Example 2: Least Norm Solution

Least norm solution: $A \in \mathbb{R}^{m\times n}$, $n \geq m$

$$\min_{x\in\mathbb{R}^n} \quad \frac{1}{2}x^\top x$$
$$\text{s.t.} \quad Ax = b. : \mu \in \mathbb{R}^m, \ \underline{A \text{ is full-rank}}.$$

Find L and d.

$$L(x,\mu) = \frac{1}{2}x^T x + \mu^T(Ax-b)$$

$$\underline{d(\mu)} = \inf_{x\in\mathbb{R}^n} L(x,\mu) = \inf_{x\in\mathbb{R}^n}\left[\frac{1}{2}x^T x + \mu^T A x - \mu^T b\right]$$

necessary condition
$$\begin{cases} \nabla_x L(x,\mu) = x + A^T\mu = 0 \\ \nabla_x^2 L(x,\mu) = I \succ 0 \end{cases}$$
$$\Downarrow$$
$$\bar{x} = -A^T\mu$$
minimizes $L(x,\mu)$.

then, $d(\mu) = \frac{1}{2}(-A^T\mu)^T(-A^T\mu) + \mu^T A(-A^T\mu) - \mu^T b$

$$= \frac{1}{2}\mu^T A A^T \mu - \mu^T A A^T \mu - \mu^T b$$

<u>Dual optimization problem:</u> $\max_{\mu\in\mathbb{R}^m} \boxed{-\frac{1}{2}\mu^T A A^T \mu - \mu^T b}$

$$\cong \min_{\mu\in\mathbb{R}^m} \frac{1}{2}\mu^T A A^T \mu + \mu^T b.$$

optimal dual solution μ^* satisfies $\nabla_\mu d(\mu) = 0$

$\nabla_\mu^2 d(\mu) = -AA^T$ and AA^T is positive definite. $\Rightarrow AA^T\mu^* + b = 0$

$$\Rightarrow \underline{\mu^* = -(AA^T)^{-1}b.}$$

Then $x^* = -A^T(-(AA^T)^{-1}b)$

$$x^* = A^T(AA^T)^{-1}b.$$

Lecture 16: 6th Feb

Towards Optimality Conditions

Consider the primal optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, i \in [m] := \{1,2,\dots,m\}, \\ & h_j(x) = 0, j \in [p]. \end{aligned}$$

Let the dual function be defined as

$$d(\lambda,\mu) := \inf_x L(x,\lambda,\mu) = \inf_x \left[f(x) + \sum_{i\in[m]} \lambda_i g_i(x) + \sum_{j\in[p]} \mu_i h_j(x) \right].$$

The corresponding dual optimization problem is:

$$\begin{aligned} \max_{\lambda\in\mathbb{R}^m,\mu\in\mathbb{R}^p} \quad & d(\lambda,\mu) \\ \text{s.t.} \quad & \lambda \ge 0, \\ & (\lambda,\mu) \in \mathtt{dom}(d). \end{aligned}$$

Consequently, for any $(\bar x, \bar\lambda, \bar\mu)$ with $\bar x$ being primal feasible and $\bar\lambda \ge 0$, we have

$$d(\lambda^*,\mu^*) \qquad d(\bar\lambda,\bar\mu) = \inf_x \left[f(x) + \sum_{i\in[m]} \bar\lambda_i g_i(x) + \sum_{j\in[p]} \bar\mu_j h_j(x) \right]$$

$$= \quad \le \left[f(\bar x) + \underbrace{\sum_{i\in[m]} \underset{\ge 0}{\bar\lambda_i}\, \underset{\le 0}{g_i(\bar x)}}_{\le 0} + \sum_{j\in[p]} \bar\mu_j \underset{=0}{h_j(\bar x)} \right]$$

$$= \quad \le f(\bar x). \qquad = f(x^*)$$

When do both the above inequalities hold with equality?

- the last inequality holds with equality when

$$\sum_{i\in[m]} \bar\lambda_i g_i(\bar x) = 0 \iff \bar\lambda_i g_i(\bar x) = 0 \quad \forall i = 1,2,..m$$

- the first inequality holds with equality when
  - $\bar x$ is the minimizer of $L(x,\bar\lambda,\bar\mu)$.
    - necessary condition $\nabla_x L(\bar x,\bar\lambda,\bar\mu) = 0$
    - sufficient condition $\nabla_x^2 L(\bar x,\bar\lambda,\bar\mu) \succeq 0$.

