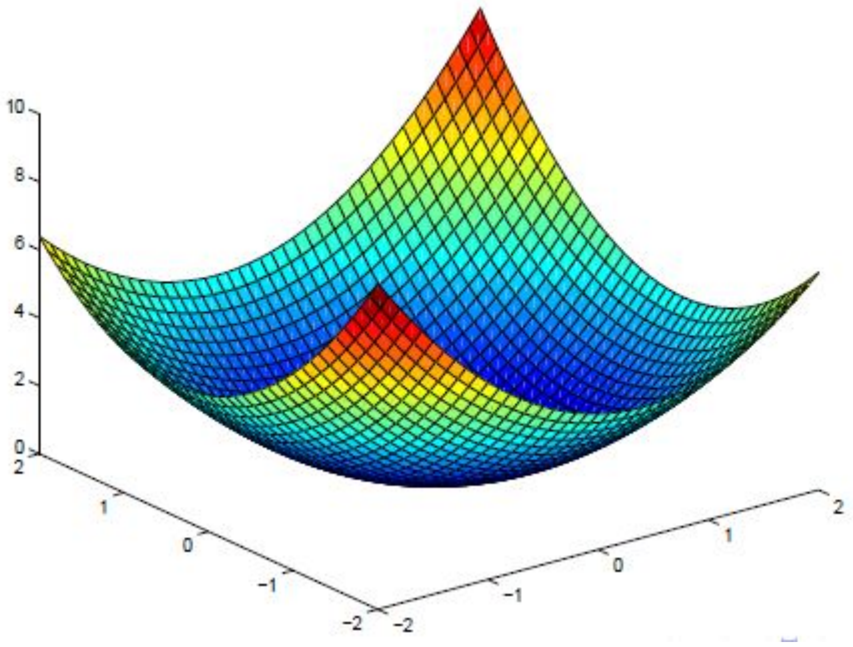
Introduction to Convex Optimization



Mrinmay Maharaj

Office: MB 113

mrinmay.mj@rkmvu.ac.in

ML & Optimization

- Machine learning and Artificial Intelligent systems (such as search engines, recommendation platforms, and speech and image recognition software)have become an indispensable part of life

- These are rooted in statistics, rely on efficiency of numerical algorithms to solve data drive problem

- One of the pillars of machine learning is mathematical optimization

ML Setup

- **Feature vector**: a_j, $j=1,2,\ldots,m$
- Outcome/observation: y_j for each a_j.

- The outcomes could be
  - ✓ y_j real : **regression**
  - ✓ y_j is a label indicating a_j lies in one of N (N>=2) classes: **classification**
  - ✓ Multiple labels: classify according to multiple criteria
  - ✓ No labels (y_j is null) : Partition a_j into few clusters: **clustering**

ML Setup

- Find a function $\Phi(a_j)$ that approximately maps a_j to y_j for each j : $\Phi(a_j) \approx y_j$ for *j* = *1; 2; : : : ;m*

- We define $\Phi(.)$ in terms of some parameter vector **x**

- Identification of $\Phi(.)$ becomes a data-fitting problem: Find the best **x**.

- Objective function in this problem is built up of *m* terms that capture mismatch between predictions and observations for each (a_j; y_j).

- The process of finding $\Phi(.)$ is called learning or training.

- Prediction: Given new data vectors a_k predict output $y_k \rightarrow \Phi(a_k)$

Text Classification via Convex Optimization

Suppose we want to find if an article discuss food

Naive approach:

by observing that the document contains the names of food items

Statistical machine learning approach

Begins with the collection of a sizable set of examples

$$\{(x_1, y_1), \ldots, (x_n, y_n)\},$$

where for each $i \in \{1, \ldots, n\}$ the vector x_i represents the features of a text document (e.g., the words it includes)

The scalar y_i is a label indicating whether the document belongs ($y_i = 1$) or not ($y_i = -1$) to a particular class (i.e., topic of interest).

Text Classification via Convex Optimization

- We can construct a classification program, defined by a prediction function ***h***, and measure its performance by counting how often the program prediction ***h***(x_i) differs from the correct prediction y_i .

- The prediction function ***h*** can be taken as the function that minimizes the frequency of observed misclassifications (the empirical risk of misclassification):

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left[h(x_i) \neq y_i\right],$$

where 1[.] is an indicator function defined as

$$\mathbb{1}[A] = \begin{cases} 1 & \text{if } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

Convex Optimization Problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in C \end{array}$$

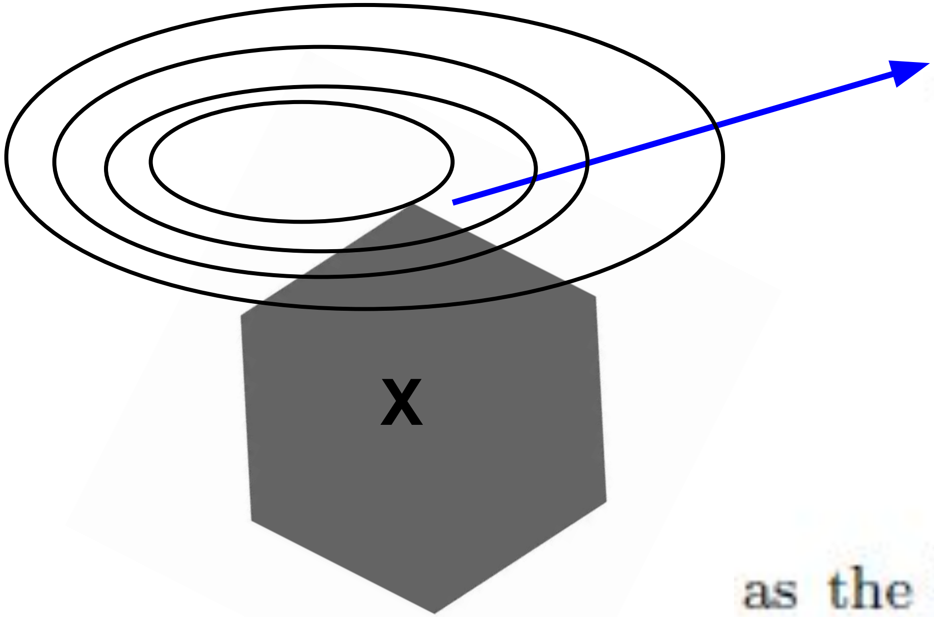
f is a convex function,
C is a convex set,

$$\begin{array}{lll} \text{minimize} & f(x) & \\ \text{subject to} & g_i(x) \leq 0, & i = 1, \ldots, m \\ & h_i(x) = 0, & i = 1, \ldots, p \end{array}$$

f is convex function, g_i are convex functions, and h_i are affine functions, and x is the optimization variable.

