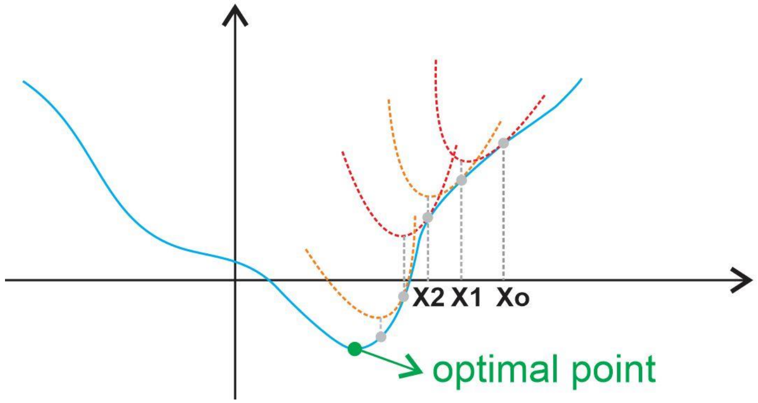
AA203
Optimal and Learning-based Control

Optimization theory

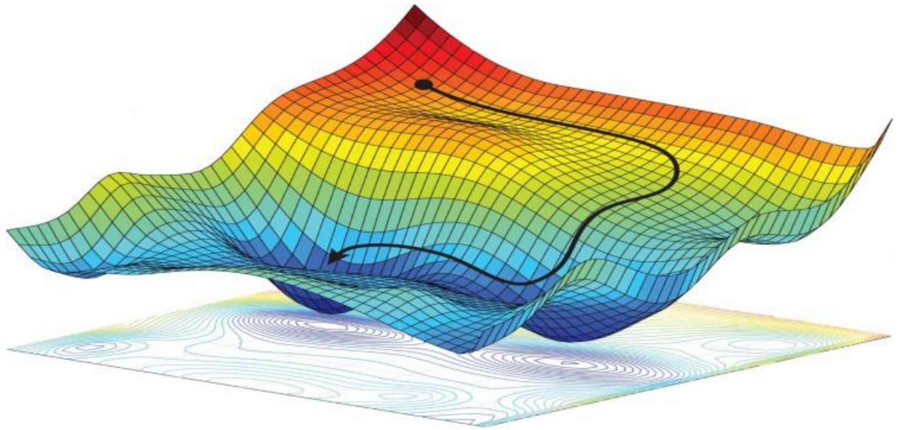
Stanford University

Autonomous Systems Lab

Optimization in many dimensions

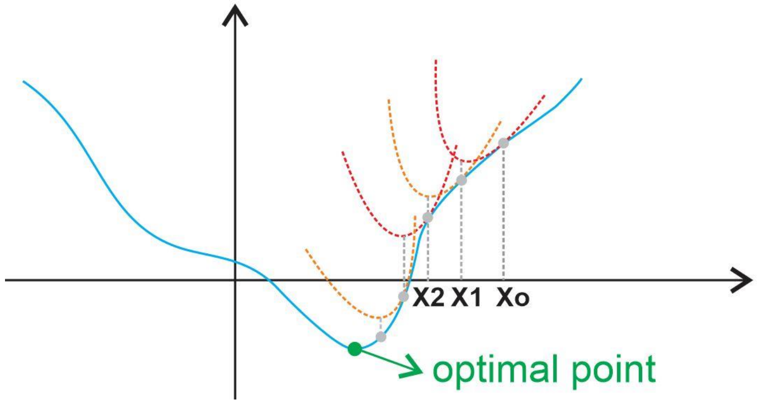


1-D

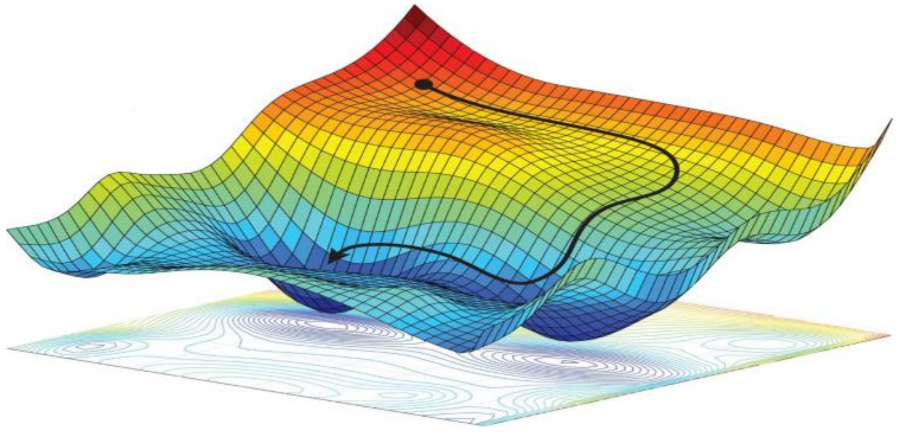


2-D

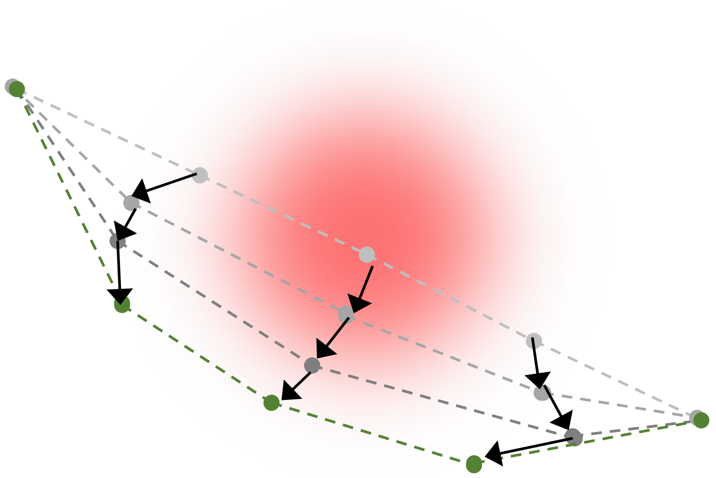
Optimization in many dimensions



1-D

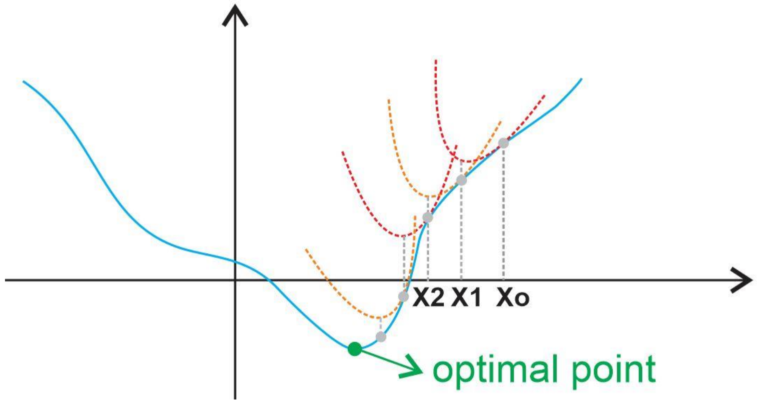


2-D

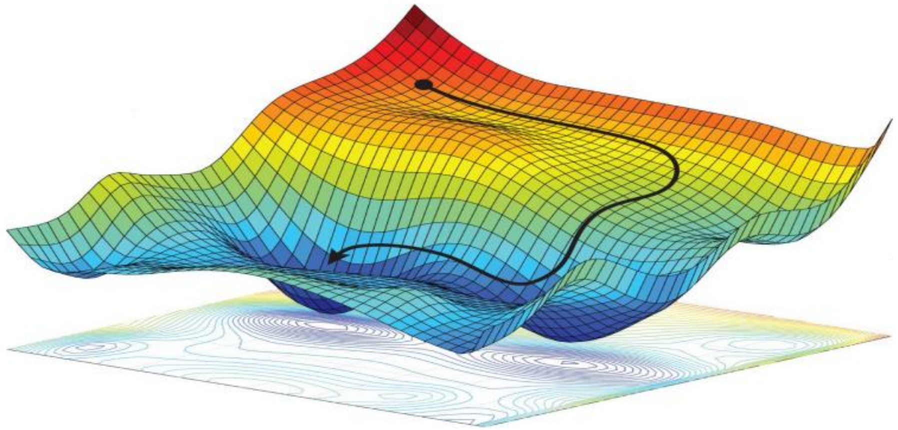


10-D

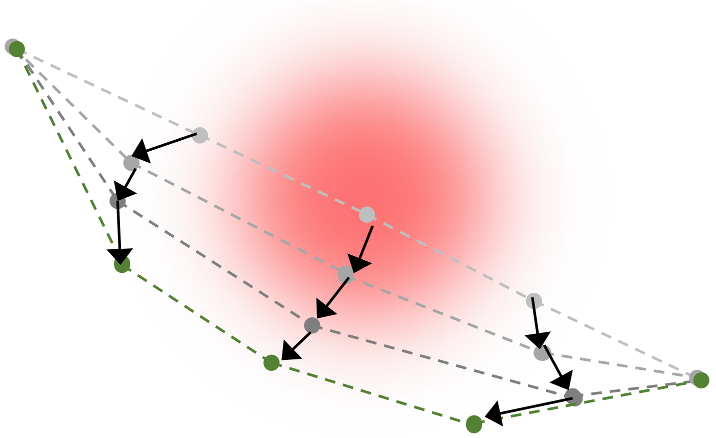
Optimization in many dimensions



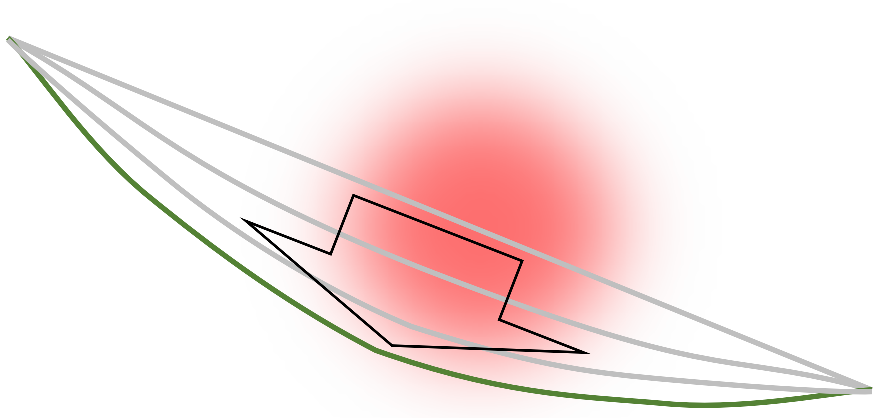
1-D



2-D



10-D



∞-D

Outline

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

Unconstrained optimization

Unconstrained non-linear program

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

- f usually assumed continuously differentiable (and often twice continuously differentiable)

Local and global minima

- A vector $\mathbf{x}^*$ is said to be an unconstrained *local* minimum if $\exists \epsilon > 0$ such that

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \qquad \forall \mathbf{x} \,\big|\, \|\mathbf{x} - \mathbf{x}^*\| < \epsilon$$

- A vector $\mathbf{x}^*$ is said to be an unconstrained *global* minimum if

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathbb{R}^n$$

- $\mathbf{x}^*$ is a strict local/global minimum if the inequality is strict

Necessary conditions for optimality

Key idea: compare cost of a vector with cost of its close neighbors

- Assume $f \in C^1$, by using Taylor series expansion

$$f(\mathbf{x}^* + \Delta\mathbf{x}) - f(\mathbf{x}^*) \approx \nabla f(\mathbf{x}^*)'\Delta\mathbf{x}$$

- If $f \in C^2$

$$f(\mathbf{x}^* + \Delta\mathbf{x}) - f(\mathbf{x}^*) \approx \nabla f(\mathbf{x}^*)'\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}'\nabla^2 f(\mathbf{x}^*)\Delta\mathbf{x}$$

