Beyond optimisation in 1D, we will study two directions. First, the equivalent in n-th dimension, \( x^* \in \mathbb{R}^n \) such that \( f(x^*) \leq f(x) \) for all \( x \in \mathbb{R}^n \). Second, constrained optimisation, i.e. \( x^* \in \mathbb{R}^n \) such that \( f(x^*) \leq f(x) \) for all \( x \in \mathbb{R}^n \) where \( g_i(x) \leq 0, i = 1 \ldots m \). For arbitrary \( f, g_i : \mathbb{R}^n \to \mathbb{R} \), this is undecidable. We hence focus on (in some sense) smooth \( f, g_i \), where it is still NP-Hard to decide, whether a point is a local optimum. Only for smooth and convex \( f, g_i \) and under additional assumptions, can one reason about global optima. The methods presented are used throughout all of modern machine learning and much of operations research. For an excellent treatment, see Nesterov [2004].

Key concepts to illustrate include: Constrained minimisation: \( x^* \in \mathbb{R}^n \) such that \( f(x^*) \leq f(x) \) for all \( x \in \mathbb{R}^n \) where \( g_i(x) \leq 0, i = 1 \ldots m \). Jacobian \( \nabla f \): the \( m \times n \) matrix of all first-order partial derivatives of a vector-valued function \( g : \mathbb{R}^n \to \mathbb{R}^m \). Hessian \( H \): a square matrix of second-order partial derivatives of a scalar-valued function \( f, H(f)(x) = J(\nabla f)(x) \). Gradient methods: consider \( f(x + \Delta x) \approx f(x) + \nabla f(x)\Delta x \) and go in the “antigradient direction”. Newton-type methods: consider the quadratic approximation \( f(x + \Delta x) \approx f(x) + \nabla f(x)\Delta x + \frac{1}{2} \Delta x^T H(x) \Delta x \) and multiply the “antigradient direction” with the inverse Hessian \( A \) witness: Checking whether a point \( x^* \in \mathbb{R} \) satisfies \( f'(x^*) = 0 \) is beyond 1D much easier than checking \( x^* \) is a local (!) minimum.

1. Multi-Variate Calculus

1.1. Revision. We will be dealing with scalar- and vector-valued functions in dimension \( n \), and so we need some basic facts from multivariate calculus.

1.1.1. Smooth Function.

- Function \( f : \mathbb{R}^n \to \mathbb{R} \) is Lipschitz-continuous with constant \( L \), \( L \) finite, if and only if: \( ||f(x) - f(y)|| \leq L ||x - y|| \) for any \( x, y \in \mathbb{R}^n \).
- Any Lipschitz-continuous function can be approximated by an infinitely differentiable function within arbitrarily small accuracy.
- We denote by \( C^{0,p}_L(Q) \) the class of functions defined on \( Q \subseteq \mathbb{R}^n \), which are \( k \) times continuously differentiable on \( Q \) and whose \( p \)th derivative is Lipschitz-continuous on \( Q \) with constant \( L \).
- Function \( f \) belongs to \( C^{2,1}_L(\mathbb{R}^n) \) if and only if \( ||f''(x)|| \leq L \) for all \( x \in \mathbb{R}^n \).

1.1.2. First Derivative (Gradient and Jacobian).

**Definition 6.1.** If a scalar-valued function \( f : \mathbb{R}^n \to \mathbb{R} \) has first-order partial derivatives with respect to each \( x_i \), then the \( n \)-dimensional equivalent of the first derivative \( f'(x) \) is the gradient vector

\[
\nabla f(x) = \nabla f = \left( \frac{\partial f(x)}{\partial x_1} \right) = \left( \begin{array}{c} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{array} \right)
\]

When the partial derivatives are not well-defined, we may consider:
Definition 6.2. Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \). The directional derivative of \( f \) at \( x \in \mathbb{R}^n \) in the direction \( v \) is
\[
d_v f(x) = \frac{\partial f}{\partial v} := v \cdot \nabla f(x) \quad \text{(dot product)}
\]
where \( v = (v_1, \ldots, v_n)^t \in \mathbb{R}^n \).

