Consider an example (with equality constraints)

$$\min_{(x_1, x_2)} \; x_1 + x_2 \quad \text{s.t.} \quad x_1^2 + x_2^2 = a^2$$

(The points are on a circle)

$f(x) = x_1 + x_2$

$\underline{x} \in \mathbb{R}^2$

$\nabla f = \left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2}\right)$

$\nabla f = (1 \quad 1)$

$= (1, 1)$

$c = x_1^2 + x_2^2 - a^2 \in E$

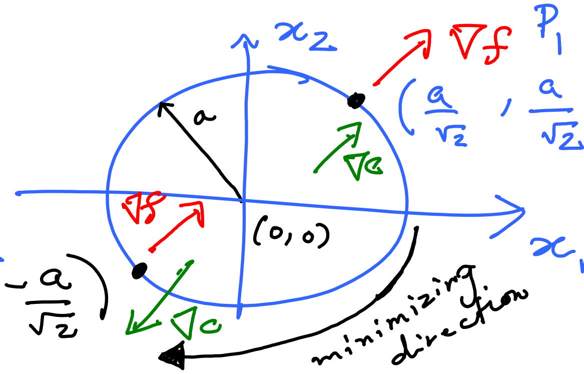
$= 0$

$\nabla c = \left(\frac{\partial c}{\partial x_1} \quad \frac{\partial c}{\partial x_2}\right)$

$\nabla c = (2x_1 \quad 2x_2)$

$\nabla f_{\left(-\frac{a}{\sqrt{2}}, -\frac{a}{\sqrt{2}}\right)} = $

$\nabla c_{\left(-a/\sqrt{2}, -\frac{a}{\sqrt{2}}\right)} = \left(-\sqrt{2}\,a, -\sqrt{2}\,a\right)$



$\overline{\mathrm{III}}^{rd}$ quadrant has both x_1 and x_2 $-$ve

$\Longrightarrow$ Soln lies there

Just ∇f does not suffice for minima !

From the figure,

$$\nabla f(\underline{x}^*) = \lambda_1^* \nabla c(\underline{x}^*)$$

$$\lambda_1^* = \frac{-1}{a\sqrt{2}}$$

Note that : ∇f is a scalar multiple of ∇c @ the point of maxima as well i.e., $\left(\frac{a}{\sqrt{2}}, \frac{a}{\sqrt{2}}\right)$

Let us analyze this issue through a Taylor Series expansion around the constraint.

$$c(\underline{x}) = 0$$

(C: Equality constraint)

$$c(\underline{x}+\underline{d}) = 0$$

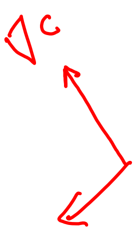
(To maintain feasibility w.r.t. $c(\underline{x}) = 0$)

$$c(\underline{x}+\underline{d}) \approx c(\underline{x}) + \underbrace{\nabla c^T(\underline{x})\,\underline{d}}_{\text{inner product}}$$

(With a first order approx.)

$$c(\underline{x}) + \nabla c^T(\underline{x})\,\underline{d} = 0$$

(C: $c(\underline{x}) = 0$)



$$\Rightarrow \boxed{\nabla c^T(\underline{x})\,\underline{d} = 0}$$

A

|||ly the direction of optimization must produce a decrease in f

$$\underbrace{f(\underline{x}+\underline{d})}_{f(\underline{x})+\nabla^T f(\underline{x})\cdot d} - f(\underline{x}) < 0$$

By doing a Taylor expansion around $\underline{x}$ using a 1st order approx.

$$\nabla^T f(\underline{x})\,\underline{d} < 0 \qquad \text{(B)}$$

If $\exists$ a $\underline{d}$ satisfying (A) and (B), an improvement exists

There are 2 cases to consider here

(1) Such a direction does not exists

(2) Such a direction exists

Case 1: When such a direction $\nexists$, ∇f and ∇c are scalar multiples of each other

i.e., $\nabla f \nearrow \; \swarrow \nabla c$ or $\nabla f \nearrow \; \nearrow \nabla c$.

i.e., ∇f and ∇c can point in the same or opposite directions

$$\boxed{\nabla f = \lambda \nabla c}$$

Ponder why: When $\nabla f = \lambda \nabla G$ Ⓐ and Ⓑ 'do not' simultaneously hold

$\therefore$ forming a Lagrangian

Lagrange multiplier

$$L = f \pm \lambda G$$

$$\nabla L = 0 \implies \nabla f = \mp \lambda G$$

$\therefore$ Sign of the "constraint in the Lagrangian" does not matter!

