arg min \frac{1}{2} || \mathbf{w} \|^2 \\
\mathbf{w}, \mathbf{b} \\
subject to \ \mathbf{t}_n (\mathbf{w}^T \mathbf{q}(\mathbf{x}_n) + \mathbf{b}) \geq 1, \ n = 1, \ldots, N

Quadratic Programming with Linear Inequality Constraints

How to analyze and solve such an optimization problem?

Primal Optimization Problem
Given functions \( f, h_n, n = 1, \ldots, N \), defined on a domain \( \mathcal{X} \subseteq \mathbb{R}^d \)

minimize \( f(\mathbf{w}) \), \( \mathbf{w} \in \mathcal{X} \),
subject to \( h_n(\mathbf{w}) = 0, \ n = 1, \ldots, N \)

where \( f(\mathbf{w}) \) is called the objective function, and the equalities regarding \( h_n \) are called equality constraints.

The Lagrangian function is defined as

\[
\mathcal{L}(\mathbf{w}, \lambda) = f(\mathbf{w}) + \sum_{n=1}^{N} \lambda_n h_n(\mathbf{w})
\]

\( \lambda_n \) are called the Lagrange multipliers.

Lagrange Theorem: A necessary condition for a normal point \( \mathbf{w}^* \) to be a minimum of \( f(\mathbf{w}) \) subject to \( h_n(\mathbf{w}) = 0, \ n = 1, \ldots, N \) is

\[
\frac{\partial \mathcal{L}(\mathbf{w}, \lambda^*)}{\partial \mathbf{w}} = 0 \quad \text{and} \quad \frac{\partial \mathcal{L}(\mathbf{w}, \lambda^*)}{\partial \lambda_n} = 0
\]

for some value \( \lambda^* \).

\[\text{(Note: the conditions are sufficient if } f(\mathbf{w}) \text{ is convex)}\]

how about inequality constraints?

minimize \( f(\mathbf{w}) \) \( \mathbf{w} \in \mathcal{X} \)
subject to \( g_n(\mathbf{w}) \leq 0, \ n = 1, \ldots, N \)

the Lagrangian function is \( \mathcal{L}(\mathbf{w}, \lambda) = f(\mathbf{w}) + \sum_{n=1}^{N} \lambda_n g_n(\mathbf{w}) \)

The Lagrangian dual problem of the primal problem is

maximize \( \Theta(\lambda) \)
subject to \( \lambda_n \leq 0, \ n = 1, \ldots, N \)

where \( \Theta(\lambda) = \inf_{\mathbf{w} \in \mathcal{X}} \mathcal{L}(\mathbf{w}, \lambda) \).

Kuhn-Tucker Theorem

Given an optimization problem with convex domain \( \mathcal{X} \subseteq \mathbb{R}^d \)

minimize \( f(\mathbf{w}) \), \( \mathbf{w} \in \mathcal{X} \)
subject to \( g_n(\mathbf{w}) \leq 0, \ n = 1, \ldots, N \)

with \( f \) \( \mathcal{C} \) convex and \( g_n \) affine, the necessary and sufficient conditions for a normal point \( \mathbf{w}^* \) to be an optimum are the existence of \( \lambda_n \) such that

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

(KKT complementary conditions)

\[
\lambda_n^* g_n(\mathbf{w}^*) = 0, \ n = 1, \ldots, N
\]

\( g_n(\mathbf{w}^*) \leq 0, \ n = 1, \ldots, N \)

\( \lambda_n^* \geq 0, \ n = 1, \ldots, N \)
A Simple Example

\[
\begin{align*}
\text{minimize} & \quad (x_1 - 2)^2 + (x_2 - 3)^2 \\
\text{subject to} & \quad x_1 + x_2 - 2 \leq 0 \\
& \quad x_1 - 2 \leq 0 \\
& \quad -x_1 - 2 \leq 0
\end{align*}
\]

\[
L(x, a) = (x_1 - 2)^2 + (x_2 - 3)^2 + a_1 (x_1 + x_2 - 2) + a_2 (x_1 - 2) + a_3 (-x_1 - 2)
\]

