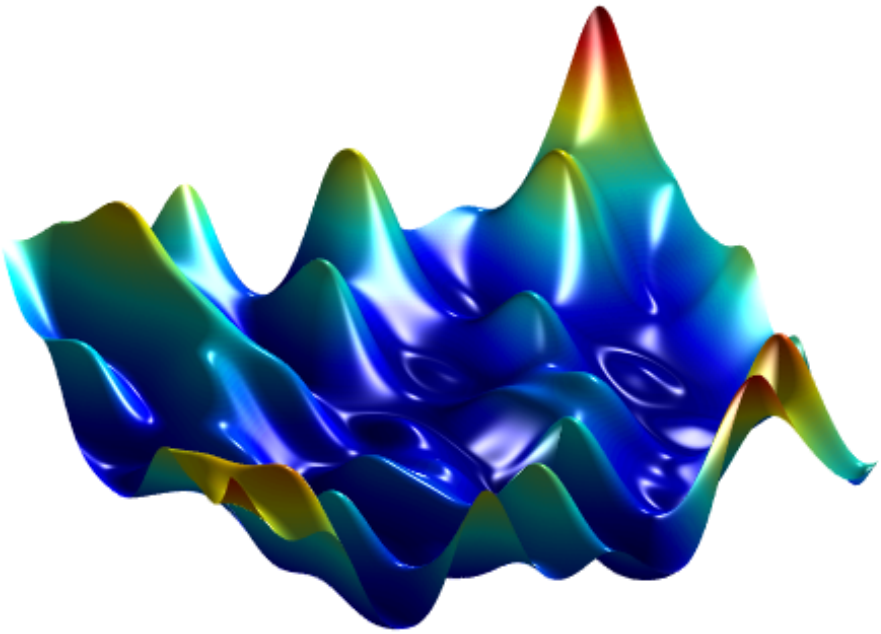
MATH60005/70005: Optimisation (Autumn 23-24)

Last update: January 2, 2024

For overview only, the official material for every week is available in Blackboard, these notes will be updated as we progress.

Dr Dante Kalise & Dr Estefanía Loayza-Romero
Department of Mathematics
Imperial College London, United Kingdom
{dkaliseb,kloayzar}@imperial.ac.uk



Notation for this module 5

I. Mathematical Preliminaries 6

Vector Spaces 6
Inner Products and Norms 7
Eigenvalues and Eigenvectors 10
Basic Topological Concepts 10
Directional Derivatives and Gradients 12
Gradient and Hessian of Quadratic Functions 14
Notation for this module 17

II. Unconstrained Optimization 18

Global Minimum and Maximum 18
Local Minima and Maxima 19
Second Order Optimality Conditions 22
Global Optimality Conditions 26
Appendix: Classification of Matrices 28

III. Linear and Nonlinear Least Squares Problems 35

A Bit of History 35
Formulating the Linear Least Squares Problem 36
Data Fitting 37
Regularized Least Squares 39
Denoising 39
Nonlinear Least Squares 40
A Case Study: Circle Fitting 41

IV. The Gradient Descent Algorithm 44

Descent Directions Methods 44
Taking the Direction of Minus the Gradient 46
Convergence of the Gradient Method 49
The Condition Number 53
Scaled Gradient Method 55
The Gauss-Newton Method 60

An Example of a Gradient Method: The Fermat-Weber Problem and Weiszfeld's Method . . . 63
Newton's Method (Un-assessed) . . . 64

V. Stochastic Gradient Descent **69**
The Kaczmarz Algorithm . . . 69
Stochastic Gradient Descent . . . 71
Revisiting the Nonlinear Regression Example . . . 73

VI. Convex Sets and Functions **75**
Convex Sets . . . 75
The Convex Hull . . . 78
Convex Functions . . . 81
First-order Characterization of Convex Functions . . . 82
Second-order Characterization of Convex Functions . . . 84
Further Results for Convex Functions . . . 85

VII. Convex Optimization **89**
Convex Optimization Problems . . . 89
Optimization over a Convex Set and Stationarity . . . 90
The Orthogonal Projection Operator . . . 92
The Gradient Projection Method . . . 94

VIII. Optimality Conditions **95**
Separation Theorem . . . 95
KKT Conditions for Linearly Constrained Problems . . . 97
Orthogonal projections . . . 99
KKT conditions for nonlinear problems . . . 100
KKT conditions for nonlinear convex problems . . . 101

IX. Duality **103**
The Primal and Dual Problems . . . 103
Weak and Strong Duality . . . 104
Three Important Examples of Duality Use . . . 108

X. Optimal Control 111
What is Mathematical Control Theory? 111
What is Optimal Control? 114
Using Calculus of Variations 118
Pontryagin's Maximum Principle 120
Optimal Feedback Control & Dynamic Programming 125

Notation for this module

Symbol	Object
$\mathbb{R}^n$	the space of $n-$dimensional real column vectors
$\mathbf{x}$	a vector in $\mathbb{R}^n$
x_i	the $i-$th coordinate of a vector $\mathbf{x}$
$\langle \mathbf{x}, \mathbf{y} \rangle$	inner product of $\mathbf{x}$ and $\mathbf{y}$
$[\mathbf{x}, \mathbf{y}]$	closed line segment between $\mathbf{x}$ and $\mathbf{y}$
$(\mathbf{x}, \mathbf{y})$	open line segment between $\mathbf{x}$ and $\mathbf{y}$
$B(\mathbf{c}, r)$	open ball with center $\mathbf{c}$ and radius r
$B[\mathbf{c}, r]$	closed ball with center $\mathbf{c}$ and radius r
$\mathbb{R}^{m \times n}$	space of $m \times n$ real-valued matrices
$\mathbf{A}^T$	transpose of $\mathbf{A}$
$\mathbf{e}_i$	$i-$th vector in the standard basis of $\mathbb{R}^n$
$\mathbf{e}$	vector of all ones
$\mathbf{0}$	vector of all zeros
$\lvert \cdot \rvert$	absolute value of scalar x
$\lVert \cdot \rVert_p$	ℓ_p-norm for $\mathbf{x} \in \mathbb{R}^n$
Δ_n	unit simplex
$\mathbb{R}^n_+$	nonnegative orthant
$\mathbb{R}^n_{++}$	positive orthant
$\lVert \mathbf{A} \rVert_F$	Frobenius norm of $\mathbf{A}$
$\lVert \mathbf{A} \rVert_{ab}$	induced norm of $\mathbf{A} \in \mathbb{R}^{m \times n}$
$\lVert \mathbf{A} \rVert_2$	spectral norm of $\mathbf{A}$
$\lambda_{max}(\mathbf{A})$	maximum eigenvalue of a symmetric matrix $\mathbf{A}$
$\lambda_{min}(\mathbf{A})$	minimum eigenvalue of a symmetric matrix $\mathbf{A}$
$int(S)$	interior of set S
$f'(\mathbf{x}; \mathbf{d})$	directional derivative of f at x in the direction $\mathbf{d}$
$\nabla f(\mathbf{x})$	gradient of f at $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$	Hessian of $f(\mathbf{x})$ at $\mathbf{x}$
$C_L^{1,1}(D)$	class of $L-$smooth functions over D
$\mathbf{I}_n$	identity matrix in $R^{n \times n}$
$\mathbf{0}_{mn}$	zero matrix in $R^{m \times n}$
$diag(\mathbf{x})$	diagonal matrix with entries $\mathbf{x}$

Part I.
Mathematical Preliminaries

Vector Spaces

The space $\mathbb{R}^n$ is the set of n -dimensional column vectors $\mathbf{x}$ with real components endowed with the component-wise addition operator:

$$\mathbf{x}+\mathbf{y}=\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}+\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}x_1+y_1\\x_2+y_2\\\vdots\\x_n+y_n\end{pmatrix},$$

and the scalar-vector product

$$\lambda\mathbf{x}=\lambda\begin{pmatrix}x_1\\x_2\\\vdots\\x_n\end{pmatrix}=\begin{pmatrix}\lambda x_1\\\lambda x_2\\\vdots\\\lambda x_n\end{pmatrix}.$$

The vectors $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ denote the standard/canonical basis, and $\mathbf{e}$ and $\mathbf{0}$ denote all ones and all zeros column vectors, respectively.

Important subsets of $\mathbb{R}^n$.

- Nonnegative orthant:

$$\mathbb{R}^n_+=\left\{(x_1,x_2,\ldots,x_n)^\top : x_1,x_2,\ldots,x_n\geq 0\right\}$$

- Positive orthant:

$$\mathbb{R}^n_{++}=\left\{(x_1,x_2,\ldots,x_n)^\top : x_1,x_2,\ldots,x_n> 0\right\}.$$

- If $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$, the closed line segment between x and y is given by

$$[\mathbf{x},\mathbf{y}]=\{\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x}) : \alpha\in[0,1]\}.$$

- The open line segment $(\mathbf{x},\mathbf{y})$ is similarly defined as

$$(\mathbf{x},\mathbf{y})=\{\mathbf{x}+\alpha(\mathbf{y}-\mathbf{x}) : \alpha\in(0,1)\}$$

  for $\mathbf{x}\neq\mathbf{y}$ and $(\mathbf{x},\mathbf{x})=\emptyset$.

- The unit-simplex:

$$\Delta_n=\left\{\mathbf{x}\in\mathbb{R}^n : \mathbf{x}\geq\mathbf{0}, \mathbf{e}^\top\mathbf{x}=1\right\}.$$

The space $\mathbb{R}^{m\times n}$. The set of all real valued $m \times n$ matrices is denoted by $\mathbb{R}^{m\times n}$. The $n \times n$ identity matrix is denoted by $\mathbf{I}_n$. The $m \times n$ zero matrix is denoted by $\mathbf{O}_{m\times n}$.

Inner Products and Norms

Definition (Inner Product). *An inner product on $\mathbb{R}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ with the following properties:*

1. *Symmetry: $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*
2. *Additivity: $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$ for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$.*
3. *Homogeneity: $\langle \lambda \mathbf{x}, \mathbf{y} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle$ for any $\lambda \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*
4. *Positive definiteness: $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\langle \mathbf{x}, \mathbf{x} \rangle = 0$ if and only if $\mathbf{x} = 0$.*

Examples:

- The usual "dot product":

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^{n} x_i y_i \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n .$$

- The "weighted dot product":

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathbf{w}} = \sum_{i=1}^{n} w_i x_i y_i$$

  where $\mathbf{w} \in \mathbb{R}^n_{++}$.

Definition (Vector Norms). *A norm $\| \cdot \|$ on $\mathbb{R}^n$ is a function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$ satisfying:*

1. *Nonnegativity: $\|\mathbf{x}\| \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = 0$.*
2. *Positive homogeneity: $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$ for any $\mathbf{x} \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$.*
3. *Triangle inequality: $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

A natural way to generate a norm on $\mathbb{R}^n$ is to take any inner product $\langle \cdot, \cdot \rangle$ defined on $\mathbb{R}^n$, and define the associated norm

$$\|\mathbf{x}\| \equiv \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}, \text{ for all } \mathbf{x} \in \mathbb{R}^n .$$

The norm associated with the dot-product is the so-called Euclidean norm or ℓ_2-norm:

$$\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \text{ for all } \mathbf{x} \in \mathbb{R}^n .$$

The ℓ_p-norm ($p \geq 1$) is defined by

$$\|\mathbf{x}\|_p \equiv \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} .$$

The ℓ_∞-norm is

$$\|\mathbf{x}\|_\infty \equiv \max_{i=1,2,\ldots,n} |x_i| .$$

It can be shown that

$$\|\mathbf{x}\|_\infty = \lim_{p \to \infty} \|\mathbf{x}\|_p .$$

Why $\ell_{1/2}$ is not a norm?

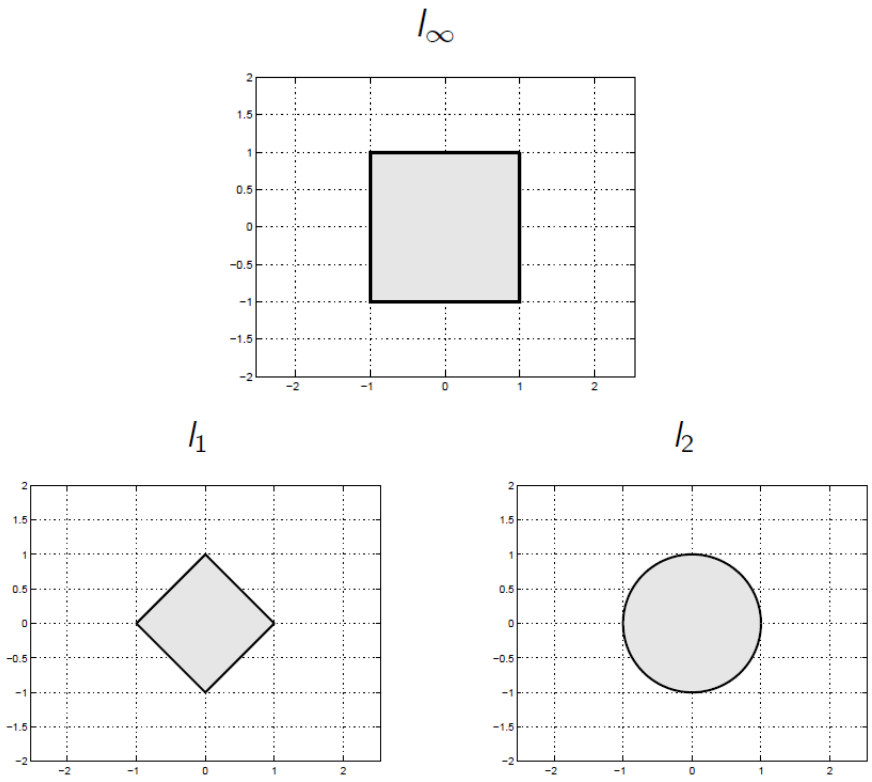


Figure 1: Different unit balls $\|\mathbf{x}\|_p \leq 1$ in $\mathbb{R}^2$.

Theorem (Cauchy-Schwartz Inequality). *For any* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\left| \mathbf{x}^\top \mathbf{y} \right| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$$

Proof. For any $\lambda \in \mathbb{R}$:

$$\|\mathbf{x} + \lambda \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + 2\lambda \langle \mathbf{x}, \mathbf{y} \rangle + \lambda^2 \|\mathbf{y}\|^2 ,$$

leading to (why?)

$$\langle \mathbf{x}, \mathbf{y} \rangle^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 ,$$

establishing the desired result. □

Definition (Matrix Norms). *A norm* $\|\cdot\|$ *on* $\mathbb{R}^{m\times n}$ *is a function* $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ *satisfying*

1. *Nonnegativity:* $\|\mathbf{A}\| \geq 0$ *for any* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\|\mathbf{A}\| = 0$ *if and only if* $\mathbf{A} = \mathbf{0}$.
2. *Positive homogeneity:* $\|\lambda\mathbf{A}\| = |\lambda|\|\mathbf{A}\|$ *for any* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\lambda \in \mathbb{R}$.
3. *Triangle inequality:* $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ *for any* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$.

Induced Norms. Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and two norms $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, the induced matrix norm $\|\mathbf{A}\|_{a,b}$ (called (a,b)-norm) is defined by

$$\|\mathbf{A}\|_{a,b} = \max_{\mathbf{x}} \{\|\mathbf{A}\mathbf{x}\|_b : \|\mathbf{x}\|_a \leq 1\} ,$$

from where it follows that

$$\|\mathbf{A}\mathbf{x}\|_b \leq \|\mathbf{A}\|_{a,b}\|\mathbf{x}\|_a .$$

An induced norm is a norm (satisfies nonnegativity, positive homogeneity and triangle inequality). We refer to the matrix-norm $\|\cdot\|_{a,b}$ as the (a,b)-norm. When $a = b$, we will simply refer to it as an a-norm.

Relevant cases:

- **The spectral norm.** If $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_2$, the induced $(2,2)$-norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the maximum singular value of $\mathbf{A}$

  $$\|\mathbf{A}\|_2 = \|\mathbf{A}\|_{2,2} = \sqrt{\lambda_{\max}(\mathbf{A}^\top\mathbf{A})} \equiv \sigma_{\max}(\mathbf{A}) .$$

  This norm is called the spectral norm.

- **The ℓ_1-norm.** When $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_1$, the induced $(1,1)$-matrix norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is given by

  $$\|\mathbf{A}\|_1 = \max_{j=1,2,\ldots,n} \sum_{i=1}^{m} |A_{i,j}| .$$

- **The ℓ_∞-norm.** When $\|\cdot\|_a = \|\cdot\|_b = \|\cdot\|_\infty$, the induced (∞,∞)-matrix norm of matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is given by

  $$\|\mathbf{A}\|_\infty = \max_{i=1,2,\ldots,m} \sum_{j=1}^{n} |A_{i,j}|$$

- **The Frobenius norm.**

  $$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}^2}, \quad \mathbf{A} \in \mathbb{R}^{m\times n}$$

  The Frobenius norm is not an induced norm. Why is it a norm?

Eigenvalues and Eigenvectors

Definition (Eigenvalues and eigenvectors)**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$*. Then a nonzero vector* $v \in \mathbb{R}^n$ *is called an eigenvector of* $\mathbf{A}$ *if there exists a* $\lambda \in \mathbb{C}$ *for which*

$$\mathbf{Av} = \lambda \mathbf{v}\,.$$

The scalar λ *is the eigenvalue corresponding to the eigenvector* $\mathbf{v}$*.*

In general, real-valued matrices can have complex eigenvalues, but when the matrix is symmetric the eigenvalues are necessarily real. The eigenvalues of a symmetric $n \times n$ matrix $\mathbf{A}$ are denoted by

$$\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots \geq \lambda_n(\mathbf{A})\,.$$

The maximum eigenvalue is also denoted by $\lambda_{\max}(\mathbf{A})$ $(= \lambda_1(\mathbf{A}))$ and the minimum eigenvalue is also denoted by $\lambda_{\min}(\mathbf{A})$ $(= \lambda_n(\mathbf{A}))$.

Theorem (Spectral Factorization Theorem)**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be an* $n \times n$ *symmetric matrix. Then there exists an orthogonal matrix* $\mathbf{U} \in \mathbb{R}^{n\times n}$ $(\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}_n)$ *and a diagonal matrix* $\mathbf{D} = \text{diag}\,(d_1, d_2, \ldots, d_n)$ *for which*

$$\mathbf{U}^\top\mathbf{A}\mathbf{U} = \mathbf{D}\,.$$

The columns of the matrix $\mathbf{U}$ constitute an orthogonal basis comprising eigenvectors of $\mathbf{A}$ and the diagonal elements of $\mathbf{D}$ are the corresponding eigenvalues. A direct result is that $\text{Tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i(\mathbf{A})$ and $\det(\mathbf{A}) = \Pi_{i=1}^n \lambda_i(\mathbf{A})$.

Another important consequence of the spectral decomposition theorem is the bounding of the so-called Rayleigh quotient. For a symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, the Rayleigh quotient is defined by

$$R_{\mathbf{A}}(\mathbf{x}) = \frac{\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\|\mathbf{x}\|^2} \text{ for any } \mathbf{x} \neq 0$$

We use the spectral decomposition theorem to show that if $\mathrm{A} \in \mathbb{R}^{n\times n}$ is symmetric, then

$$\lambda_{\min}(\mathbf{A}) \leq R_{\mathbf{A}}(\mathbf{x}) \leq \lambda_{\max}(\mathbf{A}) \text{ for any } \mathbf{x} \neq 0\,.$$

Basic Topological Concepts

The open ball with center $c \in \mathbb{R}^n$ and radius r :

$$B(c, r) = \{\mathbf{x} : \|\mathbf{x} - c\| < r\}\,.$$

The closed ball with center c and radius r :

$$B[c, r] = \{\mathbf{x} : \|\mathbf{x} - c\| \leq r\}\,.$$

Definition (Interior Point). *Given a set $U \subseteq \mathbb{R}^n$, a point $\mathbf{c} \in U$ is called an interior point of U if there exists $r > 0$ for which $B(\mathbf{c}, r) \subseteq U$. The set of all interior points of a given set U is called the interior of the set and is denoted by* $\operatorname{int}(U)$*:*

$$\operatorname{int}(U) = \{\mathbf{x} \in U : B(\mathbf{x}, r) \subseteq U \text{ for some } r > 0\}.$$

Examples:

$$\operatorname{int}\left(\mathbb{R}^n_+\right) = \mathbb{R}^n_{++}$$
$$\operatorname{int}(B[\mathbf{c}, r]) = B(\mathbf{c}, r) \quad (\mathbf{c} \in \mathbb{R}^n, r \in \mathbb{R}_{++})$$
$$\operatorname{int}([\mathbf{x}, \mathbf{y}]) = ?$$

An open set is a set that contains only interior points. Meaning that

$$U = \operatorname{int}(U).$$

Examples of open sets are open balls (hence the name...) and the positive orthant $\mathbb{R}^n_{++}$. The union of any number of open sets is an open set and the intersection of a finite number of open sets is open.

Closed Sets. A set $U \subseteq \mathbb{R}^n$ is closed if it contains all the limits of convergent sequences of vectors in U, that is, if $\{\mathbf{x}_i\}_{i=1}^{\infty} \subseteq U$ satisfies $\mathbf{x}_i \to \mathbf{x}^*$ as $i \to \infty$, then $\mathbf{x}^* \in U$. A known result states that U is closed iff its complement U^c is open. Examples of closed sets are the closed ball $B[\mathbf{c}, r]$, closed lines segments, the nonnegative orthant $\mathbb{R}^n_+$ and the unit simplex Δ_n. What about $\mathbb{R}^n$?$\emptyset$?

Boundary Points. Given a set $U \subseteq \mathbb{R}^n$, a boundary point of U is a vector $\mathbf{x} \in \mathbb{R}^n$ satisfying the following: any neighbourhood of $\mathbf{x}$ contains at least one point in U and at least one point in its complement U^c. The set of all boundary points of a set U is denoted by $\operatorname{bd}(U)$.

Examples:

$$(\mathbf{c} \in \mathbb{R}^n, r \in \mathbb{R}_{++}), \operatorname{bd}(B(\mathbf{c}, r)) =$$
$$(\mathbf{c} \in \mathbb{R}^n, r \in \mathbb{R}_{++}), \operatorname{bd}(B[\mathbf{c}, r]) =$$
$$\operatorname{bd}\left(\mathbb{R}^n_{++}\right) =$$
$$\operatorname{bd}\left(\mathbb{R}^n_{+}\right) =$$
$$\operatorname{bd}\left(\mathbb{R}^n\right) =$$
$$\operatorname{bd}\left(\Delta_n\right) =$$

Definition (Closure). *The closure of a set $U \subseteq \mathbb{R}^n$ is denoted by* $\operatorname{cl}(U)$ *and is defined to be the smallest closed set containing U :*

$$\operatorname{cl}(U) = \bigcap \{T : U \subseteq T, T \text{ is closed}\}.$$

Another equivalent definition of $\operatorname{cl}(U)$ *is:*

$$\operatorname{cl}(U) = U \cup \operatorname{bd}(U)$$

Examples:

$$\text{cl}\left(R^n_{++}\right) =$$
$$(\mathbf{x} \neq \mathbf{y}), \text{cl}((\mathbf{x}, \mathbf{y})) =$$

Boundedness and Compactness. A set $U \subseteq \mathbb{R}^n$ is called bounded if there exists $M > 0$ for which $U \subseteq B(0, M)$. A set $U \subseteq \mathbb{R}^n$ is called compact if it is closed and bounded. Examples of compact sets are: closed balls, unit simplex, and closed line segments.

Directional Derivatives and Gradients

Definition (Directional Derivative)**.** *Let f be a function defined on a set $S \subseteq \mathbb{R}^n$. Let $\mathbf{x} \in \text{int}(S)$ and let $\mathbf{d} \in \mathbb{R}^n$. If the limit*

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t}$$

exists, then it is called the directional derivative of f at $\mathbf{x}$ along the direction $\mathbf{d}$ and is denoted by $f'(\mathbf{x}; \mathbf{d})$. For any $i = 1, 2, \ldots, n$, if the limit

$$\lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x})}{t}$$

exists, then its value is called the i-th partial derivative and is denoted by $\frac{\partial f}{\partial x_i}(\mathbf{x})$. If all the partial derivatives of a function f exist at a point $\mathbf{x} \in \mathbb{R}^n$, then the gradient of f at $\mathbf{x}$ is

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

Definition (Continuous Differentiability)**.** *A function f defined on an open set $U \subseteq \mathbb{R}^n$ is called continuously differentiable over U if all the partial derivatives exist and are continuous on U. In that case,*

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d}, \quad \mathbf{x} \in U, \mathbf{d} \in \mathbb{R}^n.$$

Proposition. *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is continuously differentiable over U. Then*

$$\lim_{\mathbf{d} \to 0} \frac{f(\mathbf{x} + \mathbf{d}) - f(\mathbf{x}) - \nabla f(\mathbf{x})^\top \mathbf{d}}{\|\mathbf{d}\|} = 0, \quad \text{for all } \mathbf{x} \in U.$$

Another way to write the above result is as follows:

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + o(\|\mathbf{y} - \mathbf{x}\|)$$

where $o(\cdot) : \mathbb{R}_+ \to \mathbb{R}$ is a one-dimensional function satisfying $\frac{o(t)}{t} \to 0$ as $t \to 0$.

Twice Differentiability and the Hessian

The partial derivatives $\frac{\partial f}{\partial x_1}$ are themselves real-valued functions that can be partially differentiated. The (i, j)-partial derivatives of f at $\mathbf{x} \in U$ (if exists) is defined by

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial \left(\frac{\partial f}{\partial x_j}\right)}{\partial x_i}(\mathbf{x}) .$$

A function f defined on an open set $U \subseteq \mathbb{R}^n$ is called twice continuously differentiable over U if all the second order partial derivatives exist and are continuous over U. In that case, for any $i \neq j$ and any $\mathbf{x} \in U$:

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x})$$

Definition (The Hessian). *The Hessian of f at a point $\mathbf{x} \in U$ is the $n \times n$ matrix:*

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & & \vdots \\ \vdots & \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}$$

For twice continuously differentiable functions, the Hessian is a symmetric matrix.

Theorem (Linear Approximation Theorem). *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq R^n$. Suppose that f is twice continuously differentiable over U. Let $\mathbf{x} \in U$ and $r > 0$ satisfy $B(\mathbf{x}, r) \subseteq U$. Then for any $\mathbf{y} \in B(\mathbf{x}, r)$, there exists $\xi \in [\mathbf{x}, \mathbf{y}]$ such that:*

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\xi)(\mathbf{y} - \mathbf{x}) .$$

Theorem (Quadratic Approximation Theorem). *Let $f : U \to \mathbb{R}$ be defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U. Let $\mathbf{x} \in U$ and $r > 0$ satisfy $B(\mathbf{x}, r) \subseteq U$. Then for any $\mathbf{y} \in B(\mathbf{x}, r)$:*

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + o\left(\|\mathbf{y} - \mathbf{x}\|^2\right) .$$

The *little "o"* notation above can be interpreted as $o(\|\mathbf{x} - \mathbf{y}\|^2)$ goes faster to zero than $\|\mathbf{x} - \mathbf{y}\|^2$.

Gradient and Hessian of Quadratic Functions

Gradient of Linear Function

Consider a linear function of the form

$$f(\mathbf{w}) = \mathbf{a}^\top \mathbf{w}$$

where $\mathbf{a}$ and $\mathbf{w} \in \mathbb{R}^n$. We can derive the gradient in matrix notation as follows:

1. Convert to summation notation:

$$f(\mathbf{w}) = \sum_{j=1}^{n} a_j w_j \,.$$

2. Take the partial derivative with respect to a generic element k :

$$\frac{\partial}{\partial w_k}\left[\sum_{j=1}^{n} a_j w_j\right] = a_k z \,.$$

3. Assemble the partial derivatives into a vector:

$$\nabla f(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_n} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \mathbf{a}$$

So our final result is that

$$\nabla f(\mathbf{w}) = \mathbf{a} \,.$$

This generalizes the scalar case where $\frac{d}{dw}[\alpha w] = \alpha$. We can also consider general linear functions of the form

$$f(\mathbf{w}) = \mathbf{a}^\top w + \beta$$

for a scalar β, but in this case we still have $\nabla f(\mathbf{w}) = \mathbf{a}$ since β does not depend on $\mathbf{w}$.

Gradient of Quadratic Function

Consider a quadratic function of the form

$$f(\mathbf{w}) = \mathbf{w}^\top A \mathbf{w}$$

where $\mathbf{w} \in \mathbb{R}^n$ and A in a matrix in $\mathbb{R}^{n \times n}$. We can derive the gradient in matrix notation as follows:

1. Convert to summation notation:

$$f(\mathbf{w}) = \mathbf{w}^\top \begin{bmatrix} \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n a_{nj} w_j \end{bmatrix} = \sum_{i=1}^n \sum_{j=1}^n w_i a_{ij} w_j$$

   where a_{ij} is the element in row i and column j of A. To help with computing the partial derivatives, it is useful to re-write it in the form

$$f(\mathbf{w}) = \sum_{i=1}^n \sum_{j=1}^n w_i a_{ij} w_j = \sum_{i=1}^n \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right)$$

2. Take the partial derivative with respect to a generic element k :

$$\frac{\partial}{\partial w_k} \left[\sum_{i=1}^n \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right) \cdot \right] = 2 a_{kk} w_k + \sum_{j \neq k} w_j a_{jk} + \sum_{j \neq k} a_{kj} w_j$$

   The first term comes from the a_{kk} term that is quadratic in w_k, while the two sums come from the terms that are linear in w_k. We can move one $a_{kk} w_k$ into each of the sums to simplify this to

$$\frac{\partial}{\partial w_k} \left[\sum_{i=1}^n \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right) \cdot \right] = \sum_{j=1}^n w_j a_{jk} + \sum_{j=1}^n a_{kj} w_j$$

3. Assemble the partial derivatives into a vector:

$$\nabla f(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_n} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n w_j a_{j1} + \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n w_j a_{j2} + \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n w_j a_{jn} + \sum_{j=1}^n a_{nj} w_j \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n w_j a_{j1} \\ \sum_{j=1}^n w_j a_{j2} \\ \vdots \\ \sum_{j=1}^n w_j a_{jn} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n a_{dj} w_j \end{bmatrix}$$

4. Convert to matrix notation:

$$\nabla f(w) = \begin{bmatrix} \sum_{j=1}^n w_j a_{j1} \\ \sum_{j=1}^n w_j a_{j2} \\ \vdots \\ \sum_{j=1}^n w_j a_{jn} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n a_{nj} w_j \end{bmatrix} = A^\top \mathbf{w} + A\mathbf{w} = \left(A^\top + A \right) \mathbf{w}$$

So our final result is that

$$\nabla f(\mathbf{w}) = \left(A^\top + A \right) \mathbf{w}$$

Note that if A is symmetric $(A^\top = A)$ then we have $(A^\top + A) = (A + A) = 2A$ so we have

$$\nabla f(\mathbf{w}) = 2A\mathbf{w}$$

This generalizes the scalar case where $\frac{d}{dw}\left[\alpha w^2\right] = 2\alpha w$. We can also consider general quadratic functions of the form

$$f(w) = \frac{1}{2}\mathbf{w}^\top A\mathbf{w} + \mathbf{b}^\top w + \gamma$$

Using the above results we have

$$\nabla f(\mathbf{w}) = \frac{1}{2}\left(A^\top + A\right)\mathbf{w} + \mathbf{b}$$

and if A is symmetric then

$$\nabla f(\mathbf{w}) = A\mathbf{w} + \mathbf{b}$$

Hessian of a Quadratic Form

For a quadratic function of the form,

$$f(\mathbf{w}) = \mathbf{w}^\top A\mathbf{w}$$

we have shown that the partial derivatives are given by linear functions,

$$\frac{\partial f}{\partial w_k} = \sum_{j=1}^{n} w_j a_{jk} + \sum_{j=1}^{n} a_{kj} w_j \,.$$

The second partial derivatives are thus constant functions of the form

$$\frac{\partial^2 f}{\partial w_k \partial w_{k'}} = a_{k'k} + a_{kk'}$$

which means that the Hessian matrix has a simple form

$$\nabla^2 f(\mathbf{w}) = \begin{bmatrix} \frac{\partial}{\partial w_1 \partial w_1} f(\mathbf{w}) & \frac{\partial}{\partial w_1 \partial w_2} f(\mathbf{w}) & \cdots & \frac{\partial}{\partial w_1 \partial w_n} f(\mathbf{w}) \\ \frac{\partial w_2 \partial w_1}{\partial w_2 \theta w_2} f(\mathbf{w}) & \frac{\partial}{\partial w_2 \partial w_2} f(\mathbf{w}) & \cdots & \frac{\partial}{\partial w_2 \partial w_n} f(\mathbf{w}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_n \partial w_1} f(\mathbf{w}) & \frac{\partial}{\partial w_n \partial w_2} f(\mathbf{w}) & \cdots & \frac{\partial}{\partial w_n \partial w_n} f(\mathbf{w}) \end{bmatrix} = \begin{bmatrix} a_{11} + a_{11} & a_{21} + a_{12} & \cdots & a_{n1} + a_{1n} \\ a_{12} + a_{21} & a_{22} + a_{22} & \cdots & a_{n2} + a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} + a_{n1} & a_{2n} + a_{n2} & \cdots & a_{nn} + a_{nn} \end{bmatrix}$$

This gives a result of

$$\nabla^2 f(\mathbf{w}) = A + A^\top \,,$$

and if A is symmetric this simplifies to

$$\nabla^2 f(\mathbf{w}) = 2A \,.$$

Notation for this module

Symbol	Object
$\mathbb{R}^n$	the space of $n-$dimensional real column vectors
$\mathbf{x}$	a vector in $\mathbb{R}^n$
x_i	the $i-$th coordinate of a vector $\mathbf{x}$
$\langle \mathbf{x}, \mathbf{y} \rangle$	inner product of $\mathbf{x}$ and $\mathbf{y}$
$[\mathbf{x}, \mathbf{y}]$	closed line segment between $\mathbf{x}$ and $\mathbf{y}$
$(\mathbf{x}, \mathbf{y})$	open line segment between $\mathbf{x}$ and $\mathbf{y}$
$B(\mathbf{c}, r)$	open ball with center $\mathbf{c}$ and radius r
$B[\mathbf{c}, r]$	closed ball with center $\mathbf{c}$ and radius r
$\mathbb{R}^{m \times n}$	space of $m \times n$ real-valued matrices
$\mathbf{A}^T$	transpose of $\mathbf{A}$
$\mathbf{e}_i$	$i-$th vector in the standard basis of $\mathbb{R}^n$
$\mathbf{e}$	vector of all ones
$\mathbf{0}$	vector of all zeros
$\lvert \cdot \rvert$	absolute value of scalar x
$\lVert \cdot \rVert_p$	ℓ_p-norm for $\mathbf{x} \in \mathbb{R}^n$
Δ_n	unit simplex
$\mathbb{R}^n_+$	nonnegative orthant
$\mathbb{R}^n_{++}$	positive orthant
$\lVert \mathbf{A} \rVert_F$	Frobenius norm of $\mathbf{A}$
$\lVert \mathbf{A} \rVert_{ab}$	induced norm of $\mathbf{A} \in \mathbb{R}^{m \times n}$
$\lVert \mathbf{A} \rVert_2$	spectral norm of $\mathbf{A}$
$\lambda_{max}(\mathbf{A})$	maximum eigenvalue of a symmetric matrix $\mathbf{A}$
$\lambda_{min}(\mathbf{A})$	minimum eigenvalue of a symmetric matrix $\mathbf{A}$
$int(S)$	interior of set S
$f'(\mathbf{x}; \mathbf{d})$	directional derivative of f at x in the direction $\mathbf{d}$
$\nabla f(\mathbf{x})$	gradient of f at $\mathbf{x}$
$\nabla^2 f(\mathbf{x})$	Hessian of $f(\mathbf{x})$ at $\mathbf{x}$
$C_L^{1,1}(D)$	class of $L-$smooth functions over D
$\mathbf{I}_n$	identity matrix in $R^{n \times n}$
$\mathbf{0}_{mn}$	zero matrix in $R^{m \times n}$
$diag(\mathbf{x})$	diagonal matrix with entries $\mathbf{x}$

Part II.
Unconstrained Optimization

Global Minimum and Maximum

Definition (Global Minimum and Maximum). *Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$. Then:*

1. $\mathbf{x}^* \in S$ *is a global minimum point of* f *over* S *if* $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ *for any* $\mathbf{x} \in S$.
2. $\mathbf{x}^* \in S$ *is a strict global minimum point of* f *over* S *if* $f(\mathbf{x}) > f(\mathbf{x}^*)$ *for any* $\mathbf{x}^* \neq \mathbf{x} \in S$.
3. $\mathbf{x}^* \in S$ *is a global maximum point of* f *over* S *if* $f(\mathbf{x}) \leq f(\mathbf{x}^*)$ *for any* $\mathbf{x} \in S$.
4. $\mathbf{x}^* \in S$ *is a strict global maximum point of* f *over* S *if* $f(\mathbf{x}) < f(\mathbf{x}^*)$ *for any* $\mathbf{x}^* \neq \mathbf{x} \in S$.

Some definitions:

- We denote by **global optimum** the global minimum or maximum.
- The maximal value of f over S:

$$\sup\{f(\mathbf{x}) : \mathbf{x} \in S\}.$$

- The minimal value of f over S :

$$\inf\{f(\mathbf{x}) : \mathbf{x} \in S\}.$$

Note that the minimal and maximal values are always unique.

Example: Find the global minimum and maximum points of $f(x_1, x_2) = x_1 + x_2$ over the unit ball in $\mathbb{R}^2$, $S = B[0,1] = \left\{(x_1, x_2)^\top : x_1^2 + x_2^2 \leq 1\right\}$.

Another example: What about the global maximizers and minimizers in $\mathbb{R}^2$ for

$$f(x_1, x_2) = \frac{x_1 + x_2}{x_1^2 + x_2^2 + 1} \quad ?$$

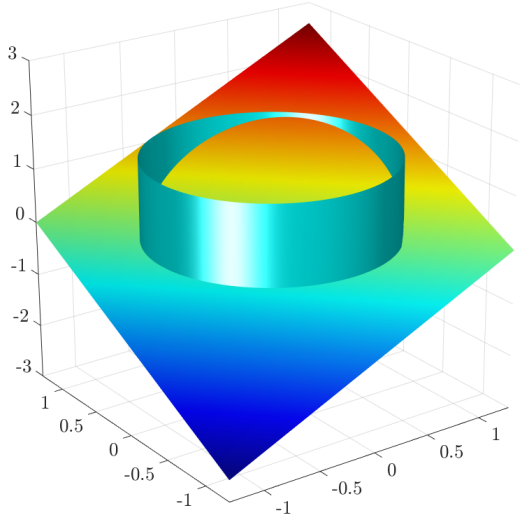


Figure 2: The function $f(x_1, x_2) = x_1 + x_2$ constrained over the unit ball.

Local Minima and Maxima

Definition (Local Minimum and Maximum). *Let $f : S \to \mathbb{R}$ be defined on a set $S \subseteq \mathbb{R}^n$. Then:*

1. $\mathbf{x}^* \in S$ *is a local minimum of* f *over* S *if there exists* $r > 0$ *for which* $f(\mathbf{x}^*) \leq f(\mathbf{x})$ *for any* $\mathbf{x} \in S \cap B(\mathbf{x}^*, r)$.
2. $\mathbf{x}^* \in S$ *is a strict local minimum of* f *over* S *if there exists* $r > 0$ *for which* $f(\mathbf{x}^*) < f(\mathbf{x})$ *for any* $\mathbf{x} \neq \mathbf{x}^*$ *in* $S \cap B(\mathbf{x}^*, r)$.
3. $\mathbf{x}^* \in S$ *is a local maximum of* f *over* S *if there exists* $r > 0$ *for which* $f(\mathbf{x}^*) \geq f(\mathbf{x})$ *for any* $\mathbf{x} \in S \cap B(\mathbf{x}^*, r)$.
4. $\mathbf{x}^* \in S$ *is a strict local maximum of* f *over* S *if there exists* $r > 0$ *for which* $f(\mathbf{x}^*) > f(\mathbf{x})$ *for any* $\mathbf{x} \neq \mathbf{x}^*$ *in* $S \cap B(\mathbf{x}^*, r)$.

Of course, a global minimum (maximum) point is also a local minimum (maximum) point.

