Maximum Margin Classifiers: Support Vector Machines

Machine Learning and Pattern Recognition:
Lecture 14

Sumit Chopra

Outline of the Talk

- Quick Tutorial on Optimization
  - Basic idea behind Support Vector Machines
  - Optimization concepts and terminology
- Support Vector Machines in Detail
  - Given by Fu Jie Huang

Binary Classification Problem

- Given: Training data generated according to the distribution D

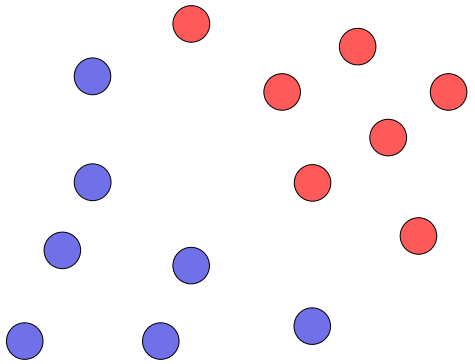
$$(x_1, y_1), \ldots, (x_p, y_p) \in \Re^n \times \{-1, 1\}$$

- Problem: Find a classifier (a function) $h(x) : \Re^n \to \{-1, 1\}$ such that it generalizes well on the test set obtained from the same distribution D

- Solution:
  - Linear Approach: linear classifiers - perceptron and many other.
  - Non Linear Approach: non-linear classifiers - neural nets and many other.

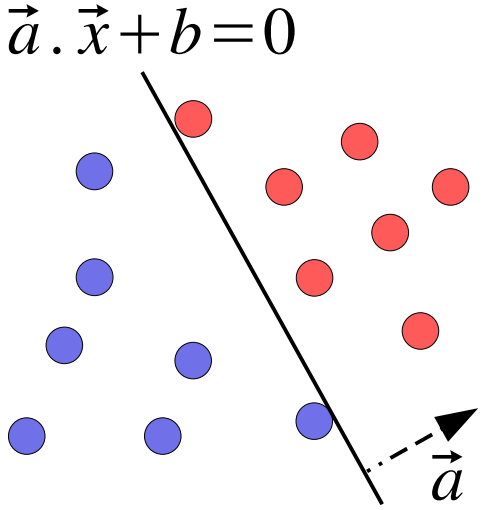
Linearly Separable Data

- Assume that the training data is linearly separable



Linearly Separable Data

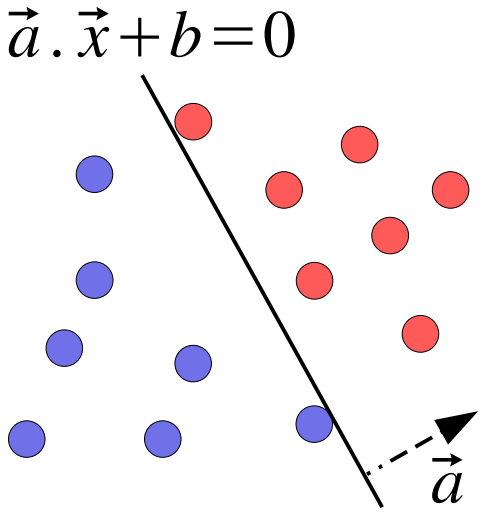
- Assume that the training data is linearly separable