KKT:

\[
\begin{align*}
\frac{\partial L}{\partial x_1} &= 2(x_1 - 2) + a_1 + a_2 - a_3 = 0 \\
\frac{\partial L}{\partial x_2} &= 2(x_2 - 3) + a_1 = 0 \\
2x_n - 2 &\leq 0, \; x_1 - 2 \leq 0, \; -x_1 - 2 \leq 0 \\
a_1 (x_1 + x_2 - 2) &= 0 \\
a_2 (x_1 - 2) &= 0 \\
a_3 (-x_1 - 2) &= 0
\end{align*}
\]

Case 1: no constraint is tight

\( x_1 + x_2 - 2 < 0, \; x_1 - 2 < 0, \; -x_1 - 2 < 0 \)

by KKT \( a_1 = a_2 = a_3 = 0 \Rightarrow 2(x_1 - 2) = 0 \) and \( 2(x_2 - 3) = 0 \)

we get \( (x_1, x_2) = (2, 3) \) not a feasible solution \( X \)

Case 2: only \( x_1 - 2 = 0 \) is tight

by KKT \( a_1 = 0 \Rightarrow 2(x_2 - 3) = 0 \Rightarrow (2, 3) \) not a solution.

Case 3: only \( x_1 + x_2 - 2 = 0 \) is tight

by KKT \( a_1 = 0, \; a_3 = 0, \; x_1 + x_2 - 2 = 0 \) (a, w)

\( 2(x_1 - 2) + a_2 = 0, \; x_1 + x_2 = 0, \) and \( 2(x_2 - 3) + a_3 = 0 \)

\( (\frac{1}{2}, \frac{3}{2}) \) is a global solution \( (a_1 = 3) \)

Apply Lagrange multipliers to large margin optimization

Lagrangian function:

\[
L(w, b, a) = \frac{1}{2} ||w||^2 - \sum_{n=1}^{N} a_n \left( w^T \phi(x_n) + b \right) - 1 \]

maximize w.r.t. \( a \), minimize w.r.t. \( w, b \)

\[
\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N} a_n t_n \phi(x_n)
\]

\[
\frac{\partial L}{\partial b} = 0 \Rightarrow b = \sum_{n=1}^{N} a_n t_n
\]

Substitute \( w \) and \( b \) in \( L(w, b, a) \)

(maximize) \( z(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \phi(x_n, x_m) \)

\( a_n \geq 0, \; \sum_{n=1}^{N} a_n t_n = 0 \)

\( k \) is positive definite \( \Rightarrow z(a) \) is bounded below

\[
y(x) = \sum_{n=1}^{N} a_n t_n \phi(x, x_n) + b
\]

KKT \( a_n \geq 0 \)

\[
t_n \left( w^T \phi(x_n) + b \right) - 1 = 0 \Rightarrow t_n y(x_n) - 1 = 0
\]

\[
a_n \left( t_n (w^T \phi(x_n) + b) - 1 \right) = 0 \Rightarrow a_n \left( t_n y(x_n) - 1 \right) = 0
\]

Either \( a_n = 0 \) or \( t_n y(x_n) = 1 \)

\( a_n \neq 0 \) support vectors \( \Rightarrow t_n y(x_n) = 1 \) \( \Rightarrow \) on maximum margin hyperplane
Suppose we have solved for \( \mathbf{a} \), \( \mathbf{x}_n \) satisfies \( t_n y(\mathbf{x}_n) = 1 \), \( \mathbf{x}_n \) is a support vector.

\[
(n_m \mathbf{a}_m) \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) + b = t_n \quad \Rightarrow \quad \mathbf{a}_m \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) + b = t_n
\]

where \( S \) is the index set of support vectors. 

To decide \( b \), multiply the above equation by \( t_n \)

\[
\begin{align*}
\mathbf{a}_m \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) + b &= t_n \\
\Rightarrow \quad \sum_{\mathbf{n} \in S} \mathbf{a}_m \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) + \frac{1}{N_S} \sum_{\mathbf{n} \in S} t_n &= t_n \\
\Rightarrow \quad b &= \frac{1}{N_S} \sum_{\mathbf{n} \in S} t_n \\
\Rightarrow \quad y(\mathbf{x}) = \sum_{\mathbf{n} \in S} \mathbf{a}_m \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) + b
\end{align*}
\]

is decided.

For a new input \( \mathbf{x} \), we can use \( y(\mathbf{x}) \) to predict the class of \( \mathbf{x} \) according to \( \text{sign}(y(\mathbf{x})) \).

