LECTURE 6

OPTIMIZATION THEORY
LAGRANGIANS
DUALITY

HOW APPLIED TO SUPPORT VECTOR MACHINES

MORE ON KERNELS
NO CONSTRAINTS:

GIVEN A FUNCTION $f$ DEFINED ON A DOMAIN $\mathcal{D} \subset \mathbb{R}^n$.

MINIMIZE $f(w)$ $w \in \mathcal{D}$

NECESSARY CONDITION FOR MINIMUM

$$\frac{\partial f(w)}{\partial w} = 0$$

SUFFICIENT CONDITION FOR MINIMUM

$$\frac{\partial f(w)}{\partial w} = 0 \quad \text{AND} \quad \frac{\partial^2 f(w)}{\partial w^2} \quad \text{POSITIVE SEMI-DEFINITE}$$

STRICT MIN. IF POSITIVE DEFINITE

A SYMMETRIC POS. DEFINITE IF

1. SYMMETRIC
2. A UNIT VECTORS $u$, $u^T A u \geq 0$

OR 2'. ALL EIGENVALUES $\geq 0$
EXAMPLE RIDGE REGRESSION (LEAST SQUARES)

\[ f(w) = \frac{1}{2} \| w \|_2^2 + \lambda \sum_{t=1}^{L} \left( \bar{w} \cdot \bar{x}_t - y_t \right)^2 \]

**Regularization**

**Loss**

\( \eta \) is trade-off parameter

**Remark:**
- \( \eta \to \infty \) no regularization only the loss minimized
- \( \eta > 0 \): solution unique & finite
- Non-homogeneous HP is homogeneous HP in one dim. larger

\[ w \cdot x + b = 0 \text{ if } (w, b) \cdot (x, 1) = 0 \]
\[
\frac{\partial f(w)}{\partial w_i} = \frac{1}{\eta} w_i + \sum_{t=1}^{L} x_{t,i} (w \cdot x_t - y_t)
\]

or
\[
\frac{\partial f(w)}{\partial w} = \frac{1}{\eta} w + \sum_{t=1}^{L} x_{t} (x_t^T w - y_t)
\]

\[
= \left( \frac{1}{\eta} I + \sum_{t=1}^{L} x_t x_t^T \right) w - \sum_{t=1}^{L} x_t y_t
\]

\[
\frac{\partial f(w)}{\partial w} \bigg|_{w=w^*} = 0
\]

\[
w^* = \left( \frac{1}{\eta} I + \sum_{t=1}^{L} x_t x_t^T \right)^{-1} \sum_{t=1}^{L} x_t y_t
\]

**IN MATRICES NOTATION:**

\[
f(w) = \frac{1}{2} \frac{1}{\eta} \|w\|^2 + \frac{1}{2} \|X^T w - y \|^2
\]

\[
\text{RESIDUAL}
\]

\[
X = \begin{bmatrix} x_1 & x_2 & \cdots & x_L \end{bmatrix}
\]

\[
X \text{ is } n \times L
\]
\[ \frac{\partial f(w)}{\partial w} = \frac{1}{\eta} w + X (X^T w - y) \]
\[ = \frac{1}{\eta} w + XX^T w - X y \]
\[ = \left( \frac{1}{\eta} I + XX^T \right) w - X y \]

\[ w^* = \left( \frac{1}{\eta} I + XX^T \right)^{-1} X y \]

\[ \frac{\partial^2 f(w)}{\partial (w^T) w} = \left( \frac{1}{\eta} I + XX^T \right) \]

**Strictly Pos. Def. When \( \eta > 0 \) & Finite**

So **Strict Min. At** \( w^* \)

\[ \lim_{\lambda \to 0} (\lambda I + A)^{-1} = A^+ \text{ Pseudo Inverse} \]

\[ w^* = (XX^T)^+ X y \text{ is shortest sol minimizing } f(w) \]

**Second Deriv. Pos Definite**

**Not Necessary Condition for Minima**

\[ f(x) = x^4 \]
\[ f'(x) = 4x^3 \]
\[ f''(x) = 12x^2 \]
EQUALITY CONSTRAINTS

GIVEN FUNCTIONS $f_i, h_i$ (1 ≤ i ≤ m) DEFINED ON A DOMAIN $J \subseteq \mathbb{R}^n$

MINIMIZE $f(w)$, $w \in J$

SUBJECT TO: $h_i(w) = 0$ i = 1, ..., m

f OBJECTIVE FUNCTION
h CONSTRAINTS

LAGRANGIAN:

$L(w, \beta) = f(w) + \sum_{i=1}^{m} \beta_i h_i(w)$

$= f(w) + \beta^T h(w)$

w PRIMAL VARIABLES
\beta LAGRANGIAN OR DUAL VARS.
(ONE PER CONSTRAINT)
Necessary condition

\[ w^* \min \text{ of } f(w) \ \text{subject to } = \text{constraints} \]

Then

\[ \frac{\partial L(w^*, \beta^*)}{\partial w} = 0 \]

\[ \frac{\partial L(w^*, \beta^*)}{\partial \beta} = 0 \]

"Primal constraints"

For some values of \( \beta^* \)

If \( L(w, \beta^*) \) convex in \( w \) then the above condition also sufficient

Example: Find largest vol. box with surface area \( a \) convex

\[ \min : \quad -wuv \]

Subject to: \( wu + uv + vw = c/2 \)

\[ L(u, v, w, \beta) = -wuv + \beta(wu + uv + vw - \frac{c}{2}) \]
\[
\frac{\partial L}{\partial w} = -uv + \beta (u + v) = 0 \\
\frac{\partial L}{\partial u} = -vw + \beta (v + w) = 0 \\
\frac{\partial L}{\partial v} = -wu + \beta (w + u) = 0 \\
\frac{\partial L}{\partial \beta} = wu + uv + vw - \frac{c}{2} = 0
\]

PRIMAL CONSTRAINT

ONLY ONE SOLUTION: \( u = v = w = \sqrt{\frac{c}{2}} \)

\[ \Rightarrow \text{MAX VOL. IS CUBE} \]

MAXIMUM ENTROPY DISTR.

\[
-H(p) = \min \left\{ \sum p_i \ln p_i \right\}
\]

SUBJ. TO: \( \sum p_i = 1 \)

\[
L(p, \beta) = \sum p_i \ln p_i + \beta \left( \sum p_i - 1 \right)
\]

\[
\frac{\partial L}{\partial p_i} = \ln p_i + 1 + \beta = 0
\]

\[
p_i = e^{-1 - \beta}
\]

ALL \( p_i \) IDENTICAL

CONSTRAINT: \( \sum p_i = 1 \)

IMPLIES: \( p_i = \frac{1}{n} \)

PLUG IN

\[
\sum \frac{1}{n} \ln \frac{1}{n} = -\sum \frac{1}{n} \ln m
\]
INEQUALITY CONSTRAINTS

\[
\begin{align*}
\text{MINIMIZE} & \quad f(w) & \quad w \in \mathbb{R} \\
\text{SUBJECT TO} & \quad g_i(w) \leq 0 & \quad 1 \leq i \leq k \\
& \quad h_i(w) = 0 & \quad 1 \leq i \leq m \\
\end{align*}
\]

\{ \text{PRIMAL} \}

LAGRANGIAN:

\[
L(w, \lambda, \beta) = f(w) + \lambda^T g(w) + \beta^T h(w)
\]

\{ \text{PRIMAL VARS OR DUAL VARS} \}

LAGRANGIAN DUAL PROBLEM

\[
\begin{align*}
\text{MAXIMIZE} & \quad \Theta(w, \beta) \\
\text{SUBJECT TO} & \quad \lambda \geq 0
\end{align*}
\]

\{ \text{DUAL} \}

WHERE \( \Theta(w, \beta) = \inf_{w \in \mathbb{R}} L(w, \lambda, \beta) \)

\( \beta \) UNCONSTRAINED
\[
\min \sum_{p_i} \ln p_i
\]
\[
\sum_{p_i} = 1
\]

\[
L(p_i, \beta) = \sum_{p_i} \ln p_i + \beta \left( \sum_{i} p_i - 1 \right)
\]

\[
\frac{\partial L}{\partial p_i} = \ln p_i + 1 + \beta = 0
\]
\[
p_i^* = e^{1-\beta}
\]
OPTIMAL PRIMAL

\[
L(p^*, \beta) = \sum_{i} e^{-1-\beta} (1 - e^{-\beta}) + \beta \left( \sum_{i} e^{-\beta} - 1 \right)
\]
\[
= -n e^{-1-\beta} - \beta
\]

DUAL:

\[
\max \quad -n e^{-1-\beta} - \beta
\]
\[
\beta \quad \Theta(\beta)
\]
\[\text{UNCONST.}\]

SOLVE DUAL FOR OPTIMUM \(\beta^*\)

\[
\frac{\partial \Theta(\beta)}{\partial \beta} = -n e^{-1-\beta} (1) - 1 = 0
\]
\[
e^{-1-\beta^*} = \frac{1}{n}
\]
\[
-1-\beta^* = -\ln n
\]
\[
\beta^* = (\ln n) - 1
\]

\[
\Theta(\beta^*) = -n e^{-1-\beta^*} - \ln n + 1
\]
\[
= -n \frac{1}{n} - \ln n + 1
\]
SAME VALUE AS PRIMAL
Example 5.24 (Quadratic programme) We demonstrate the practical use of duality by applying it to the important special case of a quadratic objective function.