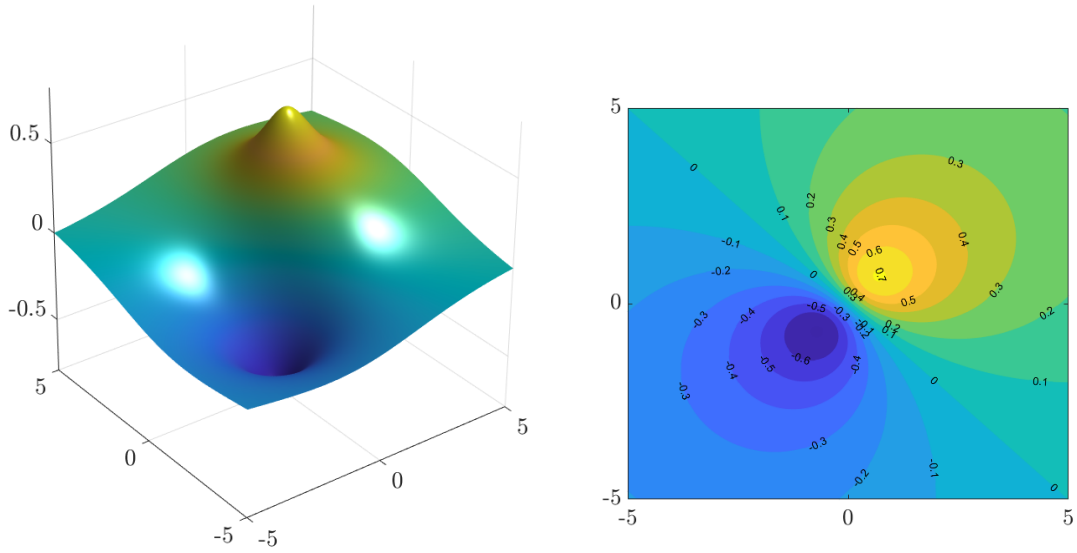


Figure 3: The function $f(x_1, x_2) = \frac{x_1+x_2}{x_1^2+x_2^2+1}$ (left) and its contour plot (right).

Example: Identify the different (strict) local minima and maxima of the function

$$f(x) = \begin{cases} (x-1)^2 + 2, & -1 \le x \le 1 \\ 2, & 1 \le x \le 2 \\ -(x-2)^2 + 2, & 2 \le x \le 2.5 \\ (x-3)^2 + 1.5, & 2.5 \le x \le 4 \\ -(x-5)^2 + 3.5, & 4 \le x \le 6 \\ -2x + 14.5, & 6 \le x \le 6.5 \\ 2x - 11.5, & 6.5 \le x \le 8 \end{cases} \tag{1}$$

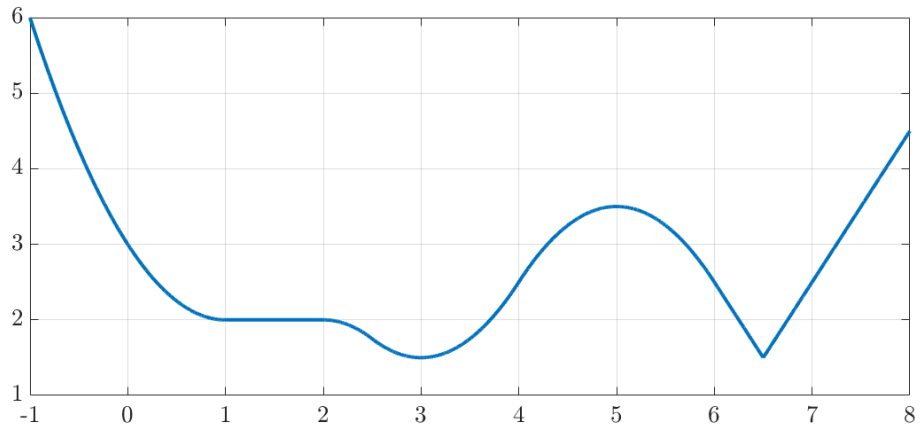


Figure 4: The function described in (1) has several minima and maxima.

Theorem (Fermat's Theorem: First Order Optimality Conditions). *Let $f : U \to \mathbb{R}$ be a*

function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\mathbf{x}^* \in \text{int}(U)$ is a local optimum point and that all the partial derivatives of f exist at $\mathbf{x}^*$. Then $\nabla f(\mathbf{x}^*) = 0$.

Proof. Let $i \in \{1, 2, \ldots, n\}$ and consider the 1-D function $g(t) = f(\mathbf{x}^* + t\mathbf{e}_i)$. Then, $\mathbf{x}^*$ is a local optimum point of f implies that $t = 0$ is a local optimum of g, and hence $g'(0) = 0$. Thus, $\frac{\partial f}{\partial x_i}(\mathbf{x}^*) = g'(0) = 0$.

Definition (Stationary Points)**.** *Let $f : U \longrightarrow \mathbb{R}$ be a function defined on a set $U \subseteq \mathbb{R}^n$. Suppose that $\mathbf{x}^* \in \text{int}(U)$ and that all the partial derivatives of f are defined at $\mathbf{x}^*$. Then $\mathbf{x}^*$ is called a stationary point of f if $\nabla f(\mathbf{x}^*) = 0$.*

Figure 5: Pierre de Fermat (1607-1665), French mathematician (and lawyer!) who made significant contributions to differential calculus and number theory.

Example:

$$\min \left\{ f(x_1, x_2) = \tfrac{x_1 + x_2}{x_1^2 + x_2^2 + 1} : x_1, x_2 \in \mathbb{R} \right\}$$

$$\nabla f(x_1, x_2) = \frac{1}{\left(x_1^2 + x_2^2 + 1\right)^2} \begin{pmatrix} \left(x_1^2 + x_2^2 + 1\right) - 2(x_1 + x_2)x_1 \\ \left(x_1^2 + x_2^2 + 1\right) - 2(x_1 + x_2)x_2 \end{pmatrix}$$

Stationary points are those satisfying:

$$\begin{aligned} -x_1^2 - 2x_1x_2 + x_2^2 &= -1 \\ x_1^2 - 2x_1x_2 - x_2^2 &= -1\,. \end{aligned}$$

Hence, the stationary points are $(1/\sqrt{2}, 1/\sqrt{2})$ and $(-1/\sqrt{2}, -1/\sqrt{2})$, with $(1/\sqrt{2}, 1/\sqrt{2})$ a global maximum, and $(-1/\sqrt{2}, -1/\sqrt{2})$ a global minimum.

Second Order Optimality Conditions

Theorem (Necessary Second Order Optimality Conditions). *Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$ Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point. Then*

1. *if $\mathbf{x}^*$ is a local minimum point, then $\nabla^2 f(\mathbf{x}^*) \succeq 0$.*
2. *if $\mathbf{x}^*$ is a local maximum point, then $\nabla^2 f(\mathbf{x}^*) \preceq 0$.*

Proof. We prove 1. There exists a ball $B(\mathbf{x}^*, r) \subseteq U$ for which $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for all $\mathbf{x} \in B(\mathbf{x}^*, r)$. Next, let $\mathbf{d} \in \mathbb{R}^n$ be a nonzero vector. For any $0 < \alpha < \frac{r}{\|\mathbf{d}\|}$, we have $\mathbf{x}^*_\alpha \equiv \mathbf{x}^* + \alpha\mathbf{d} \in B(\mathbf{x}^*, r)$ and for any such α, $f\left(\mathbf{x}^*_\alpha\right) \geq f(\mathbf{x}^*)$.

On the other hand, there exists a vector $\mathbf{z}_\alpha \in \left[\mathbf{x}^*, \mathbf{x}^*_\alpha\right]$ such that

$$f\left(\mathbf{x}^*_\alpha\right) - f(\mathbf{x}^*) = \frac{\alpha^2}{2}\mathbf{d}^\top \nabla^2 f(\mathbf{z}_\alpha)\,\mathbf{d}\,.$$

This implies that for any $\alpha \in \left(0, \frac{r}{\|\mathbf{d}\|}\right)$ the inequality $\mathbf{d}^\top \nabla^2 f(\mathbf{z}_\alpha)\,\mathbf{d} \geq 0$ holds. Since $\mathbf{z}_\alpha \to \mathbf{x}^*$ as $\alpha \to 0^+$, we obtain that $\mathbf{d}^\top \nabla^2 f(\mathbf{x}^*)\,\mathbf{d} \geq 0$, which leads to $\nabla^2 f(\mathbf{x}^*) \succeq 0$. The proof of 2. follows analogously. □

Theorem (Sufficient Second Order Optimality Conditions). *Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point. Then*

1. *if $\nabla^2 f(\mathbf{x}^*) \succ 0$, then $\mathbf{x}^*$ is a strict local minimum point of f over U.*
2. *if $\nabla^2 f(\mathbf{x}^*) \prec 0$, then $\mathbf{x}^*$ is a strict local maximum point of f over U.*

Proof. We prove 1. There exists a ball $B(\mathbf{x}^*, r) \subseteq U$ for which $\nabla^2 f(\mathbf{x}) \succ 0$ for any $\mathbf{x} \in B(\mathbf{x}^*, r)$. Then, by the Linear Approximation Theorem (week 1), there exists a vector $\mathbf{z}_\mathbf{x} \in [\mathbf{x}^*, \mathbf{x}]$ (and hence $\mathbf{z}_\mathbf{x} \in B(\mathbf{x}^*, r)$) for which

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{z}_\mathbf{x})(\mathbf{x} - \mathbf{x}^*)\,.$$

From this, it follows that $\nabla^2 f(\mathbf{z}_\mathbf{x}) \succ 0$, implying that for any $\mathbf{x} \in B(\mathbf{x}^*, r)$ such that $\mathbf{x} \neq \mathbf{x}^*$, the inequality $f(\mathbf{x}) > f(\mathbf{x}^*)$ holds. Therefore $\mathbf{x}^*$ is a strict local minimum point of f over U. □

Saddle Points

Definition (Saddle Point). *Let $f : U \to \mathbb{R}$ be a continuously differentiable function defined on an open set $U \subseteq \mathbb{R}^n$. A stationary point $\mathbf{x}^* \in U$ is called a saddle point of f over U if it is neither a local minimum point nor a local maximum point of f over U.*

Theorem (Sufficient Condition for Saddle Points). *Let $f : U \to \mathbb{R}$ be a function defined on an open set $U \subseteq \mathbb{R}^n$. Suppose that f is twice continuously differentiable over U and that $\mathbf{x}^*$ is a stationary point. If $\nabla^2 f(\mathbf{x}^*)$ is an indefinite matrix, then $\mathbf{x}^*$ is a saddle point of f over U.*

Proof. The Hessian $\nabla^2 f(\mathbf{x}^*)$ has at least one positive eigenvalue $\lambda > 0$, corresponding to a normalized eigenvector denoted by $\mathbf{v}$. There exists a radius $r > 0$ such that $\mathbf{x}^* + \alpha\mathbf{v} \in U$ for any $\alpha \in (0, r)$. By the Quadratic Approximation Theorem (week 1), there exists a function $g : \mathbb{R}_{++} \to \mathbb{R}$ satisfying

$$\frac{g(t)}{t} \to 0 \text{ as } t \to 0,$$

such that for any $\alpha \in (0, r)$

$$f(\mathbf{x}^* + \alpha\mathbf{v}) = f(\mathbf{x}^*) + \frac{\lambda\alpha^2}{2}\|\mathbf{v}\|^2 + g\left(\|\mathbf{v}\|^2\alpha^2\right).$$

Recalling that $\mathbf{v}$ is normalized, we write

$$f(\mathbf{x}^* + \alpha\mathbf{v}) = f(\mathbf{x}^*) + \frac{\lambda\alpha^2}{2} + g\left(\alpha^2\right)$$

By the properties of g, it follows that there exists an $\varepsilon_1 \in (0, r)$ such that $g\left(\alpha^2\right) > -\frac{\lambda}{2}\alpha^2$ for all $\alpha \in (0, \varepsilon_1)$ and hence $f(\mathbf{x}^* + \alpha\mathbf{v}) > f(\mathbf{x}^*)$ for all $\alpha \in (0, \varepsilon_1)$. This shows that $\mathbf{x}^*$ cannot be a local maximum point of f over U. A similar argument-exploiting an eigenvector of $\nabla^2 f(\mathbf{x}^*)$ corresponding to a negative eigenvalue-shows that $\mathbf{x}^*$ cannot be a local minimum point of f over U, establishing the desired result that $\mathbf{x}^*$ is a saddle point. □

Attainment of Minimal/Maximal Points

Theorem (Weierstrass' Theorem). *Let f be a continuous function defined over a nonempty compact set $C \subseteq \mathbb{R}^n$. Then there exists a global minimum point of f over C and a global maximum point of f over C.*

When the underlying set is not compact, Weierstrass theorem does not guarantee the attainment of the solution, but certain properties of the function f can imply attainment of the solution.

Figure 6: Karl Weierstrass (1815-1897), German mathematician considered one of the founding figures of modern analysis. He formalized the definition of continuity of a function! If you want to pursue a research career in analysis, you might want to work/visit the Weierstrass Institute in Berlin.

Definition (Coerciveness)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function over $\mathbb{R}^n$. f is called coercive if*

$$\lim_{\|\mathbf{x}\| \to \infty} f(\mathbf{x}) = \infty$$

Theorem (Attainment of Global Optima Points for Coercive Functions)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous and coercive function and let $S \subseteq \mathbb{R}^n$ be a nonempty closed set. Then f attains a global minimum point on S.*

Proof. Let $\mathbf{x}_0 \in S$ be an arbitrary point in S. Since the function is coercive, it follows that there exists an $M > 0$ such that

$$f(\mathbf{x}) > f(\mathbf{x}_0) \text{ for any } \mathbf{x} \text{ such that } \|\mathbf{x}\| > M\,.$$

Since any global minimizer $\mathbf{x}^*$ of f over S satisfies $f(\mathbf{x}^*) \leq f(\mathbf{x}_0)$, it follows that the set of global minimizers of f over S is the same as the set of global minimizers of f over $S \cap B[0, M]$. The set $S \cap B[0, M]$ is compact and nonempty, and thus by the Weierstrass theorem, there exists a global minimizer of f over $S \cap B[0, M]$ and hence also over S. □

Example. Classify the stationary points of the function $f(x_1, x_2) = -2x_1^2 + x_1x_2^2 + 4x_1^4$. The gradient is given by

$$\nabla f(x) = \begin{pmatrix} -4x_1 + x_2^2 + 16x_1^3 \\ 2x_1x_2 \end{pmatrix},$$

and the stationary points are the solutions to

$$\begin{aligned} -4x_1 + x_2^2 + 16x_1^3 &= 0 \\ 2x_1x_2 &= 0 \end{aligned}.$$

The stationary points are $(0,0)$, $(0.5,0)$, and $(-0.5,0)$. We compute te Hessian

$$\nabla^2 f\left(x_1, x_2\right) = \begin{pmatrix} -4 + 48x_1^2 & 2x_2 \\ 2x_2 & 2x_1 \end{pmatrix},$$

and evaluating at the stationary points leads to

$$\nabla^2 f(0.5,0) = \begin{pmatrix} 8 & 0 \\ 0 & 1 \end{pmatrix}, \quad \nabla^2 f(-0.5,0) = \begin{pmatrix} 8 & 0 \\ 0 & -1 \end{pmatrix}, \quad \nabla^2 f(0,0) = \begin{pmatrix} -4 & 0 \\ 0 & 0 \end{pmatrix}.$$

Classify each stationary point.

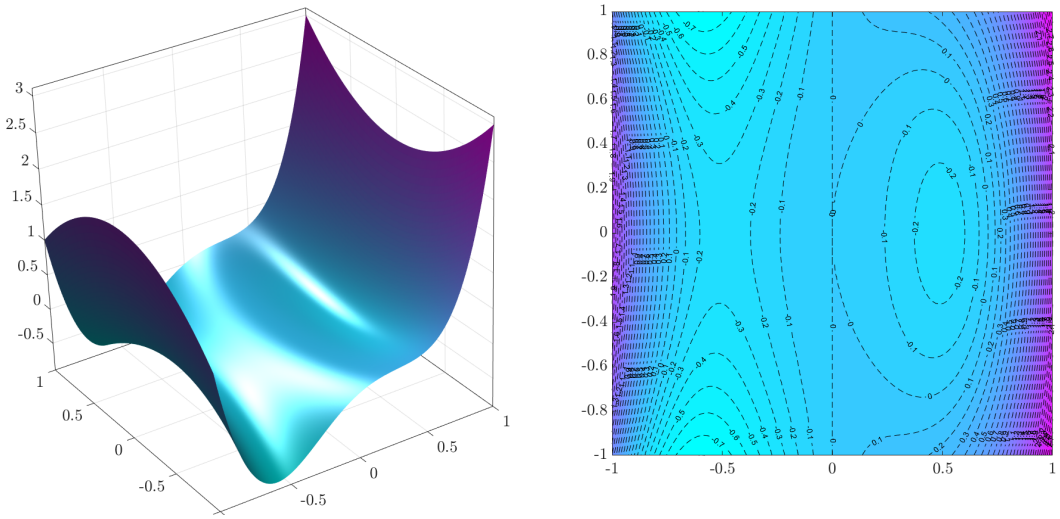


Figure 7: The function $f(x_1, x_2) = -2x_1^2 + x_1x_2^2 + 4x_1^4$ and its contour plot.

Global Optimality Conditions

Theorem (Global Optimality Condition)**.** *Let f be a twice continuously differentiable function defined over $\mathbb{R}^n$. Suppose that $\nabla^2 f(\mathbf{x}) \geq 0$ for any $\mathbf{x} \in \mathbb{R}^n$. Let $\mathbf{x}^* \in \mathbb{R}^n$ be a stationary point of f. Then $\mathbf{x}^*$ is a global minimum point of f.*

Proof. By the Linear Approximation Theorem, it follows that for any $\mathbf{x} \in \mathbb{R}^n$, there exists a vector $\mathbf{z_x} \in [\mathbf{x}^*, \mathbf{x}]$ for which

$$f(\mathbf{x}) - f(\mathbf{x}^*) = \frac{1}{2} (\mathbf{x} - \mathbf{x}^*)^\top \nabla^2 f(\mathbf{z_x}) (\mathbf{x} - \mathbf{x}^*) .$$

Since $\nabla^2 f(\mathbf{z_x}) \geq 0$, we have that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$, establishing the fact that $\mathbf{x}^*$ is a global minimum point of f. □

Example. The function $f(\mathbf{x}) = x_1^2 + x_2^2 + x_3^2 + x_1x_2 + x_1x_3 + x_2x_3 + \left(x_1^2 + x_2^2 + x_3^2\right)^2$ has a gradient

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2x_1 + x_2 + x_3 + 4x_1\left(x_1^2 + x_2^2 + x_3^2\right) \\ 2x_2 + x_1 + x_3 + 4x_2\left(x_1^2 + x_2^2 + x_3^2\right) \\ 2x_3 + x_1 + x_2 + 4x_3\left(x_1^2 + x_2^2 + x_3^2\right) \end{pmatrix},$$

and the Hessian

$$\nabla^2 f(x) = \begin{pmatrix} 2 + 4\left(x_1^2 + x_2^2 + x_3^2\right) + 8x_1^2 & 1 + 8x_1x_2 & 1 + 8x_1x_3 \\ 1 + 8x_1x_2 & 1 + 8x_2x_3 2 + 4\left(x_1^2 + x_2^2 + x_3^2\right) + 8x_2^2 & \\ 1 + 8x_1x_3 & 1 + 8x_2x_3 & 2 + 4\left(x_1^2 + x_2^2 + x_3^2\right) + 8x_3^2 \end{pmatrix}.$$

We observe that $\mathbf{x} = 0$ is a stationary point. Check that $\nabla^2 f(\mathbf{x}) \geq 0$ for all $\mathbf{x}$. It follows that $\mathbf{x} = 0$ is the global minimum point.

Quadratic Functions

A **quadratic function** over $\mathbb{R}^n$ is a function of the form

$$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{c} \in \mathbb{R}$. From here onwards, we adopt the convention of assuming $\mathbf{A}$ to be symmetric. [1]

Check(!) that

$$\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x} + 2\mathbf{b}$$
$$\nabla^2 f(\mathbf{x}) = 2\mathbf{A}$$

[1] Otherwise, work with $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{W}\mathbf{x} + \mathbf{b}^\top \mathbf{x} + c$, where $\mathbf{W} = \frac{\mathbf{A} + \mathbf{A}^\top}{2}$ (it's the same!).

Proposition. *Let* $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c}$, *with* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *symmetric*, $\mathbf{b} \in \mathbb{R}^n$, *and* $\mathbf{c} \in \mathbb{R}$. *Then*

1. $\mathbf{x}$ *is a stationary point of* f *iff* $\mathbf{A}\mathbf{x} = -\mathbf{b}$.
2. *if* $\mathbf{A} \succeq 0$, *then* $\mathbf{x}$ *is a global minimum point of* f *iff* $\mathbf{A}\mathbf{x} = -\mathbf{b}$.
3. *if* $\mathbf{A} \succ 0$, *then* $\mathbf{x} = -\mathbf{A}^{-1}\mathbf{b}$ *is a strict global minimum point of* f.

Proof.

1. The proof follows immediately from the formula of the gradient of f.
2. since $\nabla^2 f(\mathbf{x}) = 2\mathbf{A} \succeq 0$, it follows from global optimality conditions that the global minimum points are exactly the stationary points, which combined with part 1. implies the result.
3. When $\mathbf{A} \succ 0$, the vector $\mathbf{x} = -\mathbf{A}^{-1}\mathbf{b}$ is the unique solution to $\mathbf{A}\mathbf{x} = -\mathbf{b}$, and hence by parts 1. and 2., it is the unique global minimum point of f.

□

Two Important Theorems on Quadratic Functions

Lemma (Coerciveness of Quadratic Functions)**.** *Let* $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}x + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c}$ *where* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is symmetric,* $\mathbf{b} \in \mathbb{R}^n$ *and* $\mathbf{c} \in \mathbb{R}$. *Then* f *is coercive if and only if* $\mathbf{A} \succ \mathbf{0}$.

Theorem (Characterization of the Nonnegativity of Quadratic Functions)**.** *Consider the quadratic form* $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c}$, *where* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *is symmetric,* $\mathbf{b} \in \mathbb{R}^n$ *and* $\mathbf{c} \in \mathbb{R}$. *Then the following two claims are equivalent:*

1. $f(\mathbf{x}) \equiv \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c} \geq 0$, *for all* $\mathbf{x} \in \mathbb{R}^n$.
2. *The augmented matrix* $\begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & \mathbf{c} \end{pmatrix} \succeq 0$.

Proof. Suppose that 2. holds. Then, in particular for any $\mathbf{x} \in \mathbb{R}^n$ the inequality

$$\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}^\top \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & c \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} \geq 0$$

holds, which is the same as the inequality $\mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c} \geq 0$, proving the validity of 1. Now, assume that 1. holds. We begin by showing that $\mathbf{A} \succeq 0$. Suppose in contradiction that $\mathbf{A}$ is not positive semidefinite. Then there exists an eigenvector $\mathbf{v}$ corresponding to a negative eigenvalue $\lambda < 0$ of $\mathbf{A}$:

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}\,.$$

Thus, for any $\alpha \in \mathbb{R}$

$$f(\alpha \mathbf{v}) = \lambda \|\mathbf{v}\|^2 \alpha^2 + 2\left(\mathbf{b}^\top \mathbf{v}\right)\alpha + \mathbf{c} \longrightarrow -\infty$$

as $\alpha \longrightarrow -\infty$, contradicting the nonnegativity of f.

Now our objective is to prove 2.; that is, we want to show that for any $\mathbf{y} \in \mathbb{R}^n$ and $t \in \mathbb{R}$

$$\begin{pmatrix} \mathbf{y} \\ t \end{pmatrix}^\top \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}^\top & \mathbf{c} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ t \end{pmatrix} \geq 0$$

which is equivalent to

$$\mathbf{y}^\top \mathbf{A}\mathbf{y} + 2t\mathbf{b}^\top \mathbf{y} + \mathbf{c}t^2 \geq 0 \tag{2}$$

To show the validity of eq. (2) above for any $\mathbf{y} \in \mathbb{R}^n$ and $t \in \mathbb{R}$, we consider two cases. If $t = 0$, then (2) reads as $\mathbf{y}^\top \mathbf{A}\mathbf{y} \geq 0$, which is a valid inequality since we have shown that $\mathbf{A} \succeq 0$. The second case is when $t \neq 0$. To show that (2) holds in this case, note that (2) is the same as the inequality

$$t^2 f\left(\frac{\mathbf{y}}{t}\right) = t^2 \left[\left(\frac{\mathbf{y}}{t}\right)^\top \mathbf{A} \left(\frac{\mathbf{y}}{t}\right) + 2\mathbf{b}^\top \left(\frac{\mathbf{y}}{t}\right) + \mathbf{c}\right] \geq 0\,,$$

which holds true by the nonnegativity of f. □

Appendix: Classification of Matrices

Definition (Positive Definiteness). *1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called positive semidefinite, denoted by $\mathbf{A} \succeq \mathbf{0}$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$ for every $\mathbf{x} \in \mathbb{R}^n$.*

2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called positive definite, denoted by $\mathbf{A} \succ \mathbf{0}$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0$ for every $\mathbf{x} \neq \mathbf{0} \in \mathbb{R}^n$.

Exercise: check for

$$\mathbf{A} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}.$$

Proposition. *Let $\mathbf{A}$ be a positive definite (semidefinite) matrix. Then the diagonal elements of $\mathbf{A}$ are positive (nonnegative).*

Definition (Negative (Semi) Definiteness, Indefiniteness). *1. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called negative semidefinite, denoted by $\mathbf{A} \preceq \mathbf{0}$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} \leq 0$ for every $\mathbf{x} \in \mathbb{R}^n$.*

2. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called negative definite, denoted by $\mathbf{A} \prec \mathbf{0}$, if $\mathbf{x}^\top \mathbf{A}\mathbf{x} < 0$ for every $\mathbf{0} \neq \mathbf{x} \in \mathbb{R}^n$.

3. A symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called indefinite if there exist $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ such that $\mathbf{x}^\top \mathbf{A}\mathbf{x} > 0, \mathbf{y}^\top \mathbf{A}\mathbf{y} < 0$.

Remark.

- *$\mathbf{A}$ is negative (semi)definite if and only if $-\mathbf{A}$ is positive (semi)definite.*
- *A matrix is indefinite if and only if it is neither positive semidefinite nor negative semidefinite.*

- *A symmetric matrix is with positive and negative elements in the diagonal is indefinite.*
- *The sum of two positive(negative) (semi)definite matrices is positive(negative) (semi)definite.*

Theorem (Eigenvalue Characterization)**.** *Let* $\mathbf{A}$ *be a symmetric* $n \times n$ *matrix. Then:*

a) $\mathbf{A}$ *is positive definite iff all its eigenvalues are positive.*

b) $\mathbf{A}$ *is positive semidefinite iff all its eigenvalues are nonnegative.*

c) $\mathbf{A}$ *is negative definite iff all its eigenvalues are negative.*

d) $\mathbf{A}$ *is negative semidefinite iff all its eigenvalues are nonpositive.*

e) $\mathbf{A}$ *is indefinite iff it has at least one positive eigenvalue and at least one negative eigenvalue.*

Proof. Part a). There exists orthogonal $\mathbf{U} \in \mathbb{R}^{n \times n}$ such that

$$\mathbf{U}^\top \mathbf{A}\mathbf{U} = \mathbf{D} \equiv \operatorname{diag}(d_1, d_2, \ldots, d_n) \ ,$$

where $d_i = \lambda_i(\mathbf{A})$. Making the linear change of variables $\mathbf{x} = \mathbf{U}\mathbf{y}$, we have

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = \mathbf{y}^\top \mathbf{U}^\top \mathbf{A}\mathbf{U}\mathbf{y} = \mathbf{y}^\top \mathbf{D}\mathbf{y} = \sum_{i=1}^{n} d_i y_i^2 \ .$$

Therefore, $\mathbf{x}^\top A\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$ iff

$$\sum_{i=1}^{n} d_i y_i^2 > 0 \text{ for any } \mathbf{y} \neq 0 \ .$$

The latter holds iff $d_i > 0$ for all i. (why?) □

Trace and Determinant. Let $\mathbf{A}$ be a positive semidefinite (definite) matrix. Then the trace $Tr(\mathbf{A})$ and the determinant $\det(\mathbf{A})$ are nonnegative (positive). This follows directly recalling that the trace of a matrix is the sum of its eigenvalues and the determinant its product.

Exercises. Let $\mathbf{A}$ be a symmetric 2×2 matrix. Then $\mathbf{A}$ is positive semidefinite (definite) if and only if $\operatorname{Tr}(\mathbf{A}), \det(\mathbf{A}) \geq 0$ ($\operatorname{Tr}(\mathbf{A}), \det(\mathbf{A}) > 0$).

Classify the matrices

$$\mathbf{A} = \begin{pmatrix} 4 & 1 \\ 1 & 3 \end{pmatrix}, \mathbf{B} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 0.1 \end{pmatrix}.$$

The Principal Minors Criteria

Given an $n \times n$ matrix, the determinant of the upper left $k \times k$ submatrix is called the k-th principal minor and is denoted by $D_k(\mathbf{A})$.

Example:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

$$D_1(\mathbf{A}) = a_{11}, D_2(\mathbf{A}) = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, D_3(\mathbf{A}) = \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

Proposition. *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix. Then* $\mathbf{A}$ *is positive definite if and only if*

$$D_1(\mathbf{A}) > 0, D_2(\mathbf{A}) > 0, \ldots, D_n(\mathbf{A}) > 0\,.$$

Proposition. *Let* $\mathbf{A}$ *be an* $n \times n$ *symmetric matrix. Then* $\mathbf{A}$ *is negative definite if and only if*

$$(-1)^k D_k(\mathbf{A}) > 0\,, \quad \textit{for all } \; k = 1, \ldots, n.$$

This is equivalent to check that $-\mathbf{A}$ *is positive definite.*

Exercise. Classify the matrices

$$\mathbf{A} = \begin{pmatrix} 4 & 2 & 3 \\ 2 & 3 & 2 \\ 3 & 2 & 4 \end{pmatrix}, \; \mathbf{B} = \begin{pmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & -1 \end{pmatrix}, \; \mathbf{C} = \begin{pmatrix} -4 & 1 & 1 \\ 1 & -4 & 1 \\ 1 & 1 & -4 \end{pmatrix}.$$

Diagonal Dominance

Definition (Diagonal Dominance)**.** *Let* A *be a symmetric* $n \times n$ *matrix.*

a) $\mathbf{A}$ *is called diagonally dominant if*

$$|\mathbf{A}_{ii}| \geq \sum_{j \neq i} |\mathbf{A}_{ij}| \;\; \forall i = 1, 2, \ldots, n$$

b) $\mathbf{A}$ *is called strictly diagonally dominant if*

$$|\mathbf{A}_{ii}| > \sum_{j \neq i} |\mathbf{A}_{ij}| \;\; \forall i = 1, 2, \ldots, n$$

Theorem (Positive definiteness of diagonally dominant matrices)**.** *a) If the matrix* $\mathbf{A}$ *is symmetric, diagonally dominant with nonnegative diagonal elements, then* $\mathbf{A}$ *is positive semidefinite.*

b) *If* $\mathbf{A}$ *is symmetric, strictly diagonally dominant with positive diagonal elements, then* $\mathbf{A}$ *is positive definite.*

Proof. a) Suppose in contradiction that there exists a negative eigenvalue λ of $\mathbf{A}$, and let u be a corresponding eigenvector. Let $i \in \{1, 2, \ldots, n\}$ be an index for which $|u_i|$ is maximal among $|u_1|, |u_2|, \ldots, |u_n|$. Then by the equality $\mathbf{A}\mathrm{u} = \lambda \mathrm{u}$ we have

$$|\mathbf{A}_{ii} - \lambda| \cdot |u_i| = \left| \sum_{j \neq i} \mathbf{A}_{ij} u_j \right| \leq \left(\sum_{j \neq i} \left| \mathbf{A}_{ij} \right| \right) |u_i| \leq |\mathbf{A}_{ii}||u_i|$$

implying that $|bA_{ii} - \lambda| \leq |\mathbf{A}_{ii}|$, which is a contradiction to the negativity of λ and the nonnegativity of $\mathbf{A}_{ii}$.

b) since by part (a) we know that $\mathbf{A}$ is positive semidefinite, all we need to show is that A has no zero eigenvalues. Suppose in contradiction that there is a zero eigenvalue, meaning that there is a vector $u \neq 0$ such that $\mathbf{A}u = 0$. Then, similarly to the proof of part (a), let $i \in \{1, 2, \ldots, n\}$ be an index for which $|u_i|$ is maximal among $|u_1|, |u_2|, \ldots, |u_n|$ and we obtain

$$|\mathbf{A}_{ii}| \cdot |u_i| = \left| \sum_{j \neq i} \mathbf{A}_{ij} u_j \right| \leq \left(\sum_{j \neq i} \left| \mathbf{A}_{ij} \right| \right) |u_i| < |\mathbf{A}_{ii}||u_i|$$

which is obviously impossible, establishing the fact that $\mathbf{A}$ is positive definite.

□

Appendix: Gradient and Hessian of Quadratics

Gradient of Linear Function

Consider a linear function of the form

$$f(w) = a^\top w$$

where a and w are length-d vectors. We can derive the gradient in matrix notation as follows:

1. Convert to summation notation:

$$f(w) = \sum_{j=1}^{d} a_j w_j$$

   where a_j is element j of a and w_j is element j of w

2. Take the partial derivative with respect to a generic element k :

$$\frac{\partial}{\partial w_k}\left[\sum_{j=1}^{d} a_j w_j\right] = a_k$$

3. Assemble the partial derivatives into a vector:

$$\nabla f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_d \end{bmatrix} = a$$

So our final result is that

$$\nabla f(w) = a$$

This generalizes the scalar case where $\frac{d}{dw}[\alpha w] = \alpha$. We can also consider general linear functions of the form

$$f(w) = a^\top w + \beta$$

for a scalar β. But in this case we still have $\nabla f(w) = a$ since β does not depend on w.

Gradient of Quadratic Function

Consider a quadratic function of the form

$$f(w) = w^\top A w$$

where w is a length- d vector and A is a d by d matrix. We can derive the gradient in matrix notation as follows

1. Convert to summation notation:

$$f(w) = w^\top \begin{bmatrix} \sum_{j=1}^n a_{1j} w_j \\ \sum_{j=1}^n a_{2j} w_j \\ \vdots \\ \sum_{j=1}^n a_{dj} w_j \end{bmatrix} = \sum_{i=1}^d \sum_{j=1}^d w_i a_{ij} w_j$$

   where a_{ij} is the element in row i and column j of A. To help with computing the partial derivatives, it helps to re-write it in the form

$$f(w) = \sum_{i=1}^d \sum_{j=1}^d w_i a_{ij} w_j = \sum_{i=1}^d \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right)$$

2. Take the partial derivative with respect to a generic element k :

$$\frac{\partial}{\partial w_k} \left[\sum_{i=1}^d \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right) \cdot \right] = 2 a_{kk} w_k + \sum_{j \neq k} w_j a_{jk} + \sum_{j \neq k} a_{kj} w_j$$

   The first term comes from the a_{kk} term that is quadratic in w_k, while the two sums come from the terms that are linear in w_k. We can move one $a_{kk} w_k$ into each of the sums to simplify this to

$$\frac{\partial}{\partial w_k} \left[\sum_{i=1}^d \left(a_{ii} w_i^2 + \sum_{j \neq i} w_i a_{ij} w_j \right) \cdot \right] = \sum_{j=1}^d w_j a_{jk} + \sum_{j=1}^d a_{kj} w_j$$

3. Assemble the partial derivatives into a vector:

$$\nabla f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1} \\ \frac{\partial}{\partial w_2} \\ \vdots \\ \frac{\partial}{\partial w_d} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} + \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d w_j a_{j2} + \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} + \sum_{j=1}^d a_{dj} w_j \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} \\ \sum_{j=1}^d w_j a_{j2} \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d a_{dj} w_j \end{bmatrix}$$

4. Convert to matrix notation:

$$\nabla f(w) = \begin{bmatrix} \sum_{j=1}^d w_j a_{j1} \\ \sum_{j=1}^d w_j a_{j2} \\ \vdots \\ \sum_{j=1}^d w_j a_{jd} \end{bmatrix} + \begin{bmatrix} \sum_{j=1}^d a_{1j} w_j \\ \sum_{j=1}^d a_{2j} w_j \\ \vdots \\ \sum_{j=1}^d a_{dj} w_j \end{bmatrix} = A^\top w + A w = \left(A^\top + A \right) w$$

So our final result is that

$$\nabla f(w) = \left(A^\top + A \right) w$$

Note that if A is symmetric ($A^\top = A$) then we have $(A^\top + A) = (A + A) = 2A$ so we have

$$\nabla f(w) = 2Aw$$

This generalizes the scalar case where $\frac{d}{dw}\left[\alpha w^2\right] = 2\alpha w$. We can also consider general quadratic functions of the form

$$f(w) = \frac{1}{2} w^\top A w + b^\top w + \gamma$$

Using the above results we have

$$\nabla f(w) = \frac{1}{2}\left(A^\top + A\right) w + b$$

and if A is symmetric then

$$\nabla f(w) = Aw + b$$

Hessian of a Quadratic Form

4 Hessian of Quadratic Function For a quadratic function of the form,

$$f(w) = w^\top A w$$

we show above the partial derivatives are given by linear functions,

$$\frac{\partial f}{\partial w_k} = \sum_{j=1}^{d} w_j a_{jk} + \sum_{j=1}^{d} a_{kj} w_j$$

The second partial derivatives are thus constant functions of the form

$$\frac{\partial^2 f}{\partial w_k \partial w_{k'}} = a_{k'k} + a_{kk'}$$

which means that the Hessian matrix has a simple form

$$\nabla^2 f(w) = \begin{bmatrix} \frac{\partial}{\partial w_1 \partial w_1} f(w) & \frac{\partial}{\partial w_1 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_1 \partial w_d} f(w) \\ \frac{\partial w_2 \partial w_1}{\partial w_2 \theta w_2} f(w) & \frac{\partial}{\partial w_2 \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_2 \partial w_d} f(w) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_d \partial w_1} f(w) & \frac{\partial}{\partial w_d \partial w_2} f(w) & \cdots & \frac{\partial}{\partial w_d \partial w_d} f(w) \end{bmatrix} = \begin{bmatrix} a_{11} + a_{11} & a_{21} + a_{12} & \cdots & a_{d1} + a_{1d} \\ a_{12} + a_{21} & a_{22} + a_{22} & \cdots & a_{d2} + a_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1d} + a_{d1} & a_{2d} + a_{d2} & \cdots & a_{dd} + a_{dd} \end{bmatrix}$$

This gives a result of

$$\nabla^2 f(w) = A + A^\top$$

and if A is symmetric this simplifies to

$$\nabla^2 f(w) = 2A$$

Part III.
Linear and Nonlinear Least Squares Problems

A Bit of History

In January 1, 1801, the Italian monk Giuseppe Piazzi, discovered a faint, nomadic object through his telescope in Palermo, correctly believing it to reside in the orbital region between Mars and Jupiter. Piazzi watched the object for 41 days but then fell ill, and shortly thereafter the wandering star strayed into the halo of the Sun and was lost to observation. The newly-discovered planet had been lost, and astronomers had a mere 41 days of observation covering a tiny arc of the night from which to attempt to compute an orbit and find the planet again.

Figure 8: Giusepe Piazzi (1746-1826), Italian priest, mathematician, and astronomer. His observations led to the discovery of the dwarf planet Ceres.

The dean of the French astrophysical establishment, Pierre-Simon Laplace, declared that the orbit recovery simply could not be done. In Germany, the 24 years old German mathematician Car Friedrich Gauss had considered that this type of problem, to determine a planet's orbit from a limited handful of observations- "commended itself to mathematicians by its difficulty and elegance." Gauss discovered a method for computing the planet's orbit using only three of the original observations and successfully predicted where Ceres might be found (now considered to be a dwarf planet). The prediction catapulted him to worldwide acclaim.

More than 200 years later, in 2019 American computer scientist Katie Bouman used similar mathematical methods, which heavily rely on optimization and large-scale astronomical observation datasets, to recover the first image of a black hole.

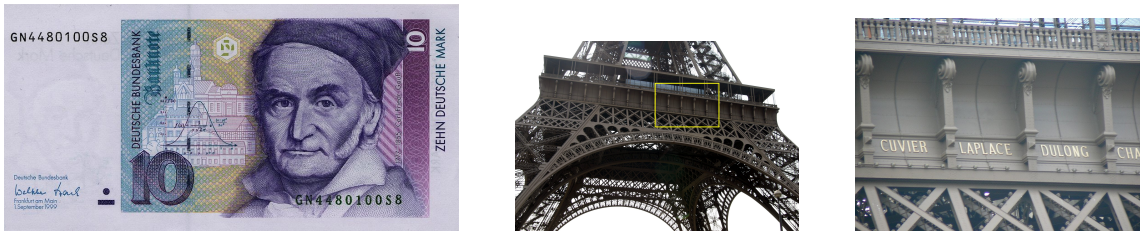


Figure 9: Carl Friedrich Gauss (1777-1855) and Pierre-Simon Laplace (1749-1827) need no introduction. On the left, a 10 Deutsche Mark bank note with Gauss. Laplace has his name engrave in the Eiffel Tower. How cool is that!

Figure 10: 200 years later, Katie Bouman (1990-) used similar optimization techniques as Gauss (and a bit of supercomputing power) to recover the first image of a black hole. After this module, you could be next!

