Support Vector Machines: Maximum Margin Classifiers

Machine Learning and Pattern Recognition:
September 23, 2010

Piotr Mirowski
Based on slides by Sumit Chopra, Fu-Jie Huang and Mehryar Mohri

Outline

- What is behind Support Vector Machines?
  - Constrained optimization
  - Lagrange constraints
  - “Dual” solution
- Support Vector Machines in detail
  - Kernel trick
  - LibSVM demo

Binary Classification Problem

- Given: Training data generated according to the distribution D

$$(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \Re^n \times \{-1, 1\}$$

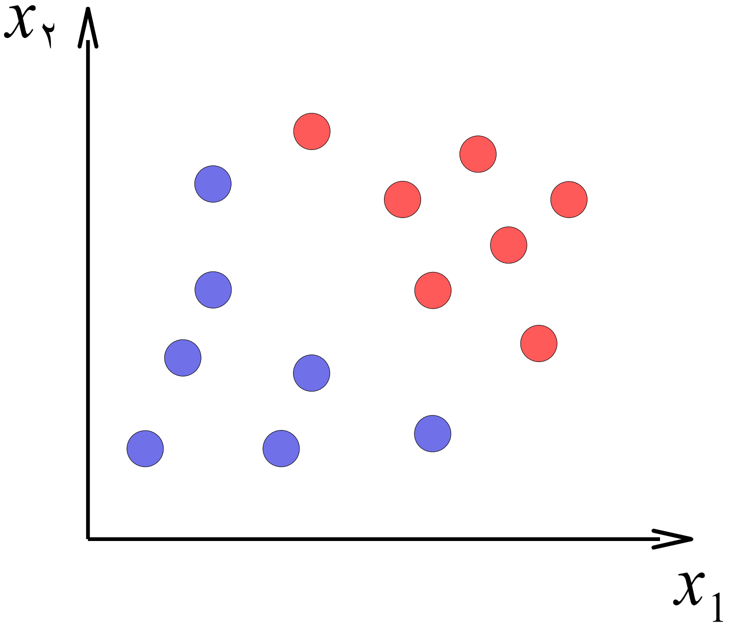
input label input space label space

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

- Solution:
  - Linear Approach: linear classifiers (e.g. logistic regression, Perceptron)
  - Non Linear Approach: non-linear classifiers (e.g. Neural Networks, SVM)