- If f and g_i are convex function —- convex problem,
- If all g_i are differentiable –- smooth problem
- if any g_i not differentiable — non smooth problem
- If m =0, p=0, — unconstrained problem

Convex Optimization: Optimality



Set of optimal solutions

$$X^* = \cap_{k=0}^{\infty}\{x \in X \mid f(x) \leq \gamma_k\},$$

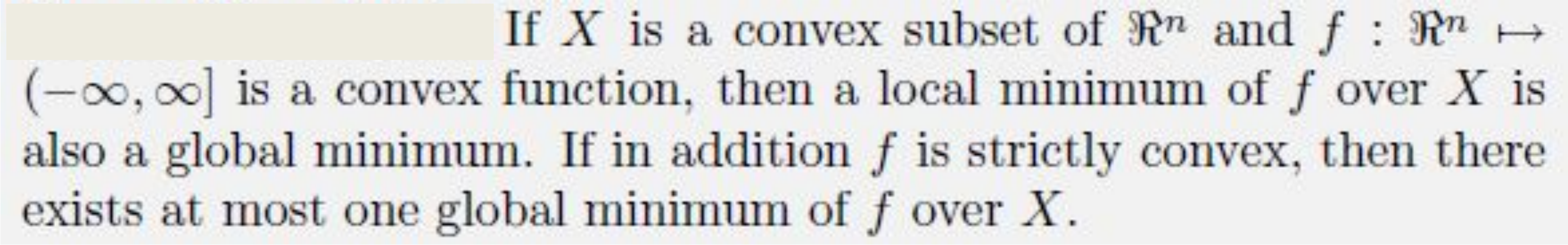
$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}$$

as the intersection of all the nonempty level sets of the form

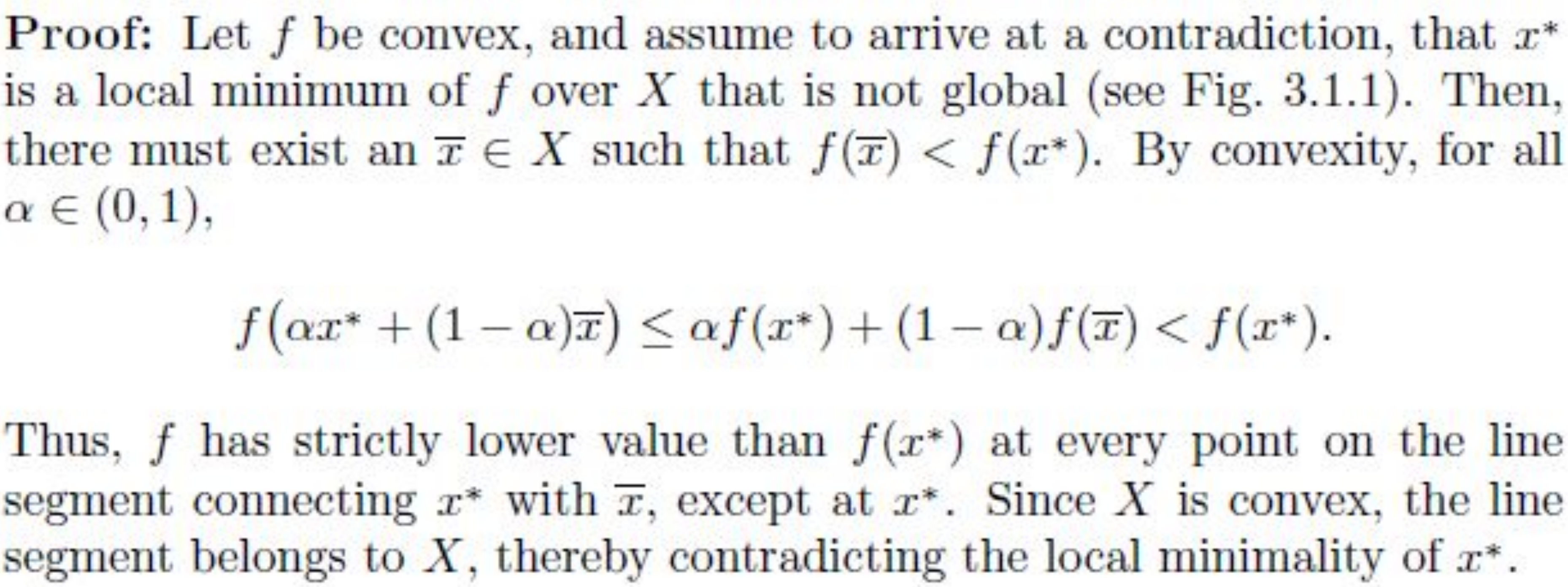
$$\{x \in X \mid f(x) \leq \gamma\}, \quad \gamma \in \Re.$$

Def. A point $x^* \in \mathcal{X}$ is **locally optimal** if $f(x^*) \leq f(x)$ for all x in a **neighborhood** of x^*. **Global** if $f(x^*) \leq f(x)$ for **all** $x \in \mathcal{X}$.

Convex Optimization: Local = Global



If X is a convex subset of $\Re^n$ and $f : \Re^n \mapsto (-\infty, \infty]$ is a convex function, then a local minimum of f over X is also a global minimum. If in addition f is strictly convex, then there exists at most one global minimum of f over X.



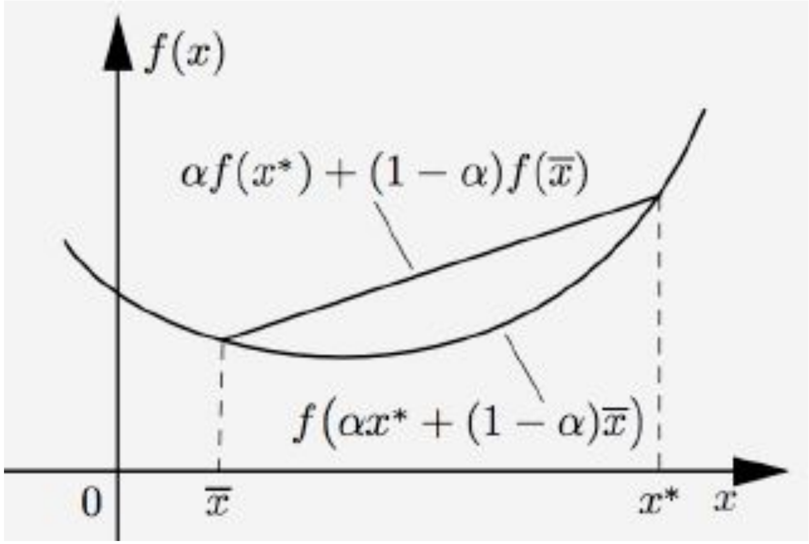
Proof: Let f be convex, and assume to arrive at a contradiction, that x^* is a local minimum of f over X that is not global (see Fig. 3.1.1). Then, there must exist an $\overline{x} \in X$ such that $f(\overline{x}) < f(x^*)$. By convexity, for all $\alpha \in (0, 1)$,

$$f\big(\alpha x^* + (1-\alpha)\overline{x}\big) \leq \alpha f(x^*) + (1-\alpha) f(\overline{x}) < f(x^*).$$

Thus, f has strictly lower value than $f(x^*)$ at every point on the line segment connecting x^* with $\overline{x}$, except at x^*. Since X is convex, the line segment belongs to X, thereby contradicting the local minimality of x^*.

Convex Optimization: Local=Global

For convex problems, local $\Longrightarrow$ global!

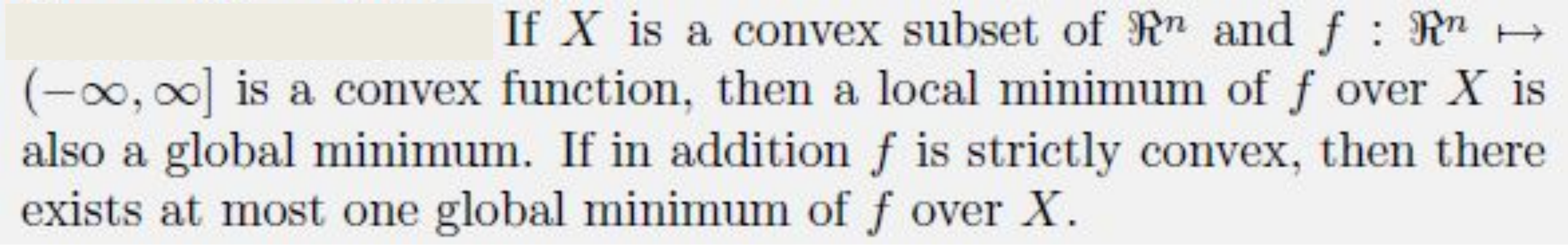

$f(x)$
$\alpha f(x^*) + (1-\alpha) f(\overline{x})$
$f\big(\alpha x^* + (1-\alpha)\overline{x}\big)$
0
$\overline{x}$
x^*
x

Given x^* and $\overline{x}$ with $f(\overline{x}) < f(x^*)$, every point of the form