Formulating the Linear Least Squares Problem

Consider the linear system

$$\mathbf{S}\mathbf{x} \approx \mathbf{b}, \quad \left(\mathbf{S} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^{m}\right)$$

We assume that $\mathbf{S}$ has a full column rank, that is, $rank(\mathbf{S}) = n$. However, when $m > n$, that is when you have more equations than unknowns, the system is usually inconsistent and a common approach for finding an approximate solution is to pick the solution of the problem

$$\min_{\mathbf{x}} \|\mathbf{S}\mathbf{x} - \mathbf{b}\|^2, \tag{LLS}$$

which is the same as (check!)

$$\min_{\mathbf{x} \in \mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \mathbf{x}^\top \mathbf{S}^\top \mathbf{S}\mathbf{x} - 2\mathbf{b}^\top \mathbf{S}\mathbf{x} + \|\mathbf{b}\|^2 \right\}$$

Note that $\nabla^2 f(\mathbf{x}) = 2\mathbf{S}^\top\mathbf{S} \succ 0$ since $rank(\mathbf{S}) = n$ and $m > n$. Therefore, the unique optimal solution $\mathbf{x}_{LS}$ is the solution $\nabla f(\mathbf{x}) = 0$, namely,

$$\left(\mathbf{S}^\top\mathbf{S}\right) \mathbf{x}_{\mathrm{LS}} = \mathbf{S}^\top \mathbf{b} \leftarrow \text{ Normal Equations},$$

leading to

$$\mathbf{x}_{\mathrm{LS}} = \left(\mathrm{S}^\top \mathrm{S}\right)^{-1} \mathrm{S}^\top \mathbf{b}\,.$$

A Numerical Example. Consider the inconsistent linear system

$$\begin{aligned} x_1 + 2x_2 &= 0 \\ 2x_1 + x_2 &= 1 \\ 3x_1 + 2x_2 &= 1 \end{aligned}$$

To find the least squares solution, we will solve the normal equations:

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}^\top \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 2 \end{pmatrix}^\top \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

which is the same as

$$\begin{pmatrix} 14 & 10 \\ 10 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 3 \end{pmatrix} \Rightarrow \mathbf{x}_{LS} = \begin{pmatrix} 15/26 \\ -8/26 \end{pmatrix}.$$

Note that $\mathbf{A}\mathbf{x}_{LS} = (-0.038; 0.846; 1.115)$, so that the errors are

$$\mathbf{b} - \mathrm{S}\mathbf{x}_{LS} = \begin{pmatrix} 0.038 \\ 0.154 \\ -0.115 \end{pmatrix} \Rightarrow \text{ sq. error } = 0.038^2 + 0.154^2 + (-0.115)^2 = 0.038\,.$$

What is the core numerical task for solving LLS?

Data Fitting

We revisit, from an optimization viewpoint, a fundamental problem in statistics, namely, linear regression. Assume we have a dataset $(s_i, b_i)\,, i = 1, 2, \ldots, m$, where $s_i \in \mathbb{R}^n$ and $b_i \in \mathbb{R}$. Assume that an approximate linear relation holds:

$$b_i \approx \mathbf{s}_i^\top \mathbf{x}, \quad i = 1, 2, \ldots, m\,.$$

The corresponding least squares/regression problem reads

$$\min_{\mathbf{x}\in\mathbb{R}^n} \sum_{i=1}^{m} \left(\mathbf{s}_i^\top \mathbf{x} - b_i\right)^2,$$

or equivalently

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathrm{S}\mathbf{x} - \mathbf{b}\|^2\,.$$

where

$$S = \begin{pmatrix} s_1^\top \\ s_2^\top \\ \vdots \\ s_m^\top \end{pmatrix}, \; t = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

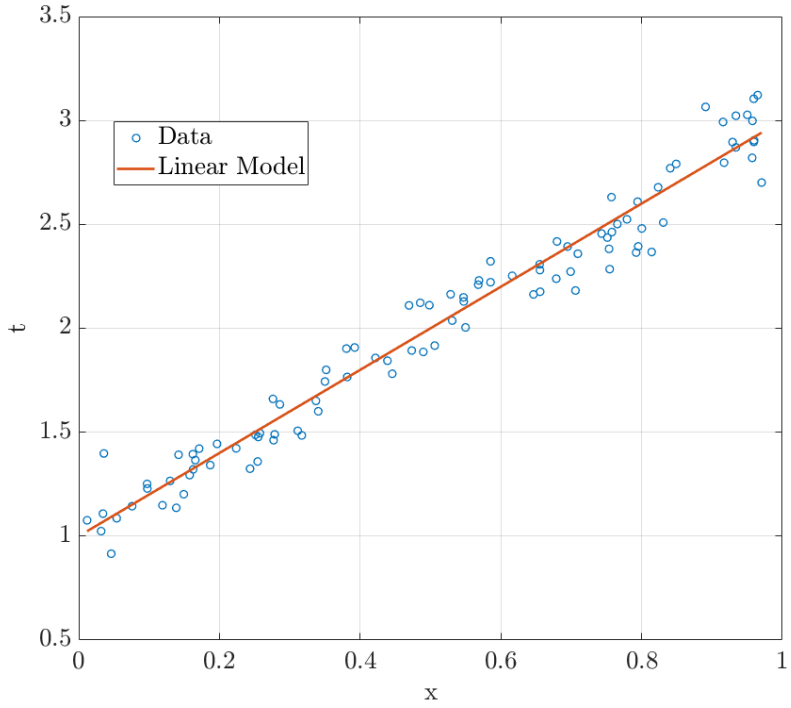


Figure 11: The typical situation in linear regression. A set of scattered data (s_i, t_i) (in blue) suggests a linear relation between s and t. Among all the possible linear models, there is an optimal choice (in red) which minimizes the square error between the model and the measurements.

Polynomial Fitting

Another problem relevant in statistics that can be cast as a linear least squares problems is the polynomial fit of data. Given a set of points in $\mathbb{R}^2$: (u_i, y_i) , $i = 1, 2, \ldots, m$, for which the following approximate relation holds for some $a_0, \ldots, a_d$:

$$\sum_{j=0}^{d} a_j u_i^j \approx y_i, \quad i = 1, \ldots, m$$

The linear system associated to the relation reads:

$$\underbrace{\begin{pmatrix} 1 & u_1 & u_1^2 & \cdots & u_1^d \\ 1 & u_2 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & u_m & u_m^2 & \cdots & u_m^d \end{pmatrix}}_{\mathrm{U}} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_d \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_m \end{pmatrix}$$

The least squares solution is of course well defined if the $m \times (d + 1)$ matrix is of full column rank. This is true when all the u_i 's are different from each other (why?).

Regularized Least Squares

There are several situations in which the least squares solution does not give rise to a good estimate of the "true" vector $\mathbf{x}$. In these cases, a regularized problem (called regularized least squares (RLS)) is often solved:

$$(\mathbf{RLS}) \qquad \min_{\mathbf{x}} \|\mathbf{S}\mathbf{x} - \mathbf{b}\|^2 + \lambda R(\mathbf{x}) \,.$$

Here, λ is the regularization parameter and $R(\cdot)$ is the regularization function (also called a penalty function). A common choice is a quadratic regularization function:

$$\min \|\mathbf{S}\mathbf{x} - \mathbf{b}\|^2 + \lambda \|\mathbf{D}\mathbf{x}\|^2 \,.$$

The optimal solution of the above problem is (why?)

$$\mathbf{x}_{\mathbf{RLS}} = \left(\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{D}^\top \mathbf{D}\right)^{-1} \mathbf{S}^\top \mathbf{b} \,.$$

What kind of assumptions are needed to assure that $\mathbf{S}^\top \mathbf{S} + \lambda \mathbf{D}^\top \mathbf{D}$ is invertible? (answer: $\text{Null}(\mathbf{S}) \cap \text{Null}(\mathbf{D}) = \{0\}$)[2].

Denoising

A very important application of linear least squares and regularization techniques is the denoising of signals (acoustic, images). Suppose that a noisy measurement of a signal $\mathbf{x} \in \mathbb{R}^n$ is given:

$$\mathbf{b} = \mathbf{x} + \mathbf{w} \,,$$

where $\mathbf{x}$ is the "true" unknown signal, $\mathbf{w}$ is the unknown noise and $\mathbf{b}$ is the (known) measures vector. Note that the least squares problem:

$$\min_{\mathbf{x}} \|\mathbf{x} - \mathbf{b}\|^2 \,,$$

is meaningless, as we would trivially recover $\mathbf{x} = \mathbf{b}$, without any denoising. We need to enrich the optimization problem by adding a suitable regularization term, exploiting some a priori information of the signal. For example, if the we know that the signal is "smooth" in some sense, then $R(\cdot)$ can be chosen as a penalization of the signal "sudden variations"

$$R(\mathbf{x}) = \sum_{i=1}^{n-1} (x_i - x_{i+1})^2 \,,$$

[2] here $\text{Null}(\mathbf{S})$ refers to the nullspace of the matrix or kernel, $\text{Ker}(\mathbf{S}) := \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{S}\mathbf{x} = \mathbf{0}\}$.

where the regularization $R(\cdot)$ can also be written as $R(\mathbf{x}) = \|\mathbf{L}\mathbf{x}\|^2$ where $\mathbf{L} \in \mathbb{R}^{(n-1)\times n}$ is given by

$$\mathbf{L} = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{pmatrix}.$$

The resulting regularized least squares problem is

$$\min_{\mathbf{x}} \underbrace{\|x - b\|^2}_{\text{fitting}} + \underbrace{\lambda \|Lx\|^2}_{\text{denoisig}}.$$

The direct solution of this problem leads to (why?)

$$\mathbf{x}_{\mathrm{RLS}}(\lambda) = \left(\mathbf{I} + \lambda \mathbf{L}^\top \mathbf{L}\right)^{-1} \mathbf{b}.$$

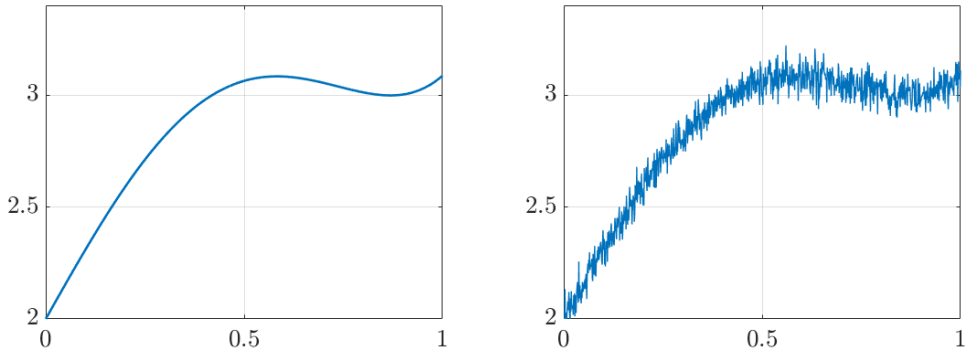


Figure 12: Signal denoising. We only have access to the noisy signal (left), and we would like to recover a "clear" signal (right) by solving a regularized least squares problem.

Nonlinear Least Squares

The least squares problem $\min \|\mathbf{S}\mathbf{x} - \mathbf{b}\|^2$ is often called linear least squares. In some applications we are given a set of nonlinear equations:

$$f_i(\mathbf{x}) \approx b_i, \quad i = 1, 2, \ldots, m.$$

The nonlinear least squares (NLS) problem is the one of finding an $\mathbf{x}$ solving the problem

$$\min_{\mathbf{x}} \sum_{i=1}^{m} \left(f_i(\mathbf{x}) - b_i\right)^2. \tag{NLS}$$

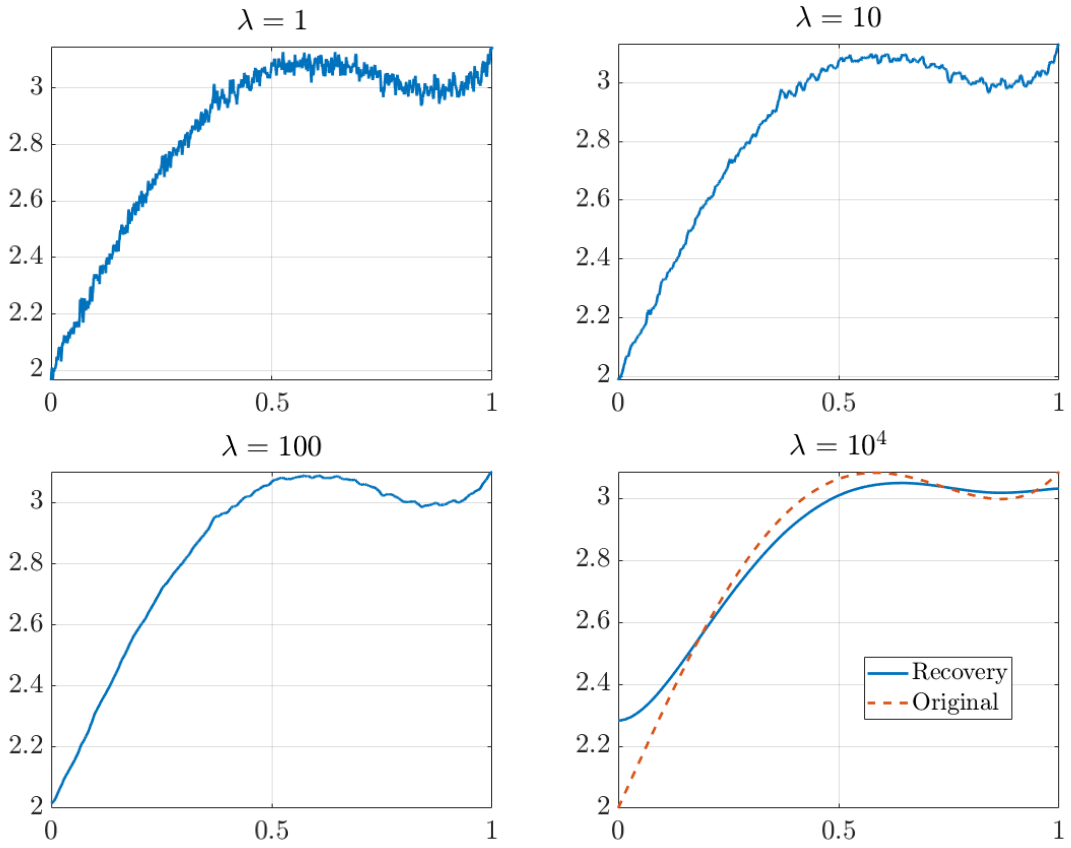


Figure 13: Signal denoising for different regularization parameters. Increasing the regularization parameters leads to better denoising, however, as the parameter becomes too large, the fit with the original signal is lost. What is the limit as λ grows?

As opposed to linear least squares, there is no easy way to to solve NLS problems. However, there are some dedicated algorithms for this problem, which we will explore later on in this module.

A Case Study: Circle Fitting

Given m points $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \in \mathbb{R}^n$, the circle fitting problem seeks to find a circle

$$C(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\| = r\}$$

that best fits the m points.

We formulate a set of equations to match the center of the circle and its radius:

$$\|\mathbf{x} - \mathbf{a}_i\| \approx r, \quad i = 1, 2, \ldots, m\,.$$

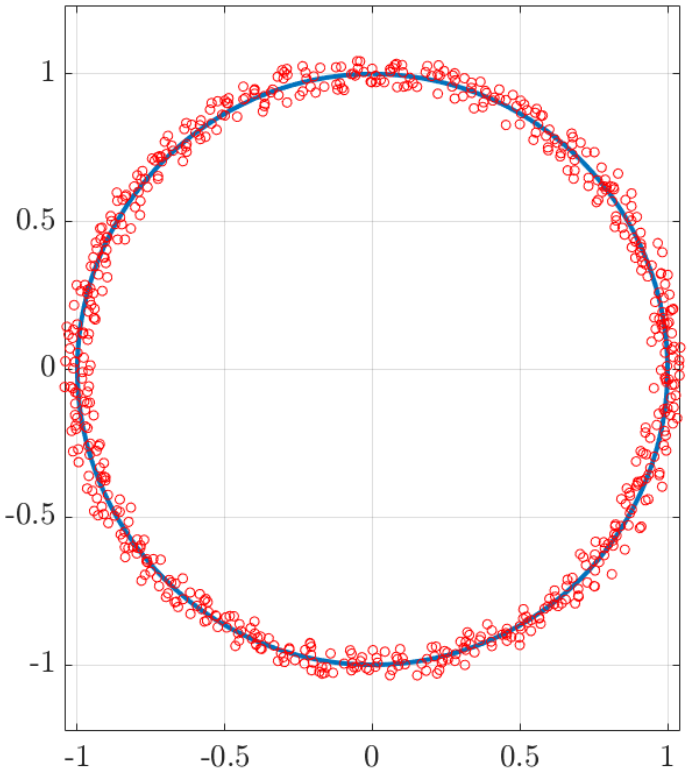


Figure 14: In the circle fitting problem, we look for a circle to match a set of measurements.

To avoid nondifferentiability, consider the squared version:

$$\|\mathbf{x} - \mathbf{a}_i\|^2 \approx r^2, \quad i = 1, 2, \ldots, m\,.$$

This leads to a Nonlinear Least Squares formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^n,\, r\in\mathbb{R}_+} \sum_{i=1}^{m} \left(\|\mathbf{x} - \mathbf{a}_i\|^2 - r^2\right)^2 ,$$

or expanding

$$\min_{\mathbf{x},r} \left\{ \sum_{i=1}^{m} \left(-2\mathbf{a}_i^\top \mathbf{x} + \|\mathbf{x}\|^2 - r^2 + \|\mathbf{a}_i\|^2\right)^2 : \mathbf{x} \in \mathbb{R}^n, r \in \mathbb{R} \right\}$$

We will reduce this problem to a Linear Least Squares. Making the change of variables $R = \|x\|^2 - r^2$, the above problem reduces to

$$\min_{\mathbf{x}\in\mathbb{R}^n, R\in\mathbb{R}} \left\{ f(\mathbf{x}, R) \equiv \sum_{i=1}^{m} \left(-2\mathbf{a}_i^\top \mathbf{x} + R + \|\mathbf{a}_i\|^2\right)^2 : \|\mathbf{x}\|^2 \geq R \right\}$$

The constraint $\|\mathbf{x}\|^2 \geq R$ can be dropped, and therefore the problem is equivalent to the LS problem

$$\min_{\mathbf{x},R} \left\{ \sum_{i=1}^{m} \left(-2\mathbf{a}_i^\top \mathbf{x} + R + \|\mathbf{a}_i\|^2 \right)^2 : \mathbf{x} \in \mathbb{R}^n, R \in \mathbb{R} \right\} . \tag{CF-LS}$$

Redundancy of the Constraint $\|\mathbf{x}\|^2 \geq R$. We will show that any optimal solution $(\hat{\mathbf{x}}, \hat{R})$ of eq. (CF-LS) automatically satisfies $\|\hat{\mathbf{x}}\|^2 \geq \hat{R}$. Otherwise, if $\|\hat{\mathbf{x}}\|^2 < \hat{R}$, then

$$-2\mathbf{a}_i^\top \hat{\mathbf{x}} + \hat{R} + \|\mathbf{a}_i\|^2 > -2\mathbf{a}_i^\top \hat{\mathbf{x}} + \|\hat{\mathbf{x}}\|^2 + \|\mathbf{a}_i\|^2 = \|\hat{\mathbf{x}} - \mathbf{a}_i\|^2 \geq 0, \ \ i = 1, \ldots, m ,$$

thus contradicting the optimality of $(\hat{\mathbf{x}}, \hat{R})$.

Part IV.
The Gradient Descent Algorithm

Descent Directions Methods

Recall that our objective is to find an optimal solution of the problem

$$\min \{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\} .$$

We have seen that in some particular cases, such as the Linear Least Squares problem from Week 3, it is possible to obtain a direct solution by solving the normal equations. Unfortunately, this is not the case for an arbitrary nonlinear objective $f(\mathbf{x})$. We will study the solution of this optimization problem by using iterative algorithms of the form

$$\mathbf{x}^{k+1} = \mathbf{x}^k + t^k \mathbf{d}^k , \ k = 0, 1, \ldots ,$$

where $\mathbf{d}^k \in \mathbb{R}^n$ is a **direction** and $t^k \in \mathbb{R}$ is called the **stepsize**. We will see that a careful selection of $\mathbf{d}^k$ and t^k will generate a sequence $\{\mathbf{x}^k\}_{k=0}^{\infty}$ converging to a stationary point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$ (a very good candidate for solving our original problem). A first important concept is the one of descent direction.

Definition (Descent Direction). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function over $\mathbb{R}^n$. A vector $\mathbf{0} \neq \mathbf{d} \in \mathbb{R}^n$ is called a descent direction of f at $\mathbf{x}$ if the directional derivative $f'(\mathbf{x}; \mathbf{d})$ is negative, meaning that*

$$f'(\mathbf{x}; \mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d} < 0 .$$

The Descent Property of Descent Directions

Lemma. *Let f be a continuously differentiable function over $\mathbb{R}^n$, and let $\mathbf{x} \in \mathbb{R}^n$. Suppose that $\mathbf{d}$ is a descent direction of f at $\mathbf{x}$. Then there exists $\varepsilon > 0$ such that*

$$f(\mathbf{x} + t\mathbf{d}) < f(\mathbf{x})$$

for any $t \in (0, \varepsilon]$.

Proof. Since $f'(\mathbf{x}; \mathbf{d}) < 0$, it follows from the definition of the directional derivative that

$$\lim_{t \to 0^+} \frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} = f'(\mathbf{x}; \mathbf{d}) < 0 .$$

Therefore, there exists $\varepsilon > 0$ such that

$$\frac{f(\mathbf{x} + t\mathbf{d}) - f(\mathbf{x})}{t} < 0$$

for any $t \in (0, \varepsilon]$, which readily implies the desired result. □

Figure 15: Augustin-Louis Cauchy (1789-1857), French mathematician, physicist and engineer. He made pioneering contributions to several branches of mathematics. In his 1847 paper " Méthode générale pour la résolution desde systèmes simultanées" he presented the first formulation of the gradient descent method.

A first version of the descent direction algorithm is presented below.

Algorithm 1: Schematic Descent Direction Method

Initialization: pick $\mathbf{x}^0 \in \mathbb{R}^n$ arbitrarily.

General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Pick a descent direction $\mathbf{d}^k$.

2 Find a stepsize t^k satisfying $f\left(\mathbf{x}^k + t^k\mathbf{d}^k\right) < f\left(\mathbf{x}^k\right)$.

3 Set $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\mathbf{d}^k$.

4 If a stopping criteria is satisfied, then STOP and $\mathbf{x}^{k+1}$ is the output.

Of course, many details are missing in the above schematic algorithm:

- What is the starting point?
- How to choose the descent direction?
- What stepsize should be taken?
- What is the stopping criteria?

Stepsize Selection Rules

Let's assume for one second that a descent direction has been found. How do we choose the stepsize? There are several options available.

- Constant stepsize: $t^k = \bar{t}$ for any k.

- Exact stepsize: t^k is a minimizer[3] of f along the ray $\mathbf{x}^k + t\mathbf{d}^k$:

$$t^k \in \operatorname*{argmin}_{t \geq 0} f(\mathbf{x}^k + t\mathbf{d}^k)$$

- Backtracking (or Armijo rule): the method requires three parameters: $s > 0, \alpha \in (0,1)$, and $\beta \in (0,1)$. Here we start with an initial stepsize $t^k = s$. While

$$f(\mathbf{x}^k) - f(\mathbf{x}^k + t^k\mathbf{d}^k) < -\alpha t^k \nabla f(\mathbf{x}^k)^\top \mathbf{d}^k$$

  set $t^k := \beta t^k$, iterating until achieving the **Sufficient Decrease Property**

$$f(\mathbf{x}^k) - f(\mathbf{x}^k + t^k\mathbf{d}^k) \geq -\alpha t^k \nabla f(\mathbf{x}^k)^\top \mathbf{d}^k \,.$$

Finding the right t^k is referred in the literature as line search.

Exercise: exact line search for quadratic functions. Find the exact stepsize when $f(\mathbf{x})$ is a quadratic function $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + \mathbf{c}$ where $\mathbf{A}$ is an $n \times n$ positive definite matrix, $\mathbf{b} \in \mathbb{R}^n$ and $\mathbf{c} \in \mathbb{R}$. Let $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{d} \in \mathbb{R}^n$ be a descent direction of f at $\mathbf{x}$. The objective is to find a solution to

$$\min_{t \geq 0} f(\mathbf{x} + t\mathbf{d}) \,.$$

Taking the Direction of Minus the Gradient

In the gradient method we make the choice $\mathbf{d}^k = -\nabla f\left(\mathbf{x}^k\right)$. This is a descent direction as long as $\nabla f\left(\mathbf{x}^k\right) \neq 0$ since

$$f'(\mathbf{x}^k; -\nabla f(\mathbf{x}^k)) = -\nabla f(\mathbf{x}^k)^\top \nabla f(\mathbf{x}^k) = -\|\nabla f(\mathbf{x}^k)\|^2 < 0$$

In addition to being a descent direction, minus the gradient is also the steepest descent direction.

Lemma. *Let f be a continuously differentiable function and let $\mathbf{x} \in \mathbb{R}^n$ be a non-stationary point ($\nabla f(\mathbf{x}) \neq 0$). Then an optimal solution of*

$$\min_{\mathbf{d}} \{f'(\mathbf{x}; \mathbf{d}) : \|\mathbf{d}\| = 1\}$$

is $\mathbf{d} = -\nabla f(\mathbf{x}) / \|\nabla f(\mathbf{x})\|$.

[3] We call the operation of selecting the element that minimizes instead of just computing the minimum value the argmin.

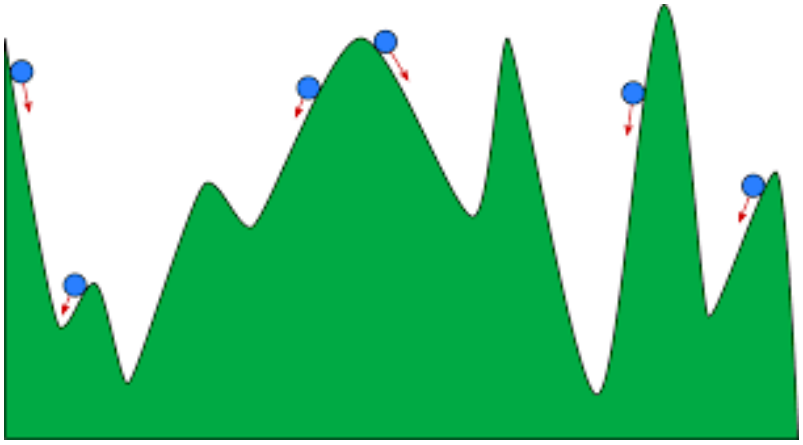


Figure 16: The landscape is $f(\mathbf{x})$ and the balls are ready to move in the direction of $-\nabla f(\mathbf{x})$. What can you say about the points where the balls are expected to end?

Proof. Since $f'(\mathbf{x};\mathbf{d}) = \nabla f(\mathbf{x})^\top \mathbf{d}$, we write the problem as

$$\min_{\mathbf{d}\in\mathbb{R}^n} \left\{\nabla f(\mathbf{x})^\top \mathbf{d} : \|\mathbf{d}\| = 1\right\}$$

By the Cauchy-Schwarz inequality we have

$$\nabla f(\mathbf{x})^\top \mathbf{d} \geq -\|\nabla f(\mathbf{x})\| \cdot \|\mathbf{d}\| = -\|\nabla f(\mathbf{x})\|\,.$$

Thus, $-\|\nabla f(\mathbf{x})\|$ is a lower bound on the optimal value of the directional derivative. On the other hand, using the direction $\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$ we obtain that

$$f'\left(\mathbf{x}, -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\right) = -\nabla f(\mathbf{x})^\top \left(\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}\right) = -\|\nabla f(\mathbf{x})\|$$

and we thus come to the conclusion that the lower bound $-\|\nabla f(\mathbf{x})\|$ is attained at $\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$, which readily implies that this is an optimal solution for the descent direction. □

Algorithm 2: The Gradient Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in \mathbb{R}^n$.

General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Pick a stepsize t^k by a line search procedure on the function

$$g(t) = f(\mathbf{x}^k - t\nabla f(\mathbf{x}^k))\,.$$

2 Set $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla f\left(\mathbf{x}^k\right)$.

3 If $\left\|\nabla f\left(\mathbf{x}^{k+1}\right)\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

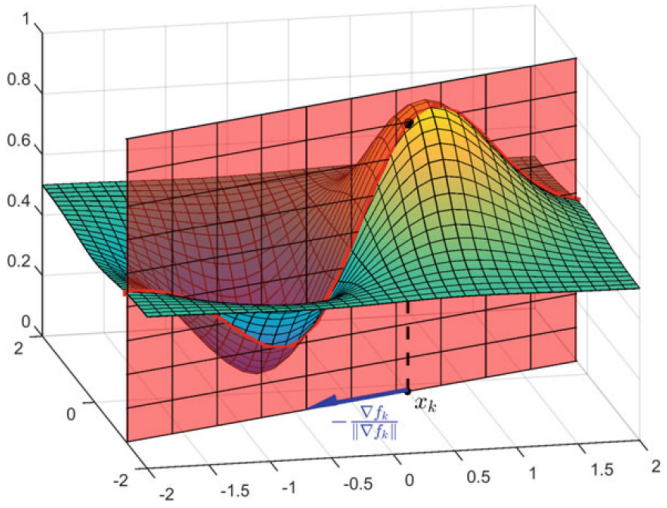


Figure 17: At a given iteration, the current point $\mathbf{x}^k$ and the descent direction $-\frac{\nabla f(\mathbf{x}^k)}{\|\nabla f(\mathbf{x}^k)\|}$ define a plane along which the next iterate $\mathbf{x}^{k+1}$ is sought. The line search procedure defines how far do we move in this direction. A constant stepsize will move a fixed distance along the red line, whereas an exact stepsize will look for the minimizer of the surface constrained to the plane.

Example. Solve

$$\min x_1^2 + 2x_2^2$$

using a gradient descent with exact line search. Use the starting guess $\mathbf{x}^0 = (2, 1)$, and a stopping tolerance of $\varepsilon = 10^{-5}$. The iteration converges in 13 steps, and the convergence history is shown in Figure 18. After you try to code it by yourself, you can check the MATLAB code at the end of this document.

The Zig-Zag Effect

An evident behavior of the gradient method as illustrated in Figure 18 is the "zig-zag" effect, meaning that the direction found at the k-iterarion $\mathbf{x}^{k+1} - \mathbf{x}^k$ is orthogonal to the direction found at the $(k+1)$-th iteration $\mathbf{x}^{k+2} - \mathbf{x}^{k+1}$. We now establish this result.

Lemma. *Let $\{\mathbf{x}^k\}_{k>0}$ be the sequence generated by the gradient method with exact line search for solving a problem of minimizing a continuously differentiable function f. Then for any $k = 0, 1, 2, \ldots$*

$$(\mathbf{x}^{k+2} - \mathbf{x}^{k+1})^\top (\mathbf{x}^{k+1} - \mathbf{x}^k) = 0\,.$$

Proof. First, we write $\mathbf{x}^{k+1} - \mathbf{x}^k = -t^k \nabla f\left(\mathbf{x}^k\right)$, and $\mathbf{x}^{k+2} - \mathbf{x}^{k+1} = -t^{k+1} \nabla f\left(\mathbf{x}^{k+1}\right)$. Therefore, we need to prove that $\nabla f\left(\mathbf{x}^k\right)^\top \nabla f\left(\mathbf{x}^{k+1}\right) = 0$. The exact line search is by definition

$$t^k \in \operatorname*{argmin}_{t \geq 0} \left\{ g(t) \equiv f(\mathbf{x}^k - t \nabla f(\mathbf{x}^k)) \right\}$$

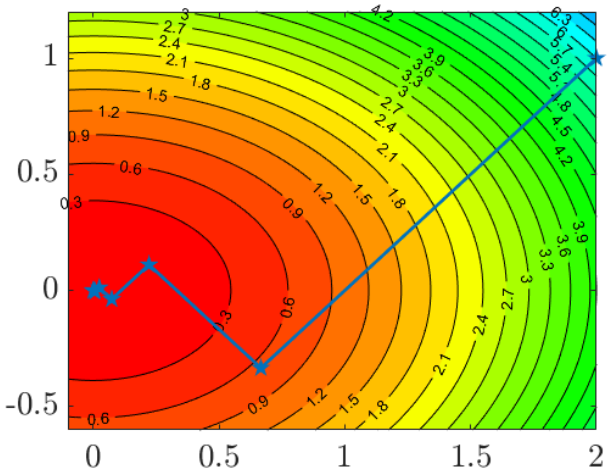


Figure 18: Gradient descent with exact line search for $f(\mathbf{x}) = x_1^2 + 2x_2^2$. After 13 iterations, the method converges to the optimum $(0, 0)$.

Hence,

$$\begin{aligned} g'(t^k) &= 0\,, \\ -\nabla f(\mathbf{x}^k)^\top \nabla f(\mathbf{x}^k - t^k \nabla f(\mathbf{x}^k)) &= 0\,, \\ -\nabla f(\mathbf{x}^k)^\top \nabla f(\mathbf{x}^{k+1}) &= 0\,. \end{aligned}$$

□

Convergence of the Gradient Method

We begin this discussion with a computational example.

Example. Solve

$$\min x_1^2 + 2x_2^2$$

using gradient descent with constant stepsize. Set $\mathbf{x}^0 = (2, 1)$, $\varepsilon = 10^{-5}$, and $\bar{t} = 0.1$. The MATLAB code for this method can be found the end of the document. The iteration sequence reads

iter_number = 1	norm_grad = 4.000000	fun_val = 3.280000
iter_number = 2	norm_grad = 2.937210	fun_val = 1.897600
⋮	⋮	⋮
iter_number = 3	norm_grad = 2.222791	fun_val = 1.141888
iter_number = 56	norm_grad = 0.000015	fun_val = 0.000000
iter_number = 57	norm_grad = 0.000012	fun_val = 0.000000
iter_number = 58	norm_grad = 0.000010	fun_val = 0.000000

achieving convergence after 58 iterations. If we increase the stepsize parameter to $\bar{t} = 10$, we observe the history below

iter_number = 1	norm_grad = 1783.488716	fun_val = 476806.000000
iter_number = 2	norm_grad = 656209.693339	fun_val = 56962873606.00
iter_number = 3	norm_grad = 256032703.004797	fun_val = 8318300807
⋮	⋮	⋮
iter_number = 119	norm_grad = NaN	fun_val = NaN

In this case, the sequence diverges. This leads us to a very important question: how can we choose the constant stepsize so that convergence is guaranteed?

Lipschitz Continuity of the Gradient

Definition (Lipschitz Gradient)**.** *Let f be a continuously differentiable function over $\mathbb{R}^n$. We say that f has a Lipschitz gradient if there exists $L \geq 0$ for which*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|,$$

*for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. L is called the **Lipschitz constant**.*

Some relevant remarks:

- If ∇f is Lipschitz with constant L, then it is also Lipschitz with constant $\tilde{L}$ for all $\tilde{L} \geq L$.
- The class of functions with Lipschitz gradient with constant L is denoted by $C_L^{1,1}(\mathbb{R}^n)$ or just $C_L^{1,1}$. When the constant is not relevant, we simply denote the class by $C^{1,1}$.
- **Linear functions** - Given a $\in \mathbb{R}^n$, the function $f(\mathbf{x}) = a^\top x$ is in $C_0^{1,1}$.
- **Quadratic functions** - Let $\mathbf{A}$ be a symmetric $n \times n$ matrix, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then the function $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$ is a $C^{1,1}$ function. The smallest Lipschitz constant of ∇f is $2\|\mathbf{A}\|_2$ (why?).

Theorem (Equivalence to Boundedness of the Hessian)**.** *Let f be a twice continuously differentiable function over $\mathbb{R}^n$. Then the following two claims are equivalent:*

1. $f \in C_L^{1,1}(\mathbb{R}^n)$.
2. $\left\|\nabla^2 f(\mathbf{x})\right\| \leq L$ *for any* $\mathbf{x} \in \mathbb{R}^n$.

Proof. (2) $\Rightarrow$ (1). Suppose that $\left\|\nabla^2 f(\mathbf{x})\right\| \leq L$ for any $\mathbf{x} \in \mathbb{R}^n$. Then by the fundamental theorem of calculus we have for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$\begin{aligned}\nabla f(\mathbf{y}) &= \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))(\mathbf{y} - \mathbf{x}) dt \\ &= \nabla f(\mathbf{x}) + \left(\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) dt\right) \cdot (\mathbf{y} - \mathbf{x})\end{aligned}.$$

Thus,

$$\begin{aligned}\|\nabla f(\mathbf{y}) - \nabla f(\mathbf{x})\| &= \left\|\left(\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) dt\right) \cdot (\mathbf{y} - \mathbf{x})\right\| \\ &\leq \left\|\int_0^1 \nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) dt\right\| \|\mathbf{y} - \mathbf{x}\| \\ &\leq \left(\int_0^1 \left\|\nabla^2 f(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))\right\| dt\right) \|\mathbf{y} - \mathbf{x}\| \\ &\leq L\|\mathbf{y} - \mathbf{x}\|\end{aligned}$$

establishing the desired result $f \in C_L^{1,1}$.

Now we prove (1) $\Rightarrow$(2). Suppose now that $f \in C_L^{1,1}$. Then by the fundamental theorem of calculus for any $\mathbf{d} \in \mathbb{R}^n$ and $\alpha > 0$ we have

$$\nabla f(\mathbf{x} + \alpha \mathbf{d}) - \nabla f(\mathbf{x}) = \int_0^a \nabla^2 f(\mathbf{x} + t\mathbf{d}) \mathbf{d} dt$$

Thus,

$$\left\|\left(\int_0^\alpha \nabla^2 f(\mathbf{x} + t\mathbf{d}) dt\right) \mathbf{d}\right\| = \|\nabla f(\mathbf{x} + \alpha \mathbf{d}) - \nabla f(\mathbf{x})\| \leq \alpha L \|\mathbf{d}\|$$

Dividing by α and taking the limit $\alpha \longrightarrow 0^+$, we obtain

$$\left\|\nabla^2 f(\mathbf{x}) \mathbf{d}\right\| \leq L\|\mathbf{d}\|$$

implying that $\left\|\nabla^2 f(\mathbf{x})\right\| \leq L$. □

Example. Show that $f(\mathbf{x}) = \sqrt{1 + x^2} \in C_L^{1,1}$.

Main Convergence Result

We now state an important result connecting functions in $C_L^{1,1}(\mathbb{R}^n)$ and gradient descent.

Lemma (Sufficient decrease of the gradient method). *Let $f \in C_L^{1,1}(\mathbb{R}^n)$. Let $\left\{\mathbf{x}^k\right\}_{k \geq 0}$ be the sequence generated by the gradient method for solving*

$$\min_{\mathbf{x} \in \mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- *constant stepsize* $\bar{t} \in \left(0, \frac{2}{L}\right)$,
- *exact line search,*
- *backtracking procedure with parameters* $s \in \mathbb{R}_{++}$, $\alpha \in (0,1)$, *and* $\beta \in (0,1)$.

Then

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M \left\| \nabla f(\mathbf{x}^k) \right\|^2$$

where

$$M = \begin{cases} \bar{t}\left(1 - \frac{\bar{t}L}{2}\right) & \text{constant stepsize,} \\ \frac{1}{2L} & \text{exact line search,} \\ \alpha \min\left\{s, \frac{2(1-\alpha)\beta}{L}\right\} & \text{backtracking.} \end{cases}$$

We will now show the convergence of the norms of the gradients $\left\|\nabla f\left(\mathbf{x}^k\right)\right\|$ to zero.

Theorem (Convergence of the Gradient Method)**.** *Let* $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ *be the sequence generated by the gradient descent method for solving*

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x})$$

with one of the following stepsize strategies:

- *constant stepsize* $\bar{t} \in \left(0, \frac{2}{L}\right)$,
- *exact line search,*
- *backtracking procedure with parameters* $s > 0$ *and* $\alpha, \beta \in (0,1)$.

Assume that

- $f \in C_L^{1,1}(R^n)$.
- f *is bounded below over* $\mathbb{R}^n$*, that is, there exists* $m \in \mathbb{R}$ *such that* $f(\mathbf{x}) > m$ *for all* $\mathbf{x} \in \mathbb{R}^n$.

Then:

1. *for any* k, $f\left(\mathbf{x}^{k+1}\right) < f\left(\mathbf{x}^k\right)$ *unless* $\nabla f\left(\mathbf{x}^k\right) = 0$.
2. $\nabla f\left(\mathbf{x}^k\right) \to 0$ *as* $k \to \infty$.

Proof. (1) By the sufficient decrease of the gradient method (previous lemma) we have that

$$f(\mathbf{x}^k) - f(\mathbf{x}^{k+1}) \geq M \left\| \nabla f(\mathbf{x}^k) \right\|^2 \geq 0$$

for some constant $M > 0$, and hence the equality $f\left(\mathbf{x}^k\right) = f\left(\mathbf{x}^{k+1}\right)$ can hold only when $\nabla f\left(\mathbf{x}^k\right) = 0$.

(2) Since the sequence $\left\{f\left(\mathbf{x}^k\right)\right\}_{k\geq 0}$ is nonincreasing and bounded below, it converges. Thus, in particular $f\left(\mathbf{x}^k\right) - f\left(\mathbf{x}^{k+1}\right) \to 0$ as $k \to \infty$, which combined with the sufficient decrease of the gradient method implies that $\left\|\nabla f\left(\mathbf{x}^k\right)\right\| \to 0$ as $k \to \infty$. □

Example. Implement our good old example

$$\min x_1^2 + 2x_2^2$$

using gradient descent with backtracking and parameters $\mathbf{x}^0 = (2, 1)$, $s = 2$, $\alpha = 0.25$, $\beta = 0.5$, and $\varepsilon = 10^{-5}$. Using the code at the end of the document you should see that it converges in only two iterations! However, if you try to solve

$$\min 0.01x_1^2 + x_2^2$$

with the same method and parameters, it will take about 200 iterations to reach the optimum $(0, 0)$. Can we detect key properties of the objective function that imply fast/slow convergence? We will see that in the quadratic case, a fundamental characterization is given by the condition number of the associated quadratic form.