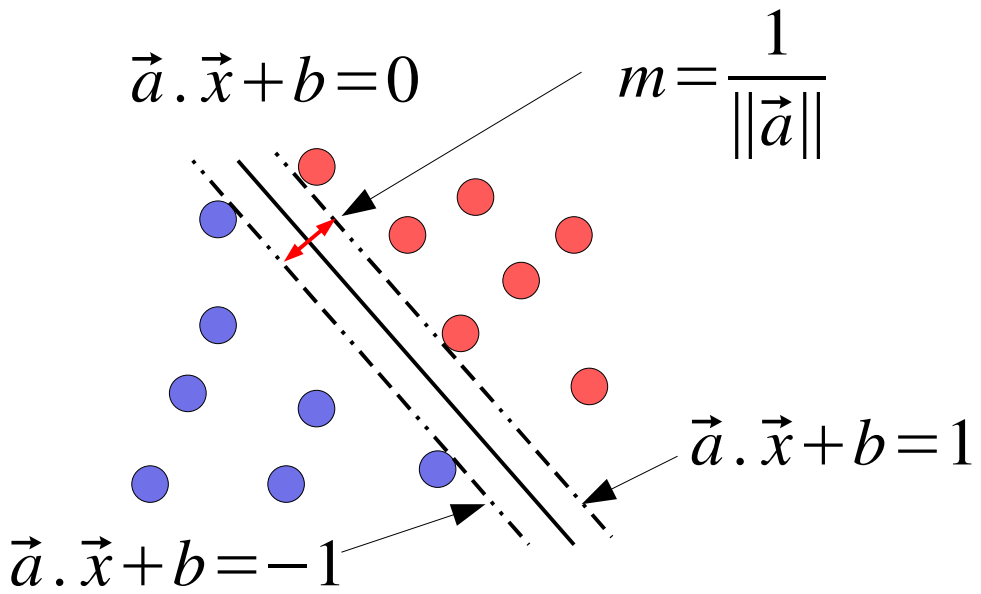


- Then the classifier is: $h(x) = \vec{a}.\vec{x}+b \qquad where\ a \in \Re^n, b \in \Re$
- Inference: $sign(h(x)) \in \{-1,1\}$

Linearly Separable Data

- Assume that the training data is linearly separable





- For the Closest Points: $h(x) = \vec{a}.\vec{x}+b \in -1,1$
- Margin: $m = \frac{1}{\|\vec{a}\|}$

Optimization Problem

- Its a Constrained Optimization Problem

$$\min_{x} \frac{1}{2}\|\vec{a}\|^2$$

$$s.t.:$$

$$y_i(\vec{x}_i.\vec{a}+b) \geqslant 1, \quad i=1,\ldots,p$$

- A convex optimization problem
- Constraints are affine hence convex

Optimization: Some Theory

- The problem:

$$\min_{x} f_0(x) \quad \leftarrow \text{objective function}$$

$$s.t.:$$

$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m \quad \leftarrow \text{inequality constraints}$$

$$h_i(x) = 0, \quad i = 1, \ldots, p \quad \leftarrow \text{equality constraints}$$

- Solution of problem: x^o
  - Global Optimum – if the problem is convex
  - Local Optimum – if the problem is not convex

Optimization: Some Theory

- Example: Standard Linear Program (LP)

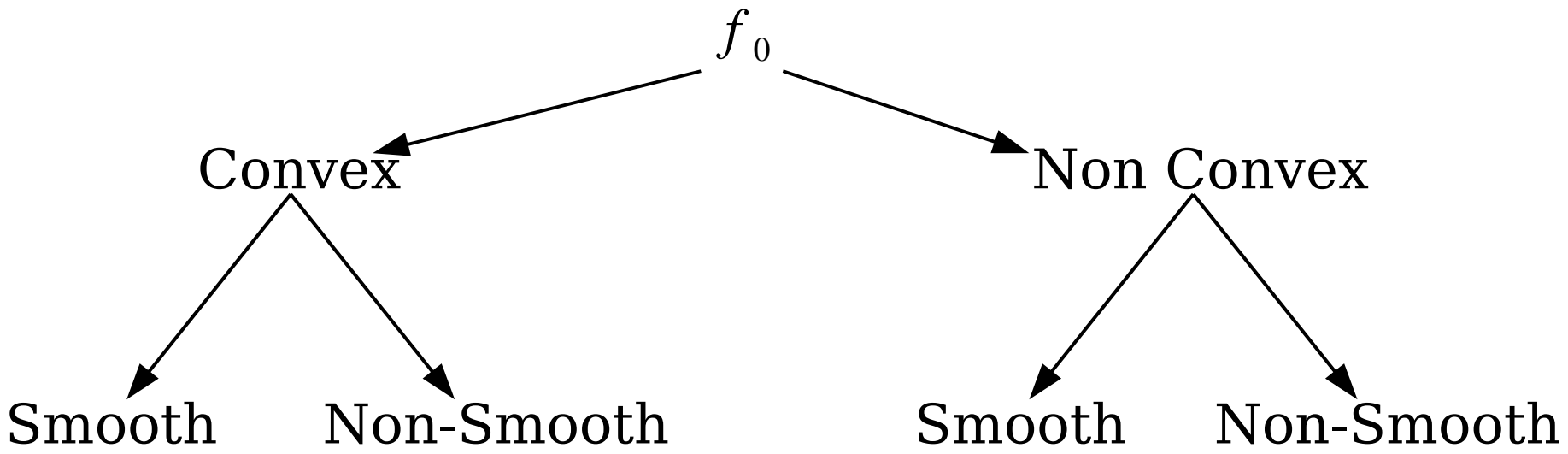
$$\begin{aligned} &\min_{x} \quad c^T x \\ &s.t.: \\ &Ax = b \\ &x \geqslant 0 \end{aligned}$$