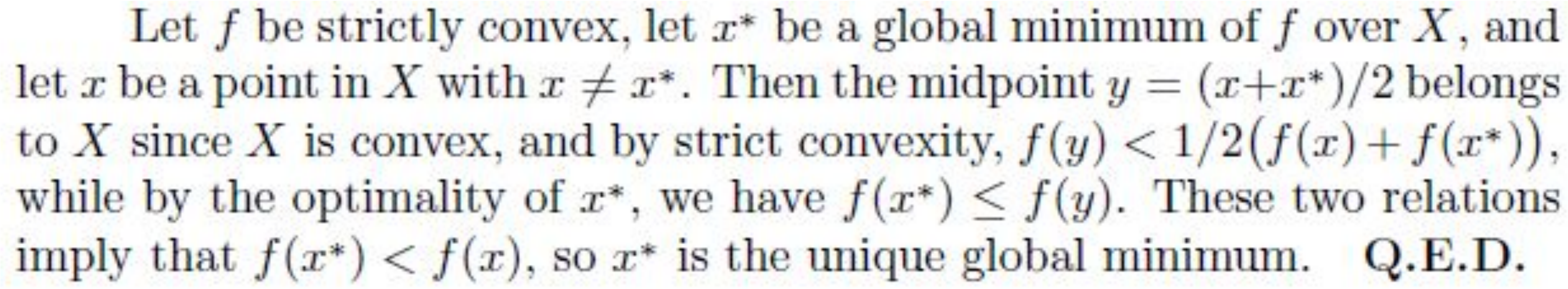
$$x_\alpha = \alpha x^* + (1-\alpha)\overline{x}, \qquad \alpha \in (0,1),$$

satisfies $f(x_\alpha) < f(x^*)$. Thus x^* cannot be a local minimum that is not global.

Convex Optimization: Global is unique

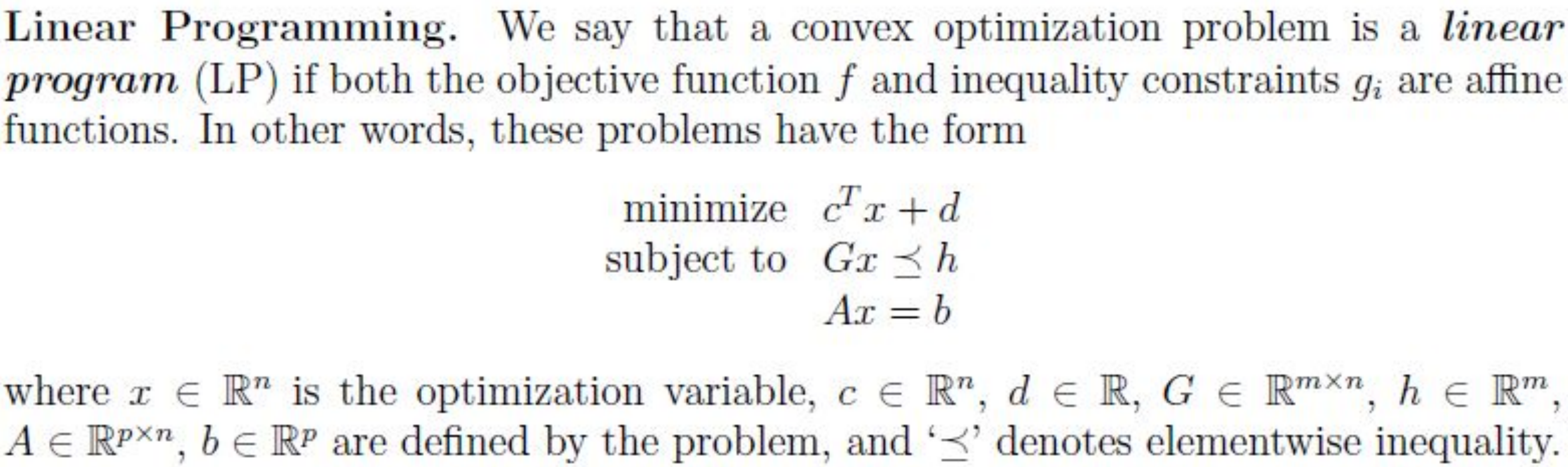


If X is a convex subset of $\Re^n$ and $f : \Re^n \mapsto (-\infty, \infty]$ is a convex function, then a local minimum of f over X is also a global minimum. If in addition f is strictly convex, then there exists at most one global minimum of f over X.



Let f be strictly convex, let x^* be a global minimum of f over X, and let x be a point in X with $x \neq x^*$. Then the midpoint $y = (x+x^*)/2$ belongs to X since X is convex, and by strict convexity, $f(y) < 1/2\big(f(x) + f(x^*)\big)$, while by the optimality of x^*, we have $f(x^*) \leq f(y)$. These two relations imply that $f(x^*) < f(x)$, so x^* is the unique global minimum. **Q.E.D.**

Special Convex Optimization Problems: LP



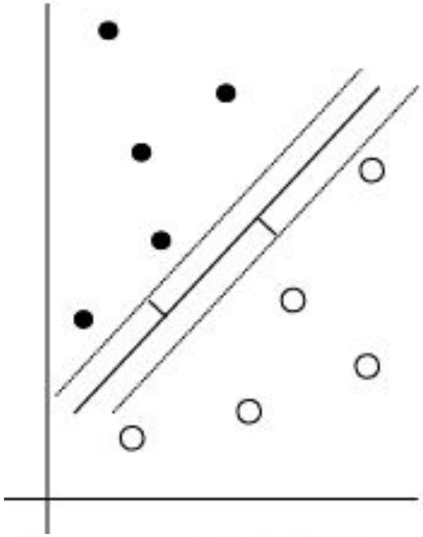
Linear Programming. We say that a convex optimization problem is a ***linear program*** (LP) if both the objective function f and inequality constraints g_i are affine functions. In other words, these problems have the form

$$\begin{aligned} \text{minimize} \quad & c^T x + d \\ \text{subject to} \quad & Gx \preceq h \\ & Ax = b \end{aligned}$$

where $x \in \mathbb{R}^n$ is the optimization variable, $c \in \mathbb{R}^n$, $d \in \mathbb{R}$, $G \in \mathbb{R}^{m \times n}$, $h \in \mathbb{R}^m$, $A \in \mathbb{R}^{p \times n}$, $b \in \mathbb{R}^p$ are defined by the problem, and '$\preceq$' denotes elementwise inequality.

Example of LP formulation in ML

Linear classifier: We are given data points and labels on the data-points {0;1}. The task is to find a hyperplane which separates the two kind of labels. Such a separating hyperplane is called a linear classier. The input is a function $\mathbf{R}^d \rightarrow \{0,1\}$, where d is the dimension of the input. We are required to find a hyperplane matching (classifying) the output of our function.



label 1 for data points i = 1 to n,
label 0 for i = n + 1 to m.

$$\alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta > 0 \ \ \forall i \in \{1, 2, \cdots n\},$$

$$\alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta < 0 \ \ \forall i \in \{n+1, n+2, \cdots, m\}.$$

Example of LP formulation in ML

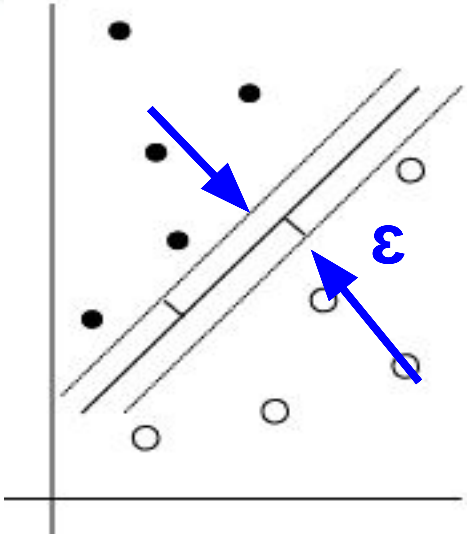
$$\alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta > 0 \ \ \forall i \in \{1, 2, \cdots n\},$$

$$\alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta < 0 \ \ \forall i \in \{n+1, n+2, \cdots, m\}.$$

$$\max \quad \epsilon$$

$$\text{s.t. } \alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta - \epsilon \geq 0 \ \ \forall i \in \{1, 2, \cdots n\}$$

$$\alpha_1 r_i + \alpha_2 a_i + \alpha_3 w_i + \alpha_4 f_i + \beta + \epsilon < 0 \ \ \forall i \in \{n+1, n+2, \cdots m\}$$



Special Convex Optimization Problems: QP

We say that a convex optimization problem is a **quadratic program (QP)** if the inequality constraints g_i are still all affine, but if the objective function f is a convex quadratic function.

