

# AA203

# Optimal and Learning-based Control

Optimization theory



**Stanford**  
University



# Outline

1. Computational methods for unconstrained optimization
2. Optimization with equality constraints
3. Optimization with inequality constraints

# Outline

1. Computational methods for unconstrained optimization
2. Optimization with equality constraints
3. Optimization with inequality constraints

# Computational methods (unconstrained case)

**Goal:** find “numerical recipes” to solve optimization problem

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

**Key idea:** iterative descent. We start at some point  $\mathbf{x}^0$  (initial guess) and successively generate vectors  $\mathbf{x}^1, \mathbf{x}^2, \dots$  such that  $f$  is decreased at each iteration, i.e.,

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k), \quad k = 0, 1, \dots$$

The hope is to decrease  $f$  all the way to the minimum

# Gradient methods

Given  $\mathbf{x} \in \mathbb{R}^n$  with  $\nabla f(\mathbf{x}) \neq 0$ , consider the half line of vectors

$$\mathbf{x}_\alpha = \mathbf{x} - \alpha \nabla f(\mathbf{x}), \quad \forall \alpha \geq 0$$

From first order Taylor expansion ( $\alpha$  small)

$$f(\mathbf{x}_\alpha) \approx f(\mathbf{x}) + \nabla f(\mathbf{x})'(\mathbf{x}_\alpha - \mathbf{x}) = f(\mathbf{x}) - \alpha \|\nabla f(\mathbf{x})\|^2$$

So for  $\alpha$  small enough  $f(\mathbf{x}_\alpha)$  is smaller than  $f(\mathbf{x})$ !

# Gradient methods

Carrying this idea one step further, consider the half line of vectors

$$\mathbf{x}_\alpha = \mathbf{x} + \alpha \mathbf{d}, \quad \forall \alpha \geq 0$$

where  $\nabla f(\mathbf{x})' \mathbf{d} < 0$  (angle  $> 90^\circ$ )

By Taylor expansion

$$f(\mathbf{x}_\alpha) \approx f(\mathbf{x}) + \alpha \nabla f(\mathbf{x})' \mathbf{d}$$

For small enough  $\alpha$ ,  $f(\mathbf{x} + \alpha \mathbf{d})$  is smaller than  $f(\mathbf{x})$ !

# Gradient methods

Broad and important class of algorithms: **gradient methods**

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \alpha^k \mathbf{d}^k, \quad k = 0, 1, \dots$$

where if  $\nabla f(\mathbf{x}^k) \neq 0$ ,  $\mathbf{d}^k$  is chosen so that

$$\nabla f(\mathbf{x}^k)' \mathbf{d}^k < 0$$

and the stepsize  $\alpha$  is chosen to be positive

# Gradient descent

Most often the stepsize is chosen so that

$$f(\mathbf{x}^k + \alpha^k \mathbf{d}^k) < f(\mathbf{x}^k), \quad k = 0, 1, \dots$$

and the method is called **gradient descent**. “Tuning” parameters:

- selecting the descent direction
- selecting the stepsize

# Selecting the descent direction

General class

$$\mathbf{d}^k = -D^k \nabla f(\mathbf{x}^k), \quad \text{where } D^k > 0$$

(Obviously,  $\nabla f(\mathbf{x}^k)' \mathbf{d}^k < 0$ )

Popular choices:

- **Steepest descent:**  $D^k = I$
- **Newton's method:**  $D^k = (\nabla^2 f(\mathbf{x}^k))^{-1}$  provided  $\nabla^2 f(\mathbf{x}^k) > 0$

# Selecting the stepsize

- **Minimization rule:**  $\alpha^k$  is selected such that the cost function is minimized along the direction  $\mathbf{d}^k$ , i.e.,

$$f(\mathbf{x}^k + \alpha^k \mathbf{d}^k) = \min_{\alpha \geq 0} f(\mathbf{x}^k + \alpha \mathbf{d}^k)$$

- **Constant stepsize:**  $\alpha^k = s$ 
  - the method might diverge
  - convergence rate could be very slow
- **Diminishing stepsize:**  $\alpha^k \rightarrow 0$  and  $\sum_{k=0}^{+\infty} \alpha^k = \infty$ 
  - it does not guarantee descent at each iteration

# Undiscussed in this class

Mathematical analysis:

- convergence (to stationary points)
- termination criteria
- convergence rate

Derivative-free methods, e.g.,

- coordinate descent
- Nelder-Mead

# Next: constrained optimization

- constraint set usually specified in terms of equality and inequality constraints
- sophisticated collection of optimality conditions, involving some auxiliary variables, called Lagrange multipliers

## Viewpoints:

- penalty viewpoint: we disregard the constraints and we add to the cost a high penalty for violating them
- feasibility direction viewpoint: it relies on the fact that at a local minimum there can be no cost improvement when traveling a small distance along a direction that leads to feasible points

# Outline

1. Computational methods for unconstrained optimization
2. Optimization with equality constraints
3. Optimization with inequality constraints