- Example: Least Squares Solution of Linear Equations

$$\begin{aligned} &\min_{x} \quad x^T x \\ &s.t.: \\ &Ax = b \end{aligned}$$

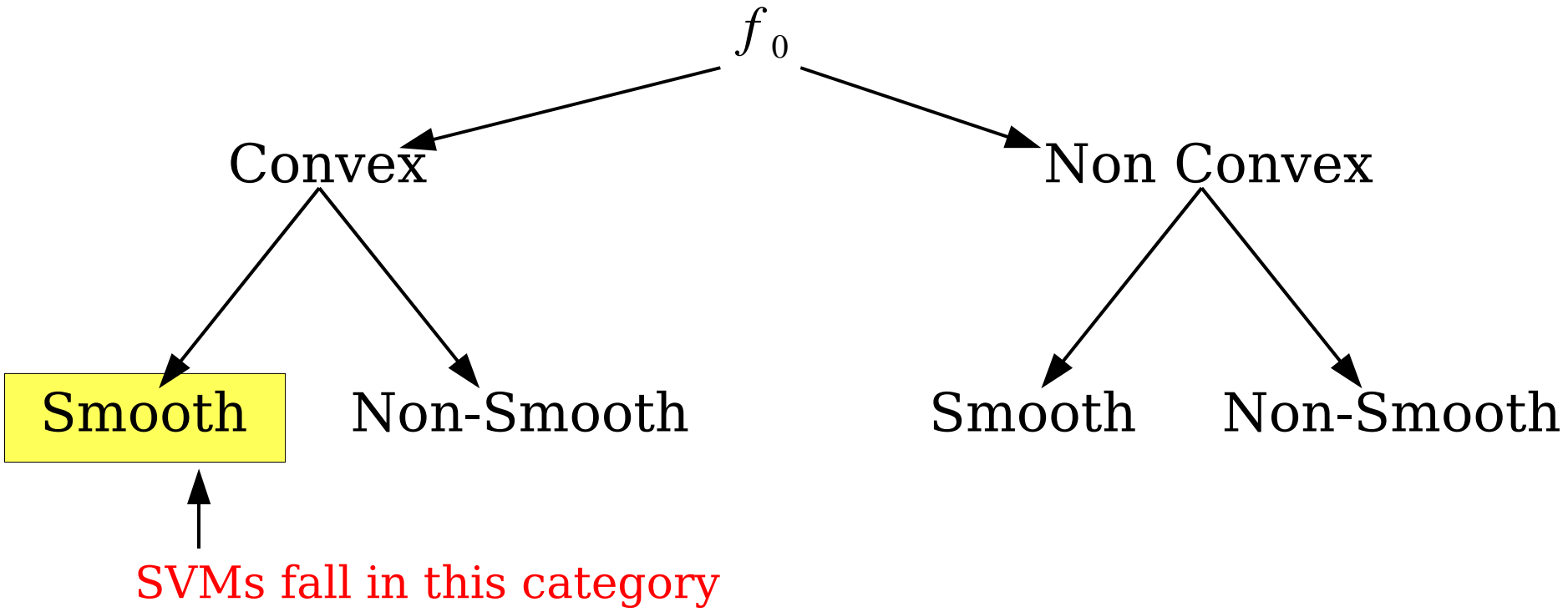
The Big Picture

- Constrained / Unconstrained Optimization
- Hierarchy of object function



The Big Picture

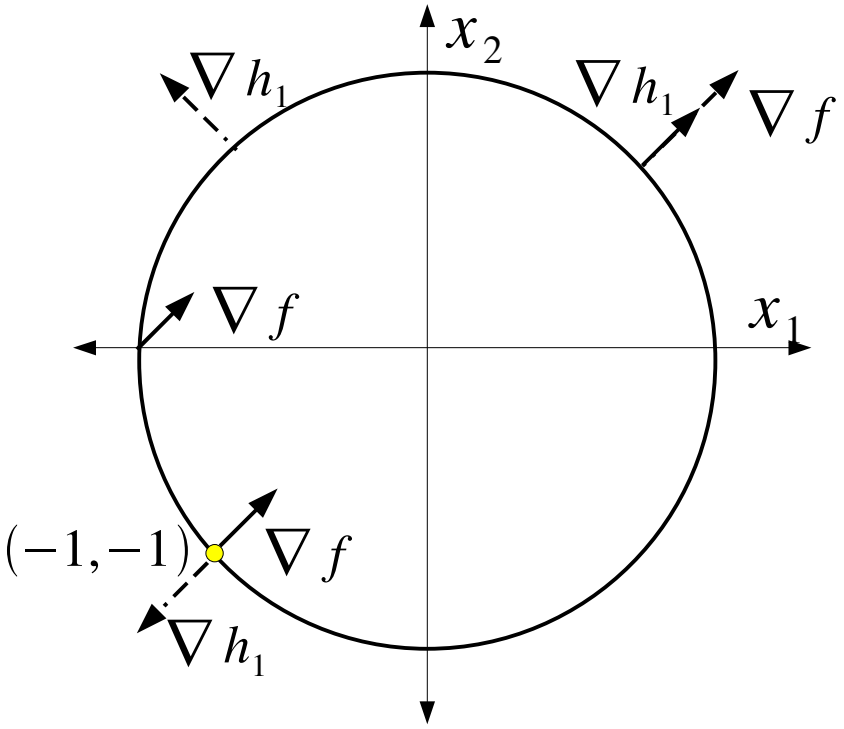
- Constrained / Unconstrained Optimization
- Hierarchy of object function



A Toy Example: Equality Constraint

- Example 1: $\quad min \quad x_1 + x_2$

$$s.t.: \quad x_1^2 + x_2^2 - 2 = 0 \qquad \equiv h_1$$



