BGSM/CRM
AL&DNN

Convex optimization techniques

S. Xambó

UPC & IMTech

5/10/2021

Sketches on convex optimization theory.

We will define what is meant by a constrained optimization problem, introduce some relevant concepts (mainly the Lagrangian formalism) and study some useful results, with the aim to prove the Karush-Kuhn-Tucker theorem and, under suitable assumptions, the equivalence of the primal and dual versions of a convex optimization problem.

Main references: [1], [2], [3].

The historical unfolding of optimization techniques and their significance on multiple fields are fascinating topics.

A good reference for the facts (researchers, results, impacts) is the book [4], as it collects eighteen classical papers headed by an extensive historical paper by the editors.

In particular, we can find the celebrated (although unpublished) master thesis of W. Kurush (University of Chicago, 1939) and seminal papers by J. Lagrange (1759), F. John (1948), W. Fenchel (1949), M. Slater (1950) and the landmark paper by H. W. Kuhn and A. W. Tucker (1951).

It also includes an historical overview up to 1976 by H. W. Kuhn.

The bibliography notes at the end of each chapter of the treatise [1] are also an excellent source of facts and ideas about those topics.

The space of real $r \times c$ matrices (r rows and c columns) will be denoted by $\mathbf{R}_c^r$; the space of row vectors of length r, by $\mathbf{R}^r$ (it is isomorphic to $\mathbf{R}_1^r$); and the space of column vectors of length c, by $\mathbf{R}_c$ (it is isomorphic to $\mathbf{R}_c^1$).

We have a bilinear map $\mathbf{R}_c^r \times \mathbf{R}_s^c \to \mathbf{R}_s^r$ given by the product of matrices.

The square matrices of order k, $\mathbf{R}(k) = \mathbf{R}_k^k$, form an $\mathbf{R}$-algebra.

The algebra $\mathbf{R}(1)$ is isomorphic to $\mathbf{R}$.

Note that the map $\mathbf{R}^r \times \mathbf{R}_r \to \mathbf{R}$, $(x, y) \mapsto xy$ is a duality: it allows to identify $\mathbf{R}^r$ as the dual of $\mathbf{R}_r$ and $\mathbf{R}_r$ as the dual of $\mathbf{R}^r$.

If m is a positive integer, $[m]$ denotes the sequence $1, \cdots, m$.

Optimizations problems

If $\mathcal{X} \subseteq \mathbf{R}^n$ and we have functions $f, g_i : \mathcal{X} \to \mathbf{R}$ ($i \in [m]$), then we can consider the problem

$$p^* = \min_{x \in \mathcal{X}} f(x) \text{ ⅄ } g_i(x) \leqslant 0 \ (i \in [m]). \tag{1}$$

(we use the symbol ⅄ to be read as "subject to")

Such problems are said to be *constrained optimization problems* (in *primal* form). If there are no constraints g_i, the problem $\min_{x \in X} f(x)$ is said to be *unconstrained*.

The set $X = \{x \in \mathcal{X} \,|\, g(x) \leqslant 0\}$ is called the *feasible set*.

If $X = \emptyset$, the problem (1) is inconsistent, so we will assume that X is non-empty.

The function f is called the *cost* or *objective* of the problem. Any $x^* \in X$ such that $f(x^*) = p^*$ will be said to be an *optimal solution* of (1).

Other problems that are amenable to the form (1) are also called optimization problems.

For example, conditions such as $a_i \leqslant g_i(x) \leqslant b_i$ (a_i, b_i constants) can be expressed as $g_i(x) - b_i \leqslant 0$ and $a_i - g_i(x) \leqslant 0$.

Likewise, $g_i(x) = 0$ is equivalent to $g_i(x) \leqslant 0$ and $-g_i(x) \leqslant 0$.

Note also that $\max_{x \in X} f(x) = -\min_{x \in X} -f(x)$, and hence max problems are also optimization problems. In such cases, f is usually called by names suggested by the context.

- $\min_{x\in\mathbf{R}^n} \|xA-b\|$, where $x\in\mathbf{R}^n$, $A\in\mathbf{R}^n_k$, and $b\in\mathbf{R}^k$.

This unconstrained optimization problem is equivalent to

$$\begin{aligned}\min_{x\in\mathbf{R}^n} \|xA-b\|^2 &= \min_{x\in\mathbf{R}^n}(xA-b)(A^Tx^T-b^T)\\ &= \min_{x\in\mathbf{R}^n}(xAA^Tx^T-2bA^Tx^T+bb^T).\end{aligned}$$

If the linear system $xA=b$ has solutions, then the minimum is 0 and it is achieved for all points on the affine subset of solutions.

Otherwise, we have to minimize a quadratic equation in x whose gradient with respect to x is twice xAA^T-bA^T. If AA^T is invertible, then there is a unique solution, namely $x=(bA^T)(AA^T)^{-1}$.

Otherwise, the solutions of $xAA^T=bA^T$ can be obtained by means of the matrix $(AA^T)^\dagger$, the pseudoinverse of AA^T, as discussed in numerical analysis texts (see, for example, [5, Lecture 11], and Appendix SVD, page 53, for a synopsis).

A map $\phi : \mathcal{X} \to \mathbf{R}_{n'}$ is called a *feature map* (more on this notion in the session 10-13).

Given a labeled dataset $\mathcal{D} = \{(x^i, y^i) \mid i \in [m], x^i \in \mathcal{X}, y^i \in \mathbf{R}\}$, let $L_{\mathcal{D}} : \mathbf{R} \times \mathbf{R}^{n'} \to \mathbf{R}$ be defined by

$$L_{\mathcal{D}}(w_0, w') = \sum_{i=1}^{m} (w'\phi(x^i) + w_0 - y^i)^2.$$

The quest for a solution $w = (w_0, w')$ to the optimization problem $\min_w L_{\mathcal{D}}(w)$ can be called a ϕ-*based linear regression problem* (this generalizes the notion of linear regression discussed earlier). It can be solved by means of least squares as follows.

If we let $y = (y^1, \cdots, y^m)$ and

$$A = \begin{bmatrix} 1 & \cdots & 1 \\ \phi(x^1) & \cdots & \phi(x^m) \end{bmatrix} \in \mathbf{R}_m^{n'+1}$$

then we have

$$\begin{aligned} L_{\mathcal{D}}(w) &= \|wA - y\|^2 \\ &= (wA - y)(A^T w^T - y^T) \\ &= wAA^T w^T - 2yA^T w^T + yy^T. \end{aligned}$$

Thus we can proceed much as in the previous example.

With the notations of the previous example, *ridge regression* is the unconstrained minimization problem of the form $\min_w L(w) + \lambda\|w\|^2$, where λ is some positive constant.

Since the gradient of $\lambda\|w\|^2 = \lambda w w^T$ is $2\lambda w$, we are led to solve

$$w(AA^T + \lambda \mathrm{Id}_{n'+1}) = yA^T,$$

which can be approached much as in the previous Example.

The role of the term $\lambda\|w\|^2$ is to favor solutions with small $\|w\|$. For $\lambda = 0$, we have linear regression. The higher the value of the chosen λ, the more it favors smaller $\|w\|$.

Ridge regression is also called *penalized*, or *regularized*, linear regression.

The Lagrangian formalism

The *Lagrangian* associated to the problem

$$p^* = \min_{x \in \mathcal{X}} f(x) \text{ ⅄ } g_i(x) \leqslant 0 \ (i \in [m])$$

is the function $L : \mathcal{X} \times \mathbf{R}^m_+ \to \mathbf{R}$ defined by the formula

$$L(x, u) = f(x) + u \cdot g(x) = f(x) + \sum_{i \in [m]} u_i g_i(x). \tag{2}$$

The symbols $u = u_1, \ldots, u_m$ are called *Lagrange* or *dual variables*.

Remark. For a short introduction to the *Lagrange multipliers* in analysis and differential geometry, see Appendix C, page 56.

By definition, the dual variable u_i of an inequality constraint $g_i(x) \leqslant 0$ is required to be non-negative.

Now recall that equality constraints $h(x) = 0$ can be handled as two inequalities, $h(x) \leqslant 0$ and $-h(x) \leqslant 0$. To these inequalities there correspond two dual (non-negative) variables, say u^+ and u^-, respectively.

They contribute to the Lagrangian with the expression

$$u^+ h(x) - u^- h(x) = (u^+ - u^-)h(x) = vh(x),$$

where $v = u^+ - u^- \in \mathbf{R}$ is free.