Sign does not matter

We can arrive at a saddle point here

We still need the sign of the Hessian

to proceed & assess the validity.

Case 2 : When such a direction exists

$$\underline{d} = -\left(I - \frac{\nabla c \; \nabla c^T}{\| \nabla c \|^2} \right) \nabla f \qquad \text{(I)}$$

Let us verify if (I) satisfies (A) and (B)

$$\underline{d} = -\nabla f + \frac{\overbrace{\nabla c \; \nabla c^T}^{\text{outer product}} \nabla f}{\underbrace{\nabla c^T \; \nabla c}_{\text{Inner product}}} \qquad \text{(1)}$$

Let us consider (A)

Pre-multiply (1) by ∇c^T;

$$\nabla c^T \underline{d} = -\nabla c^T \nabla f + \frac{\overset{(\text{scalar})}{\cancel{\nabla c^T \nabla c}}\, \nabla c^T \nabla f}{\underset{(\text{scalar})}{\cancel{\nabla c^T \nabla c}}} = 0$$

Let us <u>Consider B</u>

$$\nabla f^T \underline{d}$$

Plug in $\underline{d}$

$$= -\nabla f^T \left(\nabla f - \frac{\nabla c \, \nabla c^T \nabla f}{\| \nabla c \|^2} \right)$$

$$= -\underbrace{\nabla f^T \nabla f}_{\text{1st term}} + \underbrace{\frac{\nabla f^T \nabla c \, \nabla c^T \nabla f}{\| \nabla c \|^2}}_{\text{2nd term}}$$

$$= -\|\nabla f\|^2 + \frac{\|\nabla^T f \nabla c\|^2}{\|\nabla c\|^2} < 0$$

($\because$ Cauchy Schwartz inequality)

The equality is ruled out due to Case A
($\because \nabla f \neq \lambda \nabla c$)

$\Rightarrow$ $\underline{d}$ is the direction satisfying the constraints.

Single inequality constraint

$c(\underline{x}) \geq 0$

$$0 \leq c(\underline{x}+\underline{d}) \approx c(\underline{x}) + \nabla^T c(\underline{x})\,\underline{d}$$

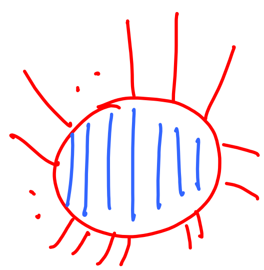
Observe the '$\leq$' as against '$=$' in equality constraint

Feasibility of $\underline{d}$ is retained while still improving the objective if

$$\underbrace{c(\underline{x})}_{\geq 0} + \nabla^T c(\underline{x})\,\underline{d} \geq 0$$

C

Considering the example from the circular constraint with inequality conditions, we are optimizing over all points lying on & inside the circle.



$$\underbrace{x_1^2 + x_2^2}\le a^2$$

$$-(x_1^2 + x_2^2) \geq -a^2$$

$$c(x) \geq 0 \longrightarrow a^2 - x_1^2 - x_2^2 \geq 0$$

We have 2 cases

Case A : The strict inequality holds i.e., $c(\underline{x}) > 0$

Whenever $\nabla f(\underline{x}) \neq 0$ i.e., when we have not yet reached optimum points

II
$$\begin{cases} \nabla^T f(\underline{x})\, \underline{d} < 0 & \text{———} \quad (\because \text{ B }) \\ c(\underline{x}) + \nabla^T c(\underline{x})\, \underline{d} \geq 0 & \text{———} \quad (\because \text{ C }) \end{cases}$$

A $\underline{d}$ that satisfies the constraints is

$$\underline{d} = -\frac{c(\underline{x})\, \nabla f(\underline{x})}{\|\nabla f(\underline{x})\|\,\|\nabla c(\underline{x})\|} \quad \text{———} \quad D$$

