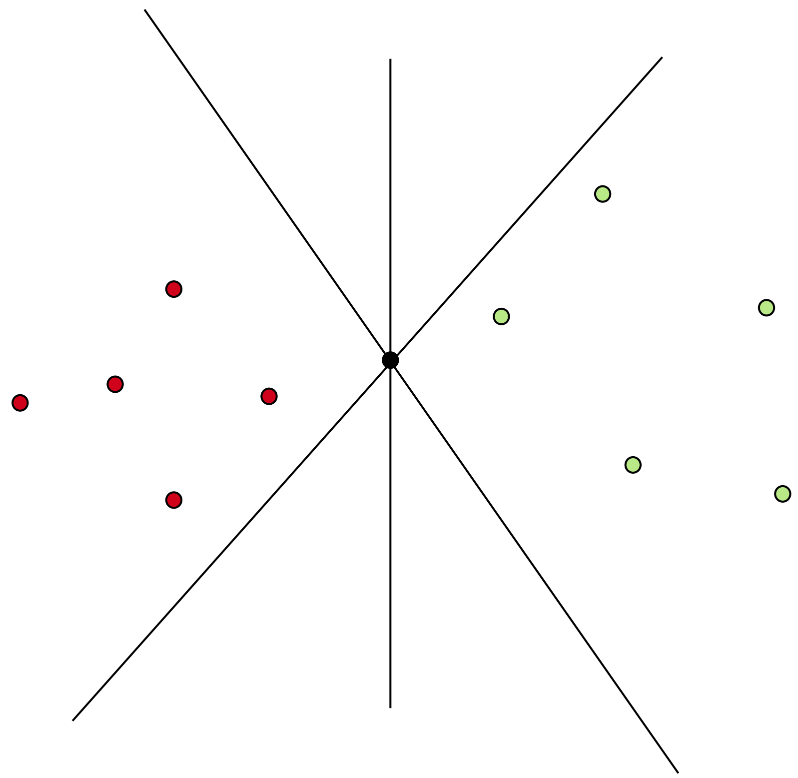
Support Vector Machines

Machine Learning Techniques

Outline

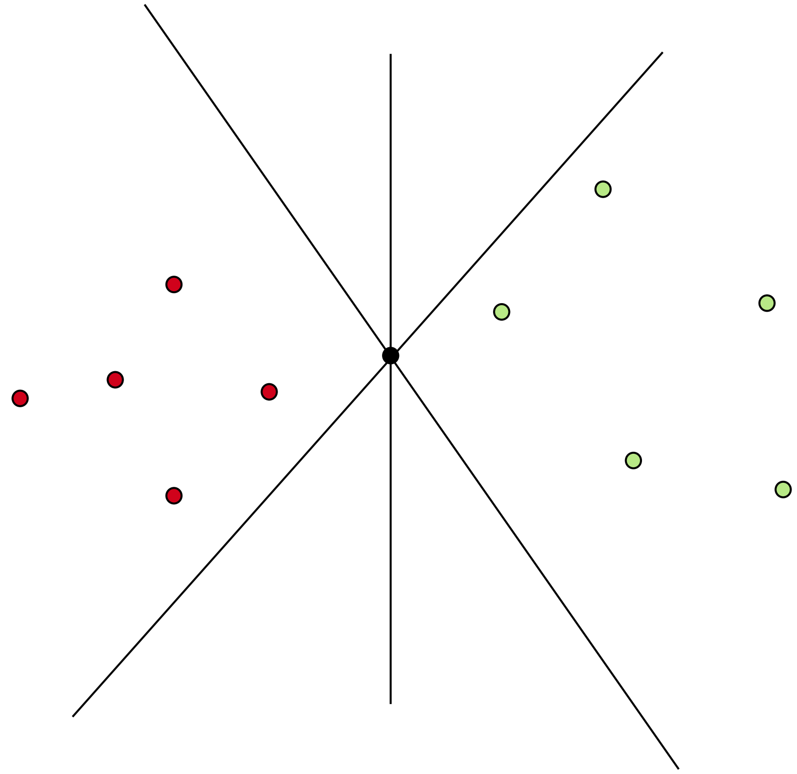
- Margin
- Max-margin classifier
- Duality
- Weight vector
- Support vectors
- Hard-margin Linear-SVM
- Soft-margin SVM

From Perceptrons to SVM



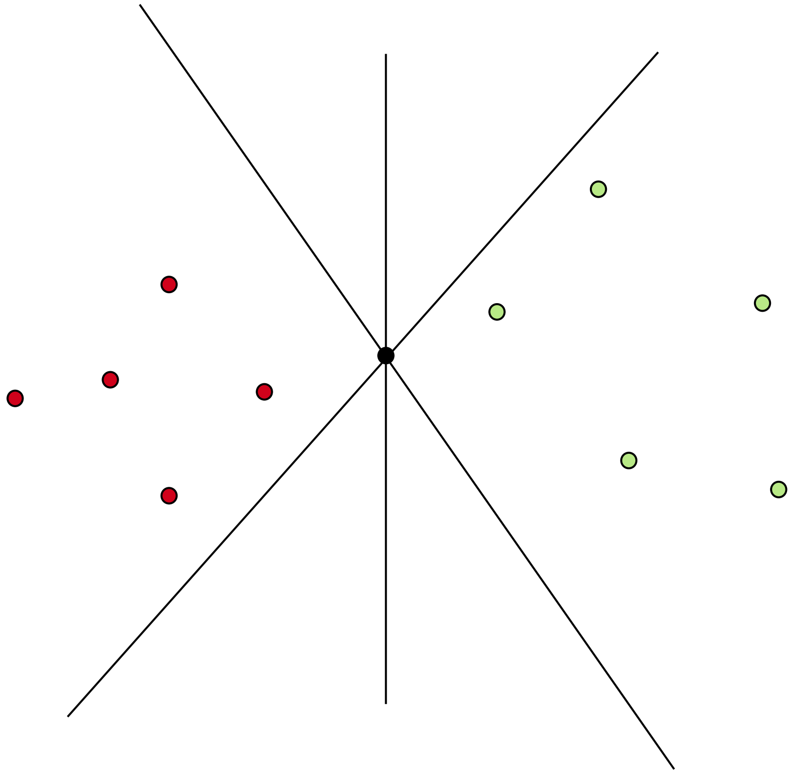
For linearly separable data with γ margin:

From Perceptrons to SVM



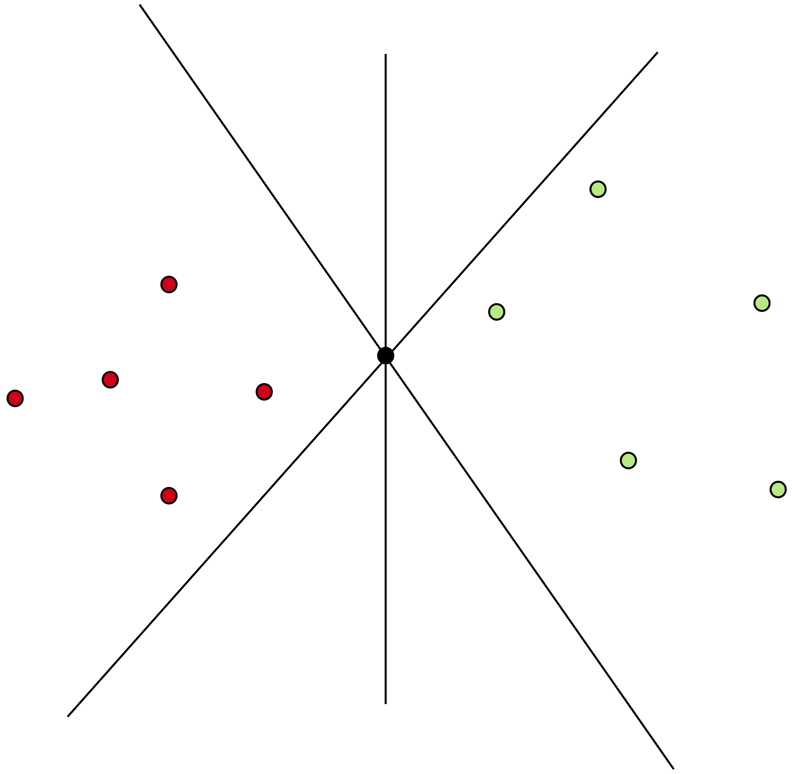
For linearly separable data with γ margin:
(1) Infinite number of valid linear classifiers.

From Perceptrons to SVM



For linearly separable data with γ margin:
(1) Infinite number of valid linear classifiers.
(2) Perceptron returns a valid linear classifier.

From Perceptrons to SVM



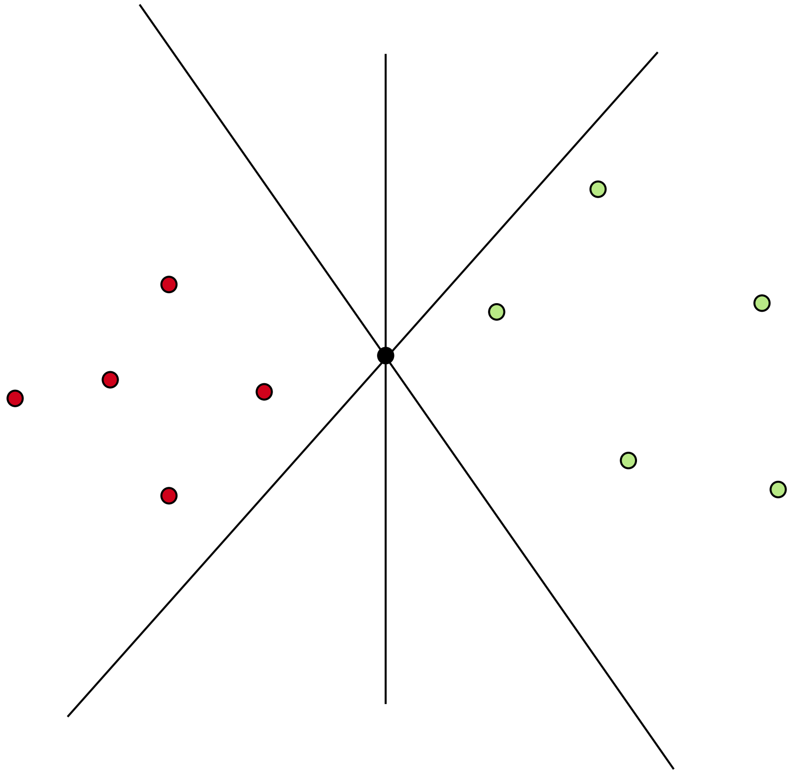
For linearly separable data with γ margin:

(1) Infinite number of valid linear classifiers.

(2) Perceptron returns a valid linear classifier.

(3) Is it the "best"?

From Perceptrons to SVM



For linearly separable data with γ margin:

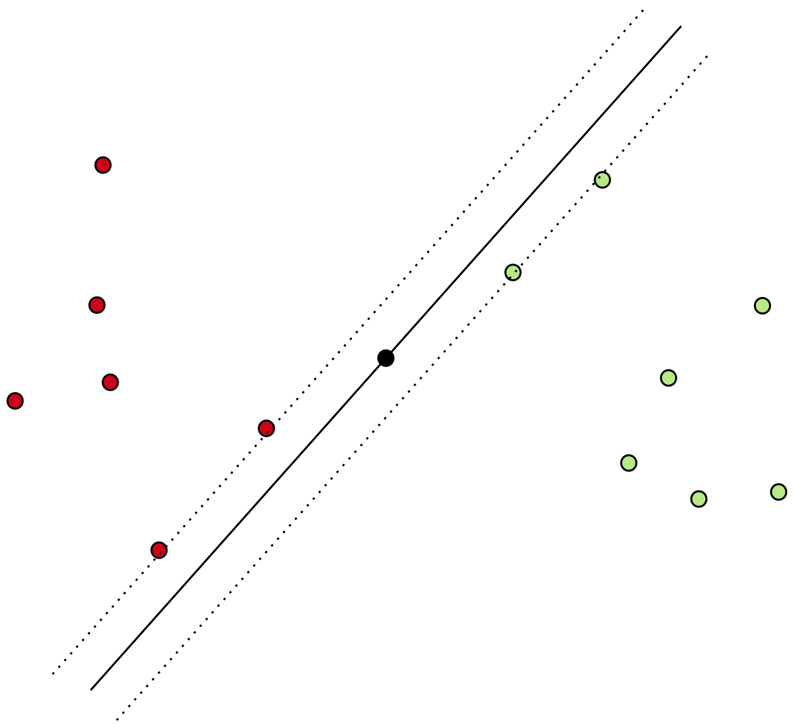
(1) Infinite number of valid linear classifiers.

(2) Perceptron returns a valid linear classifier.

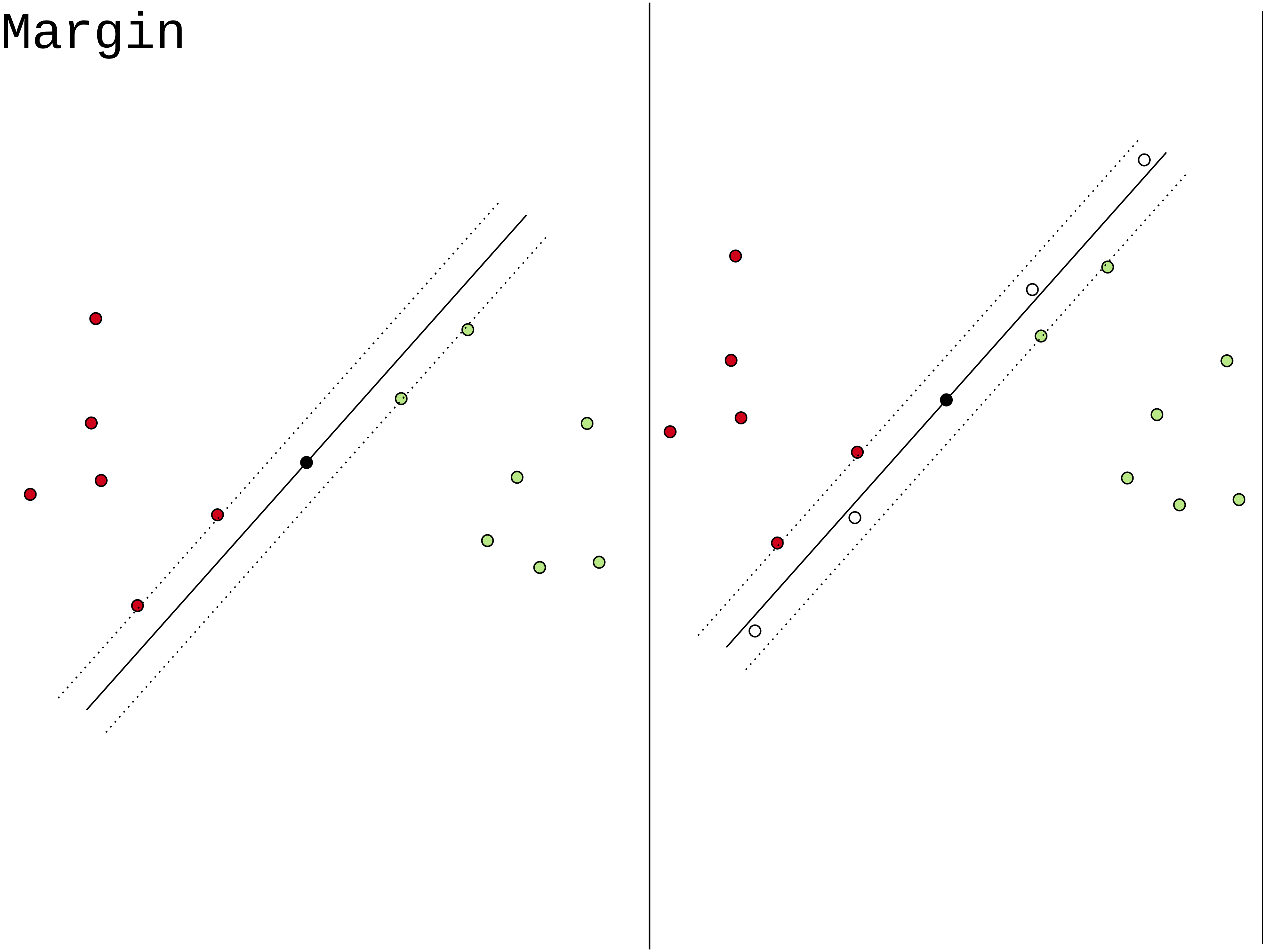
(3) Is it the "best"?

(4) What is a good notion of "best"?