In general, if the constraints of a problem include equalities $h_j(x) = 0$ ($j \in [p]$) in addition to the inequalities $g_i(x) \leqslant 0$ ($i \in [m]$), then the Lagrangian $L(x, u, v)$ ($u = u_1, \cdots, u_m$ non-negative real variables, $v = v_1, \cdots, v_p$ free real variables) is expressed by the formula

$$L(x, u, v) = f(x) + u \cdot g(x) + v \cdot h(x).$$

The function

$$f^*(u,v) = \inf_{x\in\mathcal{X}} L(x,u,v)$$
$$= \inf_{x\in\mathcal{X}} \left(f(x) + \sum_{i=1}^{m} u_i g_i(x) + \sum_{j=1}^{p} v_j h_j(x) \right)$$

is called the *Lagrangian dual function*.

The *dual* problem of the *primal* problem

$$p^* = \min_{x\in\mathcal{X}} f(x) \text{ s.t. } g_i(x) \leqslant 0, h_j(x) = 0 \ (i \in [m], j \in [p])$$

is

$$d^* = \max f^*(u,v) \text{ s.t. } u \in \mathbf{R}^m_+, v \in \mathbf{R}^p.$$

In particular, the dual problem of (1) is

$$d^* = \max f^*(u) \text{ s.t. } u \in \mathbf{R}^m_+.$$

When there are only equality constraints $h(x)$, we will simply write $L(x, v) = f(x) + v \cdot h(x)$. In this case the dual function is $f^*(v) = \min_{x \in \mathcal{X}} L(x, v)$ and the dual problem is $d^* = \max_v f^*(v)$.

Remark. In what follows our focus will be mainly on Lagrangians $L(x, u)$ with only inequality constraints. This symplifies the exposition and is not really restrictive, for all results are valid, with straightfoward adaptation of their proofs, for the more general Lagrangians $L(x, u, v)$ (see Appendix D, page 58).

Theorem (Weak duality) $d^* \leqslant p^*$.

Proof. For any feasible x ($x \in X$) and $u \geqslant 0$, we have

$$f^*(u) \leqslant L(x,u) = f(x) + u \cdot g(x) \leqslant f(x),$$

as $u \cdot g(x) \leqslant 0$. Thus $f^*(u) \leqslant \inf_{x \in X} f(x) = p^*$ and hence $d^* = \max_{u \geqslant 0} f^*(u) \leqslant p^*$. □

A point $(\bar{x}, \bar{u}) \in \mathcal{X} \times \mathbf{R}^m_+$ is said to be a *saddle point* of the Lagrangian if it satisfies the inequalities:

$$L(\bar{x}, u) \leqslant L(\bar{x}, \bar{u}) \leqslant L(x, \bar{u}) \tag{3}$$

for all $x \in \mathcal{X}$ and $u \in \mathbf{R}^m_+$. See figure 7.1.

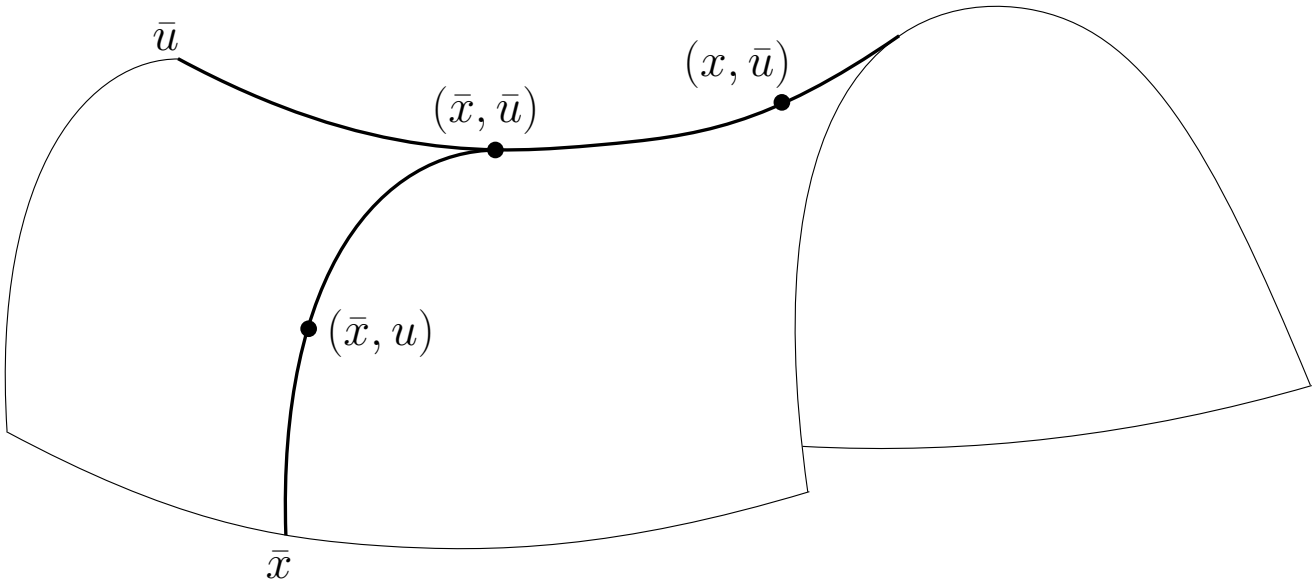


Figure 7.1: Saddle point of the Lagrangian.

Theorem (Strong duality) Suppose $(\bar{x}, \bar{u}) \in \mathcal{X} \times \mathbf{R}^m_+$ is a saddle point of the Lagrangian. Then $\bar{x}$ and $\bar{u}$ solve the primal and dual problems, respectively, and $f(\bar{x}) = f^*(\bar{u})$.

Proof. From the first inequality in (3), $f(\bar{x}) + u \cdot g(\bar{x}) \leqslant L(\bar{x}, \bar{u})$, which holds for all $u \geqslant 0$, we infer immediately that $g(\bar{x}) \leqslant 0$, hence that $\bar{x}$ is feasible.

Still from that first inequality we get, setting $u = 0$, that $f(\bar{x}) \leqslant f(\bar{x}) + \bar{u} \cdot g(\bar{x})$, which yields $\bar{u} \cdot g(\bar{x}) \geqslant 0$ and hence $\bar{u} \cdot g(\bar{x}) = 0$ (this is called *slack equation of the saddle point*). Notice that it is equivalent to $\bar{u}_i g_i(\bar{x}) = 0$ for all $i \in [m]$ (*slack relations*).

Now look at the second inequality in (3). From $f(\bar{x}) = L(\bar{x}, \bar{u})$ (by slack equation), we get $f(\bar{x}) \leqslant L(x, \bar{u}) = f(x) + \bar{u} \cdot g(x) \leqslant f(x)$ for all $x \in X$, which shows that $\bar{x}$ is optimal and $f(\bar{x}) \leqslant f^*(\bar{u})$. Thus $p^* = f(\bar{x}) \leqslant f^*(\bar{u}) \leqslant d^*$. By weak duality, these inequalities must be equalities and this completes the proof. □

It is thus desirable to seek conditions ensuring that the Lagrangian has a saddle point.

For this to happen, the strategy is to find hypotheses on the constraints (generally called *constraint qualifications*) that are sufficient to guarantee that the Lagrangian has a saddle point.

And to solve this in a setting that is suitable for our purposes we need some background notions to which we turn next.

Convex sets and functions

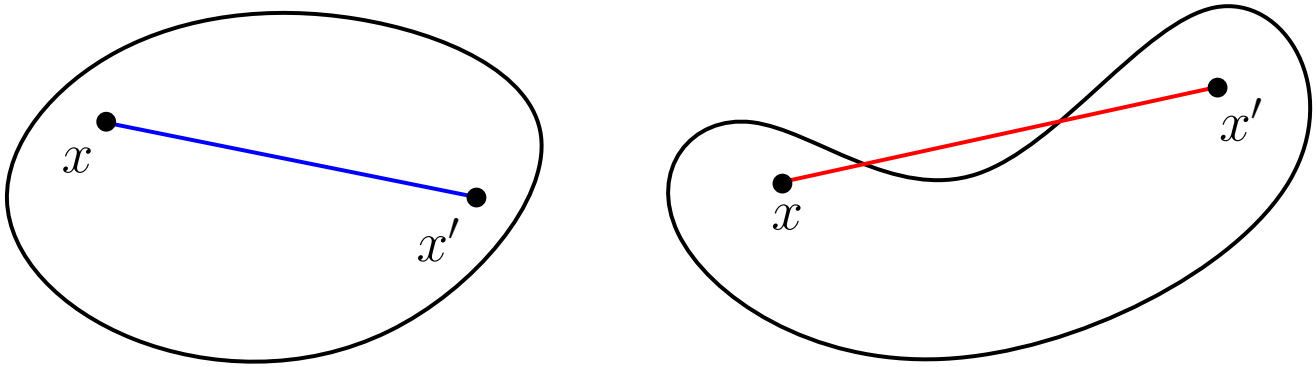


Figure 9.1: Left: Convex set. Right: Nonconvex set.