Definition 6.3. If the function \( g : \mathbb{R}^n \rightarrow \mathbb{R}^m \) has first-order partial derivatives with respect to each \( x_i \), then the \( m \times n \) matrix:
\[
J = \frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}
\]
is the Jacobian.

1.1.3. Second Derivative or Hessian.

Definition 6.4. The \( n \)-dimensional equivalent of the second derivative \( f''(x) \) is the Hessian matrix:
\[
H_f(x) = \left( \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \right) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \cdots & \ddots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{bmatrix}
\]
Note that \( H_f(x^*) \) is a symmetric matrix since \( \frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i} \) for all \( i, j \). We omit the subscript where not needed.

1.1.4. Taylor Series in \( n \) Dimensions.

Definition 6.5. The Taylor series expansion of \( f(x) \) about some \( x_k \in \mathbb{R}^n \) is:
\[
f(x) \approx f(x_k) + (\nabla f(x_k))^t (x - x_k) + \frac{1}{2} (x - x_k)^t H_f(x_k) (x - x_k) + \cdots,
\]
where \( f(x) \in \mathbb{R} \) and \( x_k \in \mathbb{R}^n \), and \( H_f(x_k) \in M_n \mathbb{R} \).

Also we define

Definition 6.6.
\[
\frac{\partial^2 f}{\partial v^2} := \sum_{i=1}^n v_i \frac{\partial f}{\partial v} \left( \frac{\partial f}{\partial x_i} \right) = \sum_{i=1}^n v_i \left( \sum_{j=1}^n v_j \frac{\partial^2 f}{\partial x_i \partial x_j} \right) = \sum_{i,j=1}^n v_i v_j \frac{\partial^2 f}{\partial x_i \partial x_j} = v^t H_f(x) v.
\]

The main theoretical result on the existence of local extrema and saddle points is:

Theorem 6.7. Let \( U \) be an open subset of \( \mathbb{R}^n \), \( f : U \rightarrow \mathbb{R} \) be a twice continuously differentiable function on \( U \), and let \( x^* \) be a critical point of \( f \), i.e., \( \nabla f(x^*) = 0 \). Then

(a) \( x^* \) is a local maximum of \( f \) if \( \frac{\partial^2 f}{\partial v^2} < 0 \) for all nonzero \( v \in \mathbb{R}^n \);
(b) $x^*$ is a local minimum of $f$ if $\frac{\partial^2 f}{\partial x^2} > 0$ for all nonzero $v \in \mathbb{R}^n$; 
(c) $x^*$ is a saddle point of $f$ if there exist $v, w \in \mathbb{R}^n$ such that 
$$\frac{\partial^2 f}{\partial v^2} < 0 < \frac{\partial^2 f}{\partial w^2}.$$ 

It is clear that this involves examining the sign of $v^T H_f(x)v$ for various $v$. It can be shown that this theorem leads to a practical test as follows.

### 2. Derivative-free Methods

For functions for $x^* \in \mathbb{R}^n$, the convergence of derivative-free methods is provably slow.

**Theorem 6.8 (1.2 in Nesterov [2004]).** For $L$-Lipschitz function $f$, $\epsilon \leq \frac{1}{2} L$, and $f$ provided as an oracle that allows $f(x)$ to be evaluated for $x$, derivative-free methods require 
$$\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor \right)^n$$ 
calls to the oracle to reach $\epsilon$ accuracy.

That is: If the number of calls to the oracle is less than $\left(\left\lfloor \frac{L}{2\epsilon} \right\rfloor \right)^n$, the accuracy of the result cannot be better than $\epsilon$.

