Optimization for Learning

Anders Hansson and Martin Andersen

Linköping University and Technical University of Denmark

March 26, 2024

Contents of Lectures

1. Optimization Theory
2. Optimization Problems
3. Optimization Methods I
4. Optimization Methods II
5. Optimization Methods III
6. Probabilites
7. Unsupervised Learning I
8. Unsupervised Learning II
9. Unsupervised Learning III
10. Supervised Learning I
11. Supervised Learning II
12. System Identification

Contents

Optimization Problem

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i \in \mathbf{N}_m \\ & h_i(x) = 0, & i \in \mathbf{N}_p \end{array} \tag{1}$$

$f_0 : \mathbf{R}^n \to \mathbf{R}$ *objective function*

$f_i : \mathbf{R}^n \to \mathbf{R}, i \in \mathbf{N}_m$ *inequality constraint functions*

$h_i : \mathbf{R}^n \to \mathbf{R},\ i \in \mathbf{N}_p$ *equality constraint functions*

$\mathcal{D} = \mathbf{dom}\, f_0 \cap_{i=1}^m \mathbf{dom}\, f_i \cap_{i=1}^p \mathbf{dom}\, h_i$ *domain*

$$f(x) = (f_1(x), \ldots, f_m(x)); \qquad h(x) = (h_1(x), \ldots, h_p(x))$$

Assumption: $p < n$

Feasibility

A point $x \in \mathcal{D}$ is *feasible* if $f(x) \leq 0$ and $h(x) = 0$.

A feasible point is *strictly feasible* if $f(x) < 0$.

The set X of all feasible points is called the *feasible set*.

If $X = \emptyset$, then the optimization problem is *infeasible*. Otherwise it is *feasible*.

Optimality

The *optimal value* of (1) is

$$p^\star = \inf \{f_0(x) | f(x) \leq 0,\ h(x) = 0\}$$

If (1) is infeasible, then $p^\star = \infty$.

If $\exists x_k \in \mathcal{D}$: $\lim_{k\to\infty} f_0(x_k) = -\infty$, then the optimization problem is *unbounded from below*, and $p^\star = -\infty$.

$x^\star$ is an *optimal point* or *solution* if $x^\star$ is feasible and $f_0(x^\star) = p^\star$.

The set of all optimal points is called the *optimal set*.
If the optimal set is nonempty we say that the optimal value is *attained* or *achieved*. Otherwise it is not attained or achieved.

A feasible point x for which $f_0(x) \leq p^\star + \epsilon$, where $\epsilon > 0$, is called ϵ- *suboptimal*.

Local Optimality

A feasible point x is called *locally optimal* if $\exists R > 0$:

$$f_0(x) = \inf \{f_0(z) |\ f(z) \leq 0,\ h(z) = 0,\ \|z - x\|_2 \leq R\}$$

A locally optimal point is not necessarily optimal. However, the converse is always true.

To distinguish a local optimal point from an optimal point we sometimes say that an optimal point is *globally optimal.*