Theorem (Convex domains) A set $\mathcal{X} \subseteq \mathbf{R}^n$ is said to be *convex* if $\lambda x + (1-\lambda)x' \in \mathcal{X}$ for all $x, x' \in \mathcal{X}$ and $\lambda \in [0,1]$.

Example: If $g_i : \mathbf{R}^n \to \mathbf{R}$ are affine, then $\mathcal{X} = \{x \in \mathbf{R}^n \mid g(x) \leqslant 0\}$ is convex, for in this case, for all $i \in [m]$,

$$g_i(\lambda x + (1-\lambda)x') = \lambda g_i(x) + (1-\lambda)g_i(x') \leqslant 0$$

if $g_i(x), g_i(x') \leqslant 0$.

Theorem (Convex functions) If $\mathcal{X} \subseteq \mathbf{R}^n$, a function $f : \mathcal{X} \to \mathbf{R}$ is said to be *convex* if $\mathcal{X}$ is convex and

$$f\left(\lambda x + (1-\lambda)x'\right) \leqslant \lambda f(x) + (1-\lambda)f(x') \tag{4}$$

for all $x, x' \in \mathcal{X}$ and $\lambda \in [0,1]$.

If the inequality is strict when $x \neq x'$ and $\lambda \in (0,1)$, then f is said to be *strictly convex*.

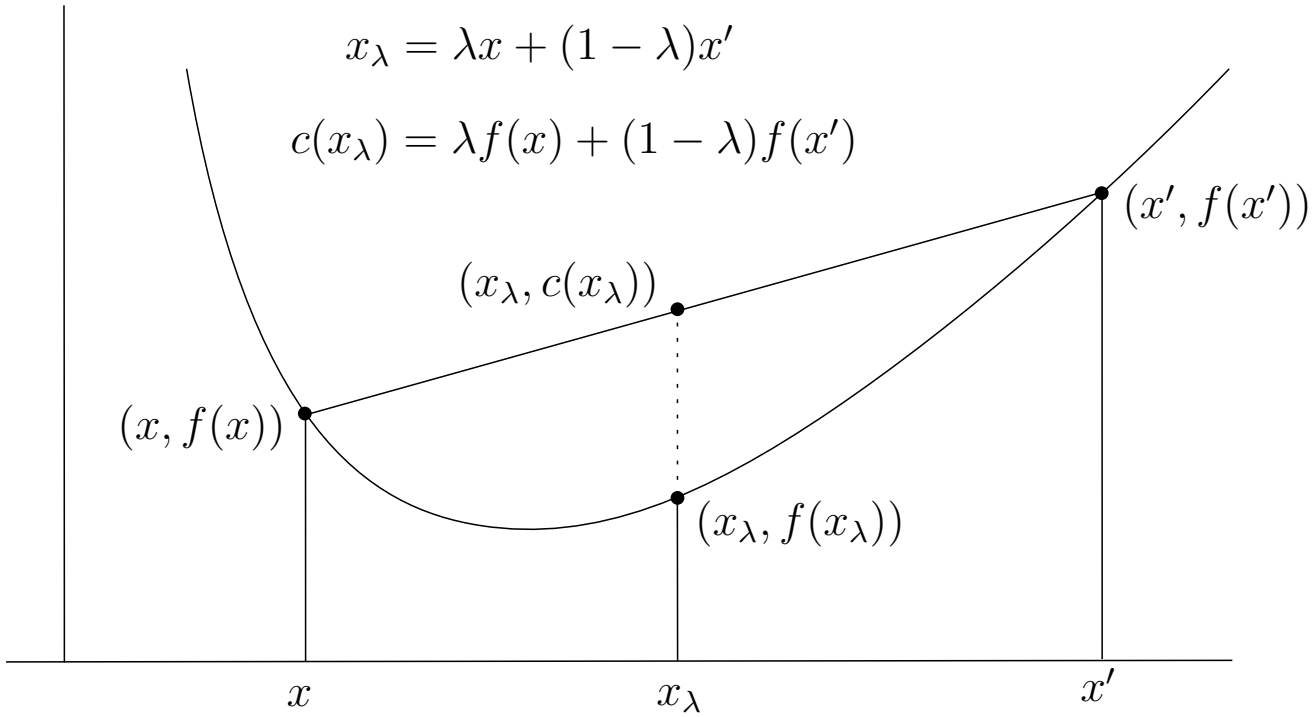


Figure 9.2: Geometric illustration of the convexity condition for a function $f : \mathcal{X} \to \mathbf{R}$, $\mathcal{X} \subseteq \mathbf{R}$ an interval: The chord segment joining $(x, f(x))$ and $(x', f(x'))$ lies above the graph of f.

If $\mathcal{X}$ and $f : \mathcal{X} \to \mathbf{R}$ are convex, and L is a line intersecting $\mathcal{X}$, then the restriction $f : \mathcal{X} \cap L \to \mathbf{R}$ is convex. Conversely, if for any $x \in X$ and $v \in \mathbf{R}^n$ the function $t \mapsto f(x + tv)$ is convex (on the intersection $\mathcal{X} \cap L_{x,v}$, $L_{x,v} = \{x + tv\}_{t \in \mathbf{R}}$), then f is convex.

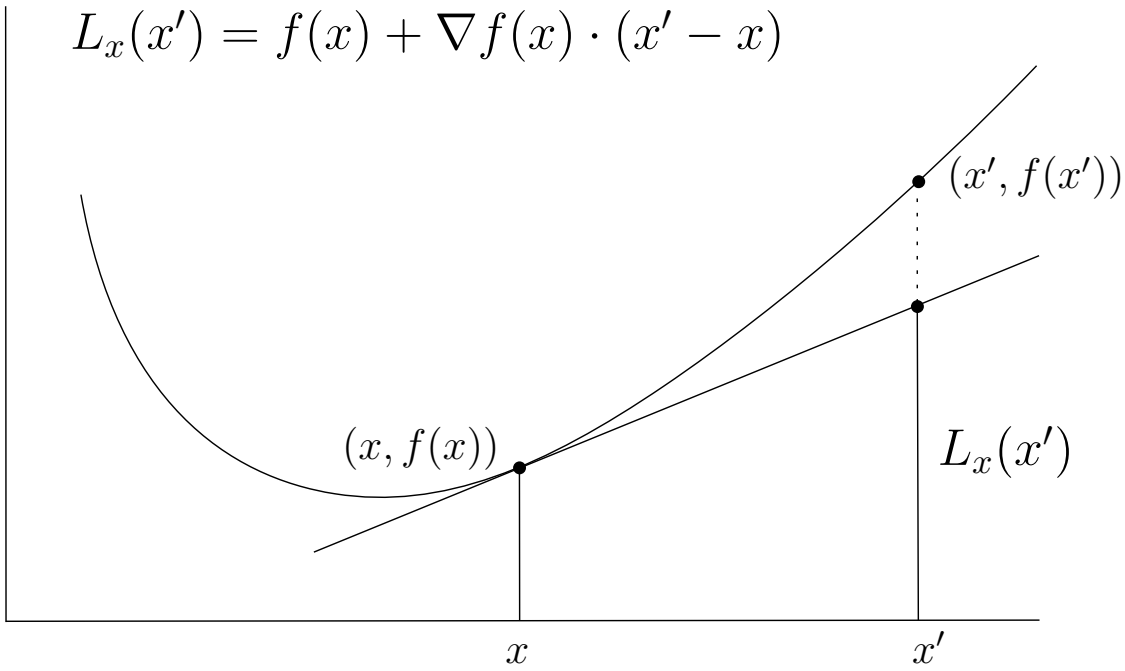


Figure 9.3: The tangent (a line in the illustration) at any point of the graph of a differentiable convex function f lies below the graph of f. This property is also sufficient to ensure that f is convex.

Theorem Assume that $\mathcal{X}$ is convex and let $f : \mathcal{X} \to \mathbf{R}$ be differentiable. Then f is convex if and only if

$$f(x') - f(x) \geqslant \nabla f(x) \cdot (x' - x). \tag{5}$$

for all $x, x' \in \mathcal{X}$ (see Fig. 9.3).

If f is twice differentiable, then f is convex if and only if

$$\nabla^2 f(x) \geqslant 0. \tag{6}$$

for all $x \in X$ (in words, the Hessian of f is semi-positive).

Proof. See [1, §§3.1.3, 3.1.4]. □

Example. If $A \in \mathbf{R}(n)$ is symmetric and semi-positive ($A \geqslant 0$), then the quadratic function $q(x) = xAx^T$ is convex.