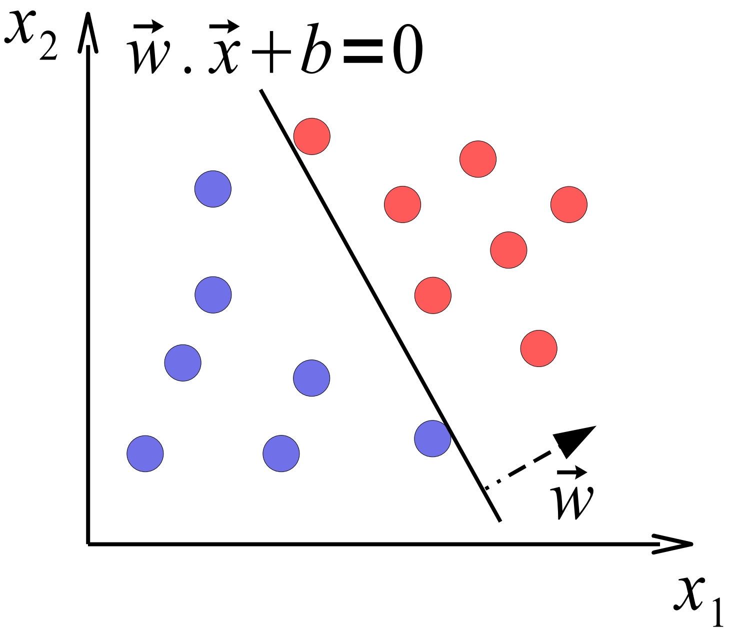
Linearly Separable Data

- Assume that the training data is linearly separable



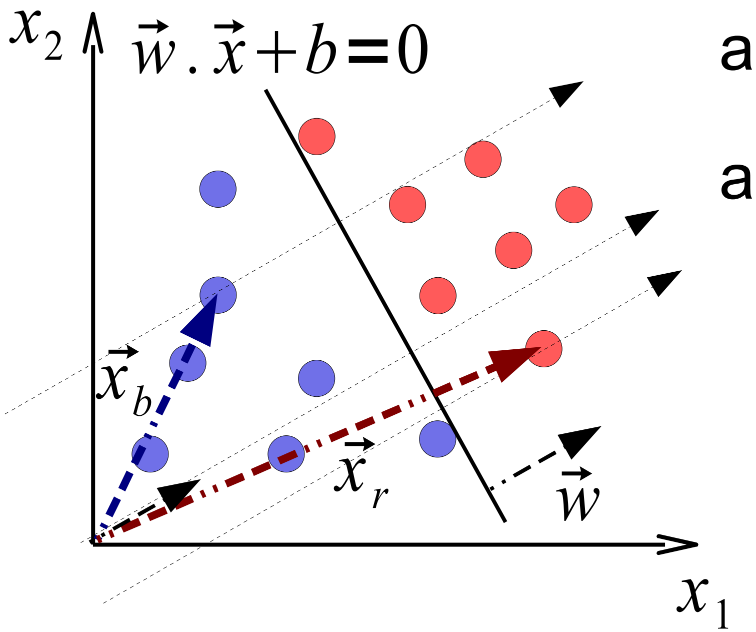
Linearly Separable Data

- Assume that the training data is linearly separable



Linearly Separable Data

- Assume that the training data is linearly separable



abscissa on axis parallel to $\vec{w}$

abscissa of origin *0* is b

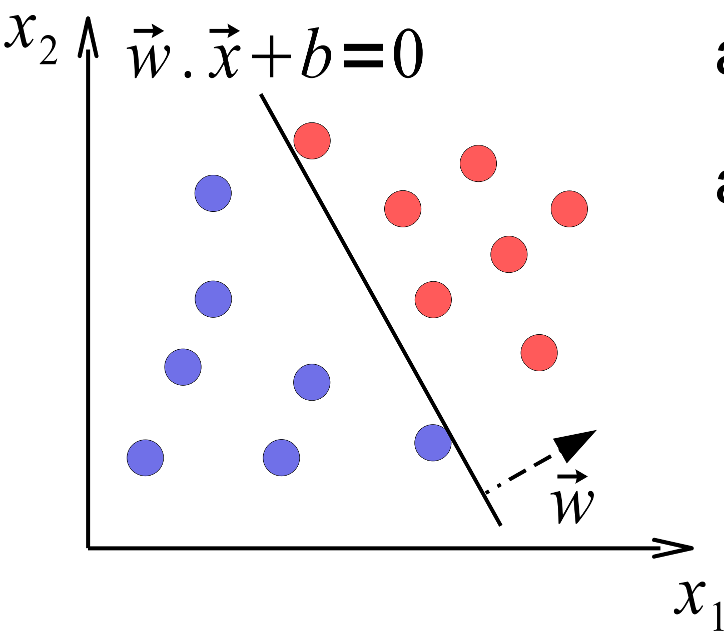
$$\vec{w}.\vec{x}_b+b=\tilde{y}_b$$

$$\vec{w}.\vec{x}_r+b=\tilde{y}_a$$

$$\vec{w}.\vec{O}+b=b$$

Linearly Separable Data

- Assume that the training data is linearly separable



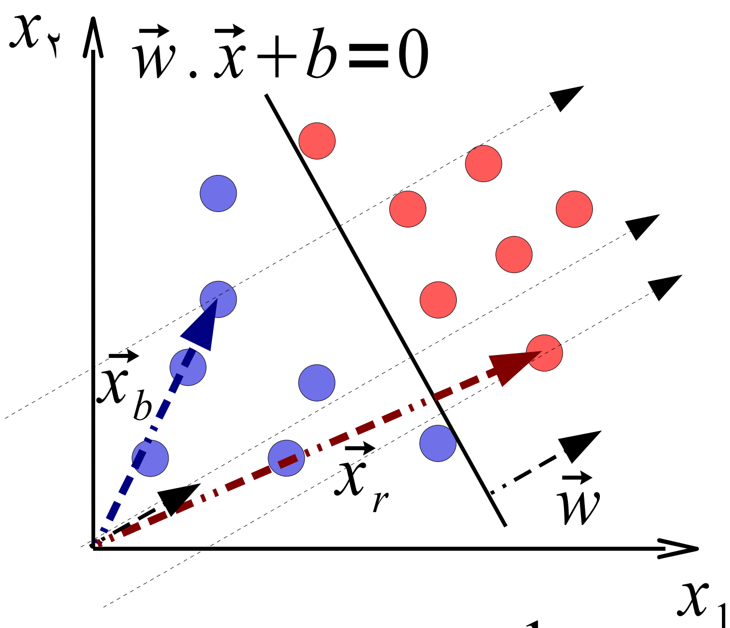
abscissa on axis parallel to $\vec{w}$

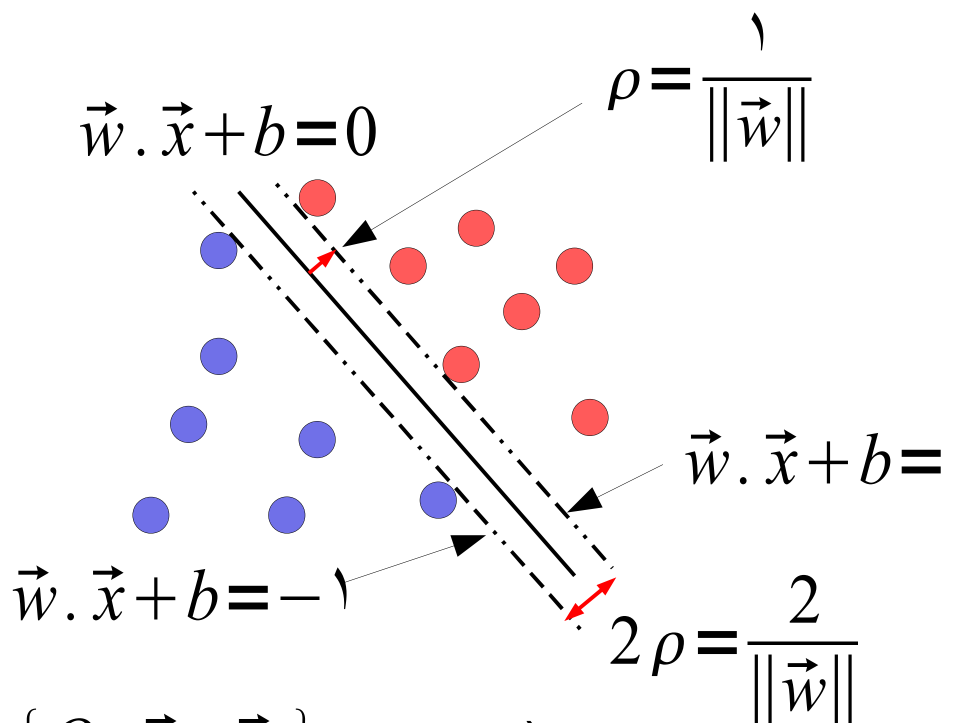
abscissa of origin *0* is b

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

Linearly Separable Data

- Assume that the training data is linearly separable





- Margin $\rho=\frac{1}{\|\vec{w}\|}$ (in the $\{O,\vec{x}_1,\vec{x}_2\}$ space)

- Maximize margin ρ (or 2ρ) so that:

  For the closest points: $h(x) = \vec{w}.\vec{x}+b \in \{-1,1\}$

Optimization Problem

- A Constrained Optimization Problem

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t.:$$

$$y_i(\boldsymbol{w}.\boldsymbol{x}_i + b) \geqslant 1, \quad i = 1,\ldots,m$$

label input

- Equivalent to maximizing the margin $\rho = \frac{1}{\|\boldsymbol{w}\|}$
- A convex optimization problem:
  - Objective is convex
  - Constraints are affine hence convex
  - Therefore, admits an unique optimum at $\boldsymbol{w}_0$

