6. Support Vector Machines (SVM)

Given a training set $(x_1, y_1), \dots, (x_n, y_n)$

$x_i \in \mathbb{R}^p$: data points

$y_i \in \{-1, +1\}$: class membership (2 classes/groups)

Assume: Exists a separating hyperplane H

s.t. $\{x_i \mid y_i = 1\}$ and $\{x_i \mid y_i = -1\}$

are separated by H. (will be released later)

[Fig 1]

6.1. Hyperplanes and Margins

Representing hyperplanes in $\mathbb{R}^p$:

a) Given $a \in \mathbb{R}^p$

$\{x \in \mathbb{R}^p \mid a^T x = 0\}$ is the $(p-1)$-dim linear space orth. to a.

[Fig 2]

b) Given $a \in \mathbb{R}^p, b \in \mathbb{R}$. Consider

$$\{x \in \mathbb{R}^p \mid a^T x - b = 0\}$$

It holds

$$a^T x - b = 0 \iff a^T x - \frac{a^T a}{\|a\|^2} b = 0$$

$$\iff a^T\left(x - \frac{b}{\|a\|^2} a\right) = 0$$

[Fig 3]

Hence: $\{x \mid a^T x - b \overset{!}{=} 0\}$ is a lin. subspace shifted by $\frac{b}{\|a\|^2} a$, a hyperplane.