Feasibilty Problem

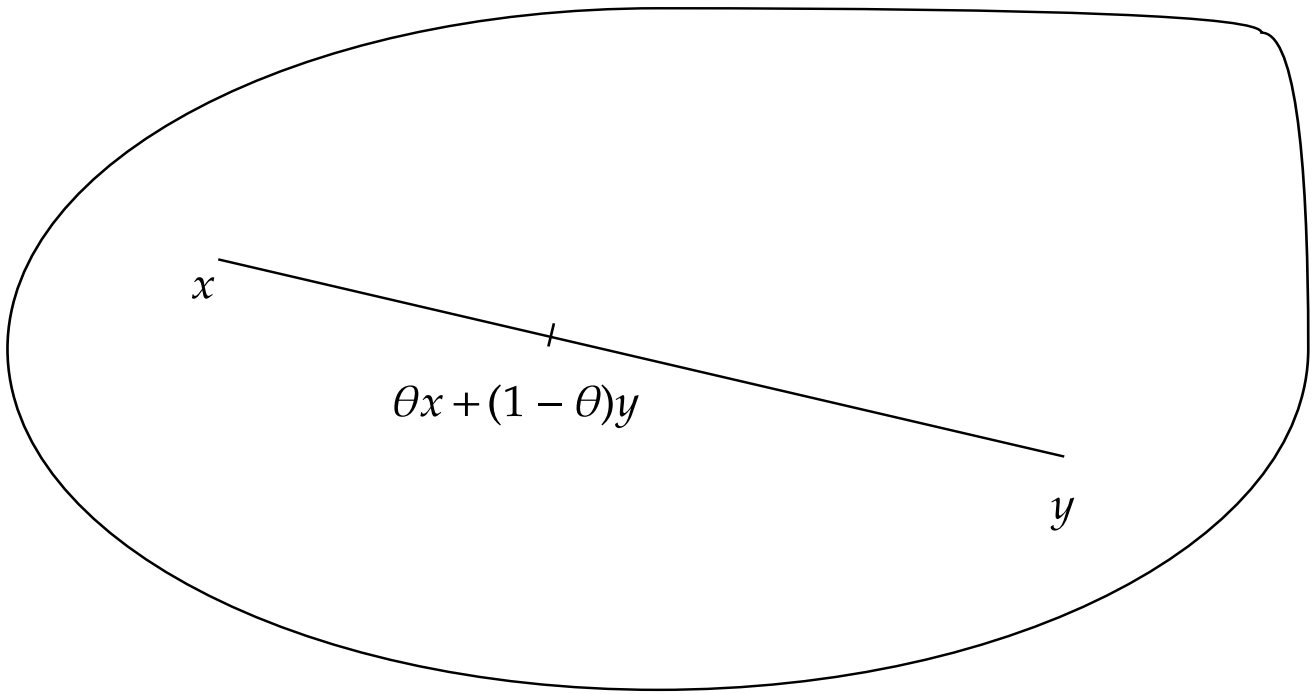
When $f_0(x) = 0 \ \forall x \in \mathcal{D}$, then $p^\star = 0$ if $X \neq \emptyset$, and $p^\star = \infty$ if $X = \emptyset$.

Such an optimization problem is called a *feasibility problem* and solving it amounts to finding an $x \in \mathcal{D}$ that satisfies the constraints.