Necessary conditions for optimality

- We expect that if $\mathbf{x}^*$ is an unconstrained local minimum, the first order cost variation due to a small variation $\Delta\mathbf{x}$ is nonnegative, i.e.,

$$\nabla f(\mathbf{x}^*)'\Delta\mathbf{x} = \sum_{i=1}^{n} \frac{\partial f(\mathbf{x}^*)}{\partial x_i}\Delta x_i \geq 0$$

- By taking $\Delta\mathbf{x}$ to be positive and negative multiples of the unit coordinate vectors, we obtain conditions of the type

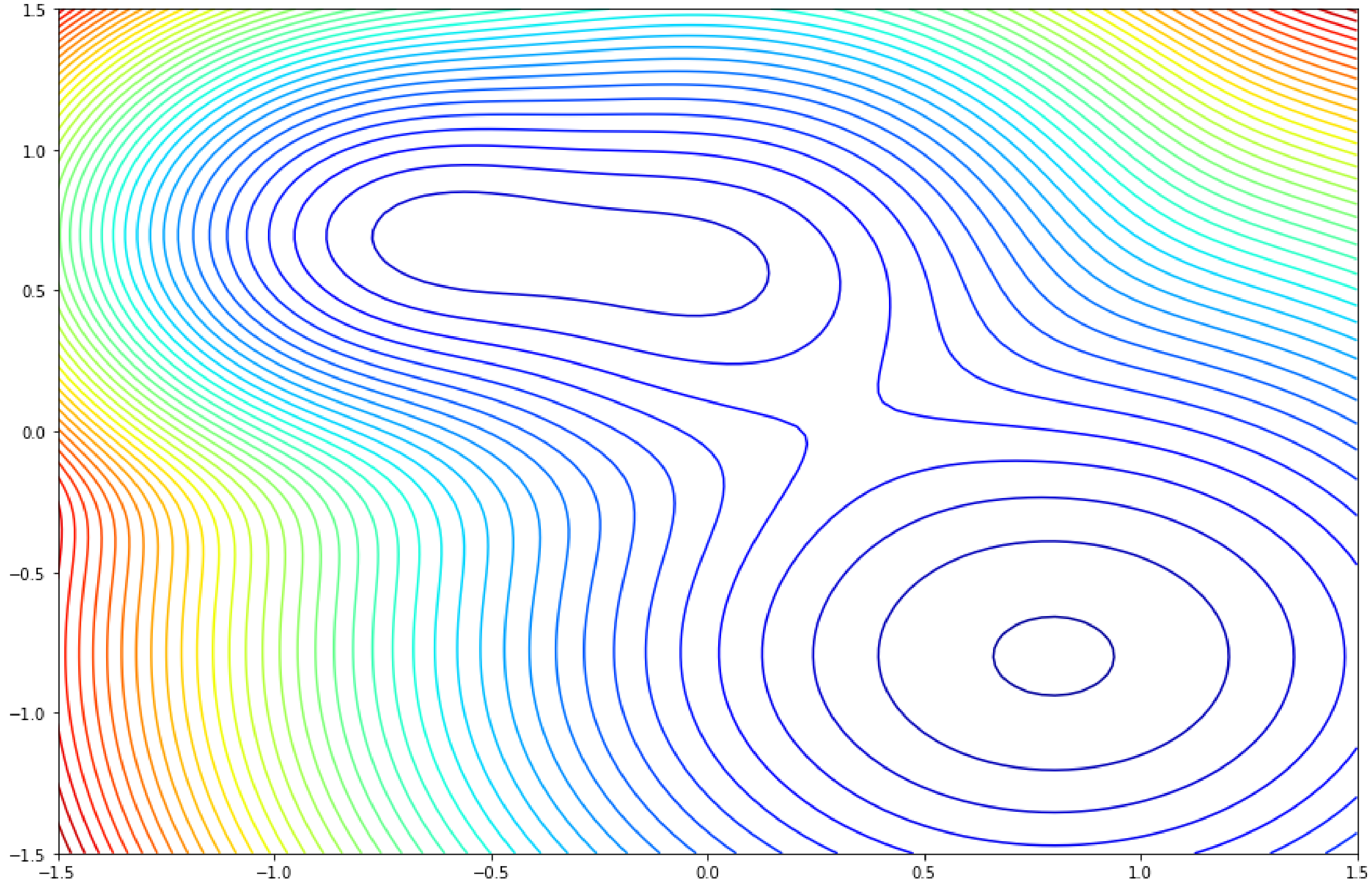
$$\frac{\partial f(\mathbf{x}^*)}{\partial x_i} \geq 0, \quad \text{and} \quad \frac{\partial f(\mathbf{x}^*)}{\partial x_i} \leq 0$$

