Part I

Theory of Stochastic Optimization and Learning

Chapter 1

Gradient-based Methods for Deterministic Continuous Optimization

This chapter presents a summary of salient results in deterministic optimization, particularly focusing on numerical methods. For basic definitions and results we refer to standard textbook. For a brief review of fundamental results we refer to the Appendix.

1.1 Unconstrained Optimization

Consider a cost function $J(\theta)$, with $J\colon \Theta \subset \mathbb{R}^d \to \mathbb{R}$, where θ is a decision vector. Throughout this monograph, we seek to the find the minimum of $J(\theta)$ for $\theta \in \Theta$. As is standard in the literature, we are not only interested in the value of the global minimum (if it exists) but also its location, i.e., we seek the solution θ^* to the problem

$$\arg\min_{\theta\in\mathbb{R}^d} J(\theta). \tag{1.1}$$

In case that the global minimum is attained at several locations, θ^* is one of these locations. In case that $J(\cdot)$ is an (affine) linear mapping, the above optimization problem is called a *linear problem* and it can be addressed with methods from the theory of linear optimization. See, for example, [23, 24, 8, 9] for details. In case that $J(\cdot)$ is a general "smooth" continuous real-valued function, the above problem is called a *non linear problem* and it is referred to as an NLP. The theory presented in this monograph is devoted to the study of NLP's. It is worth noting that while results presented here can also be applied to linear problems, there are often more efficient methods available for linear problem exploiting the linear nature of the problem.

We assume that $\mathbb{R}^d$ is equipped with a norm denoted by $||\cdot||$. Most results presented in the following are independent of the choice of $||\cdot||$. Occasionally, we will work with the Euclidean norm on $\mathbb{R}^d$ given by

$$\|x\| = \sqrt{x_1^2 + \ldots + x_d^2},$$

and when results only hold for this particular norm it will be stated in the text.

A particular class of applications arises when an input data vector x and corresponding output data vector $h(x)$ is available. Letting $f(\theta, x)$ denote some parametrized mapping proposed for re-

placing the unknown mapping $h(x)$, considering

$$J(\theta, x) = \|f(\theta, x) - h(x)\|^2$$

and solving (1.1) for given x, yields then he best fit to the output. This is called *supervised learning* in the literature. In this monograph we will discuss classical optimization as well as learning applications.

Definition 1.1 *The* level sets *of a function* $J\colon \mathbb{R}^d \to \mathbb{R}$ *are defined for every level* $\alpha \in \mathbb{R}$ *as:*

$$\mathcal{L}_\alpha(J) = \{\theta \in \mathbb{R}^d \colon J(\theta) \leq \alpha\}.$$

When no confusion arises, the notation will be simplified to $\mathcal{L}_\alpha$.

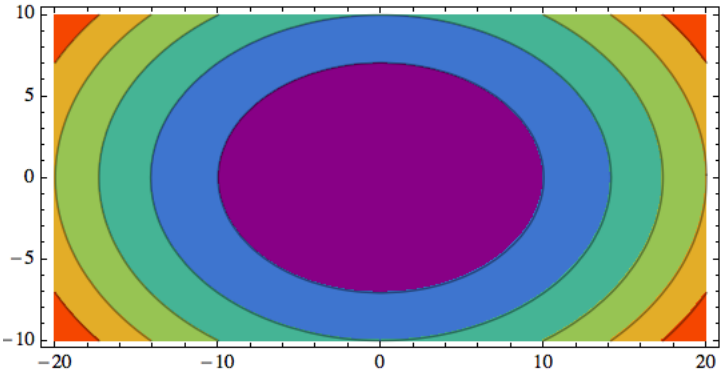


Figure 1.1: Plot showing various level curves for the function $x^2 + 2y^2$.

Denote the n-times continuously differentiable mappings from $\mathbb{R}^d$ to $\mathbb{R}$ by $\mathcal{C}^n$. For $J \in \mathcal{C}^1$, we denote the gradient of $J(\cdot)$ by $\nabla J(\cdot)$, and for $J \in \mathcal{C}^2$, we denote the Hessian of $J(\cdot)$ by $HJ(\cdot) = \nabla^2 J(\cdot)$. Following standard notation, vectors in $\mathbb{R}^d$ are *column* vectors. For $x \in \mathbb{R}^d$, we denote the i-th element of x by x_i. In case of a sequence of vectors $\{x_n\}$, with $x_n \in \mathbb{R}^d$, we denote the i-th element of x by $x_{n,i}$. The gradient is a *row* vector with components $\partial/\partial\theta_k, k = 1, \ldots d$. The Hessian is a $d \times d$ matrix with (i, j)-components $\partial^2/\partial\theta_j\partial\theta_i$. For a vector $v \in \mathbb{R}^d$ we write $v \geq 0$ if $v_i \geq 0$ for all components $i = 1, \ldots, d$.

A matrix $B \in \mathbb{R}^{d\times d}$ is negative (positive) definite if $d^\top Bd < (>)0$ for all $d \in \mathbb{R}^d$ with $d \neq 0$, where $B^\top$ denotes the transposed of matrix B. It is called "semi"-definite if equality is replaced by inequality. The notation $B < (>)0$ is often used. A square matrix B is called symmetric if $B = B^\top$. For symmetric matrices the following characterization of positive definiteness exists: if B is symmetric, then $B > 0$ if and only if all its eigenvalues are strictly positive.

REMARK. The visual interpretation of the gradient of a function will be very useful in the sequel. Refer to Figure 1.1. This is a "topographical" visualization of a two dimensional function, where colors indicate height. Each of the level sets defines a boundary (in the example, they are ellipses). The gradient of the function (in this case $x^2 + 2y^2$) records the rate of growth of the function along each of the axes. Now, take any point on a level set (refer to Figure 1.2). Because the function does

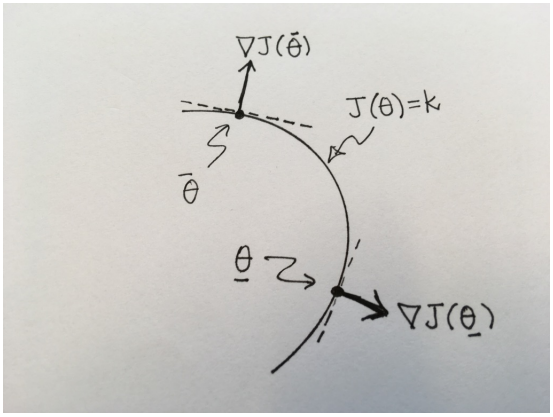


Figure 1.2: Illustration of the gradient of a convex function at two points $\bar{\theta}$ and $\underline{\theta}$.

not change *along* this curve, then necessarily the gradient ∇J must point *perpendicular* to the curve (i.e., the projection of the gradient on the curve is zero). For the example it points outwards, in the direction of growth.

Definition 1.2 *A function* $J\colon \mathbb{R}^d \to \mathbb{R}$ *is called* concave (convex) *if for all* $x, y \in \mathbb{R}^d$ *and* $\alpha \in [0,1]$,

$$J(\alpha x + (1-\alpha)y) \geq (\leq)\ \alpha J(x) + (1-\alpha)J(y)\ .$$

Strict *concavity (convexity) is obtained when the above inequalities are strict. An equivalent condition is that the Hessian of the function be negative (positive) semi-definite:* $\nabla^2 J(\cdot) \leq (\geq)\ 0$, *and strict concavity (convexity) follow when the Hessian is negative (positive) definite throughout the domain of* J.

Definition 1.3 *A point* $\theta^* \in \mathbb{R}^d$ *can be characterized as follows:*

- *If* $J(\theta^*) \leq J(\theta)$, *for all* $\theta \in \mathbb{R}^d$, *then* θ^* *is a called* global *minimum. It is called a* local minimum *if there is a* $\rho > 0$ *such that* $||\theta - \theta^*|| \leq \rho$ *implies* $J(\theta) \geq J(\theta^*)$.
- *If* $J(\theta^*) \geq J(\theta)$, *for all* $\theta \in \mathbb{R}^d$, *then* θ^* *is called a* global *maximum. It is called a* local maximum *if there is a* $\rho > 0$ *such that* $||\theta - \theta^*|| \leq \rho$ *implies* $J(\theta) \leq J(\theta^*)$.
- *If the function may increase or decrease in a small neighborhood of the point, depending on the direction of motion, then* $\theta^* \in \mathbb{R}^d$ *is a* saddle point.

Let $\alpha^* \stackrel{\text{def}}{=} \min J(\theta)$, with $\alpha^* = J(\theta^*)$. Then α^* is called the *value* of the minimum and θ^* the *location of the minimum*. Note that the value of the minimum is unique (provided it exists) but there may be more than one location yielding the same minimal value of $J(\theta)$. In fact the level set of $J(\theta)$ for level α^*, denoted as $\mathcal{L}_{\alpha^*}$, yields the set of all locations of the global minima of $J(\theta)$. Figure 1.3 shows an example with various minima, maxima and saddle points. The function is $\cos(\theta_1) + \sin(\theta_2)$.

For a twice continuously differentiable function $J(\theta)$, i.e., $J \in \mathcal{C}^2$, the Taylor series expansion yields the following approximation of the value of $J(\theta)$ at a point $\theta + td$, for $d \in \mathbb{R}^d$ and $t \in \mathbb{R}$:

$$J(\theta + td) = J(\theta) + t\,\nabla J(\theta) d + \frac{t^2}{2} d^\top \nabla^2 J(\theta)\, d + o(t^3). \tag{1.2}$$

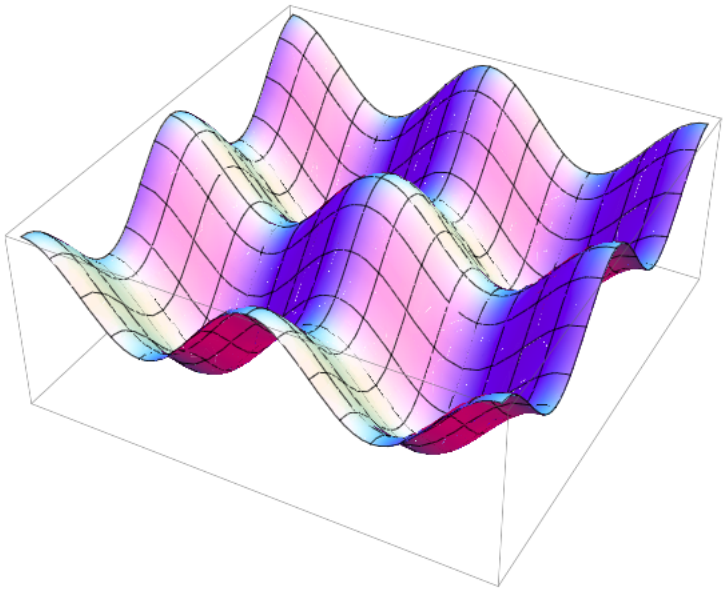


Figure 1.3: Example of a function with several maxima, minima and saddle points.

Definition 1.4 *The points $\bar{\theta} \in \mathbb{R}^d$ that satisfy $\nabla J(\bar{\theta}) = 0$ are called* stationary *points of J.*

The following theorem provides conditions for deciding the type of a stationary point.