Equivalent Problems

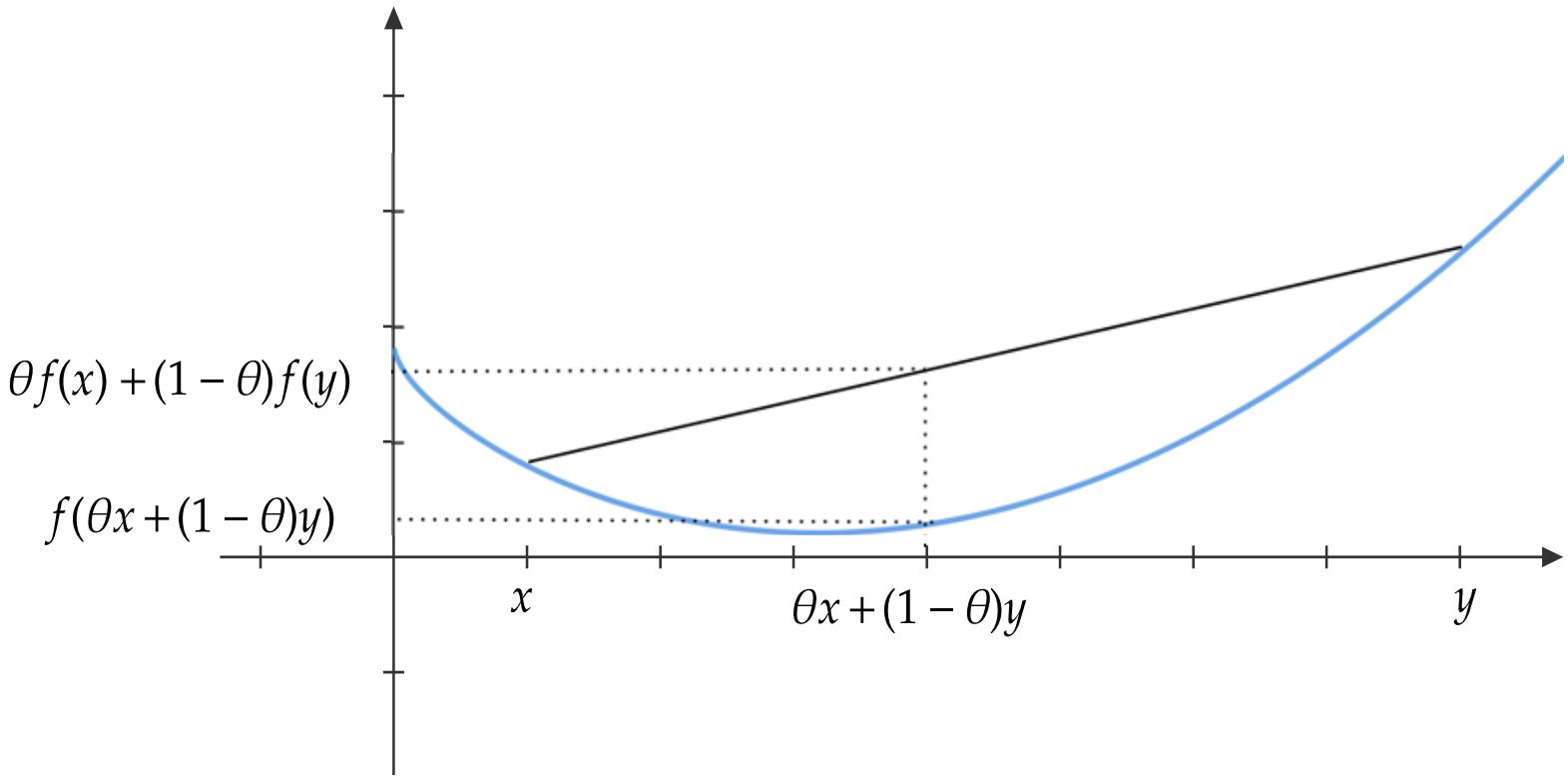
Two optimization problems are said to be *equivalent* if the solution from one of them can be found from the other one with simple manipulations and vice versa.

Convex Set



A set C is called *convex* if $\forall x, y \in C$ and $\forall \theta \in [0, 1]$ it holds that $\theta x + (1 - \theta)y \in C$.

Convex Function



We say that a function $f : \mathbf{R} \to \mathbf{R}$ is *convex* if its domain is convex and if $\forall \theta \in [0, 1]$ it holds that

$$f\left(\theta x + (1-\theta) y\right) \leq \theta f(x) + (1-\theta) f(y)$$

Convex Function ctd.

The function is called *strictly convex* if the inequality holds strictly whenever $x \neq y$ and $\theta \in (0, 1)$.

We say that a function f is *concave* of $-f$ is convex and similarly for strict concavity.

A vector valued function is convex if all its component functions are convex.