Optimization Problem

- Compare:

$$\min_{w} \frac{1}{2}\|\boldsymbol{w}\|^2$$

objective

$$s.t.:$$

$$y_i(\boldsymbol{w}.\boldsymbol{x_i}+b) \geqslant 1, \quad i=1,\ldots,m$$

constraints

- With:

$$\min_{w}\left(\sum_{i=1}^{m}\left(-y_i(\boldsymbol{w}.\boldsymbol{x_i}+b)\right)+\frac{\lambda}{2}\|\boldsymbol{w}\|^2\right)$$

energy/errors

regularization

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 (unique) optimum – if the problem is convex
  - Local optimum – if the problem is not convex

(notation change: the parameters to optimize are noted x)

Optimization: Some Theory

- Example: Standard Linear Program (LP)

$$\min_x \quad c^T x$$
$$s.t.:$$
$$Ax = b$$
$$x \geqslant 0$$

- Example: Least Squares Solution of Linear Equations (with L_2 norm regularization of the solution x)
  i.e. Ridge Regression

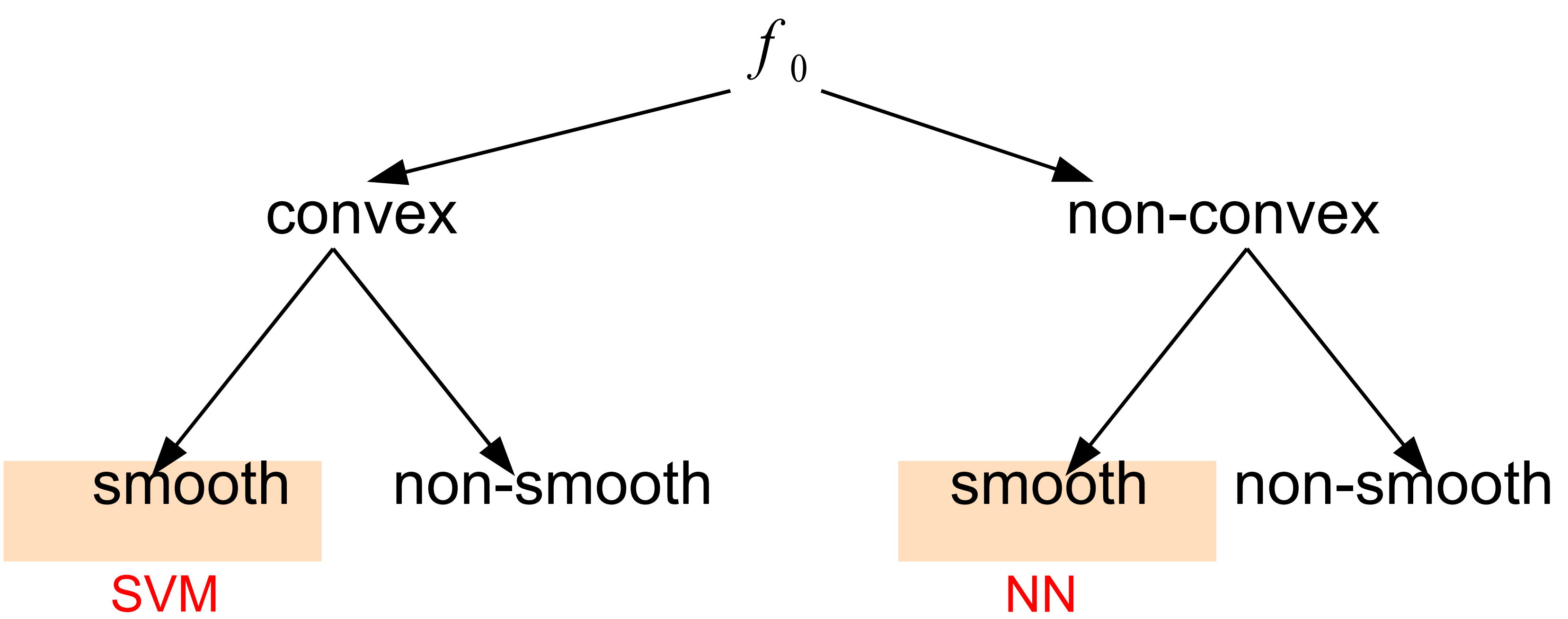
$$\min_x \quad x^T x$$
$$s.t.:$$
$$Ax = b$$