Theorem 1.1 *Let $J \in \mathcal{C}^2$.*

- *A local minimum (local maximum) θ^* of J is a stationary point, that is, it satisfies the* first order optimality condition*:*
$$\nabla J(\theta^*) = 0. \tag{1.3}$$
- *A decision value θ^* is a local minimum (local maximum) of J if in addition to (1.3), the following is also satisfied*
$$\nabla^2 J(\theta^*) > 0 \qquad \left(\nabla^2 J(\theta^*) < 0\right). \tag{1.4}$$
Equation (1.4) is called second order optimality condition.
- *If J is a* convex *(concave) function, then (1.3) is necessary and sufficient for θ^* being a global minimum (maximum).*

Proof: Use a Taylor series approximation around θ^*, and the definition of a positive definite matrix, which implies that $d^\top \nabla^2 J(\theta^*) d > 0$ for all $d \in \mathbb{R}^d$. The details are left as an exercise.

QED

EXAMPLE 1.1. A well known historical problem is that of explaining the phenomenon of refraction of light when traversing two different media. Since Ptolemy (circa 140 AD), scientists were concerned with finding the relationship between the angles of refraction and the media's characteristics. The law of refraction was described by Ibn Sahl of Baghdad (978) who used it to shape lenses, and in 1621

by the Dutch astronomer Snellius. It is now called Snell's Law in English. In 1637 Descartes found the same principle, using conservation of moments, and in French it is called the Law of Decartes-Snell. We are mostly interested in Fermat, who in 1657 used variational calculus and his principle of "least time" to derive this law through an optimization problem.

The version of the problem that we give here is the pedagogical version of Richard Feynman. Imagine that you are walking on the beach when you see a person drowning and shouting for help. To get from where you are to the drowning person in the fastest way, you should not move along the straight line, because you run faster on the sand than you can swim. The distance from your position to the water is a, the distance from the water to the person is b, and the length of shoreline between the two points is d, as shown in Figure 1.4. In Cartesian coordinates the drowning person is at $B^\top = (d, -b)$ and your position is $A^\top = (0, a)$. Here we assume that the water front is a straight line for simplicity. The speed on sand is v_1 and in the water v_2, with $v_2 < v_1$. We call θ the crossing point.

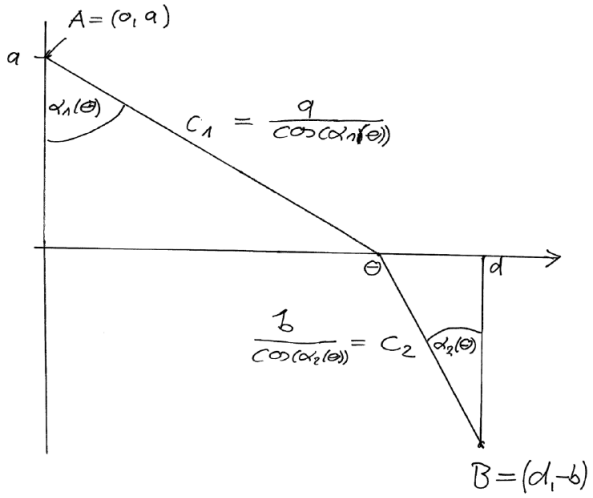


Figure 1.4: Problem: find the optimal crossing point, indicated by x in diagram.

Fermat's reasoning was that light chooses not the shortest path but the one that saves more energy, which is the *fastest* path. It is easy to argue existence of a solution for this problem: any value of θ to the left of the crossing point of the straight line between points A and B will give a slower path than the straight line, because $v_2 < v_1$. On the other hand, any point $(0, \theta)$, with $\theta > d$, will require unnecessary additional travel time compared to crossing at $(0, d)$, therefore there must be a minimum point between these two points. That the travel times are continuously differentiable follows from the linear relationships between distance and time.

At speed v, the distance traveled in time t is vt, so the total travel time can be expressed as:

$$\begin{aligned} J(\theta) &= \frac{1}{v_1}\frac{a}{\cos(\alpha_1(\theta))} + \frac{1}{v_2}\frac{b}{\cos(\alpha_2(\theta))} \\ &= \frac{a}{v_1}\sec(\alpha_1(\theta)) + \frac{b}{v_2}\sec(\alpha_2(\theta)), \end{aligned}$$

where the angles $\alpha_i(\theta)$ are as labeled in the diagram of Figure 1.4. Figure 1.5 shows the plot of the time as a function of the crossing point θ.

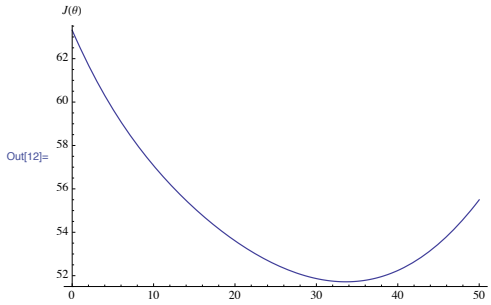


Figure 1.5: Plot of the function $J(\theta)$.

According to Theorem 1.1 we now find the stationary points of $J(\theta)$. We use the following identities:

$$\tan\alpha_1(\theta) = \frac{\theta}{a}, \tag{1.5a}$$

$$\tan\alpha_2(\theta) = \frac{d-\theta}{b}, \tag{1.5b}$$

$$\frac{d}{d\alpha}\tan(\alpha) = \sec^2(\alpha) = \cos^{-2}(\alpha), \tag{1.5c}$$

$$\frac{d}{d\alpha}\sec(\alpha) = \sec(\alpha)\tan(\alpha). \tag{1.5d}$$

To obtain the first order optimality condition, differentiate $J(\theta)$ and set it equal to zero:

$$J'(\theta) = \frac{a}{v_1}\,\sec(\alpha_1(\theta))\tan(\alpha_1(\theta))\,\frac{d\alpha_1(\theta)}{d\theta} + \frac{b}{v_2}\,\sec(\alpha_2(\theta))\tan(\alpha_2(\theta))\,\frac{d\alpha_2(\theta)}{d\theta} = 0.$$

By identity (1.5) it holds $\tan(\alpha_1(\theta))/\theta = a = \text{constant}$. Differentiating both sides of this equation with respect to θ yields

$$\frac{1}{\theta}\sec^2(\alpha_1(\theta))\left(\frac{d\alpha_1(\theta)}{d\theta}\right) - \frac{\tan\alpha_1(\theta)}{\theta^2} = 0 \;\Rightarrow\; \frac{d\alpha_1(\theta)}{d\theta} = \frac{1}{\theta}\sin(\alpha_1(\theta))\,\cos(\alpha_1(\theta)).$$

Similarly,

$$\frac{d\alpha_2(\theta)}{d\theta} = -\frac{1}{d-\theta}\sin(\alpha_2(\theta))\,\cos(\alpha_2(\theta)).$$

Replacing these values in $J'(\theta)$ and using again the identities in (1.5), one reaches the conclusion that $J'(\theta^*) = 0$ is achieved at the unique point that satisfies:

$$\frac{\sin(\alpha_1(\theta^*))}{\sin(\alpha_2(\theta^*))} = \frac{v_1}{v_2}, \tag{1.6}$$

known as Snell's Law of refraction. Going back to the person at the beach, knowing Snell's Law is not very useful because he or she still has to determine the optimal crossing point $(0, \theta^*)$, however (1.6) gives it as an implicit solution.

⁂

When solving a problem of the form (1.1) analytically, one first looks for all points that satisfy (1.3). After the set of candidates is determined, one then evaluates the Hessian $\nabla^2 J(\theta)$ to verify which are local minima. If several local minima are found and if, in addition, it can be shown that $J(\theta)$ tends to ∞ as $||\theta||$ tends to infinity, then the location of the global minimum θ^* can be found by comparing the values of the local minima. It is worth noting that if there exists a unique stationary point this analysis can be simplified. Indeed, if θ^* is the unique stationary point of $J(\theta)$ and if $J(\theta)$ tends to ∞ as $||\theta||$ tends to infinity, then θ^* is the unique location of the global mimimum of $J(\theta)$.

Numerical Methods

In most cases as in Example 1.1, it is impossible to solve the inversion problem $\nabla J = 0$ analytically and iterative numerical methods are used for finding a root θ^* of $\nabla_\theta J = 0$. Generally speaking, a numerical algorithm for approximating the solution θ^* of $\nabla_\theta J = 0$ is a recursion of the form:

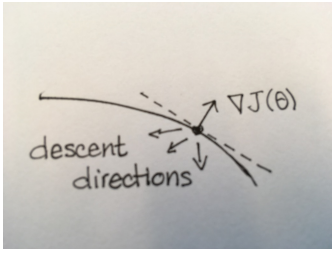
$$\theta_{n+1} = \theta_n + \epsilon_n d(\theta_n), \tag{1.7}$$

where, for each n, ϵ_n is called the *stepsize* or *gain sequence* and $d(\theta_n)$ is called the *direction* of the algorithm.

Methods for approximating θ^* can be classified according to the choice of the stepsize rule and the directions. Together with an initial value θ_0 and a *stopping rule*, (1.7) constitutes a numerical algorithm that terminates hopefully close to the true optimum, where "closeness" has to be defined appropriately.

Definition 1.5 *A descent direction of a differentiable function J at a point $\theta \in \mathbb{R}^d$ is any vector $d(\theta)$ such that $\nabla J(\theta)\, d(\theta) < 0$.*

A descent direction $d(\theta)$ is pointing away from the direction $\nabla J(\theta)$. Indeed $(-\nabla J(\theta)\, d(\theta)) > 0$ implies that there is an angle of less than 90 degrees between $d(\theta)$ and $-\nabla J(\theta)$. The figure to the right depicts the situation. Recall that $\nabla J(\theta)$ points towards the direction of growth of the function. For a detailed geometric interpretation of the gradient, we refer to Section A.1 in the Appendix.



Gradient-based methods for optimization use $d(\theta_n) = -\nabla J(\theta_n)^\top$ as a direction in the algorithm, which is clearly a descent direction. There a many methods available for gradient-based optimization. Typically these algorithms are tailored to specific classes of function such as the conjugate-gradient method and variations thereof, which is a popular method for optimization of quadratic functions. The next section will discuss Newton's Method which is a general method applying to smooth functions.

*Newton's Method

One of the most efficient methods for unconstrained optimization is the method developed by Newton (published in 1685) and Raphson (1690). It was originally designed to find the zeroes of a polynomial. In the context of finding stationary points of $J(\theta)$, $\theta \in \mathbb{R}^d$, let the vector $G(\theta)$ represent

the gradient $\nabla J(\theta)^\top$. From a point θ_n, use a linear approximation of $G(\theta)$, that is, using Taylor's expansion,

$$G(\theta_{n+1}) \approx G(\theta_n) + \nabla G(\theta)(\theta_{n+1} - \theta_n),$$

where now $\nabla G(\theta) = \nabla^2 J(\theta)$ is a $d \times d$ matrix for each θ.

To obtain the zero (of the approximation) in one step, simply set $G(\theta_{n+1}) = 0$ in the approximation and solve the RHS for θ_{n+1}. Assuming that the inverse matrix of $\nabla G(\theta)$ exists, this results in:

$$\theta_{n+1} = \theta_n - [\nabla G(\theta_n)]^{-1} G(\theta_n). \tag{1.8}$$

In words, Newton's method is a gradient decent method with step-size sequence $\epsilon_n = [\nabla G(\theta_n)]^{-1}$.