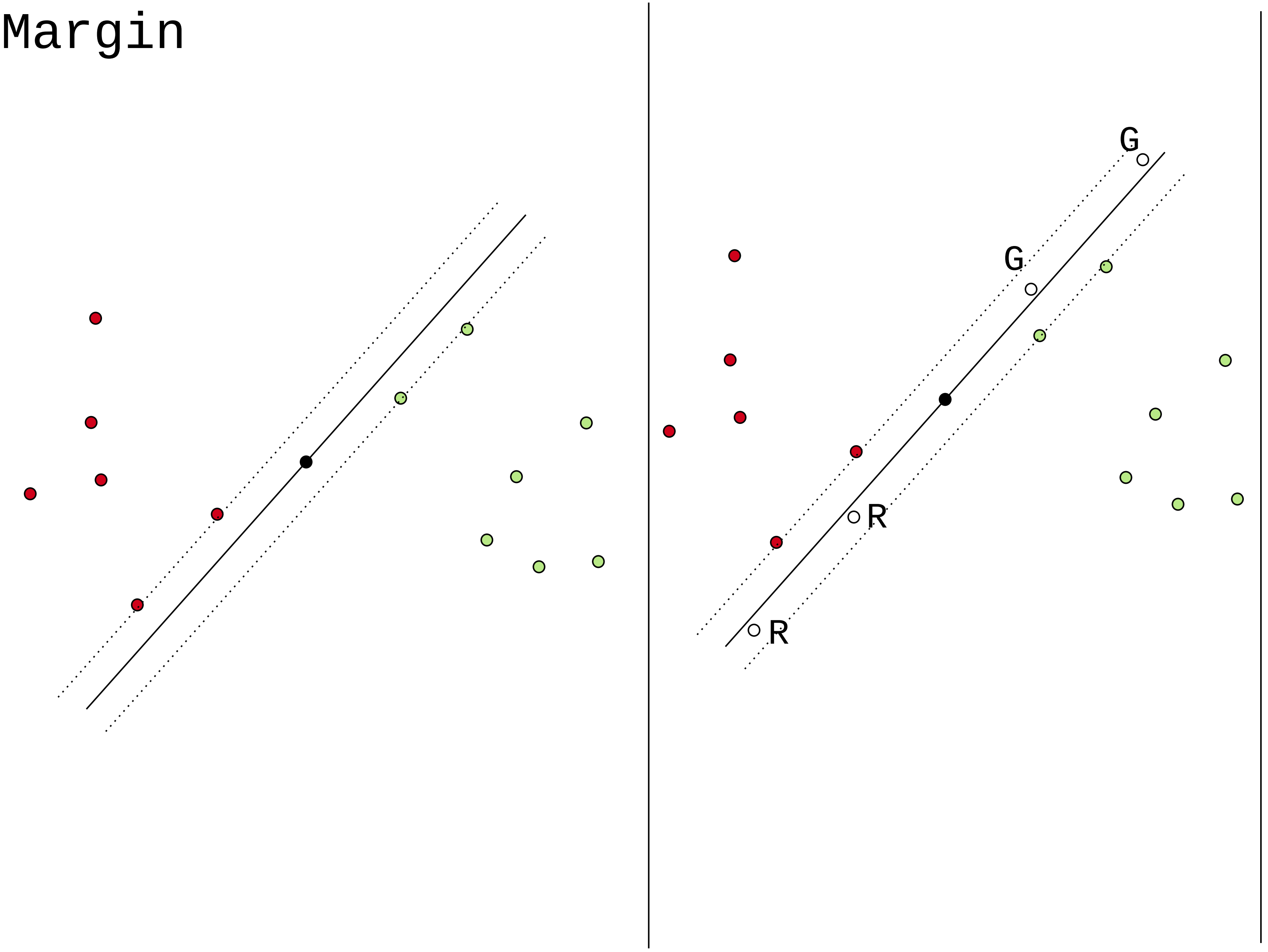
Margin



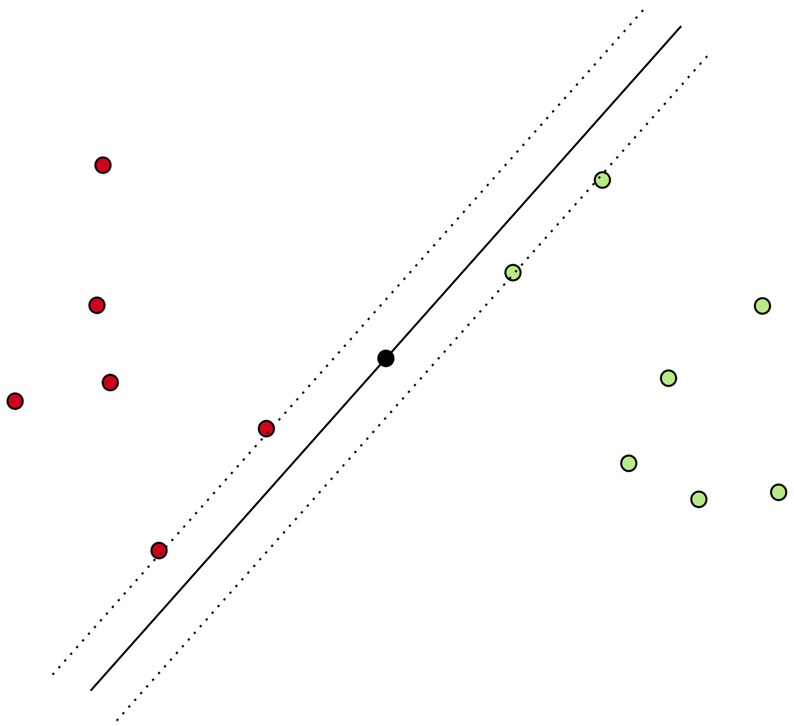
Margin

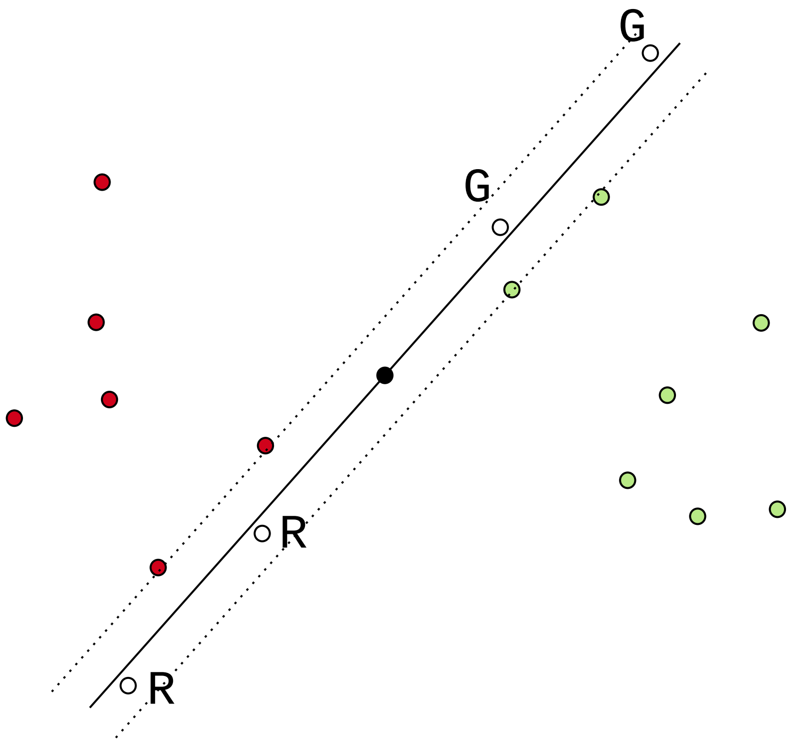


Margin



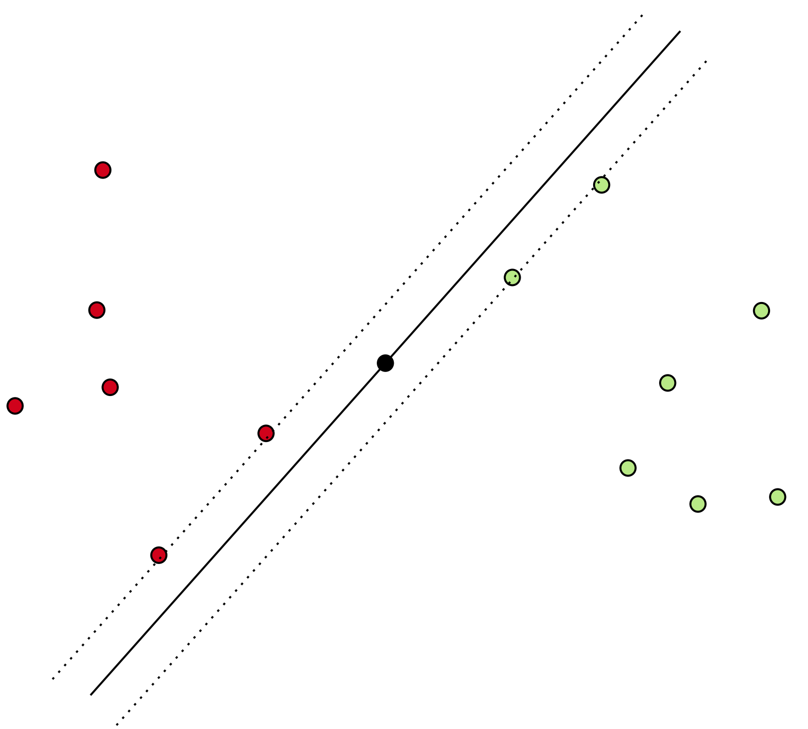
Margin

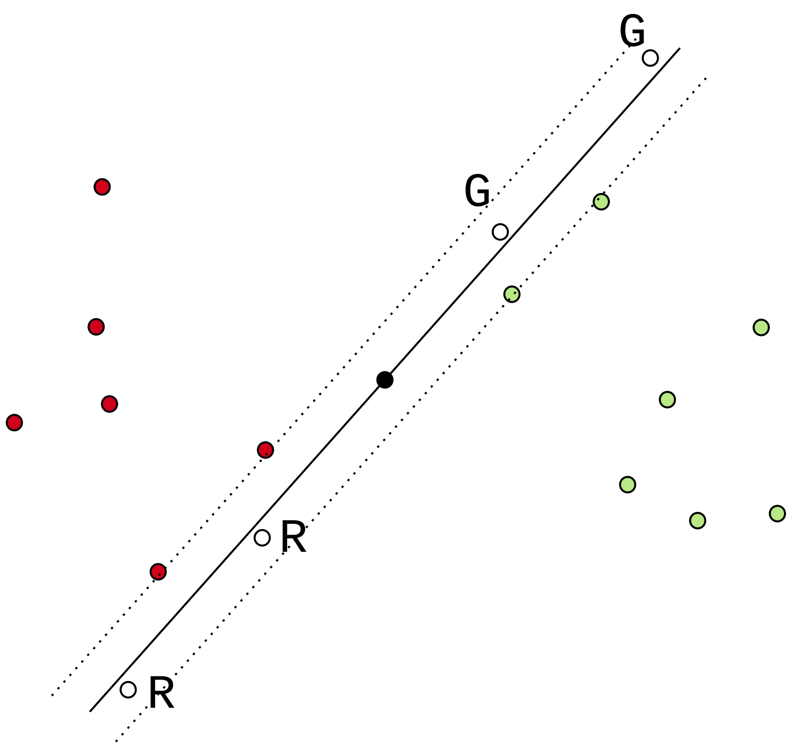




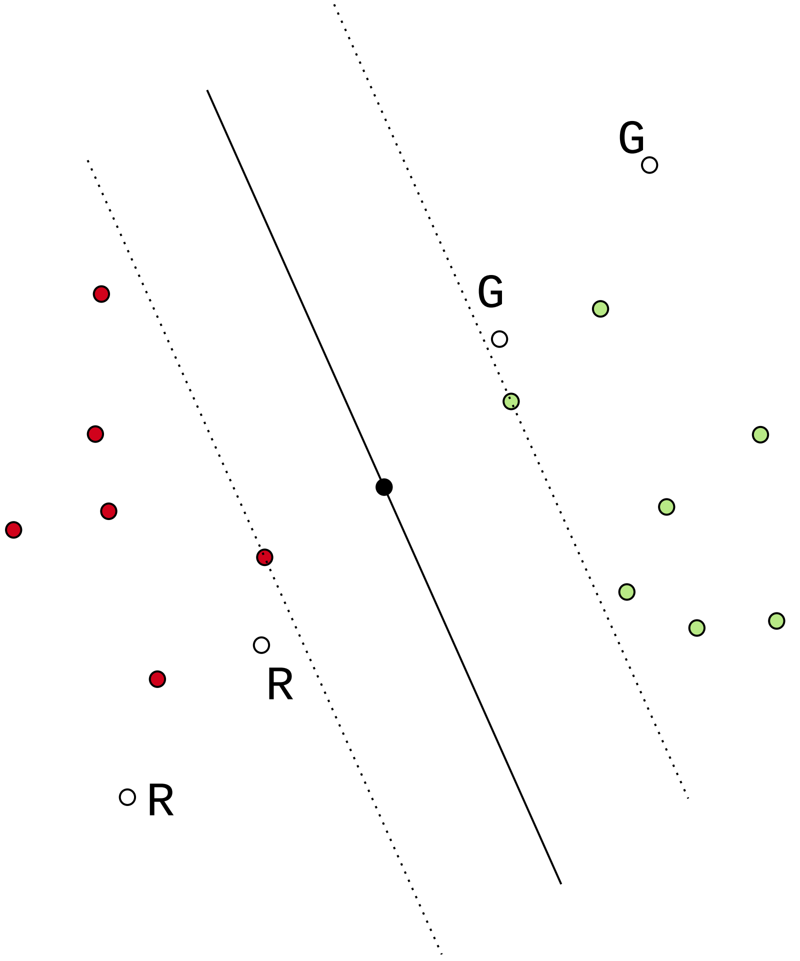
Small margin
Doesn't generalize well

Margin



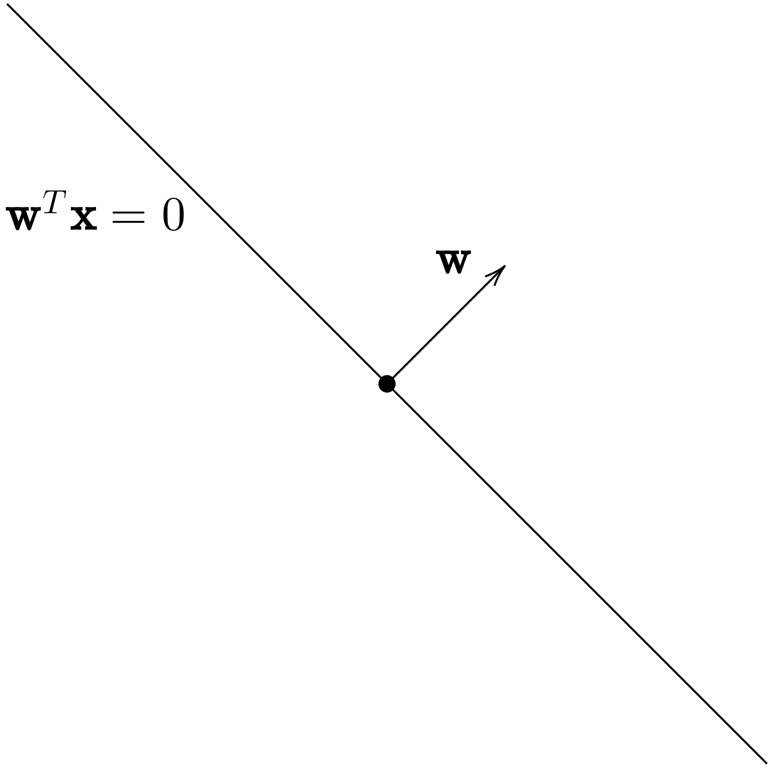


Small margin
Doesn't generalize well



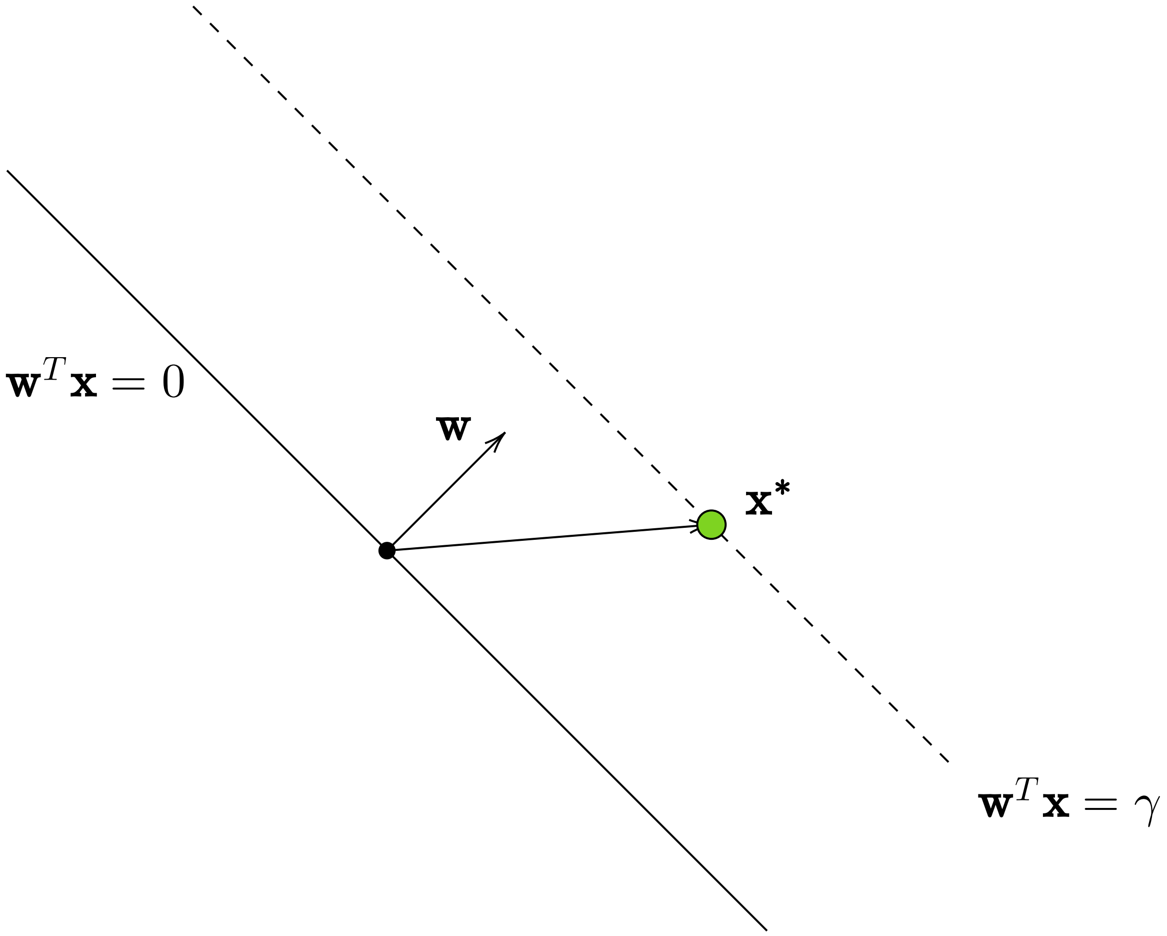
Large margin
Better generalization

Computing the Margin



For any linear classifier represented by $\mathbf{w}$:

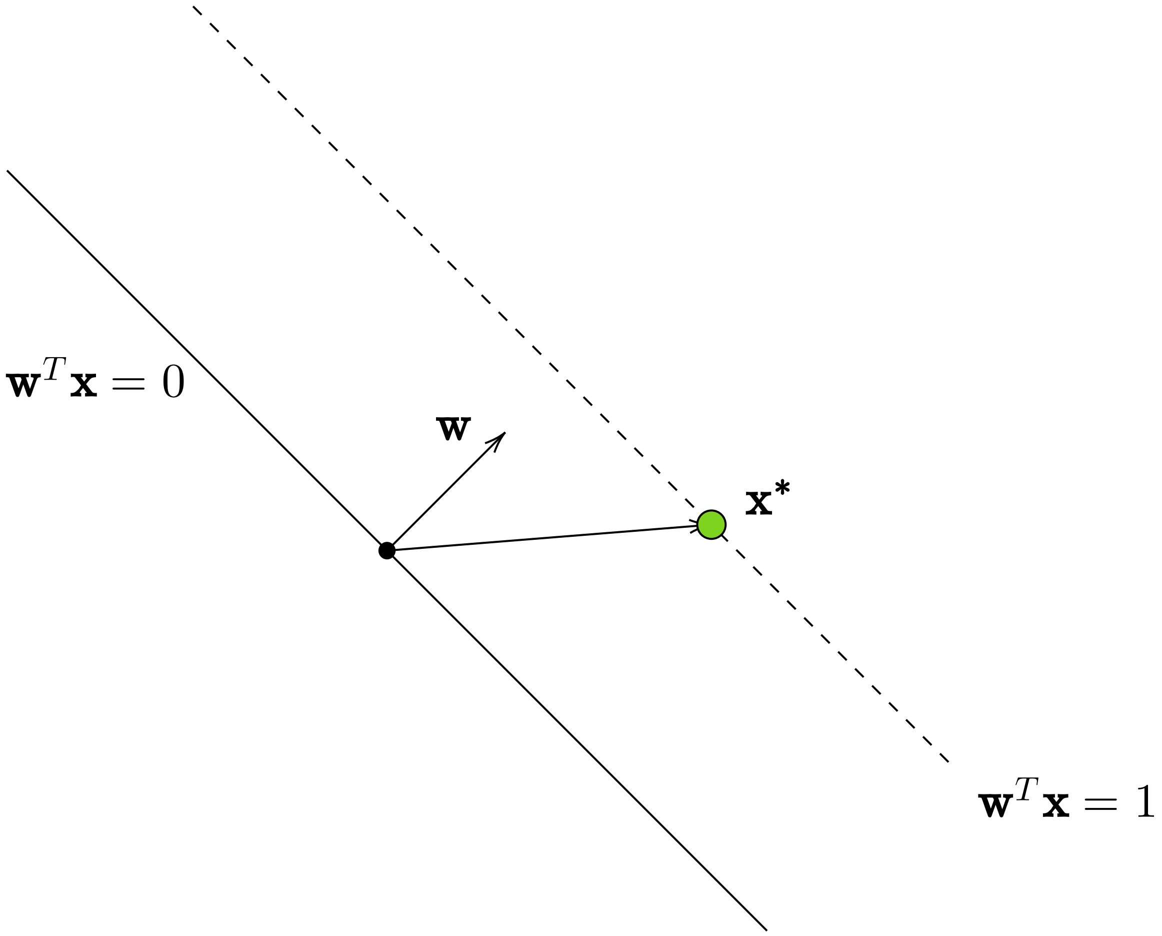
Computing the Margin



For any linear classifier represented by $\mathbf{w}$:

(1) Find the point closest to it → $\mathbf{x}^*$

Computing the Margin

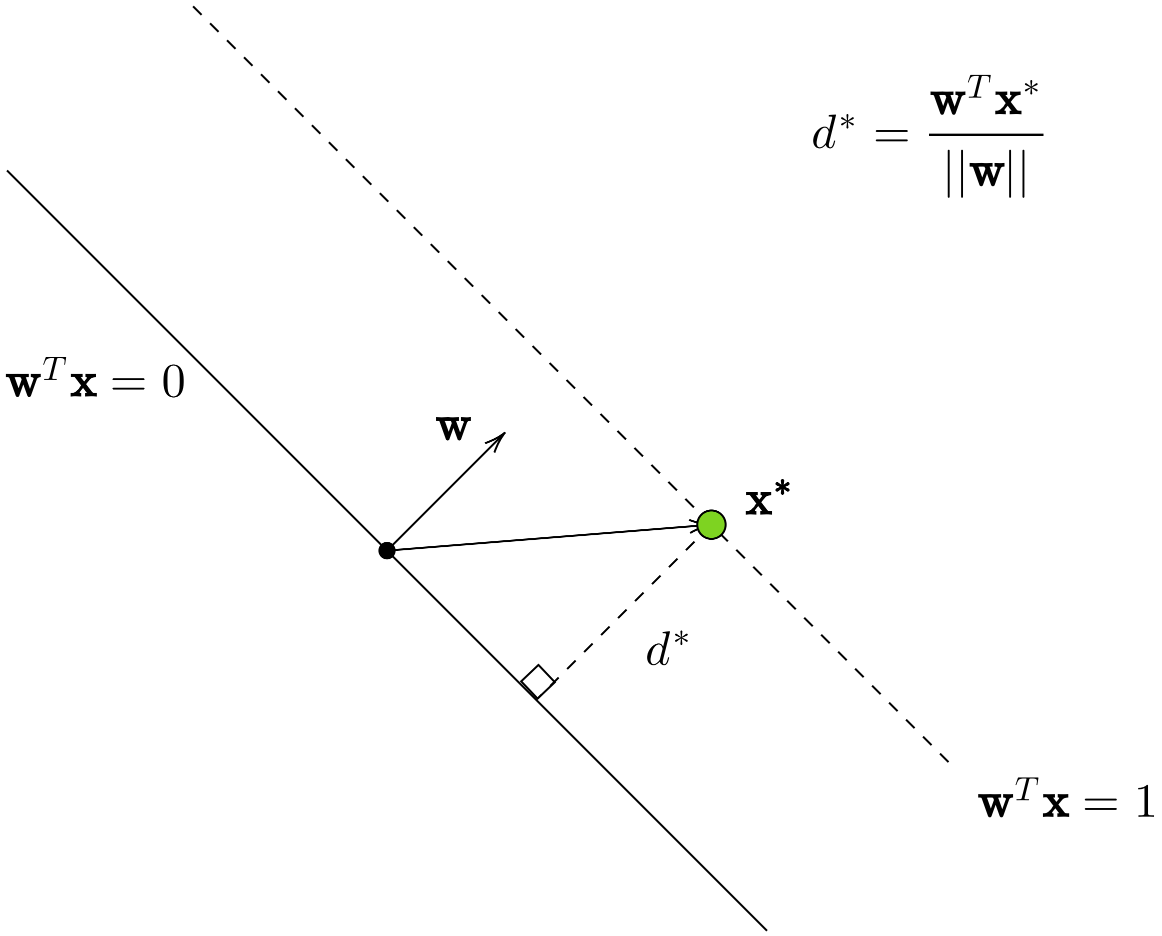


For any linear classifier represented by $\mathbf{w}$:

(1) Find the point closest to it → $\mathbf{x}^*$

(2) Scale $\mathbf{w}$ such that $\mathbf{x}^*$ lies on $\mathbf{w}^T\mathbf{x} = 1$

Computing the Margin



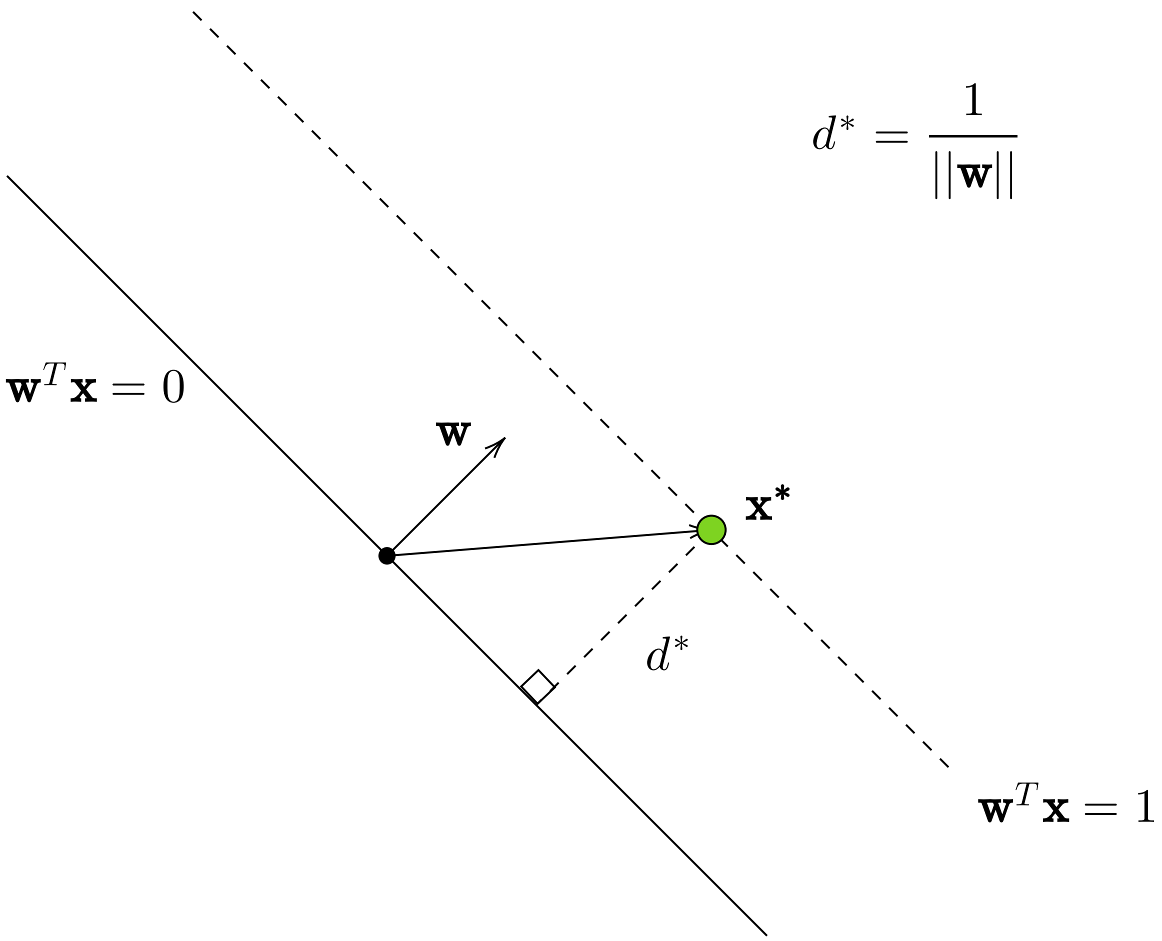
For any linear classifier represented by $\mathbf{w}$:

(1) Find the point closest to it → $\mathbf{x}^*$

(2) Scale $\mathbf{w}$ such that $\mathbf{x}^*$ lies on $\mathbf{w}^T\mathbf{x} = 1$

(3) Distance of $\mathbf{x}^*$ from the line is $\frac{1}{||\mathbf{w}||}$

Computing the Margin



For any linear classifier represented by $\mathbf{w}$:

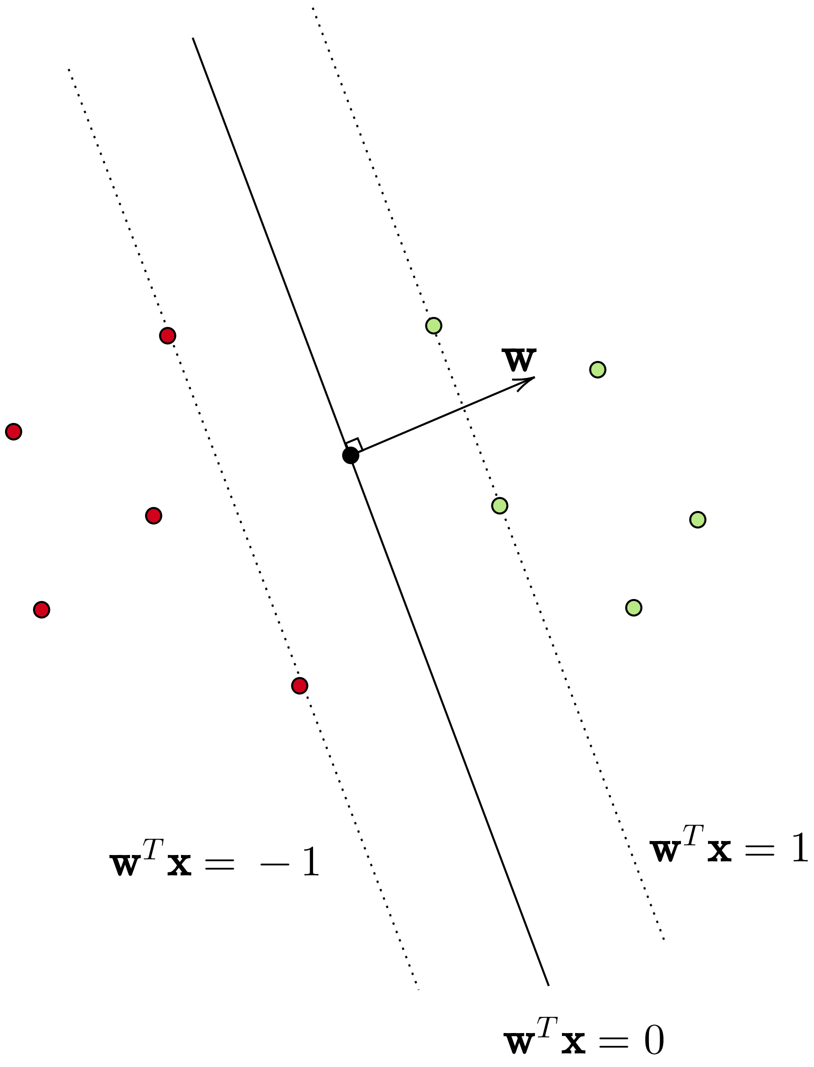
(1) Find the point closest to it → $\mathbf{x^*}$

(2) Scale $\mathbf{w}$ such that $\mathbf{x^*}$ lies on $\mathbf{w}^T\mathbf{x} = 1$

(3) Distance of $\mathbf{x^*}$ from the line is $\frac{1}{||\mathbf{w}||}$

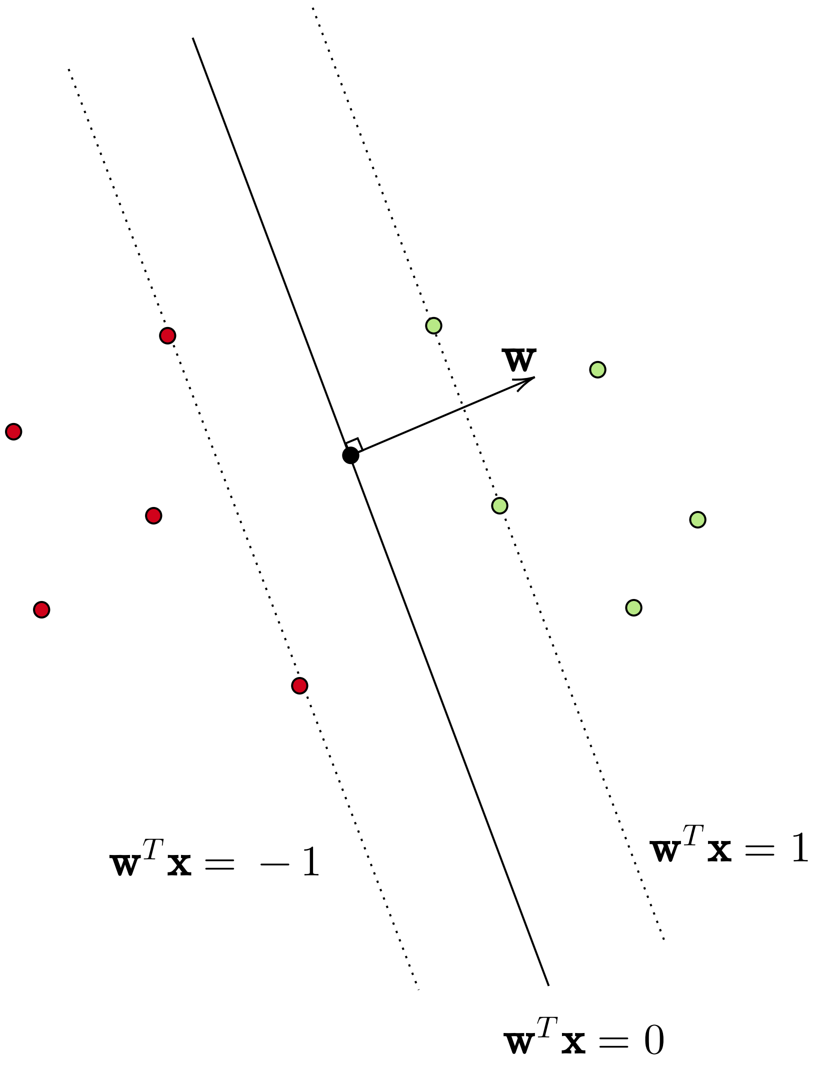
(4) This is the (geometric) margin for this linear classifier.

Beyond the "margin"



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

Max-Margin Classifier

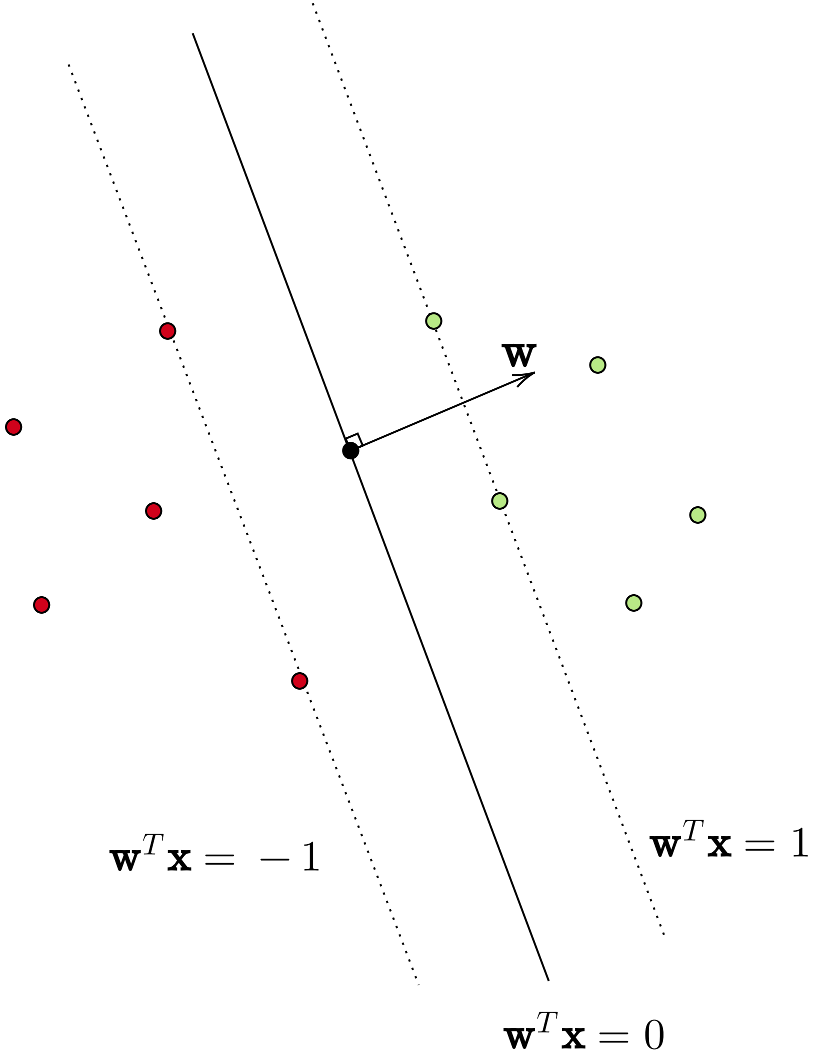


$$\max_{\mathbf{w}} \quad \frac{1}{||\mathbf{w}||}$$

sub. to

$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

Max-Margin Classifier



$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2}$$

sub. to

$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

Primal and Dual

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to

$$g(\mathbf{w}) \leqslant 0$$

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to

$$g(\mathbf{w}) \leqslant 0$$

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to $\qquad \equiv \qquad \min_{\mathbf{w}} \left[\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) \right]$

$$g(\mathbf{w}) \leqslant 0$$

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to $\equiv$ $\min_{\mathbf{w}} \left[\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) \right]$ $\equiv$ $\max_{\alpha \geqslant 0} \; \min_{\mathbf{w}} f(\mathbf{w}) + \alpha g(\mathbf{w})$

$$g(\mathbf{w}) \leqslant 0$$

Strong Duality

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to $\qquad \equiv \qquad \min_{\mathbf{w}} \left[\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) \right] \qquad \equiv \qquad \max_{\alpha \geqslant 0} \; \min_{\mathbf{w}} f(\mathbf{w}) + \alpha g(\mathbf{w})$

$$g(\mathbf{w}) \leqslant 0$$

Strong Duality

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2}$$

sub. to

$$\left(\mathbf{w}^T \mathbf{x}_i\right) y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to

$$g(\mathbf{w}) \leqslant 0$$

$$\equiv \quad \min_{\mathbf{w}} \left[\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) \right] \quad \equiv \quad \max_{\alpha \geqslant 0} \ \min_{\mathbf{w}} f(\mathbf{w}) + \alpha g(\mathbf{w})$$

Strong Duality

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2}$$

sub. to

$$\left(\mathbf{w}^T \mathbf{x}_i\right) y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

$$\equiv \quad \min_{\mathbf{w}} \ \max_{\boldsymbol{\alpha} \geqslant 0} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \Big[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i \Big]$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

Primal and Dual

$$\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) = \begin{cases} f(\mathbf{w}), & g(\mathbf{w}) \leqslant 0 \\ \infty, & g(\mathbf{w}) > 0 \end{cases}$$

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

sub. to

$$g(\mathbf{w}) \leqslant 0$$

$$\equiv \quad \min_{\mathbf{w}} \left[\max_{\alpha \geqslant 0} \quad f(\mathbf{w}) + \alpha g(\mathbf{w}) \right] \quad \equiv \quad \max_{\alpha \geqslant 0} \; \min_{\mathbf{w}} f(\mathbf{w}) + \alpha g(\mathbf{w})$$

Strong Duality

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2}$$

sub. to

$$\left(\mathbf{w}^T \mathbf{x}_i\right) y_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

$$\equiv \quad \min_{\mathbf{w}} \; \max_{\boldsymbol{\alpha} \geqslant 0} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \Big[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i \Big] \quad \equiv \quad \max_{\boldsymbol{\alpha} \geqslant 0} \; \min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \Big[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i \Big]$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

Formulating the Dual

$$\max_{\boldsymbol{\alpha} \geq 0} \boxed{\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \Big[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i \Big]}$$

$$\boxed{\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i y_i}$$

Formulating the Dual

$$\max_{\boldsymbol{\alpha}\geq 0} \boxed{\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \Big[1 - \big(\mathbf{w}^T \mathbf{x}_i\big) y_i\Big]}$$

$$\boxed{\mathbf{w} = \sum_{i=1}^{n} \alpha_i \mathbf{x}_i y_i}$$

$$d \times n$$

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

$$n \times n$$

$$\mathbf{Y} = \begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}$$

$$\mathbf{X}\mathbf{Y}\boldsymbol{\alpha} = \begin{bmatrix} | & & | \\ y_1\mathbf{x}_1 & \cdots & y_n\mathbf{x}_n \\ | & & | \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \sum_{i=1}^{n} \alpha_i (y_i \mathbf{x}_i) = \mathbf{w}$$

Formulating the Dual

$$\max_{\boldsymbol{\alpha}\geq 0} \boxed{\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + \sum_{i=1}^{n} \alpha_i \left[1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i\right]}$$

$$\boxed{\mathbf{w} = \mathbf{X}\mathbf{Y}\boldsymbol{\alpha}}$$

$d \times n$

$$\mathbf{X} = \begin{bmatrix} | & & | \\ \mathbf{x}_1 & \cdots & \mathbf{x}_n \\ | & & | \end{bmatrix}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

$n \times n$

$$\mathbf{Y} = \begin{bmatrix} y_1 & & 0 \\ & \ddots & \\ 0 & & y_n \end{bmatrix}$$

$$\boxed{\max_{\boldsymbol{\alpha}\geq 0} \ \boldsymbol{\alpha}^T \mathbf{1} - \frac{\boldsymbol{\alpha}^T \left(\mathbf{Y}^T \mathbf{X}^T \mathbf{X}\mathbf{Y}\right) \boldsymbol{\alpha}}{2}}$$

Advantages of the Dual

$$\boxed{\max_{\boldsymbol{\alpha} \geqslant 0} \ \boldsymbol{\alpha}^T \mathbf{1} - \frac{\boldsymbol{\alpha}^T \left(\mathbf{Y}^T \mathbf{X}^T \mathbf{X} \mathbf{Y} \right) \boldsymbol{\alpha}}{2}}$$

(1) $\boldsymbol{\alpha} \in \mathbb{R}^n$ and $\mathbf{w} \in \mathbb{R}^d$, if $n \ll d$, we are solving for fewer variables in the dual

(2) Simpler constraints, just bounds.