Indeed, $\nabla q(x) = 2xA$, $\nabla^2 q(x) = 2A \geqslant 0$, and (6) proves the claim.

The criterion also yields that $q(x)$ is not convex if A is not semi-positive.

Theorem (Separating hyperplane) If $\mathcal{X} \subseteq \mathbf{R}^n$ is convex, and $0 \notin \mathcal{X}$, then there exists $u \in \mathbf{R}^n$ such that $u \cdot x > 0$ for all $x \in \mathcal{X}$.

Proof. See [1, §2.5.1], which actually establishes a more general result for two disjoint convex non-empty sets $\mathcal{X}$ and $\mathcal{X}'$, namely that there is a hyperplane H such that $\mathcal{X} \subseteq H^+$ and $\mathcal{X}' \subseteq H^-$, where H^+ and H^- are the two closed half-spaces defined by H. □

Constraint qualifications

Let $\mathcal{X} \subseteq \mathbf{R}^n$ be a convex set, $g = g_1, \cdots, g_m : \mathcal{X} \to \mathbf{R}$ convex functions, and

$$X = \{x \in \mathcal{X} \,|\, g(x) \leqslant 0\},$$

the *feasible set* of the constraints $g_i(x) \leqslant 0$.

Definition. The constraints $g(x) \leqslant 0$ satisfy the *Slater condition* (or *qualification*) if there exists $\bar{x} \in X$ such that $g(\bar{x}) < 0$ (meaning that $g_i(\bar{x}) < 0$ for all $i \in [m]$).

For other useful constraint qualifications, see Appendix E, page 61.

Theorem

Assume that the functions f and g in (1) are convex and that $\bar{x} \in X$ is a solution. If the constraints satisfy the Slater condition, then there exits $\bar{u} \in \mathbf{R}^m_+$ such that $(\bar{x}, \bar{u})$ is a saddle point of the Lagrangian. In particular, in this context we have strong duality (page 19).

Proof. See [1], §5.3.2. Since in this text the proof is developed only under special assumptions, we have included a proof in Appendix E, page 65. □

Quadratic example. Let $A \in \mathbf{R}(n)$ be symmetric and positive, and $b \in \mathbf{R}^n$. Then $f(x) = xAx^T + 2xb^T$ is (strictly) convex. Therefore $p^* = \min f(x) \text{ ⅄ } xx^T \leqslant 1$ is a convex problem.

The value of $\min_x f(x)$ is achieved when $\nabla f(x) = 0$, which yields $x = -bA^{-1}$. So we will assume that $xx^T = bA^{-2}b^T \leqslant 1$ holds, and hence $p^* = -bA^{-1}b^T$.

We are going to show directly that this problem satisfies strong duality.

The Lagrangian of the problem is

$$L(x,u) = f(x) + u(xx^T - 1) = x(A + uI_n)x^T + 2xb^T - u.$$

Since $L(x,u)$ is (strictly) convex for each u, $\min_x L(x,u)$ is achieved when

$$\nabla_x L(x,u) = 2x(A + uI_n) + 2b = 0, \text{ or } x = -b(A + uI_n)^{-1}.$$

On substituting this value in $L(x,u)$, we get the following expression for the dual function $f^*(u)$:

$$f^*(u) = -b(A + uI_n)^{-1}b^T - u.$$

Then we have $p^* = -bA^{-1}b^T = f^*(0) \leqslant d^*$, which together with weak duality establishes that $p^* = m^*$.

Remark. Strong duality in this example is valid even when A is not semi-positive (and hence $f(x)$ is not convex), but the argument is more delicate, already for the case in which A is semi-positive but not positive, mainly because the pseudo-inverses $A^\dagger$ and $(A + uI_n)^\dagger$ have to replace A^{-1} and $(A + uI_n)^{-1}$ in the arguments. See, for example, [1, p. 229].

Quadratic minimization with linear constraints

Consider the problem $\min xx^T \text{ 丄 } xA = b$, where $x \in \mathbf{R}^n$, $A \in \mathbf{R}^n_k$, $b \in \mathbf{R}^k$, and assume $\text{rank}(A) = k$, so that the k linear constraints are feasible.

The solucion p^* of this problem is the square of the distance of the origin to the linear affine set $\{x \mid xA = b\}$, which by analytic geometry turns out to be $bA^T Ab^T$.

Now we can reproduce this result by solving the optimization problem (and its dual). The Lagrangian is

$$L(x, v) = xx^T + v \cdot (xA - b) = xx^T + v(A^T x^T - b^T).$$

Since $\nabla_x L(x, v) = 2x + vA^T$, the dual Lagrangian is obtained by replacing $x = -\frac{1}{2} vA^T$ in $L(x, v)$, which yields

$$f^*(v) = -\tfrac{1}{4} vA^T Av^T - vb^T.$$

The maximum d^* of this quadratic equation in v is attained for $v = -2b(A^TA)^{-1}$, and this yields

$$\begin{aligned} d^* &= -b(A^TA)^{-1}(A^TA)(A^TA)^{-1}b^T + 2b(A^TA)^{-1}b^T \\ &= b(A^TA)^{-1}b^T = p^*. \end{aligned}$$

Theorem (Karush-Kuhn-Tucker, KKT)

Assume that f and the g_i in (1) are convex and differentiable, and that the constraints satisfy the Slater condition. Then $\bar{x} \in X$ is a solution of (1) if and only if there exists $\bar{u} \in \mathbf{R}^m_+$ such that

(a) $\nabla_x L(\bar{x}, \bar{u}) = \nabla_x f(\bar{x}) + \bar{u} \cdot \nabla_x g(\bar{x}) = 0$

(b) $\nabla_u L(\bar{x}, \bar{u}) = g(\bar{x}) \leqslant 0$

(c) $\bar{u} \cdot g(\bar{x}) = 0 \Leftrightarrow \forall i, \bar{u}_i g_i(\bar{x}) = 0$ *(slack equation/conditions)*

Proof. Let $\bar{x}$ be a solution of (1). By the theorem on page 30, there exists $\bar{u} \geqslant 0$ such that $(\bar{x}, \bar{u})$ is a saddle point of the Lagrangian and this implies the three conditions: (a) follows from the definition of a saddle point, as $L(\bar{x}, \bar{u})$ is minimum if we only move $x \in \mathcal{X}$; (b) is obvious; and (c) follows from the proof of strong duality, page 19.

Conversely:

$$
\begin{aligned}
f(x) - f(\bar{x}) &\geqslant \nabla_x f(\bar{x}) \cdot (x - \bar{x}) \quad (f \text{ convex}) \\
&= -\sum_{i=1}^{m} \bar{u}_i \nabla_x g_i(\bar{x}) \cdot (x - \bar{x}) \quad (\text{by } (a)) \\
&\geqslant -\sum_{i=1}^{m} \bar{u}_i (g_i(x) - g_i(\bar{x})) \quad (g_i \text{ convex}) \\
&= -\sum_{i=1}^{m} \bar{u}_i g_i(x) \quad (\text{by } (c)) \\
&\geqslant 0 \quad (\text{by } (b)),
\end{aligned}
$$

and this ends the proof. □

Corollary. For an unconstrained problem (with f differentiable and convex), $\bar{x} \in X$ is a solution of (1) if and only if $\nabla_x f(\bar{x}) = 0$. □

Support Vector Machines

We will look more closely at the problem of linear classification, defined earlier, by means of *support vector techniques*.

These techniques, usually known as *support vector machines* (SVM), were introduced by V. Vapnik and his school in 1992 (see [6], no doubt a classic of algorithmic learning, where they were actually called Support-Vector Networks).

An early application was the recognition of hand-written digits with an accuracy not less than the best systems available at the time. Excellent treatments of this topic can be found in [3, Ch. 5] and [7, Ch. 14].

Assume that the input space $\mathcal{X}$ is a subset of $\mathbf{R}^n$ ($n \geqslant 1$) and that the output space is $\mathcal{Y} = \{-1, 1\}$.

A dataset of the form $\mathcal{D} = \{(x^1, y^1), \ldots, (x^m, y^m)\}$, with $x^j \in \mathcal{X}$ and $y^j \in \mathcal{Y}$, is *linearly separable* if there is a hyperplane $h(x) = w \cdot x + b$ ($w \in \mathbf{R}^n$, $b \in \mathbb{R}$)) such that $h(x^j) > 0$ or $h(x^j) < 0$ according to whether $y^j = +1$ or $y^j = -1$.

These conditions are equivalent to say that $y^j h(x^j) > 0$ for $j \in [m]$.