So far we assumed the training data are linearly separable. What if the data are not linearly separable?

### Soft Margin Optimization

Slack variables \( \xi_n \geq 0 \) \( n=1, \ldots, N \)

\[
\begin{array}{c}
\xi_n \\
\xi_n = 0 \\
\xi_n = \frac{1}{N_S} \sum_{\mathbf{n} \in S} t_n
\end{array}
\]

The constraints become \( t_n y(\mathbf{x}_n) \geq 1 - \xi_n \) \( n=1, \ldots, N \).

The is called "soft margin" optimization.

Some of the training data are allowed to be misclassified.

\[
\begin{align*}
\text{minimize} & \quad C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \| \mathbf{w} \|^2 \\
\text{subject to} & \quad \xi_n \geq 0, \quad t_n y(\mathbf{x}_n) \geq 1 - \frac{1}{2}, \quad n=1, \ldots, N
\end{align*}
\]

The Lagrangian function is

\[
L(\mathbf{w}, b, \xi, \mathbf{a}, \mu) = \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{n=1}^{N} \xi_n - \sum_{n=1}^{N} \sum_{m=1}^{N} \mathbf{a}_m \mathbf{k}(\mathbf{x}_n, \mathbf{x}_m) t_n y(\mathbf{x}_n) - \sum_{n=1}^{N} \mathbf{a}_m \xi_n - \sum_{n=1}^{N} \mu_n \xi_n
\]

\( \mu_n, \mathbf{a}_m \) are Lagrange multipliers.
$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{n=1}^{N} a_n \phi(x_n)$$
$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} a_n = 0$$
$$\frac{\partial L}{\partial a_n} = 0 \Rightarrow a_n = C - \mu_n$$

KKT

$\begin{align*}
a_n &\geq 0, \quad \mu_n \geq 0, \quad \beta_n \geq 0 \\
t_n y(x_n) - 1 + \beta_n &\geq 0 \\
a_n (t_n y(x_n) - 1 + \beta_n) &\leq 0 \\
\mu_n \beta_n &\leq 0
\end{align*}$

$$Z(a) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m \phi(x_n, x_m)$$

$a_n = C - \mu_n, \quad \mu_n \geq 0 \Rightarrow 0 \leq a_n \leq C$

maximize $Z(a)$
subject to $0 \leq a_n \leq C, \quad \sum_{n=1}^{N} a_n t_n = 0, \quad n = 1, \ldots, N.$

To decide $b$ for those support vectors $x_n$ with $0 < a_n < C$:
$$t_n y(x_n) - 1 = 0$$
$$\sum_{m=1}^{N} a_m t_m \phi(x_n, x_m) + b = 1$$
$$b = \frac{1}{N} \sum_{n=1}^{N} (t_n - \sum_{m=1}^{N} a_m t_m \phi(x_n, x_m))$$

$$M = \{n \mid 0 < a_n < C\}$$

**SLT09 Week 11 第 7 頁**

**Safe margin SVM**

$$a_n = 0 \quad \text{useless}$$
$$a_n > 0 \quad \text{support vectors} \quad t_n y(x_n) = 1 - \beta_n$$
$$\begin{align*}
\sum_{n=1}^{N} a_n &\geq 0 \\
(\sum_{n=1}^{N} a_n) t_n &\geq 0 \\
(\sum_{n=1}^{N} a_n) t_n y(x_n) &\geq \beta_n
\end{align*}$$

$$\text{on margin} \quad a_n = C \Rightarrow \mu_n = 0 \Rightarrow \beta_n \geq 0 \Rightarrow \text{margin error}$$

**\( \nu \)-SVM**

$$\begin{align*}
\text{minimize} & \quad \frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{N} \sum_{n=1}^{N} \beta_n \\
\text{subject to} & \quad t_n (w^T \phi(x_n) + b) \geq \rho - \beta_n \\
& \quad \beta_n \geq 0, \quad \rho > 0
\end{align*}$$

$$\begin{align*}
L(w, \beta, \rho, a_n, \mu_n, \delta) &\equiv \frac{1}{2} \|w\|^2 - \nu \rho + \frac{1}{N} \sum_{n=1}^{N} \beta_n - \sum_{n=1}^{N} \left\{ t_n (w^T \phi(x_n) + b) - \rho - \beta_n \right\} \\
&\quad - \sum_{n=1}^{N} \mu_n \beta_n - \delta \rho \\
&\quad a_n, \mu_n, \delta \geq 0, \quad n = 1, \ldots, N
\end{align*}$$