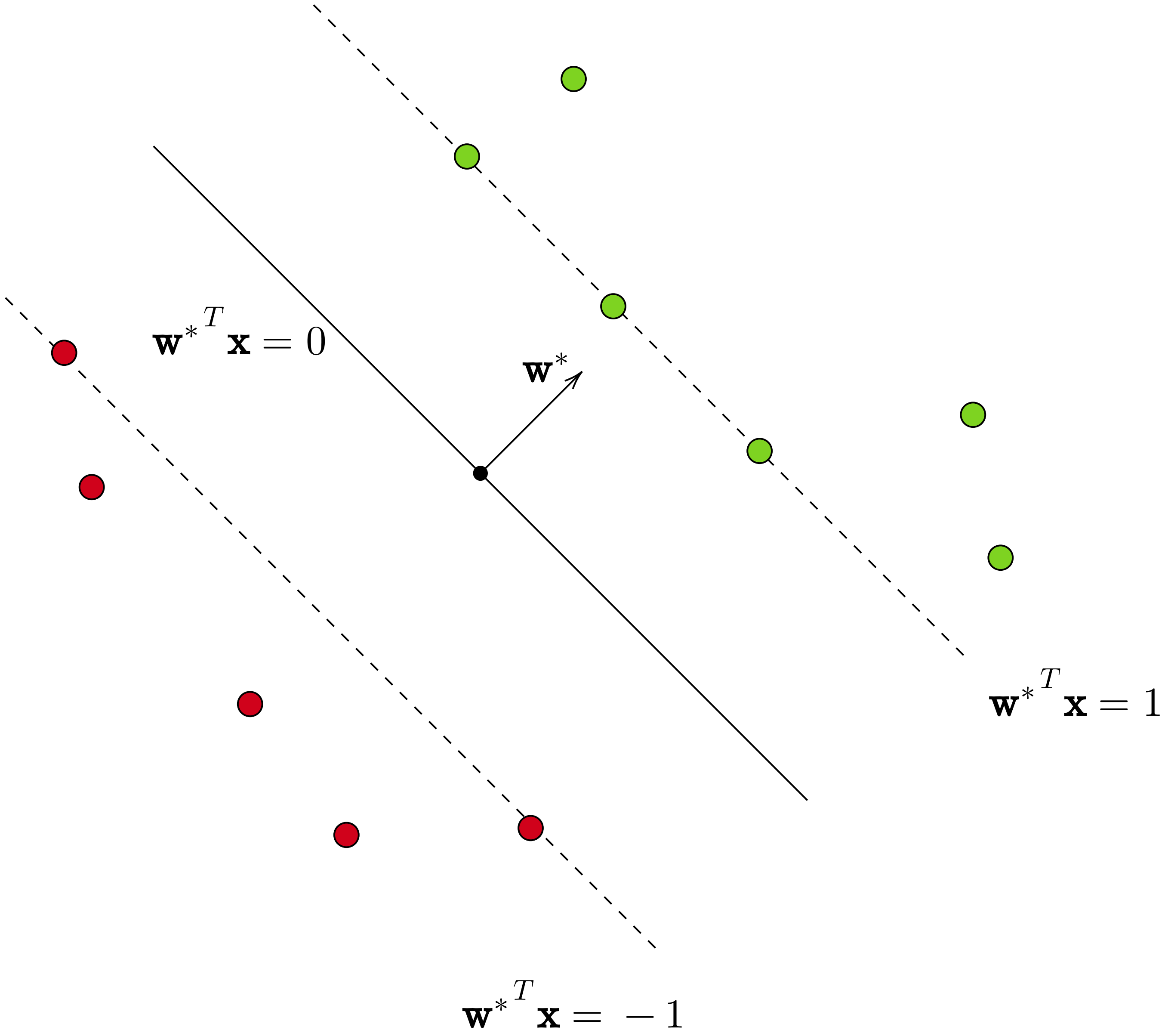
(3) The appearance of $\mathbf{X}^T \mathbf{X} \implies$ kernels.

Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Support Vectors

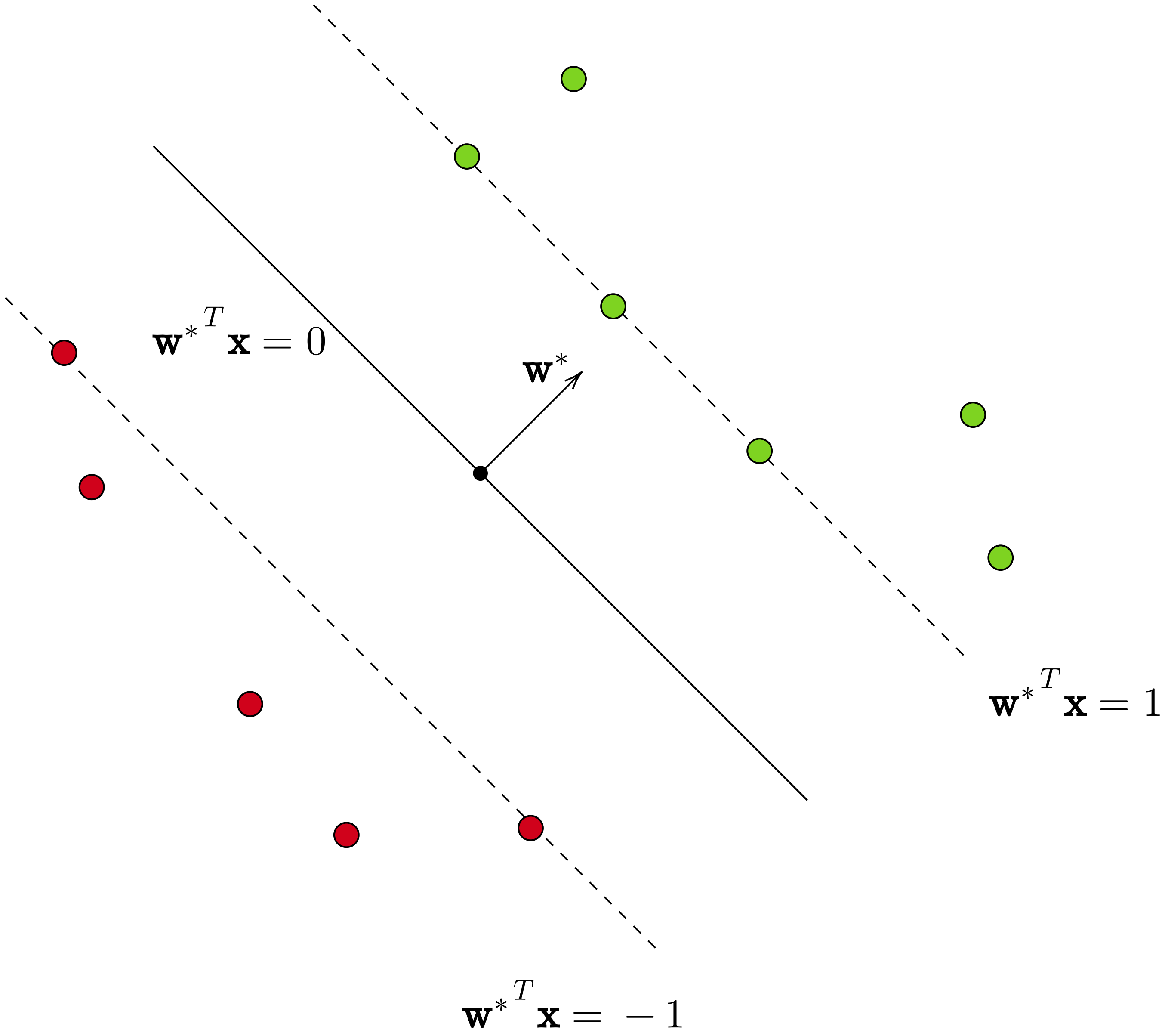
$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$



Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

$$\alpha_i^* \left[1 - \left(\mathbf{w}^{*T} \mathbf{x}_i \right) y_i \right] = 0$$



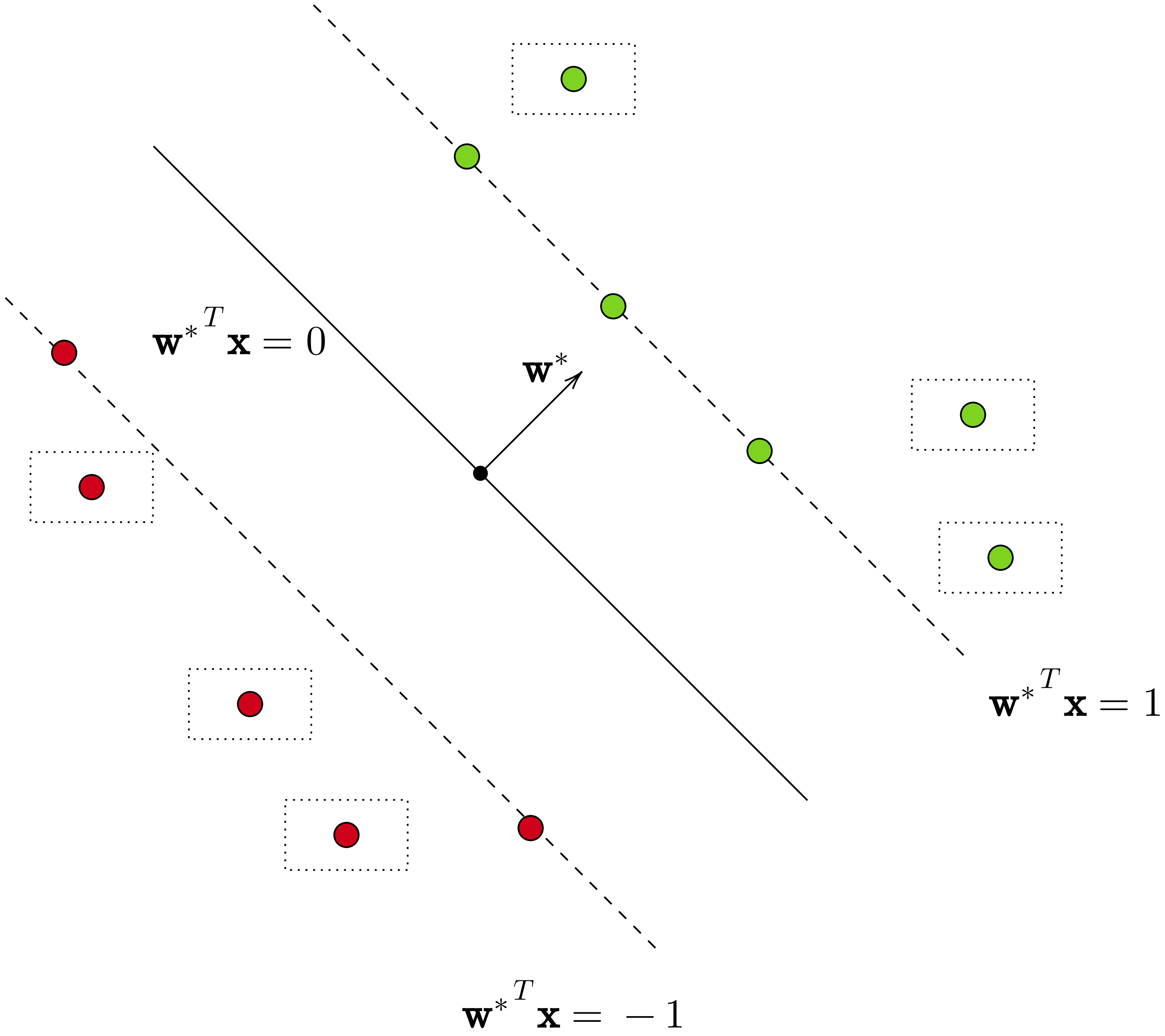
Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

$$\alpha_i^* \left[1 - \left({\mathbf{w}^*}^T \mathbf{x}_i \right) y_i \right] = 0$$

$$y_i \left({\mathbf{w}^*}^T \mathbf{x}_i \right) \neq 1$$

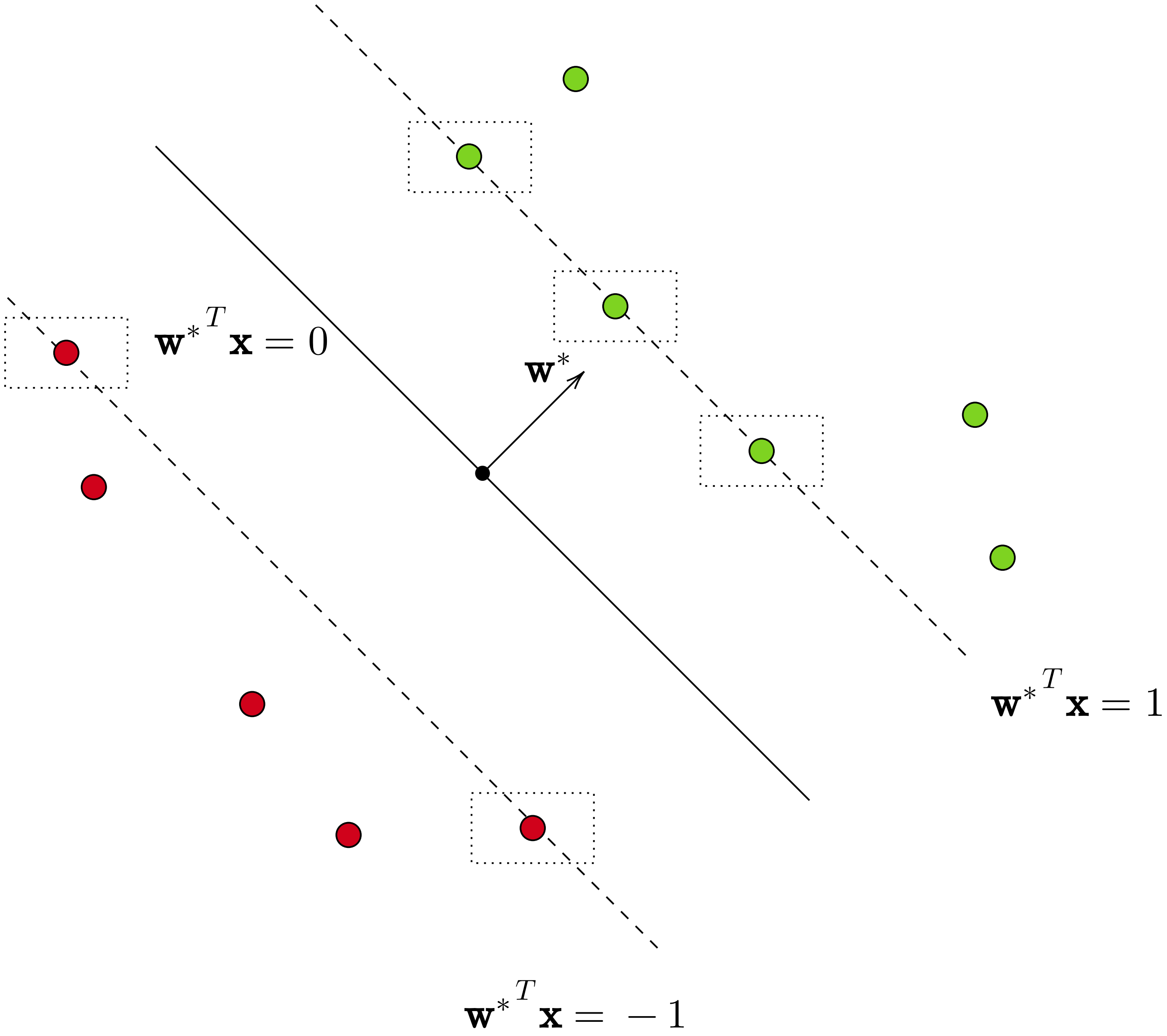
$$\alpha_i^* = 0$$



Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

$$\begin{tikzcd}
& \alpha_i^*\Big[1 - \big(\mathbf{w}^{*T}\mathbf{x}_i\big) y_i\Big] = 0 \arrow[dl, "y_i\left(\mathbf{w}^{*T}\mathbf{x}_i\right) \neq 1"'] \arrow[dr, "y_i\left(\mathbf{w}^{*T}\mathbf{x}_i\right) = 1"] & \\
\alpha_i^* = 0 & & \alpha_i^* \geqslant 0
\end{tikzcd}$$