\[
\begin{align*}
\text{minimise} & \quad \frac{1}{2} w^T Q w - k^T w, \\
\text{subject to} & \quad X w \leq c,
\end{align*}
\]

where $Q$ is a positive definite $n \times n$ matrix, $k$ is an $n$-vector; $c$ an $m$-vector, $w$ the unknown, and $X$ an $m \times n$ matrix. Assuming that the feasible region is not empty, this problem can be rewritten as

\[
\max_{x \geq 0} \left( \min_w \left( \frac{1}{2} w^T Q w - k^T w + x^T (X w - c) \right) \right).
\]

The minimum over $w$ is unconstrained, and is attained at $w = Q^{-1}(k - X'x)$. Resubstituting this back in the original problem, one obtains the dual:

\[
\begin{align*}
\text{maximise} & \quad -\frac{1}{2} x^T P x - x^T d - \frac{1}{2} k^T Q k, \\
\text{subject to} & \quad x \geq 0,
\end{align*}
\]

where $P = XQ^{-1}X'$, and $d = c - XQ^{-1}k$. Thus, the dual of a quadratic program is another quadratic programme but with simpler constraints.
AN EXAMPLE WITH INEQUALITIES

MINIMIZE \( \frac{1}{2} w^T Q w - k^T w \)  
SUBJECT TO \( X w \leq c \)  
\( Q \) pos. def.

\( L (w, \alpha) = \frac{1}{2} w^T Q w - k^T w + \alpha^T (Yw - c) \)

\( \frac{\partial L}{\partial w} = Qw - k + X^T \alpha = 0 \)

\( w = Q^{-1}(k - X^T \alpha) \)

SUBSTITUTING INTO L GIVES

\( \frac{1}{2} (Q^{-1}(k - X^T \alpha))^T Q^{-1} (k - X^T \alpha) - k^T Q^{-1} (k - X^T \alpha) \)

\( -\alpha^T (X Q^{-1} (k - X^T \alpha) - c) \)

\( \frac{\alpha}{\epsilon} \)

MAX \( -\frac{1}{2} x^T P x - x^T 01 - \frac{1}{2} k^T Q k \)  
SUBJECT TO \( \alpha \geq 0 \)

DUAL \( \) SIMPLER

WHERE \( P = X Q^{-1} X \)

\( d = c - X Q^{-1} k \)
**Weak Duality Th:**

If \( w \in R \) is a feasible sol. to primal

\( (x, \beta) \quad " \quad \text{dual} \)

Then \( f(w) \geq \theta(x, \beta) \)

\[ \begin{align*}
\begin{array}{c}
\text{minimize} \quad \text{maximize} \\
\end{array}
\end{align*} \]

**Proof:**

\[ \theta(x, \beta) = \inf_{\tilde{w} \in R} L(\tilde{w}, x, \beta) \]

\[ \leq L(w, x, \beta) \]

\[ \leq f(w) + x^T g(w) + \beta^T h(w) \]

\[ \geq \begin{align*}
&\geq 0 \quad \leq 0 \\
&\text{feasibility of } x \quad \text{feasibility of } w \\
&= 0 \quad \leq f(w)
\end{align*} \]

\( \square \)
CONCLUSION:

VALUE OF DUAL UPPER BOUNDED BY VALUE OF PRIMAL

\[ \inf \{ f(w) : g(w) \leq 0, h(w) = 0 \} \]

\[ \sup \{ \theta(x, \beta) : x \geq 0 \} \]

WHEN THE GAP IS 0

THEN THIS GIVES FEASIBLE SOLUTIONS

\[ f(w^*) = \theta(x^*, \beta^*) \]

\[ x^* \geq 0 \quad g(w^*) \leq 0 \quad h(w^*) = 0 \]

THEN \( w^* \) & \( (x^*, \beta^*) \) SOLVE THE PRIMAL AND DUAL PROBLEM, RESPECTIVELY.

ALSO:

\[ x_i^* g_i(w^*) = 0 \quad \text{for } 1 \leq i \leq k \]

WHY? \((*)\) IS TIGHT IFF THE ABOVE HOLDS
For us, the constraints are always affine:

\[ h_i(w) = (\tilde{\omega}_i^T \tilde{w} - \tilde{a}_i = 0) \]

\[ g_i(w) = (\tilde{c}_i^T w - \tilde{a}_i \leq 0) \]

Also, if always convex.