Geometrically, the points x^j with $y^j = 1$ lie on the positive half-space defined by h and those with $y^j = -1$ lie on the negative half-space (see Fig. 13.1).

In general, there are (if any) infinitely many separating hyperplanes, so that we can envisage to impose additional constraints, like the condition that the points on either side are as far as possible from the hyperplane.

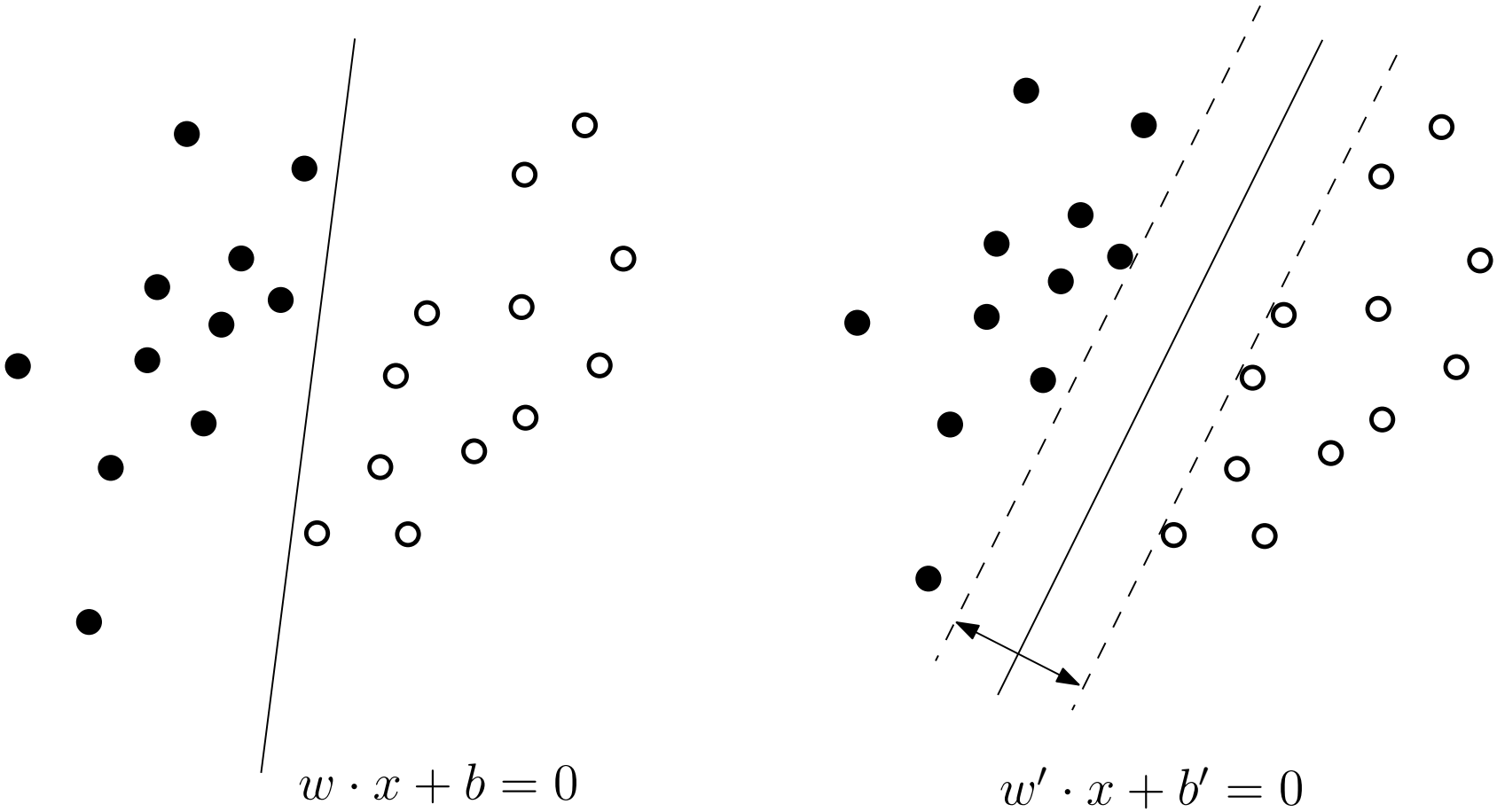


Figure 13.1: Separation by hyperplanes and support vectors. On the left, the white points ($+1$) and the black points (-1) are linearly separable. On the right we see the same set of points and the greatest margin separator, which is computed by the SVM.

Given a separating hyperplane $h(x) = w \cdot x + b$, let

$$s = \min_{j\in[m]} |w \cdot x^j + b| > 0.$$

On dividing h by s, an artifice that does not change the hyperplane, we may assume the *normalization condition*

$$\min_{j\in[m]} |w \cdot x^j + b| = 1.$$

In this case, $1/\|w\|$ is the distance to the hyperplane of any x^j at minimum distance from it, for this distance is $|w \cdot x^j + b|/\|w\|$.

The quantity $1/\|w\|$ is the *margin* of the hyperplane and the x^j at a margin distance are its *support vectors*.

Thus it is clear that to maximize the margin of $h \Leftrightarrow$ to minimize $\|w\|$, or, more conveniently, $\frac{1}{2}\|w\|^2$, which is *strictly convex*, as its gradient and hessian are w and Id (the identity), respectively. This shows that *the problem is equivalent to minimizing* $\|w\|$ *under the constraints*

$$y^j(w \cdot x^j + b) \geqslant 1.$$

So the optimization problem to solve is

$$\min_{w,b} \tfrac{1}{2}\|w\|^2 \text{ s.t. } y^j(w \cdot x^j + b) \geqslant 1,\ j \in [m].$$

The constraints $g_j(w, b) = 1 - y^j(w \cdot x^j + b) \leqslant 0$ are affine in w, b, hence qualified. So the problem has a *unique solution*.

The Lagrangian of the problem is

$$L(w, b, u) = \tfrac{1}{2}\|w\|^2 - \sum_{j\in[m]} u_j \left(y^j(w \cdot x^j + b) - 1\right).$$

The KKT conditions for this case yield the following:

(a) $\nabla_w L = w - \sum_{j\in[m]} u_j y^j x^j = 0$, or $w = \sum_{j\in[m]} u_j y^j x^j$

(b) $\nabla_b L = -\sum_{j\in[m]} u_j y^j = 0$, or $\sum_{j\in[m]} u_j y^j = 0$.

(c) For all $j \in [m]$, $u_j(w \cdot x^j + b) - 1 = 0$, or $u_j = 0 \vee y^j(w \cdot x^j + b) = 1$.

The solution w is a *linear combination* of the x^j such that $u_j \neq 0$. They belong to the *marginal hyperplanes*, as $w \cdot x^j + b = y^j = \pm 1$, and are called *support vectors* (they suffice to construct the maximal-margin hyperplane).

Duality. On plucking the expression $w = \sum_{j\in[m]} u_j y^j x^j$ into the Lagrangian we find, after some algebra with the KKT conclusions, that

$$L = \sum_{j\in[m]} u_j - \tfrac{1}{2} \sum_{j,j'} u_j u_{j'} y^j y^{j'} (x^j \cdot x^{j'})$$

Thus the dual problem, which we know is equivalent to the primal problem, is:

$$d^* = \max_{u\in\mathbf{R}^m_+} \left(\sum_{j\in[m]} u_j - \tfrac{1}{2} \sum_{j,j'} u_j u_{j'} y^j y^{j'} (x^j \cdot x^{j'}) \right)$$
$$\curlywedge\ u \geqslant 0 \wedge \sum_{j\in[m]} u_j y^j = 0.$$

The classifier h returned by the SVM algorithm can be expressed in terms of the solution u to the dual problem:

$$h(x) = \operatorname{sgn}(w \cdot x + b) = \operatorname{sgn}\left(\sum_{j\in[m]} u_j y^j (x^j \cdot x) + b\right).$$

Since any support vector $x^{j'}$ satisfies $w \cdot x^{j'} + b = y^{j'}$, b can also be obtained from u:

$$b = y^{j'} - \sum_{j\in[m]} u_j y^j (x^j \cdot x^{j'}).$$

Remark. The predictor h provided by the SVM algorithm shows that it *only needs the scalar products of vectors, not on the vectors themselves*. The significance of this fact is related to the kernel methods studied later.

For the study of the non-separable case, see Appendix F, page 67.

Appendices

Principal Component Analysis
Singular Value Decomposition
Classical Lagrange multipliers
On the Lagrangian $L(x, u, v)$
Other constraint qualifications
Linear predictors, the non-separable case

Let $X \in \mathbf{R}_n^m$. We regard X as the result of observing m characteristics of n objects, so that row X^j contains the n observations $(x_1^j, \ldots, x_n^j)$ of the j-th characteristic ($j \in [m]$).