Theorem 1.2 *Let $J\colon \mathbb{R}^d \to \mathbb{R} \in \mathcal{C}^2$ be a convex function, and assume that the Hessian is invertible (thus, it is positive definite around any local minimum). Choose an initial point $\theta_0 \in \mathbb{R}^d$ and let $\{\theta_n\}$ be the sequence defined by (1.8), with $G = \nabla J^\top$, that is,*

$$\theta_{n+1} = \theta_n - [\nabla^2 J(\theta_n)]^{-1} \nabla J(\theta_n)^\top.$$

Suppose that $\bar{\theta}$ is an accumulation point of the sequence $\{\theta_n\}$ such that $\nabla^2 J(\bar{\theta}) > 0$, then $\bar{\theta}$ is a local minimum of $J(\theta)$ and the rate of convergence *is superlinear, that is, there exists a sequence $\{c_n\}$ such that c_n tends to zero as n tends to ∞ and for some finite N it holds that*

$$\|\theta_{n+1} - \bar{\theta}\| \leq c_n \|\theta_n - \bar{\theta}\|, \quad n \geq N.$$

Furthermore, if $J \in \mathcal{C}^3$ then the rate of convergence is quadratic, that is, there exists a constant $c > 0$ such that, for large n,

$$\|\theta_{n+1} - \bar{\theta}\| \leq c \|\theta_n - \bar{\theta}\|^2.$$

The proof of the result uses Taylor's approximation. In addition, positive semi-definiteness of the Hessian due to convexity implies Newton's direction exists and it is an descent direction at every point. Although very efficient for convex functions, Newton's method has a number of practical problems when applied as a general purpose optimization method:

- Newton's method finds zeros of the gradient, which may be locations of minima or inflection points for general functions, and consequently it cannot be guaranteed that the Hessian is positive definite at every stationary point.
- The Hessian may not be invertible at every point.
- It needs calculation of gradients, Hessians, and Hessian inversion, all of which may be lengthy numerical operations, rendering the method slow.
- Finally, when a closed form expression does not exist, approximations to gradients via finite differences may make the method very slow, due to the sheer amount of function calculations required for Hessian evaluation (2^{2d}).

For deterministic problems, the *efficiency* of a method is defined in terms of CPU time to achieve a given precision δ. A number of algorithms have been proposed under the common name of "quasi-Newton" methods, which attempt to increase the efficiency of the method, overcoming the problems pointed out above. In particular, if the function $J(\theta)$ is not convex, then for points θ_n in regions which are not convex Newton's direction is not a descent direction. To ensure proper behavior of the algorithm, one can choose the negative gradient, which is always a descent direction.

*Cauchy's Method

The method known as *steepest descent* (or Cauchy's) for minimization of a cost function $J(\theta)$ chooses $d(\theta_n) = -\nabla_\theta J(\theta_n)$ at each iteration of (1.7). Originally proposed by Cauchy in 1847, instead of premultiplying by the matrix $[\nabla^2 J(\theta_n)]^{-1}$, the method chooses the step size to "jump" along the direction $d(\theta_n)$ to reach the minimum on that line, that is,

$$\epsilon_n = \arg\min_{\epsilon>0}(J(\theta_n - \epsilon \nabla J(\theta_n)^\top)) .$$

In this section we present results that use the steepest descent direction $d(\theta_n) = -\nabla J(\theta_n)^\top$ and focus on *decreasing* stepsizes of the simplest form.

Gradient-based Methods: non-adaptive step sizes

As mentioned before, Newton's method has a good convergence rate but every iteration may require too much computational time. Cauchy's method can have slow convergence due to possible zig-zagging of the iterations, and several modifications have been proposed for adaptive step sizes (where ϵ_n depends on $\theta_n, J(\theta_n), \nabla J(\theta_n)$, etc). Common methods use Wolfe's conditions [32, 33] and Armijo's rules [2], which ensure that all accumulation points are local minima. For deterministic optimisation adaptive step sizes are undoubtedly superior to non-adaptive step sizes. However, the focus of the present text is to extend the basic methodology for deterministic optimisation to problems where the observations of the function $J(\theta)$ and its gradients (if available) are noisy, and the noise models may be very complex. For such a scenario it is often a better idea to use algorithm parameters that will not be corrupted by the noise, which is why we only study non-adaptive step sizes here. The gradient-based methods use $d(\theta) = -\nabla J(\theta)$ as the direction of the algorithm, and the step sizes can be of two kinds: either decreasing: $\epsilon_n \downarrow 0$, or constant: $\epsilon_n \equiv \epsilon$.

Without any detailed analysis, inspecting the mere structure of (1.7) allows us already to deduce properties of the stepsize sequence. To see this, insert the expression for θ_n on the right-hand side of (1.7), which yields $\theta_{n+1} = \theta_{n-1} + \epsilon_n d(\theta_n) + \epsilon_{n-1} d(\theta_{n-1})$ and continuing the recurrence,

$$\theta_{n+1} = \theta_0 + \sum_{i=0}^{n} \epsilon_i \, d(\theta_i).$$

Suppose that $d(\cdot)$ is bounded. Then, for the algorithm to find θ^*, the stepsizes have to satisfy

$$\sum_{n=1}^{\infty} \epsilon_n = \infty, \tag{1.9}$$

so that the sequence $\{\theta_n\}$ is not confined to some bounded set (or, equivalently, will cover any bounded set as it can potentially reach any point in $\mathbb{R}^d$). Further conditions are required in order to ensure convergence of the algorithm to the optimal θ^*, as we will see now.

Theorem 1.3 *Let $J \in \mathcal{C}^2$ and assume that ∇J is Lipschitz continuous. For given initial value θ_0, let $\{\theta_n\}$ be given through the algorithm*

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla J(\theta_n)^\top, \tag{1.10}$$

where the gain sequence satisfies:

$$\sum_{n=1}^{\infty} \epsilon_n = +\infty, \quad \sum_{n=1}^{\infty} \epsilon_n^2 < \infty. \tag{1.11}$$

If $\{\|\nabla J(\theta_n)\| : n \geq 0\}$ is bounded, then every accumulation point of $\{\theta_n\}$ is a local minimum of J, provided that θ_0 is not a stationary point.

To prove this theorem, we need first an auxiliary result that will become useful later on:

Lemma 1.1 *Consider the real-valued recursion:*

$$x_{n+1} = x_n - g_n + h_n, \quad x_0 \in \mathbb{R}$$

where $g_n \geq 0$ for all n, and the sequence h_n is summable, i.e., $|\sum_n h_n| < \infty$. Then either $x_n \to -\infty$ or x_n converges to a finite value and $\sum_n g_n$ converges.

Proof: Note that

$$\begin{aligned} x_{n+2} &= x_{n+1} - g_{n+1} + h_{n+1} \\ &= x_n - (g_n + g_{n+1}) + (h_n + h_{n+1}). \end{aligned}$$

Repeating this argument m times yields the telescopic sum:

$$x_{m+n} = x_n - \sum_{i=n}^{m+n-1} g_i + \sum_{i=n}^{m+n-1} h_i. \tag{1.12}$$

By assumption $g_n \geq 0$, which implies

$$x_{m+n} \leq x_n + \sum_{i=n}^{m+n-1} h_i < \infty. \tag{1.13}$$

Use now $-\infty < H = \sum_{i=1}^{\infty} h_i < \infty$ to show that for all n

$$\limsup_{m\to\infty} \sum_{i=n}^{m+n-1} h_i = \lim_{m\to\infty} \sum_{i=n}^{m+n-1} h_i = \sum_{i=n}^{\infty} h_i < \infty \tag{1.14}$$

and

$$\liminf_{n\to\infty} \sum_{i=n}^{\infty} h_i = \lim_{n\to\infty} \sum_{i=n}^{\infty} h_i = 0. \tag{1.15}$$

By (1.14), taking the limit superior on both sides of the inequality (1.13) as m tends to ∞ yields for all n

$$\limsup_{m\to\infty} x_{m+n} \leq x_n + \sum_{i=n}^{\infty} h_i.$$

By (1.15), taking the limit inferior on both sides of the above inequality gives

$$\limsup_{m\to\infty} x_m \leq \liminf_{n\to\infty} x_n < \infty,$$

which implies that either x_n converges, so that $x_n \to \bar{x} \in \mathbb{R}$ for some $\bar{x} \in \mathbb{R}$, or $x_n \to -\infty$.

In the case that $\lim_n x_n = \bar{x} \in \mathbb{R}$, letting $n = 0$ in (1.12) yields

$$\sum_{i=0}^{m-1} g_i = \sum_{i=0}^{m-1} h_i - x_m + x_0,$$

which converges as $m \to \infty$ to a finite value, proving the claim.

QED

Proof: [of Theorem 1.3] Approximating $J(\theta_{n+1})$ via a Taylor series expansion developed at θ_n (i.e., let $d = \theta_{n+1} - \theta_n$ and $t = 1$ in (1.2)), yields

$$J(\theta_{n+1}) = J(\theta_n) + \nabla J(\theta_n)(\theta_{n+1} - \theta_n) + \frac{1}{2}(\theta_{n+1} - \theta_n)^\top \nabla^2 J(\xi)(\theta_{n+1} - \theta_n), \tag{1.16}$$

where $\xi = \alpha\,\theta_n + (1-\alpha)\theta_{n+1}$ for some $\alpha \in [0,1]$. Inserting (1.10) into the above representation of $J(\theta_{n+1})$ yields

$$J(\theta_{n+1}) = J(\theta_n) - \epsilon_n |\nabla J(\theta_n)|^2 + \frac{\epsilon_n^2}{2} \nabla J(\theta_n) \nabla^2 J(\xi) \nabla J(\theta_n)^\top, \tag{1.17}$$

recall that $|\cdot|$ denotes the Euclidean norm. Call $g_n = \epsilon_n |\nabla J(\theta_n)|^2$ and $h_n = \epsilon_n^2 \nabla J(\theta_n)^\top \nabla^2 J(\xi) \nabla J(\theta_n)/2$, then

$$J(\theta_{n+1}) = J(\theta_n) - g_n + h_n.$$

From Lipschitz continuity of $\nabla J(\theta)$ and assumption (1.20) it follows (see Exercise 1.7 below), that

$$|h_n| \le \frac{\epsilon_n^2}{2} L |\nabla J(\theta_n)|^2,$$

for some finite constant L. Boundedness of the gradient along the trajectory toegther with $\sum \epsilon_n^2 < \infty$, shows that h_n is absolutely summable, so we can apply Lemma 1.1 to conclude that $J(\theta_n)$ either diverges to $-\infty$ or it converges.