The Condition Number

We recall the definition of condition number.

Definition (Condition Number)**.** *Let* $\mathbf{A}$ *be an* $n \times n$ *positive definite matrix. Then the condition number of* $\mathbf{A}$ *is defined by*

$$\kappa(\mathbf{A}) = \frac{\lambda_{\max}(\mathbf{A})}{\lambda_{\min}(\mathbf{A})}$$

where $\lambda_{\max}(\mathbf{A})$ *and* $\lambda_{\min}(\mathbf{A})$ *are the largest and smaller eigenvalues, respectively.*

Matrices (or quadratic functions) with large condition number are called ill-conditioned. Matrices with small condition number are called well-conditioned. Among other things, the condition number gives an idea of the error amplification when working with the matrix $\mathbf{A}$. For an ill-conditioned matrix, small perturbations in the matrix entries can lead to large errors when solving a linear system, or computing $\mathbf{A}^{-1}$. We continue with a technical lemma.

Lemma (Kantorovich Inequality)**.** *Let* $\mathbf{A}$ *be a positive definite* $n \times n$ *matrix. Then for any* $0 \neq \mathbf{x} \in \mathbb{R}^n$ *the inequality*

$$\frac{(\mathbf{x}^\top \mathbf{x})^2}{(\mathbf{x}^\top \mathbf{A}\mathbf{x})\,(\mathbf{x}^\top \mathbf{A}^{-1}\mathbf{x})} \geq \frac{4\lambda_{\max}(\mathbf{A})\lambda_{\min}(\mathbf{A})}{(\lambda_{\max}(\mathbf{A}) + \lambda_{\min}(\mathbf{A}))^2}$$

holds.

We are now in position to state a precise result regarding the minimization of quadratic functions via gradient descent and its rate of convergence based on the condition number of the matrix $\mathbf{A}$.

Theorem (Gradient Method for Minimizing $\mathbf{x}^\top \mathbf{A}\mathbf{x}$). *Let $\{\mathbf{x}^k\}_{k\geq 0}$ be the sequence generated by the gradient method with exact line search for solving the problem*

$$\min_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^\top \mathbf{A}\mathbf{x} \quad (\mathbf{A} \succ 0),$$

Then for any $k = 0, 1, \ldots$

$$f(\mathbf{x}^{k+1}) \leq \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}^k),$$

where $M = \lambda_{\max}(\mathbf{A})$, and $m = \lambda_{\min}(\mathbf{A})$.

Proof. The gradient descent iteration with exact line search for a quadratic function reads $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \mathbf{d}^k$, where $t^k = \frac{(\mathbf{d}^k)^\top \mathbf{d}^k}{2(\mathbf{d}^k)^\top \mathbf{A}_{\mathbf{d}^k}}$, and $\mathbf{d}^k = 2\mathbf{A}\mathbf{x}^k$.

Plugging in the expression for t^k

$$\begin{aligned} f(\mathbf{x}^{k+1}) &= (x^k)^\top \mathbf{A} x^k - \frac{1}{4}\frac{\left((\mathbf{d}^k)^\top \mathbf{d}^k\right)^2}{(\mathbf{d}^k)^\top \mathbf{A}\mathbf{d}^k} \\ &= (\mathbf{x}^k)^\top \mathbf{A}\mathbf{x}^k \left(1 - \frac{1}{4}\frac{\left((\mathbf{d}^k)^\top \mathbf{d}^k\right)^2}{\left((\mathbf{d}^k)^\top \mathbf{A}\mathbf{d}^k\right)\left((\mathbf{x}^k)^\top \mathbf{A}\mathbf{A}^{-1}\mathbf{A}\mathbf{x}^k\right)}\right) \\ &= \left(1 - \frac{\left((\mathbf{d}^k)^\top \mathbf{d}^k\right)^2}{\left((\mathbf{d}^k)^\top \mathbf{A}\mathbf{d}^k\right)\left((\mathbf{d}^k)^\top \mathbf{A}^{-1}\mathbf{d}^k\right)}\right) f(\mathbf{x}^k). \end{aligned}$$

Finally, using Kantorovich's Inequality:

$$f(\mathbf{x}^{k+1}) \leq \left(1 - \frac{4Mm}{(M+m)^2}\right) f(\mathbf{x}^k) = \left(\frac{M-m}{M+m}\right)^2 f(\mathbf{x}^k) = \left(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1}\right)^2 f(\mathbf{x}^k)$$

□

From the expression above, we deduce that large condition number implies a large number of iterations of the gradient method. A small condition number (slightly greater than 1) implies a small number of iterations of the gradient method. For a non-quadratic function, the asymptotic rate of convergence of $\mathbf{x}^k$ to a stationary point $\mathbf{x}^*$ is usually determined by the condition number of $\nabla^2 f(\mathbf{x}^*)$.

Example A severely ill-conditioned function is the so called Rosenbrock function:

$$\min \left\{ f(x_1, x_2) = 100\left(x_2 - x_1^2\right)^2 + (1 - x_1)^2 \right\}$$

The optimal solution to this problem is easily found as $(x_1, x_2) = (1, 1)$, with optimal value 0. The gradient is given by

$$\nabla f(\mathbf{x}) = \begin{pmatrix} -400x_1\left(x_2 - x_1^2\right) - 2(1 - x_1) \\ 200\left(x_2 - x_1^2\right) \end{pmatrix},$$

and the Hessian

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} -400x_2 + 1200x_1^2 + 2 & -400x_1 \\ -400x_1 & 200 \end{pmatrix}.$$

Evaluating the Hessian at the optimum we obtain

$$\nabla^2 f(1,1) = \begin{pmatrix} 802 & -400 \\ -400 & 200 \end{pmatrix},$$

with a high condition number: 2508. Solving the Rosenbrock problem with gradient descent and backtracking stepsize selection and parameters $\mathbf{x}^0 = (2,5), s = 2, \alpha = 0.25, \beta = 0.5, \varepsilon = 10^{-5}$, leads to 6890(!!) iterations. Figure 19 depicts the slow convergence due to poor conditioning of the Hessian around the optimum.

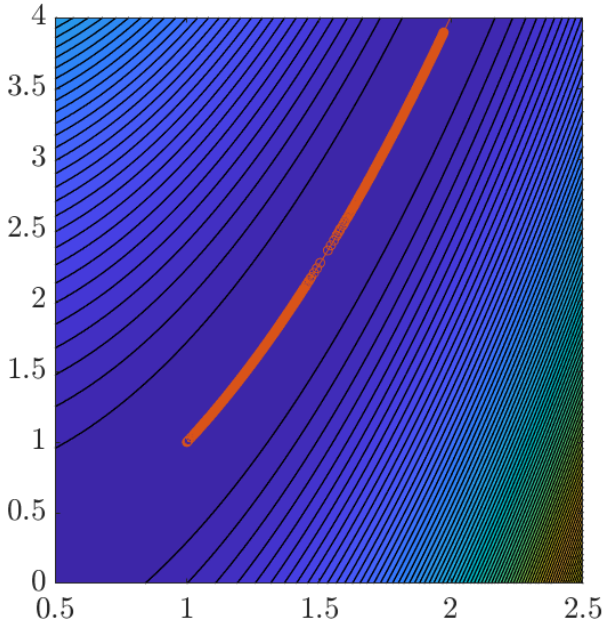


Figure 19: Gradient descent with backtracking line search for the Rosenbrock problem. Convergence towards the optimum $(1,1)$ is achieved after several thousands iterations. The method is extremely slow due to poor conditioning of the Hessian around the optimum.

Scaled Gradient Method

A way to mitigate the slow convergence due to poor conditioning of the Hessian is to formulate a rescaled version of the problem. Consider the minimization problem

$$\min \{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\}$$

For a given nonsingular matrix $\mathbf{S} \in \mathbb{R}^{n\times n}$, we make the linear change of variables $\mathbf{x} = \mathbf{S}\mathbf{y}$, and obtain the equivalent problem

$$\min \{g(\mathbf{y}) \equiv f(\mathbf{S}\mathbf{y}) : \mathbf{y} \in \mathbb{R}^n\} .$$

Since $\nabla g(\mathbf{y}) = \mathbf{S}^\top \nabla f(\mathbf{S}\mathbf{y}) = \mathbf{S}^\top \nabla f(\mathbf{x})$, the gradient method for the rescaled problem reads

$$\mathbf{y}^{k+1} = \mathbf{y}^k - t^k \mathbf{S}^\top \nabla f(\mathbf{S}\mathbf{y}^k) .$$

Multiplying the latter equality by $\mathbf{S}$ from the left, and using the notation $\mathbf{x}^k = \mathbf{S}\mathbf{y}^k$

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \mathbf{S}\mathbf{S}^\top \nabla f(\mathbf{x}^k)$$

Defining $\mathbf{D} = \mathbf{S}\mathbf{S}^\top$, we obtain the scaled gradient method:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \mathbf{D}\nabla f(\mathbf{x}^k) .$$

Note that $\mathbf{D} \succ \mathbf{0}$, so the direction $-\mathbf{D}\nabla f\left(\mathbf{x}^k\right)$ is a descent direction:

$$f'(\mathbf{x}^k; -\mathbf{D}\nabla f(\mathbf{x}^k)) = -\nabla f(\mathbf{x}^k)^\top \mathbf{D}\nabla f(\mathbf{x}^k) < 0 .$$

We also allow different scaling matrices at each iteration.

Algorithm 3: Scaled Gradient Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in \mathbb{R}^n$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Pick a scaling matrix $\mathbf{D}^k > 0$
2 Pick a stepsize t^k by a line search procedure on the function

$$g(t) = f\left(\mathbf{x}^k - t\mathbf{D}^k \nabla f\left(\mathbf{x}^k\right)\right)$$

3 Set $\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \mathbf{D}^k \nabla f\left(\mathbf{x}^k\right)$.
4 If $\left\|\nabla f\left(\mathbf{x}^{k+1}\right)\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

Choosing the Scaling Matrix $\mathbf{D}^k$. The scaled gradient method with scaling matrix $\mathbf{D}$ is equivalent to the gradient method employed on the function $g(\mathbf{y}) = f\left(\mathbf{D}^{1/2}\mathbf{y}\right)$, where $\mathbf{D}^{1/2}$ is a matrix such that $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$. Note that the gradient and Hessian of g are given by

$$\nabla g(\mathbf{y}) = \mathbf{D}^{1/2} f\left(\mathbf{D}^{1/2}\mathbf{y}\right) = \mathbf{D}^{1/2}\nabla f(\mathbf{x})$$
$$\nabla^2 g(\mathbf{y}) = \mathbf{D}^{1/2}\nabla^2 f\left(\mathbf{D}^{1/2}\mathbf{y}\right)\mathbf{D}^{1/2} = \mathbf{D}^{1/2}\nabla^2 f(\mathbf{x})\mathbf{D}^{1/2}$$

The objective is usually to pick $\mathbf{D}^k$ so as to make $(\mathbf{D}^k)^{1/2}\nabla^2 f\left(\mathbf{x}^k\right)(\mathbf{D}^k)^{1/2}$ as well-conditioned as possible. A well known choice (Newton's method) is to pick $\mathbf{D}^k = \left(\nabla^2 f\left(\mathbf{x}^k\right)\right)^{-1}$. Another alternative is to use a diagonal scaling: $\mathbf{D}^k$ is picked to be diagonal. For example,

$$\left(\mathbf{D}^k\right)_{ii} = \left(\frac{\partial^2 f\left(\mathbf{x}^k\right)}{\partial x_i^2}\right)^{-1}$$

Using diagonal scaling can be very effective when the decision variables are of different magnitudes.

Example Revisit the Rosenbrock example (p. 12) using a suitable scaling.

```
function [hist,x,fun_val]=gradient_method_quadratic(A,b,x0,epsilon);
% INPUT
% ======================
% A ....... the positive definite matrix associated with the ...
    objective function
% b ....... a column vector associated with the linear part of ...
    the objective function
% x0 ...... starting point of the method
% epsilon . tolerance parameter
% OUTPUT
% ======================
% hist.....convergence history of the iterations
% x ....... an optimal solution (up to a tolerance) of ...
    min(x^{\top} A x+2 b^{\top} x)
% fun_val . the optimal function value up to a tolerance


x=x0;
iter=0;
grad=2*(A*x+b);
hist=x0;
while (norm(grad)>epsilon)
iter=iter+1;
t=norm(grad)^2/(2*grad'*A*grad);
x=x-t*grad;
grad=2*(A*x+b);
fun_val=x'*A*x+2*b'*x;
hist=[hist x];
fprintf('iter_number = %3d norm_grad = %2.6f fun_val = %2.6f ...
    \n',iter,norm(grad),fun_val);
end
```

```
function [hist,x,fun_val]=gradient_method_constant(f,g,x0,t,epsilon)
% Gradient method with constant stepsize
%
% INPUT
%=======================================
% f ......... objective function
% g ......... gradient of the objective function
% x0......... initial point
% t ......... constant stepsize
% epsilon ... tolerance parameter
% OUTPUT
%=======================================
% x ......... optimsl solution (up to a tolerance)
% of min f(x)
% fun_val ... optimal function value
x=x0;
grad=g(x);
iter=0;
hist=x0;
while (norm(grad)>epsilon)
iter=iter+1;
x=x-t*grad;
fun_val=f(x);
grad=g(x);
hist=[hist x];
fprintf('iter_number = %3d norm_grad = %2.6f fun_val = %2.6f ...
    \n',iter,norm(grad),fun_val);
end
```

```
function ...
    [x,fun_val]=gradient_method_backtracking(f,g,x0,s,alpha,beta,epsilon)
% Gradient method with backtracking stepsize rule
%
% INPUT
%=======================================
% f ......... objective function
% g ......... gradient of the objective function
% x0......... initial point
% s ......... initial choice of stepsize
% alpha ..... tolerance parameter for the stepsize selection
% beta ...... the constant in which the stepsize is multiplied
% at each backtracking step (0<beta<1)
% epsilon ... tolerance parameter for stopping rule
% OUTPUT
%=======================================
% x ......... optimal solution (up to a tolerance)
% of min f(x)
% fun_val ... optimal function value
x=x0;
grad=g(x);
fun_val=f(x);
iter=0;
while (norm(grad)>epsilon)
iter=iter+1;
t=s;
while (fun_val-f(x-t*grad)<alpha*t*norm(grad)^2)
t=beta*t;
end
x=x-t*grad;
fun_val=f(x);
grad=g(x);
fprintf('iter_number = %3d norm_grad = %2.6f fun_val = %2.6f ...
    \n',iter,norm(grad),fun_val);
end
```

The Gauss-Newton Method

We use the gradient method from last week to build an algorithm for solving nonlinear least squares problem of the type:

$$\text{(NLS):}\ \min_{\mathbf{x}\in\mathbb{R}^n}\left\{g(\mathbf{x})\equiv\sum_{i=1}^{m}\left(f_i(\mathbf{x})-c_i\right)^2\right\}$$

$f_1, \ldots, f_m$ are continuously differentiable over $\mathbb{R}^n$ and $c_1, \ldots, c_m \in \mathbb{R}$.

Denote:

$$F(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) - c_1 \\ f_2(\mathbf{x}) - c_2 \\ \vdots \\ f_m(\mathbf{x}) - c_m \end{pmatrix}$$

Then the problem becomes:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|F(\mathbf{x})\|^2$$

Figure 20: Sir Isaac Newton (1642–1727) the protagonist of this week, needs no introduction. On the left, he discovers Pink Floyd's "The Dark Side of the Moon" album. On the right, a recreation of a little domestic incident involving fire and his notes on Optics.

The Gauss-Newton method is an algorithm for solving this particular problem, and it reads as follows. Given the k th iterate $\mathbf{x}^k$, the next iterate is chosen to minimize the sum of squares of the linearized terms, that is,

$$\mathbf{x}^{k+1} = \underset{\mathbf{x} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \sum_{i=1}^{m} \left[f_i(\mathbf{x}^k) + \nabla f_i(\mathbf{x}^k)^\top (\mathbf{x} - \mathbf{x}^k) - c_i \right]^2 \right\}$$

The general step actually consists of solving the linear LS problem

$$\min \left\| \mathbf{A}^k \mathbf{x} - \mathbf{b}^k \right\|^2$$

where

$$\mathbf{A}^k = \begin{pmatrix} \nabla f_1 \left(\mathbf{x}^k\right)^\top \\ \nabla f_2 \left(\mathbf{x}^k\right)^\top \\ \vdots \\ \nabla f_m \left(\mathbf{x}^k\right)^\top \end{pmatrix} = J(\mathbf{x}^k),$$

is the Jacobian matrix, assumed to have full column rank, and

$$\mathbf{b}^k = \begin{pmatrix} \nabla f_1\left(\mathbf{x}^k\right)^\top \mathbf{x}^k - f_1\left(\mathbf{x}^k\right) + c_1 \\ \nabla f_2\left(\mathbf{x}^k\right)^\top \mathbf{x}^k - f_2\left(\mathbf{x}^k\right) + c_2 \\ \vdots \\ \nabla f_m\left(\mathbf{x}^k\right)^\top \mathbf{x}^k - f_m\left(\mathbf{x}^k\right) + c_m \end{pmatrix} = J(\mathbf{x}^k)\mathbf{x}^k - F(\mathbf{x}^k)\,.$$

The Gauss-Newton method can thus be written as:

$$\mathbf{x}^{k+1} = \left(J(\mathbf{x}^k)^\top J(\mathbf{x}^k)\right)^{-1} J(\mathbf{x}^k)^\top \mathbf{b}^k\,.$$

Note that the gradient of the objective function $g(\mathbf{x}) = \|F(\mathbf{x})\|^2$ is

$$\nabla g(\mathbf{x}) = 2J(\mathbf{x})^\top F(\mathbf{x})\,.$$

The Gauss-Newton method can be rewritten as follows:

$$\begin{aligned} \mathbf{x}^{k+1} &= \left(J(\mathbf{x}^k)^\top J(\mathbf{x}^k)\right)^{-1} J(\mathbf{x}^k)^\top \left(J(\mathbf{x}^k)\mathbf{x}^k - F(\mathbf{x}^k)\right) \\ &= \mathbf{x}^k - \left(J(\mathbf{x}^k)^\top J(\mathbf{x}^k)\right)^{-1} J(\mathbf{x}^k)^\top F(\mathbf{x}^k) \\ &= \mathbf{x}^k - \frac{1}{2}\left(J(\mathbf{x}^k)^\top J(\mathbf{x}^k)\right)^{-1} \nabla g(\mathbf{x}^k) \end{aligned}$$

that is, it is a scaled gradient method with a special choice of scaling matrix:

$$\mathbf{D}^k = \frac{1}{2}\left(J(\mathbf{x}^k)^\top J(\mathbf{x}^k)\right)^{-1}\,.$$

The Damped Gauss-Newton Method

The Gauss-Newton method does not incorporate a stepsize, which might cause it to diverge. A well known variation of the method incorporating stepsizes is the **damped Gauss-Newton method**.

Algorithm 4: The Damped Gauss-Newton Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in \mathbb{R}^n$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Set $\mathbf{d}^k = -\left(J\left(\mathbf{x}^k\right)^\top J\left(\mathbf{x}^k\right)\right)^{-1} J\left(\mathbf{x}^k\right)^\top \mathrm{F}\left(\mathbf{x}^k\right)$.

2 Set t^k by a line search procedure on the function

$$h(t) = g(\mathbf{x}^k + t\mathbf{d}^k)\,.$$

3 Set $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k \mathrm{d}_k$.

4 If $\left\|\nabla g\left(\mathbf{x}^{k+1}\right)\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

An Example of a Gradient Method: The Fermat-Weber Problem and Weiszfeld's Method

The Fermat-Weber problem is stated as follows. Given m points in $\mathbb{R}^n$: $\mathbf{a}_1, \ldots, \mathbf{a}_m$, also called *anchor points*, and m weights $\omega_1, \omega_2, \ldots, \omega_m > 0$, find a point $\mathbf{x} \in \mathbb{R}^n$ that minimizes the weighted distance of $\mathbf{x}$ to each of the points $\omega_1, \omega_2, \ldots, \omega_m > 0$, that is

$$\min_{\mathbf{x}\in\mathbb{R}^n} \left\{ f(\mathbf{x}) \equiv \sum_{i=1}^{m} \omega_i \left\|\mathbf{x} - \mathbf{a}_i\right\| \right\}$$

Note that the objective function is not differentiable at the anchor points $\mathbf{a}_1, \ldots, \mathbf{a}_m$ (why?).

In terms of applications, this is one of the simplest instances of facility location problems. In 1937, with only 16 years old, the Hungarian mathematician Endre Weiszfeld proposed a method for solving this problem, unsurprisingly known as *Weiszfeld's Method*. Under the assumption that the optimum $\mathbf{x}$ is not an anchor point, we write the stationarity condition

$$\begin{aligned}
\nabla f(\mathbf{x}) &= 0\,, \\
\sum_{i=1}^{m} \omega_i \frac{\mathbf{x} - \mathbf{a}_i}{\left\|\mathbf{x} - \mathbf{a}_i\right\|} &= 0\,, \\
\left(\sum_{i=1}^{m} \frac{\omega_i}{\left\|\mathbf{x} - \mathbf{a}_i\right\|} \right) \mathbf{x} &= \sum_{i=1}^{m} \frac{\omega_i \mathbf{a}_i}{\left\|\mathbf{x} - \mathbf{a}_i\right\|}\,.
\end{aligned}$$

The stationarity condition can be written as a fixed point $\mathbf{x} = T(\mathbf{x})$, where T is the operator

$$T(\mathbf{x}) \equiv \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\left\|\mathbf{x}-\mathbf{a}_i\right\|}} \sum_{i=1}^{m} \frac{\omega_i \mathbf{a}_i}{\left\|\mathbf{x} - \mathbf{a}_i\right\|}\,.$$

Weiszfeld's method is the fixed point iteration:

$$\mathbf{x}^{k+1} = T(\mathbf{x}^k)\,,$$

which can be interpreted as a gradient method since

$$\begin{aligned}
\mathbf{x}^{k+1} &= \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\left\|\mathbf{x}^k-\mathbf{a}_i\right\|}} \sum_{i=1}^{m} \frac{\omega_i \mathbf{a}_i}{\left\|\mathbf{x}^k - \mathbf{a}_i\right\|} \\
&= \mathbf{x}^k - \frac{1}{\sum_{i=1}^{m} \frac{1}{\left\|\mathbf{x}^k-\mathbf{a}_i\right\|}} \sum_{i=1}^{m} \omega_i \frac{\mathbf{x}^k - \mathbf{a}_i}{\left\|\mathbf{x}^k - \mathbf{a}_i\right\|} \\
&= \mathbf{x}^k - \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\left\|\mathbf{x}^k-\mathbf{a}_i\right\|}} \nabla f(\mathbf{x}^k)\,.
\end{aligned}$$

Therefore, it corresponds to a gradient method with a special choice of stepsize:

$$t^k = \frac{1}{\sum_{i=1}^{m} \frac{\omega_i}{\|\mathbf{x}^k - \mathbf{a}_i\|}}$$

Newton's Method (Un-assessed)

We now discuss a different class of algorithms for solving the problem

$$\min \{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\} .$$

Assuming that f is twice continuously differentiable over $\mathbb{R}^n$, looking for the optimum begins with the search of a stationary point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$. This can be framed as finding a root (zero) for $g(\mathbf{x}) \equiv \nabla f(\mathbf{x})$. A classical algorithm for finding the zeros of a function is Newton's method. For $g : \mathbb{R} \to \mathbb{R}$ it reads

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{g(\mathbf{x}^k)}{g'(\mathbf{x}^k)} .$$

We will utilise this idea for minimizing a function in several variables. In particular, going back to $f(\mathbf{x})$ and assuming it is twice continuously differentiable over $\mathbb{R}^n$, we generate a sequence $\{\mathbf{x}^k\}_{k=1}^{\infty}$ converging to a stationary point of $f(\mathbf{x})$ through the iteration

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^\top (\mathbf{x} - \mathbf{x}^k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}^k)^\top \nabla^2 f(\mathbf{x}^k)(\mathbf{x} - \mathbf{x}^k) \right\} .$$

This expression is interpreted as follows. Given a current value $\mathbf{x}^k$, we built a quadratic approximation of $f(\mathbf{x})$ around $\mathbf{x}^k$ (recall the quadratic approximation theorem), and find the minimizer of this parabolic approximation. In the case the minimizer exists, this is the next point of our sequence $\mathbf{x}^{k+1}$, where we move and repeat. This minimization problem not well-defined in general. However, if the Hessian is positive definite at $\mathbf{x}^k$, $\nabla^2 f\left(\mathbf{x}^k\right) > 0$, then the update reads

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\nabla^2 f(\mathbf{x}^k))^{-1} \nabla f(\mathbf{x}^k)$$

The vector

$$-(\nabla^2 f(\mathbf{x}^k))^{-1} \nabla f(\mathbf{x}^k)$$

is called Newton's direction. This algorithm is called Pure Newton's Method and it is summarized below.

Algorithm 5: Pure Newton's Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in \mathbb{R}^n$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Compute the Newton direction $\mathbf{d}^k$, which is the solution to the linear system

$$\nabla^2 f(\mathbf{x}^k)\mathbf{d}^k = -\nabla f(\mathbf{x}^k) .$$

2 Set $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$.

3 if $\left\|\nabla f\left(\mathbf{x}^{k+1}\right)\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

Convergence of Newton's Method

At the very least, Newton's method requires that $\nabla^2 f(\mathbf{x}) \succ 0$ for every $\mathbf{x} \in \mathbb{R}^n$, which in particular implies that there exists a unique optimal solution $\mathbf{x}^*$. However, this is not enough to guarantee convergence.

Example. Analyse what happens when $f(x) = \sqrt{1+x^2}$.

A lot of assumptions are required to be made in order to guarantee convergence of the method. However, Newton's method does have one very attractive feature. Under certain assumptions one can prove local quadratic rate of convergence, which means that near the optimal solution the errors $e^k = \left\|\mathbf{x}^k - \mathbf{x}^*\right\|$ satisfy an inequality $e^{k+1} \leq M(e^k)^2$ for some positive $M > 0$. This property essentially means that the number of accuracy digits is doubled at each iteration. This is in contrast to the gradient method in which the convergence theorems are rather independent in the starting point, but only "relatively" slow linear convergence is assured.

Theorem (Quadratic Local Convergence of Newton's Method)**.** *Let f be a twice continuously differentiable function defined over $\mathbb{R}^n$. Assume that*

- *there exists $m > 0$ for which $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$, for any $\mathbf{x} \in \mathbb{R}^n$,*
- *there exists $L > 0$ for which $\left\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\right\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.*

Let $\left\{\mathbf{x}^k\right\}_{k\geq 0}$ be the sequence generated by Newton's method and let $\mathbf{x}^*$ be the unique minimizer of f over $\mathbb{R}^n$. Then for any $k = 0, 1, \ldots$ the inequality

$$\left\|\mathbf{x}^{k+1} - \mathbf{x}^*\right\| \leq \frac{L}{2m}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2$$

holds. In addition, if $\left\|\mathbf{x}^0 - \mathbf{x}^*\right\| \leq \frac{m}{L}$, then:

$$\left\|\mathbf{x}^k - \mathbf{x}^*\right\| \leq \frac{2m}{L}\left(\frac{1}{4}\right)^{2^k}, \quad k = 0, 1, 2, \ldots$$

Proof. We prove the first part of the result. Let k be a nonnegative integer. Then

$$\begin{aligned}
\mathbf{x}^{k+1} - \mathbf{x}^* =& \mathbf{x}^k - (\nabla^2 f(\mathbf{x}^k))^{-1}\nabla f(\mathbf{x}^k) - \mathbf{x}^* \\
=& \mathbf{x}^k - \mathbf{x}^* + (\nabla^2 f(\mathbf{x}^k))^{-1}(\nabla f(\mathbf{x}^*) - \nabla f(\mathbf{x}^k)) \\
=& \mathbf{x}^k - \mathbf{x}^* + (\nabla^2 f(\mathbf{x}^k))^{-1}\int_0^1 \left[\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}^* - \mathbf{x}^k))\right](\mathbf{x}^* - \mathbf{x}^k)dt \\
=& (\nabla^2 f(\mathbf{x}^k))^{-1}\int_0^1 \left[\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}^* - \mathbf{x}^k)) - \nabla^2 f(\mathbf{x}^k)\right](\mathbf{x}^* - \mathbf{x}^k)dt
\end{aligned}$$

Combining the latter equality with the fact that $\nabla^2 f(\mathbf{x}^k) \succeq m\mathbf{I}$, implies that

$$\left\|(\nabla^2 f(\mathbf{x}^k))^{-1}\right\| \leq \frac{1}{m},$$

Hence,

$$\begin{aligned}
\left\|\mathbf{x}^{k+1}-\mathbf{x}^*\right\| &\leq \left\|(\nabla^2 f(\mathbf{x}^k))^{-1}\right\| \left\|\int_0^1 \left[\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}^* - \mathbf{x}^k)) - \nabla^2 f(\mathbf{x}^k)\right](\mathbf{x}^* - \mathbf{x}^k)dt\right\| \\
&\leq \left\|(\nabla^2 f(\mathbf{x}^k))^{-1}\right\| \int_0^1 \left\|\left[\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}^* - \mathbf{x}^k)) - \nabla^2 f(\mathbf{x}^k)\right](\mathbf{x}^* - \mathbf{x}^k)\right\| dt \\
&\leq \left\|(\nabla^2 f(\mathbf{x}^k))^{-1}\right\| \int_0^1 \left\|\nabla^2 f(\mathbf{x}^k + t(\mathbf{x}^* - \mathbf{x}^k)) - \nabla^2 f(\mathbf{x}^k)\right\| \cdot \left\|\mathbf{x}^* - \mathbf{x}^k\right\| dt \\
&\leq \frac{L}{m}\int_0^1 t\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 dt = \frac{L}{2m}\left\|\mathbf{x}^k - \mathbf{x}^*\right\|^2 .
\end{aligned}$$

The second part of the theorem can be shown by induction (try it!). □

Numerical Example. Consider the minimization problem

$$\min 100x^4 + 0.01y^4 .$$

Compare Newton's method against a gradient descent with backtracking. Repeat for

$$\min \sqrt{x_1^2+1} + \sqrt{x_2^2+1}.$$

Note that ins this case, the Hessian of the function is

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{1}{(x_1^2+1)^{3/2}} & 0 \\ 0 & \frac{1}{(x_2^2+1)^{3/2}} \end{pmatrix} > 0$$

but there does not exists an $m > 0$ for which $\nabla^2 f(x) \geq m\mathbf{I}$. In this case using pure's Newton method will diverge. An alternative is to use a damped version of Newton's method, using backtracking, as follows.

Algorithm 6: Damped Newton's Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in \mathbb{R}^n$. (α, β) parameters for the backtracking procedure $(\alpha \in (0,1), \beta \in (0,1))$.

General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Compute the Newton direction d^k, which is the solution to the linear system

$$\nabla^2 f(\mathbf{x}^k)\mathbf{d}^k = -\nabla f(\mathbf{x}^k).$$

2 Set $t_k = 1$. **while** $\left(f\left(\mathbf{x}^k\right) - f\left(\mathbf{x}^k + t^k\mathbf{d}^k\right) < -\alpha t^k \nabla f\left(\mathbf{x}^k\right)^\top \mathbf{d}^k\right)$ **do**

3 set $t^k := \beta t^k$.

4 Set $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\mathbf{d}^k$.

5 If $\left\|\nabla f\left(\mathbf{x}^{k+1}\right)\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

Repeat the last numerical example using damped Newton's method starting from $(10, 10)$.

```
function x=pure_newton(f,g,h,x0,epsilon)
% Pure Newton's method
%
% INPUT
% ==============
% f .......... objective function.
% g .......... gradient of the objective function
% h .......... Hessian of the function
% x0.......... initial point
% epsilon ..... tolerance
% OUTPUT
% ==============
% x - solution obtained by Newton's method (up to some tolerance)

if (nargin<5)
epsilon=1e-5;
end

x=x0;
gval=g(x);
hval=h(x);
iter=0;
while ((norm(gval)>epsilon)&&(iter<10000))
iter=iter+1;
x=x-hval\gval;
fprintf('iter= %2d f(x)=%10.10f\n',iter,f(x))
gval=g(x);
hval=h(x);
end

if (iter==10000)
fprintf('did not converge')
end
```

```
function x=newton_backtracking(f,g,h,x0,alpha,beta,epsilon)
% Newton's method with backtracking
%
% INPUT
%=======================================
% f ......... objective function
% g ......... gradient of the objective function
% h ......... hessian of the objective function
% x0......... initial point
% alpha ..... tolerance parameter for the stepsize selection strategy
% beta ...... the proportion in which the stepsize is multiplied
% at each backtracking step (0<beta<1)
% epsilon ... tolerance parameter for stopping rule
% OUTPUT
%=======================================
% x ......... optimal solution (up to a tolerance)
% of min f(x)
% fun_val ... optimal function value

x=x0;
gval=g(x);
hval=h(x);
d=hval\gval;
iter=0;
while ((norm(gval)>epsilon)&&(iter<10000))
iter=iter+1;
t=1;
while(f(x-t*d)>f(x)-alpha*t*gval'*d)
t=beta*t;
end
x=x-t*d;
fprintf('iter= %2d f(x)=%10.10f\n',iter,f(x))
gval=g(x);
hval=h(x);
d=hval\gval;
end

if (iter==10000)
fprintf('did not converge\n')
end
```

Part V.
Stochastic Gradient Descent

The Kaczmarz Algorithm

We begin our discussion by studying a classical algorithm proposed by the Polish mathematician Stefan Kaczmarz in 1937, and which was later re-discovered in the 1970s in image processing. This technique was implemented in the very first medical scanners. The Kaczmarz algorithm solves the linear system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

by iterating projections along the $i-th$ for of the matrix $\mathbf{A}$, denoted by $\mathbf{a}_i^\top$:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \frac{b_i - \mathbf{a}_i^\top \mathbf{x}^k}{\|\mathbf{a}_i\|^2}\mathbf{a}_i\,. \qquad \text{(Kaczmarz)}$$

Note that this algorithm does not require to compute $\mathbf{A}^{-1}$! In the original Kaczmarz algorithm, the $i-$th row that is chosen at the $k-$th iteration of the algorithm is cycled periodically through all the rows of the matrix $\mathbf{A}$, i.e.

$$i = \mod(k, m) + 1,$$

where m is the number of rows of $\mathbf{A}$. Provided that the system is consistent, that is, there exists at least one solution of the system, the iteration $\mathbf{x}^k$ converges to the minimum norm solution of the problem, assuming that $\mathbf{x}^0 = \mathbf{0}$. The convergence analysis of this algorithm remained an open problem until probabilistic methods were introduced by 2009. Nowadays, we can show that the Kaczmarz algorithm converges exponential (and independently of the number of rows) if at the $k-$th iteration, the $i-$th row is chosen randomly. This algorithm is known as *Randomized Kaczmarz Algorithm*. We can sample uniformly among the rows, or with a probability that is proportional to the squared row norm $\|\mathbf{a}_i\|^2$, known as *importance sampling*. Figure 21[4] shows a comparison between cycling (standard Kaczmarz), uniform (simple randomized Kaczmarz), and importance sampling (randomized Kaczmarz) among the rows of a 300 by 100 linear system. The plot shows the evolution of the least squares error, $\|\mathbf{A}\mathbf{x}^k - \mathbf{b}\|$, against the number of iterations (projections) for each algorithm. For every variant of the Kaczmarz algorithm, the method converges to the solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$, however the method converges faster if a randomized row selection is selected instead of a deterministic cycling. For large-scale systems, avoiding the inversion of $\mathbf{A}$ is desirable, so this method can be applied in this case. We recall that solving the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ can be cast as the optimization

[4] Taken from Strohmer, Thomas; Vershynin, Roman (2009), "A randomized Kaczmarz algorithm for linear systems with exponential convergence", Journal of Fourier Analysis and Applications, 15 (2): 262–278, arXiv:math/0702226

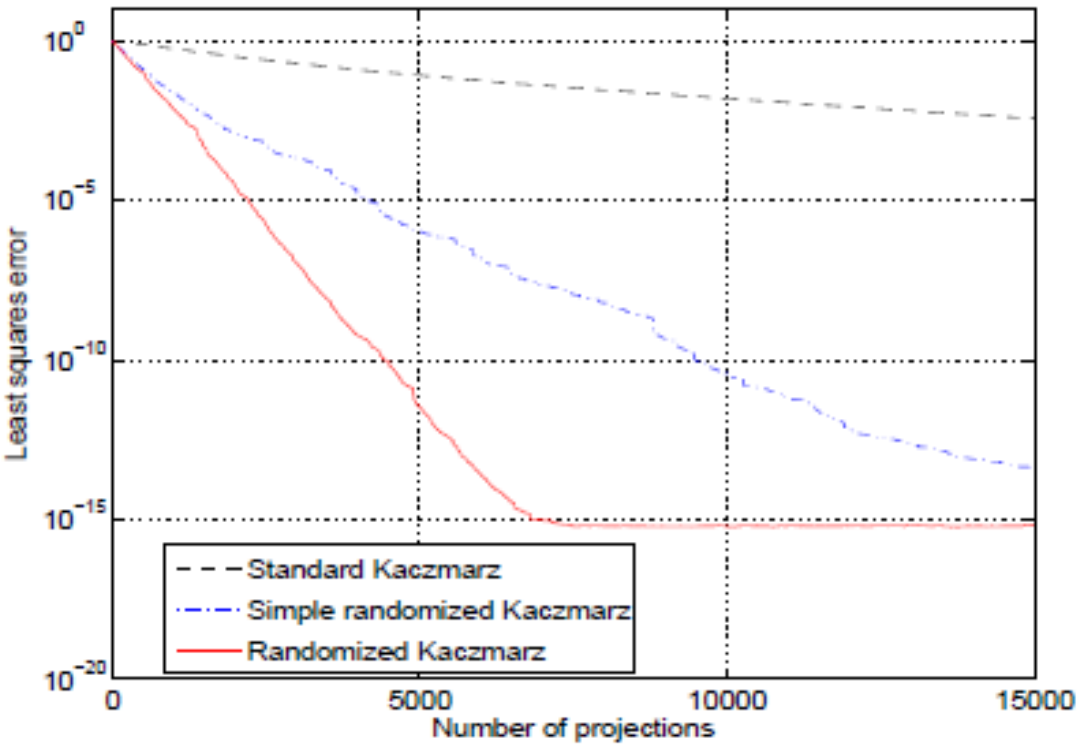


Figure 21: Iteration of the Kaczmarz algorithm for a 300 by 100 linear system of equations, under three different sampling procedures for the rows of $\mathbf{A}$.

problem

$$\min_{\mathbf{x}} \frac{1}{2m}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 = \frac{1}{2m}\sum_{i=1}^{m}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2\,,$$

for which a gradient descent method can be constructed as

$$\mathbf{x}^{k+1} = \mathbf{x}^k - \frac{t}{m}\mathbf{A}^\top(\mathbf{A}\mathbf{x}-\mathbf{b})\,,$$

and comparing against (Kaczmarz) we can interpret it as a gradient descent where at each iteration, instead of computing the full gradient of

$$\frac{1}{2m}\sum_{i=1}^{m}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2\,,$$

we sample a single element of the cost and work with the gradient of

$$\frac{1}{2m}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2\,.$$

Note that in the case of a uniform sampling among the rows, we can write

$$\frac{1}{2m}\sum_{i=1}^{m}(\mathbf{a}_i^\top \mathbf{x} - b_i)^2 = \frac{1}{2}\mathbb{E}_i[(\mathbf{a}_i^\top \mathbf{x} - b_i)^2]\,,$$

where the expected value is among a uniform distribution when choosing the vector $[\mathbf{a}_i^\top | b_i]$ uniformly from the rows of the augmented matrix $[\mathbf{A}|\mathbf{b}]$. The generalization of this idea for nonlinear regression problems is what we will discuss in the next section as stochastic gradient descent.

Stochastic Gradient Descent

Stochastic Gradient Descent (SGD) is a fundamental algorithm in machine learning. It is naturally meant for solving nonlinear regression/estimation problems where the cost is of the type

$$\min_{\mathbf{x}} \quad \frac{1}{m}\sum_{i=1}^{m} Q_i(\mathbf{x})\,. \tag{3}$$

A natural setting for such an optimization framework arises in nonlinear regression problems.