The Epigraph and the Perspective Function

The *epigraph* of a function f is

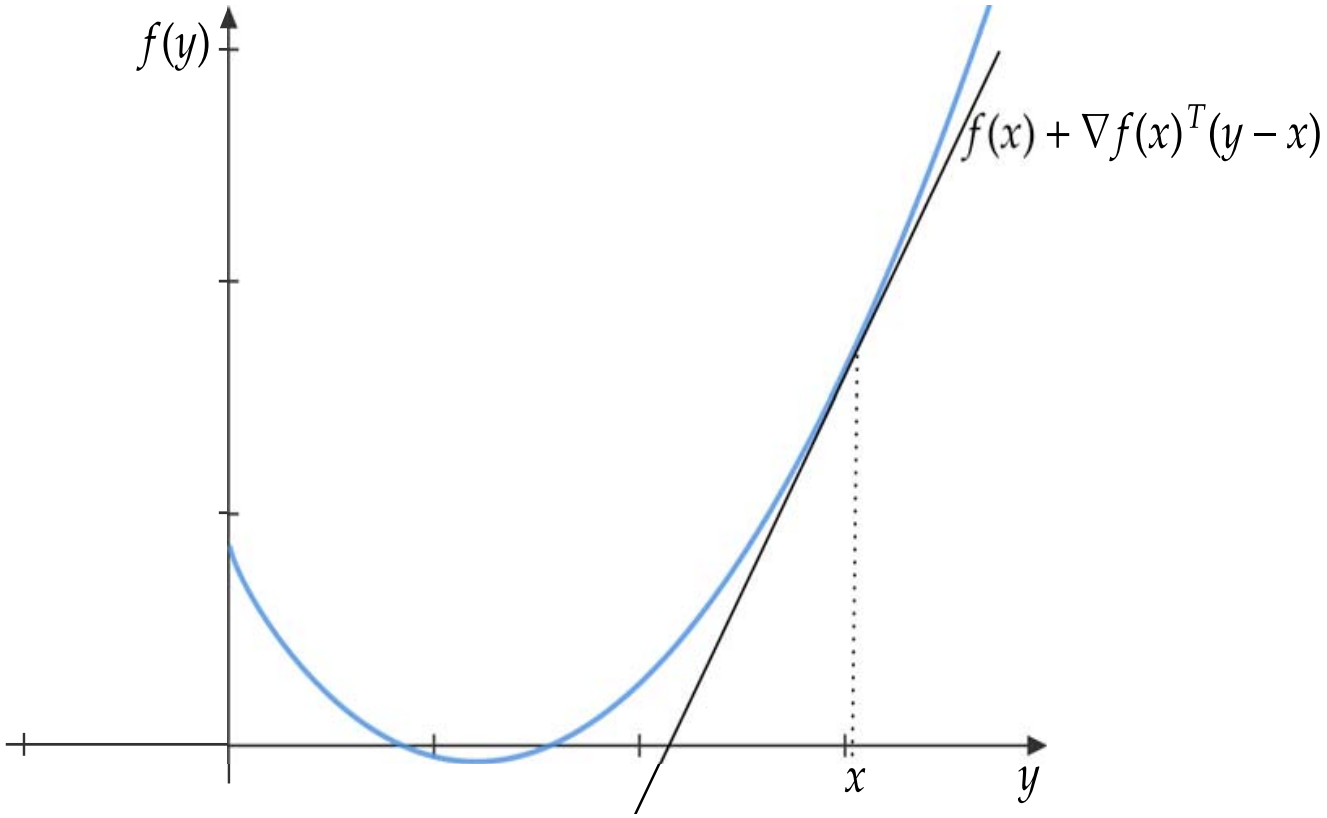
$$\mathbf{epi}\, f = \{(x,t)|\ f(x) \leq t\}$$

A function f is convex if and only if $\mathbf{epi}\, f$ is a convex set.

The *perspective* of a function $f : \mathbf{R}^n \to \mathbf{R}$ is $g : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}$ for which $g(x,y) = yf(x/y)$ with domain $\{(x,y)|\ y > 0, x/y \in \mathbf{dom}\, f\ \}$.

If f is convex so is its perspective.

Conditions for Convexity



Assume f differentiable. Then f is convex if and only if **dom** f convex and

$$f(y) \geq f(x) + \nabla f(x)^T(y - x), \quad \forall x, y \in \textbf{dom}\, f \tag{2}$$

Conditions for Convexity ctd.

Strict convexity holds if and only if the inequality holds strictly whenever $x \neq y$.

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

$$\nabla^2 f(x) \succeq 0$$

$\forall x \in$ **dom** f.

If $\nabla^2 f(x) \succ 0$ $\forall x \in$ **dom** f then f is strictly convex. The converse is however not true.

Convex Optimization Problem

The optimization problem in (1), i.e.

$$\begin{array}{lll} \text{minimize} & f_0(x) & \\ \text{subject to} & f_i(x) \leq 0, & i \in \mathbf{N}_m \\ & h_i(x) = 0, & i \in \mathbf{N}_p \end{array}$$

is a *convex optimization problem* if f_0, and f are convex functions, and if $h(x)$ is an *affine* function, i.e. both convex and concave.