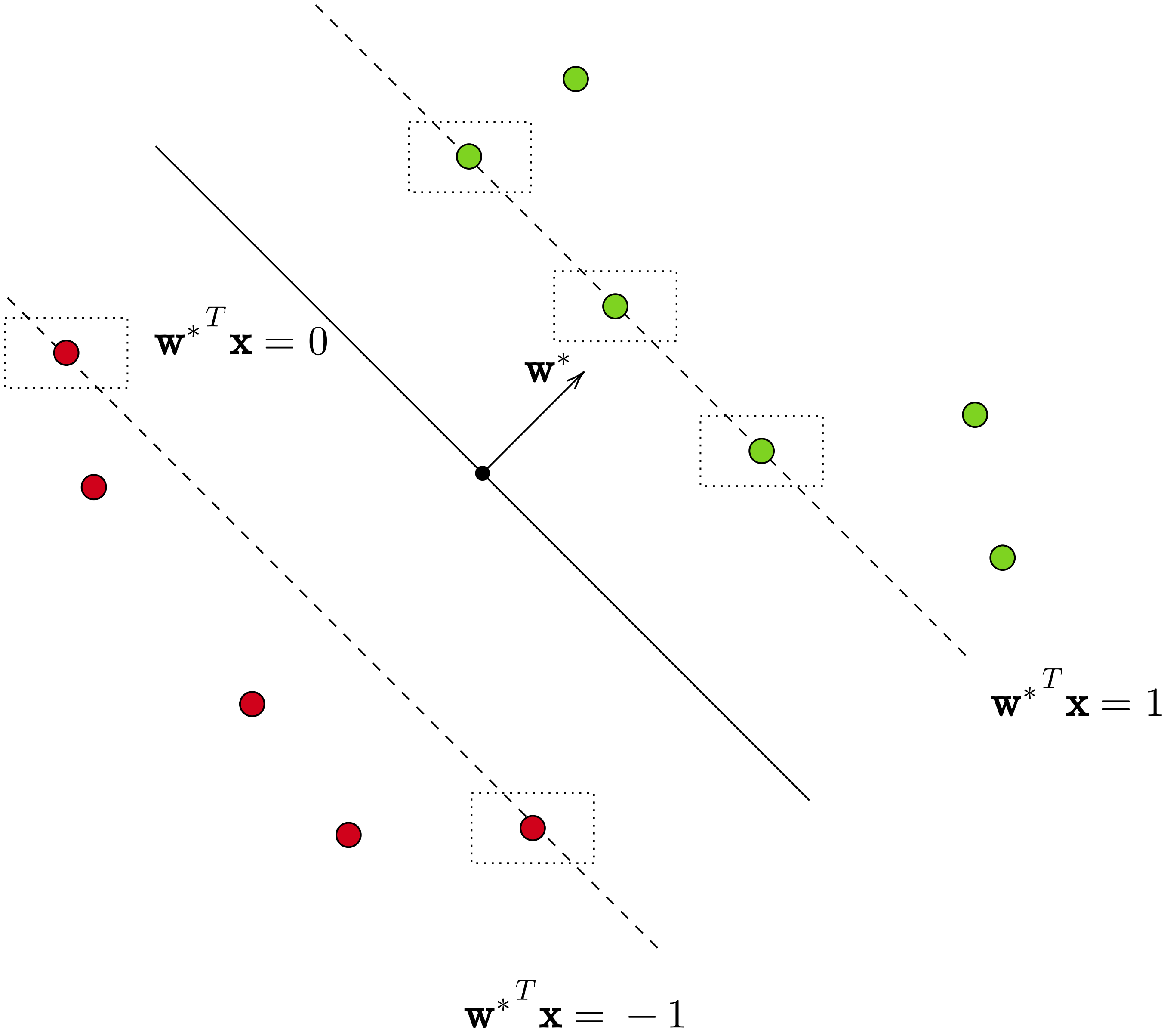
Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

$$\alpha_i^* \left[1 - \left({\mathbf{w}^*}^T \mathbf{x}_i \right) y_i \right] = 0$$

$$\alpha_i^* > 0$$

$$y_i \left({\mathbf{w}^*}^T \mathbf{x}_i \right) = 1$$



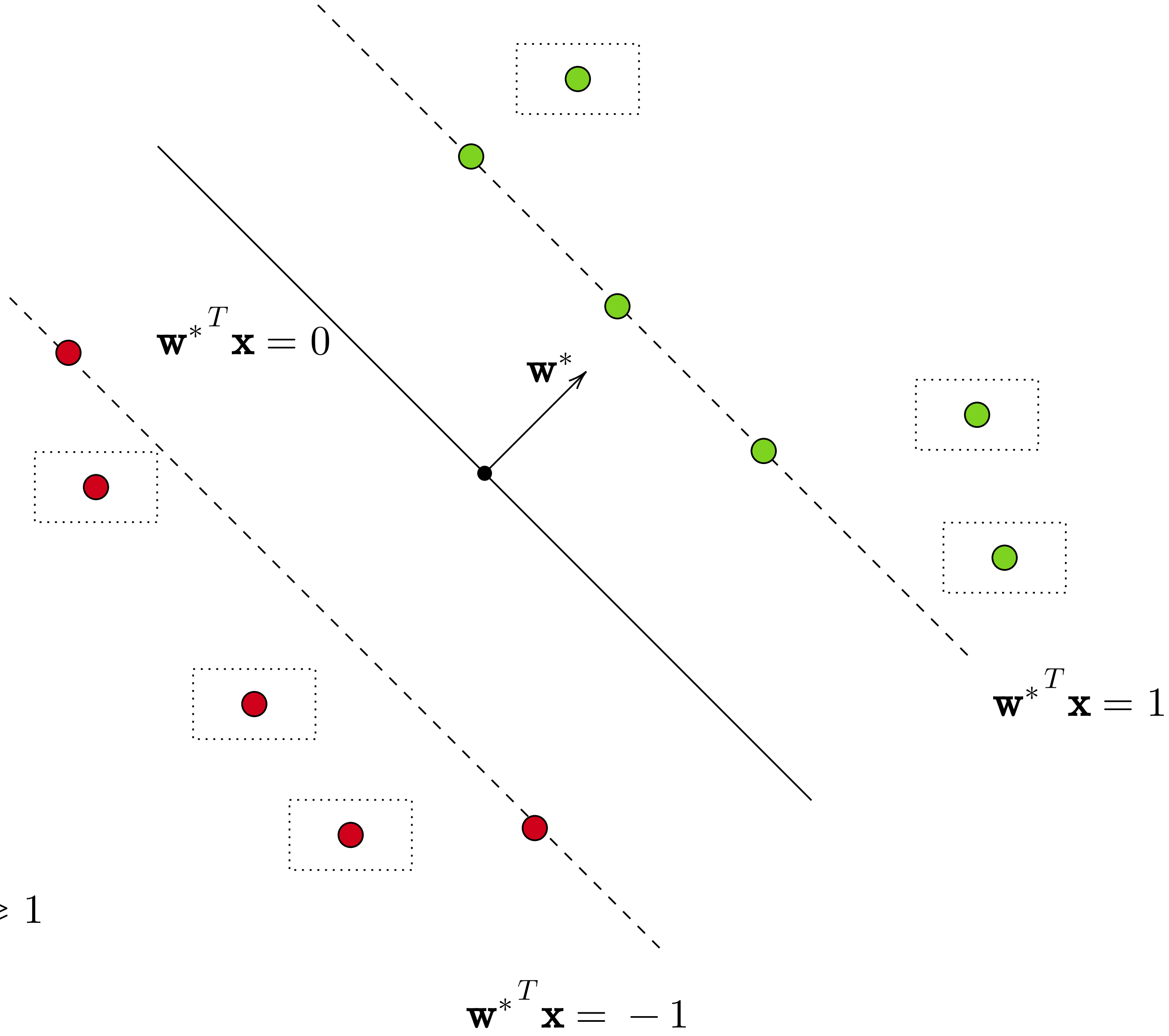
Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

$$\alpha_i^* \left[1 - \left({\mathbf{w}^*}^T \mathbf{x}_i \right) y_i \right] = 0$$

$$\alpha_i^* > 0 \implies y_i \left({\mathbf{w}^*}^T \mathbf{x}_i \right) = 1$$

$$\alpha_i^* = 0 \implies y_i \left({\mathbf{w}^*}^T \mathbf{x}_i \right) \geqslant 1$$



Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix}$$

$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Definition: A support vector is a point for which $\alpha_i^* > 0$

Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Definition: A support vector is a point for which $\alpha_i^* > 0$

$$\alpha_i^* g_i\left(\mathbf{w}^*\right) = 0 \Longrightarrow \alpha_i^*\left[1 - \left(\left(\mathbf{w}^*\right)^T \mathbf{x}_i\right) y_i\right] = 0$$

Complementary Slackness

Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Definition: A support vector is a point for which $\alpha_i^* > 0$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0 \Longrightarrow \alpha_i^* \left[1 - \left((\mathbf{w}^*)^T \mathbf{x}_i \right) y_i \right] = 0$$

Complementary Slackness

Every support vector lies on one of the two supporting hyperplanes $(\mathbf{w}^*)^T \mathbf{x} = \pm 1$

Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Definition: A support vector is a point for which $\alpha_i^* > 0$

$$\alpha_i^* g_i(\mathbf{w}^*) = 0 \Longrightarrow \alpha_i^* \left[1 - \left(\left(\mathbf{w}^*\right)^T \mathbf{x}_i\right) y_i\right] = 0$$

Complementary Slackness

Every support vector lies on one of the two supporting hyperplanes $\left(\mathbf{w}^*\right)^T \mathbf{x} = \pm 1$

Every point that is **not** on one of the two supporting hyperplanes has $\alpha_i^* = 0$.

Support Vectors

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix} \qquad \mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Definition: A support vector is a point for which $\alpha_i^* > 0$

$$\alpha_i^* g_i\left(\mathbf{w}^*\right) = 0 \Longrightarrow \alpha_i^*\left[1 - \left(\left(\mathbf{w}^*\right)^T \mathbf{x}_i\right) y_i\right] = 0$$

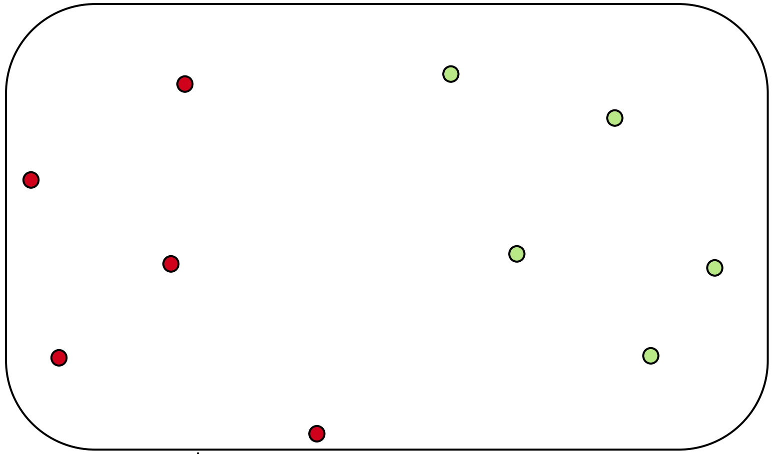
Complementary Slackness

Every support vector lies on one of the two supporting hyperplanes $\left(\mathbf{w}^*\right)^T \mathbf{x} = \pm 1$

Is every point on one of the two supporting hyperplanes a support vector?

Every point that is **not** on one of the two supporting hyperplanes has $\alpha_i^* = 0$.

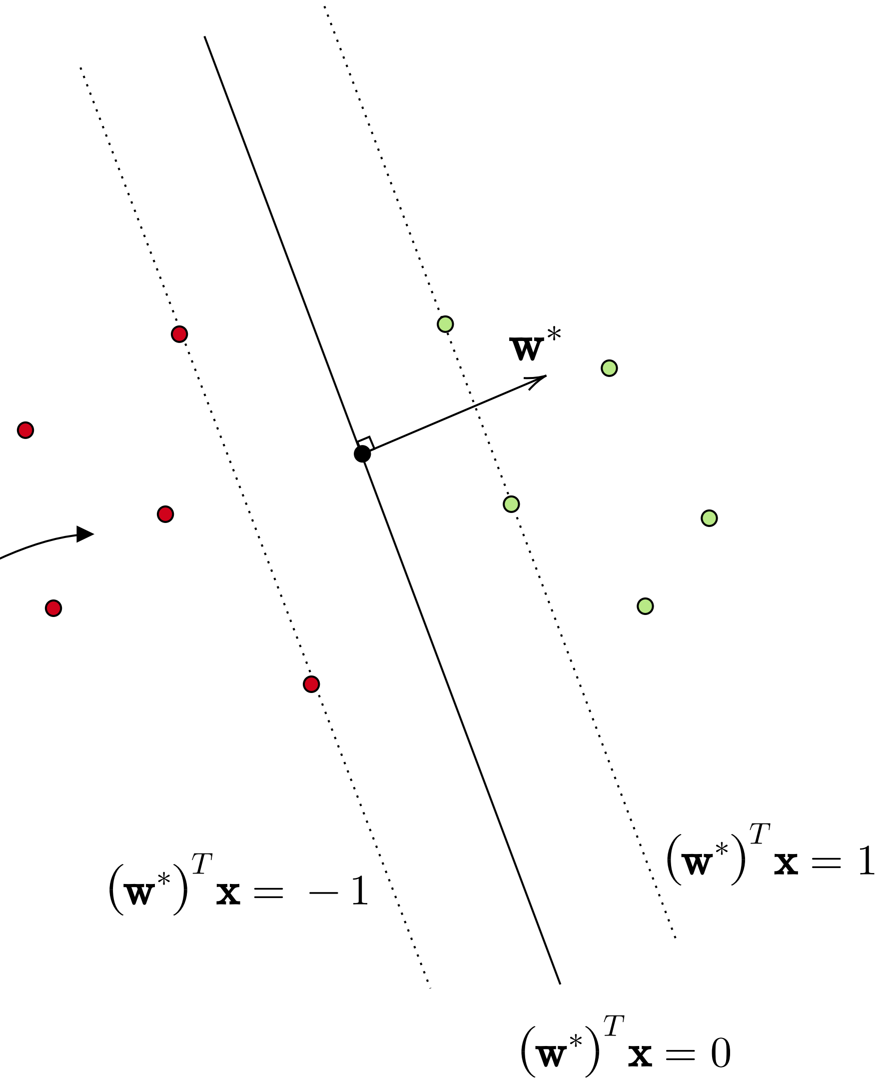
Hard-Margin, Linear-SVM



$$\widehat{y} = \begin{cases} 1, & (\mathbf{w}^*)^T \mathbf{x} \geqslant 0 \\ -1, & (\mathbf{w}^*)^T \mathbf{x} < 0 \end{cases}$$

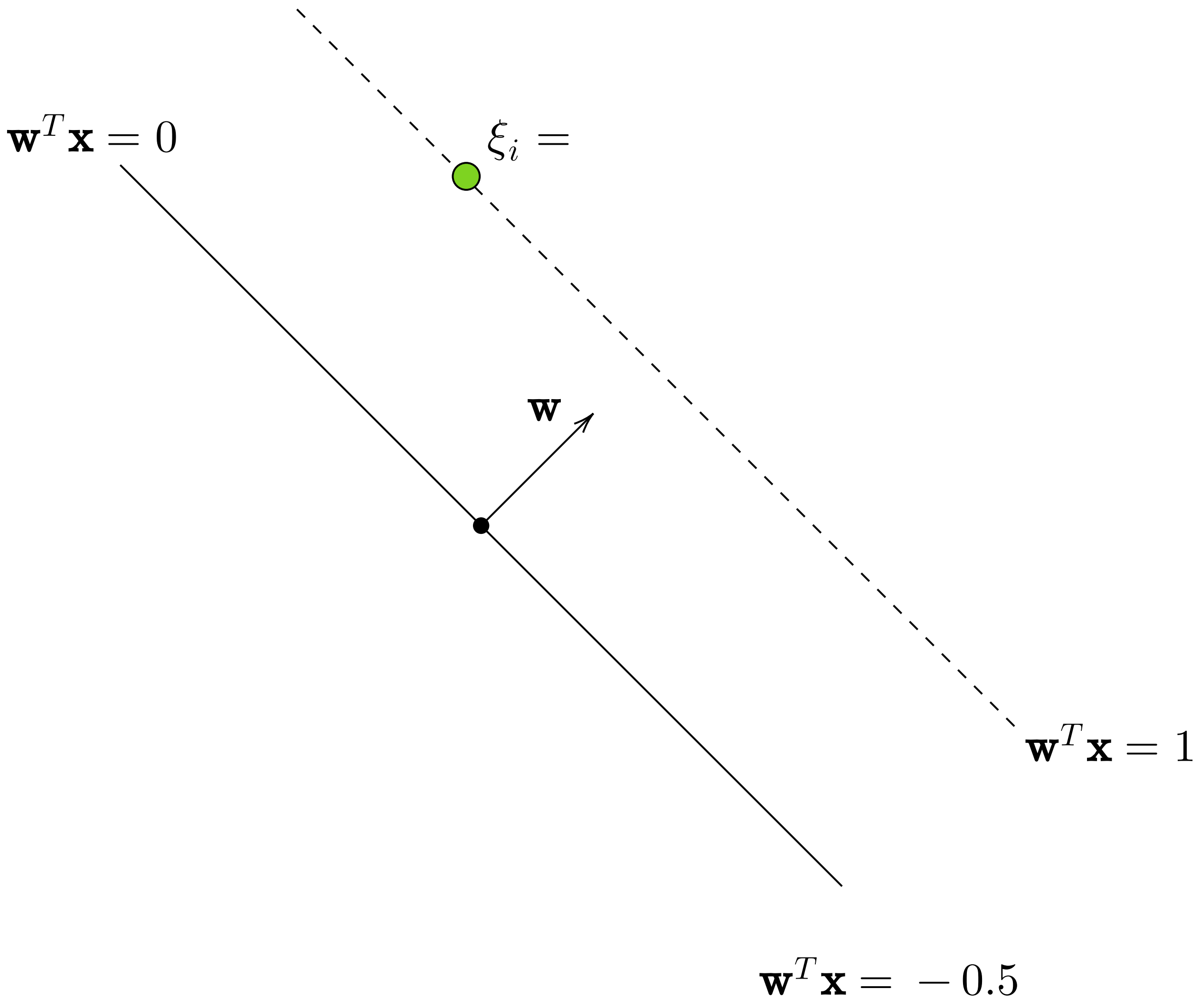
SVM Solver

$$\boldsymbol{\alpha}^* = \begin{bmatrix} \alpha_1^* \\ \vdots \\ \alpha_n^* \end{bmatrix}$$



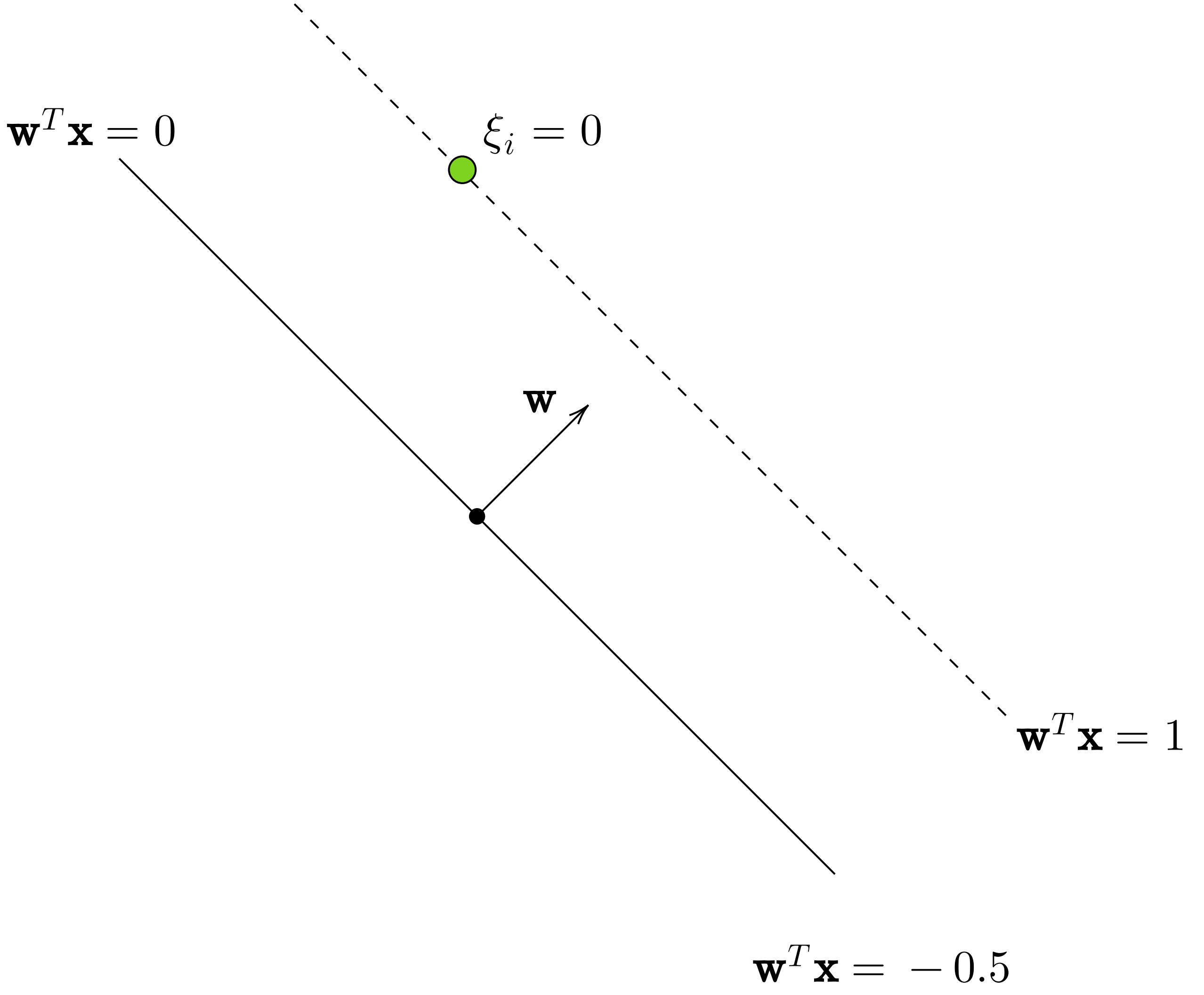
$$\mathbf{w}^* = \sum_{i=1}^{n} \alpha_i^* \mathbf{x}_i y_i$$