- At Optimal Solution: $\nabla f(x^o) = \lambda_1^o \nabla h_1(x^o)$

A Toy Example: Equality Constraint

- x is not an optimal solution, if there exists an $s \neq 0$ such that

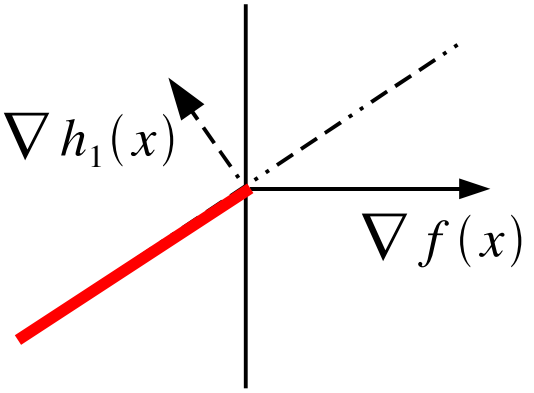
$$h_1(x+s) = 0$$
$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

$$h_1(x+s) = h_1(x) + \nabla h_1(x)^T s = \nabla h_1(x)^T s = 0 \quad (1)$$

$$f(x+s) - f(x) = \nabla f(x)^T s < 0 \quad (2)$$

- Such an s can exist only when $\nabla h_1(x)$ and $\nabla f(x)$ are not parallel



