 1$, and $\epsilon = 0.05$. Hint: the derivative can be re written as:

$$J'(\theta) = \frac{1}{v_1}\frac{\theta}{\sqrt{\theta^2 + a^2}} - \frac{1}{v_2}\frac{d-\theta}{\sqrt{(d-\theta)^2 + b^2}}.$$

EXERCISE 2.6. Consider the one dimensional case where $d = p = 1$ and $q = 0$ (no equality constraints). Evaluate the eigenvalues of $\mathbb{A}$ and prove that if $J(\cdot)$ is strictly convex and $g(\cdot)$ is convex, then $\mathbb{A}$ is Hurwitz, provided that $g'(\theta^*) \neq 0$. This condition follows if we require that the constraint qualifications hold for this problem.

EXERCISE 2.7. Prove Theorem 2.7, as follows. Consider a strictly convex non linear problem:

$$\min_{\theta\in\Theta} J(\theta),$$
$$\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0, h(\theta) = 0\},$$

and assume that $\nabla J(\theta), \nabla g(\theta)$ are continuous and bounded on bounded intervals (for example, with uniformly bounded Lipshitz constant), and let $x_0 \in \mathbb{R}^d, y_0 > 0 (\in \mathbb{R}^p), \eta_0 \in \mathbb{R}^q$ be an initial point of the Arrow-Hurwicz algorithm. Then this algorithm has a local vector field described by the ODE (2.23), whose stable points are saddle points of the Lagrangian.