Soft-Margin



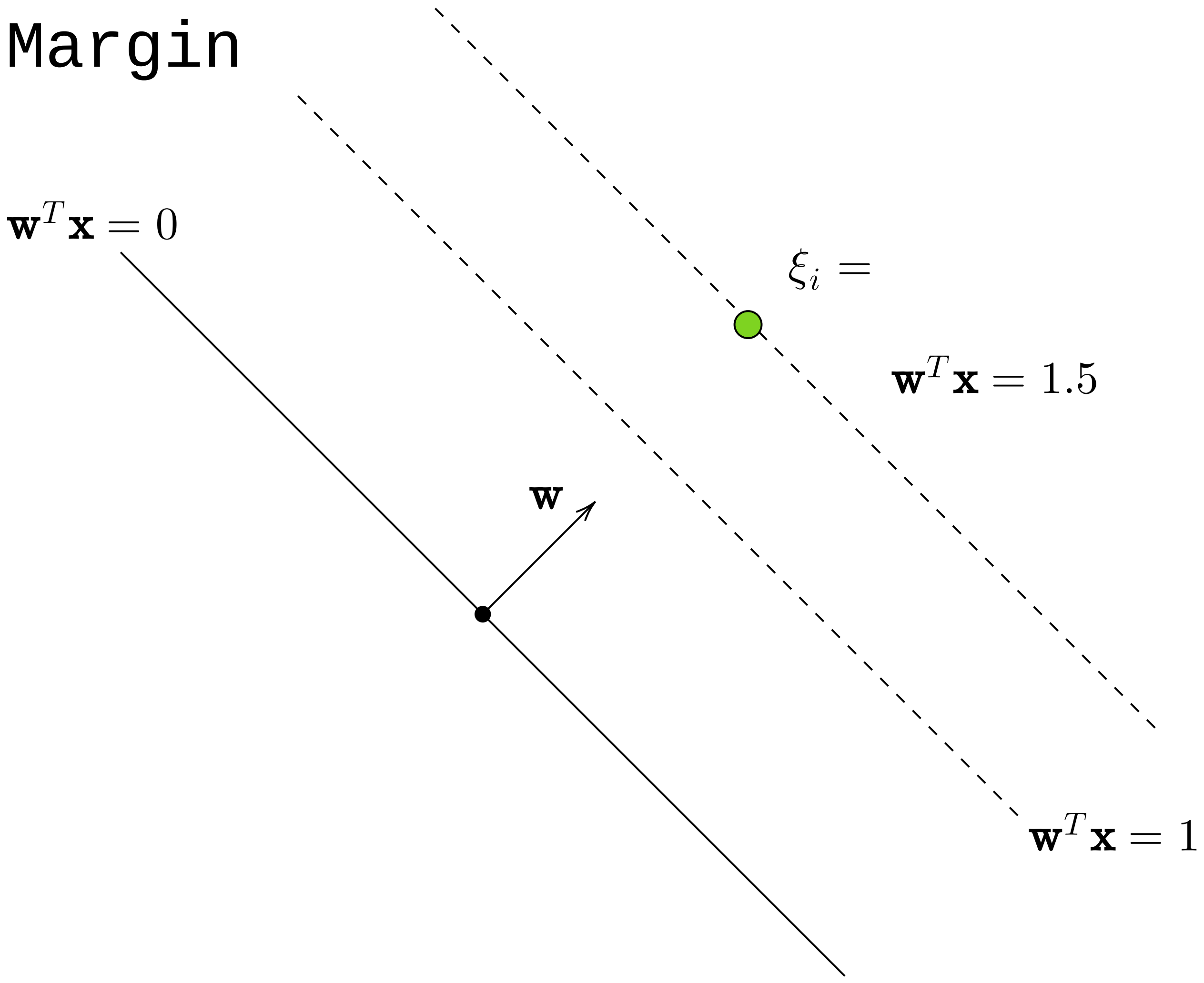
$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

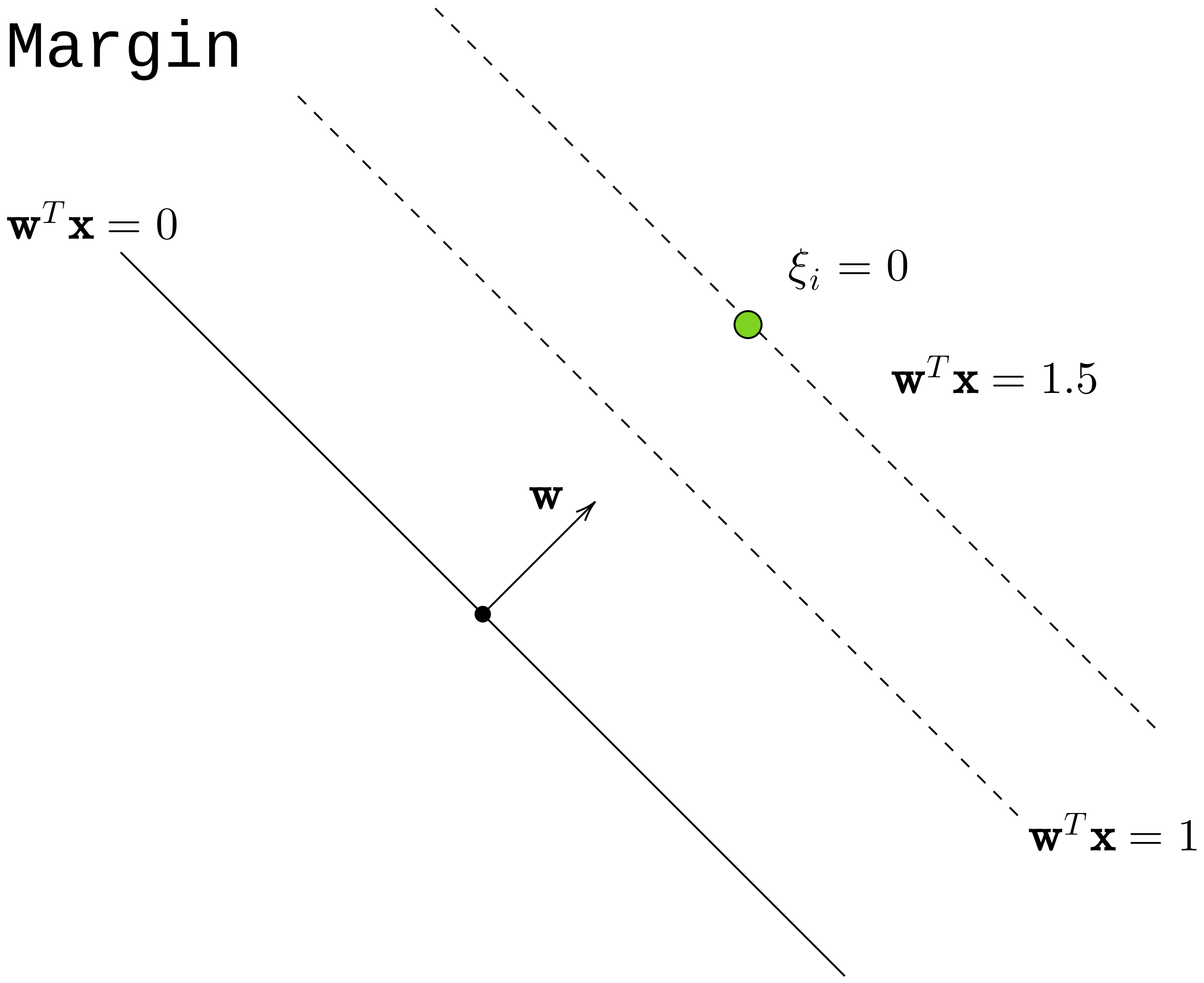
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

$$\xi_i \geqslant 0$$

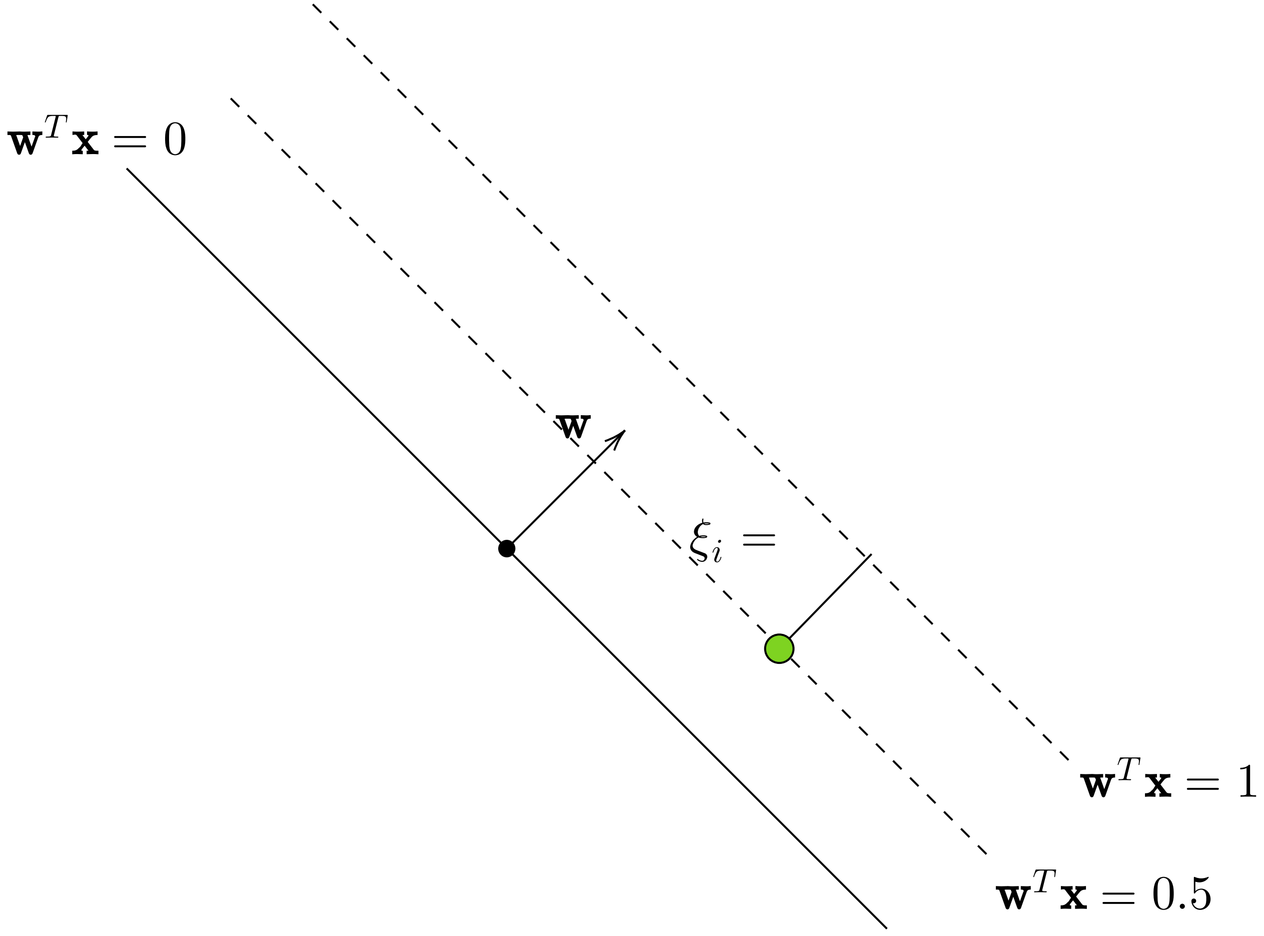
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

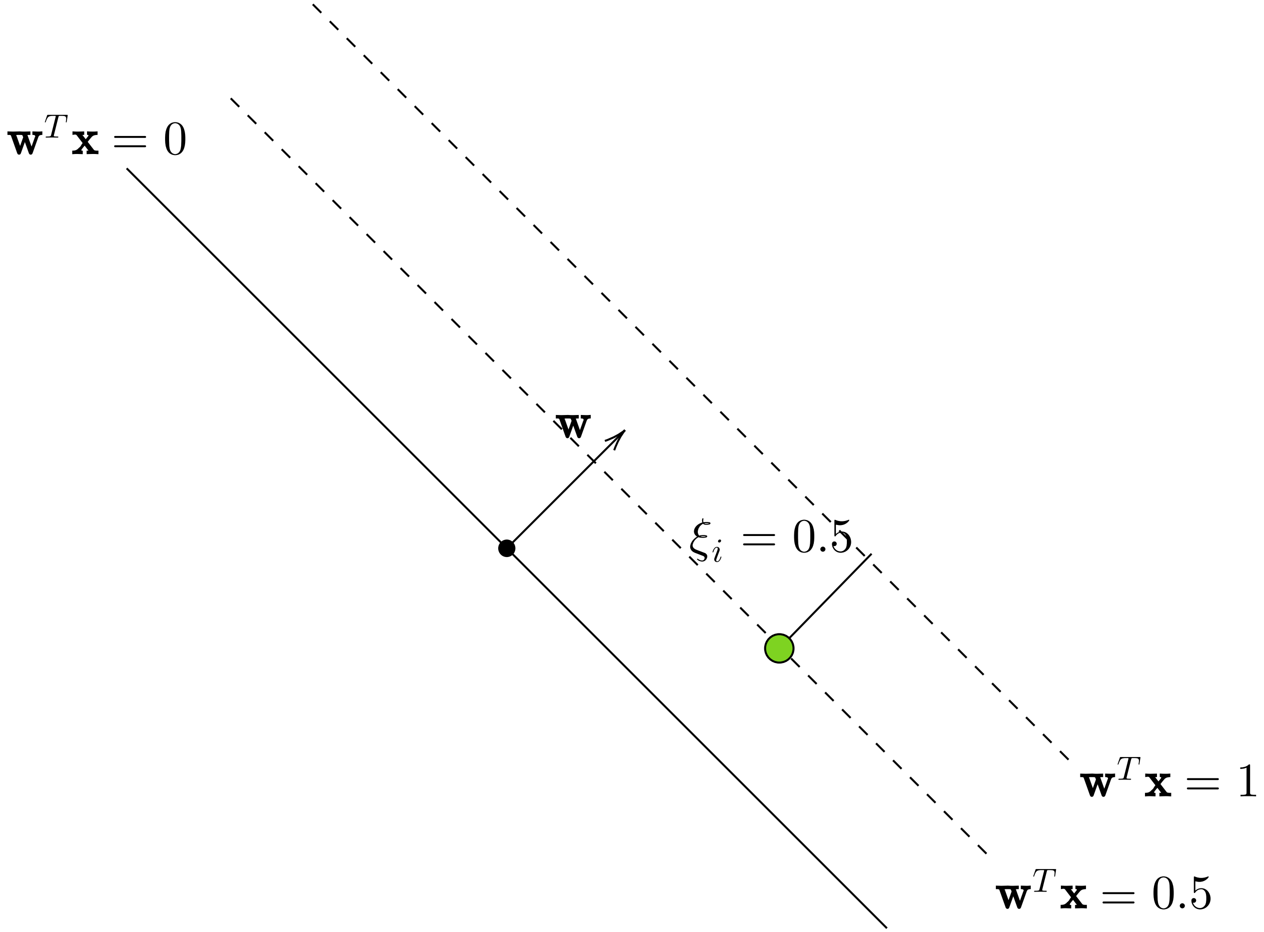
$$\xi_i \geqslant 0$$

Soft-Margin



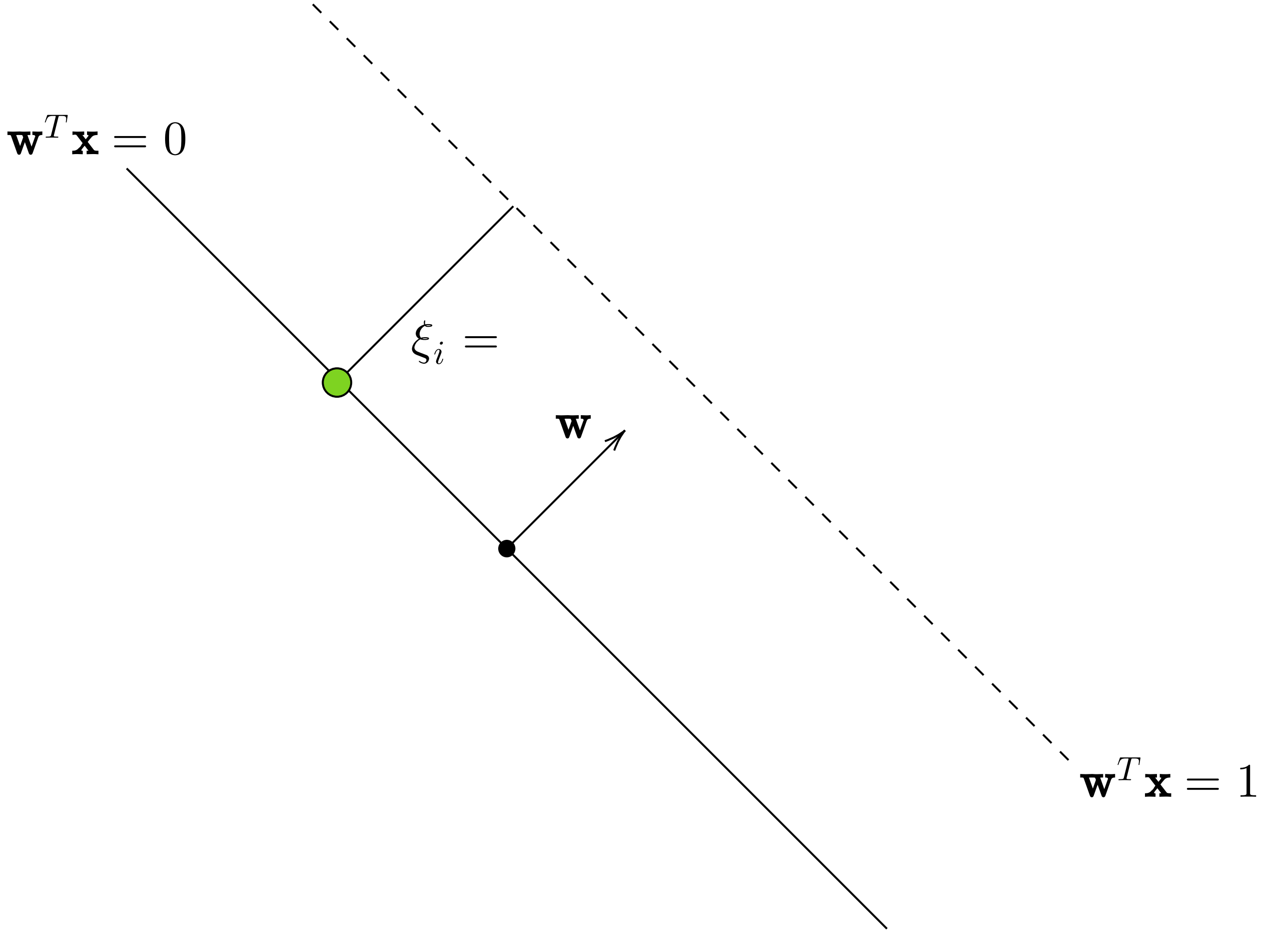
$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