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^T P x + c^T x + d \\ \text{subject to} & Gx \preceq h \\ & Ax = b \end{array}$$

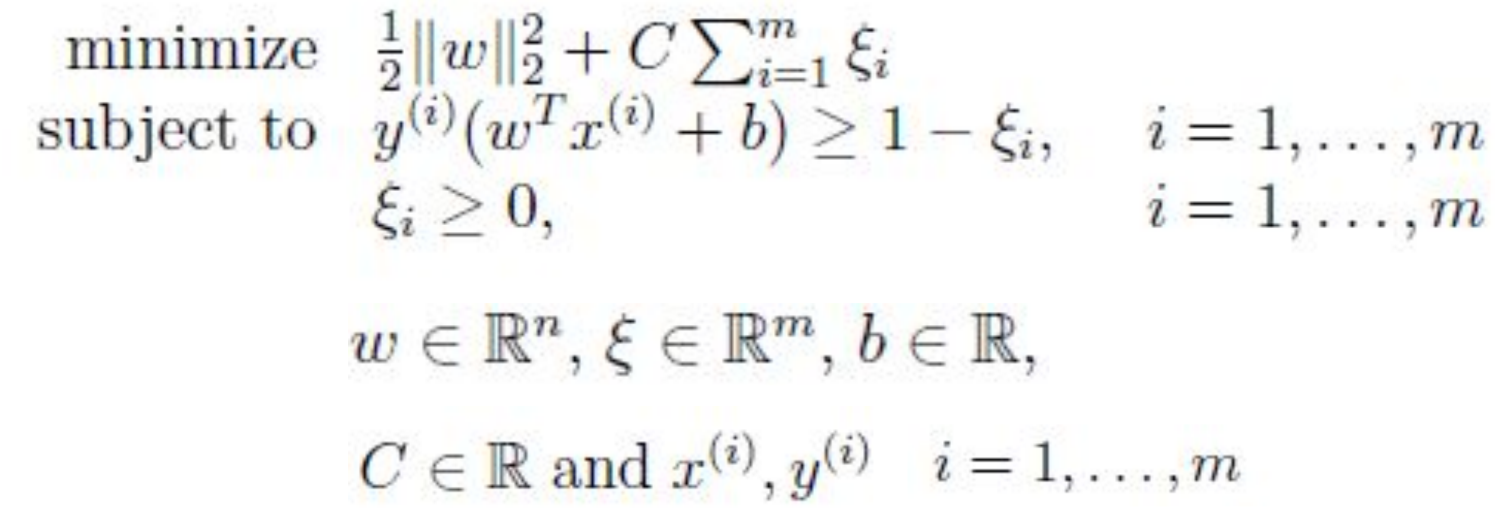
$$c \in \mathbb{R}^n,\ d \in \mathbb{R},\ G \in \mathbb{R}^{m\times n},\ h \in \mathbb{R}^m,$$

$$A \in \mathbb{R}^{p\times n},\ b \in \mathbb{R}^p$$

$P \in \mathbb{S}^n_+,$ symmetric positive semidefinite matrix.

Support Vector Machines

Support vector machines (SVM) optimization problem formulation with slack variables



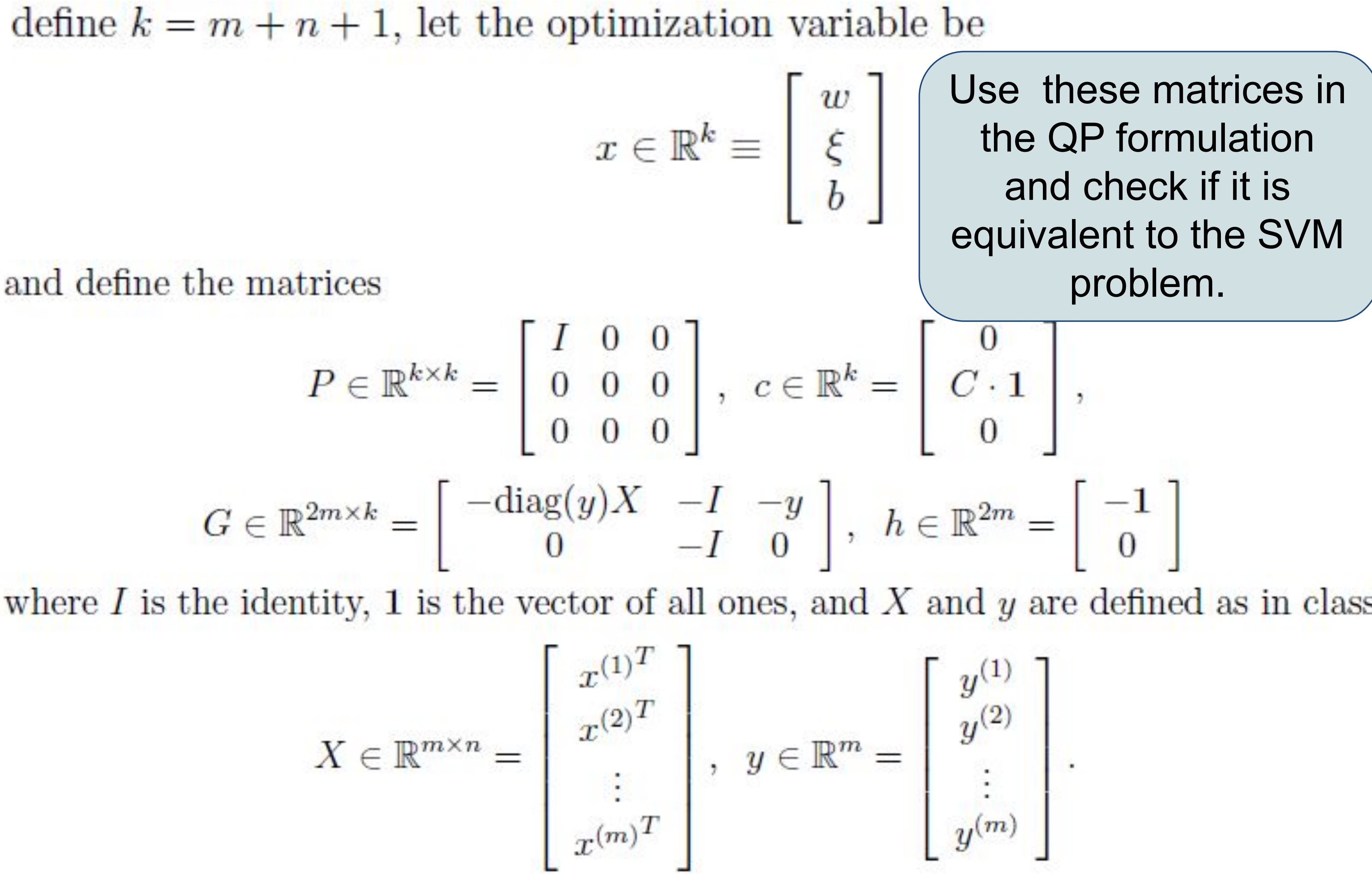
$$\begin{array}{lll} \text{minimize} & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^m \xi_i & \\ \text{subject to} & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, & i = 1, \ldots, m \\ & \xi_i \geq 0, & i = 1, \ldots, m \end{array}$$

$$w \in \mathbb{R}^n, \xi \in \mathbb{R}^m, b \in \mathbb{R},$$

$$C \in \mathbb{R} \text{ and } x^{(i)}, y^{(i)} \quad i = 1, \ldots, m$$

it is easy to see that there the SVM optimization problem has a quadratic objective and linear constraints,