Necessary Conditions for Optimality

Theorem 3

Suppose strong duality holds. Let $x^\star$ be the optimal solution of the primal problem, and $\lambda^\star, \mu^\star$ be the optimal solution of the dual problem. Then, the following conditions are satisfied.

- **Primal Feasibility**: $g_i(x^\star) \le 0, i \in [m], h_j(x^\star) = 0, j \in [p]$.
- **Dual Feasibility**: $\lambda^\star \ge 0$.
- **Complementary Slackness**: $\lambda_i^\star g_i(x^\star) = 0$ for all $i \in [m]$.
- **Lagrangian Stationarity**:

$$\nabla_x L(x^\star, \lambda^\star, \mu^\star) = \nabla_x f(x^\star) + \sum_{i \in [m]} \lambda_i^\star \nabla_x g_i(x^\star) + \sum_{j \in [p]} \mu_i^\star \nabla_x h_j(x^\star) = 0.$$

The above four conditions are called Karush–Kuhn–Tucker (KKT) optimality conditions.

When are the above conditions sufficient for optimality?

Essentially, the Lagrangian stationarity conditions must imply that $x^\star$ is the optimal solution of the Lagrangian problem. When it is true?

Suppose $(\hat{x}, \hat{\lambda}, \hat{\mu})$ satisfying the KKT conditions. Then,

$$d(\hat{\lambda}, \hat{\mu}) = \inf_x L(x, \hat{\lambda}, \hat{\mu})$$

$$\le L(\hat{x}, \hat{\lambda}, \hat{\mu}) = f(\hat{x}) + \sum_i \underbrace{\hat{\lambda}_i g_i(\hat{x})}_{=0} + \sum_j \mu_j \underbrace{h_j(\hat{x})}_{=0}$$

$$= f(\hat{x})$$

Sufficient Condition for Optimality for Convex Problems

Theorem 4

Suppose the primal optimization problem is **convex**. Let $\bar{x}, \bar{\lambda}$ and $\bar{\mu}$ satisfy KKT conditions stated above. Then,

- $d(\bar{\lambda}, \bar{\mu}) = f(\bar{x})$ (strong duality holds).
- $\bar{x}$ is optimal solution of primal problem.
- $(\bar{\lambda}, \bar{\mu})$ are optimal solution of dual problem.

It is still possible for a convex optimization problem to have an optimal solution but no KKT points. We need additional conditions to make sure that optimal solutions satisfy KKT conditions.

Let $\bar{x}, \bar{\lambda}$ and $\bar{\mu}$ satisfy KKT conditions stated above. From primal and dual feasibility we have

$$\begin{aligned} d(\bar{\lambda}, \bar{\mu}) &= \inf_x \left[f(x) + \sum_{i \in [m]} \bar{\lambda}_i g_i(x) + \sum_{j \in [p]} \bar{\mu}_i h_j(x) \right] \\ &\leq f(\bar{x}) + \sum_{i \in [m]} \bar{\lambda}_i g_i(\bar{x}) + \sum_{j \in [p]} \bar{\mu}_i h_j(\bar{x}) \leq f(\bar{x}). \end{aligned}$$

Further, both inequalities hold with equality.

Constraint Qualification and Strong Duality

Theorem 5

Suppose the primal optimization problem is convex which satisfies Slater's constraint qualification condition: there exists $\bar{x} \in \texttt{int}(\mathcal{D})$ in the domain of the optimization problem for which $g_i(\bar{x}) < 0$ for all $i \in [m]$ and $h_i(\bar{x}) = 0$ for all $i \in [p]$. Then, strong duality holds with $f^* = d^*$. Moreover, if $f^* > -\infty$, then, there exist $(\lambda^\star, \mu^\star)$ such that $g(\lambda^\star, \mu^\star) = d^* = f^*$.

- The point $\bar{x}$ need not be an optimal solution. It is any arbitrary feasible point.

- **Relaxed Slater Condition:** If some of the inequality constraints are affine, then they need not hold with strict inequality. It is sufficient to find $\bar{x} \in \texttt{relint}(\mathcal{D})$ such that $g_i(\bar{x}) < 0$ for all g_i that are not affine.

- We now have the following result.

  Proposition 1. *Suppose the primal problem is convex and satisfies Slater's condition. Then, if a feasible solution x^* is optimal, then there exist λ^*, μ^* such that (x^*, λ^*, μ^*) satisfy KKT conditions.*

- Note that sufficiency part holds even without Slater's condition.