Equivalently, the column $X_k = (x_k^1, \ldots, x_k^m)^T$ ($k \in [n]$) contains the values of the m characteristics of the k-th object.

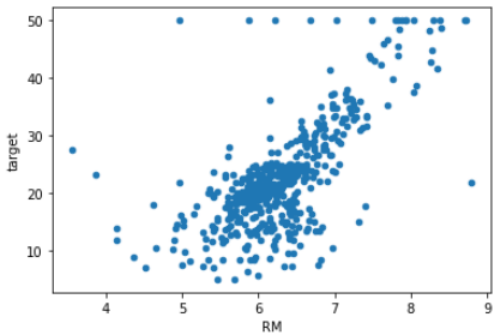
In any case, with the obvious meaning of the expressions,
$X = [X_1, \ldots, X_n] = [X^1, \ldots, X^m]$.

See Fig. 15.1 for an example os such a matrix.

```
In [3]: import pandas as pd
        from sklearn.datasets import load_boston
        boston = load_boston()
        dataset = pd.DataFrame(boston.data,columns=boston.feature_names)
        dataset['target'] = boston.target
        print(boston.feature_names)

['CRIM' 'ZN' 'INDUS' 'CHAS' 'NOX' 'RM' 'AGE' 'DIS' 'RAD' 'TAX' 'PTRATIO'
 'B' 'LSTAT']
```

```
In [2]: %matplotlib inline
        scatter = dataset.plot(kind='scatter', x='RM', y='target')
```



There are **14** attributes in each case of the dataset. They are:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

Figure 15.1: A Dataset derived from information collected by the U.S. Census Service concerning housing in the area of Boston Mass. It has 14 features or attibutes and 506 cases.

Set $\mu^j = E[X^j] = (\sum X^j)/n$, the mean of X^j, and $\sigma_{jk} = \mathrm{Cov}(X^j, X^k) = E[X^j X^k] - \mu^j \mu^k$, the covariance of X^j and X^k.

The symmetric matrix $\Sigma = \mathrm{Cov}(X) = (\sigma_{jk}) \in \mathbf{R}(m)$ is semi-positive and is called the *covariance matrix* of X. Let us note that $\sigma_{jj} = \mathrm{Var}(X^j)$ (*variance* of X^j), and that $\mathrm{Var}(X^j) = \sigma_j^2$, where $\sigma_j \geqslant 0$ is the standard deviation of X^j.

Given a unit vector $u \in \mathbf{R}^m$, it turns out that $\mathrm{Var}(uX) = u\Sigma u^T$, and that *this value is maximum precisely when u is an eigenvector of Σ with maximum eigenvalue*.

Under these conditions, $u_1 = u$ is said to be the *main component* of X. Note that uX is the vector formed with the projections of the columns of X on u, so that these projections pick up the maximum variability of X in one direction.

The second main component of X is the unit eigenvector u_2 corresponding to the second eigenvalue (*in non-ascending order*) of Σ. This vector maximizes $\mathrm{Var}(uX) = u\Sigma u^T$ for unit vectors u perpendicular to u_1.

Continuing this process, we obtain an orthonormal basis $u_1, \ldots, u_m$ of $\mathbf{R}^m$ such that $\mathrm{Var}(u_r X) = u_r \Sigma u_r^T$ is maximum for unit vectors perpendicular to $u_1, \ldots, u_{r-1}$.

If we put $U_r \in \mathbf{R}_m^r$ to denote the matrix formed by the vectors $u_1, \ldots, u_r$, the matrix $U_r X$ is of type $r \times n$ and incorporates the variability of X attributable to the first r eigenvalues (in non-ascending order) of Σ.

The technique of the main components is a first example of *dimensional reduction*. It can be understood as a form of unsupervised learning.

It should also be noted that $U_r X$ can be used as a form of preprocessing applicable to the data X before it is subjected to a supervised learning procedure. The value of r is chosen so that $u_{r+1}, \ldots, u_m$ play a negligible role in explaining the variability of X. See Figure 15.2 for an illustration.

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60

```
In [21]: import pandas as pd
         url = "https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data"
         # load dataset into Pandas DataFrame
         df = pd.read_csv(url, names=['sepal length','sepal width','petal length','petal width','target'])
         print(df)
```

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
..	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

[150 rows x 5 columns]

```
In [24]: from sklearn.preprocessing import StandardScaler
         features = ['sepal length', 'sepal width', 'petal length', 'petal width']
         # Separating out the features
         x = df.loc[:, features].values
         # Separating out the target
         y = df.loc[:,['target']].values
         # Standardizing the features
         x = StandardScaler().fit_transform(x)
```

```
In [6]: pca.explained_variance_ratio_
Out[6]: array([0.72770452, 0.23030523])
```

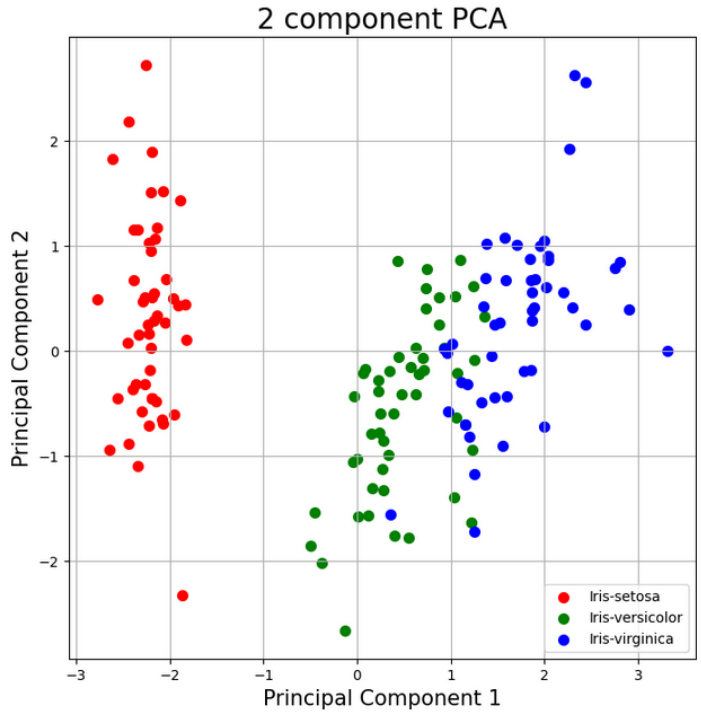


Figure 15.2: PCA of 150 cases of iris flowers with four metrical features and one label for the species. It reproduces the computations in a 1936 paper by R. A Fisher (*The use of multiple measurements in taxonomic problems*). The first principal compoent explains 72.77% of the variation in the data, while the second explains 23.03%. Note that the units on the vertical axis are longer than on the horizontal axis.

Let $X \in \mathbf{R}_n^m$, which we think of as data in the style of what we saw for PCA (page 47), and let r be its rank.

The **SVD theorem** establishes that there are real values $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > 0$ and orthonormal matrices $U \in \mathbf{R}(m)$ and $V \in \mathbf{R}(n)$ such that $A = U\Lambda V^T$, where the only non-zero vectors of Λ are $\Lambda_{ii} = \lambda_i$ for $i \in [r]$.

Note that $\Lambda\Lambda^T \in \mathbf{R}(m)$ and $\Lambda^T\Lambda \in \mathbf{R}(n)$ are diagonal and that their only non-zero elements are the same: $\lambda_1^2, \cdots, \lambda_r^2$ down the main diagonal.

The equality $XX^T = (U\Lambda V^T)(V\Lambda^T U^T) = U(\Lambda\Lambda^T)U^T$ shows that $\lambda_1^2, \cdots, \lambda_r^2$ are the only non-zero eigenvectors of XX^T and that U is the corresponding matrix of orthonormal eigenvectors.

Denote by u_j the eigenvector corresponding to λ_j^2, $j \in [r]$.

Similarly, the relation $X^T X = (V\Lambda^T U^T)(U\Lambda V^T) = V(\Lambda^T\Lambda)V^T$ yields that $\lambda_1^2, \cdots, \lambda_r^2$ are the only non-zero eigenvectors of $X^T X$ and that V is the corresponding matrix of orthonormal eigenvectors.

Denote by v_j the eigenvector corresponding to λ_j^2, $j \in [r]$.

Notice that the relation $Av_j = \lambda_j u_j$ holds.