To put the lower bound into perspective, consider a single computer, which can sustain the performance of $10^{11}$ operations per second (“100 gigaFLOPS”) and a function, which can be evaluated in $n$ operations:

For $L = 2, n = 10, 10\%$ accuracy, you need $10^{11}$ operations, or 1 second. For $L = 2, n = 10, 1\%$ accuracy, you need $10^{21}$ operations, or 325 years. For $L = 2, n = 10, 0.1\%$ accuracy, you need $10^{31}$ operations, or 10$^{12}$ years. For $L = 2, n = 100, 1\%$ accuracy, you need $10^{201}$ operations, or 10$^{982}$ years.

### 3. Gradient Methods

Let us consider a local, unconstrained minimum $x^*$ of a multi-variate function, i.e., $f(x^*) \leq f(x)\forall x$ with $\|x - x^*\| \leq \epsilon$. From the definition, at a local minimum $x^*$, we expect the variation in $f$ due to a small variation $\Delta x$ in $x$ to be non-negative: $\nabla f(x^*) \Delta x = \sum_{i=1}^n \frac{\partial f(x^*)}{\partial x_i} \Delta x_i \geq 0$. By considering $\Delta x$ coordinate wise, we get: $\nabla f(x^*) = 0$.

In gradient methods, you consider $x^{k+1} = x^k - h^k \nabla f(x^k)$, where $h^k$ is one of:

- Constant step $h^k = h$ or $h^k = h/\sqrt{k+1}$
- Full relaxation $h^k = \arg \min_{t \geq 0} f(x^k - \nabla f(x^k))$
- Armijo line search: find $x^k+1$ such that the ratio $\frac{\nabla f(x^k)(x^k-x^k+1)}{f(x^k)-f(x^k+1)}$ is within some interval

For all of the choices above with $f \in C^1_1(R^n)$, one has $f(x^k) - f(x^{k+1}) \geq \frac{\epsilon}{L} \|\nabla f(x_k)\|^2$. We hence want to bound the norm of the gradient. It turns out:

$$\min_{i=0}^k \|\nabla f(x_i)\| \leq \frac{1}{\sqrt{k+1}} \left[ \frac{1}{\omega} L (f(x_0) - f^*) \right]^{1/2}$$

This means that the norm of the gradient is less than $\epsilon$, if the number of iterations is greater than $\frac{L}{\omega \epsilon} (f(x_0) - f^*) - 1$.

**Theorem 6.9 (1.2.4 in Nesterov [2004]).** For $f \in C^2_{M}(R^n)$, $\|H \| \geq H(f^*) \geq L I_n$, a certain gradient method starting from $x^0, r^0 = \|x^0 - x^*\| \leq \frac{2L}{\epsilon} := \bar{r}$ converges as follows:

$$\|x^k - x^*\| \leq \frac{\bar{r} v^0}{\bar{r} - v^0} \left( 1 - \frac{2L}{L + 3L} \right)^k$$

This is called the (local) linear (rate of) convergence.
4. Newton-Type Methods

In finding a solution to a system of non-linear equations \( F(x) = 0, \ x \in \mathbb{R}^n, F : \mathbb{R}^n \to \mathbb{R}^n \), we can define displacement \( \Delta x \) as a solution to \( F(x) + \nabla F(x) \Delta x = 0 \), which is known as the Newton system. Assuming \( \nabla F \) exists, we can use: \( x^{k+1} = x^k - \left[ \nabla F(x^k) \right]^{-1} F(x^k) \). When we move from finding zeros of \( F(x) \) to minimising \( f(x), x \in \mathbb{R}^n, f : \mathbb{R}^n \to \mathbb{R} \) by finding zeros of \( \nabla f(x) = 0 \), we obtain: \( x^{k+1} = x^k - \left[ \nabla^2 f(x^k) \right]^{-1} \nabla f(x^k) \).