- An alternative condition is **Linear Independence Constraint Qualification (LICQ)** which holds at a feasible solution x^* if the vectors

$$\begin{aligned} \nabla h_j(x^*), &\quad j \in [p], \\ \nabla g_i(x^*), &\quad i \in \{k \in [m] | g_k(x^*) = 0\} \end{aligned}$$

are linearly independent. If LICQ holds at a point x^*, then KKT conditions are necessary for the local optimality of x^*.

Convex Theorem of the Alternative

Consider the following general form of optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \le 0, i \in [m] := \{1, 2, \dots, m\}, \end{aligned}$$

where f and g_i are convex functions.

Theorem 6

Let the constraint functions g_i satisfy **slater's** condition: there exists $\bar{x}$ such that $g_i(\bar{x}) < 0$ for all $i \in [m]$. Then, exactly one of the following two sets must be empty.

- $S = \{x \in \mathbb{R}^n | f(x) < 0, g_i(x) \le 0, i \in [m]\}$
- $T = \{\lambda \in \mathbb{R}^m | \lambda \ge 0, \ \inf_{x\in\mathbb{R}^n}[f(x) + \sum_{i\in[m]} \lambda_i g_i(x)] \ge 0\}$.

Case 1: If T is non-empty, then S is empty. Suppose S is not empty.

$\Rightarrow \bar{x} \in S \iff f(\bar{x}) < 0, \ g_i(\bar{x}) \le 0 \ \forall i$.

Case 2: If S is empty, then T is non-empty. Can be shown via separating hyperplane theorem, but bit more involved.

Skipped

$f(\bar{x}) + \sum_i \lambda_i g_i(\bar{x}) < 0$ when $\lambda_i \geq 0$

which means

$\inf_{x\in\mathbb{R}^n} \left[f(x) + \sum_i \lambda_i g(x)\right] < 0 \quad \forall \lambda \geq 0$

$\Rightarrow T$ is empty, contradicting initial hypothesis.

Strong Duality Theorem

weak duality:
$d^* \leq f^*$

Consider the following general form of optimization problem:

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g_i(x) \leq 0, i \in [m] := \{1, 2, \ldots, m\}, \end{aligned}$$

where f and g_i are convex functions satisfying Slater's condition.

Theorem 7

x^* is an optimal solution to the above problem if and only if there exists $\lambda^* \geq 0$ such that $\inf_{x\in\mathbb{R}^n}[f(x) + \sum_{i\in[m]} \lambda_i^* g_i(x)] \geq f(x^*)$. $\Rightarrow d^* = f^*$.

Since x^* is an optimal solution, the set

$$S = \{x \in \mathbb{R}^n | f(x) - f(x^*) < 0, g_i(x) \leq 0, i \in [m]\}$$

is infeasible.

It follows from the above theorem that the set

$$T = \{\lambda \in \mathbb{R}^m | \lambda \geq 0, \quad \inf_{x\in\mathbb{R}^n} [f(x) - f(x^*) + \sum_{i\in[m]} \lambda_i g_i(x)] \geq 0\}$$

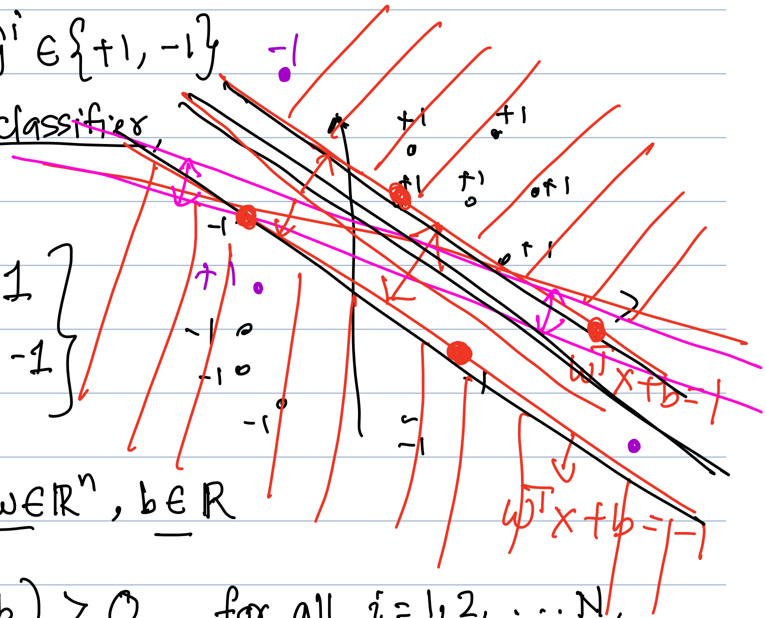
is feasible.