Fig. 1

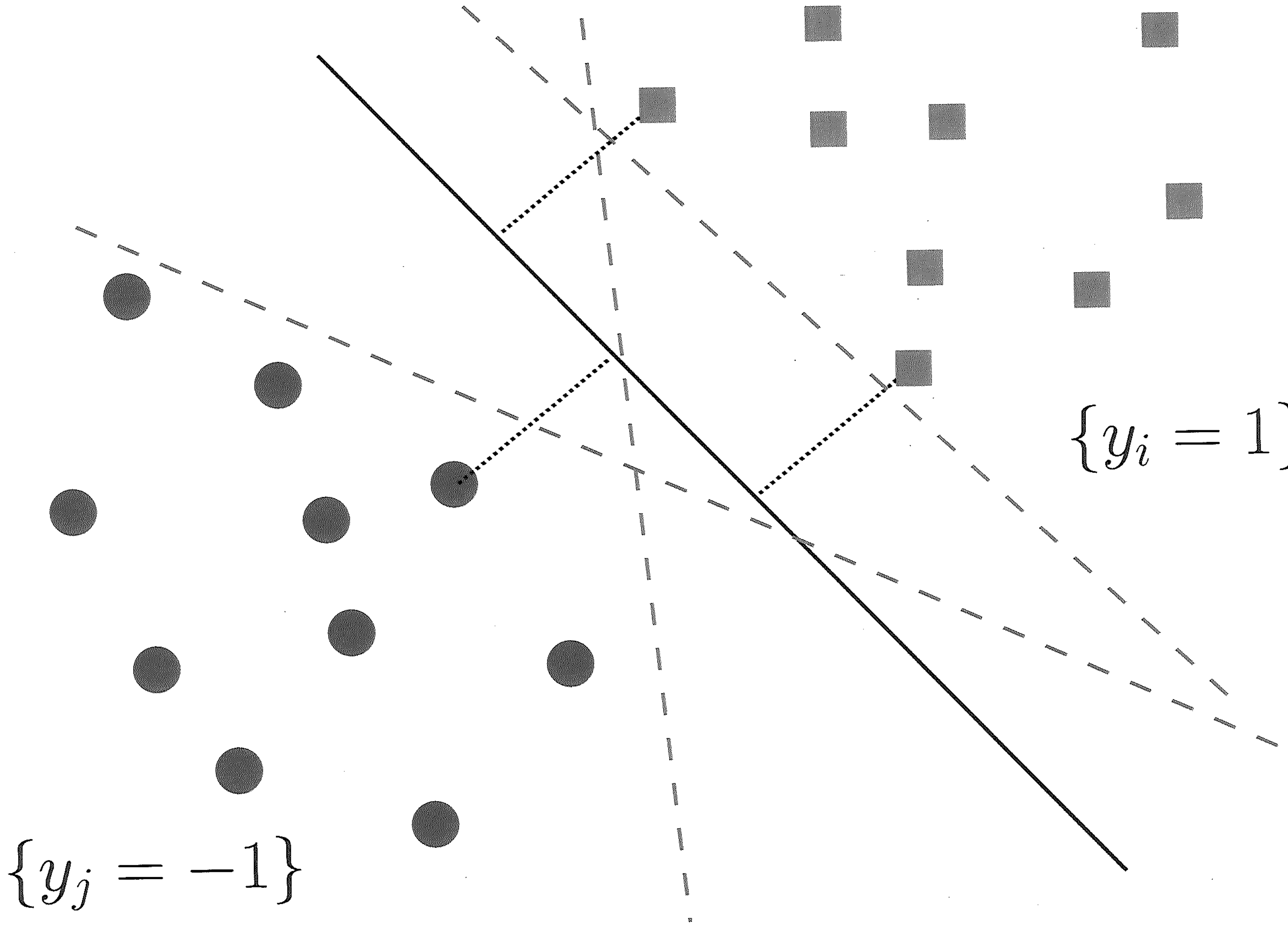


Fig 2

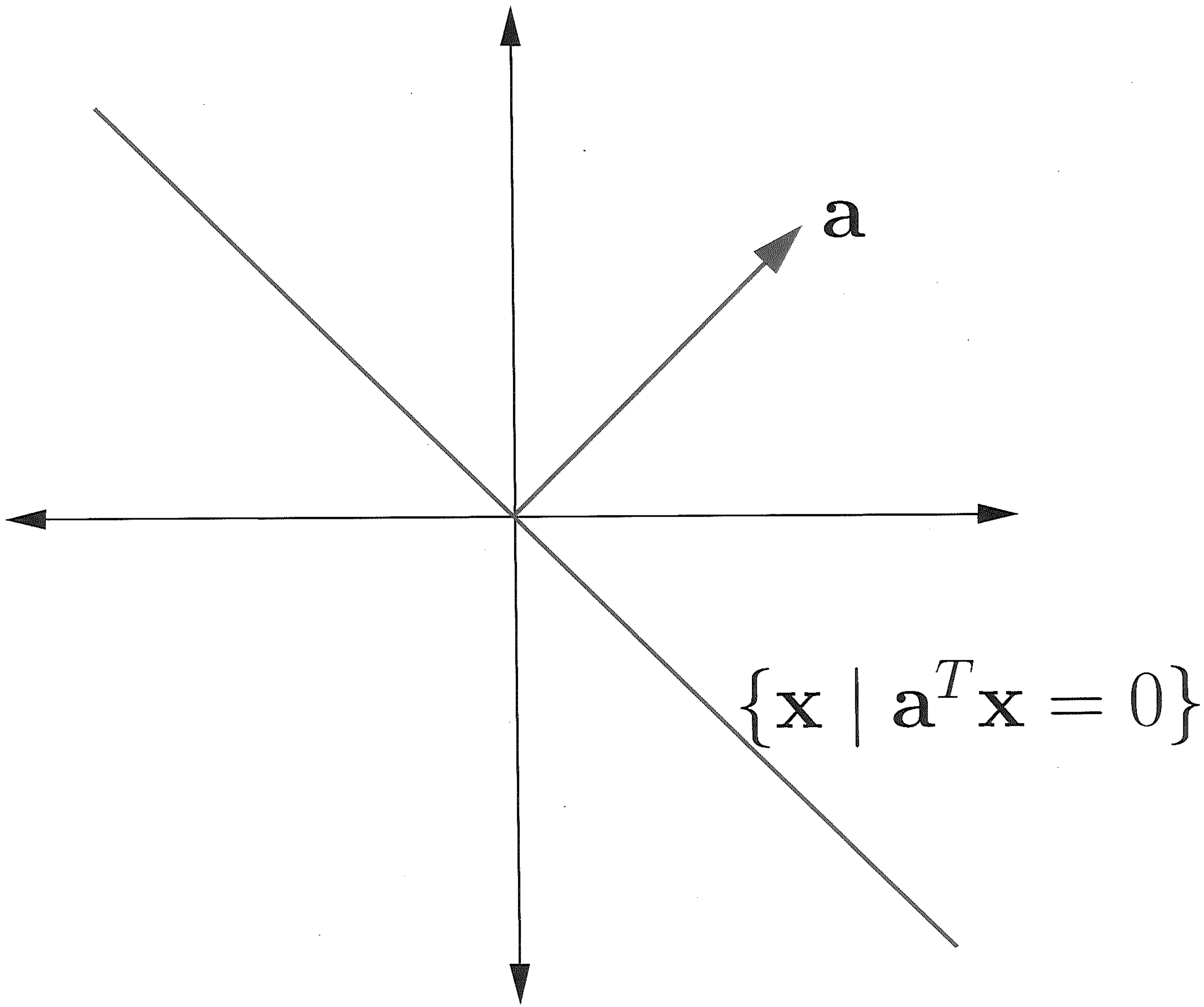