Nonlinear regression example. Consider the nonlinear model in $\theta \in \mathbb{R}$

$$f(\theta;\mathbf{x}) = x_1 e^{x_2\theta}\cos(x_3\theta + x_4)\,,$$

with parameters $\mathbf{x} \in \mathbb{R}^4$ for which we want to find the optimal value $\mathbf{x}^*$ minimizing the norm of the ℓ_2-error with respect to m observations of the *true* model

$$\hat{f}_i := f(\theta_i)\,, \qquad i = 1,\dots,m\,.$$

We formulate this problem as a nonlinear least squares problem

$$\min_{\mathbf{x}} g(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m}(f(\theta_i;\mathbf{x}) - \hat{f}_i)^2\,. \tag{NLS}$$

Note that the $\frac{1}{m}$ scaling does not affect the minimizer. Setting

$$Q_i(\mathbf{x}) = (f(\theta_i;\mathbf{x}) - \hat{f}_i)^2$$

we can see how nonlinear regression problems lead to costs or **loss functions** similar to eq. (3). In general, while setting a gradient descent iteration for this problem

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k\nabla g(\mathbf{x}^k) = \mathbf{x}^k - \frac{t^k}{m}\sum_{i=1}^{m}\nabla Q_i(\mathbf{x}^k)$$

is straightforward (following exactly what we discussed in the previous week), when the number of observations m is large[5], and the parameter space is high-dimensional ($\mathbf{x} \in \mathbb{R}^n$, $n >>> 1$), the computation of $\sum_{i=1}^{m}\nabla Q_i(\mathbf{x}^k)$ is overwhelmingly expensive. In the

[5] as in a *big data* framework -think about image datasets, Spotify songs-

spirit of the Kaczmarz algorithm, stochastic gradient descent circumvents this limitation. Stochastic Gradient Descent dates back to 1951 to the paper by Robbins and Munro *"A stochastic approximation method"*, and instead of computing the gradient with all m points, only a single data point is randomly selected and used. At the next iteration, another randomly selected point is used to compute the gradient and update the solution

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla Q_i(\mathbf{x}^k)\,,$$

where the index i is sampled in each gradient iteration. In machine learning applications, the parameter t^k is known as the **learning rate**. Note that if the samples are drawn uniformly, then

$$g(\mathbf{x}) := \frac{1}{m}\sum_{i=1}^{m} Q_i(\mathbf{x}) = \mathbb{E}_i[Q_i(\mathbf{x})]\,.$$

and that a random sample constitutes and unbiased estimator of $\nabla g(\mathbf{x})$. The convergence of SGD is stated in the following theorem.

Theorem (Convergence of SGD)**.** *Assume:*

- *The cost* $g(\mathbf{x})$ *is such that*

$$\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\|\,, \quad \textit{and} \quad \nabla^2 g(\mathbf{x}) \succeq \mu \mathbf{I}\,.$$

- *The sample gradient* $\nabla Q_i(\mathbf{x}^k)$ *is an unbiased estimate of* $\nabla g(\mathbf{x}^k)$*.*
- *For all* $\mathbf{x}$*,*

$$\mathbb{E}_i[\|Q_i(\mathbf{x})\|^2] \le \sigma^2 + c\|\nabla g(\mathbf{x})\|^2\,.$$

Then, if $t^k \equiv t \le \frac{1}{Lc}$*, then SGD achieves*

$$\mathbb{E}[g(\mathbf{x}^k) - g(\mathbf{x}^*)] \le \frac{tL\sigma^2}{2\mu} + (1-t\mu)^k (g(\mathbf{x}^0) - g(\mathbf{x}^*))\,.$$

The result above implies:

1. Fast (linear) convergence during the first iterations.
2. Convergence to a neighbourhood of $\mathbf{x}^*$, without further progress.
3. If gradient computation is noiseless, that is $\sigma = 0$, then linear convergence to optimal points.
4. A smaller stepsize t yield better converging points.

The algorithm may require multiple passes through all the data to converge, but each step is now easy to evaluate versus the full computation of the gradient. If instead of a single point, a different subset of points is sampled at each iteration, then we have a **batch gradient descent** algorithm:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - t^k \nabla g(\mathbf{x}^k) = \mathbf{x}^k - \frac{t^k}{|K|}\sum_{i\in K} \nabla Q_i(\mathbf{x}^k)\,,$$

where K denotes a set of p randomly selected datapoints.

Revisiting the Nonlinear Regression Example

If we go back to the model

$$f(\theta; \mathbf{x}) = x_1 e^{x_2 \theta} \cos(x_3 \theta + x_4)\,,$$

during week 5 we explored the Gauss-Newton method for nonlinear regression. We shall modify this setting and implement SGD. Results shown in Figure 22 show the different behaviours observed for different learning rates. Notice the large number of iterations required.

Model:

```
function [val]=model(t,x)
val= x(1).*exp(x(2).*t).*cos(x(3).*t+x(4)); %%model
```

Gradient of the model with respect to $\mathbf{x}$:

```
function [val]=modelgrad(t,x)
val(1,1)=exp(x(2).*t).*cos(x(3).*t+x(4));
val(2,1)=x(1).*exp(x(2).*t).*t.*cos(x(3).*t+x(4));
val(3,1)=-x(1).*exp(x(2).*t).*sin(x(3).*t+x(4)).*t;
val(4,1)=-x(1).*exp(x(2).*t).*sin(x(3).*t+x(4));
```

Cost function:

```
function [val]=g1(x,tm,fn)
m=length(tm);
val=norm(model(tm,x)-fn)^2/m;
```

Gradient of the cost function:

```
function[val]=gradg1(x,tm,fn)
m=length(tm);
ind=randi([1 m]); %% sampling among the dataset
val=2*(model(tm(ind),x)-fn(ind))*modelgrad(tm(ind),x);
```

Main:

```
clear all

xt=[1;2;pi;0]; %% true parameters
tm=[-1:0.001:1]'; %% measurements of the independent variable
```

```
randn('seed',666);
ft=model(tm,xt);%% true model measurements
m=length(tm);
maxiter=10^4;
x0=[1;1;1;1];
xm=x0;
t=0.01; %% learning rate
hist=[g1(xm,tm,fn)];
for i=1:maxiter
xp=xm-t*gradg1(xm,tm,ft); %%the random sampling is inside gradg1
hist=[hist g1(xp,tm,ft)];
xm=xp;
end
```

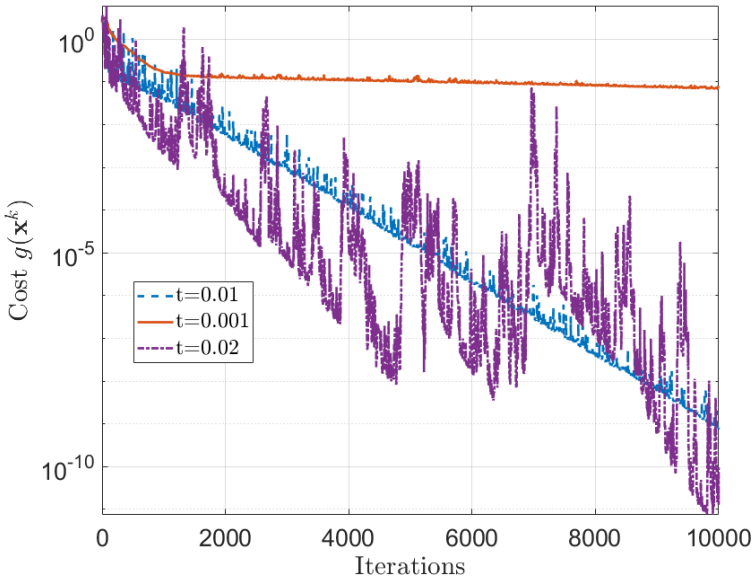


Figure 22: Convergence of the SGD iteration for different learning rates.

Part VI.
Convex Sets and Functions

Convex Sets

We begin discussing convexity by defining what a convex set is.

Definition (Convex Set). *A set $C \subseteq \mathbb{R}^n$ is called convex if for any $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0,1]$ the point $\lambda\mathbf{x} + (1-\lambda)\mathbf{y}$ belongs to C.*

In other words, the above definition is equivalent to saying that for any $\mathbf{x}, \mathbf{y} \in C$, the line segment $[\mathbf{x}, \mathbf{y}]$ is also in C.

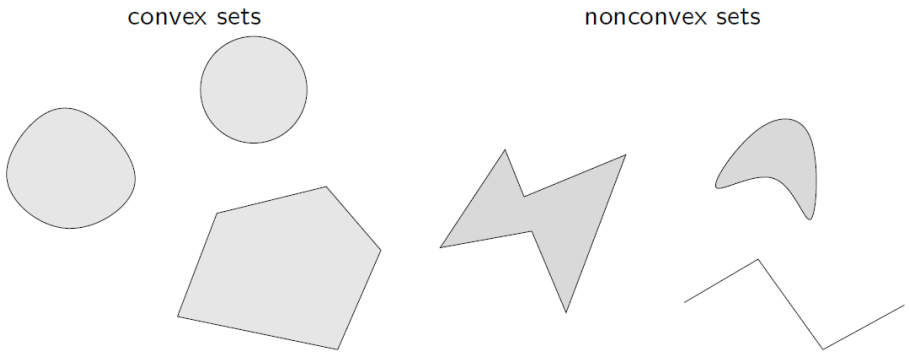


Figure 23: Examples of convex and nonconvex sets.

Very Important Convex Sets

- A line in $\mathbb{R}^n$ is a set of the form

$$L = \{\mathbf{z} + t\mathbf{d} : t \in \mathbb{R}\}.$$

  where $\mathbf{z}, \mathbf{d} \in \mathbb{R}^n$ and $\mathbf{d} \neq \mathbf{0}$.
- $[\mathbf{x}, \mathbf{y}]$, $(\mathbf{x}, \mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n (\mathbf{x} \neq \mathbf{y})$, $\emptyset$, and $\mathbb{R}^n$.
- A hyperplane is a set of the form

$$H = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\right\} \quad (\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}, b \in \mathbb{R}).$$

- The associated half-space is the set

$$H^- = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} \leq b\right\}.$$

  Both hyperplanes and half-spaces are convex sets.

- The open ball

$$B(\mathbf{c}, r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{c}\| < r\} ,$$

and the closed ball

$$B[\mathbf{c}, r] = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x} - \mathbf{c}\| \leq r\} ,$$

are convex. Note that the norm is an arbitrary norm defined over $\mathbb{R}^n$.

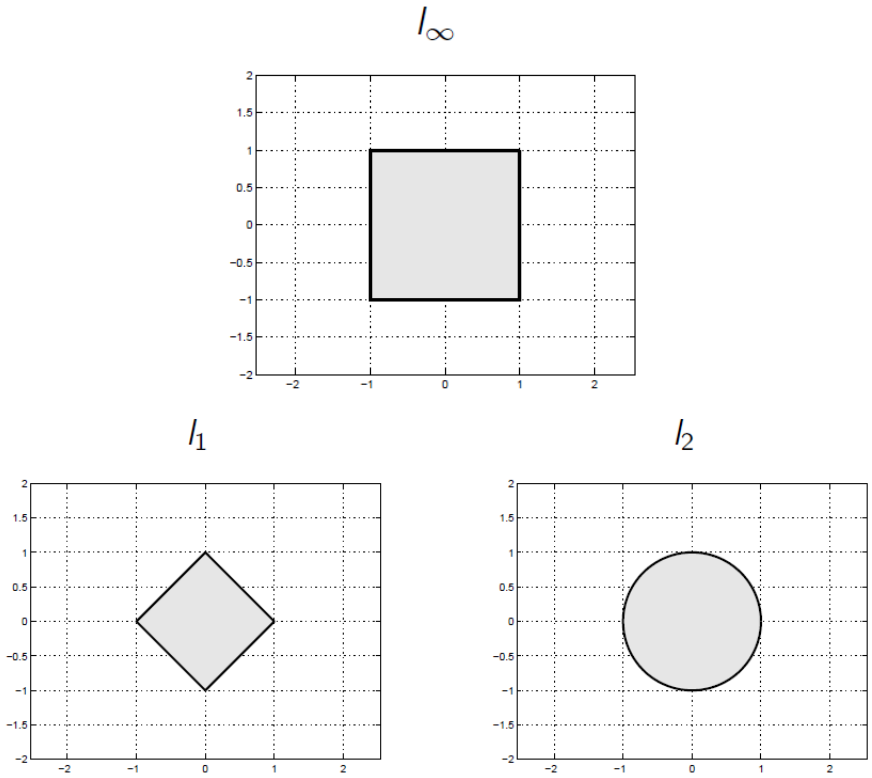


Figure 24: Different unit balls $\|\mathbf{x}\|_p \leq 1$ in $\mathbb{R}^2$.

A relevant result in optimization is the convexity of ellipsoids. An ellipsoid is a set of the form

$$E = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}^\top \mathbf{Q}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c \leq 0\right\} ,$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is positive semidefinite, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$.

Lemma. *E is convex.*

Proof. Write E as $E = \{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \leq 0\}$ where $f(\mathbf{x}) \equiv \mathbf{x}^\top \mathbf{Q}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$. Then, take $\mathbf{x}, \mathbf{y} \in E$ and $\lambda \in [0, 1]$, and $f(\mathbf{x}) \leq 0, f(\mathbf{y}) \leq 0$. The vector $z = \lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$ satisfies

$$\mathbf{z}^\top \mathbf{Q}\mathbf{z} = \lambda^2 \mathbf{x}^\top \mathbf{Q}\mathbf{x} + (1 - \lambda)^2 \mathbf{y}^\top \mathbf{Q}\mathbf{y} + 2\lambda(1 - \lambda)\mathbf{x}^\top \mathbf{Q}\mathbf{y} ,$$

and using Cauchy-Schwartz it follows that

$$\mathbf{x}^\top \mathbf{Q}\mathbf{y} \leq \left\|\mathbf{Q}^{1/2}\mathbf{x}\right\| \cdot \left\|\mathbf{Q}^{1/2}\mathbf{y}\right\| = \sqrt{\mathbf{x}^\top \mathbf{Q}\mathbf{x}}\sqrt{\mathbf{y}^\top \mathbf{Q}\mathbf{y}}$$
$$\leq \frac{1}{2}\left(\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{y}^\top \mathbf{Q}\mathbf{y}\right),$$

which implies that

$$\mathbf{z}^\top \mathbf{Q}\mathbf{z} \leq \lambda \mathbf{x}^\top \mathbf{Q}\mathbf{x} + (1-\lambda)\mathbf{y}^\top \mathbf{Q}\mathbf{y}.$$

Finally,

$$\begin{aligned} f(\mathbf{z}) &= \mathbf{z}^\top \mathbf{Q}\mathbf{z} + 2\mathbf{b}^\top \mathbf{z} + c \\ &\leq \lambda \mathbf{x}^\top \mathbf{Q}\mathbf{x} + (1-\lambda)\mathbf{y}^\top \mathbf{Q}\mathbf{y} + 2\lambda \mathbf{b}^\top \mathbf{x} + 2(1-\lambda)\mathbf{b}^\top \mathbf{y} + c \\ &= \lambda\left(\mathbf{x}^\top \mathbf{Q}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c\right) + (1-\lambda)\left(\mathbf{y}^\top \mathbf{Q}\mathbf{y} + 2\mathbf{b}^\top \mathbf{y} + c\right) \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \leq 0 \end{aligned}$$

establishing the desired result that $\mathbf{z} \in E$. □

Algebraic Operations with Convex Sets

Lemma (Intersection of convex sets is convex). *Let $C_i \subseteq \mathbb{R}^n$ be a convex set for any $i \in I$ where I is an index set (possibly infinite). Then the set $\bigcap_{i\in I} C_i$ is convex.*

A direct consequence of the above is that convex polytopes of the form

$$P = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \leq \mathbf{b}\},$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$ are convex since they are generated as the intersection of m half-spaces $\mathbf{a}_i^\top \mathbf{x} \leq b_i$.

Some important algebraic properties of convex sets are summarized in the following result:

Theorem.

1. *Let $C_1, C_2, \ldots, C_k \subseteq \mathbb{R}^n$ be convex sets and let $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}$. Then the set $\mu_1 C_1 + \mu_2 C_2 + \ldots + \mu_k C_k$ is convex.*
2. *Let $C_i \subseteq \mathbb{R}^{k_i}$, $i = 1, \ldots, m$ be convex sets. Then the cartesian product*

$$C_1 \times C_2 \times \cdots \times C_m = \{(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) : \mathbf{x}_i \in C_i, i = 1, 2, \ldots, m\}$$

is convex.

3. *Let $M \subseteq \mathbb{R}^n$ be a convex set and let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then the set*

$$\mathbf{A}(M) = \{\mathbf{A}\mathbf{x} : \mathbf{x} \in M\}$$

is convex.

4. *Let $D \subseteq \mathbb{R}^m$ be convex and let $\mathbf{A} \in \mathbb{R}^{m\times n}$. Then the set*

$$\mathbf{A}^{-1}(D) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} \in D\}$$

is convex.

The Convex Hull

Definition (Convex Combinations). *Given m points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in \mathbb{R}^n$, a convex combination of these m points is a vector of the form $\lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \cdots + \ldots + \lambda_m\mathbf{x}_m$, where $\lambda_1, \lambda_2, \ldots, \lambda_m$ are nonnegative numbers satisfying $\lambda_1 + \lambda_2 + \ldots + \lambda_m = 1$.*

A convex set is defined by the property that any convex combination of two points from the set is also in the set. We will now show that a convex combination of any number of points from a convex set is in the set.

Theorem. *Let $C \subseteq \mathbb{R}^n$ be a convex set and let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in C$. Then for any $\boldsymbol{\lambda} \in \Delta_m$, the relation $\sum_{i=1}^m \lambda_i\mathbf{x}_i \in C$ holds.*

Proof. Proof by induction on m. For $m = 1$ the result is obvious. The induction hypothesis is that for any m vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m \in C$ and any $\boldsymbol{\lambda} \in \Delta_m$, the vector $\sum_{i=1}^m \lambda_i\mathbf{x}_i$ belongs to C. We will now prove the theorem for $m+1$ vectors. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{m+1} \in C$ and that $\boldsymbol{\lambda} \in \Delta_{m+1}$. We will show that $\mathbf{z} \equiv \sum_{i=1}^{m+1} \lambda_i\mathbf{x}_i \in C$. For this, if $\lambda_{m+1} = 1$, then $\mathbf{z} = \mathbf{x}_{m+1} \in C$ and the result obviously follows. Otherwise, if $\lambda_{m+1} < 1$, then

$$\begin{aligned}\mathbf{z} &= \sum_{i=1}^{m} \lambda_i\mathbf{x}_i + \lambda_{m+1}\mathbf{x}_{m+1} \\ &= (1-\lambda_{m+1}) \sum_{i=1}^{m} \frac{\lambda_i}{1-\lambda_{m+1}}\mathbf{x}_i + \lambda_{m+1}\mathbf{x}_{m+1}\,.\end{aligned}$$

Since $\sum_{i=1}^m \frac{\lambda_i}{1-\lambda_{m+1}} = \frac{1-\lambda_{m+1}}{1-\lambda_{m+1}} = 1$, it follows that $\mathbf{v} = \sum_{i=1}^m \frac{\lambda_i}{1-\lambda_{m+1}}\mathbf{x}_i$ is a convex combination of m points from C, and hence by the induction hypotheses we have that $\mathbf{v} \in C$. Thus, by the definition of a convex set, $\mathbf{z} = (1-\lambda_{m+1})\,\mathbf{v} + \lambda_{m+1}\mathbf{x}_{m+1} \in C$. □

Definition (The Convex Hull). *Let $S \subseteq \mathbb{R}^n$. The convex hull of S, denoted by $\mathrm{conv}(S)$, is the set comprising all the convex combinations of vectors from S:*

$$\mathrm{conv}(S) \equiv \left\{ \sum_{i=1}^{k} \lambda_i\mathbf{x}_i : \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in S, \boldsymbol{\lambda} \in \Delta_k \right\}$$

The convex hull $\mathrm{conv}(S)$ is "smallest" convex set containing S, in the sense that if another convex set T contains S, then $\mathrm{conv}(S) \subset T$.

The following well-known result, called the Carethéodory theorem, states that any element in the convex hull of a subset of a given set $S \subset \mathbb{R}^n$ can be expressed as a **convex combination** of no more than $n+1$ vectors from S.

Theorem (Carathéodory). *Let $S \subseteq \mathbb{R}^n$ and let $\mathbf{x} \in \mathrm{conv}(S)$. Then, there exist $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1} \in S$ such that $\mathbf{x} \in \mathrm{conv}\left(\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n+1}\}\right)$, that is, there exist $\lambda \in \Delta_{n+1}$ such that*

$$\mathbf{x} = \sum_{i=1}^{n+1} \lambda_i\mathbf{x}_i$$

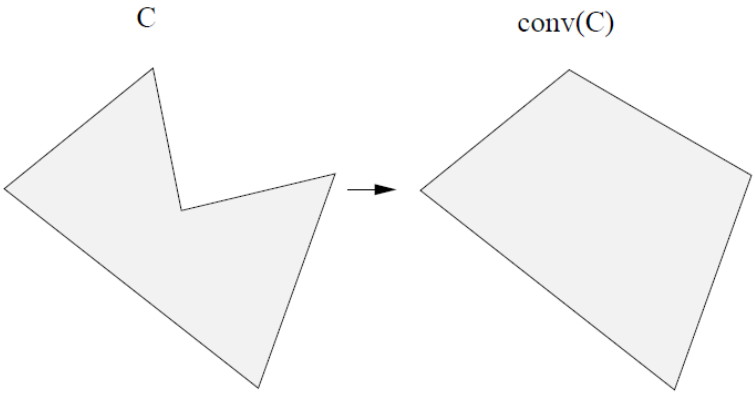


Figure 25: The convex hull of a non-convex set.

We present this proof as it provides a construction mechanism.

Proof. Let $\mathbf{x} \in \text{conv}(S)$. Then, there exist vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in S$ and $\boldsymbol{\lambda} \in \Delta_k$ such that

$$\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \,.$$

We can assume that $\lambda_i > 0$ for all $i = 1, 2, \ldots, k$. If $k \leq n+1$, the result is proven. Otherwise, if $k \geq n+2$, then the vectors $\mathbf{x}_2 - \mathbf{x}_1, \mathbf{x}_3 - \mathbf{x}_1, \ldots, \mathbf{x}_k - \mathbf{x}_1$, being more than n vectors in $\mathbb{R}^n$, are necessarily linearly dependent implying that there exist $\mu_2, \mu_3, \ldots, \mu_k$ not all zeros such that

$$\sum_{i=2}^{k} \mu_i \left(\mathbf{x}_i - \mathbf{x}_1\right) = \mathbf{0} \,.$$

Defining $\mu_1 = -\sum_{i=2}^{k} \mu_i$, we obtain that

$$\sum_{i=1}^{k} \mu_i \mathbf{x}_i = \mathbf{0} \,.$$

Note that not all of the coefficients $\mu_1, \mu_2, \ldots, \mu_k$ are zeros and $\sum_{i=1}^{k} \mu_i = 0$. There exists an index i for which $\mu_i < 0$. Let $\alpha \in \mathbb{R}_+$. Then,

$$\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i + \alpha \sum_{i=1}^{k} \mu_i \mathbf{x}_i = \sum_{i=1}^{k} \left(\lambda_i + \alpha \mu_i\right) \mathbf{x}_i \,.$$

We have $\sum_{i=1}^{k} \left(\lambda_i + \alpha \mu_i\right) = 1$, so the equation above is a convex combination representation if and only if

$$\lambda_i + \alpha \mu_i \geq 0 \text{ for all } i = 1, \ldots, k$$

but since $\lambda_i > 0$ for all i, it follows that these inequalities are satisfied for all $\alpha \in [0, \varepsilon]$ where $\varepsilon = \min_{i:\mu_i<0} \left\{-\frac{\lambda_i}{\mu_i}\right\}$. If we substitute $\alpha = \varepsilon$, then the inequalities still hold, but $\lambda_j + \varepsilon\mu_j = 0$ for $j \in \underset{i:\mu_i<0}{\operatorname{argmin}} \left\{-\frac{\lambda_i}{\mu_i}\right\}$. This means that we found a representation of $\mathbf{x}$ as a convex combination of $k-1$ (or less) vectors. This process can be carried on until a representation of $\mathbf{x}$ as a convex combination of no more than $n+1$ vectors is derived. □

Example For $n = 2$, consider the four vectors

$$\mathbf{x}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{x}_2 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \mathbf{x}_3 = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \mathbf{x}_4 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

and let $\mathbf{x} \in \operatorname{conv}(\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\})$ be given by

$$\mathbf{x} = \frac{1}{8}\mathbf{x}_1 + \frac{1}{4}\mathbf{x}_2 + \frac{1}{2}\mathbf{x}_3 + \frac{1}{8}\mathbf{x}_4 = \begin{pmatrix} \frac{13}{8} \\ \frac{11}{8} \end{pmatrix}$$

Find a representation of $\mathbf{x}$ as a convex combination of no more than 3 vectors.

Definition (Extreme Point). *Let $S \subseteq \mathbb{R}^n$ be a convex set. A point $\mathbf{x} \in S$ is called an extreme point of S if there do not exist $\mathbf{x}_1, \mathbf{x}_2 \in S$ $(\mathbf{x}_1 \neq \mathbf{x}_2)$ and $\lambda \in (0,1)$, such that $\mathbf{x} = \lambda\mathbf{x}_1 + (1-\lambda)\mathbf{x}_2$. The set of extreme point is denoted by* $\operatorname{ext}(S)$.

For example, the set of extreme points of a convex polytope consists of all its vertices.

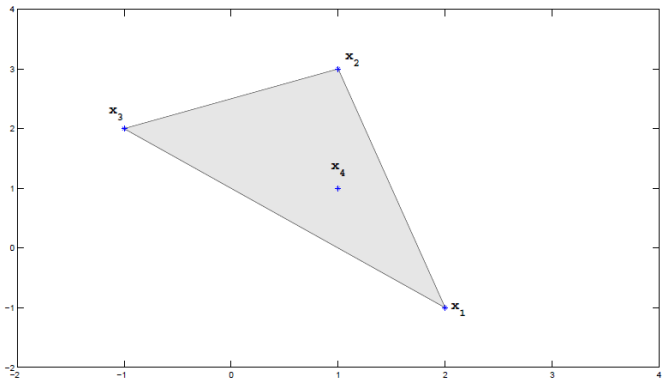


Figure 26: The extreme points of this triangle are given by its vertices $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$. The point $\mathbf{x}_4$ is not an extreme point.

Theorem (The Krein-Milman Theorem). *Let $S \subseteq \mathbb{R}^n$ be a compact convex set. Then*

$$S = \operatorname{conv}(\operatorname{ext}(S)) .$$

Convex Functions

We begin by giving a definition of convex function.

Definition (Convex Function). *A function $f : C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is called convex (or convex over C) if*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \text{ for any } \mathbf{x}, \mathbf{y} \in C, \lambda \in [0,1]$$

Convexity, Strict Convexity and Concavity

In case where no domain is specified, we naturally assume that f is defined over the entire space $\mathbb{R}^n$.

Definition (Strict Convexity). *A function $f : C \to \mathbb{R}$ defined on a convex set $C \subseteq \mathbb{R}^n$ is called strictly convex if*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \text{ for any } \mathbf{x} \neq \mathbf{y} \in C, \lambda \in (0,1)$$

Definition (Concavity). *A function is called concave if $-f$ is convex. Similarly, f is called strictly concave if $-f$ is strictly convex. We can also define concavity directly: a function f is concave if and only if for any $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in [0,1]$*

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$$

Examples of Convex Functions

- Affine Functions. $f(\mathbf{x}) = a^\top \mathbf{x} + b$, where $a \in \mathbb{R}^n$ and $b \in \mathbb{R}$. Take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} f(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) &= \mathbf{a}^\top (\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) + b \\ &= \lambda \left(\mathbf{a}^\top \mathbf{x}\right) + (1-\lambda) \left(\mathbf{a}^\top \mathbf{y}\right) + \lambda b + (1-\lambda) b \\ &= \lambda \left(\mathbf{a}^\top \mathbf{x} + b\right) + (1-\lambda) \left(\mathbf{a}^\top \mathbf{y} + b\right) \\ &= \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) \,. \end{aligned}$$

- Norms. $g(\mathbf{x}) = \|\mathbf{x}\|$, take $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then

$$\begin{aligned} g(\lambda \mathbf{x} + (1-\lambda)\mathbf{y}) &= \|\lambda \mathbf{x} + (1-\lambda)\mathbf{y}\| \\ &\leq \|\lambda \mathbf{x}\| + \|(1-\lambda)\mathbf{y}\| \\ &= \lambda \|\mathbf{x}\| + (1-\lambda)\|\mathbf{y}\| \\ &= \lambda g(\mathbf{x}) + (1-\lambda) g(\mathbf{y}) \,. \end{aligned}$$

We now state a fundamental result for convex functions.

Theorem (Jensen's Inequality). *Let $f : C \to \mathbb{R}$ be a convex function where $C \subseteq \mathbb{R}^n$ is a convex set. Then, for any $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in C$ and $\boldsymbol{\lambda} \in \Delta_k$, the following inequality holds:*

$$f\left(\sum_{i=1}^{k} \lambda_i \mathbf{x}_i\right) \leq \sum_{i=1}^{k} \lambda_i f\left(\mathbf{x}_i\right) .$$

An important set associated with convex functions is the *epigraph.*

Definition (epigraph). *Let $f : \mathbb{R}^n \to \mathbb{R}$. Then the epigraph set $\operatorname{epi}(f) \subseteq \mathbb{R}^{n+1}$ is defined by*

$$\operatorname{epi}(f) = \left\{ \begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} : f(\mathbf{x}) \leq t \right\} .$$

An example of an epigraph is shown in Figure 27.

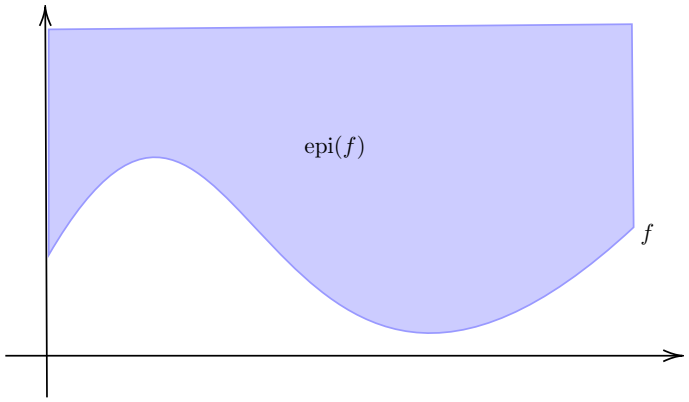


Figure 27: The epigraph of a one-dimensional function.

The definition of the epigraph allows us to connect the notions of convex sets and convex functions through the following theorem.

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}$. Then the function f is convex if and only if its epigraph is a convex set.*

First-order Characterization of Convex Functions

Convex functions are not necessarily differentiable (think about $f(x) = |x|$), but in case they are, we can generalize first and a second-order optimality conditions for functions such as those from Week 3.

Theorem (The Gradient Inequality). *Let $f : C \to \mathbb{R}$ be a continuously differentiable function defined on a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if*

$$f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) \text{ for any } \mathbf{x}, \mathbf{y} \in C .$$

An analogous result holds for strictly convex functions (with a strict inequality).

Proof. Suppose first that f is convex. Let $\mathbf{x}, \mathbf{y} \in C$ and $\lambda \in (0, 1]$. If $\mathbf{x} = \mathbf{y}$, then the inequality trivially holds. We will therefore assume that $\mathbf{x} \neq \mathbf{y}$. The

$$\frac{f(\mathbf{x} + \lambda(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\lambda} \leq f(\mathbf{y}) - f(\mathbf{x}) .$$

Taking $\lambda \to 0^+$, we obtain

$$f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) \leq f(\mathbf{y}) - f(\mathbf{x}) .$$

Since f is continuously differentiable, $f'(\mathbf{x}; \mathbf{y} - \mathbf{x}) = \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x})$, and the inequality follows.

To prove the reverse direction, assume that that the gradient inequality holds. Let $\mathbf{z}, \mathbf{w} \in C$, and let $\lambda \in (0, 1)$. We will show that $f(\lambda \mathbf{z} + (1 - \lambda)\mathbf{w}) \leq \lambda f(\mathbf{z}) + (1 - \lambda) f(\mathbf{w})$, Let $\mathbf{u} = \lambda \mathbf{z} + (1 - \lambda)\mathbf{w} \in C$. Then

$$\mathbf{z} - \mathbf{u} = \frac{\mathbf{u} - (1 - \lambda)\mathbf{w}}{\lambda} - \mathbf{u} = -\frac{1 - \lambda}{\lambda}(\mathbf{w} - \mathbf{u}) .$$

We have

$$f(\mathbf{u}) + \nabla f(\mathbf{u})^\top (\mathbf{z} - \mathbf{u}) \leq f(\mathbf{z}) ,$$
$$f(\mathbf{u}) - \frac{\lambda}{1 - \lambda} \nabla f(\mathbf{u})^\top (\mathbf{z} - \mathbf{u}) \leq f(\mathbf{w}) .$$

Thus,

$$f(\mathbf{u}) \leq \lambda f(\mathbf{z}) + (1 - \lambda) f(\mathbf{w})$$

□

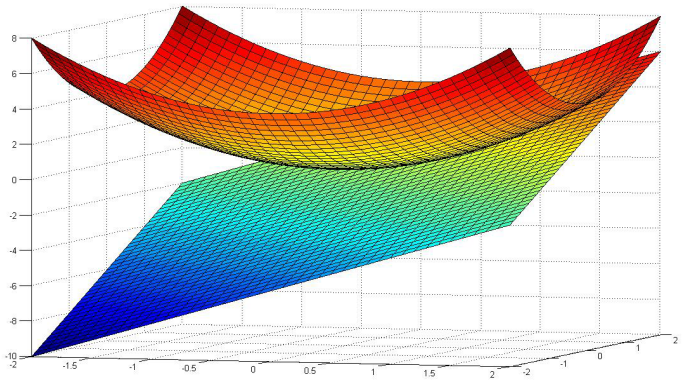


Figure 28: For a convex function f, the tangent plane at every point is always below f.

Convexity + Stationarity ⇒ Global Optimality! A direct result of the gradient inequality is that the first order optimality condition $\nabla f(\mathbf{x}^*) = 0$ is sufficient for global optimality.

Theorem (Stationarity Implies Global Optimality)**.** *Let f be a continuously differentiable function which is convex over a convex set $C \subseteq \mathbb{R}^n$. Suppose that $\nabla f(\mathbf{x}^*) = \mathbf{0}$ for some $\mathbf{x}^* \in C$. Then $\mathbf{x}^*$ is the global minimizer of f over C.*

We now revisit optimality conditions for quadratic functions.

Theorem (Convexity of Quadratic Functions with Positive Semidefinite Matrices)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be the quadratic function given by $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c$ where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is symmetric, $\mathbf{b} \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Then f is (strictly) convex if and only if $\mathbf{A} \succeq \mathbf{0} (\mathbf{A} \succ \mathbf{0})$.*

Proof. The convexity of f is equivalent to

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^n .$$

This is the same as stating that

$$\mathbf{y}^\top \mathbf{A}\mathbf{y} + 2\mathbf{b}^\top \mathbf{y} + c \geq \mathbf{x}^\top \mathbf{A}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} + c + 2(\mathbf{A}\mathbf{x} + \mathbf{b})^\top (\mathbf{y} - \mathbf{x}) ,$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, from where it follows that

$$(\mathbf{y} - \mathbf{x})^\top \mathbf{A}(\mathbf{y} - \mathbf{x}) \geq 0 ,$$

for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. This is equivalent to the inequality $\mathbf{d}^\top \mathbf{A}\mathbf{d} \geq 0$ for any $\mathbf{d} \in \mathbb{R}^n$, which is the same as $\mathbf{A} \succeq \mathbf{0}$. Similar arguments show that strict convexity is equivalent to

$$\mathbf{d}^\top \mathbf{A}\mathbf{d} > 0 \text{ for any } \mathbf{0} \neq \mathbf{d} \in \mathbb{R}^n ,$$

namely to $\mathbf{A} \succ \mathbf{0}$. □

Theorem (Monotonicity of the Gradient)**.** *Suppose that f is a continuously differentiable function over a convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if*

$$(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{x} - \mathbf{y}) \geq 0 \text{ for any } \mathbf{x}, \mathbf{y} \in C .$$

Second-order Characterization of Convex Functions

We can now extend our link between convexity and optimality conditions to second-order characterizations.

Theorem (Second-Order Characterization of Convexity)**.** *Let f be a twice continuously differentiable function over an open convex set $C \subseteq \mathbb{R}^n$. Then f is convex over C if and only if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ for any $\mathbf{x} \in C$.*

Example. Convexity of the log-sum-exp function:

$$f(\mathbf{x}) = \log\left(e^{x_1} + e^{x_2} + \ldots + e^{x_n}\right), \quad \mathbf{x} \in \mathbb{R}^n .$$

The gradient is given by:

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, \quad i = 1, 2, \ldots, n .$$

Therefore, the Hessian is computed as

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \begin{cases} -\frac{e^{x_i} e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2}, & i \neq j \\ -\frac{e^{x_i} e^{x_j}}{\left(\sum_{j=1}^n e^{x_j}\right)^2} + \frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}, & i = j \end{cases} .$$

We can thus write the Hessian matrix as

$$\nabla^2 f(\mathbf{x}) = \operatorname{diag}(\mathbf{w}) - \mathbf{w}\mathbf{w}^\top, \qquad \text{with} \qquad \mathbf{w} = \left(\frac{e^{x_i}}{\sum_{j=1}^n e^{x_j}}\right)_{i=1}^n \in \Delta_n .$$

For any $\mathbf{v} \in \mathbb{R}^n$:

$$\mathbf{v}^\top \nabla^2 f(\mathbf{x}) \mathbf{v} = \sum_{i=1}^n w_i v_i^2 - \left(\mathbf{v}^\top \mathbf{w}\right)^2 \geq 0 ,$$

since defining $s_i = \sqrt{w_i} v_i$, $t_i = \sqrt{w_i}$, we have

$$\left(\mathbf{v}^\top \mathbf{w}\right)^2 = \left(\mathbf{s}^\top \mathbf{t}\right)^2 \leq \|\mathbf{s}\|^2 \|\mathbf{t}\|^2 = \left(\sum_{i=1}^n w_i v_i^2\right)\left(\sum_{i=1}^n w_i\right) = \sum_{i=1}^n w_i v_i^2 .$$

Thus, $\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$ and hence f is convex over $\mathbb{R}^n$.

Example Show the convexity of the quad-over-lin function

$$f(x_1, x_2) = \frac{x_1^2}{x_2}$$

defined over $\mathbb{R} \times \mathbb{R}_{++} = \{(x_1, x_2) : x_2 > 0\}$.

Further Results for Convex Functions

Operations Preserving Convexity

- Let f be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$ and let $\alpha \geq 0$. Then αf is a convex function over C.
- Let $f_1, f_2, \ldots, f_p$ be convex functions over a convex set $C \subseteq \mathbb{R}^n$. Then the sum function $f_1 + f_2 + \ldots + f_p$ is convex over C.

- Let f be a convex function defined on a convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{b} \in \mathbb{R}^n$. Then the function g defined by

  $$g(\mathbf{y}) = f(\mathbf{A}\mathbf{y} + \mathbf{b})$$

  is convex over the convex set $D = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{A}\mathbf{y} + \mathbf{b} \in C\}$.

- Let $f : C \to \mathbb{R}$ be a convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Let $g : I \to \mathbb{R}$ be a one-dimensional nondecreasing convex function over the interval $I \subseteq \mathbb{R}$. Assume that the image of C under f is contained in $I : f(C) \subseteq I$. Then the composition of g with f defined by

  $$h(\mathbf{x}) \equiv g(f(\mathbf{x}))$$

  is convex over C.

Several Examples of Convex Functions Using these Properties

- The generalized quad-over-lin function

  $$g(\mathbf{x}) = \frac{\|\mathbf{A}\mathbf{x} + \mathbf{b}\|^2}{\mathbf{c}^\top \mathbf{x} + d} \quad \left(\mathbf{A} \in \mathbb{R}^{m \times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{c} \in \mathbb{R}^n, d \in \mathbb{R}\right)$$

  is convex over $D = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{c}^\top \mathbf{x} + d > 0\}$.

- $f(x_1, x_2) = -\log(x_1 x_2)$, over $\mathbb{R}^2_{++}$.
- $f(x_1, x_2) = x_1^2 + 2x_1 x_2 + 3x_2^2 + 2x_1 - 3x_2 + e^{x_1}$.
- $h(\mathbf{x}) = e^{\|\mathbf{x}\|^2}$.

Theorem (Point-Wise Maximum of Convex Functions). *Let $f_1, f_2, \ldots, f_p : C \to \mathbb{R}$ be p convex functions over the convex set $C \subseteq \mathbb{R}^n$. Then the maximum function*

$$f(\mathbf{x}) \equiv \max_{i=1,2,\ldots,p} \{f_i(\mathbf{x})\}$$

is convex over C.

Examples

- $f(\mathbf{x}) = \max\{x_1, x_2, \ldots, x_n\}$ is convex.
- For a given vector $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top \in \mathbb{R}^n$, let $x_{[i]}$ denote the i -th largest value in $\mathbf{x}$. For any $k \in \{1, 2, \ldots, n\}$ the function

  $$h_k(\mathbf{x}) = x_{[1]} + x_{[2]} + \ldots + x_{[k]}$$

  is convex.

Theorem (Preservation of Convexity Under Partial Minimization). *Let $f : C \times D \to \mathbb{R}$ be a convex function defined over the set $C \times D$ where $C \subseteq \mathbb{R}^m$ and $D \subseteq \mathbb{R}^n$ are convex sets. Let*

$$g(\mathbf{x}) = \min_{\mathbf{y} \in D} f(\mathbf{x}, \mathbf{y}), \quad \mathbf{x} \in C$$

where we assume that the minimum is finite. Then g is convex over C.