We can verify that (D) satisfies both the constraints in (II)

$$\text{(i)} \quad \nabla^T f(\underline{x})\, \underline{d} = -\nabla^T f(\underline{x}) \cdot \overset{\text{scalar}}{c(\underline{x})} \frac{\nabla f(\underline{x})}{\|\nabla f(\underline{x})\| \, \|\nabla c(\underline{x})\|}$$

$$= -\underset{>0}{c(\underline{x})} \frac{\nabla^T f(\underline{x}) \nabla f(\underline{x})}{\|\nabla f(\underline{x})\| \, \|\nabla c(\underline{x})\|}$$

Evaluates to $\|\nabla f\|$

$$-\frac{c(x) \|\nabla f\|}{\|\nabla c\|}$$

$\Longrightarrow$ First constraint in (II) is satisfied i.e., < 0

(ii) Consider

$$c(\underline{x}) + \nabla^T c(\underline{x})\, \underline{d}$$

$$= c(\underline{x}) + \nabla^T c(\underline{x}) \left[- \frac{c(\underline{x})\, \nabla f(\underline{x})}{\|\nabla f(\underline{x})\|\, \|\nabla c(\underline{x})\|} \right]$$

$$= c(\underline{x}) - \underbrace{c(\underline{x})}_{\geq 0} \frac{\nabla^T c(\underline{x})\, \nabla f(\underline{x})}{\|\nabla f(\underline{x})\|\, \|\nabla c(\underline{x})\|} < 1$$

$$|\cdot| < 1$$

Unless $\nabla f(\underline{x}) \neq \lambda \nabla c(\underline{x})$,

$$\nabla^T c(\underline{x}) \nabla f(\underline{x}) < \|\nabla f(\underline{x})\| \, \|\nabla c(\underline{x})\|$$

We have $\quad c(\underline{x}) + \nabla^T c(\underline{x}) \, \underline{d}$

$$\Longrightarrow \quad c(\underline{x}) - c(\underline{x}) \, \alpha$$

α can be +ve or −ve

$$c(\underline{x})(1-\alpha) \geq 0$$

(The equality is only over the case when $\alpha = 1$)

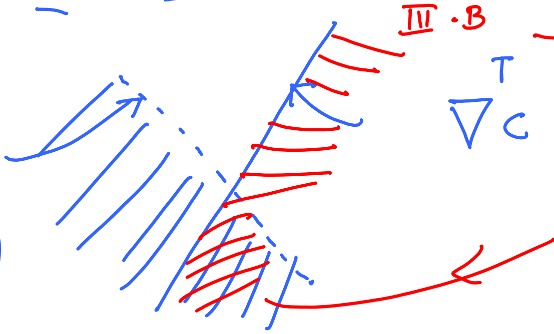
Case B: When $\underline{x}$ is on the boundary of the constraint eqn i.e., $c(\underline{x}) = 0$

We have

III $\left\{ \begin{array}{ll} \nabla^T f(\underline{x})\,\underline{d} < 0 & \text{—— III. A —— (B)} \\ \nabla^T c(\underline{x})\,\underline{d} \geq 0 & \text{—— III. B —— (C)} \end{array} \right.$

boundary case $c(\underline{x}) = 0$ plug into (C)

FEASIBLE SOLN REGION (GEOMETRY)



$\nabla^T c(\underline{x})\,\underline{d} \geq 0$

$\nabla^T f(\underline{x}) . \underline{d} < 0$ (open half plane)

Region where the constraints in III are met.

∇f ↗ ↗ ∇c

When $\nabla f = \lambda \nabla c$ and
∇f and ∇c point in the same direction
the regions from III do not intersect. !

∇f ↗ ↙ ∇c

When $\nabla f = -\lambda \nabla c$ where $\lambda > 0$
the constrained regions satisfying III overlap into
an entire half space! (Fully intersect!)

Forming the Lagrangian for $\lambda > 0$

($\lambda = 0$ $\Rightarrow$ No constraint)

If $L = f - \lambda c$ When $\lambda > 0$

$\nabla L = \nabla f - \lambda \nabla c = 0$ $\Rightarrow$ The search stops since constraints are not met

$\nabla f = + \lambda \nabla c$

$\Longrightarrow$

With $c(x) \geq 0$ While forming the "Lagrangian" with inequality constraints, have a '−' sign before the constraint scaled by '$\lambda \geq 0$'!