A Toy Example: Equality Constraint

- Thus we have

$$\nabla f(x^o) = \lambda_1^o \nabla h_1(x^o)$$

- The Lagrangian

Lagrange multiplier or dual variable for h_1

$$L(x, \lambda_1) = f(x) - \lambda_1 h_1(x)$$

- Thus at the solution

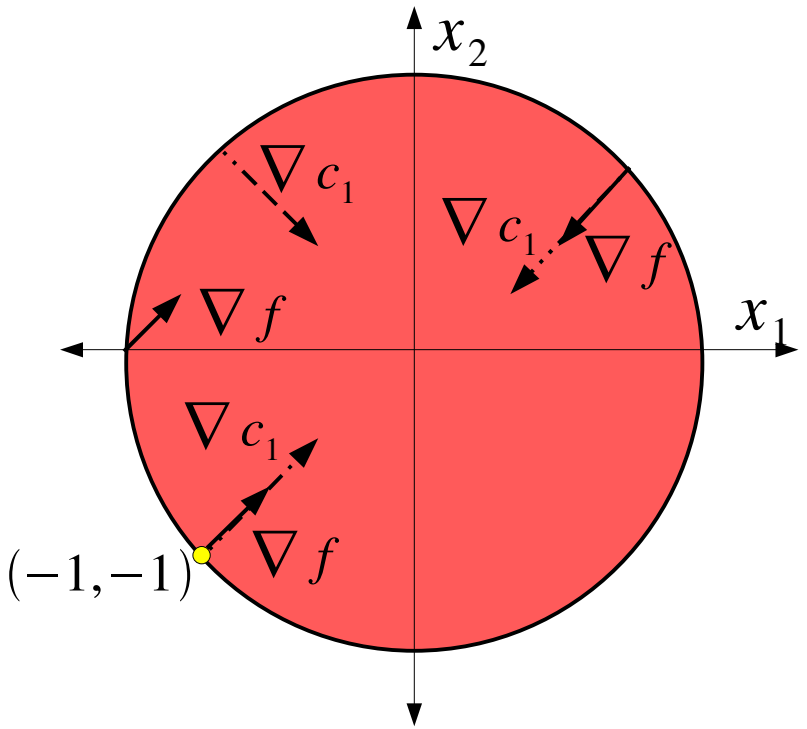
$$\nabla_x L(x^o, \lambda_1^o) = \nabla f(x^o) - \lambda_1^o \nabla h_1(x^o) = 0$$

- This is just a necessary condition and not a sufficient condition.