Strong Duality Theorem:

For a convex domain \( \mathcal{D} \),

- \( g_i, h_i \) affine

Then the following optim. problem has duality gap 0:

\[
\begin{align*}
\min_{w \in \mathcal{D}} & \quad f(w) \\
\text{subj. to:} & \quad g_i(w) \leq 0 & 1 \leq i \leq n \\
& \quad h_i(w) = 0 & 1 \leq i \leq m
\end{align*}
\]
A TH. YOU SHOULD REMEMBER

\[
\begin{align*}
\min & \quad f(w) \quad \text{WHERE} \quad w \in \mathbb{R}^n \\
\text{SUBJ. TO:} \quad & g_i(w) \leq 0 \quad 1 \leq i \leq n \\
& h_i(w) = 0 \quad 1 \leq i \leq m
\end{align*}
\]

\text{S CONVEX, } \quad g_i, h_i \text{ AFFINE}

\text{NEC. & SUFF. CONDITIONS FOR } w^* \text{ TO BE AN OPTIMUM ARE EXISTENCE OF } x^*, \beta^* \text{ S.T.}

\[
\begin{align*}
\frac{\partial}{\partial w} L(w^*, x^*, \beta^*) &= 0 \\
& h_i(w^*) = 0 \quad 1 \leq i \leq m \\
g_i(w^*) \leq 0 \quad 1 \leq i \leq n
\end{align*}
\]

\[
\alpha_i^* g_i(w^*) = 0 \quad 1 \leq i \leq n
\]

\[
\alpha_i^* \geq 0 \quad \text{KARUSH-KUHN-TUCKER CONDITIONS}
\]

\text{CONST. } g_i(w^*) \leq 0 \quad \text{ACTIVE IF } g_i(w^*) = 0 \quad \text{INACTIVE}

\text{ACTIVE } \Rightarrow \alpha_i^* \geq 0

\text{INACTIVE } \Rightarrow \alpha_i^* = 0
SUPPORT VECTOR MACHINES

SEPARATE DATA INTO 2 CLASSES
BASED ON A HYPERPLANE

\[ w \cdot x + b \geq 0 \]
\[ \forall x : w \cdot x + b = 0^3 = \frac{1}{2} (x, 1) : (w, b), (x, 1) - 0^3 \]

\[ w \cdot x + b \leq 0 \]

IF \( b \geq 0 \) THEN ORIGIN ON \( \geq \) SIDE

WHAT IS DISTANCE TO ORIGIN ?

INTERSECTION BETWEEN LINE \( \lambda \overline{w} \) \& PLANE:

\[ w \cdot (\lambda w) + b = 0 \]
\[ C = \frac{-b}{\|w\|^2} \]

\[ \|C\| = \|C\| \|w\| \]
\[ = \frac{|C|}{\|w\|} \]

DISTANCE THE SIGNED OF ORIGIN TO \( \forall x : w \cdot x + b = 0^3 \)

\[ -\frac{1}{\|w\|} \] WHEN \( b = 0 \) THEN DISTANCE 0
NOTE THAT \[ \vec{w} \cdot \vec{x} + b = 0 \]

\[ \Rightarrow q \vec{w} \cdot \vec{x} + q b = 0 \quad \text{FOR SCALAR } q \neq 0 \]

\[ \Rightarrow \frac{\vec{w}}{||w||} \cdot \vec{x} + \frac{b}{||w||} = 0 \quad \text{SAME PLANE} \]

UNIT

DOT PRODUCT: \[ \vec{w} \cdot \vec{x} = ||w|| ||x|| \cos \theta \]

\[ \cos \theta = \frac{||\vec{p}||}{||x||} \]

\[ ||\vec{p}|| = ||x|| \cos \theta \]

\[ ||\vec{p}|| = \frac{\vec{w} \cdot \vec{x}}{||w||} \]
\[
\frac{|\mathbf{w} \cdot \mathbf{x} + b|}{\|\mathbf{w}\|} \quad \text{DISTANCE OF } \mathbf{x} \text{ TO PLANE}
\]

\[
\frac{\mathbf{w} \cdot \mathbf{x} + b}{\|\mathbf{w}\|} \quad \text{SIGNED DISTANCE OF PLANE TO } \mathbf{x}
\]

\[
\text{WANT} \quad w \cdot x_i + b > 0 \quad \text{IF } y_i = +1 \\
< 0 \quad \text{IF } y_i = -1 \\
\]

\[
\text{WANT} \quad y_i (w \cdot x_i + b) > 0 \\
\]

\[\delta_i\]
FUNCTIONAL MARGIN OF EXAMPLE \((x_i, y_i)\) WRT \((w_1, b)\) IS \(\delta_i = y_i (w^T x_i + b)\)

GEOMETRIC MARGIN

\[
\delta_i \quad \text{subject to} \quad \|w\|_H
\]

GEOMETRIC MARGINS ARE SCALED DISTANCES

IF \(\|w\|_H = 1\) \(\Rightarrow\) NO NORMAL, NECC

WANT PLANE S.T. ALL EXAMPLES HAVE POSITIVE MARGIN
Planes \((w, b)\) and \((\lambda w, \lambda b)\) for \(\lambda > 0\) are the same.