Binary Classification

we are given $(\hat{x}^1, \hat{y}^1), (\hat{x}^2, \hat{y}^2), \ldots (\hat{x}^N, \hat{y}^N)$,

where $\hat{x}^i \in \mathbb{R}^n$, $\hat{y}^i \in \{+1, -1\}$

the goal is to find a linear classifier, $w^T x + b$ such that



$$\begin{cases} w^T \hat{x}^i + b > 0 & \text{whenever } \hat{y}^i = 1 \\ w^T \hat{x}^i + b < 0 & \text{whenever } \hat{y}^i = -1 \end{cases}$$

decision variables: (w, b), $w \in \mathbb{R}^n$, $b \in \mathbb{R}$

constraints:

$$\hat{y}^i (w^T \hat{x}^i + b) > 0 \quad \text{for all } i = 1, 2, \ldots N.$$

$$\Leftrightarrow \hat{y}^i (w^T \hat{x}^i + b) \geq \beta \quad \text{where } \beta > 0, \ \forall i$$

$$\Leftrightarrow \hat{y}^i \left(\left(\frac{w}{\beta}\right)^T \hat{x}^i + \frac{b}{\beta} \right) \geq 1 \quad \forall i$$

$$\Leftrightarrow \boxed{\hat{y}^i (w^T \hat{x}^i + b) \geq 1} \quad \forall i = 1, 2 \ldots N$$

without loss of generality.

$\Updownarrow$

cost function:

$w^T \hat{x}^i + b \geq 1$ for all i with $\hat{y}^i = +1$

$w^T \hat{x}^i + b \leq -1$ for all i with $\hat{y}^i = -1$

For maximum margin classifier, we try to maximize the distance between the two hyperplanes

$$H_1 = \{x \in \mathbb{R}^n \mid w^T x + b = 1\}$$
$$H_2 = \{x \in \mathbb{R}^n \mid w^T x + b = -1\}$$

$$\boxed{\text{dist}(H_1, H_2) = \frac{2}{\|w\|_2}}.$$ (try to prove yourself).

we choose the cost function $\frac{1}{2}\|w\|_2^2$ in order to maximize $\text{dist}(H_1, H_2)$.

optimization problem:

$$\min_{w \in \mathbb{R}^n, b \in \mathbb{R}^n} \quad \frac{1}{2} w^T w$$

$$\text{s.t.} \quad 1 - \hat{y}^i (w^T \hat{x}^i + b) \leq 0, \quad i = 1, 2 \cdots N.$$

$\rightarrow \lambda_i, \ \lambda \in \mathbb{R}^N$

$$L(w, b, \lambda) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i \left[1 - \hat{y}^i (w^T \hat{x}^i + b)\right]$$

$$= \left[\frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i \hat{y}^i w^T \hat{x}^i\right] - \sum_{i=1}^{N} \lambda_i \hat{y}^i b + \sum_{i=1}^{N} \lambda_i$$

Let us find $(\bar{w}, \bar{b})$ that minimize $L(w, b, \lambda)$.

$$\nabla_w L(w, b, \lambda) = w - \sum_{i=1}^{N} \lambda_i \hat{y}^i \hat{x}^i$$

$$\nabla_w^2 L(w, b, \lambda) = I \succeq 0$$

$$\Rightarrow \boxed{\bar{w} = \sum_{i=1}^{N} \lambda_i \hat{y}^i \hat{x}^i} \text{ minimizes } L(w, b, \lambda).$$

dual function: $d(\lambda) = -\infty$ if $\sum_{i=1}^{N} \lambda_i \hat{y}^i \neq 0$

and $d(\lambda) = \frac{1}{2} (\bar{w})^T \bar{w} - \sum_{i=1}^{N} \lambda_i \hat{y}^i (\hat{x}^i)^T (\bar{w}) + \sum_{i=1}^{N} \lambda_i$ when $\sum_{i=1}^{N} \lambda_i \hat{y}^i = 0$

$$\Rightarrow d(\lambda) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i \hat{y}^i \hat{x}^i\right)^T \left(\sum_{j=1}^{N} \lambda_j \hat{y}^j \hat{x}^j\right) - \sum_{i=1}^{N} \lambda_i \hat{y}^i (\hat{x}^i)^T \left(\sum_{j=1}^{N} \lambda_j \hat{y}^j \hat{x}^j\right) + \sum_{i=1}^{N} \lambda_i$$

$$= \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \hat{y}^i \hat{y}^j \underline{(\hat{x}^i)^T \hat{x}^j} - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \hat{y}^i \hat{y}^j (\hat{x}^i)^T(\hat{x}^j) + \sum_{i=1}^{N} \lambda_i$$

$$= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j (\hat{y}^i)(\hat{y}^j)(\hat{x}^i)^T(\hat{x}^j) + \sum_{i=1}^{N} \lambda_i$$