If the inequality was $c(\underline{x}) \leq 0$, we can form a $g(\underline{x}) \geq 0$ such that

$$g(\underline{x}) = -c(\underline{x}) \geq 0$$

Support Vector Machines

SVMs : Another class of algorithms for pattern classification and non linear regression.

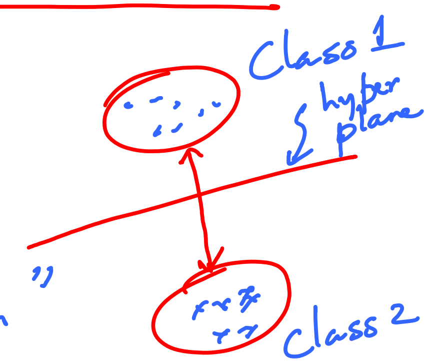
It is a linear machine

$w^T x + b$

Roots to SVMs : Vladmir Vapnik

Very elegant theory with firm roots in convex optimization

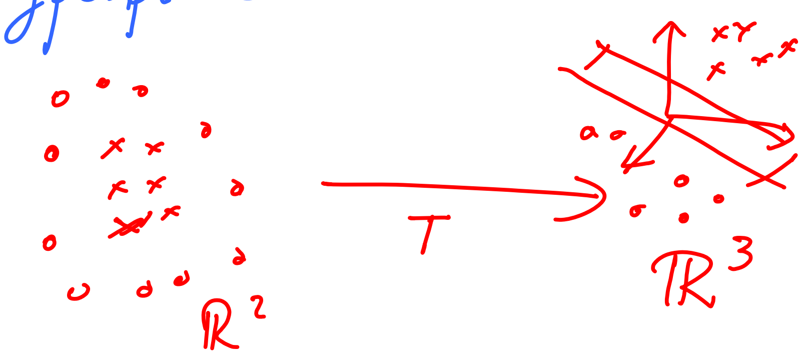
Idea: Construct a hyperplane as the decision surface in such a way that the margin of separation between the 2 classes is maximized



Idea of deriving the hyperplane stems from "structural risk minimization"

In the case of <u>linearly separable patterns,</u> we need to derive a hyperplane that solves our objective.

In the case of <u>non-linearly separable patterns,</u> we need to lift the data points to a higher dimension so that we can still derive a hyperplane that solves our objective.



A notion central to the SVM is the "inner product kernel" between a support vector $\underline{x}_i^{(s)}$ and a vector $\underline{x}$ drawn from the input space.

The support vectors are a small subset of vectors extracted of the training set by the algo.