Convex optimization problems have convex feasible sets. The converse is however not true.

Condition for Optimality

For the convex optimization problem (1) assume f_0 differentiable.

Then x is optimal if and only if $x \in X$ and

$$\nabla f_0(x)^T (y - x) \geq 0, \qquad \forall\, y \in X \tag{3}$$

where X is the feasible set.

Equality Constrained Optimization Problem

Consider

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax = b \end{array}$$

where $f_0 : \mathbf{R}^n \to \mathbf{R}$ differentiable and $A \in \mathbf{R}^{p \times n}$ $b \in \mathbf{R}^p$.

Define *Lagrangian* $L : \mathbf{R}^n \times \mathbf{R}^p \to \mathbf{R}$ by

$$L(x, \mu) = f_0(x) + \mu^T h(x)$$

where $h(x) = Ax - b$. Then x is optimal if and only if there exist μ such that x satisfies

$$\begin{aligned} \nabla_x L(x, \mu) &= 0 \\ h(x) &= 0 \end{aligned}$$

Lagrangian for General Non-Convex Problem

Consider (1), and define Lagrangian $L : \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ via

$$L(x, \lambda, \mu) = f_0(x) + \lambda^T f(x) + \mu^T h(x) \tag{4}$$

with domain $\mathcal{D} \times \mathbf{R}^m \times \mathbf{R}^p$.

Here λ and μ are called the *Lagrange multipliers* or *dual variables*.

The *Lagrange dual function* $g : \mathbf{R}^m \times \mathbf{R}^p \to \mathbf{R}$ is defined as

$$g(\lambda, \mu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \mu) \tag{5}$$

In case L is not bounded from below for some (λ, μ), it takes the value $-\infty$.

The function g is concave whether (1) is a convex optimization problem or not.

Duality

For any $\lambda \geq 0$ and any μ

$$g(\lambda, \mu) \leq p^\star$$

We say that (λ, μ) are *dual feasible* if $\lambda \geq 0$ and they belong to the domain of g.

The Conjugate Function

The *conjugate function* $f^\star : \mathbf{R}^n \to \mathbf{R}$ of a function $f : \mathbf{R}^n \to \mathbf{R}$ is defined as

$$f^\star(y) = \sup_{x \in \mathbf{dom}\, f} \left(y^T x - f(x) \right)$$

Application

Consider

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & Ax \leq b \\ & Cx = d \end{array} \tag{6}$$

where $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, $C \in \mathbf{R}^{p \times n}$, and $d \in \mathbf{R}^p$. Then

$$\begin{aligned} g(\lambda, \mu) &= \inf_{x \in \mathcal{D}} \left(f_0(x) + \lambda^T (Ax - b) + \mu^T (Cx - d) \right) \\ &= -\lambda^T b - \mu^T d + \inf_{x \in \mathcal{D}} \left(f_0(x) + \left(A^T \lambda + C^T \mu \right)^T x \right) \\ &= -\lambda^T b - \mu^T d - f_0^\star \left(-A^T \lambda - C^T \mu \right) \end{aligned}$$

where the domain of g is $\{(\lambda, \mu) | -A^T \lambda - C^T \mu \in \mathbf{dom}\, f_0^\star\}$.

Lagrange Dual Problem

The *Lagrange dual problem* associated with (1) is defined as

$$\begin{array}{ll} \text{maximize} & g(\lambda, \mu) \\ \text{subject to} & \lambda \geq 0 \end{array} \tag{7}$$

This provides a lower bound $p^\star$ of (1), called the *primal problem*.

For optimal value of the dual problem $d^\star$ it holds that $d^\star \leq p^\star$, called *weak duality*.

Optimal solution to (7) denoted $(\lambda^\star, \mu^\star)$, called *dual optimal* or *optimal Lagrange multipliers*.

Lagrange dual problem is convex optimization problem.

Stopping Criteria

For any dual feasible (λ, μ) and primal feasible x it holds that $g(\lambda, \mu) \leq p^\star$ implies that

$$f_0(x) - p^\star \leq f_0(x) - g(\lambda, \mu)$$

and hence x is ϵ-suboptimal with $\epsilon = f_0(x) - g(\lambda, \mu)$.