<u>Dual optimization:</u>

$$\boxed{\begin{aligned} &\max_{\lambda \in \mathbb{R}^N} \quad d(\lambda) \\ &\text{s.t.} \quad \lambda \geq 0,\ \sum_{i=1}^{N} \lambda_i \hat{y}^i = 0 \end{aligned}}$$

<u>Recovering optimal solution of the primal from optimal dual solution</u>

Let λ^* be the optimal solution of the dual.

Let (w^*, b^*) be the optimal solution of the primal.

KKT conditions are given by:

i) $\lambda^* \geq 0, \ \sum_{i=1}^{N} \lambda_i^* \hat{y}^i = 0$

ii) $1 - \hat{y}^i\left((w^*)^T \hat{x}^i + b^*\right) \leq 0 \quad , \forall i = 1, 2 .. N$

iii) $\lambda_i \left[1 - \hat{y}^i\left((w^*)^T \hat{x}^i + b^*\right)\right] = 0 \quad , \forall i = 1, 2 .. N$

iv) $\nabla_w L(w^*, b^*, \lambda^*) = 0 \Leftrightarrow \underline{w^* = \sum_{i=1}^{N} \lambda_i^* \hat{y}^i \hat{x}^i}$

Finding w^* is straightforward.

to find b^*, first find an index τ s.t. $\lambda^*_\tau \neq 0$

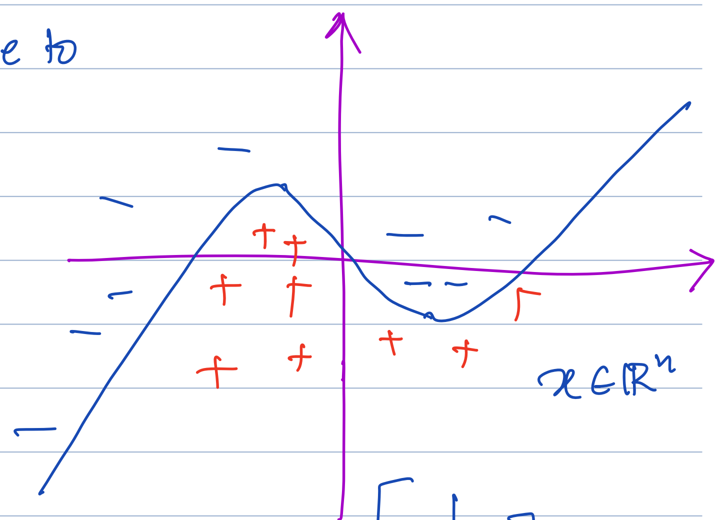
then
$$1 - \hat{y}^\tau\left[(w^*)^T \hat{x}^\tau + b^*\right] = 0$$

$$\Rightarrow (w^*)^T \hat{x}^\tau + b^* = \frac{1}{\hat{y}^\tau}$$

$$\Rightarrow b^* = \frac{1}{\hat{y}^\tau} - (w^*)^T \hat{x}^\tau$$

If the points are amenable to a linear classifier, we can try to define a feature vector
$$\phi : \mathbb{R}^n \to \mathbb{R}^K,$$
$K \gg n$. and try to find (w, b) s.t.

$w^T\phi(x) + b$ classifies the points. $w \in \mathbb{R}^K$



$x \in \mathbb{R}^n$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix}, \quad K = 5$$

$w \in \mathbb{R}^5$

If $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $\phi(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{bmatrix}$

We can now formulate the optimization problem

Lecture -18
8th Feb.

$$(P)\begin{cases} \min\limits_{\substack{w\in\mathbb{R}^k,\, b\in\mathbb{R}\\ \varepsilon\in\mathbb{R}^N}} & \frac{1}{2}w^Tw + C\sum_{i=1}^{N}\varepsilon_i \\ \text{s.t.} & 1-\hat{y}^i(w^T\phi(\hat{x}^i)+b)\le\varepsilon_i \;;\quad i=1,2\ldots N \quad : \lambda_{1,i},\ \lambda_1\in\mathbb{R}^N \\ & \varepsilon_i \geq 0,\quad i=1,2\ldots N \\ & -\varepsilon_i\le 0 \quad : \lambda_{2,i}\quad \lambda_2\in\mathbb{R}^N \end{cases}$$

C : large constant that penalizes non-zero ε values.