Big Picture

- Constrained / unconstrained optimization
- Hierarchy of objective function:

  smooth = infinitely derivable

  convex = has a global optimum



Introducing the concept of Lagrange function on a toy example

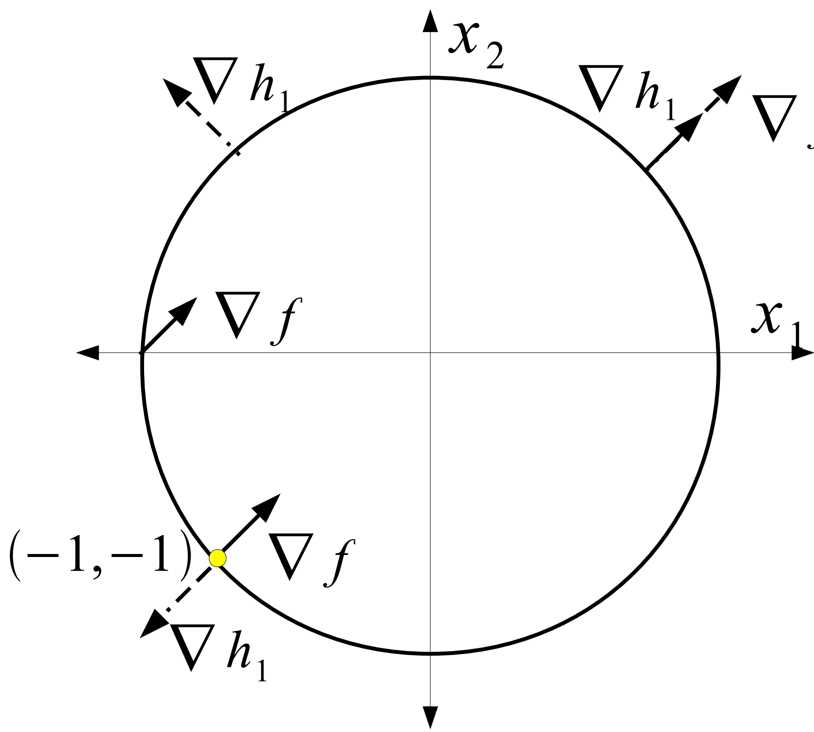
Toy Example: Equality Constraint

- Example 1:

$$\min \quad x_1 + x_2 \qquad \equiv f$$

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

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix}$$



$$\nabla h_1 = \begin{pmatrix} \frac{\partial h_1}{\partial x_1} \\ \frac{\partial h_1}{\partial x_2} \end{pmatrix}$$

- At Optimal Solution:

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

Toy Example: Equality Constraint

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

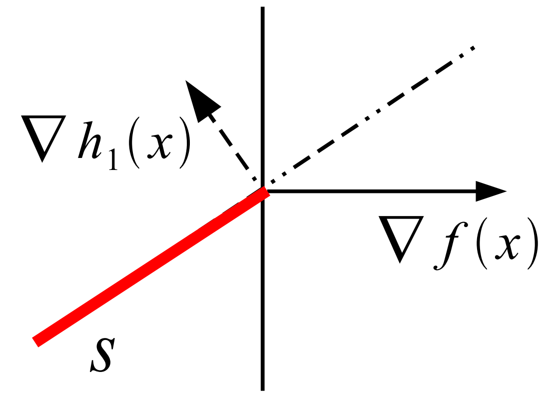
$$h_1(x+s) = 0$$
$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

$$\cancel{h_1(x+s)} = \cancel{h_1(x)} + \nabla h_1(x)^T s = \nabla h_1(x)^T s = 0 \qquad (1)$$