A Toy Example: Inequality Constraint

- Example 1: $min \quad x_1 + x_2$

  $s.t.: \quad 2 - x_1^2 - x_2^2 \geqslant 0 \quad \equiv c_1$



A Toy Example: Inequality Constraint

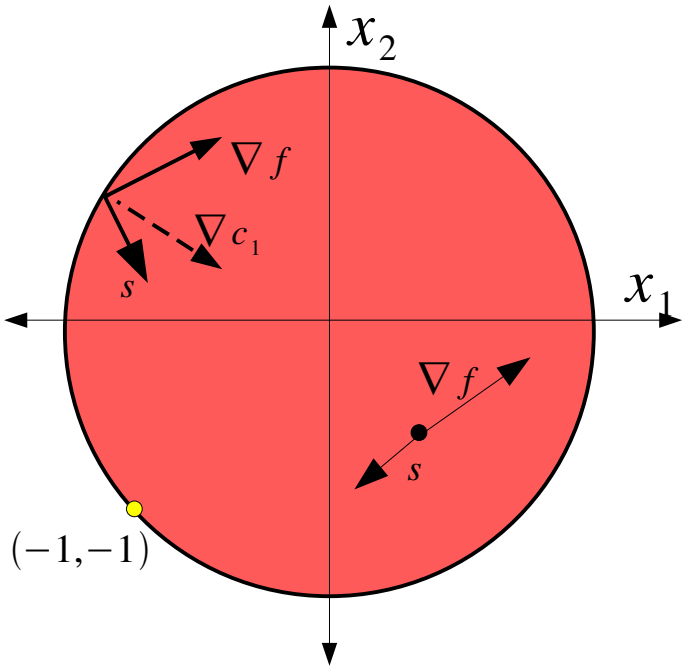
- x is not an optimal solution, if there exists an such that

$$c_1(x+s) \geqslant 0$$
$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

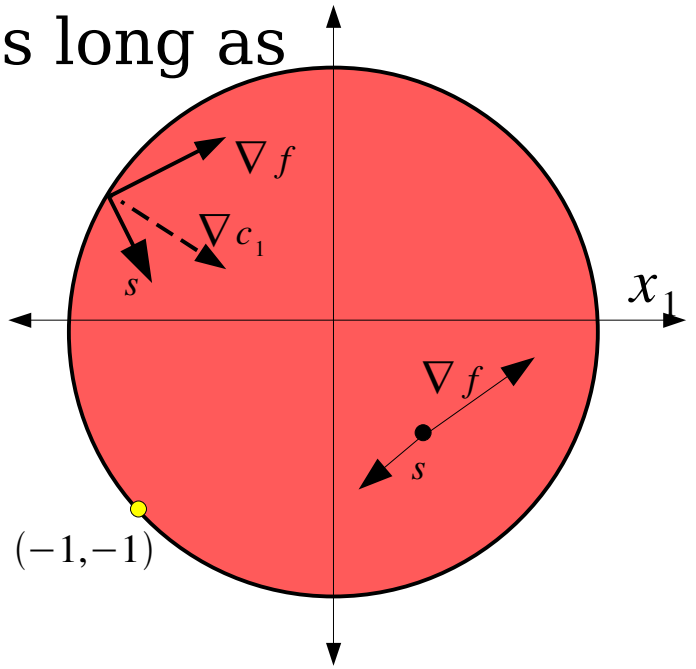
$$c_1(x+s) = c_1(x) + \nabla c_1(x)^T s \geqslant 0 \qquad (1)$$

$$f(x+s) - f(x) = \nabla f(x)^T s < 0 \qquad (2)$$



A Toy Example: Inequality Constraint

- Case 1: Inactive Constraint $c_1(x) > 0$
  - Any sufficiently small s would do as long as $\nabla f_1(x) \neq 0$
  - Thus $s = -\alpha \nabla f(x) \quad where \ \alpha > 0$

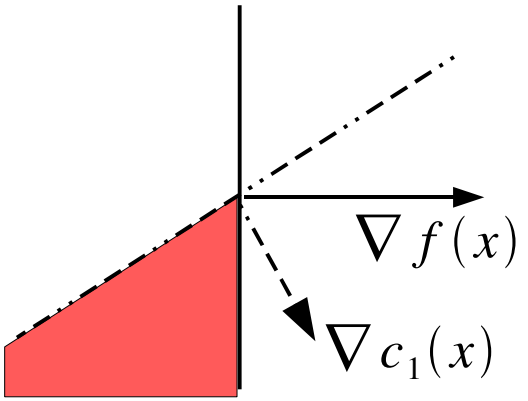


- Case 2: Active Constraint $c_1(x) = 0$

$$\nabla c_1(x)^T s \geq 0 \qquad (1)$$

$$\nabla f(x)^T s < 0 \qquad (2)$$

$$\nabla f(x) = \lambda_1 \nabla c_1(x), \qquad where \ \lambda_1 \geq 0$$



A Toy Example: Inequality Constraint

- Thus we have the Lagrangian (as before)