Homework : derive dual of (P).
from optimal dual solution, recover the optimal primal solution.

$$L(w,b,\varepsilon,\lambda_1,\lambda_2) = \frac{1}{2}w^Tw + C\sum_{i=1}^{N}\varepsilon_i + \sum_{i=1}^{N}\lambda_{1,i}\left(1-\hat{y}^i(w^T\phi(\hat{x}^i)+b)-\varepsilon_i\right) + \sum_{i=1}^{N}\lambda_{2,i}(-\varepsilon_i)$$

$$= \left[\frac{1}{2}w^Tw - \sum_{i=1}^{N}\lambda_{1,i}\hat{y}^i w^T\phi(\hat{x}^i)\right] - \sum_{i=1}^{N}\lambda_{1,i}\hat{y}^i b + \sum_{i=1}^{N}\lambda_{1,i} + \sum_{i=1}^{N}\varepsilon_i\left(C-\lambda_{1,i}-\lambda_{2,i}\right)$$

$$\nabla_w L = w - \sum_{i=1}^{N}\lambda_{1,i}\hat{y}^i\phi(\hat{x}^i) = 0 \Rightarrow \boxed{\bar{w} = \sum_{i=1}^{N}\lambda_{1,i}\hat{y}^i\phi(\hat{x}^i)}$$

to obtain a finite value for $\inf_{\omega, b, \varepsilon} L(\omega, b, \varepsilon, \lambda_1, \lambda_2)$,
we need to set

$$C - \lambda_{1,i} - \lambda_{2,i} = 0 \quad \forall i = 1, 2 .. N$$

$$\sum_{i=1}^{N} \lambda_{1,i} \hat{y}^i = 0$$

The dual optimization problem is given by:

$$\max_{\lambda_1, \lambda_2} d(\lambda_1, \lambda_2)$$

$$\text{s.t.} \quad \lambda_1 \geq 0, \lambda_2 \geq 0$$

$$C - \lambda_{1,i} - \lambda_{2,i} = 0 \quad \forall i = 1, 2 .. N$$

$$\sum_{i=1}^{N} \lambda_{1,i} \hat{y}^i = 0$$

$$C - \lambda_{1,i} = \lambda_{2,i} \geq 0$$

$$d(\lambda_1, \lambda_2) = \frac{1}{2} (\bar{\omega})^T \bar{\omega} - \sum_{i=1}^{N} \lambda_{1,i} \hat{y}^i (\bar{\omega})^T \phi(\hat{x}^i) + \sum_{i=1}^{N} \lambda_{1,i}$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_{1,i} \lambda_{1,j} \hat{y}^i \hat{y}^j \phi(\hat{x}^i)^T \phi(\hat{x}^j) + \sum_{i=1}^{N} \lambda_{1,i}$$

Equivalently:

(dual)

$$\max_{\lambda_1 \in \mathbb{R}^N} d(\lambda_1)$$

$$\text{s.t.} \quad \lambda_1 \geq 0, \quad C - \lambda_{1,i} \geq 0 \quad \forall i,$$

$$\sum_{i=1}^{N} \lambda_{1,i} \hat{y}^i = 0$$

Suppose the optimal dual solution λ_1^* is given. Let us try to recover the optimal primal solution $(\omega^*, b^*, \varepsilon^*)$.

$$\omega^* = \sum_{i=1}^{N} \lambda_{1,i}^* \hat{y}^i \phi(\hat{x}^i)$$

Let us state complementarity slackness conditions.

$$\lambda^*_{1,i}\left(1-\hat{y}^i\left((w^*)^T\phi(\hat{x}^i)+b^*\right)-\varepsilon^*_i\right)=0 \qquad \text{(CS)}$$

$$\varepsilon^*_i\,(\cancel{\lambda^*_{2,i}})\,(C-\lambda^*_{1,i})=0$$

Simply finding an index τ at which $\lambda^*_{1,\tau}>0$ is not enough since if $\lambda^*_{1,\tau}=C$, we can't determine ε^*_i.
Hence, we choose index τ s.t $0<\lambda^*_{1,\tau}<C$,
which implies

$$\varepsilon^*_\tau=0 \quad\text{and}\quad 1-\hat{y}^\tau\left((w^*)^T\phi(\hat{x}^\tau)+b^*\right)-\varepsilon_\tau=0$$

$$\Rightarrow b^*=\frac{1}{\hat{y}^\tau}-(w^*)^T\phi(\hat{x}^\tau).$$

After finding (w^*, b^*), we can find ε^*_i from (CS) conditions.