Suppose that $\bar{\theta} \in \mathbb{R}^d$ is an accumulation point of the algorithm. By continuity $J(\theta_{n_m})$ converges to $J(\bar{\theta})$ and $\nabla J(\theta_{n_m})$ converges to $\nabla J(\bar{\theta})$ for any subsequence $\{\theta_{n_m}\}$ that converges to an accumulation point of $\{\theta_n\}$. It follows from Lemma 1.1 that in this case, J and $\nabla J(\theta)$ converge and g_n is summable, that is,

$$\lim_{n\to\infty} \sum_{i=1}^{n} \epsilon_{m_i} |\nabla J(\theta_{m_i})|^2 < \infty,$$

which, by assumption (1.20), implies that $|\nabla J(\bar{\theta})| = \lim_{i\to\infty} |\nabla J(\theta_{m_i})| = 0$. This shows that any accumulation point is a *stationary* point. It only remains to show that such a point must be a local minimum. To this end, use the bound $|\nabla J(\theta_n) \nabla^2 J(\xi) \nabla J(\theta_n)^\top| \le L |\nabla J(\theta_n)|^2$ (see Exercise 1.7) to establish, from (1.17), that:

$$J(\theta_{m_{i+1}}) \le J(\theta_{m_i} - (\epsilon_{m_i} - L\epsilon_{m_i}^2)) |\nabla J(\theta_{m_i})|^2.$$

Since ϵ_{m_i} tends to zero as i tends to infinity, we have for sufficiently large i that $L\epsilon_{m_i} < 1$, and thus $(\epsilon_{m_i} - L\epsilon_{m_i}^2)|\nabla J(\theta_{m_i}|^2 > 0$. This implies that $J(\theta_{m_i})$ becomes is a monotone decreasing sequence in i for i sufficiently large. The fact that $\bar{\theta}$ is an accumulation point of $\{\theta_{m_i}\}$, implies that in any arbitrarily small neighborhood of $\bar{\theta}$ there exists a dense set of of values θ such that $J(\theta^*) \le J(\theta)$. Continuity of

$J(\theta)$ then implies that $J(\theta^*)$ is a local minimum.

QED

Theorem 1.3 provides sufficient conditions under which an accumulation point of a sequence $\{\theta_n\}$ obtained via a descent algorithm is the location of a local minimum. Next to more generic conditions such as the choice of the step size and sufficient smoothness of $J(\theta)$, the key condition is that boundedness of the gradient along the trajectory $\{\theta_n\}$. This condition is, for given $J(\theta)$, not straightforward to check and several approach to establishing boundedness along trajectories are provided, described in the subsequent remark.

REMARK. The following generic arguments for establishing boundedness of the gradient along trajectories are available. Assuming that the gradient of $J(\theta)$ is bounded is certainly a sufficient condition but also rather restrictive as this conditions rules out even quadratic mappings. A nontrivial example with bounded gradient is, e.g., the mapping $J(\theta) = 1-2/(2+\theta^2)$. Note that $\lim_{|\theta|\to\infty} J'(\theta) = 0$.

Alternatively, one tries to control the algorithm "away from the minimizer." For illustration, conisder $J(\theta) = \theta^2 + c$, for some constant c. The minimization problem has unique solution $\theta^* = 0$ and, by computation,

$$|\theta_{n+1}| = |\theta_n - \epsilon_n J'(\theta_n)| = |\theta_n - 2\epsilon_n\theta_n| = |\theta_n(1-2\epsilon_n)| = |\theta_n|\,|1-2\epsilon_n|.$$

So, as soon as $\epsilon_n < 1/2$ for some n, we see that $|\theta_{m+1}| < |\theta_m|$ for all $m \geq n$, and the trajectory stays inside a bounded set, which implies finiteness of the gradient along the trajectory. In Section 1.3, we will provide a more through discussion of arguments like the above.

On occasion, it is possible to measure the outcome $J(\theta)$ of the performance of a system but the gradient $\nabla J(\cdot)$ is analytically unavailable. Instead of a gradient, some methods use a finite difference approximation. More generally, suppose that the algorithm is driven by a biased approximation of the gradient:

$$\theta_{n+1} = \theta_n - \epsilon_n\big(\nabla J(\theta_n)^\top + \beta_n(\theta_n)\big), \tag{1.18}$$

where the decreasing bias terms satisfy $\beta_n(\theta_n) \to 0$. Lemma 1.2 provides an important extension of Theorem 1.3 to biased algorithms.

Lemma 1.2 *Let $J \in \mathcal{C}^2$ be such that the gradient is a Lipschitz continuous function and consider the biased algorithm :*

$$\theta_{n+1} = \theta_n - \epsilon_n(\nabla J(\theta_n)^\top + \beta_n(\theta_n)), \tag{1.19}$$

where the decreasing bias and the stepsize sequence satisfy:

$$\sum_{n=1}^{\infty}\epsilon_n = +\infty, \quad \sum_{n=1}^{\infty}\epsilon_n\,\|\beta_n(\theta_n)\| < \infty, \quad \sum_{n=1}^{\infty}\epsilon_n^2 < \infty. \tag{1.20}$$

If $\{|\nabla J(\theta_n)| : n \geq 0\}$ is bounded, then every accumulation point of $\{\theta_n\}$ is a stationary point of J.

Proof: The method of proof for this Lemma is exactly the same as for Theorem 1.3. Under the conditions put forward in Theorem 1.3, $\|\nabla J(\theta)\|^2$ is bounded and under the additional assumption that $\sum_n \epsilon_n\|\beta_n(\theta_n)\| < \infty$, we obtain that $\|\beta_n(\theta_n)\|$ tends to zero as n tends to infinity, which in turn gives that $\|\beta_n(\theta_n)\|^2$ is bounded along the sequence $\{\theta_n\}$. Consequently, the norms of the update sequence $\{|Y_n(\theta)|^2\}$ are bounded. We now revisit the Taylor series approximation in (1.16) where we

let $\theta_{n+1} - \theta_n = \epsilon_n Y_n(\theta_n)$. Argue just as in the proof of Theorem 1.3 to show that θ_n converges to a solution θ^* of $\nabla J(\theta^*) = 0$. The details are left as exercise, see Exercise **??**.

QED

EXAMPLE 1.2. Suppose that we do not know the function $J(\cdot)$ analytically, but for any point θ it is possible to obtain the numerical value of $J(\theta)$. In this situation, $\nabla J(\theta)$ is not available in closed form either. A commonly used approximation to the derivative is given by finite differences (FD), which require that $J \in \mathcal{C}^3$. In this example we will use a "centered" version of the approximation as follows. For simplicity, let $\theta \in \mathbb{R}$ and use a Taylor expansion around θ to obtain:

$$\frac{J(\theta_n + c_n) - J(\theta_n)}{2c_n} = \frac{J'(\theta_n)}{2} + \frac{1}{4}J''(\theta_n)\, c_n + \beta_+(\theta_n, c_n)$$
$$\frac{J(\theta_n) - J(\theta_n - c_n)}{2c_n} = \frac{J'(\theta_n)}{2} - \frac{1}{4}J''(\theta_n)\, c_n + \beta_-(\theta_n, c_n)$$

so that the centered, or two-sided FD satisfies:

$$\frac{J(\theta_n + c_n) - J(\theta_n - c_n)}{2c_n} = J'(\theta_n) + \beta_n(\theta_n, c_n),$$

where $\beta_n(\theta, x) = \beta_+(\theta, x) + \beta_-(\theta, x) = \mathcal{O}(x^2)$, for fixed θ. Note that the terms containing $J''(\theta_n)$ cancel out.

When implementing FD in the descent algorithm, it is necessary to show that $\lim_{n\to\infty} \beta_n(\theta_n, c_n) = 0$ to conclude that the algorithm converges to the optimal value. Note that the main problem in showing convergence lies in the fact that we do not know beforehand the sequence $\{\theta_n\}$ visited by the algorithm. To establish convergence in (1.18), we need to verify either (a) that the third derivative $J'''(\cdot)$ is uniformly bounded in θ, or (b) that θ_n remains within a compact set along the sequence, which would imply that $J'''(\theta_n)$ is uniformly bounded (as $n \to \infty$). When either (a) or (b) hold, we know that $\beta_n(\theta_n, x) \to 0$ for any sequence $\{\theta_n\}$ visited by (1.18) as long as $x \to 0$. Hence, we can choose $c_n = \mathcal{O}(n^{-c})$ for some constant $c > 0$, which implies $\beta_n(\theta_n, c_n) = \mathcal{O}(n^{-2c})$. In general the choice of c_n will depend on how fast $\epsilon_n \to 0$. Assume hat $\epsilon_n = \mathcal{O}(n^{-\gamma})$, so that (1.20) holds for $\gamma \in (0, 1]$. From Lemma 1.2 it follows that Theorem 1.3 can be extended to finite difference algorithms provided that

$$\sum_{n\geq 1} \epsilon_n \beta_n < \infty \implies \sum_{n\geq 1} n^{-(\gamma+2c)} < \infty,$$

so that we need $\gamma + 2c > 1$ for the algorithm to converge. When $\gamma = 1$, positive c is sufficient.

⁂

While gradient-based methods of type (1.19) ensure convergence for functions with only one global minimum (called "unimodal") and which are continuously differentiable functions, the rate of convergence may be much slower than Newton's method. In particular, the steepest descent method has linear convergence, i.e., there is a constant $c \in (0, 1)$ such that $\|\theta_{n+1} - \theta^*\| \leq c\|\theta_n - \theta^*\|$, whereas Newton's method in general has quadratic convergence, see Theorem 1.2. On the other hand, the gradient descent algorithm shows remarkable resilience even for distorted gradient measurements as long as the size of the distortion decreases as $n \to \infty$, as shown in Lemma 1.2.

It is worth noting that the limit θ^* can be written in algebraic form as

$$\theta^* = \theta_0 - \sum_n \epsilon_n \nabla J(\theta_n)^\top + \sum_n \epsilon_n \beta_n(\theta_n).$$

We complete this discussion by providing the equivalent statement to Theorem 1.3 for constant step size. A proof of this result can be found in [5] and we do not repeat it here.

Theorem 1.4 *Let $J \in \mathcal{C}^2$. Assume that $\nabla J(\theta)$ is has Lipschitz constant L. Consider the constant step size algorithm*

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

If $\epsilon < 2/L$ then every accumulation point of $\{\theta_n : n \geq 0\}$ is a local minimum of J, provided that θ_0 is not a stationary point.

Typically, the Lipschitz constant for the gradient is hard to bound, and one applies the algorithm for ϵ "small".

1.2 Constrained Optimization

In this section we consider the more general problem:

$$\min_{\theta \in \Theta} J(\theta), \tag{1.21}$$
$$\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0, h(\theta) = 0\}$$

for a "smooth" continuous real-valued function $J \colon \mathbb{R}^d \to \mathbb{R}$, and convex continuous functions $g \colon \mathbb{R}^d \to \mathbb{R}^p, h \colon \mathbb{R}^d \to \mathbb{R}^q$, that is, there are p inequality and q equality constraints that must be satisfied.

Definition 1.6 *The non-linear problem (1.21) is called a* (strictly) convex non-linear problem *if $J(\theta)$ and each $g_i(\theta), i = 1, \ldots, p$, are (strictly) convex, and each $h_j(\theta), j = 1, \ldots, q$, is an affine function (linear plus a constant).*

When we want to stress that a gradient or a Hessian is taken with respect to θ in a mapping with more arguments, we write ∇_θ and ∇^2_θ, respectively.

Definition 1.7 *The* Lagrangian $\mathcal{L} \colon \mathbb{R}^d \times \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$ *associated with the problem (1.21) is defined as:*

$$\mathcal{L}(\theta, \lambda, \eta) = J(\theta) + \lambda^\top g(\theta) + \eta^\top h(\theta). \tag{1.22}$$

Definition 1.8 *A constraint g_i is said to be* active at a feasible point $\theta \in \Theta$ *if $g_i(\theta) = 0$. Otherwise it is said to be* inactive. *The set $A(\theta)$ of active constraints at θ contains all indices i for which $g_i(\theta) = 0$. The* constraint qualification *condition for problem (1.21) at a feasible point θ requires that the set of vectors $\{\nabla_\theta g_i(\theta), i \in A(\theta); \nabla_\theta h_j(\theta), j = 1, \ldots, q\}$ be linearly independent, and that there exist a vector $v \in \mathbb{R}^d$, $v \neq 0$, such that:*

(a) $\nabla h_j(\theta) v = 0, \quad 1 \leq j \leq q,$

(b) for all $i \in A(\theta)$ it holds that $\nabla g_i(\theta) v < 0$.