Similarly (λ, μ) is ϵ-suboptimal for the dual problem.

We call ϵ the *duality gap*. For any dual feasible (λ, μ) and primal feasible x it holds that

$$p^\star \in [g(\lambda, \mu), f_0(x)]\,; \qquad d^\star \in [g(\lambda, \mu), f_0(x)]$$

Strong Duality

The difference $p^\star - d^\star$ called the *optimal duality gap*.

If $d^\star = p^\star$, we then say that *strong duality* holds—not true in general.

Often need *constraint qualifications* and convexity for strong duality.

Slater's condition: There exist $x \in \mathbf{relint}\,\mathcal{D} \cap X$ such that $f(x) < 0$, i.e. *strict feasibility*.

Slater's condition also implies that the dual optimal value is attained when $d^\star > -\infty$, i.e. there exist $(\lambda^\star, \mu^\star)$ such that $g(\lambda^\star, \mu^\star) = d^\star$.

Karush-Kuhn-Tucker (KKT) conditions

Consider the primal problem in (1) and the corresponding dual problem in (7) and asssume that f_0, f and h are differentiable. Given primal and dual optimal points $x^\star$ and $(\lambda^\star, \mu^\star)$ with zero duality gap it holds that

$$\nabla L(x^\star, \lambda^\star, \mu^\star) = 0 \tag{8}$$

$$f(x^\star) \leq 0 \tag{9}$$

$$h(x^\star) = 0 \tag{10}$$

$$\lambda^\star \geq 0 \tag{11}$$

$$\lambda_i^\star f_i(x^\star) = 0, \qquad i \in \mathbf{N}_m \tag{12}$$

The above conditions are also sufficient if the primal problem is convex.

Necessary Conditions for Nonconvex Problems

For non-convex problems there rarely exist primal and dual optimal points $x^\star$ and $(\lambda^\star, \mu^\star)$ with zero duality gap.

Consider (1) where we assume that f_0, f and h are differentiable.

Let $x^\star$ be a local optimal point and let $\mathcal{A} \subset \mathbf{N}_m$ be the set of i such that $f_i(x^\star) = 0$, i.e. the set of *active inequality constraints*. If $\nabla f_i(x^\star)$, $i \in \mathcal{A}$ and $\nabla h_i(x^\star)$, $i \in \mathbf{N}_p$ are linearly independent, then there exist $\lambda \in \mathbf{R}^m$ and $\mu \in \mathbf{R}^p$ such that

$$\begin{aligned} \nabla L(x^\star, \lambda, \mu) &= 0 \\ f(x^\star) &\leq 0 \\ h(x^\star) &= 0 \\ \lambda &\geq 0 \\ \lambda_i f_i(x^\star) &= 0, \qquad i \in \mathbf{N}_m \end{aligned}$$

Sufficient Conditions for Nonconvex Problems

Let

$$M = \{y \in \mathbf{R}^n \mid \nabla h_i(x^\star)y = 0,\ \nabla f_j(x^\star)y = 0,\ \forall i \in \mathbf{N}_p,\ \forall j \in J\}$$

where

$$J = \{j \in \mathbf{N}_m : f_j(x^\star) = 0,\ \lambda_j > 0\}$$

Assume that $f_0,\ f_i,\ i \in \mathbf{N}_m$ and $h_j,\ j \in \mathbf{N}_p$ are twice continuously differentiable.

Sufficient condition for $x^\star$ to be a local optimal point is $\exists(\lambda, \mu)$:

$$\begin{aligned} \nabla L(x^\star, \lambda, \mu) &= 0 \\ f(x^\star) &\leq 0 \\ h(x^\star) &= 0 \\ \lambda &\geq 0 \\ \lambda_i f_i(x^\star) &= 0, \qquad i \in \mathbf{N}_m \end{aligned}$$

and that $y\nabla^2 L(x^\star, \lambda, \mu)y^T > 0$ for all $y \in M$.