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

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



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

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

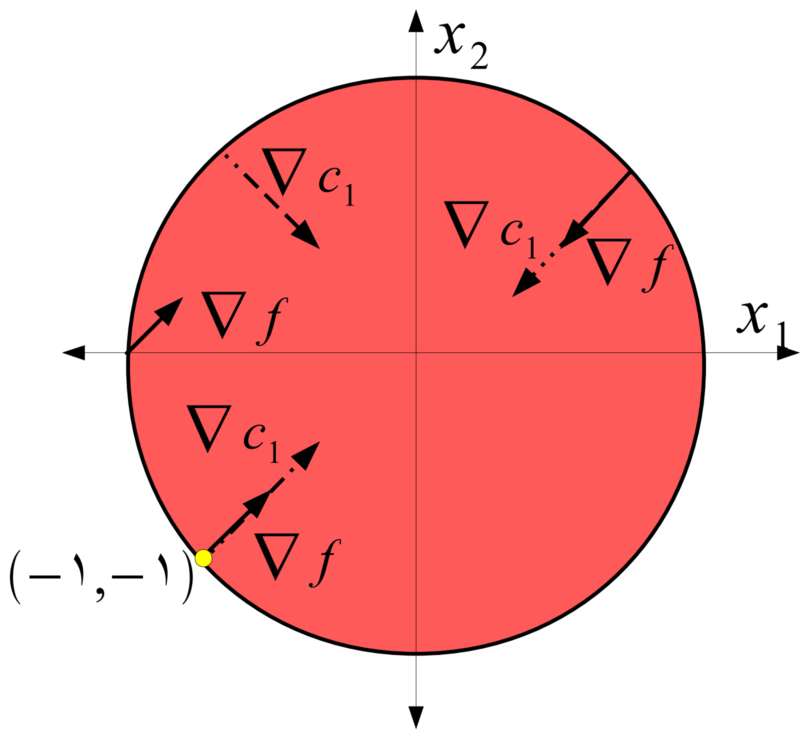
- This is just a necessary (not a sufficient) condition"
  x solution implies $\nabla h_1(x) \parallel \nabla f(x)$

Toy Example: Inequality Constraint

- Example 2:

$$\min \quad x_1 + x_2 \qquad \equiv f$$

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



$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{pmatrix}$$

$$\nabla c_1 = \begin{pmatrix} \frac{\partial c_1}{\partial x_1} \\ \frac{\partial c_1}{\partial x_2} \end{pmatrix}$$