Definition 1.9 *A* stationary point $(\theta^*, \lambda^*, \eta^*)$ *of (1.21) is a point that satisfies the* first order *Karush Kuhn-Tucker (KKT) conditions if*

$$\nabla_\theta \mathcal{L}(\theta^*, \lambda^*, \eta^*) = 0 \tag{1.23a}$$
$$\nabla_\lambda \mathcal{L}(\theta^*, \lambda^*, \eta^*) = g(\theta^*)^\top \leq 0, \lambda^* \geq 0, \text{ and } \forall i : \lambda_i^* g_i(\theta^*) = 0 \tag{1.23b}$$
$$\nabla_\eta \mathcal{L}(\theta^*, \lambda^*, \eta^*) = h(\theta^*)^\top = 0; \tag{1.23c}$$

where $\nabla_\lambda \mathcal{L}(\theta, \lambda, \eta)$ denotes the gradient of $\mathcal{L}(\theta, \lambda, \eta)$ with respect to λ and $\nabla_\eta \mathcal{L}(\theta, \lambda, \eta)$ the gradient with respect to η.

Condition (1.23b) is called the *complementary slackness* property, from this property it follows that $i \notin A(\theta)$ implies $\lambda_i = 0$.

Theorem 1.5 *Let $J(\theta), g(\theta), h(\theta) \in \mathcal{C}^1$ and assume that the constraint qualification holds at a local minimum θ^* of $J(\theta)$ in (1.21). Then there exist $\lambda^* \in \mathbb{R}^p, \eta^* \in \mathbb{R}^q$, such that $(\theta^*, \lambda^*, \eta^*)$ is a stationary point. The vectors λ^* and η^* are called* Lagrange multipliers.

If in addition, if the problem is a strictly convex problem, then the KKT conditions hold at θ^* if and only if θ^* is the global minimum.

EXAMPLE 1.3. Many canned products in the supermarket come in cans of similar shape, where the height is the same as the diameter of the container. What is the reason for this? Allegedly, a similar question haunted Galileo about the leather bags used by traders. Here is the answer: if a fixed volume of a given good has to be canned, the containers should be produced at minimal cost (in particular using minimal amount of material), in order to maximize your profit.

This problem can be formulated as a surface minimization problem under the fixed volume constraint. Call $\theta = (r, y)^T$, where r is the radius and y is the height of the (cylindrical) can. Then we want to find:

$$\begin{aligned} \min_{r,h} J(\theta) &\stackrel{\text{def}}{=} 2(\pi r^2) + 2\pi r\, y \\ \text{subject to:} \qquad & \pi r^2 y = V, \end{aligned}$$

where we have expressed the total surface as the rectangular surface for the side of the can, plus the two covers. The volume V is fixed. Call $h(\theta) = \pi r^2 y - V$.

We will show how to apply Theorem 1.5 in practice. The problem fails to be convex, as neither is $J(\theta)$ convex nor is h affine, and the second part of the theorem cannot be used. Instead, we proceed as follows. First, we find the KKT points that satisfy (1.23), and then we determine which one (if several) is the global optimizer. The Lagrangian is:

$$\mathcal{L}(\theta; \eta) = \mathcal{L}(r, y; \eta) = 2(\pi r^2) + 2\pi r\, y + \eta(\pi r^2 y - V).$$

Condition (1.23a) for a KKT points reads:

$$\frac{\partial}{\partial r}\mathcal{L}(r, y; \eta) = 4\pi r + 2\pi y + \eta\, 2\pi r y = 0 \tag{1.24}$$

$$\frac{\partial}{\partial y}\mathcal{L}(r, y; \eta) = 2\pi r + \eta\, \pi r^2 = 0. \tag{1.25}$$

From the second equality we get $\eta^* = -2/r^*$, replacing this value in the first we get: $2r + y - 2y = 2r - y = 0$, so that $y^* = 2r^*$, which is the actual proportion found in many commercial cans.

To illustrate the mathematical method, we will finish the example. Using (1.23c), i.e., $h(\theta) = 0$, we replace $y = V/\pi r^2$ to obtain the actual solution to (1.23), namely $(r^*)^3 = V/2\pi$ and $y^* = 2r^*$. The constraint qualification holds at this (unique) KKT point. Indeed there is only one constraint, and it satisfies:

$$\nabla h\,(\theta) = (2\pi r y,\ \pi r^2),$$

which is non zero at r^*, y^*, as required. Observe that taking $v^\top = (r/2, -y)$ yields $\nabla h(\theta)\, v = 0$.

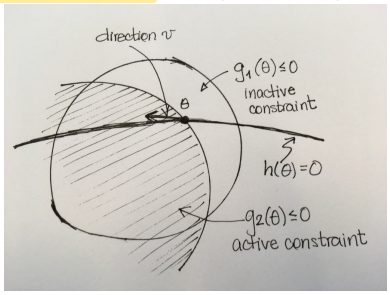
Because this is the only KKT point, it is the only candidate for the solution. To see that the KKT point is indeed a local minimum, note that $r \in (0, \infty)$ and as r either tends to 0 or to ∞, the value of $J(c, r)$ tends to ∞, so that we can conclude that J has to have a minimum for some value of $r \in (0, \infty)$. Since the only candidates for the location of a minimum are the KKT points, it follows from the uniqueness of the solution, that the KKT point is the location of the global minimum.

The theorem below provides the second order conditions that help in determining if a KKT point is indeed a local minimum along the feasible set under no convexity.

REMARK. Of course this example is academic, in order to illustrate the use of the theory. A more direct solution is readily obtained by direct substitution $y = V/\pi r$ into J to obtain a function of only one variable $f(r) = 2\pi(r^2 + V/\pi r^2)$. That this is convex follows from $f'(r) = 2\pi(2r - V/\pi r^2)$, and $f''(r) = 2\pi(2 + 2\,V/\pi r^3) > 0$ for all $r > 0$. The unique zero of $f'(r), r \geq 0$ is exactly at r^*.

⁂

Definition 1.10 *Let $(\theta^*, \lambda^*, \eta^*)$ be a stationary point of (1.21). The* critical cone $\mathbf{C}(\theta^*, \lambda^*)$ *is:*



$$\mathbf{C}(\theta^*, \lambda^*) = \Big\{ v \in \mathbb{R}^d \colon \nabla g_i(\theta^*) v \leq 0, \text{ if } i \in A(\theta^*), \lambda_i^* = 0, \\ \nabla g_i(\theta^*) v = 0, \text{ if } \lambda_i^* > 0, \\ \nabla h(\theta^*) v = 0 \Big\}.$$

This cone defines the set of directions v that move along the active and equality constraints, as well as those that move "inside" the feasible set if the active constraint has a null multiplier.

Theorem 1.6 *Suppose that $J(\theta), g(\theta), h(\theta) \in \mathcal{C}^2$ and that the constraint qualifications hold for $g(\theta), h(\theta)$ at θ^*. If $(\theta^*, \lambda^*, \eta^*)$ is a stationary point, and for all $0 \neq v \in \mathbf{C}(\theta^*, \lambda^*)$, then the following* second order condition *holds:*

$$v^\top \nabla^2_\theta \mathcal{L}(\theta^*, \lambda^*, \eta^*) v > 0, \tag{1.26}$$

then θ^* is a local minimum of (1.21), where $\nabla^2_\theta \mathcal{L}$ denotes the Hessian of $\mathcal{L}$ with respect to θ.

Note that if the domain $\{\theta \in \mathbb{R}^d : g_i(\theta) \leq 0, 1 \leq i \leq p; h_j(\theta) = 0, 1 \leq j \leq q\}$ is compact, then the use of the second order condition can be avoided as continuity of $J(\theta)$ already implies existence of a global maximum and minimum on a compact set. Evaluating all stationary points then solves the optimization problem. See, for example, [6]. Lagrange multipliers frequently have an interpretation in practical contexts. In Economics, they can often be interpreted in terms of prices for constraints, while in Physics they can represent concrete physical quantities. Mathematically, Lagrange multipliers can be viewed as *rates of change* of the optimal cost as the level of constraint changes. These type of results are called *envelop theorems* in the literature. Next we state a typical envelop theorem.

Theorem 1.7 *Assume that $J(\theta) \in \mathcal{C}^2$ is convex and consider (1.21) with no inequality constraints ($p = 0$), and let (θ^*, η^*) be a local minimum and Lagrange multiplier (LM), respectively, satisfying the KKT conditions and condition (1.26). Consider the family of continuous non-linear problems*

$$\min J(\theta), \theta \in \mathbb{R}^d \tag{1.27}$$

$$s.t. \quad h(\theta) = u \tag{1.28}$$

parameterized by $u \in \mathbb{R}^q$. Then there exists an open sphere S centered at $u = 0$ such that for every $u \in S$, there exist $\theta(u) \in \mathbb{R}^d$ and $\eta(u) \in \mathbb{R}^q$ which are a local minimum and LM for the corresponding problem. Furthermore, $\theta(u), \eta(u)$ are continuously differentiable functions within S and we have $\theta(0) = \theta^*, \eta(0) = \eta^*$. In addition, for all $u \in S$,

$$\nabla_u F(u) = -\eta(u),$$

where $F(u) = J(\theta(u))$ is the optimal cost of the problem at value u.

In the case of inequality constraints, evidently $\{\theta \colon g(\theta) \leq 0\} \subset \{\theta \colon g(\theta) \leq u\}$ for $u > 0$. Thus, the optimal cost value of the modified problem must satisfy $F(u) \leq F(0)$, for $F(u)$ defined as in Theorem 1.7. For all inactive inequality constraints, $\lambda_i = 0$, and for all active constraints, $\lambda_i > 0$, indicating a potential marginal *decrease* in the cost function as a result of increased resources.

EXAMPLE 1.4. A company has a budget of $\$10,000$ for advertising, all of which must be spent. It costs $\$3,000$ dollars per minute to advertise on television and $\$1,000$ per minute to advertise on radio. If the company buys x minutes of television advertising and y minutes of radio advertising, its revenue in thousands of dollars is determined by the company's data mining oracle/statistician to be reasonably approximated by the function

$$f(x, y) = -2x^2 - y^2 + xy + 8x + 3y.$$

We can find the best solution to maximise profit solving the minimisation problem:

$$\begin{aligned} \min_{x,y \in \mathbb{R}} \quad & f(x, y) = 2x^2 + y^2 - xy - 8x - 3y \\ \text{s.t.} \quad & h(x, y) = 3x + y - 10 = 0 && (1) \\ & g_1(x, y) = -x \leq 0 && (2) \\ & g_2(x, y) = -y \leq 0 && (3), \end{aligned}$$

where f and h are expressed in units of thousands of dollars. The Lagrangian is

$$\mathcal{L}(x, y, \lambda, \eta) = 2x^2 + y^2 - xy - 8x - 3y + \lambda_1(-x) + \lambda_2(-y) + \eta(3x + y - 10)$$

and $\nabla_{(x,y)}\mathcal{L}(x, y, \lambda, \eta) = (4x - y - 8 - \lambda_1 + 3\eta, 2y - x - 3 - \lambda_2 + \eta)^T$. By the first KKT condition, a local minimum (x^*, y^*) satisifies $\nabla_{(x,y)}\mathcal{L}(x^*, y^*, \lambda^*, \eta^*) = 0$, which gives the following simultaneous equations

$$\begin{aligned} 4x^* - y^* - 8 - \lambda_1^* + 3\eta^* &= 0 \\ 2y^* - x^* - 3 - \lambda_2^* + \eta^* &= 0\,. \end{aligned}$$

There are four combinations of $g_1(x)$ and $g_2(x)$ being active/inactive.

