

AA 203

Optimal and Learning-Based Control

Nonlinear optimization theory

Autonomous Systems Laboratory

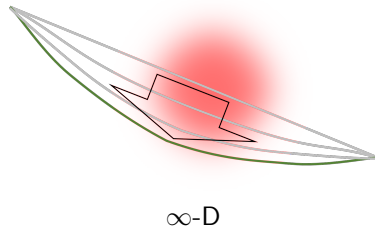
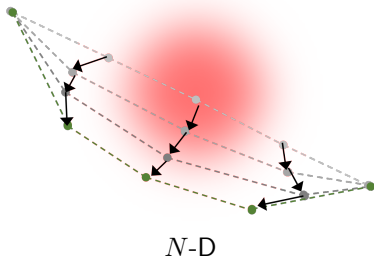
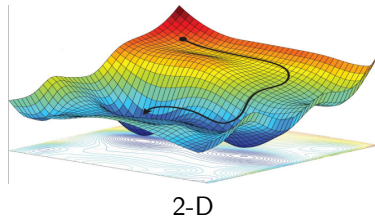
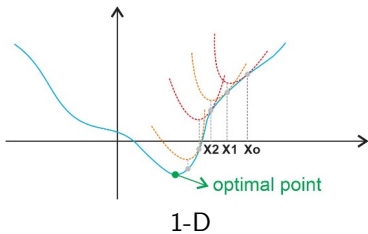
Stanford University

April 5, 2023
(last updated May 2, 2023)



Stanford
University

Optimization in many dimensions



Agenda

1. Unconstrained optimization
2. Descent methods for unconstrained problems
3. Equality-constrained optimization
4. Inequality-constrained optimization

1. Unconstrained optimization
2. Descent methods for unconstrained problems
3. Equality-constrained optimization
4. Inequality-constrained optimization

Unconstrained optimization

Given an objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we denote an *unconstrained nonlinear program* with the notation

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} f(x).$$

We usually assume either $f \in \mathcal{C}^1$ (i.e., “continuously differentiable”) or $f \in \mathcal{C}^2$ (i.e., “twice continuously differentiable”).

A solution candidate $x^* \in \mathbb{R}^n$ can be a:

local minimum $\exists \varepsilon > 0 : f(x^*) \leq f(x), \forall x : \|x - x^*\| \leq \varepsilon$

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

If the inequality is strict, i.e., “ $<$ ”, then x^* is a strict unconstrained local/global minimum. Any (strict) global minimum is also a (strict) local minimum.

There can be many minima, or none at all!

First-order necessary optimality condition

Let x^* be a local minimum.

Suppose $f \in \mathcal{C}^1$. Then near x^* we have must have

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^\top \Delta x \geq 0$$

For each i , take $\Delta x = \delta e^{(i)}$ and $\Delta x_i = -\delta e^{(i)}$ for small $\delta > 0$, where

$$e^{(i)} := (\underbrace{0, \dots, 0}_{i-1}, 1, 0, \dots, 0) \in \{0, 1\}^n.$$

Then we get

$$\frac{\partial f}{\partial x_i}(x^*)\delta \geq 0, \quad -\frac{\partial f}{\partial x_i}(x^*)\delta \geq 0 \iff \frac{\partial f}{\partial x_i}(x^*) = 0.$$

Overall, we have $\nabla f(x^*) = 0$, i.e., x^* must be a *stationary point*.

Second-order necessary optimality condition

Let x^* be a local minimum.

Suppose $f \in \mathcal{C}^2$. Then near x^* we have must have

$$f(x^* + \Delta x) - f(x^*) \approx \nabla f(x^*)^\top \Delta x + \frac{1}{2} \Delta x^\top \nabla^2 f(x^*) \Delta x \geq 0$$

We know $\nabla f(x^*) = 0$, so we must have

$$\frac{1}{2} \Delta x^\top \nabla^2 f(x^*) \Delta x \geq 0.$$

Since we can choose Δx arbitrarily within an ε -sized ball around x^* , we must have $\nabla^2 f(x^*) \succeq 0$, i.e., the Hessian of f at x^* is a *positive semi-definite* matrix.