Alternatively, let us consider a quadratic approximation of \( f \) at point \( x^k \), i.e., \( f(x^k) + \nabla f(x^k) \cdot (x - x^k) + \frac{1}{2} \left[ H(x^k) \right] (x - x^k) \cdot (x - x^k) \). Assuming that \( H(x^k) \succeq 0 \), one should like to choose \( x^{k+1} \) by minimising the approximation, i.e., solving \( \nabla f(x^k) + H(x^k)(x^{k+1} - x^k) = 0 \) for \( x^{k+1} \). Thereby, one obtains: \( x^{k+1} = x^k - \left[ \nabla^2 f(x^k) \right]^{-1} \nabla f(x^k) \).

**Example 6.10.** Study the behaviour of Newton method on \( x/\sqrt{1+t^2} \), where the root is clearly 0. ◊

**Theorem 6.11** (1.2.5 in Nesterov Nesterov [2004]). For \( f \in C^2_\infty(\mathbb{R}^n) \), where there exists a local minimum with positive definite Hessian \( H(f^*) \succeq I_n \), \( x_0 \) close enough to \( x^* \), i.e., \( ||x_0 - x^*|| < 2M \), Newton method starting from \( x_0 \) converges as follows:

\[
||x^{k+1} - x^*|| \leq \frac{M||x^k - x^*||^2}{2I^2 - 2M||x^k - x^*||}
\]

This is called the (local) quadratic (rate of) convergence.

Newton method is only locally convergent, but the region of convergence is similar for gradient and Newton methods. One can try to address the possible divergence by considering dampening: \( x^{k+1} = x^k - h^k \left[ \nabla^2 f(x^k) \right]^{-1} \nabla f(x^k) \), where \( h^k \geq 0 \) is a step-size, which usually goes to 1 as \( k \) goes to infinity, or other “regularisations”. One can try to make a single iteration cheaper by either exploiting sparsity of the Hessian or by considering some approximation of its inverse.

5. Quasi-Newton Methods

Quasi-Newton methods build up a sequence of approximations \( H^k \) of the inverse of the Hessian and use it in computing the step. Starting with \( H^0 = I_n \) and some \( x^0 \), in each iteration: \( x^{k+1} = x^k + h^k H^k \nabla f(x^k) \) for some step-length \( h^k \) and \( H^{k+1} = H^k + \Delta H^k \) where:

- In rank-one methods, \( \Delta H^k = \frac{(\delta^k - H^k \gamma^k)(\delta^k - H^k \gamma^k)^T}{(\delta^k - H^k \gamma^k)^T H^k} \)

- In Broyden-Fletcher-Goldfarb-Shanno (BFGS): \( \Delta H^k = \frac{H^k \gamma^k (\delta^k)^T + \delta^k (\gamma^k)^T H^k}{(\gamma^k)^T H^k - \beta k H^k \gamma^k (\gamma^k)^T H^k} \)

where \( \delta^k = x^{k+1} - x^k, \gamma^k = \nabla f(x^{k+1}) - \nabla f(x^k) \), and \( \beta k = 1 + \langle \gamma^k, \delta^k \rangle / (H^k \gamma^k, \gamma^k) \). These methods are very successful in practice, although their rates of convergence are very hard to bound.

6. Constrained Optimisation

Let us introduce the extension to constrained optimisation by example, which is based on the treatment of Gondzio.

Consider a constrained minimisation problem \( \min f(x) \) subject to \( g(x) \leq 0, x \in \mathbb{R}^n \) where \( f(x) : \mathbb{R}^n \to \mathbb{R}, g(x) : \mathbb{R}^n \to \mathbb{R}^m \) and the global optimum at \( x^* \) is \( f^* \). Let us consider \( g \) as \( m \) inequalities \( g_i(x) \leq 0, i = 1 \ldots m \) and let us introduce Lagrange multipliers (also known as dual variables) \( y = (y_1, y_2, \ldots, y_m)^T, y_i \geq 0 \), one scalar for each inequality \( g_i \). The Lagrangian of the constrained minimisation problem is: \( L(x, y) = f(x) + y^T g(x) \). One can extend this to an additional constraint \( x \in X \subseteq \mathbb{R}^n \).