Suppose both inequality constraints are inactive, so that complementary slackness gives $\lambda_1^* = \lambda_2^* = 0$. Together with the equality constraint, this gives three equations in three unknowns x^*, y^*, η^*. Their solutions yields the KKT point $(x^*, y^*, \lambda_1^*, \lambda_2^*, \eta^*)^\top = (\frac{69}{28}, \frac{73}{28}, 0, 0, \frac{1}{4})^\top$. This point satisfies a constraint qualification since the function $h(x, y)$ is linear, so it is a KKT point, and is thus a candidate for a local minimum. Furthermore, we have

$$\nabla^2 f(x, y) = \begin{pmatrix} 4 & -1 \\ -1 & 2 \end{pmatrix},$$

which is positive definite, and therefore is positive semi-definite (a sufficient condition for a function to be convex), thus f is convex and the KKT point is the unique global minimum of f, and is therefore the unique global maximum of the original maximisation problem. The company can therefore maximise its revenue by purchasing $\frac{69}{28}$ minutes of television time and $\frac{73}{28}$ minutes of radio time. Since we have found the unique global maximum of the optimisation problem, we do not need to search for any other KKT points.

Now suppose you have in front of you this solution and the company boss puts you "on the spot" during a meeting and asks for an estimate of the extra revenue which would be generated if she spent an extra $\$1000$ on advertising, what would be a reasonable answer ?

Instead of solving again the problem with the budget changed to $\$11000$, you can use Theorem 1.7: $-\eta$ is the instantaneous rate of change of the minimum cost function value $F(u)$ as a function of the change in the level of constraint. Here $-\eta^* = -\eta(0) = -0.25$. In terms of the original maximization problem, this translates to an *increase* of 250 dollars to the maximum revenue that can be generated if the the advertising budget is increased by 1000 dollars. Thus, knowing $\eta^* = .25$ will be enough for you to answer promptly "Madam, an extra expense of $1000 can only provide an extra revenue around $250. Actually, we would be better off *decreasing* the advertising budget".

⁂

Numerical Methods

Consider (1.21) again, that is,

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

It should be apparent that, even for seemingly small dimensions, finding all KKT points of a constrained non-linear problem may be an infeasible task. As in the case of unconstrained optimization, one often uses numerical iterative procedures to approximate the solution. We will now mention some of the methods that extend the simple recursive procedure (1.19). The main idea of the methods is to either approximate or reformulate the problem in terms of unconstrained optimization and then use an appropriate numerical algorithm.

Penalty Methods. These methods modify the original performance function to penalize the extent to which the constraints are not satisfied. Recall that $|\cdot|$ denotes the Euclidean norm, then the penalized function is defined:

$$J_\alpha(\theta) = J(\theta) + \frac{\alpha}{2}\big(|g(\theta)_+|^2 + |h(\theta)|^2\big),$$

where $g(\theta)_+ = (g_1(\theta)_+, \ldots, g_j(\theta)_+)^\top$, and $g_i(\theta)_+ = \max(0, g_i(\theta))$.

Theorem 1.8 *Suppose that $J(\theta), g(\theta), h(\theta) \in \mathcal{C}^1$ and let $\{\alpha_n\}$ be an increasing sequence such that $\lim_n \alpha_n = \infty$. For α_n given, let θ_n be the location of the minimum of $J_{\alpha_n}(\theta)$, i.e., let $\theta_n = \arg\min J_{\alpha_n}(\theta)$. If $\{\theta_n\}$ has an accumulation point θ^* and a constraint qualification holds at θ^*, then θ^* is (feasible and) stationary for the non-linear problem (1.21). Moreover, if λ^*, η^* are the Lagrange multipliers for θ^*, then*

$$\lambda^* = \lim_{n\to\infty} \alpha_n (g(\theta_n))_+, \quad \eta^* = \lim_{n\to\infty} \alpha_n h(\theta_n),$$

and for each $i = 1, \ldots, j$, $\lambda_i^* \geq 0$ and $\lambda_i^* = 0$ if $g_i(\theta^*) < 0$.

Numerical methods that use Theorem 1.8 typically use $\theta_{n+1} = \theta_n - \epsilon_n \nabla J_{\alpha_k}(\theta_n)^\top$, with $n = 1, 2, \ldots$, for minimizing $J_{\alpha_k}(\theta)$ with respect to θ where α_k is kept constant for a number of iterations, and then it is increased. Specifically, let T_i denote the index of the i-th update in $\{\alpha_k\}$, and suppose that $(T_{i+1} - T_i) \to \infty$. Then consider the algorithm:

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla J_{\alpha_n}(\theta_n)^\top \tag{1.29a}$$

$$\alpha_{n+1} = \alpha_n + \delta_n \mathbf{1}_{\{n \in \{T_i\}\}} \tag{1.29b}$$

where $\sum \delta_n = +\infty$, and

$$\nabla J_{\alpha_n}(\theta_n) = \nabla_\theta J(\theta_n) + \alpha_n \big(g(\theta_n)^\top \nabla g(\theta_n) \mathbf{1}_{\{\|g(\theta_n)\| > 0\}} + h(\theta_n)^\top \nabla h(\theta_n) \big). \tag{1.30}$$

Under appropriate conditions, the sequence θ_n will converge to the constrained optimum. Different schemes yield different overall rates of convergence. Alternatively, one can introduce a two-time scale method. Let $T_i = i$ and suppose that δ_n grows in such a controlled way that α_n "looks" constant for the iteration in θ_n when using Taylor expansions. For example, if $\delta_n = \epsilon_n$ (i.e., $\{\delta_n\}$ is equal to the gain sequence), we know that $\alpha_n \to \infty$ under (1.20), yet from one iteration to the next, the contribution of the term δ_n is in $\mathcal{O}(\epsilon_n^2)$. The two-time scale algorithm is:

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla J_{\alpha_n}(\theta_n)^\top \tag{1.31a}$$

$$\alpha_{n+1} = \alpha_n + \delta_n, \tag{1.31b}$$

with $\delta_n \epsilon_n \to 0$, $\sum \delta_n = +\infty$, and $\nabla J_{\alpha_n}(\theta_n)$ as in (1.30). Convergence of this method may be much slower than (1.29) because α_n grows very slowly. On the other hand, while δ_n can be even increasing in (1.29), it is necessary that $T_{i+1} - T_i$ grows so as to ensure convergence.

Projection Methods. Gradient projection methods iterate successive solutions in the direction of improvement of the cost function (descent directions), but *remaining always feasible*. The algorithm is in general form:

$$\tilde{\theta}_{n+1} = \theta_n - \epsilon_n \nabla J(\theta_n)^\top \tag{1.32}$$

$$\theta_{n+1} = \Pi_\Theta \Big(\tilde{\theta}_{n+1} \Big), \tag{1.33}$$

where $\Pi_\Theta(v)$ is the projection of the vector $v \in \mathbb{R}^d$ onto the set Θ. The projection version of the gradient descent algorithm then reads

$$\theta_{n+1} = \theta_n - \epsilon_n Z(\theta_n), \tag{1.34}$$

where

$$\nabla J(\theta_n)^\top = Z(\theta_n)$$

if $\theta_n - \epsilon_n \nabla J(\theta_n)^\top \in \Theta$ and the projection on Θ otherwise.

The actual evaluation of the projection operation is usually the main computational burden for each step in the algorithm. See, for example, [7], where the projection onto a simplex is provided. The simplest case is that of projection on a hypercube, which is detailed in the following example.

EXAMPLE 1.5. In case Θ is a d-dimensional hypercube, i.e., $\Theta = [-M, M]^d$ for some finite M, the projection is easily obtained through

$$\Pi_\Theta(\theta) = \Big(\max(\theta_i, -M) \mathbf{1}_{\{\theta_i \le 0\}} + \min(\theta_i, M) \mathbf{1}_{\{\theta_i \ge 0\}}) : 1 \le i \le d \Big)^\top.$$

In the special case of the projection on a hypercube we call the projection *truncation*. We call the constraint set Θ *box constraints*.

⁂

In the follwong we consider the slightly more restricitive case of a convex set $\Theta \subset \mathbb{R}^d$ such that there exist differentiable mappings $g_i : \Theta \to \mathbb{R}$, $1 \leq i \leq p$, characterizing Θ, i.e., $\theta \in \Theta$ if and only if $g_i(\theta) \leq 0$, for $1 \leq i \leq p$. The constraint optimization problem (1.21) becomes

$$\begin{aligned} &\min_{\theta \in \Theta} J(\theta), \\ &\Theta = \{\theta \in \mathbb{R}^d \colon g(\theta) \leq 0\} \end{aligned} \tag{1.35}$$

and solutions are characterized by the KKT conditions.

Before stating (a version) of the convergence result, we will motivate the result by the following consideration. Suppose that $\{\theta_n\}$ is obtained via (1.32), then the following cases can occur: (i) the minimizer θ^* is an inner point of Θ and the algorithm will (after possibly finitely many projections) stay inside Θ and will behave just like the unconstraint version; (ii) the **unconstrained** minimizer $\tilde{\theta}^*$ lies outside of Θ (or on the boundary of Θ) and the algorithm will eventually converge to θ^* on the boundary of Θ closest to $\tilde{\theta}^*$; and finally (iii) the problem may be ill-posed so that θ_n has no accumulation points at all (e.g., minimizing $J(\theta) = -\theta^2$.) Note that case (iii) is ruled out if we assume Θ to be bounded. Before turning to the proof of the algorithm, we provide some details on case (ii). Suppose that θ^* lies outside Θ, and suppose that the algorithm reaches a point θ' such that $Z(\theta') = 0$. Assume, for simplicity, that only one constraint g_i is active at θ', i.e., $g(\theta') := g_i(\theta') = 0$. Then, the descent direction in θ' is $-\nabla J(\theta')$ and since the projection of $-\nabla J(\theta')$ onto the surface $g(\theta) = 0$ is zero, this means that $-\nabla J(\theta')$ is perpendicular to the tangent line of $g(\theta)$ at θ'. As $\nabla g(\theta')$ is a normal vector for the tangent line, it follows that $\nabla g(\theta')$ and $-\nabla J(\theta')$ are co-linear, i.e., there exists $\lambda \neq 0$ such that $\lambda \nabla g(\theta') = -\nabla J(\theta')$. For background on the geometrical interpretation of the gradient we refer to Section A.1 in the Appendix.

INSERT FIGURE HERE

To summarize, for $\nabla J(\theta') \neq 0$ and $g(\theta') = 0$, $Z(\theta') = 0$ implies $\lambda \nabla g(\theta') = -\nabla J(\theta')$, which shows that θ' is a KKT point for (1.35) and under appropriate smoothness conditions a local minimizer for (1.35). We now present the theorem.