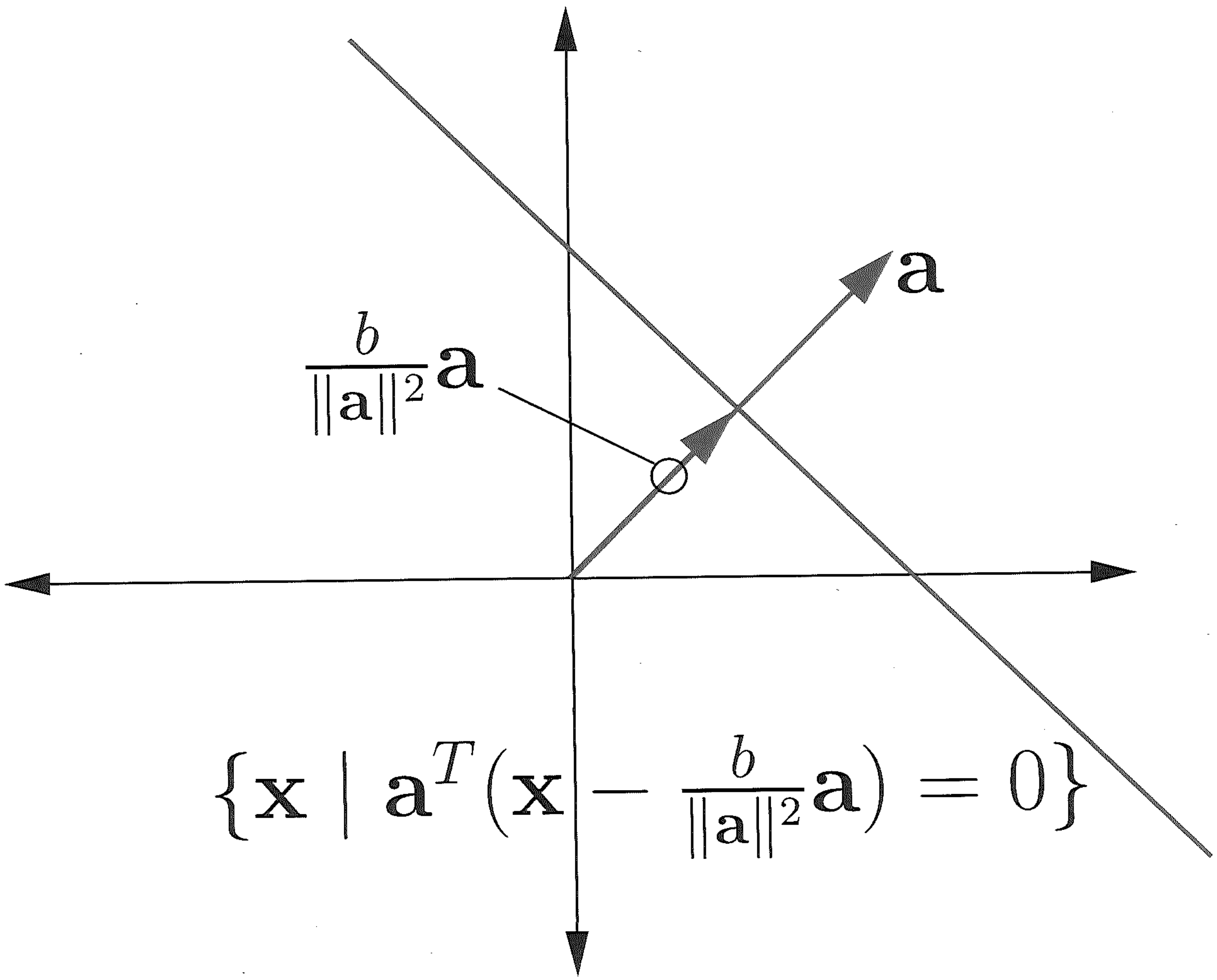
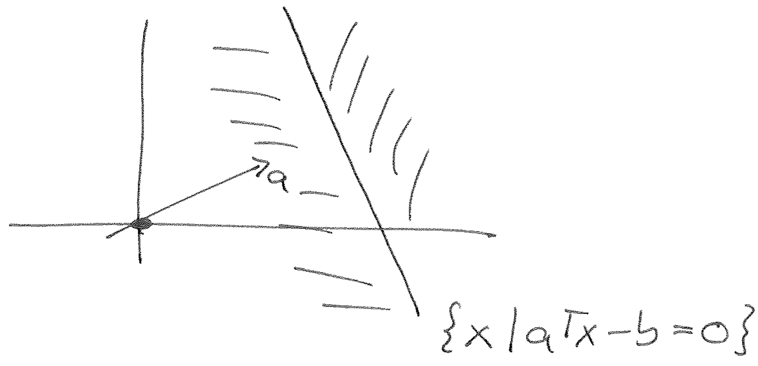