Example. The distance function from a convex set $d_C(\mathbf{x}) \equiv \inf_{\mathbf{y}\in C} \|\mathbf{x}-\mathbf{y}\|$ is convex.

Level Sets of Convex Functions

Let us start with the definition of a *level set.*

Definition (level sets). *Let $f: S \to \mathbb{R}$ be a functions defined over a set $S \subseteq \mathbb{R}^n$. Then the level set of f with level α is given by*

$$\text{Lev}(f,\alpha) = \{\mathbf{x} \in S: f(\mathbf{x}) \leq \alpha\} \ .$$

An example of a level set of a one-dimensional function f at level α is shown in Figure 29.

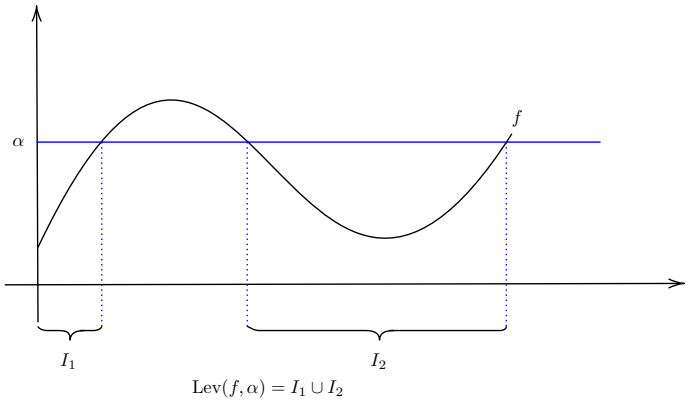


Figure 29: Example of a level set of a one-dimensional function f at level α.

In what follows, we present a fundamental property of convex functions with respect to their level sets.

Theorem (convexity of level sets of convex functions). *Let $f: C \to \mathbb{R}$ be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$. Then for any $\alpha \in \mathbb{R}$ the level set $\text{Lev}(f,\alpha)$ is convex.*

The previous theorem states that all convex functions have all their levels sets convex. However, the opposite is not true. In fact, there exists nonconvex functions whose level sets are all convex: for example, consider the function $f(\mathbf{x}) = \sqrt{|\mathbf{x}|}$.

Four Important Theorems for Convex Functions

We now state four important results for convex functions.

Theorem (Continuity of Convex Functions). *Let $f: C \to \mathbb{R}$ be a convex function defined over a convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x}_0 \in \text{int}(C)$. Then there exist $\varepsilon > 0$ and $L > 0$ such that $B[\mathbf{x}_0, \varepsilon] \subseteq C$ and*

$$|f(\mathbf{x}) - f(\mathbf{x}_0)| \leq L\|\mathbf{x}-\mathbf{x}_0\| \ \text{ for any } \mathbf{x} \in B[\mathbf{x}_0, \varepsilon]$$

Theorem (Existence of Directional Derivatives of Convex Functions). *Let $f : C \to \mathbb{R}$ be a convex function over the convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x} \in \text{int}(C)$. Then for any $\mathbf{d} \neq 0$, the directional derivative $f'(\mathbf{x}; \mathbf{d})$ exists.*

The last two theorems relate to the problem of **maximizing** a non-constant convex function over a convex set.

Theorem (No Maximum Inside the Convex Set). *Let $f : C \to \mathbb{R}$ be convex and non-constant over the nonempty convex set $C \subseteq \mathbb{R}^n$. Then f does not attain a maximum at a point in $int(C)$.*

Finally, we state that maximum of convex function over compact convex sets can be found at the extreme points of the set.

Theorem (Maximum of a Convex Function Over a Compact Convex Set). *Let $f : C \to \mathbb{R}$ be convex over the nonempty convex and compact set $C \subseteq \mathbb{R}^n$. Then there exists at least one maximizer of f over C that is an extreme point of C.*

Proof. Let $\mathbf{x}^*$ be a maximizer of f over C. If $\mathbf{x}^*$ is an extreme point of C, then the result is established. Otherwise, by Krein-Milman, $C = \text{conv}(\text{ext}(C))$ implies the existence of $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k \in \text{ext}(C)$ such that

$$\mathbf{x}^* = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i$$

By the convexity of f,

$$f(\mathbf{x}^*) \leq \sum_{i=1}^{k} \lambda_i f(\mathbf{x}_i)$$

or equivalently

$$\sum_{i=1}^{k} \lambda_i \left(f(\mathbf{x}_i) - f(\mathbf{x}^*) \right) \geq 0 \,.$$

Since $\mathbf{x}^*$ is a maximizer of f over C, we have $f(\mathbf{x}_i) \leq f(\mathbf{x}^*)$ for all $i = 1, \ldots, k$. This implies that $f(\mathbf{x}_i) = f(\mathbf{x}^*)$. Consequently, the extreme points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are all maximizers of f over C. □

Part VII.
Convex Optimization

Convex Optimization Problems

A convex optimization problem (or just a convex problem) is a problem consisting of minimizing a convex function $f(\mathbf{x})$ over a convex set C:

$$\begin{array}{ll} \min & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{x} \in C \end{array} \tag{CVX}$$

A functional form of a convex problem can written as

$$\begin{array}{lll} \min & f(\mathbf{x}) & \\ \text{s.t.} & g_i(\mathbf{x}) \leq 0, & i = 1, 2, \ldots, m \\ & h_j(\mathbf{x}) = 0, & j = 1, 2, \ldots, p\,, \end{array}$$

where $f, g_1, \ldots, g_m : \mathbb{R}^n \to \mathbb{R}$ are convex functions, and $h_1, h_2, \ldots, h_p : \mathbb{R}^m \to \mathbb{R}$ are affine functions. The functional form does fit into the general formulation (CVX). A very important feature of convex optimization problems is that local minima are global minima!

Theorem (Local minima are global in CVX.)**.** *Let $f : C \to \mathbb{R}$ be a convex function defined on the convex set $C \subseteq \mathbb{R}^n$. Let $\mathbf{x}^* \in C$ be a local minimum of f over C. Then x^* is a global minimum of f over C.*

Proof. Assume $\mathbf{x}^*$ is a local minimum of f over C. This implies that there exists $r > 0$ such that $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ for any $\mathbf{x} \in C \cap B[\mathbf{x}^*, r]$. Let $\mathbf{x}^* \neq \mathbf{y} \in C$. We will show that $f(\mathbf{y}) \geq f(\mathbf{x}^*)$. Let $\lambda \in (0,1)$ be such that $\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*) \in B[\mathbf{x}^*, r]$. Since $\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*) \in B[\mathbf{x}^*, r]$, it follows that $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*))$ and hence by Jensen's inequality:

$$f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \lambda(\mathbf{y} - \mathbf{x}^*)) \leq (1 - \lambda) f(\mathbf{x}^*) + \lambda f(\mathbf{y})\,.$$

Thus, the desired inequality $f(\mathbf{x}^*) \leq f(\mathbf{y})$ follows. □

A small variation of the proof of the last theorem yields the following.

Theorem. *Let $f : C \to \mathbb{R}$ be a strictly convex function defined on the convex set C. Let $x^* \in C$ be a local minimum of f over C. Then x^* is a strict global minimum of f over C.*

Another important and easily deduced property of convex problems is that set of optimal solutions is also convex.

Theorem. *Let $f : C \to \mathbb{R}$ be a convex function defined over the convex set $C \subseteq \mathbb{R}^n$. Then the set of optimal solutions of the problem*

$$\min\{f(\mathbf{x}) : \mathbf{x} \in C\}$$

is convex. If, in addition, f is strictly convex over C, then there exists at most one optimal solution of the problem.

Examples:

- A Convex Problem:

$$\begin{array}{ll} \min & -2x_1 + x_2 \\ \text{s.t.} & x_1^2 + x_2^2 \leq 3 \end{array}$$

- A Nonconvex Problem:

$$\begin{array}{ll} \min & x_1^2 - x_2 \\ \text{s.t.} & x_1^2 + x_2^2 = 3 \end{array}$$

- Linear Programming

$$\begin{array}{lll} & \min & \mathbf{c}^\top \mathbf{x} \\ (\mathbf{LP}): & \text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{b} \\ & & \mathbf{B}\mathbf{x} = \mathbf{g} \end{array}$$

  (LP) is a convex optimization problem (constraints and objective function are linear / affine and hence convex). It is also equivalent to a problem of maximizing a convex (linear) function subject to a convex constraints set. Hence, if the feasible set is compact and nonempty, then there exists at least one optimal solution which is an extreme point, or equivalently, a basic feasible solution.

- Convex Quadratic Problems consist of minimizing a convex quadratic function subject to affine constraints. The general form is

$$\begin{array}{ll} \min & \mathbf{x}^\top \mathbf{Q}\mathbf{x} + 2\mathbf{b}^\top \mathbf{x} \\ \text{s.t.} & \mathbf{A}\mathbf{x} \leq \mathbf{c} \end{array}$$

  $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is positive semidefinite, $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{c} \in \mathbb{R}^m$.

Optimization over a Convex Set and Stationarity

We will consider the constrained optimization problem given by

$$\min_{\mathbf{x}} \quad \{f(\mathbf{x}) : \mathbf{x} \in C\}, \tag{P}$$

where C is a closed convex subset of $\mathbb{R}^n$, and f is continuously differentiable over C, not necessarily convex. To characterize optimality in the presence of convex constraints, we define the concept of stationarity.

Definition (Stationarity). *Let f be a continuously differentiable function over a closed and convex set C. Then $\mathbf{x}^*$ is called a stationary point of* (P) *if*

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0\,, \quad \text{for any } \mathbf{x} \in C\,.$$

Theorem (Stationarity as a Necessary Optimality Condition). *Let f be a continuously differentiable function over a nonempty closed convex set C, and let $\mathbf{x}^*$ be a local minimum of* (P). *Then $\mathbf{x}^*$ is a stationary point of* (P).

Proof. Let $\mathbf{x}^*$ be a local minimum of (P), and assume in contradiction that $\mathbf{x}^*$ is not a stationary point of (P). This implies that there exists $\mathbf{x} \in C$ such that

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) < 0\,.$$

Thus, $f'(\mathbf{x}^*; \mathbf{d}) < 0$, where $\mathbf{d} = \mathbf{x} - \mathbf{x}^*$. Therefore, there exists $\varepsilon \in (0,1)$ such that $f(\mathbf{x}^* + t\mathbf{d}) < f(\mathbf{x}^*)$, $\forall t \in (0, \varepsilon)$. Finally, since $\mathbf{x}^* + t\mathbf{d} = (1-t)\mathbf{x}^* + t\mathbf{x} \in C$, $\forall t \in (0, \varepsilon)$, we conclude that $\mathbf{x}^*$ is not a local optimum point of $(\mathbf{P})$. Contradiction. □

Examples of Stationarity Conditions

- For $C = \mathbb{R}^n$, $\mathbf{x}^*$ is a stationary point of (P) iff

  $$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n\,.$$

  We will show that the above condition is equivalent to $\nabla f(\mathbf{x}^*) = 0$, Indeed, if $\nabla f(\mathbf{x}^*) = 0$, then obviously the inequality is satisfied. In the other direction, suppose that the inequality holds. Plugging $\mathbf{x} = \mathbf{x}^* - \nabla f(\mathbf{x}^*)$ in the above implies

  $$-\|\nabla f(\mathbf{x}^*)\|^2 \geq 0\,.$$

  Thus, $\nabla f(\mathbf{x}^*) = 0$.

- For $C = \mathbb{R}^n_+$, $\mathbf{x}^*$ is a stationary point iif

  $$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0 \quad \forall \mathbf{x} \in \mathbb{R}^n_+\,.$$

  This is equivalent to say $\nabla f(\mathbf{x}^*)^\top \mathbf{x} - \nabla f(\mathbf{x}^*)^\top \mathbf{x}^* \geq 0$ for all $\mathbf{x} \geq 0$. Or equivalently, $\nabla f(\mathbf{x}^*) \geq 0$ and $\nabla f(\mathbf{x}^*)^\top \mathbf{x}^* \leq 0$. The latter is equivalent to claim that

  $$\nabla f(\mathbf{x}^*) \geq 0 \quad \text{and} \quad x_i^* \frac{\partial f}{\partial x_i}(\mathbf{x}^*) = 0, \quad i = 1, 2, \ldots, n\,,$$

  which we summarize as

  $$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) \begin{cases} = 0 & x_i^* > 0 \\ \geq 0 & x_i^* = 0 \end{cases}$$

Some important stationarity conditions are summarized in the table below:

Feasible Set	Explicit Stationarity Condition
$\mathbb{R}^n$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$
$\mathbb{R}^n_+$	$\frac{\partial f}{\partial x_i}(\mathbf{x}^*) \begin{cases} = 0 & x_i^* > 0 \\ \geq 0 & x_i^* = 0 \end{cases}$
$\{\mathbf{x} \in \mathbb{R}^n : \mathbf{e}^\top \mathbf{x} = 1\}$	$\frac{\partial f}{\partial x_1}(\mathbf{x}^*) = \ldots = \frac{\partial f}{\partial x_n}(\mathbf{x}^*)$
$B[\mathbf{0}, 1]$	$\nabla f(\mathbf{x}^*) = \mathbf{0}$ or $\|\mathbf{x}^*\| = 1$ and $\exists \lambda \leq 0 : \nabla f(\mathbf{x}^*) = \lambda \mathbf{x}^*$

For convex problems, stationarity is a necessary and sufficient condition.

Theorem (Stationarity in Convex Optimization)**.** *Let f be a continuously differentiable convex function over a nonempty closed and convex set $C \subseteq \mathbb{R}^n$. Then $\mathbf{x}^*$ is a stationary point of* (P) *iff $\mathbf{x}^*$ is an optimal solution of* (P)*.*

Proof. If $\mathbf{x}^*$ is an optimal solution of (P), then we already showed that it is a stationary point of (P). Assume that $\mathbf{x}^*$ is a stationary point of (P). Let $\mathbf{x} \in C$. Then

$$f(\mathbf{x}) \geq f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq f(\mathbf{x}^*)$$

establishing the optimality of $\mathbf{x}^*$. □

The Orthogonal Projection Operator

Definition (Orthogonal Projection)**.** *Given a nonempty closed convex set C, the orthogonal projection operator $P_C : \mathbb{R}^n \to C$ is defined by*

$$P_C(\mathbf{x}) = \operatorname{argmin} \left\{ \|\mathbf{y} - \mathbf{x}\|^2 : \mathbf{y} \in C \right\} .$$

We state two important results concerning orthogonal projections.

Theorem (The First Projection Theorem)**.** *Let $C \subseteq \mathbb{R}^n$ be a nonempty closed and convex set. Then for any $\mathbf{x} \in \mathbb{R}^n$, the orthogonal projection $P_C(\mathbf{x})$ exists and is unique.*

Theorem (The Second Projection Theorem)**.** *Let C be a nonempty closed convex set and let $\mathbf{x} \in \mathbb{R}^n$. Then $\mathbf{z} = P_C(\mathbf{x})$ if and only if*

$$(\mathbf{x} - \mathbf{z})^\top (\mathbf{y} - \mathbf{z}) \leq 0 , \quad \text{for any } \mathbf{y} \in C$$

Examples of Orthogonal Projections:

- For $C = \mathbb{R}^n_+$,

$$P_{\mathbb{R}^n_+}(\mathbf{x}) = [\mathbf{x}]_+$$

where $[\mathbf{v}]_+ = (\max\{v_1, 0\}, \max\{v_2, 0\}, \ldots, \max\{v_n, 0\})^\top$.

- A box is a subset of $\mathbb{R}^n$ of the form

$$B = [\ell_1, u_1] \times [\ell_2, u_2] \times \cdots \times [\ell_n, u_n] = \{\mathbf{x} \in \mathbb{R}^n : \ell_i \le x_i \le u_i\},$$

where $\ell_i \le u_i$ for all $i = 1, 2, \ldots, n$. For this set

$$[P_B(\mathbf{x})]_i = \begin{cases} u_i & x_i \ge u_i \\ x_i & \ell_i < x_i < u_i \\ \ell_i & x_i \le \ell_i \end{cases}$$

- For the closed ball in $\mathbb{R}^n$, $C = B[0, r]$, it holds

$$P_{B[0,r]} = \begin{cases} \mathbf{x} & \|\mathbf{x}\| \le r \\ r\frac{\mathbf{x}}{\|\mathbf{x}\|} & \|\mathbf{x}\| > r \end{cases}$$

A very important result in convex optimization is the representation of stationarity using the orthogonal projection operator.

Theorem (Representation of Stationarity via the Orthogonal Projection Operator)**.** *Let f be a continuously differentiable function over the nonempty closed convex set C, and let $s > 0$. Then $\mathbf{x}^*$ is a stationary point of (P) if and only if*

$$\mathbf{x}^* = P_C\left(\mathbf{x}^* - s\nabla f(\mathbf{x}^*)\right).$$

Proof. By the second projection theorem, $\mathbf{x}^* = P_C\left(\mathbf{x}^* - s\nabla f(\mathbf{x}^*)\right)$ iff

$$\left(\mathbf{x}^* - s\nabla f(\mathbf{x}^*) - \mathbf{x}^*\right)^\top (\mathbf{x} - \mathbf{x}^*) \le 0 \text{ for any } \mathbf{x} \in C,$$

which is equivalent to

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \ge 0 \text{ for any } \mathbf{x} \in C,$$

namely, the definition of stationarity. □

The Gradient Projection Method

It is convenient to define the gradient mapping as

$$G_L(\mathbf{x}) = L\left[\mathbf{x} - P_C\left(\mathbf{x} - \frac{1}{L}\nabla f(\mathbf{x})\right)\right],$$

where $L > 0$. In the unconstrained case $G_L(\mathbf{x}) = \nabla f(\mathbf{x})$. Otherwise, $G_L(\mathbf{x}) = \mathbf{0}$ if and only if $\mathbf{x}$ is a stationary point of (P). This means that we can consider $\|G_L(\mathbf{x})\|^2$ to be optimality measure. We use the orthogonal projection operator to introduce a gradient descent type algorithm for solving convex optimization problems.

Algorithm 7: The Gradient Projection Method

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in C$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Pick a stepsize t^k by a line search procedure.
2 Set $\mathbf{x}^{k+1} = P_C\left(\mathbf{x}^k - t^k \nabla f\left(\mathbf{x}^k\right)\right)$.
3 If $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

There are several strategies for choosing the stepsizes t^k. When $f \in C_L^{1,1}$, we can choose t^k to be constant and equal to $\frac{1}{L}$. An alternative is to include backtracking.

Algorithm 8: The Gradient Projection Method with Backtracking

Initialization: A tolerance parameter $\varepsilon > 0$ and $\mathbf{x}^0 \in C$. Parameters $s > 0$, $\alpha \in (0, 1)$, and $\beta \in (0, 1)$.
General Step: for any $k = 0, 1, 2, \ldots$ execute the following steps:

1 Pick $t^k = s$.
2 While $f(\mathbf{x}^k) - f(P_C(\mathbf{x}^k - t^k \nabla f(\mathbf{x}^k))) < \alpha t^k \left\|G_{\frac{1}{t^k}}(\mathbf{x}^k)\right\|^2$, set $t^k := \beta t^k$.
3 Set $\mathbf{x}^{k+1} = P_C\left(\mathbf{x}^k - t^k \nabla f\left(\mathbf{x}^k\right)\right)$.
4 If $\left\|\mathbf{x}^k - \mathbf{x}^{k+1}\right\| \leq \varepsilon$, then STOP and $\mathbf{x}^{k+1}$ is the output.

Theorem (Convergence of the Gradient Projection Method). *Let $\{\mathbf{x}^k\}$ be the sequence generated by the gradient projection method for solving problem (P) with either a constant stepsize $\bar{t} \in \left(0, \frac{2}{L}\right)$, where L is a Lipschitz constant of ∇f or a backtracking stepsize strategy. Assume that f is bounded below. Then:*

1. *The sequence $\left\{f\left(\mathbf{x}^k\right)\right\}$ is nonincreasing.*
2. *$G_d\left(\mathbf{x}^k\right) \to 0$ as $k \to \infty$, where*

$$d = \begin{cases} 1/\bar{t} & \text{constant stepsize} \\ 1/s & \text{backtracking.} \end{cases}$$

Part VIII.
Optimality Conditions

Separation Theorem

To begin our study of optimality conditions for linearly constrained problems we need first some technical results, known as *Alternative and Separation Theorems.*

A hyperplane

$$H = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\right\} \quad (\mathbf{a} \in \mathbb{R}^n \backslash \{\mathbf{0}\}, b \in \mathbb{R})$$

is said to strictly separate a point $\mathbf{y} \notin S$ from S if

$$\mathbf{a}^\top \mathbf{y} > b\,,$$

and

$$\mathbf{a}^\top \mathbf{x} \leq b \text{ for all } \mathbf{y} \in S\,.$$

Theorem (Separation of a Point from a Closed and Convex Set)**.** *Let $C \subseteq \mathbb{R}^n$ be a nonempty closed and convex set, and let $\mathbf{y} \notin C$. Then there exists $\mathbf{p} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and $\alpha \in \mathbb{R}$ such that $\mathbf{p}^\top \mathbf{y} > \alpha$ and $\mathbf{p}^\top \mathbf{x} \leq \alpha$ for all $\mathbf{x} \in C$.*

Proof. By the second orthogonal projection theorem, the vector $\overline{\mathbf{x}} = P_C(\mathbf{y}) \in C$ satisfies

$$(\mathbf{y} - \overline{\mathbf{x}})^\top (\mathbf{x} - \overline{\mathbf{x}}) \leq 0 \quad \text{for all } \mathbf{x} \in C\,,$$

which is the same as

$$(\mathbf{y} - \overline{\mathbf{x}})^\top \mathbf{x} \leq (\mathbf{y} - \overline{\mathbf{x}})^\top \overline{\mathbf{x}} \quad \text{for all } \mathbf{x} \in C\,.$$

Denote $\mathbf{p} = \mathbf{y} - \overline{\mathbf{x}} \neq \mathbf{0}$ and $\alpha = (\mathbf{y} - \overline{\mathbf{x}})^\top \overline{\mathbf{x}}$. Then,

$$\mathbf{p}^\top \mathbf{x} \leq \alpha \text{ for all } \mathbf{x} \in C\,.$$

On the other hand, we have

$$\mathbf{p}^\top \mathbf{y} = (\mathbf{y} - \overline{\mathbf{x}})^\top \mathbf{y} = (\mathbf{y} - \overline{\mathbf{x}})^\top (\mathbf{y} - \overline{\mathbf{x}}) + (\mathbf{y} - \overline{\mathbf{x}})^\top \overline{\mathbf{x}} = \|\mathbf{y} - \overline{\mathbf{x}}\|^2 + \alpha > \alpha\,.$$

□

Lemma (Farkas' Lemma - an Alternative Theorem)**.** *Let $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{m \times n}$. Then exactly one of the following systems has a solution:*

I. $\mathbf{A}\mathbf{x} \leq \mathbf{0}, \mathbf{c}^\top \mathbf{x} > 0$.

II. $\mathbf{A}^\top \mathbf{y} = \mathbf{c}, \mathbf{y} \geq 0$.

An alternative formulation is the following:

Lemma (Farkas' Lemma - Second Formulation)**.** *Let* $\mathbf{c} \in \mathbb{R}^n$ *and* $\mathbf{A} \in \mathbb{R}^{m \times n}$*. Then the following two claims are equivalent:*

(**A**) *The implication* $\mathbf{Ax} \leq \mathbf{0} \Rightarrow \mathbf{c}^\top \mathbf{x} \leq 0$ *holds true.*

(**B**) *There exists* $\mathbf{y} \in \mathbb{R}^m_+$ *such that* $\mathbf{A}^\top \mathbf{y} = \mathbf{c}$.

Proof. Suppose that system (**B**) is feasible. Then, there exists $\mathbf{y} \in \mathbb{R}^m_+$ such that $\mathbf{A}^\top \mathbf{y} = \mathbf{c}$. To see that the implication (**A**) holds, suppose that $\mathbf{Ax} \leq 0$ for some $\mathbf{x} \in \mathbb{R}^n$. Multiplying this inequality from the left by $\mathbf{y}^\top$ we obtain:

$$\mathbf{y}^\top \mathbf{Ax} \leq 0\,,$$

and hence,

$$\mathbf{c}^\top \mathbf{x} \leq 0\,.$$

Now, suppose that the implication (**A**) is satisfied, and let us show that the system (**B**) is feasible. Suppose in contradiction that system (**B**) is infeasible. Consider the following closed and convex set

$$S = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x} = \mathbf{A}^\top \mathbf{y} \text{ for some } \mathbf{y} \in \mathbb{R}^m_+\right\}\,.$$

Note that $\mathbf{c} \notin S$. By the separation theorem, there exists $\mathbf{p} \in \mathbb{R}^n \backslash \{\mathbf{0}\}$ and $\alpha \in \mathbb{R}$ such that $\mathbf{p}^\top \mathbf{c} > \alpha$ and

$$\mathbf{p}^\top \mathbf{x} \leq \alpha \text{ for all } \mathbf{x} \in S\,.$$

$\mathbf{0} \in S$ implies that $\alpha \geq 0 \Rightarrow \mathbf{p}^\top \mathbf{c} > 0$, and the inequality above is equivalent to

$$\mathbf{p}^\top \mathbf{A}^\top \mathbf{y} \leq \alpha \quad \text{for all } \mathbf{y} \geq \mathbf{0}$$

or to

$$(\mathbf{Ap})^\top \mathbf{y} \leq \alpha \quad \text{for all } \mathbf{y} \geq \mathbf{0}\,.$$

Therefore, $\mathbf{Ap} \leq \mathbf{0}$, which is a contradiction to the assertion that implication (**A**) holds. □

Example. What does this mean for $\mathbf{A} = \begin{pmatrix} 1 & 5 \\ -1 & 2 \end{pmatrix}$, $\mathbf{c} = \begin{pmatrix} -1 \\ 9 \end{pmatrix}$?

Theorem (Gordan's Alternative Theorem)**.** *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$*. Then exactly one of the following two systems has a solution:*

I. $\mathbf{Ax} < \mathbf{0}$.

II. $\mathbf{p} \neq \mathbf{0}, \mathbf{A}^\top \mathbf{p} = \mathbf{0}, \mathbf{p} \geq \mathbf{0}$.

KKT Conditions for Linearly Constrained Problems

The Karush-Kuhn-Tucker (or KKT) conditions are a set of fundamental charaterizations of the solution of convex optimization problems. Here we study some variants for the linearly constrained case.

Theorem (KKT conditions for Linearly Constrained Problems - Necessary Optimality Conditions)**.** *Consider the minimization problem*

$$\begin{cases} & \min f(\mathbf{x}) \\ \textit{subject to} & \mathbf{a}_i^\top \mathbf{x} \leq b_i, \quad i = 1, 2, \ldots, m \end{cases} \tag{LCP}$$

where f is continuously differentiable over $\mathbb{R}^n$, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m \ \mathbb{R}^n, b_1, b_2, \ldots, b_m \in \mathbb{R}$ and let $\mathbf{x}^*$ be a local minimum point of (LCP)*. Then, there exist $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \mathbf{a}_i = \mathbf{0},$$

and

$$\lambda_i \left(\mathbf{a}_i^\top \mathbf{x}^* - b_i \right) = 0, \quad i = 1, 2, \ldots, m.$$

Proof. If $\mathbf{x}^*$ is a local minimum, this implies $\mathbf{x}^*$ is a stationary point, meaning

$$\nabla f(\mathbf{x}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq 0$$

for every $\mathbf{x} \in \mathbb{R}^n$ satisfying $\mathbf{a}_i^\top \mathbf{x} \leq b_i$ for any $i = 1, 2, \ldots, m$. Now, denote the set of active constraints by

$$I(\mathbf{x}^*) = \left\{ i : \mathbf{a}_i^\top \mathbf{x}^* = b_i \right\}.$$

Making the change of variables $\mathbf{y} = \mathbf{x} - \mathbf{x}^*$, we have $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$ for any $\mathbf{y} \in \mathbb{R}^m$ satisfying

$$\mathbf{a}_i^\top (\mathbf{y} + \mathbf{x}^*) \leq b_i, \quad i = 1, 2, \ldots, m.$$

Or equivalently $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$ for any $\mathbf{y}$ satisfying

$$\begin{array}{ll} \mathbf{a}_i^\top \mathbf{y} \leq 0 & i \in I(\mathbf{x}^*) \\ \mathbf{a}_i^\top \mathbf{y} \leq b_i - \mathbf{a}_i^\top \mathbf{x}^* & i \notin I(\mathbf{x}^*). \end{array}$$

The second set of inequalities can be removed, that is, we will prove that

$$\mathbf{a}_i^\top \mathbf{y} \leq 0 \ \text{ for all } i \in I(\mathbf{x}^*) \Rightarrow \nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0.$$

For this, suppose that $\mathbf{y}$ satisfies $\mathbf{a}_i^\top \mathbf{y} \leq 0$ for all $i \in I(\mathbf{x}^*)$. Since $b_i - \mathbf{a}_i^\top \mathbf{x}^* > 0$ for all $i \notin I(\mathbf{x}^*)$, it follows that there exists a small enough $\alpha > 0$ for which $\mathbf{a}_i^\top (\alpha \mathbf{y}) \leq b_i - \mathbf{a}_i^\top \mathbf{x}^*$. Thus, since in addition $\mathbf{a}_i^\top (\alpha \mathbf{y}) \leq 0$ for any $i \in I(\mathbf{x}^*)$, it follows by the stationarity condition that $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$. Therefore, we have shown $\mathbf{a}_i^\top \mathbf{y} \leq 0$ for all $i \in I(\mathbf{x}^*)$

implies that $\nabla f(\mathbf{x}^*)^\top \mathbf{y} \geq 0$. Now, by Farkas' lemma, there exists $\lambda_i \geq 0, i \in I(\mathbf{x}^*)$ such that

$$-\nabla f(\mathbf{x}^*) = \sum_{i \in I(x^*)} \lambda_i \mathbf{a}_i \,.$$

Finally, defining $\lambda_i = 0$ for all $i \notin I(\mathbf{x}^*)$ we get that $\lambda_i \left(\mathbf{a}_i^\top \mathbf{x}^* - b_i\right) = 0$ for all $i \in \{1, 2, \ldots, m\}$ and

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \mathbf{a}_i = \mathbf{0} \,.$$

□

The previous theorem can be improved to necessary and sufficient conditions when f is convex.

Theorem (KKT Conditions for Convex Linearly Constrained Problems - Necessary and Sufficient Optimality Conditions)**.** *Consider the minimization problem* (LCP) *where in addition f is a convex continuously differentiable function over $\mathbb{R}^n$, and let $\mathbf{x}^*$ be a feasible solution. Then $\mathbf{x}^*$ is an optimal solution if and only if there exist $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \mathbf{a}_i = 0$$

and

$$\lambda_i \left(\mathbf{a}_i^\top \mathbf{x}^* - b_i\right) = 0, \quad i = 1, 2, \ldots, m$$

Proof. Necessity was already proven. For sufficiency,suppose that $\mathbf{x}^*$ is a feasible solution of (LCP) satisfying the optimality conditions. Let $\mathbf{x}$ be a feasible solution of (LCP). Define the function

$$h(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \left(\mathbf{a}_i^\top \mathbf{x} - b_i\right) \,.$$

The condition $\nabla h(\mathbf{x}^*) = \mathbf{0}$ implies that $\mathbf{x}^*$ is a minimizer of h over $\mathbb{R}^n$. From here, it follows that

$$f(\mathbf{x}^*) = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \left(\mathbf{a}_i^\top \mathbf{x}^* - b_i\right) \leq f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \left(\mathbf{a}_i^\top \mathbf{x} - b_i\right) \leq f(\mathbf{x}) \,,$$

that is, $\mathbf{x}^*$ is a minimizer of f. □

We conclude by stating the KKT conditions associated to linear problems with equality and inequality constraints.

Theorem (KKT conditions for Linearly Constrained Problems). *Consider the minimization problem*

$$\begin{cases} & \min f(\mathbf{x}) \\ \textit{subject to} & \mathbf{a}_i^\top \mathbf{x} \le b_i, \quad i = 1, 2, \dots, m \\ & \mathbf{c}_j^\top \mathbf{x} = d_j, \quad j = 1, 2, \dots, p \end{cases} \tag{LCPI}$$

where f is continuously differential, $\mathbf{a}_i, \mathbf{c}_j \in \mathbb{R}^n, b_i, d_j \in \mathbb{R}$.

(i) (necessity of the KKT conditions) If x^* is a local minimum of (LCPI) *then there exist $\lambda_1, \lambda_2, \dots, \lambda_m \ge 0$ and $\mu_1, \mu_2, \dots, \mu_p \in \mathbb{R}$ such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \mathbf{a}_i + \sum_{j=1}^{p} \mu_j \mathbf{c}_j = 0$$
$$\lambda_i \left(\mathbf{a}_i^\top \mathbf{x}^* - b_i\right) = 0, \quad i = 1, 2, \dots, m\,.$$

(ii) (sufficiency in the convex case) If f is convex over $\mathbb{R}^n$ and $\mathbf{x}^*$ is a feasible solution of (LCPI) *for which there exist $\lambda_1, \dots, \lambda_m \ge 0$ and $\mu_1, \dots, \mu_p \in \mathbb{R}$ such that the conditions are satisfied, then $\mathbf{x}^*$ is an optimal solution of* (LCPI).

Examples: Solve the problem

$$\begin{array}{ll} \min & \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right) \\ \text{s.t.} & x_1 + x_2 + x_3 = 3\,. \end{array}$$

Orthogonal projections

Using KKT conditions, show the following results!

Orthogonal Projection onto Affine Spaces

Let C be the affine space $C = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = \mathbf{b}\}$, where $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Then,

$$P_C(\mathbf{y}) = \mathbf{y} - \mathbf{A}^\top \left(\mathbf{A}\mathbf{A}^\top\right)^{-1} (\mathbf{A}\mathbf{y} - \mathbf{b})\,.$$

Orthogonal Projection onto Hyperplanes

Consider the hyperplane

$$H = \left\{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\right\} \quad (\mathbf{0} \ne \mathbf{a} \in \mathbb{R}^n, b \in \mathbb{R})\,.$$

Then by the projection on an affine space result (above):

$$P_H(\mathbf{y}) = \mathbf{y} - \mathbf{a}\left(\mathbf{a}^\top \mathbf{a}\right)^{-1} \left(\mathbf{a}^\top \mathbf{y} - b\right) = \mathbf{y} - \frac{\mathbf{a}^\top \mathbf{y} - b}{\|\mathbf{a}\|^2}\mathbf{a}\,.$$

Lemma (distance of a point from a hyperplane). *Let* $H = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} = b\}$, *where* $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$. *Then*

$$d(\mathbf{y}, H) = \frac{|\mathbf{a}^\top \mathbf{y} - b|}{\|\mathbf{a}\|}$$

Proof.

$$d(\mathbf{y}, H) = \|\mathbf{y} - P_H(\mathbf{y})\| = \left\|\mathbf{y} - \left(\mathbf{y} - \frac{\mathbf{a}^\top \mathbf{y} - b}{\|\mathbf{a}\|^2}\mathbf{a}\right)\right\| = \frac{|\mathbf{a}^\top \mathbf{y} - b|}{\|\mathbf{a}\|}$$

□

Similarly, it follows the computation of the orthogonal projection onto half-spaces.

Lemma. *Let* $H^- = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{a}^\top \mathbf{x} \leq b\}$ *where* $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ *and* $b \in \mathbb{R}$. *Then*

$$P_{H^-}(\mathbf{x}) = \mathbf{x} - \frac{[\mathbf{a}^\top \mathbf{x} - b]_+}{\|\mathbf{a}\|^2}\mathbf{a}\,.$$

KKT conditions for nonlinear problems

Now, we extend the notion of KKT conditions to the general nonlinear case. We will start by giving an alternative notion of necessary optimality conditions in terms of **feasible descent directions**, which we introduce in what follows. Consider the problem

$$\min \quad f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in C \tag{G}$$

with $C \subset \mathbb{R}^n$ convex and f a continuously differentiable function over C. A vector $\mathbf{d} \neq \mathbf{0}$ is called a feasible descent direction at $\mathbf{x} \in C$ if

i) $\nabla f(\mathbf{x})^\top \mathbf{d} < 0$, and

ii) there exists $\varepsilon > 0$ such that $\mathbf{x} + t\mathbf{d} \in C$ for all $t \in [0, \varepsilon]$.

Using the notion of feasible descent directions, we reformulate the necessary optimality conditions as given in the following Lemma.

Lemma. *If* $\mathbf{x}^*$ *is a local optimal solution of* (G); *then, there are no feasible descent directions at* $\mathbf{x}^*$.

Now, we will extend this result to problems of the type.

$$\begin{cases} \min f(\mathbf{x}) \\ \text{subject to} \quad g_i(\mathbf{x}) \leq 0, \quad i = 1, 2, \ldots, m \end{cases} \tag{NLP}$$

We introduce the notion of **active constraints**. We say the i-th constraint is active at $\tilde{\mathbf{x}}$ if $g_i(\tilde{\mathbf{x}}) = 0$, i.e., when the constraints are satisfies as equalities. Moreover, the set

$$I(\tilde{\mathbf{x}}) = \{i = \{1, \ldots, m\} : gi(\tilde{\mathbf{x}}) = 0\}$$

is called the set of active constraints at $\tilde{\mathbf{x}}$.

Lemma. *Let $\mathbf{x}^*$ be a local minimum of* (NLP)*, where f and $g_1, \ldots, g_m$ are continuously differentiable functions over $\mathbb{R}^n$. Let $I(\mathbf{x}^*)$ be the set of active constraints at $\mathbf{x}^*$. Then, there does not exists a vector $d \in \mathbb{R}^n$ such that*

i) $\nabla f(\mathbf{x}^*)^\top \mathbf{d} < 0$*, and*

ii) $\nabla g_i(\mathbf{x}^*)^\top \mathbf{d} < 0$*, for all* $i \in I(\mathbf{x}^*)$.

The previous theorem shows that a necessary optimality condition for local optimality is the **infeasibility** of certain system of strict inequalities. More treatable assumptions over the constraints can be imposed to guarantee the infeasibilty of previous system, and they are commonly referred to in the literature as *constraint qualifications*. In the next section, we will show one constraint qualification which works in particular in the convex case.

KKT conditions for nonlinear convex problems

KKT conditions can be extended to study nonlinear constraints. Unfortunately, the amount of detail required for this goes beyond the scope of our module. However, under convexity assumptions, we can state results regarding necessary and sufficient optimality conditions.

Theorem (Sufficiency of the KKT conditions for convex optimization problems)**.** *Let $\mathbf{x}^*$ be a feasible solution of*

$$\left\{ \begin{array}{lll} & \min f(\mathbf{x}) & \\ \textit{subject to} & g_i(\mathbf{x}) \leq 0, & i = 1, 2, \ldots, m \\ & h_j(\mathbf{x}) = 0, & j = 1, 2, \ldots, p\,, \end{array} \right. \tag{NLP}$$

where $f, g_1, \ldots, g_m$ are continuously differentiable convex functions over $\mathbb{R}^n$ and $h_1, \ldots, h_p$ are affine functions. Suppose that there exist multipliers $\lambda_1, \ldots, \lambda_m \geq 0$ and $\mu_1, \ldots, \mu_p \in \mathbb{R}$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \mu_j \nabla h_j(\mathbf{x}^*) = 0\,, \tag{4}$$

$$\lambda_i g_i(\mathbf{x}^*) = 0\,, \quad i = 1, 2 \ldots, m\,. \tag{5}$$

Then $\mathbf{x}^*$ is an optimal solution of (NLP).

A more refined result is stated for necessity of KKT conditions, that is, whether optimal solutions do satisfy the KKT system.