Suppose we are given a new data point x^{new}.
We can predict its label by evaluating

$$(w^*)^T\phi(x^{new})+b^*.$$

If $\operatorname{sign}\left[\;\cdot\;\right] >0$, then assign label $+1$
<0, " " " -1.

Observe that:

$$b^*+(w^*)^T\phi(x^{new}) = b^*+\left(\sum_{i=1}^N \lambda^*_{1,i}\,\hat{y}^i\,\phi(\hat{x}^i)\right)^T\phi(x^{new})$$

$$=\left[\frac{1}{\hat{y}^\tau}-\sum_{i=1}^N\lambda^*_{1,i}\,\hat{y}^i\,\underbrace{\phi(\hat{x}^i)^T\phi(\hat{x}^\tau)}_{K(\hat{x}^i,\,\hat{x}^\tau)}\right]+\sum_{i=1}^N\lambda^*_{1,i}\,\hat{y}^i\,\underbrace{\phi(\hat{x}^i)^T\phi(x^{new})}_{K(\hat{x}^i,\,x^{new})}$$

– Note that the above expression does not depend on primal decision variables.

- In addition, the dependence on ϕ is in terms of $\phi(x)^T \phi(y)$ for various x, y.

- In practice, it could be challenging to try out different choice of ϕ, and a larger value of K ($\dim \phi(x)$), increases the complexity of the primal optimization problem.

- To avoid the above issues, often Kernel methods are deployed. $K(x_1, x_2)$, e.g.: $K(x_1, x_2) = \exp\left(\frac{\|x_1 - x_2\|^2}{2M}\right)$,

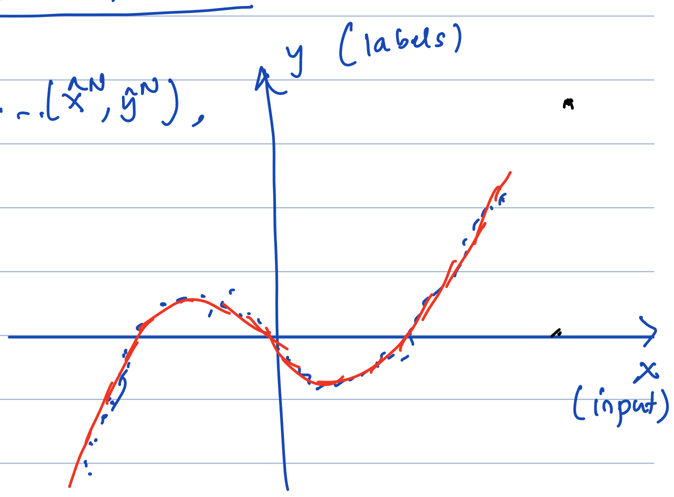
- we replace $\phi(x_1)^T \phi(x_2)$ terms by $K(x_1, x_2)$.

- The dual optimization problem can be written as

$$
\begin{cases}
\max\limits_{\lambda_1 \in \mathbb{R}^N} & -\frac{1}{2} \sum\limits_{i=1}^{N} \sum\limits_{j=1}^{N} \lambda_{1,i} \lambda_{1,j} \hat{y}^i \hat{y}^j K(\hat{x}_i, \hat{x}_j) + \sum\limits_{i=1}^{N} \lambda_{1,i} \\
\text{s.t.} & \lambda_1 \geq 0, \quad \lambda_{1,i} \leq C \quad \forall i \\
& \sum\limits_{i=1}^{N} \lambda_{1,i} \hat{y}_i = 0
\end{cases}
$$

Regression Problems

As before, we are given $(\hat{x}^1, \hat{y}^1) \ldots (\hat{x}^N, \hat{y}^N)$,
$\hat{y}^i \in \mathbb{R}$, $\hat{x}^i \in \mathbb{R}^n$.



we wish to find $w \in \mathbb{R}^K$ s.t.

$$\underline{w^T \phi(\hat{x}^i) \simeq \hat{y}^i} \quad \forall i = 1, 2 \ldots N.$$

$\phi : \mathbb{R}^n \to \mathbb{R}^K$ feature map.