Fig 3



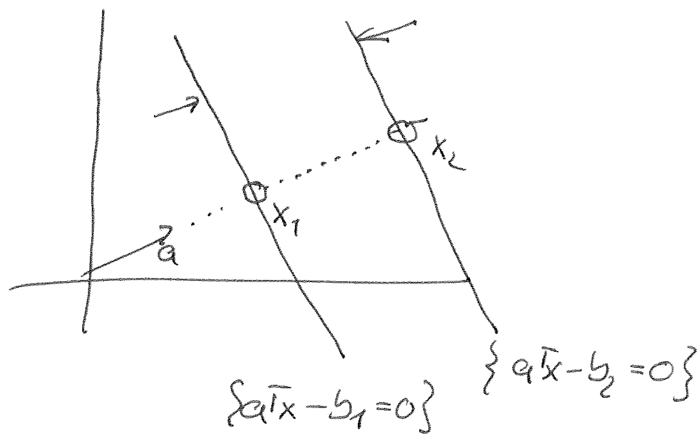
RWTH AACHEN UNIVERSITY

$\{x \in \mathbb{R}^p \mid a^T x \gtrless b\}$ is called half-space :



c) Given $a \in \mathbb{R}^p$, $b_1, b_2 \in \mathbb{R}$

Distance between $H_1 = \{a^T x - b_1 = 0\}$, $H_2 = \{a^T x - b_2 = 0\}$



Both hyperplanes are parallel and orthogonal to a.