SVM is a QP



define $k = m + n + 1$, let the optimization variable be

$$x \in \mathbb{R}^k \equiv \begin{bmatrix} w \\ \xi \\ b \end{bmatrix}$$

and define the matrices

$$P \in \mathbb{R}^{k \times k} = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad c \in \mathbb{R}^k = \begin{bmatrix} 0 \\ C \cdot \mathbf{1} \\ 0 \end{bmatrix},$$

$$G \in \mathbb{R}^{2m \times k} = \begin{bmatrix} -\mathrm{diag}(y)X & -I & -y \\ 0 & -I & 0 \end{bmatrix}, \quad h \in \mathbb{R}^{2m} = \begin{bmatrix} -\mathbf{1} \\ 0 \end{bmatrix}$$

where I is the identity, $\mathbf{1}$ is the vector of all ones, and X and y are defined as in class

$$X \in \mathbb{R}^{m \times n} = \begin{bmatrix} x^{(1)^T} \\ x^{(2)^T} \\ \vdots \\ x^{(m)^T} \end{bmatrix}, \quad y \in \mathbb{R}^m = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix}.$$

Use these matrices in the QP formulation and check if it is equivalent to the SVM problem.

Constrained least squares is QP

Consider the least squares problem, with a constraint that the entries in the solution **x** has to lie within some predefined ranges.

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||Ax - b||_2^2 \\ \text{subject to} & l \preceq x \preceq u \end{array}$$

$$A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m, l \in \mathbb{R}^n, \text{ and } u \in \mathbb{R}^n.$$

Note: The least squares problem can actually be solved analytically via the normal equations but it turns out that there will no longer be an analytical solution to the constrained least square problem.
This problem is a **quadratic program,** with matrices defined by

$$P \in \mathbb{R}^{n \times n} = \frac{1}{2}A^T A, \quad c \in \mathbb{R}^n = -b^T A, \quad d \in \mathbb{R} = \frac{1}{2}b^T b,$$

$$G \in \mathbb{R}^{2n \times 2n} = \left[\begin{array}{cc} -I & 0 \\ 0 & I \end{array} \right], \quad h \in \mathbb{R}^{2n} = \left[\begin{array}{c} -l \\ u \end{array} \right].$$

Special Convex Optimization Problems: QCQP

We say that a convex optimization problem is a quadratically constrained quadratic program (**QCQP**) if both the objective f and the inequality constraints g_i are convex quadratic functions,

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^T P x + c^T x + d \\ \text{subject to} & \frac{1}{2}x^T Q_i x + r_i^T x + s_i \leq 0, \quad i = 1, \ldots, m \\ & Ax = b \end{array}$$

$$c \in \mathbb{R}^n, d \in \mathbb{R}, A \in \mathbb{R}^{p \times n}, b \in \mathbb{R}^p,$$

$$P \in \mathbb{S}^n_+, \quad Q_i \in \mathbb{S}^n_+$$

symmetric positive semidefinite matrix.

$$r_i \in \mathbb{R}^n, \ s_i \in \mathbb{R}, \text{ for } i = 1, \ldots, m.$$

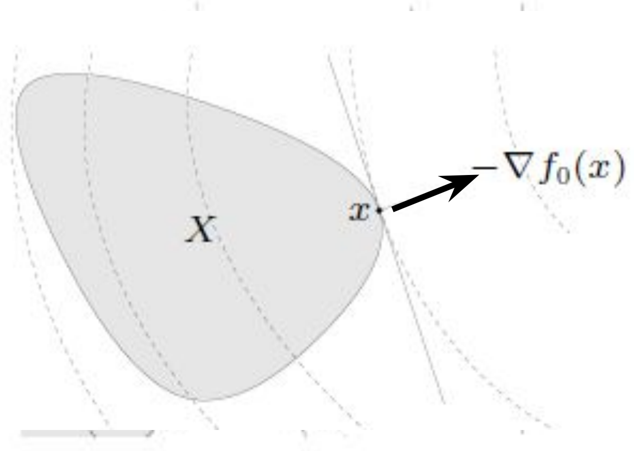
Example: Kernel regression models are QCQP (Advanced ML !)

First-order optimality condition

$$\min_x \ f(x) \quad \text{subject to} \quad x \in C$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

Important special case: if $C = \mathbb{R}^n$ (unconstrained optimization), then optimality condition reduces to familiar $\nabla f(x) = 0$

First-order optimality condition: Example

$$f(x) = \frac{1}{2}x^T Q x + b^T x + c$$

where $Q \succeq 0$. The first-order condition says that solution satisfies

$$\nabla f(x) = Qx + b = 0$$

- if $Q \succ 0$, then there is a unique solution $x = -Q^{-1}b$
- if Q is singular and $b \notin \mathrm{col}(Q)$, then there is no solution (i.e., $\min_x \ f(x) = -\infty$)
- if Q is singular and $b \in \mathrm{col}(Q)$, then there are infinitely many solutions

  $$x = -Q^+ b + z, \quad z \in \mathrm{null}(Q)$$

  where Q^+ is the pseudoinverse of Q

First-order optimality condition: Example

Consider the equality-constrained convex problem:

$$\min_{x} \; f(x) \;\; \text{subject to} \;\; Ax = b$$

According to first-order optimality, solution x satisfies $Ax = b$ and

$$\nabla f(x)^T (y - x) \geq 0 \;\; \text{for all } y \text{ such that } Ay = b$$

This is equivalent to

$$\nabla f(x)^T v = 0 \;\; \text{for all } v \in \text{null}(A)$$

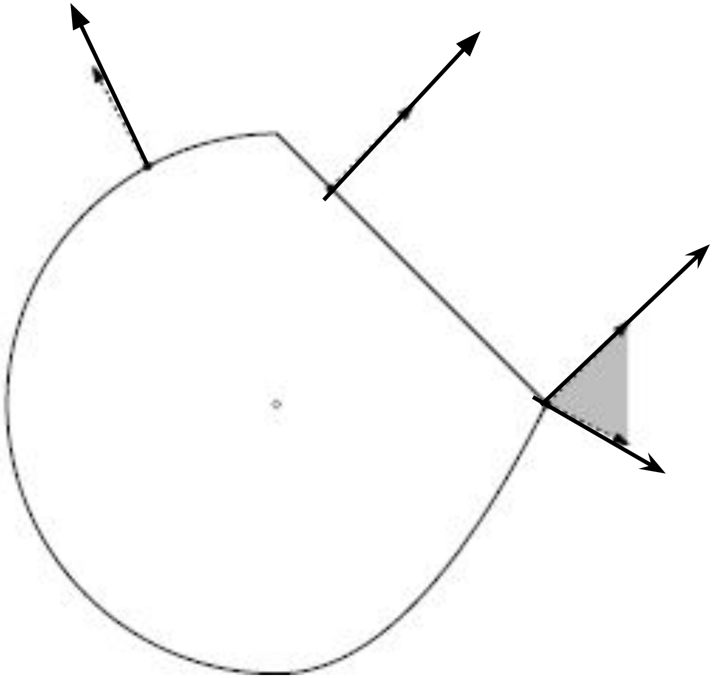
$$\text{null}(A)^{\perp} = \text{row}(A)$$