Now $U\Lambda = [\lambda_1 u_1, \cdots, \lambda_r u_r, \overbrace{\mathbf{0}, \ldots, \mathbf{0}}^{m-r}]$, so

$$A = U\Lambda V^T = \lambda_1 u_1 v_1^T + \cdots + \lambda_r u_r v_r^T,$$

which expresses A as a linear combination of the rank-one matrices $u_j v_j^T$.

In fact, it turns out that for $k = 1, \ldots, r$, the matrix

$$M_k = \lambda_1 u_1 v_1^T + \cdots + \lambda_r u_k v_k^T$$

is the *optimal approximation of X by rank k matrices* (Eckart-Young theorem).

This result is then also a form of *dimensional reduction*. The value of k is chosed so that $X - M_k$ is negligible in the context where it is applied.

An important application of the singular decomposition is that the solution a of $Xa = b$ by *least-squares* is $a = X^\dagger b$, where $X^\dagger$ is the *Moore-Penrose pseudoinverse* of X, namely $X^\dagger = V \Lambda^\dagger U^T$, with $\Lambda^\dagger_{jj} = \lambda_j^{-1}$ for $j = 1, \ldots, r$ and all other entries zero.

For other dimensional reduction procedures, see [8, Ch. 6]: (Fischer's) Linear Discriminant Analysis (LDA), PCA, kernel PCA, Independent Components Analysis (ICA), Multidimensional Scaling (MDS), ...

Let $\mathcal{X}$ be an open set in $\mathbf{R}^n$ and $f, g_1, \cdots, g_r : \mathcal{X} \to \mathbf{R}$ differentiable functions. Let $X = \{x \in \mathcal{X} \,|\, g_1(x) = \cdots = g_r(x) = 0\}$.

The question is to find *necessary* conditions for $f|_X$ (the restriction of f to X) to have an *extremum* at a point $x_0 \in X$.

For this, assume that $d_{x_0}g_1, \cdots, d_{x_0}g_r$ are linearly independent. Then X is a manifold of dimension $n - r$ around x_0 and the usual necessary condition is that $d_{x_0}(f|_X) = 0$. But since $d_{x_0}(f|_X) = (d_{x_0}f)|_X$, the condition becomes $(d_{x_0}f)|_X = 0$. This means that $(d_{x_0}f)$ has to be in the kernel K_{x_0} of the restriction map $D_{\mathcal{X},x_0} \to D_{X,x_0}$, where $D_{\mathcal{X},x_0}$ is the vector space of differentials of $\mathcal{X}$ at x_0, with a similar meaning for D_{X,x_0}.

Now $d_{x_0} g_j \in K_{x_0}$ $(j = 1, \cdots, r)$, because $(d_{x_0} g_j)|_X = d_{x_0}(g_j|_X) = 0$.

Under our assumptions, it turns out that actually

$$K_{x_0} = \langle d_{x_0} g_1, \cdots, d_{x_0} g_r \rangle.$$

Consequently, the necessary condition we are looking for is that *there exist real numbers* $\lambda_1, \cdots, \lambda_r$ *such that*

$$d_{x_0} f = \lambda_1 d_{x_0} g_1 + \cdots + \lambda_r d_{x_0} g_r.$$

In conclusion, the search for extrema points of $f|_X$ leads to the *Lagrange conditions*:

$$d_x f = \lambda_1 d_x g_1 + \cdots + \lambda_r d_x g_r, \;\; g_1(x) = \cdots = g_r(x) = 0,$$

where the unknowns are $x \in \mathcal{X}$ and $\lambda_1, \cdots, \lambda_r \in \mathbf{R}$.

Primal problem:

$p^* = \min_{x \in \mathcal{X}} f(x) \curlywedge g_i(x) \leqslant 0, h_j(x) = 0 \ (i \in [m], j \in [p])$.

Feasible set: $X = \{x \in \mathcal{X} \mid g(x) \leqslant 0, h(x) = 0\}$, so that $p^* = \min_{x \in X} f(x)$.

Lagrangian: $L(x, u, v) = f(x) + u \cdot g(x) + v \cdot h(x)$,
$u = u_1, \cdots, u_m$, $v = v_1, \cdots, v_p$.

Dual Lagrangian: $f^*(u, v) = \min_{x \in \mathcal{X}} L(x, u, v)$.

Dual problem: $d^* = \max f^*(u, v) \curlywedge u \geqslant 0$.

Weak duality: $d^* \leqslant p^*$ (in general they are not equal).

Dual gap: The difference $p^* - d^*$.

Strong duality: When $d^* = p^*$.

Theorem (Saddle point implies strong duality) Assume f, g, h smooth. Suppose there exist $\bar{x} \in X$, $\bar{u} \in \mathbf{R}^m_+$ and $\bar{v} \in \mathbf{R}^p$ such that for all $x \in \mathcal{X}$, $u \in \mathbf{R}^m_+$, $v \in \mathbf{R}^p$ we have (*saddle condition*)

$$L(\bar{x}, u, v) \leqslant L(\bar{x}, \bar{u}, \bar{v}) \leqslant L(x, \bar{u}, \bar{v}). \tag{7}$$

Then $\bar{x}$ and $(\bar{u}, \bar{v})$ solve the primal and dual problems, respectively, and $f(\bar{x}) = f^*(\bar{u}, \bar{v})$.

Proof. From the first inequality in (7), namely

$$f(\bar{x}) + u \cdot g(\bar{x}) + v \cdot h(\bar{x}) \leqslant L(\bar{x}, \bar{u}, \bar{v}),$$

which holds for all $u \geqslant 0$ and all v, we infer immediately that $g(\bar{x}) \leqslant 0$ and $h(\bar{x}) = 0$, which show that $\bar{x}$ is feasible.

Still from the first inequality we get, setting $u = 0$ and using that $v \cdot h(\bar{x}) = 0$ for any v, that $f(\bar{x}) \leqslant f(\bar{x}) + \bar{u} \cdot g(\bar{x})$, which yields $\bar{u} \cdot g(\bar{x}) \geqslant 0$. Since $\bar{u} \geqslant 0$ and $g(\bar{x}) \leqslant 0$, we must have $\bar{u} \cdot g(x) = 0$, which actually is equivalent to $\bar{u}_i g_i(\bar{x}) = 0$ for all $i \in [m]$ (these relations are called *complementary slack conditions*).

Now look at the second inequality in (7). Using that $f(\bar{x}) = L(\bar{x}, \bar{u}, \bar{v})$, we get $f(\bar{x}) \leqslant L(x, \bar{u}, \bar{v}) = f(x) + \bar{u} \cdot g(x) \leqslant f(x)$ for all $x \in X$, which shows that $\bar{x}$ is optimal and $f(\bar{x}) \leqslant f^*(\bar{u}, \bar{v})$. Thus $p^* = f(\bar{x}) \leqslant f^*(\bar{u}, \bar{v}) \leqslant d^*$. By weak duality, the inequalities must be equalities and this completes the proof.

Karlin condition. For all $u \in \mathbf{R}^m_+$, $u \neq 0$, there exists $x \in \mathcal{X}$ such that $u \cdot g(x) < 0$.

The negation of this condition is that there exists $u \in \mathbf{R}^m_+$, $u \neq 0$, such that for all $x \in \mathcal{X}$ we have $u \cdot g(x) \geqslant 0$.

Theorem *The Slater and Karlin conditions are equivalent*.

Proof. Assume that $x \in \mathcal{X}$ satisfies $g_i(x) < 0$ for all $i \in [m]$. Then for all $u \in \mathbf{R}^m_+$ we have $u \cdot g(x) < 0$ if $u \neq 0$.

This shows that Slater $\Rightarrow$ Karlin.

For the converse, namely that Karlin $\Rightarrow$ Slater, we will show that the Karlin condition cannot be satisfied if the Slater condition is not satisfied.

We will proceed in several steps.

Suppose then that the Slater condition is not satisfied. This means that for all $x \in \mathcal{X}$ there is an index i such that $g_i(x) \geqslant 0$.

Let $Z = \{z \in \mathbf{R}^m \,|\, \exists x \in \mathcal{X} \text{ with } z > g(x)\}$.

It is clear that $Z \neq \emptyset$ and $0 \notin Z$.

Moreover, it is easy to see, using that the g_i are convex, that Z *is convex*.

So 0 and Z can be separated by a hyperplane (page 27): there is $u \in \mathbf{R}^m$, $u \neq 0$, such that $u \cdot z \geqslant 0$ for all $z \in Z$. This implies that $u \geqslant 0$ (if we had $u_i < 0$, we would find a $z' \in Z$ with $u \cdot z' < 0$ by letting $z_i \uparrow \infty$ starting with any $z \in Z$).

Now let $\delta = \inf_{x \in \mathcal{X}} u \cdot g(x)$ and $\delta' = \inf_{z \in Z} u \cdot z \geqslant 0$. We will see that $\delta = \delta'$, and so $\delta \geqslant 0$, which means that the Karlin condition is not satisfied.