Pick x_1, x_2 such that

$x_1 = \lambda_1 a$	$x_2 = \lambda_2 a$
$a^T x_1 - b_1 = 0$	$a^T x_2 - b_2 = 0$

Then

$\lambda_1 a^T a - b_1 = 0$	$\lambda_2 a^T a - b_2 = 0$
$\lambda_1 \lVert a \rVert^2 - b_1 = 0$	$\lambda_2 \lVert a \rVert^2 - b_2 = 0$
$\lambda_1 = \frac{b_1}{\lVert a \rVert^2}$	$\lambda_2 = \frac{b_2}{\lVert a \rVert^2}$

and

$$\|x_2 - x_1\| = \|\lambda_2 a - \lambda_1 a\| = |\lambda_2 - \lambda_1| \, \|a\|$$

$$= \left| \frac{b_2}{\|a\|^2} - \frac{b_1}{\|a\|^2} \right| \|a\| = \frac{1}{\|a\|} |b_2 - b_1|$$

Hence the distance between H_1 and H_2 is $\frac{1}{\|a\|} |b_2 - b_1|$

d) Given $a \in \mathbb{R}^p$, $b \in \mathbb{R}$, $x_0 \in \mathbb{R}^p$

Distance between $H = \{x \mid a^T x - b = 0\}$ and point x_0.

Consider auxiliary hyperplane containing x_0.

$H_0 = \{x \mid a^T x - b_0 = 0\} = \{x \mid a^T x - a^T x_0 \}$

(~~since~~ $b_0 = a^T x_0$ since $a^T x_0 - b_0 = 0$)

By c), the distance between H and H_0 is

$$\frac{1}{\|a\|} |b - a^T x_0| .$$

This distance is called <u>margin</u> of x_0.

6.2 The optimal margin classifier

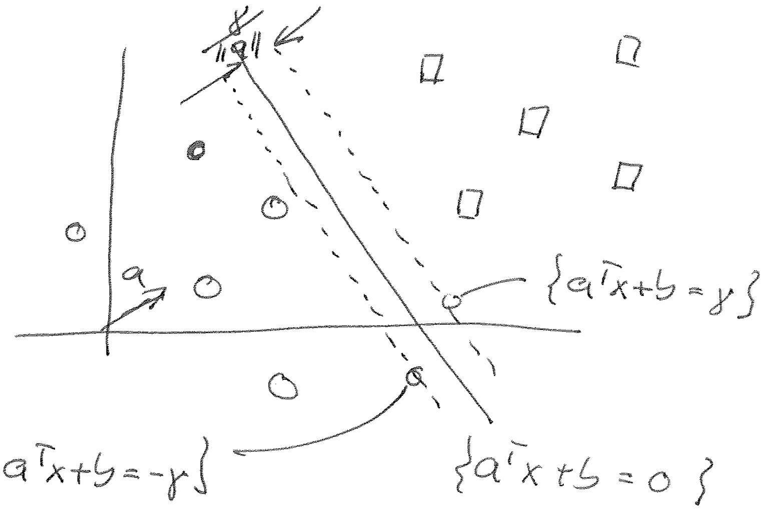
Given a training set $(x_1, y_1), \dots, (x_n, y_n)$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$

Assume there exists a separating hyperplane.

$$\{x \mid a^T x + b = 0\}$$

Then
$$\left.\begin{array}{ll} y_i = +1 & \Rightarrow a^T x_i + b \geq \gamma \\ y_i = -1 & \Rightarrow a^T x_i + b \leq -\gamma \end{array}\right\} \text{ for some } \gamma \geq 0$$

Hence
$$y_i (a^T x_i + b) \geq \gamma \quad \text{for some } \gamma \geq 0 \text{ for all } i = 1, \dots, n$$



Objective: Find a hyperplane $\{x \mid a^T x + b = 0\}$ such that the minimum margin is maximum.

$$\max_{\substack{\gamma \geq 0 \\ a \in \mathbb{R}^p, b \in \mathbb{R}}} \frac{\gamma}{\|a\|} \quad \text{s.t.} \quad y_i (a^T x_i + b) \geq \gamma$$