Toy Example: Inequality Constraint

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

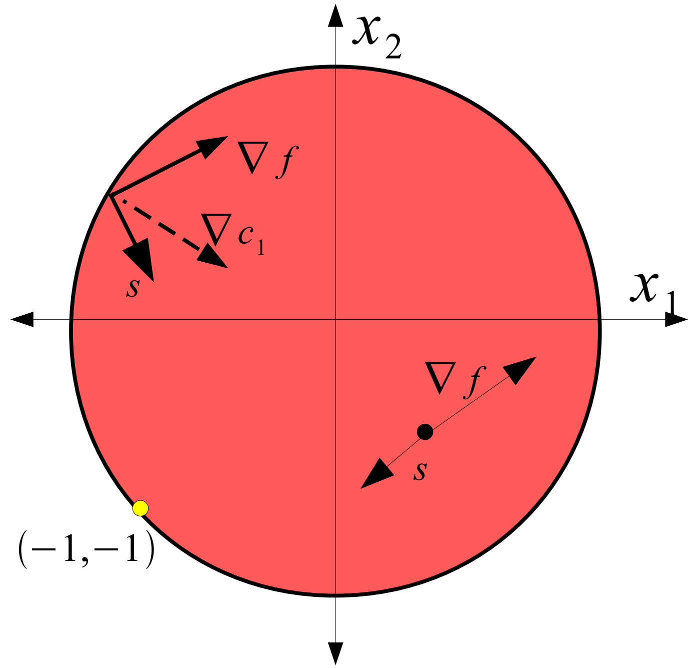
$$c_1(x+s) \geqslant 0$$

$$f(x+s) < f(x)$$

- Using first order Taylor's expansion

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

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



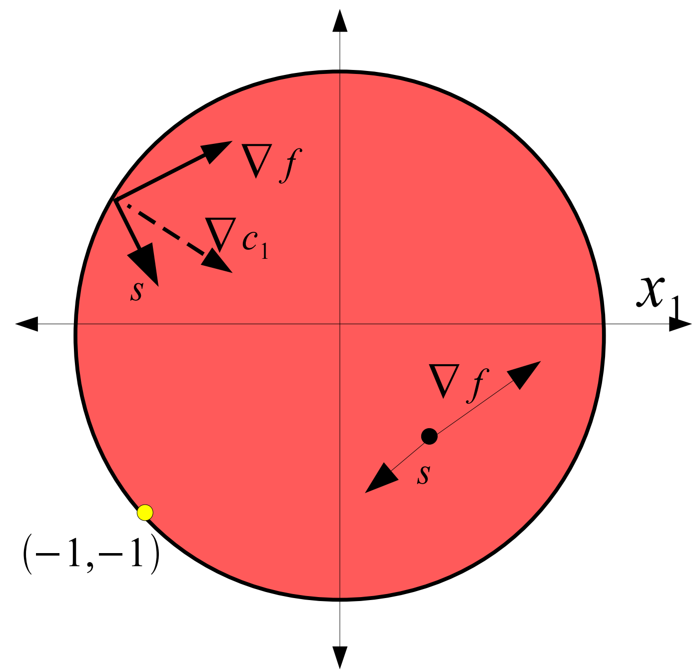
Toy Example: Inequality Constraint

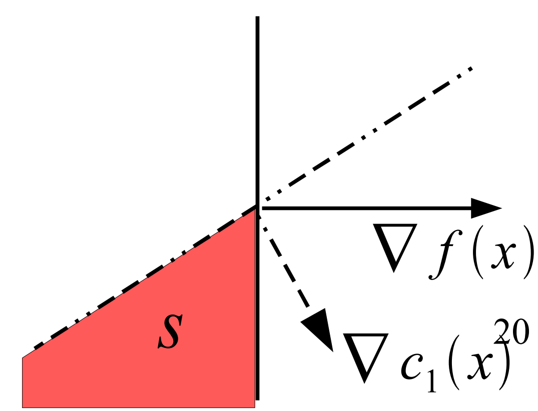
- Case 1: Inactive constraint $c_1(x) > 0$
  - Any sufficiently small s as long as $\nabla f_1(x) \neq 0$
  - Thus $s = -\alpha \nabla f(x) \quad where\ \alpha > 0$
- Case 2: Active constraint $c_1(x) = 0$

$$\nabla c_1(x)^T s \geqslant 0 \quad (1)$$
$$\nabla f(x)^T s < 0 \quad (2)$$

In that case, $s = 0$ when:

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





Toy Example: Inequality Constraint

- Thus we have the Lagrange function (as before)

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

Lagrange multiplier or dual variable for c_1

- The optimality conditions

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

and

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

Complementarity condition

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

(active) (inactive)

Same Concepts in a More General Setting

Lagrange Function

- The Problem

$$\min_x f_0(x)$$ objective function

$$s.t.:$$

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

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

- Standard tool for constrained optimization: the Lagrange Function

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

dual variables or Lagrange multipliers

Lagrange Dual Function

- Defined, for λ, ν as the minimum value of the Lagrange function over x

m inequality constraints
p equality constraints

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

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

Lagrange Dual Function

- Interpretation of Lagrange dual function:
  - Writing the original problem as unconstrained problem but with hard indicators (penalties)