$$L(x,\lambda_1)=f(x)-\lambda_1 c_1(x)$$

Lagrange multiplier or dual variable for c_1

- The optimality conditions

$$\nabla_x L(x^o,\lambda_1^o)=\nabla f(x^o)-\lambda_1^o \nabla c_1(x^o) = 0 \quad \textit{for some} \quad \lambda_1 \geqslant 0$$

and

$$\lambda_1^o c_1(x^o) = 0$$

Complementarity condition

Same Concepts in a More General Setting

The Lagrangian

- The Problem

$$\min_x f_0(x)$$
$$s.t.:$$
$$f_i(x) \leqslant 0, \quad i = 1, \ldots, m$$
$$h_i(x) = 0, \quad i = 1, \ldots, p$$

- The Lagrangian associated with the problem

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

dual variables or Lagrangian multipliers

The Lagrange Dual Function

- Defined as the minimum value of the Lagrangian over x

$$g : \Re^m \times \Re^p \to \Re$$

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_1 f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

The Lagrange Dual Function

- Interpretation of Lagrange dual function:
  - Writing the original problem as unconstrained problem

$$\underset{x}{minimize} \quad \left(f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right)$$

where

$$I_0(u) = \begin{cases} 0 & u \leqslant 0 \\ \infty & u > 0 \end{cases} \qquad I_1(u) = \begin{cases} 0 & u = 0 \\ \infty & u \neq 0 \end{cases}$$

indicator functions

The Lagrange Dual Function

- Interpretation of Lagrange dual function:
  - The Lagrange multipliers in Lagrange dual function can be seen as “softer” version of indicator (penalty) function.

$$minimize \quad \left(f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right)$$

$$\inf_{x \in D} \quad \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right)$$

The Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem.

$$g(\lambda, \nu) \leqslant p^o$$

- Proof: Let $\hat{x}$ be a feasible point and let $\lambda \geqslant 0$. Then we have:

$$f_i(\hat{x}) \leqslant 0 \qquad i = 1, \ldots, m$$
$$h_i(\hat{x}) = 0 \qquad i = 1, \ldots, p$$

The Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem.

$$g(\lambda, \nu) \leqslant p^o$$

- Proof: Let $\hat{x}$ be a feasible point and let $\lambda \geqslant 0$.
  Then we have:

$$f_i(\hat{x}) \leqslant 0 \qquad i = 1, \ldots, m$$
$$h_i(\hat{x}) = 0 \qquad i = 1, \ldots, p$$

- Thus

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \sum_{i=1}^{m} \lambda_i f_i(\hat{x}) + \sum_{i=1}^{p} \nu_i h_i(\hat{x}) \leqslant f_0(\hat{x})$$

The Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem.

$$g(\lambda, \nu) \leqslant p^o$$

- Proof: Let $\hat{x}$ be a feasible point and let $\lambda \geqslant 0$.
  Then we have:

$$f_i(\hat{x}) \leqslant 0 \qquad i = 1, \ldots, m$$
$$h_i(\hat{x}) = 0 \qquad i = 1, \ldots, p$$

- Thus

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \sum_{i=1}^{m} \lambda_i f_i(\hat{x}) + \sum_{i=1}^{p} \nu_i h_i(\hat{x}) \leqslant f_0(\hat{x}) \qquad . \leqslant 0$$

The Lagrange Dual Function

- Lagrange dual function gives a lower bound on optimal value of the problem.

$$g(\lambda, \nu) \leqslant p^o$$

- Proof: Let $\hat{x}$ be a feasible point and let $\lambda \geqslant 0$.
  Then we have:

$$\begin{aligned} f_i(\hat{x}) &\leqslant 0 \qquad i = 1, \ldots, m \\ h_i(\hat{x}) &= 0 \qquad i = 1, \ldots, p \end{aligned}$$