First-order optimality condition: Example

$$\min_x \ \|a - x\|_2^2 \ \text{ subject to } \ x \in C$$

First-order optimality condition says that the solution x satisfies

$$\nabla f(x)^T (y - x) = (x - a)^T (y - x) \geq 0 \ \text{ for all } y \in C$$



Equivalently, this says that

$$a - x \in \mathcal{N}_C(x)$$

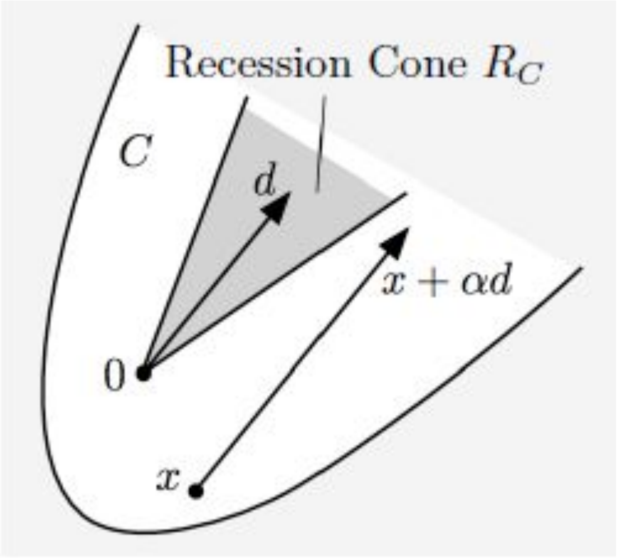
where recall $\mathcal{N}_C(x)$ is the normal cone to C at x

Recession Cone and recession direction

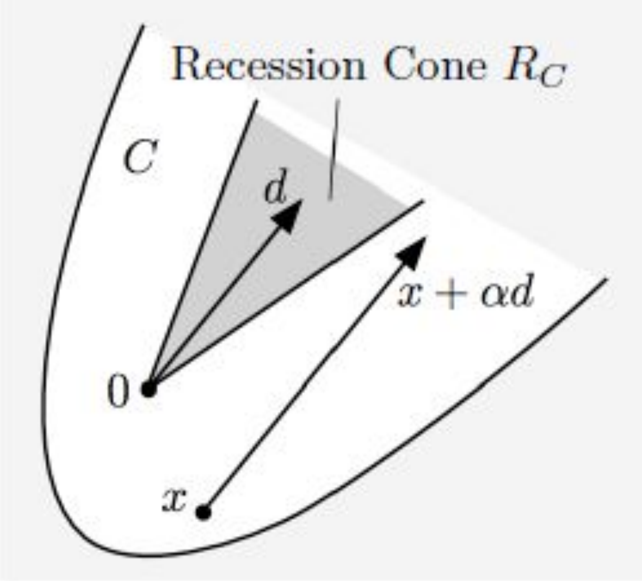
Given a nonempty convex set C, we say that a vector **d** is a **recession direction** of C if $\mathbf{x} + \alpha\mathbf{d} \in$ C for all $\mathbf{x} \in$ C and $\alpha \geq 0$. Thus, **d** is a direction of recession of C if starting at any **x** in C and going indefinitely along **d**, we never cross the relative boundary of C to points outside C.

The set of all directions of recession is a cone containing the origin. It is called the **recession cone** of C and it is denoted by R_C

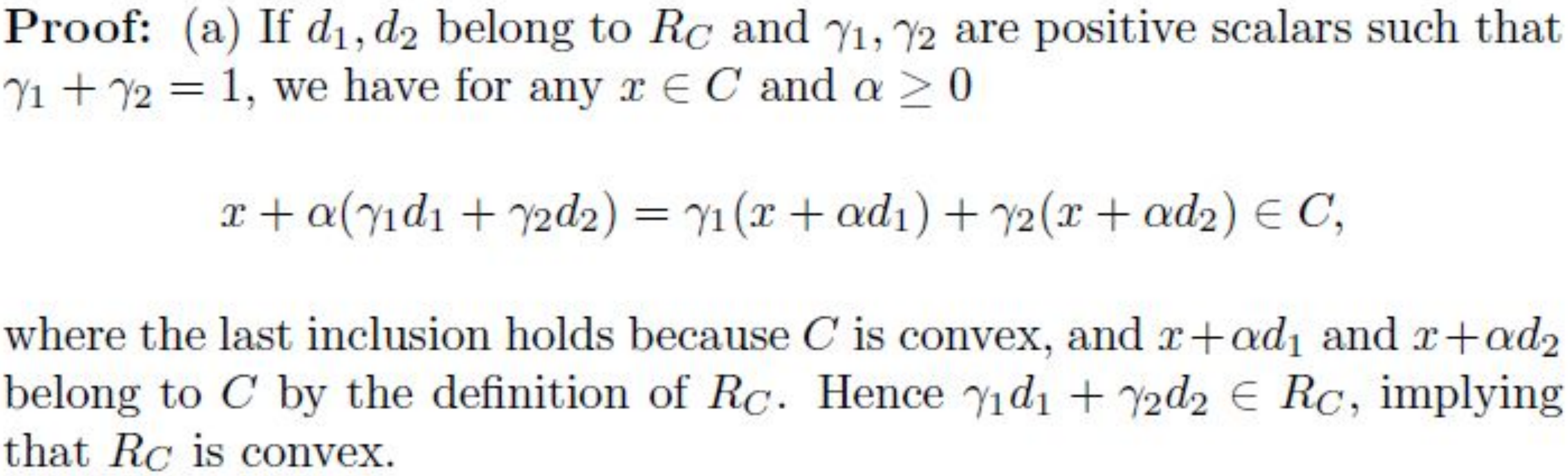
Thus $\mathbf{d} \in R_C$ if $\mathbf{x} + \alpha\mathbf{d} \in$ C for all $\mathbf{x} \in$ C and $\alpha \geq 0$.



Recession cone of a convex set



Theorem: If C is a nonempty closed convex set. then the recession cone R_C is also closed and convex.



Proof: (a) If d_1, d_2 belong to R_C and γ_1, γ_2 are positive scalars such that $\gamma_1 + \gamma_2 = 1$, we have for any $x \in C$ and $\alpha \geq 0$

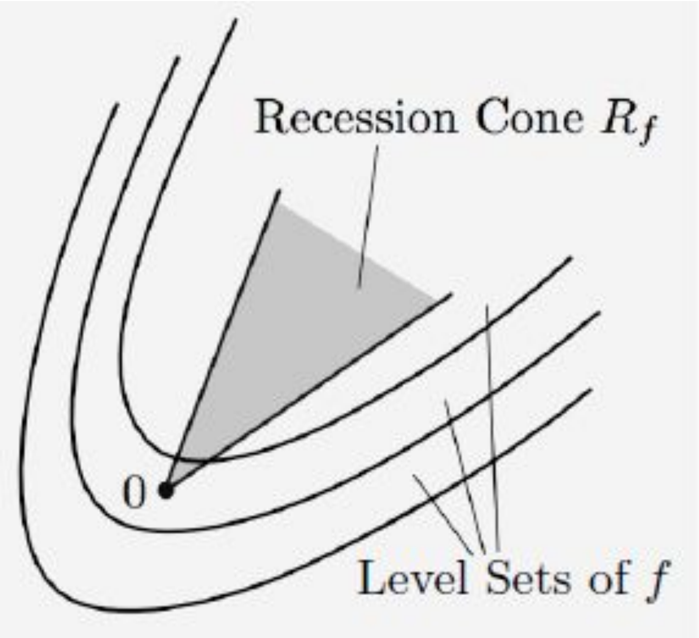
$$x + \alpha(\gamma_1 d_1 + \gamma_2 d_2) = \gamma_1(x + \alpha d_1) + \gamma_2(x + \alpha d_2) \in C,$$

where the last inclusion holds because C is convex, and $x + \alpha d_1$ and $x + \alpha d_2$ belong to C by the definition of R_C. Hence $\gamma_1 d_1 + \gamma_2 d_2 \in R_C$, implying that R_C is convex.