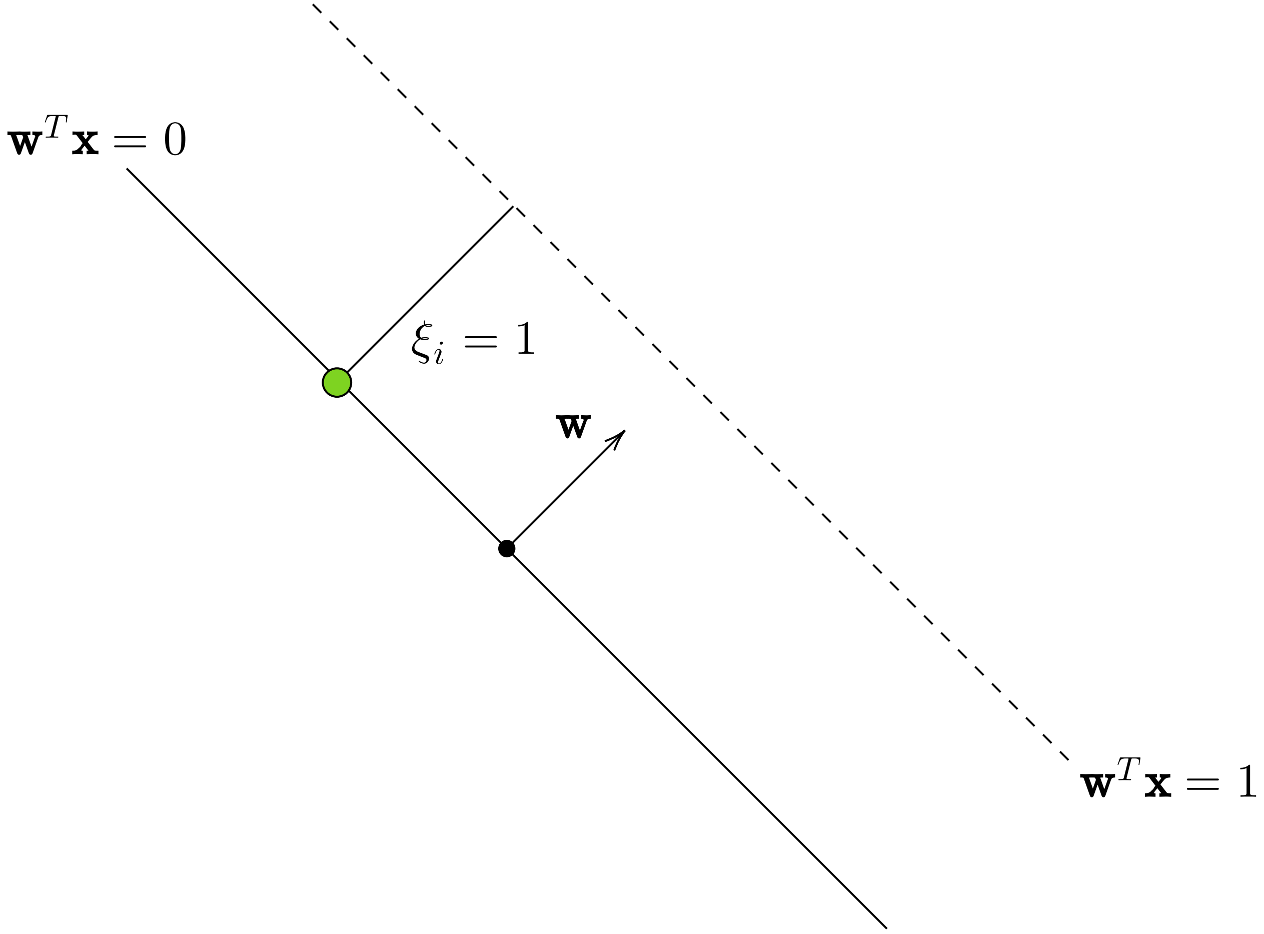
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

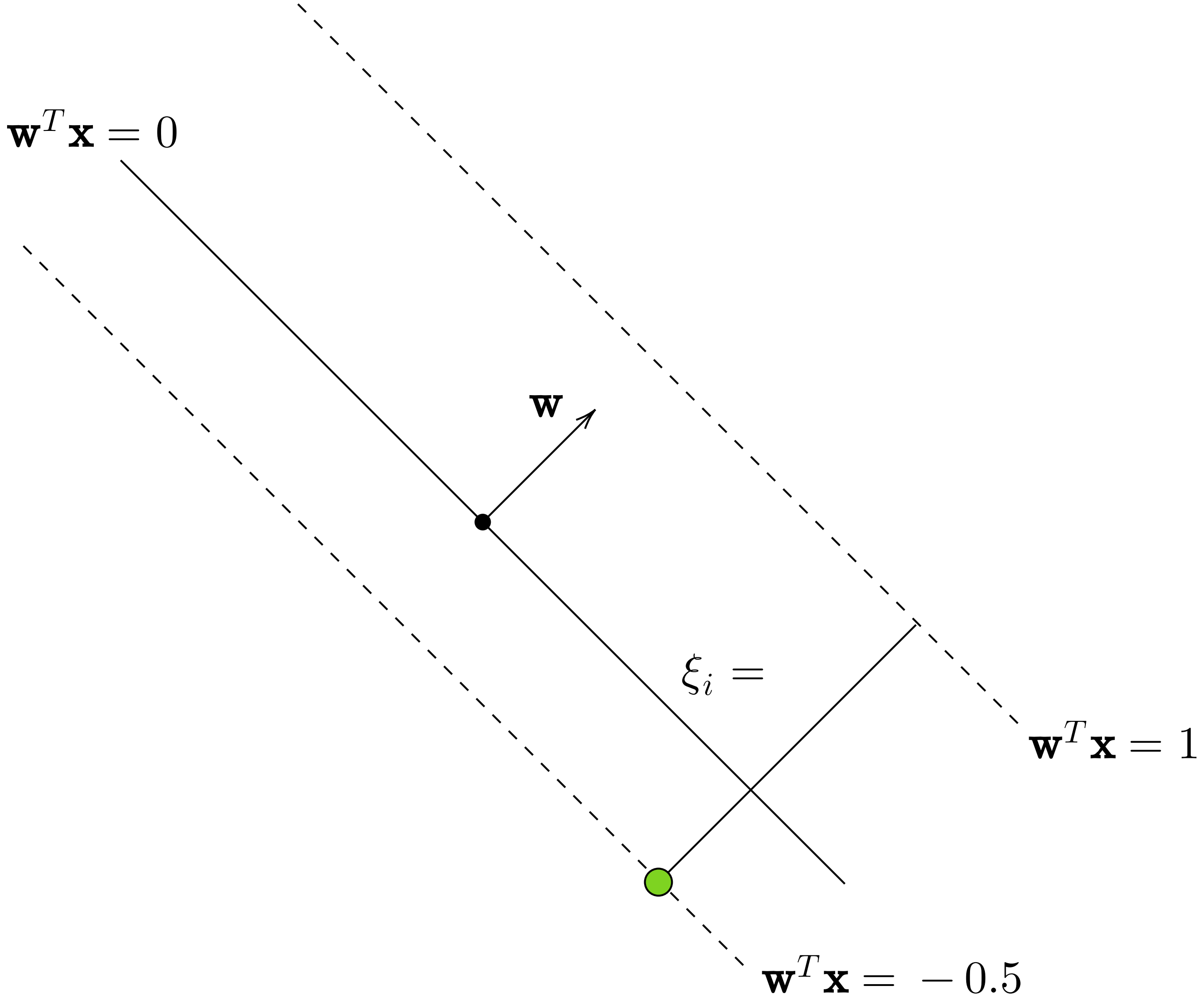
$$\xi_i \geqslant 0$$

Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

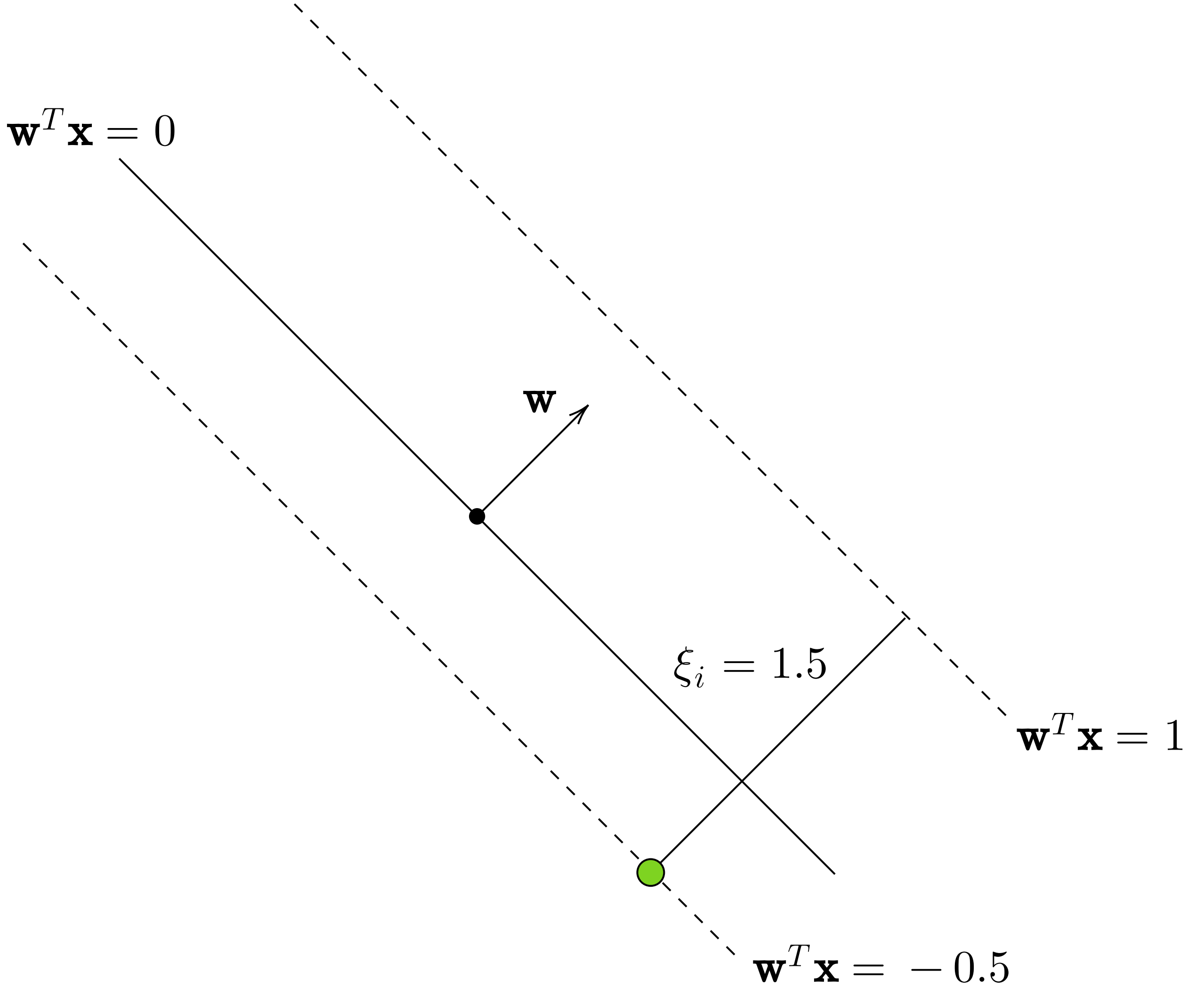
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

$$\xi_i \geqslant 0$$

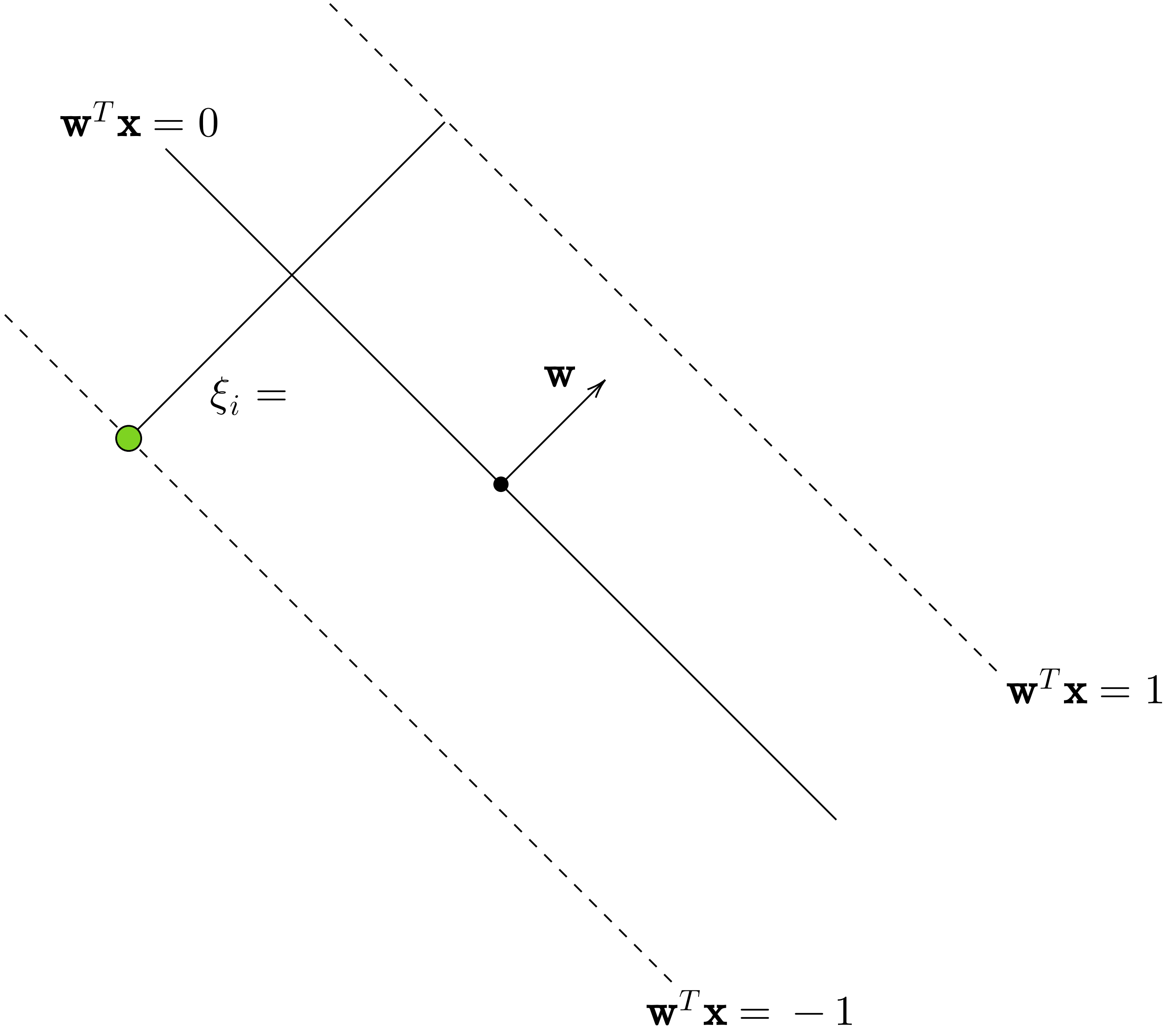
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

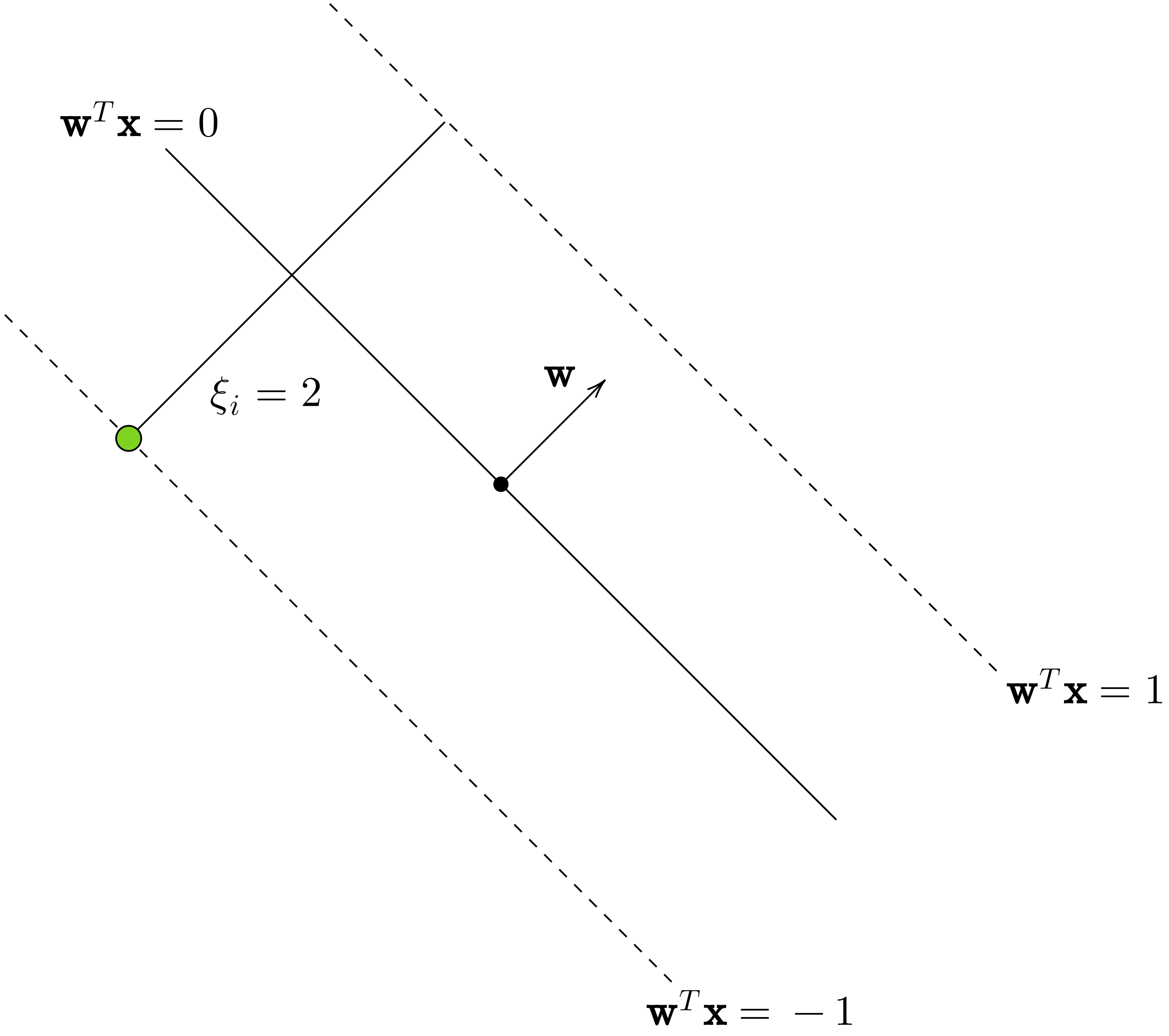
$$\xi_i \geqslant 0$$

Soft-Margin



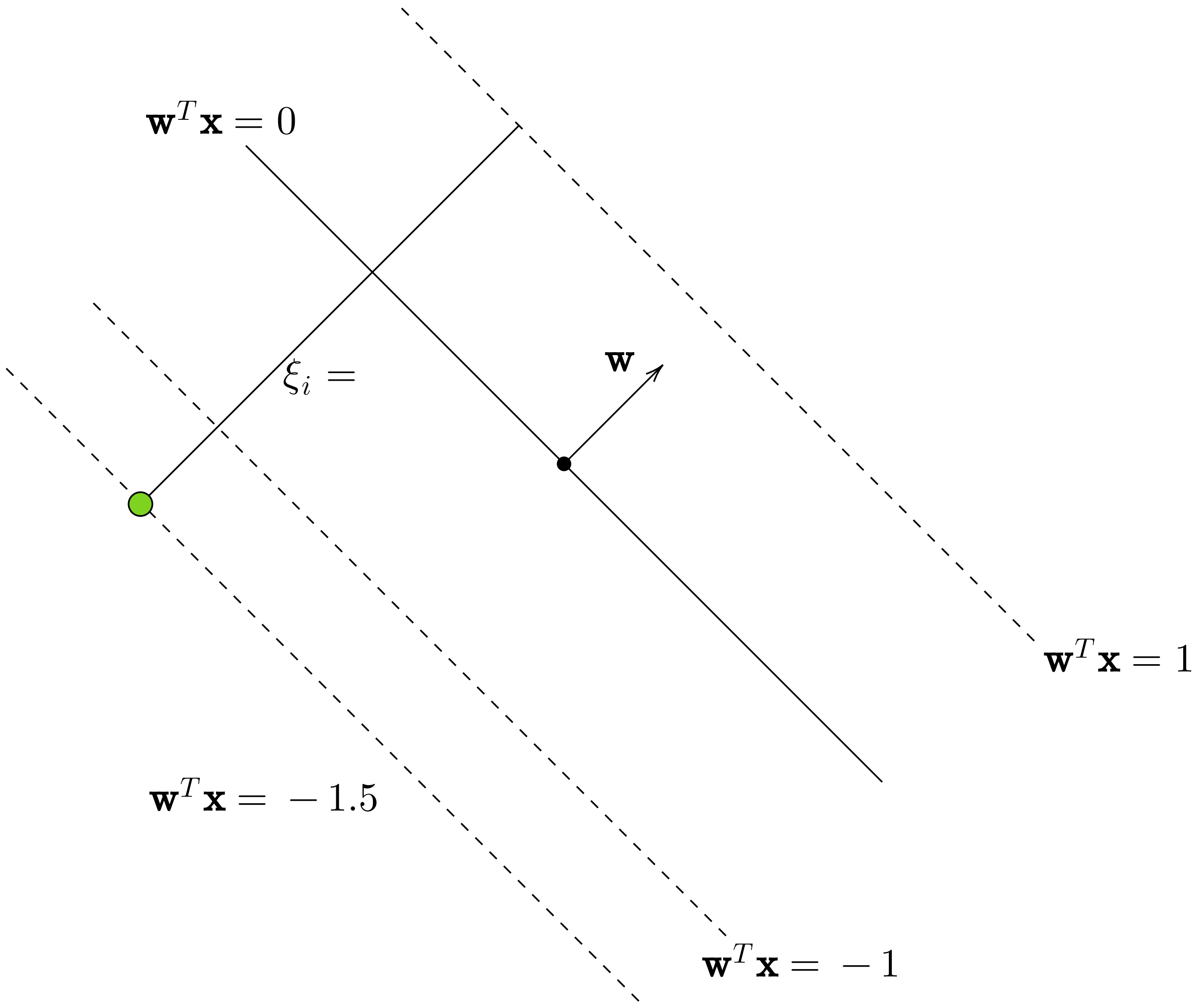
$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