Theorem (NOCs for unconstrained problems)

Suppose $x^ \in \mathbb{R}^n$ is an unconstrained local minimum of $f : \mathbb{R}^n \rightarrow \mathbb{R}$.*

- If $f \in \mathcal{C}^1$ on an open set $\mathcal{X} \subseteq \mathbb{R}^n$ containing x^* , then $\nabla f(x^*) = 0$.*
- If $f \in \mathcal{C}^2$ on \mathcal{X} , then $\nabla^2 f(x^*) \succeq 0$.*

Sufficient optimality conditions (SOCs) for unconstrained problems

If $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$, then

$$f(x^* + \Delta x) - f(x^*) \approx \frac{1}{2} \Delta x^\top \nabla^2 f(x^*) \Delta x > 0$$

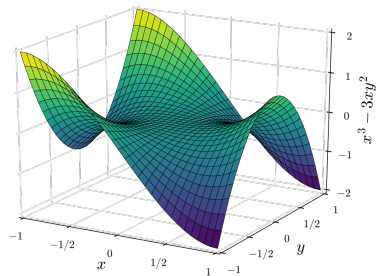
for small Δx .

Theorem (SOCs for unconstrained problems)

Suppose $f \in \mathcal{C}^2(\mathcal{X}, \mathbb{R})$ on some open set $\mathcal{X} \subseteq \mathbb{R}^n$. If $x^* \in \mathcal{X}$ satisfies

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) \succ 0,$$

then x^* is an unconstrained strict local minimum of f .



We cannot just use $\nabla^2 f(x^*) \succeq 0$ due to saddle points.

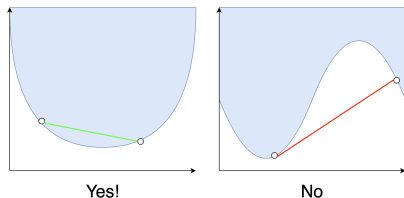
Convex sets and convex functions

A set $\mathcal{X} \subseteq \mathbb{R}^n$ is *convex* if

$$\alpha x + (1 - \alpha)y \in \mathcal{X}, \quad \forall x, y \in \mathcal{X}, \quad \forall \alpha \in [0, 1].$$

A function $f : \mathcal{X} \rightarrow \mathbb{R}^n$ is *convex* on \mathcal{X} if

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y), \\ \forall x, y \in \mathcal{X}, \quad \forall \alpha \in [0, 1].$$



If the inequality is strict, then f is *strictly convex*.

A function $f \in \mathcal{C}^2$ is convex on \mathcal{X} if and only if $\nabla^2 f(x) \succeq 0$ for all $x \in \mathcal{X}$. If $\nabla^2 f(x) \succ 0$ for all $x \in \mathcal{X}$, then f is strictly convex.

Important examples of convex functions for this course are:

Quadratic $f(x) = x^T Q x$ (where $Q \succeq 0$)

Affine $f(x) = Ax + b$ (both convex and concave)

Theorem (NOCs are SOC for unconstrained convex problems)

Let $f : \mathcal{X} \rightarrow \mathbb{R}$ be a convex function over a convex set $\mathcal{X} \in \mathbb{R}^n$.

- If $x^* \in \mathcal{X}$ is local minimum of f , then it is also a global minimum over \mathcal{X} .
- If f is strictly convex, then there exists at most one global minimum of f over \mathcal{X} .
- Suppose additionally that \mathcal{X} is open and $f \in \mathcal{C}^1(\mathcal{X}, \mathbb{R})$. Then $\nabla f(x^*) = 0$ if and only if x^* is a global minimum of f over \mathcal{X} .

Agenda

1. Unconstrained optimization
2. Descent methods for unconstrained problems
3. Equality-constrained optimization
4. Inequality-constrained optimization

Descent methods for unconstrained problems

Iterative descent methods start at an initial guess $x^{(0)}$, and try to successively generate vectors $\{x^{(1)}, x^{(2)}, \dots\}$ such that the objective decreases at each iteration, i.e.,

$$f(x^{(k+1)}) \leq f(x^{(k)}), \quad \forall k \in \{0, 1, 2, \dots\}.$$

The hope is that we can decrease f all the way to a minimum.