Direction of Recession of Convex function

Note:

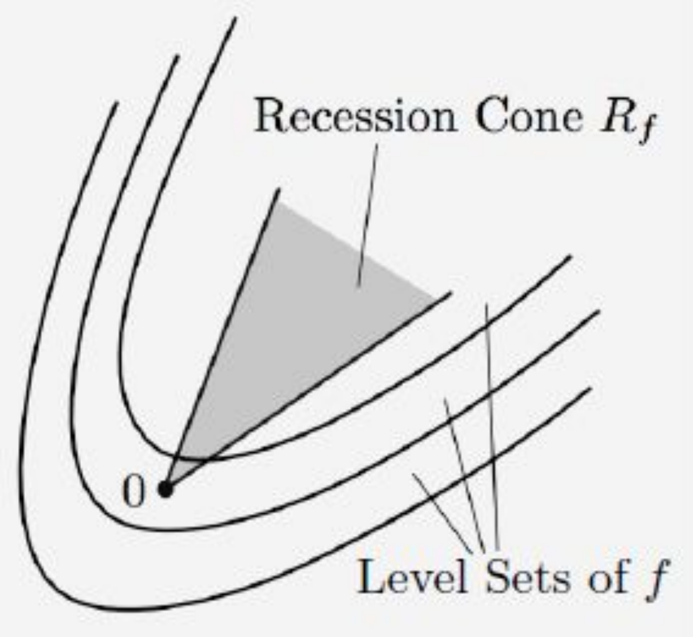
- A function f is convex if and only if its epigraph is a convex set.
- The recession cone of epi(f) can be used to obtain the directions along which f does not increase monotonically.
- The directions in the recession cone of epi(f) correspond to the directions along which the level sets $\{x \mid f(x) \leq \gamma\}$ are unbounded.
- Along these directions, f is monotonically nonincreasing.



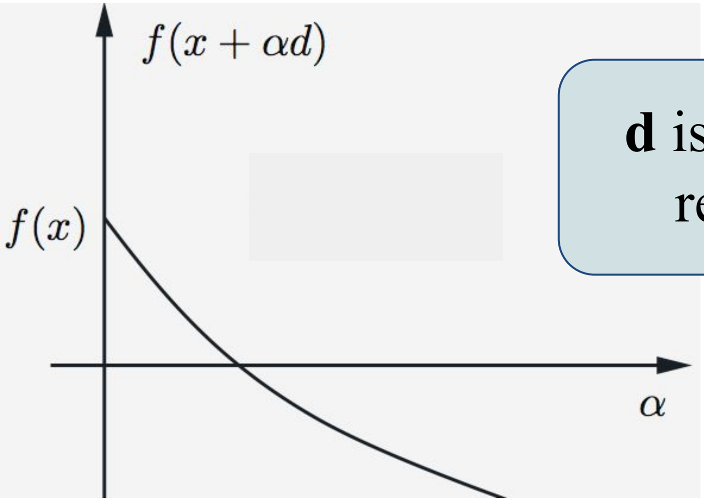
Direction of Recession of Convex function

Note:

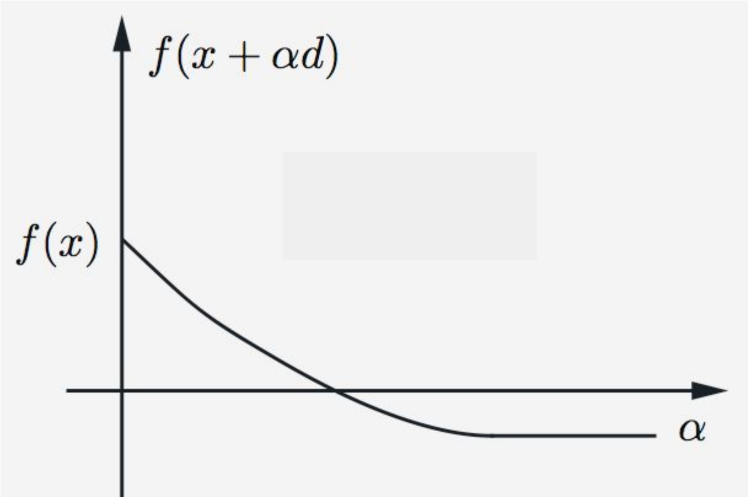
- Let $f : \Re^n \to (-\infty,\infty]$ be closed proper convex function.
- Let S ={**x** | f (**x**) ≤ γ } be the nonempty level sets
- All level sets have common recession direction, Let R_f be the cone of recession directions of the nonempty level sets S.
- If we start at any **x** ∈ dom(f) and move indefinitely along a direction of recession, we must stay within each level set that contains **x**, or equivalently we must encounter exclusively points **z** with f (**z**) ≤ f (**x**).
- In words, a direction of recession of f is a direction of continuous nonascent for f



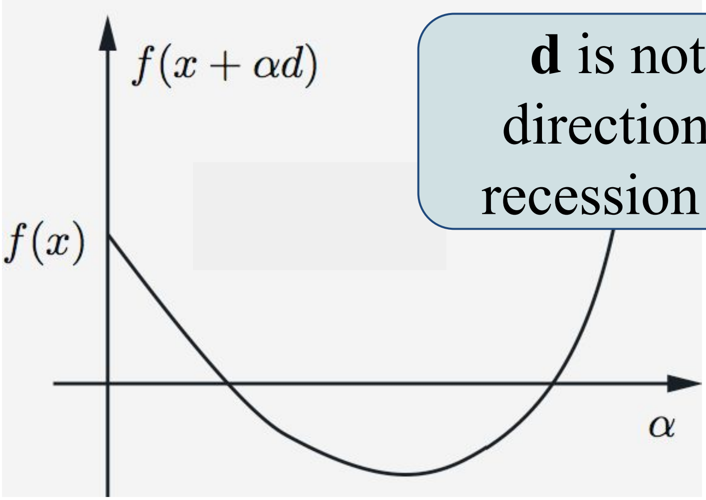
Direction of Recession of Convex function



d is a direction of recession of *f*



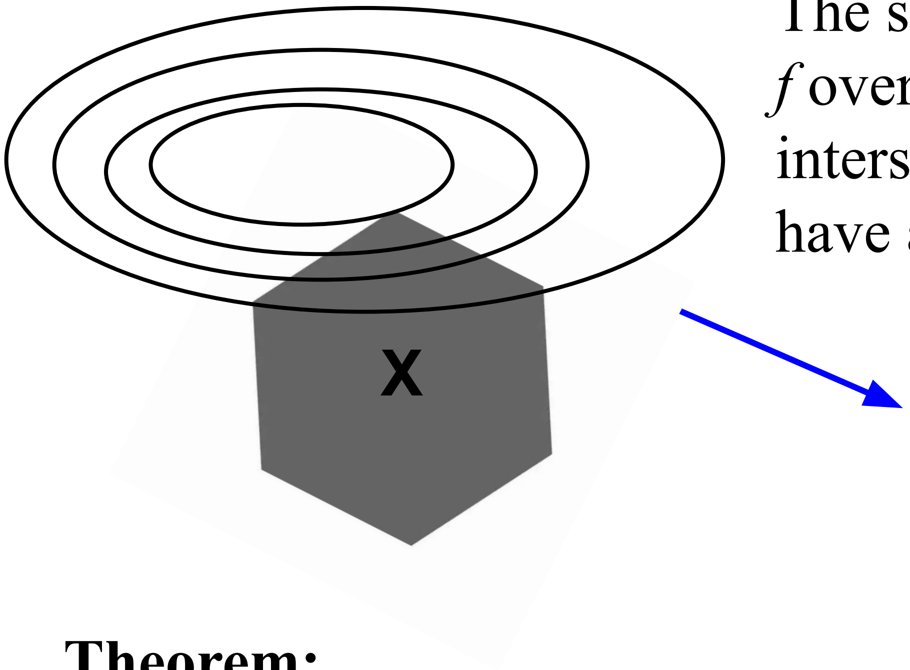
Ascent/descent behavior of a closed proper convex function starting at some **x** $\in$ dom(*f*) and moving along a direction **d**.