Optimal hyperplane for linearly separable patterns

Consider the training samples $\{\underline{x}_i, d_i\}_{i=1}^{N}$

$\underline{x}_i$: i/p pattern for the i^{th} example

d_i : target

Assume that the patterns represented by $d_i = \{+1, -1\}$ is linearly separable

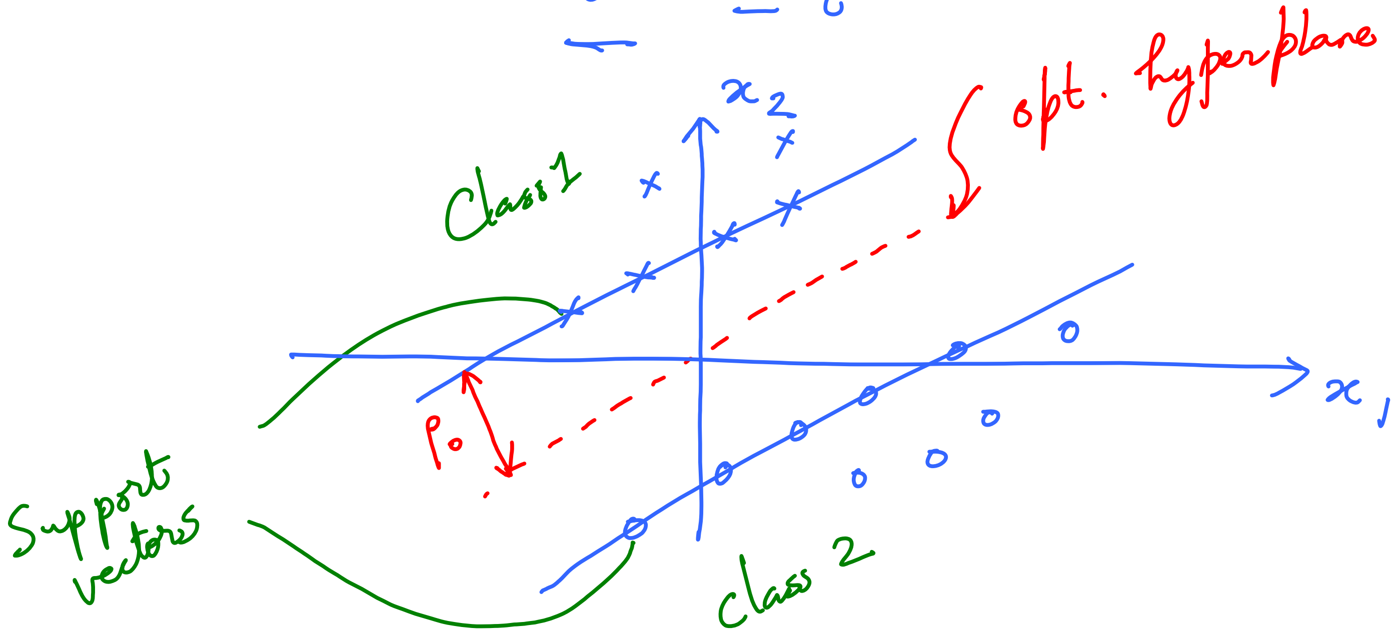
The eqn of the decision surface is

$$\underline{w}^T \underline{x} + b = 0$$

Now,

$$\underline{w}^T \underline{x}_i + b \geq 0 \quad \text{for } d_i = +1$$

$$\underline{w}^T \underline{x}_i + b < 0 \quad \text{for } d_i = -1$$



Let $\underline{w}_0$ and b_0 be the opt. values of the weight vector and the bias

$$\underline{w}_0^T \underline{x} + b_0 = 0$$

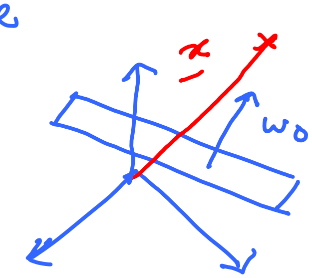
Eqn of the decision boundary

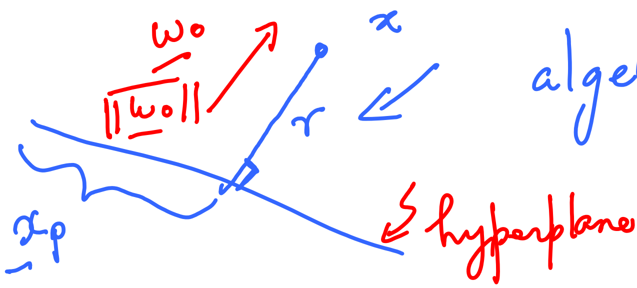
Let us write the discriminant function as

$$g(\underline{x}) = \underline{w}_0^T \underline{x} + b_0$$

From our notion of the normal to a plane

$$\underline{x} = \underline{x}_p + r \frac{\underline{w}_0}{\|\underline{w}_0\|}$$





algebraic distance of the point $\underline{x}$ w.r.t plane

$$\underline{x} = \underline{x}_p + r \frac{\underline{w}_0}{\|\underline{w}_0\|}$$

normal projection of $\underline{x}$ on to the hyperplane

r is +ve if $\underline{x}$ is on the +ve side of the hyperplane

— " — -ve — " — -ve side of the hyperplane

$g(\underline{x}_p) = 0 \quad \left(\because \underline{x}_p \text{ lies on the } \underline{\text{discriminant boundary}} \right)$

$g(\cdot)$ is an $\underline{\text{affine map}}$ $\quad g(x) = (\underline{w}_0^T \underline{x} + b_0)$ $\quad b_0 = 0$ (Linear map)

$$g(\underline{x}) = g\left(\underline{x}_p + r \frac{\underline{w}_0}{\|\underline{w}_0\|}\right) = \underbrace{\underline{w}_0^T\left(\underline{x}_p + r \frac{\underline{w}_0}{\|\underline{w}_0\|}\right) + b_0}$$

$$g(\underline{x}) = \underbrace{\underline{w}_0^T \underline{x}_p + b_0}_{g(\underline{x}_p) = 0} + r \cdot \frac{\underline{w}_0^T \underline{w}_0}{\|\underline{w}_0\|} \qquad \|\underline{w}_0\|^2 = r\,\|\underline{w}_0\|$$

$$\therefore \quad r = \frac{g(\underline{x})}{\|\underline{w}_0\|}$$

Relationship between the alg. distance, $g(\underline{x})$, $\underline{w}_0$

Now, the distance from the origin to the hyperplane $\dfrac{b_0}{\|\underline{w}_0\|}$