**SLT09 Week 11 第 8 頁**
\[ \frac{2}{\sum w} = 0 \Rightarrow w = \sum a_n \mathbf{x}_n \neq (x_n) \]
\[ \frac{2}{b} = 0 \Rightarrow a_n + \mathbf{m}_n = \frac{1}{N} \]
\[ \frac{2}{b} = 0 \Rightarrow o = \sum \mathbf{a}_n t_n \]
\[ \frac{2}{b} = 0 \Rightarrow \sum \mathbf{a}_n - \delta = \nu \]

\[ \Gamma(a) = -\frac{1}{2} \sum_{m=1}^{N} \mathbf{a}_n \mathbf{a}_m \mathbf{x}_n \mathbf{x}_m \mathbf{k}(x_n, x_m) \]
subject to
\[ \sum a_n = \frac{1}{N} \]
\[ \sum a_n t_n = 0 \]
\[ \sum a_n \geq \nu \]

By KKT
\[ p \delta = 0, \quad p > 0 \Rightarrow \delta = 0 \Rightarrow \frac{1}{N} \sum a_n = \nu \]

\[ \frac{\text{# of margin errors}}{N} \leq \sum_{\text{the margin errors}} \mathbf{a}_n + \sum_{\text{the remaining} \mathbf{S}_N^+} \]
\[ = \sum_{\mathbf{S}_N^+} \mathbf{a}_n = \nu \]

so there are at most \( \nu \) fraction of training data with \( \mathbf{b} > 0 \)

\[ \frac{\# \text{ of SVs}}{N} \geq \sum_{\mathbf{S}_N^+} a_n = \nu \]

\( \nu \) is the lower bound of the fraction of support vectors to the training data.

To decide \( b \)

by KKT

for \( x_n \) with \( 0 < a_n < \frac{1}{N} \), \( t_n \mathbf{y}(x_n) - \rho = 0 \) (since \( \mathbf{1}_n = 0 \))

\[ t_n \left( \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) + b \right) - \rho = 0 \]

Consider two index sets \( S^+ \), \( S^- \)

\[ S^+ = \{ n \mid 0 < a_n < \frac{1}{N}, t_n = +1 \} \]
\[ S^- = \{ n \mid 0 < a_n < \frac{1}{N}, t_n = -1 \} \]

\[ \left\{ \begin{array}{l}
\sum_{n \in S^+} \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) + b |S^+| - |S^-| | \rho | | S^- | = 0 \\
- \sum_{n \in S^-} \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) - b |S^-| - |S^+| | \rho | | S^+ | = 0 \\
\end{array} \right. \]

Assume \( |S^+| = |S^-| = N_0 \)

\[ \left\{ \begin{array}{l}
b = -\frac{1}{2N_0} \sum_{n \in S^+} \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) \\
\rho = -\frac{1}{2N_0} \left( \sum_{n \in S^+} \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) - \sum_{n \in S^-} \sum_{m=1}^{N} a_m t_m \mathbf{k}(x_n, x_m) \right) \\
\end{array} \right. \]

\[ \mathbf{y}(x) = \sum_{n=1}^{N} a_n t_n \mathbf{k}(x, x_n) + b \]
MULTICLASS SVMs

Fundamentally, SVMs are two-class classifiers. Various approaches to applying SVMs to multiclass classification:

1. one-versus-the-rest (one-against-all)
   - Training: $K$ separate SVMs
   - Prediction: $\hat{y}(x) = \max_k \hat{y}_k(x)$
   - Drawbacks:
     1. $\hat{y}_k(x)$ may have different scales
     2. Imbalanced training set

2. one-versus-one (one-against-one)
   - Requires more training and test time
   - Training: $\frac{K(K-1)}{2}$ SVMs
   - Prediction by majority voting

   A variant for speeding up test time: DagsVM

3. Error-correcting output code (ECOC)

4. Single-class SVM

Figure 1 (a): The decision diagram for finding the best class out of five classes. The organization lists the nodes in a circle and in the order from top to bottom. (b) A diagram of the input phase of a five-class problem. A $K=3$ SVM can only exclude one class from consideration.