Theorem (necessity of the KKT conditions under the generalized Slater's condition)**.** *Let* $\mathbf{x}^*$ *be an optimal solution of the problem*

$$\begin{cases} & \min f(\mathbf{x}) & \\ \textit{subject to} & g_i(\mathbf{x}) \leq 0, & i = 1, 2, \ldots, m \\ & h_j(\mathbf{x}) \leq 0, & j = 1, 2, \ldots, p\,, \\ & s_k(\mathbf{x}) = 0, & k = 1, 2, \ldots, q\,, \end{cases} \tag{NLP2}$$

where $f, g_1, \ldots, g_m$ *are continuously differentiable convex functions over* $\mathbb{R}^n$ *and* $h_1, \ldots, h_p, s_1 \ldots, s_k$ *are affine functions. Suppose that there exists a point* $\hat{\mathbf{x}}$ *satisfying the generalized Slater's condition*

$$g_i(\hat{\mathbf{x}}) < 0, \quad i = 1, 2, \ldots, m \tag{6}$$
$$h_j(\hat{\mathbf{x}}) \leq 0, \quad j = 1, 2, \ldots, p\,, \tag{7}$$
$$s_k(\hat{\mathbf{x}}) = 0, \quad k = 1, 2, \ldots, q\,. \tag{8}$$

Then, the exist multipliers $\lambda_1, \ldots, \lambda_m, \eta_1, \ldots, \eta_p, \geq 0$ *and* $\mu_1, \ldots, \mu_q \in \mathbb{R}$ *such that*

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{p} \eta_j \nabla h_j(\mathbf{x}^*) + \sum_{k=1}^{q} \mu_k \nabla s_k(\mathbf{x}^*) = 0\,, \tag{9}$$
$$\lambda_i g_i(\mathbf{x}^*) = 0\,, \quad i = 1, 2 \ldots, m\,, \tag{10}$$
$$\eta_j h_j(\mathbf{x}^*) = 0\,, \quad i = 1, 2 \ldots, p\,. \tag{11}$$

Part IX.
Duality

The Primal and Dual Problems

Consider the problem

$$\begin{aligned} f^* := \min\ & f(\mathbf{x}) \\ \text{s.t.}\quad & g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m \\ & h_j(\mathbf{x}) = 0, j = 1, 2, \ldots, p, \\ & \mathbf{x} \in X, \end{aligned} \tag{Primal}$$

where $f, g_i, h_j (i = 1, 2, \ldots, m, j = 1, 2, \ldots, p)$ are functions defined on the set $X \subseteq \mathbb{R}^n$. This is the "usual" optimization problem, and we will refer to it as the **primal** problem. As discussed in last week, the Lagrangian associated to this problem is

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x}) \quad \left(\mathbf{x} \in X, \boldsymbol{\lambda} \in \mathbb{R}^m_+, \boldsymbol{\mu} \in \mathbb{R}^p\right).$$

The **dual** objective function $q : \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$ is defined to be

$$q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

The domain of the dual objective function is

$$\operatorname{dom}(q) = \left\{(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \mathbb{R}^m_+ \times \mathbb{R}^p : q(\boldsymbol{\lambda}, \boldsymbol{\mu}) > -\infty\right\}.$$

The **dual problem** is given by

$$\begin{aligned} q^* := \max\ & q(\boldsymbol{\lambda}, \boldsymbol{\mu}) \\ \text{s.t.}\quad & (\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \operatorname{dom}(q) \end{aligned} \tag{Dual}$$

In many optimization problems it is useful to study the properties of the dual problem, and even resorting to solving the dual problem instead. This is an open-ended question that we shall explore in this week, trying to understand when and why is this good idea, and illustrating with some relevant examples. We begin by stating some relevant properties of the dual problem.

Theorem. *Consider the primal problem* (Primal) *with* $f, g_i, h_j (i = 1, 2, \ldots, m, j = 1, 2, \ldots, p)$ *being functions defined on the set* $X \subseteq \mathbb{R}^n$, *and let* q *be the dual function defined in* (Dual). *Then:*

a) $\operatorname{dom}(q)$ *is a convex set.*

b) q *is a concave function over* $\operatorname{dom}(q)$.

Proof. (a) Take $(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1), (\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) \in \operatorname{dom}(q)$ and $\alpha \in [0,1]$. Then

$$\min_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1) > -\infty,$$
$$\min_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}_2, \boldsymbol{\mu}_2) > -\infty.$$

Therefore, since the Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is affine w.r.t. $\boldsymbol{\lambda}, \boldsymbol{\mu}$

$$\begin{aligned} q\left(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2\right) &= \min_{\mathbf{x}\in X} L\left(\mathbf{x}, \alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\boldsymbol{\mu}_1 + (1-\alpha)\boldsymbol{\mu}_2\right) \\ &= \min_{\mathbf{x}\in X} \left\{\alpha L\left(\mathbf{x}, \boldsymbol{\lambda}_1, \boldsymbol{\mu}_1\right) + (1-\alpha) L\left(\mathbf{x}, \boldsymbol{\lambda}_2, \boldsymbol{\mu}_2\right)\right\} \\ &\geq \alpha \min_{\mathbf{x}\in X} L\left(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}_1\right) + (1-\alpha) \min_{\mathbf{x}\in X} L\left(\mathbf{x}, \boldsymbol{\lambda}_2, \boldsymbol{\mu}_2\right) \\ &= \alpha q\left(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1\right) + (1-\alpha) q\left(\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2\right) \\ &> -\infty \end{aligned}$$

Hence, $\alpha\left(\boldsymbol{\lambda}_1, \boldsymbol{\mu}_1\right) + (1-\alpha)\left(\boldsymbol{\lambda}_2, \boldsymbol{\mu}_2\right) \in \operatorname{dom}(q)$, and the convexity of $\operatorname{dom}(q)$ is established. (b) $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ is an affine function w.r.t. $(\boldsymbol{\lambda}, \boldsymbol{\mu})$. In particular, it is a concave function w.r.t. $(\boldsymbol{\lambda}, \boldsymbol{\mu})$. Hence, since q is the minimum of concave functions, it must be concave. □

Weak and Strong Duality

A first important consequence for optimization is the week duality theorem, which establishes a lower bound for the primal optimal value with respect to the dual optimal value.

Theorem (Weak Duality Theorem)**.** *Consider the primal problem* (Primal) *and its dual problem* (Dual)*. Then*

$$q^* \leq f^*$$

where f^*, q^* are the primal and dual optimal values respectively.

Proof. The feasible set of the primal problem is

$$S = \left\{\mathbf{x} \in X : g_i(\mathbf{x}) \leq 0, h_j(\mathbf{x}) = 0, i = 1, 2, \ldots, m, j = 1, 2, \ldots, p\right\}.$$

Then for any $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \operatorname{dom}(q)$ we have

$$\begin{aligned} q(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x}\in X} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) &\leq \min_{\mathbf{x}\in S} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= \min_{\mathbf{x}\in S} \left\{f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \mu_j h_j(\mathbf{x})\right\} \\ &\leq \min_{\mathbf{x}\in S} f(\mathbf{x}) = f^*. \end{aligned}$$

Taking the maximum over $(\boldsymbol{\lambda}, \boldsymbol{\mu}) \in \operatorname{dom}(q)$, the result follows. □

Example:

$$\begin{array}{ll} \min & x_1^2 - 3x_2^2 \\ \text{s.t.} & x_1 = x_2^3 \end{array}$$

While the weak duality theorem is useful to obtain a lower bound for the optimal value of the primal problem, a more powerful result can be proven known as **strong duality**. For this, we need to cast a nonlinear variant of Farkas' Lemma, which we briefly state.

Theorem (Supporting Hyperplane Theorem). *Let $C \subseteq \mathbb{R}^n$ be a convex set and let $\mathbf{y} \notin C$. Then there exists $\mathbf{0} \neq \mathbf{p} \in \mathbb{R}^n$ such that*

$$\mathbf{p}^T\mathbf{x} \leq \mathbf{p}^T\mathbf{y} \text{ for any } \mathbf{x} \in C\,.$$

Theorem (Separation of Two Convex Sets). *Let $C_1, C_2 \subseteq \mathbb{R}^n$ be two nonempty convex sets such that $C_1 \cap C_2 = \emptyset$. Then there exists $\mathbf{0} \neq \mathbf{p} \in \mathbb{R}^n$ for which*

$$\mathbf{p}^T\mathbf{x} \leq \mathbf{p}^T\mathbf{y} \text{ for any } \mathbf{x} \in C_1, \mathbf{y} \in C_2\,.$$

Theorem (Nonlinear Farkas Lemma). *Let $X \subseteq \mathbb{R}^n$ be a convex set and let $f, g_1, g_2, \ldots, g_m$ be convex functions over X. Assume that there exists $\hat{x} \in X$ such that*

$$g_1(\hat{\mathbf{x}}) < 0, g_2(\hat{\mathbf{x}}) < 0, \ldots, g_m(\hat{\mathbf{x}}) < 0\,.$$

Let $c \in \mathbb{R}$. Then the following two claims are equivalent:

a) The following implication holds:

$$\mathbf{x} \in X, g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m \Rightarrow f(\mathbf{x}) \geq c\,.$$

b) There exist $\lambda_1, \lambda_2, \ldots, \lambda_m \geq 0$ such that

$$\min_{x \in X} \left\{ f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right\} \geq c\,.$$

With these technical results, we are in position to state the strong duality result.

Theorem (Strong Duality of Convex Problems with Inequality Constraints). *Consider the optimization problem*

$$\begin{array}{rll} f^* = \min & f(\mathbf{x}) & \\ s.t. & g_i(\mathbf{x}) \leq 0, & i = 1, 2, \ldots, m, \\ & \mathbf{x} \in X & \end{array}$$

where X is a convex set and $f, g_i, i = 1, 2, \ldots, m$ are convex functions over X. Suppose that there exists $\hat{\mathbf{x}} \in X$ for which $g_i(\hat{\mathbf{x}}) < 0, i = 1, 2, \ldots, m$. If this problem has a finite optimal value, then

a) the optimal value of the dual problem is attained.

b) the primal and dual problems have the same optimal value, $f^* = q^*$.

Proof. Since $f^* > -\infty$ is the optimal value of the primal problem, it follows that the following implication holds:

$$\mathbf{x} \in X, g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m \Rightarrow f(\mathbf{x}) \geq f^* .$$

By the nonlinear Farkas Lemma, there exists $\tilde{\lambda}_1, \tilde{\lambda}_2, \ldots, \tilde{\lambda}_m \geq 0$ such that

$$q(\tilde{\lambda}) = \min_{\mathbf{x} \in X} \left\{ f(\mathbf{x}) + \sum_{j=1}^{m} \tilde{\lambda}_j g_j(\mathbf{x}) \right\} \geq f^* .$$

By the weak duality theorem,

$$q^* \geq q(\tilde{\boldsymbol{\lambda}}) \geq f^* \geq q^* .$$

Hence, $f^* = q^*$ and $\tilde{\boldsymbol{\lambda}}$ is an optimal solution of the dual problem. □

The result above indicates that under the convexity assumptions of the theorem, it is possible to obtain the solution of the primal problem by solving its dual.

Example:

$$\begin{array}{ll} \min & x_1^2 - x_2 \\ \textbf{s.t.} & x_2^2 \leq 0 . \end{array}$$

Theorem (Complementary Slackness Conditions)**.** *Consider the optimization problem*

$$f^* := \min \{ f(\mathbf{x}) : g_i(\mathbf{x}) \leq 0, i = 1, 2, \ldots, m, \mathbf{x} \in X \} ,$$

and assume that $f^* = q^*$ where q^* is the optimal value of the dual problem. Let $\mathbf{x}^*, \boldsymbol{\lambda}^*$ be feasible solutions of the primal and dual problems. Then $\mathbf{x}^*, \boldsymbol{\lambda}^*$ are optimal solutions of the primal and dual problems iff

$$\mathbf{x}^* \in \operatorname*{argmin}_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*)$$
$$\lambda_i^* g_i(\mathbf{x}^*) = 0, i = 1, 2, \ldots, m$$

Proof. We have

$$q(\lambda^*) = \min_{\mathbf{x} \in X} L(\mathbf{x}, \lambda^*) \leq L(\mathbf{x}^*, \lambda^*) = f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) \leq f(\mathbf{x}^*) .$$

By strong duality, $\mathbf{x}^*, \boldsymbol{\lambda}^*$ are optimal iff $f(\mathbf{x}^*) = q(\boldsymbol{\lambda}^*)$. This is equivalent to $\min_{\mathbf{x} \in X} L(\mathbf{x}, \boldsymbol{\lambda}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*)$, and $\sum_{i=1}^{m} \lambda_i^* g_i(\mathbf{x}^*) = 0$, which in turn is equivalent to the slackness conditions in the theorem. □

We conclude these results with a more general duality theorem including convex affine inequality and equality contraints.

Theorem (General Strong Duality Theorem). *Consider the optimization problem*

$$\begin{aligned} f^* = \min \ & f(\mathbf{x}) \\ s.t.\ g_i(\mathbf{x}) &\leq 0, \quad i = 1, 2, \ldots, m \\ h_j(\mathbf{x}) &\leq 0, \quad j = 1, 2, \ldots, p \\ s_k(\mathbf{x}) &= 0, \quad k = 1, 2, \ldots, q \\ \mathbf{x} &\in X, \end{aligned}$$

where X is a convex set and $f, g_i, i = 1, 2, \ldots, m$ are convex functions over X. The functions h_j, s_k are affine functions. Suppose that there exists $\hat{\mathbf{x}} \in \text{int}(X)$ for which $g_i(\hat{\mathbf{x}}) < 0$, $h_j(\hat{\mathbf{x}}) \leq 0$, and $s_k(\hat{\mathbf{x}}) = 0$. Then if the problem has a finite optimal value, then the optimal value of the dual problem

$$q^* = \max\{q(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) : (\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) \in \text{dom}(q)\}$$

where

$$q(\boldsymbol{\lambda}, \boldsymbol{\eta}, \boldsymbol{\mu}) = \min_{\mathbf{x} \in X} \left[f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{p} \eta_j h_j(\mathbf{x}) + \sum_{k=1}^{q} \mu_k s_k(\mathbf{x}) \right]$$

is attained, and $f^* = q^*$.

Example. Consider the problem

$$\begin{aligned} \min\ & x_1^3 + x_2^3 \\ & x_1 + x_2 \geq 1 \\ & x_1, x_2 \geq 0\,. \end{aligned}$$

We will show that a problem can have different dual formulations with different duality gaps. It depends on our choice of X. Through the usual KKT conditions, we can easily find that $\left(\frac{1}{2}, \frac{1}{2}\right)$ is the optimal solution of the primal problem with an optimal value $f^* = \frac{1}{4}$. A first dual problem is constructed by taking $X = \{(x_1, x_2) : x_1, x_2 \geq 0\}$. The primal problem is $\min\left\{x_1^3 + x_2^3 : x_1 + x_2 \geq 1, (x_1, x_2) \in X\right\}$. Strong duality holds for the problem and hence in particular $q^* = \frac{1}{4}$. A second dual is constructed by taking $X = \mathbb{R}^2$. In this case, the objective function is not convex, implying that strong duality is not necessarily satisfied. In this case, the Lagrangian is given by

$$L\left(x_1, x_2, \lambda, \eta_1, \eta_2\right) = x_1^3 + x_2^3 - \lambda\left(x_1 + x_2 - 1\right) - \eta_1 x_1 - \eta_2 x_2\,.$$

In this case, $q\left(\lambda, \eta_1, \eta_2\right) = -\infty$ for all $\left(\lambda, \eta_1, \eta_2\right) \Rightarrow q^* = -\infty$, and the duality gap is infinite. It is important to make a good choice of X to obtain useful information about the solution of the primal problem.

Three Important Examples of Duality Use

Linear Programming

Consider the linear programming problem

$$\begin{aligned} \min \quad & \mathbf{c}^T\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \end{aligned}$$

where $\mathbf{c} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. We assume that the problem is feasible, implying that strong duality holds. The Lagrangian is given by

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T\mathbf{x} + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \left(\mathbf{c} + \mathbf{A}^T\boldsymbol{\lambda}\right)^T \mathbf{x} - \mathbf{b}^T\boldsymbol{\lambda},$$

and the dual objective function is

$$q(\boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}\in\mathbb{R}^n} \left(\mathbf{c} + \mathbf{A}^T\lambda\right)^T \mathbf{x} - \mathbf{b}^T\boldsymbol{\lambda} = \begin{cases} -\mathbf{b}^T\boldsymbol{\lambda} & \mathbf{c} + \mathbf{A}^T\boldsymbol{\lambda} = 0 \\ -\infty & \text{else.} \end{cases}$$

Therefore, the dual problem is formulated as

$$\begin{aligned} \max \quad & -\mathbf{b}^T\boldsymbol{\lambda} \\ \text{s.t.} \quad & \mathbf{A}^T\boldsymbol{\lambda} = -\mathbf{c} \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned}$$

Strictly Convex Quadratic Programming

Consider the strictly convex quadratic programming problem

$$\begin{aligned} \min \quad & \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{f}^T\mathbf{x} \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} \leq \mathbf{b}, \end{aligned}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$ positive definite, $\mathbf{f} \in \mathbb{R}^n, \mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^m$. The Lagrangian (recall $\boldsymbol{\lambda} \in \mathbb{R}^m_+$) is given by:

$$L(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\mathbf{f}^T\mathbf{x} + 2\boldsymbol{\lambda}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = \mathbf{x}^T\mathbf{Q}\mathbf{x} + 2\left(\mathbf{A}^T\boldsymbol{\lambda} + \mathbf{f}\right)^T \mathbf{x} - 2\mathbf{b}^T\boldsymbol{\lambda}.$$

The minimizer of the Lagrangian is attained at $\mathbf{x}^* = -\mathbf{Q}^{-1}\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right)$. With this, we work over the dual objective,

$$\begin{aligned} q(\boldsymbol{\lambda}) &= L\left(\mathbf{x}^*, \boldsymbol{\lambda}\right) \\ &= \left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right)^T \mathbf{Q}^{-1}\mathbf{Q}\mathbf{Q}^{-1}\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right) - 2\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right)^T \mathbf{Q}^{-1}\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right) - 2\mathbf{b}^T\boldsymbol{\lambda} \\ &= -\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right)^T \mathbf{Q}^{-1}\left(\mathbf{f} + \mathbf{A}^T\boldsymbol{\lambda}\right) - 2\mathbf{b}^T\boldsymbol{\lambda} \\ &= -\boldsymbol{\lambda}^T\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\boldsymbol{\lambda} - 2\mathbf{f}^T\mathbf{Q}^{-1}\mathbf{A}^T\boldsymbol{\lambda} - \mathbf{f}^T\mathbf{Q}^{-1}\mathbf{f} - 2\mathbf{b}^T\boldsymbol{\lambda} \\ &= -\boldsymbol{\lambda}^T\mathbf{A}\mathbf{Q}^{-1}\mathbf{A}^T\boldsymbol{\lambda} - 2\left(\mathbf{A}\mathbf{Q}^{-1}\mathbf{f} + \mathbf{b}\right)^T\boldsymbol{\lambda} - \mathbf{f}^T\mathbf{Q}^{-1}\mathbf{f}. \end{aligned}$$

While this might look complicated, the resulting dual problem $\max\{q(\boldsymbol{\lambda}) : \boldsymbol{\lambda} \geq 0\}$ it is in fact another convex optimization problem, with a simpler feasible set than the primal problem.

Computing the Orthogonal Projection onto the Unit Simplex

Given a vector $\mathbf{y} \in \mathbb{R}^n$, we would like to compute the orthogonal projection of the vector $\mathbf{y}$ onto Δ_n. The corresponding optimization problem is

$$\begin{array}{ll} \min & \|\mathbf{x} - \mathbf{y}\|^2 \\ \text{s.t.} & \mathbf{e}^T \mathbf{x} = 1 \\ & \mathbf{x} \geq 0 \,. \end{array}$$

We will associate a Lagrange multiplier $\lambda \in \mathbb{R}$ to the linear equality constraint $\mathbf{e}^T \mathbf{x} = 1$ and obtain the Lagrangian function

$$\begin{aligned} L(\mathbf{x}, \lambda) &= \|\mathbf{x} - \mathbf{y}\|^2 + 2\lambda \left(\mathbf{e}^T \mathbf{x} - 1 \right) = \|\mathbf{x}\|^2 - 2(\mathbf{y} - \lambda \mathbf{e})^T \mathbf{x} + \|\mathbf{y}\|^2 - 2\lambda \\ &= \sum_{j=1}^{n} \left(x_j^2 - 2 \left(y_j - \lambda \right) x_j \right) + \|\mathbf{y}\|^2 - 2\lambda \,. \end{aligned}$$

The arising problem is therefore separable with respect to the variables x_j and hence the optimal x_j is the solution to the one-dimensional problem

$$\min_{x_j \geq 0} \left[x_j^2 - 2 \left(y_j - \lambda \right) x_j \right]$$

The optimal solution to the above problem is given by

$$x_j = \begin{cases} y_j - \lambda, & y_j \geq \lambda \\ 0 & \text{else} \end{cases} = \left[y_j - \lambda \right]_+$$

and the optimal value is $- \left[y_j - \lambda \right]_+^2$. The dual problem is therefore

$$\max_{\lambda \in \mathbb{R}} \left\{ g(\lambda) \equiv - \sum_{j=1}^{n} \left[y_j - \lambda \right]_+^2 - 2\lambda + \|\mathbf{y}\|^2 \right\}$$

By the basic properties of dual problems, the dual objective function is concave. In order to actually solve the dual problem, we note that

$$\lim_{\lambda \to \infty} g(\lambda) = \lim_{\lambda \to \infty} g(\lambda) = -\infty$$

Therefore, since $-g$ is a coercive and differentiable function, it follows that there exists an optimal solution to the dual problem attained at a point λ in which

$$g'(\lambda) = 0 \,,$$

meaning that

$$\sum_{j=1}^{n} \left[y_j - \lambda \right]_+ = 1 \,.$$

The function $b(\lambda) = \sum_{j=1}^{n} \left[y_j - \lambda \right]_+ - 1$ is a nonincreasing function over $\mathbb{R}$ and is in fact strictly decreasing over $\left(-\infty, \max_j y_j \right]$. In addition, by denoting $y_{\max} = \max_{j=1,2,\ldots,n} y_j$, and $y_{\min} = \min_{j=1,2,\ldots,n} y_j$, we have

$$h\left(y_{\max}\right) = -1$$
$$b\left(y_{\min} - \frac{2}{n}\right) = \sum_{j=1}^{n} y_j - n y_{\min} + 2 - 1 > 0\,,$$

and we can therefore invoke a bisection procedure to find the unique root λ of the function h over the interval $\left[y_{\min} - \frac{2}{n}, y_{\max} \right]$ and then define $P_{\Delta_n}(\mathbf{y}) = [\mathbf{y} - \lambda \mathbf{e}]_+$.

Part X.
Optimal Control

What is Mathematical Control Theory?

Mathematical control theory is the area of applied mathematics that deals with the analysis, design, and computation of control systems. Controlling a system means to influence its behavior to achieve a desired goal. This simple idea underpins **all our technology**. Engines, autopilot systems, satellites, computer networks, and chemical reactors are just a handful of systems that exist thanks to control mechanisms. From a mathematical viewpoint, control theory is at the interface of many areas including:

- Dynamical Systems (ODEs and PDEs)
- Optimization, Calculus of Variations, and Operations Research
- Game Theory
- Computational Mathematics
- Data Science

Moreover, many fundamentals aspects of mathematical control theory make extensive use of real and complex analysis, algebra, and geometry. A colourful 2020 panorama of control theory is presented in Figure 30.

Before embarking in our journey, we must understand the central object of study in control theory. Given a dynamical system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0 \in \mathbb{R}^n, \tag{12}$$

mathematicians are often concerned about well-posedness (existence, uniqueness, continuous dependence with initial conditions) and characterization of equilibria. There exists the expectation that, from a given initial condition and without external intervention, the state variable of the system $\mathbf{x}(t)$ will evolve and exhibit a certain behavior as in the motion of the planets, or even our atmosphere before we started messing with it. **This is radically different in the control realm**. We will make extensive use of the knowledge we have of the system dynamics, however, we will introduce an external forcing represented by a **control variable** $\mathbf{u}(t)$ in $\mathbb{R}^m$, leading to a control system

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)). \tag{13}$$

The control paradigm is to **synthesize** a control law $\mathbf{u}(t)$ to enforce a desired objective, which can include:

- the stabilization of unstable dynamics (mechanics, chemical reactors),
- the rejection/mitigation of external disturbances (active noise cancelling),

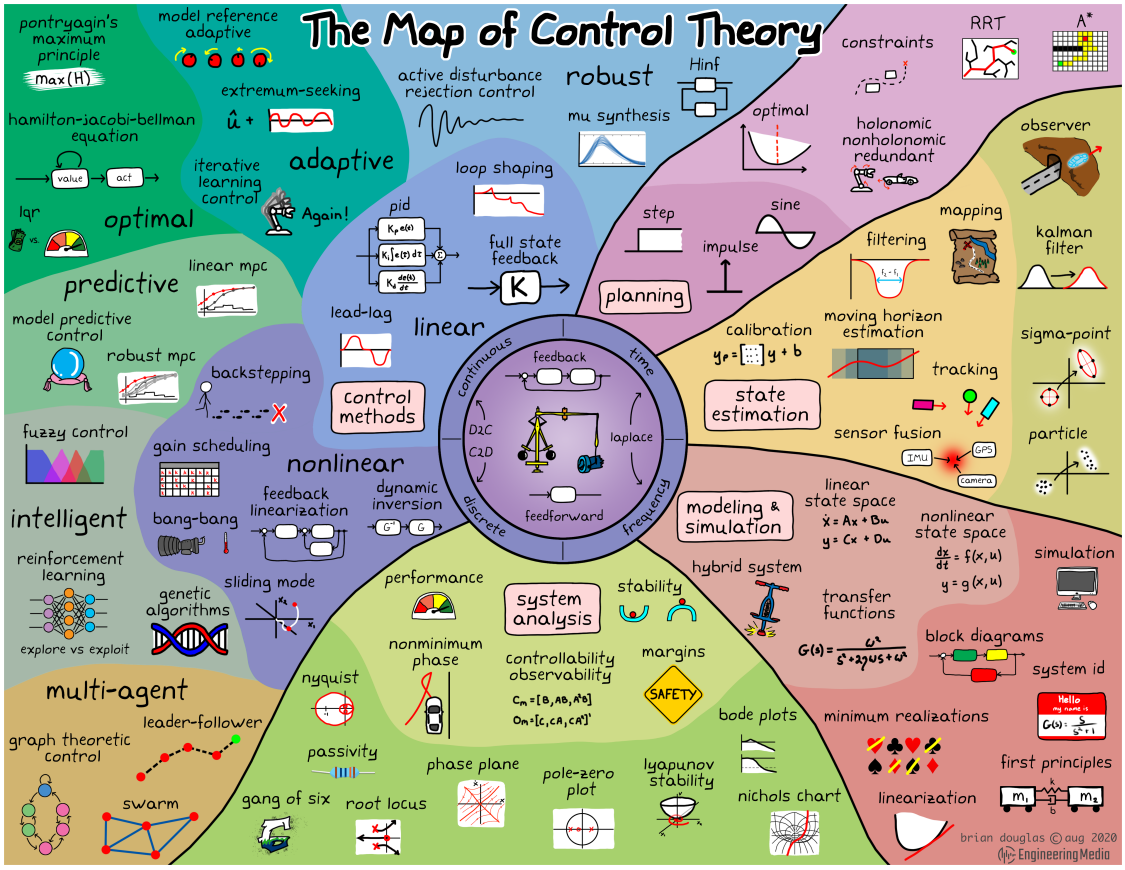


Figure 30: Control Theory Map by Brian Douglas, https://engineeringmedia.com/. In dynamic optimization , we move around the upper/middle left part of the map, including optimal control, model predictive control, and reinforcement learning.

- the tracking of a reference trajectory (autopilots, satellites),
- accelerating the convergence towards stable equilibria (molecular dynamics).

A quintessential example: the pendulum. This problem is present in every single control book. Consider the dynamics of a rigid pendulum of unit length and a ball of mass m given by

$$m\ddot{\theta}(t) + mg\sin\theta(t) = \mathbf{u}(t)\,, \tag{14}$$

where the state of the system is determined by the pendulum angle and angular velocity denoted by $\theta(t)$ and $\dot{\theta}(t)$, respectively. The control action $\mathbf{u}(t)$ is a motor placed at the pivot, as depicted in Figure 31. Setting $x_1 = \theta$ and $x_2 = \dot{\theta}$, we write the dynamics (14) as a first-order system (assumming $m = g = 1$.)

$$\dot{\mathbf{x}}(t) = \frac{d}{dt}\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} x_2(t) \\ -\sin(x_1(t)) + \mathbf{u}(t) \end{bmatrix}, \tag{15}$$

If we study the uncontrolled dynamics, we can see that the vertical position at rest $\theta = \pi, \dot{\theta} = 0$ is an equilibrium of the system, however, any small perturbation from this state will result in unstable motion. A more realistic model would include a dissipation effect which would cause all the initial states, except for $\theta = \pi, \dot{\theta} = 0$, to converge to $\theta = \dot{\theta} = 0$. In this case, we can state as a control objective, our desire to steer a given configuration at $t = 0$ towards the vertical stationary position in minimum time, or to compensate external disturbances to stabilize around this reference configuration. A common idea in control theory is to assume the initial state close enough to the reference configuration and linearize the system. In this case,

$$\theta \approx \pi \Rightarrow \sin\theta = -(\theta - \pi) + o(\theta - \pi) \tag{16}$$

and the dynamics are approximated by

$$\ddot{\phi}(t) - \phi(t) = \mathbf{u}(t), \tag{17}$$

where the new state $\phi(t)$ is the departure angle from the position $\theta = \pi$. As you can

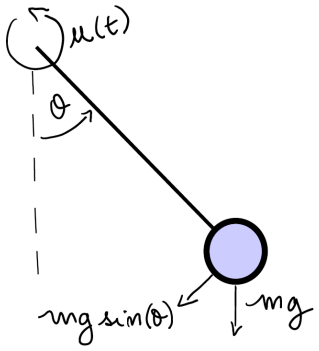


Figure 31: A rigid pendulum is controlled through the action of an engine $\mathbf{u}(t)$ acting at the pivot. The control objective is steering the state of the system, given by the pendulum angle and its velocity ($\theta(t)$ and $\dot{\theta}(t)$, respectively), towards a reference configuration.

see in the control map, there exists a big splitting in the control theory realm between linear and nonlinear synthesis methods. Here, we will work with the nonlinear dynamics directly, avoiding any kind of linearization.

In this case, it is possible to synthesize a **control law** simply by having a close look at the dynamics. If the pendulum is at the left of the vertical position, say $\phi = \theta - \pi > 0$, then we wish to push ϕ to the right, and the opposite happens when $\phi < 0$. Therefore, we can prescribe a **linear feedback control**, proportional to the angle, given by

$$\mathbf{u}(t) = \mathbf{u}(\phi(t)) = -\alpha\phi(t). \tag{18}$$

The implementation of such a control law would require the tuning a suitable parameter α and some observation mechanism to, at every time t, recover the angle $\phi(t)$ to provide the law $\mathbf{u}(t)$. Regardless of the implementation, we can always do an stability analysis by

studying the **closed-loop**, that is, the dynamics resulting from applying the feedback law. In our case, they read

$$\ddot{\phi}(t) - \phi(t) = -\alpha\phi(t)\,. \tag{19}$$

This sort of linear stability analysis allows us to understand whether such a control law is a suitable mechanism to stabilize the dynamics around the equilibrium.

Exercise. Do this linear stability analysis. You should arrive to the conclusion that this is a bad idea no matter what value of α you choose. This leads to consider feedback laws that also include the velocity,

$$\mathbf{u}(t) = -\alpha\phi(t) - \beta\dot{\phi}(t)\,, \tag{20}$$

and to determine that if $\beta^2 > 4(\alpha - 1)$, then convergence to the vertical position can be achieved without oscillations. All this analysis strongly relies on using the linearized version of the dynamics, therefore this control law is only expected to work **locally**, that is for small perturbations from the reference position.

What is Optimal Control?

While the previous approach can be applied in some specific examples, it strongly relies on the physics of the problem or some further structural assumptions. The construction of suitable feedback control signals is rarely that evident. In optimal control, we express our control goal by means of an objective function to be optimized, i.e., as a **dynamic optimization** problem such as

$$\min_{\mathbf{u}(\cdot)\in\mathcal{U}} \int_0^T L(\mathbf{x}(s), \mathbf{u}(s))\, ds + \Phi(\mathbf{x}(T))$$

subject to

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\,, \\ \mathbf{x}(0) &= \mathbf{x}_0\,. \end{aligned}$$

Here, the control signal lives in a space of **admissible controls** $\mathcal{U}$, representing the control constraints of our problem. The running cost $L(\mathbf{x}, \mathbf{u})$ is a **running cost**, expressing our wish to achieve an objective with a certain control budget. Finally, $\Phi(\mathbf{x}(T))$ is a **final time penalty**, encoding the fact that when our optimization finishes at time $t = T$, we expect to find the system in the reference position. Here, the **time horizon** can be fixed (T or $+\infty$) or variable, i.e. treated as an additional optimization variable. Going back to our pendulum example, some suitable choice could be:

- $\mathcal{U} := L^\infty([0, +\infty[; U)$, where U is a compact subset of $\mathbb{R}$.
- $L(\mathbf{x}, \mathbf{u}) := \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \frac{\gamma}{2}\|\mathbf{u}\|^2$, where $\bar{\mathbf{x}}$ is the vertical position and $\gamma > 0$.
- $\Phi(\mathbf{x}) = \|\mathbf{x} - \bar{\mathbf{x}}\|^4$, i.e. some stronger penalization of the objective.

- The horizon could be fixed T, $+\infty$ or minimized (minimum arrival time to the vertical position).

The formulation above can be seen as an optimization problem with a dynamical constraint (hence the dynamic optimization name). At least formally, it is possible to proceed as in a constrained optimization framework, using Lagrange multipliers for the constraints and deriving first and second-order optimality conditions. We will focus on the study of optimal control problems of this type, with an emphasis on characterizing optimal solutions.

Open and Closed-loop Control, and Learning

The control theory map in Figure 30 does not capture a very relevant aspect dividing synthesis methods in optimal control theory. If we think about the pendulum example, it is somehow natural for humans to think in terms of feedback laws. If you look at Figure 32, you will see the Chavo trying to balance a broom with his foot. For us, it comes as a very natural act to compensate the deviation angle and its velocity based on our current perception of the state of the system. This is a feedback mechanism at its purest, also known as **closed-loop control**. Our control synthesis is directly expressed as a function of the current state, and the initial condition becomes irrelevant. On the other hand, our formalization of the balancing problem as a dynamic optimization problem requires a rigid setting under which the optimal control will be computed. In particular, the initial condition $\mathbf{x}_0$ has been fixed. For a different initial condition, we cast a different optimization problem, and we assume the optimal control signal will be plugged into the system from 0 to T in a perfect way. We call this design a **open-loop control**. This is theoretically and computationally sound, but

- is it reasonable to think that we will be able to execute the control signal without any error or external perturbation?
- is it really necessary to determine an optimal action for every possible initial state of the system?

The ultimate answer to these questions, is the synthesis of an **optimal feedback map**, perhaps the most challenging problem in optimal control theory. The computation of optimal feedback maps is a problem of formidable computational complexity. The current understanding of the subject leads to the fascinating topic of **reinforcement learning**. Somehow, the way we learn to balance a broom is through extensive trial and error where our attempts are driven by our objective of optimizing a performance measure. This information is used to improve the construction of a feedback map, which becomes harder and harder to optimize.

Why Optimal Control?

In the previous chapter we have discussed some essential aspects of control theory, and we have seen examples of control design based on physical considerations. We have

Figure 32: El Chavo del Ocho, a former student of Optimization applying deep reinforcement learning techniques to balance a broom. Note how his graceful stance allows him to smoothly compensate external disturbances in real time.

proposed the use of optimization-based controllers as suitable synthesis strategy. Is this really the case? In fact, it is estimated that about 90% of industrial control systems currently operating in industry **do not** use any kind of optimal control design.The majority of the control systems correspond to *PID* controllers, which use simple design principles similar to those discussed in the previous section. That is, a linear feedback proportional to the state $\mathbf{x}$, its derivative $\dot{\mathbf{x}}$, and an integral term. Tuning the influence of these terms is an art of its own, and strongly relies on experts' knowledge[6] On the other hand, data-driven methods like reinforcement learning (RL) exhibit and unparalleled degree of automation, however this comes at the expense of huge sampling sets and a very inefficient use of data in general. Here, we will study the use of optimal control theory as a compromise or trade-off between data efficiency and automation level. On the one hand, we will assume the dynamical system we want to control can be represented as a dynamical system for which the governing equations are known, as opposed to a pure RL framework. On the other hand, we will avoid further parametrizations of the control action as in the simplest *ad-hoc* designs. Instead, we will cast the control synthesis as a dynamic optimization problem, for which we will derive optimality conditions to be numerically realized by a computer. In summary, we will use as much theory as possible,pushing the introduction of computational methods for the synthesis until the very end of our design process.

[6] In this context, experts' knowledge is the opposite of blind automation, as a the amount of information that can be extracted from a single sample is large compared to random sampling in, for example, stochastic gradient descent.

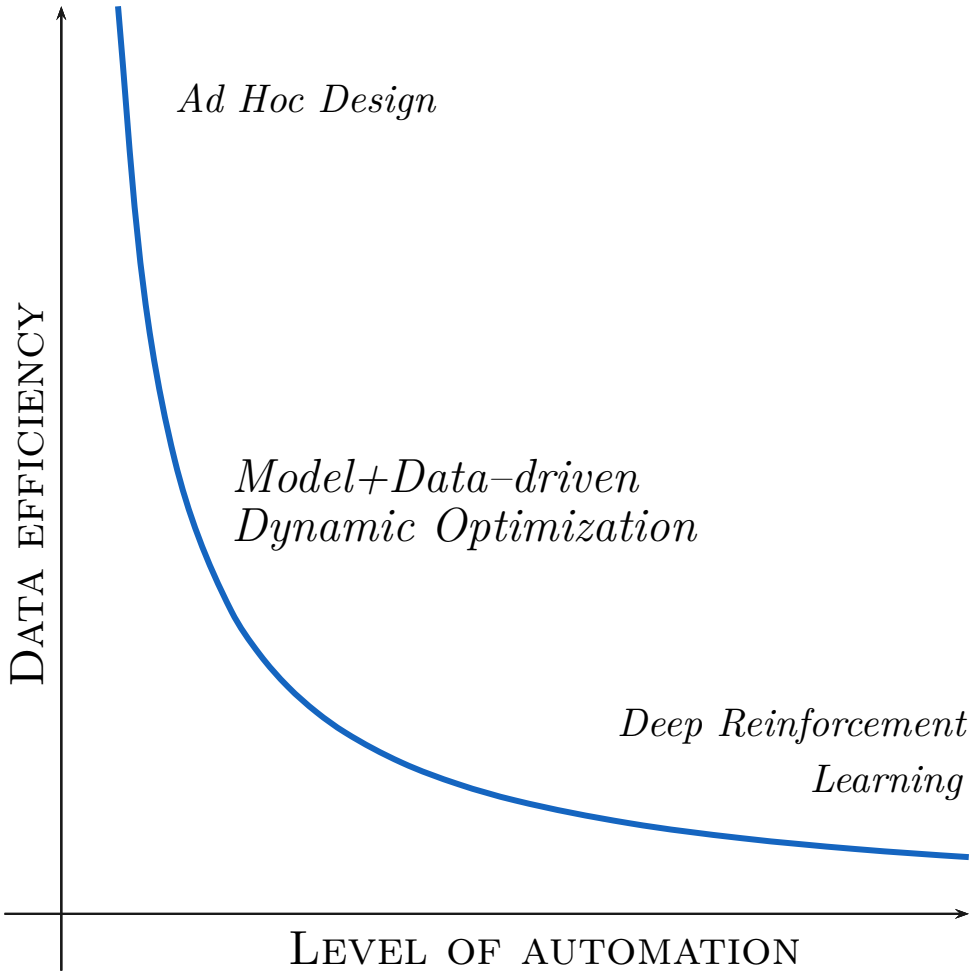


Figure 33: In a graph displaying automation level against data efficiency, the use of model-based optimization and control techniques represents a trade-off between ad-hoc designs and data-intensive synthesis methods such as reinforcement learning.

The Optimal Control Formulation

Our starting point is the control of a nonlinear dynamical system of the form

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n, \quad t_0 \leq t \leq t_f, \tag{21}$$

over a time frame $[t_0, t_f]$, and where the initial condition $\mathbf{x}_0$ is known. This system is manipulated through the action of an external control signal $\mathbf{u}(t)$ to be computed. We aim characterizing the design of such a signal by means of calculus of variations and

optimization methods. For this, we write a cost functional expressing our control goals

$$\mathcal{J}(\mathbf{x}, \mathbf{u}) := \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} L(\mathbf{x}(t), \mathbf{u}(t))\, dt \tag{22}$$

by means of a running cost $L(\mathbf{x}(t), \mathbf{u}(t))$ and a terminal penalty $\Phi(\mathbf{x}(t_f), t_f)$. We cast the control synthesis as a nonlinear optimization problem

$$\min_{\mathbf{u}(\cdot)} \mathcal{J}(\mathbf{x}(\cdot), \mathbf{u}(\cdot)), \quad \text{subject to the dynamics (21).} \tag{23}$$

We need to understand what is the true nature of the optimization problem above. A first observation is that including the nonlinear system (21) as a constraint generates a dynamic optimization problem, which is considerably different to standard *static* optimization problems[7]. A second observation, which will be used for the construction of numerical techniques, is to understand that for a fixed initial condition $\mathbf{x}_0$, the controlled trajectory $\mathbf{x}(\cdot)$ can be uniquely determined from the control signal, that is $\mathbf{x} = \mathbf{x}(\mathbf{u}(\cdot))$, expressing the cost as $\mathcal{J}(\mathbf{u}(\cdot))$, known in the literature as a **reduced objective**. In the following, we will discuss two methods to characterize the synthesis of the signal $\mathbf{u}^*(\cdot)$ based on the solution of the dynamic optimization problem (23).

Using Calculus of Variations

It is possible to compute the optimal solution of problem (23) by using the Euler-Lagrange equations from calculus of variations. If you are not familiar with this technique, it is safe to skip to the next section on Pontryagin's Maximum Principle and attempt the example of this section using this technique.