Better goal:

Want maximum margin hyperplane geometrical margin of \(x_i\) invariant to scaling

\[
\frac{w \cdot x_i + b}{\|w\|_2}
\]
HOW TO DEFINE MAX. MARGIN HYPERPLANE

\[
\max_{w,b,\xi} \quad \delta \quad \text{FUNCTIONAL}
\]

\[
\text{SUBJECT TO} \quad y_i (w \cdot x_i + b) \geq \delta
\]

\[
1 \leq i \leq l
\]

\[
= \infty
\]

MAX. MARGIN HYPERPLANE

\[
\max_{w,b,\xi} \quad \delta \quad \text{GEOMETRIC}
\]

\[
\text{SUBJECT TO} \quad y_i \left( \frac{w}{||w||} \cdot x_i + \frac{b}{||w||} \right) \geq \delta
\]

\[
1 \leq i \leq l
\]

GOOD DEF. BUT HARD TO OPTIMIZE
Hyperplanes with functional margin 1 are known as canonical hyperplanes.

\[(x^+, 1) \, \text{pos example} \quad \} \quad \text{with functional margin 1} \]
\[(x^-, -1) \, \text{neg} \quad \]

i.e.
\[
w \cdot x^+ + b = +1
\]
\[
w \cdot x^- + b = -1
\]

Geometric margin of \(x^+\) and \(x^-
\]
\[
\delta = \frac{1}{2} \left( \frac{w}{\|w\|_2} \cdot x^+ + \frac{b}{\|w\|_2} - \frac{w}{\|w\|_2} \cdot x^- - \frac{b}{\|w\|_2} \right)
\]
\[
= \frac{1}{\|w\|_2} \left( w \cdot x^+ + 1 - \left( w \cdot x^- + 1 \right) \right)
\]
\[
= \frac{1}{\|w\|_2}
\]

**EQUIVALENT OPTIMIZATION PROBLEM**

\[
\text{MIN}_{w, b} \frac{1}{2} \|w\|_2^2
\]

**SUBJECT TO** \( y_i (w \cdot x_i + b) \geq 1 \)

**REALIZES MAXIMUM MARGIN HYPERPLANE WITH GEOMETRIC MARGIN \( \frac{1}{\|w\|_2} \)**
DUAL OPTIMIZATION PROBLEM?

\[ L(w, b, \alpha) = \frac{1}{2} (\mathbf{w} \cdot \mathbf{w}) + \sum_{i=1}^{l} \alpha_i \left[ y_i \left( (\mathbf{w} \cdot \mathbf{x}_i) + b \right) - 1 \right] \]

\[ \frac{\partial L}{\partial w} = w + \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \]

\[ w = -\sum_{i} y_i \alpha_i \mathbf{x}_i \]

KEY:
- ONLY DOT PRODUCTS MATTER
- KERNELIZABLE
DUAL PROBLEM

MAXIMISE \[ W(\mathbf{x}) = \sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i^* x_j \]

SUBJECT TO \[ \sum_i y_i \alpha_i = 0 \]
\[ \alpha_i \geq 0 \]

IF \[ \mathbf{x}^* \] IS SOLUTION TO DUAL THEN \[ W^* = \sum_i y_i \alpha_i^* \mathbf{x}_i \]
\[ \mathbf{w}^* = \frac{\max \left( \mathbf{w}^* \cdot \mathbf{y}_i + \min \left( \mathbf{w}^* \cdot \mathbf{x}_i \right) \right)}{2} \]

GEOMETRIC MARGIN
\[ \frac{1}{\|\mathbf{w}^*\|_2} \]

KUesktop Kuhn-Tucker Condition
\[ \alpha_i^* \left[ y_i \left( \mathbf{w}^* \cdot \mathbf{x}_i + b^* \right) - 1 \right] = 0 \]

\[ \text{IF MARGIN} \quad 0 \text{ JEFF MARGIN 1} \]
\[ \text{MARGIN} \quad > 0 \text{ JEFF MARGIN 1} \]
\[ \text{}> 1 \]