- Equivalently we have the necessary condition

$$\boxed{\nabla f(\mathbf{x}^*) = 0} \qquad (\mathbf{x}^* \text{ is said a stationary point})$$

Necessary conditions for optimality

$$\nabla f(\mathbf{x}^*) = 0 \qquad (\mathbf{x}^* \text{ is said a stationary point})$$



Necessary conditions for optimality

- Of course, also the second order cost variation due to a small variation $\Delta\mathbf{x}$ must be non-negative

$$\nabla f(\mathbf{x}^*)'\Delta\mathbf{x} + \frac{1}{2}\Delta\mathbf{x}'\nabla^2 f(\mathbf{x}^*)\Delta\mathbf{x} \geq 0$$

- Since $\nabla f(\mathbf{x}^*)'\Delta\mathbf{x}=0$, we obtain $\Delta\mathbf{x}'\nabla^2 f(\mathbf{x}^*)\Delta\mathbf{x} \geq 0$. Hence

$$\boxed{\nabla^2 f(\mathbf{x}^*) \text{ has to be positive semidefinite}}$$

Necessary conditions for optimality

Theorem: NOC

Let $\mathbf{x}^*$ be an unconstrained local minimum of $f: \mathbb{R}^n \mapsto \mathbb{R}$ and assume that f is C^1 in an open set S containing $\mathbf{x}^*$. Then

$$\nabla f(\mathbf{x}^*) = 0 \qquad \text{(first order NOC)}$$

If in addition $f \in C^2$ within S,

$$\nabla^2 f(\mathbf{x}^*) \text{ positive semidefinite} \qquad \text{(second order NOC)}$$

Sufficient conditions for optimality

- Assume that $\mathbf{x}^*$satisfies the first order NOC

$$\nabla f(\mathbf{x}^*) = 0$$

- and also assume that the second order NOC is strengthened to

$$\nabla^2 f(\mathbf{x}^*) \quad \text{positive } \textit{definite}$$

- Then, for all $\Delta\mathbf{x} \neq 0$, $\Delta\mathbf{x}'\nabla^2 f(\mathbf{x}^*)\Delta\mathbf{x} > 0$. Hence, f tends to increase *strictly* with small excursions from $\mathbf{x}^*$, suggesting SOC…

Sufficient conditions for optimality

Theorem: SOC

Let $f: \mathbb{R}^n \mapsto \mathbb{R}$ be C^2 in an open set S. Suppose that a vector $\mathbf{x}^* \in S$ satisfies the conditions

$$\nabla f(\mathbf{x}^*) = 0 \qquad \text{and} \qquad \nabla^2 f(\mathbf{x}^*) \text{ positive definite}$$

Then $\mathbf{x}^*$ is a strict unconstrained local minimum of f

Special case: convex optimization

A subset C of $\mathbb{R}^n$ is called convex if

$$\alpha \mathbf{x} + (1-\alpha)\mathbf{y} \in C, \quad \forall \mathbf{x}, \mathbf{y} \in C, \forall \alpha \in [0,\ 1]$$

Let C be convex. A function $f: C \rightarrow \mathbb{R}$ is called convex if

$$f(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y})$$

Special case: convex optimization

Let $f: C \to \mathbb{R}$ be a convex function over a convex set C

- A local minimum of f over C is also a global minimum over C. If in addition f is strictly convex, then there exists at most one global minimum of f
- If f is in C^1 and convex, and the set C is open, $\nabla f(\mathbf{x}^*) = 0$ is a necessary and sufficient condition for a vector $\mathbf{x}^* \in C$ to be a global minimum over C

Discussion

- Optimality conditions are important to filter candidates for global minima
- They often provide the basis for the design and analysis of optimization algorithms
- They can be used for sensitivity analysis

Outline

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

Computational methods (unconstrained case)

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

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

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

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

From first order Taylor expansion (α 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 α small enough $f(\mathbf{x}_{\boldsymbol{\alpha}})$ is smaller than $f(\mathbf{x})$!