Formally, we adjoin the nonlinear constraints through a time-dependent Lagrange multiplier $\mathbf{p}(t)$ leading to

$$\bar{\mathcal{J}} := \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} L(\mathbf{x}(t), \mathbf{u}(t)) + \mathbf{p}^\top(t) \left\{ \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)) - \dot{\mathbf{x}}(t) \right\} dt\,. \tag{24}$$

At this point, it is useful to define the **Hamiltonian**

$$\mathcal{H}(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t)) := L(t, \mathbf{x}(t), \mathbf{u}(t)) + \mathbf{p}^\top \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\,. \tag{25}$$

Integration by parts of the term $\mathbf{p}^\top \dot{\mathbf{x}}$ in the Lagrangian (24) yields

$$\bar{\mathcal{J}} := \Phi(\mathbf{x}(t_f), t_f) - \mathbf{p}^\top(t_f)\mathbf{x}(t_f) + \mathbf{p}^\top(t_0)\mathbf{x}(t_0) + \int_{t_0}^{t_f} H(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t)) + \dot{\mathbf{p}}^\top \mathbf{x}(t)\, dt\,. \tag{26}$$

[7] Although a natural approach for its numerical treatment would be to discretize the dynamics and treat (23) as a large scale nonlinear optimization problem.

Now, we compute the variation $\delta\bar{\mathcal{J}}$ due to variations in $\delta\mathbf{x}$ and $\delta\mathbf{u}$ obtaining

$$\delta\bar{\mathcal{J}} = \left[\left(\frac{\partial\Phi}{\partial\mathbf{x}} - \mathbf{p}^\top\right)\delta\mathbf{x}\right]_{t=t_f} + \left[\mathbf{p}^\top\delta\mathbf{x}\right]_{t=t_0} + \int_{t_0}^{t_f}\left[\left(\frac{\partial\mathcal{H}}{\partial\mathbf{x}} + \dot{\mathbf{p}}^\top\right)\delta\mathbf{x} + \frac{\partial\mathcal{H}}{\partial\mathbf{u}}\delta\mathbf{u}\right]dt\,. \tag{27}$$

Finding a stationary point of this functional by imposing $\delta\bar{\mathcal{J}} = 0$ over arbitrary variations $\delta\mathbf{x}$ and $\delta\mathbf{u}$ leads to the conditions

$$\mathbf{p}^\top = \frac{\partial\Phi}{\partial\mathbf{x}}, \quad \text{at } t = t_f\,, \tag{28}$$

$$-\dot{\mathbf{p}}^\top = \frac{\partial\mathcal{H}}{\partial\mathbf{x}}, \quad \text{for } t_0 \le t \le t_f\,, \tag{29}$$

$$\frac{\partial\mathcal{H}}{\partial\mathbf{u}} = 0\,, \quad \text{at every } t_0 \le t \le t_f\,. \tag{30}$$

Let's analyse the structure of the optimality system. Our first intention was to compute the optimal control $\mathbf{u}$, however we have introduced an additional adjoint variable $\mathbf{p}$. The optimality conditions tell us that the adjoint equation is in turn governed by the **backward** dynamical system (29) with terminal condition (28). This system is closed by the optimality condition (30) that must be satisfied at all times. In principle, the complete system provides a set of forward dynamics for $\mathbf{x}$ (the state equation with initial condition), backward dynamics for $\mathbf{p}$, and a final **static** relation to compute $\mathbf{u}$. However, these equations are fully coupled in a forward-backward structure, and need to be solved simultaneously in order to synthesize $\mathbf{u}(\cdot)$. In general, a problem of this type does not have a closed-form solution and we must resort to the use of computational methods for its realization. In the following we discuss a simple example where optimality conditions can be solved explicitly.

A simple one-dimensional problem. Consider the linear scalar system

$$\dot{\mathbf{x}} = a\mathbf{x} + b\mathbf{u}\,, \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}\,,$$

where a and b are constants, and the cost is given by

$$\mathcal{J} = \frac{1}{2}\int_{t_0}^{t_f}\mathbf{u}^2\,dt + \frac{1}{2}c\mathbf{x}(t_f)^2\,, \quad c > 0\,.$$

In this case it is easy to see that $L(\mathbf{x},\mathbf{u}) = \frac{1}{2}\mathbf{u}^2$ and that $\Phi(\mathbf{x}(t_f)) = \frac{1}{2}c\mathbf{x}(t_f)^2$. Therefore, the optimality system reads

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= a\mathbf{x}(t) + b\mathbf{u}(t)\,,\\
\mathbf{x}(t_0) &= \mathbf{x}_0\,,\\
-\dot{\mathbf{p}}(t) &= a\mathbf{p}(t)\,,\\
\mathbf{p}(t_f) &= c\mathbf{x}(t_f)\,,\\
0 &= \mathbf{u}(t) + b\mathbf{p}(t)\,, \quad \text{for all } t\,.
\end{aligned}$$

From the equations for $\mathbf{p}$ we can integrate and find

$$\mathbf{p}(t) = c\mathbf{x}(t_f)e^{a(t_f-t)} \Rightarrow \mathbf{u}(t) = -b\mathbf{p}(t) = -bc\mathbf{x}(t_f)e^{a(t_f-t)},$$

which using the forward equation leads to

$$\mathbf{x}(t) = \mathbf{x}(t_0)e^{a(t-t_0)} + \frac{b^2c}{2a}\mathbf{x}(t_f)\left[e^{a(t_f-t)} - e^{a(t+t_f-2t_0)}\right].$$

Evaluating this expression at $t = t_f$ allows to determine $\mathbf{x}(t_f)$,

$$\mathbf{x}(t_f) = \frac{2a\mathbf{x}(t_0)e^{a(t_f-t_0)}}{2a - b^2c(1-e^{2a(t_f-t_0)})}$$

obtaining the optimal control $\mathbf{u}^*(t)$ and the optimal state $\mathbf{x}^*(t)$

$$\mathbf{u}^*(t) = -\frac{2abc\mathbf{x}(t_0)e^{a(2t_f-t_0-t)}}{2a - b^2c(1-e^{2a(t_f-t_0)})} \tag{31}$$

$$\mathbf{x}(t) = \mathbf{x}(t_0)e^{a(t-t_0)} + \frac{b^2c\mathbf{x}(t_0)e^{a(t_f-t_0)}}{2a - b^2c(1-e^{2a(t_f-t_0)})}\left[e^{a(t_f-t)} - e^{a(t+t_f-2t_0)}\right]. \tag{32}$$

It is possible to analyse the influence of the cost function in the trajectories. For example, in the case $c \to \infty$, which means a strong penalty on the terminal state, we can see that from the optimal trajectory

$$\lim_{c\to\infty} \mathbf{x}^*(t_f) = 0,$$

regardless of t_f.

Pontryagin's Maximum Principle

The derivation of optimality conditions through Euler-Lagrange equations dates back to 1750, approximately. A refined version of this result derived about 200 years later in the middle of the Cold War in the Soviet side, where there was a growing interest in control theory driven by ballistic applications. This result is known as Pontryagin's Maximum Principle (or PMP), and characterizes optimality conditions for a problem of the type

$$\dot{\mathbf{x}}(t) = \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t)), \quad \mathbf{x}(t_0) = \mathbf{x}_0 \in \mathbb{R}^n, \mathbf{u} \in \mathcal{U} \subset \mathbb{R}^m, \quad t_0 \le t \le t_f, \tag{33}$$

with a cost functional, also known as the *Bolza problem*[8]

$$\mathcal{J}(\mathbf{x}, \mathbf{u}) := \Phi(\mathbf{x}(t_f), t_f) + \int_{t_0}^{t_f} L(\mathbf{x}(t), \mathbf{u}(t))\, dt \tag{34}$$

[8] There exists a formulation called the *Mayer problem* where the $\mathcal{J} := \mathcal{J}(\mathbf{x}(t_f), \mathbf{u})$, solely depending on the terminal state. Both formulations are equivalent by augmenting with an auxiliary state $\dot{\mathbf{z}}(t) = L(\mathbf{x}(t), \mathbf{u}(t))$.

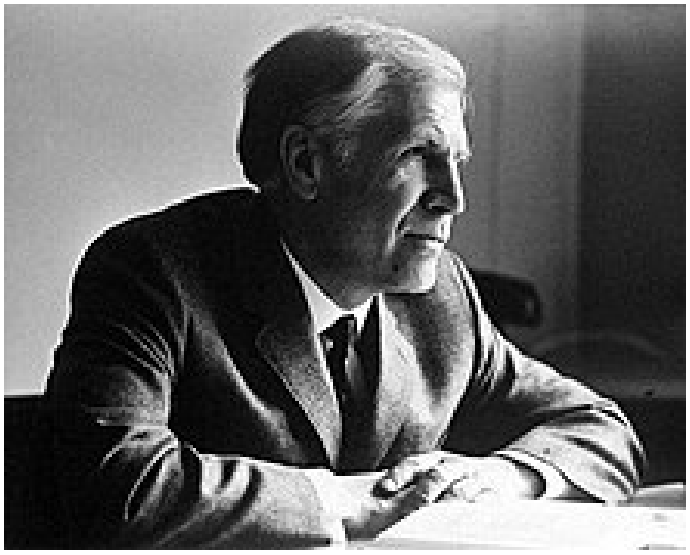


Figure 34: Lev Pontryagin (1908-1988), Soviet mathematician and one of the central figures of modern optimal control theory. He lost his eyesight when he was 14, and he also made remarkable contributions to algebraic and differential topology.

and terminal constraints

$$\Psi(\mathbf{x}(t_f)) = 0\,, \quad \Psi : \mathbb{R}^n \to \mathbb{R}^q\,. \tag{35}$$

There are three important differences from the first optimal control formulation we have discussed earlier. From now on, the terminal time t_f is allowed to be free, and we express a set of q terminal constraint for the final state through $\Psi(\mathbf{x}(t_f))$. However, the most remarkable difference in the PMP formulation, is the existence of a space of **admissible controls** $\mathcal{U}$ where we restrict our optimal control signal. Recalling Hamiltonian

$$\mathcal{H}(t, \mathbf{x}(t), \mathbf{u}(t), \mathbf{p}(t)) := L(t, \mathbf{x}(t), \mathbf{u}(t)) + \mathbf{p}^\top \mathbf{f}(t, \mathbf{x}(t), \mathbf{u}(t))\,, \tag{36}$$

the optimality conditions now read

$$\begin{aligned}
\dot{\mathbf{x}}(t) &= \frac{\partial \mathcal{H}}{\partial \mathbf{p}}\,,\\
-\dot{\mathbf{p}}(t) &= \frac{\partial \mathcal{H}}{\partial \mathbf{x}}\,,\\
\mathbf{x}(t_0) &= \mathbf{x}_0\,,\\
\Psi(\mathbf{x}(t_f)) &= 0\,,\\
\mathbf{p}(t_f) &= \left[\frac{\partial \Phi}{\partial \mathbf{x}} + \nu^\top \frac{\partial \Psi}{\partial \mathbf{x}}\right]_{t=t_f}\,,\\
\left[\frac{\partial \Phi}{\partial t} + \mathbf{p}^\top \mathbf{f} + L\right]_{t=t_f} &= 0\,,\\
\mathcal{H}(t, \mathbf{x}^*, \mathbf{u}^*(t), \mathbf{p}^*(t)) &\leq \mathcal{H}(t, \mathbf{x}^*, \mathbf{u}(t), \mathbf{p}^*(t))\,, \quad \text{for all } \mathbf{u} \in \mathcal{U}
\end{aligned} \tag{PMP}$$

The last condition states that the optimal control is the minimizer of the Hamiltonian *ceteris paribus*. In the smooth, unconstrained control case, this is equivalent to the previous

condition $\frac{\partial \mathcal{H}}{\partial \mathbf{u}} = 0$, but since we include the constraint $\mathbf{u} \in \mathcal{U}$, this condition must be realized as

$$\mathbf{u}^* \in \operatorname*{argmin}_{\mathbf{w} \in \mathcal{U}} \mathcal{H}(t, \mathbf{x}^*, \mathbf{w}, \mathbf{p}^*) \,.$$

We conclude with a simple linear example illustrating the differences arising in the constrained control case.

Bang-bang control of linear systems. We study a time-optimal control problem for linear dynamics. Consider the control system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^n$ and the control variable is constrained to $|\mathbf{u}| \leq 1$. Our objective is to minimize the time t_f to reach the origin departing from $\mathbf{x}(0) = \mathbf{x}_0$, that is $\mathbf{x}(t_f) = 0$. The cost functional and the terminal constraint are chosen as

$$\mathcal{J} := \int_0^{t_f} 1 \, dt \,, \quad \Psi(\mathbf{x}(t_f)) = \mathbf{x}(t_f) \,.$$

We begin by assembling the Hamiltonian

$$\mathcal{H} = 1 + \mathbf{p}^\top (\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}) \,,$$

and the optimality conditions read

$$\begin{aligned} \dot{\mathbf{x}} &= \frac{\partial \mathcal{H}}{\partial \mathbf{p}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} \,, \\ -\dot{\mathbf{p}} &= \frac{\partial \mathcal{H}}{\partial \mathbf{x}} = \mathbf{A}^\top \mathbf{p} \,, \\ \mathbf{u} &= \operatorname*{argmin}_{|\mathbf{w}| \leq 1} \{\mathbf{p}^\top \mathbf{B}\mathbf{w}\} = -\operatorname{sgn}(\mathbf{p}^\top \mathbf{B}) \,, \end{aligned}$$

from where we can infer that the optimal control will always take values in the boundary of the control set, i.e. $\mathbf{u}^*(t) = \{-1, 1\}$. In the upcoming section we will discuss how to realize the synthesis of optimal controllers via PMP in a more general way by resorting to computational optimization and numerical analysis tools.

The Linear-Quadratic Case

A very important instance class of control problems can be formulated assuming linear dynamics of the form

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \,, \quad \mathbf{x} \in \mathbb{R}^n \,, \mathbf{u} \in \mathbb{R}^m \,,$$

where $\mathbf{A} \in \mathbb{R}^{n \times n}$, $\mathbf{B} \in \mathbb{R}^{m \times n}$, and the cost function is quadratic

$$\mathcal{J}(\mathbf{x}, \mathbf{u}) = \frac{1}{2} \int_0^T \left[\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{u}^\top \mathbf{R}\mathbf{u} \right] \, dt + \frac{1}{2} \mathbf{x}(T)^\top \mathbf{S}\mathbf{x}(T) \,, \quad \mathbf{x}(0) = \mathbf{x}_0 \,.$$

We further assume that the matrices $\mathbf{Q} \in \mathbb{R}^{n\times n}$, $\mathbf{R} \in \mathbb{R}^{m\times m}$, and $\mathbf{S} \in \mathbb{R}^{n\times n}$ are symmetric, $\mathbf{Q}$ and $\mathbf{S}$, are positive semi-definite and $\mathbf{R}$ is positive definite (do you have an intuition about this condition?). We assume the control space $\mathcal{U}$ is the class of all measurable and a.e. bounded controls, and the horizon T is fixed. The Hamiltonian is given by

$$\mathcal{H} = \frac{1}{2}\left[\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{u}^\top \mathbf{R}\mathbf{u}\right] + \mathbf{p}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}),$$

and the optimality conditions read

$$\dot{\mathbf{x}} = \frac{\partial \mathcal{H}}{\partial \mathbf{p}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}, \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{37}$$

$$-\dot{\mathbf{p}} = \frac{\partial \mathcal{H}}{\partial \mathbf{x}} = \mathbf{Q}\mathbf{x} + \mathbf{A}^\top \mathbf{p}, \quad \mathbf{p}(T) = \mathbf{S}\mathbf{x}(T), \tag{38}$$

$$\mathbf{u} = \underset{\mathbf{w}}{\operatorname{argmin}} \{\frac{1}{2}\mathbf{w}^\top \mathbf{R}\mathbf{w} + \mathbf{p}^\top \mathbf{B}\mathbf{w}\} = -\mathbf{R}^{-1}\mathbf{B}^\top \mathbf{p}. \tag{39}$$

To solve this optimality system we can plug the optimal control expression (39) back to the forward eq. (37), leading to

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top \mathbf{p}, \\ -\dot{\mathbf{p}} &= \mathbf{Q}\mathbf{x} + \mathbf{A}^\top \mathbf{p}, \\ \mathbf{x}(0) &= \mathbf{x}_0, \\ \mathbf{p}(T) &= \mathbf{S}\mathbf{x}(T). \end{aligned} \tag{TPBVP}$$

These type of problems are known as two-point boundary value problems (TPBVP), and in general, are difficult to solve due to the nonlinear coupling between the variables. Note that the terminal state $\mathbf{x}(T)$ determines the terminal condition for the backward adjoint equation by $\mathbf{p}(T) = \mathbf{S}\mathbf{x}(T)$. The literature offer different alternatives for the computational solution of this optimality system, most notably:

1. Solving the TPBVP using some type of multiple-shooting method. You can check MATLAB's `bvp4c`, which is comprehensive package for the solution of TPBVP arising in control and elsewhere.
2. Discretizing the forward-backward dynamics $(\mathbf{x}(\cdot), \mathbf{p}(\cdot)) \approx \{\mathbf{x}_k, \mathbf{p}_k\}_{k=0}^{N_T}$, and solve a large-scale nonlinear system $\mathcal{N}(\mathbf{x}_k, \mathbf{p}_k) = 0$ using a Newton-type method.
3. Write a reduced cost $\mathcal{J}(\mathbf{x}, \mathbf{u}) = \mathcal{J}(\mathbf{x}(\mathbf{u}), \mathbf{u})$ and apply gradient descent for $\mathbf{u}$. Computing $\nabla_{\mathbf{u}}\mathcal{J}(\mathbf{x}(\mathbf{u}), \mathbf{u})$ can be done using the adjoint equation (38) and the optimality condition (39).

However, in the linear quadratic case we can follow a different approach, known in the literature as the Linear Quadratic Regulator (LQR). Since the dynamics are linear, we can assume that the adjoint is expressed as $\mathbf{p}(t) = \Pi(t)\mathbf{x}(t)$, where $\Pi(t) \in \mathbb{R}^{n\times n}$ and symmetric. Substituting in the adjoint this leads to

$$-\dot{\mathbf{p}} = -(\dot{\Pi}\mathbf{x} + \Pi\dot{\mathbf{x}}) = -\dot{\Pi}\mathbf{x} - \Pi(\mathbf{A}\mathbf{x} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top \mathbf{p}) = \mathbf{Q}\mathbf{x} + \mathbf{A}^\top \mathbf{p},$$

which is a relation to be satisfied for all $\mathbf{x}$, and hence it is equivalent to the **Differential Riccati Equation**

$$-\dot{\Pi} = \Pi\mathbf{A} + \mathbf{A}^\top\Pi - \Pi\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top\Pi + \mathbf{Q}\,, \quad \Pi(T) = \mathbf{S}\,. \tag{DRE}$$

This is a matrix differential equation, and a first numerical naive treatment is to use a numerical integrator componentwise for n^2 entries (it can be halved by symmetry), which can become very costly for large scale dynamics. More sophisticated approaches exploit low-rank properties of the matrices involved. Once the solution $\Pi(t)$ has been computed, the optimal control can be recovered as

$$\mathbf{u}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\mathbf{p}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\Pi(t)\mathbf{x}(t)\,.$$

However, appreciating the true nature of the optimal control $\mathbf{u}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\Pi(t)\mathbf{x}(t)$, we observe it corresponds to an **optimal linear feedback map** of $\mathbf{x}$, which is computed regardless of the initial state $\mathbf{x}_0$. In the linear-quadratic case, the use of the PMP leads to an optimal feedback control.

The Algebraic Riccati Equation

Let us recall that for a finite horizon problem with linear dynamics and quadratic cost function, the optimal control law is obtained by solving the Differential Riccati Matrix Equation

$$-\dot{\Pi} = \Pi\mathbf{A} + \mathbf{A}^\top\Pi - \Pi\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top\Pi + \mathbf{Q}\,, \quad \Pi(T) = \mathbf{S}\,. \tag{DRE}$$

Once the solution $\Pi(t)$ has been computed, the optimal control is given by

$$\mathbf{u}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\mathbf{p}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\Pi(t)\mathbf{x}(t)\,.$$

As we have already noted, the optimal control $\mathbf{u}(t) = -\mathbf{R}^{-1}\mathbf{B}^\top\Pi(t)\mathbf{x}(t)$ is expressed in feedback form. This is somehow better than what we were originally looking for, as this control can computed regardless of the initial state $\mathbf{x}_0$. This is a very particular instance of a problem where the PMP technique yields an **optimal feedback control**. An important asymptotic case is obtained when $T \longrightarrow \infty$ (with $\mathbf{S} = 0$), here the optimal control is given by

$$\mathbf{u}(\mathbf{x}) = -\mathbf{K}\mathbf{x} = -\mathbf{R}^{-1}\mathbf{B}^\top\Pi\mathbf{x}\,,$$

where the **Kalman** gain $\mathbf{K}$ is determined by Π, which is the steady state solution of (DRE), also known as the **Algebraic Riccati Equation**

$$\Pi\mathbf{A} + \mathbf{A}^\top\Pi - \Pi\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top\Pi + \mathbf{Q} = 0\,. \tag{ARE}$$

In general, there's no unique solution to this equation, as we need further structural assumptions to guarantee the existence of a unique stabilizing Π. Stabilizing here means that the spectrum of the closed-loop operator $(\mathbf{A} - \mathbf{B}\mathbf{R}^{-1}\mathbf{B}^\top\Pi)$ is in the left half-plane.

Figure 35: On the right, Rudolf Emil Kálmán (1930-2016), Hungarian-American engineer and mathematician, who made seminal contributions to control theory including the LQR. His control theoretical methods were implemented in the navigation system of the Apollo developed by Margareth Hamilton (b. 1936). In 2009, he received the National Medal of Science from Barack Obama.

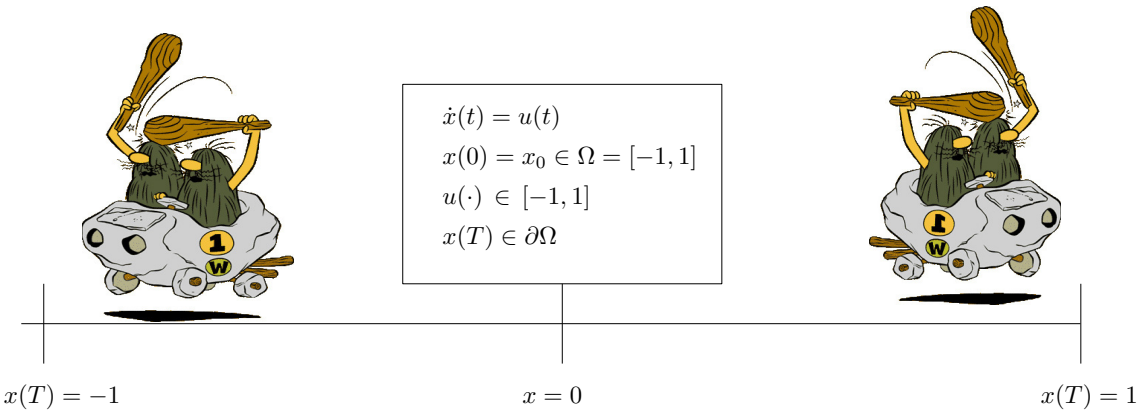


Figure 36: Given the linear dynamics $\dot{\mathbf{x}}(t) = \mathbf{u}(t)$, we want to minimize the arrival time to $\partial\Omega = \{-1, 1\}$ with a control $\mathbf{u}(t) \in [-1, 1]$.

Optimal Feedback Control & Dynamic Programming

Let's start by revisiting the bang-bang control example depicted in Figure 36. Here, the cost function we want to minimize is the arrival time of the cart to the boundary of $\Omega = [-1, 1]$, which we we express as

$$\min_{\mathbf{u}(\cdot)\in\mathcal{U}} T(\mathbf{u})$$
$$\text{subject to} \qquad \dot{\mathbf{x}}(t) = \mathbf{u}(t)\,, \quad \mathbf{x}(0) = \mathbf{x}_0\,, \quad \mathbf{x}(T) \in \partial\Omega\,, \quad \mathbf{u}(t) \in [-1, 1]\,.$$

The two questions we want to solve from an optimal control viewpoint are: what is the optimal cost T and the optimal control law $\mathbf{u}$? We have already studied the answer to these questions through calculus of variations and Pontryagin's Maximum Principle. We will now adopt a **Dynamic Programming** perspective, relating these questions to the

solution of the partial differential equation

$$\|\nabla T(\mathbf{x})\| = 1, \quad \mathbf{x} \in \Omega = [-1, 1]\,. \tag{40}$$

It is easy to see that for the simple dynamics $\dot{\mathbf{x}} = \mathbf{u}(t)$ the solution of the minimum time control problem is to drive to the left at full speed $\mathbf{u} = -1$ is the initial position is negative, and the right at speed $\mathbf{u} =$ if the initial position is positive. From here it follows that the optimal cost is the distance function to the boundary of Ω. This distance function can be in fact identified with the solution of the PDE (40), as depicted in Figure 54.

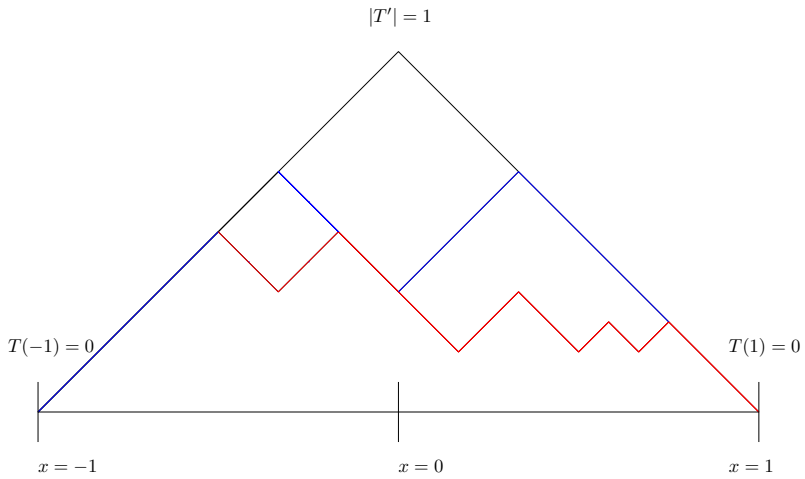


Figure 37: The Eikonal equation $\|\nabla T(\mathbf{x})\| = 1$ poses different PDE-type difficulties: no classical solutions and non-uniqueness of weak solutions. It's solution must be understood in the **viscosity solution** sense.

However, we must proceed with care. Derivatives are not properly defined at $\mathbf{x} = -1, 0, 1$ and in fact there are infinitely many sawtooth-type functions that satisfy both the boundary conditions $T(-1) = T(1) = 0$ and the PDE $\|\nabla T(\mathbf{x})\| = 1$ in the weak sense, and none of them correspond to the distance function we are looking for. To define a concept of solution in this context, we must talk about viscosity solutions, that is, solutions to the PDE in a vanishing viscosity limit

$$\|\nabla T(\mathbf{x})\| = 1 + \epsilon \Delta T_\epsilon \longrightarrow T_\epsilon(\mathbf{x}) = 1 - \|\mathbf{x}\| + \epsilon \left(e^{-1/\epsilon} - e^{-\|\mathbf{x}\|/\epsilon} \right)\,.$$

Having identified a relation between solving a PDE and finding the optimal cost of the problem, we now turn our attention to finding the optimal control law based on the PDE. We observe that in fact, is $T(\mathbf{x})$ is the hat function depicted in Figure (54), then its gradient (wherever it exists) coincides with the optimal control. Moreover, because of this relation, the optimal control is naturally expressed as a function of $\mathbf{x}$, i.e., as a feedback law. This synthesizes the fundamental philosophy behind the dynamic approach: we will establish a link between the solution of an optimal control problem and the solution of a nonlinear partial differential equation, and after solving the PDE, we will recover the optimal feedback law as a by-product.

The Hamilton-Jacobi-Bellman PDE

A central idea behind the synthesis of optimal feedback laws is the formulation of a partial differential equation governing the optimal cost of our problem, from which the optimal feedback map is directly recovered as a by-product. This very particular partial differential equation is called the Hamilton-Jacobi-Bellman PDE, and is central object of study in optimal control theory and reinforcement learning. It is often written as

$$\begin{cases} \mathcal{H}(\mathbf{x}, V, \nabla V) = 0 & \text{in } \Omega \subset \mathbb{R}^d, \\ V(\mathbf{x}) = b(\mathbf{x}) & \text{in } \partial\Omega, \end{cases}$$

where $\mathcal{H}(\mathbf{x}, V, \nabla V)$ is the Hamiltonian (replacing $\mathbf{p}$ by ∇V), whose selection gives origin to different PDEs/control problems:

- Eikonal equations:
$$\mathcal{H}(\mathbf{x}, \nabla V) = \|\nabla V\| - I(\mathbf{x}) .$$
- Minimum time control:
$$\mathcal{H}(\mathbf{x}, \nabla V) = \sup_{\mathbf{u} \in U} \left[-f(\mathbf{x}, \mathbf{u}) \cdot \nabla V - 1 \right] .$$
- Infinite horizon optimal control:
$$\mathcal{H}(\mathbf{x}, V, \nabla V) = \sup_{\mathbf{u} \in U} \left[\lambda V - f(\mathbf{x}, \mathbf{u}) \cdot \nabla V - \ell(\mathbf{x}, \mathbf{u}) \right] .$$
- Isaacs equation (differential games):
$$\mathcal{H}(\mathbf{x}, V, \nabla V) = \sup_{\mathbf{u} \in U} \inf_{\mathbf{w} \in W} \left[\lambda V - f(\mathbf{x}, \mathbf{u}, \mathbf{w}) \cdot \nabla V - \ell(\mathbf{x}, \mathbf{u}, \mathbf{w}) \right] .$$

In general, whenever we attempt the synthesis of an optimal feedback law for a deterministic continuous dynamical system, we first find the associated HJB PDE, and then we proceed to numericall approximate its solution either by classical, grid-based techniques (finite differences, semi-Lagrangian schemes) or by deep neural networks (deep reinforcement learning). From a computational perspective, there are two fundamental difficulties in the numerical of solution of HJB PDEs: they are fully nonlinear PDEs (minimization with respect to u), and they are written over a domain $\Omega \subset \mathbb{R}^d$, where $\mathbb{R}^d$ is the state space of the control system, which can be arbitrarily large. Richard Bellman, the founding father of dynamic programming, saw this limitation at a very early stage, coining the term *curse of dimensionality*, when referring to the overwhelming computational complexity associated to the numerical solution of HJB PDEs.

Figure 38: Richard Bellman (1920-1984), a central figure in modern control theory. Regarding the term dynamic programming -which he invented- he said: *I thought "dynamic programming" was a good name. It was something not even a Congressman could object to. So I used it as an umbrella for my activities* (from R.E. Bellman, " Eye of the Hurricane").

Dynamic Programming and Minimum Time Control

We consider nonlinear system dynamics of the form

$$\begin{cases} \dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t), \mathbf{u}(t)), \\ \mathbf{y}(0) = \mathbf{x}, \end{cases}$$

where $\mathbf{f} : \mathbb{R}^d \times U \to \mathbb{R}^d$ is continuous. The situation is depicted in Figure 39. Given a target set $\mathcal{T} \subset \mathbb{R}^d$, we want to find the optimal control signal $\mathbf{u}(t)$ such that its boundary $\partial\mathcal{T}$ is reached in minimum time from a departure point $\mathbf{x}$. To study this problem with

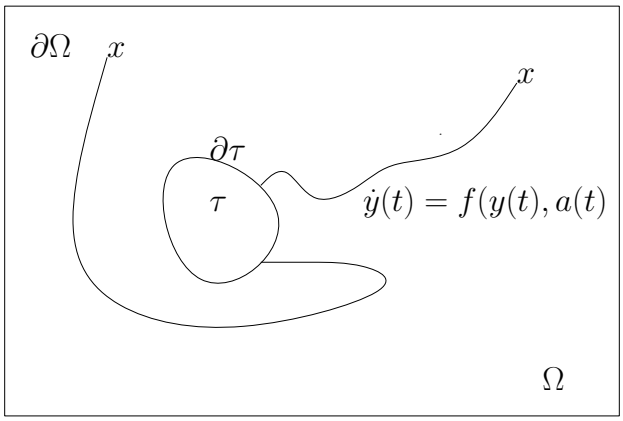


Figure 39: The minimum time control problem.

dynamic programming techniques, we must define the minimum time function:

$$T(\mathbf{x}) = \inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \{ t > 0 : \mathbf{y}_\mathbf{x}(t, \mathbf{u}) \in \mathcal{T} \},$$

where the notation $\mathbf{y}_{\mathbf{x}}(t, \mathbf{u})$ stands for the trajectory followed by $\mathbf{y}$ departing from $\mathbf{y}(0) = \mathbf{x}$ through the action of the control signal $\mathbf{u}$. Under the local controllability condition

$$\inf_{\mathbf{u} \in U} \mathbf{f}(\mathbf{x}, \mathbf{u}) \cdot n(\mathbf{x}) < 0, \quad \forall \mathbf{x} \in \partial \mathcal{T},$$

where $n(\mathbf{x})$ is the vector normal to $\partial \mathcal{T}$, the minimum time function $T(\mathbf{x})$ is continuous over the reachable set $\mathcal{R}$, the set of states with $T(\mathbf{x}) < \infty$. In dynamic programming, **we look for a functional relation characterizing** $T(\mathbf{x})$, (that is, a global approach). How do we find a HJB-type PDE? Here, we follow the steps of Bellman and the dynamic programming principle. If we look at the isochrone map in Figure 40, it depicts the fastest travel times from anywhere in the world to London in 1914. This a minimum time function.

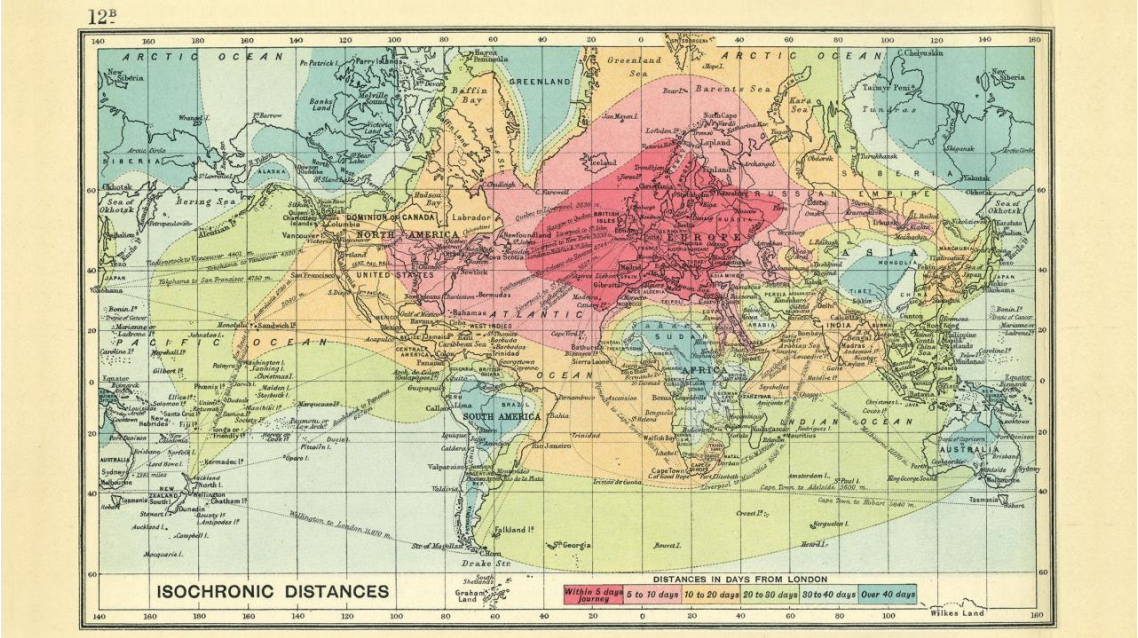


Figure 40: An isochrone map from 1914 indicating the travel times to London (in days), assuming the shortest path is taken. What is the optimal direction as a feedback map based on the minimum time function?

How can we generate such a plot? An extremely inefficient way to generate this plot would be to solve the fastest travel route (and compute its travel time) for each point in the globe. Instead, Bellman proposed the use of a **backward induction** procedure. Let's start in London. Travel time from London to London is 0. Next, we identify the set of points whose fastest travel route can reach London in one hour. This defines a **level set** $T(\mathbf{x}) = 1$ for the minimum time function. Now, if we want to compute the set of points reaching London (with an optimal route) within two hours, rather than blindly solving this problem, we solve the problem of finding the outer points whose optimal travel time to the level set $T(\mathbf{x}) = 1$ is of one hour. This is Bellman's optimality principle: optimal trajectories are such that they can be split at any point along the trajectory, and the second half is always the optimal trajectory for this departure point (otherwise, the trajectory wouldn't be optimal). In mathematical language, this is expressed as:

Theorem (Dynamic Programming Principle for Minimum Time Control)**.** *For all* $\mathbf{x}$ *in the*

reachable set $\mathcal{R}$ *and* $\tau \in (0, T(\mathbf{x}))$*, it holds*

$$T(\mathbf{x}) = \inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \{\tau + T(\mathbf{y}_\mathbf{x}(\tau, \mathbf{u}))\}\,.$$

Formally speaking, we can take divide this equation by τ and taking the limit $\tau \to 0$, the expression

$$\lim_{\tau\to 0} \frac{T(\mathbf{y}_\mathbf{x}(\tau, \mathbf{u})) - T(\mathbf{x})}{\tau} = \nabla T(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u})\,,$$

i.e., the directional derivative along the dynamics. This leads to a more general version of the Eikonal PDE, namely, the HJB equation for minimum time control:

$$\sup_{\mathbf{u}\in U}\{-\nabla T(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u})\} = 1 \quad \text{in } \Omega\,, \qquad T(\mathbf{x}) = 0 \quad \text{in } \partial\mathcal{T}\,.$$

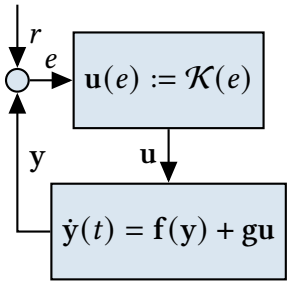
Once we solve this partial differential equation, the optimal feedback map corresponds to the minimizer of the Hamiltonian above, that is

$$\mathbf{u}^*(x) = \operatorname*{argmax}_{\mathbf{u}\in U}\{-\nabla T(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}, \mathbf{u})\}\,.$$

Infinite Horizon Optimal Control

The same idea can be applied for the infinite horizon optimal control problem:

$$\begin{aligned} \underset{\mathbf{u}(\cdot)\in\mathcal{U}}{\text{minimize}} \quad & \mathcal{J}(\mathbf{u}(\cdot), \mathbf{x}) := \int_0^\infty \ell(\mathbf{y}(t), \mathbf{u}(t))\, dt \\ \text{subject to} \quad & \dot{\mathbf{y}}(t) = \mathbf{f}(\mathbf{y}(t)) + \mathbf{g}\mathbf{u}(t) \\ & \mathbf{y}(0) = \mathbf{x} \in \mathbb{R}^d \end{aligned}$$



Dynamic Programming: the value function $V(\mathbf{x}) := \inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \mathcal{J}(\mathbf{u}, \mathbf{x})$ can be decomposed in

$$\begin{aligned} V(\mathbf{x}) &= \inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \left\{ \int_0^\tau \ell(\mathbf{y}(t), \mathbf{u}(t))\, dt + \int_\tau^\infty \ell(\mathbf{y}(t), \mathbf{u}(t))\, dt \right\}, \\ &= \inf_{\mathbf{u}(\cdot)\in\mathcal{U}} \left\{ \int_0^\tau \ell(\mathbf{y}(t), \mathbf{u}(t))\, dt + V(\mathbf{y}_\mathbf{x}(\tau, \mathbf{u})) \right\}, \end{aligned}$$

and dividing by τ and taking the limit $\tau \to 0$ leads to the the **Hamilton-Jacobi-Bellman** equation

$$\min_{\mathbf{u}\in U} \left[(\mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u})^\top \nabla V(\mathbf{x}) + \ell(\mathbf{x}, \mathbf{u})\right] = 0\,, \quad V(\mathbf{0}) = 0\,.$$

The optimal control is given in **feedback** (state-dependent) form:

$$u^*(\mathbf{x}) = \mathcal{K}(\mathbf{x}) := \operatorname*{argmin}_{\mathbf{u}\in U} \left[(\mathbf{f}(\mathbf{x}) + \mathbf{g}\mathbf{u})^\top \nabla V(\mathbf{x}) + \ell(\mathbf{x}, \mathbf{u})\right]\,.$$

Exercise. Do the calculations for the LQ case, for the HJB PDE you should arrive to the Algebraic Riccati Equation. You need to assume $f(\mathbf{x}) = A\mathbf{x}$, $g(\mathbf{x}) = B$, $\ell(\mathbf{x}, \mathbf{u}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{u}^\top R\mathbf{u}$, $V(\mathbf{x}) = \mathbf{x}^\top \Pi \mathbf{x}$ and $U = \mathbb{R}^m$.

References

[1] M. Bardi and I. Capuzzo-Dolcetta *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations.* Birkhäuser, Boston, 2008.

[2] A. Bressan and B. Piccoli *Introduction to the Mathematical Theory of Control.* AIMS Series on Applied Mathematics, 2007.

[3] Richard M. Murray *Optimization-based Control.* available online.