SUPPORT VECTORS:
ALL \[ \mathbf{x}_i \] FOR WHICH \[ \alpha_i > 0 \]
REPRESENTING MAX. MARGIN HYPERPLANE IN DUAL

\[ w^* \cdot x + b^* = \sum_{i=1}^{l} y_i x_i x + b^* \]

\[ = \sum_{i \in S^U} y_i x_i x + b^* \]

FOR \( j \in S^U \):

\[ w^* \cdot x_j + b^* = y_j \]

ALSO:

\[ w^* \cdot w^* = \sum_{i,j} y_i y_j x_i x_j \]

\[ = \sum_{j \in S^U} x_j y_j \sum_{i \in S^U} y_i x_i x_j \]

\[ = \sum_{j \in S^U} x_j (1 - y_j b^*) \]

\[ = \sum_{j \in S^U} x_j - \sum_{j \in S^U} x_j y_j b^* \]

\[ = 0 \]

MARGIN: \( \frac{1}{\|w^*\|_2} = \frac{1}{\sqrt{\sum_{j} x_j^2}} \)
KERNEL TRICK

$\mathbf{x}_i$ EXPANDED TO $\phi(\mathbf{x}_i)$

$\phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) = K(\mathbf{x}_i, \mathbf{x}_j)$

\[ \begin{align*}
\mathbf{w}^* \cdot \phi(\mathbf{x}) + b^* &= \sum_{i \in SV} y_i x_i^* \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b^* \\
&= \sum_{i \in SV} y_i x_i^* K(\mathbf{x}_i, \mathbf{x}) + b^*
\end{align*} \]

ONE WEIGHT PER FEATURE

\[ \begin{align*}
|SV| \text{ USUALLY } \ll \text{ DIMENSION OF FEATURE SPACE}
\end{align*} \]

GENERALIZATION ERROR GOOD

MARGIN LARGE

OR $|SV|$ SMALL
NON-SEPARABLE CASE

\[ \text{MINIMIZE } \mathbf{w}^\top \mathbf{w} + C \sum_{i} \xi_i \]

\[ \text{SUBJECT TO } y_i (\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i \]

\[ \xi_i \geq 0 \]

ONE SLACK VARIABLE PER EXAMPLE

COST IS LINEAR IN SLACK

ALTERNATE DEF.

\[ \text{MINIMIZE } \mathbf{w}^\top \mathbf{w} + C \sum_{i} \left(1 - y_i (\mathbf{w}_i^\top \mathbf{x}_i + b)\right)_+ \]

\[
\begin{align*}
\text{DUAL} & \\
\text{MAXIMIZE: } & \sum_{i} \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j \mathbf{x}_i \mathbf{x}_j \\
\text{SUBJECT TO: } & \sum_{i} y_i \alpha_i = 0 \\
& \alpha_i \in [0, C] \\
& \uparrow \text{NEW} \\
\end{align*}
\]

BEFORE \( C = \infty \)
QUADRATIC HINGE

\[ \text{MINIMIZE: } w^T w + C \sum \xi_i^2 \]
\[ \text{SUBJECT TO } y_i (w^T x_i + b) \geq 1 - \xi_i \]
\[ \xi_i \geq 0 \]

ALTERNATE FORMULATION:
\[ \text{MINIMIZE } w^T w + C \sum \left(1 - y_i (w^T x_i + b)\right)^2 \]

DUAL
\[ \text{MAXIMIZE } \sum \alpha_i - \frac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j x_i \cdot x_j \]
\[ - \frac{1}{2C} \sum \alpha_i \]
\[ \text{SUBJECT TO } \sum y_i \alpha_i = 0 \]
\[ \alpha_i \geq 0 \]
6.2 Support Vector Regression

Figure 6.6: The linear $\varepsilon$-insensitive loss for zero and non-zero $\varepsilon$

Figure 6.7: The quadratic $\varepsilon$-insensitive loss for zero and non-zero $\varepsilon$
KERNEL METHODS

WHEN APPLICABLE?