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

where

satisfied

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

unsatisfied

satisfied

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

unsatisfied

indicator functions

Lagrange Dual Function

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

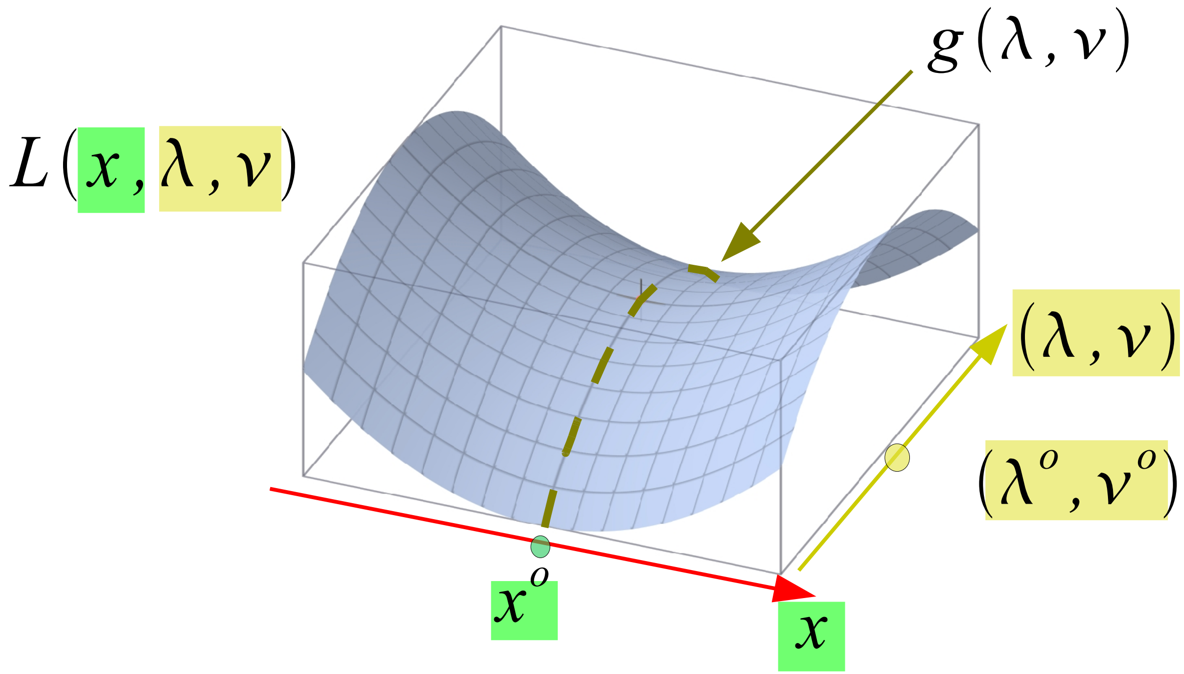
$$\underset{x}{\text{minimize}} \left(f_0(x) + \sum_{i=1}^{m} I_0(f_i(x)) + \sum_{i=1}^{p} I_1(h_i(x)) \right)$$

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

Sufficient Condition

- If $\left(x^o, \lambda^o, \nu^o\right)$ is a saddle point, i.e. if

$$\forall x \in \Re^n, \quad \forall \lambda \geqslant 0, \quad L\left(x^o, \lambda, \nu\right) \leqslant L\left(x^o, \lambda^o, \nu^o\right) \leqslant L\left(x, \lambda^o, \nu^o\right)$$



- ... then $\left(x^o, \lambda^o, \nu^o\right)$ is a solution of the primal problem p^o

Lagrange Dual Problem

- Lagrange dual function gives a lower bound on the optimal value of the problem.
- We seek the "best" lower bound to minimize the objective:

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

- 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:

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

- Dual problem:

$$d^o \qquad \begin{aligned} &\max_{\lambda, \nu} \quad g(\lambda, \nu) \\ &s.t.: \quad \lambda \geqslant 0 \\ &g(\lambda, \nu) = \inf_{x \in D} \left(f_0(x) + \sum_{i=1}^{m} \lambda_1 f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x) \right) \end{aligned}$$

Optimality Conditions: First Order

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

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

- KKT conditions are
  - necessary in general (local optimum)
  - necessary and sufficient in case of convex problems (global optimum)