Consider the update rule

$$x^{(k+1)} = x^{(k)} + \alpha^{(k)} d^{(k)},$$

where $\alpha^{(k)} > 0$ is the *step-size* and $d^{(k)} \in \mathbb{R}^n$ is the *descent direction*. Then

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \alpha^{(k)} \nabla f(x^{(k)})^\top d^{(k)}.$$

The goal is to choose $\alpha^{(k)} > 0$ and $d^{(k)} \in \mathbb{R}^n$ such that this approximation is appropriate and $\nabla f(x^{(k)})^\top d^{(k)} < 0$.

Gradient descent directions

Let $d^{(k)} = -D^{(k)} \nabla f(x^{(k)})$, where $D^{(k)} \succ 0$. Then

$$\begin{aligned} f(x^{(k+1)}) &\approx f(x^{(k)}) + \alpha^{(k)} \nabla f(x^{(k)})^\top d^{(k)} \\ &= f(x^{(k)}) - \alpha^{(k)} \nabla f(x^{(k)})^\top D^{(k)} \nabla f(x^{(k)}) \end{aligned}$$

Since $D^{(k)} \succ 0$, we have that $f(x^{(k+1)}) \leq f(x^{(k)})$ for small enough $\alpha^{(k)} > 0$.

Popular choices for the descent scaling $D^{(k)}$ are

steepest $D^{(k)} = I$.

Newton $D^{(k)} = \nabla^2 f(x^{(k)})^{-1}$, provided that the inverse exists.

The Newton descent direction analytically minimizes the quadratic approximation

$$f(x^{(k+1)}) \approx f(x^{(k)}) + \nabla f(x^{(k)})^\top d^{(k)} + \frac{1}{2} d^{(k)\top} \nabla^2 f(x^{(k)}) d^{(k)}$$

at each iteration k , assuming $\nabla^2 f(x^{(k)})$ is invertible.

Selecting the step-size

Constant Choose $\alpha^{(k)} \equiv \alpha > 0$. Convergence can be slow, or the iterates could diverge if α is too large.

Diminishing Ensure $\alpha^{(k)} \rightarrow 0$ and $\sum_{k=0}^{\infty} \alpha^{(k)} = \infty$. This does not guarantee descent at each iteration, but it can avoid diverging iterates.

Line search Given the current iterate $x^{(k)}$ and a descent direction $d^{(k)}$, compute

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

exactly if possible. Otherwise, do *backtracking line search*

initialize $\alpha^{(k)} = 1$

while $f(x^{(k)} + \alpha d^{(k)}) > f(x^{(k)}) + \gamma \alpha^{(k)} \nabla f(x^{(k)})^\top d^{(k)}$
 $\alpha^{(k)} \leftarrow \beta \alpha^{(k)}$

where $\gamma \in (0, 0.5)$ and $\beta \in (0, 1)$ are hyperparameters.

There is a wealth of mathematical analyses of descent methods involving:

- guarantees for convergence to a stationary point
- good convergence criteria (e.g., $\|x^{(k)} - x^{(k-1)}\| < \varepsilon$, $|f(x^{(k)}) - f(x^{(k-1)})| < \varepsilon$, $\|\nabla f(x^{(k)})\| < \varepsilon$)
- convergence rates (e.g., $f(x^{(k)}) - f(x^*) \lesssim \frac{1}{k} \|x^{(0)} - x^*\|_2^2$)

There are other descent methods that can be implemented “derivative-free”, such as

- coordinate descent
- Nelder-Mead algorithms

Agenda

1. Unconstrained optimization
2. Descent methods for unconstrained problems
3. Equality-constrained optimization
4. Inequality-constrained optimization

Given an objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and a *constraint function* $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$, we denote an *equality-constrained nonlinear program* with the notation

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & h(x) = 0 \end{array}$$

We assume $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ and $h \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^m)$.

Lagrange multipliers for equality-constrained problems

Define the *Lagrangian* function

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

where $\lambda \in \mathbb{R}^m$ is a vector of *Lagrange multipliers*.