Theorem 1.9 *Consider (1.35) with Θ being a compact convex set, and let $J(\theta) \in \mathcal{C}^2$ with L denoting the uniform Lipschtiz constant of ∇J on Θ. If either*

$$\sum_n \epsilon_n = \infty \quad \textit{and} \quad \sum_n \epsilon_n^2 < \infty$$

or $\epsilon_n = \epsilon$ for all n, with $\epsilon < 2/L$, then every accumulation point of $\{\theta_n\}$ is the location of a local minimum of problem (1.32).

Proof: As Θ is closed and bounded, then by the Bolzano-Weierstrass Theorem, $\{\theta_n\}$ has accumulation points.

Let θ^* be an accumulation point of $\{\theta_n\}$ and assume that θ^* is an inner point of Θ. Let $\theta_m := \theta_{n_m}$ denote the sub-sequene converging towards θ^*. Then, for N sufficiently large, $\theta_m \in \hat{\Theta}$, for $m \geq N$,

for some compact proper subset $\hat{\Theta}$ of Θ (i.e., $\hat{\Theta}$ contains no boundary points of Θ). Continuity of the gradient and the Hessian, implies that the gradient as well as the Hessian are bounded on $\hat{\Theta}$. We now apply the arguments put forward in the proof of Theorem 1.3 for the decreasing ϵ case and Theorem 1.4 for the fixed ϵ case, to show that

$$\lim_{m\to\infty} \nabla J(\theta_m) = 0 = \nabla J(\theta^*),$$

which shows that θ^* is a stationary point of $J(\theta)$. We have assumed that θ^* is an inner point of Θ, so that $g_i(\theta^*) < 0$ for all i, and it follows that θ^* is a KKT point for (1.35).

Now consider the case that θ^* lies on the boundary of Θ. Convergence of θ_m implies that $\|Z(\theta_n)\|$ converges towards zero. Note that projection on a convex set is continuous; see Exercise 1.5. By continuity of gradient and projection it holds that

$$\lim_{m\to\infty} \|Z(\theta_m)\| = \|Z(\theta^*)\| = 0. \tag{1.36}$$

From $Z(\theta^*) = 0$ we conclude that either $\nabla J(\theta^*) = 0$ and $g_i(\theta) = 0$, or $\nabla J(\theta^*) \neq 0$ in which case the negative gradient points outwards from Θ. For the projection to become the zero vector, the negative gradient has to be perpendicular to the g_i's at θ^*. This implies that $-\nabla J(\theta^*) = \lambda_i \nabla g_i(\theta^*)$ for some constants λ_i where i runs trough the indicies of the active constraints. This shows that θ^* is a KKT point for (1.35).

For the decreasing ϵ we take N such that $\epsilon_n \leq 2/L$ for $n \geq N$, and we use $\epsilon_n \|Z(\theta_n)\| \leq \epsilon \|Z(\theta_n)\|$ for $n \geq N$. The proof then follows from (1.36).

To conclude the proof, we evoke Theorem 1.5, to show that any KKT point for (1.35) is the location of a local minimum for the optimization problem (1.35).

QED

REMARK. While practically of limited value, the projection method provides an analytically attractive tool. Indeed, many technical assumptions become much less restrictive if the algorithm is applied only on a bounded set. For example, the condition that the gradient is bounded along trajectories in Theorem 1.3 rules out even a quadratic form of $J(\theta)$, which is certainly not desirable. However, we may consider the version of our algorithm restricted to some hypothetical large hypercube. Then, we switch for theoretical arguments to the projection version of the algorithm while in practice use the unconstrained version.

We conclude this section on the projection method with a discussion on finding (approximate) projections. In the case that the constraint function g is affine, Example 4.3 in Chapter 3 of [5] shows that the dual of this problem leads to a much simpler optimization problem with only positivity constraints. We now refer to [31] where a method is proposed for general $g \in \mathcal{C}^2$ using the fact that $\theta_n \in \Theta$, and $\tilde{\theta}_{n+1} - \theta_n$ is of order ϵ_n so a Taylor approximation can be used to linearize the constraints around θ_n:

$$g(x) \approx g(\theta_n) + \nabla g(\theta_n)(x - \theta_n).$$

Let $v = \tilde{\theta}_{n+1}$. The constraint $x \in \Theta$ is approximated by the constraint

$$g(\theta_n) + \nabla g(\theta_n) x \leq \nabla g(\theta_n)\theta_n.$$

Call $\mathbb{A} = \nabla g(\theta_n)^T$, and $b = \mathbb{A}\theta_n - g(\theta_n)$. The approximated, or "surrogate" subsidiary problem becomes

$$\min_x \left(\frac{1}{2} x^\top x - v^\top x \right), \tag{1.37}$$

$$\text{s.t.} \quad \mathbb{A}x \leq b, \tag{1.38}$$

The Lagrangian for this subsidiary problem is:

$$\mathcal{L}(x, \mu) = \frac{1}{2} x^\top x - v^\top x + \mu^\top \mathbb{A} x - \mu^\top b.$$

Using Lagrange duality (Theorem 1.11), we seek $\max_{\mu \geq 0} \left(\min_{x \in \mathbb{R}^d} \mathcal{L}(x, \mu) \right)$. Because of the quadratic form, we can solve the minimisation step analytically by setting the gradient to zero, which readily yields $x^*(\mu) = v - \mathbb{A}^\top \mu$. Then the subsidiary problem is:

$$\max_{\mu \geq 0} \left(\frac{1}{2} (v - \mathbb{A}^\top \mu)^\top (v - \mathbb{A}^\top \mu) - b^\top \mu \right),$$

which, after replacing the appropriate values, gives another quadratic maximization problem with a projection to the positive real numbers. Finding the zero of the derivative of this function, however, requires now inversion of matrices depending on $g(\theta_n)$ and $\nabla g(\theta_n)$, which is generally computationally expensive. Instead, [31] propose a recursive gradient method to solve for μ. With this, one sets $\theta_{n+1} = v - \mathbb{A}^\top \mu^* = \tilde{\theta}_{n+1} - \nabla g(\theta_n)^\top \mu^*$.

Let T_i denote the index of the i-th update in η, and suppose that $(T_{i+1} - T_i) \to \infty$. Then the algorithm is:

$$\tilde{\theta}_{n+1} = \theta_n - \epsilon_n \nabla J(\theta_n) \, \mathbf{1}_{\{n \in \{T_i\}\}} \tag{1.39a}$$

$$\mu_{n+1} = \max \left(0, \mu_n + \delta_n \left(\nabla g(\theta_n) \nabla g(\theta_n)^\top \mu_n + \nabla g(\theta_n)(\tilde{\theta}_{n+1} - \theta_n) - g(\theta_n) \right) \right) \tag{1.39b}$$

$$\theta_{n+1} = \theta_n \, \mathbf{1}_{\{n \notin \{T_i\}\}} + \left(\tilde{\theta}_{n+1} - \nabla g(\theta_n) \mu_{n+1} \right) \mathbf{1}_{\{n \in \{T_i\}\}}. \tag{1.39c}$$

Because of the approximation of the non-linear constraint, this new point may be infeasible, but under appropriate conditions on $T_i, \epsilon_n, \delta_n$, this algorithm will converge. It is also possible to adapt this algorithm to a two-time scale version.

Multiplier Methods. These methods are based on the following result for equality constraint problems; for a proof we refer to [5].

Theorem 1.10 *Consider an equality constrained problem. Let*

$$\theta_n^* = \arg\min_\theta \mathcal{L}(\theta, \eta_n)$$

$$\eta_{n+1} = \eta_n + \rho_n h(\theta_n^*),$$

for a sequence $\rho_n \to \infty$, then $(\theta_n^*, \eta_n) \to (\theta^*, \eta^*)$ a local minimum and a KKT point of (1.21).

The *inexact* multipliers methods use an approximation to θ_n^* via Theorem 1.3. Let T_i denote the index of the i-th update in η, and suppose that $(T_{i+1} - T_i) \to \infty$. Then the algorithm is:

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla_\theta \mathcal{L}(\theta_n, \eta_n)^\top = \theta_n - \epsilon_n \left(\nabla_\theta J(\theta_n)^\top + \nabla_\theta h(\theta_n) \eta_n \right) \tag{1.40a}$$

$$\eta_{n+1} = \eta_n + \rho_n h(\theta_n) \mathbf{1}_{\{n \in \{T_i\}\}}. \tag{1.40b}$$

This method can be applied to inequality constraints as well, via a transformation (see [5]).

A "two time-scale" algorithm can be implemented here, as for the penalty method, updating at every iteration, but making ρ_n grow "slower" than ϵ_n decreases so that the primal variable behaves locally in bounded intervals as if it was driven with a constant dual variable.

Lagrange Duality Methods. In both penalty and multiplier methods, the theory establishes convergence only when an exact minimization takes place for given multiplier values. The numerical approximations often use inexact minimization by updating the decision variable θ_n for T_n iterations, and then updating the multipliers. However, there is no guarantee that the algorithm will converge, and it is not clear how to tune the parameter T_n for better convergence.

An important class of methods is based on Lagrange Duality Theory. It is straightforward to note that the solution to (1.21) is the same as the solution of the minmax problem:

$$\min_{\theta\in\mathbb{R}^d}\max_{\lambda\geq 0,\eta}\mathcal{L}(\theta;\lambda,\eta)=\min_{\theta\in\mathbb{R}^d}\begin{cases}J(\theta) & \text{if } g(\theta)\leq 0, h(\theta)=0\\ +\infty & \text{otherwise.}\end{cases}$$

However the above minmax problem is clearly not useful for an iterative algorithm. Instead, we use the following strong result.

Theorem 1.11 (Saddle Point Theorem) *If (1.21) is a convex NLP, then a triplet $(\theta^*,\lambda^*,\eta^*)$ is a KKT point if and only if it is a saddle point of the Lagrangian, that is,*

$$\mathcal{L}(\theta^*,\lambda,\eta)\leq\mathcal{L}(\theta^*,\lambda^*,\eta^*)\leq\mathcal{L}(\theta,\lambda^*,\eta^*)$$

for every $\theta\in\mathbb{R}^d,\lambda(\geq 0)\in\mathbb{R}^p,\eta\in\mathbb{R}^q$. Furthermore,

$$\min_{\theta\in\mathbb{R}^d}\max_{\lambda\geq 0,\eta}\mathcal{L}(\theta;\lambda,\eta)=\max_{\lambda\geq 0,\eta}\min_{\theta\in\mathbb{R}^d}\mathcal{L}(\theta;\lambda,\eta).$$

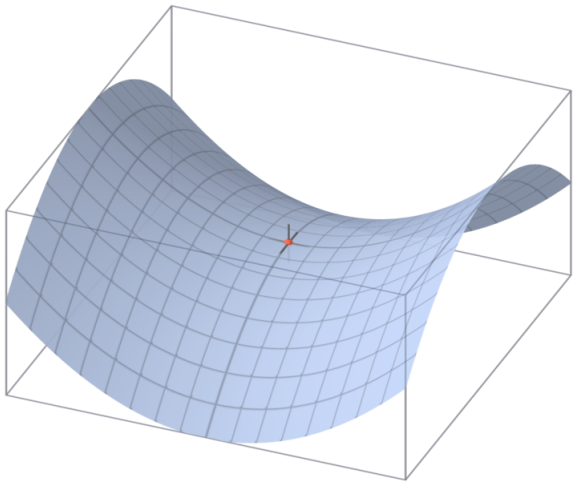


Figure 1.6: Saddle point illustration, for θ,λ on the axes, no equality constraints.