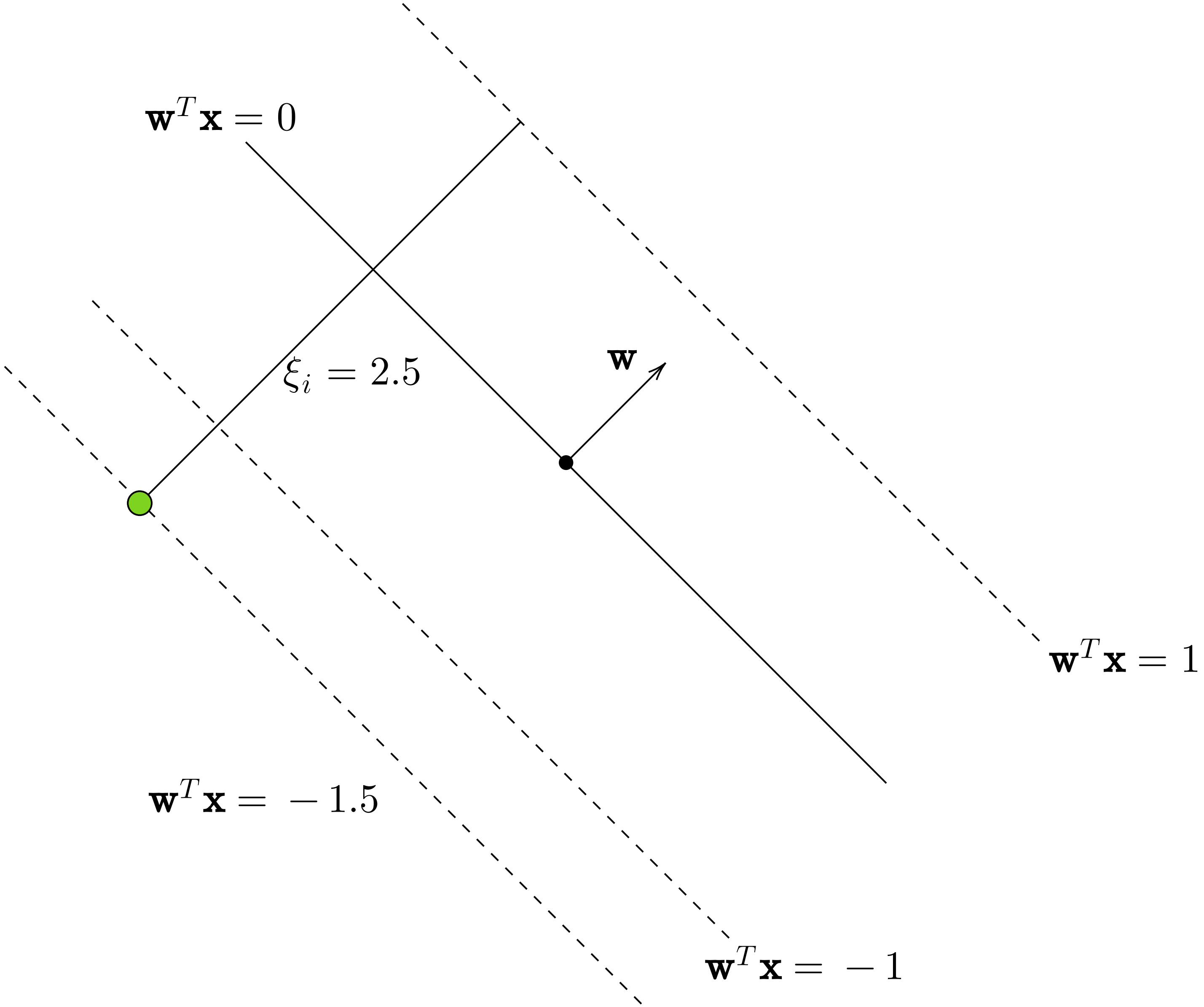
Soft-Margin



$$\left(\mathbf{w}^T\mathbf{x}_i\right)y_i + \xi_i \geqslant 1$$

$$\xi_i \geqslant 0$$

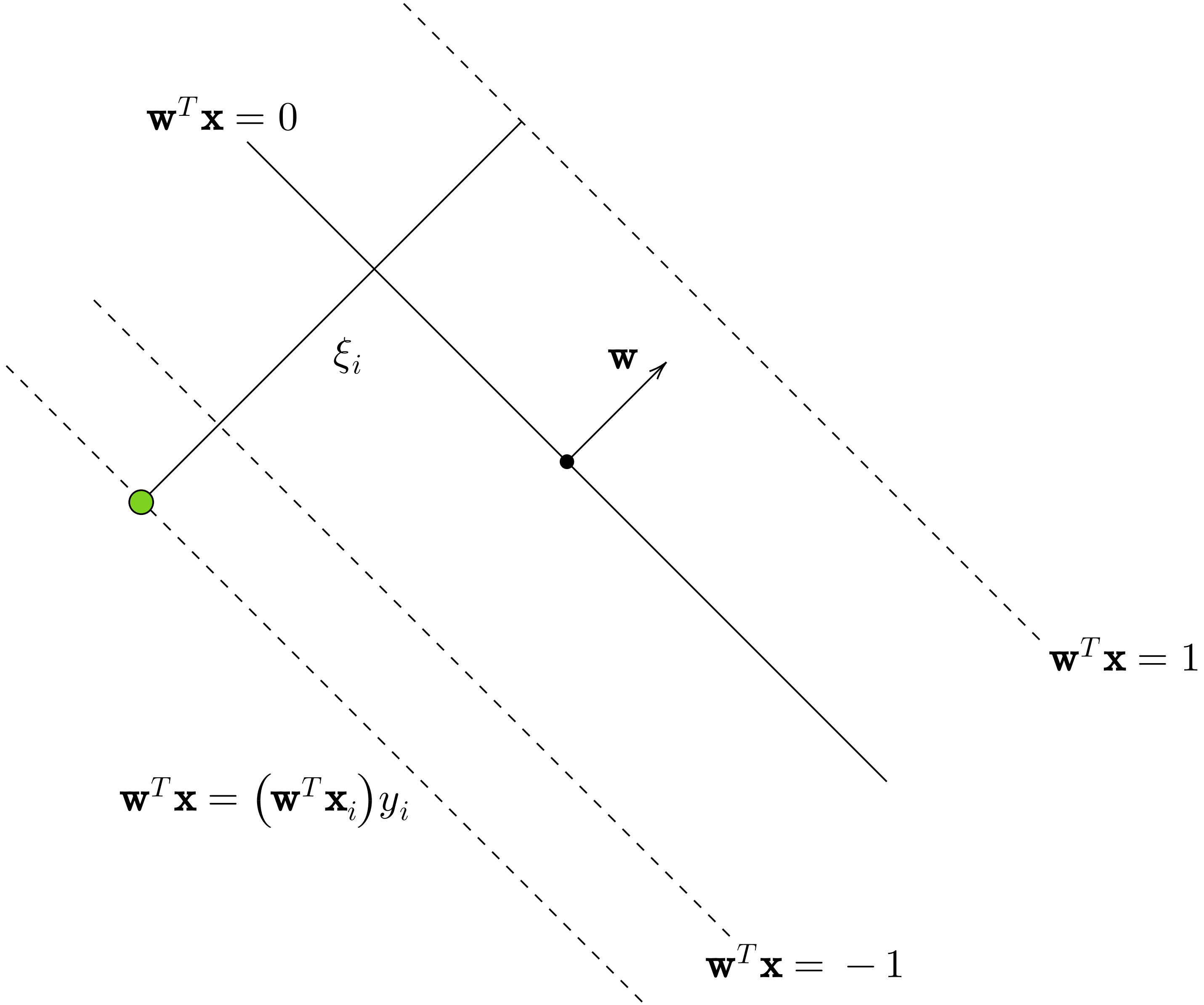
Soft-Margin



$$\big(\mathbf{w}^T\mathbf{x}_i\big)y_i + \xi_i \geqslant 1$$

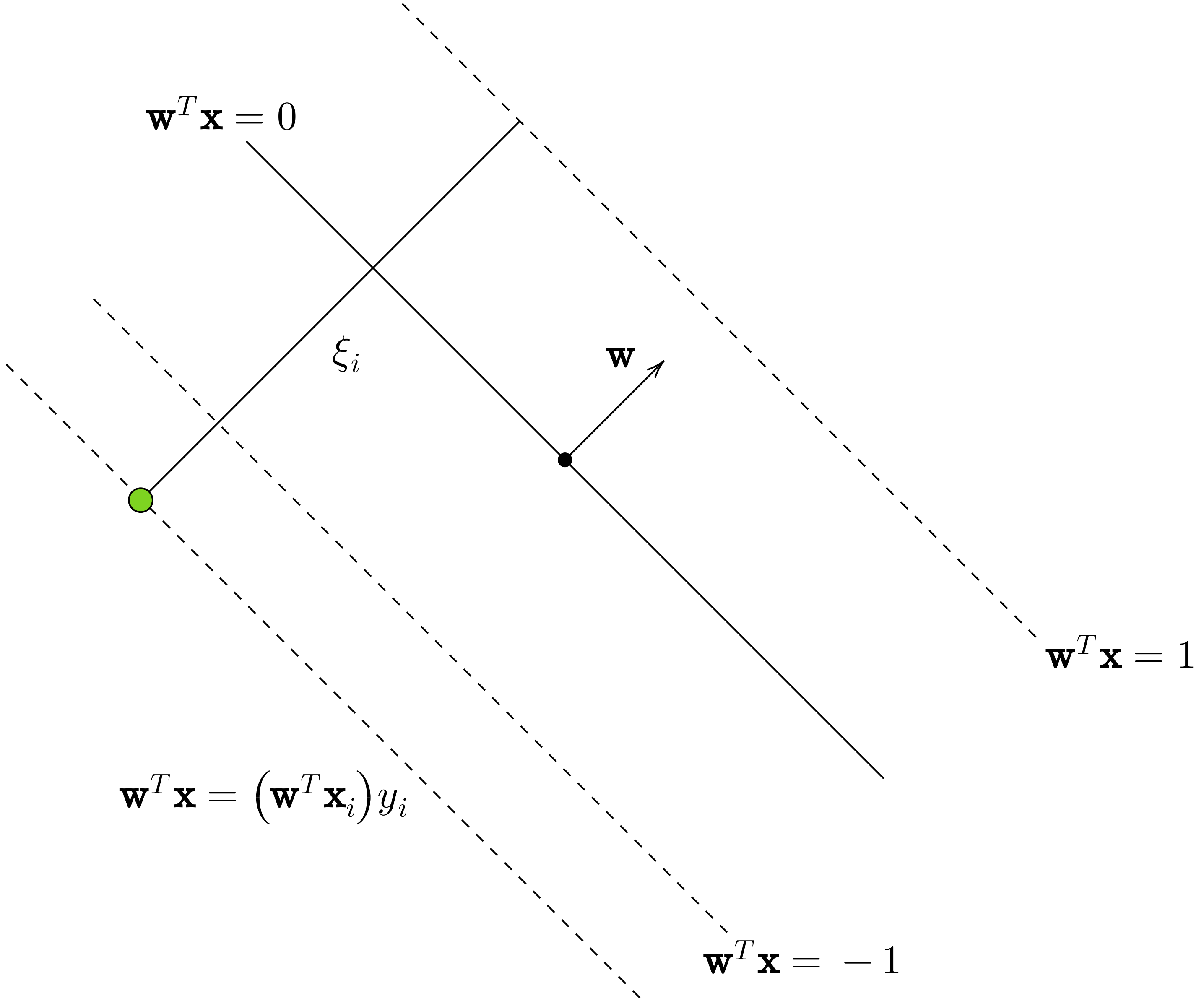
$$\xi_i \geqslant 0$$

Soft-Margin



$$\big(\mathbf{w}^T\mathbf{x}_i\big)y_i + \xi_i \geqslant 1$$
$$\xi_i \geqslant 0$$

Soft-Margin



$$\big(\mathbf{w}^T\mathbf{x}_i\big)y_i + \xi_i \geqslant 1$$

$$\xi_i \geqslant 0$$

$$\xi_i = \max\Big(1 - \big(\mathbf{w}^T\mathbf{x}_i\big)y_i, 0\Big)$$

Soft-Margin, Linear-SVM

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \xi_i$$

sub. to

$$\begin{aligned} \big(\mathbf{w}^T \mathbf{x}_i\big) y_i + \xi_i \geqslant 1, &\quad 1 \leqslant i \leqslant n \\ \xi_i \geqslant 0, &\quad 1 \leqslant i \leqslant n \end{aligned}$$

Soft-Margin, Linear-SVM: Hinge-loss formulation

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \xi_i$$

sub. to

$$\left(\mathbf{w}^T \mathbf{x}_i\right) y_i + \xi_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

$$\xi_i \geqslant 0, \quad 1 \leqslant i \leqslant n$$

$$\equiv$$

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \max\Big(0,\ 1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i\Big)$$

Soft-Margin, Linear-SVM: Hinge-loss formulation

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \xi_i$$

sub. to

$$\left(\mathbf{w}^T \mathbf{x}_i\right) y_i + \xi_i \geqslant 1, \quad 1 \leqslant i \leqslant n$$

$$\xi_i \geqslant 0, \quad 1 \leqslant i \leqslant n$$

$$\equiv$$

Model | Data

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \max\Big(0,\ 1 - \left(\mathbf{w}^T \mathbf{x}_i\right) y_i\Big)$$

Regularization | Hinge Loss

Soft-Margin, SVM: Hinge-loss formulation

$$\min_{\mathbf{w}} \quad \underbrace{\frac{||\mathbf{w}||^2}{2}}_{(1)} + C \cdot \underbrace{\sum_{i=1}^{n} \max\Big(0,\ 1 - \big(\mathbf{w}^T \mathbf{x}_i\big) y_i\Big)}_{(2)}$$

Terms in the objective:

(1) $\frac{||\mathbf{w}||^2}{2}$ controls the width of the margin

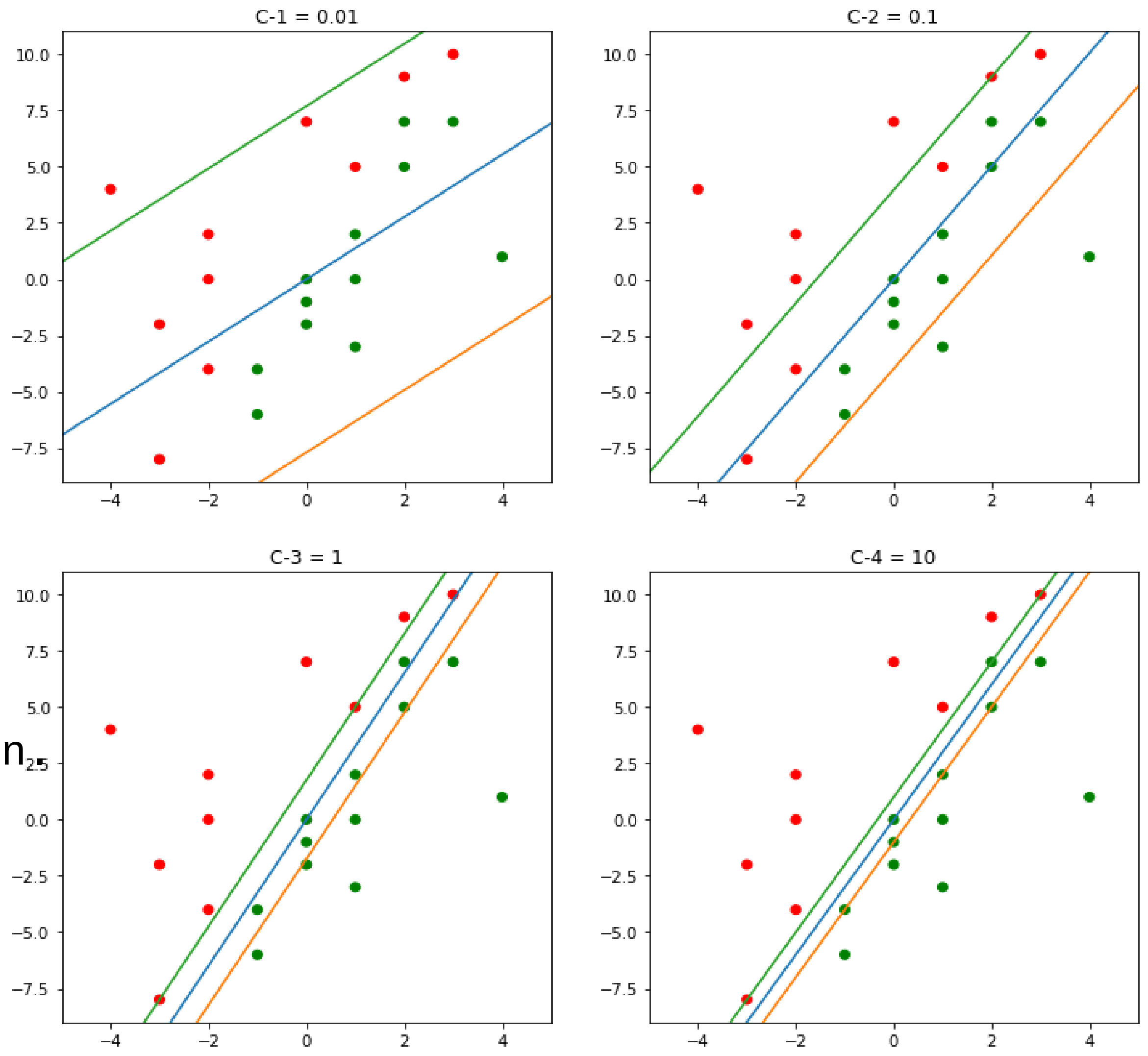
Smaller the value of $||\mathbf{w}||$, wider the margin

(2) $\sum_{i=1}^{n} \max\Big(0,\ 1 - \big(\mathbf{w}^T \mathbf{x}_i\big) y_i\Big)$ is the hinge-loss. Wider the margin, larger the loss.

Soft-Margin, SVM: Hinge-loss formulation

$$\min_{\mathbf{w}} \quad \frac{||\mathbf{w}||^2}{2} + C \cdot \sum_{i=1}^{n} \max\Big(0,\ 1 - \big(\mathbf{w}^T \mathbf{x}_i\big) y_i\Big)$$

(1) (2)



- (1) and (2) work in opposite directions
- If $||\mathbf{w}||$ decreases, the margin becomes wider, which increases the hinge-loss.
- C controls the tradeoff between (1) and (2):
  - If C is small, we are fine with a wide margin.
  - If C is large, we prefer a narrow margin.
  - If $C \to \infty$, we do not tolerate bribery at all.

Miscellaneous

Terminology used

(1) Hard-margin, Linear-SVM
(2) Hard-margin, Kernel-SVM
(3) Soft-margin, Linear-SVM
(4) Soft-margin, Kernel-SVM

Additional points

- Discriminative model
- Weight vector is a sparse linear combination of data-points
- The dual is a quadratic programming problem