# Optimization with equality constraints

$$\begin{aligned} & \min f(\mathbf{x}) \\ & \text{subject to } h_i(\mathbf{x}) = 0, \quad i = 1, \dots, m \end{aligned}$$

- $f: \mathbb{R}^n \rightarrow \mathbb{R}$  and  $h_i: \mathbb{R}^n \rightarrow \mathbb{R}$  are  $C^1$
- notation:  $\mathbf{h} := (h_1, \dots, h_m)$

# Lagrange multipliers

- **Basic Lagrange multiplier theorem:** for a given local minimum  $\mathbf{x}^*$  there exist scalars  $\lambda_1, \dots, \lambda_m$  called Lagrange multipliers such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla h_i(\mathbf{x}^*) = 0$$

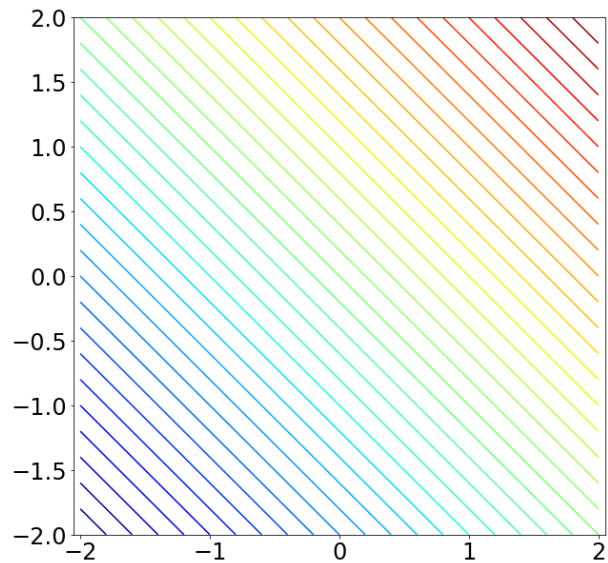
- Example

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 2 \end{aligned}$$

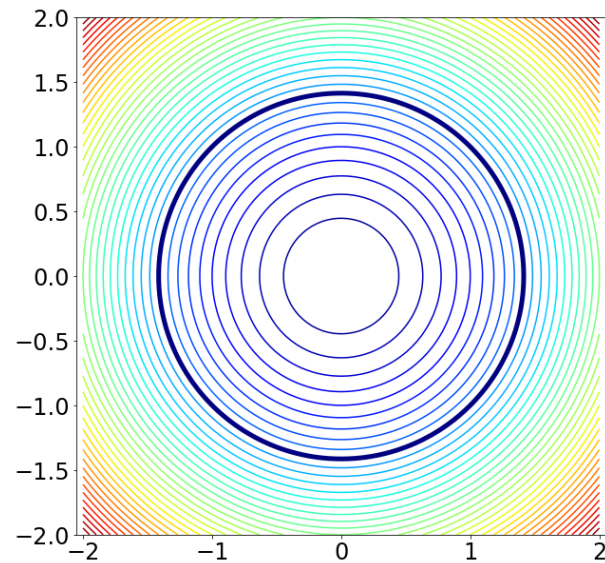
# Lagrange multipliers

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 = 2 \end{aligned}$$

$$f(\mathbf{x}) = x_1 + x_2$$



$$h(\mathbf{x}) = x_1^2 + x_2^2 - 2$$



# Lagrange multipliers

- **Basic Lagrange multiplier theorem:** for a given local minimum  $\mathbf{x}^*$  there exist scalars  $\lambda_1, \dots, \lambda_m$  called Lagrange multipliers such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla h_i(\mathbf{x}^*) = 0$$

- Example

$$\begin{array}{ll} \min & x_1 + x_2 \\ \text{subject to} & x_1^2 + x_2^2 = 2 \end{array} \quad \text{Solution: } \mathbf{x}^* = (-1, -1)$$

# Lagrange multipliers

## Interpretations:

1. The cost gradient  $\nabla f(\mathbf{x}^*)$  belongs to the subspace spanned by the constraint gradients at  $\mathbf{x}^*$ . That is, the constrained solution will be at a point of tangency of the constrained cost curves and the constraint function
2. The cost gradient  $\nabla f(\mathbf{x}^*)$  is orthogonal to the subspace of first order feasible variations

$$V(\mathbf{x}^*) = \{ \Delta \mathbf{x} \mid \nabla h_i(\mathbf{x}^*)' \Delta \mathbf{x} = 0, \quad i = 1, \dots, m \}$$

This is the subspace of variations  $\Delta \mathbf{x}$  for which the vector  $\mathbf{x} = \mathbf{x}^* + \Delta \mathbf{x}$  satisfies the constraint  $\mathbf{h}(\mathbf{x}) = 0$  up to first order. Hence, at a local minimum, the first order cost variation  $\nabla f(\mathbf{x}^*)' \Delta \mathbf{x}$  is zero for all variations  $\Delta \mathbf{x}$  in this subspace