HYPOTHESIS MUST BE DEFINED I.T.O. DOT PRODUCT

\[ \hat{y} = h_w(x) = f(w \cdot x) \]

PREDICTION ON NEW X

WHERE W IS LINEAR COMBINATION OF INSTANCES IN TRAINING SET

\[ \text{i.e.: } w = \sum_{q=1}^{t} \alpha_q x_q \]

Called Linear Form

EXAMPLES:

- Widrow Hoff
- Perceptron
- Backprop
- Support Vector Machines

\{ GD FAMILY \}
W IN LINEAR FORM WHEN
- PARAMETER DIVERGENCE IS \( \| w \|_2^2 \)
- LOSS DEPENDS ON DOT PRODUCT

\[ w_{t+1} = \text{ARGINF}_w \left( \| w - w_t \|_2^2 + \eta \sum_y L_{yt} (w \cdot x_t) \right) \]

\[
\Rightarrow w_{t+1} = \sum_{q=1}^t x_{q,t} x_q \quad \text{LINEAR FORM} \]

GD FAMILY

MAIN OTHER FAMILY

\[ w_{t+1} = \text{ARGINF}_w \left( \sum_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i + \eta \sum_y L_{yt} (w \cdot x_t) \right) \quad w \geq 0 \]

\[
\ln w_{t+1,i} = \sum_{q=1}^t x_{q,t} x_{q,i} \quad \text{EXPONENTIAL FORM} \]

NOW...
EGU FAMILY
Assume prediction $\hat{y}_t$ is linear ($\hat{y}_t = w_t \cdot x_t$) or linear threshold function ($\hat{y}_t = (w^T x_t > \theta)$).

Linear form good because after expanding instances still efficient.

\[\overline{x} = (x_1, \ldots, x_n) \quad \rightarrow \quad \phi(\overline{x}) = (\phi_1(x), \ldots, \phi_N(x))\]

$\phi(x) \in \mathbb{R}$

\[\{\phi(x) : x \in X\}\]

Input space

Feature space

For example:

$\phi(x_1, x_2) = (x_1^2, x_2^2, x_1 x_2)$
REASONS FOR EXPANSION:

- LINEAR MODELS IN FEATURE SPACE ARE CONVENIENT NON-LINEAR MODELS OF INPUT SPACE

- DIMENSIONALITY REDUCTION (I.E. \( N < n \))

\[
\sum_{q=1}^{t-1} \alpha_q^t \phi(x_q) \cdot \phi(x_t)
\]

WEIGHT VECTOR \( w_t \) IN LINEAR FORM

\[
\sum_{q=1}^{t-1} \alpha_q^t \phi(x_q) \cdot \phi(x_t)
\]

COMPUTE DOT PRODUCTS WITH PAST EXAMPLES

OFTEN EFFICIENT COMPUTATION OF \( \phi(x) \cdot \phi(z) \) EVEN WHEN \( N \gg n \)
\( (x_1, x_2, x_3) \rightarrow (1, x_1, x_2, x_3, x_1 x_2, x_1 x_3, x_2 x_3, x_1 x_2 x_3) \)

8 FEATURES

\[ \phi(x) \cdot \phi(z) = 1 + x_1 z_1 + x_2 z_2 + x_3 z_3 + \]

\[ + x_1 x_2 z_1 z_2 + x_1 x_3 z_1 z_3 + x_2 x_3 z_2 z_3 \]

\[ + x_1 x_2 x_3 z_1 z_2 z_3 \]

\[ = (1 + x_1 z_1) \cdot (1 + x_2 z_2) \cdot (1 + x_3 z_3) \]

8 TERMS : \[ Z \times Z \times Z \]
\[ \phi(\mathbf{x}) \]
\[ (x_1, x_2, \ldots, x_n) \]
\[ (1, x_i, x_i x_j, x_i x_j x_k, \ldots) \]
\[ \forall i \neq j \neq \ldots \]
\[ 2^n \text{ MONOMIALS} \]

\[ \phi(x) \cdot \phi(z) = \sum_{\mathbf{I} \subseteq \{1, \ldots, n\}} \prod_{i \in \mathbf{I}} x_i \prod_{i \notin \mathbf{I}} z_i \]
\[ = \prod_{i=1}^{2^n} (1 + x_i z_i) \]
\[ O(2^n) \text{ TIME} \]

\textbf{EFFICIENCY:}
\[ \left( \sum_{q=1}^{2^t} x_q^t \phi(q) \right) \cdot \phi(x_t) \]
\[ \text{DIMENSION } 2^n \]
\[ \text{SEEMINGLY TIME } O(2^n) \]
\[ = \sum_{q=1}^{2^t} x_q^t \phi(q) \cdot \phi(x_t) \]
\[ \text{TIME } O(2^n) \]
\[ \text{TIME } O(t \cdot 2^n) \]
MORE EXAMPLES

\[ (x_1, \ldots, x_n) \mapsto \left( \frac{1}{n} \sum_{i=1}^{n} x_i x_j \right)_{1 \leq i, j \leq n} \]

\[ \phi (\bar{x}) \]

\[ = \phi (\bar{\bar{x}}) \]

\[ = \frac{1}{n} \sum_{i,j} (x_i x_j) (z_i z_j) \]

\[ = \left( \sum_{i=1}^{n} x_i z_i \right)^2 \]

\[ = (x \cdot \bar{z})^2 \]
IMPLICIT MAPPING INTO FEATURE SPACE

A KERNEL IS A FUNCTION $K$ S.T. FOR ALL $x, z$

$$K(x, z) = \phi(x) \cdot \phi(z)$$

\[ \text{implicit} \]