Theorem (First-order NOC for equality-constrained problems)

Suppose $x^ \in \mathbb{R}^n$ is a local minimum of $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ subject to $h(x^*) = 0$ with $h \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^m)$. Moreover, assume $\{\nabla h_i(x^*)\}_{i=1}^m$ are linearly independent. Then there exists a unique $\lambda^* \in \mathbb{R}^m$ such that*

$$\nabla_x L(x^*, \lambda^*) = \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) = 0.$$

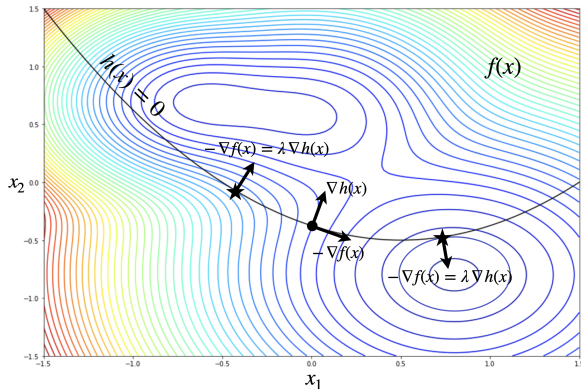
Second-order NOCs and SOCs for constrained problems are discussed in [AA203-Notes](#) and ([Bertsekas, 2016](#)).

First-order NOC visualized

Re-arrange $\nabla_x L(x^*, \lambda^*) = 0$ to get

$$-\nabla f(x^*) = \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*).$$

Further reduction of the objective value would produce a change in the constraint function, thereby violating $h(x) = 0$.



The first-order NOC required that x^* is a *regular* point, i.e., that $\{\nabla h_i(x^*)\}_{i=1}^m$ are linearly independent vectors. Since $\nabla h_i(x^*) \in \mathbb{R}^n$, this implicitly requires $m \leq n$ (i.e., you cannot find more than n linearly independent vectors in \mathbb{R}^n).

Solving $\min_{x: h(x)=0} f(x)$ can be viewed as solving for n variables subject to m constraints.

The proof of the first-order NOC relies on eliminating m variables to arrive at an unconstrained problem in $n - m$ variables, which in turn relies on $\{\nabla h_i(x^*)\}_{i=1}^m$ being linearly independent to apply the implicit function theorem.

See ([Bertsekas, 2016](#), §4.1.2) for further details.

1. Unconstrained optimization
2. Descent methods for unconstrained problems
3. Equality-constrained optimization
4. Inequality-constrained optimization

Inequality-constrained optimization

Given an objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and *constraint functions* $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^r$, we denote an *inequality-constrained nonlinear program* with the notation

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && h(x) = 0 \\ & && g(x) \preceq 0 \end{aligned}$$

We assume $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$, $h \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^m)$, and $g \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^r)$. We use “ \preceq ” to denote element-wise inequality in this scenario.

For any feasible point x , i.e., such that $h(x) = 0$ and $g(x) \preceq 0$, define the set of *active inequality constraints* by

$$\mathcal{A}_g(x) := \{j \in \{1, 2, \dots, r\} \mid g_j(x) = 0\}.$$

Karush-Kuhn-Tucker (KKT) NOC conditions

With Lagrangian multipliers $\lambda \in \mathbb{R}^m$ and $\mu \in \mathbb{R}^r$, define the Lagrangian

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

Theorem (First-order NOC for inequality-constrained problems)

Suppose $x^ \in \mathbb{R}^n$ is a local minimum of $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ subject to $h(x^*) = 0$ and $g(x^*) \preceq 0$ with $h \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^m)$ and $g \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^r)$. Moreover, assume*

$$\{\nabla h_i(x^*)\}_{i=1}^m \cup \{\nabla g_j(x^*)\}_{j \in \mathcal{A}_g(x^*)}$$

are linearly independent. Then there exist unique $\lambda^ \in \mathbb{R}^m$ and $\mu^* \in \mathbb{R}^r$ such that*

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0, \quad \mu^* \succeq 0, \quad \mu_j^* = 0, \quad \forall j \notin \mathcal{A}_g(x^*).$$

We can also write the last condition succinctly as $\mu^{*\top} g(x^*) = 0$.