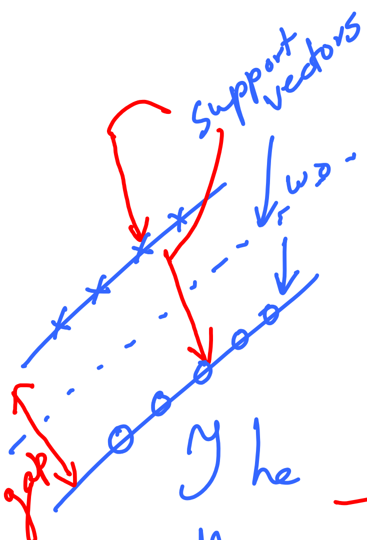
If $b_0 > 0$; the origin is on the +ve side of the hyperplane

— " — $b_0 < 0$; the origin — " — -ve side — " —

If $b_0 = 0$, the opt. hyperplane passes through the origin!

Our training set comprises of $\mathcal{T} = \{\underline{x}_i, d_i\}_{i=1}^{N}$

Opt. hyperplane $\underline{w}_o^T x + b_o = 0$



$$\begin{cases} \underline{w}_o^T \underline{x}_i + b_o \geq 1 & d_i = +1 \\ \underline{w}_o^T \underline{x}_i + b_o \leq 1 & d_i = -1 \end{cases} \quad (A)$$

The particular data points $(\underline{x}_i, d_i)$ for which the eqns in (A) are satisfied with equality are the "support vectors"!

$\underline{x}_i^{(s)} \leftarrow$ support vect.

Consider a support vector $\underline{x}^{(s)}$

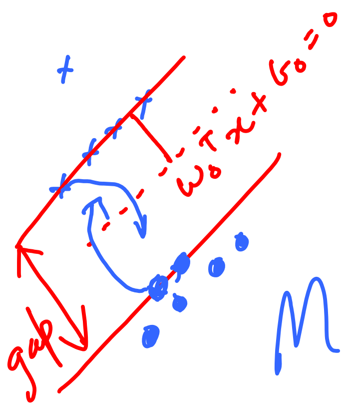
$$g(\underline{x}^{(s)}) = \underline{w}_o^T \underline{x}^{(s)} + b_o = \mp 1 \quad \text{for} \quad \underline{d}^{(s)} = \mp 1$$

The algebraic distance from the support vector $\underline{x}^{(s)}$ to the opt. hyperplane is

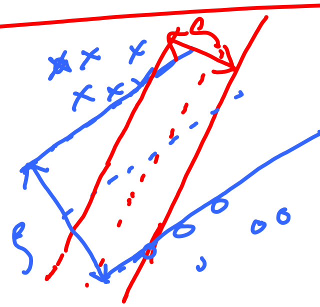
$$r = \frac{g(\underline{x}^{(s)})}{\|\underline{w}_o\|} = \begin{cases} \dfrac{1}{\|\underline{w}_o\|} & \text{if } d^{(s)} = +1 \\ -\dfrac{1}{\|\underline{w}_o\|} & \text{if } d^{(s)} = -1 \end{cases}$$

Let ρ be the opt. value of margin of separation

$$\rho = 2r \quad \text{where} \quad r = \frac{1}{\|\underline{w}_0\|}$$



Max. margin of separation "ρ" $\Longrightarrow$ Min the Euclidean norm of $\underline{w}_0$