WHERE $\phi$ IS A MAPPING FROM THE INPUT SPACE $X$ TO A FEATURE SPACE $F$

WHICH HAS AN INNER PRODUCT

EXAMPLES SO FAR

$$K(x, z) = \prod_{i=1}^{n} (1 + x_i z_i)$$

$$K(x, z) = (x \cdot z)^d$$

ALSO:

$$K(x, z) = (x \cdot z + c)^d$$

T-TRAINING EXAMPLES

$w = \frac{1}{2} x y y$

HYPOTHESIS REPRESENTED BY $\phi(x)$

$w \cdot x + \text{kernel comp}$

PROPERITIES OF KERNEL FUNCTION:

$$K(x, z) = \phi(x) \cdot \phi(z)$$

$$= \phi(z) \cdot \phi(x)$$

$$= K(z, x)$$

\[ \text{symmetry} \]

$$K(x, z)^2 = (\phi(x) \cdot \phi(z))^2$$

$$\leq ||\phi(x)||^2 ||\phi(z)||^2$$

$$= (\phi(x) \cdot \phi(x)) (\phi(z) \cdot \phi(z))$$

$$= K(x, x) K(z, z)$$

\[ \text{cauchy-schwartz} \]
CHARACTERIZATION OF KERNELS

MERCER'S TH (FINITE CASE):

LET \( X = \{x_1, \ldots, x_n\} \) BE A FINITE INPUT SPACE. THEN \( K(x, z) \) IS A KERNEL FUNCTION IFF

\[
K = \left( K(x_i, x_j) \right)_{i,j=1}^n
\]

IS SYMMETRIC POSITIVE DEFINITE (HAS NON-NEG. EIGENVALUES).

THERE ARE CONTINUOUS VERSIONS OF THIS THEOREM.

WHAT ARE GOOD KERNELS FOR A GIVEN PROBLEM?
Special kernel functions:

\[ k(x, z) = k(x - z) \]

are translation invariant.

Example:

\[ k(n) = \sum_{m=0}^{\infty} a_n \cos(n \pi u) \]

\[ k(x - z) = a_0 + \sum_{n=1}^{\infty} a_n \sin(n \pi x) \sin(n \pi z) \]

\[ + \sum_{n=1}^{\infty} b_n \cos(n \pi x) \cos(n \pi z) \]

\[ \phi_i(x) = (1, \sin(nx), \cos(nx), \sin(2nx), \cos(2nx), \ldots, \sin(nx), \cos(nx)) \]

ORTHOGONAL FEATURES

\[ 2\pi \]

\[ \int_{0}^{2\pi} \sin(x) \cos(3x) = 0 \]

Why? : \[ (R(x - z))^2 = (R(x - z))^T R(x - z) \]

\[ = (x - z)^T R^T R (x - z) \]

\[ = (x - z)^T (x - z) \]

\[ = (x - z)^2 \]
Proposition 3.12 Let $K_1$ and $K_2$ be kernels over $X \times X$, $X \subseteq \mathbb{R}^n$, $a \in \mathbb{R}^+$, $f(\cdot)$ a real-valued function on $X$,

$$\phi : X \rightarrow \mathbb{R}^m$$

with $K_3$ a kernel over $\mathbb{R}^m \times \mathbb{R}^m$, and $K$ a symmetric positive semi-definite $n \times n$ matrix. Then the following functions are kernels:

1. $K(x, z) = K_1(x, z) + K_2(x, z)$,
2. $K(x, z) = aK_1(x, z)$,
3. $K(x, z) = K_1(x, z)K_2(x, z)$,
4. $K(x, z) = f(x)f(z)$,
5. $K(x, z) = K_3(\phi(x), \phi(z))$,
6. $K(x, z) = x'Bz$.

Corollary 3.13 Let $K_1(x, z)$ be a kernel over $X \times X$, $x, z \in X$, and $p(x)$ a polynomial with positive coefficients. Then the following functions are also kernels:

1. $K(x, z) = p(K_1(x, z))$,
2. $K(x, z) = \exp(K(x, z))$,
3. $K(x, z) = \exp(-\|x - z\|^2 / \sigma^2)$.