The “Lagrangian primal” is \( L_P(x) = \max_{y \geq 0} L(x, y) \), with \( L_P(x) = \infty \) if any inequality is violated. The “Lagrangian dual” is \( L_D(y) = \min_{x \in X} L(x, y) \). Its value clearly depends on the choice of \( y \).

For any \( y \geq 0 \), however, \( f^* \geq L_D(y) \), i.e., \( f^* \geq \max_{y \geq 0} \min_{x \in X} L(x, y) \), and by considering first, a violated inequality \( g_i \) at \( x \) leads to \( \max_{y \geq 0} L(x, y) = \infty \), while no violated inequality leads to \( y_i g_i(x) = 0 \), \( i = 1, 2, \ldots, m \), \( \max_{y \geq 0} L(x, y) = f(x), f^* = \min_{x \in X} \max_{y \geq 0} L(x, y) \). Thereby, \( \min_{x \in X} \max_{y \geq 0} L(x, y) \geq \max_{y \geq 0} \min_{x \in X} L(x, y) \) (“weak duality”). One usually interprets this as follows: Any primal feasible
solution provides an upper bound for the dual problem, and any dual feasible solution provides a lower bound for the primal problem.

Assuming differentiability of $L$, Karush-Kuhn-Tucker (KKT) conditions are composed of stationarity ($\nabla_x L(x, y) = 0$), primal feasibility ($g(x) \leq 0, x \in X$), dual feasibility ($y \geq 0$), and "complementarity slackness" ($y_i g_i(x) = 0$). Under some "regularity" assumptions (also known as constraint qualifications), we are guaranteed that a point $x$ satisfying the KKT condition exists. If $X \subseteq \mathbb{R}^n$ is convex, $f$ and $g$ are convex, optimum $f^*$ is finite, and the regularity assumptions hold, then we have $\min_{x \in X} \max_{y \geq 0} L(x, y) = \max_{y \geq 0} \min_{x \in X} L(x, y)$ ("strong duality") and KKT conditions guarantee global optimality. For example, Slater’s constraint qualifications is: $\exists x \in \text{int}(X)$ such that $g(x) < 0$. If $f$ and $g$ are linear, no further constraint qualifications is needed and KKT conditions suffice.

Notice that the Lagrangian, as defined above, is not Lipschitz-continuous and is not differentiable. Let $x$ and $y$ be feasible for the primal problem. A barrier is a continuous function, such that $\sum_{i=1}^m \max\{g_i(x), 0\} \leq 0$ (non-smooth), $\sum_{i=1}^m (\max\{g_i(x), 0\})^2$ (smooth). A barrier $\phi$ for $G$ is a continuous function, such that $\phi(x) \to \infty$ as $x$ approaches the boundary of $G$ and is bounded from below elsewhere. E.g. $\sum_{i=1}^m \frac{1}{|g_i(x)|^p} \geq 1$ (power), $-\sum_{i=1}^m \ln(-g_i(x))$ (logarithmic). One can consider (variants of) the Lagrangian of a constrained problem, which involve a barrier for the inequalities. Using such Lagrangians, one can develop interior-point methods.

Remark 6.12. Some authors use the word "Lagrangian" very loosely. See Liu et al. [2015] for an extensive discussion of many different Lagrangians of a single problem in power systems.