Quadratic Optimization for finding opt. hyperplane

Given $\mathcal{T} = \{\underline{x}_i, d_i\}_{i=1}^{N}$, find the opt. hyperplane

subject to

$$d_i(\underline{w}^T \underline{x}_i + b) \geq 1 \quad \text{for } i = 1, 2, \ldots, N$$

and the weight vector that minimizes the cost function

$$\phi(\underline{w}) = \frac{1}{2} \underline{w}^T \underline{w}$$

NOTE :
a) Cost function is convex
b) Constraints are linear in $\underline{w}$

Set up the Lagrangian function

$$J(\underline{w}, b, \underline{\alpha}) = \frac{1}{2}\underline{w}^T\underline{w} - \sum_{i=1}^{N} \alpha_i \left[d_i (\underline{w}^T \underline{x}_i + b) - 1 \right]$$

wt. vector, bias, vector of all Lag. multipliers for each constraint

Observe the sign flip required for inequality constraints

Lagrange multiplier for each constraint 'i'

Conditions

1) $$\frac{\partial J(\underline{w}, b, \underline{\alpha})}{\partial \underline{w}} = 0$$

2) $$\frac{\partial J(\underline{w}, b, \underline{\alpha})}{\partial b} = 0$$

3) Trivially $\dfrac{\partial J(\underline{w}, b, \underline{\alpha})}{\partial \alpha_i}$ gives us the constraints

Evaluating the partial derivatives

Condition 1 gives us,

$$\underline{w} = \sum_{i=1}^{N} \alpha_i d_i \underline{x}_i \qquad \text{———} \quad (1)$$

Condition 2 gives us,

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad \text{———} \quad (2)$$

Due to the nature of the convex opt. set up, soln is unique

NOTE :

1) It is important to note that, at the saddle point, for each Lagrange multiplier α_i, the product of that multiplier with the constraint vanishes

i.e., $\alpha_i \left[d_i \left(\underline{w}^T \underline{x}_i + b \right) - 1 \right] = 0 \quad \forall \ i = 1, \ldots, N$

(Home Work)

$\alpha_i \neq 0$

$$\Rightarrow \boxed{d_i \left(\underline{w}^T \underline{x}_i + b \right) - 1 = 0}$$

Primal & dual problems

1) If the primal problem has an optimal solution, the dual too has, and the corresponding opt. values are equal. (For convex problems)

2) In order to find $\underline{w}_{opt}$ for the primal problem, we may need to find an alternative variable that optimizes the dual problem

$$J(\underline{w}, b, \underline{\alpha}) = \frac{1}{2} \underbrace{\underline{w}^T \underline{w}}_{1} - \underbrace{\sum_{i=1}^{N} \alpha_i d_i \underline{w}^T \underline{x}_i}_{2}$$

$$\underbrace{- b \sum_{i=1}^{N} \alpha_i d_i}_{3} + \underbrace{\sum_{i=1}^{N} \alpha_i}_{4}$$

(Expanding from the primal problem)

From the optimality conditions,

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \qquad \left(\frac{\partial J(\cdot)}{\partial b} = 0 \right)$$

Also, $\underline{w}^T \underline{w} = \sum_{i=1}^{N} \alpha_i d_i \underline{w}^T \underline{x}_i$

$\left(\because \text{Condition } 1 \quad \frac{\partial J(\cdot)}{\partial \underline{w}} = 0 \right)$

$$\therefore \underline{w}^T \underline{w} = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i d_i \alpha_j d_j \underline{x}_i^T \underline{x}_j$$

α_i's are non-negative

Our dual objective function is $Q(\alpha)$ given by

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \underline{x}_i^T \underline{x}_j$$

Statement of the dual problem

Given training samples $\{\underline{x}_i, d_i\}_{i=1}^{N}$, find Lagrange multipliers $\{\alpha_i\}_{i=1}^{N}$ that maximize $Q(\alpha)$

Subject to the conditions

1) $\sum_{i=1}^{N} \alpha_i d_i = 0$

2) $\alpha_i \geq 0 \quad \forall \; i = 1, \ldots, N$

Note that the dual problem is recast completely in terms of training data!

Having obtained the opt. Lagrange multipliers, denoted by $\alpha_{opt,i}$, (each constraint $i = 1, \dots, N$) we may compute the opt. weight $\underline{w}_{opt}$ and write it as

$$\underline{w}_{opt} = \sum_{i=1}^{N} \alpha_{opt,i} \, d_i \, \underline{x}_i$$

Opt. bias $b_0 = 1 - \underline{w}_0^T \underline{x}^{(s)}$ for $d^{(s)} = 1$