# NOC

## Theorem: NOC

Let  $\mathbf{x}^*$  be a local minimum of  $f$  subject to  $\mathbf{h}(\mathbf{x}) = 0$  and assume that the constraint gradients  $\nabla h_1(\mathbf{x}^*), \dots, \nabla h_m(\mathbf{x}^*)$  are linearly independent. Then there exists a unique vector  $(\lambda_1, \dots, \lambda_m)$ , called a Lagrange multiplier vector, such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i \nabla h_i(\mathbf{x}^*) = 0$$

(2<sup>nd</sup> order NOC and SOC are provided in [AA203-Notes](#))

# Discussion

- A feasible vector  $\mathbf{x}$  for which  $\{\nabla h_i(\mathbf{x})\}_i$  are linearly independent is called *regular*
- Proof relies on transforming the constrained problem into an unconstrained one
  1. penalty approach: we disregard the constraints while adding to the cost a high penalty for violating them  $\rightarrow$  extends to inequality constraints
  2. elimination approach: we view the constraints as a system of  $m$  equations with  $n$  unknowns, and we express  $m$  of the variables in terms of the remaining  $n - m$ , thereby reducing the problem to an unconstrained problem
- There may not exist a Lagrange multiplier for a local minimum that is not regular

# The Lagrangian function

- It is often convenient to write the necessary conditions in terms of the Lagrangian function  $L: \mathbb{R}^{n+m} \rightarrow \mathbb{R}$

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i h_i(\mathbf{x})$$

- Then, if  $\mathbf{x}^*$  is a local minimum which is regular, the NOC conditions are compactly written

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*) = 0$$

$$\nabla_{\lambda} L(\mathbf{x}^*, \lambda^*) = 0$$

System of  $n + m$  equations  
with  $n + m$  unknowns

# Outline

1. Computational methods for unconstrained optimization
2. Optimization with equality constraints
3. Optimization with inequality constraints

# Optimization with inequality constraints

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) = 0, \quad i = 1, \dots, m \\ & g_j(\mathbf{x}) \leq 0, \quad j = 1, \dots, r \end{aligned}$$

- $f, h_i, g_j$  are  $C^1$
- In compact form (ICP problem)

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{subject to} \quad & \mathbf{h}(\mathbf{x}) = 0 \\ & \mathbf{g}(\mathbf{x}) \leq 0 \end{aligned}$$

# Active constraints

For any feasible point, the set of active inequality constraints is denoted

$$A(\mathbf{x}) := \{j \mid g_j(\mathbf{x}) = 0\}$$

If  $j \notin A(\mathbf{x})$ , then the constraint is *inactive* at  $\mathbf{x}$ .

## Key points

- if  $\mathbf{x}^*$  is a local minimum of the ICP, then  $\mathbf{x}^*$  is also a local minimum for the identical ICP without the inactive constraints
- at a local minimum, active inequality constraints can be treated to a large extent as equalities

# Active constraints

- Hence, if  $\mathbf{x}^*$  is a local minimum of ICP, then  $\mathbf{x}^*$  is also a local minimum for the **equality** constrained problem

$$\begin{aligned} & \min && f(\mathbf{x}) \\ & \text{subject to} && \mathbf{h}(\mathbf{x}) = 0 \\ & && g_j(\mathbf{x}) = 0, \quad \forall j \in A(\mathbf{x}^*) \end{aligned}$$

# Active constraints

- Thus if  $\mathbf{x}^*$  is regular, there exist Lagrange multipliers  $(\lambda_1, \dots, \lambda_m)$  and  $\mu_j^*, j \in A(\mathbf{x}^*)$ , such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^n \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j \in A(\mathbf{x}^*)} \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

- or equivalently

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^n \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(\mathbf{x}^*) = 0$$

$$\mu_j^* = 0 \quad \forall j \notin A(\mathbf{x}^*) \quad (\text{indeed } \mu_j^* \geq 0)$$

# Karush-Kuhn-Tucker NOC

Define the Lagrangian function

$$L(\mathbf{x}, \lambda, \mu) := f(\mathbf{x}) + \sum_{i=1}^n \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^r \mu_j g_j(\mathbf{x})$$

## Theorem: KKT NOC

Let  $\mathbf{x}^*$  be a local minimum for ICP where  $f, h_i, g_j$  are  $C^1$  and assume  $\mathbf{x}^*$  is regular (equality + active inequality constraints gradients are linearly independent). Then, there exist unique Lagrange multiplier vectors  $(\lambda_1^*, \dots, \lambda_m^*), (\mu_1^*, \dots, \mu_m^*)$  such that

$$\nabla_{\mathbf{x}} L(\mathbf{x}^*, \lambda^*, \mu^*) = 0$$

$$\mu_j^* \geq 0, \quad j = 1, \dots, r$$

$$\mu_j^* = 0 \quad \forall j \notin A(\mathbf{x}^*)$$

# Example

$$\begin{aligned} \min \quad & x^2 + y^2 \\ \text{s. t.} \quad & 2x + y \leq 2 \end{aligned}$$

Solution: (0,0)

# Next time

Calculus of variations  
(infinite-dimensional optimization!)