(not scale invariant)

RWTH AACHEN UNIVERSITY

$$\Leftrightarrow \min_{\gamma, a, b} \frac{\|a\|}{\gamma} \quad \text{s.t.} \quad y_i\left(\frac{a^T}{\gamma} x_i + \frac{b}{\gamma}\right) \geq 1$$

$$\Leftrightarrow \min_{\substack{a \in \mathbb{R}^p \\ b \in \mathbb{R}}} \|a\| \quad \text{s.t.} \quad y_i(a^T x_i + b) \geq 1$$

$$\Leftrightarrow \min_{a \in \mathbb{R}^p, b \in \mathbb{R}} \frac{1}{2}\|a\|^2 \quad \text{s.t.} \quad y_i(a^T x_i + b) \geq 1 \quad i = 1, \dots, n$$

In summary (OMC) (opt. margin classifier):

Given $(x_1, y_1), \dots, (x_n, y_n)$, $x_i \in \mathbb{R}^p$, $y_i \in \{-1, 1\}$

$$\min_{a \in \mathbb{R}^p, b \in \mathbb{R}} \frac{1}{2}\|a\|^2 \quad \text{s.t.} \quad y_i(a^T x_i + b) \geq 1, \; i = 1, \dots, n$$

Quadratic optimization problem with linear constraints, special case of a convex optimization problem.

- Assume a^* is an optimum solution of (OMC) and x_k some point with minimum margin

Then $y_k(a^{*T} x_k + b^*) = 1$

$$\Leftrightarrow (a^{*T} x_k + b^*) = y_k \quad (\text{since } y_k^2 = 1)$$

$$\Leftrightarrow b^* = y_k - a^{*T} x_k$$

Hence, $b^* = y_k - a^{*T} x_k$ is the optimum b-value.

Thinking the Future
Zukunft denken

- The solution (a^*, b^*) is called the <u>optimal margin classifier</u>. [Fig 4]

- Use commercial or public domain software to solve (OMC).

Problem solved? Yes and no!

Consider: • Smarter way to solve (OMC)
• Non-seperability

6.3. SVM and Lagrange Duality

Brief excursion on convex optimization

- Convex optimization problem:

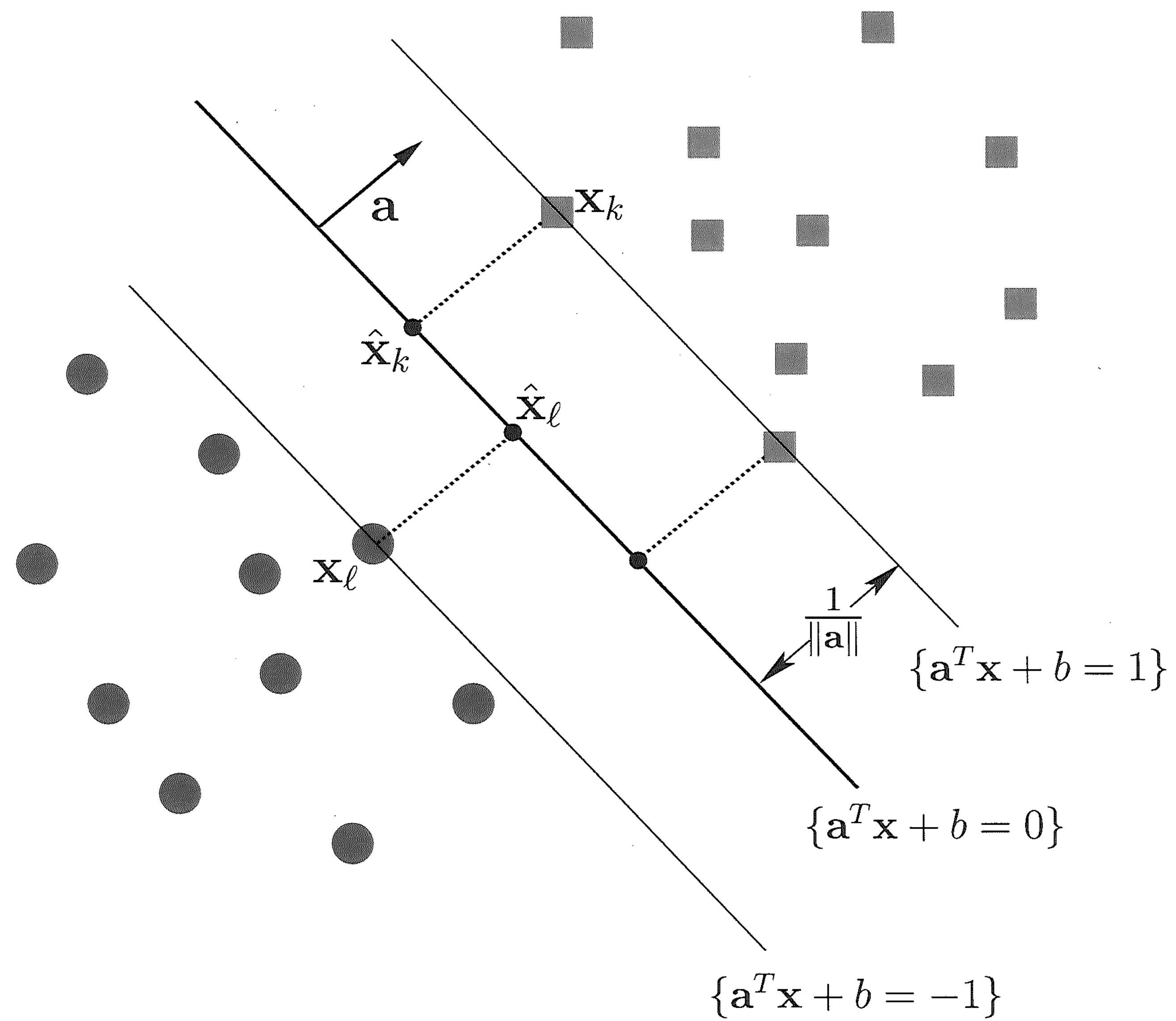
$$(P) \quad \text{minimize } f_0(x)$$
$$\text{s.t.} \quad f_i(x) \leq 0, \ i = 1, \dots, m$$
$$h_i(x) = 0, \ i = 1, \dots, p$$

f_0, f_i are convex, h_i are linear.

- Lagrangian: (prime function)

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

Fig 4



$\mathbf{a}$
$\mathbf{x}_k$
$\hat{\mathbf{x}}_k$
$\hat{\mathbf{x}}_\ell$
$\mathbf{x}_\ell$
$\frac{1}{\|\mathbf{a}\|}$
$\{\mathbf{a}^T\mathbf{x} + b = 1\}$
$\{\mathbf{a}^T\mathbf{x} + b = 0\}$
$\{\mathbf{a}^T\mathbf{x} + b = -1\}$

- Lagrangian dual function

$$g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu)$$

$$\mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom}(f_i) \cap \bigcap_{i=1}^{p} \operatorname{dom}(h_i)$$

- Lagrangian dual problem:

$$(\mathcal{D}) \qquad \max \; g(\lambda, \nu) \quad \text{s.t. } \lambda \geq 0$$

- <u>Weak duality</u> theorem:

$$g(\lambda^*, \nu^*) \leq f_0(x^*)$$

λ^*, ν^* opt. solutions of $(\mathcal{D})$, x^* opt. solution of (P).

- Strong duality:

$$g(\lambda^*, \nu^*) = f_0(x^*)$$

- If the constraints are linear the "Slater's condition" holds, which implies that $g(\lambda^*, \nu^*) = f_0(x^*)$, "strong duality" holds, the "duality gap is $\emptyset$"