KKT conditions for convex problems

Consider when f is convex, each $g_j(x)$ is convex, and $h(x)$ is affine, i.e., $h(x) = Ax - b$. Then we have

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && Ax = b \\ & && g(x) \preceq 0 \end{aligned}$$

for which the feasible set $\mathcal{X} := \{x \in \mathbb{R}^n \mid Ax = b, g(x) \preceq 0\}$ is convex.

Theorem (KKT conditions are NOCs and SOCs for convex problems)

Suppose $f \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R})$ and $g \in \mathcal{C}^1(\mathbb{R}^n, \mathbb{R}^r)$ are convex, and that there exists at least one strictly feasible point $x \in \mathcal{X}$, i.e., $Ax = b$ and $g(x) \prec 0$. Then (x^, λ^*, μ^*) describe a global minimum if and only if*

$$Ax^* = b, \quad g(x^*) \preceq 0, \quad \nabla_x L(x^*, \lambda^*, \mu^*) = 0, \quad \mu^* \succeq 0, \quad \mu^{*\top} g(x^*) = 0.$$

Example: Maximal rectangle inside a circle

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

We have $f(x) = -x_1 - x_2$ (for minimization) with $h(x) = x_1^2 + x_2^2 - r^2$, so

$$L(x, \lambda) = -x_1 - x_2 + \lambda(x_1^2 + x_2^2 - r^2).$$

The first-order NOC at a local minimum (x^*, λ^*) is

$$\nabla_x L(x^*, \lambda^*) = \begin{pmatrix} -1 + 2\lambda^* x_1^* \\ -1 + 2\lambda^* x_2^* \end{pmatrix} \stackrel{!}{=} 0 \iff x_1^* = x_2^* = \frac{1}{2\lambda^*}.$$

Substitute into $x_1^{*2} + x_2^{*2} = r^2$ to get $\lambda^* = \pm \frac{1}{\sqrt{2}r} \implies x_1^* = x_2^* = \pm \frac{1}{\sqrt{2}}r$.

Of the two possible solutions, $x_1^* = x_2^* = \frac{1}{\sqrt{2}}r$ is the global maximum (i.e., a square).

Why should we care about characterizing optimality conditions?

- Even just NOCs can form a filter for distilling local minima from feasible points.
- NOCs and SOC's can serve as a means for “measuring progress” towards optimality during an optimization procedure, particularly for convex problems.
- Problem structure (e.g., quadratic objective with linear constraints) coupled with convexity and the KKT conditions can be leveraged to implement efficient solvers with good convergence properties ([Boyd and Vandenberghe, 2004](#)).
- Even for non-convex problems, convex solvers can be used in iterative convex sub-problems that can converge to a local minimum.

Preview: Sequential Convex Programming (SCP)

Consider the non-convex problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x) \\ & \text{subject to} && h(x) = 0, \quad g(x) \preceq 0 \end{aligned}$$

The basic idea of *sequential convex programming (SCP)* is to maintain an estimate $x^{(k)}$ and iteratively solve for $x^{(k+1)}$ via the convex sub-problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && \hat{f}^{(k)}(x) \\ & \text{subject to} && \hat{h}^{(k)}(x) := \hat{A}^{(k)}x - \hat{b}^{(k)} = 0, \quad \hat{g}^{(k)}(x) \preceq 0, \quad x \in \mathcal{T}^{(k)} \end{aligned}$$

where $(\hat{f}^{(k)}, \hat{g}^{(k)})$ and $\hat{h}^{(k)}$ are convex and affine, respectively, *approximations* of (f, g) and h , respectively, over a convex *trust region* constructed around $x^{(k)}$, e.g.,

$$\mathcal{T}^{(k)} := \{x \mid \|x - x^{(k)}\|_{\infty} \leq \rho\},$$

for some $\rho > 0$.

Pontryagin's maximum principle and indirect methods for optimal control
(i.e., applying NOCs to optimal control problems)

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 3 edition, 2016.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.