- Thus

$$L(\hat{x}, \lambda, \nu) = f_0(\hat{x}) + \sum_{i=1}^{m} \lambda_i f_i(\hat{x}) + \sum_{i=1}^{p} \nu_i h_i(\hat{x}) \leqslant f_0(\hat{x})$$

- Hence

$$g(\lambda, \nu) = \inf_{x \in D} L(x, \lambda, \nu) \leqslant L(\hat{x}, \lambda, \nu) \leqslant f_0(\hat{x})$$

The Lagrange Dual Problem

- Lagrange dual function gives a lower bound on optimal value of the problem.
- It is natural to seek the "best" lower bound.

$$\begin{aligned} &\textit{maximize} \quad g(\lambda, \nu) \\ &\textit{s.t.:} \quad \lambda \succcurlyeq 0 \end{aligned}$$

- Dual feasibility:

$$(\lambda, \nu): \quad \lambda \succcurlyeq 0, \;\; g(\lambda, \nu) \geq -\infty$$

- The dual optimal value and solution:

$$d^o = g(\lambda^o, \nu^o)$$

- The Lagrange dual problem is convex even if the original problem is not.

Primal / Dual Problems

- Primal problem:

$$\begin{aligned} &\min_{x} f_0(x) \\ &s.t.: \qquad\qquad\qquad\qquad\qquad\qquad p^o \\ &f_i(x) \leqslant 0, \quad i = 1, \ldots, m \\ &h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

- Dual problem:

$$\begin{aligned} &\max_{\lambda, \nu} \quad g(\lambda, \nu) \qquad\qquad\qquad d^o \\ &s.t.: \quad \lambda \geqslant 0 \end{aligned}$$

Weak Duality

- Weak duality theorem:

$$d^o \leqslant p^o$$

- Optimal duality gap:

$$p^o - d^o \geqslant 0$$

- This bound is sometimes used to get an estimate on the optimal value of the original problem that is difficult to solve.

Strong Duality

- Strong Duality:

$$d^o = p^o$$

- Strong duality does not hold in general.
- Slater's Condition: If $x \in relint\, D$, that it is strictly feasible.

$$f_i(x) < 0 \quad for \quad i = 1, \ldots m$$
$$h_i(x) = 0 \quad for \quad i = 1, \ldots p$$

- Strong duality theorem: Strong duality holds if Slater's condition holds.
- It also implies that the dual optimal value is attained.

$$\exists (\lambda^o, \nu^o) \quad with \quad g(\lambda^o, \nu^o) = d^o = p^o$$

Optimality Conditions: First Order

- Complementary slackness: If strong duality holds, then at optimality

$$\lambda_i^o f_i(x^o) = 0 \qquad i=1,\dots m$$

- Proof: We have

$$\begin{aligned} f_0(x^o) &= g(\lambda^o, \nu^o) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^o f_i(x) + \sum_{i=1}^p \nu_i^o h_i(x) \right) \\ &\leqslant f_0(x^o) + \sum_{i=1}^m \lambda_i^o f_i(x^o) + \sum_{i=1}^p \nu_i^o h_i(x^o) \\ &\leqslant f_0(x^o) \end{aligned}$$

less than 0

- The result follows

Optimality Conditions: First Order

- Karush-Kuhn-Tucker (KKT) Conditions: If the strong duality holds, then at optimality

$$
\begin{aligned}
f_i(x^o) &\leqslant 0, \quad i=1,\ldots,m \\
h_i(x^o) &= 0, \quad i=1,\ldots,p \\
\lambda_i^o &\geqslant 0, \quad i=1,\ldots,m \\
\lambda_i^o f_i(x^o) &= 0, \quad i=1,\ldots,m \\
\nabla f_0(x^o) + \sum_{i=1}^{m} \lambda_i^o \nabla f_i(x^o) + \sum_{i=1}^{p} \nu_i^o \nabla h_i(x^o) &= 0
\end{aligned}
$$

- KKT conditions are necessary in general and necessary and sufficient in case of convex problems.