For a fixed w, the residual error for ith data point: $\underline{w^T \phi(\hat{x}^i) - \hat{y}^i}$

residual error vector : $\begin{bmatrix} w^T \phi(\hat{x}^1) - \hat{y}^1 \\ \vdots \\ w^T \phi(\hat{x}^N) - \hat{y}^N \end{bmatrix} \in \mathbb{R}^N = \underline{\phi(\hat{x}) w - \hat{y}}$,

where $\hat{y} \in \mathbb{R}^N$, $(\hat{y})_i = \hat{y}^i$,
$\underline{\phi(\hat{x}) \in \mathbb{R}^{N \times K}}$, $\phi(\hat{x}^i)$ is the i-th row of $\phi(\hat{x})$ matrix.

<u>Regression problem</u> : $\min_{w \in \mathbb{R}^K} \| \phi(\hat{x}) w - \hat{y} \|_{2/1/\infty}$

<u>Least squares problem</u> : $\min_{w \in \mathbb{R}^K} \| \underline{\phi(\hat{x})} w - \hat{y} \|_2^2$

$$= (\phi(\hat{x}) w - \hat{y})^T (\phi(\hat{x}) w - \hat{y})$$
$$= w^T \phi(\hat{x})^T \phi(\hat{x}) w - 2 (\hat{y})^T \phi(\hat{x}) w + (\hat{y})^T \hat{y}$$

the above problem is convex

$\Uparrow$

Hessian: $\underline{\phi(\hat{x})^T \phi(\hat{x}) \succeq 0}$

If there are constraints on $w \in W$, then the problem remains a convex optimization problem if W is a convex set.

Now, let us tackle the Regression problem with respect to the ∞-norm. Recall $\|x\|_\infty = \underset{1 \le i \le n}{\max} |x_i|$

$$\min_{w \in \mathbb{R}^K} \left[\| \phi(\hat{x}) w - \hat{y} \|_\infty \right] = \min_{w \in \mathbb{R}^K} \left[\max_{1 \le i \le N} \left| \phi(\hat{x}^i)^T w - \hat{y}^i \right| \right]$$

$$= \min_{\substack{w \in \mathbb{R}^K \\ t \in \mathbb{R}}} t$$

$$\text{s.t} \quad \max_{1 \le i \le N} \left| \phi(\hat{x}^i)^T w - \hat{y}^i \right| = t$$

equivalently:

$$\min_{\begin{bmatrix} w \in \mathbb{R}^K \\ t \in \mathbb{R} \end{bmatrix}} t$$

$$\text{s.t} \quad \left| \phi(\hat{x}^i)^T w - \hat{y}^i \right| \le t, \quad \forall i = 1, 2 \dots N$$

$$= \left[\begin{array}{ll} \min_{\substack{w \in \mathbb{R}^K \\ t \in \mathbb{R}}} & t \\ \text{s.t} & \phi(\hat{x}^i)^T w - \hat{y}^i \le t \quad \forall i = 1, 2 \dots N \\ & -\phi(\hat{x}^i)^T w + \hat{y}^i \le t \quad \forall i = 1, 2 \dots N. \end{array} \right.$$

is a Linear programming problem.

$$\|x\|_1 = \sum_{i=1}^{N} |x_i|$$

Regression problem w.r.t 1-norm

$$\min_{w \in \mathbb{R}^K} \| \phi(\hat{x}) w - \hat{y} \|_1 = \min_{w \in \mathbb{R}^K} \sum_{i=1}^{N} \underbrace{| \phi(\hat{x}^i)^T w - \hat{y}^i |}_{t_i}$$

$$= \min_{\substack{w \in \mathbb{R}^K \\ t \in \mathbb{R}^N}} \sum_{i=1}^{N} t_i$$

$$\text{s.t.} \quad | \phi(\hat{x}^i)^T w - \hat{y}^i | = t_i \qquad \forall\, i = 1, 2 \ldots N$$

equivalently:

$$\min_{\begin{bmatrix} w \in \mathbb{R}^K \\ t \in \mathbb{R}^N \end{bmatrix}} \sum_{i=1}^{N} t_i$$

$$\text{s.t.} \quad \phi(\hat{x}^i)^T w - \hat{y}^i \le t_i$$

$$-\phi(\hat{x}^i)^T w + \hat{y}^i \le t_i$$

$$t_i \geq 0$$

which is a Lineare programming problem.

Class Test on Wednesday during class hours.

↳ Syllabus: all topics covered so far.