To finish, we will show first that $\delta' \geqslant \delta$ and then that $\delta \geqslant \delta'$.

Indeed, for any $z \in Z$, there is $x \in \mathcal{X}$ such that $z > g(x)$, so $u \cdot z \geqslant u \cdot g(x) \geqslant \delta$, hence $\delta' \geqslant \delta$.

On the other hand, for any $x \in \mathcal{X}$ and any arbitrarily small positive vector $\epsilon \in \mathbf{R}^m$, we have that $z = g(x) + \epsilon \in Z$, and for this z

$$\delta' \leqslant u \cdot z = u \cdot g(x) + u \cdot \epsilon,$$

which implies that $\delta' \leqslant \delta + u \cdot \epsilon$. Since $u \cdot \epsilon \geqslant 0$ is arbitrarily small, we conclude that $\delta' \leqslant \delta$. □

Strict constraint qualification. It is satisfied if $|X| > 1$ and there is $x \in X$ such that the g_i are strictly convex at x within X (this means that for any $x' \in X$, $x' \neq x$, and any $\lambda \in (0,1)$, we have $\lambda g_i(x) + (1-\lambda) g_i(x') > g_i(\lambda x + (1-\lambda)x')$).

Theorem

The strict constraint qualification implies the Slater (and hence also the Karlin) condition.

Proof. With the notations on page 64, if $\bar{x} = \lambda x + (1-\lambda)x'$ then

$$g_i(\bar{x}) < \lambda g_i(x) + (1-\lambda) g_i(x') \leqslant 0.$$

□

Theorem

Let $\mathcal{X} \subset \mathbf{R}^n$ be convex. Let $f, g_i : \mathcal{X} \to \mathbf{R}$ $(i \in [m])$ be convex functions and set $X = \{x \in \mathcal{X} \mid g_i(x) \leqslant 0\}$ (*feasible set*). Assume that the g_i satisfy the Slater condition ($\Leftrightarrow \exists\, x \in X$ such that $g_i(x) < 0$, $i \in [m]$) and that $\bar{x}$ is an optimal solution of the problem $\min_{x \in \mathcal{X}} f(x) \curlywedge g(x) \leqslant 0$. Then there exists $\bar{u} \in \mathbf{R}^m_+$ such that $(\bar{x}, \bar{u})$ is a saddle point of the Lagrangian.

Proof. For $x \in \mathcal{X}$, consider the conditions

$$f(x) - f(\bar{x}) \leqslant 0, g_1(x) \leqslant 0, \cdots, g_m(x) \leqslant 0.$$

These conditions do not satisfy Slater's condition, as $f(x) - f(\bar{x}) \geqslant 0$ (by the optimality of $\bar{x}$). Therefore they do not satisfy the Karlin condition. This means that we can find $(\bar{u}_0, \bar{u}) \in \mathbf{R}^{m+1}_+$ such that

$$\bar{u}_0(f(x) - f(\bar{x})) + \sum_{i=1}^{m} \bar{u}_i g_i(x) \geqslant 0. \tag{8}$$

In particular, for $x = \bar{x}$ we get $\sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}) \geqslant 0$. Since $\bar{u}_i g_i(\bar{x}) \leqslant 0$ for $i \in [m]$, we actually have $\sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}) = 0$. Together with (8), we can write

$$\bar{u}_0 f(x) + \sum_{i=1}^{m} \bar{u}_i g_i(x) \geqslant \bar{u}_0 f(\bar{x}) = \bar{u}_0 f(\bar{x}) + \sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}).$$

If we had $\bar{u}_0 = 0$, then we would also have $\sum_{i=1}^{m} \bar{u}_i g_i(x) \geqslant 0$, which contradicts the Karlin qualification, hence also the Slater qualification of the $g_i(x)$. On dividing by $\bar{u}_0$, and redefining $\bar{u}_i$ as $\bar{u}_i / \bar{u}_0$, we obtain that $L(\bar{x}, \bar{u}) \leqslant L(x, \bar{u})$, which is the second inequality in the definition of a saddle point.

On the other hand, for all $u \in \mathbf{R}^m_+$, we have $\sum_{i=1}^{m} u_i g_i(\bar{x}) \leqslant 0$, for $g_i(\bar{x} \leqslant 0$ for all $i \in [m]$. Therefore

$$L(\bar{x}, u) = f(\bar{x}) + u \cdot g(\bar{x}) \leqslant f(\bar{x}) = f(\bar{x}) + \sum_{i=1}^{m} \bar{u}_i g_i(\bar{x}) = L(\bar{x}, \bar{u}),$$

and this ends the proof. □

Deviation variables. If the data are not linearly separable, we cannot assume that $y^j(w \cdot x^j + b) \geqslant 1$ for all $j \in [m]$.

In view of applying optimization tools, introduce non-negative *deviation* or *slack variables* $t_1, \ldots, t_n$ and consider the relaxed constraints $y^j(w \cdot x^j + b) \geqslant 1 - t_j$.

In this situation, a convenient modification of the function to be minimized is

$$\tfrac{1}{2}\|w\|^2 + \lambda \sum t_j,$$

where $\lambda \in \mathbf{R}_+$ is a constant.

The hyperplane produced with this minimization separates correctly the x^j with margin $1/\|w\|$ except the *outliers*, points that fall either on the incorrect half-space or within the ribbon $-1 < w \cdot x^j + b < +1$ (see, for example, [9, Ch. 9]).

Another extension is to multi-class classifications (*ibidem*).

Feature mappings. Finally, let us mention the very useful device consisting of applying linear separation after mapping the input space to a higher dimension by means of a *non-linear* map. Roughly, it works as follows.

A *feature map* of the space $\mathcal{X}$ is a map $\phi : \mathcal{X} \to \mathbf{R}^{n'}$, where n' can be arbitrary. Usually n' and ϕ are chosen to facilitate that the data $\psi(x)$ appear to be linearly separable in $\mathbf{R}^{n'}$ when this condition is not satisfied in $\mathcal{X}$.

If we manage to obtain a linear separator h' of the $\phi(x^i) \in \mathbf{R}^{n'}$, then $h(x) = h'(\phi(x))$ is a non-linear separator of the x^i in $\mathbf{R}^n$ and the hypersurface $\{h(x) = 0\}$ is the *decision boundary*.

A relevant point is that these techniques lead naturally to the notion *kernels* (we will study them in the session 10-13 on RKHS), which rely on the pairing $\kappa(x, x') = \phi(x) \cdot \phi(x')$ or, more specifically, on the *kernel matrix* $\kappa(x^i, x^j)$, which is sufficient, as remarked before, to run the support vector algorithms (in $\mathbf{R}^{n'}$), and the *kernel trick* amounts to the realization that often the values $\kappa(x^i, x^j)$ can be judiciously specified with no reference to ϕ (more in session 10-13).

One more point is that there are also interesting cases in which $n' < n$, and then we speak of *dimension reduction*. As noted, the PCA and SVD techniques mentioned earlier fall under this notion. A quite interesting achievement is the *t*-SNE separation algorithm developed in [10] and [11] mapping images of hand-written digits (dimension $n = 28^2$) to $\mathbf{R}^2$.

References I

[1] S. Boyd and L. Vandenberghe, *Convex optimization*.

Cambridge University Press, 2009.

Seventh printing with corrections. xiv +716 p.
`https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf`.
First published 2004.

[2] B. Schölkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*.

MIT Press, 2002.

xviii + 626 pp.

References II

[3] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*.

MIT press, 2018.

xvi + 486 pp.

[4] G. Giorgi and T. H. Kjeldsen (editors), *Traces and emergence of nonlinear programming*.

Birkhäuser, 2014.

A collection of selected classical papers on nonlinear optimization, with the introductory piece "A historical view of nonlinear programming: traces and emergence" by the editors G. Giorgi and H. Kjeldsen.

References III

[5] L. N. Trefethen and D. Bau III, *Numerical linear algebra*.

Siam, 1997.

[6] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[7] E. Alpaydin, *Introduction to machine learning (fourth edition)*.

Adaptive computation and machine learning, MIT press, 2020.

xxiv + 682 pp. 1st edition: 2004; 2nd, 2010; 3rd, 2014.

[8] S. Marsland, *Machine learning: an algorithmic perspective (second edition)*.

CRC press, 2015.

References IV

[9] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning (corrected at 4th printing)*, vol. 112 of *Springer Texts in Statistics*.

Springer, 2014.

[10] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[11] L. Van Der Maaten, "Accelerating t-SNE using tree-based algorithms," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014.