### 7. Interior-Point Methods

Interior-point methods solve progressively less relaxed first-order optimality conditions of a problem, which is equivalent to a constrained optimisation problem and uses barriers. Let us introduce interior-point methods by example, which is based on the treatment of Benzi et al. [2005], which is in turn based on the treatment of Bergamaschi et al. [2004]. Consider a constrained minimisation $\min f(x)$ subject to $g(x) \leq 0$ where $f(x) : \mathbb{R}^n \to \mathbb{R}, g(x) : \mathbb{R}^n \to \mathbb{R}^m$ are convex and twice differentiable. A nonnegative slack variable $z \in \mathbb{R}^m$ can be used to replace the inequality by equality $g(x) + z = 0$. Negative $z$ can be avoided by using a barrier $\mu \sum_{i=1}^m \ln z_i$. The (variant of) Lagrangian is: $L(x, y, z; \mu) = f(x) + y^T (g(x) + z) + \mu \sum_{i=1}^m \ln z_i$.

Now we can differentiate:

$$\nabla_x L(x, y, z; \mu) = \nabla f(x) + \nabla g(x)^T y$$  \hspace{1cm} (6.1)

$$\nabla_y L(x, y, z; \mu) = g(x) + z$$  \hspace{1cm} (6.2)

$$\nabla_z L(x, y, z; \mu) = y u Z^{-1} e,$$  \hspace{1cm} (6.3)

where $Z = \text{diag}(z_1, z_2, \ldots, z_m)$ and $e = [1, 1, \ldots, 1]^T$.

The first-order optimality conditions obtained by setting the partial derivatives to zero are:

$$\nabla f(x) + \nabla g(x)^T y = 0$$  \hspace{1cm} (6.4)

$$g(x) + z = 0$$  \hspace{1cm} (6.5)

$$Y Z e = \mu e$$  \hspace{1cm} (6.6)

$$y, z \geq 0$$  \hspace{1cm} (6.7)

where $Y = \text{diag}(y_1, y_2, \ldots, y_m)$ and the parameter $\mu$ is reduced to 0 in the large number of iterations.

This can be solved using the Newton method, where at each step, one solves a linear system:

$$\begin{bmatrix}
-H(x, y) & B(x)^T \\
B(x) & X Y^{-1}
\end{bmatrix}
\begin{bmatrix}
\Delta x \\
-\Delta y
\end{bmatrix}
= \begin{bmatrix}
\nabla f(x) + B(x)^T y \\
-g(x) - \mu Y^{-1} e
\end{bmatrix}$$

where $H(x, y) = \nabla^2 f(x) + \sum_{i=1}^m y_i \nabla^2 g_i(x) \in \mathbb{R}^{n \times n}$ and $B(x) = \nabla g(x) \in \mathbb{R}^{m \times n}$. This is a saddle point system, which has often positive semidefinite $A$. For convex $f, g$, $H(x, y)$ is positive semidefinite and diagonal matrix $Z Y^{-1}$ is also positive definite. A variety of methods works very well. For an excellent overview, see Benzi et al. [2005].
8. Condition Numbers

Assume the instance \( d := (A; b; c) \) is given. One can formalise the following notion, due to Renegar:

\[
C(d) := \frac{||d||}{\inf\{||\Delta d|| : \text{instance } d + \Delta d \text{ is infeasible or unbounded} \}}.
\]

The system (6.4–6.7) will have the condition number \( C(d)/\mu \).

There are a variety of other methods for constrained optimisation. For a variety of methods, including:

- Interior-point methods (Renegar)
- Ellipsoid method (Freund and Vera [1999])
- Perceptron method (Belloni et al. [2009])
- Von Neumann method (Epelman and Freund [2000])

assuming \( A \) is invertible, one can show a bound on the number of iterations is logarithmic in \( C(d) \), with further dependence on the dimension \( n \). This highlights the need for preconditioners, c.f., Bergamaschi et al. [2004], as mentioned in Chapter 4.

Constrained optimisation is the work-horse of operations research. Interior-point methods have been used on problems in dimensions \( 10^9 \). Still, there are many open problems, including Smale’s 9th problem: Is feasibility of a linear system of inequalities \( Ax \geq b \) in \( P \) over reals, i.e., solvable in polynomial time on the BSS machine?
Bibliography