Gradient methods

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

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

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

By Taylor expansion

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

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

Gradient methods

Broad and important class of algorithms:
gradient methods

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

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

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

and the stepsize α is chosen to be positive

Gradient descent

Most often the stepsize is chosen so that

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

and the method is called gradient descent.
"Tuning" parameters:

- selecting the descent direction
- selecting the stepsize

Selecting the descent direction

General class

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

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

Popular choices:

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

Selecting the stepsize

- Minimization rule: α^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: $\alpha^k \to 0$ 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

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

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:

- <u>Penalty viewpoint</u>: we disregard the constraints and we add to the cost a high penalty for violating them
- <u>Feasibility direction viewpoint</u>: 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. Unconstrained optimization
2. Computational methods for unconstrained optimization
3. Optimization with equality constraints
4. Optimization with inequality constraints

Optimization with equality constraints

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

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

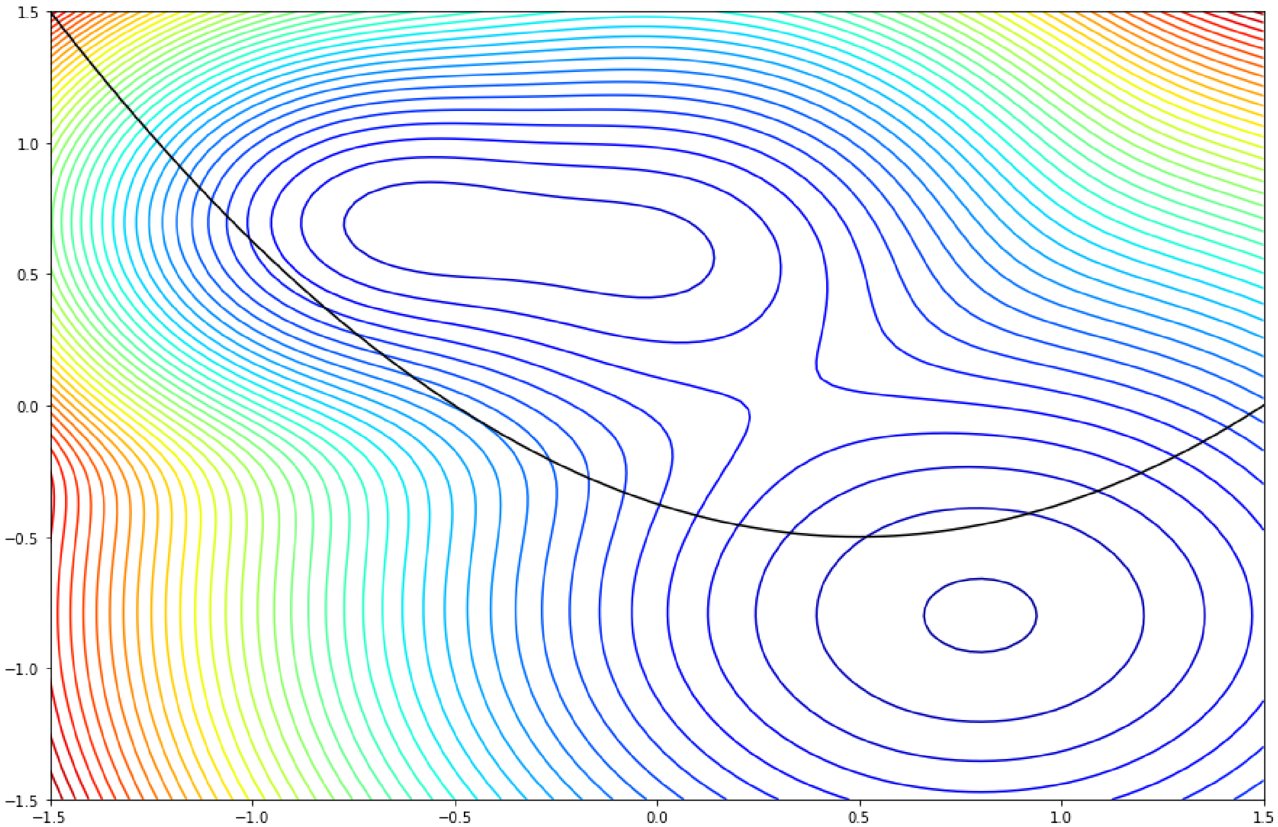
Lagrange multipliers

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

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

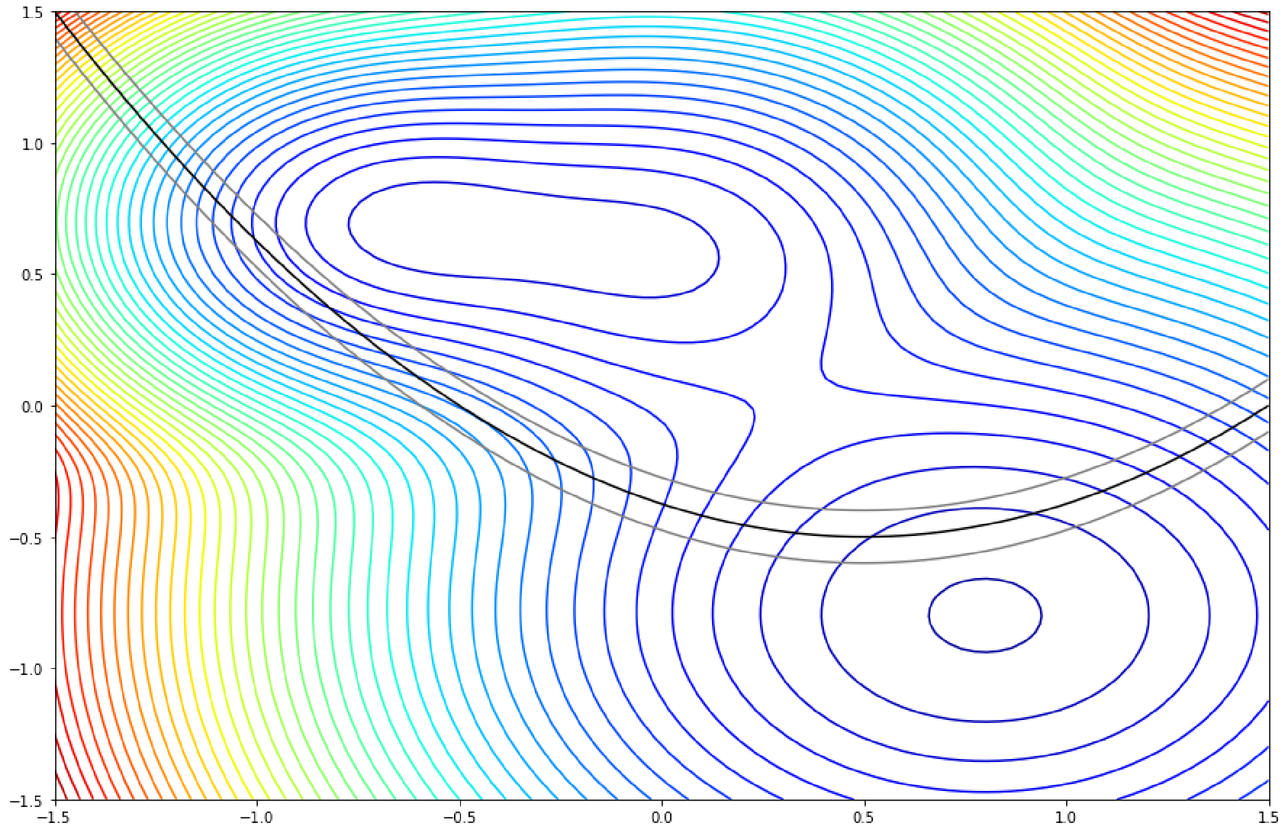
Lagrange multipliers

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



Lagrange multipliers

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



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

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

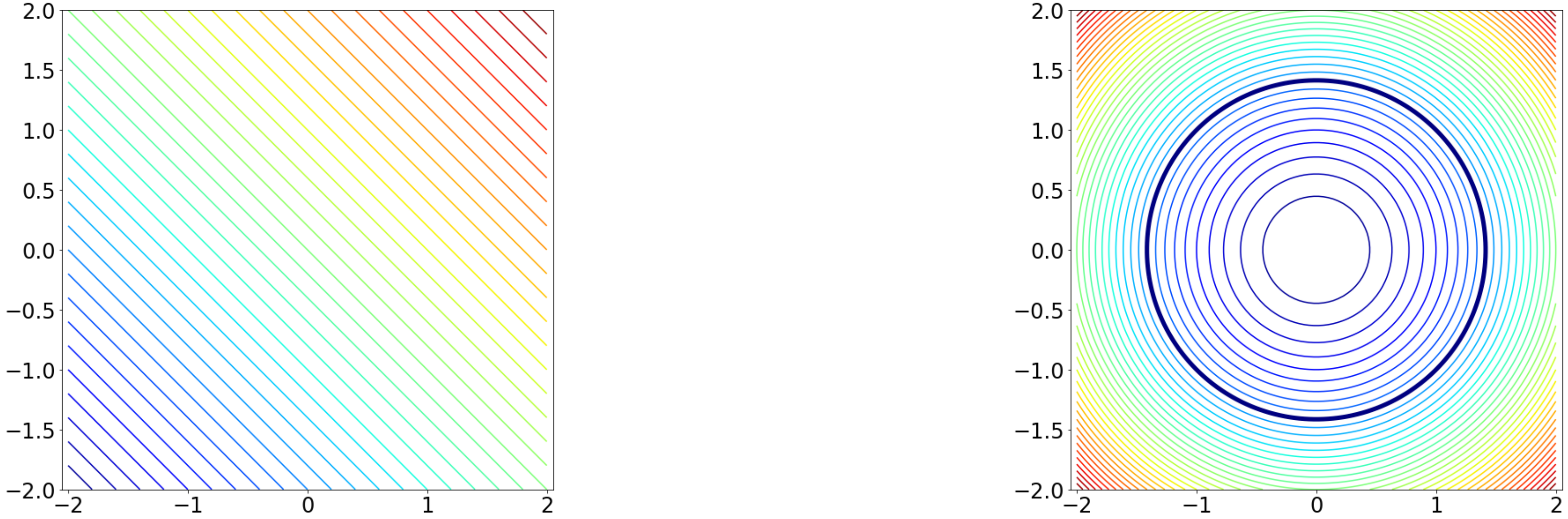
- Example

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

Lagrange multipliers

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

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



Lagrange multipliers

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

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

- Example

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

Solution: $\mathbf{x}^*$= (-1, -1)

Lagrange multipliers

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

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

Lagrange multipliers

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

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

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

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 $\nabla f(\mathbf{x}^*)' \Delta \boldsymbol{x}$ 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^{nd} 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 → 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 Lagrange multipliers 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 $L: \mathbb{R}^{n+m} \to \mathbb{R}$

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

- 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. Unconstrained optimization
2. Computational methods for unconstrained optimization
3. Optimization with equality constraints
4. Optimization with inequality constraints

Optimization with inequality constraints

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

- f, h_i, g_j are C^1
- Inequality Constrained Problem (ICP) in compact form

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

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 $\mathbf{x}$.

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

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

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

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

- or equivalently

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

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

Karush-Kuhn-Tucker NOC

Define the Lagrangian function

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

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 $(\lambda_1^*, \dots, \lambda_m^*)$, $(\mu_1^*, \dots, \mu_r^*)$ such that

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

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

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

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!)