The saddle point Theorem can be used to maximize first over the multipliers, and then perform a minimization over the decision variables. This is the motivation for [30] the *Uzawa* algorithm:

$$\theta_{n+1} = \arg\min_{\theta} \mathcal{L}(\theta, \lambda_n, \eta_n) \tag{1.41a}$$

$$\lambda_{n+1} = \lambda_n + \max(0, \lambda_n + \epsilon_n \nabla_\lambda \mathcal{L}(\theta_{n+1}, \lambda_n, \eta_n)^\top) \tag{1.41b}$$

$$\eta_{n+1} = \eta_n + \epsilon_n \nabla_\eta \mathcal{L}(\theta_{n+1}, \lambda_n, \eta_n)^\top, \tag{1.41c}$$

where $\nabla_\lambda \mathcal{L}(\theta, \lambda, \eta)^\top = g(\theta)$ and $\nabla_\lambda \mathcal{L}(\theta, \lambda, \eta)^\top = h(\theta)$. The $\max(0, \cdot)$ is a component-wise max operation on the vector.

Instead of exact minimization, the so-called *Arrow-Hurwicz iterative algorithm* [3] can be used for convex NLP's:

$$\theta_{n+1} = \theta_n - \epsilon_n \nabla_\theta \mathcal{L}(\theta_n, \lambda_n, \eta_n)^\top \tag{1.42a}$$

$$\lambda_{n+1} = \max\left(0, \lambda_n + \epsilon_n \nabla_\lambda \mathcal{L}(\theta_n, \lambda_n, \eta_n)^\top\right) \tag{1.42b}$$

$$\eta_{n+1} = \eta_n + \epsilon_n \nabla_\eta \mathcal{L}(\theta_n, \lambda_n, \eta_n)^\top, \tag{1.42c}$$

with

$$\nabla_\theta \mathcal{L}(\theta, \lambda, \eta) = \nabla J(\theta) + \lambda^\top \nabla g(\theta) + \eta^\top \nabla h(\theta),$$

and where the $\max(0, \cdot)$ is again the component-wise max operation on the vector.

We will summarize the convergence properties of this algorithm in the following section.

REMARK. As with the unconstrained methods, the above numerical methods can be (and are often) implemented with constant step sizes ($\epsilon_n \equiv \epsilon$, $\delta_n \equiv \delta$ and $\rho_n \equiv \rho$). The following chapter presents general iterative algorithms and provides the analysis of the behavior of such algorithms both for constant and decreasing step sizes.

1.3 Practical Considerations

All the assumptions in the theorems are there for a reason. However, some conditions are hard to verify analytically. In this section we illustrate how modeling/meta-arguments can be used to justify that the conditions are verified.

Any optimization algorithm can only be successful if the optimization problem is well-posed, and part of the conditions in the theorems are there to ensure exactly this: there actually is a solution. For example, if $J(\theta)$ is convex, then the minimization problem is well-posed and posses a unique global solution. For more general mappings, the theory identifies analytical conditions that characterize candidates for the global solution (stationary points, KKT conditions). Additional arguments are then typically required to identify the nature of a candidate. There is no one sizes fits all theory/algorithm, and building meaningful models that are well-posed for optimization is what constitutes the *art of modeling*.

Generally speaking it is recommendable to first get a good impression of the behavior of the algorithm by comparing several trajectories. For this, start the algorithm for several randomized initial values and let the algorithm run for some time. With today's computer technology, computing many trajectories in parallel is feasible. The following cases can be distinguished:

- If the trajectories tend to the same point, it is a good guess that the trajectories converged to the solution of the problem.
- If the limit point seems to be dependent on the initial value, then the problem may have several local minima (constrained or unconstrained) in which case finding the overall solution requires methods for exploration (which are outside the scope of this book). In practice one explores the best solution (e.g., by evaluating the cost function at neighboring points).
- If some of the trajectories seem to wander off to infinity, this indicates that the algorithm is numerically unstable which indicates that the problem might be ill-posed.

The golden rule in optimization can be phrased like this: *When something goes wrong in the numerical tests your main suspects for explaining the wrongful behavior are those conditions in the theorems which you were not able to establish thoroughly. In solving real problems, mathematical theory and the art of modeling complement each other, neither one dominates the other. Know your theory very well, but do not be restrained by all the technical and often hard to prove conditions. Beware of minimizing* $-\theta^2$.

In the following we address the two conditions for validity of the algorithms that deserve some further elaboration: the choice of the gain sequence, and establishing boundedness of the gradient along trajectories.

Choice of the Gain Sequence

The key issue is here that the stepsize should be such that it allows the algorithm to cover the relevant search space but should not lead the algorithm astray. Fixed step-sizes are very useful in understanding the way the algorithm works for a given problem. Visualizing $\{(\theta_n, J(\theta_n))\}$ and comparing outcomes for various initial values and various choices for ϵ, provides valuable insight into the algorithm. If your control variable θ_n is in high dimensions, θ_n maybe replaced by $\|\theta_n\|$. For a fixed ϵ algorithm, such a pre-analysis is done to see whether ϵ is small enough to yield convergence and large enough to avoid unnecessary long trajectories. In case of decreasing ϵ_n the key question is whether the decrease is too slow yielding rather erratic behaviour of the algorithm, or too small having as consequence that $\epsilon_n \nabla J(\theta_n)$ approaches zero only due to ϵ_n becoming very small. This is the problem of the *vanishing update*.

Boundedness along Trajectories

Suppose that we can argue that the solution set of the original problem is unaltered when θ is confined to some hyper-cube $[-M, M]^d$, for $M > 0$. Then, we can apply the projection method. As the constant M can be chosen arbitrarily large, we may in practice simply neglect the projection to $[-M, M]^d$. An indication that this line of though is not applicable to the problem under consideration is when a trajectory is found that seem to wanders off to infinity. In this situation, one has to go back to the modeling table and rethink the problem formulation. If $J(\theta)$ tends to infinity as $\|\theta\|$ tends to infinity, then provided that $J(\theta)$ is continuous, it can be argued by the Weierstrass theorem that M exists satisfying the above condition. A formal approach to this heuristic is provided in [1], where it is shown that under appropriate set of conditions a growing sequence of hyper-cubes can be constructed that contain a hyper-cube where all the trajectories eventually lie.

Fewer Constraints are better

Sometimes it can be argued that a constraint is not active for the given problem and therefore can be discarded. Alternatively, if the solution to the problem can be found with a constraint disregarded, and the solution remains feasible under the constraint, then the constraint has no influence on the solution (it is not active at the solution). Consider, for example, a simple economic model where θ denotes a production volume, the profit is increasing in θ, and cost of resources are also increasing in θ, so that one would like to choose θ as large as possible, while $g(\theta) \leq b$, for some budget b. From the model it is clear that one will use the maximal budget, i.e., $\{g(\theta) = 0\}$. In words, the constraint can be argued to be an equality constraint. Furthermore , it can be argued that the constraint $\theta \geq 0$ is not active at the solution. However, this line of argument is only possible in case of a physical interpretation of the model.

1.4 Exercises

EXERCISE 1.1. Show that a convex function has *convex* level sets, that is, if $x, y \in \mathcal{L}_\alpha(J)$ then any convex combination of x and y is also in the set $\mathcal{L}_\alpha(J)$.

EXERCISE 1.2. Let $\bar{\theta}$ be a stationary point for which $\nabla^2 J(\bar{\theta})$ has at least one negative and one positive eigenvalue. Using the relationship $\nabla^2 J(\bar{\theta}) v_i = \lambda_i v_i, i = 1, \ldots, d$, for eigenvalues λ_i and eigenvectors $v_i \in \mathbb{R}^d$, and Taylor's expansion around $\bar{\theta}$ in the direction of appropriate eigenvectors, show that $\bar{\theta}$ is neither a local maximum nor a local minimum.

EXERCISE 1.3. Use a Taylor series expansion to show that for any descent direction $d(\theta)$ at a point θ of a twice continuously differentiable function, there exists $\epsilon_0 > 0$ such that

$$J(\theta + \epsilon d(\theta)) \leq J(\theta), \text{ for all } 0 \leq \epsilon \leq \epsilon_0.$$

EXERCISE 1.4. Provide an example of a non-convex set Θ such that projection on Θ fails to be continuous.

EXERCISE 1.5. Give an example of a non-convex set such that the projection on the set is discontinuous. Hint: Consider $\mathbb{R}^2$ with an elipsoid-shaped area removed.

EXERCISE 1.6. Consider $f(x) : \mathbb{R} \to \mathbb{R}$. Show that if f is Lipschtiz with Lipschitz constant L and differentiable, then $|f'(x)|$ is bounded by L on $\mathbb{R}$.

EXERCISE 1.7. For $J \in \mathcal{C}^2$, prove that if $\nabla J(\cdot)$ is Lipschitz continuous on $\mathbb{R}^d$, that is, there is a constant $0 \leq L < \infty$ such that for every $x, y \in \mathbb{R}^d$ $\|\nabla J(x) - \nabla J(y)\| \leq L\|x - y\|$, then $-L\|x\|^2 \leq x^\top \nabla^2 J(\theta) x \leq L\|x\|^2$ for any $x \in \mathbb{R}^d$, where $\|A\| = \max_{\|x\|=1} \|Ax\|$. Use this result to show that $|\sum_n h_n| < \infty$ in the proof of Theorem 1.3.

EXERCISE 1.8. Consider (1.19) and suppose that θ^* is the unique global minimum, so that $J(\theta) \geq J(\theta^*) > -\infty$ for all $\theta \in \mathbb{R}^d$. Argue that in this case, $\lim_{n\to\infty} \nabla J(\theta_n) = 0$. Use this to show that $\{\theta_n\}$ is a Cauchy sequence and conclude that for each initial point $\theta_0 \in \mathbb{R}^d$, θ_n converges.

EXERCISE 1.9. Let $F(u)$ be defined as in Theorem 1.7. Prove that $\nabla_u F(0) = -\eta^*$ in the special case of affine constraints $h(\theta) = a^\top \theta - b$.

EXERCISE 1.10. Consider a single machine that can operate one piece at a time, and let the service times of the machine constitute a sequence of iid exponentially distributed random variables. Parts arrive to the machine according to a Poisson process with unit rate. In other words, the time between arrivals of parts is exponentially distributed with mean 1 and that interarrival times are mutually independent. In order for the system to be stable, assume that the expected service time is strictly less than one. Let $C(\theta) = 1/\theta^2$ be the cost of operating the system at service mean θ. Let $P(\theta)$ denote the stationary probability that the queue length is larger than or equal to a threshold b.

(a) Find the solution to the constrained problem:

$$\min C(\theta), \quad \text{s.t. } P(\theta) \leq \alpha.$$

Interpret the constraint qualifications, the second order condition and the Lagrange multiplier. Hint: Use the fact that the probability that the stationary queue length equals n is given by $(1-\theta)\theta^n$, for $n \in \mathbb{N}$.

(b) Program two different numerical methods to solve this problem for $\alpha = .01$ and $b = 10$. Plot the consecutive values of θ_n and discuss your results, comparing withe the theoretical answer in part (a).