d is not a direction of recession of *f*

Existence of optimal solution to COP

The set of minima of a real-valued function *f* over a nonempty set **X**, is equal to the intersection of **X** and the level sets of *f* that have a common point with **X**:.



$$X^* = \cap_{k=0}^{\infty} \{x \in X \mid f(x) \leq \gamma_k\},$$

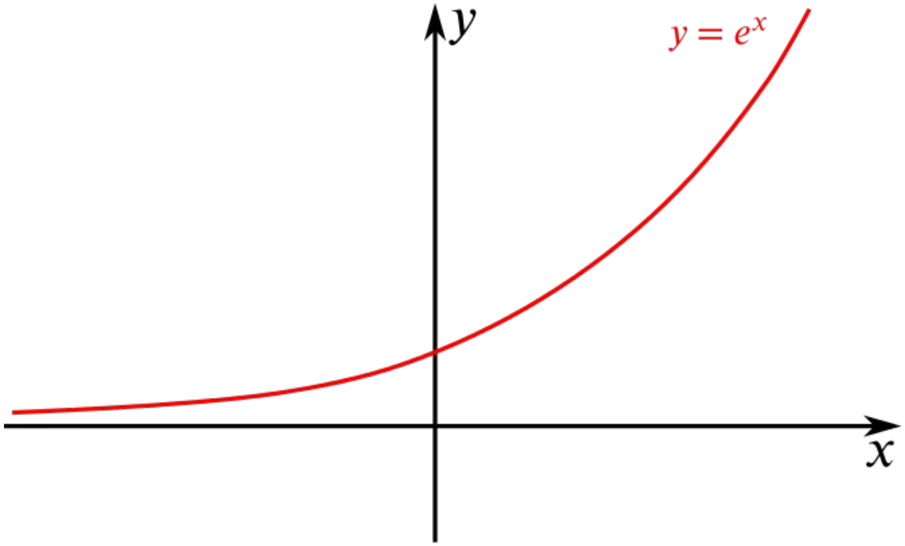
Theorem:

Let **X** be a closed convex subset of $\Re^n$, and let $f : \Re^n \rightarrow (-\infty,\infty]$ be a closed convex function with **X** ∩dom(*f*) ≠ Ø.

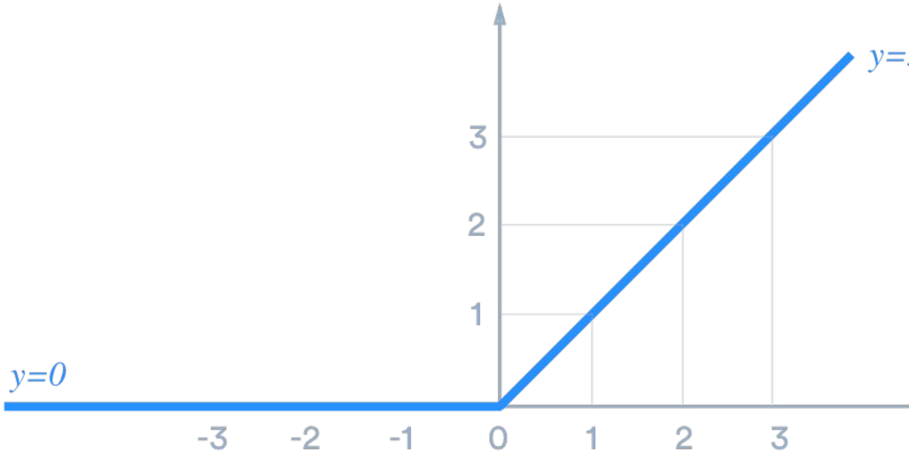
The set of minima of *f* over **X** is nonempty and compact if and only if **X** and *f* have no common nonzero direction of recession

Existence of optimal solution to COP

What if **X** and *f* have a common direction of recession?



$\mathbf{X} = \Re$ and $f(\mathrm{x}) = \mathrm{e}^{\mathrm{x}}$
The optimal solution set is empty



$\mathbf{X} = \Re$ and $f(\mathrm{x}) = \max\{0, \mathrm{x}\}$
The optimal solution set is nonempty and unbounded

Partial minimization

Consider a function $F : \Re^{n+m} \rightarrow (-\infty,\infty]$ and the function $f : \Re^{n} \rightarrow [-\infty,\infty]$ defined by

$$f(x) = \inf_{z \in \Re^m} F(x,z).$$

If F is convex, then f is also convex.

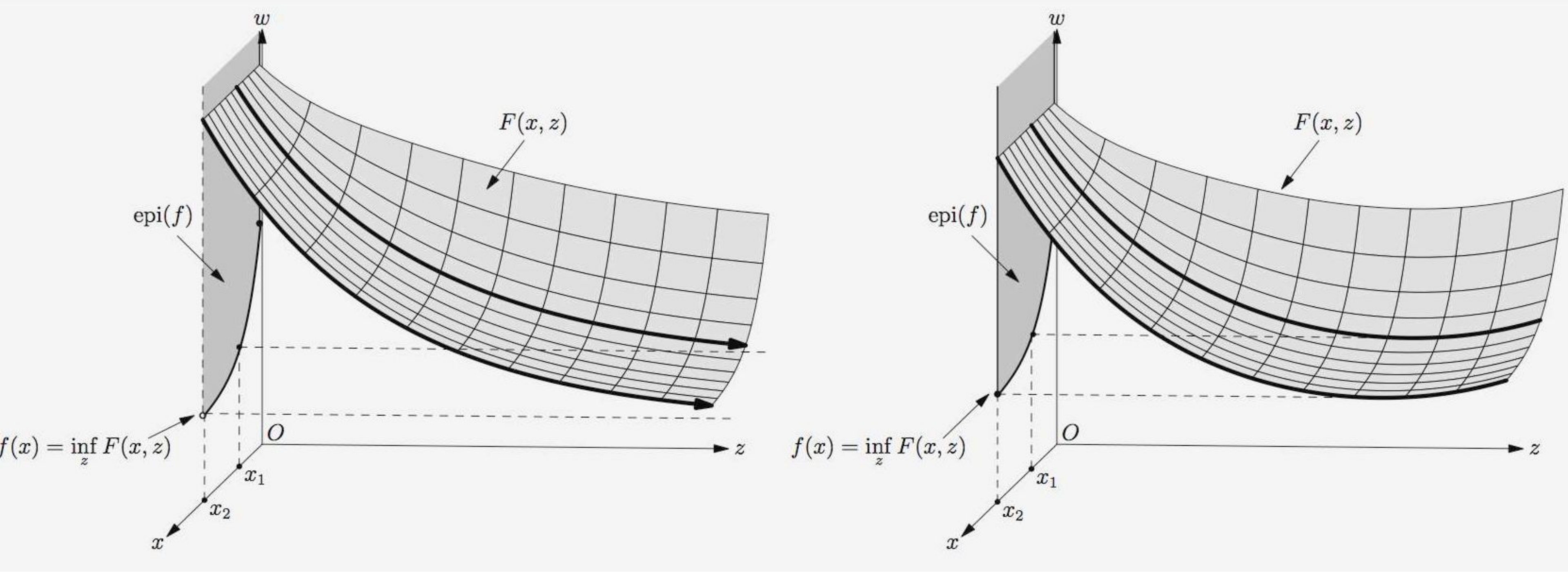
Example:

$f(x,y) = x^T A x + 2x^T B y + y^T C y$ with

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0, \qquad C \succ 0$$

minimizing over y gives $g(x) = \inf_y f(x,y) = x^T(A - BC^{-1}B^T)x$

Partial minimization



for a fixed x, F(x, z) attains a minimum over z if and only if { x, f(x) }belongs to P{epi(F)}